Time series mixtures of generalized t experts *

Alexandre Carvalho † Georgios Skoulakis ‡

September 30, 2004

Abstract

We propose and analyze a new class of nonlinear time series models based on local mixtures of linear regressions, referred to as experts, with thick-tailed disturbances. The mean function of each expert is an affine function of covariates that may include lags of the dependent variable and lags of external predictors. The mixing of the individual experts is determined by a latent variable, the distribution of which depends on the same covariates appearing in the expert regressions. The error terms in the experts are modelled to follow a generalized t distribution, which is a rather flexible parametric form encompassing the standard t and normal distributions as special cases and allowing separate modelling of scale and kurtosis. We show consistency and asymptotic normality of the maximum likelihood estimator and establish the stochastic stability of the model in the case of autoregressive experts. Further, we provide Monte Carlo evidence on the performance of standard model selection criteria in selecting the number of experts. As an illustration, we employ the model to describe the conditional dynamics of daily stock returns and evaluate its performance.

1 Introduction

During the last decades, a great deal of research has been dedicated to nonlinear time series analysis, in order to build models capable of capturing several features of commonly encountered data sets, not captured by classical linear models. For example, the usual autoregressive (AR) model of order 1, $y_t = \alpha_0 + \alpha_1 y_{t-1} + \epsilon_t$, assumes a linear relationship

*We wish to thank Joel Horowitz and Ernsrt Schaumburg for their comments and suggestions. Any errors or omissions are our responsibility.

†University of British Columbia, Department of Statistics, e-mail: carvalho@stat.ubc.ca

‡Northwestern University, Department of Finance, e-mail: g-skoulakis@northwestern.edu

between the current observed value y_t and the previous observation y_{t-1}. Besides, the innovations ϵ_t are commonly assumed to have a normal distribution with constant variance σ^2. In practice, however, it is very common to find data sets presenting non-linearities in the conditional mean $E[y_t|y_{t-1}]$ and innovations ϵ_t with non-constant distributions.

Many nonlinear time series model have been proposed in the literatures, and they have shown to be useful in various practical situations. For a good survey see Tong (1990) and Franses and Dijk (2000). Most of the proposed models have focused on modelling the nonlinear relationships in the conditional mean $E[y_t|\mathcal{I}_{t-1}]$, where $\mathcal{I}_{t-1}$ is the information set available at time $t-1$. However, during the most recent years, there has been a great interest in modelling the entire conditional distribution $F(y_t|\mathcal{I}_{t-1})$. In fact, several methods to evaluate density forecasts have been studied, in order to check the performance of the available nonlinear models (see, for example, Diebold, Gunther and Tay, 1998, Berkowitz, 2000, and Hong, 2000 and 2002). For some complex data sets, most of the models construct to flexibly capture nonlinearities in $E[y_t|\mathcal{I}_{t-1}]$ may not take into account certain features of the conditional distribution, such as multiple modes, heteroscedasticity, and heavy tails.

In this paper, we study a class of nonlinear models, based on the mixtures-of-experts structure, as proposed by Jacobs, Jordan, Nowlan and Hinton (1991). The idea is to have a family of models, which is flexible enough to capture not only the nonlinearities in the conditional means $E[y_t|\mathcal{I}_{t-1}]$, but also capture other complexities in the conditional distribution, specially heavy tails. Although, heavy tail time series are commonly encountered in financial data, the proposed model can be applied to time series in any other area, where one wants to have enough flexibility to build accurate density forecasts.

Mixtures-of-experts have been extensively studied during the last decade, and it has been shown to be very appropriate in several applications. In these models, the dependent variable $y_t \in \Re$ is assumed to have the following conditional density specification

$$f(y_t|\mathbf{x}_t, \theta) = \sum_{j=1}^{J} g_j(\mathbf{x}_t; \boldsymbol{\lambda}) \pi(y_t|\mathbf{x}_t; \boldsymbol{\zeta}_j), \tag{1}$$

where $\mathbf{x}_t$ is a s-dimensional vector of covariates, and $\pi(y_t|\mathbf{x}_t; \boldsymbol{\zeta}_j)$ is some know parametric family of distributions. The specification in (1) describes a mixture model, with J components, where the weights $g_j(\mathbf{x}_t; \boldsymbol{\lambda}) \in (0, 1)$ are also functions of the covariates $\mathbf{x}_t$.

Kurnik, Oliver, Waterhouse et al. (1999) apply mixtures-of-experts to signal processing

in a noninvasive glucose monitoring system. Gutta, Huang, Jonathon and Wechsler (2000) describe an application of mixtures-of-experts to gender and ethnic classification of human faces. Lin, McCulloch, Turnbull et al. (2000) present the use of mixtures-of-experts to uncover subpopulation structure for both biomarker trajectories and the probability of disease outcome in highly unbalanced longitudinal data. Jeffries and Pfeiffer (2001) present an application of mixtures-of-experts in modelling hourly measurements of rain rates. The flexibility of mixtures-of-experts has been demonstrated by Jiang and Tanner (1999a and 1999b). They show that the use of mixtures-of-experts is capable of approximating any conditional mean function in a Sobolev space.

In terms of mixtures-of-experts of time series models, the literature is very extensive, especially when treating mixtures of Gaussian processes. An application of this kind of model in finance is presented in Weigend, Mangeas and Srivastava (1995). In their paper, the authors apply mixtures-of-experts, what they call "gated experts", to forecast the density of stock returns. They work with mixtures of Gaussian experts, where the experts are autoregressive processes and the gates are functions of external lagged covariates. Liehr, Pawelzi, Kohlmorgen and Muller (1999) present a unification of the mixtures-of-experts and the hidden Markov models with an input-dependent transition matrix.

Wong and Li (2001) study models, which they call LMARX, where the probabilities of observing each regime, or expert, are functions of the lags of the observed outcome and also of external covariates. Basically, they mix two linear autoregressive Gaussian models, allowing for the inclusion of external covariates in the vector of predictors $\mathbf{x}_t$. After applying the LMARX to two real time series, they compare its prediction performance against other models for nonlinear time series and conclude that mixtures-of-experts may constitute a valuable alternative in modelling time series. Zeevi, Meir and Adler (2000) study a class of models referred to as mixtures of local autoregressive models (MixAR), where the covariate vector $\mathbf{x}_t$ includes only lags of y_t. They present conditions to guarantee stochastic stability and present several applications for their model.

Carvalho and Tanner (2002, 2002a, 2002b, 2003) generalize the models in Wong and Li (2001), Weigend, Mangeas and Srivastava (1995), and Zeevi, Meir and Adler (2000), and study mixtures-of-experts of generalized linear time series. In their models, the densities $\pi(y_t|\mathbf{x}_t;\boldsymbol{\zeta}_j)$ belong to the exponential family, and the covariate vector $\mathbf{x}_t$ may include trans-

formations of lags of the response y_t and lags of external covariates. They give conditions for consistency and asymptotic normality of the maximum likelihood estimator under correctly specified and misspecified models, provide some results for stochastic stability, discuss issues of model specification, and present some applications to real data.

In this paper, we introduce a class of mixtures models where the mixed components $\pi(y_t|\mathbf{x}_t;\boldsymbol{\zeta}_t)$ in (1) are generalized t densities. These mixed densities allow for flexibility in modelling heavy tail time series, encountered in many practical situations. Giacomini, Gottschling, Haefke and White (2002) apply mixture models, with constant weights $g_j(\mathbf{x}_t;\boldsymbol{\lambda}) = \beta_j$, based on the generalized t distributions, denoted by them as mixtures of hypernormal densities. They apply the proposed model to obtain density forecasts of the U.S. inflation.

The remainder of the paper is organized as follows. In Section 2 we introduce the class of time series mixtures of generalized t experts. Section 3 establishes the identifiability of the parameters under appropriate conditions. We address the issue of stochastic stability and existence of moments for the proposed mixture construction in Section 4. Section 5 contains results on the asymptotic behavior of the maximum likelihood estimator which is shown to be consistent and asymptotically normal. In Section 6 we present simulation results on the performance of the AIC and BIC criteria in selecting the number of experts. Section 7 provides an application of the model to two time series of daily stock returns along with an evaluation of the model's ability to adequately describe the conditional dynamics of the data. In the final section we offer some concluding remarks. An overview of the generalized t distribution and all the proofs are contained in the Appendices.

2 Description of the model

The purpose of the class of models we introduce is to provide a flexible description of the conditional dynamics of a univariate nonlinear time series with heavy-tailed innovations. Our formulation builds on the idea of mixtures of distributions and extends it to a time series setting. Denote by $\mathbf{y} = \{y_t\}_{t=0}^{\infty}$ the time series of interest. For each time t, a vector of explanatory variables, termed as covariate vector and denoted by $\mathbf{x}_{t-1}$, is available. The vector $\mathbf{x}_{t-1}$ is assumed to be observed at time $t-1$ and may include lags of the response variable y_t. We further assume that there is a number of local linear models, called experts,

that are mixed to produce the conditional distribution of y_t given the covariate vector $\mathbf{x}_{t-1}$. The role of the covariates is twofold. First, the weights according to which the experts are mixed are particular functions of $\mathbf{x}_{t-1}$. Second, the conditional mean $E_k[y_t|\mathbf{x}_{t-1}]$ for each individual local linear model k, $k = 1, \ldots, J$, is modelled as an affine function of $\mathbf{x}_{t-1}$, $E_k[y_t|\mathbf{x}_{t-1}] = a + \mathbf{b}'\mathbf{x}_{t-1}$.

The conditional density function for y_t given $\mathbf{x}_{t-1}$ is a mixture-of-experts with heavy-tailed noise and has the form

$$f(y_t|\mathbf{x}_{t-1}, \boldsymbol{\theta}) = \sum_{j=1}^{J} g_j(\mathbf{x}_{t-1}; \boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1}; \boldsymbol{\zeta}_j). \tag{2}$$

Here, J is the number of experts being mixed and $j \in \{1, 2, \ldots, J\}$ is the label specifying each expert. The covariate vector $\mathbf{x}_{t-1}$ has dimension S and so the regression coefficients $(a, \mathbf{b}')'$ take values in $\mathbb{R}^{S+1}$. Note that we wrote $\mathbf{x}_{t-1}$ in (2) instead of $\mathbf{x}_t$, as in (1), to emphasize that the covariate vector is observed at time $t-1$. Throughout the paper, we will assume the condition below, which will be important for identifiability of the proposed model and asymptotic normality of the maximum likelihood estimator.

Condition 1 *The covariate vector $\mathbf{x}_{t-1}$ takes values on a open set $\mathbf{X} \subseteq \mathbb{R}^S$ and has a positive density with respect to Lebesgue measure.*

The local weights $g_j(\cdot;\cdot)$, which are referred to as the gating functions, are positive and always sum to one. Our choice of the gating functions follows the original formulation of mixtures-of-experts in the seminal paper by Jacobs et al (1991). Specifically, we consider the standard logistic-type of gating functions

$$g_j(\mathbf{x}_{t-1}; \boldsymbol{\lambda}) = \frac{\exp(\xi_j)}{\sum_{k=1}^{J} \exp(\xi_k)} \tag{3}$$

where $\xi_j = \upsilon_j + \mathbf{u}_j'\mathbf{x}_{t-1} \equiv \boldsymbol{\omega}_j'\mathbf{z}_{t-1}$, $\upsilon_j \in \mathbb{R}$, $\mathbf{u}_j \in \mathbb{R}^S$, $\boldsymbol{\omega}_j = (\upsilon_j, \mathbf{u}_j')'$ for $j = 1, 2, \ldots, J$, $\mathbf{z}_{t-1} = (1, \mathbf{x}_{t-1}')'$. In order to avoid identification problems, it is customary to assume that $\upsilon_J = 0$ an $\mathbf{u}_J = \mathbf{0}_S$. The issue of identifiability is addressed thoroughly in Section 3. The vector $\boldsymbol{\lambda}$ of gating parameters is the union of all the individual gating parameters as follows

$$\boldsymbol{\lambda} = (\upsilon_1, \mathbf{u}_1', \upsilon_2, \mathbf{u}_2', \ldots, \upsilon_{J-1}, \mathbf{u}_{J-1}')' \in \mathbb{R}^{(J-1)(S+1)}. \tag{4}$$

To model the distributions of the experts we employ a three-parameter density specification which is obtained from the generalized t density by setting the peakedness parameter equal to 2. For an overview of the generalized t distribution see Appendix A. In the present context, the density of the experts assumes the functional form

$$\pi(y|\mathbf{x};\boldsymbol{\zeta}) = \frac{1}{\gamma B(1/2,\alpha/2)}\left[1+(y-\mu)^2/\gamma^2\right]^{-\frac{\alpha+1}{2}} \tag{5}$$

where $\boldsymbol{\zeta} = (\boldsymbol{\delta}', \gamma, \alpha)' \in \mathbb{R}^{S+3}$, $\boldsymbol{\delta} = (a, \mathbf{b}')' \in \mathbb{R}^{S+1}$, $\mathbf{z} = (1, \mathbf{x}')'$ and $\mu = \boldsymbol{\delta}'\mathbf{z} \equiv a + \mathbf{b}'\mathbf{x}$. The grand vector of parameters for the whole model is $\boldsymbol{\theta} = (\boldsymbol{\zeta}_1', \ldots, \boldsymbol{\zeta}_J', \boldsymbol{\lambda}')' \in \boldsymbol{\Theta} \subset \mathbb{R}^K$ and the dimension of the parameter space $\boldsymbol{\Theta}$ is $K = J(S+3)+(J-1)(S+1) = (2J-1)(S+1)+2J$.

Alternatively, we can think of the resulting y_t as being generated by a two-stage random process. At the first stage, from each local linear model j, $j = 1, 2, \ldots, J$, values $y_{j,t}$ are generated. Then, at the second stage, we have a latent multinomial random variable $z \in \{1, 2, \ldots, J\}$, the distribution of which is determined by the gating functions g_j, that determines which $y_{j,t}$ will be observed.

The special case in which the covariate vector consists of lagged values of y_t deserves some further discussion. Suppose that $\mathbf{x}_{t-1} = (y_{t-1}, \ldots, y_{t-p})' \equiv \mathbf{y}_{t,p}$ where $p \geq 1$. Then, it follows that the process $\mathbf{y}$ is Markov of order p with transition probability measure given by

$$P(y_t \in B|\mathbf{y}_{t,p} = \mathbf{x};\boldsymbol{\theta}) = \sum_{j=1}^{J} g_j(\mathbf{x};\boldsymbol{\lambda})\Pi(B|\mathbf{x};\boldsymbol{\zeta}_j).$$

where $\Pi(\cdot|\mathbf{x};\boldsymbol{\zeta}_j)$ is the probability measure induced on $\mathbb{R}$ by the conditional density $\pi(\cdot|\mathbf{x};\boldsymbol{\zeta}_j)$ and B is any Borel subset of $\mathbb{R}$. In this case, our model corresponds to a mixture of autoregressive models with thick-tailed noise. In Section 4 we address the issue of stochastic stability for this model. A detailed account of the corresponding model with Gaussian noise is given in Zeevi, Meir and Adler (2001).

3 Identifiability

Given the functional form of the expert densities, identifiability of the parameters is not immediately obvious. However, one can show the identifiability for the mixtures of generalized t experts employing the analysis in Jiang and Tanner (1999c). It turns out that, identifiability follows under the standard assumptions that the mixtures-of-experts are ordered and

initialized. Next, we provide the relevant definitions and precisely state assumptions that guarantee identification of the proposed model.

