Score-Driven Models: Methodology and Theory*

Mariia Artemova(1,2), Francisco Blasques(1,2),
Janneke van Brummelen(1,2), Siem Jan Koopman(1,2,3)

[1]*School of Business and Economics, Vrije Universiteit Amsterdam*

[2]*Tinbergen Institute Amsterdam*

[3]*CREATES, Aarhus University*

January 2022

Summary

The score-driven model belongs to a wider class of observation-driven time series models which are used intensively in empirical studies in economics and finance. A defining feature of the model is its mechanism of updating time-varying parameters by means of the score function of the predictive likelihood function. The class of score-driven models contains many other well known observation-driven models as a special case, but also many new models have been developed based on the score-driven principle. It provides a general way of parameter updating, or filtering, in which all relevant features of the observation density function are considered. In case of models with fat-tailed observation densities, the score-driven updates become robust to large observations in time series. This kind of robustness is a convenient feature of score-driven models and makes them suitable for applications in finance and economics where noisy data sets are regularly encountered. Parameter estimation for score-driven models is straightforward when using the method of maximum likelihood. In many cases theoretical results are available under rather general conditions.

Keywords: Score-driven model, Time-varying parameter, Observation-driven model, Location and scale.

Subjects: Dynamic Econometrics, Time Series and Forecasting, Economics, Finance

[*] This is a draft of an article that has been published by Oxford University Press in the Oxford Research Encyclopedia of Economics and Finance on 19 October 2022.

Background and Motivation of Score-Driven Models

In empirical research in economics and finance, it is widely acknowledged that parameters in reduced-form representations of structural models are subject to various instabilities due to model misspecifications and approximations. These instabilities are partly caused by the limitations of linear model specifications which are typically used in applied econometrics. Furthermore, empirical data sets of interest in economics and finance are typically subject to changing behaviors of economic agents (possibly aimed at returning to some level of equilibrium), endogenous and exogenous shocks, systemic innovations, fiscal policy changes, and more. To account for such instabilities and distortions, linear model specifications are often extended by replacing fixed parameters with time-varying parameters. These generalizations can apply to parameters related to both the mean or location (constant and regression coefficients) and the variance or scale (error variance) equations of the model.

In the context of the standard linear regression model, time-varying parameters can be empirically detected by means of the *recursive least squares* method. Although the method was already developed in the original work of Gauss, it is oftentimes credited to Plackett (1950) who has provided an elegant derivation based on matrix algebra. The recursive least squares method is a set of equations that provides the least squares estimates based on a set of observations that is growing with individual observations, sequentially over time. The standard errors of the recursive estimates tend to become smaller after each update because the sample size increases. The usual assumptions of the regression model are also applicable in the context of recursive estimation, including the assumption that the regression coefficients are constant over time. The relaxation of this assumption has been considered in the work of Rudolf E. Kalman where in effect each regression parameter can potentially follow a linear dynamic process; see Kalman (1960). The celebrated *Kalman filter* can be regarded as the corresponding recursive least squares method that allows for time-varying regression parameters. It should be emphasized that Kalman's work was developed in the context of control and system theory, relevant in engineering and mathematics. Due to the work of Andrew C. Harvey in the 1970s, the statistical impact of Kalman's work for linear regression, autoregressive moving average (ARMA) and other linear dynamic models has been revealed and acknowledged; see Harvey (1981, Chapter 4) for a textbook treatment. His work has been instrumental in the recognition of the Kalman filter and its relevance in econometrics. In particular, a strong case is provided by the notion that almost all linear dynamic models can be represented as a *state space model* which consists of an observation equation and an updating equation for the state vector with the time-varying parameters. Once the model of interest is represented in state space form, the Kalman filter can be applied straightforwardly. The output of the Kalman filter enables the estimation of the time-varying parameters but also the evaluation of the likelihood function using the prediction error decomposition.

In the same period, Sir David Cox recognized two different classes of statistical time series models with time-varying parameters in his seminal article Cox (1981): *observation-driven models* and *parameter-driven models*. In the latter class of models, the parameters are treated as dynamic processes with their own source of errors. Due to this additional source of errors, the time-varying parameters are not perfectly predictable, even if the time-varying parameters are analyzed conditional on past and concurrent observations. The state space model as discussed above belongs clearly to the class of parameter-driven models. For the class of observation-driven models, time-varying parameters are treated as functions of lagged dependent variables as well as exogenous variables. In such model specifications, when conditioning on past and concurrent observations, the time-varying parameters are perfectly predictable. It also implies that likelihood evaluation is relatively straightforward. Given these features and characteristics, observation-driven models have become popular in econometrics. Typical examples of observation-driven models are the generalized autoregressive conditional heteroskedasticity (GARCH) models of Engle (1982) and Bollerslev (1986)), the exponential GARCH (EGARCH) model of Nelson (1991), and the autoregressive conditional duration (ACD) and intensity (ACI) models of Engle and Russell (1998).

A sub-class of observation-driven models is the class of score-driven models which are proposed and developed by Creal et al. (2011, 2013) and Harvey (2013). For this class of models, the score of the conditional observational density is used to update the time-varying parameters in the model. Here, the score refers to the first derivate of the log likelihood function with respect to the parameter. A more precise definition of the derivative in the context of score-driven models is provided in the section ‘Model specification’. The score-driven models are also known as *generalized autoregressive score* (GAS) models and as *dynamic conditional score* (DCS) models. Various factors have given rise to the use of score-driven models in empirical studies in economics and finance. For example, the score function provides in many cases an intuitive driving mechanism for the time-varying parameter. Generally, the score indicates in which ‘direction’ the time-varying parameter must ‘move’ in order to improve the fit in terms of a local (predictive) density. It can therefore be used 'naturally' for designing new updating equations for time-varying parameters; especially in settings where it is not immediately obvious what is a good choice for a parameter driving mechanism. The score-driving mechanism is of a particular convenience because it has the important advantage that the entire predictive density structure is exploited. In other words, all features of the available information are used in the updating of the time-varying parameter and the information is not limited to means and/or higher order moments of the predictive density. A more formal motivation to use the score as parameter driving mechanism is provided by Blasques et al. (2015) where it is argued that score-driven models are optimal in approximating the conditional observation density, even when the model is not correctly specified.

In the initial work of both Creal et al. (2013) and Harvey (2013), it has been argued that score-driven models also lead naturally to robust updating equations for parameters in models where fat-tailed and asymmetric densities are present. This feature of score-driven models make them quite appealing for modeling noisy economic and financial data. In particular, high-frequency data such as weekly, daily or intra-daily time series are often contaminated with much noise and outliers. Another key feature of score-driven models is that they are easy to implement and fast in computations. Score-driven models do not rely on complex algorithms or simulation-based methods for the estimation of fixed parameters and the filtering of time-varying parameters.

More specifically, a convenient feature of observation-driven models is that the likelihood has a closed-form expression that can be constructed from the prediction error decomposition. Hence, this class of models offers a convenient framework for its use in empirical work. This is in contrast to parameter-driven models where the unobserved time-varying parameter process has its own source of errors which need to be integrated out from the joint density function to obtain the likelihood. In many cases of empirical interest, integration cannot rely on closed-form expressions and the solution is typically found in numerical simulation-based methods which are computationally intensive. Still, it was shown by Koopman et al. (2016) that a score-driven model has a similar forecasting performance as the corresponding parameter-driven model, even when the latter is the true data generation process. Furthermore, it is relatively straightforward to extend observation-driven models to settings where, for example, nonlinear equations, asymmetric densities, and long memory dynamic processes are present. It may be concluded that score-driven models can be applied straightforwardly to a wide range of different and elaborate models.

The score-driven class of models encompasses many of the well known observation-driven models such as GARCH, ACD and ACI models. It has also given rise to a wide range of newly developed models which have proven to be useful for economic and financial applications. An overview of the literature on these more advanced score-driven models is provided in Artemova et al. (2022). Here, a thorough but still practical introduction of the methodology for score-driven models is presented. The remainder of the chapter consists of three parts. First, the general model specification and corresponding methodology are introduced. Second, the various aspects of score-driven models are examined for a basic location model. Third, a thorough discussion is presented for score-driven scale models with an emphasis on the Student's t conditional volatility model.

Score-Driven Model Specification

This section provides the statistical specification of the score-driven model and discusses its most important features for a univariate time series of T observations which is denoted by $y_1, y_2, y_3, \ldots, y_T$. Furthermore, it discusses the estimation of the fixed parameters in the model, the filtering of the time-varying parameters and the predictions and forecasting of the observations, in-sample and out-of-sample.

