Blasques, Francisco F.; Gorgi, Paolo; Koopman, Siem Jan S.J.

Working Paper

Accelerating GARCH and Score-Driven Models: Optimality, Estimation and Forecasting

Tinbergen Institute Discussion Paper, No. 17-059/III

Provided in Cooperation with:
Tinbergen Institute, Amsterdam and Rotterdam

Suggested Citation: Blasques, Francisco F.; Gorgi, Paolo; Koopman, Siem Jan S.J. (2017) : Accelerating GARCH and Score-Driven Models: Optimality, Estimation and Forecasting, Tinbergen Institute Discussion Paper, No. 17-059/III, Tinbergen Institute, Amsterdam and Rotterdam

This Version is available at:
https://hdl.handle.net/10419/177627

TI 2017-059/III
Tinbergen Institute Discussion Paper

Accelerating GARCH and Score-Driven Models: Optimality, Estimation and Forecasting

Francisco (F.) Blasques[1]
Paolo Gorgi[2]
Siem Jan (S.J.) Koopman[3]

1: VU Amsterdam, The Netherlands; Tinbergen Institute, The Netherlands
2: VU Amsterdam, The Netherlands
3: VU Amsterdam, The Netherlands; CREATES, Aarhus University, Denmark

Accelerating GARCH and Score-Driven Models: Optimality, Estimation and Forecasting *

F. Blasques[a,b], P. Gorgi[a], and S. J. Koopman[a,b,c]

[a]Vrije Universiteit Amsterdam, The Netherlands

[b]Tinbergen Institute, The Netherlands

[c]CREATES, Aarhus University, Denmark

July 5, 2017

*Corresponding author: P. Gorgi. Email address: p.gorgi@vu.nl

Accelerating GARCH and Score-Driven Models: Optimality, Estimation and Forecasting

by F. Blasques, P. Gorgi and S.J. Koopman

Abstract

We first consider an extension of the generalized autoregressive conditional heteroskedasticity (GARCH) model that allows for a more flexible weighting of financial squared-returns for the filtering of volatility. The parameter for the squared-return in the GARCH model is time-varying with an updating function similar to GARCH but with the squared-return replaced by the product of the volatility innovation and its lagged value. This local estimate of the first order autocorrelation of volatility innovations acts as an indicator of the importance of the squared-return for volatility updating. When recent volatility innovations have the same sign (positive autocorrelation), the current volatility estimate needs to adjust more quickly than in a period where recent volatility innovations have mixed signs (negative autocorrelation). The empirical relevance of the accelerated GARCH updating is illustrated by forecasting daily volatility in return series of all individual stocks present in the Standard & Poor's 500 index. Major improvements are reported for those stock return series that exhibit high kurtosis. The local adjustment in weighting new observational information is generalised to score-driven time-varying parameter models of which GARCH is a special case. It is within this general framework that we provide the theoretical foundations of accelerated updating. We show that acceleration in updating is more optimal in terms of reducing Kullback-Leibler divergence and in comparison to fixed updating. The robustness of our proposed extension is highlighted in a simulation study within a misspecified modelling framework. The score-driven acceleration is also empirically illustrated with the forecasting of US inflation using a model with time-varying mean and variance; we report significant improvements in the forecasting accuracy at a yearly horizon.

Key words: GARCH models, Kullback-Leibler divergence, score-driven models, S&P 500 stocks, time-varying parameters, US inflation.

1 Introduction

Economic and financial time series often exhibit intricate dynamic features. When the time series is analysed by use of a parametric dynamic model, it needs to be sufficiently flexible to describe its salient features. Oftentimes, time-varying parameter models provide the necessary flexibility. However, for such dynamic models, the estimation and forecasting can become subject to particular challenges. A possible challenge for parameter estimation (or filtering) is to account for the varying amount of information that is contained in past observations. For example, in the case of analysing time series of daily financial returns and filtering its time-varying volatility by means of the well-known generalized autoregressive conditional heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986), the parameters can be tuned in such a way that the conditional volatility change slowly over time. But after specific events such as a financial crisis or a major news event, one may need to change the updating process such that the conditional volatility adapts to its new level quickly rather than slowly. We then need to change the values of the parameters temporarily to accommodate such changes to a new volatility level. For such and other purposes, we introduce a dynamic specification in order to update the time-varying parameter (in the example, conditional volatility) quickly when the data is informative and slowly when the data is less informative. The introduction of time-varying coefficients in the GARCH model has been considered elsewhere but for different purposes and motivations. For example, Engle and Lee (1999) have introduced a time-varying intercept in the GARCH model with the motivation to introduce a long-run, slowly evolving component in conditional volatility. More recently, Quaedvlieg et al. (2016) introduce time-varying parameters in realized volatility models by having them as functions of precision measures that are computed from high-frequency data. In our case, we do not require additional or external information. In the context of GARCH models, we only require other sample statistics from the daily financial returns as we argue below.

Our proposed extension is first considered for the GARCH model with the purpose of having more flexible weighting functions of past financial squared-returns for the filtering and forecasting of volatility. The weighting parameter of the squared-return in the GARCH model is made time-varying with an updating function that is similar to GARCH itself but with the squared-return replaced by the product of the volatility innovation and its lagged value. This local estimate of the first-order autocorrelation of volatility innovations provides an indication of the importance of the

squared-return for volatility updating. When recent volatility innovations have the same sign, the adjustment of the current volatility estimate needs to accelerate faster than in a period where these innovations have mixed signs. The former hints towards a positive first-order autocorrelation in innovations while the latter hints towards a negative autocorrelation. To let the weighting parameter be a function of the local estimate of the correlation, the updating can be accelerated when a set of consecutive innovations have the same sign. In this case, the adjustment to a new volatility level will rapidly materialise. The empirical relevance of this accelerated GARCH (aGARCH) model is investigated for all stocks present in the Standard & Poor's 500 index. We present in-sample and out-of-sample performance measures. Large improvements are reported for stock return series with high kurtosis.

This mechanism of accelerated updating can also be introduced to other observation-driven time-varying parameter models. For example, the class of score-driven time-varying parameter models of Creal et al. (2013) encompasses many well-known dynamic models, including GARCH and related models but also facilitates the formulation of new dynamic models. Recent examples of score-driven models are provided by Harvey and Luati (2014) and Andres (2014) where they consider location and scale models for fat-tailed distributions, Creal et al. (2014) where dynamic factor models are explored, and Creal et al. (2011), Oh and Patton (2017) and De Lira Salvatierra and Patton (2015) who adopt different dynamic copula models with time-varying coefficients. A collection of all recent developments on score-driven models, also known as generalised autoregressive score (GAS) models of Creal et al. (2013) or dynamic conditional score (DCS) models of Harvey (2013), is provided online at `http://gasmodel.com`. The extension of having a dynamic parameter in the GARCH model is an illustration of the flexible framework that score-driven models can provide. Within the updating of the parameter of interest, the score function provides a sensible and optimal formulation of how the actual updating can be accelerated.

In the general setting, we propose a generalisation of the class of score-driven or GAS models: the accelerating GAS (aGAS) models. We will discuss the intuition behind this specification and provide a theoretical justification for our proposed method. In particular, we follow Blasques et al. (2015) and show that acceleration in updating is more optimal in terms of reducing Kullback-Leibler divergence when compared to fixed updating. Furthermore, we present a simulation study to illustrate the role that our approach can play, the provision of more flexible models and the improved approximation of an unknown data generating process. Finally, in the context of location

and scale models, we consider an empirical application for the modelling and forecasting of the quarterly time series of US CPI inflation. Our proposed model is based on a fat-tailed density with time-varying conditional mean and volatility. The accelerating updating equation renders our aGAS model capable of jointly describing the fast changes in the inflation level during the 1970's and 1980's, but also, the smooth and slow dynamic behaviour of the conditional mean during the great moderation of two decades that followed the early 1980s.

The paper is structured as follows. Section 2 introduces the aGARCH model. Section 3 presents the general aGAS framework. Section 4 develops the optimal properties for the aGAS models. Section 5 discuses the results of a simulation experiment. Section 6 gives evidence of the empirical relevance of aGAS models for the S&P500 stock returns and for the US inflation series. Section 7 concludes.

2 Accelerated GARCH model

In this section we motivate our extension for observation driven time-varying parameter models by introducing the accelerated updating mechanism for the GARCH model. A natural updating function of current and past squared-returns is proposed and discussed. The empirical relevance of our extension for the GARCH model is investigated in a volatility forecasting study concerning 460 stock return series of U.S. companies from the S&P500 stock index.

2.1 Model formulation

The generalized autoregressive conditional heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986) treats the clustering of large, but also small, shocks in time series of *financial returns* $\{y_t\}_{t\in\mathbb{Z}}$, with time index t. The variable y_t typically represents daily differences of logged closure prices of stocks traded at financial markets. A time series of financial returns can also be based on stock indices, exchange rates, commodity prices and related variables. The basic GARCH model is given by

$$y_t = \sqrt{h_t}\varepsilon_t, \qquad h_{t+1} = \omega + \bar{\alpha} y_t^2 + \bar{\beta} h_t, \tag{1}$$

where the *volatility* $\{h_t\}_{t\in\mathbb{Z}}$ is the time-varying scaling for y_t and the locally scaled return $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is assumed to be an independent identically distributed (i.i.d) sequence of random variables with

zero mean and unity variance. The coefficients $\omega > 0$, $\bar{\alpha} > 0$ and $\bar{\beta} > 0$ are treated as fixed unknown parameters that need to be estimated; they determine the overall variance, the amount of changes in volatility, the persistence of volatility, and other features of y_t. The restriction bounds on the coefficients ensure $h_t > 0$ while $\bar{\alpha} + \bar{\beta} < 1$ ensures a weak stationary process for h_t: see the discussion in Nelson (1990). We focus in particular on the coefficient $\bar{\alpha}$ that determines the level of changes in the volatility h_t; the coefficient determines how fast the volatility responds to changes in the amount of clustering in the time series of returns y_t. For given values of ω and $\bar{\beta}$, we may question whether the constancy of $\bar{\alpha}$ is appropriate when we need to determine locally how quickly the volatility h_t must adapt to changes in the amount of clustering, especially when we have a longer time series. A relatively small value for $\bar{\alpha}$ can be appropriate when the current level of volatility is appropriate. But in a more turbulent period, the $\bar{\alpha}$ may need to be larger so that h_t can adjust faster to new information. To address this empirical feature in financial time series, we present an extension of the GARCH model in which $\bar{\alpha}$ is allowed to vary over time. We propose a specification for the time-varying $\bar{\alpha}$, we investigate the consequences for the volatility h_t and we study the statistical properties of the new model. We refer to this extended GARCH model as the accelerated GARCH model, or the aGARCH model.

