An autoregressive model based on the generalized hyperbolic distribution

Henri Karttunen

University of Helsinki, Finland
Department of Mathematics and Statistics

Abstract We define a non-linear autoregressive time series model based on the generalized hyperbolic distribution in an attempt to model time series with non-Gaussian features such as skewness and heavy tails. We show that the resulting process has a simple condition for stationarity and it is also ergodic. An empirical example with a forecasting experiment is presented to illustrate the features of the proposed model.

Key words: *Autoregressive model, conditional heteroscedasticity, generalized hyperbolic distribution, non-linear time series, skewness.*

1 Introduction

The generalized hyperbolic distribution is a normal mean-variance mixture distribution introduced by Barndorff-Nielsen (1977) and further studied by, for example, Blæsild (1981). It has versatile control over the shape of its density function through its parameters and is, therefore, an attractive choice for an error distribution of an autoregressive time series model when modelling the distribution of the data is of interest.

In this paper a univariate autoregressive process based on the generalized hyperbolic distribution is proposed. Previously, the generalized hyperbolic distribution has been used for applications in economics and finance by McNeil et al. (2005), Prause (1999), Hu (2005) and others. Besides being used in its general form, the generalized hyperbolic distribution has many well-known special cases such as the normal inverse Gaussian (NIG) distribution which was introduced by Barndorff-Nielsen (1997) who used it for applications in turbulence and finance. Other special and limiting cases include distributions such as the Student's t distribution and the hyperbolic distribution.

The generalized hyperbolic distribution could be used as the conditional distribution of a conventional autoregressive (AR) process by defining the distribution of the independent and identically distributed errors as a generalized hyperbolic distribution instead of a Gaussian distribution. This approach would, however, only affect the shape of the conditional density function of the process and is thus not considered in this paper.

Instead, we use the same approach as Meitz et al. (2018) in the case of a mixture autoregression based on the (symmetric) Student's t-distribution. Following their approach we define a process we call the generalized hyperbolic autoregressive (GHAR) process by employing the properties of the multivariate generalized hyperbolic distribution in conjunction with basic Markov chain theory. A distinctive property of the GHAR process is that its error terms are not independent but a conditionally heteroscedastic martingale difference sequence.

For many non-linear and non-Gaussian processes the stationarity conditions and the possible stationary distributions are troublesome if not impossible to derive without imposing excessive restrictions on the parameters of the process (see, e.g., Teräsvirta et al. 2010). For the GHAR process this is not the case because its stationarity is a straightforward consequence of the properties of the multivariate generalized hyperbolic distribution and the stationarity condition

of an auxiliary AR process used to define the GHAR process. Therefore, the GHAR process has a simple explicit condition for stationarity and, moreover, for a pth order GHAR process the stationary distributions up to dimension $p + 1$ follow multivariate generalized hyperbolic distributions and are fully known.

This paper is organized as follows. The multivariate generalized hyperbolic distribution is discussed briefly in Section 2 along with suitable parameter restrictions to avoid any identifiability issues. The density function of the univariate generalized hyperbolic distribution is also plotted with selected values of the parameters to illustrate the role of different parameters.

In Section 3.1 we define the GHAR process by using the theory of Markov chains and discuss its stationarity and ergodicity. The connection between the GHAR process and conventional linear AR processes is highlighted in Section 3.2 by expressing the GHAR process in the form of an equation which involves the conditional expectation and a martingale difference error term. Here we also provide more details of the properties of the GHAR process. In Section 3.3 we discuss the maximum likelihood (ML) estimation of the GHAR process and show that the ML estimator is strongly consistent.

An empirical example on a particle matter measurement data is presented in Section 4. The results show that the GHAR process performs well in comparison to the DAR model of Ling (2007) and the Gaussian AR process in terms of both in-sample fit and out-of-sample forecasting performance. Finally, conclusions are provided in Section 5, and proofs and other technical material are presented in appendices.

2   Multivariate generalized hyperbolic distribution

The density function of the generalized hyperbolic distribution has a general representation without any parameter restrictions and it is given, for example, in McNeil et al. (2005, p. 78). In the general representation there is, however, an identification problem which can be addressed by imposing suitable parameter restrictions. Identification can be achieved in several ways and, for example, McNeil et al. (2005, p. 79) restrict the determinant of the dispersion matrix to one.

Rather than restricting the dispersion matrix, which for our purposes will be defined using covariance matrices of a Gaussian AR process, the restrictions used by Browne & McNicholas (2015) are adopted. Thus, the restricted density function of the d dimensional generalized hyperbolic distribution we employ has the expression

$$
\mathsf{gh}_d(y; \lambda, \delta, \mu, \Gamma, \gamma) = c \frac{K_{\lambda - \frac{d}{2}}\left(\sqrt{(\delta + (y - \mu)'\Gamma^{-1}(y - \mu))(\delta + \gamma'\Gamma^{-1}\gamma)}\right) e^{\gamma'\Gamma^{-1}(y-\mu)}}{\left(\sqrt{(\delta + (y - \mu)'\Gamma^{-1}(y - \mu))(\delta + \gamma'\Gamma^{-1}\gamma)}\right)^{\frac{d}{2} - \lambda}}, \quad (1)
$$

where $y \in \mathbb{R}^d$ and K_λ denotes the (modified) Bessel function (of the third kind) with order λ and

$$
c = \frac{(\delta + \gamma'\Gamma^{-1}\gamma)^{\frac{d}{2} - \lambda}}{(2\pi)^{\frac{d}{2}}\sqrt{\det(\Gamma)}K_\lambda(\delta)}.
$$

The parameters μ and γ are real vectors of dimension d, λ is real, $\delta > 0$ and Γ is a positive definite $(d \times d)$ dimensional matrix. Clearly, the density function is univariate if $d = 1$ is selected, in which case we use the notation $\Gamma = \sigma^2$.

For a detailed discussion of the Bessel functions and their properties, see, for example, Watson (1944). The function K_λ is hereafter referred to as *the Bessel function* as there is no risk of confusion.

The expected value of a random variable y following the d dimensional generalized hyperbolic distribution is (see, e.g., McNeil et al. 2005, p. 77)

$$
E(y) = \mu + \frac{K_{\lambda+1}(\delta)}{K_\lambda(\delta)}\gamma \tag{2}
$$

and its covariance matrix is

$$\mathsf{Cov}(y) = \left(\frac{K_{\lambda+2}\left(\delta\right)}{K_\lambda\left(\delta\right)} - \left(\frac{K_{\lambda+1}\left(\delta\right)}{K_\lambda\left(\delta\right)} \right)^2 \right) \gamma\gamma' + \frac{K_{\lambda+1}\left(\delta\right)}{K_\lambda\left(\delta\right)}\Gamma. \tag{3}$$

The one dimensional density function gh_1 is plotted in Appendix A.7.1 to demonstrate how different parameter values affect the shape of the distribution. In addition to controlling the variance, the parameter σ^2 along with λ and δ affect the kurtosis and the heaviness of the tails. Also, changing the value of σ^2 has no obvious effect on the mode of the distribution whereas changing the value of λ and δ has. Setting $\gamma = 0$ leads to a symmetric distribution while non-zero values of γ produce a skewed distribution. The parameter μ only shifts the density to left or right.

3 The GHAR process

3.1 Definition of the GHAR process

In this section we construct an autoregressive process of order $p \geq 1$ which has a generalized hyperbolic distribution as its conditional distribution. We show that a condition similar to that in a conventional Gaussian AR(p) process guarantees the stationarity of the process and that stationary distributions of consecutive observations up to dimension $p + 1$ are multivariate generalized hyperbolic distributions. Throughout the rest of this paper we use the following simplified notations: $\boldsymbol{y}_t = (y_t, \ldots, y_{t-p+1})$ and $\boldsymbol{y}_t^+ = (y_t, \boldsymbol{y}_{t-1}) = (\boldsymbol{y}_t, y_{t-p})$. Note that the dimension of both $\boldsymbol{y}_t$ and $\boldsymbol{y}_t^+$ depend on the chosen order p even though it is not explicitly expressed in the notation.

It is not immediately obvious that an autoregressive process with properties discussed above exists. However, suppose that $\{y_t\}$, $t \geq 1$ is such a process. Then the (p dimensional) random vector $\boldsymbol{y}_t$, for example, must follow a generalized hyperbolic distribution and by stationarity its covariance matrix, given by equation (3), is a Toeplitz matrix. This implies that, in the definition of the GHAR process below, restrictions have to be imposed on the parameters of the generalized hyperbolic distribution for the properties stated above to hold. Parameter restrictions related to the covariance matrix Γ of $\boldsymbol{y}_t$ in equation (3) will be formulated using an auxiliary linear AR(p) process denoted by $\{z_t\}$.

Thus, let $\{z_t\}$ be a conventional (zero mean) stationary linear AR(p) process so that, for each t, the random variables z_t have the representation

$$z_t = \sum_{i=1}^p \phi_i z_{t-i} + \eta_t, \tag{4}$$

where $\{\eta_t\}$ is a sequence of independent zero mean Gaussian random variables with finite variance $\sigma^2 > 0$. As indicated above, we also assume that the autoregressive coefficients ϕ_i, $i = 1, \ldots, p$, satisfy the stationarity condition

$$1 - \phi_1 z - \ldots - \phi_p z^p \neq 0, \quad |z| \leq 1, \quad z \in \mathbb{C}. \tag{5}$$

Related to the process $\{z_t\}$ we define the vectors $\boldsymbol{z}_{t-1} = (z_{t-1}, z_{t-2}, \ldots, z_{t-p})$ and $\boldsymbol{z}_t^+ = (z_t, \boldsymbol{z}_{t-1})$ and denote the covariance matrices of $\boldsymbol{z}_t^+$ and $\boldsymbol{z}_{t-1}$ as Γ_{p+1} and Γ_p respectively. Because $\{z_t\}$ is a stationary AR(p) process both Γ_{p+1} and Γ_p are Toeplitz matrices and depend only on the variance parameter σ^2 and the coefficient vector $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)$ (see, e.g., Lütkepohl, 2006, p. 28–29). Thus, the parameter vector $(\boldsymbol{\phi}, \sigma^2)$ of the AR(p) process will be used, along with other parameters, to construct the GHAR process.

Choosing $d = 1$ in (1) we introduce the conditional density function

$$\mathsf{gh}_1(y_t; \tilde{\lambda}, \tilde{\delta}_t, \tilde{\mu}_t, \tilde{\sigma}_t^2, \tilde{\gamma}_t) = \mathsf{gh}_1(y_t; \tilde{\theta}_t) \tag{6}$$

where the components of the time-varying parameter vector $\tilde{\theta}_t = (\tilde{\lambda}, \tilde{\delta}_t, \tilde{\mu}_t, \tilde{\sigma}_t^2, \tilde{\gamma}_t)$ depend on the vector $\boldsymbol{y}_{t-1}$ as well as time-invariant parameter vector $\theta = (\boldsymbol{\phi}, \lambda, \delta, \mu, \sigma^2, \gamma)$. Here the parameters $\boldsymbol{\phi}$ and σ^2 are as defined above so that the components of $\boldsymbol{\phi}$ satisfy the stationarity condition (5), and the remaining parameters satisfy $\lambda, \mu, \gamma \in \mathbb{R}$ and $\delta > 0$. The components of $\tilde{\theta}_t$ are defined as

- $\tilde{\lambda} = \lambda - p/2$,

- $\tilde{\delta}_t = \sqrt{\left(\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)' \Gamma_p^{-1} (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)\right) \left(\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p\right)}$,

- $\tilde{\mu}_t = \mu + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)' \boldsymbol{\phi}$,

- $\tilde{\sigma}_t^2 = \dfrac{\tilde{\delta}_t}{\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p} \sigma^2$,

- $\tilde{\gamma}_t = \dfrac{\tilde{\delta}_t}{\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p} \left(\gamma - \boldsymbol{\phi}' \boldsymbol{\gamma}_p\right)$.

The covariance matrix Γ_p used above is constructed by using parameters $\boldsymbol{\phi}$ and σ^2 as for the auxiliary AR process $\{z_t\}$ and, letting $\mathbf{1}_p$ denote the p dimensional vector of ones, we have set $\boldsymbol{\mu}_p = \mu \mathbf{1}_p$ and $\boldsymbol{\gamma}_p = \gamma \mathbf{1}_p$.

Denoting by $\mathcal{F}_{t-1}$ the sigma algebra generated by the set $\{y_{t-i}, i \geq 1\}$, $t \geq 1$, the density function can be viewed as the conditional density function of y_t given $\mathcal{F}_{t-1}$ or, more precisely, given the vector $\boldsymbol{y}_{t-1}$. For any fixed value of $\boldsymbol{y}_{t-1}$, $\mathsf{gh}_1(y_t; \tilde{\theta}_t)$ is the density function of a generalized hyperbolic distribution so that it defines a transition probability measure of a pth order Markov chain on $\mathbb{R}$, which is equivalent to viewing the process $\{\boldsymbol{y}_t\}$ as a Markov chain on $\mathbb{R}^p$.

Definition 1. *We define the GHAR process $\{y_t\}$, $t \geq 1$, as a pth order Markov chain on $\mathbb{R}$ whose transition probability measure is determined by the density function (6) of a univariate generalized hyperbolic distribution.*

The motivation for the particular definition of the parameter vector $\tilde{\theta}_t$ is that it implies stationarity of the process $\{y_t\}$. This is a straightforward consequence of choosing a $p+1$ dimensional generalized hyperbolic distribution with parameter vector $(\lambda, \delta, \boldsymbol{\mu}_{p+1}, \Gamma_{p+1}, \boldsymbol{\gamma}_{p+1})$ for $(y_t, \ldots, y_{t-p})$ and then conditioning y_t on $(y_{t-1}, \ldots, y_{t-p})$ (here $\boldsymbol{\mu}_{p+1}$ and $\boldsymbol{\Gamma}_{p+1}$ are $p+1$ dimensional analogs of $\boldsymbol{\mu}_p$ and $\boldsymbol{\Gamma}_p$ respectively). For more detailed derivation of the parameters, see Appendix A.6. Now we can state the stationarity result anticipated in the preceding discussion. The proof is given in Appendix A.1.

