Accelerating score-driven time series models

Blasques, F.; Gorgi, P.; Koopman, S. J.

published in
Journal of Econometrics
2019

DOI (link to publisher)
10.1016/j.jeconom.2019.03.005

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Blasques, F., Gorgi, P., & Koopman, S. J. (2019). Accelerating score-driven time series models. *Journal of Econometrics*, *212*(2), 359-376. https://doi.org/10.1016/j.jeconom.2019.03.005

Journal of Econometrics 212 (2019) 359–376

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Accelerating score-driven time series models

F. Blasques [a,b], P. Gorgi [a], S.J. Koopman [a,b,c,*,1]

[a] Vrije Universiteit Amsterdam, The Netherlands
[b] Tinbergen Institute, The Netherlands
[c] CREATES, Aarhus University, Denmark

ARTICLE INFO

Article history:
Received 7 July 2017
Received in revised form 11 February 2019
Accepted 4 March 2019
Available online 11 April 2019

Keywords:
GARCH models
Kullback–Leibler divergence
Score-driven models
S&P 500 stocks
Time-varying parameters
US inflation

ABSTRACT

We propose a new class of score-driven time series models that allows for a more flexible weighting of score innovations for the filtering of time varying parameters. The parameter for the score innovation is made time-varying by means of an updating equation that accounts for the autocorrelations of past innovations. We provide the theoretical foundations for this acceleration method by showing optimality in terms of reducing Kullback–Leibler divergence. The empirical relevance of this accelerated score-driven updating method is illustrated in two empirical studies. First, we include acceleration in the generalized autoregressive conditional heteroskedasticity model. We adopt the new model to extract volatility from exchange rates and to analyze daily density forecasts of volatilities from all individual stock return series in the Standard & Poor's 500 index. Second, we consider a score-driven acceleration for the time-varying mean and use this new model in a forecasting study for US inflation.

1. Introduction

Economic and financial time series often exhibit intricate dynamic features. When the time series is analyzed 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, when filtering the conditional volatility of daily financial returns by means of the well-known GARCH model of Engle (1982) and Bollerslev (1986), we can describe well smooth changes in volatility. However, in the event of a financial crisis or a major news event, the volatility level may change suddenly and the GARCH model may not be suited to properly describe such changes. We introduce a dynamic specification which allows us 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.

We base our developments on score-driven time series models which are also referred to as generalized autoregressive score (GAS) models in Creal et al. (2013) and dynamic conditional score (DCS) models in Harvey (2013). The class of GAS models 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) who consider location and scale models for fat-tailed distributions, Creal et al. (2014) who explore

* Correspondence to: Department of Econometrics, Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, The Netherlands.
E-mail address: s.j.koopman@vu.nl (S.J. Koopman).

[1] Koopman acknowledges the support from CREATES, the Center for Research in Econometric Analysis of Time Series (DNRF78) at Aarhus University, Denmark, funded by the Danish National Research Foundation.

https://doi.org/10.1016/j.jeconom.2019.03.005
0304-4076/

dynamic factor models, 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 recent developments on score-driven models is provided online at http://gasmodel.com.

We propose a generalization of the GAS model by introducing a more flexible weighting scheme for the score innovation: the accelerating GAS (aGAS) model. This key generalization allows the weighting parameter of the score innovation in GAS models to be time-varying with an updating function that is similar to GAS itself but with the score innovation replaced by the product of the score innovation and its lagged value. This product is the ingredient for the estimate of the first-order autocorrelation of score innovations and it provides some indication of the importance of the score innovation for volatility updating. When recent score innovations have the same sign, the adjustment of the current dynamic parameter 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 allow the weighting parameter to be a function of the local estimate of autocorrelation, the updating will accelerate when a set of consecutive innovations have the same sign.

We discuss the intuition behind the accelerating mechanism through several illustrations and provide the theoretical justification for the 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 how acceleration can be effective, how it lets the models become more flexible, and how it can improve the approximation of the unknown data generating process. The empirical relevance of aGAS models is first investigated for GARCH models with the applications of extracting volatility from exchange rate time series and of investigating daily density forecasts of volatilities from all stocks present in the Standard & Poor's 500 index. The two applications show that the accelerating mechanism can be useful in capturing sudden changes in the volatility level. The method is also shown to provide benefits in terms of density forecasts of daily stock returns. Finally, in the context of location and scale models, we consider an empirical application for the modeling 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 1970s and 1980s, but also, the smooth and slow dynamic behavior of the conditional mean during the great moderation of two decades that followed the early 1980s.

The paper is structured as follows. Section 2 presents the general aGAS framework. Section 3 develops the optimality properties for the aGAS model. Section 4 discusses the results of our simulation study. Section 5 presents an application for an Asian exchange rate series and for all stock returns in the S&P500 index. Section 6 presents an application to US inflation. Section 7 concludes.

2. Accelerated score-driven time series models

We introduce the accelerated score-driven model which generalizes the score-driven time series model of Creal et al. (2013) and Harvey (2013). For a time series variable $\{y_t\}_{t\in\mathbb{Z}}$, the basic GAS model is given by

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

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).

The accelerated GAS (aGAS) model is defined as the GAS model (1) with a time-varying α_λ coefficient that we specify as

$$\alpha_{\lambda,t} = g(f_{t+1};\theta), \qquad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f s_{f,t}, \tag{2}$$

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 depends on the first derivative of the conditional density contribution at time t, that is

$$s_{f,t} = S_{f,t}\, u_{f,t}, \qquad u_{f,t} = \frac{\partial \log p(y_t|\lambda_t;\theta)}{\partial f_t}, \tag{3}$$

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 $\{y_{t-1}, y_{t-2}, \ldots\}$. It is straightforward to show that the innovation $s_{f,t}$ in (3) can be expressed as

$$s_{f,t} = C_{f,t} u_{\lambda,t} u_{\lambda,t-1}, \tag{4}$$

where $C_{f,t}$ is a positive scaling factor and is a function of the scaling factors $S_{\lambda,t}$ and $S_{f,t}$. Note that the derivation of this expression follows from the fact the derivative of the log density is taken with respect to f_t and therefore unfolding the GAS recursion in (3) is not needed. This is the equivalent of implicitly taking $\partial\lambda_{t-1}/\partial f_t = 0$ in the calculation of the derivative of the unfolded process. The expression (4) 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 (4) 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. 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 implication that the variability of the innovation of the autoregressive process in (1) is determined solely by $\alpha_{\lambda,t}$.

3. Optimality properties

We provide a theoretical justification for the aGAS specification in (1) and (2). 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 (4) 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 distribution of the DGP. Section 3.1 introduces this framework, Section 3.2 delivers the optimality of the score update for $\alpha_{\lambda,t}$ and Section 3.3 shows how flexible GAS models can outperform standard GAS models.

3.1. 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 (1), $y_t \sim p(y_t|\lambda_t;\theta)$, 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

$$\mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_t;\theta)\right) = \int_Y p_t^o(y) \log \frac{p_t^o(y)}{p(y|\lambda_t;\theta)} dy, \tag{5}$$

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 $\{\lambda_t^*\}_{t\in\mathbb{N}}$.