Definition 1 *(***Initialized mixtures-of-experts***)* *Initialized mixtures-of-experts are those satisfying the restrictions* $\upsilon_J = 0$ *and* $\mathbf{u}_J = \mathbf{0}_S$.

Definition 2 *(***Ordered mixtures-of-experts***)* *Ordered mixtures-of-experts in the case of generalized t experts are those with all the parameters* $\boldsymbol{\zeta}_j = (\boldsymbol{\delta}_j', \gamma_j, \alpha_j)'$, $j = 1, \ldots, J$, *completely ordered, i.e.* $\boldsymbol{\zeta}_1 \prec \boldsymbol{\zeta}_2 \prec \ldots \prec \boldsymbol{\zeta}_J$ *according to some order relation, so there is no invariance caused by the permutation of expert indices.*

Although the lexicographic ordering has been the standard choice for the complete order mentioned above, we will use a slightly less flexible ordering based only on the tail-thickness parameters. Let $\mathbb{G}$ be the family of generalized t densities $\pi(y; \mu, \gamma, \alpha)$ where $\mu \in \mathbb{R}$, $\gamma, \alpha \in (0, \infty)$ such that no two densities have the same value of the parameter α. We assume that the J experts possess densities that are distinct elements of $\mathbb{G}$. For all $j, k \in \{1, \ldots, J\}$, we define $\boldsymbol{\zeta}_j \prec \boldsymbol{\zeta}_k$ if and only if $\alpha_j < \alpha_k$. Clearly then, given our choice of $\mathbb{G}$, we have that $\prec$ is a complete order. While the choice of $\mathbb{G}$ and the use of this ordering are, from a technical point of view, most probably unnecessary, they will simplify our analysis without restricting the applicability of the model. Next we state a condition, assumed to be in effect for the rest of the paper, that is sufficient for identification of the time-series mixture of generalized t experts.

Condition 2 *The generalized t densities of the J experts are distinct elements of the set* $\mathbb{G}$. *In addition, the following two restrictions on the parameter* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *hold: (a)* $\alpha_1 < \cdots < \alpha_J$ *and (b)* $\upsilon_J = 0$ *and* $\mathbf{u}_J = \mathbf{0}_S$.

Condition 2 implies that the mixture of experts we consider is both ordered and initialized. The last component needed in dealing with the identifiability issue is a nondegeneracy condition for the expert conditional densities, which is guaranteed by the following lemma. The proof is given in subsection B.1 of Appendix B.

Lemma 3.1 *Consider a subset* $\{\pi(\cdot; \mu_j, \gamma_j, \alpha_j) : j = 1, \ldots, M\}$ *of the set* $\mathbb{G}$ *consisting of M distinct generalized t density functions, where M is any positive integer. Then the functions* $\pi(y; \mu_j, \gamma_j, \alpha_j)$, $j = 1, \ldots, M$, *are linearly independent functions of y.*

Given the discussion above, we are now able to state the theorem that fully addresses the issue of identifiability. The proof of the theorem follows by repeating the arguments in the proof of Theorem 1 in Jiang and Tanner (1999c). Although their theorem is developed in the context of experts in an exponential family, careful examination reveals that the same proof goes through in our case as well.

Theorem 3.1 *Under Condition 1 and the nondegeneracy condition on the expert densities provided by Lemma 3.1, the equality of the density functions for two ordered and initialized mixtures-of-experts is equivalent to the equality of the two sets of ordered and initialized parameters.*

According to Theorem 3.1, Conditions 1 and 2 guarantee the identifiability of time-series mixture of generalized t experts. As will be discussed in Section 5, identifiability plays a key role for consistency and asymptotic normality of the maximum likelihood estimator.

4 Stochastic stability

In this section, we discuss the probabilistic properties of the mixtures-of-experts of generalized t distributions introduced in Section 2, in the case where the vector of covariates $\mathbf{x}_{t-1}$ consists of lags up to order p of the response variable y_t. In other words, we consider a class of mixtures of autoregressive models with heavy-tailed innovations. The results are based on Carvalho and Skoulakis (2003) and follow the methodology developed in Meyn and Tweedie (1993), which is the standard reference on stochastic stability of Markov chains on general state spaces.

The description of the class of mixtures of AR models we are concerned with is given by

$$y_t = \sum_{j=1}^{J} I_{j,t} \left[\boldsymbol{\delta}_j' \mathbf{Y}_{t-1} + W_{j,t} \right] \tag{6}$$

where $\boldsymbol{\delta}_j = (\delta_{j,1}, \ldots, \delta_{j,p})'$, $\mathbf{Y}_{t-1} = (y_{t-1}, \ldots, y_{t-p})'$ and p is the lag of the autoregressive experts. The innovation terms $W_{j,t}$ are independent random variables with identical generalized t density functions $f_{\mathrm{GT}}(w; \mu_j, \gamma_j, 2, \alpha_j)$, across t, for each $j = 1, \ldots, J$ (see Appendix A). Note that the above representation does not include intercepts and thus the innovation means μ_j are not necessarily zero. The indicator variables $I_{j,t}$ determine the choice of expert

at time t. The jth expert is chosen at time t, and so we have $I_{j,t} = 1$ and $I_{k,t} = 0$ for $k \neq j$, with conditional probability $g_j(\mathbf{Y}_{t-1}; \boldsymbol{\lambda})$ as presented in (3). To simplify the notation, we will write $g_j(\mathbf{x}_{t-1})$ instead of $g_j(\mathbf{x}_{t-1}; \boldsymbol{\lambda})$. For each time t, the terms $W_{j,t}$ are independent of the indicator vector $\mathbf{I}_t = (I_{1,t}, \ldots, I_{J,t})$. We denote by $\pi_j(\cdot|\mathbf{Y}_{t-1})$ the conditional density of $y_{j,t} = \boldsymbol{\delta}_j' \mathbf{Y}_{t-1} + W_{j,t}$ given $\mathbf{Y}_{t-1}$.

Initially, we consider the one-lag case where $p = 1$, i.e. $\mathbf{x}_{t-1} = y_{t-1}$ and $\boldsymbol{\delta}_j = \delta_j \in \mathbb{R}$, $j = 1, \ldots, J$, so that the time series $\mathbf{y} = \{y_t\}_{t=0}^{\infty}$ is a first order Markov process, for which the transition measure is expressed as $P(x, dy) = \sum_{j=1}^{J} g_j(x) P_j(x, dy)$, where $P_j(x, dy) = \pi_j(y|x)dy$ is the transition measure for the jth expert. The following theorem provides conditions under which the time series described by (6), possesses a unique invariant measure under the standard assumption that guarantees stationarity for each autoregressive expert.

Theorem 4.1 *Consider a time series* $\mathbf{y} = \{y_t\}_{t=0}^{\infty}$ *admitting the representation (6), with* $p = 1$, *and assume that the following conditions are satisfied: (a)* $\overline{\delta} = \max_{j=1,\ldots,J} |\delta_j| < 1$, *and (b)* $\underline{\alpha} \equiv \min_{j=1,\ldots,J} \alpha_j > \xi > 1$. *Then* $\mathbf{y}$ *is positive Harris recurrent and* V*-uniformly ergodic where* $V(x) = 1 + |x|^{\xi}$. *Specifically, there exist* $r > 1$ *and* $R < \infty$ *such that*

$$|||P^n - P^{\infty}|||_V \leq Rr^{-n}, \tag{7}$$

where P^{∞} *is the unique invariant probability measure. Besides,* P^{∞} *satisfies* $\int_{\mathbb{R}} |x|^{\xi} P^{\infty}(dx) < \infty$, *so that* $\{y_t\}_{t=0}^{\infty}$ *has finite* ξ*th moment.*

The assumption that $\underline{\alpha} \equiv \min_{j=1,\ldots,J} \alpha_j > \xi > 1$ implies the existence of the ξ moment for all innovations $W_{j,t}$, so that Theorem 4.1 is an immediate consequence of Theorem 3.1 in Carvalho and Skoulakis (2003).

When the covariate vector consists of lags of up to order p greater than 1, the time series with representation (6) is not a first order Markov process. However, we can use the standard trick and augment y_t with its lagged values up to order $p - 1$ so that we obtain a vector process that is Markov of order 1. Specifically, we define

$$\mathbf{Y}_t = \begin{bmatrix} y_t \\ y_{t-1} \\ y_{t-2} \\ \vdots \\ y_{t-p+1} \end{bmatrix}, \quad \mathbf{F}_j = \begin{bmatrix} \delta_{j,1} & \delta_{j,2} & \cdots & \delta_{j,p-1} & \delta_{j,p} \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}, \quad \mathbf{W}_{j,t} = \begin{bmatrix} W_{j,t} \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

Then we can rewrite expression (6) in the form

$$\mathbf{Y}_t = \sum_{j=1}^{J} I_{j,t} \left[\mathbf{F}_j \mathbf{Y}_{t-1} + \mathbf{W}_{j,t} \right] \tag{8}$$

and therefore $\mathbf{Y} = \{\mathbf{Y}_t\}_{t=0}^{\infty}$ is Markov of order 1. Let Q denote the transition measure of the process $\{\mathbf{Y}_t\}_{t=0}^{\infty}$. The following theorem gives conditions for geometric ergodicity and existence of moments. It is an application of Theorem 3.2 in Carvalho and Skoulakis (2003).

Theorem 4.2 *Consider a time series* $\mathbf{y} = \{y_t\}_{t=0}^{\infty}$ *admitting the representation (6), with* $p > 1$*, and assume that the following conditions are satisfied: (a) the roots of the equation*

$$\zeta^p - \sum_{k=1}^{p} \delta_{j,k} \zeta^{p-k} = 0, \tag{9}$$

are contained strictly inside the unit circle on the complex plane for all $j = 1, 2, \ldots, J$*, and (b)* $\underline{\alpha} \equiv \min_{j=1,\ldots,J} \alpha_j > \xi > 1$*. Then the Markov process* $\mathbf{Y} = \{\mathbf{Y}_t\}_{t=0}^{\infty}$ *is positive Harris recurrent and geometrically ergodic, namely there exists* $r > 1$ *such that*

$$\sum_{n=1}^{\infty} r^n ||Q^n(\mathbf{x}, \cdot) - Q^{\infty}|| < \infty, \text{ for all } \mathbf{x} \in \mathbb{R}^p \tag{10}$$

where Q^{∞} *is the unique invariant probability measure. Besides,* Q^{∞} *satisfies* $\int_{\mathbb{R}^p} ||\mathbf{x}||^{\xi} Q^{\infty}(d\mathbf{x}) < \infty$*, so that* $\{y_t\}_{t=0}^{\infty}$ *has finite* ξ*th moment.*

In the next section, we provide conditions that guarantee the consistency and the asymptotic normality of the maximum likelihood estimator. Among other requirements, the conditions involve stationarity and applicability of the law of large numbers to functions of the process as well as the existence of some of the moments for the vector of covariates $\mathbf{x}_{t-1}$. For consistency and asymptotic normality, we will require the finiteness of $E[||\mathbf{x}_{t-1}||]$ and $E[||\mathbf{x}_{t-1}||^3]$ respectively. When $\mathbf{x}_{t-1}$ includes only lags of the response variable y_t, it suffices to ensure the stationarity of $\mathbf{y}$ and that $E[|y_t|^k] < \infty$, where $k > 3$. In this case, according to Theorems 4.1 and 4.2, if every mixed generalized t expert has tail parameter greater than 3, namely possesses finite absolute kth moments, with $k > 3$, the sufficient conditions for the consistency and asymptotic normality of the MLE are satisfied. For more details on the concepts used in this section, see Meyn and Tweedie (1993).

5 Maximum Likelihood Estimation

An estimate for the parameter vector $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ can be obtained by maximizing the partial loglikelihood function (Wong, 1986) given by

$$\log L_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}) = \sum_{t=1}^{T} \log\Big(\sum_{j=1}^{J} g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\Big). \tag{11}$$

In this case, the maximum likelihood estimator (MLE) based on a sample of size T, denoted by $\widehat{\boldsymbol{\theta}}_T$, is defined as

$$\widehat{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}). \tag{12}$$

In this section, we initially discuss the EM algorithm used in the optimization problem related to the estimation of mixtures of generalized t experts. Additionally, we present some technical results that guarantee consistency and asymptotic normality for the MLE.

5.1 The EM Algorithm

For simple problems, where the parameter space is low dimensional, maximization of loglikelihood function can be performed directly by using some standard optimization algorithm, such as Newton-Raphson. However, in most of the practical problems, the dimension of $\boldsymbol{\Theta}$ is high enough so that the usual optimization methods become very unstable. The alternative, commonly used in mixture of distribution models, is the EM algorithm, proposed by Dempster, Laird and Rubin (1977). The use of the EM algorithm for mixture of experts models is thoroughly described in Jordan and Jacobs (1994) and Huerta, Jiang and Tanner (2001), and that is the procedure used here for estimation.

To initialize the EM algorithm, choose an starting value $\boldsymbol{\theta}^0$ for the parameter vector $\boldsymbol{\theta} = (\boldsymbol{\zeta}_1', \ldots, \boldsymbol{\zeta}_J', \boldsymbol{\lambda}')'$. Then, obtain the sequence $\{\boldsymbol{\theta}^i\}$ iterating between the E-step (expectation step) and the M-step (maximization step). For $i = 0, 1, 2, \ldots$,

E-step: Construct

$$Q^i(\boldsymbol{\theta}) = \sum_{t=1}^{T}\sum_{j=1}^{J} h_{j,t}(\boldsymbol{\theta}^i)\log\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j) + \sum_{t=1}^{T}\sum_{j=1}^{J} h_{j,t}(\boldsymbol{\theta}^i)\log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}), \tag{13}$$

where

$$h_{j,t}(\boldsymbol{\theta}) = \frac{g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)}{\sum_{l=1}^{J}\Big[g_l(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_l)\Big]}. \tag{14}$$

M-step: Find $\boldsymbol{\theta}^{i+1} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} Q^i(\boldsymbol{\theta})$.

Note that, at each iteration i, maximization of $Q^i(\boldsymbol{\theta})$ in (13) can be obtained by maximizing separately the J terms Q_j^i, corresponding to parameters for each generalized t expert individually,

$$Q_j^i(\boldsymbol{\theta}_j) = \sum_{t=1}^{T} h_{j,t}(\boldsymbol{\theta}^i)\log\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j) \tag{15}$$

and the term $Q_{\mathrm{g}ates}^i$, corresponding to the parameter vector $\boldsymbol{\lambda}$ for the gating functions,

$$Q_{\mathrm{g}ates}^i(\boldsymbol{\lambda}) = \sum_{t=1}^{T}\sum_{j=1}^{J} h_{j,t}(\boldsymbol{\theta}^i)\log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}), \tag{16}$$

Therefore, the EM algorithm in our case consists of calculating, at each iteration i, the weights $h_{j,t} \in (0,1)$, $j = 1,\ldots,J$, $t = 1,\ldots,T$, and then maximizing the functions $Q_1^i(\boldsymbol{\theta}_1)$, ..., $Q_J^i(\boldsymbol{\theta}_J)$, $Q_{\mathrm{g}ates}^i(\boldsymbol{\lambda})$, to find the new value $\boldsymbol{\theta}^{i+1}$. The limit of the sequence $\{\boldsymbol{\theta}^i\}$, denoted by $\widehat{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)$, is a root of the first-order condition $\partial_{\boldsymbol{\theta}}\log L_T(\boldsymbol{\theta}) = 0$ (see Tanner, 1996). When the loglikelihood function is multimodal, the limits $\widehat{\boldsymbol{\theta}}(\boldsymbol{\theta}^0)$ may not correspond to the global maximum of the loglikelihood function, so we used multiple starting points to initialize the algorithm. The maximum likelihood estimator $\widehat{\boldsymbol{\theta}}$ is the root corresponding to the largest likelihood value $L_T(\widehat{\boldsymbol{\theta}}(\boldsymbol{\theta}^0))$.