Before we introduce a time-varying coefficient for $\bar{\alpha}$, we express the GARCH updating in its innovation form, that is

$$\begin{aligned} h_{t+1} &= \omega + \bar{\alpha}(y_t^2 - h_t) + \beta\, h_t \\ &= \omega + \beta\, h_t + \alpha\, h_t\,(\varepsilon_t^2 - 1), \end{aligned} \tag{2}$$

where we have replaced y_t^2 by $h_t\varepsilon_t^2$ as implied by the model for y_t and with $\beta = \bar{\alpha} + \bar{\beta}$ and $\alpha \equiv \bar{\alpha}$. We refer to prediction error $y_t^2 - h_t$ as the *volatility innovation* and the term $\varepsilon_t^2 - 1$ as *scaled volatility innovation* (SVI) since $\varepsilon_t^2 - 1 = h_t^{-1}(y_t^2 - h_t)$. Assuming normality for ε_t, the SVI has some convenient properties as it is a χ^2 distributed variable with mean equal to zero and variance equal to two. Its properties do not relate to h_t or its updating function which is useful for the development below. In this specification we require $\alpha < \beta$ to ensure that $h_t > 0$ for all t. We therefore specify the time-varying coefficient through the link function

$$\alpha_t = \beta \cdot \text{logit}(f_{t+1}), \tag{3}$$

where variable $\{f_t\}_{t\in\mathbb{Z}}$ is a time-varying scalar coefficient and logit$(\cdot)$ is the logistic function such that logit$(a) = \exp(a)\,/\,(1+\exp(a))$ for any $a \in \mathbb{R}$. Given this link function, $f_t \in \mathbb{R}$ can be any unbounded process. Other such link functions can also be considered. Since we let f_t be a nonlinear function of past observations $y_1, \ldots, y_{t-1}$, in a similar way as h_t, we have the update $h_{t+1} = \omega + \beta\, h_t + \alpha_t\,(y_t^2 - h_t)$ for which y_t is available and hence α_t can be a function of f_{t+1} that partly relies on y_t.

For the time-varying process f_t, we consider a similar updating scheme as for GARCH but with the product of the current value of SVI, on time t, and its previous value, on time $t-1$, as the innovation term. More specifically, we propose the updating equation

$$f_{t+1} = \omega_f + \beta_f f_t + \alpha_f(\varepsilon_t^2 - 1)(\varepsilon_{t-1}^2 - 1), \tag{4}$$

where the coefficients ω_f, α_f and β_f have similar roles as ω, $\bar{\alpha}$ and $\bar{\beta}$, respectively, in the GARCH model presented in equation (1). The equation (2) with a time-varying α given by (3) and (4) is the accelerated GARCH model. The product of the contemporaneous and lagged SVI is treated as indicative of whether or not α_t needs to change more quickly or slowly. When two consecutive values of SVI have the same sign, it may indicate that the level of volatility is either too low or too high and that the model needs to adapt to this change more quickly. Hence a larger value for α_t is necessary. The resulting model is a straightforward extension of the GARCH model with the additional updating equation (4) and the addition of two coefficients only, α_f and β_f, since ω_f is effectively replacing the static $\bar{\alpha}$ coefficient in the GARCH model (1). Next we discuss some further details of our aGARCH model.

2.2 Discussion of the aGARCH model

We have argued that the coefficient $\bar{\alpha}$ in the GARCH model (1) is of key importance to determine how much information in the most recent squared returns, that is y_t^2, must be provided to h_{t+1}. The time-varying α_t facilitates the possibility that for some time periods the squared returns may be more informative than in other periods. For instance, the necessity for a time-varying α_t can be due to a break in the level of the variance. Before the break, the variance may be changing slowly and therefore the magnitude of the innovations should be small. After the break, however, the new observations are informative about the new variance level and thus the parameter α_t should

increase in order to give more prominence to the information in y_t^2. To illustrate the motivation for the aGARCH model further and the role of a time-varying α_t, we consider the estimation of the true variance that we observe as the squared return y_t^2 which is subject to error. The time series of squared returns are filtered by means of GARCH updating (1), with a small and a large value $\bar{\alpha}$, and with $\bar{\alpha}$ replaced by a time-varying α_t based on equations (3) – (4). Figure 1 illustrates the effects on the filtered variance paths for the three different cases. For a fixed $\bar{\alpha}$ we observe the trade-off between the fact of being exposed to the disturbance component and the need to update quickly when the level of the variance has changed. We can compare the volatility paths for a small and a large value of $\bar{\alpha}$ in Figure 1. The small-$\bar{\alpha}$ path is far too slow in adapting to the new variance level after the break while the large-$\bar{\alpha}$ path is too volatile in the pre- and post-break periods. The advantage of our aGARCH time-varying α_t volatility path: it adapts quickly to the variance level after the break and is robust against the disturbance component in periods when the true variance is constant.

A further convenient property of the aGARCH updating scheme becomes apparent when α_t approaches or gets close to the value of β. In this case the aGARCH model mimics the first-order autoregressive conditional heteroskedasticity (ARCH) model of Engle (1982). It implies that the filtered variance depends only on the most recent observations. Whereas, when α_t is close to zero, the impact of the most recent squared return y_t^2 is lower since the effect is averaged with the contributions of the earlier (lagged) squared returns. As a result, a large α_t after the break leads to a shorter memory of the filtered variance. This effect is highly intuitive since these recent observations are very informative about the new variance level while the filtered variance h_t, constructed from the squared returns before the break, is not very informative.

The updating mechanism of the time-varying α_t in equation (4) has an intuitive interpretation. In particular, it is driven by products of SVIs. Therefore, α_t increases when past innovations are positively correlated, decreases when the correlation is negative and remains constant when the correlation is zero. A positive correlation indicates that for repeated observations the SVI tend to be either above or below its expectation. This is indeed an indication that the variance should be updated more quickly. In the same way, a negative correlation indicates that consecutive SVIs tend to have opposite signs. This may indicate that the variance is being updated too quickly as the disturbance component affect the path of the filtered variance and hence SVIs are more likely to have opposite sign. Finally, a correlation equal to zero suggests a situation of equilibrium where

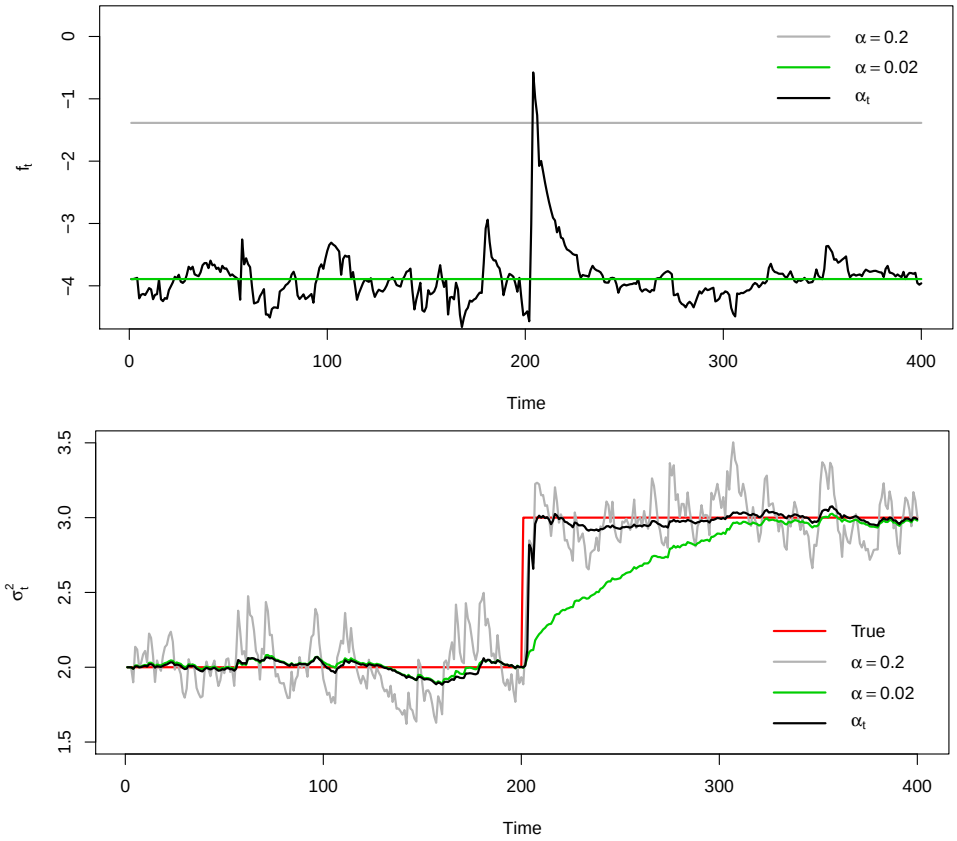

Figure 1: *Filtered estimates of time-varying parameter f_t and volatility h_t from the aGARCH model. In the upper graph, the filtered estimate of f_t is presented with reference to fixed values of f_t corresponding to $\bar{\alpha} = 0.2$ and $\bar{\alpha} = 0.02$. In the lower graph, the true volatility path is presented together with filtered estimates of volatility h_t from the aGARCH model with time-varying $\alpha_t = logit(f_t)$ and from GARCH models with $\bar{\alpha} = 0.2$ and $\bar{\alpha} = 0.02$.*

the variance is being updated in the right way. In Section 4, we also show for a more general case that the updating mechanism considered for α_t has an optimality property.

Time-varying coefficients in the GARCH model have also been considered by Engle and Lee (1999) who introduce a time-varying parameter for ω with the motivation to introduce a long-run component in conditional volatility. Engle and Lee show that the GARCH model with a time-varying ω can be formulated as a higher-order GARCH model, with two lags for both y_t^2 and h_t. In our case, the aGARCH model does not have a higher-order GARCH representation because the variance recursion becomes a nonlinear function of past y_t^2 when α_t is time-varying. The features of the two different extensions will be explored next in a volatility forecasting study for a large collection of time series of daily financial returns.

2.3 Illustration: volatility forecasting for all series in S&P500

We evaluate the performance of the aGARCH model through a comparison with other GARCH models using the stocks that are currently in the S&P500 index. Daily stock returns from 2008 to 2015 are considered. The series of the S&P500 that are not available since 2008 are excluded from the study. The resulting number of time series is 460. The performances of the GARCH models are evaluated both in-sample and out-of-sample. The in-sample evaluation is based on fit and the Akaike information criterion (AIC). We have opted for the AIC statistic because GARCH models can be viewed as filters in a misspecified modeling framework for which the AIC provides a meaningful interpretation. The out-of-sample evaluation is based on the log-score criterion as given by $n^{-1}\sum_{i=1}^{n}\log p_{T+i}(y_{T+i})$, where T is the in-sample time series length, n is the out-of-sample length and $p_t(\cdot)$ is the conditional density of y_t given the past observations up to $t-1$. This criterion is widely known and is regularly used in the context of evaluating density forecasts; see, for example, Geweke and Amisano (2011). The in-sample evaluation is based on the whole sample, whereas, the out-of-sample period consists of all daily observations in 2015. For the out-of-sample evaluation, the training sample used for estimation is from 2008 to 2014. The static parameters are estimated only once; we are not using expanding or rolling sample windows for estimation.

	Full Dataset				Top 10% Kurtosis			
	In-sample		Out-of-sample		In-sample		Out-of-sample	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
GARCH	59	12.8%	66	14.3%	0	0.0%	10	21.7%
ELGARCH	183	39.8%	156	33.9%	4	8.7%	11	23.9%
aGARCH	84	18.3%	96	20.9%	27	58.7%	17	37.0%
aELGARCH	134	29.1%	142	30.9%	15	32.6%	8	17.4%
Total	460	100.0%	460	100.0%	46	100.0%	46	100.0%

Table 1: *The number and percentage of series in the S&P 500 index where each model outperforms the others. The in-sample performance is based on the Akaike information criterion. The out-of-sample performance is based on a log-score criterion. In the first panel, all 460 stocks in the S&P500 index are considered. In the second panel, the 46 stocks with the highest in-sample kurtosis are considered.*

In our illustration we consider standard GARCH models with time-varying parameter extensions: the Engle and Lee (1999) with a time-varying ω, denoted by ELGARCH, our aGARCH

model with the time-varying α_t and the accelerating ELGARCH model, denoted by aELGARCH, which includes both time varying α_t and ω. Table 1 reports the number of series in the S&P500 index where a model outperforms the others. When considering all stocks, the in-sample fit of the aGARCH model produces the smallest AIC for 18.3% of the series, whereas the aELGARCH in-sample fit produces the smallest AIC in 29.1% of the cases. The out-of-sample performance, as measured by the smallest log-score value in 2015, is similar, although aGARCH and aELGARCH have a slightly larger number of stocks in which they perform best. We can say that the aGARCH together with the aELGARCH model have the best in-sample and out-of-sample performance for about 50% of the series. Furthermore, the aGARCH and aELGARCH models appear to perform particularly well for stock returns that have a high kurtosis (or have heavy-tailed densities). For example, when we consider the 10% of stocks with the highest in-sample kurtosis for their GARCH residuals, the aGARCH and aELGARCH models perform best for more than 90% of the series. This interesting finding is further highlighted in Figure 2. Figure 2 presents the percentage of stocks for which a particular model performs best for a finer sub-selection of stocks with the highest in-sample kurtosis. We can clearly conclude that the performance of our aGARCH model increases as we select the series with fatter tailed return densities. We may conclude that in general the aGARCH and aELGARCH models outperform the GARCH and ELGARCH models when the distribution of the daily return series has fat tails.

