Beyond location and dispersion models: The Generalized Structural Time Series Model with Applications

Djennad, Abdelmajid and Rigby, Robert and Stasinopoulos, Dimitrios and Voudouris, Vlasios and Eilers, Paul

Statistics, Operational Research and Mathematics (STORM) Research Centre, London Metropolitan University, Department of Biostatistics, Erasmus University, The Netherlands, ABM Analytics Ltd, ESCP Europe Business School

12 March 2015

Beyond location and dispersion models: The Generalized Structural Time Series Model with Applications

Abdelmajid Djennad[a], Robert Rigby[a], Dimitrios Stasinopoulos[a], Vlasios Voudouris[a,b,d], Paul H. C. Eilers[c]

[a] *Statistics, Operational Research and Mathematics (STORM) Research Centre, London Metropolitan University.*
[b] *ABM Analytics Ltd, London, UK.*
[c] *Department of Biostatistics, Erasmus University, The Netherlands.*
[d] *ESCP Europe Business School, London, UK.*

Abstract

In many settings of empirical interest, time variation in the distribution parameters is important for capturing the dynamic behaviour of time series processes. Although the fitting of heavy tail distributions has become easier due to computational advances, the joint and explicit modelling of time-varying conditional skewness and kurtosis is a challenging task. We propose a class of parameter-driven time series models referred to as the generalized structural time series (GEST) model. The GEST model extends Gaussian structural time series models by a) allowing the distribution of the dependent variable to come from any parametric distribution, including highly skewed and kurtotic distributions (and mixed distributions) and b) expanding the systematic part of parameter-driven time series models to allow the joint and explicit modelling of all the distribution parameters as structural terms and (smoothed) functions of independent variables. The paper makes an applied contribution in the development of a fast local estimation algorithm for the evaluation of a penalised likelihood function to update the distribution parameters over time *without* the need for evaluation of a high-dimensional integral based on simulation methods.

Keywords: non-Gaussian parameter-driven time series, fast local estimation algorithm, time-varying skewness, time-varying kurtosis.

Preprint submitted to STORM (London Met University) Working Paper 1-15, 2015. *March 12, 2015*

1. Introduction

Many observable business and economic variables are characterized by high skewness and heavy tails. In many settings of empirical interest, the need for joint and explicit modelling of time-varying skewness and kurtosis (as a way of capturing the dynamic behavior of univariate time series processes) has become more apparent in recent years, particularly since the aftermath of the 2008 financial crisis. For example, since the early work of Fama (1965) and Mandelbrot (1963), the failure of the Gaussian distribution to accurately model (high frequencies) financial returns has been extensively discussed in econometric and financial literature. The departure from normality constitutes an important issue in quantifying market risk since it means that extreme movements in the variables are more likely than a normal distribution would predict. Although the fitting of heavy tail distributions has become easier due to computational advances, the joint and explicit modelling of time-varying conditional skewness and kurtosis is a challenging task.

Our main motivation is to develop a highly flexible structural time series modelling framework for the estimation, analysis and forecasting of the dynamic behaviour of univariate time series processes. In our empirical analysis we focus on developing different structural time series models to analyse (rather than forecast) a number of different data series such as the S&P 500 stock index, the Pound sterling and US dollar exchange rate and the number of Van drivers killed in the UK (count data). In particular, the analysis of the S&P 500 stock index aims to stochastically analyse the stylised facts of the series based upon time-varying estimates of the skew student-t (SST) distribution parameters, namely μ (location), σ (scale), ν (skewness) and τ (kurtosis) (see Appendix 7.1). This is a distinguishing feature of our approach. It allows the expansion of the systematic part of parameter-driven time series models to allow the joint and explicit stochastic modelling of all of the distribution parameters as structural terms and (if necessary) linear, non-linear and smooth functions of independent variables (see section 3). Thus, we propose a class of parameter-driven time series model referred to as the generalized structural time series (GEST) model. A fast new estimation approach through a local likelihood function Q given in Appendix 7.2 is explained (avoiding computing the likelihood function through the evaluation of a high-dimensional integral based on simulation methods such as importance sampling and Markov chain Monte Carlo; see Shephard and Pitt, 1997) and explains why the proposed class of parameter-driven GEST models have the potential to become popular in the applied statistics and econometrics literature.

The GEST modelling framework is entirely *parameter driven*. A key advance of the parameter-driven models is that they are flexible and can be easily adjusted in new settings; see Cox (1981) for a more detailed discussion of the two classes of (observation- and parameter-driven) models. In particular, in the GEST model, the structural terms for each distribution parameter of the conditional distribution can be a random walk or autoregressive term (of any order) and can include seasonal and/or leverage effects. The GEST modelling framework proposed here seems to be the first parameter-driven approach to allow the joint and explicit modelling of time-varying skewness and kurtosis. The GEST model allows parameters to vary over time as functions of lagged *distribution parameters* and exogenous variables. In the model estimation there is no need to evaluate a high-dimensional integral since model estimation is achieved by maximizing the local likelihood function Q generalizing Lee *et al.* (2006). This is a local estimation method, which is much faster in practice, and has been called penalized quasi likelihood (PQL) (Breslow and Clayton, 1993).

The alternative class of observation-driven models, by contrast, allows parameters to vary over time as functions of lagged *dependent variable* values and exogenous variables. By way of an example, the recently introduced Generalized Autoregressive Score (GAS) models (Creal *et al.*, 2013), also known as Dynamic Conditional Score (DCS) models, also provide a general framework for modelling time variation in parametric models as functions of lagged dependent variables and exogenous variables (see also Creal *et al.*, 2011). Thus, the GAS model is an observation-driven time series model assuming that we can compute the score of the parametric conditional observation density with respect to the time varying parameter. Although parameter-driven models, such as the GEST model, have two or more error terms while observation-driven models, such as the GAS model, have one error term, heuristically, for example, two error terms could have a smaller total variance than a single error term. Thus, in principle it is quite possible for the parameter-driven models to give better fits (when the focus of the analysis is on explaining stylized facts of the past) or forecasts. In any case, observation-driven models and parameter-driven models represent two different classes of models for the estimation, analysis and forecasting of the dynamic behaviour of time series processes.

Non-Gaussian parameter-driven time series models that rely on parametric theoretical conditional distributions offer a way of modelling economic and financial observations. Previous non-Gaussian time series models are based on the structural model for the mean and the stochastic volatility model for the variance. For example, the stochastic volatility (SV) model; see Shephard (2005) for a detailed discussion; the stochastic intensity models of Bauwens and Hautsch (2006) and Koopman et al. (2008); the Bayesian perspective of West *et al.* (1985) using Kalman filtering to model response observations from an exponential family distribution; the treatment of both filtering and smoothing non-Gaussian data based on approximating non-Gaussian densities by Gaussian mixtures; see Kitagawa (1987) and Kitagawa (1996). Durbin and Koopman (2000) model the mean of an exponential family distribution with a state space model and separately model the variance as a stochastic volatility model. Although non-Gaussian time series models relax the assumption of the conditional Gaussian distribution, they usually model the conditional mean and occasionally the conditional variance of the non-Gaussian distribution, but rarely both. Effectively, the systematic part of these models is limited to modelling explicitly the mean or variance which are usually two of the distribution parameters. In our GEST modelling framework for time-varying parameters, many of the existing parameter-driven models are encompassed. In addition, new models can be formulated and investigated.

Thus, there are several important points to make here as a way of justifying the qualification 'generalized' of the class of parameter-driven time series model proposed here:

- A GEST model is fitted using a local penalised likelihood estimation algorithm. We will argue that the proposed estimation algorithm is an effective choice, as it exploits the complete density structure, for introducing a fast driving mechanism for time-varying distribution parameters for parameter-driven time series models. Effectively, the GEST model expands the systematic part of parameter-driven time series models to allow the joint and explicit stochastic modelling of all of the distribution parameters as structural terms. The structural terms for each distribution parameter of the conditional distribution can be a random walk or autoregressive (of any order) and can include seasonal and/or leverage effects. Thus, extensions to time-varying skewness and kurtosis and other more complicated dynamics can be considered without introducing further complexities.

- The GEST model allows the use of a flexible parametric distribution $\mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ for the dependent variable, including highly skewed and/or kurtotic distributions such as the generalized beta type 2 (including the special case of the generalized Pareto) of McDonald and Xu (1995), power exponential of Nelson (1991), Johnson's SU of Johnson et al. (1994), Gumbel of Crowder et al. (1991), Box-Cox Cole-Green of Cole and Green (1992), Sinh-arcsinh of Jones and Pewsey (2009) and (skewed) t-family distributions. The GEST model also allows the use of discrete and mixed distributions (a mixed distribution is a continuous distribution with extra discrete points. e.g. a gamma distribution with possible values at zero).

- Because of the use of a parametric distribution, the use of a variety of diagnostic tools (from both the econometric or standard statistical literature) for model checking and selection is supported; see, for example, section 5.1.2.

- The GEST process has properly defined stochastic properties (see section 2 where we also illustrate a simulation example). We also present two theorems to show that under certain circumstances the GEST process is stationary with well defined marginal mean and variance. Understanding of the properties of a stochastic process helps to understand the evolution of the fitted conditional distribution through time, among other things.

The rest of the paper is organized as follows. In Section 2 we introduce the GEST process, and provide a simulated example of the GEST stochastic process. In Section 3 the full flexibility of the GEST model is presented. In Section 4 we describe how the GEST model is estimated using the proposed local maximum likelihood estimation algorithm. Section 5 illustrate the flexibility of the GEST model through the analysis (and model checking and selection) of the S&P 500 index where the skew Student t (SST) distribution has been selected - this example is also used to compare the flexibility of the GEST model with popular observation-driven models. We also demonstrate how the GEST model can be used for the modelling of stochastic volatility of the Pound sterling and US dollar exchange rate as well as the use of the Poison distribution to model the number of Van drivers killed in the UK. The latter two examples are used to compare the GEST model with other parameter-driven models. Section 6 provides a conclusion.

2. The GEST process

This section defines the GEST process essential for the modelling framework definition of the next section. The GEST process assumes that the random variable Y_t is derived from a distribution $\mathcal{D}(\boldsymbol{\theta}_t)$ with a probability (density) function $f_{Y_t}(y_t|\boldsymbol{\theta}_t)$ conditional on $\boldsymbol{\theta}_t$ where ${\boldsymbol{\theta}_t}^\top = (\theta_{1,t}, \ldots, \theta_{K,t})$ is a vector of distribution parameters for $f_{Y_t}()$.

Hence we may write $Y_t|\boldsymbol{\theta}_t \sim \mathcal{D}(\boldsymbol{\theta}_t)$ where each $\theta_{k,t}$ is generated by a random process given by

$$g_k(\theta_{k,t}) = \beta_{k,0} + \gamma_{k,t} \tag{1}$$

for $t = 1, 2, \ldots, T$, where

$$\gamma_{k,t} = \sum_{j=1}^{J_k} \phi_{k,j}\gamma_{k,t-j} + b_{k,t} \tag{2}$$

for $t = J+1, J+2, \ldots, T$, where, for $k = 1, 2, \ldots, K$, function $g_k()$ is a specified link function, $\gamma_{k,t}$ for $t = 1, 2, \ldots, T$ is an individual structural time series random process and $b_{k,t}$ are random errors, independent from each other mutually and serially, and normally distributed with expected values equal to zero and variance $\sigma^2_{b_k}$. Thus $\mathbf{b_k} \sim N_{n-J_k}(0, \sigma^2_{b_k}\mathbf{I}_{n-J_k})$, where $\mathbf{b_k}^\top = (b_{k,J_k+1}, \ldots, b_{k,T})$ for $k = 1, 2, \ldots, K$.

There are several important points to be made here about a GEST process.