5.2 Consistency and asymptotic normality of the MLE

In this section, we show that, under correct model specification, the maximum likelihood estimator is consistent and asymptotically normal for the true parameter vector $\boldsymbol{\theta}_0$. We will assume that the parameter space Θ is compact and some the moments of the covariate vector $\mathbf{x}_{t-1}$ exist. Although the hypothesis of compactness of Θ may seem to be restrictive, it is convenient for theoretical reasons and is still general enough in terms of applications. Besides, it guarantees the existence of a global maximum of likelihood function in the mixture structure (see discussion in Hamilton, 1994, Chapter 22). Consider initially the following conditions.

Condition 3 *The time series* $\{(y_t, \mathbf{x}'_{t-1})'\}_{t=1}^{\infty}$ *is stationary and ergodic.*

Condition 4 *The first moments of the time series* $\mathbf{x}_{t-1}$ *are finite, i.e.* $E\left[\|\mathbf{x}_{t-1}\|\right] < \infty$.

Theorem 5.1 *(***Consistency of the MLE***) For the mixtures of generalized t experts defined above assume that*
(a) the model is correctly specified with true parameter vector $\boldsymbol{\theta}_0$*, and identifiability holds,*
(b) the parameter space $\boldsymbol{\Theta}$ *is a compact and convex subset of* $\mathbb{R}^K$ *and* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$*,*
(c) Conditions 3 and 4 hold, and
(d) $\min_{j=1,\ldots,J} \alpha_{0,j} \equiv \underline{\alpha_0} > 1$.
Then the MLE $\widehat{\boldsymbol{\theta}}_T$ *is a consistent estimator of* $\boldsymbol{\theta}_0$*, that is*

$$\widehat{\boldsymbol{\theta}}_T \xrightarrow{P} \boldsymbol{\theta}_0 \quad as \quad T \to \infty. \tag{17}$$

Condition (d) in Theorem 5.1 is equivalent to the existence of the first moment of the heavy-tailed innovations for all experts $j = 1, \ldots, J$ (see Appendix A). We now establish the asymptotic normality of the maximum likelihood estimator. This requires two additional conditions that we state next.

Condition 5 *The third moments of the time series* $\mathbf{x}_t$ *are finite, i.e.* $E[\|\mathbf{x}_t\|^3] < \infty$.

We are in a position to state the result for the asymptotic normality of MLE.

Theorem 5.2 *(***Asymptotic normality of the MLE***) For the mixtures of generalized t experts defined above assume that*
(a) the model is correctly specified with true parameter vector $\boldsymbol{\theta}_0$*, and identifiability holds,*
(b) the parameter space $\boldsymbol{\Theta}$ *is a compact and convex subset of* $\mathbb{R}^K$ *and* $\boldsymbol{\theta}_0 \in \text{interior}(\boldsymbol{\Theta})$*,*
(c) Conditions 1, 3 and 5 hold, and
(d) $\min_{j=1,\ldots,J} \alpha_{0,j} \equiv \underline{\alpha_0} > 3$.
Then the maximum likelihood estimator $\widehat{\boldsymbol{\theta}}_T$ *has an asymptotic normal distribution, given by*

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}_K, \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\right) \quad as \quad T \to \infty \tag{18}$$

where $\mathbf{I}(\boldsymbol{\theta}_0) \equiv -E[\partial_{\boldsymbol{\theta}} \partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0)]$ *is the Fisher information matrix.*

6 Empirical performance of standard model selection criteria

Although extremely important practical issue, the selection of the number of experts does not admit an easy theoretical treatment. Loglikelihood ratio tests are not applicable in this case, for under the null hypothesis of fewer experts, the alternative hypothesis implies a non-identified problem, as pointed out by Quinn, McLachlan and Hjort (1987). In the absence of theoretical results, we resort to numerical simulations to obtain evidence on the effectiveness of widely used model selection criteria. In this section, we present the results of Monte Carlo simulations to evaluate the performance of the Bayesian Information Criterion (BIC) (Schwarz, 1978) and the Akaike Information Criterion (AIC) (Akaike, 1973 and 1981) in selecting the number of mixed experts. The expressions for the two criteria are

$$BIC = -2\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1}; \widehat{\boldsymbol{\theta}}_T) + K \log T, \quad AIC = -2\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1}; \widehat{\boldsymbol{\theta}}_T) + 2K$$

where K is the dimension of the parameter vector $\boldsymbol{\theta}$ and T is the sample size. Wood, Jiang and Tanner (2002) and Carvalho and Tanner (2002b) use the BIC to select the number of experts for spatially adaptive nonparametric regression models. For well-behaved models, it is known that the BIC is consistent for model selection, since, with probability approaching one as the sample size increases, the smallest BIC value will correspond to the true model. However, when the model is overidentified, the standard regularity conditions required to support this result are not satisfied. Fortunately, Wood, Jiang and Tanner (2002) present some evidence that, even in the case of overidentified models, the BIC may still be consistent for model selection.

We perform simulations under true models with one, two and three experts and, for each generated data set, we estimate mixture models with various numbers of mixed generalized t experts. Beforehand, we expect that at least one of the two criteria will present the smallest value for the estimated model with the same number of experts as the simulated true model. The sample sizes selected for the simulation experiment are $T = 500$ and $T = 1,000$ and for each choice 1,000 data sets are generated. Each model includes an external covariate x_{t-1}, which was generated as an AR(1) process, with autoregressive coefficient equal to 0.5, unconditional mean equal to 3.0 and noise variance equal to 5.0. The expressions for the

experts and the gating functions, assuming one, two and three experts in the true model, are presented next. The true model with one expert is

$$y_t = 3.0 + 0.4y_{t-1} + 1.2x_{t-1} + \epsilon_t,$$

where ϵ_t has a generalized t distribution, with scale parameter $\gamma = 2.0$ and tail parameter $\alpha = 8.0$. The true model with two experts is

$$\begin{aligned} y_{1t} &= 1.0 - 0.7y_{t-1} + 1.2x_{t-1} + \epsilon_{1t}, \\ y_{2t} &= -0.8 + 0.8y_{t-1} - 1.1x_{t-1} + \epsilon_{2t}, \end{aligned}$$

with the gating function determined by

$$\xi_{1,t} = 0.8 - 0.7y_{t-1} + 0.8x_{t-1},$$

where $\epsilon_{1t}, \epsilon_{2t}$ have generalized t distributions, with scale parameters $\gamma_1 = 2.0, \gamma_2 = 1.5$, and tail parameters $\alpha_1 = 8.0, \alpha_2 = 6.0$, respectively. The true model with three experts is

$$\begin{aligned} y_{1t} &= 0.5 + 0.4y_{t-1} + 0.8x_{t-1} + \epsilon_{1t}, \\ y_{2t} &= -0.8 + 0.7y_{t-1} - 0.7x_{t-1} + \epsilon_{2t}, \\ y_{3t} &= 1.0 - 0.6y_{t-1} + 0.5x_{t-1} + \epsilon_{3t}, \end{aligned}$$

with the gating function determined by

$$\begin{aligned} \xi_{1t} &= -0.8 - 0.8y_{t-1} + 0.3x_{t-1}, \\ \xi_{2t} &= 0.7 + 0.9y_{t-1} - 0.8x_{t-1}, \end{aligned}$$

where $\epsilon_{1t}, \epsilon_{2t}, \epsilon_{3t}$ have generalized t distributions, with scale parameters $\gamma_1 = 2.0, \gamma_2 = 1.5, \gamma_3 = 1.0$, and tail parameters $\alpha_1 = 4.0, \alpha_2 = 6.0, \alpha_3 = 9.0$, respectively. The results for all the Monte Carlo simulations experiments are summarized in Table 1.

Simple observation of the table suggests that the BIC is a rather accurate criterion in selecting the number of experts. On the other hand, the AIC tends to lead to overparameterization. This evidence is consistent with the performance of the BIC and the AIC in a number of different settings (see, for example, Carvalho and Tanner, 2002b). Even for the small sample size considered, namely $T = 500$, the BIC selects the correct number of experts at least 98% of the time. Related evidence on the superior performance of the BIC is presented by Wong and Li (2001) where they use it to select the number of lags in a mixture of time series model consisting of two Gaussian experts. Overall, the simulation results suggest that the BIC offers a promising tool in tackling the difficult problem of model selection in the context of time series of mixtures-of-experts.

Table 1: Simulation results on the performance of AIC and BIC in selecting the number of experts. The simulations are repeated 1,000 times and the various numbers of experts are selected by the two criteria with the frequencies reported in the table.

True model with 1 expert			**True model with 2 experts**			**True model with 3 experts**		
Selected number of experts	T = 500		Selected number of experts	T = 500		Selected number of experts	T = 500	
	BIC	AIC		BIC	AIC		BIC	AIC
1	987	235	1	0	0	1	0	0
2	13	765	2	993	167	2	0	0
			3	7	833	3	980	61
						4	20	939
Total	1,000	1,000	Total	1,000	1,000	Total	1,000	1,000
True model with 1 expert			**True model with 2 experts**			**True model with 3 experts**		
Selected number of experts	T = 1,000		Selected number of experts	T = 1,000		Selected number of experts	T = 1,000	
	BIC	AIC		BIC	AIC		BIC	AIC
1	994	246	1	0	0	1	0	0
2	6	754	2	997	181	2	0	0
			3	3	819	3	986	107
						4	14	893
Total	1,000	1,000	Total	1,000	1,000	Total	1,000	1,000

7 Application to modelling conditional densities of stock returns

In this section, we present an application of mixtures of generalized t experts to describe the behavior of the logarithm of daily returns for two individual stocks. The individual stocks considered were Hewlett-Packard and Intel, henceforth referred to by HP and INTEL respectively. We used 2,100 daily observations, from September 19, 1990, to January 8, 1999. As an external covariate, we used a proxy for the stock trading activity. This proxy variable was constructed based on the volume traded of each stock. The logarithm of the returns and the logarithm of the traded volume are plotted in Figures 1 and 2.

It is apparent from Figures 1 and 2 that there is positive trend in the time series of log volume for both stocks, suggesting the need for some detrending procedure. Since the positive trend is clearly nonlinear, we estimated a random walk without drift model (see

Harvey, 1994). More specifically, we estimate the following model

$$\begin{aligned} \log v_t &= \mu_t + x_t, \\ \mu_t &= \mu_{t-1} + \eta_t, \end{aligned} \tag{19}$$

where v_t is the volume traded on day t, μ_t provides an estimate for the nonlinear trend and we use $x_t = \log v_t - \mu_t$ as an external covariate to express the market activity. If $x_t > 0$, the market is more active than on average. In the model given in (19), one important variable is the variance ratio $q = \sigma_x^2/\sigma_\eta^2$, where σ_x^2 and σ_η^2 are the variance of x_t and η_t respectively. This ratio determines the smoothness of μ_t. In our case, q was chosen to allow some positive autocorrelation in x_t, so that we have some persistence in the market activity indicator. The smoothness of μ_t is noticed at both Figures 1 and 2.

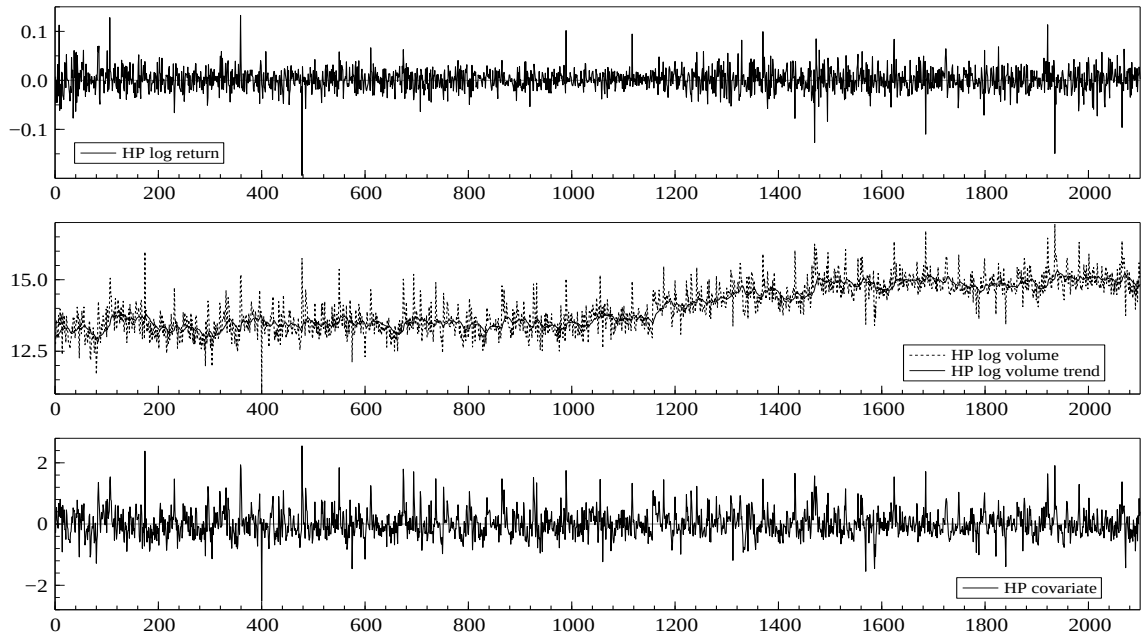

Figure 1: Log return, log volume, log volume trend and covariate time series - HP stock.

To avoid numerical complications, we standardized both the response variable y_t and the predictor x_t. To select the number of experts and the number of lags of the response y_t and the covariate x_t used in each expert and each gating function, we used only the first 1,500 observations, in order to be consistent with the density forecast evaluation described in Subsection 7.1. We initially tried different number of experts for a maximum number of 4 lags

for both the response y_t and the covariate x_t. For both stocks, according to the BIC criterion, we selected two experts. At a second stage, using two experts, we sequentially deleted the number of lags until the resulting model had only maximum lags with corresponding t-statistics greater than 1.0. Although the value 1.0 for the threshold in the model selection procedure may be low, we believed the resulting models were parsimonious enough and still able to capture the dynamics in the data.

The final model for HP, estimated using all the 2,100 data points and keeping the same model specification described above, is a mixture of two generalized t distributions with fixed means $\mu_1 = 0.12890$, and $\mu_2 = -0.0060508$. The scale parameters are $\gamma_1 = 1.5606$ and $\gamma_2 = 1.5227$, and the tail parameters are $\alpha_1 = 4.1057$ and $\alpha_2 = 6.2231$. The resulting expression for the gating function is given by

$$\begin{aligned}
&\xi_{1t} = -5.0057 - 6.7222 y_{t-1} - 4.3548 y_{t-2} - 0.33468 y_{t-3} + 3.5587 x_{t-1}, \\
&g_{1t} = \tfrac{\exp(\xi_{1t})}{1+\exp(\xi_{1t})}, \quad g_{2t} = 1 - g_{1t}.
\end{aligned}$$

For INTEL stock returns, the fitted model is very similar to the one obtained for HP. The final model, also estimated using all 2,100 observations, is a mixture of two generalized t distributions with fixed means $\mu_1 = 0.016185$ and $\mu_2 = 0.027721$, scale coefficients $\gamma_1 = 2.7446$ and $\gamma_2 = 25.405$, and tail parameters $\alpha_1 = 6.8270$ and $\alpha_2 = 1503.3$. The expression for the gating function is given by

$$\begin{aligned}
&\xi_{1t} = -1.5068 - 0.48935 y_{t-1} - 0.48667 y_{t-2} - 0.29949 y_{t-3} + 0.63840 x_{t-1}, \\
&g_{1t} = \tfrac{\exp(\xi_{1t})}{1+\exp(\xi_{1t})}, \; g_{2t} = 1 - g_{1t}.
\end{aligned}$$

For both stocks, the first experts have lower tail parameters α, and therefore thicker tails, than the second ones. Looking at the estimated parameters for the gating functions, we can see that negative returns and high volume traded seem to increase the weight for the first expert and hence antecipate outcomes with heavier tails. Note that both experts for HP stocks have heavy tails, with $\alpha_1, \alpha_2 < 7.0$. For INTEL stocks, the second expert presents a very high tail parameter, which indicates a distribution close the normal case.