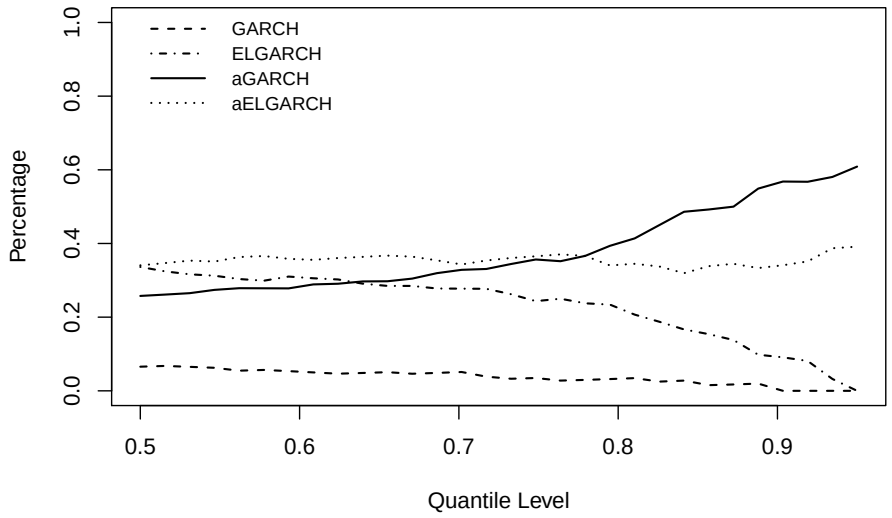

Figure 2: *Percentage of series where each model outperform the others in terms of AIC. The percentage is computed only for those series with their GARCH residual kurtosis above a certain quantile. The quantile levels are indicated on the horizontal axis.*

3 Accelerated score-driven time series models

In the previous section we have introduced a time-varying coefficient in the GARCH model and have discussed the merits of this extension. In this section we introduce a similar time-varying coefficient for the general class of score-driven time series models of Creal et al. (2013) and Harvey (2013). We refer to this class of models as generalised autoregressive score (GAS) models. In the introductory Section 1 we have given a short review of GAS models with further references. We introduce the accelerated GAS updating equation in a similar way as it is done for the aGARCH model. The aGARCH model is a special case of the accelerated GAS (aGAS) model.

For a time series variable $\{y_t\}_{t\in\mathbb{Z}}$, the GAS model is given by

$$y_t \sim p(y_t|\lambda_t;\theta), \qquad \lambda_{t+1} = \omega_\lambda + \beta_\lambda \lambda_t + \alpha_\lambda s_{\lambda,t}, \tag{5}$$

where $p(\cdot|\lambda_t;\theta)$ is a parametric conditional density with λ_t as the time-varying parameter of interest and θ as an unknown vector containing all static parameters in the model, including ω_λ, β_λ, and α_λ, and $s_{\lambda,t}$ is an innovation term. The time-varying parameter evolves as an autoregressive process of order 1 with intercept ω_λ, autoregressive coefficient β_λ and scale parameter α. The distinguishing feature of a GAS model is the choice of the innovation $s_{\lambda,t}$ as the local score or gradient of density $p(y_t|\lambda_t;\theta)$ with respect to λ_t. We specify the scaled innovation by

$$s_{\lambda,t} = S_{\lambda,t}\, u_{\lambda,t}, \qquad u_{\lambda,t} = \frac{\partial \log p(y_t|\lambda_t;\theta)}{\partial \lambda_t},$$

where $S_{\lambda,t}$ is a strictly positive scaling factor and $u_{\lambda,t}$ is the innovation term defined as the first derivative of the conditional density contribution for a single observation at time t. Many standard models can be derived from this framework as is shown by Creal et al. (2013). For example, in case of having $p(y_t|\lambda_t;\theta)$ as the normal distribution with time-varying mean λ_t and some static variance, we obtain the autoregressive moving average model, of order $(1,1)$. When we switch mean and variance, that is, we have some static mean and time-varying variance λ_t, we obtain the GARCH$(1,1)$ model as considered in Section 2. This flexible framework can be used in a general and flexible manner for the introduction of a time-varying parameter in a model. Various theoretical properties of GAS models are studied. For example, Blasques et al. (2015) show that score-driven updating is optimal in terms of reducing Kullback-Leibler divergence.

The accelerated GAS (aGAS) model is defined as the GAS model (5) with a time-varying α_λ coefficient that we specify as

$$\alpha_{\lambda,t} = g(f_{t+1}; \theta), \qquad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{f,t}, \tag{6}$$

where $g(\cdot)$ is a strictly increasing link function and the time-varying variable f_{t+1} determines the time-variation of $\alpha_{\lambda,t}$, for all time indices t, and it evolves according to an autoregressive process of order 1 with innovation term $s_{f,t}$, intercept ω_f, autoregressive coefficient β_f, and scale parameter α_f. The time-varying $\alpha_{\lambda,t}$ is subject to a link function but the overall framework is similar to the GAS model itself. The scaled innovation term relies on the first derivative of the conditional density contribution at time t, that is

$$s_{f,t} = S_{f,t}\, u_{f,t}, \qquad u_{f,t} = \frac{\partial \log p(y_t|\lambda_t; \theta)}{\partial f_t}\,, \tag{7}$$

where $S_{f,t}$ is a strictly positive scaling factor and $u_{f,t}$ is the innovation term defined as the first derivative of the conditional density contribution for a single observation at time t. The time index t for the time-varying parameters λ_t and f_t indicates that the parameters are functions of past observations up to time $t-1$, that is $\{y_{t-1}, y_{t-2}, \dots\}$. It is straightforward to show that the innovation $s_{f,t}$ in (7) can be expressed as

$$s_{f,t} = C_{f,t} u_{\lambda,t} u_{\lambda,t-1}, \tag{8}$$

where $C_{f,t}$ is a positive scaling factor and is a function of the scaling factors $S_{\lambda,t}$ and $S_{f,t}$. The expression (8) for $s_{f,t}$ is highly convenient as it is expressed directly in terms of $u_{\lambda,t}$ and hence there is no need to derive and compute other derivatives. Perhaps even more importantly, the expression (8) shows that the score-driven update is a local estimate of the first-order autocorrelation of the innovation term of the time-varying parameter of interest λ_t. The innovation term $u_{f,t}$ of the dynamic f_t is driven by the standardized product of current and past score innovations. The parameter α_t increases when there is positive autocorrelation in past score innovations. The same intuitive interpretation for the aGARCH model applies. Positive correlation means that past score innovations tend to have the same sign. Therefore, it is natural to think that the step size α_t should increase as this is an indication that the parameter α_t is being updated too slowly.

The use of scaling factors $S_{\lambda,t}$ and $S_{f,t}$ for the score innovation terms is standard practice in analyses based on GAS models. The choice typically depends on the model at hand. Creal et al. (2013) propose the use of the Fisher information $\mathcal{I}_t$ to account for the curvature of the score. For example, we can consider the inverse of the Fisher Information, the square root of the Fisher Information inverse or simply the identity matrix as scaling factors. The use of the inverse of the square root of $\mathcal{I}_t$ as the scaling factor implies that the conditional variance of the score innovation equals the unity matrix. It has the convenient effect that the variability of the innovation of the autoregressive process in (5) is determined solely by α_t.

4 Optimality properties

We next provide a theoretical justification for the aGAS specification in (5) and (6). Blasques et al. (2015) have developed a framework from which optimality features for the GAS updating can be derived. We build on these developments and show that the use of the score-based innovation in (8) for α_t has an optimality justification. Furthermore, we show that, under certain conditions, the updating mechanism of the aGAS model outperforms standard GAS updating in terms of its local Kullback-Leibler (KL) divergence reduction. The results are based on a misspecified model setting where the objective is to consider the dynamic specification that allows to minimize the KL divergence between a postulated conditional distribution and the unknown true distribution of the DGP. The Section is structured as follows: Section 4.1 introduces the framework considered, Section 4.2 delivers the optimality of the score update for α_t and Section 4.3 shows how flexible GAS models can outperform classic GAS models.

4.1 A general updating mechanism

Assume that the sequence of observed data $\{y_t\}_{t=1}^T$ with values in $\mathcal{Y} \subseteq \mathbb{R}$ is generated by an unknown stochastic process that satisfies

$$y_t \sim p_t^o(y_t), \quad t \in \mathbb{N},$$

where p_t^o is the true unknown conditional density. We consider a conditional density for the observations as in (5), $y_t \sim p(y_t|\lambda_t;\theta)$, where $\theta \in \Theta$ is a static parameter and λ_t is a time-varying

parameter that takes values in $\Lambda \subseteq \mathbb{R}$. Note that also the model density $p(\cdot|\lambda_t;\theta)$ is allowed to be misspecified and there may not exist a true λ_t^o and θ_0 such that $p_t^o = p(\cdot|\lambda_t^o;\theta_0)$.

The objective is to specify the dynamics of the time-varying parameter λ_t in such a way that the conditional density $p(\cdot|\lambda_t;\theta)$ implied by the model is as close as possible to the true conditional density p_t^o. To evaluate the distance between these two conditional densities, a classical approach is to consider the Kullback-Leibler (KL) divergence introduced in Kullback and Leibler (1951) as a measure of divergence, or distance, between probability distributions. The KL divergence plays an important role in information theoretic settings Jaynes (1957, 2003) as well as in the world of statistics (Kullback, 1959; Akaike, 1973). The importance of the KL divergence in econometric applications is reviewed in Maasoumi (1986) and Ullah (1996, 2002).

The ideal specification of λ_t minimizes the KL divergence between the true conditional density p_t^o and the model-implied conditional density $p(\cdot|\lambda_t;\theta)$. In other words, a sequence $\{\lambda_t\}_{t\in\mathbb{N}}$ is *optimal* if for each $t \in \mathbb{N}$, the value of λ_t minimizes the following KL divergence

$$\mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_t;\theta)\right) = \int_Y p_t^o(y) \log \frac{p_t^o(y)}{p(y|\lambda_t;\theta)} dy, \tag{9}$$

where Y denotes the set over which the *local* KL divergence is evaluated; see Hjort and Jones (1996), Ullah (2002) and Blasques et al. (2015) for applications of the local KL divergence. Assuming that $\{\lambda_t^*\}_{t\in\mathbb{N}}$ is an optimal sequence that minimizes the KL divergence for any $t \in \mathbb{N}$, we would like our model to deliver a filtered time-varying parameter $\{\lambda_t\}_{t\in\mathbb{N}}$ that approximates arbitrarily well the trajectory of $\{\lambda_t^*\}_{t\in\mathbb{N}}$.

