VU Research Portal

Finite Sample Optimality of Score-Driven Volatility Models

Blasques, Francisco; Lucas, André; van Vlodrop, Andries C.

published in

Econometrics and Statistics
2021

DOI (link to publisher)

[10.1016/j.ecosta.2020.03.010](https://doi.org/10.1016/j.ecosta.2020.03.010)

document version

Publisher's PDF, also known as Version of record

document license

Article 25fa Dutch Copyright Act

[Link to publication in VU Research Portal](#)

citation for published version (APA)

Blasques, F., Lucas, A., & van Vlodrop, A. C. (2021). Finite Sample Optimality of Score-Driven Volatility Models: Some Monte Carlo Evidence. *Econometrics and Statistics*, 19, 47-57.
<https://doi.org/10.1016/j.ecosta.2020.03.010>

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal ?

Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

E-mail address:

vuresearchportal.ub@vu.nl

Finite Sample Optimality of Score-Driven Volatility Models: Some Monte Carlo Evidence

Francisco Blasques^a, André Lucas^a, Andries C. van Vlodrop^{a,*}

Vrije Universiteit Amsterdam and Tinbergen Institute Netherlands

ARTICLE INFO

Article history:

Received 4 December 2017

Revised 25 March 2020

Accepted 29 March 2020

Available online 28 May 2020

Keywords:

volatility models

score-driven dynamics

finite samples

Kullback-Leibler divergence

optimality

ABSTRACT

Optimality properties are studied in finite samples for time-varying volatility models driven by the score of the predictive likelihood function. Available optimality results for this class of models suffer from two drawbacks. First, they are only asymptotically valid when evaluated at the pseudo-true parameter. Second, they only provide an optimality result 'on average' and do not provide conditions under which such optimality prevails. Using finite sample Monte Carlo experiments, it is shown that score-driven volatility models have optimality properties when they matter most. Score-driven models perform best when the data are fat-tailed and robustness is important. Moreover, they perform better when filtered volatilities differ most across alternative models, such as in periods of financial distress. These simulation results are supplemented by an empirical application based on U.S. stock returns.

© 2020 EcoSta Econometrics and Statistics. Published by Elsevier B.V. All rights reserved.

1. Introduction

The class of score-driven models introduced in Creal et al. (2011, 2013) and Harvey (2013) has gained considerable popularity in the recent statistical literature. Score-driven models are typically appreciated for their robustness properties since the models flexibly adapt themselves to the distribution of the innovations. Despite being relatively new, a wide range of applications of score-driven models is already available in the literature; see e.g. Harvey and Luati (2014) for location and scale for fat-tailed data; Creal et al. (2014) for mixed measurement dynamic factor models; Opschoor et al. (2017) for a multivariate dynamic covariance matrix; Blasques et al. (2016) and Catania and Billé (2017) for applications to spatial models; and Janus et al. (2014), Lucas et al. (2014, 2017), Salvatierra and Patton (2015), and Oh and Patton (2017) for recent applications to score-driven copula models.

Blasques et al. (2015) – from now on referred to as BKL2015 – provide one of the theoretical motivations for score-driven models. They highlight the unique properties of score-driven models by showing that these models are optimal in an information theoretic sense. Their results, however, are only valid *asymptotically* when the model is evaluated at the limiting pseudo-true parameter. Furthermore, their supporting Monte Carlo findings only illustrate that score-driven models are optimal *on average*; see also the corresponding theoretical results in Creal et al. (2018). Therefore, little is known about the finite sample optimality properties of score-driven models.

In this paper, we try to fill this gap by answering two main questions using a Monte Carlo study. First, we study how score-driven models perform when analyzed at a (finite-sample) parameter estimate rather than at the pseudo-true param-

* Corresponding author. Tel.: Andries van Vlodrop, SBE/FIN Vrije Universiteit, De Boelelaan 1105, 1081HV Amsterdam, Netherlands Tel.: +31 20 598 3739. E-mail addresses: f.blasques@vu.nl (F. Blasques), a.lucas@vu.nl (A. Lucas), a.c.van.vlodrop@vu.nl (A.C. van Vlodrop).

eter. Second, we investigate whether the better performance of a score-driven volatility model depends on the properties of the data and the level of the time-varying parameter itself: does the model perform better during either periods of high (or low) volatility, and in the presence or absence of outliers and fat tails? As our benchmark we use the well-known Generalized Autoregressive Conditional Heteroskedasticity (GARCH) model of Engle (1982) and Bollerslev (1986), both with normally and Student's t distributed error terms. We find that the score-driven volatility model not only outperforms the GARCH models asymptotically, but also in finite samples when evaluated at estimated parameter values, despite the considerable variation in these estimates. We also find that the score-driven model performs better precisely when it matters most, i.e., when volatility is high, data are fat-tailed, and robustness is most important. Finally, we find empirically that score-driven volatility models perform best in periods of financial distress. This finding is particularly useful in practice since it indicates that the optimality results apply when we need them most.

The properties of score-driven models in this finite sample context are of course shared by alternative models that possess similar news impact curves, i.e., where future volatility reacts similarly to current observations. A prime example of this is the BIP-GARCH model of Laurent et al. (2016), which is based on an outlier robust, bounded influence perspective. As the impact curve of this model is quite similar to that of a Student's t based score-driven volatility model, we also expect the BIP-GARCH model and similar models to do better in an information theoretic sense than the standard GARCH model if the true errors are fat-tailed. Our paper may therefore also provide a further theoretical motivation for such alternative specifications.

Section 2 provides a further motivation and establishes the main concepts. Section 3 studies the finite sample properties of score-driven volatility models evaluated at estimated parameter values. Section 4 characterizes the volatility scenarios for which score-driven models perform best. Section 5 applies the results to U.S. stock returns. Section 6 concludes.