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 that the time-varying update

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

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 $t \in \mathbb{N}$.*

3.2. Optimality of score innovations

In practice, the problem is how to specify the dynamics of f_t. We will address this issue by providing a theoretical justification for the score-based update for f_t. 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) show 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

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

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 (2), 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 is 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 justifies 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 (2) 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 (2) setting $(\omega_f, \beta_f) \approx (0, 1)$,

$$f_{t+1} = f_t + \alpha_f s_{f,t}. \tag{6}$$

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

$$\Delta_{f,t}^{t+1} = \mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_t(f_{t+1});\theta)\right) - \mathrm{KL}_Y\left(p_t^o, p(\cdot|\lambda_t(f_t);\theta)\right).$$

A parameter update for f_t is said to be optimal in local realized KL divergence if and only if $\Delta_{f,t}^{t+1} < 0$ almost surely for any $(f_t, \theta) \in \mathcal{F} \times \Theta$.

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 parameter f_t is always updated, and it also has 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 (6) is optimal in local realized KL divergence as long as α_f is positive.*

The next proposition stresses that only the score $u_{f,t}$ delivers 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 $f_t \in \mathcal{F}$.*

3.3. 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 results 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 as 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 (6) dominates 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 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 $\lambda_{t+1}(f_t)$.

We can summarize the optimality properties derived in this section as follows. Proposition 1 shows that the GAS model with f_t updated using (6) has locally a smaller KL divergence with respect to the true DGP, than the model with a non-updated f_t. Then, Proposition 2 shows that the any parameter update that delivers this optimal result must be locally equivalent to (6), i.e. the sign has to be the same as the score in (6). Finally, Proposition 3 extends the results in Proposition 1 by studying the optimality in KL divergence of the overall update of λ_t instead of f_t only. The proofs are presented in Appendix A.

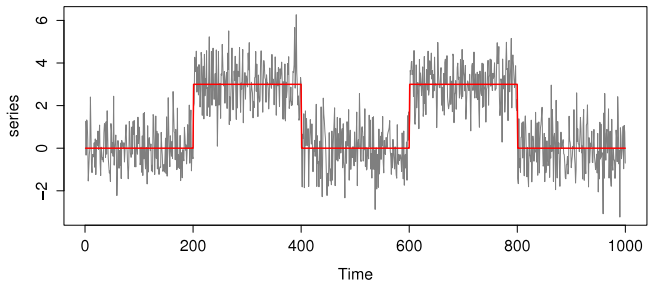

Fig. 1. A realized time series of length $T = 1000$ from the DGP (7)–(8) with $\delta = 3$ and $\gamma = 2$. The solid thick (red) line represents the deterministic mean μ_t^o.

4. 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

$$y_t = \mu_t^o + \eta_t, \quad t \in \mathbb{Z}, \tag{7}$$

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

$$\mu_t^o = \begin{cases} 0 & \text{if} \quad \sin\left(\gamma^{-1}10^{-2}(\pi t - 1)\right) \geq 0 \\ \delta & \text{if} \quad \sin\left(\gamma^{-1}10^{-2}(\pi t - 1)\right) < 0. \end{cases} \tag{8}$$

Fig. 1 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.

To estimate the time-varying mean μ_t^o from each simulated series, we consider the GAS model (1) 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

$$y_t = \mu_t + \epsilon_t, \qquad \epsilon_t \overset{\text{i.i.d}}{\sim} N(0, \sigma^2), \tag{9}$$

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 innovations $s_{\mu,t}$ and $s_{f,t}$ are obtained as special cases of the general treatment for (1) and (2). 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 $t \in \mathbb{Z}$.

In our Monte Carlo study we generate 1,000 time series of sample size $T = 1,000$ from the data generation process, DGP, (7) 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 (9) 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 1. 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 $\delta = 0$.

Table 1
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.

	$\gamma = 1.0$		$\gamma = 1.5$		$\gamma = 2.0$		$\gamma = 2.5$	
	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**
$\delta = 2.0$	45.78	**43.50**	41.05	**37.62**	36.81	**31.97**	34.21	**28.47**
$\delta = 3.0$	57.38	**53.09**	51.18	**45.02**	45.75	**37.58**	42.40	**32.83**

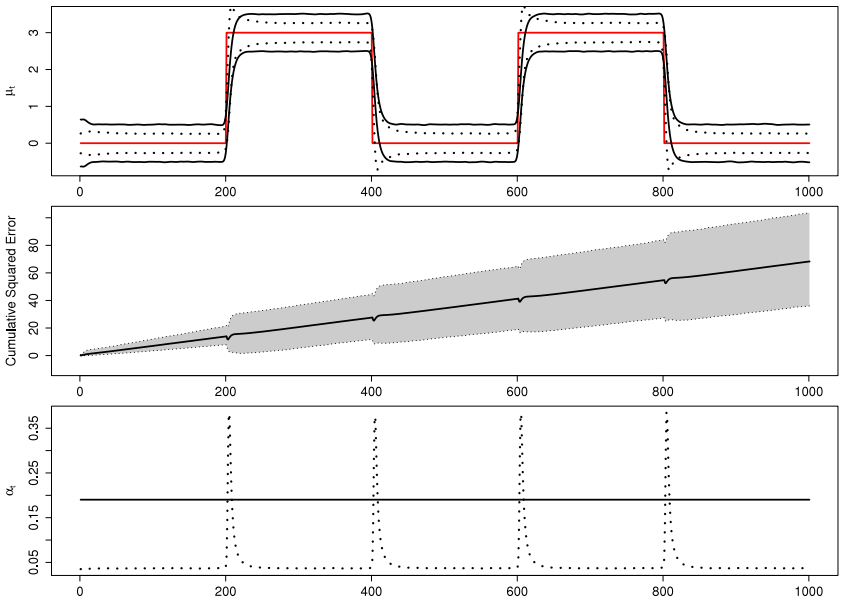

Fig. 2. First plot: the solid lines represent μ_t^o (in red) and the 90% variability bounds for the GAS μ_t, and the dotted lines represent 90% variability bounds for the aGAS μ_t. Second plot: cumulative squared error difference between the aGAS and the GAS. The shadowed area denotes a 90% confidence region. Third plot: the solid line is the average estimate of α for the GAS, and the dotted line is the average estimate of α_t for the aGAS. All estimates are smoothed to reduce Monte Carlo uncertainty.

To gain more insights into the effect of the dynamic parameter α_t, Fig. 2 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 Fig. 2, 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.

5. Accelerating GARCH and related models

In this section we motivate our extension for score-driven time-varying parameter models by introducing the accelerated updating mechanism for some GARCH models. A natural updating function of current and past squared-returns is proposed and discussed. The empirical relevance of our extension is investigated in an illustrative application concerning an Asian exchange rate and in a forecasting study concerning 436 stock return series for US companies that are present in the S&P500 stock index.