7.1 One-step-ahead density forecast evaluation

We now address the conditional density prediction performance for the mixtures of generalized t experts fitted to the two stock return data sets. The approach used here is based on

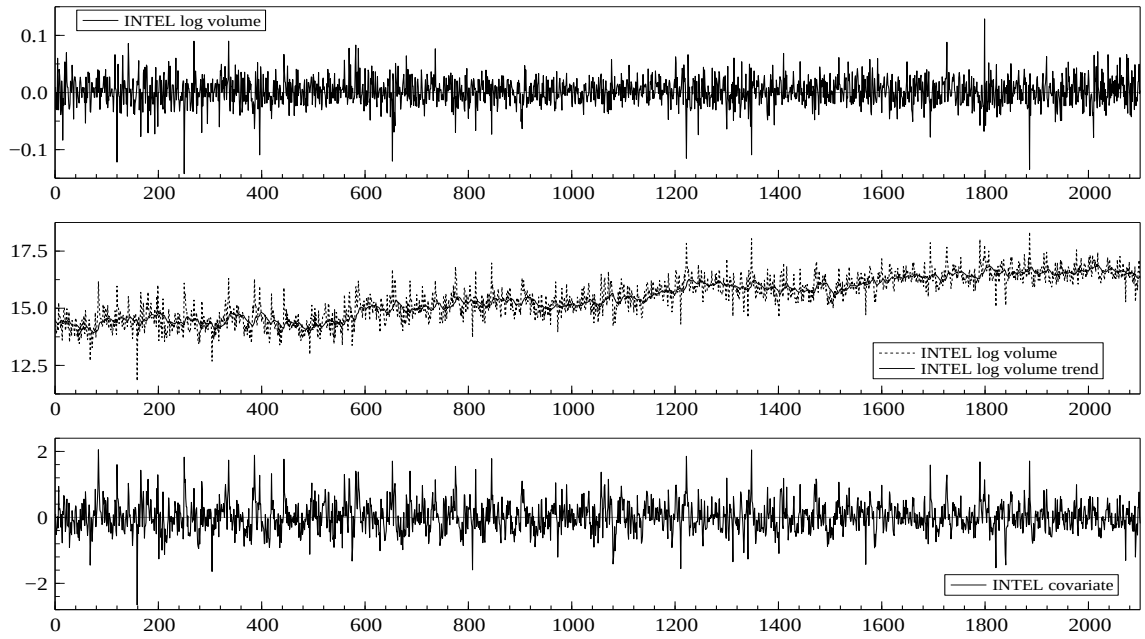

Figure 2: Log return, log volume, log volume trend and covariate time series - INTEL stock.

the probability integral transform, initially defined by Rosenblatt (1952), which has been employed by a number of recent papers such as Diebold, Gunther and Tay (1998) and Berkowitz (2000). The analysis is based on the relationship between the data generating process $f_t(y_t)$, for the response variable y_t, and the sequence of predicted densities $\hat{p}_t(y_t)$, obtained by using the mixture model.

The forecasting evaluation was performed based on the one-step-ahead density prediction. Initially we used 1,500 observations (from September 19, 1990, to August 22, 1996) to estimate the model parameters and to select the lags of the response y_t and the covariate x_t used in the experts and in the gate functions. We then used the estimated parameters to construct the predicted conditional density $\hat{p}_1(y_1)$ for the next day, August 23, 1996. To obtain $\hat{p}_2(y_2)$, we deleted the first observation, corresponding to September 19, 1990, and added the observation for August 23, 1996, so that the new model parameters are estimated based on a rolling window always with 1,500 data points. We proceeded with the rolling window approach and obtained a series of one-step ahead predicted conditional densities $\hat{p}_t(y_t)$, $t = 1, \ldots, T_{eval} = 600$, with the last prediction $\hat{p}_{600}(y_{600})$ corresponding to January 8,

1999.

The probability integral transform u_t, $t = 1, \ldots, T_{eval}$ is the cumulative density function corresponding to the density $\hat{p}_t(y_t)$ evaluated at the actual observed value y_t,

$$u_t = \int_{-\infty}^{y_t} \hat{p}_t(x)dx \equiv \hat{P}_t(y_t).$$

We then have the following fact, a proof of which can be found in Diebold, Gunther and Tay (1998), which is the backbone for the model checking analysis is this paper: *If a sequence of density estimates* $\{\hat{p}_t(y_t)\}_{t=1}^{T_{eval}}$ *coincides with the true data generating process* $\{f_t(y_t)\}_{t=1}^{T_{eval}}$, *then under the usual conditions of nonzero Jacobian with continuous partial derivatives, the sequence of probability integral transforms* $\{u_t\}_{t=1}^{T_{eval}}$ *of* $\{y_t\}_{t=1}^{T_{eval}}$ *with respect to* $\{\hat{p}_t(y_t)\}_{t=1}^{T_{eval}}$ *is i.i.d.* $U(0,1)$.

In this paper, instead of working directly with the sequence $\{u_t\}_{t=1}^{T_{eval}}$, we will follow the suggestion of Berkowitz (2000) and work with the transformation $\{\Phi^{-1}(u_t)\}_{t=1}^{T_{eval}}$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function. In this case, the aforementioned fact implies that $\{\Phi^{-1}(u_t)\}_{t=1}^{T_{eval}}$ is a standard normal i.i.d. sequence. Therefore, after estimating the mixtures of generalized t experts, we can evaluate the model fitting by testing for the hypothesis of independence and standard normality for the constructed series $\{z_t\}_{t=1}^{T_{eval}}$, where $z_t = \Phi^{-1}(u_t)$, $t = 1, \ldots, T$. Diebold, Gunther and Tay suggest a number of graphical methods for assessing goodness-of-fit and forecast ability. In their spirit, a simple first-step analysis can be done by plotting the histogram for the series z_t and comparing it to the standard normal density function. Figures 3 and 4 present the histograms and the PP-plots of $\{z_t\}_{t=1}^{T_{eval}}$ for HP and INTEL stocks respectively. Note that in both cases the forecasted density is quite close to benchmark $N(0,1)$ density.

Additionally, we can plot the autocorrelation function for the series $z_t - \bar{z}$, $(z_t - \bar{z})^2$, $(z_t - \bar{z})^3$ and $(z_t - \bar{z})^4$, as suggested by Diebold, Gunther and Tay (1998), where $\bar{z}$ is the sample mean for $\{z_t\}_{t=1}^{T}$. Figures 5 and 6 contain the plots of autocorrelation functions for the 4 series above for HP and INTEL stocks respectively along with the corresponding 95% confidence bands. The ACF plots seem to provide support for the hypothesis of independence for $\{z_t\}_{t=1}^{T}$. Again, the graphical analysis supports the validity of the mixtures of heavy tail experts in approximating the true conditional density.

Besides the informal - but rather informative - previous graphical analysis, methods can be advanced to formally test the hypothesis of i.i.d. $N(0,1)$. Berkowitz (2000) proposes a

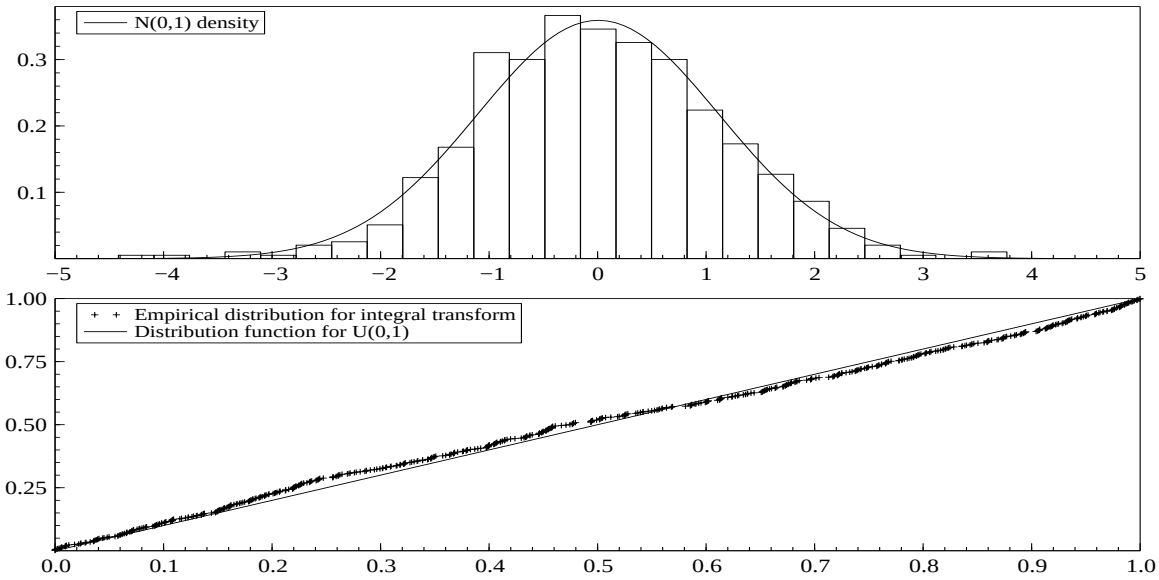

Figure 3: Estimated density for the transformed series z_t for HP stock returns.

likelihood ratio (LR) approach, while Hong (2000) proposes a test based on a generalized spectral approach. We employ the LR test of Berkowitz and we further develop a simple, but powerful, test based on moment conditions. Berkowitz's LR test amounts to fitting an autoregressive model of the form

$$z_t - \mu = \rho(z_{t-1} - \mu) + \varepsilon_t$$

and observing that the null hypothesis of i.i.d. $N(0,1)$ implies $\mu = 0$, $\sigma^2 = 1$ and $\rho = 0$ where σ^2 is the variance of the noise term ε_t. Let us denote by $L(\mu, \sigma^2, \rho)$ the loglikelihood function and by $(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})$ the maximum likelihood estimator under normality of the noise ε_t. It then follows by standard arguments that the likelihood ratio statistic

$$LR = -2\left(L(0,1,0) - L(\widehat{\mu}, \widehat{\sigma}^2, \widehat{\rho})\right)$$

has an asymptotic χ^2 distribution with 3 degrees of freedom under the hypothesis $(\mu, \sigma^2, \rho) = (0,1,0)$. The resulting LR test statistics are equal to 0.1689 and 3.8691 for HP and INTEL respectively with the p-values being equal to 0.9824 and 0.2760. Thus, in both cases the null hypothesis of i.i.d. $N(0,1)$ is supported.

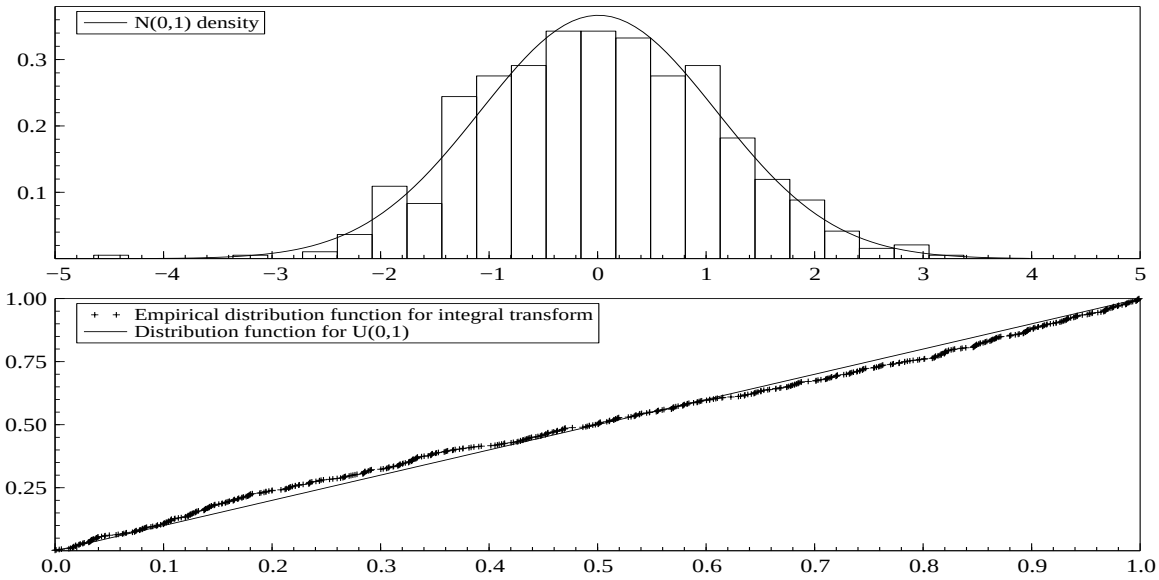

Figure 4: Estimated density for the transformed series z_t for INTEL stock returns.