2. Score-driven volatility models and optimality

Let $y_1, y_2, \dots, y_T$ be a sample of financial returns with true time-varying conditional density $p_t^0(y_t | Y_{t-1})$, where $Y_{t-1} = \{y_1, \dots, y_{t-1}\}$. While the true unknown conditional density p_t^0 may be potentially complex and nonparametric (i.e. known only to belong to an infinite dimensional class of distributions), we attempt to describe the distribution of the data using a parametric model that generates a much smaller and simpler class of conditional densities. In particular, we consider observation-driven models with a time-varying volatility component,

$$y_t = \sigma_t \varepsilon_t, \quad \sigma_{t+1}^2 = h(\sigma_t^2, y_t; \theta), \quad (1)$$

where $\{\varepsilon_t\}_{t \in \mathbb{Z}}$ is a sequence of independent and identically distributed innovations with marginal densities $p(\varepsilon_t)$ in which ε_t has zero mean and unit variance, $h(\cdot)$ is some measurable function, and θ is a vector of parameters. The function h in (1) depends on one lag of σ_t^2 and y_t only, but can easily be generalized to include more lags. The key feature of (1) is that σ_t^2 depends on past values of y_t only, which facilitates estimation of θ by maximum likelihood (ML) using a standard prediction error decomposition. Let $\hat{\theta}_T$ denote the ML estimator (MLE) based on a sample of size T .

The class of models in (1) includes many well-known models such as the ARCH model of Engle (1982), the GARCH model of Bollerslev (1986), the EGARCH model of Nelson (1991), the TGARCH model of Zakoian (1994), the QGARCH model of Sentana (1995), and many more. Each of these models implies a parametric model density $p(y_t | \sigma_t^2; \theta_T) = p(y_t | Y_{t-1}; \theta_T)$ that aims to approximate the true unknown conditional density $p_t^0(y_t | Y_{t-1})$. The typical concrete objective is to minimize a discrepancy $D(\cdot, \cdot)$ between the two distributions,

$$D\left(p_t^0(\cdot | Y_{t-1}), p(\cdot | Y_{t-1}; \hat{\theta}_T)\right). \quad (2)$$

Different models with different specifications for h in (1) can be compared in terms of their ability to reduce the discrepancy in (2).

BKL2015 showed that if one uses the familiar Kullback-Leibler (KL) divergence as the discrepancy measure in (2), then we can identify a class of observation-driven models that possesses unique fundamental approximation properties that distinguish them from other models. This class of models is called score-driven models.

Score-driven models were introduced by Creal et al. (2011, 2013) and Harvey (2013). They define the conditional density of y_t given Y_{t-1} implicitly through a time-varying parameter f_t that is updated using the derivative of the time t log-likelihood function, also known as *the score*:

$$y_t \sim p(y_t | Y_{t-1}; \theta), \quad f_{t+1} = \omega + \alpha s_t + \beta f_t \quad (3)$$

$$s_t := S_t \cdot \frac{\partial \log p(y_t | f_t; \theta)}{\partial f},$$

where S_t is a scaling matrix that is known given the data up to time $t - 1$. The form of the conditional density $p(y_t | Y_{t-1}; \theta)$ is assumed such that all its dependence on past data operates through the time-varying parameter f_t , i.e., $p(y_t | Y_{t-1}; \theta) = p(y_t | f_t; \theta)$. This suffices for our current purpose of studying volatility models, but can be generalized to $p(y_t | Y_{t-1}, f_t; \theta)$ in case of for instance an autoregressive model with time-varying coefficient f_t . Also the transition equation for f_{t+1} can be generalized further by including more lags of f_t and/or s_t , or by scaling the score by an estimate of its local curvature via the

matrix S_t . A typical concrete example of this model that we use throughout this paper is based on a setting with $f_t = \sigma_t^2$ and assuming a Student's t distribution with $\tau > 2$ degrees of freedom and unit variance for ε_t in (1). Scaling the score by a factor proportional to the time t inverse conditional Fisher information matrix to account for the curvature, we then obtain a non-linear recurrence $h(\cdot)$ in (1) of the form

$$f_{t+1} = \omega + \alpha (w_t y_t^2 - f_t) + \beta f_t, \quad w_t = \frac{(1 + \tau^{-1})}{1 - 2\tau^{-1} + \tau^{-1} y_t^2 / f_t}, \quad (4)$$

where $\theta = (\omega, \alpha, \beta, \tau)$; see Creal et al. (2013) for further details. To ensure positivity of σ_t^2 at all times, we assume $\sigma_1^2 > 0$, $\omega > 0$, and $\beta > \alpha > 0$. If ε_t is normally distributed, $\tau^{-1} = 0$ and $w_t = 1$, such that (4) reduces to the standard GARCH(1,1) recursion $f_{t+1} = \omega + \alpha (y_t^2 - f_t) + \beta f_t = \omega + \alpha y_t^2 + (\beta - \alpha) f_t$. If $\tau^{-1} > 0$, however, ε_t is modeled as a fat-tailed Student's t distribution and the recursion in (4) is fundamentally different from the GARCH(1,1) model with Student's t distributed innovations. In particular, outlying observations y_t^2/f_t are downweighted via w_t , thus lending the score-driven model a robustness feature. Intuitively, the score-driven dynamics account for the fact that ε_t is fat-tailed. A large value of y_t thus need not automatically be attributed to an increase in volatility, but can also be due to the fat-tailed nature of the innovation distribution. This results in a mitigated impact of such observations on the volatility dynamics in the score-driven approach.