Of course, from the outset, there is no reason to suppose that the score driven recursion

$$\lambda_{t+1} = \omega_\lambda + \beta_\lambda \lambda_t + \alpha_\lambda s_{\lambda,t}$$

would ever deliver such a result. Lemma 1, reminds us however of the simple fact that, a time-varying updating of the type

$$\lambda_t(f_t) = \omega_\lambda + \beta_\lambda \lambda_{t-1} + g(f_t) s_{\lambda,t-1},$$

could deliver a better approximation to $\{\lambda_t^*\}_{t\in\mathbb{N}}$.

Lemma 1. *If an optimal sequence $\{\lambda_t^*\}_{t\in\mathbb{N}}$ exists, then for any given initialization, $\lambda_0 \in \Lambda$ there*

exists a sequence $\{f_t\}_{t\in\mathbb{N}}$ *of points such that* $\lambda_t(f_t) = \lambda_t^* \ \forall t \in \mathbb{N}$. *Moreover,* f_t *is almost surely constant if and only if there is some* $c \in \mathbb{R}$ *such that* $s_{\lambda,t} = (\lambda_{t+1}^* - \omega_\lambda - \beta_\lambda \lambda_t)/g(c)$ *almost surely for every* $t \in \mathbb{N}$.

In practice, the problem is however how to specify the dynamics of f_t. Below, we will address the issue by providing a theoretical justification for the score-based update for f_t.

4.2 Optimality of score innovations

We build on the work of Blasques et al. (2015) that provides optimality arguments for a score-based updating equation. Specifically, Blasques et al. (2015) shows that considering an updating scheme of the form

$$\lambda_{t+1} = \lambda_t + \alpha_\lambda s_{\lambda,t}$$

reduces locally the KL divergence between the model density and the true probability density, in particular, they show that the variation in the KL divergence obtained by updating the time-varying parameter from λ_t to λ_{t+1} satisfies

$$\mathrm{KL}_Y\big(p_t^o, p(\cdot|\lambda_{t+1};\theta)\big) - \mathrm{KL}_Y\big(p_t^o, p(\cdot|\lambda_t;\theta)\big) < 0,$$

when the update is local $\lambda_t \approx \lambda_{t+1}$ and the set Y is a neighborhood of y_t. This result is subject to the fact that the parameter α_λ has to be positive because otherwise the information provided by the score is distorted. Clearly, as this optimality concept regards only the direction of the update we can conclude that the optimality holds also when α_λ is time-varying as long as it is positive. This justifies the use of a positive link function g in (6), which ensures the positivity of $g(f_t)$.

It is also worth mentioning that the optimality concept in Blasques et al. (2015) is shown to hold for $(\omega_\lambda, \beta_\lambda) \approx (0, 1)$. This because the reduction of local KL divergence from the update is considered with respect to p_t^o. In practice, what we really want is to reduce the KL divergence with respect to p_{t+1}^o as the updated time-varying parameter λ_{t+1} is used to specify the conditional probability measure of y_{t+1}. The problem is that λ_t is updated using information from p_t^o and therefore, without imposing any restriction on the true sequence of conditional densities, it is impossible to say if the updating scheme makes any sense with respect to p_{t+1}^o. Blasques et al. (2015) show that having $(\omega_\lambda, \beta_\lambda) \approx (0, 1)$ is optimal also with respect to the density p_{t+1}^o only if

the true conditional density varies sufficiently smoothly over time. This justify the possibility that in practice it may be reasonable to consider also $(\omega_\lambda, \beta_\lambda) \neq (0, 1)$.

We now add to the results of Blasques et al. (2015) by considering the more flexible updating scheme in (6) for the time-varying parameter f_t and showing that it has a similar optimality justification. More specifically, we provide an optimality reasoning for the updating scheme in (6) setting $(\omega_f, \beta_f) \approx (0, 1)$,

$$f_{t+1} = f_t + \alpha_f s_{f,t}. \tag{10}$$

At time $t-1$, the parameter f_t is used to update λ_{t-1} by the recursion in (5), namely

$$\lambda_t(f_t) = \omega_\lambda + \beta_\lambda \lambda_{t-1} + g(f_t) s_{\lambda,t-1},$$

then, at time t we observe y_t and the parameter f_t is updated to f_{t+1}. We consider optimal an updating mechanism that process properly the information provided by y_t. The idea is that f_t has to be updated in such a way that the model density with the updated f_t is closer to the true density p_t^o than the model density $p(\cdot|\lambda_t(f_t); \theta)$. We consider the following definition

Definition 1. *The realized KL variation for the parameter update from f_t to f_{t+1} is*

$$\Delta_{f,t}^{t+1} = KL_Y\big(p_t^o, p(\cdot|\lambda_t(f_{t+1}); \theta)\big) - KL_Y\big(p_t^o, p(\cdot|\lambda_t(f_t); \theta)\big).$$

A parameter update for f_t is said to be optimal in local realized KL divergence if and only if $\Delta_{f,t} < 0$ for any $(f_t, \theta) \in \mathcal{F} \times \Theta$ and almost every $y_t \in \mathcal{Y}$.

The results we present are local in the sense that we will show that at each step the score update gives the right direction to reduce a local realized KL divergence. As in Blasques et al. (2015), we focus on sets of the form

$$Y = B(y_t, \epsilon_y) = \{y \in \mathcal{Y} : |y_t - y| < \epsilon_y\}$$
$$F = B(f_t, \epsilon_f) = \{f_{t+1} \in \mathbb{R} : |f_t - f_{t+1}| < \epsilon_f\}.$$

We set some regularity assumptions on the score $s_{\lambda,t}$. In particular, the score is nonzero with probability 1 to ensure that the parameter f_t is always updated, and it also have some differentiability properties.

Assumption 1. *The score $u_{\lambda,t} = u_\lambda(y_t, \lambda_t, \theta)$ is continuously differentiable in y_t and λ_t, and almost surely $u_\lambda(y_t, \lambda_t, \theta) \neq 0$ for any $(\lambda_t, \theta) \in \Lambda \times \Theta$ and $t \in \mathbb{N}$.*

The next proposition states that the score update for f_t is optimal in the sense of Definition 1

Proposition 1. *Let Assumption 1 hold, then the update from f_t to f_{t+1} in (10) is optimal in local realized KL divergence as long as α_f is positive.*

The next proposition stress the fact that only the score $u_{f,t}$ provides the right direction to update f_t

Proposition 2. *Let Assumption 1 hold, then any parameter update from f_t to f_{t+1} is optimal in local realized KL divergence if and only if sign$(f_{t+1} - f_t)$ = sign$(s_{f,t})$ almost surely for any $f_t \in \mathcal{F}$.*

4.3 Relative optimality

The optimality concept developed in the previous section is only related to the update of f_t, but, in practice, the update of f_t is only a tool to improve the update of $\lambda_t(f_t)$. The idea is to compare the score update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ with the score update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$. As before, the quality of the updates is measured in terms of KL reduction. We are thus interested in comparing the variation in KL divergence obtained by updating the parameter from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$,

$$\Delta^{t+1}_{\lambda,t+1} = \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_{t+1}(f_{t+1}); \theta)\big) - \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_t(f_t); \theta)\big),$$

against the variation in KL divergence obtained under the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$

$$\Delta^{t}_{\lambda,t+1} = \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_{t+1}(f_t); \theta)\big) - \mathrm{KL}_Y\big(p^o_t, p(\cdot|\lambda_t(f_t); \theta)\big).$$

Clearly, the first type of update is better if it can ensure a greater reduction in KL divergence.

Definition 2. *The parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ is said to dominate the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ in local realized KL divergence, if and only if*

$$\Delta^{t+1}_{\lambda,t+1} - \Delta^{t}_{\lambda,t+1} < 0.$$

The notion of dominance in local realized KL divergence in Definition 2 provides a line of comparison for the parameter updates. We can say that the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ outperforms the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ if $\Delta^{t+1}_{\lambda,t+1} < \Delta^{t}_{\lambda,t+1}$. The result we obtain are local in the sense that the KL divergence is evaluated locally and the innovations $s_{\lambda,t-1}$ and $s_{\lambda,t}$ are in a neighborhood of zero. Moreover, we also impose that the observation y_t lies in a neighborhood of y_{t-1}. More formally, the realized KL divergence in Definition 1 is evaluated is a sets of the form

$$Y = B(y_t, \epsilon_y) = \{y \in \mathcal{Y} : |y_t - y| < \epsilon_y\},$$

with $y_t \in B(y_{t-1}, \epsilon_y)$ and $s_{\lambda,t-1}, s_{\lambda,t} \in B(0, \epsilon_\lambda)$. The result is stated in the following proposition

Proposition 3. *Let Assumption 1 hold. Then, the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ generated by (10) dominates the the parameter update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ in local realized KL reduction for every $\lambda_{t-1} \in \Lambda$ and $f_t \in \mathbb{R}$.*

The result in Proposition 3 is related to the fact that when the updating steps are small enough and the information provided by the data changes smoothly, y_{t-1} is close to y_t, then the the update from λ_{t-1} to $\lambda_t(f_t)$ and the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ are in the same direction. In this situation, the score update for f_t leads to $f_{t+1} > f_t$ and therefore an update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ in the same direction as the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$ but larger in absolute value. This means that for some small enough $s_{\lambda,t}$ and $s_{\lambda,t+1}$ the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_{t+1})$ reduces the local KL divergence more than the update from $\lambda_t(f_t)$ to $\lambda_{t+1}(f_t)$.

5 Monte Carlo experiment

We present a simulation study as an intuitive illustration of the role that the time-varying parameter α_t can play. The simulation study has a simple design. We generate time series from a stochastic process and we subsequently compare the predictive ability of GAS and aGAS models. The time series are generated by the data generation process (DGP) as given by

$$y_t = \mu^o_t + \eta_t, \quad t \in \mathbb{Z}, \tag{11}$$

where μ_t^o is a deterministic mean and $\{\eta_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Gaussian random variables with zero mean and unit variance. The deterministic mean μ_t^o takes values in $\{0, \delta\}$, $\delta > 0$, and is defined to switch every $\gamma \times 10^2$ time periods from 0 to δ and vice versa. More formally, μ_t^o is specified as

$$\mu_t^o = \begin{cases} 0 & \text{if} \quad \sin\left(\gamma^{-1}10^{-2}(\pi t - 1)\right) \geq 0 \\ \delta & \text{if} \quad \sin\left(\gamma^{-1}10^{-2}(\pi t - 1)\right) < 0. \end{cases} \tag{12}$$

Figure 3 shows a realization from the DGP with $\delta = 3$ and $\gamma = 2$. We consider this particular DGP to provide an intuition for the circumstances under which the time-varying α_t of the aGAS model can be relevant. In time periods where the true μ_t^o is constant, the noise component η_t should not affect the filtered path of the mean very much. This situation requires a small value for α_t. On the other hand, when a break in the level occurs, we need to attain a new level of μ_t^o rapidly. This situation requires a relatively large value for α_t.

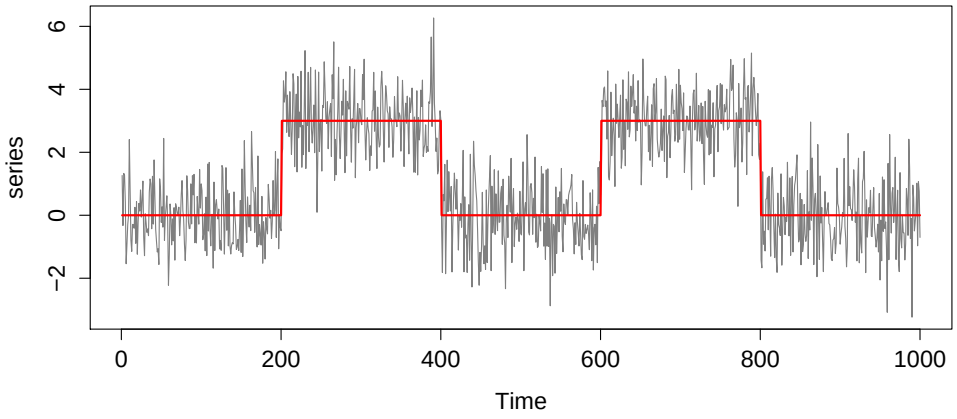

Figure 3: *A realised time series of length* $T = 1000$ *from the DGP* (11) – (12) *with* $\delta = 3$ *and* $\gamma = 2$. *The solid thick (red) line represents the deterministic mean* μ_t^o.