Theorem 1. *Let $\{y_t\}$ be a GHAR process with parameter vector θ as defined after equation (6) with $\boldsymbol{\phi}$ satisfying the stationarity condition (5). Assume that the initial value $\boldsymbol{y}_0$ follows the generalized hyperbolic distribution such that $\boldsymbol{y}_0 \sim \mathsf{gh}_p(\lambda, \delta, \boldsymbol{\mu}_p, \Gamma_{p+1}, \boldsymbol{\gamma}_p)$, with the parameters defined above. Then, the process $\{y_t\}$ is (strictly) stationary.*

Note that, if the skewness parameter γ equals zero the value of the time-varying parameter $\tilde{\gamma}_t$, defined above, is equal to zero as well. Then the marginal distributions and the conditional distribution are both symmetric and the expressions for the time-varying parameters of the conditional distribution (6) are simplified considerably.

If stationary initial values are assumed it is clear by stationarity that, for every $t \geq 1$, we have $\boldsymbol{y}_t \sim \mathsf{gh}_p(\lambda, \delta, \boldsymbol{\mu}_p, \Gamma_p, \boldsymbol{\gamma}_p)$ and, by equation (A2) in Appendix A.1.1, it holds for the random vector $\boldsymbol{y}_t^+ = (y_t, \boldsymbol{y}_{t-1})$ that $\boldsymbol{y}_t^+ \sim \mathsf{gh}_{p+1}(\lambda, \delta, \boldsymbol{\mu}_{p+1}, \Gamma_{p+1}, \boldsymbol{\gamma}_{p+1})$. From Proposition 3.13 of McNeil et al. (2005) it therefore follows that all $1 \leq s \leq p+1$ dimensional marginal distributions of the random vectors $\boldsymbol{y}_t^+$ follow the $\mathsf{gh}_s(\lambda, \delta, \boldsymbol{\mu}_s, \Gamma_s, \boldsymbol{\gamma}_s)$ distribution with Γ_s being equal to the $(s \times s)$ dimensional upper-left corner of Γ_{p+1}. Moreover, as all components of $\boldsymbol{\mu}_{p+1}$ are equal, and the

same applies to $\boldsymbol{\gamma}_{p+1}$, their lower-dimensional counterparts have the representations $\boldsymbol{\mu}_s = \mu \mathbf{1}_s$ and $\boldsymbol{\gamma}_s = \gamma \mathbf{1}_s$.

On the other hand, if the initial value $\boldsymbol{y}_0$ does not follow the generalized hyperbolic distribution, the process $\{y\}$ is not stationary. However, even if $\{y_t\}$ is not stationary, we have the following stability result proved in Appendix A.1.2.

Theorem 2. *The Markov chain $\{\boldsymbol{y}_t\}$ is ergodic.*

This ergodicity in the sense of Meyn & Tweedie (2009, Chapter 13) means that, starting from any fixed initial value $\boldsymbol{y}_0$, the n-step transition probability measure of the GHAR process tends to its stationary distribution in total variation norm as n $\to \infty$ (see the proof of Theorem 2). As an implication of this, the usual asymptotic results, such as the strong law of large numbers, apply.

From a practical point of view it is convenient that the stationary distribution of the GHAR process is known. The family of generalized hyperbolic distributions contains many well known special cases (see, e.g., McNeil et al. 2005, p. 80) and either the conditional distribution or the stationary distribution can be set to some of these special cases by restricting the values of the parameter λ. Some well known special cases with their stationary and conditional distributions are presented in Table 1. Limiting cases (as limits of one or more parameters) include such distributions as Student's t distribution and the variance-gamma distribution.

Table 1: Different choices of λ and the conditional and marginal (stationary) distributions (see, e.g., McNeil et al. 2005, p. 80)

λ	$\tilde{\lambda}$	p-dimensional stationary	Conditional
$-1/2$	$-(p+1)/2$	Normal inverse Gaussian	Generalized hyperbolic
1	$-(p-2)/2$	Hyperbolic	Generalized hyperbolic
$(p-1)/2$	$-1/2$	Generalized hyperbolic	Normal inverse Gaussian
$(p+2)/2$	1	Generalized hyperbolic	Hyperbolic

3.2 Autoregression

To demonstrate how the GHAR process, defined in the previous section, is related to conventional AR processes we next show that the GHAR process also has a representation where, for each $t \geq 1$, the random variable y_t is expressed as a function of the random vector $\boldsymbol{y}_{t-1}$ and a martingale difference error term. To this end, we first introduce some notation.

Substituting the parameters of the conditional distribution of y_t given in (6) for their counterparts in equations (2) and (3) results in the following expressions for the conditional mean and conditional variance of y_t:

$$\mu_t := \mathsf{E}(y_t | \boldsymbol{y}_{t-1}) = \tilde{\mu}_t + R_1(\boldsymbol{y}_{t-1})\tilde{\gamma}_t \tag{7}$$

and

$$\sigma_t^2 := \mathsf{Var}(y_t | \boldsymbol{y}_{t-1}) = R_1(\boldsymbol{y}_{t-1})\tilde{\sigma}_t^2(\theta) + \left(R_2(\boldsymbol{y}_{t-1}) - \left(R_1(\boldsymbol{y}_{t-1}) \right)^2 \right) \tilde{\gamma}_t^2, \tag{8}$$

where $R_k(\boldsymbol{y}_t) = K_{\tilde{\lambda}+k}(\tilde{\delta}_t)/K_{\tilde{\lambda}}(\tilde{\delta}_t)$ is a ratio of two Bessel functions and other notations are as in (6). As $\boldsymbol{y}_t$ is a Markov chain we have $\mathsf{E}(y_t | \mathcal{F}_{t-1}) = \mathsf{E}(y_t | \boldsymbol{y}_{t-1})$ and similarly for the conditional variance. Note also that, although there is no general solution available, it is possible to calculate the first $p+1$ autocorrelations using the covariance matrix Γ_{p+1} and Equation (3).

With μ_t and σ_t^2 as defined above we now define the sequence $\{\varepsilon_t\}$ of error terms as

$$\varepsilon_t := \frac{y_t - \mu_t}{\sigma_t}, \quad t \geq 1, \tag{9}$$

5

which, by solving for y_t, yields the representation

$$y_t = \mu_t + \sigma_t \varepsilon_t. \tag{10}$$

The sequence $\{\varepsilon_t\}$ is not independent but we show in Appendix A.3 that the sequence $\{\varepsilon_t, \mathcal{F}_{t-1}\}$ is a well-defined martingale difference sequence with unit variance. Moreover, the conditional distribution of ε_t given $\mathcal{F}_{t-1}$ only depends on $\boldsymbol{y}_{t-1}$ and, by Proposition 3.13 of McNeil et al. (2005), has the expression

$$\varepsilon_t | \mathcal{F}_{t-1} \sim \mathsf{gh}_1 \left(\tilde{\lambda}, \, \tilde{\delta}_t, \, \frac{\tilde{\mu}_t - \mu_t}{\sigma_t}, \, \frac{\tilde{\sigma}_t^2}{\sigma_t^2}, \, \frac{\tilde{\gamma}_t}{\sigma_t} \right), \tag{11}$$

where the dependence of the parameters on $\boldsymbol{y}_{t-1}$ is obvious by the definition of $\tilde{\theta}_t$ after the equation (6).

The definitions of the terms on the right-hand side of (10) indicate that the GHAR process can be viewed as a non-linear and conditionally heteroscedastic autoregression. In order to illustrate this issue in more detail it appears convenient to first discuss the symmetric case where the parameter γ is restricted to zero. In this special case the expressions of the conditional mean μ_t and conditional variance σ_t^2 are considerably simplified because, as noted in the previous section, then also $\tilde{\gamma}_t = 0$ for every $t \geq 1$.

Thus, suppose that $\gamma = 0$. Using the definition of $\tilde{\mu}_t$ (see the equations following (6)) and the notation $\phi_0 = \mu \left(1 - \boldsymbol{\phi}' \mathbf{1}_p \right)$ we can write (10) as

$$y_t = \phi_0 + \boldsymbol{\phi}' \boldsymbol{y}_{t-1} + \sigma_t \varepsilon_t = \phi_0 + \sum_{j=1}^{p} \phi_j y_{t-j} + \sigma_t \varepsilon_t. \tag{12}$$

This equation and the martingale difference property of ε_t together imply that the conditional expectation of y_t is a linear function of its p lags whereas the conditional variance, see (8), has the expression

$$\sigma_t^2 = \sigma^2 R_1(\boldsymbol{y}_{t-1}) \sqrt{\frac{\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)' \Gamma_p^{-1} (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)}{\delta}}. \tag{13}$$

Thus, even in the symmetric case the conditional variance of y_t is time-variant depending on p lags of y_t. Note also, that the magnitude of the factor $R_1(\boldsymbol{y}_{t-1})$ on the right-hand side of (13) can be controlled via the shape parameter $\tilde{\lambda} = \lambda - p/2$. By the properties of the Bessel function, $R_1(\cdot) > 1$ for all $\tilde{\lambda} > -1/2$ and $0 < R_1(\cdot) < 1$ for $\tilde{\lambda} < -1/2$. Choosing $\tilde{\lambda} = -1/2$, which makes the conditional distribution normal inverse Gaussian (see Table 1), implies $R_1(\cdot) = 1$ so that the non-linearity in the conditional variance simplifies to $\tilde{\sigma}_t^2$ (see the definition following (6) and, for the properties of the Bessel function, see, e.g., Watson (1944)).

While in the symmetric case the conditional mean is constant when all components of ϕ are set to zero, equation (13) shows that the conditional variance σ_t^2 cannot be made constant by restricting parameters. Also, if we restrict the autoregressive coefficients to zero, the covariance matrix Γ_p becomes the scalar matrix $\sigma^2 I_p$ with I_p being a $p \times p$ identity matrix. This implies that the error sequence $\{\varepsilon_t\}$ and hence the process $\{y_t\}$ are, even in this case, conditionally heteroscedastic. Thus, in the symmetric case, the conditional variance is the major difference between a conventional linear AR process and GHAR process. For a linear AR process the conditional variance is always a constant by definition, but the above discussion shows that for the GHAR process it is always time-variant.

Now, consider the skewed case where the value of the parameter γ is non-zero. Then even the conditional mean is non-linear and the skewness also has an effect on the non-linearity of the conditional variance. Inserting the definition of μ_t (see the equations following (6)) into the representation (10) yields

$$y_t = \tilde{\mu}_t + \sigma_t \varepsilon_t + R_1(\boldsymbol{y}_{t-1})\tilde{\gamma}_t$$

$$= \phi_0 + \sum_{j=1}^{p} \phi_j y_{t-j} + \sigma_t \varepsilon_t + R_1(\boldsymbol{y}_{t-1}) \frac{\sqrt{\left(\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)' \Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)\right)}}{\sqrt{\left(\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p\right)}} \left(\gamma - \boldsymbol{\phi}' \boldsymbol{\gamma}_p\right). \quad (14)$$

Equation (14) shows that in the asymmetric case even the conditional expectation cannot be made constant by means of parameter restrictions. In fact, if $\gamma \neq 0$, setting the components of $\boldsymbol{\phi}$ to zero may still result in very strong autocorrelation. To illustrate this fact, consider the case $p = 1$ where $\phi_1 = \mu = 0$ and $\gamma > 0$. For simplicity, suppose that the conditional distribution is normal inverse Gaussian so that $\tilde{\lambda} = -1/2$ and hence $R_1(\boldsymbol{y}_t) = 1$. Then we also have $\mu_t = \tilde{\gamma}_t$ and the definition of $\tilde{\gamma}_t$ shows that $\tilde{\gamma}_t \to |y_t|$ as $\delta \to 0$. Hence, for any small enough δ we have $\mu_t \approx |y_t|$ which implies that the process is strongly autocorrelated.

In the skewed case the expression of the conditional variance given in (8) is rather complicated, and we only note that simplifications occur when the conditional distribution is normal inverse Gaussian because then $\tilde{\lambda} = -1/2$ and, in addition to $R_1(\boldsymbol{y}_{t-1}) = 1$, an explicit expression for $R_2(\boldsymbol{y}_{t-1})$ is also available (see, for example, Watson (1944, p. 80, Eqs. (12) and (13))).

Different values of γ control the non-linearity of the conditional expectation, where small absolute values of γ lead to almost linear conditional expectation. On the other hand, the conditional expectation can even be made everywhere positive (or negative) by choosing γ to be big (or small) enough, depending on the value of ϕ. See Figure 4 in Appendix A.7.2 where both the conditional mean and conditional variance are plotted as a function of y_{t-1} with different values of γ and ϕ_1.

3.3 Likelihood function

Assuming that the observed data $(y_{-p+1}, \ldots, y_0, y_1, \ldots, y_n)$ is generated by a GHAR process, we next derive an expression for the log-likelihood function and show that the ML estimator is strongly consistent. We define the parameter space Θ as a subset of the set of all vectors $(\boldsymbol{\phi}, \lambda, \delta, \mu, \sigma^2, \gamma) \in \mathbb{R}^{p+5}$ where $\delta, \sigma^2 > 0$ and the vector $\boldsymbol{\phi}$ satisfies the stationarity condition (5). As the log-likelihood function is rather complicated, its analysis in detail is difficult and, therefore, we choose to assume that the parameter space Θ is compact and the true parameter value $\theta_0 \in \Theta$. In practice, however, the parameter space needs to be searched widely for an optimal solution.

Denote the likelihood function of the single observation y_t by

$$L_t(\theta) = \mathsf{gh}_1(y_t; \tilde{\theta}_t),$$

where the parameter vector $\tilde{\theta}_t$ is defined in equation (6) and the subsequent discussion. With the assumption that the initial values satisfy $\boldsymbol{y}_0 \sim \mathsf{gh}_p(\lambda, \delta, \mu \mathbf{1}_p, \Gamma_p, \gamma \mathbf{1}_p)$, the exact likelihood function can now be expressed as

$$L(\theta) = \mathsf{gh}_p(\boldsymbol{y}_0; \lambda, \delta, \mu \mathbf{1}_p, \Gamma_p, \gamma \mathbf{1}_p) \prod_{t=1}^{n} L_t(\theta),$$

where the effect of the initial values vanishes as the sample size n tends to infinity. Thus, if the assumption of stationary initial values is questionable, conditioning on the initial values leads to the conditional likelihood function

$$L^c(\theta) = \prod_{t=1}^{n} L_t(\theta). \quad (15)$$

Applying the logarithm transformation on both sides we get the conditional log-likelihood function

$$\ell^c(\theta) = \sum_{t=1}^{n} \log\left(L_t(\theta)\right) := \sum_{t=1}^{n} \ell_t(\theta). \tag{16}$$

The log-likelihood function could be expressed just as the logarithm of the density function of the generalized hyperbolic distribution with parameters being functions of $\boldsymbol{y}_{t-1}$. However, as the resulting expressions are rather complicated, we write the conditional log-likelihood function in another form, which is more suitable for our purposes.