Let θ_0^* denote the MLE's pseudo-true parameter, i.e., the probability limit of the MLE for θ in (1) under the true (unknown) distribution $p_t^0(y_t | Y_{t-1})$. Then BKL2015 derive an optimality property for score-driven models like that in Equation (4). They show that under appropriate regularity conditions an observation-driven time-varying parameter model improves the local KL divergence from time t to $t+1$ if and only if the volatility filter behaves like a score-driven filter. The results in BKL2015 are quite limited, not only because they are local in f_t and local in y_t , but also because they hold only in areas where the true conditional density of y_t is large relative to the score model's conditional density. In this sense, BKL2015 only show that, under certain restrictive conditions, a small score update from f_t to f_{t+1} will increase the conditional density of the model in a neighborhood of y_t . Creal et al. (2018) generalize BKL2015 to estimators based on moment conditions. Following their results, we obtain that *in expectation* small score-based steps improve the KL divergence, i.e.,

$$\mathbb{E} \left[D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_{t+1}; \theta_0^*) \right) | Y_{t-1} \right] < D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_t; \theta_0^*) \right), \quad (5)$$

if and only if f_t is updated in the direction suggested by the score s_t as in (3), where $D(\cdot, \cdot)$ denotes the KL divergence, and where the conditional expectation $\mathbb{E}[\cdot | Y_{t-1}]$ is taken with respect to the new observation y_t , defining f_{t+1} via Equation (4). This improvement in KL divergence holds even if the model density $p(\cdot | f_t; \theta)$ is severely mis-specified. To see the intuition for this result, one can take a first-order approximation of the KL divergence around f_t , omit higher order terms, and obtain

$$\mathbb{E} \left[D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_{t+1}; \theta_0^*) \right) - D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_t; \theta_0^*) \right) | Y_{t-1} \right] \quad (6)$$

$$\begin{aligned} &\approx -\mathbb{E} \left[\int \frac{\partial p(z | f_t; \theta_0^*)}{\partial f_t} (f_{t+1} - f_t) p_t^0(z | Y_{t-1}) dz | Y_{t-1} \right] \\ &\approx -\int \frac{\partial p(z | f_t; \theta_0^*)}{\partial f_t} p_t^0(z | Y_{t-1}) dz \cdot \mathbb{E}[f_{t+1} - f_t | Y_{t-1}] \\ &\approx -\alpha \cdot \left(\int \frac{\partial p(z | f_t; \theta_0^*)}{\partial f_t} p_t^0(z | Y_{t-1}) dz \right)^2 < 0, \end{aligned} \quad (7)$$

where the last step uses $\omega = 0$, $\beta = 1$, and $\alpha > 0$ small in (3), or small so-called Newton score steps; see also Creal et al. (2018) and BKL2015.

The theoretical optimality results in BKL2015 are supported by a Monte Carlo study for the score-driven volatility model in (4). These simulations show that the score-driven Student's t based volatility model outperforms the popular GARCH and t-GARCH specifications in terms of global expected KL divergence. This suggests that a score-driven approach can lead to a better description of the conditional data density, globally, and over time. The results reported in BKL2015, however, are subject to two important limitations. First, the model comparisons are performed by evaluating each parametric model at its (asymptotically) best possible parameter, namely, the *pseudo-true* parameter θ_0^* . Moreover, the result in (7) is valid for small so-called 'Newton' steps only, i.e., $\omega = 0$, $\beta = 1$, and α is small and positive. There is no reason to assume that the optimality results continue to hold if the models are evaluated at the estimated parameter $\hat{\theta}_T$ in a finite sample. To illustrate the potential extent of this problem, consider Fig. 1. The figure shows the results of an extensive simulation experiment where we generated $T = 1,000$ data points from a stochastic volatility data generating process (DGP),

$$y_t = \sigma_t \varepsilon_t = \sqrt{f_t} \varepsilon_t, \quad \log f_{t+1} = a + b \log f_t + v_t, \quad (8)$$

where $\varepsilon_t \sim t(\tau)$, i.e., ε_t follows a Student's t distribution with unit variance and τ degrees of freedom, and $v_t \sim N(0, \sigma_v^2)$ is serially independent and independent from ε_t . For this stochastic volatility DGP we took the same parameterization as in BKL2015, $(a, b, \sigma_v) = (0.00, 0.98, 0.065)$, and considered $\tau \in \{3, 4, 5, 7, 9\}$. For each generated series, we estimated $\theta =$

Fig. 1. Estimated α parameters of the Student's t score-driven volatility model (t-GAS) against the standard t-GARCH model. Pseudo-true values are plotted as a circle with plus.

$(\omega, \alpha, \beta, \tau)$ by the MLE $\hat{\theta}_T$ for both the score-driven model (t-GAS) and the t-GARCH model. We repeated the process $N = 1,000$ times for each DGP. This simulation setup forms the basis for our further analysis.

Fig. 1 presents the results for $\hat{\alpha}_T$ for both models for the case $\tau = 5$. Results for the other parameters ($\hat{\omega}_T, \hat{\beta}_T, \hat{\tau}_T$) and for other degrees of freedom (τ) for the DGP reveal a similar message. The figure also plots the pseudo-true values α_0^* for both models as the circle with a plus around the point (0.020, 0.015), so that α_0^* is slightly higher for the t-GAS model. We clearly see that there is substantial variability in the MLE point estimates obtained from a t-GARCH and a t-GAS model. In particular, the point estimates in finite samples may differ substantially from the asymptotic pseudo-true value as used in BKL2015. Sometimes the finite sample estimate of the t-GARCH model is closer to its pseudo-true value, while at other times the t-GAS estimate is closer. It is therefore difficult to say from the figure whether the asymptotic optimality results carry over to the finite sample setting, either on average or conditionally. In Section 3 we therefore demonstrate that the score-driven model still outperforms the standard GARCH and the Student's t based GARCH models in a KL divergence sense in our finite sample Monte Carlo study.

A second drawback of the simulation results in BKL2015 is that they only show that the score-driven model outperforms its competitors *on average*. It is not clear whether the outperformance is due to the less interesting low-volatility periods, or to the more interesting high-volatility periods. Particularly in periods of distress / high-volatility, we would appreciate a model that fits the true DGP better.