Model Specification

The score-driven model is specified as in Creal et al. (2013) and is given by

$$\begin{aligned} y_t &\sim p_y(y_t|f_t, \mathcal{F}_{t-1};\theta)\,, & f_{t+1} &= \omega + \alpha s_t + \beta f_t, \\ s_t &= S_t \cdot \nabla_t\,, & \nabla_t &= \frac{\partial \log p_y(y_t|f_t, \mathcal{F}_{t-1};\theta)}{\partial f_t}\,, \end{aligned} \tag{1}$$

where y_t is the time series observation at time t, for $t = 1, \ldots, T$, $p_y(y_t|f_t, \mathcal{F}_{t-1};\theta)$ denotes the predictive conditional density for observation y_t, f_t is the time-varying parameter for the conditional density, with $\mathcal{F}_{t-1}$ denoting the information set based on the observations $y_1, \ldots, y_{t-1}$, and θ is the parameter vector that contains fixed and unknown coefficients, including ω, α, and β, but also parameters that index the predictive density p_y. It is assumed that the observations become available sequentially over time. When the observation y_t arrives, the time-varying parameter f_t is updated using the *updating equation* $f_{t+1} = \omega + \alpha s_t + \beta f_t$. The *innovation* term for the updating equation is the scaled score $s_t = S_t \cdot \nabla_t$ where ∇_t denotes the score of the predictive conditional density

$p_y(y_t|f_t, \mathcal{F}_{t-1}; \theta)$ with respect to f_t and $S_t \equiv S(f_t, \mathcal{F}_{t-1}; \theta)$ is a scaling function that is discussed in the section ‘Possible Scaling Measures for the Score’. A similar formulation of the score-driven model is given by Harvey (2013).

The updating of f_t using the score of the conditional predictive density is intuitive. Namely, if a local ascent algorithm was used to find the local maximum of the predictive density at time t over f_t for a given parameter vector θ , then the score ∇_t indicates the direction in which the time-varying parameter must move, given the current position of f_t, according to the algorithm. For this reason, it is natural to use the score to determine the optimal updating step of f_t. This is underlined by Blasques et al. (2015), who prove the information-theoretic optimality of a simple class of score-driven models under certain regularity conditions. In particular, they show that time-varying parameter updates based on the score always reduce the local Kullback-Leibler divergence in expectation and in every step. These results hold for misspecified models as well. Also, they argue that any parameter update that is *not* based on the score does not have this property. These are all limit results. However, Blasques et al. (2020) show in a Monte Carlo study that the optimality also holds in finite samples for score-driven conditional volatility models.

Creal et al. (2008, 2013) discuss more options for generalizing the specification in (1). For example, the model can be specified with a more extensive lag structure and with exogenous covariates x_t in the updating equation for f_t. Also, the model can include other forms of nonlinearity in the updating equation, such as a regime-switching process and a long-memory process. Furthermore, a setting where the predictive density depends on past values of y_t and on some exogenous covariates x_t can be considered. Finally, y_t and f_t can be vectors, or even matrices, rather than scalars. In case f_t is a time-varying parameter vector, the score s_t is a vector and the scaling factor S_t is a matrix. Having an observation vector y_t results in having a multivariate score-driven model. A range of extensions including multivariate score-driven models is discussed in Artemova et al. (2022).

Possible Scaling Measures for the Score

The scaling function S_t allows us to determine how the score ∇_t impacts the updating at time t, to obtain the next f_{t+1}. It must be emphasized that a different choice for S_t results in an inherently different model, with a different set of statistical properties. Hence, the most suitable scaling function for a specific setting must be determined with some care.

Typically, a natural choice for the scaling factor is a function of the variance of the score ∇_t. Then the scaling is based on the curvature of the conditional log density of the t-th observation. More specifically, it is common to use the inverse asymptotic variance of the score, which is equal to the conditional information matrix under standard regularity conditions; so $S_t = \mathcal{I}_{t|t-1}^{-1}$ where $\mathcal{I}_{t|t-1} = \mathbb{E}_{t-1}[\nabla_t \nabla_t']$. This leads to the scaled score s_t having a variance of $\mathcal{I}_{t|t-1}^{-1}$. It is also an intuitive scaling choice, because in this case the updating equation of f_t in (1) can be interpreted as a Gauss-Newton algorithm for estimating f_t over time, as is pointed out by Creal et al. (2011). For this choice of S_t, well known models become special cases of the score-driven model. For example, a score-driven volatility model $y_t = f_t \varepsilon_t$, with error ε_t being standard normally distributed, reduces to the GARCH model of Bollerslev (1986) for this choice of scaling. Another example of a model that is encompassed by score-driven models for this choice of scaling is the multiplicative error (MEM) model of Engle and Gallo (2006), which in turn encompasses the autoregressive conditional duration and intensity (ACD and ACI) models of Engle and Russell (1998) and Russell (2001), respectively; see Creal et al. (2013) for a more extensive discussion.

Another possible scaling matrix is $S_t = \mathcal{J}_{t|t-1}$ where $\mathcal{J}_{t|t-1}$ is the square root of the (pseudo-)inverse information matrix such that $\mathcal{J}_{t|t-1}' \cdot \mathcal{J}_{t|t-1} = \mathcal{I}_{t|t-1}^{-1}$. This choice of S_t is particularly convenient, because it standardizes the score ∇_t such that s_t itself has a unit variance, which improves the tractability of the statistical properties of the model.

Finally, a typical last-resort option is to set $S_t = I$. For this choice of scaling, or essentially the lack of scaling, the statistical properties of the model will typically become more complicated. This is a clear disadvantage. However, when the information matrix $\mathcal{I}_{t|t-1}$ is a constant (scalar), or it does not depend on the time-varying parameter, the scaling is less relevant as the scaled score is scaled again through its multiplication by α in the updating equation. Such cases include log-scale models of the form $y_t = \exp(0.5 f_t)\varepsilon_t$ and location models of the form $y_t = f_t + \varepsilon_t$ where $S_t = \mathcal{I}_{t|t-1}^{-1}$ is proportional to $S_t = 1$.

Estimation, Filter Invertibility and Asymptotic Properties

As highlighted before, a convenient property of observation-driven models, and hence of score-driven models, is that an explicit expression of the log likelihood is available. Hence the estimation of the static parameter vector θ via the method of maximum likelihood (ML) is straightforward. Consider a sample of T observations $\{y_t\}_{t=1}^T$ generated as described in (1) under some true parameter θ_0. The true values of $\{f_t\}_{t=1}^T$ that are used to generate the observations are not observed. Therefore, a filtered sequence $\{\hat{f}_t(\theta)\}_{t=1}^T$ is constructed recursively using the updating equation of f_t for some value of θ and some starting value $\hat{f}_1$. The application of the prediction error decomposition and the use of the filtered sequence $\{\hat{f}_t(\theta)\}_{t=1}^T$ provide the following maximization problem

$$\hat{\theta}_T = \arg\max_{\theta} L_T(\theta)\,, \qquad \text{where} \quad L_T(\theta) = \sum_{t=1}^{T} \log p_y(y_t | \hat{f}_t(\theta); \theta)\,.$$

It is straightforward to evaluate the value of the log likelihood for some θ because it only requires two steps: (i) calculating the filtered time-varying parameter $\{\hat{f}_t(\theta)\}_{t=1}^T$ via the score-driven updating equation in (1), and (ii) calculating the log likelihood contribution of every y_t given $\hat{f}_t(\theta)$, for $t = 1, \ldots, T$.

Before turning to the asymptotic properties of this ML estimator, an important property that must be examined when using observation-driven filters is *filter invertibility*. Invertibility essentially means that the filtered sequence $\{\hat{f}_t(\theta)\}_{t=1}^T$ will 'forget' its starting value in the limit. The starting value of the time-varying parameter f_t is unobserved. Hence, the true starting point f_1 is unknown and is not available for initializing the sequence $\{\hat{f}_t(\theta)\}_{t=1}^T$. It needs to be replaced by some arbitrary value $\hat{f}_1$. Sometimes it is possible to estimate the starting value alongside the other parameters, but typically this is not preferred, especially when f_t represents a high-dimensional vector. It is important to notice that when filter invertibility fails, the true path $f_t(\theta_0)$ will not be retrieved in the limit, even if the true static parameter θ_0 is known. Clearly, this will be problematic if the filtered values $\hat{f}_t(\theta)$ are used to construct the log likelihood that is used for ML. This crucial point is sometimes overlooked by those who base their theoretical results on the implicit assumption that the value of f_1 is known exactly. Also, this assumption does not appear to be a realistic because the rest of the sequence $f_2, f_3, \ldots$ is assumed to be unobserved.

The concept of filter invertibility is discussed in Straumann and Mikosch (2006) who stress that it is a condition needed for applicability of the quasi ML estimation of a general class of GARCH models. The importance of filter invertibility for the EGARCH model is highlighted by Wintenberger (2013). Blasques et al. (2018) provide a way to determine invertibility regions when no feasible analytical invertibility conditions on the parameters are available. Blasques et al. (2022) give sufficient conditions for filter invertibility for score-driven models. As is usual in the stationary observation-driven literature, invertibility is established by showing that the filtered process converges almost surely to some unique stationary and ergodic limit process. They follow the approach of Straumann and Mikosch (2006), which adapts Bougerol (1993, Theorem 3.1) to obtain low-level conditions for filter invertibility for score filters and their first and second derivative processes. The latter results are useful for establishing asymptotic normality of the ML estimator.

After establishing the invertibility results for the score-driven model, Blasques et al. (2022) further provide sufficient conditions for consistency and asymptotic normality of the ML estimator $\hat{\theta}_T$. Two conditions are crucial in this development. First, the updating equation of the 'true' f_t must be contracting on average to ensure that the true time-varying parameter is stationary and ergodic. Second, the filtering equation of $\hat{f}_t$ must be uniformly contracting to ensure filter invertibility. The asymptotic properties derived by Blasques et al. (2022) are global and they rely on low-level conditions in terms of 'building blocks' of score-driven models which are represented by the equations provided in (1). For example, the derivatives of the score with respect to the parameters must have bounded moments up to some specific order. These results are particularly helpful for researchers who want to establish theoretical properties of the ML estimator for a specific score-driven model. Given certain conditions for the observations, the asymptotic properties remain to be applicable under potential model misspecification. These theoretical results do not trivially generalise to settings where the observations y_t and/or the time-varying parameter f_t are no longer univariate. However, specific asymptotic results can be established for particular multivariate score-driven models; see the discussions in Blasques et al. (2016) and Bazzi et al. (2017).