5.1. Model formulation

The GARCH model of Engle (1982) and Bollerslev (1986) treats the clustering of large, but also small, shocks in time series of *financial returns* $\{y_t\}_{t\in\mathbb{Z}}$. The variable y_t typically represents daily differences of logged closure prices of assets 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

$$y_t = \sqrt{h_t}\varepsilon_t, \qquad h_{t+1} = \omega + \beta h_t + \alpha(y_t^2 - h_t), \tag{10}$$

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 model in (10) is a special case of GAS models when the error distribution is standard normal and the score is rescaled with the inverse of the Fisher information, see the discussion in Creal et al. (2013).

In the following we focus on the coefficient α 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 β, we may question whether the constancy of α 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 α can be appropriate when the current level of volatility is appropriate. But in a more turbulent period, the α 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 α is allowed to vary over time. We refer to this extended GARCH model as the accelerated GARCH model, or the aGARCH model.

We introduce a time-varying coefficient for α_t and we express the GARCH updating equation in (10) in its innovation form, that is,

$$h_{t+1} = \omega + \beta\, h_t + \alpha_t\, h_t\, (\varepsilon_t^2 - 1), \tag{11}$$

where we have replaced y_t^2 by $h_t\varepsilon_t^2$ as implied by the model for y_t. The time-varying coefficient α_t is specified using the accelerating GAS framework introduced in Section 2 as

$$\alpha_t = \beta\, \text{logit}(f_{t+1}), \quad f_{t+1} = \omega_f + \beta_f f_t + \alpha_f(\varepsilon_t^2 - 1)(\varepsilon_{t-1}^2 - 1), \tag{12}$$

where $\text{logit}(\cdot)$ is the logistic function such that $\text{logit}(a) = \exp(a)/(1+\exp(a))$ for any $a \in \mathbb{R}$.

The aGARCH model (11) and (12) is a special case of the aGAS model (1) and (2) when considering λ_t as the time-varying variance of the Gaussian conditional density $p(y_t|\lambda_t;\theta)$. The time-varying process for α_t is driven by the product of current and lagged volatility innovations. The product of the contemporaneous and lagged standardized volatility innovations is treated as indicative of whether or not α_t needs to change more quickly or slowly. When two consecutive volatility innovations 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 parameters α_f and β_f determine the relative importance of past products of scaled volatility innovations. For instance, if β_f is close to 1 and α_f close to zero, then α_t is driven by all the accumulated history of products of volatility innovations. Instead, if β_f is close to zero, then α_t is driven only by the most recent products of volatility innovations. The parameter ω_f determines the average level of α_t. In particular, the unconditional expectation of f_t is given by $\omega_f/(1-\beta_f)$. The resulting model is a straightforward extension of the GARCH model with the addition of two coefficients only, α_f and β_f, since ω_f is effectively replacing the static α coefficient in the GARCH model (10). In case $\alpha_f = \beta_f = 0$, the aGARCH model reduces to the standard GARCH model.

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 (1) with $\lambda_t = h_t$ 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 $y_t^2 - h_t$. 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

$$\begin{aligned} y_t &= \sqrt{h_t}\varepsilon_t, \qquad & h_{t+1} &= \omega + \beta h_t + \alpha_t h_t s_{h,t}, \\ \alpha_t &= \beta\, \text{logit}(f_{t+1}), \qquad & f_{t+1} &= \omega_f + \beta_f f_t + \alpha_f s_{h,t} s_{h,t-1}, \end{aligned}$$

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 $s_{h,t}$ has the following expression

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

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.

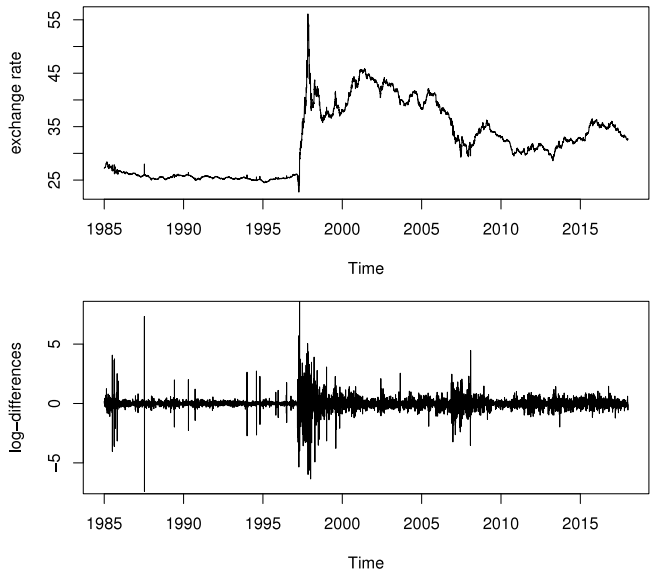

Fig. 3. The first plot shows the exchange rate between the Thai Bat and the US dollar from January 1985 to December 2017. The second plot shows the log differences of the exchange rate.

Table 2
Parameter estimates for all considered models with their standard errors in brackets. The last three columns contain respectively the maximized log-likelihood value, the BIC and the AIC criteria. The parameters δ and δ_f are given by $\delta = \omega/(1-\beta)$ and $\delta_f = \omega_f/(1-\beta_f)$.

	δ	β	δ_f	β_f	α_f	ν	log-lik	BIC	AIC
at-GAS	0.170	0.996	−1.843	0.939	0.061	3.029	−438.1	**930.3**	**888.2**
	(0.053)	(0.001)	(0.123)	(0.026)	(0.020)	(0.113)			
t-GAS	0.264	0.997	−1.550	–	–	2.989	−449.5	935.1	907.1
	(0.073)	(0.001)	(0.085)			(0.110)			
at-GARCH	0.006	0.910	−2.211	0.740	0.004	3.016	−474.1	1002.3	960.2
	(0.001)	(0.007)	(0.066)	(0.027)	(0.001)	(0.129)			
t-GARCH	0.007	0.876	−2.121	–	–	2.998	−567.4	1170.9	1142.8
	(0.001)	(0.008)	(0.080)			(0.116)			
a-GARCH	0.758	0.992	−1.436	0.902	0.002	–	−2457.3	4959.6	4924.6
	(0.117)	(0.001)	(0.068)	(0.012)	(0.000)				
GARCH	0.697	0.990	−1.333	–	–	–	−2517.5	5062.0	5041.0
	(0.128)	(0.002)	(0.069)						

5.2. Application for an asian exchange rate time series

We illustrate how the accelerating parameter α_t can enhance the performance of GARCH models in an empirical application for an Asian exchange rate. Exchange rates play an important role in understanding macroeconomic policies. It is well known that most Asian exchange rates were in fact pegged to the US dollar before the Asian crisis in 1997; see for instance Patnaik et al. (2011), Pan et al. (2007) and Frankel and Wei (2007). After the crisis hit, most Asian countries gave up the peg with the US dollar and the regime of their exchange rates suddenly changed not only in the level but also in volatility. A similar shift of regime, though with a milder impact, is also encountered in the financial crisis in 2008. In the following we show that accelerating GARCH models can better capture these sudden changes in volatility regimes. In particular, we consider the daily exchange rate between the Thai Bat and the US dollar from January 1985 to December 2017. A similar behavior can be encountered in several other Asian exchange rates; see Patnaik et al. (2011). Fig. 3 presents the time series of exchange rates and the log differences. The rapidly evolving break before July 1997 in the exchange rate series, both for level and volatility, can be clearly observed.