First, we introduce the following shorthand notations: $\xi_{t-1,1} = \delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)' \Gamma_p^{-1} (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)$, $\xi_{t,2} = (y_t - \phi_0 - \boldsymbol{\phi}' \boldsymbol{y}_{t-1})/\sigma$, $\zeta_1 = \delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p$ and $\zeta_2 = \left(\gamma - \boldsymbol{\phi}' \boldsymbol{\gamma}_p\right)/\sigma$. Using these notations we show in Appendix A.2 that the conditional log-likelihood function of a single observation has the expression

$$\begin{aligned}
\ell_t(\theta) = &-\frac{1}{2} \ln\left(2\pi\sigma^2\right) + \xi_{t,2}\zeta_2 \\
&- \tilde{\lambda}\left(\ln\left(\sqrt{\xi_{t-1,1}}\right) - \ln\left(\sqrt{\zeta_1}\right)\right) \\
&- \left(\tilde{\lambda} - \frac{1}{2}\right)\left(\ln\left(\sqrt{\zeta_1 + \zeta_2^2}\right) - \ln\left(\sqrt{\xi_{t-1,1} + \xi_{t,2}^2}\right)\right) \\
&- \ln\left(K_{\tilde{\lambda}}\left(\sqrt{\xi_{t-1,1}\zeta_1}\right)\right) + \ln\left(K_{\tilde{\lambda} - \frac{1}{2}}\left(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)}\right)\right).
\end{aligned} \tag{17}$$

where $\tilde{\lambda} = \lambda - p/2$. Now we can state the following result:

Theorem 3. *Assume that the parameter space Θ is compact. Then, the (exact) ML estimator $\hat{\theta}$ is strongly consistent.*

The proof of Theorem 3 is based on the uniform strong law of large numbers applied to the standardized version of the log-likelihood function (16) and the identifiability result of Browne & McNicholas (2015) which implies that the expectation of the log-likelihood function has a unique maximum in the point of the true parameter value θ_0. The proof is given in Appendix A.1.3. As noted above, the role of the initial values becomes negligible in large samples so that the above statement holds for the conditional ML estimator as well.

Although the asymptotic normality of the ML estimator seems very plausible, it is difficult to prove due to the complexity of the likelihood function and therefore left for future research.

4 Empirical example

Next we fit the GHAR model to daily means of particle matter PM_{10} measurements in Helsinki (Finland), Kallio, during 2013 and 2014. The original data was requested from HSY (Helsinki Region Environmental Services Authority HSY, Air protection unit) and it consists of hourly measurements of particle matter ($\mu g/m^3$) during the years 2013 – 2015. This particular data set is chosen for its evident non-Gaussian features which can be used to demonstrate the ability of the GHAR model to take these features into account. Similar data has been previously modelled using, for example, neural networks (see, e.g., Perez & Reyes (2006), Voukantsis et al. (2011)).

The measurements are aggregated to daily means (Figure 1) discarding missing values, and therefore some days may consist of less than 24 observations. The last day of 2013 is missing all measurements and is therefore left out of the data set. Hence, using the observations from the first two years for parameter estimation leads to sample size $n = 729$. The last year of the data is used in a forecasting experiment where we use the double AR (DAR) model of Ling (2007) for comparison.

A pth order DAR(p) process y_t is defined as

$$y_t = \sum_{i=1}^{p} \phi_i y_{t-i} + \eta_t \sqrt{\omega + \sum_{i=1}^{p} \alpha_i y_{t-i}^2},$$

where $\{\eta_t\}$ is a sequence of independent standard Gaussian random variables. The DAR model is conditionally heteroscedastic and, like the GHAR process, its conditional variance is a function of preceding values of the process itself.

In addition, a linear Gaussian AR model is considered as a benchmark model even though it is not considered to be an ideal choice for the data.
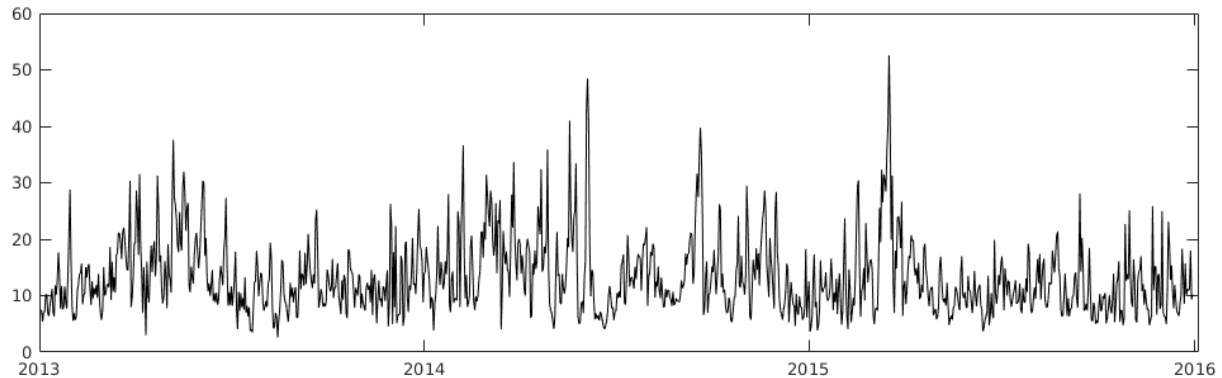

Figure 1: PM$_{10}$ measurements in Kallio from January 2013 to December 2015 as daily means.

Estimation of the GHAR and DAR models is carried out using the genetic algorithm combined with a gradient based method, an approach suggested by Dorsey & Mayer (1995). Matlab 2017b equipped with the optimization and parallel toolboxes is employed to compute the estimates for the GHAR model. The key idea in this two step estimation is that the genetic algorithm finds reasonable starting points for a gradient based method which is then expected to converge more efficiently to a point where the gradient of the log-likelihood is zero. Multiple optimization rounds are carried out in order to ensure that the procedure performs well enough to find a global maximum. Estimation and diagnostics of the DAR and AR models are carried out using the R software (R Core Team 2017) and the stats package.

The model selection procedure for the GHAR model is as follows: First we estimate the unrestricted GHAR model for all candidate values of p, for example $1 \le p \le 20$. We select the order p based on information criteria (AIC, HQC and BIC) and after the best fitting unrestricted model is found, we restrict the shape parameter λ to some of the special cases given in Table 1 in an attempt to increase the reliability of the estimation. Restricting λ is not a necessary step but it is expected to lead to an adequate and in, any case, a more parsimonious model.

Table 2: Estimation results for different GHAR models where the bolded zeros represent the restricted parameters. The standard errors are given in parenthesis below the estimates. The lowest values in the information criteria columns are bolded.

Model	ϕ_1	λ	δ	μ	σ^2	γ	$\hat{\ell}^c$	AIC	HQC	BIC
GHAR(1) $\lambda=0$	0.405 (0.07)	**0**	0.930 (0.23)	7.730 (0.64)	15.106 (1.48)	4.386 (0.80)	-2125.7	4256.4	4260.8	4284.4
GHAR(1) $\gamma=0$	0.684 (0.03)	−0.655 (0.17)	0.631 (0.14)	11.015 (0.43)	32.285 (3.68)	**0**	-2159.8	4324.6	4329.1	4352.6
GHAR(1)	0.194 (0.10)	1.794 (0.24)	0.004 (0.12)	6.214 (0.37)	0.020 (0.59)	0.009 (0.26)	-2121.6	**4249.2**	**4254.5**	**4282.7**
AR(1)	0.64 (0.03)			13.8 (0.53)	26.47		-2228.8	4460.6	4477.4	4463.2

Note that, fitting a GHAR model of order $p \ge 2$ with the pth autoregressive coefficient ϕ_p restricted to zero does not correspond to fitting a GHAR model of order $p-1$ (see discussion in Section 3.2). This also means that choosing a model with a larger order may give a poorer fit and, unlike with the Gaussian autoregressive model, reduce the maximum value of the likelihood function.

The estimation results are given in Table 2 for both the linear Gaussian AR model and the

GHAR model with order $p = 1$, suggested by the BIC (4441.1 for the DAR model) for both models, and two different restricted GHAR models. For the DAR(p) model, the case $p = 1$ yielded the best results with estimates

$$y_t = \underset{(0.08)}{0.66}\, y_{t-i} + \eta_t \sqrt{\underset{(2.71)}{20.47} + \underset{(0.06)}{0.14}\, y_{t-i}^2}.$$

The residual diagnostics for the DAR(1) model are presented in the Supplementary appendix. The residuals refer to a non-Gaussian distribution implying that the skewness is taken inadequately into account. In addition, there is some minor autocorrelation left in the first few lags of the residual series.

The GHAR model is also fitted with two different restrictions: Firstly, the shape parameter λ is restricted to zero so that the one-dimensional conditional distributions of the GHAR model follow the normal inverse Gaussian distribution (see Table 1 for the list of special cases). Secondly, despite the evident skewness of the data, a symmetric GHAR model is estimated by restricting the skewness parameter γ to zero.

The error sequence of the GHAR process is not independent and identically distributed so that, instead of usual residual diagnostics, we employ quantile residuals and the related tests developed by Kalliovirta (2012). The GHAR model restricted with $\lambda = 0$ performed slightly better than the unrestricted model in terms of residual diagnostics, showing no signs of misspecification (see Supplementary appendix), and hence we mainly focus on the restricted GHAR model. As anticipated above, the linear AR model is not capable of filtering out the conditional heteroscedasticity sufficiently (see Supplementary appendix).

In addition, the one-dimensional estimated marginal distributions of the DAR and GHAR processes and the kernel density estimate of the observed data in Figure 2 suggest that the GHAR model is able to model the estimated distribution more precisely than the DAR model, mostly due to the skewness observed in the series.
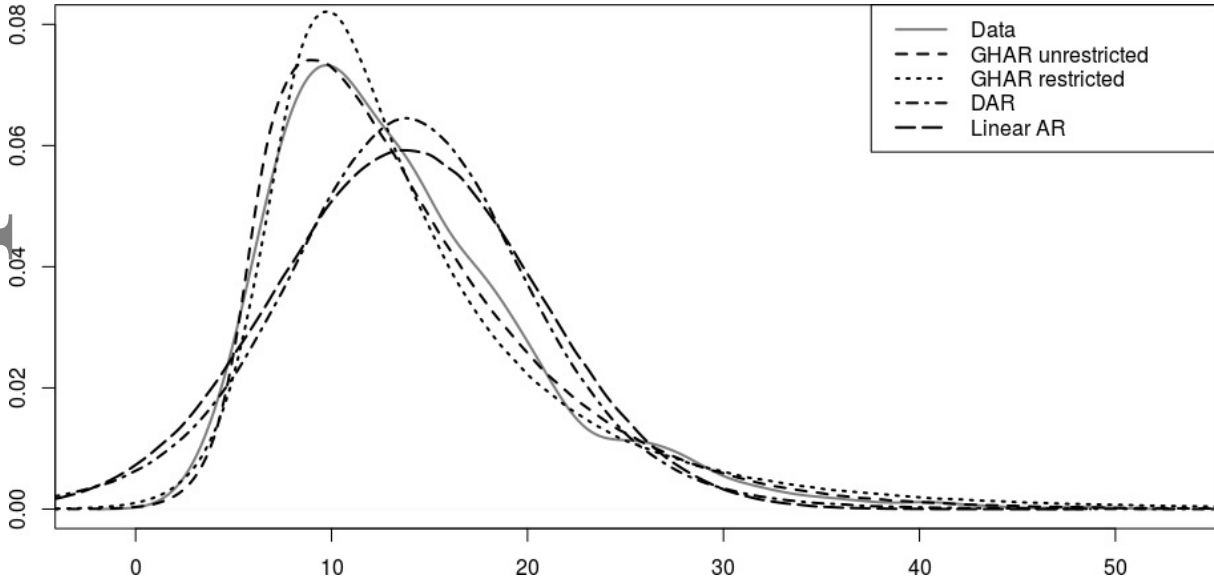

Figure 2: Estimated marginal densities based on simulation for the GHAR and DAR models and the kernel density estimate for the data.

Next we consider the out-of-sample forecasting performance of the GHAR, DAR and AR models by comparing the accuracy of point forecasts and forecast intervals. We use the estimated models to compute forecasts for the year 2015 with multiple forecast horizons. As forecasting particle concentrations in air is considered, it seems desirable to make negative forecast errors (i.e. forecast smaller than the observed value) more costly than positive ones. Therefore, we choose to compare the point forecasts by using the linex loss function. Linex has been used previously, for example, to compare forecasts of New Zealand stock market volatility by Yu

Table 3: Forecast errors for the estimated GHAR and DAR models using the MSE and linex loss functions divided by the corresponding forecast errors of the Gaussian AR model.

Loss	Model	Forecast horizon (days)								
		1	2	3	4	5	6	7	14	30
MSE	GHAR	1.01	0.98	0.98	0.99	0.99	0.99	0.99	1.01	1.01
	GHAR$_{\lambda=0}$	1.01	0.97	0.97	0.98	0.98	0.98	0.98	1.01	1.02
	DAR	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00	1.00
Linex	GHAR	0.13	0.09	0.21	0.18	0.21	0.42	0.34	0.83	0.87
	GHAR$_{\lambda=0}$	0.03	0.01	0.06	0.04	0.04	0.17	0.10	0.83	0.75
	DAR	0.73	0.64	0.72	0.73	0.75	0.83	0.84	1.00	1.00

Table 4: Proportion of observations in the year 2015 below the α sample quantiles in 250000 simulations computed for every observation and at each of the forecast horizons.