Score-Driven Location Models

This section illustrates the workings of score-driven models in a basic setting. The model for location is a natural starting point to obtain further insight into score-driven models. These models focus on the mean equation and are typically used for analysing macroeconomic time series.

Location Model Specification

When the aim is to filter the conditional mean $f_t = \mathbb{E}(y_t|\mathcal{F}_{t-1})$ of a (univariate) sample of observed data $\{y_t\}_{t=1}^{T}$, consider the score-driven filtering model as given by

$$y_t = f_t + \varepsilon_t \,, \tag{2}$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is assumed to be an independent and identically distributed (i.i.d.) random sequence with mean zero and probability density function p_ε, and where f_t is updated according to (1). Assume that ε_t is Gaussian with variance σ^2. Then $S_t = \mathcal{I}_{t|t-1}^{-1} = \sigma^2$, leads to a scaled score $s_t = y_t - f_t$ which yields the linear updating equation for f_t as given by

$$f_{t+1} \;=\; \omega + \alpha s_t + \beta f_t \;=\; \omega + \alpha(y_t - f_t) + \beta f_t\,. \tag{3}$$

The updating for f_t reduces to an exponentially weighted moving average (EWMA) recursion when coefficients have values $\omega = 0$, $\beta = 1$ and $0 < \alpha < 1$. When $\alpha > 0$, the updating of the conditional expectation f_t is intuitive since f_t is effectively the one-step ahead forecast of y_t and $y_t - f_t$ is the corresponding forecast error. Hence, the updating equation (3) takes into account the forecast error to construct f_{t+1}, which is defined as the one-step ahead forecast of the next observation y_{t+1}. Specifically, for $\alpha > 0$, if f_t has a lower (higher) value than y_t, then the scaled error $\alpha(y_t - f_t)$ will ensure that f_{t+1} increases (decreases) compared to f_t.

Substituting (2) into this equation shows that $\{f_t\}_{t\in\mathbb{Z}}$ follows an autoregressive process of order 1, an AR(1) process, with autoregressive coefficient β. Hence, a necessary and sufficient condition for stationarity of this sequence is that $|\beta| < 1$. The updating equation (3) implies that f_t is a weighted average of all past observations, where observation y_{t-j} has weight $\alpha(\beta-\alpha)^{j-1}$, for $j = 1, \ldots, t-1$. It follows immediately that the filtered sequence $\hat{f}_t$ is stationary in the limit, if and only if $|\beta - \alpha| < 1$. This is also the condition for filter invertibility.

Finally, by substituting (3) into (2), it follows that y_t is implicitly generated by an ARMA(1,1) process as given by

$$y_t = \omega + \beta y_{t-1} + \varepsilon_t + (\alpha - \beta)\varepsilon_{t-1}\,,$$

with autoregressive coefficient β and moving average coefficient $\alpha - \beta$. This is an interesting result and it applies to any distribution for ε_t. Notice that the process reduces to an AR(1) process when $\alpha = \beta$, and to an MA(1) process when $\beta = 0$. Higher order ARMA processes are obtained when higher order lags for f_t and $s_t = y_t - f_t$ are considered for the updating equation f_{t+1} in (1) or, more specifically, in (3).

Robust Filtering

The properties of the score-driven location model are almost identical to those of a parameter-driven Gaussian (stationary) signal plus noise model, which is obtained by replacing s_t with a Gaussian random error sequence in the updating equation (3). As shown by Harvey and Luati (2014), this parallel vanishes when considering a non-Gaussian predictive conditional density $p_y(y_t|f_t, \mathcal{F}_{t-1}; \theta)$ in (1). For instance, consider a fat-tailed conditional density such as the Student's t. The data generation process for this model will lead to many observations that would be referred to as "outliers" in a Gaussian context. The updating of $f_{t+1} = \omega + \alpha(y_t - f_t) + \beta f_t$ will not work well as large observations from previous times will be incorporated in the filtered level f_t: it will take time for the outlying observations to "work through the system". Hence, this dynamic model is not *robust* to large observations which are induced by a heavy-tailed conditional distribution. Therefore, it is preferred to adopt a model that accounts for these fat-tails. Score-driven models are designed to do this by considering fat-tailed densities for $p_y(y_t|f_t, \mathcal{F}_{t-1}; \theta)$ in (1). A more elaborate discussion is provided by Caivano et al. (2016, Section 2.3 and 2.4) where it is demonstrated that the conditional score reflects the tail shape of distributions and how this connects to robustness.

For instance, let ε_t in (2) be Student's t distributed with $\nu > 0$ degrees of freedom and scale parameter $\sigma > 0$. Then the conditional observation distribution becomes

$$f(y_t|f_t, \mathcal{F}_{t-1}; \theta) = \frac{\Gamma(\frac{\nu+1}{2})}{\sqrt{\pi\nu}\Gamma(\frac{\nu}{2})\sigma} \left(1 + \frac{(y_t - f_t)^2}{\nu\sigma^2}\right)^{-\frac{\nu+1}{2}},$$

implying that the score-driven updating equation will become

$$f_{t+1} = \omega + \alpha s_t + \beta f_t\,, \qquad s_t = \frac{y_t - f_t}{1 + \nu^{-1}\sigma^{-2}(y_t - f_t)^2}\,, \tag{4}$$

according to (1), using $S_t = (1 + \nu^{-1})^{-1}\sigma^2$ which is proportional to the inverse of the conditional information matrix $\mathcal{I}_{t|t-1} = (\nu+1)\sigma^{-2}/(\nu+3)$. Equations (2) and (4) together form the score-driven Student's t model proposed by Harvey (2013, Section 3.1) and Harvey and Luati (2014). Notice that the degrees of freedom parameter ν is only required to be positive, which is why the model is referred to as a location model and not as a mean model. If the degrees of freedom $\nu \to \infty$, the update becomes identical to the Gaussian update of (3), which is not surprising because in that case the Student's t distribution approaches a normal distribution. The score is plotted for various different choices of ν

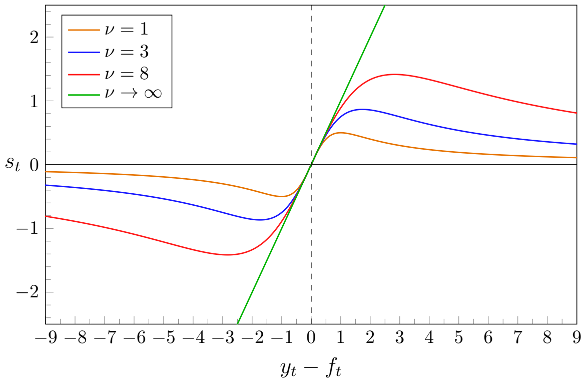

Figure 1: Student's t score plots for the location model. The score s_t of the Student's t location model in (4) for scale $\sigma = 1$ and different degrees of freedom ν.

in Figure 1. For finite values of ν, it is clear that the update in (4) downweights large values of $y_t - f_t$ and this downweighting is more severe for smaller values of ν. In that way the updating takes into account the heavy tails of the disturbances ε_t, because it implies that large prediction errors might be due to the nature of the innovations. As $|y_t - f_t|$ goes to infinity, the score s_t goes to zero. Because of the redescending nature of the score, Caivano and Harvey (2014) refer to the weighting induced by this score-driven filter as a parametric form of trimming. In the robustness literature, trimming is a common technique in which observations above some threshold receive a weight of zero. Hence, the Student's t location filter essentially induces a soft form of trimming.

Harvey and Luati (2014) provide asymptotic results for the ML estimator of this model including an explicit asymptotic variance matrix in terms of the model's parameters, but invertibility of the filter is not explicitly discussed. Blasques et al. (2022) have formulated conditions for consistency and asymptotic normality of the ML estimator of this model while taking account of the invertibility conditions. Blasques et al. (2018) have developed a weaker version of the invertibility condition for this model. It is also possible to adopt a non-stationary version of the score-driven local level model as given by

$$y_t = f_t + \varepsilon_t \,, \qquad f_{t+1} = f_t + \alpha s_t \,,$$

where s_t is defined as in (4); this model is suggested by Harvey and Luati (2014). It is a special case of the model in (2) and (4) with $\omega = 0$ and $\beta = 1$. This updating equation for f_t also implies an EWMA scheme but now with time-varying weights that account for the shape of the distribution; see Caivano et al. (2016). Furthermore, this model can be extended to a more general unobserved components

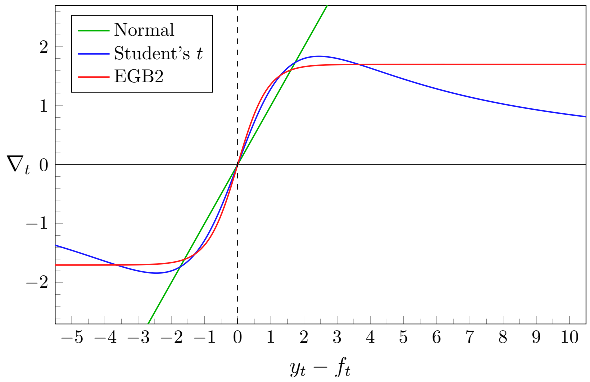

Figure 2: Plots of the score ∇_t corresponding to the location model for different distributions. The unscaled score ∇_t corresponding to the location model (2) for the Gaussian distribution (with $\sigma^2 = 1$), the Student's t distribution (with $\sigma^2 = 6/8$ and $\nu = 8$) and the symmetric EGB2 distribution (with $\sigma^2 = 1$ and $\xi = \varsigma \approx 0.7689$) which all have variance 1 and excess kurtosis 1.5 (except for the Gaussian distribution).

model with score-driven time-varying trend and seasonal components. Such specifications can also be extended with explanatory variables and fixed effects; see Harvey and Luati (2014).