Table 2 reports the parameter estimates of the GARCH, t-GARCH and t-GAS models and their accelerating counterparts. The best model in terms of AIC and BIC is the at-GAS model. The reported results highlight that the accelerating parameter α_t is useful to describe some of the dynamic features in the time series. The t-GAS models perform overall better than the GARCH models in terms of log-likelihood, AIC and BIC. This finding is most likely due to the provided robustness in variance updating against outliers in the time series.

Finally, Fig. 4 illustrates how the at-GAS model is capable of better adapting to the rapidly evolving change in the volatility level that occurs in the period before the Thai Bat loses the peg against the US dollar (2 July 1997). In particular,

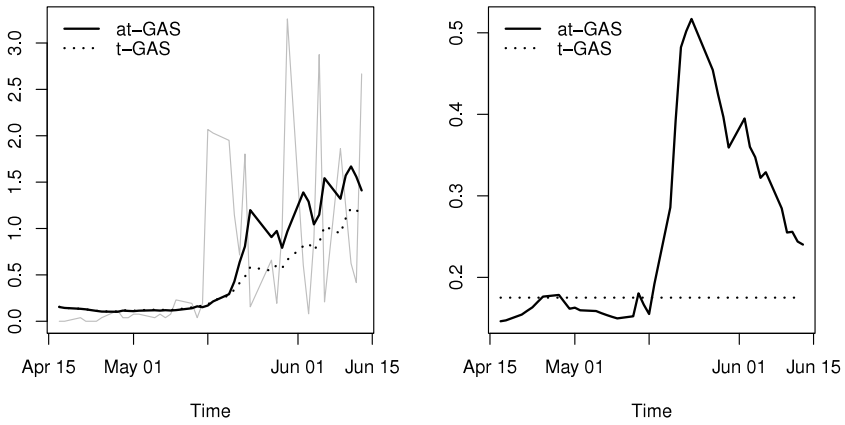

Fig. 4. The first plot shows the absolute log differences of the exchange rate and the filtered standard deviations $\sqrt{h_t}$ for the t-GAS and at-GAS models for the period between 15 April and 15 June, 1997. The second plot shows the corresponding estimated α_t for the two models.

the first plot in Fig. 4 shows the absolute log variations of the exchange rate and the filtered standard deviations from the at-GAS and the t-GAS models during the period between 15 April and 15 June, 1997. When the sudden increase in the volatility level occurs, the at-GAS filter is able to adapt to the new volatility level faster than the standard t-GAS filter. The second plot presents the estimated time-varying coefficient α_t of the at-GAS model against the estimated static coefficient of the t-GAS model. The estimate of α_t shows an increase when the change in volatility occurs and this enables the filter of the at-GAS model to adjust quicker that the one of the t-GAS model. This empirical illustration shows that the accelerated GAS models can be useful to describe sudden changes in volatility levels for GARCH and related models.

5.3. Application to all series in the S&P500 stock index

We evaluate the performance of the accelerated GARCH models through a comparison using all the stocks that are currently in the S&P500 index. Daily stock returns from January 2005 to December 2017 are considered. The series of the S&P500 that are not available since 2005 are excluded from the study. The resulting number of time series is 436. 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 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 exercise consists of comparing density forecasts of daily log-returns. The performance 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 log-score criterion delivers a consistent ranking of models in terms of KL divergence under standard regularity conditions, see Lemma 2 in Appendix B. Hence our results do not suffer from the inconsistency problem as discussed in Patton (2011) for volatility forecasts. The in-sample evaluation is based on the whole sample, whereas, the out-of-sample is based on three yearly periods: the daily observations in 2015, 2016 and 2017. For the out-of-sample evaluation, the "training sample" is based on a rolling window approach where all models are re-estimated monthly (every 20 working days).

Table 3 reports the number of series in the S&P500 index where a model outperforms the others. The degrees of freedom parameter ν of the Student-t models are estimated along with the other parameters. The results are reported separately for models with Gaussian and Student-t innovations since Student-t models have a better performance than Gaussian models for all series. This is due to the fact that the Student-t distribution can account for the heavy tails of stock return data. However, the estimation of models with a Gaussian distribution can be regarded as quasi maximum likelihood (QML). QML has the advantage of delivering consistent estimates of the conditional volatility even when the error distribution is misspecified. Instead, this is not the case when a Student-t likelihood is used; see Straumann (2005).

When we focus on models with Gaussian innovations, the GARCH and aGARCH models, we can conclude that the aGARCH model performs better than the GARCH model for the majority of the series. This improvement can only be due to the accelerating mechanism that allows fast adjustment towards a new volatility level after a break, in a similar way as we have discussed for the exchange rate series. The same argument applies when outliers are present the series; in this case, the linear filter of the GARCH model is heavily affected and it takes possibly a long period to return to its natural level. The aGARCH model helps in this respect by making the impact of the outlier to vanish faster. Also, when we consider the models with Student-t errors, the t-GARCH, at-GARCH, t-GAS and at-GAS models, we find that the accelerating mechanism gives better results for a significant proportion of the 436 daily return series, both in-sample and out-of-sample. In this setting, the improvement is mainly due to sudden changes in volatility levels because outliers

Table 3
The number and percentage of series in the S&P500 index where each model outperforms the others. The in-sample performance is measured by AIC. The out-of-sample performance is based on a log-score criterion and considered separately for the years 2015, 2016 and 2017. In the first panel, the comparison is among models with Gaussian error distributions. In the second panel, the comparison is among models with Student-t error distributions.

	In-sample		Out-of-sample					
	2005–2017		2015		2016		2017	
	No.	Pct.	No.	Pct.	No.	Pct.	No.	Pct.
GARCH	134	30.7%	233	53.4%	250	57.3%	219	50.2%
aGARCH	302	69.3%	203	46.6%	186	42.7%	217	49.8%
Total	436	100.0%	436	100.0%	436	100.0%	436	100.0%
t-GARCH	68	15.6%	118	27.1%	83	19.1%	56	12.8%
at-GARCH	31	7.1%	115	26.4%	62	14.2%	49	11.2%
t-GAS	246	56.4%	103	23.6%	202	46.3%	182	41.8%
at-GAS	91	20.9%	100	22.9%	89	20.4%	149	34.2%
Total	436	100.0%	436	100.0%	436	100.0%	436	100.0%

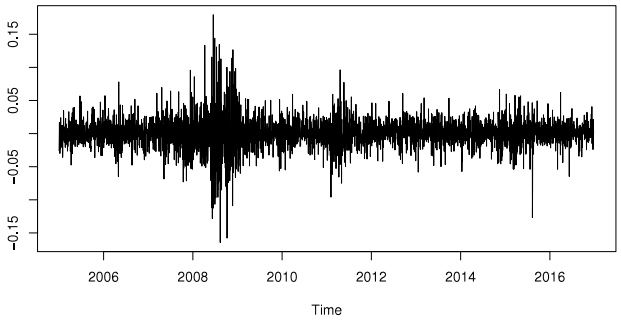

Fig. 5. Daily log returns of stock Charles Schwab Corp. from January 2005 to December 2017.

are already handled by the robust filter of t-GAS models. In particular, the impact of extreme return observations on h_t is attenuated as discussed in Creal et al. (2013). Therefore, the number series where the accelerating parameter is useful is lower than in the Gaussian case.