Model	α	Forecast horizon (days)								
		1	2	3	4	5	6	7	14	30
	0.950	0.951	0.964	0.959	0.964	0.961	0.961	0.964	0.966	0.964
GHAR(1)	0.975	0.970	0.975	0.981	0.986	0.981	0.978	0.978	0.983	0.982
	0.990	0.986	0.989	0.997	0.994	0.997	0.997	0.992	0.991	0.991
	0.995	0.992	0.997	0.997	0.997	0.997	0.997	0.997	0.991	0.997
	0.950	0.951	0.967	0.967	0.967	0.964	0.964	0.969	0.972	0.970
GHAR(1)$_{\lambda=0}$	0.975	0.973	0.978	0.981	0.989	0.989	0.989	0.989	0.989	0.988
	0.990	0.989	0.994	1.000	0.997	1.000	0.997	1.000	0.994	0.997
	0.995	0.992	1.000	1.000	1.000	1.000	1.000	1.000	0.997	0.997
	0.950	0.946	0.958	0.952	0.949	0.943	0.946	0.946	0.946	0.946
DAR(1)	0.975	0.964	0.973	0.967	0.964	0.958	0.964	0.964	0.964	0.964
	0.990	0.976	0.985	0.982	0.985	0.979	0.979	0.976	0.979	0.979
	0.995	0.982	0.991	0.991	0.991	0.991	0.988	0.991	0.988	0.988
	0.950	0.945	0.948	0.948	0.947	0.950	0.950	0.950	0.949	0.946
AR(1)	0.975	0.959	0.961	0.970	0.967	0.961	0.964	0.964	0.966	0.964
	0.990	0.970	0.975	0.972	0.975	0.975	0.972	0.972	0.972	0.973
	0.995	0.984	0.986	0.981	0.983	0.981	0.981	0.978	0.980	0.979

(2002). The linex loss at time t has the expression $\exp\{a(y_t - y_t^*)\} - a(y_t - y_t^*) - 1$, $a \neq 0$, where y_t^* is the forecasted value for time t (see, for example, Patton & Timmermann (2007)). Choosing $a > 0$ leads to greater loss for positive forecast errors than for negative errors and thus $a = 1$ is chosen.

We use estimates of the conditional expectation, computed by simulating 250000 series of length 30 starting from each day, as point forecasts. All forecasts are computed using the same simulation procedure to ensure comparability between the forecasts. The parameter estimates are updated for each day using the optim function in R software (stats package), setting the previous esimates as the starting point. The employed method is used for simplicity but (in general) it does not result in optimal forecasts under the linex loss function (for linex optimal point forecasts see, for example, Christoffersen & Diebold (1997) and Patton & Timmermann (2007)). Simulating the generalized hyperbolic distribution is carried out by simulating values from the generalized inverse Gaussian distribution and then from the normal distribution as suggested by McNeil et al. (2005). The simulations are computed using the R software (R Core Team 2017) and GIGrvg library (Leydold & Hormann 2017).

Results in Table 3 show that in the sense of mean squared error (MSE) loss, all the considered models perform equally well with only minor differences. When comparing the forecasts under

the asymmetric linex loss, we see that all forecast errors for the GHAR models are considerably smaller than for the other models, especially the AR model, even for the one week and one month ahead forecasts. This is likely due to the GHAR model being more capable of taking the non-Gaussianity of the distribution into account.

We also consider forecasting the right tails to forecast probabilities that certain limits are exceeded in the near future by comparing sample quantiles of the simulated series to the corresponding theoretical probabilities. Proportions of observations in the data that remain below the α quantiles are shown in Table 4 (for more results, see Supplementary appendix).

The DAR and AR models seem to perform exceptionally well near the 0.95 quantile but for smaller values of α the sample quantiles are significantly oversized and for $\alpha = 0.99$ (and above) the sample quantiles are undersized. The GHAR models, on the other hand, perform more uniformly and their sample quantiles are relatively close to the theoretical quantiles even for the longer forecast horizons and all α. Figure B4 in Supplementary appendix shows the 0.95 quantiles for January 2015 for the selected forecast horizons. As a reference, the figure also shows the sample quantiles of the simulated AR process which settle to the corresponding quantiles of its stationary distribution considerably faster than the sample quantiles of the GHAR processes as the length of the forecast horizon grows.

5 Conclusions

In this paper we defined a non-linear non-Gaussian autoregressive model by utilizing the co-variance structure of a Gaussian AR model in combination with the multivariate generalized hyperbolic distribution. The resulting GHAR model has multiple interesting properties such as an explicit stationarity condition and known low order stationary marginal distributions.

The GHAR model was fitted to a PM_{10} particle measurement data, consisting of daily means of hourly measurements in Helsinki, Finland. Even though the data is notably skewed and conditionally heteroscedastic the GHAR model was able to model the data, including the shape of the density, reasonably well. Forecasting performance of the GHAR model was also compared against the Gaussian AR model and the DAR model of Ling (2007) by using both symmetric and asymmetric loss functions. The results obtained indicate that taking the non-Gaussianity of the data into account leads to better forecasts when the goal is to forecast the tails of the distribution or when there is a reason to treat positive and negative forecast errors differently.

Aknowledgements

References

Anderson, T. W. (1984), *An introduction to multivariate statistical analysis*, Wiley series in probability and mathematical statistics, Probability and mathematical statistics, 2nd edn, Wiley, New York.

Barndorff-Nielsen, O. (1977), 'Exponentially decreasing distributions for the logarithm of particle size', *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* **353**(1674), 401–419.

Barndorff-Nielsen, O. E. (1997), 'Normal inverse gaussian distributions and stochastic volatility modelling', *Scand. J. Stat.* **24**(1), 1–13.

Blæsild, P. (1981), 'The two-dimensional hyperbolic distribution and related distributions, with an application to johannsen's bean data', *Biometrika* **68**(1), 251–263.

Browne, R. P. & McNicholas, P. D. (2015), 'A mixture of generalized hyperbolic distributions', *Canad. J. Statist.* **43**(2).

Brychkov, Y. A. (2016), 'Higher derivatives of the bessel functions with respect to the order', *Integral Transforms Spec. Funct.* **27**(7), 566–577.

Christoffersen, P. F. & Diebold, F. X. (1997), 'Optimal prediction under asymmetric loss', *Econometric Theory* **13**(6), 808–817.

Davidson, J. (2002), *Stochastic limit theory*, 2nd edn, Oxford University Press, New York, N.Y.

Dorsey, R. E. & Mayer, W. J. (1995), 'Genetic algorithms for estimation problems with multiple optima, nondifferentiability, and other irregular features', *J. Bus. Econom. Statist.* **13**(1), 53–66.

Gaunt, R. E. (2014), 'Inequalities for modified bessel functions and their integrals', *J. Math. Anal. Appl.* **420**(1), 373 – 386.

Hu, W. (2005), Calibration of multivariate generalized hyperbolic distributions using the EM algorithm, with applications in risk management, portfolio optimization and portfolio credit risk, Dissertation, Florida state university.

Ismail, M. E. H. & Muldoon, M. E. (1978), 'Monotonicity of the zeros of a cross-product of bessel functions', *SIAM J. Math. Anal.* **9**(4), 759–767.

Kalliovirta, L. (2012), 'Misspecification tests based on quantile residuals', *Econom. J.* **15**(2), 358–393.

Kalliovirta, L., Meitz, M. & Saikkonen, P. (2015), 'A gaussian mixture autoregressive model for univariate time series', *Journal of Time Series Analysis* **36**(2), 247–266.

Laforgia, A. & Natalini, P. (2010), 'Some inequalities for modified bessel functions', *J. Inequal. Appl.* **2010**(1), 10.

Leydold, J. & Hormann, W. (2017), *GIGrvg: Random Variate Generator for the GIG Distribution*. R package version 0.5.
URL: *https://CRAN.R-project.org/package=GIGrvg*

Ling, S. (2007), 'A double ar (p) model: Structure and estimation', *Statist. Sinica* **17**(1), 161–175.

Loève, M. (1955), *Probability theory : foundations, random sequences*, Univ. Ser. Higher Math., Van Nostrand, Toronto.

Lütkepohl, H. (2006), *New introduction to multiple time series analysis*, Springer, Berlin. Corr. 2nd pr. 2007.

McNeil, A. J., Embrechts, P. & Frey, R. (2005), *Quantitative risk management : concepts, techniques and tools*, Princet. Ser. Finance, Princeton University Press, Princeton.

Meitz, M., Preve, D. & Saikkonen, P. (2018), A mixture autoregressive model based on student's t–distribution. Pre-print version arXiv:1805.04010v1.

Meyn, S. & Tweedie, R. L. (2009), *Markov chains and stochastic stability*, 2nd edn, Cambridge University Press, Cambridge.

Patton, A. J. & Timmermann, A. (2007), 'Properties of optimal forecasts under asymmetric loss and nonlinearity', *J. Econometrics* **140**(2), 884 – 918.

Perez, P. & Reyes, J. (2006), 'An integrated neural network model for pm10 forecasting', *Atmospheric Environment* **40**(16), 2845 – 2851.

Pötscher, B. & Prucha, I. (1997), *Dynamic Nonlinear Econometric Models: Asymptotic Theory*, Springer-Verlag Berlin Heidelberg.

Prause, K. (1999), The Generalized Hyperbolic Model: Estimation, Financial Derivatives, and Risk Measures, Disseration, University of Freiburg.

R Core Team (2017), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Ranga Rao, R. (1962), 'Relations between weak and uniform convergence of measures with applications', *The Annals of Mathematical Statistics* **33**(2), 659–680.

Soni, R. P. (1965), 'On an inequality for modified bessel functions', *Journal of Mathematics and Physics* **44**(1-4), 406–407.

Stout, W. (1974), *Almost sure convergence*, Academic Press, New York, N.Y.

Straumann, D. & Mikosch, T. (2006), 'Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach', *Ann. Statist.* **34**(5), 2449–2495.

Teräsvirta, T., Tjøstheim, D. & Granger, C. W. J. (2010), *Modelling nonlinear economic time series*, Advanced texts in econometrics, Oxford University Press, New York, N.Y.

Tierney, L. (1994), 'Markov chains for exploring posterior distributions', *Ann. Statist.* **22**(4), 1701–1728.

Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A. & Kolehmainen, M. (2011), 'Intercomparison of air quality data using principal component analysis, and forecasting of pm10 and pm2.5 concentrations using artificial neural networks, in thessaloniki and helsinki', *Science of The Total Environment* **409**(7), 1266 – 1276.

Watson, G. N. (1944), *A treatise on the theory of Bessel functions*, 2nd edn, Cambridge University Press, Cambridge.

Yu, J. (2002), 'Forecasting volatility in the new zealand stock market', *Applied Financial Economics* **12**(3), 193–202.

Department of Mathematics and Statistics, University of Helsinki, Finland

E-mail address: henri.c.karttunen@helsinki.fi

A Appendix

A.1 Proofs

A.1.1 Proof of Theorem 1

To show that the GHAR process is stationary, denote $\boldsymbol{y}_t^+ = (y_t, \boldsymbol{y}_{t-1}) = (\boldsymbol{y}_t, y_{t-p})$ and assume that $\boldsymbol{y}_0$ follows a generalized hyperbolic distribution with density function $\mathsf{gh}_p(\boldsymbol{y}_0; \lambda, \delta, \boldsymbol{\mu}_p, \Gamma_p, \boldsymbol{\gamma}_p)$, where the parameters as well as the notation $\tilde{\theta}_t$ to be used below are defined in Section 3.1. It follows from the Yule-Walker equations that the covariance matrix Γ_{p+1}, defined after the equation (5), can be partitioned as

$$\Gamma_{p+1} = \begin{bmatrix} \sigma^2 + \boldsymbol{\phi}'\Gamma_p\boldsymbol{\phi} & \boldsymbol{\phi}'\Gamma_p \\ \Gamma_p\boldsymbol{\phi} & \Gamma_p \end{bmatrix}. \tag{A1}$$

From the derivation of the conditional density function of the generalized hyperbolic distribution in Appendix A.6 (see especially equation (A41)) and the definition of the conditional density function $\mathsf{gh}_1(y_t; \tilde{\theta}_t)$ in (6) we can now conclude that

$$\mathsf{gh}_p(\boldsymbol{y}_0; \lambda, \delta, \boldsymbol{\mu}_p, \Gamma_p, \boldsymbol{\gamma}_p)\mathsf{gh}_1(y_1; \tilde{\theta}_t) = \mathsf{gh}_{p+1}(\boldsymbol{y}_1^+; \lambda, \delta, \boldsymbol{\mu}_{p+1}, \Gamma_{p+1}, \boldsymbol{\gamma}_{p+1}), \tag{A2}$$

where $\boldsymbol{\mu}_{p+1} = \mu\mathbf{1}_{p+1}$ and $\boldsymbol{\gamma}_{p+1} = \gamma\mathbf{1}_{p+1}$.

The stationarity of the process $\{z_t\}$ implies that Γ_p and Γ_{p+1} are positive definite Toeplitz matrices. Hence, for all t, $\Gamma_p = \mathsf{Cov}(\boldsymbol{z}_t) = \mathsf{Cov}(\boldsymbol{z}_{t-1})$ equals both the upper left corner and lower right corner of Γ_{p+1}, the covariance matrix of $\boldsymbol{z}_t^+$. Marginalizing y_{-p+1} out of the distribution of $\boldsymbol{y}_1^+$ and using the fact that all the components of μ_{p+1} and γ_{p+1} are equal we now find by Proposition 3.13 of McNeil et al. (2005) that the random vectors $\boldsymbol{y}_0$ and $\boldsymbol{y}_1$ are identically distributed.

As $\{\boldsymbol{y}_t\}$, $t \geq 1$, is a Markov chain, the above 'first step stationarity' implies stationarity of the whole process, provided that the initial distribution is chosen to be $\boldsymbol{y}_0 \sim \mathsf{gh}_p(\lambda, \delta, \mu\mathbf{1}_p, \Gamma_p, \gamma\mathbf{1}_p)$ (see, e.g., Meyn & Tweedie 2009, p. 230–231).

A.1.2 Proof of Theorem 2

We show that the process $\{\boldsymbol{y}_t\}$ is ergodic in the sense that the n step transition probability measure tends to the stationary distribution of the process. In Section 3.1 the GHAR process was defined as a pth order Markov chain and it was noted that the process has a stationary distribution, which we denote here by π. The argumentation below follows Kalliovirta et al. (2015) and is based on Chapter 13 of Meyn & Tweedie (2009).