To estimate the time-varying mean μ_t^o from each simulated series, we consider the GAS model (5) with $p(y_t|\lambda_t;\theta)$, for any $t \in \mathbb{Z}$, as a Gaussian density with time-varying mean $\lambda_t = \mu_t$ and time-invariant variance σ^2. The full specification of this GAS model is given by

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \overset{\text{iid}}{\sim} N(0, \sigma^2), \tag{13}$$

with time-varying mean μ_t given by the updating equation

$$\mu_{t+1} = \mu_t + \alpha_\mu s_{\mu,t},$$

where α_μ is a fixed unknown coefficient and $s_{\mu,t}$ is the scaled score function which reduces to the scaled prediction error $s_{\mu,t} = y_t - \mu_t$. This local level GAS model can be represented as an ARIMA$(0,1,1)$ model; we can show this by taking first differences and by observing that we obtain the MA(1) model $y_t - y_{t-1} = (\alpha_\mu - 1)\epsilon_{t-1} + \epsilon_t$. The accelerated GAS model replaces α_μ by a time-varying parameter that, after a transformation, has the updating equation

$$\alpha_t = \exp(f_{t+1}/2), \quad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{f,t},$$

where ω_f, β_f and α_f are treated as fixed unknown coefficients, $s_{\mu,t} = y_t - \mu_t$ and $s_{f,t} = s_{\mu,t}s_{\mu,t-1}$. These expressions for the innovations $s_{\mu,t}$ and $s_{f,t}$ are obtained as special cases of the general treatment for (5) and (6). In this model specification, the Fisher information is constant and therefore the scaling of the score is irrelevant as it only leads to a reparametrization of the model. The GAS model is simply obtained by treating α_t of the aGAS model as a static parameter, that is $\alpha_t = \alpha_\mu$ for any $t \in \mathbb{Z}$.

	$\gamma = 1.0$		$\gamma = 1.5$		$\gamma = 2.0$		$\gamma = 2.5$	
	GAS	aGAS	GAS	aGAS	GAS	aGAS	GAS	aGAS
$\delta = 0.0$	**3.86**	3.99	**3.86**	3.99	**3.86**	3.99	**3.86**	3.99
$\delta = 0.5$	22.34	**22.33**	20.19	**20.19**	18.17	**18.13**	17.05	**16.94**
$\delta = 1.0$	31.69	**31.40**	28.57	**28.07**	25.70	**24.91**	23.99	**22.89**
$\delta = 1.5$	39.21	**38.13**	35.25	**33.56**	31.66	**29.14**	29.48	**26.31**
$\delta = 2.0$	45.78	**43.50**	41.05	**37.62**	36.81	**31.97**	34.21	**28.47**
$\delta = 2.5$	51.78	**48.29**	46.30	**41.26**	41.45	**34.64**	38.47	**30.60**
$\delta = 3.0$	57.38	**53.09**	51.18	**45.02**	45.75	**37.58**	42.40	**32.83**
$\delta = 3.5$	62.71	**58.05**	55.80	**48.98**	49.79	**40.91**	46.08	**35.54**

Table 2: *We present the square root of the mean squared error (MSE) where the error is between the true μ_t^o and the filtered parameter μ_t from GAS and aGAS models, for different true values of δ and γ. The mean is over all time points t and all Monte Carlo replications.*

In our Monte Carlo study we generate $1,000$ time series of sample size $T = 1,000$ from the data generation process, DGP, (11) for different values of δ and γ. For each of the $1,000$ generated series, we estimate by maximum likelihood the parameters in the aGAS model (13) and its standard GAS counterpart. To evaluate the performance of the models, the filtered means for μ_t of these two models are compared with the true mean μ_t^o. We compute the square root of the mean square error (MSE) between the filtered μ_t and true mean μ_t^o, over all time points t and all

Monte Carlo replications. The results of the experiment are collected in Table 2. We learn from these results that the aGAS model can outperform the GAS model. In particular, the MSE of the aGAS model is smaller for all DGPs except for the DGP with $\delta = 0$. This indicates that the aGAS filter is able to better approximate the true μ_t^o in terms of quadratic error. In case $\delta = 0$, the true mean μ_t^o is constant for all t and hence there are no benefits in using a time-varying α_t but there is only the drawback of having more parameters in the model and hence more estimation uncertainty. Similarly, we learn from Table 2 that the improvement due to the dynamic parameter α_t tends to increase as the size of the jumps increases.

To gain more insights into the effect of the dynamic parameter α_t, Figure 4 reports various simulation results for the DGP with $\delta = 3$ and $\gamma = 2$. In the upper graph, we can see that the 90% variability bounds for the aGAS are narrower than those for the GAS in time periods when μ_t^o is constant. It implies that the true mean is predicted with greater accuracy by the aGAS model and that the corresponding filter is less exposed to the noise component. The opposite situation occurs right after the breaks: the variability bounds of the aGAS are larger for a few time periods. It is a consistent finding as the aGAS filter is reacting faster to the change in the level and is then more exposed to the noise component. In the middle graph of Figure 4, the mean squared errors tend to be larger for the GAS model in most time periods. Furthermore, the 90% level confidence bounds show that the aGAS model seems to outperform the GAS not only on average but for almost all individual Monte Carlo draws. Finally, the bottom graph illustrates the behavior of the time-varying α_t. In particular, the dashed line is the average filtered α_t from the aGAS model and the solid line is the average estimate of the static α_μ from the GAS model. The dynamic α_t is close to zero when μ_t^o is constant and it increases after the breaks. The aGAS model clearly offers the flexibility for which it is designed for: it allows the filtered mean to be updated at different speeds in different time periods, where needed.

6 Applications and empirical illustrations

6.1 Conditional volatility Student's t models: at-GARCH and at-GAS

We have argued that the aGARCH model (3) and (4) is a special case of the aGAS model (5) and (6) when considering λ_t as the time-varying variance of the Gaussian conditional density

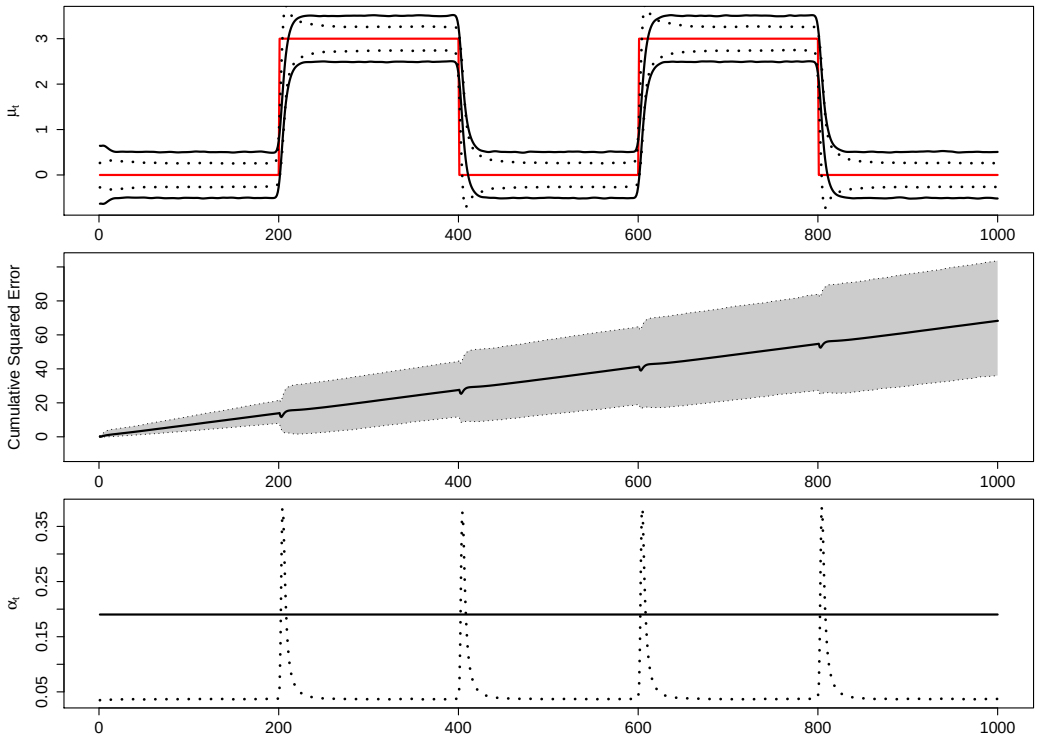

Figure 4: *First plot: the red line represents μ_t^o, the continuous lines represent 90% variability bounds for the GAS μ_t, and the dashed lines represent 90% variability bounds for the aGAS μ_t. Second plot: cumulative squared error difference between the aGAS and the GAS. The shadowed area denotes a 90% confidence region. Third plot: the continuous line is the average estimate of α for the GAS, and the dashed line is the average estimate of α_t for the aGAS.*

$p(y_t|\lambda_t;\theta)$. The time-varying process for α_t is driven by the product of current and lagged volatility innovations. Models with other densities than the Gaussian can also be considered. For example, we can replace the Gaussian by the Student's t density that has fatter tails than the normal. In the case of the GARCH model, we obtain the t-GARCH model as explored by Bollerslev (1986). The accelerated version of the t-GARCH is simply obtained by introducing the time-varying process α_t which is driven by the product of the current and lagged volatility innovation. However, when considering the GAS model (5) with $\lambda_t = \sigma_t^2$ as the time-varying variance of the Student's t conditional density $p(y_t|\lambda_t;\theta)$, we do not obtain the t-GARCH model since the score function is not simply $y_t^2 - \sigma_t^2$. In this case, we obtain the t-GAS model of Creal et al. (2013). We extend the t-GAS model by introducing a time-varying α_t to obtain the accelerated t-GAS (at-GAS) model

as given by

$$y_t = \sigma_t \varepsilon_t, \qquad \sigma_{t+1}^2 = \omega + \beta \sigma_t^2 + \alpha_t \sigma_t^2 s_{\sigma,t},$$
$$\alpha_t = \beta \, \text{logit}(f_{t+1}), \qquad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{\sigma,t} s_{\sigma,t-1},$$

where $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Student's t distributed random variables with zero mean unit variance and ν degrees of freedom. As in Creal et al. (2013), the score innovation $s_{\sigma,t}$ has the following expression

$$s_{\sigma,t} = \frac{(\nu+1)\varepsilon_t^2}{(\nu-2)+\varepsilon_t^2} - 1.$$

The limiting case of $\nu \to \infty$ for the at-GAS model coincides with the aGARCH model. Furthermore, setting $\alpha_t = \alpha$ to a static parameter reduces the model to the t-GAS model.