We finally provide a further illustration to better understand under which circumstances the accelerating mechanism delivers better results. We consider a random series from the S&P500 index for which the at-GAS outperforms the t-GAS model: the daily log returns of the stock Charles Schwab Corp. (SCHW, a bank and brokerage firm in San Francisco). The results presented below are indicative for many series in the S&P500 index. Fig. 5 shows the daily log returns of the series from January 2005 to December 2017. We observe a sudden increase in the volatility level from 2008 which is caused by the financial crisis of 2008.

The first plot in Fig. 6 presents the absolute log returns together with the filtered standard deviations from the t-GAS and the at-GAS models from July to October 2008. It is clearly visible that the filtered variance of the at-GAS model adapts more promptly than the filtered variance of the standard t-GAS model. The second plot in Fig. 6 presents the filtered estimates for the time-varying α_t and it shows a strong increase after the change in the volatility level. This increase in α_t enables the variance of the at-GAS to adjust quicker to the new volatility level.

6. Accelerated location and scale model for heavy tailed data

6.1. The model

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. An exponential link function is used for the specification of the conditional variance. The resulting model has some similarities with the stochastic volatility model of Stock and Watson (2007), see also Stock and Watson (2016) for an extension with heavy tailed distributions. 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) has considered a Student's t model with both time-varying mean and variance. The key

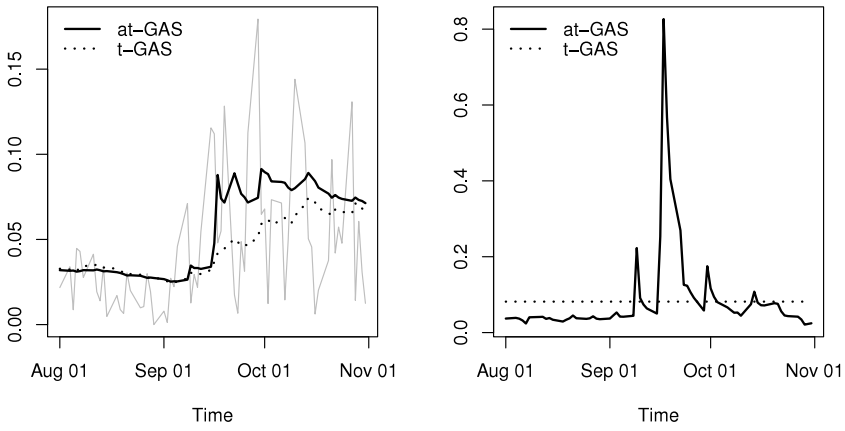

Fig. 6. The first plot shows the absolute log returns of the stock SCHW and the filtered standard deviations $\sqrt{h_t}$ for the t-GAS (dotted line) and at-GAS (solid line) models for the period between July and October 2008. The second plot shows the estimated α_t for the two corresponding models.

novelty of the model in our 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

$$y_t = \mu_t + \sigma_t \varepsilon_t, \tag{13}$$

where μ_t is the time-varying location for y_t, σ_t is the time-varying scale for y_t, and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ is an i.i.d. sequence of Student's t distributed random variables with zero mean, unit variance and ν degrees of freedom. 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. 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 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 Eq. (4); by setting $C_{f,t} = S_{\mu,t}S_{\mu,t-1}$, we obtain

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

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 update for $s_{\sigma,t}$ is the same as in the Beta-t-EGARCH model of Harvey (2013).

When the degrees of freedom of the Student's t distribution get closer to infinity, $\nu \to \infty$, the Student's t distribution approaches the standard Gaussian distribution. In this limiting case, the model (13) 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 $s_{\sigma,t} = (y_t-\mu_t)^2\sigma_t^{-2} - 1$.

6.2. Empirical illustration

In our final empirical illustration we consider the US quarterly consumer price (CP) index, which is obtained from the FRED dataset. We mention that GAS models with robust updates have already been considered in modeling US inflation,

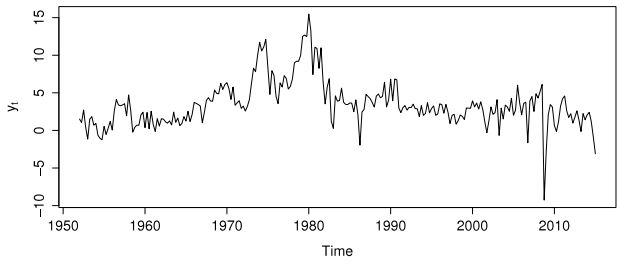

Fig. 7. Quarterly consumer price US inflation series.

Table 4
The second column indicates the specification of the model. The third column provides a possible reference for the constrained model, as a special case of our full model (13).

	Description	Reference
Model t.1	Our full model (13)	
Model t.2	$\beta_\sigma = 0$ and $\alpha_\sigma = 0$	
Model t.3	$\beta_f = 0$ and $\alpha_f = 0$	Harvey (2013, p. 139)
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 $\nu \to \infty$ for $i = 1, 2, 3, 4$	

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 Bayesian information criterion (BIC) and the Akaike information criterion (AIC). The bold values for BIC and AIC indicate their smallest values amongst all considered models. The parameters δ_f and δ_σ are given by $\delta_f = \omega_f/(1-\beta_f)$ and $\delta_\sigma = \omega_\sigma/(1-\beta_\sigma)$.

	δ_f	β_f	α_f	δ_σ	β_σ	α_σ	ν	log-lik	BIC	AIC
Model t.1	−1.518	0.967	0.258	1.055	0.861	0.215	5.571	−475.4	**989.4**	**964.7**
	(0.799)	(0.027)	(0.113)	(0.236)	(0.092)	(0.089)	(1.572)			
Model t.2	−1.493	0.914	0.294	1.182	–	–	3.826	−482.7	993.0	975.4
	(0.402)	(0.028)	(0.071)	(0.178)			(0.553)			
Model t.3	−0.468	–	–	1.080	0.869	0.163	7.583	−481.8	991.4	973.6
	(0.280)			(0.207)	(0.126)	(0.099)	(2.399)			
Model t.4	−0.305	–	–	1.111	–	–	5.639	−488.8	994.3	983.7
	(0.213)			(0.134)			(1.431)			
Model n.1	−1.366	0.969	0.182	1.169	0.937	0.088	–	−504.2	1041.6	1020.4
	(0.618)	(0.022)	(0.072)	(0.203)	(0.030)	(0.033)				
Model n.2	−0.304	0.971	0.060	1.251	–	–	–	−515.3	1052.7	1038.6
	(0.416)	(0.028)	(0.036)	(0.089)						
Model n.3	−0.231	–	–	1.213	0.939	0.054	–	−510.2	1042.5	1028.4
	(0.314)			(0.161)	(0.026)	(0.021)				
Model n.4	−0.080	–	–	1.264	–	–	–	−516.8	1044.7	1037.7
	(0.266)			(0.089)						

see for instance Delle Monache and Petrella (2017) for an application of an adaptive filter to quarterly CPI inflation. 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 Fig. 7. We consider several specifications for the aGAS model which are listed in Table 4.