Let $\boldsymbol{y} \in \mathbb{R}^p$ be an initial value. Denote the total variation norm (see, Meyn & Tweedie 2009, p. 315) as $\|\cdot\|$ and the q step transition probability measure from point $\boldsymbol{y}$ to a (Borel) measurable set $A \subset \mathbb{R}^p$ as $P^q(A, \boldsymbol{y}) = P(\boldsymbol{y}_q \in A|\boldsymbol{y}_0 = \boldsymbol{y})$. The stationary probability of the set A is denoted as $\pi(A)$. Now we show that the total variation norm of the difference $P^n(\cdot, \boldsymbol{y}) - \pi(\cdot)$ tends to zero as $n \to \infty$ which implies that $\{y_t\}$ is ergodic.

By the definition of the GHAR process the conditional density function $\mathsf{gh}_1(\cdot; \tilde{\theta}_t)$ of y_t given $\boldsymbol{y}_{t-1}$ (see Section 3.1) is strictly positive on the whole real line. Therefore, the p step transition probability

$$P^p(A, \boldsymbol{y}) = \int_A \prod_{t=1}^p \mathsf{gh}_1(y_t; \tilde{\theta}_t)d\boldsymbol{y}_p$$

is strictly positive for any Borel measurable set $A \subset \mathbb{R}^p$ with positive Lebesgue measure. Consequently, the p step transition probability measure is absolutely continuous with respect to the stationary probability measure π and, from any initial state $\boldsymbol{y}$, the process $\{\boldsymbol{y}_t\}$ can in p steps reach any set with positive Lebesgue measure. This implies that the process $\{\boldsymbol{y}_t\}$ is π-irreducible and aperiodic (for definitions see, e.g., Meyn & Tweedie 2009, Chapters 4 and 5).

Hence, it follows from Tierney (1994, Theorem 1) that, for π-almost all $\boldsymbol{y} \in \mathbb{R}^p$, the p step transition probability tends to the stationary distribution in total variation norm, that is, $\|P^{pn}(\cdot, \boldsymbol{y}) - \pi(\cdot)\| \to 0$, as $n \to \infty$. For the one step transition probability this implies that

$$\liminf_{n \to \infty} \|P^n(\cdot, \boldsymbol{y}) - \pi(\cdot)\| = 0, \tag{A3}$$

for π-almost all $\boldsymbol{y} \in \mathbb{R}^p$. As $\{\boldsymbol{y}_t\}$ is π-irreducible it follows from Corollary 1 of Tierney (1994) that (A3) is true for any initial state $\boldsymbol{y} \in \mathbb{R}^p$. As $\pi(\cdot)$ is the stationary distribution of $\boldsymbol{y}_t$, it follows from Meyn & Tweedie (2009, Proposition 13.3.2) that $\|P^n(\cdot, \boldsymbol{y}) - \pi(\cdot)\|$ is non-increasing which together with (A3) implies that

$$\|P^n(\cdot, \boldsymbol{y}) - \pi(\cdot)\| \to 0$$

for all $\boldsymbol{y} \in \mathbb{R}^p$. Thus, the process $\{\boldsymbol{y}_t\}$ is ergodic (see Meyn & Tweedie 2009, p. 313 - 316).

A.1.3 Proof of Theorem 3

For simplicity, we assume stationary initial values $\boldsymbol{y}_0$ which implies the stationarity of the process $\{\boldsymbol{y}_t\}$.

The function $\theta \mapsto \ell_t(\theta)$ is continuous (see Appendix A.5) and, as the parameter space is compact, the log-likelihood function attains its maximum in Θ. Using similar argumentation as in Appendix A.5, we see that the log-likelihood function is also a continuous function of $(y_t, \boldsymbol{y}_{t-1})$, $t = 1, \ldots, n$, for every $\theta \in \Theta$ and hence the ML estimator is measurable.

As noted in the text, the effect of the initial values on the the log-likelihood diminishes as n grows so that it is sufficient here to consider only the conditional log-likelihood function and its standardized version $\overline{\ell}^c(\theta) := n^{-1}\ell^c(\theta)$. Due to stationarity it holds that $\mathrm{E}[\overline{\ell}^c(\theta)] = \mathrm{E}[\ell_1(\theta)]$. We also note the following result, the proof of which is given in Appendix A.4.

Theorem 4. *The expected value* $\mathsf{E}(\sup_{\theta \in \Theta}\{|\ell_t(\theta; y_t|\boldsymbol{y}_{t-1})|\})$ *of the conditional log-likelihood function of the GHAR process for time t is finite.*

Next we need to show that the log-likelihood obeys the uniform strong law of large numbers. Denote by $\mathcal{X}$ the set of continuous $\mathbb{R}$ valued functions, equipped with the norm $\|f\| = \sup_{\theta \in \Theta} |f(\theta)|$. By compactness of Θ and Theorems 5.24 and 5.26 of Davidson (2002), it follows that $(\mathcal{X}, \|\cdot\|)$ is a separable Banach space. Note that the log-likelihood functions ℓ_t of a single observation are continuous functions of both the parameter θ and observations $\boldsymbol{y}_t$ and thus also belong in $\mathcal{X}$.

We have assumed stationarity for $\{y_t\}$, and therefore to show ergodicity it is sufficient to use Markov ergodicity in the sense of (Stout 1974, p. 195) instead of the ergodicity stated in Theorem 2. As $\{y_t\}$ is a stationary Markov chain with transition probability measure equivalent to the Lebesgue measure it is Markov ergodic and Theorem 3.6.5 (i) of Stout (1974) shows that every invariant set has either measure 0 or 1. Hence, the invariant sigma algebra is trivial and thus it follows by continuity of the log-likelihood function and from, for example, Proposition 2.6 of Straumann & Mikosch (2006), that the process $\{\ell_t(\theta; y_t)\}$ is also stationary and ergodic.

Having established that the process $\{\ell_t\}$ is a stationary and ergodic process with values in a separable Banach space, Theorem 4 enables us to use Theorem 6.5 of Ranga Rao (1962) and conclude that

$$\|\overline{\ell}^c(\theta) - E(\overline{\ell}^c(\theta))\| = \sup_{\theta \in \Theta} \left|\overline{\ell}^c(\theta) - E(\ell_1(\theta))\right| \to 0,$$

as $n \to \infty$.

In Appendix A.5 we show that the expected log-likelihood function has a unique maximum at the true parameter value θ_0. Now Theorem 3.1 of Pötscher & Prucha (1997) we find that the ML estimator $\hat{\theta}$ is strongly consistent so that $\hat{\theta} \to \theta_0$ almost surely as $n \to \infty$.

16

A.2 Derivation of the log-likelihood function

In this appendix we derive the expression of the (conditional) log-likelihood function given in equation (17). The log-likelihood function is based on the conditional density function derived in Appendix A.6 with $d = p + 1$ and $d_2 = p$. First, define ζ_1 and ζ_2 from the equation

$$
\tilde{\delta}_t + \left[\tilde{\sigma}_t^2\right]^{-1}(\tilde{\gamma}_t)^2 = \frac{\tilde{\delta}_t}{\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p}\left(\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p + \frac{(\gamma - \boldsymbol{\phi}' \boldsymbol{\gamma}_p)^2}{\sigma^2}\right)
$$
$$
=: \frac{\tilde{\delta}_t}{\zeta_1}\left(\zeta_1 + \zeta_2^2\right), \tag{A4}
$$

where the parameters are as defined in Section 3.1. Likewise, define $\xi_{t_1,1}$ and $\xi_{t,2}$ from

$$
\tilde{\delta}_t + (y_t - \tilde{\mu}_t)'\left[\tilde{\sigma}_t^2\right]^{-1}(y_t - \tilde{\mu}_t)
$$
$$
= \frac{\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p}{\tilde{\delta}_t}\left(\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)'\Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p) + \frac{(y_t - \phi_0 - \boldsymbol{\phi}'\boldsymbol{y}_{t-1})^2}{\sigma^2}\right)
$$
$$
=: \frac{\zeta_1}{\tilde{\delta}_t}\left(\xi_{t-1,1} + \xi_{t,2}^2\right) \tag{A5}
$$

Using the definitions in (A4) and (A5) we can write

$$
\left(\tilde{\delta}_t + (y_t - \tilde{\mu}_t)'\left[\tilde{\sigma}_t^2\right]^{-1}(y_t - \tilde{\mu}_t)\right)\left(\tilde{\delta}_t + \tilde{\gamma}_t'\left[\tilde{\sigma}_t^2\right]^{-1}\tilde{\gamma}_t\right)
$$
$$
= \left(\xi_{t-1,1} + \xi_{t,2}^2\right)\left(\zeta_1 + \zeta_2^2\right) \tag{A6}
$$

and

$$
\exp\left\{\tilde{\gamma}_t'[\tilde{\sigma}_t^2(\theta)]^{-1}(y_t - \tilde{\mu}_t)\right\} = \exp\left\{\xi_{t,2}\zeta_2\right\}. \tag{A7}
$$

With the above notation the normalizing constant c_1 of the conditional density function in equation (A41) (see Appendix A.6) can be expressed as

$$
c_1 = \frac{\left(\sqrt{\frac{\xi_{t-1,1}}{\zeta_1}}\right)^{\frac{p}{2}-\lambda}\left(\zeta_1 + \zeta_2^2\right)^{\frac{p+1}{2}-\lambda}}{(2\pi\sigma^2)^{\frac{1}{2}} K_{\lambda-\frac{p}{2}}\left(\sqrt{\xi_{t-1,1}\zeta_1}\right)}. \tag{A8}
$$

Using equations (A6) – (A8) and the conditional density function in (A41) the likelihood function (15) can be written as

$$
L^c(\theta; \mathbf{y})
$$
$$
= \prod_{t=1}^n \frac{\left(\sqrt{\frac{\xi_{t-1,1}}{\zeta_1}}\right)^{\frac{p}{2}-\lambda}\left(\sqrt{\zeta_1 + \zeta_2^2}\right)^{\frac{p+1}{2}-\lambda} K_{\lambda-\frac{p+1}{2}}\left(\sqrt{\left(\xi_{t-1,1} + \xi_{t,2}^2\right)\left(\zeta_1 + \zeta_2^2\right)}\right)}{(2\pi\sigma^2)^{\frac{1}{2}} K_{\lambda-\frac{p}{2}}\left(\sqrt{\xi_{t-1,1}\zeta_1}\right)} \frac{}{\left(\sqrt{\xi_{t-1,1} + \xi_{t,2}^2}\right)^{\frac{p+1}{2}-\lambda}} \exp\left\{\xi_{t,2}\zeta_2\right\}.
$$

Applying the logarithm function and rearranging terms we can express the conditional log-likelihood function (for time t) as in equation (17). As in (16), the conditional log-likelihood function ℓ^c can be expressed as a sum of the above log-likelihoods for time t.

A.3 Martingale property of the error sequence ε_t

In this Appendix we show that the sequence $\{\varepsilon_t, \mathcal{F}_{t-1}\}$ as defined in (9) is a martingale difference sequence with conditional variance equal to one. First we show that the process $\{\varepsilon_t\}$ has finite unconditional second moments and then that it is uncorrelated with zero mean and unit variance.

Let $\theta = (\boldsymbol{\phi}, \delta, \lambda, \mu, \sigma^2, \gamma)$ be fixed, $t \geq 1$ and the conditional parameter vector $\tilde{\theta}_t$ as defined in Section 3.1 (see the discussion after (6)). First we establish that σ_t^2, defined in equation (8), is bounded away from zero for all $\boldsymbol{y}_{t-1} \in \mathbb{R}^p$ which implies that ε_t is well defined. We begin by showing that the ratio

$$R_1(\tilde{\delta}_t) = R_1(\boldsymbol{y}_{t-1}) = \frac{K_{\tilde{\lambda}+1}(\tilde{\delta}_t)}{K_{\tilde{\lambda}}(\tilde{\delta}_t)},$$

is a bounded function of $\boldsymbol{y}_{t-1} \in \mathbb{R}^p$ (here we use a more explicit notation for the sake of clarity). We only need to consider the cases $\tilde{\lambda} > -1/2$ and $\tilde{\lambda} < -1/2$ because if $\tilde{\lambda} = -1/2$ we have $R_1(\boldsymbol{y}_t) = 1$ for any $\boldsymbol{y}_t \in \mathbb{R}^p$.

Assuming first $\tilde{\lambda} > -1/2$ it follows that $K_{\tilde{\lambda}+1}(\tilde{\delta}_t) > K_{\tilde{\lambda}}(\tilde{\delta}_t)$ (Soni 1965) and therefore $R_1(\boldsymbol{y}_t) \geq 1$ for any $\boldsymbol{y}_t \in \mathbb{R}^p$. From Gaunt (2014, Lemma 2.2) we find that $1/R_1(\boldsymbol{y}_t)$ is strictly monotone increasing for $\tilde{\lambda} > -1/2$ and, as $\tilde{\delta}_t \geq \delta$ by the definition of $\tilde{\delta}_t$, we get the upper limit

$$1 \leq R_1(\tilde{\delta}_t) \leq \frac{K_{\tilde{\lambda}+1}(\delta)}{K_{\tilde{\lambda}}(\delta)},$$

where the ratio on the right-hand side is clearly a finite constant.