- The probability distribution $f_{Y_t}(y_t|\boldsymbol{\theta}_t)$ can be a continuous or discrete distribution.
- For most practical applications, K, the number of parameters $\boldsymbol{\theta}_t$ in the distribution is less than or equal to four. We denote those four parameters as $\boldsymbol{\theta}_t = (\mu_t, \sigma_t, \nu_t, \tau_t)$ where μ_t is a time-varying location parameter, σ_t is a time-varying scale parameter and ν_t and τ_t are time-varying shape parameters, which may be related to the time-varying skewness and time-varying kurtosis of the distribution respectively.
- The link function $g_k()$ is used to ensure that the individual parameter is defined on a permissible range. For example, a log link for sigma, i.e. $g_2(\sigma_t) = \log(\sigma_t) = \gamma_{2,t}$, will ensure that $\sigma_t = \exp(\gamma_{2,t})$ is always positive.
- The $\phi_{k,j}$ in equation (2) are autoregressive parameters for the individual predictors $\gamma_{k,t}$ for $k = 1, 2, 3, 4$. Note that specific fixed values for $\phi_{k,j}$ for $j = 1, 2, \ldots, J_k$ replaces autoregressive terms with random walk terms for $\gamma_{k,t}$. For example setting $J_k = 1$ and $\phi_{k,1} = 1$ gives a random walk order 1, while setting $J_k = 2$, $\phi_{k,1} = 2$ and $\phi_{k,2} = -1$ gives a random walk order 2, for $k = 1, 2, 3, 4$.
- Note that the generation of the GEST process requires four sets of parameter values:
  (i) the constant parameters $\beta_{k,0}$ for $k = 1, 2, \ldots, K$.
  (ii) the AR parameters $\phi_{k,j}$ for $j = 1, 2, \ldots, J_k$ and $k = 1, 2, \ldots, K$,

(iii) the standard deviations σ_{b_k} of the white noises since $b_{k,t} \sim N\left(0, \sigma_{b_k}^2\right)$ for $k = 1, 2, \ldots, K$,

(iv) the initial starting values for the distribution parameters.

The GEST process is very flexible and can take familiar patterns of real data situation. Below we generate an example of a GEST stochastic process imitating the S&P500 stock index data analysed in section 5.1 by assuming that the $f_{Y_t}(y_t|\boldsymbol{\theta}_t)$ of the process is a skew Student t, $SST(\mu_t, \sigma_t, \nu_t, \tau_t)$, distribution described in Section 3 and Appendix 7.1. We simulate each of the distribution parameters of the $SST(\mu_t, \sigma_t, \nu_t, \tau_t)$ for $t = 1, 2, \ldots, n$ using a random walk order one process. The resulting GEST process is given by:

$$
\begin{aligned}
Y_t|\mu_t, \sigma_t, \nu_t, \tau_t &\sim SST(\mu_t, \sigma_t, \nu_t, \tau_t) \\
\mu_t &= \mu_{t-1} + b_{1,t} \\
\log(\sigma_t) &= \log(\sigma_{t-1}) + b_{2,t} \\
\log(\nu_t) &= \log(\nu_{t-1}) + b_{3,t} \\
\log(\tau_t - 2) &= \log(\tau_{t-1} - 2) + b_{4,t},
\end{aligned}
$$

where the simulated GEST process above is based upon equation (1) and (2) by setting $\beta_{k,0} = 0$, $J_k = 1$ and $\phi_{k,1} = 1$ for $k = 1, 2, 3, 4$ and $\gamma_{1,t} = \mu_t$, $\gamma_{2,t} = \log(\sigma_t)$, $\gamma_{3,t} = \log(\nu_t)$ and $\gamma_{4,t} = \log(\tau_t - 2)$. The initial values of the distribution parameters were $\mu_0 = 0$, $\sigma_0 = 1$, $\nu_0 = 1$, $\tau_0 = 5$ and the variances of the $b_{k,t}$ innovations were chosen to be $\sigma_{b_1}^2 = 0.0001$, $\sigma_{b_2}^2 = 0.0009$, $\sigma_{b_3}^2 = 0.0004$, and $\sigma_{b_4}^2 = 0.0004$. Note the link function $\log(\tau_t - 2)$ is used because, for the SST distribution, $\tau > 2$, ensuring it has a finite mean μ_t and standard deviation σ_t.

Figure 1 shows the simulated process y_t while Figure 2 shows the generated (black line) time-varying mean μ_t, time-varying standard deviation σ_t, time-varying skewness parameter ν_t, and time-varying reciprocal of the kurtosis parameter $1/\tau_t$. Note for the SST distribution $\nu_t < 1$ produces a negatively skewed distribution, while $\nu_t > 1$ produces a positively skewed distribution. The kurtosis increases as $\tau_t > 2$ decreases and $1/\tau_t$ increases. Figure 2 also shows the fitted GEST process (red lines) estimated using the estimation procedure described in Section 4

The GEST process can be non-stationary and potentially explosive by nature. This is not in general bad, since many economic and financial phenomena are themselves explosive. However, some statistical properties are difficult to establish unless additional assumptions about the nature of the GEST progress are made.

Here, we present two theorems to show that under certain circumstances the GEST process is stationary with well defined marginal mean and variance. In particular note Theorem 1 assumes: i) identity link function for μ_t and ii) log link function for σ_t. Theorem 2 assumes: i) log link function for μ_t and : ii) log link function for σ_t:

- ***Theorem 1***: *Let μ_t and $c\sigma_t$ (where c is a known constant) be, respectively, the conditional mean and standard deviation (assumed to exist) of the distribution* $Y_t|\mu_t, \sigma_t, \nu_t, \tau_t \sim$

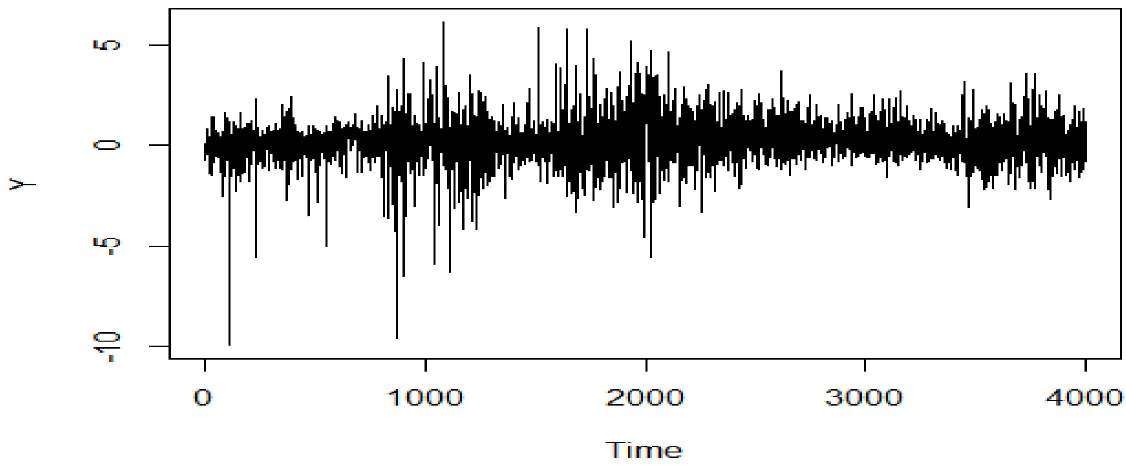


Figure 1: A GEST process realisation from a skew student t-distribution (SST)

$\mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ where $\mu_t = \beta_{1,0} + \gamma_{1,t}$ and $\log \sigma_t = \beta_{2,0} + \gamma_{2,t}$ and where

$$\gamma_{k,t} = \sum_{j=1}^{J_k} \phi_{k,j} \gamma_{k,t-j} + b_{k,t}$$

for $k = 1, 2$, where $b_{1,t}$ and $b_{2,t}$ are mutually and serially independently normally distributed with mean 0 and variances $\sigma^2_{b_1}$ and $\sigma^2_{b_2}$ respectively and hence

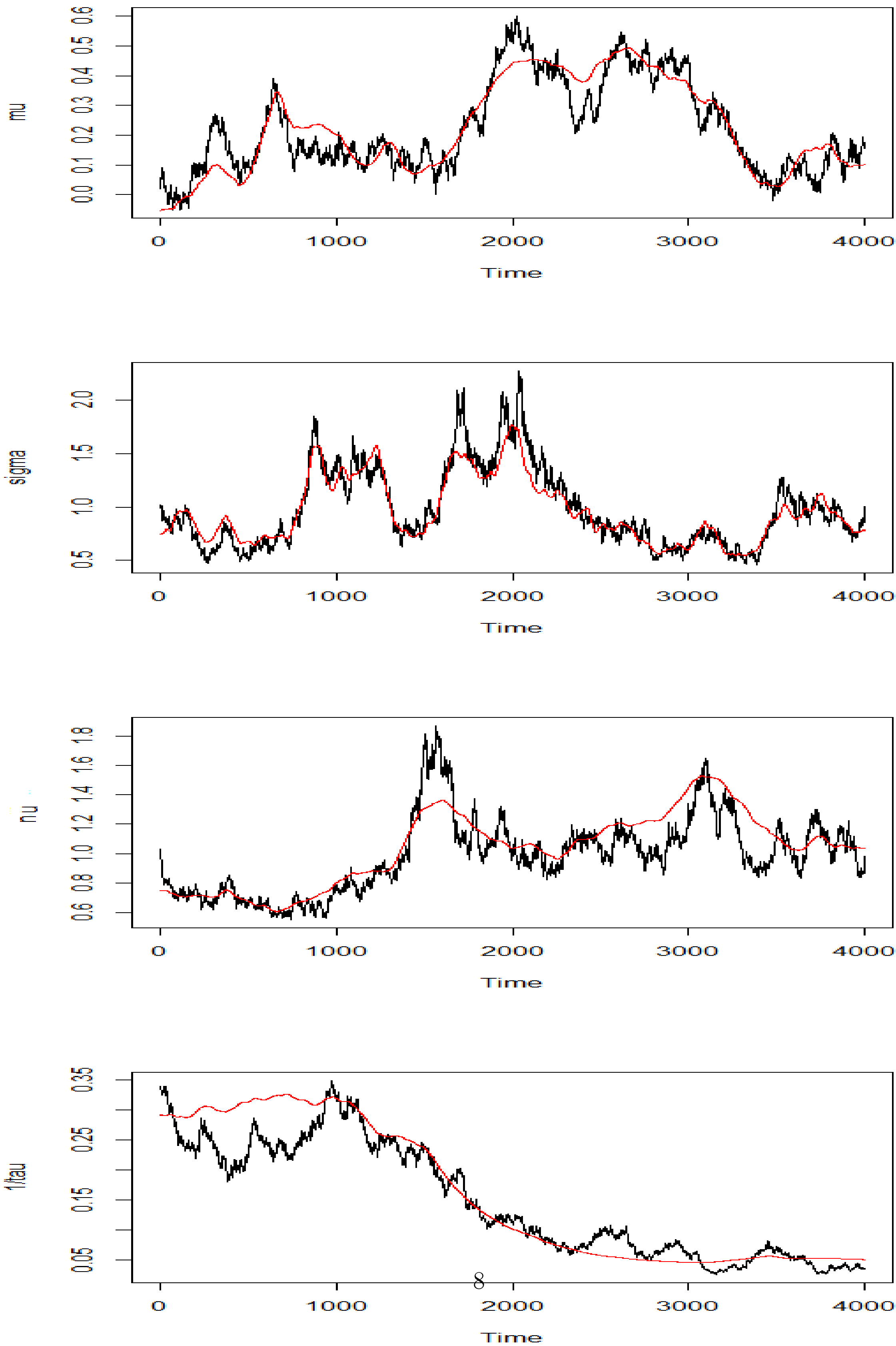
$$\begin{aligned} \gamma_{1,t} &= \Phi_1^{-1}(b_{1,t}) = \psi_1(b_{1,t}) \\ \gamma_{2,t} &= \Phi_2^{-1}(b_{2,t}) = \psi_2(b_{2,t}), \end{aligned}$$

assuming Φ_1 and Φ_1 are invertible, then the GEST process has a stationary mean and variance given by

$$\begin{aligned} E[Y_t] &= \beta_{1,0} \\ V[Y_t] &= S_1 \sigma^2_{b_1} + c^2 \exp\left(2\beta_{2,0} + 2 S_2 \sigma^2_{b_2}\right) \end{aligned}$$

respectively, where $S_k = 1 + \sum_{j=1}^{\infty} \psi^2_{k,j}$ for $k = 1, 2$ and where $\psi_k(B) = \Phi_k(B)^{-1} = 1 + \psi_{k,1} B + \psi_{k,2} B^2 + \ldots$ and provided $\Phi_k(B)$ is invertible, where $\Phi_k(B) = 1 - \phi_{k,1} B - \phi_{k,2} B^2 - \ldots - \phi_{k,J_k} B^J$ and B is the backshift time operator, $By_t = y_{t-1}$.