6.2 Illustration (ctd.): volatility forecasting for all series in S&P500

We continue our empirical study from Section 2.3 and compare a range of models based on the Student's t density in their abilities to forecast conditional volatility accurately. The models considered are the t-GAS, ELt-GAS, at-GAS and aELt-GAS, as well as the GARCH, ELGARCH, aGARCH and aELGARCH models with Student's t densities, which we denote as t-GARCH, ELt-GARCH and at-GARCH, respectively. The specific feature of the t-GAS model specification is that it takes account of the heavy tails in both the observation equation for y_t and the updating equation for the variance σ_t^2. In particular, the impact of extreme return observations on σ_t^2 is attenuated. It is discussed in Creal et al. (2013) that this feature provides various benefits in its treatment of heavy tailed financial time series. Table 3 reports the number of series in the S&P index where each model is outperforming the others, together with the corresponding percentages. The in-sample results show that the four GAS models are the best models for more that 80% of the series. However, this result does not seem to be consistent with the out-of-sample results, for which the four GAS models are the best for less than half of the series. We find that the at-GARCH, aELt-GARCH, at-GAS and aELt-GAS are the best models for a significant proportion of the series, both in-sample and out-of-sample. As in our initial analysis, we can reconsider these results for the 10% of the series with the highest kurtosis. The four GAS models are again the best in-sample specifications for all series. The out-of-sample results are also rather coherent with

this finding. Overall we can say that the the at-GARCH, aELt-GARCH, at-GAS and aELt-GAS models are the best models for a large number of series.

	Full dataset				Top 10% Kurtosis			
	In-sample		Out-of-sample		In-sample		Out-of-sample	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
t-GARCH	24	5.2%	11	2.4%	0	0.0%	0	0.0%
ELt-GARCH	22	4.8%	98	21.3%	0	0.0%	3	6.5%
at-GARCH	12	2.6%	48	10.4%	0	0.0%	5	10.9%
aELt-GARCH	17	3.7%	77	16.7%	0	0.0%	3	6.5%
t-GAS	144	31.3%	53	11.5%	20	43.5%	5	10.9%
ELt-GAS	125	27.2%	50	10.9%	11	23.9%	6	13.0%
at-GAS	21	4.6%	44	9.6%	4	8.7%	7	15.2%
aELt-GAS	95	20.6%	79	17.2%	11	23.9%	17	37.0%
Total	460	100.0%	460	100.0%	46	100.0%	46	100.0%

Table 3: *The number and percentage of series in the S&P 500 index where each model outperforms the others. The in-sample performance is based on the Akaike information criterion. The out-of-sample performance is based on a log-score criterion. In the first panel, all 460 stocks in the S&P500 index are considered. In the second panel, the 46 stocks with the highest in-sample kurtosis are considered.*

We may conclude that, for a significant portion of the S&P500 return series, the inclusion of the dynamic α_t can enhance in-sample and the out-of-sample forecast performances. This conclusion applies to the at-GARCH and aELt-GARCH models as well as the at-GAS and aELt-GAS models. For all these models, the effect of introducing a time-varying α_t appears particularly relevant for heavy-tailed daily return series. These results suggest that different specifications can be useful to obtain a better approximation of the dynamic features of such financial time series. The accelerating GAS framework provides a flexible class of models that can be useful in practical applications.

6.3 An accelerated location and scale model for heavy tailed distributions

We consider a heavy tailed distribution with a time-varying mean (location) and a time-varying variance (scale) using the score-driven approach and for which the parameter that determines the magnitude of the update of the mean process is also time-varying. More specifically, we consider

a Student's t conditional distribution for y_t where both the mean and the variance are time-varying. The resulting model has some similarities with the stochastic volatility model of Stock and Watson (2007). The Student's t distribution in a GAS framework allows us to handle outliers by attenuating their impact on the filtered parameters. Applications in the literature of the Student's t GAS models for location and scale parameters can be found in Creal et al. (2013), Harvey (2013) and Harvey and Luati (2014). In particular, Harvey (2013) have considered a Student's t model with both time-varying mean and variance. The key novelty of the model in the current study is the inclusion of a time-varying parameter α_t in order to let the time-varying location capture a wider range of dynamic specifications.

We consider the aGAS model with time-varying conditional location and scale as given by

$$y_t = \mu_t + \sigma_t \epsilon_t, \tag{14}$$

where μ_t is the time-varying location for y_t, σ_t is the time-varying scale for y_t, and $\{\epsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Student's t distributed random variables with zero mean, unit variance and ν degrees of freedom. The time-varying parameters are described by the following equations

$$\begin{aligned} & & \mu_{t+1} &= \mu_t + \alpha_t s_{\mu,t}, \\ \alpha_t &= \exp(f_{t+1}/2), & f_{t+1} &= \omega_f + \beta_f f_t + \alpha_f s_{f,t}, \\ \sigma_t &= \exp(g_t/2), & g_{t+1} &= \omega_\sigma + \beta_\sigma g_t + \alpha_\sigma s_{\sigma,t}, \end{aligned}$$

where ω_f, β_f, α_f, ω_σ, β_σ and α_σ are static unknown parameters which we estimate by maximum likelihood, and where $s_{\mu,t}$, $s_{f,t}$ and $s_{\sigma,t}$ are the score-based innovations of the processes. We graphically present the functional form of these innovations in Figure 5. The innovation $s_{\mu,t}$ of the location process μ_t is obtained by setting the scaling factor $S_{\mu,t}$ equal to the square root of the inverse Fisher information, that is $s_{\mu,t}$ takes the form

$$s_{\mu,t} = \frac{(\nu+1)(y_t-\mu_t)\sigma_t^{-1}}{(\nu-2)+(y_t-\mu_t)^2\sigma_t^{-2}}.$$

The first graph in Figure 5 presents the effect of $\varepsilon_t = (y_t-\mu_t)/\sigma_t$ on $s_{\mu,t}$. The relationship between ε_t and $s_{\mu,t}$ is nonlinear and the impact of extreme values of ε_t on $s_{\mu,t}$ is attenuated. The degree of

attenuation depends on the degrees of freedom parameter ν: a smaller value for ν delivers a lower sensitivity of $s_{\mu,t}$ on outliers; also see Harvey and Luati (2014) for a more detailed discussion. The innovation $s_{f,t}$ can be obtained from equation (8); by setting $C_{f,t} = S_{\mu,t}S_{\mu,t-1}$, we obtain

$$s_{f,t} = s_{\mu,t}s_{\mu,t-1}.$$

The second graph in Figure 5 shows the effect of ε_t and ε_{t-1} on $s_{f,t}$. We learn that $s_{f,t}$ is positive when ε_t and ε_{t-1} have the same sign and negative when ε_t and ε_{t-1} have opposite signs. Furthermore, extreme values of ε_t and ε_{t-1} are detected as outliers and their impact on $s_{f,t}$ is attenuated. The innovation of the process σ_t takes the form

$$s_{\sigma,t} = \frac{(\nu+1)(y_t-\mu_t)^2\sigma_t^{-2}}{(\nu-2)+(y_t-\mu_t)^2\sigma_t^{-2}} - 1.$$

For this case, the Fisher information is constant and so it does not affect the functional form of $s_{\sigma,t}$. The impact of ε_t on $s_{\sigma,t}$ is shown in the third graph of Figure 5. The update for $s_{\sigma,t}$ is the same as in the Beta-t-EGARCH model of Harvey (2013).

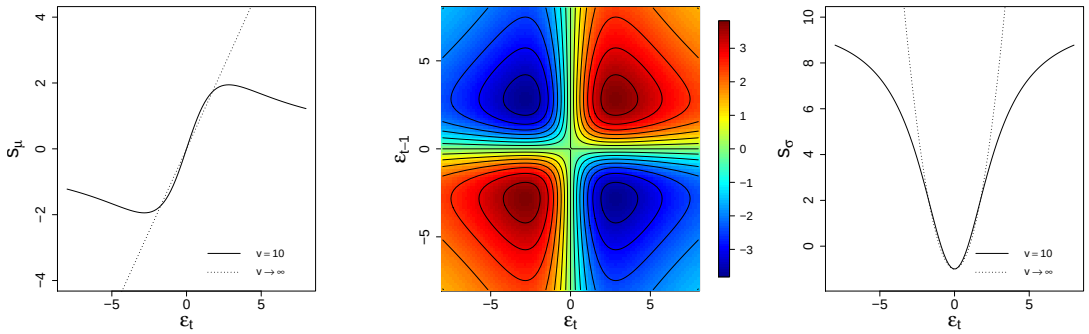

Figure 5: *In the first graph, the values taken by $s_{\mu,t}$ as a function of ε_t are presented. The second graph is a contour plot that shows the values taken by $s_{f,t}$ as a function of ε_t and ε_{t-1}. In the third graph, the values taken by $s_{\sigma,t}$ as a function of ε_t are provided. In all graphs, the degrees of freedom of the Student-t is set to $\nu = 10$.*

When the degrees of freedom of the Student's t distribution gets closer to infinity, $\nu \to \infty$, the Student's t distribution approaches the standard Gaussian distribution. In this limiting case, the model (14) reduces to a Gaussian score-driven model where the innovation for μ_t is simply given by $s_{\mu,t} = (y_t - \mu_t)\sigma_t^{-1}$ while the innovation for σ_t^2 is given by $s_{\sigma,t} = (y_t - \mu_t)^2\sigma_t^{-2} - 1$. The impact functions of the standardized observation $(y_t - \mu_t)\sigma_t^{-1}$ on $s_{\mu,t}$ and $s_{\sigma,t}$ are presented in Figure 5.

6.4 Empirical illustration

In our final empirical illustration we consider the US quarterly consumer price (CP) index, which is obtained from the FRED dataset. The inflation time series y_t is computed as the annualized log-difference of the price index series p_t, we adopt the standard transformation $y_t = 400 \log(p_t/p_{t-1})$. The inflation series is computed from the first quarter of 1952 to the first quarter of 2015. The resulting time series is presented in Figure 6. We consider several specifications for the aGAS model which are listed in Table 4.

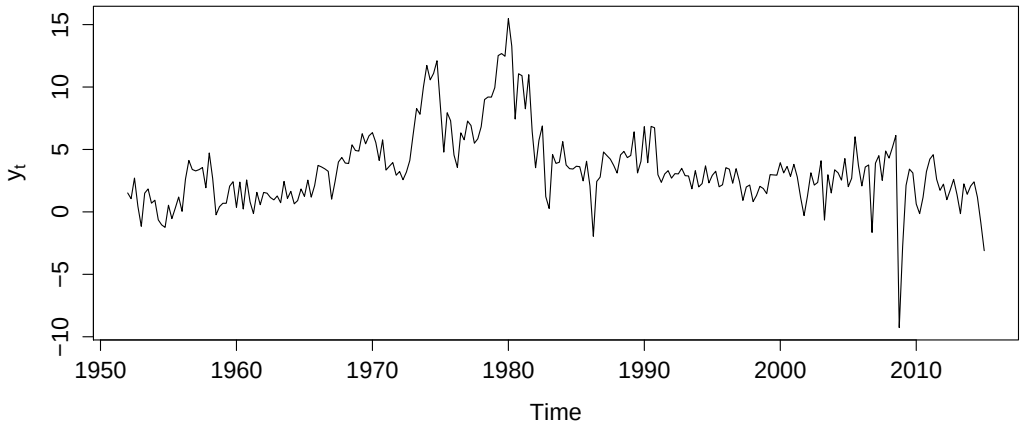

Figure 6: *Quarterly consumer price US inflation series.*

	Description	Reference
Model t.1	Our full model in (14)	
Model t.2	$\beta_\sigma = 0$ and $\alpha_\sigma = 0$	
Model t.3	$\beta_f = 0$ and $\alpha_f = 0$	Harvey (2013)
Model t.4	$\beta_\sigma = 0$, $\alpha_\sigma = 0$, $\beta_f = 0$ and $\alpha_f = 0$	Harvey and Luati (2014)
Model n.i	Limiting case of Model t.i with $v \to \infty$	$i = 1,2,3,4$

Table 4: *The second column describe the specification of the model. The third column provides some references for the specific models obtained constraining the parameters of the full model in (14).*