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, the BIC and AIC. 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 is relevant for our US inflation series. In particular, the model with the lowest BIC and AIC is Model t.1. The reported AIC and BIC 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.

Fig. 8 presents the filtered estimates of 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 hardly affects the filtered path of μ_t . 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.

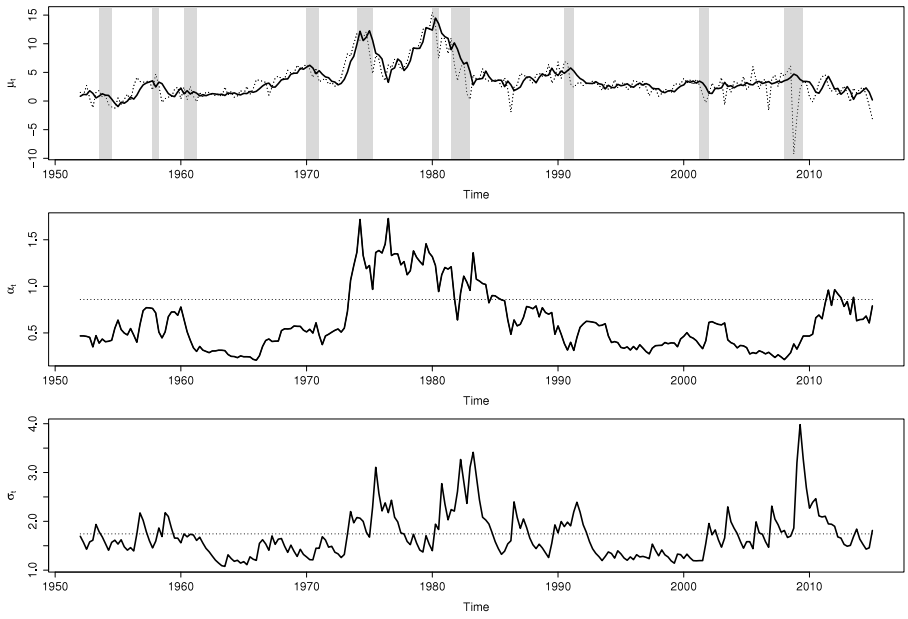

Fig. 8. The filtered estimates of the time-varying parameters in Model t.1 (solid lines): upper plot is for μ_t, with US inflation (dotted line) and NBER recession datings (gray areas) , middle plot is for α_t, with fixed α estimate from Model t.4 (dotted line, $\hat{\alpha} = \exp(-0.305/2) = 0.86$), and lower plot is for σ_t, with fixed σ estimate from Model t.4 (dotted line, $\hat{\sigma} = \exp(1.111/2) = 1.74$).

The third graph in Fig. 8 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 filtered estimate of μ_t, we present in Fig. 9 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 behavior: the first graph in Fig. 9 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 Fig. 8. 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 Fig. 9 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 Fig. 8. This allows 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 having a time-varying α_t can also be illustrated by the filtered μ_t of Model t.3 which is more noisy than the filtered μ_t of Model t.1. The two graphs in Fig. 9 show how the inclusion of the time-varying α_t allows the dynamic model to be more accurate in adapting to a changing behavior of the series. The improvements in terms of in-sample fit are also confirmed by AIC and BIC.

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 include two other models to facilitate forecast comparisons: the well-known autoregressive integrated moving average, the ARIMA(P, D, Q) model with orders $P = 4, D = 1, Q = 0$ and $P = 1, D = 1, Q = 1$. The root mean squared error (FRMSE) is computed using the last 100 observations and parameter estimation for the different model specifications is performed using a fixed rolling window. A split of the out-of-sample dataset in three sub-samples is also considered to evaluate the performance of the models in different periods. We obtain h-steps ahead forecasts, for $h = 1, 2, 3, 4$. Differences in forecast accuracy are verified by means of the Diebold 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 FRMSE as the other models against the alternative of different FRMSE.

The results are presented in Table 6. We find that either Model n.1 or Model t.1 have the best FRMSE for all forecasting horizons when we consider the entire out-of-sample window. 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 regard to the other forecasting horizons, we conclude that the differences in terms of forecasting accuracy are not significant. Finally, when we focus on the different sub-samples, we can see that results are more mixed and differences tend to be statistically insignificant.

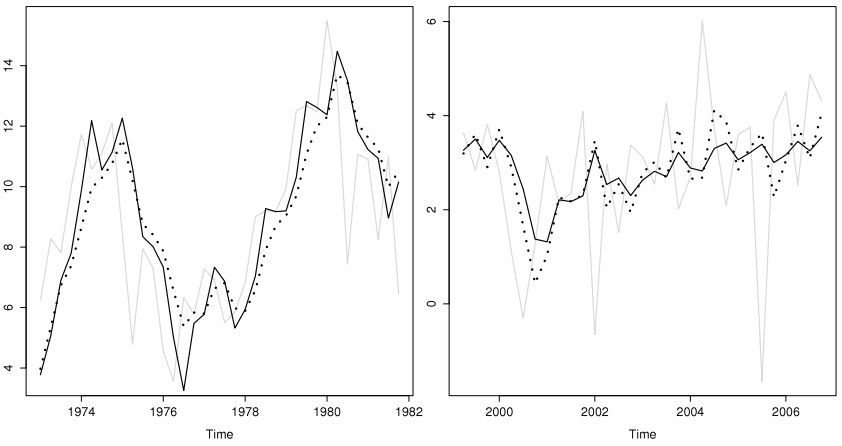

Fig. 9. The filtered estimates of μ_t from Model t.1 and Model t.3 for two different time periods. The gray line is US inflation, the dashed line is the filtered estimate μ_t from Model t.3 and the solid line is the filtered estimate μ_t from our preferred Model t.1.

Table 6
FRMSE ratio for different sub-periods of the out-of-sample window. The benchmark is Model t.1: the FRMSE of Model t.1 is the denominator of the ratios. Asterisk indicates the significance level of the Diebold Mariano test: * indicates 10% level and ** indicates 5% level.