Appendix 7.3.1 gives the proof for Theorem 1. Note that Theorem 1 is not affected by the form of the model for ν_t and τ_t. Also Theorem 1 applies to any distribution $\mathcal{D}$ in which μ_t and $c\sigma_t$ are respectively the mean and standard deviation of $\mathcal{D}$. In



S

Figure 2: The actual realisations (in black) for μ, σ, ν and $1/\tau$ for the GEST process shown in Figure 1 and the fitted process estimated using the estimation procedure described in Section 4.

particular Theorem 1 applies to the normal, $NO(\mu, \sigma)$, skew student t, $SST(\mu, \sigma, \nu, \tau)$, Power Exponential, $PE(\mu, \sigma, \nu)$, t-family parameterized so σ is the standard deviation, $TF2(\mu, \sigma, \nu)$, and Johnson's Su, $JSU(\mu, \sigma, \nu, \tau)$, distributions, where $c = 1$. It also applies to the logistic, $LO(\mu, \sigma)$, Gumbel, $GU(\mu, \sigma)$, and Reverse Gumbel, $RG(\mu, \sigma)$, where $c \neq 1$ (see Stasinopoulos *et al.*, 2008, for the parametrization of the probability density functions of the distributions).

- ***Theorem 2***: *Let the distribution of* $Y_t|\mu_t, \sigma_t, \nu_t, \tau_t \sim \mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ *have a mean* μ_t *and variance* $v(\mu_t, \sigma_t)$ *where* $\log \mu_t = \beta_{1,0} + \gamma_{1,t}$, *and* $\log \sigma_t = \beta_{2,0} + \gamma_{2,t}$ *and where*

$$\begin{aligned} \gamma_{1,t} &= \Phi_1^{-1}(b_{1,t}) = \psi_1(b_{1,t}) \\ \gamma_{2,t} &= \Phi_2^{-1}(b_{2,t}) = \psi_2(b_{2,t}) \end{aligned}$$

as defined in Theorem 1, assuming Φ_1 *and* Φ_2 *are invertible, then the following give marginal means and variances of the related process:*

  a) $E[Y_t] = E[\mu_t] = \exp\left(\beta_{1,0} + \frac{1}{2}S_1\sigma_{b_1}^2\right)$

  b) $V[Y_t] = V[\mu_t] + E[v(\mu_t, \sigma_t)]$

  c) $V[\mu_t] = \exp(2\beta_{1,0})\left[\exp\left(2S_1\sigma_{b_1}^2\right) - \exp\left(S_1\sigma_{b_1}^2\right)\right]$

  d) $E[\mu_t^r] = \exp\left(r\beta_{1,0} + \frac{1}{2}r^2S_1\sigma_{b_1}^2\right)$

  e) $E[\sigma_t^r] = \exp\left(r\beta_{2,0} + \frac{1}{2}r^2S_2\sigma_{b_2}^2\right)$

for $r > 0$.

Appendix 7.3.2 gives the proof for Theorem 2 together with a corollary for Theorem 2 providing the marginal variance of Y_t for four conditional distributions for Y_t, the negative binomial type I and type II, $NBI(\mu, \sigma)$, $NBII(\mu, \sigma)$, the gamma, $GA(\mu, \sigma)$, and inverse Gaussian, $IG(\mu, \sigma)$, distributions (see Stasinopoulos *et al.*, 2008, for the parametrization of the probability (density) functions of the distributions).

3. The GEST model

The previous section provides us with a general stochastic process with potential of modelling a variety of situations including continuous or discrete variables with possibly high or low kurtosis and/or positive or negative skewness. Here we introduce a general statistical model in which all the parameters of the assumed distribution of the dependent variable can be explicitly modelled as structural terms and (if necessary) functions of explanatory variables.

Let Y_t be the response variable for $t = 1, 2, \ldots, T$ then the GEST model is defined as:

$$
\begin{array}{rcl}
Y_t|\mu_t, \sigma_t, \nu_t, \tau_t & \sim & \mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t) \\
g_1(\mu_t) = \eta_{1,t} & = & \mathbf{x}_{1,t}^\top \boldsymbol{\beta}_1 + \gamma_{1,t} \\
g_2(\sigma_t) = \eta_{2,t} & = & \mathbf{x}_{2,t}^\top \boldsymbol{\beta}_2 + \gamma_{2,t} \\
g_3(\nu_t) = \eta_{3,t} & = & \mathbf{x}_{3,t}^\top \boldsymbol{\beta}_3 + \gamma_{3,t} \\
g_4(\tau_t) = \eta_{4,t} & = & \mathbf{x}_{4,t}^\top \boldsymbol{\beta}_4 + \gamma_{4,t}
\end{array} \tag{3}
$$

where $\mathcal{D}$ represents the conditional distribution of the response variable, g_k is a known link function (e.g., identity or log link function), $\boldsymbol{\beta}_k$ is a parameter vector of length p_k and the $\mathbf{x_{k,t}}$ are the explanatory terms and the $\gamma_{k,t}$ are defined as in equation (2) for $k = 1, 2, 3, 4$.

Regarding the GEST model, it is important to note that:

- The response variable distribution $\mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ can be a continuous or discrete distribution.
- Typically the linear term $\mathbf{x}_{k,t}^\top \boldsymbol{\beta}_k$ could include the constant, continuous or categorical explanatory variables and possibly a linear term in time or a fixed seasonal effect, for $k = 1, 2, 3, 4$.
- The explanatory variables can be different for each distribution parameter μ_t, σ_t, ν_t and τ_t.
- To account for non-linearities in the relationship between the parameters of the distribution and the explanatory variables, model (3) can be extended to include non-linear and smooth non-parametric models for the distribution parameters μ_t, σ_t, ν_t and τ_t as in Rigby and Stasinopoulos (2005). For example equation (3) can be amended to $\eta_{k,t} = \mathbf{x}_{k,t}^\top \boldsymbol{\beta}_k + \sum_{j=1}^{J_k} s_j(x_j) + \gamma_{k,t}$ where the $s_j()$ are smooth functions e.g. P-splines of Eilers and Marx (1996).
- A distribution parameter model can be extended to include a seasonal effect (with M seasons)
$$g_k(\theta_{k,t}) = \eta_{k,t} = \mathbf{x}_{k,t}^\top \boldsymbol{\beta}_k + \gamma_{k,t} + s_{k,t}$$
where $\gamma_{k,t}$ is given by (2) and
$$s_{k,t} = -\sum_{m=1}^{M-1} s_{k,t-m} + w_t$$
- The random effects $\gamma_{k,t}$ can be extended to include *persistent* explanatory variable effects,
$$\gamma_{k,t} = \sum_{j=1}^{J_k} \phi_{k,j} \gamma_{k,t-j} + v_{k,t}^\top \delta_k + b_{k,t} \tag{4}$$

where δ_k is a parameter vector of length q_k and explanatory variable vector $v_{k,t}$ is of length q_k. This term is used in the S&P 500 analysis of Section 5.1 for modelling the leverage effect using asymmetric stochastic volatility (see for example Asai and McAleer, 2005; Omori *et al.*, 2007).

An important characteristic of the GEST model is the integration of regression-type and time-series-type of models for all the distribution parameters (μ_t, σ_t, ν_t and τ_t) of the assumed parametric conditional distribution $\mathcal{D}$ of the response variable, allowing the location, scale, skewness and kurtosis of the conditional distribution $\mathcal{D}$ to change over time. Also the distribution $\mathcal{D}$ can be any parametric (continuous or discrete) distribution and is not necessarily restricted to the assumption of the exponential family distribution.

By way of two examples, we present below two GEST models by specifying two different distributions, namely the Gaussian distribution and the Skew student t distribution.

4. Model estimation

4.1. Introduction

The GEST model, defined by equation (3), has four distinct sets of parameters:

(a) $\boldsymbol{\beta}^\top = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top\right)$ the betas,

(b) $\boldsymbol{\gamma}^\top = \left(\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top, \boldsymbol{\gamma}_3^\top, \boldsymbol{\gamma}_4^\top\right)$ the gammas,

(c) $\boldsymbol{\phi}^\top = \left(\boldsymbol{\phi}_1^\top, \boldsymbol{\phi}_2^\top, \boldsymbol{\phi}_3^\top, \boldsymbol{\phi}_4^\top\right)$, the phis and

(d) $\boldsymbol{\sigma}_b^\top = (\sigma_{b_1}, \sigma_{b_2}, \sigma_{b_3}, \sigma_{b_4})$, the standard deviations of the normal variables $b_{k,t}$ for $k = 1, 2, 3, 4$,

where $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}_b$ are referred as the *hyperparameters*.

The joint distribution for all the components of the GEST model is given by:

$$f(\mathbf{y}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \boldsymbol{\phi}, \boldsymbol{\sigma}_b) = f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{\sigma}_b) f(\boldsymbol{\phi}) f(\boldsymbol{\sigma}_b) f(\boldsymbol{\beta}) \tag{5}$$

where

$$f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) = \prod_{t=1}^{T} f(y_t|\mu_t, \sigma_t, \nu_t, \tau_t)$$

is the likelihood function, based on the assumed conditional distribution $\mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ for Y_t in equation (3), $f(\boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{\sigma}_b)$ is a product of four independent multivariate normal prior distributions for $\boldsymbol{\gamma}_k$ for $k = 1, 2, 3, 4$ (assuming prior independence between the $\boldsymbol{\gamma}_k$). The terms $f(\boldsymbol{\phi}), f(\boldsymbol{\sigma}_b)$ and $f(\boldsymbol{\beta})$ are independent prior distributions for the $\boldsymbol{\phi}$, $\boldsymbol{\sigma}_b$ and $\boldsymbol{\beta}$ parameters respectively and assume independence of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ (given $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}_b$) . In a fully Bayesian inference, the posterior distribution of $\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{\phi}$ and $\boldsymbol{\sigma}_b$ can be obtained by using Markov chain Monte Carlo sampling as in Fahrmeir and Tutz (2001).

4.2. The GEST local estimation algorithm

Assuming a uniform prior for $\boldsymbol{\beta}$ from equation (5), we have the posterior distribution of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$,

$$f(\boldsymbol{\beta}, \boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{\sigma}_b, \mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}) f(\boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{\sigma}_b). \tag{6}$$

Maximizing equation (6) gives posterior mode estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, given $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}_b$. By taking the log of equation (6), maximizing (6) is equivalent to maximizing the *extended (or joint) log likelihood function* (Lee *et. al.* (2006)) for the parameters $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$, given fixed $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}_b$, defined by:

$$l_e = l + \log f\left(\boldsymbol{\gamma}|\boldsymbol{\phi}, \boldsymbol{\sigma}_b\right) \tag{7}$$

where

$$l = \log f\left(\boldsymbol{y}|\boldsymbol{\beta}, \boldsymbol{\gamma}\right) = \sum_{t=1}^{n} \log f(y_t|\mu_t, \sigma_t, \nu_t, \tau_t) \tag{8}$$

is the log likelihood function.

For the GEST estimation algorithm, we extend Rigby and Stasinopoulos (2005) Appendix B2 and C by introducing step (a)(ii)(II) to estimate the hyperparameters $\sigma^2_{b_k}$ and $\phi_{k,j}$ for $k = 1, 2, 3, 4$. In particular, Rigby and Stasinopoulos (2005) maximise the extended likelihood in (7), given fixed hyperparameters. Thus, the algorithm provides posterior mode estimates of the sets of parameters of $\boldsymbol{\beta}^\top = \left(\boldsymbol{\beta}_1^\top, \boldsymbol{\beta}_2^\top, \boldsymbol{\beta}_3^\top, \boldsymbol{\beta}_4^\top\right)$, and $\boldsymbol{\gamma}^\top = \left(\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top, \boldsymbol{\gamma}_3^\top, \boldsymbol{\gamma}_4^\top\right)$ by maximizing the extended log likelihood. Note that below we used the notation $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4) = (\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$.