The alternative simple testing procedure is based on a χ^2 test for the validity of moment conditions for the series $\{z_t\}_{t=1}^{T_{eval}}$. To test for the normality hypothesis, we test for the moment restrictions $E[z_t] = 0$, $E[z_t^2 - 1] = 0$, $E[z_t^3] = 0$ and $E[z_t^4 - 3] = 0$. To test for the hypothesis of independence, we will consider the moment conditions $E[z_t z_{t-1}] = 0$, $E[z_t z_{t-1}^2] = 0$, $E[z_t^2 z_{t-1}] = 0$, $E[z_t^2 z_{t-1}^2 - 1] = 0$, $E[z_t z_{t-2}] = 0$, $E[z_t z_{t-2}^2] = 0$, $E[z_t^2 z_{t-2}] = 0$ and $E[z_t^2 z_{t-2}^2 - 1] = 0$. In this case, let $\mathbf{m}_t$ be a 12×1 vector corresponding to the sample analogue of the 12 moment conditions just described. Under the null hypothesis of the model being correct, we have that $\{z_t\}_{t=1}^{T}$ are i.i.d. $N(0,1)$. Thus, the test statistic $S_{T_{eval}} = T_{eval} \overline{\mathbf{m}}'_{T_{eval}} \widehat{\mathbf{\Omega}}^{-1}_{T_{eval}} \overline{\mathbf{m}}_{T_{eval}}$ has an asymptotic χ^2 distribution with 12 degrees of freedom, where

$$\overline{\mathbf{m}}_{T_{eval}} = \frac{1}{T} \sum_{t=1}^{T_{eval}} \mathbf{m}_t,$$

$$\mathbf{m}'_t = [z_t, z_t^2 - 1, z_t^3, z_t^4 - 3, z_t z_{t-1}, z_t^2 z_{t-1}, z_t z_{t-1}^2, z_t^2 z_{t-1}^2 - 1, z_t z_{t-2}, z_t^2 z_{t-2}, z_t z_{t-2}^2, z_t^2 z_{t-2}^2 - 1]$$

for $t = 3, \ldots, T_{eval}$ and $\widehat{\mathbf{\Omega}}_{T_{eval}}$ is a consistent estimator of the asymptotic covariance matrix of $\overline{\mathbf{m}}_{T_{eval}}$. We use the typical choice for $\widehat{\mathbf{\Omega}}_{T_{eval}}$ which is the Newey-West estimator (see Newey and West, 1987). In our implementation we use 5 lags in the estimation of the

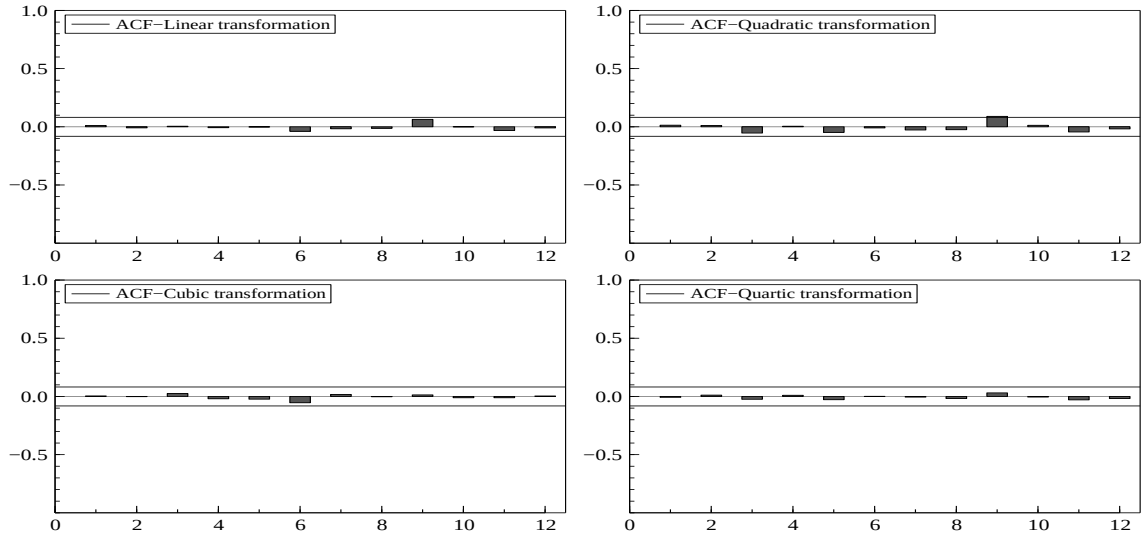

Figure 5: Autocorrelation function for the transformations of z_t - HP stock return.

asymptotic covariance matrix, but we document that the results are very robust to the choice of the number of lags. The test statistic was found to be equal to 16.034 and 11.390 for HP and INTEL respectively. The corresponding p-values are 0.18968 and 0.49584. Therefore, in both cases, the null hypothesis is not rejected.

8 Conclusion

A class of nonlinear time series models based on mixtures-of-experts is proposed and analyzed. The mean function of each expert depends on a number of covariates that also determine the probability weights according to which the experts are mixed. The parametric family of generalized t distributions is employed to model the error terms in each component. This is a quite flexible distribution that allows separation of scale and kurtosis and enables us to capture the tail-thickness of the data.

We provide simple conditions that guarantee the identifiability of the model parameters. In the case of autoregressive experts we show the uniform ergodicity of the system under appropriate simple conditions. We further present a detailed account of the asymptotic behavior of the maximum likelihood estimator (MLE). Under standard assumptions, in the context of extremum estimators, we establish consistency and asymptotically normality of the

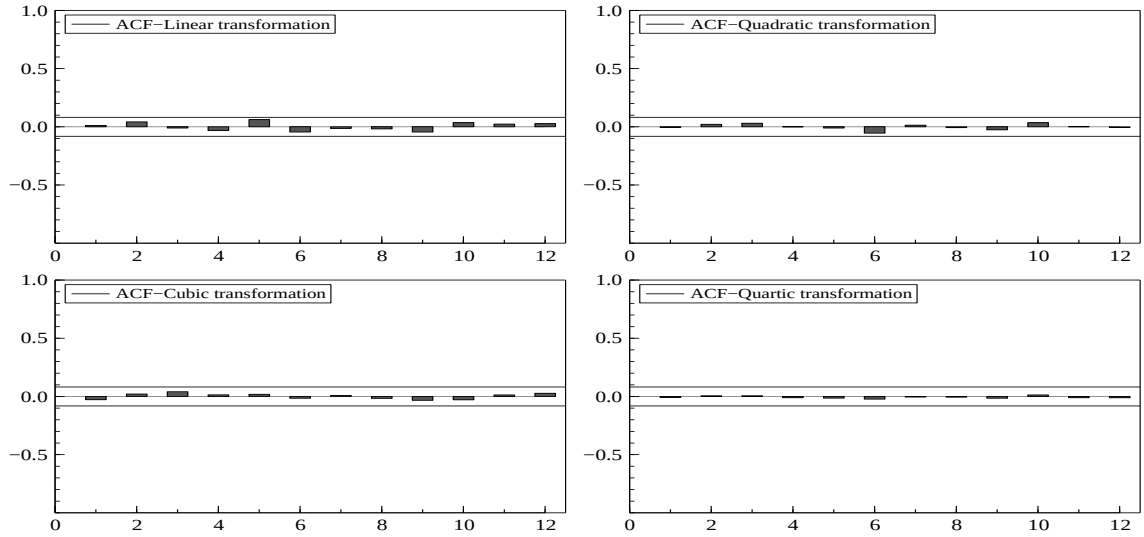

Figure 6: Autocorrelation function for the transformations of z_t - INTEL stock return.

MLE. Confidence intervals for the parameters can then be readily constructed and hypothesis testing can be performed.

On the implementation side, two issues stand out. The first issue pertains to the choice of the numbers of experts and lags to be included in the analysis. Although a procedure, with sound theoretical basis, for this purpose is not currently available, our simulation results suggest that the BIC criterion provides a reliable tool for model selection in our context. The second issue relates to the sample sizes used in practical applications. It should borne in mind that, in general, if we wish to successfully conduct inference about tail parameters then we need a sufficient amount of data - if we do not have enough data in the tails then the tail parameters will be poorly identified by the sample causing the corresponding standard errors to be rather large. The exact same principle applies to the model proposed in this paper. We apply the model to two time series of stock returns and choose a model with two experts as the most appropriate according to the BIC criterion. Employing some recently developed methodology of goodness-of-fit tests, we find that the mixture of generalized t experts provide an adequate description of conditional dynamics of the time series studied.

APPENDIX

A The generalized t (GT) distribution - an overview

The generalized t (GT) distribution was introduced in the literature by Rider (1958) under the term generalized Cauchy distribution. It is a 4-parameter symmetric distribution the density of which assumes the following form

$$f_{\mathrm{GT}}(x;\mu,\gamma,p,\alpha) = \frac{p}{2\gamma \mathrm{B}(\frac{1}{p},\frac{\alpha}{p})\left(1+|\frac{x-\mu}{\gamma}|^p\right)^{\frac{\alpha+1}{p}}} \tag{20}$$

where $-\infty < x < \infty$, $\mu \in \mathbb{R}$, $\gamma > 0$, $p > 0$, and $\alpha > 0$. The interpretation of the parameters in the above specification is clear: μ and γ are the location and scale parameters respectively (in the sense that if $X \sim \mathrm{GT}(0,1,p,\alpha)$ then $Y = \mu + \gamma X \sim \mathrm{GT}(\mu,\gamma,p,\alpha)$) whereas p describes the peakedness at the center, and α is the tail-thickness parameter. Note that $f_{\mathrm{GT}}(x;\mu,\gamma,p,\alpha) \sim |x|^{-\alpha-1}$ as $x \to \pm\infty$, i.e. the generalized t distribution with tail-thickness parameter $\alpha < 2$ is tail-equivalent to the α-stable distribution.

The GT distribution has some very important special cases. For $p = 2$ and $\alpha = 1$, GT reduces to the standard Cauchy with location parameter μ and scale parameter γ. More generally, for $\gamma = \sqrt{k}$, $p = 2$, and $\alpha = k$, where k is a positive integer, GT reduces to the standard Student's t distribution with k degrees of freedom.

The mth absolute central moment of the GT distribution, $\mu_m = E[|Y-\mu|^m]$ where $Y \sim \mathrm{GT}(\mu,\gamma,p,\alpha)$), exists if and only if $\alpha > m$ and equals

$$\mu_m = \frac{\mathrm{B}(\frac{m+1}{p},\frac{\alpha-m}{p})}{\mathrm{B}(\frac{1}{p},\frac{\alpha}{p})}\gamma^m \tag{21}$$

as can be shown by using the integration formula $\int_0^\infty (x+\beta)^{-\nu}x^{\mu-1}dx = \beta^{\mu-\nu}\mathrm{B}(\nu-\mu,\mu)$ where $\nu > \mu > 0$ (see formula 3, section 8.380 in Gradshteyn and Ryzhik (1980)). For the special case $p = 2$ this reduces to

$$\mu_{2m} = \frac{1\cdot 3\cdots(2m-1)}{(\alpha-2)\cdot(\alpha-4)\cdots(\alpha-2m)}\gamma^{2m}.$$

In general, for $\alpha > 4$, the kurtosis coefficient of GT is given by

$$\frac{\mu_4}{\mu_2^2} = \frac{\mathrm{B}(\frac{1}{p},\frac{5}{p})}{\mathrm{B}(\frac{3}{p},\frac{3}{p})}\frac{\mathrm{B}(\frac{\alpha-4}{p},\frac{\alpha}{p})}{\mathrm{B}(\frac{\alpha-2}{p},\frac{\alpha-2}{p})}.$$

Utilizing the reparameterization $\alpha = qp$ and $\gamma = \sigma q^{1/p}$ in the above density we obtain the form

$$f_{\mathrm{GT}}(x;\mu,\sigma,p,q) = \frac{p}{2\sigma q^{\frac{1}{p}}\mathrm{B}(\frac{1}{p},q)\Big(1+\frac{|x-\mu|^p}{q\sigma^p}\Big)^{q+\frac{1}{p}}}$$

which is the form used in McDonald and Newey (1988) for applying partially adaptive estimating procedures to regression models, and in McDonald (1996) for financial modelling.

B Proofs

B.1 Proof of Lemma 3.1

We use induction on M. For $M = 1$ the result holds trivially. Let $M = K \geq 2$ and assume that the result holds for all $M \leq K-1$. Consider real numbers $c_1, c_2, \ldots, c_K$ such that, for all $y \in \mathbb{R}$,

$$\sum_{j=1}^{K} c_j \pi(y;\mu_j,\gamma_j,\alpha_j) = 0 \Leftrightarrow \sum_{j=1}^{K} c_j \frac{\gamma_j^{\alpha_j}}{B(1/2,\alpha_j/2)} \frac{1}{\left[\gamma_j^2+(y-\mu_j)^2\right]^{\frac{\alpha_j+1}{2}}} = 0. \tag{22}$$

Since the expert densities are distinct elements of $\mathbb{G}$, we have $\alpha_{\rho(1)} < \ldots < \alpha_{\rho(K)}$ for some permutation ρ of the set $\{1,\ldots,K\}$. Multiplying both sides of the identity (22) by $y^{\alpha_{\rho(1)}}$ and letting $y \to \infty$, we obtain $c_{\rho(1)} = 0$. Using now the induction hypothesis completes the proof. ■

B.2 Proof of Theorem 5.1 (Consistency of the MLE)

The proof of Theorem 5.1 follows the standard line of showing the consistency of extremum estimators. It will be based on the following two lemmas. The first lemma can be found in Amemiya (1985) as Theorem 4.1.1.

Lemma B.1 *Assume that the following conditions hold:*
(a) The parameter space $\boldsymbol{\Theta}$ *is a compact subset of* $\mathbb{R}^K$ *and* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$*;*
(b) $\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ *is a continuous function in* $\boldsymbol{\theta}$ *for all* $(y_t, \mathbf{x}_{t-1}')'$*;*
(c) $\frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ *converges to* $E[\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ *in probability uniformly in* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *as* $T \to \infty$*;*

(d) $E[\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ *is strictly maximized at* $\boldsymbol{\theta}_0$ *over* $\boldsymbol{\Theta}$.
Let $\widehat{\boldsymbol{\theta}}_T$ *be defined as follows*

$$\widehat{\boldsymbol{\theta}}_T = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}). \tag{23}$$

Then $\widehat{\boldsymbol{\theta}}_T$ *converges to* $\boldsymbol{\theta}_0$ *in probability as* $T \to \infty$.

The second lemma follows from Corollary 2.2 and the subsequent discussion in Newey (1991).

Lemma B.2 *Assume that the following conditions hold:*
(a) The parameter space $\boldsymbol{\Theta}$ *is a compact and convex subset of* $\mathbb{R}^K$;
(b) For each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $\frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ *converges to* $E[\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ *in probability as* $T \to \infty$;
(c) $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})\|\right] < \infty$.
Then $\frac{1}{T}\sum_{t=1}^{T} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ *converges to* $E\left[\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})\right]$ *in probability uniformly in* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ *as* $T \to \infty$.

It follows from the density specification given by (2), (3) and (5) that assumption (b) in Lemma B.1 is satisfied. Since the true parameter value is $\boldsymbol{\theta}_0$ and the parameter is identified, the standard argument yields that $E[\log f(y_t|\mathbf{x}_t;\boldsymbol{\theta})]$ is uniquely maximized at $\boldsymbol{\theta}_0$ over $\boldsymbol{\Theta}$, i.e., assumption (d) in Lemma B.1 is also satisfied. To show that assumption (c) in Lemma B.1 is satisfied, we will employ Lemma B.2. For that, we need to establish the validity of assumptions (b) and (c) in Lemma B.2. Then, the proof of Theorem 5.1 simply follows by combining the two lemmas above.

It is more convenient to first show that $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})\|\right] < \infty$. Using the triangular inequality and the fact that $g_j(\cdot;\cdot)$ and $\pi(\cdot|\cdot;\cdot)$ are positive we obtain

$$\|\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})\| = \frac{\left\|\sum_{j=1}^{J} \partial_{\boldsymbol{\theta}} \left[g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right]\right\|}{\sum_{k=1}^{J} g_k(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_k)}$$

$$\leq \sum_{j=1}^{J} \frac{\left\|\partial_{\boldsymbol{\theta}} g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\right\| \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j) + g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}) \left\|\partial_{\boldsymbol{\theta}} \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right\|}{\sum_{k=1}^{J} g_k(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_k)}$$

$$\leq \sum_{j=1}^{J} \frac{\left\|\partial_{\boldsymbol{\theta}} g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\right\| \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j) + g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}) \left\|\partial_{\boldsymbol{\theta}} \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right\|}{g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)}$$

$$= \sum_{j=1}^{J} \left\| \partial_{\boldsymbol{\theta}} \log g_j(\mathbf{x}_{t-1}; \boldsymbol{\lambda}) \right\| + \sum_{j=1}^{J} \left\| \partial_{\boldsymbol{\theta}} \log \pi(y_t | \mathbf{x}_{t-1}; \boldsymbol{\zeta}_j) \right\|.$$

Therefore,

$$\begin{aligned} &E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})\|\right] \le \textstyle\sum_{j=1}^{J} E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\partial_{\boldsymbol{\theta}} \log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\right\|\right] \\ &+ \textstyle\sum_{j=1}^{J} E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\partial_{\boldsymbol{\theta}} \log \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right\|\right]. \end{aligned} \tag{24}$$

First, we provide a bound for the first term in the right hand side of (24). Specifically, we show that

$$E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\partial_{\boldsymbol{\theta}} \log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\right\|\right] \le J(1 + E\left[\|\mathbf{x}_{t-1}\|\right]), \text{ for all } j \in \{1,\ldots,J\}. \tag{25}$$

Consider any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $j \in \{1,\ldots,J-1\}$. Then

$$\log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}) = \log\left(\frac{e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_j}}{1+\sum_{k=1}^{J-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_k}}\right) = \mathbf{z}'_{t-1}\boldsymbol{\omega}_j - \log\left(1+\sum_{k=1}^{J-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_k}\right)$$

$$\Rightarrow \|\partial_{\boldsymbol{\theta}} \log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\| \le \|\mathbf{z}_{t-1}\| + \frac{\left\|\partial_{\boldsymbol{\theta}}\left(\sum_{k=1}^{J-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_k}\right)\right\|}{1+\sum_{m=1}^{J-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_m}}$$

$$\le \|\mathbf{z}_t\| + \sum_{k=1}^{J-1} \frac{e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_k}}{1+\sum_{m=1}^{J-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_m}} \|\mathbf{z}_{t-1}\| \le J\|\mathbf{z}_{t-1}\|.$$

It then follows that $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\|\right] \le JE\left[\|\mathbf{z}_{t-1}\|\right] \le J(1+E\left[\|\mathbf{x}_{t-1}\|\right])$, for $1 \le j \le J-1$. Using a similar derivation, we can easily conclude that

$$E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log g_J(\mathbf{x}_{t-1};\boldsymbol{\lambda})\|\right] \le (J-1)(1+E\left[\|\mathbf{x}_{t-1}\|\right]).$$

Thus the validity of (25) has been established.