On the other hand, assuming $\tilde{\lambda} < -1/2$ and denoting $\nu = -\tilde{\lambda}$ we have, as in the proof of Corollary 2.3 of Gaunt (2014), the limit

$$\lim_{x \to \infty} R_1(x) = \lim_{x \to \infty} \frac{K_{\nu-1}(x)}{K_\nu(x)} = 1,$$

and using Lemma 2.2 of Gaunt (2014) we see that

$$1 \geq R_1(\tilde{\delta}_t) \geq \frac{K_{\tilde{\lambda}+1}(\delta)}{K_{\tilde{\lambda}}(\delta)}.$$

Hence, $R_1(\tilde{\delta}_t)$ is bounded and for every fixed θ there exist positive constants m_θ and M_θ such that

$$0 < m_\theta \leq R_1(\tilde{\delta}_t) \leq M_\theta. \tag{A9}$$

Also, as $\tilde{\delta}_t \geq \delta$ we have

$$\tilde{\sigma}_t^2 = \frac{\tilde{\delta}_t}{\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p} \sigma^2 \geq \frac{\delta}{\delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p} \sigma^2 > 0,$$

which combined with (A9) implies that there exists a strictly positive constant C_θ such that

$$R_1(\tilde{\delta}_t)\tilde{\sigma}_t^2 \geq C_\theta \tag{A10}$$

for all $\boldsymbol{y}_{t-1} \in \mathbb{R}^p$. Using the Lemma 2.2 of Ismail & Muldoon (1978) we find that

$$R_2(\tilde{\delta}_t) - \left(R_1(\tilde{\delta}_t) \right)^2 > 0,$$

which together with (A10) implies that

$$\sigma_t^2 = R_1(\tilde{\delta}_t)\tilde{\sigma}_t^2 + \left(R_2(\tilde{\delta}_t) - \left(R_1(\tilde{\delta}_t) \right)^2 \right) \tilde{\gamma}_t^2$$

is bounded away from zero. Combining the above results shows that the random variables $\varepsilon_t = (y_t - \mu_t)/\sigma_t$, where μ_t is defined in (7), are well defined.

Next we show that all moments of $y_t - \mu_t$ are finite by showing that the absolute value of $y_t - \mu_t$ is dominated by a random variable that has all finite moments. If $\gamma = 0$ we have

$$y_t - \mu_t = a'\boldsymbol{y}_t^+ - \phi_0,$$

where $a = (1, -\boldsymbol{\phi})$ (see equation (12)). By Proposition 3.13 of McNeil et al. (2005) $y_t - \mu_t$ follows a one dimensional generalized hyperbolic distribution, and therefore has all moments finite.

When $\gamma \neq 0$ the definition of μ_t yields

$$y_t - \mu_t = y_t - \tilde{\mu}_t - R_1(\tilde{\delta}_t)\frac{\sqrt{(\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p)'\Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_p))}}{\sqrt{\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p}}\gamma\left(1 - \boldsymbol{\phi}'\mathbf{1}_p\right). \tag{A11}$$

Assuming $\gamma > 0$ and using equations (A9) and (A11) we find that there exists a finite constant K such that

$$a'\boldsymbol{y}_t^+ - \phi_0 + K + K\sum_{i=1}^p |y_{t-i} - \mu| \leq y_t - \mu_t \leq a'\boldsymbol{y}_t^+ - \phi_0. \tag{A12}$$

Likewise, assuming $\gamma < 0$ there exists a finite constant K such that

$$a'\boldsymbol{y}_t^+ - \phi_0 \leq y_t - \mu_t \leq a'\boldsymbol{y}_t^+ - \phi_0 - K + K\sum_{i=1}^p |y_{t-i} - \mu|. \tag{A13}$$

From Proposition 3.13 of McNeil et al. (2005) we find, again, that both $a'\boldsymbol{y}_t^+ - \phi_0 - K$ and all the $(y_{t-i} - \mu)$'s follow the one-dimensional generalized hyperbolic distribution. Therefore both sides of (A12) and (A13) are integrable which shows that $y_t - \mu_t$ is integrable. Furthermore, because the all moments of the generalized hyperbolic distribution are finite, $(y_t - \mu_t)^r$ is integrable for every finite $r > 0$. The above discussion together with ε_t being well-defined shows that all moments of ε_t are finite. Now, by definition, the sequence $\{\varepsilon_t\}$ is uncorrelated and for every $t \geq 1$ it holds that $\mathsf{E}(\varepsilon_t) = 0$ and $\mathrm{Var}(\varepsilon_t) = 1$.

A.4 Finite supremum

We next show that $\mathsf{E}(\sup_{\theta \in \Theta}\{|\ell_t(\theta; y_t|\boldsymbol{y}_{t-1})|\}) < \infty$ by showing that each row of the log-likelihood function (17) has an upper bound and a lower bound which do not depend on θ and have finite expectations. We assume that the parameter vector θ is as defined in Section 3.1 and the parameter space is compact. We begin with the last row of the log-likelihood function.

Last row of (17)

This section is split in two parts: First we show that the expectation of the last line has a finite lower bound and then that it also has a finite upper bound.

Lower bound - As the parameter space is assumed compact we can define $\lambda_{max}, \delta_{min} \in \mathbb{R}$ such that $|\tilde{\lambda}| + 1/2 \leq \lambda_{max}$ and $0 < \delta_{min} \leq \delta$. Furthermore, the stationarity assumption (5) implies that Γ_p is positive definite and therefore

$$\xi_{t-1,1}\zeta_1 = \left(\delta + (\mathbf{y}_{t-1} - \boldsymbol{\mu}_p)'\Gamma_p^{-1}(\mathbf{y}_{t-1} - \mu)\right)\left(\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\right) \geq \delta^2 \geq \delta_{min}^2.$$

Properties of the Bessel function imply that for every permissible value of λ there exists a finite constant $C_K > 0$ such that

$$K_{\tilde{\lambda}}(\sqrt{\xi_{t-1,1}\zeta_1}) \leq K_{\lambda_{max}}(\delta_{min}^2) \leq C_K. \tag{A14}$$

Properties of the Bessel function and an application of Corollary 3.4 of Gaunt (2014) yield

$$\frac{K_{\tilde{\lambda}-\frac{1}{2}}(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)})}{K_{\tilde{\lambda}}(\sqrt{\xi_{t-1,1}\zeta_1})} \geq C_K^{-1}e^{-\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)}}\left(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)} + 1\right)^{-1},$$

where the inequality $\sqrt{x + 1/2} \leq \sqrt{x} + 1$, for all $x > 0$, was also used. Applying the logarithm transformation and the inequality $\ln(y + 1) \leq y$, $y > -1$ it follows that

$$\ln\left(\frac{K_{\tilde{\lambda}-\frac{1}{2}}(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)})}{K_{\tilde{\lambda}}(\sqrt{\xi_{t-1,1}\zeta_1})}\right) \geq -\ln(C_K) - 2\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)}. \qquad (A15)$$

To show that the last row has an integrable lower bound we need to show that supremum of the second term of (A15) has a finite expectation.

Let $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_p$ be the eigenvalues of the inverse of the covariance matrix Γ_p. Positive definiteness of Γ_p implies that $\nu_i > 0$ for every $i = 1, \ldots, p$. In addition, it holds that (see, e.g., Anderson 1984, p. 590)

$$\zeta_1 = \delta + \boldsymbol{\gamma}_p' \Gamma_p^{-1} \boldsymbol{\gamma}_p \leq \delta + \nu_1 \boldsymbol{\gamma}_p' \boldsymbol{\gamma}_p. \qquad (A16)$$

By continuity of the mapping $\theta \mapsto \Gamma_p$ and compactness of the parameter space the values of the covariance matrix Γ_p lie in a compact set so that the eigenvalues $\nu_1, \ldots, \nu_p$ (as continuous functions of θ) are bounded by constants $\underline{\nu}$ and $\overline{\nu}$ such that $\overline{\nu} \geq \nu_1 \geq \cdots \geq \nu_p \geq \underline{\nu}$. Furthermore, all components of θ are bounded and the values of σ^2 are bounded away from zero. Therefore, ζ_1 and ζ_2 are bounded and there exists a finite constant C_ζ such that

$$\sup_{\theta \in \Theta}\left\{\sqrt{\zeta_1 + \zeta_2^2}\right\} \leq C_\zeta < \infty. \qquad (A17)$$

Using the Loève's c_r-inequality (Loève 1955, p. 155) and the triangle inequality we find that

$$\sqrt{\xi_{t-1,1}} \leq \sqrt{\delta} + \sqrt{\overline{\nu}}\sum_{i=1}^{p}|y_{t-i} - \mu| \leq \sqrt{\delta} + \sqrt{\overline{\nu}}\sum_{i=1}^{p}|y_{t-i}| + \sqrt{\overline{\nu}}p|\mu|. \qquad (A18)$$

Again, using the triangle inequality, we can write

$$\sqrt{\xi_{t,2}^2} \leq \sigma^{-1}\left(|y_t| + |\phi_0| + |\boldsymbol{\phi}'\mathbf{y}_{t-1}|\right). \qquad (A19)$$

Positive definiteness of the covariance matrix Γ_p implies that $\xi_{t-1,1} > 0$ and therefore combining the inequalities (A18) and (A19) and applying Minkowski's inequality we get

$$\sqrt{\xi_{t-1,1} + \xi_{t,2}^2} \leq \sqrt{\delta} + \sqrt{\overline{\nu}}\sum_{i=1}^{p}|y_{t-i}| + \sqrt{\overline{\nu}}p|\mu| + \sigma^{-1}\left(|y_t| + |\phi_0| + |\boldsymbol{\phi}'\mathbf{y}_{t-1}|\right). \qquad (A20)$$

Hence, there exists a finite constant C_ξ such that

$$\sup_{\theta \in \Theta}\left\{\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)}\right\} \leq C_\xi\left(1 + \sum_{i=0}^{p}|y_{t-i}|\right), \qquad (A21)$$

and thus we get a lower bound

$$\inf_{\theta \in \Theta}\left\{\ln\left(\frac{K_{\tilde{\lambda}-\frac{1}{2}}(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)})}{K_{\tilde{\lambda}}(\sqrt{\xi_{t-1,1}\zeta_1})}\right)\right\} \geq -\ln(C_K) - 2C_\zeta C_\xi\left(1 + \sum_{i=0}^{p}|y_{t-i}|\right). \qquad (A22)$$

The random variables y_{t-i} follow the one-dimensional generalized hyperbolic distribution and have therefore finite second moments which implies that the expectation of the lower bound (A22) is finite.

Upper bound - Now we find an upper bound for the supremum of the last row of (17) and show that it has finite expectation. We consider three cases: $\tilde{\lambda} < -1/2$, $|\tilde{\lambda}| \leq 1/2$ and $\tilde{\lambda} > 1/2$.

20

Assume first that $\tilde{\lambda} > 1/2$. Both $\xi_{t,2}^2$ and ζ_2^2 are non-negative so that

$$\sqrt{\xi_{t-1,1}\zeta_1} \leq \sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)} \tag{A23}$$

and therefore it holds for any $\tilde{\lambda}$ that

$$\frac{K_{\tilde{\lambda}-\frac{1}{2}}\left(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)}\right)}{K_{\tilde{\lambda}}\left(\sqrt{\xi_{t-1,1}\zeta_1}\right)} \leq 1,$$

which together with the properties of the Bessel function yields an upper bound zero for the last row of (17).

Assume now that $|\tilde{\lambda}| \leq 1/2$. Using properties of the Bessel function and Theorem 1.2 of Laforgia & Natalini (2010) we get an upper bound

$$\frac{K_{\tilde{\lambda}-\frac{1}{2}}(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)})}{K_{\tilde{\lambda}}(\sqrt{\xi_{t-1,1}\zeta_1})} \leq \frac{2 + \xi_{t-1,1}\zeta_1}{\sqrt{\xi_{t-1,1}\zeta_1}}. \tag{A24}$$

Here we have used the fact that $1 + \xi_{t-1,1}\zeta_1 > 1$ by positive definiteness of the covariance matrix Γ_p and $\sqrt{a} < a$, for all $a > 1$.

Assume now that $\tilde{\lambda} < -1/2$ and denote $\nu = -\lambda$. By Theorem 1.2 of Laforgia & Natalini (2010) and assuming (without loss of generality) that $\lambda_{max} \geq 1$ we find an upper bound

$$\frac{K_{\nu+\frac{1}{2}}\left(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)}\right)}{K_{\nu}(\sqrt{\xi_{t-1,1}\zeta_1})} \leq \frac{\lambda_{max} + \sqrt{\lambda_{max}^2 + \xi_{t-1,1}\zeta_1}}{\sqrt{\xi_{t-1,1}\zeta_1}} \leq \frac{2\lambda_{max}^2 + \xi_{t-1,1}\zeta_1}{\sqrt{\xi_{t-1,1}\zeta_1}}. \tag{A25}$$

On the right-hand side of (A24), as $\lambda_{max} \geq 1$, we can substitute the first term of the numerator by $2\lambda_{max}^2$ so that the upper bound in (A25) applies also in the case $|\tilde{\lambda}| \leq 1/2$.

Applying the logarithm function and supremum to both sides of (A25) we get

$$\sup_{\theta \in \Theta} \left\{ \ln \left(\frac{K_{\tilde{\lambda}-\frac{1}{2}}(\sqrt{(\xi_{t-1,1} + \xi_{t,2}^2)(\zeta_1 + \zeta_2^2)})}{K_{\tilde{\lambda}}(\sqrt{\xi_{t-1,1}\zeta_1})} \right) \right\} \leq 2\lambda_{max}^2 + \sup_{\theta \in \Theta} \{\xi_{t-1,1}\zeta_1\} - \ln(\delta_{min}),$$

where the inequality $\ln(x) < x - 1 < x$ for all $x > 0$ is also used. The first and the last terms are finite constants so that for integrability it suffices to show that $\mathsf{E}(\sup_{\theta \in \Theta} \{\xi_{t-1,1}\zeta_1\}) < \infty$.

Using the eigenvalues $\nu_1, \ldots, \nu_p$ of the inverse of Γ_p as in (A16) and Loeve's c_r inequality (Loève 1955, p. 155) we find that

$$\xi_{t-1,1} \leq \delta + \overline{\nu}(\mathbf{y}_{t-1} - \boldsymbol{\mu}_p)'(\mathbf{y}_{t-1} - \boldsymbol{\mu}_p) \leq \delta + 2\overline{\nu}\mathbf{y}_{t-1}'\mathbf{y}_{t-1} + 2p\overline{\nu}\mu^2.$$

The term ζ_1 is bounded by inequality (A16) so that, by the assumed compactness of the parameter space, there exists a finite constant $C_{\xi\zeta}$, such that

$$\sup_{\theta \in \Theta} \{\xi_{t-1,1}\zeta_1\} \leq C_{\xi\zeta} \left(1 + \mathbf{y}_{t-1}'\mathbf{y}_{t-1}\right).$$

Combining the above results we can conclude that there exists a finite positive constant C such that the last row of (17) is bounded from above by $C\left(1 + \mathbf{y}_{t-1}'\mathbf{y}_{t-1}\right)$. The random vectors $\mathbf{y}_{t-1}$ follow the p dimensional generalized hyperbolic distribution so that their second moments are finite, and therefore the last row of the log-likelihood function has an integrable supremum.