The GEST algorithm

(A) initialise $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \boldsymbol{\theta}_4)$, i.e. $(\boldsymbol{\mu}, \boldsymbol{\sigma}, \boldsymbol{\nu}, \boldsymbol{\tau})$, and set initial $\boldsymbol{\gamma}_k = 0$ for $k = 1, 2, 3, 4$

(B) start the *outer cycle* in order to fit each of the distribution parameter vectors $\boldsymbol{\theta}_k$, for $k = 1, 2, 3, 4$ sequentially until convergence [where, $\boldsymbol{\theta}_1 = \boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_T)^\top$, $\boldsymbol{\theta}_2 = \boldsymbol{\sigma}$, $\boldsymbol{\theta}_3 = \boldsymbol{\nu}$, $\boldsymbol{\theta}_4 = \boldsymbol{\tau}$],

   (a) start the following *inner cycle* (or "local scoring") for each iteration of the outer cycle in order to fit one of the distribution parameter vectors, $\boldsymbol{\theta}_k$

      (i) evaluate the current pseudo response variable $\mathbf{z}_k$ and current weights $\mathbf{W}_k$ (both defined in section 4.3)

      (ii) start the Gauss-Seidel (or "backfitting") algorithm

(I) estimate $\boldsymbol{\beta}_k$ by regressing the current partial residuals $\boldsymbol{\epsilon}_{0k} = \mathbf{z}_k - \boldsymbol{\gamma}_k$ against design matrix $\mathbf{X}_k$ using current weights $\mathbf{W}_k$.

(II) *estimate the hyperparameters σ_b^2 and $\boldsymbol{\phi}$ by maximising their local likelihood function, and then estimate $\boldsymbol{\gamma}$ using equation (11)*

(iii) end the Gauss-Seidel algorithm on convergence of $\boldsymbol{\beta}_k$ and $\boldsymbol{\gamma}_k$

(iv) update $\boldsymbol{\theta}_k$ and $\boldsymbol{\eta}_k = g(\boldsymbol{\theta}_k)$.

(b) end the inner cycle on convergence of $\boldsymbol{\theta}_k$.

(C) end the outer cycle when the global deviance ($= -2 * l$) of the estimated model converges.

This new step (a)(ii)(II) is described in section 4.3 and in Appendix 7.2. It is important to emphasise here that the outer cycle fits a specific distribution parameter vector (e.g. $\boldsymbol{\mu}$), by fixing the other distribution parameter vectors (e.g. $\boldsymbol{\sigma}, \boldsymbol{\nu}$ and $\boldsymbol{\tau}$) to their current maximum values, and the inner cycle uses a "local scoring" or Newton algorithm resulting in an iterative reweighted backfitting. Furthermore, the Gauss-Seidel algorithm in (B)(a)(ii) above is called the "backfitting" algorithm by Hastie and Tibshirani (1990) and Hastie *et al.* (2009).

4.3. Estimation of hyperparameters $\boldsymbol{\phi}$ and $\boldsymbol{\sigma}_b$ in step (a)(ii)(II) of the GEST algorithm

When the random effects hyperparameters are unknown, then in principle they can be estimated by maximizing the marginal likelihood obtained by integrating out $\boldsymbol{\gamma}$ (and also $\boldsymbol{\beta}$ for Restricted Maximum Likelihood Estimation) from l_e. This is in practice intractable so a Laplace approximation (Tierney and Kadane, 1986; Evans and Swartz, 2000, p.62) can be used to approximate the marginal likelihood, see for example Breslow and Clayton (1993), Lee and Nelder (1996), Pinheiro and Bates (2000), Pawitan (2001), p.466-467, and Rigby and Stasinopoulos (2005), Section A.2.3.

We refer to this method as the *global* estimation procedure (for the random effects hyperparameters) to distinguish it from the *local* estimation method used in this paper and described below. The local estimation method, which is much faster in practice, is based on ideas from Pinheiro and Bates (2000), Venables and Ripley (2002), p.297-298, Wood (2006), Section 6.4, and Rigby and Stasinopoulos (2013). The local method has been called penalized quasi likelihood (PQL). Note, however, that the local method described below and used in this paper uses penalized likelihood. Furthermore, the local estimation method produces almost identical results with the global estimation method in our experience.

The local estimation procedure (for the random effects hyperparameters) is step (a)(ii)(II) in the GEST algorithm. During the fit of each one of μ, σ, ν, and τ, the corresponding structural parameters [i.e. $\sigma_{b_k}^2$ and $\boldsymbol{\phi}_k$ for $k = 1, 2, 3, 4$ where $\boldsymbol{\phi}_k^\top = (\phi_{k,1}, \phi_{k,2}, \ldots, \phi_{k,J_k})$], are estimated by the internal (i.e. local) marginal maximum likelihood estimation procedure outlined below and given in detail in Appendix 7.2. To simplify the notation the subscript k is dropped from equation (3) so θ_t now represents any one of the parameters $(\mu_t, \sigma_t, \nu_t, \tau_t)$:

$$g(\theta_t) = \eta_t = \mathbf{x}_t^\top \boldsymbol{\beta} + \gamma_t \tag{9}$$

for $t = 1, 2, \ldots, T$, where γ_t is defined by (2) with subscript k omitted.

On the predictor scale (9), in the structural model fitting part of the backfitting algorithm [i.e. step (a)(ii)(II) of the GEST fitting algorithm] the following local approximate internal model is used:

$$\boldsymbol{\epsilon} = \boldsymbol{\gamma} + \boldsymbol{e}$$

where $\boldsymbol{e} \sim N_T(\mathbf{0}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} = \sigma_e^2 \mathbf{W}^{-1}$, $\boldsymbol{\epsilon} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta}$ are the current partial residuals, $\mathbf{z} = \boldsymbol{\eta} + \mathbf{W}^{-1}\mathbf{u}$ is the current pseudo response variable, $\mathbf{W}$ is a diagonal matrix of current weights given by one of the following $-\frac{\partial^2 \ell}{\partial \eta \partial \eta^\top}$, $-E\left[\frac{\partial^2 \ell}{\partial \eta \partial \eta^\top}\right]$ or $\left(\frac{\partial \ell}{\partial \eta}\right)^2$, i.e. the observed information, the expected information or the squared score function, depending respectively on whether a Newton-Raphson, Fisher scoring or quasi-Newton algorithm is used, and $\mathbf{u} = \frac{\partial \ell}{\partial \eta}$. The algorithm given in Appendix 7.2 maximises the local likelihood function Q, given below, directly over the structural model parameters $\boldsymbol{\alpha} = (\sigma_b^2, \boldsymbol{\phi})$, where $\boldsymbol{\phi}^\top = (\phi_1, \phi_2, \ldots, \phi_J)$, using a numerical algorithm.

For $t = 1, 2, \ldots, T$,

$$\epsilon_t = \gamma_t + e_t$$

where γ_t is given in (2). Hence,

$$b_t = \gamma_t - \sum_{j=1}^{J} \phi_j \gamma_{t-j}$$

for $t = J+1, J+2, \ldots, T$.

Note that Pawitan (2001) shows a *computational equivalence* between the usual estimation of random effects and their parameters (i.e. integrating out the random effects and maximizing over the fixed and random parameters) and maximizing an objective function Q (in the form of an adjusted profile extended likelihood for the random effects parameters). Given the absence of fixed effects locally, the Q function, maximized over the random effects $\boldsymbol{\gamma}$ given the random effects parameters $\boldsymbol{\alpha} = (\sigma_b^2, \boldsymbol{\phi})$, gives the local likelihood function of $\boldsymbol{\alpha}$. Here locally the random effects are $\boldsymbol{\gamma}$ with parameters $\boldsymbol{\alpha}$ and, generalizing Lee *et al.* (2006), p.277-279, the local function Q is given by

$$Q = \log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}) + \log f(\boldsymbol{\gamma}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right| + \frac{T}{2}\log 2\pi \tag{10}$$

where T is the number of observations, $\boldsymbol{\epsilon} = \mathbf{z} - \mathbf{X}\boldsymbol{\beta}$ is the vector of current partial residuals, $\mathbf{e} \sim N_T(0, \sigma_e^2 \mathbf{W}^{-1})$, where $\mathbf{e} = (e_1, e_2, \ldots, e_T)^\top$, $\boldsymbol{\Sigma} = \sigma_e^2 \boldsymbol{W}^{-1}$ and $\mathbf{D}$ is defined in Appendix 7.2. Note that assuming diffuse uniform priors for $(\gamma_1, \ldots, \gamma_J)$ in Q gives

$$f(\boldsymbol{\gamma}) = \prod_{t=J+1}^{T} f(\gamma_t | \boldsymbol{\gamma}_{t-1}) \equiv \prod_{t=J+1}^{T} f(b_t) = f(\mathbf{b})$$

where $\boldsymbol{\gamma}_{t-1} = (\gamma_1, \gamma_2, \ldots, \gamma_{t-1})$, $\mathbf{b} = (b_{J+1}, b_{J+2}, \ldots, b_T)^\top$, $\mathbf{b} \sim N_{T-J}(0, \sigma_b^2 \mathbf{I}_{T-J})$. Maximising Q over $\boldsymbol{\alpha} = (\sigma_b^2, \boldsymbol{\phi})$ gives estimates of σ_b^2 and $\boldsymbol{\phi}$. Then, from Appendix 7.2, $\boldsymbol{\gamma}$ is estimated effectively by smoothing the partial residuals using

$$\boldsymbol{\gamma} \;=\; \left[\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2} \mathbf{D}^\top \mathbf{D}\right]^{-1} \boldsymbol{\Sigma}^{-1} \; \boldsymbol{\epsilon}. \tag{11}$$

The total effective degrees of freedom of the fitted model, df, combines those of the models for $\theta_1, \theta_2, \theta_3$ and θ_4, i.e. μ, σ, ν and τ, given by df_1, df_2, df_3 and df_4 respectively. Hence,

$$df = df_1 + df_2 + df_3 + df_4$$

where

$$df_k = p_k + d_k$$

for $k = 1, 2, 3, 4$, and p_k is the length of $\boldsymbol{\beta}_k$, while d_k, the effective degrees of freedom for the random effects $\boldsymbol{\gamma}_k$, is obtained from d in Appendix A, ie

$$d_k = tr\left[\hat{\mathbf{B}}_k\right] = tr\left\{\left[\hat{\boldsymbol{\Sigma}}_k^{-1} + \hat{\sigma}_{b_k}^{-2} \hat{\mathbf{D}}_k^\top \hat{\mathbf{D}}_k\right]^{-1} \hat{\boldsymbol{\Sigma}}_k^{-1}\right\}$$

where $\hat{\mathbf{B}}_k, \hat{\boldsymbol{\Sigma}}_k, \hat{\mathbf{D}}_k$ and $\hat{\sigma}_{b_k}^2$ are the values of $\mathbf{B}, \boldsymbol{\Sigma}, \mathbf{D}$ (given in Appendix 7.2) and $\sigma_{b_k}^2$ for the model for θ_k on convergence of the GEST model fitting procedure.

5. Illustrating examples

In this section we present a number of GEST models as a way of illustrating the flexibility of the proposed class of parameter-driven time series models. We also compare the GEST model we other popular observation-driven and parameter-driven models.

First, we present a detailed analysis of the daily returns of the S&P 500 stock index where we illustrate the flexibility of the GEST model. We also demonstrate the use of variety of diagnostic tools (from both the econometric or standard statistical literature) for GEST model checking and selection. Then we present briefly how to fit a GEST stochastic volatility model to pound/dollar exchange rates' returns and how to fit a Poisson model to assess the effect of road safety measures on the development in traffic safety over time.

5.1. Standard and Poor 500 stock index

In this example, the GEST model is illustrated by an application to financial daily returns of the S&P 500 stock index. The data, taken from Yahoo.finance website, are daily closing prices of the S&P 500 stock index from 02/01/1980 to 31/12/2012, i.e. 8324 daily observations. Thus, following Harvey (1985) and Harvey and Jaeger (1993) in terms of using a structural time series model to stochastically estimate 'stylised facts' of time series observations, the flexibility of the GEST model is demonstrated here in terms of establishing a set of 'stylised facts' (rather than forecasting) associated with the returns of the S&P 500 index based upon time-varying estimates of the distribution parameters μ_t, σ_t, ν_t and τ_t.