Leptokurtosis and Asymmetry

Whereas the score-driven location filter based on Student's t innovations induces a form of trimming, the filter of a score-driven location model based on the exponential generalized beta distribution of the second kind (EGB2), as proposed by Caivano and Harvey (2014) and Caivano et al. (2016), has a *Winsorizing* property. Winsorizing is a well known approach for the treatment of large observations in the robustness literature. It entails an updating equation for f_{t+1} where s_t is set to a constant value when the observation passes some given threshold, see Maronna et al. (2006, Chapter 2).

The EGB2 distribution allows for asymmetry and leptokurtosis, but unlike the Student's t distribution, it has exponential tails instead of heavy tails. If ε_t in (2) is EGB2 distributed with mean 0, variance $\sigma^2 > 0$ and non-negative shape parameters ξ and ς, then

$$p_y(y_t|f_t, \mathcal{F}_{t-1}; \theta) = \frac{h \exp\left(\xi \left(h(y_t - f_t)/\sigma + \Delta\right)\right)}{\sigma B(\xi, \varsigma) \left(1 + \exp\left(h(y_t - f_t)/\sigma + \Delta\right)\right)^{\xi + \varsigma}}, \tag{5}$$

where $\Delta = \psi(\xi) - \psi(\varsigma)$, $h = \sqrt{\psi'(\xi) + \psi'(\varsigma)}$, $\psi(\cdot)$ is the digamma function, $\psi'(\cdot)$ is the trigamma function, and $B(\cdot, \cdot)$ is the beta function; see for instance Wang et al. (2001) for further details. The distribution is symmetric if $\xi = \varsigma$, positively skewed if $\xi > \varsigma$ and negatively skewed if $\xi < \varsigma$. The kurtosis decreases as ξ and ς increase and the distribution encompasses the normal distribution (if

$\xi = \varsigma \to \infty$) and the Laplace distribution (if $\xi = \varsigma = 0$). The scaled score s_t in (1), corresponding to the predictive density in (5), is given by

$$s_t = h\sigma \left[(\xi + \varsigma) \frac{\exp\left(h(y_t - f_t)/\sigma + \Delta\right)}{1 + \exp\left(h(y_t - f_t)/\sigma + \Delta\right)} - \xi \right] , \tag{6}$$

where the scale is set to $S_t = \sigma^2$ which is proportional to the inverse of the conditional information matrix. It is clear that the fraction in the scaled score s_t is uniformly bounded between $-h\sigma\xi$ and $h\sigma\varsigma$, and s_t converges to these values as $y_t - f_t \to -\infty$ and $y_t - f_t \to \infty$, respectively. Hence, this updating mechanism is effectively subject to a gentle or soft form of Winsorizing. The contribution of large observations is bounded, but it is not redescending like for Student's t distributed errors; see Figure 2 for a visual representation of this.

When $\xi \neq \varsigma$, the distribution is skewed and the score update becomes asymmetric. For example, if the distribution is negatively skewed, so if $\xi < \varsigma$, then large positive values get a larger weight than large negative values of the same magnitude. This is intuitive because negative skewness implies that negative spikes are more likely to occur than positive spikes and therefore the robust filter should be less sensitive to negative spikes.

Naturally, score-driven location models based on many other distributions can also be considered. This will lead to location models with different properties, because the score will take into account the shape of the distribution. Take for example a general error distribution (GED) also known as the exponential power distribution, with parameter ν, see (Harvey, 2013, Section 3.10). The GED encompasses the normal distribution (for $\nu = 2$) and the Laplace distribution (for $\nu = 1$). This GED is more peaked in comparison to the EGB2 distribution and has super-exponential tails for $\nu > 1$. This leads to the score being unbounded, unlike the score of the EGB2 distribution, but it will diverge at a lower rate than the score of the normal distribution if $\nu < 2$. Other distributions to consider for robust filtering are, for example, the Generalized t distribution of McDonald and Newey (1988) and skewed versions of all aforementioned symmetric distributions which can for instance be obtained by using the approach of Fernández and Steel (1998). The section 'Score-Driven Scale Models' considers these distributions in the context of score-driven scale models.

Illustration: Treasury Bill Rate Spreads

An empirical illustration is presented to show how the score-driven location models can perform in practice[1]. The data set under consideration consists of quarterly observations of the difference between the 3-month and the 6-month treasury bill (T-bill) rates[2]. The sample ranges from 1959:Q3 until 2021:Q1. The T-bill rate for a certain maturity is the yield received by investors for T-bills of that particular maturity in the secondary market. The difference between rates of different maturities is referred to as the spread. The left panel of Figure 3 shows a plot of the resulting T-bill rate spreads. This time series is rather noisy, as it has some sudden, and temporary spikes and drops, especially in the early 1980s. This seems to indicate the need for a robust filter if the goal is to filter some underlying location parameter.

To accommodate these concerns, score-driven models with Student's t and EGB2 innovations are considered, together with a score-driven model with Gaussian innovations for comparison. The corresponding updating functions are provided in equations (4), (6) and (3), respectively. The static parameters are estimated by the method of maximum likelihood. It is anticipated that the robust Student's t and EGB2 models will be the most suitable here, because of the occasional erratic behaviour of the data. The importance of robustness is confirmed by the estimation results reported in Table 1. Both the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) are the lowest for the model with Student's t innovations and highest for the model with the Gaussian innovations. The EGB2 model has a worse AIC and BIC value than the Student's t model, but they are both better than those of the Gaussian model. These results appear to suggest that the Student's t distribution, which has fat-tails, fits the data better than the EGB2 distribution, which has exponential tails.

Flexible distributions like the EGB2 distribution can lead to optimization problems during parameter estimation, especially for this illustration where the sample size is small. For this illustration, it appears that likelihood optimization for the EGB2 model enters into a saddlepoint, even when the process is restarted with new starting values. Hence, no standard errors are reported in this case. It appears that this numerical issue is caused by the joint estimation of ξ and ς. Therefore, Table 1 also reports the estimation results for the symmetric EGB2 model, which is obtained through the restriction $\xi = \varsigma$. For this model we do not encounter these numerical issues during the estimation process. This fitted model has a slightly worse AIC, but a slightly better BIC than its unrestricted version.

The estimated degrees of freedom parameter of the Student's t distribution is somewhat small. This estimation result implies that the error distribution is heavy-tailed which in turn causes the score s_t to become more robust against outliers. This is clearly visible in the plotted response curve of the models presented in the right panel of Figure 3. For small to moderate values, the Student's t filter reacts stronger than the Gaussian filter, but for larger values this is no longer the case. Indeed, the response of the former filter redescends back to zero in the limit, whereas that of the latter diverges to infinity. For the EGB2 models, the score follows a similar path as the one of the Student's t, but it is monotonically increasing instead of redescending. The unrestricted EGB2 model has a score function that is asymmetric, since $\xi > \varsigma$ implies a mildly positively skewed distribution (the skewness is 0.33). The left panel of Figure 3 presents the filtered location corresponding to the fitted models, where the filtered path of the symmetric EGB2 model is omitted because of its similarity to that of the asymmetric EGB2 model. It is clearly visible that the Gaussian filter reacts more to large observations than the Student's t and the EGB2 filters. The filtered paths of the Student's t and EGB2 model are very similar, but the latter tends to have a slightly stronger reaction to large shocks than the former.

Table 1: Maximum likelihood estimation results

The maximum likelihood estimates for the score-driven normal, Student's t and EGB2 location models, using data of T-bill rate spreads between 3 and 6 months maturity; see left panel of Figure 3. The symmetric EGB2 (sym.) model is subject to restriction $\xi = \varsigma$. Asymptotic standard errors are provided in parentheses. Note: * standard errors are not reported for EGB2 due to numerical issues.