First row of (17)

Applying the triangle inequality to the absolute value of the term $\xi_{t,2}\zeta_2$ we find that, by the compactness of the parameter space, there exists a finite constant C such that

$$\sup_{\theta \in \Theta} \left\{ \left| -\frac{1}{2}\ln\left(2\pi\sigma^2\right) + \xi_{t,2}\zeta_2 \right| \right\} \leq C \left(1 + \sum_{i=0}^{p} |y_{t-i}|\right). \tag{A26}$$

Arguments already used above show that the right-hand side of (A26), and hence the first row of (17), has finite expectation.

Second row of (17)

Considering the second row, we know from (A18) that $\xi_{t-1,1}$ has an integrable upper bound. Also, $\ln\left(\sqrt{\xi_{t-1,1}}\right)$ is bounded from below by the finite constant $\ln(\sqrt{\delta_{min}})$. Again, using the eigenvalues $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_p$ of Γ_p^{-1} as above, we find that $\ln\left(\sqrt{\zeta_1}\right)$ has constant upper and lower bounds because

$$\sqrt{\delta + \underline{\nu}\boldsymbol{\gamma}_p'\boldsymbol{\gamma}_p} \leq \sqrt{\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p} = \sqrt{\zeta_1} \leq \sqrt{\delta + \overline{\nu}\boldsymbol{\gamma}_p'\boldsymbol{\gamma}_p}, \tag{A27}$$

As the eigenvalues are bounded and bounded away from zero, both terms of the second row have finite constant upper and lower bounds and are therefore integrable.

Third row of (17)

Inequality (A21) provides an upper bound for $\ln\left(\sqrt{\xi_{t-1,1} + \xi_{t,2}^2}\right)$ and finding a lower bound is trivial because $\xi_{t-1,1} + \xi_{t,2}^2 \geq \xi_{t-1,1} \geq \delta \geq \delta_{min}$. Also, using the same argumentations as above we find that

$$\sqrt{\delta_{min}} \leq \sqrt{\delta} \leq \sqrt{\zeta_1} \leq \sqrt{\zeta_1 + \zeta_2^2} \leq \sqrt{\zeta_1} + \sqrt{\zeta_2^2} \tag{A28}$$

and using the same eigenvalue argumentation as in of the case (A27) we get

$$\sqrt{\zeta_1 + \zeta_2^2} \leq \sqrt{\delta + \overline{\nu}\boldsymbol{\gamma}_p'\boldsymbol{\gamma}_p + \left(\gamma - \boldsymbol{\phi}'\boldsymbol{\gamma}_p\right)^2 / \sigma^2}. \tag{A29}$$

Hence there exists finite constants C_L and C_U such that

$$C_L\left(1 + \sum_{i=0}^{p} |y_{t-i}|\right) \leq \ln\left(\sqrt{\zeta_1 + \zeta_2^2}\right) - \ln\left(\sqrt{\xi_{t-1,1} + \xi_{t,2}^2}\right) \leq C_U, \tag{A30}$$

where both the upper and lower bound are integrable by the argumentation already used above.

Conclusion

By combining the above parts of the proof we see that

$$\mathsf{E}(\sup_{\theta\in\Theta}\left\{|\ell_t(\theta; y_t|\boldsymbol{y}_{t-1})|\right\}) < \infty,$$

as the expected values of supremums of each row of the log-likelihood (17) for time t has finite upper and lower bounds.

A.5 Unique maximizer

We show that the expected log-likelihood of the GHAR model has a unique maximimum at the true parameter value, denoted by θ_0. Browne & McNicholas (2015) show that the generalized hyperbolic distribution with parametrization as in Section 2 is identifiable in the sense that

$$\mathsf{gh}_d(\cdot; \theta) = \mathsf{gh}_d(\cdot; \theta_0)$$

implies $\theta_0 = \theta$.

For the autoregressive parametrization presented in Section 3.2 we have the conditional density $\mathsf{gh}_1(y_t|\boldsymbol{y}_{t-1}; \theta) = \mathsf{gh}_1(y; \tilde{\theta}_t(\theta))$, where the vector $\tilde{\theta}_t(\theta) = (\tilde{\lambda}, \tilde{\delta}_t, \tilde{\mu}_t, \tilde{\sigma}_t^2, \tilde{\gamma}_t)$ of time-dependent parameters is defined in Section 3.1. The expected value for log-likelihood function on time t has an integral form

$$\mathsf{E}_{\theta_0}\left(\ell_t(\theta)\right) = \int_{\mathbb{R}^{p+1}} \mathsf{gh}_{p+1}(y_t, \boldsymbol{y}_{t-1}; \theta_0) \ln\left(\mathsf{gh}_1(y_t; \tilde{\theta}_t(\theta))\right) dy_t \, d\boldsymbol{y}_{t-1}$$

and therefore the difference $\mathsf{E}_{\theta_0}\left(\ell_t(\theta)\right) - \mathsf{E}_{\theta_0}\left(\ell_t(\theta_0)\right)$ can be expressed as

$$\mathsf{E}_{\theta_0}\left(\ell_t(\theta)\right) - \mathsf{E}_{\theta_0}\left(\ell_t(\theta_0)\right) = \int_{\mathbb{R}^p} \mathsf{gh}_p(\boldsymbol{y}_{t-1};\theta_0) \int_{\mathbb{R}} \mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta_0)) \ln\left(\frac{\mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta))}{\mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta_0))}\right) dy_t\, d\boldsymbol{y}_{t-1}$$

$$= -\int_{\mathbb{R}^p} \mathsf{gh}_p(\boldsymbol{y}_{t-1};\theta_0) D(\theta,\theta_0) d\boldsymbol{y}_{t-1}, \tag{A31}$$

where $D(\theta,\theta_0)$ denotes the Kullback-Leibler divergence between the densities $\mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta))$ and $\mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta_0))$.

It is shown in Appendix A.6 that $(y_t, \boldsymbol{y}_{t-1})$ follows the $p+1$ dimensional generalized hyperbolic distribution and therefore $(y_t, \boldsymbol{y}_{t-1})$ has a strictly positive density on the whole $\mathbb{R}^{p+1}$. Now by (A31) we have $E\left[\ell_t(\theta)\right] = E\left[\ell_t(\theta_0)\right]$ if and only if

$$\mathsf{gh}_1\left(y_t;\tilde{\theta}_t(\theta_0)\right) = \mathsf{gh}_1\left(y_t;\tilde{\theta}_t(\theta)\right) \tag{A32}$$

for almost all $y_t \in \mathbb{R}$ and any fixed $\boldsymbol{y}_{t-1} \in \mathbb{R}^p$. To show that the expectation of the log-likelihood has a unique maximizer we therefore need show that (A32) implies $\theta_0 = \theta$. Note that, the identifiability of the generalized hyperbolic distribution implies that if $\mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta_0)) = \mathsf{gh}_1(y_t;\tilde{\theta}_t(\theta))$ for almost all $y_t \in \mathbb{R}$ then $\tilde{\theta}_t(\theta_0) = \tilde{\theta}_t(\theta)$ for any fixed $\boldsymbol{y}_{t-1} \in \mathbb{R}^p$.

Assume that (A32) holds. Then, using the notation $\phi_0 = \mu\left(1 - \boldsymbol{\phi}'\mathbf{1}_p\right)$, we first find that

$$\tilde{\mu}_t(\theta_0) = \tilde{\mu}_t \Leftrightarrow (\phi_{0,0} - \phi_0) + (\boldsymbol{\phi}_0 - \boldsymbol{\phi})'\,\boldsymbol{y} = 0, \tag{A33}$$

which implies that $\boldsymbol{\phi}_0 = \boldsymbol{\phi}$ and $\mu = \mu_0$.

Consider now the equations $\tilde{\sigma}_t^2(\theta_0) = \tilde{\sigma}_t^2$ and $\tilde{\gamma}_t(\theta_0) = \tilde{\gamma}_t$. Because $\boldsymbol{\phi}_0 = \boldsymbol{\phi}$ we have, using the notation $\boldsymbol{\gamma}_{p,0} = \gamma_0 \mathbf{1}_p$,

$$\left(\gamma_0 - \boldsymbol{\phi}_0'\boldsymbol{\gamma}_{0,p}\right) = \left(\gamma_0 - \boldsymbol{\phi}'\boldsymbol{\gamma}_{0,p}\right) = \gamma_0\left(1 - \boldsymbol{\phi}'\mathbf{1}_p\right),$$

where $\left(1 - \boldsymbol{\phi}'\mathbf{1}_p\right) \neq 0$ is guaranteed by the assumed stationarity. Solving the equations

$$\begin{cases} \tilde{\gamma}_t = \tilde{\gamma}_t(\theta_0) \\ \tilde{\sigma}_t^2 = \tilde{\sigma}_t^2(\theta_0) \end{cases} \Leftrightarrow \begin{cases} \tilde{\delta}_t\left(\delta_0 + \boldsymbol{\gamma}_{0,p}'\Gamma_{0,p}^{-1}\boldsymbol{\gamma}_{0,p}\right)\gamma = \tilde{\delta}_t(\theta_0)\left(\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\right)\gamma_0 \\ \tilde{\delta}_t\left(\delta_0 + \boldsymbol{\gamma}_{0,p}'\Gamma_{0,p}^{-1}\boldsymbol{\gamma}_{0,p}\right)\sigma^2 = \tilde{\delta}_t(\theta_0)\left(\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\right)\sigma_0^2, \end{cases}$$

for γ yields

$$\gamma = \gamma_0 \frac{\sigma^2}{\sigma_0^2}. \tag{A34}$$

Consider next the equations $\tilde{\delta}_t(\theta_0) = \tilde{\delta}_t$ and $\tilde{\sigma}_t^2(\theta_0) = \tilde{\sigma}_t^2$ which imply that

$$\begin{cases} \tilde{\delta}_t &= \tilde{\delta}_t(\theta_0) \\ \tilde{\sigma}_t^2 &= \tilde{\sigma}_t^2(\theta_0) \end{cases} \Leftrightarrow \left(\delta_0 + \boldsymbol{\gamma}_{0,p}'\Gamma_{0,p}^{-1}\boldsymbol{\gamma}_{0,p}\right)\sigma^2 = \left(\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\right)\sigma_0^2.$$

Because $\sigma^2\Gamma_p^{-1}$ depends only on $\boldsymbol{\phi}$ (see, e.g., Lütkepohl 2006) we obtain $\sigma^2\Gamma_p^{-1} = \sigma_0^2\Gamma_{0,p}^{-1}$. Therefore, using (A34), we find that

$$\left(\delta_0 + \boldsymbol{\gamma}_{0,p}'\Gamma_{0,p}^{-1}\boldsymbol{\gamma}_{0,p}\right)\sigma^2 = \left(\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\right)\sigma_0^2 \quad \Leftrightarrow \quad \delta_0\sigma^2 + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\sigma_0^2 = \delta\sigma_0^2 + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p\sigma_0^2,$$

where the right-hand side is equivalent to $\delta_0\sigma^2 = \delta\sigma_0^2$. Thus, we have

$$\delta + \boldsymbol{\gamma}_p'\Gamma_p^{-1}\boldsymbol{\gamma}_p = \frac{\sigma^2}{\sigma_0^2}\left(\delta_0 + \boldsymbol{\gamma}_{0,p}'\Gamma_{0,p}^{-1}\boldsymbol{\gamma}_{0,p}\right)$$

and

$$\tilde{\delta}_t(\theta_0) = \tilde{\delta}_t \Leftrightarrow \sqrt{\left(\delta_0 + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_0)'\Gamma_{0,p}^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_0)\right)} = \sqrt{\frac{\sigma^2}{\sigma_0^2}\left(\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu})'\Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})\right)}.$$
(A35)

It is already shown that $\mu_0 = \mu$ and hence we find that

$$\delta_0 + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_0)'\Gamma_{0,p}^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu}_0) = \frac{\sigma^2}{\sigma_0^2}\left(\frac{\sigma_0^4}{\sigma^4}\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu})'\Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})\right).$$

Substituting into (A35) yields

$$\tilde{\delta}_t(\theta_0) = \tilde{\delta}_t \Leftrightarrow \sqrt{\left(\left(\frac{\sigma_0^4}{\sigma^4}\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu})'\Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})\right)\right)} = \sqrt{\left(\delta + (\boldsymbol{y}_{t-1} - \boldsymbol{\mu})'\Gamma_p^{-1}(\boldsymbol{y}_{t-1} - \boldsymbol{\mu})\right)}$$
(A36)

and hence we get that $\sigma_0^2 = \sigma^2$. Putting the above together with $\tilde{\lambda} = \tilde{\lambda}_0 \Rightarrow \lambda = \lambda_0$ we have established that $E\left[\ell_t(\theta)\right] = E\left[\ell_t(\theta_0)\right]$ implies $\theta = \theta_0$.

Next we need to show that the expected log-likelihood function is continuous on Θ. As σ^2 is assumed to be bounded away from zero, by assuming the parameter space compact, all the terms on the first three rows of (17) are continuous functions of $\theta \in \Theta$ so we only need to consider the terms with Bessel functions on the last row.

The Bessel function is differentiable with respect to the argument (see, e.g. Watson 1944) and the order (see, e.g. Brychkov 2016) which are both continuous functions of θ and hence the log-likelihood function (17) is continuous on the parameter space Θ. By Theorem 4 in Appendix A.1.3 we show that the conditional log-likelihood function is bounded. Now, by continuity of the log-likelihood function and the dominated convergence theorem together we see that the expected log-likelihood function is continuous.

Thus, as $E\left[\ell_t(\theta)\right]$ is continuous on the compact set Θ, it has a maximizer θ_0 which was shown above to be unique. Compactness of the parameter space together with continuity of the log-likelihood function also imply that the maximum likelihood estimator $\hat{\theta}$ is measurable (see, e.g., Pötscher & Prucha 1997, Lemma 3.4).

A.6 Conditional distribution of the multivariate generalized hyperbolic distribution

Next we derive the conditional density function of the generalized hyperbolic distribution restricted as in (1). The conditional density has been previously derived by Blæsild (1981) who used the restriction $\det(\Gamma) = 1$ for identification.