We compare different GEST models and select the best model using the Akaike information criterion (AIC). Then, in sections 5.1.1 and 5.1.2, we compare the chosen GEST model with the APARCH model using AIC and normalized probability integral transform (normalized PIT) residuals to assess the adequacy of each fitted model. It is important to note that we are not comparing the GEST model with another parameter-driven time series model as we are not aware of any other parameter-driven time series model capable of modelling all the distribution parameters. Thus, the use of the APARCH model enable us to compare the GEST model with a popular observation-driven model.

The skew student-t (SST) distribution, which is a skew heavy tailed distribution, is used in the GEST model for the conditional distribution of the S&P 500 daily returns.

Let P_t be the price at time t and $y_t = ln(P_t/P_{t-1}) * 100$ be the return of the S&P 500 over the period 02/01/1980 to 31/12/2012. The conditional probability density function $f(y_t|\mu_t, \sigma_t, \nu_t, \tau_t)$ of the S&P 500 index returns y_t is modelled by a skew t-distribution, SST described in Section 3 and Appendix 7.1, using the GEST process to model the stochastic volatility, stochastic skewness and stochastic kurtosis of the returns using an autoregressive model for $\log(\sigma_t)$, and a random walk for $\log(\nu_t)$ and $\log(\tau_t - 2)$. The model is given by

$$\begin{aligned}
Y_t|\mu_t, \sigma_t, \nu_t, \tau_t &\sim SST(\mu_t, \sigma_t, \nu_t, \tau_t) \\
\mu_t &= \beta_{1,0} \\
\log(\sigma_t) &= \beta_{2,0} + \gamma_{2,t} \\
\log(\nu_t) &= \beta_{3,0} + \gamma_{3,t} \\
\log(\tau_t - 2) &= \beta_{4,0} + \gamma_{4,t}
\end{aligned} \quad (12)$$

where

$$\begin{aligned}
\gamma_{2,t} &= \phi_1 \gamma_{2,t-1} + b_{2,t} \\
\gamma_{3,t} &= \gamma_{3,t-1} + b_{3,t} \\
\gamma_{4,t} &= \gamma_{4,t-1} + b_{4,t}
\end{aligned}$$

This is model `m1` in Table 1. Note that $\beta_{2,0}$ is the reversion line for $\log(\sigma_t)$ around which the autoregressive process $\gamma_{2,t}$ varies.

In model m2, asymmetric stochastic volatility terms were included in the model for $\log(\sigma_t)$ to account for leverage effect giving

$$\gamma_{2,t} = \phi_1 \gamma_{2,t-1} + \delta_1 v_{1,t-1} + \delta_2 v_{2,t-1} + b_{2,t} \tag{13}$$

where $v_{1,t-1} = \text{arcsinh}(y_{t-1})$ (if $y_{t-1} < 0$) and $v_{2,t-1} = \text{arcsinh}(y_{t-1})$ (if $y_{t-1} >= 0$).

The use of the transformed $\text{arcsinh}(y_{t-1})$ rather than just y_{t-1} was found to reduce occasional extreme spikes in the fitted volatility. This is model m2 in Table 1. Submodels of model m2 where ν and/or τ is constant over time are also given in Table 1. Effectively, Table 1 compares between the submodels of the GEST model m2 to check whether we need a random walk (rw) model for skewness and/or a random walk model for kurtosis or just a constant for one or both parameters. Therefore, we fit five submodels to the S&P 500 data and summarise the Akaike information criterion (AIC) (Akaike, 1983). Note that the "ar with lev." model for σ_t is given by equation (13).

Table 1: Submodels of model m2

model	μ_t	σ_t	ν_t	τ_t	df	AIC
m1	const	ar	rw	rw	349.4901	22005.87
m2	const	ar with lev.	rw	rw	311.0655	**21862.74**
m3	const	ar with lev.	rw	const	334.4506	21871.31
m4	const	ar with lev.	const	rw	299.5491	21874.33
m5	const	ar with lev.	const	const	330.4522	21879.61

The model selected with minimum AIC is m2 giving

$$\begin{aligned} Y_t|\mu_t, \sigma_t, \nu_t, \tau_t &\sim SST(\mu_t, \sigma_t, \nu_t, \tau_t) \\ \mu_t &= 0.0266 \\ \log(\sigma_t) &= 0.1362 + \gamma_{2,t} \\ \log(\nu_t) &= -0.0693 + \gamma_{3,t} \\ \log(\tau_t - 2) &= 2.451 + \gamma_{4,t} \end{aligned} \tag{14}$$

where

$$\begin{aligned} \gamma_{2,t} &= 0.9855\gamma_{2,t-1} - 0.0658 v_{1,t-1} - 0.0712 v_{2,t-1} + b_{2,t} \\ \gamma_{3,t} &= \gamma_{3,t-1} + b_{3,t} \\ \gamma_{4,t} &= \gamma_{4,t-1} + b_{4,t} \end{aligned}$$

The fitted values for the variances are $\hat{\sigma}^2_{b_2} = 0.00336, \hat{\sigma}^2_{b_3} = 3.2459e^{-05}$ and $\sigma^2_{b_4} = 0.00199$

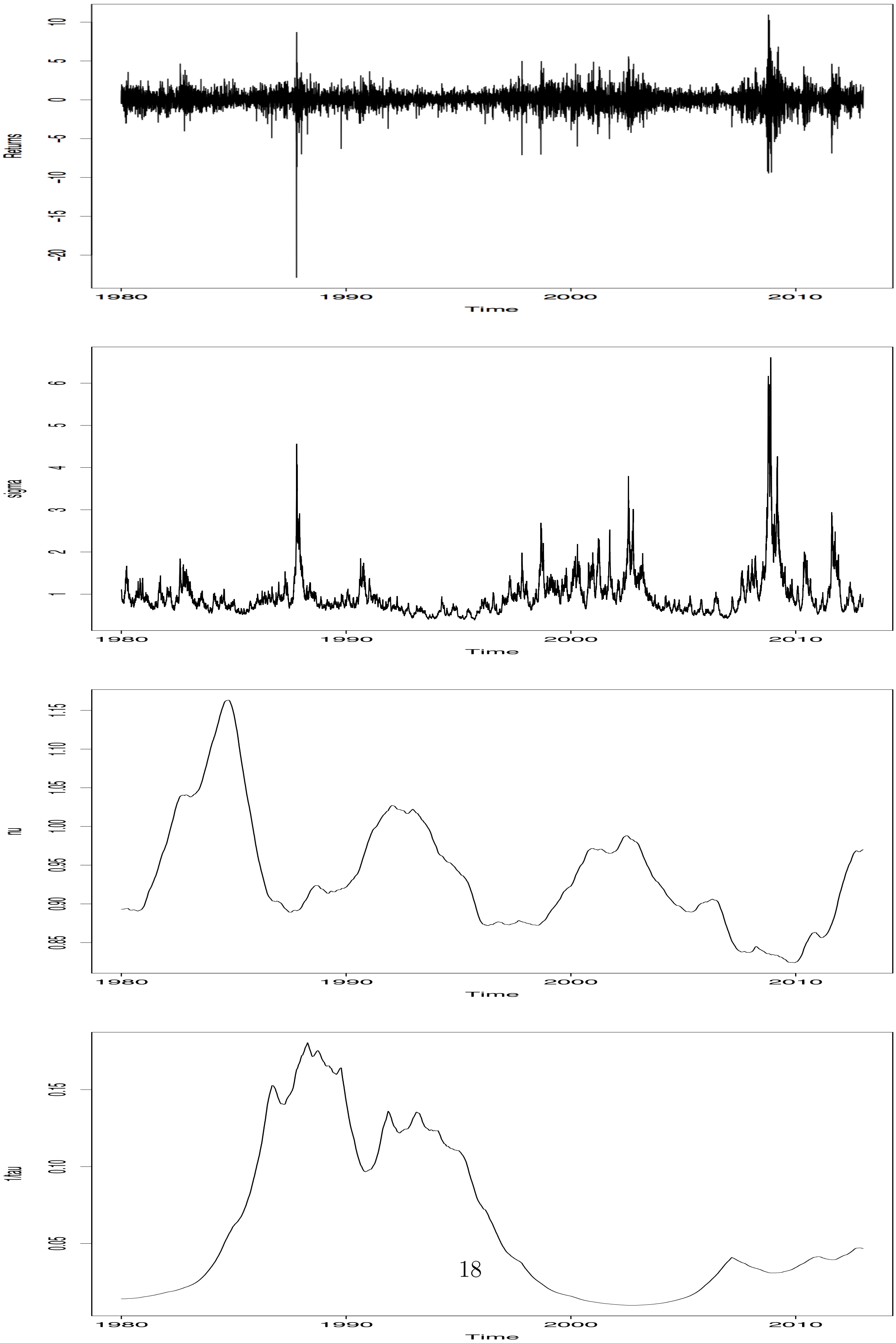


Figure 3: Returns y_t and fitted σ_t, ν_t and $1/\tau_t$ for model `m2`.

5.1.1. Comparing GEST model `m2` and APARCH(1,1) model

The chosen GEST model `m2` is now compared with alternative models. Asymmetric power ARCH (APARCH) models are used to model volatility of the S&P 500 stock index returns with the SST distribution in order to see how well they capture the asymmetry and the fat tails of the asset returns compared with the GEST model. To measure the goodness of fit, we use the global deviance (equals to - twice the maximum log-likelihood). The Akaike information criterion (AIC) was used to choose the best fitted model. Using the `fGarch` package available in `R` (Wurtz *et al.*, 2006) we compare between the GEST model, the GARCH(1,1) model introduced by Bollerslev (1986), the GJR-GARCH(1,1) model introduced by Glosten, Jagannathan and Runkle (1993) to allow for leverage effect, and the APARCH(1,1) model introduced by Ding, Granger and Engle (1993) which adds the flexibility of a varying exponent. The GEST model `m2` has the lowest AIC followed by the APARCH(1,1) model.

Table 2: Model comparison between the GEST, GARCH, GJR-GARCH, and APARCH

Information Criteria	GEST	GARCH	GJR-GARCH	APARCH
Global Deviance	21240.61	22294.64	22615.48	22141.96
AIC	**21862.74**	22306.65	22625.48	**22157.96**

5.1.2. Residual analysis for GEST model `m2` and APARCH(1,1) model

In this section we compare the residual analysis for the GEST model `m2` with that of the APARCH(1,1) model.

The residuals used here are called the normalized probability integral transform (normalized PIT) residuals, (Rosenblatt, 1952; Mitchell and Wallis, 2011) or normalized quantile residuals (Dunn and Smyth, 1996) and are defined by

$$\hat{r}_t = \Phi^{-1}(\hat{u}_t)$$

where

$$\hat{u}_t = F_{Y_t}(y_t|\hat{\mu}_t, \hat{\sigma}_t, \hat{\nu}_t, \hat{\tau}_t)$$

where $\hat{u}_t$ are the PIT residuals, F_Y is the cumulative distribution function of the conditional distribution of Y_t and Φ^{-1} is the inverse cumulative distribution function of a standard normal $N(0,1)$ variable. The reason for using these residuals is that the true residuals r_t have a standard normal distribution if the model is correct. Hence the residuals $\hat{r}_t$ can be compared with a normal distribution using the Z statistics.

5.1.3. Interpreting the Z statistics

Model diagnosis is investigated by calculating Z statistics to test the normality of the residuals within time groups (Royston and Wright, 2000).

Table 3: The guide range of Z statistics (first column), interpreted with respect to the normalized PIT residuals (seconf column) and the model response variable (third column).