Next we provide a bound for the second term in the right hand side of (24). Notice that we have

$$\begin{aligned} \left\|\partial_{\boldsymbol{\theta}} \log \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right\|^2 &= \left|\partial_{a_j} \log \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right|^2 + \left\|\partial_{\mathbf{b}_j} \log \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right\|^2 \\ &\quad + \left|\partial_{\gamma_j} \log \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right|^2 + \left|\partial_{\alpha_j} \log \pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)\right|^2 \end{aligned} \tag{26}$$

and

$$\begin{aligned} \log \pi(y_t|\mathbf{x}_{t-1},\boldsymbol{\zeta}_j) &= -\log(\gamma_j) - \log\left(B\left(1/2,\alpha_j/2\right)\right) \\ &\quad -\tfrac{\alpha_j+1}{2} \log\left(1+\left(y_t-\mu_{tj}\right)^2/\gamma_j^2\right) \end{aligned} \tag{27}$$

where $\mu_{tj} = a_j + \mathbf{b}_j' \mathbf{x}_{t-1}$. By the compactness of $\boldsymbol{\Theta}$, there exist positive $\underline{\gamma}$, $\overline{\gamma}$, $\underline{\alpha}$, $\overline{\alpha}$, $\underline{\Delta}$ and $\overline{\Delta}$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $j \in 1, 2, \ldots, J$ we have $\underline{\gamma} \leq \gamma_j \leq \overline{\gamma}$, $\underline{\alpha} \leq \alpha_j \leq \overline{\alpha}$ and $\underline{\Delta} \leq ||\boldsymbol{\delta}_j|| \leq \overline{\Delta}$. Using the fact that $d \log(B(1/2, x/2))/dx = \frac{1}{2}\left[\Psi\left(x/2\right) - \Psi\left((1+x)/2\right)\right]$ where $\Psi(x) = \frac{d \log(\Gamma(x))}{dx}$ is the digamma function, we obtain

$$\begin{aligned} \partial_{\alpha_j} \log \pi(y_t | \mathbf{x}_{t-1}; \boldsymbol{\zeta}_j) = & -\tfrac{1}{2}\left[\Psi(\alpha_j/2) - \Psi((\alpha_j+1)/2)\right] \\ & -\tfrac{1}{2} \log\left(1 + (y_t - \mu_{tj})^2/\gamma_j^2\right). \end{aligned} \tag{28}$$

Moreover, since $\log(1+x^2) < 2x$ for all $x > 0$, we have

$$\log\left(1 + (y_t - \mu_{tj})^2/\gamma_j^2\right) \leq 2\frac{|y_t - \mu_{tj}|}{\gamma_j} \leq 2\frac{|y_t| + \overline{\Delta}(1 + \|\mathbf{x}_{t-1}\|)}{\underline{\gamma}} \tag{29}$$

where the last step uses the triangular and Cauchy-Schwarz inequalities. From expressions (28) and (29), the compactness of the parameter space $\boldsymbol{\Theta}$ and the continuity of $\Psi(\cdot)$, we conclude that there exist positive constants $K_\alpha, K_\alpha', K_\alpha''$ such that

$$|\partial_{\alpha_j} \log \pi(y_t | \mathbf{x}_{t-1}; \boldsymbol{\zeta}_j)| \leq K_\alpha + K_\alpha' |y_t| + K_\alpha'' \|\mathbf{x}_{t-1}\|, \text{ for all } \boldsymbol{\theta} \in \boldsymbol{\Theta} \text{ and all } j = 1, \ldots, J. \tag{30}$$

Similarly, using the fact that $\max\{\frac{x^2}{1+x^2}, \frac{x}{1+x^2}\} \leq 1$ for all $x > 0$ and the compactness of $\boldsymbol{\Theta}$, we obtain that, for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and all $j = 1, \ldots, J$,

$$|\partial_{\gamma_j} \log \pi(y_t | \mathbf{x}_{t-1}; \boldsymbol{\zeta}_j)| = \left| -\frac{1}{\gamma_j} + \frac{\alpha_j + 1}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} \frac{(y_t - \mu_{tj})^2}{\gamma_j^3} \right| \leq K_\gamma, \tag{31}$$

$$|\partial_{a_j} \log \pi(y_t | \mathbf{x}_{t-1}; \boldsymbol{\zeta}_j)| = \left| \frac{\alpha_j + 1}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} \frac{y_t - \mu_{tj}}{\gamma_j^2} \right| \leq K_a \tag{32}$$

and

$$\|\partial_{\mathbf{b}_j} \log \pi(y_t | \mathbf{x}_{t-1}; \boldsymbol{\zeta}_j)\| = \left| \frac{\alpha_j + 1}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} \frac{y_t - \mu_{tj}}{\gamma_j^2} \right| \|\mathbf{x}_{t-1}\| \leq K_{\mathbf{b}} \|\mathbf{x}_{t-1}\| \tag{33}$$

for some positive constants K_γ, K_a and $K_{\mathbf{b}}$.

The next step involves combining the inequality (24), the bound in (25) and the expression (26) along with bounds in (30) (31), (32) and (33) to conclude that there exist universal positive constants K_1, K_2, K_3 such that

$$E\left[\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log f(y_t | \mathbf{x}_{t-1}; \boldsymbol{\theta})\|\right] \leq K_1 + K_2 E\left[|y_t|\right] + K_3 E\left[\|\mathbf{x}_{t-1}\|\right]. \tag{34}$$

Next we establish that $E\left[|y_t|\right] < \infty$. Conditioning on the gate we have

$$E\left[|y_t| \mid \mathbf{x}_{t-1}\right] = \sum_{j=1}^{J} g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}) E\left[|y_t| \mid \mathbf{x}_{t-1}, j\right] \leq \sum_{j=1}^{J} E\left[|y_t| \mid \mathbf{x}_{t-1}, j\right].$$

From assumption (d) in Theorem 5.1 and the expression for the central moments of the generalized t distribution given by (21), it follows that

$$\begin{aligned} E\left[|y_t| \mid \mathbf{x}_{t-1}, j\right] &\leq E\left[\left|y_t - \left(a_{0,j} + \mathbf{b}_{0,j}'\mathbf{x}_{t-1}\right)\right| \mid \mathbf{x}_{t-1}, j\right] + \left|a_{0,j} + \mathbf{b}_{0,j}'\mathbf{x}_{t-1}\right| \\ &\leq \frac{B(1, \frac{\alpha_{0,j}-1}{2})}{B(\frac{1}{2}, \frac{\alpha_{0,j}}{2})}\gamma_{0,j} + |a_{0,j}| + \|\mathbf{b}_{0,j}\|\|\mathbf{x}_{t-1}\|. \end{aligned}$$

Then, by the law of iterated expectations and Condition 4, it follows that $E\left[|y_t|\right] < \infty$. Using this fact and Condition 4, we obtain from (34) that $E\left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \|\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})\|\right] < \infty$.

Next we show that $E[|\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|] < \infty$ for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Since $\sum_{j=1}^{J} g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}) = 1$, $g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}) \in (0,1)$ and $\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j) > 0$ we have

$$\begin{aligned} |\log(\textstyle\sum_{j=1}^{J} g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j))| &\leq \max_{j=1,\ldots,J} |\log(\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j))| \\ &\leq \textstyle\sum_{j=1}^{J} |\log(\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j))|. \end{aligned} \tag{35}$$

Combining (27), (29) and (35), we conclude that $|\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})| \leq L_1 + L_2|y_t| + L_3\|\mathbf{x}_{t-1}\|$ for some positive constants L_1, L_2 and L_3. Thus, $E[|\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|] < \infty$ as follows from the fact $E[|y_t|] < \infty$ and Condition 4. Using the fact that $E[|\log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|] < \infty$ and Condition 3, we obtain that assumption (b) in Lemma B.2 is satisfied. This completes the proof of Theorem 5.1. ■

B.3 Proof of Theorem 5.2 (Asymptotic normality of the MLE)

The following standard lemma, which is Theorem 4.1.3 in Amemiya (1985) adopted to our setup, forms the basis of our proof. Our goal is to utilize the assumptions of Theorem 5.2 to show that all the conditions for the lemma to be applicable in our setup are satisfied.

Lemma B.3 *Let $\widehat{\boldsymbol{\theta}}_T$ be the MLE as defined by (12) and assume the following conditions hold:*
(a) $\widehat{\boldsymbol{\theta}}_T \xrightarrow{P} \boldsymbol{\theta}_0$ as $T \to \infty$;

(b) the true parameter $\boldsymbol{\theta}_0$ *is in the interior of* $\boldsymbol{\Theta}$ *and* $\partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ *exists and is a continuous function of* $\boldsymbol{\theta}$ *in the interior of* $\boldsymbol{\Theta}$ *for all* $(y_t, \mathbf{x}'_{t-1})'$*;*
(c) for any sequence $\{\boldsymbol{\theta}^*_T\}$*, such that* $\boldsymbol{\theta}^*_T \xrightarrow{P} \boldsymbol{\theta}_0$ *as* $T \to \infty$*, the probability limit of the matrix sequence* $\frac{1}{T}\sum_{t=1}^{T} \partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}^*_T)$ *is the finite nonsingular matrix*

$$\mathbf{A}(\boldsymbol{\theta}_0) = E[\partial_{\boldsymbol{\theta}}\partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0)];$$

(d) $\frac{1}{\sqrt{T}}\sum_{t=1}^{T} \partial_{\boldsymbol{\theta}} f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N(\mathbf{0}_K, \mathbf{B}(\boldsymbol{\theta}_0))$ *as* $T \to \infty$*.*
Then

$$\sqrt{T}(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} N\left(\mathbf{0}_K, \mathbf{A}(\boldsymbol{\theta}_0)^{-1}\mathbf{B}(\boldsymbol{\theta}_0)\mathbf{A}(\boldsymbol{\theta}_0)^{-1}\right) \quad as \quad T \to \infty. \tag{36}$$

Under the assumptions of Theorem 5.2, it follows from Theorem 5.1 that assumption (a) in Lemma B.3 is satisfied. It immediately follows from the form of $f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ that assumption (b) in Lemma B.3 is also satisfied. The following auxiliary lemma, which is Theorem 4.1.5 in Amemiya (1985) adopted to our framework, provides sufficient conditions for the validity of the first part of assumption (c) in Lemma B.3.

Lemma B.4 *Let* θ_i *and* θ_k *be any scalar components of* $\boldsymbol{\theta}$*. Assume that the sample mean* $\frac{1}{T}\sum_{t=1}^{T} \partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ *converges in probability to* $E[\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ *uniformly in* $\boldsymbol{\theta} \in \boldsymbol{\Theta}$*,* $\boldsymbol{\theta}_0$ *is an interior point of* $\boldsymbol{\Theta}$*,* $\boldsymbol{\theta}^*_T \xrightarrow{P} \boldsymbol{\theta}_0$ *as* $T \to \infty$ *and* $E[\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ *is continuous at* $\boldsymbol{\theta}_0$*. Then*

$$\frac{1}{T}\sum_{t=1}^{T} \partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}^*_T) \xrightarrow{P} E[\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0)] \quad as \quad T \to \infty.$$

To apply the lemma above, we need to show that $E[|\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|] < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and establish the uniform convergence in probability of $\frac{1}{T}\sum_{t=1}^{T} \partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ to $E[\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$. Once the finiteness of $E[|\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|]$ is shown, we can utilize Condition 3 to obtain that $\frac{1}{T}\sum_{t=1}^{T} \partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ converges to the expectation $E[\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ in probability for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Thus, from Corollary 2.2 in Newey (1991) and the subsequent discussion (see also Lemma B.2), it follows that, to obtain the uniform convergence, it suffices to show that $E[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|] < \infty$ for each triple of scalar components $(\theta_i, \theta_k, \theta_m)$ of $\boldsymbol{\theta}$. To ease the computation of the cross

derivatives of the loglikelihood we will write f for $f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$, g_j for $g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda})$ and π_j for $\pi(y_t|\mathbf{x}_{t-1};\boldsymbol{\zeta}_j)$. Further, define

$$\tau_j = \frac{g_j\pi_j}{\sum_{l=1}^{J} g_l\pi_l} \tag{37}$$

so that $\tau_j \in (0,1)$ and $\sum_{j=1}^{J}\tau_j = 1$. A few lines of straightforward algebra yield that

$$\begin{aligned}
&\partial_{\theta_i}\log f = \phi_i^*,\\
&\partial_{\theta_i}\partial_{\theta_k}\log f = \phi_{ik}^* - \phi_i^*\phi_k^*,\\
&\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_k}\log f = \phi_{ikm}^* - \phi_{ik}^*\phi_m^* - \phi_{im}^*\phi_k^* - \phi_{km}^*\phi_i^* + 2\phi_i^*\phi_k^*\phi_m^*,
\end{aligned} \tag{38}$$

where

$$\begin{aligned}
&\phi_i^* = \tfrac{\partial_{\theta_i} f}{f} = \textstyle\sum_{j=1}^{J}\tau_j\phi_j^i,\\
&\phi_{ik}^* = \tfrac{\partial_{\theta_i}\partial_{\theta_k} f}{f} = \textstyle\sum_{j=1}^{J}\tau_j\phi_j^{ik},\\
&\phi_{ikm}^* = \tfrac{\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m} f}{f} = \textstyle\sum_{j=1}^{J}\tau_j\phi_j^{ikm}
\end{aligned} \tag{39}$$

and

$$\begin{aligned}
&\phi_j^i = \partial_{\theta_i}\log(g_j\pi_j),\\
&\phi_j^{ik} = \partial_{\theta_i}\partial_{\theta_k}\log(g_j\pi_j) + [\partial_{\theta_i}\log(g_j\pi_j)][\partial_{\theta_k}\log(g_j\pi_j)],\\
&\phi_j^{ikm} = \partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m}\log(g_j\pi_j) + [\partial_{\theta_i}\partial_{\theta_k}\log(g_j\pi_j)][\partial_{\theta_m}\log(g_j\pi_j)]\\
&\qquad +[\partial_{\theta_i}\partial_{\theta_m}\log(g_j\pi_j)][\partial_{\theta_k}\log(g_j\pi_j)] + [\partial_{\theta_k}\partial_{\theta_m}\log(g_j\pi_j)][\partial_{\theta_i}\log(g_j\pi_j)]\\
&\qquad\qquad +[\partial_{\theta_i}\log(g_j\pi_j)][\partial_{\theta_k}\log(g_j\pi_j)][\partial_{\theta_m}\log(g_j\pi_j)].
\end{aligned} \tag{40}$$

In order to proceed we require bounds on the partial derivatives of g_j and π_j. The bounds appear in the following two lemmas.