	1990–2015 (full sample)				1990–1998				1998–2006				2006–2015			
	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$	$h=1$	$h=2$	$h=3$	$h=4$
Model t.1	1.00	**1.00**	**1.00**	**1.00**	1.00	1.00	1.00	1.00	**1.00**	**1.00**	**1.00**	**1.00**	1.00	1.00	1.00	1.00
Model t.2	1.01	1.01	1.01	1.02	1.02	1.01	0.98	**0.98****	**1.00**	**1.00**	**1.00**	1.01*	1.01	1.01	1.02	1.03**
Model t.3	1.05	1.07	1.05	1.08**	1.00	1.07	0.98	1.01	1.07	1.08	1.02	1.08*	1.06	1.07	1.07	1.08*
Model t.4	1.06*	1.08*	1.05	1.09**	1.04	1.11	1.00	1.02**	1.07*	1.08	0.98	1.07*	1.06	1.07	1.07	1.10**
Model n.1	**0.98**	**1.00**	**1.00**	1.01	0.96	1.01	**0.97**	**0.98**	1.06	1.09*	1.08*	1.10**	**0.97**	**0.99**	**0.98**	**0.98**
Model n.2	1.01	1.11	1.09	1.08**	0.97	1.05	0.98	0.99	1.07	1.08	1.03	1.09*	1.00	1.12	1.12	1.10*
Model n.3	1.02	1.07	1.05	1.07	0.98	1.05	0.97	1.00	1.12*	1.15*	1.09	1.15**	1.00	1.05	1.05	1.05
Model n.4	1.01	1.11	1.09	1.08**	0.97	1.05	0.98	0.99	1.07	1.08	1.03	1.09*	1.00	1.12	1.12	1.10*
ARMA(4,0)	1.03	1.13	1.16	1.13**	**0.92**	**0.99**	1.02	1.01	1.19	1.08	1.07	1.14	1.02	1.16	1.20	1.14**
ARMA(1,1)	0.99	1.09	1.07	1.07*	0.97	1.05	**0.97**	**0.98****	1.08	1.07	1.02	1.07	0.98	1.10	1.10	1.08

In the first sub-sample (1990–1998) there is not a clear winner and Mondel n.1 and the ARMA models have a similar performance. Instead, in the second (1998–2006) and third (2006–2015) sub-samples the best performance is given by Mondel t.1 and Mondel n.1, respectively, for all forecasting horizons. Overall, the results suggest that the accelerating models have a satisfactory forecasting performance.

7. Conclusion

We have introduced a novel class of score-driven models that allows for locally changing the weights for updating the time-varying parameters. We have provided theoretical and simulation-based evidence that these so-called accelerated GAS models can outperform corresponding GAS models with time-invariant weights for updating. We have considered two sets of illustrations: the first set is based on volatility models applied to an exchange rate series and to all daily return series that are part of the S&P 500 index; the second set is based on a time-varying location and scale model applied to an US inflation time series. For these relevant illustrations, we find that the proposed accelerating framework is capable of improving the in-sample and out-of-sample fit for GAS and related models.

Appendix A. Proofs

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

$$\begin{aligned}
\Delta_{f,t}^{t+1} &= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y, \lambda_t(\dot{f}_t)\big) u_\lambda\big(y_t, \lambda_t(f_t)\big) dy \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y_t, \lambda_t(f_t)\big)^2 dy \qquad (14) \\
&\quad -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y_t, \lambda_t(f_t)\big) \frac{\partial u_\lambda\big(\dot{y}_t, \lambda_t(\ddot{f}_t)\big)}{\partial \dot{y}_t} (y - y_t) dy \qquad (15) \\
&\quad -\int_{B(y_t,\epsilon_y)} p_t^o(y) \tilde{C}_t u_\lambda\big(y_t, \lambda_t(f_t)\big) \frac{\partial u_\lambda\big(\dot{y}_t, \lambda_t(\ddot{f}_t)\big)}{\partial \ddot{f}_t} (\dot{f}_t - f_t) dy, \qquad (16)
\end{aligned}$$

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 (14) is a.s. negative and the terms (15) and (16) 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, we first note that if $sign(f_{t+1} - f_t) = sign(s_{f,t})$ does not hold with probability 1 for any $f_t \in \mathcal{F}$, then 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. By repeated applications of the mean value theorem we obtain that

$$\begin{aligned}
\Delta_{f,t}^{t+1} &= \int_{B(y_t,\epsilon_y)} p_t^o(y) \log \frac{p(y|\lambda_t(f_t))}{p(y|\lambda_t(f_{t+1}))} dy \;=\; \int_{B(y_t,\epsilon_y)} p_t^o(y) \frac{\partial \log p(y|\lambda_t(\dot{f}_t))}{\partial \dot{f}_t} (f_t - f_{t+1}) dy \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \alpha_f S_{\lambda,t-1} u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y, \lambda_t(\dot{f}_t)\big) (f_{t+1} - f_t) dy \\
&= -\int_{B(y_t,\epsilon_y)} p_t^o(y) \alpha_f S_{\lambda,t-1} u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y_t, \lambda_t(f_t)\big) (f_{t+1} - f_t) dy \qquad (17) \\
&\quad -\int_{B(y_t,\epsilon_y)} p_t^o(y) \alpha_f S_{\lambda,t-1} u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) \frac{\partial u_\lambda\big(\dot{y}_t, \lambda_t(\ddot{f}_t)\big)}{\partial \dot{y}_t} (f_{t+1} - f_t)(y - y_t) dy \qquad (18) \\
&\quad -\int_{B(y_t,\epsilon_y)} p_t^o(y) \alpha_f S_{\lambda,t-1} u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) \frac{\partial u_\lambda\big(\dot{y}_t, \lambda_t(\ddot{f}_t)\big)}{\partial \ddot{f}_t} (f_{t+1} - f_t)(\dot{f}_t - f_t) dy, \qquad (19)
\end{aligned}$$

where $\dot{f}_t$ is a point between f_t and f_{t+1}, $\ddot{f}_t$ is a point between $\dot{f}_t$ and f_t, and $\dot{y}_t$ is a point between y and y_t. The term in (17) is positive with positive probability since $sign(f_{t+1} - f_t) \neq sign(s_{f,t})$ with positive probability and the factor $u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y_t, \lambda_t(f_t)\big)$ has the same sign as the score $s_{f,t}$. Therefore, $\Delta_{f,t}^{t+1} > 0$ holds with positive probability since the terms (18) and (19) 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. 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

$$\begin{aligned}
\Delta_{\lambda,t+1}^{t+1} - \Delta_{\lambda,t+1}^{t} &= \int_{B(y_t,\epsilon_y)} p_t^o(y) \log \frac{p(y|\lambda_{t+1}(f_t))}{p(y|\lambda_{t+1}(f_{t+1}))} dy \\
&= \int_{B(y_t,\epsilon_y)} p_t^o(y) \frac{\partial \log p(y|\lambda_{t+1}(\dot{f}_t))}{\partial \dot{f}_t} (f_t - f_{t+1}) dy
\end{aligned}$$

$$= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\alpha_f C_{f,t} S_{\lambda,t-1} u_\lambda\big(y_t, \lambda_t(f_t)\big)^2 u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y, \lambda_{t+1}(\dot{f}_t)\big) dy$$
$$= -\int_{B(y_t,\epsilon_y)} p_t^o(y)\tilde{C}_t u_\lambda\big(y_{t-1}, \lambda_{t-1}\big) u_\lambda\big(y, \lambda_t(\dot{f}_t)\big) dy,$$

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

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

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

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

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

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 $\Delta^{t+1}_{\lambda,t+1} - \Delta^t_{\lambda,t+1}$ is negative. □

Appendix B. Consistent model ranking

We have that p_t^o denotes the true conditional density of y_t given its past. In the following lemma we consider two models, Model 1 and Model 2. We suppose that these models are indexed by the parameter vectors θ^1 and θ^2, respectively. Further, we let $p_t^1(\theta^1)$ and $p_t^2(\theta^2)$ denote the conditional density of y_t as implied by models 1 and 2, under the parameters θ^1 and θ^2, respectively. Additionally, we let θ_0^1 and θ_0^2 denote the pseudo-true parameters of models 1 and 2, i.e. these are the best parameters of each model in KL divergence.