Z statistics	**Normalized PIT residuals**	**Response variable**
Z1 < -2	mean too small	mean too large
Z1 > 2	mean too large	mean too small
Z2 < -2	variance too small	variance too large
Z2 > 2	variance too large	variance too small
Z3 < -2	positive skewness	skewness too high
Z3 > 2	negative skewness	skewness too low
Z4 < -2	platykurtosis	kurtosis too high (i.e. tails too heavy)
Z4 > 2	leptokurtosis	kurtosis too low (ie. tails too light)

Let G be the number of time groups and let $\{r_{gi}, i = 1, 2, .., n_i\}$ be the residuals in time group g, with mean $\bar{r}_g$ and standard deviation s_g, for $g = 1, 2, .., G$. The following statistics $\mathsf{Z}_{g1}, \mathsf{Z}_{g2}, \mathsf{Z}_{g3}, \mathsf{Z}_{g4}$ are calculated from the residuals in group g to test whether the residuals in group g have population mean 0, variance 1, skewness 0 and kurtosis 3, where $\mathsf{Z}_{g1} = n_g^{1/2}\bar{r}_g$, $\mathsf{Z}_{g2} = \left\{ s_g^{2/3} - [1 - 2/(9n_g - 9)] \right\} / \{2/(9n_g - 9)\}^{1/2}$ and Z_{g3} and Z_{g4} are test statistics for skewness and kurtosis given by D'Agostino *et al.* (1990), in their equations (13) and (19) respectively. Provided the number of groups G is sufficiently large then the Z_{gj} values should have approximately standard normal distributions under the null hypothesis that the true residuals are standard normally distributed. We suggest as a rough guide values of $|\mathsf{Z}_{gj}|$ greater than 2 be considered as indicative of significant inadequacies in the model. Note that significant positive (or negative) values $\mathsf{Z}_{gj} > 2$ (or $\mathsf{Z}_{gj} < -2$) for $j = 1, 2, 3$ or 4 indicate respectively that the residuals in time group g have a higher (or lower) mean, variance, skewness or kurtosis than the assumed standard normal distribution. See Table 3 for the interpretation.

Table 4 gives the values of Z_{gj} obtained from the APARCH fitted model. The significant negative values of Z_{g2} are Z_{32} and Z_{52} indicating that the residual variance is too low (or equivalently that the fitted APARCH model variance or volatility is too high) within the corresponding interval of time t. The significant negative values of Z_{g3} are Z_{23} and Z_{63} indicating that the residual skewness is too low (or equivalently the model skewness is too high) while the significant positive value of Z_{g3} is Z_{13} indicating that the residual skewness is too high (or equivalently the model skewness is too low). The significant negative value of Z_{g4} is Z_{54} indicating that the residual kurtosis is too low (or equivalently the model kurtosis is too high) while the significant positive values of Z_{g4} are Z_{24} and Z_{34}, indicating that the residual kurtosis is too high (or equivalently the model kurtosis is too low). Clearly a constant skewness and constant kurtosis in APARCH model is inadequate.

Table 5 gives the values of Z_{gj} obtained from the GEST fitted `m2` model. There is only one significant value Z_{23}, indicating the residual skewness is too low (or equivalently the model skewness is too high).

In conclusion, the residual analysis shows that the GEST fitted model does fit the data better than the APARCH model.

Table 4: Z statistics of APARCH

group, g	time, t	Z1	Z2	Z3	Z4
1	0.5 to 1387.5	-0.14	-0.13	**3.44**	-1.44
2	1387.5 to 2775.5	1.27	1.51	**-2.75**	**3.14**
3	2775.5 to 4162.5	0.69	**-3.38**	0.53	**2.06**
4	4162.5 to 5549.5	0.08	1.94	-1.48	-1.85
5	5549.5 to 6937.5	-0.58	**-2.16**	-0.53	**-2.49**
6	6937.5 to 8324.5	-0.69	1.83	**-2.58**	-1.08

Table 5: Z statistics of GEST

group, g	time, t	Z1	Z2	Z3	Z4
1	0.5 to 1387.5	-0.28	-0.13	0.84	-1.71
2	1387.5 to 2775.5	1.14	0.23	**-2.16**	0.34
3	2775.5 to 4162.5	0.99	-0.66	-0.39	-0.05
4	4162.5 to 5549.5	0.20	0.38	-0.37	-1.11
5	5549.5 to 6937.5	-0.30	-0.37	-0.10	-1.82
6	6937.5 to 8324.5	-0.62	0.48	-0.49	-0.60

5.2. Pound sterling and US dollar exchange rate

The data in this example are the pound sterling and US dollar daily exchange rates from 01-10-1981 to 28-06-1985. Harvey et al. (1994), Shephard and Pitt (1997), Kim et al. (1998) and Durbin and Koopman (2000) fitted a stochastic volatility model to pound/dollar exchange rates' returns with a conditional normal distribution to model the volatility clustering effect of the returns.

Let formulate the following GEST model:

$$\begin{aligned} Y_t|\mu_t,\sigma_t &\sim NO(\mu_t,\sigma_t) \\ \mu_t &= \beta_{1,0} \\ \log(\sigma_t) &= \beta_{2,0}+\gamma_{2,t} \\ \gamma_{2,t} &= \phi\gamma_{2,t-1}+b_{2,t} \end{aligned} \qquad (15)$$

where $b_{2,t} \sim NO(0, \sigma_b^2)$.

The GEST estimation of the hyperparameters gives very similar results to Bayesian approach of Durbin and Koopman (2000):

- GEST stochastic volatility model: $\sigma_n^2 = 0.007182$ & ϕ= 0.9744
- Durbin and Koopman with Bayesian: $\sigma_n^2 = 0.007425$ & ϕ= 0.9731

Figure 4 shows the fitted stochastic volatility of the GEST model. The GEST model represents the volatility clustering effect of the pound/dollar returns with a stochastic volatility model for $log(\sigma_t)$ as an autoregressive order 1 process. Clearly, $\hat{\sigma_t}$ increases when the volatility clustering effect is high, and decreases when the clustering is low.

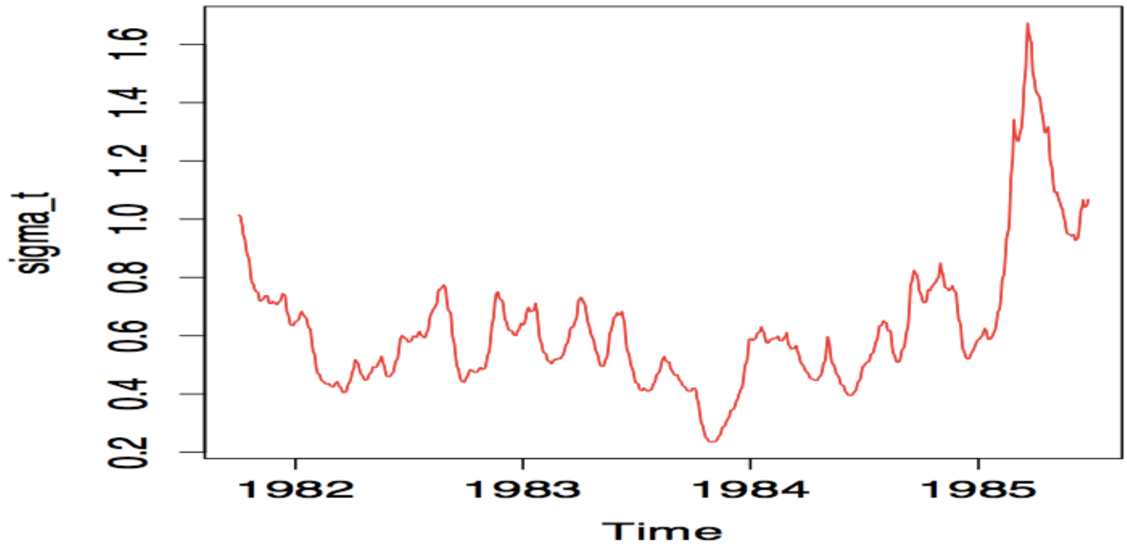


Figure 4: The fitted stochastic volatility with the GEST model for the pound/dollar daily returns

5.3. Van drivers killed in the UK

In time series analysis of road traffic safety, it is often required to assess the effect of road safety measures on the development in traffic safety over time. The data in this example are the monthly number of light goods vehicle drivers killed in road accidents from 1969 to 1984 in the UK.

The model which Durbin and Koopman (2000) fitted to the van drivers was a structural mean model with a random walk local level and stochastic seasonal for the conditional Poisson distribution. Their parameter estimates for the random walk local level and the seasonal disturbances were $\hat{\sigma_b} = 0.0245$ and $\hat{\sigma_w} = 0$ respectively, with a conclusion that the seasonal effect is constant over time. Their parameter estimate for the seat belt intervention variable was $\hat{\beta_{1,1}} = -0.280$, which corresponds to a reduction in the number of deaths of 24%. The fitted GEST model gives similar results (with $\hat{\sigma_b} = 0.02417$ and $\hat{\sigma_w} = 0.00008$):

$$\begin{aligned} Y_t|\mu_t, &\sim PO(\mu_t) \\ \mu_t &= \beta_{1,0} + \beta_{1,1}x + \gamma_{1,t} + s_{1,t} \\ \gamma_{1,t} &= \gamma_{1,t-1} + b_{1,t} \\ s_{1,t} &= -\sum_{m=1}^{M-1} s_{1,t-m} + w_{1,t} \end{aligned} \tag{16}$$

where $b_{1,t} \sim NO(0, \sigma_b^2)$ and $w_{1,t} \sim NO(0, \sigma_w^2)$ and x = (0,0,..,0,1,..,1) is an explanatory variable for the seat belt introduction.

Clearly, more complicated GEST models can be fitted, including the two parameter Negative Binomial type I distribution.

6. Conclusion

We have introduced the generalized structural time series (GEST) model as a uniformly applicable parameter-driven model specification to capture time variation in parameters. A clear advantage of the GEST model is that it expands the systematic part of parameter-driven time series models to allow the updating of all the distribution parameters over time, fitted through the development of a fast local estimation algorithm.

The proposed GEST model primarily addresses the difficulty in modelling time-varying skewness and time-varying kurtosis (beyond location and dispersion parameter-driven time series models) to better describe the non-Gaussian movements in a time series. There are several important points to make here with respect to the class of parameter-driven time series models proposed here:

- The GEST model allows the use of a flexible parametric distribution $\mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ for the dependent variable, including highly skewed and/or kurtotic distributions (mixed distributions).

- A GEST model is fitted using a *fast* penalised likelihood estimation algorithm (without computing the likelihood function though the evaluation of a high-dimensional integral based on simulation methods).

- The GEST model expands the systematic part of parameter-driven time series models to allow the joint and explicit modelling of all the distribution parameters $(\mu_t, \sigma_t, \nu_t, \tau_t)$.

- The structural terms for each distribution parameter of the conditional distribution can be a random walk or autoregressive term (of any order) and can include seasonal and/or leverage effects.

Time-varying mean, variance, skewness and kurtosis are of interest in themselves and provide information on various aspects of a time series. We argue here that the GEST model provides a useful framework within which to stochastically present 'stylised facts' on time series based upon time-varying estimates of the distribution parameters $(\mu_t, \sigma_t, \nu_t, \tau_t)$. Thus, by fitting a GEST model, we also provide a model that is intended to be taken as a full description of the conditional distribution of time-varying observations.

We demonstrate in section 5 the flexibility of the GEST model by formulating a number of GEST models. In particular, the GEST model of equation (12) demonstrates how to capture the time-varying movements of the returns of the S&P 500 index. Furthermore, a variety of diagnostic tools have also been used to compare the adequacy of the GEST model with the APARCH model for the returns of the S&P 500 index. The use of the APARCH model enables us to compare the GEST model with a popular observation-driven model. The other examples presented here demonstrate the properties of the GEST model with respect to other parameter-driven models - we are not aware of any other parameter-driven model capable of jointly and explicitly modelling time-varying skewness and kurtosis. The examples presented here should be taken as an illustration of the flexibility of a class of parameter-driven time series model referred to as the GEST model.