	ω	α	β	σ^2	ν, ξ	ς	LL	AIC	BIC
Normal	0.609 (0.154)	0.679 (0.065)	0.553 (0.087)	1.763 (0.158)			-422.2	3.437	3.494
Student's t	0.353 (0.104)	1.567 (0.145)	0.714 (0.052)	0.516 (0.059)	2.632 (0.330)		-370.6	3.029	3.100
EGB2*	0.421	0.432	0.710	1.394	0.181	0.154	-376.8	3.087	3.172
EGB2 sym.	0.319 (0.054)	0.441 (0.029)	0.714 (0.031)	1.400 (0.177)	0.172 (0.019)		-378.8	3.095	3.167

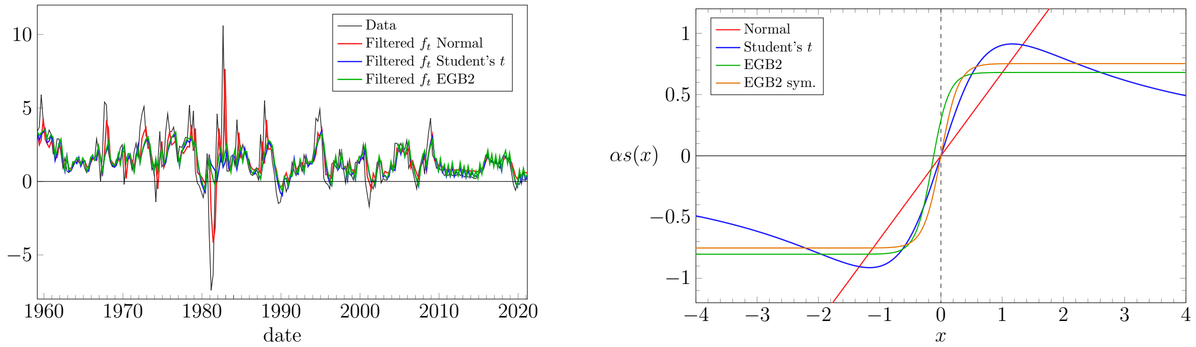

Figure 3: Data of T-bill rate spreads between 3 and 6 months maturity with the corresponding filtered location f_t of the Normal, Student's t and EGB2 models (left), and the corresponding estimated response curves (right).

Score-Driven Scale Models

Score-driven scale models have been widely studied in the literature. The conditional volatility model with the Student's t density, as discussed in both Harvey and Chakravarty (2008) and Creal et al. (2008), is one of the first compelling applications of score-driven models. In a similar way as for location models, score-driven scale models with fat-tailed innovations lead to filters that are robust against large observations. For financial data, robustness of the scale or volatility filter is especially relevant as the data tends to contain many "outliers" which do not necessarily imply a fundamental change in the underlying conditional volatility. However, it is important to emphasize that the potential robustness property of the resulting volatility filters is by no means the only motivation for considering score-driven scale models.

This section reviews univariate score-driven scale models. Artemova et al. (2022) considers multivariate scale models in which multiple volatilities and correlations can be modeled jointly.

Univariate Scale Models

For a univariate (demeaned) observation y_t, the conditional univariate scale is denoted by f_t and the basic score-driven scale model is formulated as

$$y_t = f_t^{1/2} \varepsilon_t \,, \tag{7}$$

where $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ is an i.i.d. sequence with mean zero such that $\mathbb{E}(y_t | \mathcal{F}_{t-1}) = 0$. When the variance of ε_t is equal to one, the model (7) reduces to a volatility model where f_t is equal to the conditional variance, that is $\mathrm{Var}(y_t|\mathcal{F}_{t-1}) = f_t$. However, this restriction is not necessary, the variance of ε_t could even be infinite. Instead of using demeaned observations, it is straightforward to extend the model to allow for a non-zero mean μ and for autoregressive or ARMA dynamics in the observation equation. To keep this treatment simple, assume that observations y_t are generated by model (7).

A basic illustration is obtained by taking $\varepsilon_t \sim \mathcal{N}(0,1)$ for every t. If the conditional variance f_t is updated according to the score-driven framework in (1), with scaling factor $S_t = \mathcal{I}_{t|t-1}^{-1} = 2f_t^2$, the updating equation becomes

$$f_{t+1} = \omega + \alpha(y_t^2 - f_t) + \beta f_t \tag{8}$$

where $\omega > 0$, $\alpha \geq 0$ and $\beta \geq \alpha$ to ensure positivity of f_t. It is not hard to see that the resulting model is equivalent to a regular GARCH model of Bollerslev (1986) with parameters α and $\beta - \alpha$. There are certain parameter restrictions that impose stationarity and filter invertibility (Straumann and Mikosch, 2006). If instead a Student's t distribution with $\nu > 0$ degrees of freedom is considered, without changing the updating equation of f_t, the resulting model is the GARCH-t model of Bollerslev (1987). However, in case a score-driven updating equation is used, the resulting model departs from the regular GARCH framework, because then:

$$f_{t+1} = \omega + \alpha \left(\frac{(1+\nu)\nu^{-1} y_t^2}{1 + \nu^{-1} y_t^2 / f_t} - f_t \right) + \beta f_t \,, \tag{9}$$

where $\omega > 0$, $\alpha > 0$ and $\beta \geq \alpha$, to impose positivity, and where the used scaling factor $S_t = 2f_t^2$ is proportional to the inverse of conditional information matrix; see Table 2. Notice that this model is not based on a standardized Student's t distribution and that $\nu > 2$ is not required, which is why this model is referred to as a scale model rather than a volatility model. Using a standardized Student's t distribution with unit variance is also possible; see Creal et al. (2013) for an example. Similar to the Student's t location model in (2) and (4), it is clear that for finite values of ν, large values of y_t^2 are downweighted, unlike in the Gaussian scale model. For $\nu \to \infty$ the updating function becomes identical to (8), which also follows from the fact that the Student's t distribution reduces to a Gaussian distribution in that case. Blasques et al. (2022) give (parameter) restrictions for filter invertibility, consistency and asymptotic normality of the MLE for this model in their main example.

Univariate Log-Scale Models

A point of concern for scale models is that filtered conditional volatilities should be strictly positive. In score models, the positivity of the conditional volatility can be typically ensured by imposing appropriate parameter restrictions. To avoid having to impose positivity restrictions on the parameters, it may be convenient to choose a different parametrization, where the logarithm of the scale is modeled instead of the scale itself. An advantage of this choice of parametrization is also that the stationarity condition is relatively straightforward, as it will simply be $|\beta| < 1$. So instead of modeling the scale $f_t = \sigma_t^2$ in $y_t = \sigma_t \varepsilon_t$, the log scale $f_t = \log(\sigma_t^2)$ is modeled now, which gives the following log scale model

$$y_t = \exp\left(\frac{1}{2} f_t\right) \varepsilon_t \,, \tag{10}$$

where ε_t is again i.i.d. with mean zero and where f_t is updated according to (1) based on the distribution of the innovations. So if ε_t has unit variance, then $\exp(f_t) = \mathrm{Var}(y_t|\mathcal{F}_{t-1})$. Creal et al. (2013) state that the score-driven log scale model is equivalent to the well known EGARCH model of Nelson (1991) if ε_t has an asymmetric Laplace distribution. For a selection of other distributions, Table 2 reports the score ∇_t and the conditional information matrix corresponding to the log scale parametrization. As was pointed out before, the conditional information matrix does not depend on f_t in this case, so the choice of scaling is less critical here. For example, using the Student's t distribution leads to:

$$f_{t+1} = \omega + \alpha \left((1+\nu) \frac{\nu^{-1} y_t^2 / \exp(f_t)}{1 + \nu^{-1} y_t^2 / \exp(f_t)} - 1 \right) + \beta f_t \,, \tag{11}$$

for a scaling factor equal to $S_t = 2$; see Creal et al. (2011, 2013) and Harvey (2013, Chapter 4). In the latter reference, the model is referred to as the Beta-t-EGARCH model since the fraction in the score function has a Beta distribution, if evaluated at the true parameter. This property makes it a theoretically appealing model because the asymptotic properties of the ML estimator of the parameter vector in this model can be derived in a convenient manner; see Harvey (2013, Section 4.2). This model has been adopted in many empirical studies in the literature, and it has been shown to outperform the regular GARCH and GARCH-t models in most of these studies; see, for example, Harvey and Sucarrat (2014), Blazsek et al. (2016) and Catania and Nonejad (2020).

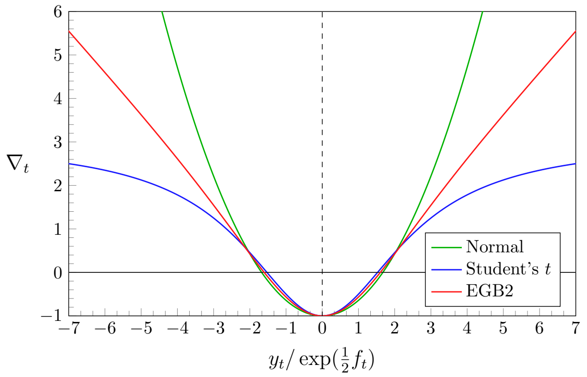

Figure 4: News impact curves of different score-driven log scale models. News impact curves (i.e. the plots of the score ∇_t as a function of the standardized observations $y_t/\exp(f_t/2)$) corresponding to the log scale model (10) for the Gaussian distribution (with $\sigma^2 = 7/5$), the Student's t distribution (with $\nu = 7$) and the symmetric EGB2 distribution (with $\xi = \varsigma \approx 1.874$), all with variance $7/5$.

Table 2 also gives the score of the EGB2 log scale model. The corresponding score-driven model is discussed briefly in Caivano and Harvey (2014), for a slightly different parametrization and with the main focus on a model where both the location and scale are score-driven.

Figure 4 plots the news impact curves (NIC) of score-driven log scale models with a normal, Student's t and EGB2 distribution with the same variance. The NIC is calculated by evaluating the score ∇_t as a function of the standardized observations $y_t/\exp(f_t/2)$. The bounded score of the Student's t distribution reflects the fact that it is a fat-tailed distribution and therefore large observations do not imply a proportionally large increase in the underlying volatility process. On the contrary, the score of the EGB2 distribution does diverge, which reflects the exponential tails of the distribution. This divergence is linear instead of a quadratic, which reflects that it has excess kurtosis. It is interesting to compare the shape of these NICs to those of the location model that were plotted in Figure 2.