We assume that the model parameters are estimated using a sample of size T and the forecasting performance is evaluated using a different sample of size N. We suppose that the pseudo-true parameters can be consistently estimated by the ML estimates $\hat{\theta}_T^1$ and $\hat{\theta}_T^2$. Finally, we have

$$\ell_N^1(\hat{\theta}_T^1) = N^{-1}\sum_{t=1}^N \log p_t^1(\hat{\theta}_T^1), \qquad \ell_N^2(\hat{\theta}_T^2) = N^{-1}\sum_{t=1}^N \log p_t^2(\hat{\theta}_T^2),$$

which are the logarithmic scoring rules evaluated at the corresponding ML estimates $\hat{\theta}_T^1$ and $\hat{\theta}_T^2$, respectively.

Lemma 2 (*Consistent Model Ranking*)**.** *Let a uniform law of large numbers apply to the logarithmic scoring rules as the sample N diverges*

(i) $\sup_{\theta^1\in\Theta^1} |\ell_N^1(\theta^1) - \mathbb{E}\log p_t^1(\theta^1)| \xrightarrow{a.s.} 0$ *as* $N \to \infty$*;*
(ii) $\sup_{\theta^2\in\Theta^2} |\ell_N^2(\theta^2) - \mathbb{E}\log p_t^2(\theta^2)| \xrightarrow{a.s.} 0$ *as* $N \to \infty$*;*

where the limit functions $\mathbb{E}\log p_t^1(\cdot)$ *and* $\mathbb{E}\log p_t^2(\cdot)$ *are continuous. Furthermore, let the ML estimates be consistent as the estimation sample T diverges to infinity*

(iii) $\hat{\theta}_T^1 \xrightarrow{a.s.} \theta_0^1$ *as* $T \to \infty$*;*
(iv) $\hat{\theta}_T^2 \xrightarrow{a.s.} \theta_0^2$ *as* $T \to \infty$*.*

Then the difference in logarithmic scoring rules converges almost surely to the difference in KL divergences between the two models,

$$\lim_{T,N\to\infty}\Big(\ell_N^1(\hat{\theta}_T^1) - \ell_N^2(\hat{\theta}_T^2)\Big) = KL\Big(p_t^o, p_t^2(\theta_0^2)\Big) - KL\Big(p_t^o, p_t^1(\theta_0^1)\Big).$$

Proof of Lemma 2. Given the uniform convergence of the logarithmic scoring rules and given the consistency of the ML estimator, the result follows immediately by letting T and N diverge to infinity either jointly or sequentially

$$\begin{aligned}\lim_{T,N\to\infty}\left(\ell_N^1(\hat{\theta}_T^1)-\ell_N^2(\hat{\theta}_T^2)\right) &= \mathbb{E}\log p_t^1(\theta_0^1)-\mathbb{E}\log p_t^2(\theta_0^2) = \mathbb{E}\log\left(\frac{p_t^o}{p_t^2(\theta_0^2)}\right)-\mathbb{E}\log\left(\frac{p_t^o}{p_t^1(\theta_0^1)}\right)\\ &= \mathrm{KL}\Big(p_t^o,p_t^2(\theta_0^2)\Big)-\mathrm{KL}\Big(p_t^o,p_t^1(\theta_0^1)\Big). \quad \square\end{aligned}$$

References

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

Andres, P., 2014. Computation of maximum likelihood estimates for score driven models for positive valued observations. Comput. Statist. Data Anal. 76, 34–43.

Blasques, F., Koopman, S.J., 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. J. Econometrics 31, 307–327.

Creal, D., Koopman, S.J., Lucas, A., 2011. A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. J. Bus. Econom. Statist. 29 (4), 552–563.

Creal, D., Koopman, S.J., Lucas, A., 2013. Generalized autoregressive score models with applications. J. Appl. Econometrics 28 (5), 777–795.

Creal, D., Schwaab, B., Koopman, S.J., Lucas, A., 2014. Observation driven mixed-measurement dynamic factor models with an application to credit risk. Rev. Econ. Stat. 96, 898–915.

De Lira Salvatierra, I., Patton, A.J., 2015. Dynamic copula models and high frequency data. J. Empir. Financ. 30, 120–135.

Delle Monache, D., Petrella, I., 2017. Adaptive models and heavy tails with an application to inflation forecasting. Int. J. Forecast. 33 (2), 482–501.

Diebold, F.X., Mariano, R.S., 1995. Comparing predictive accuracy. J. Bus. Econom. Statist. 13, 253–265.

Engle, R.F., 1982. Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. Econometrica 50, 987–1007.

Frankel, J.A., Wei, S.-J., 2007. Assessing china's exchange rate regime. Econ. Policy 22 (51), 576–627.

Geweke, J., Amisano, G., 2011. Optimal prediction pools. J. Econometrics 164, 130–141.

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

Harvey, A., Luati, A., 2014. Filtering with heavy tails. J. Amer. Statist. Assoc. 109, 1112–1122.

Hjort, N.L., Jones, M.C., 1996. Locally parametric nonparametric density estimation. Ann. Statist. 24 (4), 1433–1854.

Jaynes, E.T., 1957. Information theory and statistical mechanics. Phys. Rev. 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., Leibler, R.A., 1951. On information and sufficiency. Ann. Math. Stat. 22 (1), 79–86.

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

Oh, D.H., Patton, A.J., 2017. Time-varying systemic risk: evidence from a dynamic copula model of cds spreads. J. Bus. Econom. Statist. (forthcoming).

Pan, M.-S., Fok, R.C.-W., Liu, Y.A., 2007. Dynamic linkages between exchange rates and stock prices: evidence from east asian markets. Int. Rev. Econ. Finance 16 (4), 503–520.

Patnaik, I., Shah, A., Sethy, A., Balasubramaniam, V., 2011. The exchange rate regime in asia: from crisis to crisis. Int. Rev. Econ. Finance 20 (1), 32–43.

Patton, A.J., 2011. Volatility forecast comparison using imperfect volatility proxies. J. Econometrics 160 (1), 246–256.

Stock, Watson, 2007. Why has u.s. inflation become harder to forecast? journal of money. Credit Bank. 39, 3–33.

Stock, J.H., Watson, M.W., 2016. Core inflation and trend inflation. Rev. Econ. Stat. 98 (4), 770–784.

Straumann, D., 2005. Estimation in Conditionally Heteroschedastic Time Series Models. Springer, New York, p. 181.

Ullah, A., 1996. Entropy, divergence and distance measures with econometric applications. J. Statist. Plann. Inference 69, 137–162.

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