7. Appendix

7.1. *Skew Student t distribution*

A skewed Student t distribution was used to allow for skewness and kurtosis in the conditional distribution of financial returns initially by Hansen (1994) and subsequently by Fernandez and Steel (1998) using an alternative parametrization. Fernandez and Steel (1998) consider a shifted and scaled t distribution with τ degrees of freedom, i.e. $\mu_0 + \sigma_0 T$ where $T \sim t_\tau$, denoted here by $TF(\mu_0, \sigma_0, \tau)$, and splice together at μ_0 two differently scaled distributions, $Y_1 \sim TF(\mu_0, \sigma_0/\nu, \tau)$ below μ_0 and $Y_2 \sim TF(\mu_0, \sigma_0\nu, \tau)$ above μ_0, The resulting distribution is denoted here by $Y \sim ST3(\mu_0, \sigma_0, \nu, \tau)$. Wurtz *et al.* (2006) reparameterized the skew t distribution of Fernandez and Steel (1998) so that in the new parametrization μ is the mean and σ is the standard deviation, denoted here by $Y \sim SST(\mu, \sigma, \nu, \tau)$, where

$$
\begin{aligned}
f_Y(y|\mu,\sigma,\nu,\tau) &= \frac{2}{(1+\nu^2)}\left\{f_{Y_1}(y)I(y<\mu_0)+\nu^2 f_{Y_2}(y)I(y\geq\mu_0)\right\}\\
&= \frac{c}{\sigma_0}\left\{1+\frac{(y-\mu_0)^2}{\sigma_0^2\tau}\left[\nu^2 I(y<\mu_0)+\frac{1}{\nu^2}I(y\geq\mu_0)\right]\right\}^{-(\tau+1)/2}
\end{aligned}
$$

for $-\infty < y < \infty, -\infty < \mu < \infty, \sigma > 0, \nu > 0$, and $\tau > 2$, where $c = 2\nu/\left[(1+\nu^2)B(\frac{1}{2},\frac{\tau}{2})\tau^{1/2}\right]$,

$$\mu_0 = \mu - \sigma m/s,$$

and

$$\sigma_0 = \sigma/s$$

and

$$m = 2\tau^{1/2}(\nu^2-1)/\left[(\tau-1)\nu B(\frac{1}{2},\frac{\tau}{2})\right]$$

and

$$s^2 = \left\{\tau\left(\nu^3+\nu^{-3}\right)/\left[(\tau-2)(\nu+\nu^{-1})\right]\right\} - m^2.$$

Hence $Y \sim SST(\mu, \sigma, \nu, \tau) = ST3(\mu_0, \sigma_0, \nu, \tau)$ has mean μ and variance σ^2 since $E(Y) = \mu_0 + \sigma_0 E(Z_0) = (\mu - \sigma m/s) + (\sigma/s)m = \mu$ and $V(Y) = \sigma_0^2 V(Z_0) = (\sigma^2/s^2)s^2 = \sigma^2$, where $Z_0 = (Y-\mu_0)/\sigma_0 \sim ST3(0,1,\nu,\tau)$ and where $E(Z_0) = m$ and $V(Z_0) = s^2$ provided $\tau > 2$ from Fernandez and Steel (1998) p360. Note that $Z = (Y-\mu)/\sigma \sim SST(0,1,\nu,\tau)$ has mean 0 and variance 1.

7.2. *The algorithm for estimating* $\boldsymbol{\alpha}$

1. Select starting values for $\boldsymbol{\alpha} = (\sigma_e^2, \sigma_b^2, \phi)$.

2. Maximize Q over $\boldsymbol{\alpha}$ using a numerical algorithm, where $\boldsymbol{\gamma}$ given $\boldsymbol{\alpha}$ is obtained before calculating Q in the function evaluating Q.

3. Use the maximizing values for $\boldsymbol{\alpha}$ to calculate the maximizing values for $\boldsymbol{\gamma}$.

In step 2, let Q be given by

$$
\begin{aligned}
Q &= \log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}) + \log f(\boldsymbol{\gamma}) - \frac{1}{2}\log\left|\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right| + \frac{T}{2}\log 2\pi \\
\log f(\boldsymbol{\epsilon}|\boldsymbol{\gamma}) &= -\frac{1}{2}\log\left|2\pi\mathbf{W}^{-1}\right| - \frac{1}{2}\sigma_e^{-2}\left(\boldsymbol{\epsilon}-\boldsymbol{\gamma}\right)^\top\mathbf{W}\left(\boldsymbol{\epsilon}-\boldsymbol{\gamma}\right) - \frac{1}{2}T\log\sigma_e^2 \\
\log f(\boldsymbol{\gamma}) &= \log f(\mathbf{b}) = -\frac{1}{2}\sigma_b^{-2}\boldsymbol{\gamma}^\top\mathbf{D}^\top\mathbf{D}\boldsymbol{\gamma}^\top - \frac{1}{2}(T-J)\log\left(2\pi\sigma_b^2\right)
\end{aligned}
$$

since $\mathbf{b} = \mathbf{D}\boldsymbol{\gamma} \sim \mathbf{N}(\mathbf{0}, \sigma_\mathbf{b}^\mathbf{2}\mathbf{I_{T-J}})$, where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_T)^\top, \boldsymbol{\Sigma}^{-1} = \sigma_e^{-2}\mathbf{W}$ and

$$
D = \begin{pmatrix}
-\phi_J & -\phi_{J-1} & \ldots & \ldots & -\phi_1 & 1 & 0 & 0 & \ldots & \ldots & 0 & 0 \\
0 & -\phi_J & -\phi_{J-1} & \ldots & \ldots & -\phi_1 & 1 & 0 & \ldots & \ldots & 0 & 0 \\
 & & & & & & & & & & & \\
0 & 0 & 0 & \ldots & \ldots & 0 & -\phi_J & -\phi_{J-1} & \ldots & \ldots & -\phi_1 & 1
\end{pmatrix},
$$

and the maximizing of Q over $\boldsymbol{\gamma}$ given $\boldsymbol{\alpha}$ is given by

$$
\boldsymbol{\gamma} \ = \left[\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right]^{-1}\boldsymbol{\Sigma}^{-1}\ \boldsymbol{\epsilon} \tag{17}
$$

Note that D is a $(n-J)$ x T matrix and $\boldsymbol{\Sigma}$ is n x n.

Let $\mathbf{B} = \left[\boldsymbol{\Sigma}^{-1} + \sigma_b^{-2}\mathbf{D}^\top\mathbf{D}\right]^{-1}\boldsymbol{\Sigma}^{-1}$ and let $\hat{\mathbf{B}}, \hat{\boldsymbol{\Sigma}}, \hat{\mathbf{D}}, \hat{\boldsymbol{\gamma}}$ and $\hat{\sigma}_b^{-2}$ be the values of $\mathbf{B}, \boldsymbol{\Sigma}, \mathbf{D}, \boldsymbol{\gamma}$ and σ_b^{-2} on convergence of the GEST fitting procedure (see Section 4).

On convergence, $\hat{\boldsymbol{\gamma}} = \hat{\mathbf{B}}\boldsymbol{\epsilon}$. Hence d, the effective degrees of freedom used in the model, is

$$
d = tr\left[\hat{\mathbf{B}}\right] = tr\left\{\left[\hat{\boldsymbol{\Sigma}}^{-1} + \hat{\sigma}_b^{-2}\hat{\mathbf{D}}^\top\hat{\mathbf{D}}\right]^{-1}\hat{\boldsymbol{\Sigma}}^{-1}\right\} \tag{18}
$$

As d is difficult to calculate directly for large n, it can be calculated by setting $\partial Q/\partial\sigma_b^2 = 0$ giving on convergence $d = J + \hat{\sigma}_b^{-2}\hat{\gamma}^\top\mathbf{D}^\top\mathbf{D}\hat{\gamma}$, using the result $\frac{\partial}{\partial x}\log|xC+F| = tr\left[(xC+F)^{-1}C\right]$, where x is a scalar and C and F are r x r matrices (provided $|xC+F| \neq 0$). Hence, for each distribution parameter, d is calculated using the values $\hat{\gamma}$ and $\hat{\sigma}_b^2$ on convergence of the GEST fitting algorithm.

7.3. Proof of the Theorems

7.3.1. Theorem 1 Proof

$Y_t|\mu_t, \sigma_t, \nu_t, \tau_t \sim \mathcal{D}(\mu_t, \sigma_t, \nu_t, \tau_t)$ where $\log\mu_t = \beta_{1,0} + \gamma_{1,t}$, and $\log\sigma_t = \beta_{2,0} + \gamma_{2,t}$. Applying the law of iterated expectations, a)

$$\begin{aligned}
E[Y_t] &= E\left[E\left(Y_t|\mu_t,\sigma_t,\nu_t,\tau_t\right)\right] = E[\mu_t] \\
E[\mu_t] &= \beta_{1,0} + E\left(\gamma_{1,t}\right) \\
\Phi_1(B)\gamma_{1,t} &= b_{1,t} \\
\gamma_{1,t} &= \Phi_1(B)^{-1} b_{1,t} = \psi_1(B) b_{1,t}
\end{aligned}$$

since $\Phi_1(B)$ is assumed to be invertible

$$\begin{aligned}
E[\gamma_{1,t}] &= \psi_1(B)E(b_{1,t}) = 0 \\
E[\mu_t] &= \beta_{1,0} \\
\text{Hence E[Y}_\text{t}] &= \beta_{1,0}
\end{aligned}$$

b)

$$\begin{aligned}
V[Y_t] &= V\left[E\left(Y_t|\mu_t,\sigma_t,\nu_t,\tau_t\right)\right] + E\left[V\left(Y_t|\mu_t,\sigma_t,\nu_t,\tau_t\right)\right] \\
V[Y_t] &= V[\mu_t] + c^2 E[\sigma_t^2] \\
\text{but V}[\mu_\text{t}] &= V[\gamma_{1,t}] = V[\psi_1(B)b_{1,t}] = S_1\sigma_{b_1}^2 \\
\text{and E}[\sigma_\text{t}^2] &= E\left[\exp\left(2\beta_{2,0} + 2\gamma_{2,t}\right)\right] = \exp\left(2\beta_{2,0}\right) E\left[\exp\left(2\gamma_{2,t}\right)\right] \\
\text{where E}\left[\exp\left(2\gamma_{2,\text{t}}\right)\right] &= E\left[\exp\left(2\psi_2(B)b_{2,t}\right)\right] = \prod_{j=0}^{\infty} E\left[\exp\left(2\psi_{2,j}b_{2,t-j}\right)\right] \\
&= \prod_{j=0}^{\infty} \exp\left(2\psi_{2,j}^2\sigma_{b_2}^2\right) = \exp\left(2S_2\sigma_{b_2}^2\right) \\
\text{where S}_\text{k} &= 1 + \sum_{j=1}^{\infty} \psi_{k,j}^2 \\
\text{and } \psi_{2,0} &= 1 \\
\text{Hence V[Y}_\text{t}] &= S_1\sigma_{b_1}^2 + c^2 \exp\left(2\beta_{2,0} + 2S_2\sigma_{b_2}^2\right)
\end{aligned}$$

7.3.2. Theorem 2 Proof

a) $E[Y_t] = E\left[E\left(Y_t|\mu_t,\sigma_t,\nu_t,\tau_t\right)\right] = E[\mu_t] = \exp\left(\beta_{1,0} + \frac{1}{2}S_1\sigma_{b_1}^2\right)$, from d) below.

b) $V[Y_t] = V\left[E\left(Y_t|\mu_t,\sigma_t,\nu_t,\tau_t\right)\right] + E\left[V\left(Y_t|\mu_t,\sigma_t,\nu_t,\tau_t\right)\right] = V[\mu_t] + E\left[v(\mu_t,\sigma_t)\right]$.

c) $V[\mu_t] = E[\mu_t^2] - \{E[\mu_t]\}^2 = \exp(2\beta_{1,0})\left[\exp(2S_1\sigma_{b_1}^2) - \exp(S_1\sigma_{b_1}^2)\right]$, from d) below.

d) $E[\mu_t^r] = E\left[\exp(r\beta_{1,0} + r\gamma_{1,t})\right] = \exp(r\beta_{1,0})E\left[\exp(r\psi_1(B)b_{1,t})\right]$.

$$E\left[\exp(r\psi_1(B)b_{1,t})\right] = \prod_{j=1}^{\infty} E\left[\exp(r\psi_{1,j}b_{1,t-j}\right] = \prod_{j=1}^{\infty} \exp\left(\frac{1}{2}r^2\psi_{1,j}^2\sigma_{b_1}^2\right) = \exp\left(\frac{1}{2}r^2 S_1\sigma_{b_1}^2\right),$$

since if $b \sim N(0, \sigma_b^2), E\left[\exp(r\psi b)\right] = \int_{-\infty}^{\infty} \exp(r\psi b)\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left[-\frac{b^2}{2\sigma_b^2}\right]db$
$= \exp\left(\frac{1}{2}r^2\psi^2\sigma_b^2\right)\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}\sigma_b}\exp\left[-\frac{1}{2\sigma_b^2}(b - r\psi\sigma_b^2)^2\right]db = \exp\left(\frac{1}{2}r^2\psi^2\sigma_b^2\right).$

e) As for d).