The top panel in Fig. 2 shows that we can obtain more articulate results on when the score-driven volatility model and the GARCH alternatives differ. The figure plots the filtered volatility paths for Google stock returns for the t-GARCH and t-GAS models. The shaded areas highlight the high-volatility periods. The bottom panel shows the difference in in-sample log predictive score between the two models. Positive values represent a higher log-likelihood contribution from the t-GAS compared to the GARCH model and thus indicate a better fit of the score-driven model. It is clear that the filtered volatilities differ mostly during high-volatility periods. Moreover, during these highly volatile periods the difference in log predictive scores are on average more positive. This suggests that the score-driven approach may be most valuable during periods when it matters most. The figure also highlights a number of outliers by means of arrows. These observations appear to distort the filtered volatilities of the t-GARCH model much more than their t-GAS counterparts, again highlighting that the asymptotic optimality properties of the score-driven approach may extend to finite samples, and may be sharpened to include statements conditional on the volatility level. We investigate this in detail in Section 4.

3. Optimality in finite samples for stochastic volatility DGPs

To investigate whether the asymptotic optimality properties of the score-driven approach characterized in BKL2015 carry over to finite samples, we conduct an extensive Monte Carlo study. In this study, we compare the standard GARCH model of Engle (1982) and Bollerslev (1986), the fat-tailed t-GARCH model of Bollerslev (1987), the t-GAS model of Equation (4) as introduced by Creal et al. (2011, 2013), and the log-GAS model introduced in Harvey (2013). The latter uses an updating equation for log volatility $f_t = \log(\sigma_t^2)$. The different models and their specifications are summarized in Table 1. For nota-

Fig. 2. The top panel shows the estimated conditional volatility paths for the t-GAS and t-GARCH models. The arrows show the time points at which outliers impact the t-GARCH paths. The shaded regions show the periods of volatility clustering. The bottom panel shows the difference in in-sample log predictive score for each observation.

Table 1
Models used in the Monte Carlo comparison

Model	Observation Eq	Innovation Density	Transition Eq
GARCH	$y_t = \sqrt{\tilde{f}_t} \epsilon_t$	$\epsilon_t \sim N(0, 1)$	$\tilde{f}_{t+1} = \omega(1 - \beta) + \alpha(y_t^2 - \tilde{f}_t) + \beta \tilde{f}_t$
t-GARCH	$y_t = \sqrt{\tilde{f}_t} \epsilon_t$	$\epsilon_t \sim t(\tilde{\tau})$	$\tilde{f}_{t+1} = \omega(1 - \beta) + \alpha(y_t^2 - \tilde{f}_t) + \beta \tilde{f}_t$
t-GAS	$y_t = \sqrt{\tilde{f}_t} \epsilon_t$	$\epsilon_t \sim t(\tilde{\tau})$	$\tilde{f}_{t+1} = \omega(1 - \beta) + \alpha s_t + \beta \tilde{f}_t$
log-GAS	$y_t = \exp(\tilde{f}_t/2) \epsilon_t$	$\epsilon_t \sim t(\tilde{\tau})$	$\tilde{f}_{t+1} = \omega(1 - \beta) + \alpha s_t + \beta \tilde{f}_t$

tional purposes, we indicate the (model) filtered volatility by $\tilde{f}_t = \tilde{f}_t(\theta)$ and distinguish it from the true DGP volatility f_t . We also use this notational convention for the degrees of freedom parameter, which is $\tilde{\tau}$ for the model, and τ for the DGP. The Monte Carlo set-up is as described for Fig. 1 in Section 2, with a stochastic volatility DGP. None of the statistical models thus coincides with the DGP.

Following BKL2015, we are foremost interested in the performance of the different models in terms of their KL divergence. We evaluate the unconditionally expected KL divergence between the true conditional density $p_t^0(\cdot | Y_{t-1})$ and the model-implied conditional density $p(y_t | f_t, \hat{\theta}_T)$, where the latter is evaluated at the point estimate $\hat{\theta}_T$,

$$\mathbb{E} \left[D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_t; \hat{\theta}_T) \right) \right] = \mathbb{E} \left[\int \left(\log p_t^0(y | Y_{t-1}) - \log p(y | f_t; \hat{\theta}_T) \right) p_t^0(y_t | Y_{t-1}) dy \right]. \quad (9)$$

Given a stationary and ergodic DGP, we can approximate the KL divergence using a sample average over a very long path of $H = 500,000$ observations,

$$\frac{1}{H} \sum_{t=1}^H \log p_t^0(y_t | Y_{t-1}) - \frac{1}{H} \sum_{t=1}^H \log p(y_t | f_t; \hat{\theta}_T), \quad (10)$$

where the sample (with H observations) used to evaluate the KL divergence is different from the sample (with $T = 1,000$ observations) used to compute the MLE $\hat{\theta}_T$. The entire procedure is now as follows. For a simulated sample of $T = 1,000$ observations, we estimate θ . Using this $\hat{\theta}_T$ and a different, long sample of H observations, we estimate the unconditionally expected KL divergence as in (10). Finally, we repeat this process $N = 1,000$ times to obtain the variability of the expected KL divergence over sampling uncertainty in $\hat{\theta}_T$. The results help us to see whether the findings of BKL2015 continue to hold in finite samples and across the sampling uncertainty in the static parameter estimates. As the average involving

Fig. 3. Divergence criteria of the GARCH, t-GARCH, and log-GAS models (from left to right) relative to the t-GAS model. The left panel displays the Kullback-Leibler (KL) divergences and the right panel displays the RMSE for the DGP with degrees of freedom parameter $\tau \in \{3, 5, 9\}$. The relative KL divergence in simulation s is defined as $(KL_s(M) - KL_s(t\text{-GAS})) / \max(KL_s(M), KL_s(t\text{-GAS}))$, with model M being the GARCH, t-GARCH, or log-GAS, and KL_s computed using Equation (10). The relative RMSE values are calculated in a similar way. The divergence criteria for the pseudo-true parameter values are shown in white circles. The sample consists of $N = 1,000$ Monte Carlo simulations for each $\tau \in \{3, 5, 9\}$. Positive values indicate the t-GAS model performs better.