The parameter estimates for all Models t.1-t.4 and Models n.1-n.4 are presented in Table 5, together with the maximized log-likelihood value and the corresponding p-value of the likelihood ratio (LR) test and the Akaike information criterion (AIC). The LR test p-value is reported for each model in relation to the corresponding full model (14). We can conclude from the reported results that the inclusion of the time-varying scale σ_t as well as the time-varying α_t are highly significant

	δ_f	β_f	α_f	δ_σ	β_σ	α_σ	ν	loglik	LR	AIC
Model t.1	-1.518 (0.799)	0.967 (0.027)	0.258 (0.113)	1.055 (0.236)	0.861 (0.092)	0.215 (0.089)	5.571 (1.572)	-475.4	-	**964.7**
Model t.2	-1.493 (0.402)	0.914 (0.028)	0.294 (0.071)	1.182 (0.178)	-	-	3.826 (0.553)	-482.7	0.001	975.4
Model t.3	-0.468 (0.280)	-	-	1.080 (0.207)	0.869 (0.126)	0.163 (0.099)	7.583 (2.399)	-481.8	0.002	973.6
Model t.4	-0.305 (0.213)	-	-	1.111 (0.134)	-	-	5.639 (1.431)	-488.8	0.000	983.7
Model n.1	-1.366 (0.618)	0.969 (0.022)	0.182 (0.072)	1.169 (0.203)	0.937 (0.030)	0.088 (0.033)	-	-504.2	-	1020.4
Model n.2	-0.304 (0.416)	0.971 (0.028)	0.060 (0.036)	1.251 (0.089)	-	-	-	-515.3	0.000	1038.6
Model n.3	-0.231 (0.314)	-	-	1.213 (0.161)	0.939 (0.026)	0.054 (0.021)	-	-510.2	0.002	1028.4
Model n.4	-0.080 (0.266)	-	-	1.264 (0.089)	-	-	-	-516.8	0.000	1037.7

Table 5: *Parameter estimates for the models in Table 4, together with their standard errors in brackets. The last three columns contain respectively the maximized log-likelihood value, the p-value of the likelihood ratio (LR) test with respect to the full models and the Akaike information criterion (AIC). The parameters δ_f and δ_σ are given by $\delta_f = \omega_f/(1-\beta_f)$ and $\delta_\sigma = \omega_\sigma/(1-\beta_\sigma)$.*

for our US inflation series. In particular, we obtain that the null hypothesis of the LR test is rejected at a 1% confidence level, for all other model specifications. Also, the model with the lowest AIC is Model t.1. The reported AIC statistics also indicate that the Student's t specifications, Models t.1-t.4, have a better fit than their limiting counterparts, Models n.1-n.4. This is confirmed by the estimates for the degrees of freedom ν which are all small for the four Student's t models.

Figure 7 presents the filtered estimates of the parameters μ_t, σ_t and α_t for our preferred Model t.1. The graph of the filtered μ_t shows the robustness of the model in its handling of outliers. For example, in the fourth quarter of 2008, the extreme peak in US inflation time series does not affect the filtered path of μ_t very much. The graph of the filtered estimate of α_t shows that during the enduring period of exceptional high inflation, approximately between 1972 and 1983, also the filtered α_t takes high values. Clearly, during periods of persistent and sudden changes in the location for y_t, the parameter μ_t required fast updating to capture the changes. The time-varying α_t plays a key role in accommodating the fast updating for location. The third graph in Figure 7 indicates or suggests that the variability σ_t appears to increase in periods of lasting economic recessions in the US; see the NBER recession datings in the first graph.

To investigate in detail the effect of the inclusion of the time-varying parameter α_t on the

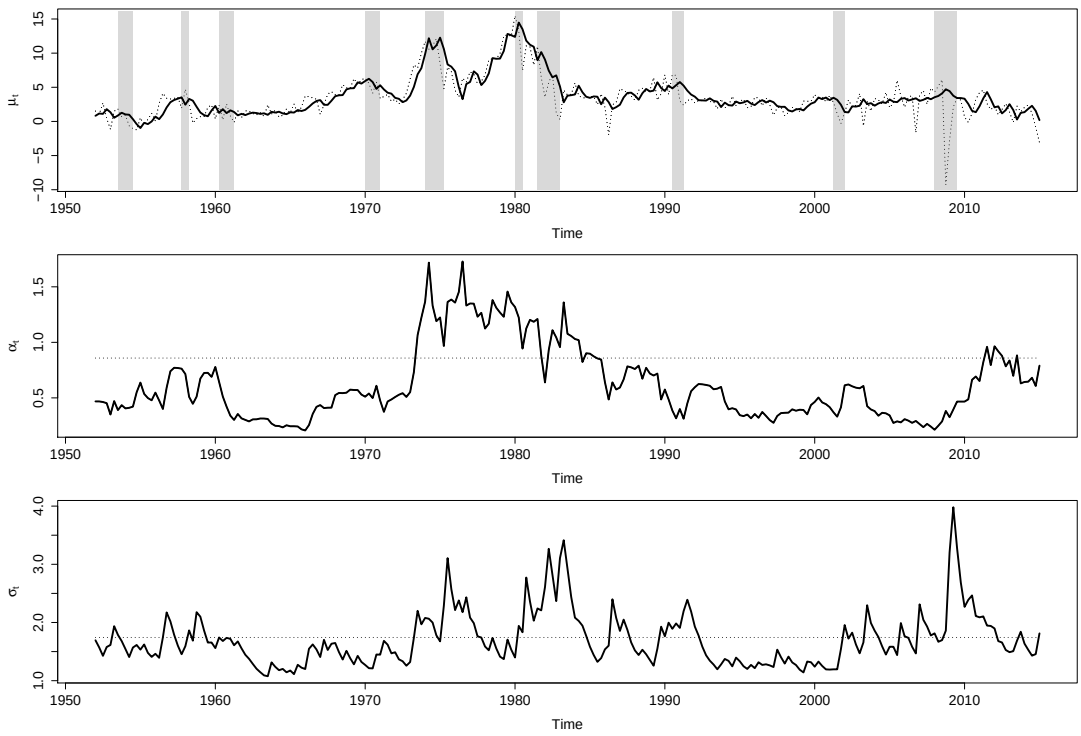

Figure 7: *The filtered estimates of the time-varying parameters in Model t.1: upper plot is μ_t, middle plot is α_t, and lower plot is σ_t.*

filtered estimate of μ_t, we present in Figure 8 the filtered estimates μ_t from Model t.1 and Model t.3. Both models include the time-varying scale σ_t, the only difference between the two models is that α_t is not time-varying in Model t.3. We consider two periods where the inflation series exhibit different behaviors: the first graph in Figure 8 is for the period from 1973 to 1982, while the second graph is for the period from 1999 to 2008. In the first period 1973 – 1982, the time series appears to be subject to a fast changing location. It may imply a low persistence in US inflation for this period. We observe that the filtered estimate of α_t contains some large values; see the second graph in Figure 7. This allows the μ_t of Model t.1 to react more promptly to the changes in the level of the series. The filtered estimate of μ_t from Model t.1 exceeds its counterpart from Model t.3 when the inflation level is increasing and vice versa when the inflation level is decreasing. For the period between 1999 and 2008, the second graph in Figure 8 shows that the inflation series seems to change location more slowly: it appears as a slow and lightly trending filtered μ_t subject to much noise. In this case, we have small values for the time-varying filtered estimate of α_t,; see the second graph in Figure 7, allow the μ_t of Model t.1 to change slowly, capturing the increasing trend but not being too much affected by the noise. The benefit of the time-varying α_t can be noted

from the plot as the filtered μ_t of Model t.3 is more noisy than the filtered μ_t of Model t.1. These two graphs in Figure 8 show how the inclusion of the time-varying α_t allows the dynamic model to be more flexible and better in adapting to changing behaviors of the series. The improvements in terms of in-sample fit are also confirmed by the likelihood ratio test and the AIC.

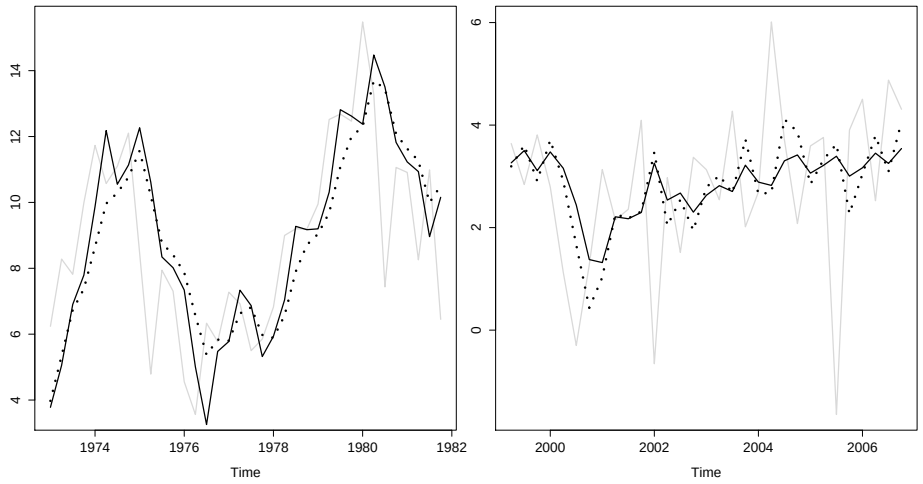

Figure 8: *The filtered estimates of μ_t from Model t.1 and Model t.3 for two different time periods. The gray line is US inflation, the dashed line is the filtered μ_t estimate from Model t.3 and the solid line is the filtered μ_t estimated from our preferred Model t.1.*

Finally, we have carried out a limited pseudo out-of-sample forecasting study to compare the performances of the models in Table 4. For this part of the study we have also included three other models to facilitate forecast comparisons: the local level model as discussed in Durbin and Koopman (2012, Chapter 2) and the well-known autoregressive integrated moving average ARIMA(P, D, Q) model with orders $P = 4, D = 1, Q = 0$ and $P = 1, D = 1, Q = 1$. The forecast mean square error (FMSE) and the forecast mean absolute error (FMAE) are computed using the last 100 observations and parameter estimation for the different model specifications is performed using a fixed rolling window. We consider h-steps ahead forecasts, for $h = 1, 2, 3, 4$. Differences in forecast accuracy are verified by means of the Diebold and Mariano (DM) test, see Diebold and Mariano (1995). The DM test is used to test the null hypothesis that Model t.1 has the same FMSE as the other models against the alternative of different FMSE. We notice that the DM test is performed for both nested and non-nested models; the asymptotic normal distribution of the DM test statistic for nested models is ensured by the scheme of a fixed rolling window; see Giacomini and White (2006). The results are presented in Table 6. We find that Model n.1 has the smallest FMSE and FMAE and Model t.1 has the best forecasting performance among the fat-

tailed models. It suggests that the inclusion of the time-varying α_t tends to enhance the forecasting performance of the GAS models. For the forecasting horizon of 1 year ($h = 4$ quarters), Model t.1 significantly outperforms most of the models at a 5% or 10% significance level. With regards to the other forecasting horizons, we conclude that we cannot reject the hypothesis that the differences in terms of forecast accuracy observed in the subsamples are not significant.