Let the parameters $\theta = (\lambda, \delta, \mu, \Gamma, \gamma)$ be as defined in Section 2 and $d = d_1 + d_2 \geq 2$, $d_1, d_2 \in \mathbb{N}$, be the dimension of the distribution. We introduce the following shorthand notations to denote the parameters of the conditional distribution

$$\tilde{\lambda} = \lambda - d_2/2,$$

$$\tilde{\mu} = \mu_1 + (x_2 - \mu_2)'\Gamma_{22}^{-1}\Gamma_{21},$$

$$\tilde{\delta} = \sqrt{\left(\delta + (x_2 - \mu_2)'\Gamma_{22}^{-1}(x_2 - \mu_2)\right)\left(\delta + \gamma_2'\Gamma_{22}^{-1}\gamma_2\right)},$$

$$\tilde{\Gamma} = \frac{\sqrt{\delta + (x_2 - \mu_2)'\Gamma_{22}^{-1}(x_2 - \mu_2)}}{\sqrt{\delta + \gamma_2'\Gamma_{22}^{-1}\gamma_2}}\left(\Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}\right),$$

$$\tilde{\gamma} = \frac{\sqrt{\delta + (x_2 - \mu_2)'\Gamma_{22}^{-1}(x_2 - \mu_2)}}{\sqrt{\delta + \gamma_2'\Gamma_{22}^{-1}\gamma_2}}\left(\gamma_1 - \Gamma_{12}\Gamma_{22}^{-1}\gamma_2\right).$$

Partition the covariance matrix Γ to conformable blocks with dimensions d_1 and d_2 as

$$\Gamma = \begin{bmatrix} \Gamma_{11} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} \end{bmatrix}. \tag{A37}$$

Denote $\Gamma_{1\cdot2} = \Gamma_{11} - \Gamma_{12}\Gamma_{22}^{-1}\Gamma_{21}$ so that the inverse matrix of Γ can be expressed as

$$\Gamma^{-1} = \begin{bmatrix} I_{d_1} & 0 \\ -\Gamma_{22}^{-1}\Gamma_{21} & I_{d_2} \end{bmatrix} \begin{bmatrix} \Gamma_{1\cdot2}^{-1} & 0 \\ 0 & \Gamma_{22}^{-1} \end{bmatrix} \begin{bmatrix} I_{d_1} & -\Gamma_{12}\Gamma_{22}^{-1} \\ 0 & I_{d_2} \end{bmatrix},$$

where I_d is a d dimensional identity matrix. Partition also x and μ conformably to Γ as $x = (x_1, x_2)$ and $\mu = (\mu_1, \mu_2)$. Using this notation we can write

$$(x - \mu)\Gamma^{-1}(x - \mu) = (x_1 - \tilde{\mu})'\Gamma_{1\cdot2}^{-1}(x_1 - \tilde{\mu}) + (x_2 - \mu_2)'\Gamma_{22}^{-1}(x_2 - \mu_2)$$

where $\tilde{\mu} = \mu_1 + (x_2 - \mu_2)'\Gamma_{22}^{-1}\Gamma_{21}$. Also denoting $\gamma_{1\cdot2} = \gamma_1 - \Gamma_{12}\Gamma_{22}^{-1}\gamma_2$ we find the expression

$$\gamma'\Gamma^{-1}\gamma = \gamma_{1\cdot2}'\Gamma_{1\cdot2}^{-1}\gamma_{1\cdot2} + \gamma_2'\Gamma_{22}^{-1}\gamma_2.$$

Combining the above we get

$$\begin{aligned} & \left(\delta + (x - \mu)\Gamma^{-1}(x - \mu)\right)\left(\delta + \gamma'\Gamma^{-1}\gamma\right) \\ & = \left(\tilde{\delta} + (x_1 - \tilde{\mu})'\tilde{\Gamma}^{-1}(x_1 - \tilde{\mu})\right)\left(\tilde{\delta} + \tilde{\gamma}'\tilde{\Gamma}^{-1}\tilde{\gamma}\right). \end{aligned} \tag{A38}$$

The normalizing constant c in the density (1) has the expression

$$c = c_1 c_2 \frac{K_{\tilde{\lambda}}\left(\tilde{\delta}\right)}{\tilde{\delta}^{\frac{d_2}{2} - \lambda}}, \tag{A39}$$

where

$$c_1 = \frac{\left(\tilde{\delta} + \tilde{\gamma}'\tilde{\Gamma}^{-1}\tilde{\gamma}\right)^{\frac{d_1}{2} - \tilde{\lambda}}}{(2\pi)^{\frac{d_1}{2}}\sqrt{\det\left(\tilde{\Gamma}\right)}K_{\tilde{\lambda}}\left(\tilde{\delta}\right)} \quad \text{and} \quad c_2 = \frac{\left(\delta + \gamma_2'\Gamma_{22}^{-1}\gamma_2\right)^{\frac{d_2}{2} - \lambda}}{(2\pi)^{\frac{d_2}{2}}\sqrt{\det(\Gamma_{22})}K_{\lambda}\left(\delta\right)}.$$

The exponential function in the numerator of the density (1) can be written as

$$\exp\left(\gamma'\Gamma^{-1}(x - \mu)\right) = \exp\left(\tilde{\gamma}'\tilde{\Gamma}^{-1}(x_1 - \tilde{\mu})\right)\exp\left(\gamma_2'\Gamma_{22}^{-1}(x_2 - \mu_2)\right). \tag{A40}$$

Using equations (A38), (A39) and (A40) the generalized hyperbolic density can be expressed as

$$\begin{aligned} & \mathsf{gh}_d(x; \lambda, \delta, \mu, \Gamma, \gamma) \\ & = c_1 \frac{K_{\tilde{\lambda} - \frac{d_1}{2}}\left(\sqrt{\left(\tilde{\delta} + (x_1 - \tilde{\mu})'\tilde{\Gamma}^{-1}(x_1 - \tilde{\mu})\right)\left(\tilde{\delta} + \tilde{\gamma}'\tilde{\Gamma}^{-1}\tilde{\gamma}\right)}\right)}{\left(\sqrt{\left(\tilde{\delta} + (x_1 - \tilde{\mu})'\tilde{\Gamma}^{-1}(x_1 - \tilde{\mu})\right)\left(\tilde{\delta} + \tilde{\gamma}'\tilde{\Gamma}^{-1}\tilde{\gamma}\right)}\right)^{\frac{d_1}{2} - \tilde{\lambda}}} \exp\left(\tilde{\gamma}'\tilde{\Gamma}^{-1}(x_1 - \tilde{\mu})\right) \\ & \times c_2 \frac{K_{\lambda - \frac{d_2}{2}}\left(\sqrt{\left(\delta + \gamma_2'\Gamma_{22}^{-1}\gamma_2\right)\left(\delta + (x_2 - \mu_2)'\Gamma_{22}^{-1}(x_2 - \mu_2)\right)}\right)}{\left(\sqrt{\left(\delta + \gamma_2'\Gamma_{22}^{-1}\gamma_2\right)\left(\delta + (x_2 - \mu_2)'\Gamma_{22}^{-1}(x_2 - \mu_2)\right)}\right)^{\frac{d_2}{2} - \lambda}} \exp\left(\gamma_2'\Gamma_{22}^{-1}(x_2 - \mu_2)\right) \\ & = \mathsf{gh}_{d_1}(x_1; \tilde{\lambda}, \tilde{\delta}, \tilde{\mu}, \Gamma_t, \tilde{\gamma})\mathsf{gh}_{d_2}(x_2; \lambda, \delta, \mu_2, \Gamma_{22}, \gamma_2). \end{aligned} \tag{A41}$$

A.7 Additional figures

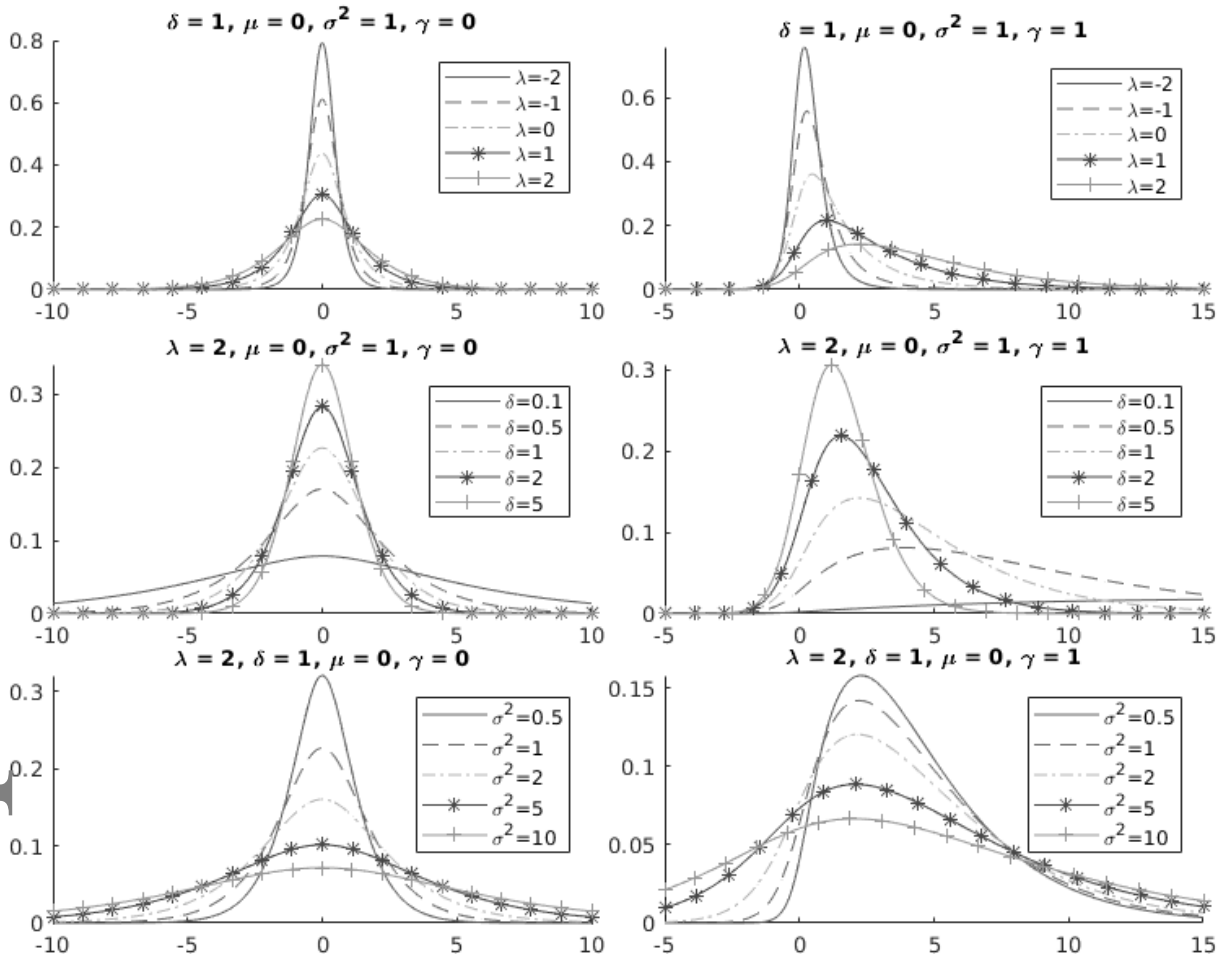### A.7.1 The generalized hyperbolic distribution

Figure 3: The effect of different parameter values on the density function of the one dimensional generalized hyperbolic distribution. On the left column the density is symmetric while for the right column $\gamma = 1$ is chosen.

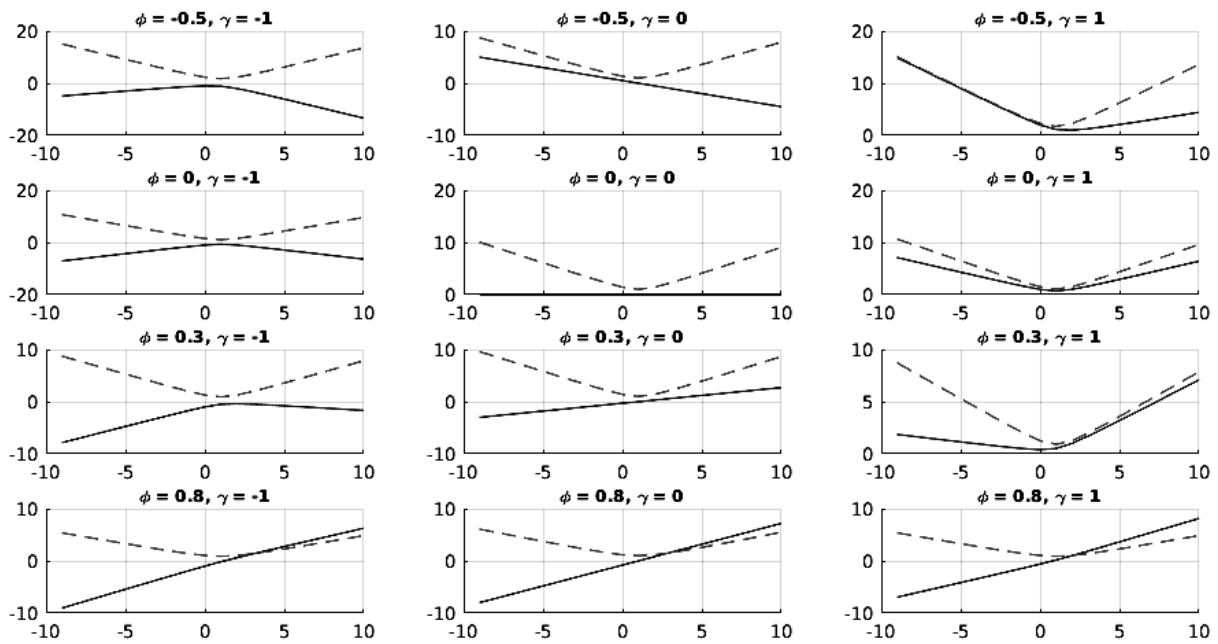### A.7.2 Conditional expectation and variance

Figure 4: The conditional expectation μ_t (solid line) and conditional covariance σ_t^2 (dashed line) with different values of γ and ϕ, and with $\delta = \sigma^2 = 1$ and $\lambda = \mu = 0$.