$\log p_t^0(y_t | Y_{t-1})$ in (10) does not depend on the estimated parameter values, we only need to compute it once. We do so using a particle filter based on the DGP specified for the simulation.

Besides the KL divergence we also consider the pathwise average root mean squared error (RMSE). We compute the RMSE based on the filtered volatility paths as

$$\widehat{\text{RMSE}}(f, \tilde{f}) = \sqrt{\frac{1}{H} \sum_{t=1}^H \left(f_t - \tilde{f}_t(\hat{\theta}_T) \right)^2}, \quad (11)$$

where f_t denotes the true volatility parameter from the DGP, and $\tilde{f}_t(\hat{\theta}_T)$ denotes the filtered volatility evaluated at the estimated parameter value $\hat{\theta}_T$ for a specific model from Table 1. The RMSE thus measures the model discrepancy in terms of the volatility parameter only, whereas the KL divergence measures the discrepancy in terms of all distributional properties.

The results for this Monte Carlo study are depicted in Fig. 3. The left-hand panel reports the distribution of the average KL divergence and RMSE of the GARCH, t-GARCH and log-GAS models relative to the t-GAS model. The relative KL divergence (as defined in the note to Fig. 3) is approximately zero if the t-GAS model and its competitor provide a similar description of the DGP. It is positive if the t-GAS model provides a better description, and negative if the t-GAS model performs worse than its competitor. The same holds for the RMSE (right-hand panel). The maximum relative KL divergence is one, in which case the t-GAS model performs infinitely better. The results are presented as box-and-whisker plots based on the $N = 1,000$ simulated time series. There are three boxes for each of the degrees of freedom considered, namely for $\tau = 3, 5, 9$.

We see that the median relative performance in finite samples coincides with the asymptotic performance (at the pseudo-true value) studied in BKL2015. There is, however, considerable spread around this median (or asymptotic) relative performance. Still, we see that the t-GAS performs considerably better in terms of KL divergence than the GARCH model, particularly if the DGP is fat-tailed ($\tau = 3$). The median relative KL divergence is far above zero, and even accounting for the spread across $N = 1,000$ different samples, we still see that the entire box (inclusive of whiskers) is far above the horizontal axis. If the DGP has lighter tails, the t-GAS model continues to perform better, but the margin of outperformance decreases somewhat. This decrease is to be expected, as for large τ also $\hat{\tau}_T$ is large, and the t-GAS and GARCH specifications coincide more and more. The t-GAS model also has better KL divergence properties than its t-GARCH counterpart. The outperformance is more modest, but still clearly visible. Though not uniform for every sample as in the case of t-GAS versus GARCH, the KL divergence is better for t-GAS versus t-GARCH for the vast majority of samples considered, at least for $\tau = 3$ and $\tau = 5$. Finally, the performance of the t-GAS and log-GAS models are highly similar, such that we can conclude that the precise form of parameterizing the time-varying parameter matters less in this case. All findings are corroborated if we consider the relative RMSEs rather than the relative KL divergences. We note that if one estimates the t-GARCH model parameters through (Gaussian) QML as is typically done in the literature, one obtains the same estimates of ω, α, β as for the GARCH model. As a result, the RMSE in the right panel of Fig. 3 for GARCH coincides with the RMSE of the t-GARCH model estimated using Gaussian QML. We see that ML estimation of the t-GARCH model performs slightly better than Gaussian QML estimation in terms of RMSE performance.

Volatility models like that in (1) are typically used to estimate economic or financial risk. Therefore, as a final exercise we also compare the different models in direct economic terms. A well-known risk measure is Value-at-Risk, or VaR, which

Fig. 4. Unconditional deviation of $\text{VaR}_t(p)$ -break probabilities $\hat{p} - p$ for the t-GAS and t-GARCH model for nominal VaR levels $p \in (0, 0.1]$. The DGP has $\tau = 3$ or $\tau = 5$ degrees of freedom. The VaR is computed under the model density $p(y_t | f_t; \theta)$, and the VaR-break probability $\hat{p}$ is computed under the true DGP as $\hat{p} = H^{-1} \sum_{t=1}^H \mathbf{1}\{y_t < -\text{VaR}_t(p)\}$ for $H = 500,000$. The left panel displays the results for the models evaluated at the pseudo-true parameters ($\theta = \theta_0^*$, estimated based on a large sample of $T = 500,000$ observations). The right panel displays the averages over the $N = 1,000$ replications of the deviations evaluated at $\theta = \hat{\theta}_T$. Results are provided for all f_t and for high and low f_t , respectively.

is defined as a quantile of the conditional distribution $p(y_t | f_t; \hat{\theta}_T)$ for profits or returns y_t , such that larger losses than $\text{VaR}_t(p)$ only occur with probability (at most) p , i.e.,

$$\mathbb{P}[y_t < -\text{VaR}_t(p) \mid Y_{t-1}] = p.$$

We compute the $\text{VaR}_t(p)$ based on the model distribution $p(y_t | f_t; \hat{\theta}_T)$ and then evaluate its accuracy under the true DGP. Note that all models considered are mis-specified for the DGP. We approximate the true probability of a loss larger than $\text{VaR}_t(p)$ by the time series average of indicator functions over a long sample H , i.e., $\hat{p} = H^{-1} \sum_{t=1}^H \mathbf{1}\{y_t < -\text{VaR}_t(p)\}$. We do this for different nominal probability levels $p \in (0, 0.1]$ and two different degrees of freedom for the DGP, $\tau = 3, 5$. Fig. 4 presents the results.