Corollary 1 Theorem 2 applies to the following distributions, the negative binomial type 1 and 2, NBI and $NBII$, respectively, the gamma, GA, and inverse Gaussian, IG, distributions. The results below for the marginal mean $E[Y_t]$ and variance $V[Y_t]$ of Y_t use the results of Theorem 2.

1. $Y_t|\mu_t, \sigma_t \sim NBI(\mu_t, \sigma_t)$ with mean μ_t and variance $\mu_t + \sigma_t\mu_t^2$, then $E[Y_t] = E[\mu_t]$ and $V[Y_t] = V[\mu_t] + E[\mu_t] + E[\sigma_t]E[\mu_t^2]$. Note $\sigma_t = 0$ gives $Y_t|\mu_t \sim PO(\mu_t)$.

2. $Y_t|\mu_t, \sigma_t \sim NBII(\mu_t, \sigma_t)$ with mean μ_t and variance $\mu_t + \sigma_t\mu_t$, then $E[Y_t] = E[\mu_t]$ and $V[Y_t] = V[\mu_t] + E[\mu_t] + E[\sigma_t]E[\mu_t]$

3. $Y_t|\mu_t, \sigma_t \sim GA(\mu_t, \sigma_t)$ with mean μ_t and variance $\sigma_t^2\mu_t^2$, then $E[Y_t] = E[\mu_t]$ and $V[Y_t] = V[\mu_t] + E[\sigma_t^2]E[\mu_t^2]$. Note $\sigma_t = 1$ gives $Y_t|\mu_t \sim EXP(\mu_t)$.

4. $Y_t|\mu_t, \sigma_t \sim IG(\mu_t, \sigma_t)$ with mean μ_t and variance $\sigma_t^2\mu_t^3$, then $E[Y_t] = E[\mu_t]$ and $V[Y_t] = V[\mu_t] + E[\sigma_t^2]E[\mu_t^3]$.

References

Akaike, H., (1983), Information measures and model selection. *Bull. Int. Statist. Inst.*, **50**, 277-290.

Asai, M. and M. McAleer, (2005), Dynamic asymmetric leverage in stochastic volatility models. *Econometric Reviews*, **24**, 317-332.

Bauwens L. and Hautsch N. (2006), Stochastic conditional intensity processes. *Journal of Financial Econometrics*, **4**, 450-493.

Bollerslev, T., (1986), Generalised autoregressive conditional heteroskedasticity. *J. Econometr*, **31**, 307-327.

Breslow, N. E. and D.G. Clayton, (1993), Approximate inference in generalized linear mixed models. *J. Am. Statist. Ass.*, **88**, 9-25.

Cole, T. J. and Green, P. J., (1992), Smoothing reference centile curves: the lms method and penalized likelihood. *Statis. Med.*, **11**, 1305-1319.

Cox, DR., (1981), Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, **8**, 93-115.

Creal, D., Koopman, S.J., Lucas, A., (2013), Generalised Autoregressive Score Models with Applications. *Journal of Applied Econometrics*, **28**, 777-795.

Creal, D., Koopman, S.J., Lucas, A., (2011), A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business and Economic Statistics*, **29**, 552-563.

Crowder, M. J., Kimber, A. C., Smith R. L. and Sweeting, T. J., (1991), *Statistical Analysis of Reliability Data.* Chapman and Hall, London.

D'Agostino, R.B., A. Balanger, and Jr R.B. D'Agostino, (1990), A suggestion for using powerful and informative tests of normality. *Am. Statist.*, **44**, 316-321.

Davidson, R., (2012), Statistical inference in the presence of heavy tails. *The Econometrics Journal*, **15**, C31-C53.

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

Ding, Z., C.W.J. Granger C.W.J., and R.F. Engle, (1993), A Long Memory Property of Stock Market Returns and a New Model. *J. Empir. Fin.*, **1**, 83-106.

Dunn, P.K. and G.K. Smyth, (1996), Randomised quantile residuals. *J. Comp. Graph. Statis.*, **5**, 236-244.

Durbin, J. and S.J. Koopman, (2000), Time Series Analysis of Non-Gaussian Observations based on State Space Models from both Classical and Bayesian Perspectives. *J. R. Statist. Soc.*, **B,62**, 3-56.

Durbin, J. and S.J. Koopman, (2012), *Time Series Analysis by State Space Methods.* Oxford University Press.

Eilers, P. H. C. and B.D. Marx, (1996), Flexible smoothing with B-splines and penalties (with comments and rejoinder). *Statis. Scien.*, **11**, 89-121.

Engle, R.F., (1982) Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987-1008.

Evans, M. and T. Swartz, (2000), *Approximating Integrals via Monte Carlo and Deterministic Methods.* Oxford: Oxford University Press.

Fahrmeir, L. and G. Tutz, (2001), *Multivariate Statistical Modelling based on Generalized Linear Models.* 2nd edn. New York: Springer.

Fama, E.F., (1965), The behavior of stock market prices. *J. Bus.*, **38**, 34-105.

Fernandez, C. and M.F.J. Steel (1998), On Bayesian modeling of fat tails and skewness. *J. Am. Statist. Ass.*, **93**, 359-371.

Giraitis L., R. Leipus, P. Robinson, and D. Surgailis, (2004), LARCH, leverage and long memory. *Journal of Financial Econometrics*, **2**, 177-210.

Glosten, L., R. Jagannathan, and D. Runkle, (1993), On the Relation Between Expected Value and the Volatility of the Nominal Excess Return on Stocks. *J. Fin.*, **48**, 1779-1801.

Granger, C.W.J, (2005), The past and future of empirical finance: some personal comments. *Journal of Econometrics*, **129**, 35-40.

Hamilton, J., (1989), A new approach to the economic analysis of nonstationary time series and the business cycle, *Econometrica*, **57**, 357-384.

Hansen, B.E., (1994), Autoregressive conditional density estimation. *Inter. Econ. Rev.*, **35**, 705-730.

Harvey, A.C., (1985), *Trends and cycles in macroeconomic time series. J. Bus. Econ. Statis.*, **3**, 216-227.

Harvey, A.C., (1989), *Forecasting, Structural Time Series Models and the Kalman Filter.* Cambridge University Press, Cambridge.

Harvey, A. C. and A. Jaeger, (1993), Detrending, stylized facts and the business cycle. *J. App. Econometr.*, **8**,231-247

Harvey, A.C., E. Ruiz, and N. Shephard, (1994), Multivariate stochastic variance models. *Review of Economic Studies*, **61**, 247-264.

Harvey, A.C. and N. Shephard, (1996), The estimation of an asymmetric stochastic volatility model for asset returns. *Journal of Business and Economic Statistics*, **14**, 429-434.

Harvey, C.R. and A. Siddique, (1999), Autoregressive Conditional Skewness. *Journal of Financial and Quantitative Analysis*, **34**, 465-487.

Hastie, T. J. and R.J. Tibshirani, (1990), *Generalized Additive Models.* London: Chapman and Hall.

Hastie, T. J., R.J. Tibshirani, and J. Friedman, (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* 2nd ed, New York, N.Y. : Springer.

Johnson, N. L., Kotz, S. and Balakrishnan, N., (1994), *Continuous Univariate Distributions*, 2nd ed. Wiley, New York.

Jones, M. C. and M. J. Faddy, (2003), A skew extension of the t-distribution, with application. *J. R. Statist. Soc.*, **B, 65**, 159-174.

Jones, M. C. and Pewsey, A, (2009), Sinh-arcsinh distributions. *Biometrika*, **96**, 761-780.

Kitagawa, G., (1987), Non-Gaussian State-Space Modelling of Nonstationary Time Series (with discussion). *J. Am. Statist. Ass.*, **82**, 1032-1063.

Kitagawa, G., (1989), Non-Gaussian Seasonal Adjustment. *Computers & Mathematics with Applications*, **18**, 503- 514.

Kitagawa, G. and W. Gersch, (1996), *Smoothness Priors Analysis of Time Series.* Springer-Verlag.

Koopman S.J., Lucas A., Monteiro A., (2008), The multi-state latent factor intensity model for credit rating transitions, *Journal of Econometrics* , **142**, 399-424.

Lee, Y. and J.A. Nelder, (1996), Hierarchical generalized linear models (with discussion). *J. R. Statist. Soc.*, **B, 58**, 619-678.

Lee, Y., J.A. Nelder, and Y. Pawitan, (2006), *Generalized Linear Models With Random Effects: Unified Analysis Via H-Likelihood.* Chapman & Hall/CRC

Mandelbrot, B. (1963), New methods in statistical economics. *J. Pol. Econ.*, **71**, 421-440.

McDonald, J. B. and Xu, Y. J., (1995), A generalisation of the beta distribution with applications. *J. Econometr*, **66**, 133-152.

Mitchell, J. and K.F Wallis, (2011), Evaluating Density Forecasts: Forecast Combinations, Model mixtures, Calibration and Sharpness. *J. Appl. Econometr*, **26**, 1023-1040.

Nelson, D. B., (1991), Conditional heteroskedasticity in asset returns: a new approach. *Econometrica*, **59**, 347-370.

Omori, Y., S. Chib, N. Shephard, and J. Nakajima, (2007), Stochastic volatility with leverage: fast likelihood inference. *J. Econometr*, **140**, 425-449.

Pawitan, Y., (2001), *In All Likelihood: Statistical Modelling and Inference Using Likelihood.* Oxford University Press.

Pinheiro, J.C. and D.M. Bates, (2000), *Mixed-Effects Models in S and S-PLUS.* New York: Springer-Verlag.

Rigby, R.A. and D.M. Stasinopoulos, (2005), Generalized Additive Models for Location, Scale and Shape (with discussion). *J. R. Statist. Soc.*, **C, 54**, 507-554.

Rigby, R.A. and D.M. Stasinopoulos, (2013), Automatic smoothing parameter selection in GAMLSS with an application to centile estimation. *Statis. Meth. Med. Res.*, **1**, 1-15.

Rosenblatt, M. (1952), Remarks on a Multivariate Transformation. *Ann. Math. Statis.*, **23**, 470-472.

Royston, P. and E.M. Wright, (2000), Goodness-of-fit statistics for age-specific reference intervals. *Statis. Med.*, **19**, 2943-2962.

Rue, H., S. Martino, and N. Chopin, (2009), Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B*, **71**, 319-392.

Shephard, N. (2005), *Stochastic Volatility: Selected Readings.* Oxford University Press, Oxford.

Shephard N., Pitt, M.K., (1997), Likelihood analysis of non-Gaussian measurement time series. *Biometrika*, **84**, 653-667.

Shumway, R.H. and D.S. Stoffer, (2011), *Time Series Analysis and Its Applications With R Examples.* Third Ed., Springer.

Stasinopoulos, D. M., R.A. Rigby, and C. Akantziliotou, (2008), *Instructions on how to use the GAMLSS package in R*, Second Edition, STORM Research Centre, London Metropolitan University, London.

Tierney, L. and J.B. Kadane, (1986), Accurate approximations for posterior moments and marginal densities. *J. Am. Statist. Ass.*, **81**, 82-86.

Van Buuren, S. and M. Fredriks, (2001), Worm plot: a simple diagnostic device for modelling growth reference curves. *Statistics in Medicine*, **20**, 1259-1277.

Venables, W.N. and B.D. Ripley, (2002), *Modern Applied Statistics with S.* Forth ed. New York: Springer-Verlag.

West, M., J. Harrison, and H.S. Migon (1985), Dynamic generalized linear models and Bayesian forecasting (with discussion) *J. Am. Statist. Ass.*, **80**, 73-97.

West, M. and J. Harrison, (1997), *Bayesian Forecasting and Dynamic Models.* second ed. Springer, New York.

Wurtz, D., Y. Chalabi, and L. Luksan, (2006), Parameter Estimation of ARMA Models with GARCH/APARCH Errors An R and SPlus Software Implementation. *J. Statis. Soft.*.

Yu, J. (2005), On leverage in a stochastic volatility model. *Journal of Econometrics*, **127**, 165-178.