A more flexible alternative to the Student's t distribution is the generalized t distribution proposed by McDonald and Newey (1988). Harvey and Lange (2017) consider a score-driven scale model based on this distribution. The corresponding predictive density for two positive shape parameters ν and h is given in Table 2. For $h = 2$ this is a regular Student's t distribution with ν degrees of freedom, while a GED distribution with parameter h is obtained if $\nu \to \infty$. The GED with $\nu = 2$ is the normal distribution, whereas $\nu = 1$ results in a Laplace distribution. Table 2 presents the predictive density, score and information matrix for a log scale model under the GED. The updating of the log scale

Table 2: Score-driven updates for scale models $y_t = \sigma_t \varepsilon_t$ for a selection of distributions and parametrizations

Parameter	Distribution	$p(y_t\|f_t)$	$\nabla_t = \partial \log p(y_t\|f_t)/\partial f_t$	$\mathcal{I}_{t\|t-1}$
$f_t = \sigma_t^2$	$\varepsilon_t \sim \mathcal{N}(0,1)$	$\frac{1}{\sqrt{2\pi f_t}} \exp(-\frac{y_t^2}{2 f_t})$	$\frac{1}{2}(\frac{y_t^2}{f_t^2} - f_t^{-1})$	$\frac{1}{2 f_t^2}$
$f_t = \sigma_t^2$	$\varepsilon_t \sim t(\nu)$	$\frac{f_t^{-1/2}}{B(\frac{1}{2},\frac{\nu}{2})\sqrt{\nu}} \left(1 + \frac{y_t^2}{\nu f_t}\right)^{-\frac{\nu+1}{2}}$	$\frac{1}{2}\left(\frac{(\nu+1)\nu^{-1} y_t^2 f_t^{-2}}{1+\nu^{-1} y_t^2 f_t^{-1}} - f_t^{-1}\right)$	$\frac{\nu}{2(3+\nu) f_t^2}$
$f_t = \log \sigma_t^2$	$\varepsilon_t \sim \mathcal{N}(0,1)$	$\frac{1}{\sqrt{2\pi} \exp(\frac{1}{2} f_t)} \exp(-\frac{y_t^2}{2\exp(f_t)})$	$\frac{1}{2}(\frac{y_t^2}{\exp f_t} - 1)$	$\frac{1}{2}$
$f_t = \log \sigma_t^2$	$\varepsilon_t \sim t(\nu)$	$\frac{\exp(-\frac{1}{2} f_t)}{B(\frac{1}{2},\frac{\nu}{2})\sqrt{\nu}} \left(1 + \frac{y_t^2}{\nu \exp(f_t)}\right)^{-\frac{\nu+1}{2}}$	$\frac{1}{2}\left(\frac{(\nu+1)\nu^{-1} y_t^2 \exp(-f_t)}{1+\nu^{-1} y_t^2 \exp(-f_t)} - 1\right)$	$\frac{\nu}{2(3+\nu)}$
$f_t = \log \sigma_t^2$	$\varepsilon_t \sim \text{GED}(\nu)^{(a)}$	$\frac{K_{\text{GE}}(\nu)}{\exp(\frac{1}{2} f_t)} \exp\left(-\frac{1}{\nu} \frac{\|y_t\|^\nu}{\exp(\frac{1}{2}\nu f_t)}\right)$	$\frac{1}{2}\left(\frac{\|y_t\|^\nu}{\exp\left(\frac{1}{2}\nu f_t\right)} - 1\right)$	$\frac{\nu}{4}$
$f_t = \log \sigma_t^2$	$\varepsilon_t \sim \text{Gen-}t(\nu,h)^{(b)}$	$\frac{K(\nu,h)}{\exp(\frac{1}{2} f_t)} \left(1 + \frac{\|y_t\|^h}{\nu \exp(\frac{h}{2} f_t)}\right)^{-\frac{\nu+1}{h}}$	$\frac{1}{2}\left(\frac{(\nu+1)(\|y_t\| \exp(-\frac{1}{2} f_t))^h/\nu}{1+(\|y_t\| \exp(-\frac{1}{2} f_t))^h/\nu} - 1\right)$	$\frac{\nu h}{4(1+h+\nu)}$
$f_t = \log \sigma_t^2$	$\varepsilon_t \sim \text{EGB2}(\xi,\varsigma)^{(c)}$	$\frac{h \exp(-\frac{1}{2} f_t) \exp\left(\xi\left(\frac{h y_t}{\exp(\frac{1}{2} f_t)} + \Delta\right)\right)}{B(\xi,\varsigma)\left(1+\exp\left(h \frac{y_t}{\exp(\frac{1}{2} f_t)} + \Delta\right)\right)^{\xi+\varsigma}}$	$\frac{1}{2}\left([(\xi+\varsigma) b_t - \xi] \frac{h y_t}{\exp(\frac{1}{2} f_t)} - 1\right)$ where $b_t = \frac{\exp\left(\frac{h y_t}{\exp(\frac{1}{2} f_t)} + \Delta\right)}{1+\exp\left(\frac{h y_t}{\exp(\frac{1}{2} f_t)} + \Delta\right)}$	$\propto 1$

The functions are defined as follows: $B(a,b) = \Gamma(a)\Gamma(b)/\Gamma(a+b)$, $K(\nu,h) = h/[2\nu^{1/h} B(1/h, \nu/h)]$ and $K_{\text{GE}}(\nu) = \nu^{1-1/\nu}/[2\,\Gamma(1/\nu)]$.

$^{(a)}$ This is the standardized general error distribution (GED) pdf of Zhu and Zinde-Walsh (2009).

$^{(b)}$ This is the generalized t distribution introduced by McDonald and Newey (1988).

$^{(c)}$ This is the standardized EGB2 distribution with mean zero and variance 1, where h and Δ are as in equation (5).

parameter f_t of the score-driven model based on the generalized t distribution is given by:

$$f_{t+1} = \omega + \alpha \left((\nu+1) \frac{\left(|y_t| \exp(-\frac{1}{2} f_t)\right)^h / \nu}{1 + \left(|y_t| \exp(-\frac{1}{2} f_t)\right)^h / \nu} - 1 \right) + \beta f_t \,,$$

using $S_t = 2$. So it is clear that this update has the same robustness properties as the Student's t scale filter in (11) as long as $\nu < \infty$, because in that case the score is bounded. Harvey and Lange (2017) provide explicit asymptotic results for the distribution of the ML estimator. They also propose Likelihood Ratio (LR) and Lagrange Multiplier (LM) tests for the null hypothesis of thin tails (that is $\nu \to \infty$) against the alternative hypothesis of fat tails (that is $\nu < \infty$).

Univariate Log-Scale Models with Skewness

Another way to introduce flexibility into the score-driven model is to use a skewed error distribution. For instance, Harvey and Sucarrat (2014) consider constructing score-driven scale models based on distributions skewed by the method of Fernández and Steel (1998). Any probability density function $p_\varepsilon(\varepsilon)$ that is unimodal and symmetric around zero, can be converted to a skewed density function as follows:

$$p(\varepsilon_t) = \frac{2}{\gamma + \gamma^{-1}} \, p_\varepsilon \left(\frac{\varepsilon_t}{\gamma^{\mathrm{sgn}(\varepsilon_t)}} \right) \,,$$

where $0 < \gamma < \infty$ is the skewing parameter, and for $\gamma = 1$, $\gamma < 1$ and $\gamma > 1$ the distribution is symmetric, left and right skewed, respectively. If this skewing method is applied to the Student's t distribution, the score is given by

$$s_t = \begin{cases} (\nu+1) \frac{y_t^2/(\nu\gamma^{-2}\exp(f_t))}{1+y_t^2/(\nu\gamma^{-2}\exp(f_t))} - 1 \,, & y_t < 0 \,, \\ (\nu+1) \frac{y_t^2/(\nu\gamma^{2}\exp(f_t))}{1+y_t^2/(\nu\gamma^{2}\exp(f_t))} - 1 \,, & y_t \geq 0 \,. \end{cases}$$

It follows that skewing the distribution directly induces asymmetry in the score and therefore in the update of the log scale parameter f_t. Figure 5 presents a plot of the news impact curve for some different choices of γ. It is clear that for $\gamma < 1$, negative values are downweighted more, whereas positive values are downweighted less than for the symmetric distribution. Hence, under negative skewness, the score update takes into account that positive values are relatively less common than negative values of the same magnitude and should therefore receive a relatively higher weight. The

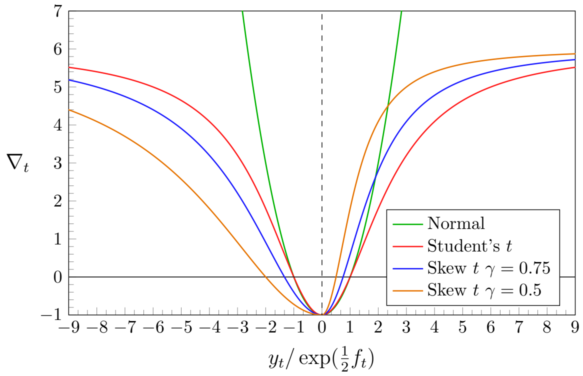

Figure 5: News impact curves of score-driven log scale model for the skewed t distribution. News impact curves (i.e. the plots of the score ∇_t as a function of the standardized observations $y_t/\exp(f_t/2)$) corresponding to the log scale model (10) for the skewed Student's t distribution with $\nu = 6$ degrees of freedom and for different values of skew parameter γ.

asymptotic results of the symmetric score-driven Student's t log scale model generalize straightforwardly to this skewed variant. The same skewing approach can be used for other distributions such as the GED.