The unconditional (All $f_t = \sigma_t^2$) VaR-break probabilities $\hat{p}$ roughly coincide with the nominal levels p for both degrees of fat-tailedness ($\tau = 3, 5$) considered, both for the t-GAS and t-GARCH model, and both at the pseudo-true parameter (estimated based on a large sample of $T = 500,000$ observations, left panel) and at the finite sample estimates (where $T = 1,000$, right panel). This is already interesting: despite the fact that both the score-driven GAS model and the GARCH model are mis-specified for the stochastic volatility DGP, they are able to reliably estimate the tail shape of the distribution by the approximating statistical model. The conditional results for high and low (true) volatility levels, however, reveal a clear difference between the performance of the t-GAS and t-GARCH model. High volatility levels refer to values of f_t above the 90th percentile of the unconditional distribution of f_t in (8). Similarly low volatility levels refer to values below the 10th percentile. Both for high and low volatility levels the difference $|\hat{p} - p|$ is closer to zero for the t-GAS model, indicating that for extreme volatility outcomes the score-driven t-GAS specification captures the tails of the distribution more accurately. This result holds at the pseudo-true value, but carries over to the finite sample setting as well. More specifically, for high volatility levels the difference $\hat{p} - p$ is higher for the t-GARCH than for the t-GAS model. This means that the t-GARCH quantile is not far enough out into the tails, implying that the t-GARCH underestimates volatility in case volatility and thus risk is high. Similarly, for low volatility levels the t-GARCH quantile is too far out into the tail and overestimates volatility and thus risk. In both cases, the t-GAS model positions the appropriate risk quantile more accurately. We investigate the origin of these differences in more detail in the next section.

4. Performance over different volatility states

Both Figs. 2 and 4 are suggestive that the t-GAS model outperforms the t-GARCH model particularly in the tails of the volatility distribution. This includes the distressed periods during which outperformance matters most in economic terms. In this section we study this phenomenon in more detail.

We first investigate the KL divergence behavior of the different models conditional on the true volatility state. For this, we set up a new set of simulations. As we want to condition on a particular volatility level, we want to simulate data in such a way that the final volatility level is high (or low), as desired. Therefore, for a grid of final volatility levels f_T , we simulate backwards a range of volatility paths $f_T, f_{T-1}, \dots, f_1$ that could have led up to the current volatility level. Given the Gaussian autoregressive structure of the stochastic volatility DGP in (8), the log-volatility process is time reversible and we can simulate past volatility scenarios by the simple time-reversed recursion

$$\log f_{t-1} = a + b \log f_t + v_t, \quad v_t \sim N(0, \sigma_v^2),$$

for $(a, b) = (0.00, 0.98)$; see Ōsawa (1988). We use 1,000 (backward) simulated volatility paths $f_T, \dots, f_1$ to draw paths of y_t as in (8). This allows us to construct estimates of the expected KL divergences at time T conditional on f_T for both the t-GAS

Fig. 5. Relative KL divergences between the t-GAS and t-GARCH models conditional on a true variance f_T at time T for the stochastic volatility model DGP. Positive values indicate the t-GAS performs better. The left panel shows the results when evaluated at the pseudo-true parameter values (estimated based on a large sample of $T = 500,000$ observations), while the right panel shows the results for the average over the finite sample estimates $\hat{\theta}_T$ (where $T = 1,000$). Number of replications is $N = 1,000$ for each of the parameter draws $\hat{\theta}_T$.

Fig. 6. Relative KL divergences between the t-GAS and t-GARCH models for four pre-specified paths of $f_t = \sigma_t^2$. Positive values indicate the t-GAS performs better. The left panels visualize the volatility paths. The right panels show the time series of expected relative KL divergences. For each path we take $\tau \in \{3, 4, 5, 9\}$. The expectations are obtained using $N_f = 2,500$ Monte Carlo simulations.

and t-GARCH model. For each model and each parameter draw $\hat{\theta}_T$ obtained earlier, the conditional (on f_T) expected relative KL divergence is estimated by the average over the KL divergences at time T for the 1,000 backward simulated volatility paths.

Fig. 5 presents the average relative KL divergences of the t-GAS versus the t-GARCH model conditional on the value of f_T . Positive values indicate that the t-GAS model does better. Both panels in Fig. 5 illustrate that the t-GAS indeed performs better in the tails of the volatility distribution (high and low values of f_T), both at the pseudo-true parameter (left panel) and at the finite sample estimates (right panel). The fatter tailed the DGP, the more pronounced is the difference. The pattern is stronger at the pseudo-true parameter, because the finite sample results are averaged across a wide range of finite sample estimates, see Fig. 1. To sum up, Fig. 5 clearly indicates the better performance of the t-GAS model in periods of high (and low) volatility. This holds both in terms of accuracy of the fit in the tails of f_t (Fig. 4) and in terms of relative KL divergence (Fig. 5).

Fig. 7. Illustration of the differences in the scaled score functions of the t-GAS and t-GARCH models. The left panel displays the shape of the scaled score functions for different values of $\tilde{\tau}$, a fixed $\tilde{f}_t = \tilde{\sigma}_t^2 = 4$, and an incoming $|y_t|$. The right panel shows the sign regions of these scaled score functions.

The results are further corroborated in Fig. 6. Here we plot the relative KL divergence over time for a number of deterministic volatility paths. Similar paths were used in Engle (2002) to show how well a model keeps track of the true time-varying parameters. The figure shows that the t-GAS outperforms the t-GARCH model in periods of high and low volatility. For example, for the sine wave pattern, we see peaks in relative KL divergence around $t = 0, 250, 500, 750, 1000$, which are precisely the times at which volatility is at a peak or a trough. Similar patterns emerge for the pulse and sawtooth case. We see a short underperformance of the t-GAS precisely at the time of the volatility break in those cases, but already quickly after the t-GAS picks up again and outperforms the t-GARCH in terms of relative KL divergence. The results are consistent across different degrees of fat-tailedness (τ) of the DGP.