Lemma B.5 *The following bounds are valid:*
(a) $|\partial_{\theta_i}\log g_j| \le 1 + \|\mathbf{x}_{t-1}\|$, *for all* i *and* $j = 1,\ldots,J$;
(b) $|\partial_{\theta_i}\partial_{\theta_k}\log g_j| \le (1+\|\mathbf{x}_{t-1}\|)^2$, *for all* (i,k) *and* $j = 1,\ldots,J$;
(c) $|\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m}\log g_j| \le 2(1+\|\mathbf{x}_{t-1}\|)^3$, *for all* (i,k,m) *and* $j = 1,\ldots,J$.

Proof of Lemma B.5. We only consider the case $1 \leq j \leq J-1$ since the case $j = J$ follows similarly. To simplify the calculations we will use the notation $(1, \mathbf{x}'_{t-1})' = \mathbf{z}_{t-1}$ and $(v_j, \mathbf{u}'_j)' = \boldsymbol{\omega}_j$. We will write $z_{t-1,s}$ and ω_{js} to denote the sth elements of $\mathbf{z}_{t-1}$ and $\boldsymbol{\omega}_j$ respectively, for $s = 1, \ldots, S+1$. Since $\log g_j = \mathbf{z}'_{t-1}\boldsymbol{\omega}_j - \log\left(1 + \sum_{k=1}^{J-1} e^{\mathbf{z}'_{t-1}\boldsymbol{\omega}_k}\right)$, we have

$$\partial_{\omega_{ks}} \log g_j = \left(\mathbb{I}_{[k=j]} - g_k\right) z_{t-1,s}, \;\; k = 1, \ldots, J-1, \;\; s = 1, \ldots, S+1$$

where $\mathbb{I}$ denotes the indicator function. Then, part (a) follows since the weights g_j lie within (0,1). Moreover, we have

$$\partial_{\omega_{mr}} g_k = \left(\mathbb{I}_{[m=k]} - g_m\right) g_k z_{t-1,r}, \;\; k, m = 1, \ldots, J-1, \;\; r = 1, \ldots, S+1 \tag{41}$$

implying

$$\partial_{\omega_{mr}} \partial_{\omega_{ks}} \log g_j = -\left(\mathbb{I}_{[m=k]} - g_m\right) g_k z_{t-1,r} z_{t-1,s}, \;\; k, m = 1, \ldots, J-1, \;\; s, r = 1, \ldots, S+1$$

which delivers the conclusion in part (b). Using the expression (41) once again yields

$$\partial_{\omega_{lq}} \partial_{\omega_{mr}} \partial_{\omega_{ks}} \log g_j = -\left[\left(\mathbb{I}_{[m=k]} - g_m\right)\left(\mathbb{I}_{[l=k]} - g_l\right) - \left(\mathbb{I}_{[l=m]} - g_l\right) g_m\right] g_k z_{t-1,q} z_{t-1,r} z_{t-1,s}$$

for all $k, m, l = 1, \ldots, J-1, \;\; s, r, q = 1, \ldots, S+1$, from which we immediately obtain the bound in part (c). ■

Lemma B.6 *Given that the parameter space $\boldsymbol{\Theta}$ is compact, the following bounds are valid:*
(a) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\partial_{\theta_i} \log \pi_j| \leq C_1 + C_2|y_t| + C_3\|\mathbf{x}_{t-1}\|$, *for all i and $j = 1, \ldots, J$;*
(b) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k} \log \pi_j| \leq D_1 + D_2\|\mathbf{x}_{t-1}\| + D_3\|\mathbf{x}_{t-1}\|^2$, *for all (i,k) and $j = 1, \ldots, J$;*
(c) $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m} \log \pi_j| \leq F_1 + F_2\|\mathbf{x}_{t-1}\| + F_3\|\mathbf{x}_{t-1}\|^2 + F_4\|\mathbf{x}_{t-1}\|^3$, *for all (i,k,m) and $j = 1, \ldots, J$*
where C_r, D_r, $r = 1, \ldots, 3$ and F_r, $r = 1, \ldots, 4$ are positive constants.

Proof of Lemma B.6. A quick inspection of the proof of Theorem 5.1 in section B.2 reveals that the bound in part (a) is valid. In what follows, however, we present all the relevant partial order derivatives up to order 3 for the sake of completeness. We will use the notation $\mathbf{z}_{t-1} = (1, \mathbf{x}'_{t-1})'$ and $\boldsymbol{\delta}_j = (a_j, \mathbf{b}'_j)'$ and further denote by $z_{t-1,s}$ and δ_{js} the sth elements of $\mathbf{z}_{t-1}$ and $\boldsymbol{\delta}_j$ respectively, for $s = 1, \ldots, S+1$. The first order derivatives are:

$$\partial_{\alpha_j} \log \pi_j = -\frac{1}{2}\left[\Psi(\alpha_j/2) - \Psi((\alpha_j+1)/2)\right] - \frac{1}{2}\log\left(1 + (y_t - \mu_{tj})^2/\gamma_j^2\right),$$

$$\partial_{\gamma_j} \log \pi_j = -\frac{1}{\gamma_j} + \frac{\alpha_j + 1}{\gamma_j} \frac{(y_t - \mu_{tj})^2/\gamma_j^2}{1 + (y_t - \mu_{tj})^2/\gamma_j^2},$$

$$\partial_{\delta_{js}} \log \pi_j = \frac{\alpha_j + 1}{\gamma_j} \frac{(y_t - \mu_{tj})/\gamma_j}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} z_{t-1,s}.$$

The second order derivatives are:

$$\partial_{\alpha_j}\partial_{\alpha_j} \log \pi_j = -\frac{1}{4}\left[\Psi'(\alpha_j/2) - \Psi'((\alpha_j+1)/2)\right], \quad \partial_{\alpha_j}\partial_{\gamma_j} \log \pi_j = \frac{1}{\gamma_j} \frac{(y_t - \mu_{tj})^2/\gamma_j^2}{1 + (y_t - \mu_{tj})^2/\gamma_j^2},$$

$$\partial_{\alpha_j}\partial_{\delta_{js}} \log \pi_j = \frac{1}{\gamma_j} \frac{(y_t - \mu_{tj})/\gamma_j}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} z_{t-1,s},$$

$$\partial_{\gamma_j}\partial_{\gamma_j} \log \pi_j = \frac{1}{\gamma_j^2} - \frac{\alpha_j + 1}{\gamma_j^2} \frac{(y_t - \mu_{tj})^2/\gamma_j^2}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} \left[1 + \frac{2}{1 + (y_t - \mu_{tj})^2/\gamma_j^2}\right],$$

$$\partial_{\gamma_j}\partial_{\delta_{js}} \log \pi_j = -\frac{2(\alpha_j + 1)}{\gamma_j^2} \frac{(y_t - \mu_{tj})/\gamma_j}{\left[1 + (y_t - \mu_{tj})^2/\gamma_j^2\right]^2} z_{t-1,s},$$

$$\partial_{\delta_{js}}\partial_{\delta_{jr}} \log \pi_j = \frac{\alpha_j + 1}{\gamma_j^2} \frac{-1 + (y_t - \mu_{tj})^2/\gamma_j^2}{\left[1 + (y_t - \mu_{tj})^2/\gamma_j^2\right]^2} z_{t-1,s} z_{t-1,r}.$$

Using the preceding expressions for the second order derivatives, the continuity of the function $\Psi'(\cdot)$, the fact that $\max\{\frac{x^2}{1+x^2}, \frac{x}{1+x^2}\} \leq 1$ for all $x > 0$, and the compactness of $\mathbf{\Theta}$, we conclude that the bound in part (b) is valid. The third order derivatives are:

$$\partial_{\alpha_j}\partial_{\alpha_j}\partial_{\alpha_j} \log \pi_j = -\frac{1}{8}\left[\Psi''(\alpha_j/2) - \Psi''((\alpha_j+1)/2)\right], \quad \partial_{\alpha_j}\partial_{\alpha_j}\partial_{\gamma_j} \log \pi_j = 0, \quad \partial_{\alpha_j}\partial_{\alpha_j}\partial_{\delta_{js}} \log \pi_j = 0,$$

$$\partial_{\alpha_j}\partial_{\gamma_j}\partial_{\gamma_j} \log \pi_j = -\frac{1}{\gamma_j^2} \frac{(y_t - \mu_{tj})^2/\gamma_j^2}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} \left[1 + \frac{2}{1 + (y_t - \mu_{tj})^2/\gamma_j^2}\right],$$

$$\partial_{\alpha_j}\partial_{\gamma_j}\partial_{\delta_{js}} \log \pi_j = -\frac{2}{\gamma_j^2} \frac{(y_t - \mu_{tj})/\gamma_j}{\left[1 + (y_t - \mu_{tj})^2/\gamma_j^2\right]^2} z_{t-1,s},$$

$$\partial_{\alpha_j}\partial_{\delta_{js}}\partial_{\delta_{jr}} \log \pi_j = \frac{1}{\gamma_j^2} \frac{-1 + (y_t - \mu_{tj})^2/\gamma_j^2}{\left[1 + (y_t - \mu_{tj})^2/\gamma_j^2\right]^2} z_{t-1,s} z_{t-1,r},$$

$$\partial_{\alpha_j}\partial_{\gamma_j}\partial_{\delta_{js}} \log \pi_j = -\frac{2}{\gamma_j^2} \frac{(y_t - \mu_{tj})/\gamma_j}{\left[1 + (y_t - \mu_{tj})^2/\gamma_j^2\right]^2} z_{t-1,s},$$

$$\partial_{\gamma_j}\partial_{\gamma_j}\partial_{\gamma_j} \log \pi_j = -\frac{2}{\gamma_j^3} + \frac{2(\alpha_j + 1)(y_t - \mu_{tj})^2/\gamma_j^2}{\gamma_j^3 + \gamma_j (y_t - \mu_{tj})^2} \left[1 + \frac{1}{1 + (y_t - \mu_{tj})^2/\gamma_j^2} + \frac{4}{\left[1 + (y_t - \mu_{tj})^2/\gamma_j^2\right]^2}\right],$$

$$\partial_{\gamma_j}\partial_{\gamma_j}\partial_{\delta_{js}} \log \pi_j = -2(\alpha_j+1)\frac{(y_t-\mu_{tj})^2/\gamma_j^2}{1+(y_t-\mu_{tj})^2/\gamma_j^2}\left[1-\frac{2}{1+(y_t-\mu_{tj})^2/\gamma_j^2}\right] z_{t-1,s},$$

$$\partial_{\gamma_j}\partial_{\delta_{js}}\partial_{\delta_{jr}} \log \pi_j = \frac{2(\alpha_j+1)}{\gamma_j^3}\frac{1-3(y_t-\mu_{tj})^2/\gamma_j^2}{\left[1+(y_t-\mu_{tj})^2/\gamma_j^2\right]^3} z_{t-1,s}z_{t-1,r},$$

$$\partial_{\delta_{js}}\partial_{\delta_{jr}}\partial_{\delta_{jq}} \log \pi_j = \frac{-2(\alpha_j+1)}{\gamma_j^3}\frac{(y_t-\mu_{tj})/\gamma_j}{1+(y_t-\mu_{tj})^2/\gamma_j^2}\frac{3-(y_t-\mu_{tj})^2/\gamma_j^2}{\left[1+(y_t-\mu_{tj})^2/\gamma_j^2\right]^2} z_{t-1,s}z_{t-1,r}z_{t-1,q}.$$

Using the preceding expressions for the third order derivatives and repeating the reasoning above we obtain that the bound in part (c) is also valid. ■

Lemma B.7 *Assume that Condition 5 and assumption (d) of Theorem 5.2 are satisfied. Then, given that* $\boldsymbol{\Theta}$ *is compact, for every triple* $(\theta_i, \theta_k, \theta_m)$ *of components of the parameter vector* $\boldsymbol{\theta}$*, both* $E[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|]$ *and* $E[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|]$ *are finite.*

Proof of Lemma B.7. We only derive the result for the third order derivative. The argument for the second order derivative is identical and actually requires only $E[\|\mathbf{x}_t\|^2] < \infty$. Using the fact that $\tau_j \in (0,1)$, expressions (38), (39), (40) and Lemmas B.5 and B.6, we conclude that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k}\partial_{\theta_m} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|$ is bounded by a polynomial consisting of terms of the form $C|y_t|^p\|\mathbf{x}_{t-1}\|^q$ where $C \geq 0$ and p, q are nonnegative integers such that $p+q \leq 3$. Under assumption (d) of Theorem 5.2, the third absolute central conditional moment, given $\mathbf{x}_{t-1}$, of y_t is finite. Indeed, imitating the last step in the proof of Theorem 5.1, it is easily shown that $E[|y_t|^p \mid \mathbf{x}_{t-1}] \leq c_p + c_p'\|\mathbf{x}_{t-1}\|^p$ where $p \leq 3$ and c_p, c_p' are some positive constants. Thus, by the law of iterated expectations, we have that $E[|y_t|^p\|\mathbf{x}_{t-1}\|^q]$ is bounded by a positive affine combination of terms of the form $E[\|\mathbf{x}_{t-1}\|^r]$ where $r \leq 3$. Finally, applying Condition 5, which states that $E[\|\mathbf{x}_t\|^3] < \infty$, delivers the desired conclusion. ■

Since $E[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})|] < \infty$ (Lemma B.7) and $\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$, the dominated convergence theorem implies that $E[\partial_{\theta_i}\partial_{\theta_k} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta})]$ is continuous at $\boldsymbol{\theta}_0$. Thus, Lemma B.7 allows us to apply Lemma B.4 to obtain that the first part of assumption (c) in Lemma B.3 is satisfied. The next lemma guarantees the validity of the second part of assumption (c) in Lemma B.3.

Lemma B.8 *Under Conditions 2 and 1, the matrix*

$$\mathbf{M}_0 \equiv E[\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0)\partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0)]$$

is nonsingular.

The proof of Lemma B.8 will use the following auxiliary lemma, the proof of which follows by using arguments similar to those made in the proof of Lemma 3.1.

Lemma B.9 *Under Condition 2,* $\bigcup_{j=1}^{J}\{\pi(y;\boldsymbol{\eta}_j), \partial_\mu \pi(y;\boldsymbol{\eta}_j), \partial_\gamma \pi(y;\boldsymbol{\eta}_j), \partial_\alpha \pi(y;\boldsymbol{\eta}_j)\}$ *is a set of linearly independent functions of* y *where* $\boldsymbol{\eta}_j = (\mu_j, \gamma_j, \alpha_j)'$, $j = 1, \ldots, J$.