It is important to realize that because of the skewness, the expectation of the innovations is no longer equal to zero, which can be solved by reformulating the observation equation of the model as

$$y_t = \exp\left(\frac{1}{2}f_t\right)(\varepsilon_t - \mu_\varepsilon) \tag{12}$$

where $\mu_\varepsilon = \mathbb{E}[\varepsilon_t]$. The score and thereby the updating equation of f_t must be altered accordingly. A more detailed discussion is provided by the study of Harvey and Sucarrat (2014, Section 4); in their empirical study they show that the score-driven skewed t model generally outperforms competing models, including alternative skewed models such as the GJR GARCH model (Glosten et al., 1993) with skewed innovations and the Normal Mixture GARCH model (Alexander and Lazar, 2006).

There are also alternative ways to skew a distribution. For example, Harvey and Lange (2017) consider a score-driven scale model based on a skewed Generalized t distribution, using the skewing method of Zhu and Galbraith (2010), which uses a slightly different parametrization than that of Fernández and Steel (1998). Furthermore, not only skewness is introduced, but also the shape parameters ν and h are allowed to have a different value above and below zero. So apart from skewness, a more explicit form of asymmetry is introduced to the distribution as well. Harvey and Lange (2017) demonstrate that the resulting score-driven model works well in practice, although it is recommended to simplify it by testing different parameter restrictions using LR, Wald or LM tests.

Volatility in Mean

Multiple extensions of score-driven scale models have been proposed in the literature. As an illustrative example, consider the volatility-in-mean model proposed in Harvey and Lange (2018). This model alters the log specification of the ARCH in mean (ARCH-M) model of Engle et al. (1987) to have score-driven updating of the standard deviation. So instead of the regular log scale specification in (10), now the time-varying scale parameter $\exp(0.5f_t)$ also occurs in the mean equation of the observations y_t:

$$y_t = \mu + \lambda \exp\left(\frac{1}{2}f_t\right) + \exp\left(\frac{1}{2}f_t\right)\varepsilon_t \,,$$

where the innovation sequence $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is i.i.d. with mean zero and fixed variance, and where the dynamics of f_t are again score-driven as in (1). Here μ and λ are fixed but unknown coefficients that determine the level and the 'ARCH-M' effect, respectively. The motivation for including the conditional scale parameter in the mean equation is that volatility of stock returns tends to be positively correlated with the level of the returns, because the equity risk premium increases if uncertainty rises.

When the innovations ε_t are Student's t distributed with ν degrees of freedom, Harvey and Lange (2018) show that,

$$s_t = (\nu+1)b_t - 1 + \lambda(1-b_t)\frac{\nu+1}{\nu}\left(\frac{y_t-\mu}{\exp(\frac{1}{2}f_t)} - \lambda\right)\,, \text{ where } \quad b_t = \frac{\left(\frac{y_t-\mu}{\exp(\frac{1}{2}f_t)} - \lambda\right)^2/\nu}{1+\left(\frac{y_t-\mu}{\exp(\frac{1}{2}f_t)} - \lambda\right)^2/\nu}\,.$$

If $\lambda = 0$ (and $\mu = 0$), the regular log scale update of (11) is recovered. In practice, the last term of the score has little impact on the filter, because λ is typically very small. Harvey and Lange (2018) derive theoretical properties for the resulting model, including moments and asymptotic results of the maximum likelihood estimator.

Illustration: Electricity Spot Price Returns

To demonstrate the robustness property of score-driven scale models with heavy-tailed distributions, this section compares the Gaussian and Student's t volatility model for a time series of electricity spot price returns[1]. The data under consideration is a sequence of daily returns of the PJM electricity

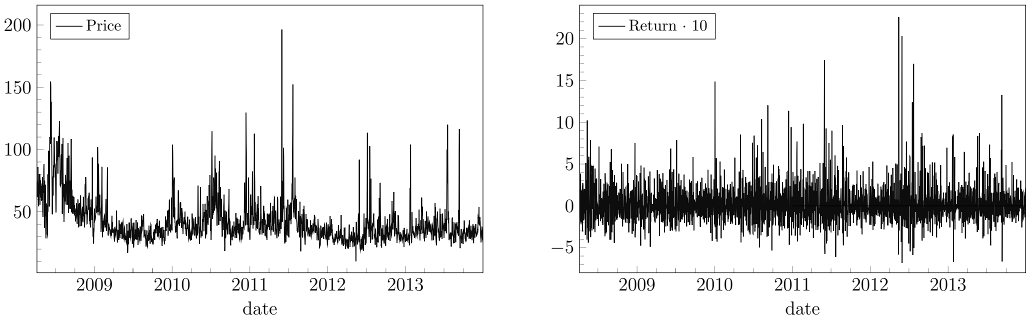

Figure 6: Electricity prices and returns. The average daily electricity price in the market of the PJM market from 06-04-2008 up to 31-12-2013 (left panel) and the corresponding returns (i.e. $(p_t - p_{t-1})/p_{t-1}$ where p_t denotes the price) scaled by a factor of ten (right panel).

market, which serves 13 states in the United States. The data[3.] span from 06-04-2008 to 31-12-2013 (2096 observations). Electricity prices are known to show occasional extreme behaviour in the form of sudden positive spikes, after which they quickly move back to the pre-spike level. The primary cause of this is the non-storability of electricity; see e.g. Escribano et al. (2011). In effect, the returns based on these prices also display these incidental "outliers". Figure 6 clearly shows that these properties are observable for the prices and returns currently under consideration. Models that are robust to sporadic extreme observations, such as the Student's t score-driven model, are thus most likely to be valuable in modeling conditional volatility dynamics of these data.

Consider the log scale specification in (10) with a non-zero mean μ: $y_t = \mu + \exp(\frac{1}{2} f_t)\varepsilon_t$. The scaling factor is chosen to be equal to the inverse of the conditional information matrix, $S_t = \mathcal{I}_{t|t-1}^{-1}$. Table 2 reports the score and information matrix of the Gaussian and Student's t distribution that were used to construct the models.

The resulting maximum likelihood estimates are reported in Table 3 alongside the corresponding likelihood-based criteria. According to both reported information criteria, Akaike's information criterion (AIC) and the Bayesian information criterion (BIC), the Student's t model has a better

Table 3: Maximum likelihood estimation results The parameter estimates from the score-driven Gaussian and Student's t volatility models, for the PJM electricity price returns as presented in Figure 6. Standard errors are in parentheses.

	μ	ω	α	β	ν	LL	AIC	BIC
Gaussian	0.316 (0.063)	0.125 (0.058)	0.074 (0.034)	0.927 (0.033)		-4770.5	4.56	4.57
Student's t	0.031 (0.052)	0.091 (0.120)	0.073 (0.048)	0.916 (0.110)	4.354 (0.459)	-4611.5	4.41	4.42

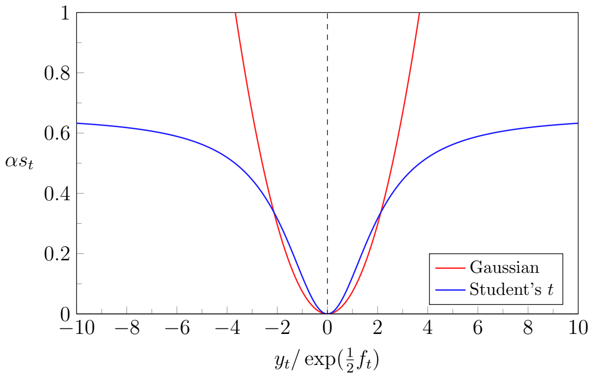

Figure 7: News impact curves of estimated models log scale models. The news impact curves (i.e. the estimated response αs_t as a function of the standardized observations $y_t/\exp(f_t/2)$) corresponding to the Gaussian and Student's t score-driven volatility models fitted to PJM electricity spot price returns.

in-sample log likelihood value than the Gaussian model, taking into account that the former model has one extra parameter. The estimates of α and β are similar for both models, but because the estimated degrees of freedom parameter ν is small, the two resulting models are inherently different.

Figure 7 shows the news impact curves of the estimated models. The estimated degrees of freedom parameter ν is low, so the estimated error density has heavy tails, and therefore the score update of the Student's t model is robust to large values. This is visible in the figure, as the response value corresponding to this model is bounded by a constant value. For the Gaussian model, this is not the case, because the response increases at a quadratic rate as y_t grows. Figure 8 shows the filtered volatility corresponding to the two fitted models alongside the absolute returns. As the news impact curves indicate, the filtered volatilities of the Gaussian model are impacted more by large observations than those of the Student's t model. After a spike in the returns, it takes a while for the Gaussian filtered volatility to move back to the pre-spike level. The filtered volatilities of the Student's t model also react to large observations, but in a less extreme manner. A spike in the returns does not indicate a large change in the underlying volatility process, given the fact that prices quickly move back to their original level. Therefore, the robust volatility filter of the score-driven Student's t model seems to be more suitable than the non-robust filter of the Gaussian model, since the former is less sensitive to these temporary increases in the returns.