The better performance of the t-GAS model should be closely related to the t-GAS model stepping more often in the correct direction than the t-GARCH model, or, at least, stepping less aggressively in the wrong direction. Fig. 7 visualizes that this is indeed the case, particularly for more fat-tailed data, i.e., for lower values of τ . The left panel plots the scaled score steps of the t-GAS model for a range of estimated degrees of freedom $\tilde{\tau}$. For $\tilde{\tau} \rightarrow \infty$, the t-GAS and the t-GARCH model coincide. The left panel shows that the t-GAS model is increasingly robust if the innovations in the model are fatter tailed (small $\tilde{\tau}$). This robustness property is what drives the better performance under fat-tailedness: outliers have a reduced impact, which causes the t-GAS model to step less quickly in the wrong direction.

The right panel in Fig. 7 illustrates when the t-GAS and t-GARCH models step in different directions. The white regions indicate where the t-GAS and the t-GARCH models step in the same direction, either by decreasing volatility (left-hand white region) or by increasing it (right-hand white region). The black region in the middle indicates where the two models step in different directions. If $\tilde{\tau}$ is large (vertical axis), the region of $|y_t|$ where the two models take a differently signed step is small. If $\tilde{\tau}$ is small, i.e., if we account for a high degree of fat-tailedness, the region is substantial and explains part of the difference in performance between the two models, as the models step into different directions more often.

5. Empirical application: Clusters of volatility in stock returns

In this section, we apply our results empirically. The challenge for an empirical application is that we do not know the true density $p_t^0(y_t | Y_{t-1})$. We can, therefore, not compute the conditionally expected KL improvement from (6). We can, however, estimate its unconditional counterpart similar as in Fig. 3 as

$$\mathbb{E} \left[D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_{t+1}; \hat{\theta}_T) \right) - D \left(p_t^0(\cdot | Y_{t-1}), p(\cdot | f_t; \hat{\theta}_T) \right) \right] \approx \frac{1}{T} \sum_{t=1}^T \hat{L}_t, \quad (12)$$

$$\hat{L}_t = \log \frac{p(y_t | f_t; \hat{\theta}_T)}{p(y_t | f_{t+1}; \hat{\theta}_T)}.$$

Effectively, $\hat{L}_t$ is negative if the model density increases at y_t after updating f_t to f_{t+1} . Although one might expect $\hat{L}_t$ to always be negative, this need not be the case. In particular, the mean-reversion effect in the dynamics of f_t based on the finite sample estimate $\hat{\theta}_T$ might work in the opposite way of the signal given by y_t . Also, the differences in the scaled score functions of the GAS and GARCH models might result in them taking opposite directions; see Fig. 7.

Fig. 8. Time t log-likelihood improvements $\hat{L}_t$ of updates from f_t to f_{t+1} for the t-GAS and t-GARCH model. The sample period is April 1, 2008 until March 19, 2012.

We illustrate this empirically for two large company stocks, namely Goldman Sachs (GS) and Google (GOOGL). For both stocks we use the total return index, which we obtain from Datastream. We take daily data for the period from April 1, 2008 until March 19, 2012, which provides us two samples consisting of $T = 1,000$ observations each. Fig. 8 shows the log-returns for both series. Two periods of high variability for both stocks are evident, namely the financial crisis (2008–2009) and part of the European sovereign debt crisis (2011–2012). Volatility appears low in the rest of the sample.

Fig. 8 shows the absolute returns, the variance paths for the variances of the GARCH and GAS model, and the instances where $\hat{L}_t$ is positive for the GARCH and GAS model, respectively. The differences in filtered variance paths are most noticeable during periods of high volatility and outliers. This is in line with the results from Section 4. During the financial crisis we see that the filtered path of the t-GAS model indicates higher levels of the conditional variance compared to the t-GARCH model. We also see that the filtered t-GARCH conditional variance can strongly increase with the occurrence of an outlier. For instance, for Google there is a peak with exponential decay in the volatility pattern at the beginning of the sample. Similar peaks with subsequent decay patterns are also seen at other times for both stocks for the t-GARCH model. The filtered conditional variances for the t-GAS model on the other hand are much more robust. This robustness property of the t-GAS is in line with the good performance of the model in the simulation setting, particularly for fat-tailed data; see Figs. 5–7.

It is interesting to see that the GAS model results in many fewer instances of $\hat{L}_t > 0$, i.e., cases where the model density decreases at y_t after y_t is observed and f_t is updated to f_{t+1} . This is in line with the differences highlighted in Fig. 7, and the overall gain in unconditionally expected KL divergence presented in Fig. 3. One of the reasons for observing $\hat{L}_t > 0$ is due to the mean reversion effect in the transition equation for f_t evaluated at the finite sample estimate $\hat{\theta}_T$. The GARCH updates appear to be much more hampered by this than those of the GAS model. Also note that though it sounds intuitive that the density at y_t should increase after observing y_t , this does not guarantee that such an update also brings the updated density closer to the true density. To guarantee this, we would have to know the true density, which we have assumed to be unknown. Nevertheless, the figure shows that the t-GARCH model makes many more ‘non-intuitive’, local likelihood decreasing updates than the t-GAS model. The differences are particularly clear in times of high volatility. There are relatively many $\hat{L}_t > 0$ for the GARCH model in these cases, whereas there are hardly any for the GAS model. This empirical finding is in line with the substantially better KL divergences in Figs. 5 and 6, and with the increased VaR accuracy in Fig. 4.

6. Conclusion

The results in this paper further characterize the optimality properties of score-driven models and extend the findings of Blasques et al. (2015) and Creal et al. (2018). We do so using Monte Carlo simulations for data generating processes (DGPs) with parameter-driven stochastic volatility dynamics. Our results do not allow us to claim finite sample optimality of score-driven models in general or even for other data generating processes, but for stochastic volatility DGPs the score-driven models appear to outperform their GARCH counterparts in several dimensions.