Proof of Lemma B.8. Assume that conclusion of the lemma does not hold. Then, there exists nonzero vector $\mathbf{e} = ((\mathbf{c}_1', p_1, r_1), \ldots, (\mathbf{c}_J', p_J, r_J), \mathbf{d}_1', \ldots, \mathbf{d}_{J-1}')' \in \mathbb{R}^K$ where $\mathbf{c}_j \in \mathbb{R}^{S+1}$, $p_j, r_j \in \mathbb{R}$, $j = 1, \ldots, J$, $\mathbf{d}_j \in \mathbb{R}^{S+1}$, $j = 1, \ldots, J-1$, such that

$$\mathbf{e}' E[\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0)\partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0)]\mathbf{e} = 0,$$

which is equivalent to $\mathbf{e}'\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0) = 0$ with probability 1. Since $f(\cdot|\cdot;\cdot)$ is always positive, this translates to $\mathbf{e}'\partial_{\boldsymbol{\theta}} f(y_t|\mathbf{x}_{t-1};\boldsymbol{\theta}_0) = 0$ with probability 1. Let us write g_j for $g_j(\mathbf{x}_{t-1};\boldsymbol{\lambda}_0)$ and π_j for $\pi(y_t|\mathbf{x}_t;\boldsymbol{\zeta}_{0,j})$ where

$$\boldsymbol{\theta}_0 = ((\boldsymbol{\delta}_{0,1}', \gamma_{0,1}, \alpha_{0,1}), \ldots, (\boldsymbol{\delta}_{0,J}', \gamma_{0,J}, \alpha_{0,J}), \boldsymbol{\omega}_{0,1}', \ldots, \boldsymbol{\omega}_{0,J-1}') \equiv (\boldsymbol{\zeta}_{0,1}', \ldots, \boldsymbol{\zeta}_{0,J}', \boldsymbol{\lambda}_0')'.$$

Then, almost everywhere, we have $\sum_{j=1}^{J}\left(g_j[\mathbf{e}'\partial_{\boldsymbol{\theta}}\pi_j] + \pi_j[\mathbf{e}'\partial_{\boldsymbol{\theta}} g_j]\right) = 0$ or

$$\sum_{j=1}^{J} g_j\left[(\mathbf{c}_j'\mathbf{z}_{t-1})\partial_{\mu_{tj}}\pi_j + p_j\partial_{\gamma_j}\pi_j + r_j\partial_{\alpha_j}\pi_j\right] + \sum_{j=1}^{J}\sum_{k=1}^{J-1}[\mathbf{d}_k'\partial_{\boldsymbol{\omega}_k} g_j]\pi_j = 0$$

where $\mu_{tj} = \boldsymbol{\delta}_{0,j}'\mathbf{z}_{t-1}$ and $\mathbf{z}_{t-1} = (1, \mathbf{x}_{t-1}')'$. Since $\mathbf{x}_{t-1}$ has a positive density on $\mathbf{X}$ with respect to Lebesgue measure and the function $\pi(y;\mu,\gamma,\alpha)$ along with its derivatives with respect to μ, γ and α are continuous functions of the vector (y, μ), it follows that

$$\sum_{j=1}^{J}[g_j(\mathbf{x};\boldsymbol{\lambda}_0)(\mathbf{c}_j'\mathbf{z})]\partial_\mu \pi(y;\boldsymbol{\delta}_{0,j}'\mathbf{z}, \gamma_{0,j}, \alpha_{0,j}) + \sum_{j=1}^{J}\left[g_j(\mathbf{x};\boldsymbol{\lambda}_0)p_j\right]\partial_\gamma \pi(y;\boldsymbol{\delta}_{0,j}'\mathbf{z}, \gamma_{0,j}, \alpha_{0,j})$$

$$+\sum_{j=1}^{J}[g_j(\mathbf{x};\boldsymbol{\lambda}_0)r_j]\partial_\alpha \pi(y;\boldsymbol{\delta}_{0,j}'\mathbf{z}, \gamma_{0,j}, \alpha_{0,j})$$

$$+\sum_{j=1}^{J}\sum_{k=1}^{J-1}[\mathbf{d}_k'\partial_{\boldsymbol{\omega}_k} g_j(\mathbf{x};\boldsymbol{\lambda}_0)]\pi(y;\boldsymbol{\delta}_{0,j}'\mathbf{z}, \gamma_{0,j}, \alpha_{0,j}) = 0$$

for all $(y, \mathbf{x}) \in \mathbb{R} \times \mathbf{X}$ where $\mathbf{z} = (1, \mathbf{x}')'$. Thus, using Lemma B.9, we obtain that, for all $j = 1, \ldots, J$, $\mathbf{c}_j'\mathbf{z} = 0$ for all $\mathbf{x} \in \mathbf{X}$, $p_j = r_j = 0$ and $\sum_{k=1}^{J-1}\mathbf{d}_k'\partial_{\boldsymbol{\omega}_k} g_j(\mathbf{x};\boldsymbol{\lambda}_0) = 0$. Using

Lemma 2 in Jiang and Tanner (2000), we obtain from the last equality that $\mathbf{d}_k'\mathbf{z} = 0$ for all k and all $\mathbf{x} \in \mathbf{X}$. Since $\mathbf{X}$ is an open subset of $\mathbb{R}^S$, this implies $\mathbf{d}_k = \mathbf{0}_{S+1}, k = 1, \ldots, J-1$. Analogously we obtain $\mathbf{c}_j = \mathbf{0}_{S+1}, j = 1, \ldots, J$. Thus, we have the contradiction that $\mathbf{e}$ is the zero vector and so the proof is complete. ■

To complete the proof of Theorem 5.2, we need to show that assumption (d) in Lemma B.3 is satisfied with $\mathbf{B}(\boldsymbol{\theta}_0) = E[\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0) \partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0)]$. For this purpose, we will use an appropriate central limit theorem for multivariate dependent random sequences. Let $\boldsymbol{\kappa}$ be any nonzero vector in $\mathbb{R}^K$ and define $U_t = \boldsymbol{\kappa}' \partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0)$. Since we assume that the model is correctly specified and the true parameter value is $\boldsymbol{\theta}_0$, it follows by the standard likelihood argument that $E[\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0)|\mathbf{x}_{t-1}] = \mathbf{0}_K$ and $\mathbf{B}(\boldsymbol{\theta}_0) = -\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0)$. Let $\mathcal{F}_t = \sigma(y_s, \mathbf{x}_s : s \leq t)$ so that U_t is adapted to the filtration $\{\mathcal{F}_t\}$ and $E[U_t|\mathcal{F}_{t-1}] = 0$. Thus, $\{U_t, \mathcal{F}_t\}$ is a martingale difference sequence. Further, the first part of Condition 3 implies that U_t is stationary. The variance of U_t is $\sigma_{\mathrm{U}}^2 = \boldsymbol{\kappa}' \mathbf{B}(\boldsymbol{\theta}_0) \boldsymbol{\kappa}$. The finiteness and nonsingularity of $\mathbf{B}(\boldsymbol{\theta}_0)$, established in Lemma B.8, implies that $0 < \sigma_{\mathrm{U}}^2 < \infty$. Using the second part of Condition 3 we now obtain $\frac{1}{T} \sum_{t=1}^{T} U_t^2 \xrightarrow{P} \sigma_{\mathrm{U}}^2$ as $T \to \infty$. Moreover, using expressions (38), (39), (40) and Lemmas B.5(a) and B.6(a) along with Condition 5, we conclude that $E[\|\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0)\|^3] < \infty$ which, in turn, implies $E[|U_t|^3] < \infty$. Applying the central limit theorem for martingale differences (see Corollary 5.26 in White (2001)) then yields $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} U_t \xrightarrow{\mathcal{D}} N(0, \boldsymbol{\kappa}' \mathbf{B}(\boldsymbol{\theta}_0) \boldsymbol{\kappa})$. According to the Cramer-Wold theorem (see Theorem 25.5 in Davidson (1994)), we then have that assumption (d) in Lemma B.3 is satisfied with $\mathbf{B}(\boldsymbol{\theta}_0) = E[\partial_{\boldsymbol{\theta}} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0) \partial_{\boldsymbol{\theta}'} \log f(y_t|\mathbf{x}_{t-1}; \boldsymbol{\theta}_0)]$. Thus, the proof of Theorem 5.2 follows from applying Lemma B.3 and using the fact $\mathbf{B}(\boldsymbol{\theta}_0) = -\mathbf{A}(\boldsymbol{\theta}_0) = \mathbf{I}(\boldsymbol{\theta}_0)$.

■

References

Akaike, H., 1973, Likelihood of a model and information criteria, *2nd International Symposium on Information Theory*, 267-281.

Akaike, H., 1981, Information theory and the extension of the maximum likelihood principle. *Journal of Econometrics*, 16 3-14.

Amemiya, T., 1985, *Advanced Econometrics*, Harvard University Press.

Berkowitz, J., 2000, Testing density forecasts, with applications to risk management, *Technical report*, University of California, Irvine, Graduate School of Management.

Carvalho, A., 2002, Mixtures-of-experts of generalized linear time series, *Ph.D. Thesis*, Department of Statistics, Northwestern University.

Carvalho, A. and M. Tanner, 2003, Hypothesis Testing in Mixtures-of-Experts of Generalized Linear Time Series, *Forthcoming in the Proceedings of the 2003 International Conference on Computational Intelligence for Financial Engineering*, Hong Kong.

Carvalho, A. and M. Tanner, 2002b, Mixtures-of-experts of generalized linear time series models, *Technical report*, University of British Columbia, Department of Statistics.

Carvalho, A. and M. Tanner, 2002c, Mixtures-of-experts of generalized linear time series: asymptotic normality and model specification, *Technical report*, University of British Columbia, Department of Statistics.

Davidson, J., 1994, *Stochastic limit theory - An introduction to econometricians*, Oxford University Press.

Dempster, A., Laird, N., and D. Rubin, 1977, Maximum likelihood from incomplete data via de EM algorithm, *Journal of the Royal Statistical Society B*, 39, 1-38.

Diebold, F., Gunther, T. and A. Tay, 1998, Evaluating density forecasts, with applications to financial risk management, *International Economic Review*, 39, 863-883.

Franses, P. and D. Dijk, 2000, *Non-linear Time Series Models in Empirical Finance*, Cambridge University Press.

Giacomini, R., Gottschling, A., Haefke, C. and H. White, 2002, Hypernormal densities, *Technical Report*, University of California, San Diego, Department of Economics.

Gradshteyn, I.S. and I.M. Ryzhik, 1980, *Tables of Integrals, Series, and Products*, Academic Press.

Gutta, S., J. Huang, P. Jonathon and H. Wechsler, 2000, Mixture of experts for classification of gender, ethnic origin, and pose of human faces, *IEEE Transactions on Neural Networks* 11 948–960.

Hamilton, J., 1994, *Time Series Analysis,* Princeton University Press.

Harvey, A., 1994, *Forecasting, Structural Time Series Models and the Kalman Filter*, Cambridge University Press.

Hong, Y., 2000, Generalized spectral tests for serial dependence, *Journal of Royal Statistical*

Society Series B, 62, 557-574.

Hong, Y., 2002, Evaluation of out-of-sample probability density forecasts, *Technical Report*, Cornell University, Department of Economics.

Horn, R. A., and Johnson C.R., 1990, *Matrix analysis*, Cambridge University Press, New York.

Huerta, G., W. Jiang and M. Tanner, 2001, Mixtures of time series models, *Journal of Computational and Graphical Statistics* 10 82–89.

Jacobs, R., M. Jordan, S. Nowlan and G. Hinton, 1991, Adaptive mixtures of local experts, *Neural Computation* 3, 79-87.

Jeffries, N. and Pfeiffer, R. 2001, A mixture model for the probability distribution of rain rate, *Environmetrics* 12, 1-10.

Jordan, M. and Jacobs, R. 1994, Hierarchical mixtures-of-experts and the EM algorithm, *Neural Computation* 6 181–214.

Jiang, W. and Tanner, M. 1998, Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation, *Technical Report*, Department of Statistics, Northwestern University.

Jiang, W. and Tanner, M. 1999a, Hierarchical mixtures-of-experts for exponential family regression models: approximation and maximum likelihood estimation, *Annals of Statistics* 27 987–1011.

Jiang, W. and Tanner, M. 1999b, On the approximation rate of hierarchical mixtures-of-experts for generalized linear models, *Neural Computation* 11 1183–1198.

Jiang, W. and Tanner, M. 1999c, On the identifiability of mixtures-of-experts, *Neural Networks* 12 1253–1258.

Jiang, W. and Tanner, M. 2000, On the asymptotic normality of hierarchical mixtures-of-experts for generalized linear models, *IEEE Transactions on Information Theory* 46 1005–1013.

Kurnik, R., Oliver, J., Waterhouse, S., Dunn, T., Jayalaksmi, Y., Lesho, M. Lopatin, M., Tamada, J. and Wei, C. 1999, Application of mixture of experts algorithm for signal processing in a noninvasive glucose monitoring system, *Sensors and Actuators B-Chemical* 60 19–26.

Liehr, S., Pawelzik, K., Kohlmorgen, J. and Muller, K. 1999, Hidden Markov mixtures of experts with an application to EEG recording from sleep, *Theory of Biosciences* 118 246–260.

Lin, H., McCulloch, C., Turnbull, B., Slate, E. and Clark, L. 2000, A latent class mixed model for analysing biomarker trajectories with irregularly scheduled observations *Statistics in Medicine* 19 1303-1318.

McDonald, J.B. and W.K. Newey, 1988, Partially adaptive estimation of regression models via the generalized t distribution, *Econometric Theory* 4, 428-457.

McDonald, J.B., 1996, Probability distributions for financial models, *Handbook of Statistics*, Vol. 14, Elsevier, 427-461.

Meyn, S. P. and R. L. Tweedie, 1993, *Markov chains and stochastic stability,* Springer-Verlag.

Newey, W., 1991, Uniform convergence in probability and stochastic equicontinuity, *Econometrica* 59 1161–1167.

Newey, W. K., and K. D. West, 1987, A simple positive semi-definite heteroscedasticity and autocorrelation consistent covariance matrix, *Econometrica* 55, 703-708.

Quinn, B., McLachlan, G., and N. Hjort, 1987, A note on the Aitkin-Rubin approach to hypothesis testing in mixture models. *Journal of the Royal Statistical Society B*, 39 311-314.

Rider, P. R., 1958, Generalized Cauchy distributions, *Annals of the Institute of Statistical Mathematics* 9, 215-223.

Rosenblatt, M., 1952, Remarks on a multivariate transformation, *Annals of Mathematical Statistics*, 23 470-472.

Rosen, O. and Tanner, M. 1999, Mixtures of proportional hazard regression models, *Statistics in Medicine* 18 1119–1131.

Schwarz, C., 1978, Estimating the dimension of a model, *Annals of Statistics*, 6 461-464.

Tanner, M., 1996, *Tools for Statistical Inference*, Springer.

Tjøstheim, D., 1990, Non-linear time series and Markov chains, *Advances in Applied Probability* 22, 587-611.

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

Weigend, A., Mangeas, M. and Srivastava, A. 1995, Nonlinear gated experts for time series: discovering regimes and avoiding overfitting, *International Journal of Neural Systems* 6 373–399.

White, H., 2001, Asymptotic theory for econometricians, Academic Press.

Wong, W. 1986, Theory of partial likelihood, *Annals of Statistics*, 14, 88–123.

Wong, C. and Li, W. 2001, On a logistic mixture autoregressive model, *Biometrika* 88 833–846.

Wood, S., Jiang, W. and M. Tanner, 2001, Bayesian mixture of splines for spatially adaptive nonparametric regression, *Technical Report*, Northwestern University, Department of Statistics.

Zeevi, M., Meir, R. and Adler, R. J. 2001, Nonlinear models for time series using mixtures of autoregressive models, working paper.