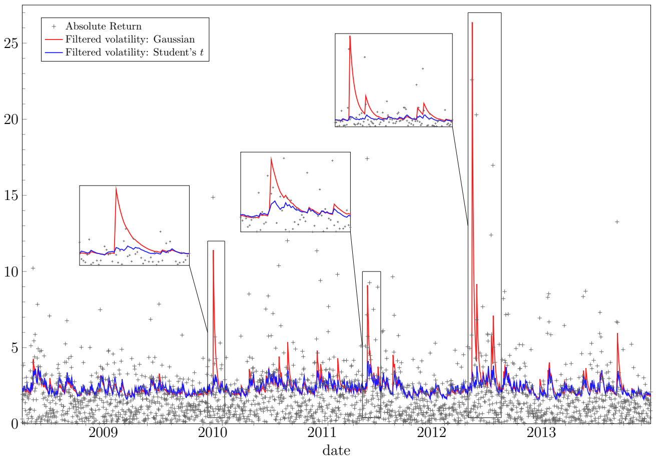

Figure 8: Filtered volatilities. The absolute PJM electricity price returns and the corresponding filtered volatilities of the Gaussian model (i.e. $\exp(\frac{1}{2}f_t)$) and Student's t model (i.e. $\sqrt{\nu/\nu-2}\exp(\frac{1}{2}f_t)$).

Acknowledgement

The authors are grateful to Ekaterina Ugulava for the outstanding support that she offered in her role as research assistant. Also, the authors thank MSc student Anthony van Veen for contributing to the scale model application.

Further reading

Ardia, D., Boudt, K., and Catania, L. (2019). Generalized autoregressive score models in R: The GAS package. *Journal of Statistical Software*, 88(6):1–28.

Artemova, M., Blasques, F., van Brummelen, J., and Koopman, S. J. (2021). Score-Driven Models: Methods and Applications. In *Oxford Research Encyclopedia of Economics and Finance*, page tba. Oxford University Press.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, volume 52. Cambridge University Press.

Harvey, A. C. (2021). Score-driven time series models. *Annual Review of Statistics and Its Application*. forthcoming.

Harvey, A. C. (2022). Score-driven time series models. *Annual Review of Statistics and Its Application*, 9:321–342.

References

Alexander, C. and Lazar, E. (2006). Normal mixture GARCH(1,1): Applications to exchange rate modelling. *Journal of Applied Econometrics*, 21(3):307–336.

Artemova, M., Blasques, F., van Brummelen, J., and Koopman, S. J. (2022). Score-Driven Models: Methods and Applications. In *Oxford Research Encyclopedia of Economics and Finance*.

Bazzi, M., Blasques, F., Koopman, S. J., and Lucas, A. (2017). Time-varying transition probabilities for Markov regime switching models. *Journal of Time Series Analysis*, 38(3):458–478.

Blasques, F., Gorgi, P., Koopman, S. J., Wintenberger, O., et al. (2018). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics*, 12(1):1019–1052.

Blasques, F., Koopman, S. J., and Lucas, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika*, 102(2):325–343.

Blasques, F., Koopman, S. J., Lucas, A., and Schaumburg, J. (2016). Spillover dynamics for systemic risk measurement using spatial financial time series models. *Journal of Econometrics*, 195(2):211–223.

Blasques, F., Lucas, A., and van Vlodrop, A. C. (2020). Finite sample optimality of score-driven volatility models: Some Monte Carlo evidence. *Econometrics and Statistics*.

Blasques, F., van Brummelen, J., Koopman, S. J., and Lucas, A. (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics*, 227(2):325–346.

Blazsek, S., Chavez, H., and Mendez, C. (2016). Model stability and forecast performance of Beta-t-EGARCH. *Applied Economics Letters*, 23(17):1219–1223.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3):307–327.

Bollerslev, T. (1987). A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics*, 69(3):542–547.

Bougerol, P. (1993). Kalman filtering with random coefficients and contractions. *SIAM Journal on Control and Optimization*, 31(4):942–959.

Caivano, M. and Harvey, A. C. (2014). Time-series models with an EGB2 conditional distribution. *Journal of Time Series Analysis*, 35(6):558–571.

Caivano, M., Harvey, A. C., and Luati, A. (2016). Robust time series models with trend and seasonal components. *SERIEs*, 7(1):99–120.

Catania, L. and Nonejad, N. (2020). Density forecasts and the leverage effect: Evidence from observation and parameter-driven volatility models. *The European Journal of Finance*, 26(2-3):100–118.

Cox, D. R. (1981). Statistical analysis of time series: Some recent developments. *Scandinavian Journal of Statistics*, 8(2):93–115.

Creal, D., Koopman, S. J., and Lucas, A. (2008). A general framework for observation driven time-varying parameter models. *Tinbergen Institute Discussion Paper 08-108/4*.

Creal, D., Koopman, S. J., and Lucas, A. (2011). A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics*, 29(4):552–563.

Creal, D., Koopman, S. J., and Lucas, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics*, 28(5):777–795.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, 50:987–1007.

Engle, R. F. and Gallo, G. M. (2006). A multiple indicators model for volatility using intra-daily data. *Journal of Econometrics*, 131(1-2):3–27.

Engle, R. F., Lilien, D. M., and Robins, R. P. (1987). Estimating time varying risk premia in the term structure: The ARCH-M model. *Econometrica*, 55:391–407.

Engle, R. F. and Russell, J. R. (1998). Autoregressive conditional duration: a new model for irregularly spaced transaction data. *Econometrica*, 66:1127–1162.

Escribano, A., Ignacio Peña, J., and Villaplana, P. (2011). Modelling electricity prices: International evidence. *Oxford Bulletin of Economics and Statistics*, 73(5):622–650.

Fernández, C. and Steel, M. F. (1998). On Bayesian modeling of fat tails and skewness. *Journal of the American Statistical Association*, 93(441):359–371.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, 48(5):1779–1801.

Harvey, A. C. (1981). *Time Series Models*. Philip Allen.

Harvey, A. C. (2013). *Dynamic models for volatility and heavy tails: With applications to financial and economic time series*, volume 52. Cambridge University Press.

Harvey, A. C. and Chakravarty, T. (2008). Beta-*t*-(E)GARCH. *University of Cambridge, Faculty of Economics, Working paper CWPE 08340.*

Harvey, A. C. and Lange, R.-J. (2017). Volatility modeling with a generalized t distribution. *Journal of Time Series Analysis*, 38(2):175–190.

Harvey, A. C. and Lange, R.-J. (2018). Modeling the interactions between volatility and returns using EGARCH-M. *Journal of Time Series Analysis*, 39(6):909–919.

Harvey, A. C. and Luati, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association*, 109(507):1112–1122.

Harvey, A. C. and Sucarrat, G. (2014). EGARCH models with fat tails, skewness and leverage. *Computational Statistics & Data Analysis*, 76:320–338.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Transactions of the ASME–Journal of Basic Engineering*, 82(Series D):35–45.

Koopman, S. J., Lucas, A., and Scharth, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *The Review of Economics and Statistics*, 98(1):97–110.

Maronna, R. A., Martin, D., and Yohai, V. (2006). *Robust Statistics: Theory and Methods*. John Wiley & Sons, Ltd.

McCracken, M. and Ng, S. (2020). FRED-QD: A quarterly database for macroeconomic research. Working Paper 2020-005, Federal Reserve Bank of St. Louis.

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

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

Plackett, R. L. (1950). Some theorems in least squares. *Biometrika*, 37(1-2):149–157.

Russell, J. R. (2001). Econometric modeling of multivariate irregularly-spaced high-frequency data. Technical report, University of Chicago.

Straumann, D. and Mikosch, T. (2006). Quasi-maximum-likelihood estimation in conditionally heteroscedastic time series: A stochastic recurrence equations approach. *The Annals of Statistics*, 34(5):2449–2495.

Wang, K.-L., Fawson, C., Barrett, C. B., and McDonald, J. B. (2001). A flexible parametric GARCH model with an application to exchange rates. *Journal of Applied Econometrics*, 16(4):521–536.

Wintenberger, O. (2013). Continuous invertibility and stable QML estimation of the EGARCH(1,1) model. *Scandinavian Journal of Statistics*, 40(4):846–867.

Zhu, D. and Galbraith, J. W. (2010). A generalized asymmetric Student-t distribution with application to financial econometrics. *Journal of Econometrics*, 157(2):297–305.

Zhu, D. and Zinde-Walsh, V. (2009). Properties and estimation of asymmetric exponential power distribution. *Journal of Econometrics*, 148(1):86–99.

Notes

1. The code developed for both illustrations are made available on the GAS website, in the code section.
2. The 6-month minus 3-month treasury bill rate [TB6M3Mx] was obtained from the FRED-QD database of McCracken and Ng (2020). The data were multiplied by a factor of 10 for scaling purposes.
3. Data were retrieved from the PJM website. Daily prices p_t are constructed as the average of the hourly prices of each day. The corresponding returns were calculated as $(p_t - p_{t-1})/p_{t-1}$ and multiplied by 10 for scaling purposes.