Using fat-tailed stochastic volatility DGPs, we found that the optimality properties of score-driven volatility models carry over from the asymptotic setting to finite samples. This is remarkable given that finite samples show considerable dispersion of the maximum likelihood estimates due to the strong non-linearity of score-driven models.

Our simulation results also indicated that score-driven models are optimal when it matters most: (a) when the data are fat-tailed and robustness is important, and (b) in periods of financial distress when true volatilities are high and filtered volatilities differ mostly across models. It is precisely then that differences are also large in terms of economic risk measures

such as distribution quantiles, or Value-at-Risk. Using new simulations, we were able to pinpoint the origin of the differences. Score-driven models in expectation step in the correct direction when updating the time-varying parameter more often than traditional competing models, such as the GARCH model. This holds particularly when volatility levels are high. The optimality properties of score-driven volatility models are shared by other models with similar news impact curves, such as the BIP-GARCH model of [Laurent et al. \(2016\)](#) in the context of the t-GAS model. Our results may thus also help to motivate such alternative model specifications.

We illustrated the results empirically for two U.S. stock return series. We found that the t-GAS model shows many fewer instances of (non-intuitive) likelihood *decreasing* updates of volatility than does the t-GARCH model. This holds particularly during highly volatile periods such as the 2008 financial crisis. Combined, the results point to a further underpinning of the use of score-driven models also in an empirical, finite sample setting.

Declaration of Competing Interest

We hereby officially state that there was no conflict of interest with respect to this paper.

Acknowledgements

We thank two anonymous referees for careful and constructive comments that helped to improve the paper. Van Vlodrop and Lucas (NWO, grant [VICI453-09-005](#)) and Blasques (NWO, grant [VI.Vidi.195.099](#)) thank the [Dutch National Science Foundation](#) for financial support.

References

- Blasques, F., Koopman, S.J., Lucas, A., 2015. Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika* 102 (2), 325–343, and “Amendments and corrections,” *Biometrika* 105(3), 753.
- Blasques, F., Koopman, S.J., Lucas, A., Schaumburg, J., 2016. Spillover dynamics for systemic risk measurement using spatial financial time series models. *Journal of Econometrics* 195 (2), 211–223.
- Bollerslev, T., 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics* 31 (3), 307–327.
- Bollerslev, T., 1987. A conditionally heteroskedastic time series model for speculative prices and rates of return. *The Review of Economics and Statistics* 69 (3), 542–547.
- Catania, L., Billé, A.G., 2017. Dynamic spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Applied Econometrics* 32, 1178–1196.
- Creal, D., Koopman, S.J., Lucas, A., 2011. A dynamic multivariate heavy-tailed model for time-varying volatilities and correlations. *Journal of Business & Economic Statistics* 29 (4), 552–563.
- Creal, D., Koopman, S.J., Lucas, A., 2013. Generalized autoregressive score models with applications. *Journal of Applied Econometrics* 28 (5), 777–795.
- Creal, D., Koopman, S.J., Lucas, A., Zamojski, M., 2018. Generalized autoregressive method of moments. Tinbergen Institute Discussion Paper TI-2015/138 (III).
- Creal, D., Schwaab, B., Koopman, S.J., Lucas, A., 2014. Observation-driven mixed-measurement dynamic factor models with an application to credit risk. *Review of Economics and Statistics* 96 (5), 898–915.
- Engle, R., 2002. Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics* 20 (3), 339–350.
- Engle, R.F., 1982. Autoregressive conditional heteroskedasticity with estimates of the variance of united kingdom inflation. *Econometrica* 50 (4), 987–1007.
- Harvey, A., Luati, A., 2014. Filtering with heavy tails. *Journal of the American Statistical Association* 109 (507), 1112–1122.
- Harvey, A.C., 2013. *Dynamic models for volatility and heavy tails: with applications to financial and economic time series*, 52. Cambridge University Press.
- Janus, P., Koopman, S.J., Lucas, A., 2014. Long memory dynamics for multivariate dependence under heavy tails. *Journal of Empirical Finance* 29, 187–206.
- Laurent, S., Lecourt, C., Palm, F.C., 2016. Testing for jumps in conditionally gaussian arma-garch models, a robust approach. *Computational Statistics & Data Analysis* 100, 383–400.
- Lucas, A., Schwaab, B., Zhang, X., 2014. Conditional euro area sovereign default risk. *Journal of Business & Economic Statistics* 32 (2), 271–284.
- Lucas, A., Schwaab, B., Zhang, X., 2017. Modeling financial sector joint tail risk in the euro area. *Journal of Applied Econometrics* 32 (1), 171–191.
- Nelson, D.B., 1991. Conditional heteroskedasticity in asset returns: A new approach. *Econometrica* 59, 347–370.
- Oh, D.H., Patton, A.J., 2017. Modeling dependence in high dimensions with factor copulas. *Journal of Business & Economic Statistics* 35 (1), 139–154.
- Opschoor, A., Janus, P., Lucas, A., Van Dijk, D., 2017. New heavy models for fat-tailed realized covariances and returns. *Journal of Business & Economic Statistics* 36, 643–657.
- Ôsawa, H., 1988. Reversibility of first-order autoregressive processes. *Stochastic processes and their applications* 28 (1), 61–69.
- Salvatierra, I.D.L., Patton, A.J., 2015. Dynamic copula models and high frequency data. *Journal of Empirical Finance* 30, 120–135.
- Sentana, E., 1995. Quadratic arch models. *The Review of Economic Studies* 62 (4), 639–661.
- Zakoian, J.-M., 1994. Threshold heteroskedastic models. *Journal of Economic Dynamics and control* 18 (5), 931–955.