	FMSE				FMAE			
	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$
Model t.1	1.00	1.00	1.00	**1.00**	**1.00**	1.00	1.00	1.00
Model t.2	1.02	1.02	1.02	1.05*	1.01	1.02	1.01	1.03*
Model t.3	1.11	1.12	1.09	1.13**	1.04	1.04	1.03	1.07**
Model t.4	1.13*	1.14*	1.09	1.16**	1.05	1.06*	1.01	1.07**
Model n.1	**0.96**	**0.99**	**0.98**	**1.00**	**1.00**	**0.98**	**0.99**	**0.99**
Model n.2	1.02	1.20	1.18	1.15*	1.02	1.09	1.07	1.05
Model n.3	1.03	1.09	1.06	1.09	1.04	1.03	1.03	1.04
Model n.4	1.02	1.20	1.18	1.15*	1.02	1.09	1.07	1.05
Local level model	1.02	1.20	1.18	1.15*	1.02	1.09	1.07	1.05
ARIMA(4,1,0)	1.06	1.25	1.32	1.25**	1.02	1.06	1.10	1.10*
ARIMA(1,1,1)	0.98	1.16	1.13	1.12	**1.00**	1.06	1.04	1.03

Table 6: *Empirical out-of-sample FMSE and FMAE ratio statistics, based on the last 100 observations of quarterly US consumer price inflation time series. The benchmark is Model t.1. The FMSE and FMAE of Model t.1 is the denominator of the ratio.*

7 Conclusion

We have introduced a novel class of observation-driven models that allows for locally changing the weights for updating the time-varying parameters. We provide both theoretical and simulation-based evidence that these so-called accelerated GAS model can outperform corresponding GAS models that have a time-invariant structure for GAS updating. Two empirical illustrations have been provided: one for the S&P 500 index series and one for US inflation series. For these highly relevant illustrations we find that the proposed accelerating framework is capable to improve the in-sample and out-of-sample fit for GAS and related models.

Appendix

Proof of Lemma 1. The first statement follows by noting that $\lambda_t(f_t) = \lambda_t^*$ if $\{f_t\}_{t\in\mathbb{N}}$ is a random sequence such that $f_t = g^{-1}\big((\lambda_{t+1}^* - \omega_\lambda - \beta_\lambda\lambda_t)/s_{\lambda,t}\big)$ for any $t \in \mathbb{N}$. As concerns the second statement, the if part is immediately proved when we notice that $s_{\lambda,t} = (\lambda_{t+1}^* - \omega_\lambda - \beta_\lambda\lambda_t)/g(c)$ implies $f_t = c$. Finally, to prove the only if part of the statement, suppose that, for some $t \in \mathbb{N}$, there exists no $c \in \mathbb{R}$ such that $s_{\lambda,t} = (\lambda_{t+1}^* - \omega_\lambda - \beta_\lambda\lambda_t)/g(c)$, then, setting $f_t = c\ \forall\ t$ implies that $\lambda_t(f_t) \neq \lambda_t^*$ for some $t \in \mathbb{N}$, for any possible $c \in \mathbb{R}$. □

Proof of Proposition 1. The proof follows the same argument as in Blasques et al. (2015). By an application of the mean value theorem, the local realized KL divergence can be expressed as

$$\begin{aligned}
\Delta_{f,t}^{t+1} &= \int_{B(y_t,\epsilon_y)} p_t^o(y) \log \frac{p(y|\lambda_t(f_t))}{p(y|\lambda_t(f_{t+1}))} dy = \\
&= \int_{B(y_t,\epsilon_y)} p_t^o(y) \frac{\partial \log p(y|\lambda_t(\dot{f}_t))}{\partial \dot{f}_t} (f_t - f_{t+1}) dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_{t-1},\lambda_{t-1}\big)^2 u_\lambda\big(y,\lambda_t(\dot{f}_t)\big) u_\lambda\big(y_t,\lambda_t(f_t)\big) dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y,\lambda_t(\dot{f}_t)\big) u_\lambda\big(y_t,\lambda_t(f_t)\big) dy,
\end{aligned}$$

where $\tilde{C}_t = \alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda(y_{t-1},\lambda_{t-1})^2$ and $\dot{f}_t$ is a point between f_t and f_{t+1}. By again applying the mean value theorem, we obtain

$$\begin{aligned}
\Delta_{f,t}^{t+1} &= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y,\lambda_t(\dot{f}_t)\big) u_\lambda\big(y_t,\lambda_t(f_t)\big) dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y_t,\lambda_t(f_t)\big)^2 dy + \qquad (15) \\
&\quad -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y_t,\lambda_t(f_t)\big) \frac{\partial u_\lambda\big(\dot{y}_t,\lambda_t(\ddot{f}_t)\big)}{\partial \dot{y}_t} (y - y_t) dy + \qquad (16) \\
&\quad -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y_t,\lambda_t(f_t)\big) \frac{\partial u_\lambda\big(\dot{y}_t,\lambda_t(\ddot{f}_t)\big)}{\partial \ddot{f}_t} (\dot{f}_t - f_t) dy, \qquad (17)
\end{aligned}$$

where $\ddot{f}_t$ is a point between $\dot{f}_t$ and f_t, and $\dot{y}_t$ is a point between y and y_t. The desired result follows since the term (15) is a.s. negative and the terms (16) and (17) can be made arbitrary small in absolute value compared to the first term by selecting the ball radius ϵ_y and ϵ_f small enough. □

Proof of Proposition 2. The if part of the proposition follows immediately from a similar argument as in the proof of Proposition 1. As concerns the only if part, if $sign(f_{t+1} - f_t) = sign(s_{f,t})$ does not hold with probability 1 for any $f_t \in \mathcal{F}$, it means that there exists an $f_t \in \mathcal{F}$ such that $sign(f_{t+1} - f_t) \neq sign(s_{f,t})$ holds with positive probability. Following a similar argument as in the proof of Proposition 1, this implies that there is a positive probability to have an y_t such that

$$\Delta_{f,t}^{t+1} = -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_t, \lambda_t(\dot{f}_t)\big)(f_{t+1} - f_t)dy > 0,$$

for small enough $\epsilon_y > 0$ and $\epsilon_f > 0$. This concludes the proof. □

Proof of Proposition 3. The line of argument is similar as in the proof of Proposition 1, the result follows by repeated applications of the mean value theorem. The difference in local KL variation can be expressed as

$$\begin{aligned}
\Delta_{\lambda,t+1}^{t+1} - \Delta_{\lambda,t+1}^{t} &= \int_{B(y_t,\epsilon_y)} p_t^o(y) \log \frac{p(y|\lambda_{t+1}(f_t))}{p(y|\lambda_{t+1}(f_{t+1}))} dy = \\
&= \int_{B(y_t,\epsilon_y)} p_t^o(y) \frac{\partial \log p(y|\lambda_{t+1}(\dot{f}_t))}{\partial \dot{f}_t}(f_t - f_{t+1})dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_t, \lambda_t(f_t)\big)^2 u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y, \lambda_{t+1}(\dot{f}_t)\big)dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y, \lambda_t(\dot{f}_t)\big)dy,
\end{aligned}$$

where $\tilde{C}_t = \alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_t, \lambda_t(f_t)\big)^2$ and $\dot{f}_t$ is a point between f_t and f_{t+1}. Applying again the mean value theorem it results

$$\begin{aligned}
\Delta_{\lambda,t+1}^{t+1} - \Delta_{\lambda,t+1}^{t} &= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y, \lambda_t(\dot{f}_t)\big) u_\lambda\big(y_{t-1}, \lambda_{t-1}\big)dy = \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t U_{1,t} U_{2,t} dy,
\end{aligned}$$

where $U_{1,t}$ and $U_{2,t}$ are respectively given by

$$U_{1,t} = u_\lambda\big(y_t, \lambda_t(f_t)\big) + \frac{\partial u_\lambda(\dot{y}_t, \dot{\lambda}_t)}{\partial \dot{\lambda}_t}(\lambda_{t+1}(\dot{f}_t) - \lambda_t(f_t)) + \frac{\partial u_\lambda(\dot{y}_t, \dot{\lambda}_t)}{\partial \dot{y}_t}(y - y_t)$$

and

$$U_{2,t} = u_\lambda\big(y_t, \lambda_t(f_t)\big) + \frac{\partial u_\lambda(\ddot{y}_t, \ddot{\lambda}_t)}{\partial \ddot{\lambda}_t}(\lambda_{t-1} - \lambda_t(f_t)) + \frac{\partial u_\lambda(\ddot{y}_t, \ddot{\lambda}_t)}{\partial \ddot{y}_t}(y_{t-1} - y_t),$$

with $\dot{y}_t$ a point between y_t and y, $\dot{\lambda}_t$ a point between $\lambda_t(f_t)$ and $\lambda_{t+1}(\dot{f}_t)$, $\ddot{y}_t$ a point between y_{t-1} and y_t and $\ddot{\lambda}_t$ a point between λ_{t-1} and $\lambda_t(f_t)$. From Assumption 1, the score $u_\lambda\big(y_t, \lambda_t(f_t)\big)$ is nonzero with probability 1, and the second and third terms in the expressions of $U_{1,t}$ and $U_{2,t}$ can be made arbitrary small in absolute value with respect to the first term by selecting the ball radius ϵ_y and ϵ_λ small enough. Hence the product $U_{1,t}U_{2,t}$ can be made positive for any $\dot{y}_t, y \in B(y_t, \epsilon_y)$. This, together with the positivity of $p_t^o(y)$ and $\tilde{C}_t$, implies that $\Delta_{\lambda,t+1}^{t+1} - \Delta_{\lambda,t+1}^{t}$ is negative. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In Petrov, B. and Caski, F., editors, *Proceedings of the Second International Symposium on Information Theory, Armenian SSR*, pages 267–281. Akademiai Kiado, Budapest.

Andres, P. (2014). Computation of maximum likelihood estimates for score driven models for positive valued observations. *Computational Statistics and Data Analysis*, 76:34–43.

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

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

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

Creal, D., Schwaab, B., Koopman, S. J., and Lucas, A. (2014). Observation driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics*, 96:898–915.

De Lira Salvatierra, I. and Patton, A. J. (2015). Dynamic copula models and high frequency data. *Journal of Empirical Finance*, 30:120–135.

Diebold, F. X. and Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, 13:253–265.

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

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

Engle, R. F. and Lee, G. G. J. (1999). A long-run and short-run component model of stock return volatility. In *Cointegration, causality and forecasting: A festschrift in honor of Clive W. J. Granger*. New York: Oxford University Press.

Geweke, J. and Amisano, G. (2011). Optimal prediction pools. *Journal of Econometrics*, 164:130–141.

Giacomini, R. and White, H. (2006). Tests of conditional predictive ability. *Econometrica*, 74:1545–1578.

Harvey, A. (2013). *Dynamic Models for Volatility and Heavy Tails: With Applications to Financial and Economic Time Series*. New York: Cambridge University Press.

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

Hjort, N. L. and Jones, M. C. (1996). Locally parametric nonparametric density estimation. *Annals of Statistics*, 24(4):1433–1854.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physics Reviews*, 106:620–630.

Jaynes, E. T. (2003). *Probability Theory: The Logic of Science*. Cambridge University Press, Cambridge.

Kullback, S. (1959). *Information Theory and Statistics*. Wiley, New York.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Maasoumi, E. (1986). The measurement and decomposition of multidimensional inequality. *Econometrica*, 54:991–997.

Nelson, D. B. (1990). Stationarity and persistence in the GARCH(1,1) model. *Econometric Theory*, 6:318–334.

Oh, D. H. and Patton, A. J. (2017). Time-varying systemic risk: Evidence from a dynamic copula model of CDS spreads. *Journal of Business and Economic Statistics*, page forthcoming.

Quaedvlieg, R., Bollerslev, T., and Patton, A. J. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, 192:1–18.

Stock and Watson (2007). Why has U.S. inflation become harder to forecast? *Journal of Money, Credit and Banking*, 39:3–33.

Ullah, A. (1996). Entropy, divergence and distance measures with econometric applications. *Journal of Statistical Planning and Inference*, 69:137–162.

Ullah, A. (2002). Uses of entropy and divergence measures for evaluating econometric approximations and inference. *Journal of Econometrics*, 107(1-2):313–326.