Biometrika (2023), **XXX**, X, *pp*. 1–15
Printed in Great Britain
Submitted for publication

On the optimality of score-driven models

BY P. GORGI

Department of Econometrics and Data Science, Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081HV Amsterdam, The Netherlands
p.gorgi@vu.nl

C.S.A. LAURIA

Department of Statistical Sciences, University of Bologna
via Belle Arti 41 Bologna 40126, ITALY
christopher.lauria2@unibo.it

AND A. LUATI

Department of Mathematics, Imperial College London
108 Queen's Gate, SW7 2AZ, London, U.K.
a.luati@imperial.ac.uk

SUMMARY

Score-driven models have been recently introduced as a general framework to specify time-varying parameters of conditional densities. The score enjoys stochastic properties that make these models easy to implement and convenient to apply in several contexts, ranging from biostatistics to finance. Score-driven parameter updates have been shown to be optimal in terms of locally reducing a local version of the Kullback–Leibler divergence between the true conditional density and the postulated density of the model. A key limitation of such an optimality property is that it holds only locally both in the parameter space and sample space, yielding to a definition of local Kullback–Leibler divergence that is in fact not a divergence measure. The current paper shows that score-driven updates satisfy stronger optimality properties that are based on a global definition of Kullback–Leibler divergence. In particular, it is shown that score-driven updates reduce the distance between the expected updated parameter and the pseudo-true parameter. Furthermore, depending on the conditional density and the scaling of the score, the optimality result can hold globally over the parameter space, which can be viewed as a generalisation of the monotonicity property of the stochastic gradient descent scheme. Several examples illustrate how the results derived in the paper apply to specific models under different easy-to-check assumptions, and provide a formal method to select the link-function and the scaling of the score.

Some key words: Kullback–Leibler divergence; pseudo-true parameters; score-driven models.

1. INTRODUCTION

A simple way to introduce dynamics in a statistical model is by allowing time variation in some features of the probability distribution. One way to do so is by letting some of the parameters that characterise the distribution itself to vary through time. Models that utilise this idea are called time-varying parameter models. Cox (1981) gives a categorisation of time-varying parameter models by dividing them into two classes: observation-driven models and parameter-driven models. The former are models where the updating equation is specified as a function of the observations, instead, the latter are models where the dynamic equation is governed by idiosyncratic innovations. In some cases, the same model can be specified both as a parameter-driven model and as an observation-driven model, see for example a linear Gaussian signal plus noise model and its corresponding filtering recursions (Durbin & Koopman, 2012; Harvey, 1989). The categorisation turns out to be useful when non linear observation-driven models are considered, such as the celebrated generalised autoregressive conditional heteroscedasticity (GARCH) model by Bollerslev (1986) and Engle (1982). We refer the reader to Koopman et al. (2016) for a discussion on strengths and weaknesses of these two classes of models.

Creal et al. (2013) and Harvey (2013) introduce a general class of observation-driven models that provide a unified framework to specify observation-driven time-varying parameters. This class of models is commonly referred to as the class of score-driven models, and it is also known as the Generalized Autoregressive Score or Dynamic Conditional Score class. The key feature of score-driven models is that the dynamic of the time-varying parameter is driven by a process that is proportional to the score of the conditional likelihood taken with respect to the parameter of interest. Score-driven models enjoy statistical properties that make these models easy to implement as well as convenient to apply in several diverse contexts, such as, for instance, robust filtering (Harvey & Luati, 2014; Gorgi, 2020), spatio-temporal modelling with applications to neuroscience (Gasperoni et al., 2023) and finance (Blasques et al., 2016; Catania & Billé, 2017), quantile estimation (Patton et al., 2019; Catania & Luati, 2022), mixture models (Catania, 2021), and survival probability models (Gorgi, 2018). An up-to-date repository on articles that use score-driven models is available at the following link, `http://www.gasmodel.com/`.

The use of the score to specify the updating equation of time-varying parameters has been motivated in the literature from a theoretical standpoint by showing that it locally reduces a local version of the Kullback–Leibler (KL) divergence between the true conditional density and the postulated density of the model. More specifically, Blasques et al. (2015) show that the sign of the score is in the direction of reducing a local KL divergence. The key limitations of this result is that it is local both in terms of the parameter space and the sample space on which the KL divergence is defined. In particular, the local KL divergence that is defined in Blasques et al. (2015) is in fact not a divergence measure as it can also take negative values. Furthermore, the result is also local in the parameter space as only the sign of the score matters and not the size of the update, leading to the fact that any parameter update with the same sign is equivalent in terms of the optimality definition in Blasques et al. (2015). Therefore, the result does not provide any theoretical insights on the optimality of the scaling of the score and on the choice of the link function as any re-scaling of the score and any monotone link function are equivalent in terms of the resulting sign of the score.

In this paper, we show that score-driven parameter updates satisfy stronger optimality properties that rely on a global definition of KL divergence. It is shown that the expected score-driven parameter update reduces the distance with respect to a pseudo-true time-varying parameter, which is defined as the time-varying parameter that minimises the global KL divergence between the true conditional density of the data generating process and the, possibly misspecified, conditional density postulated by the model. Hence, we refer to optimality in conditional expected variation. We provide different sets of conditions under which optimality can hold either globally over the parameter space or locally, requiring the size of the score update to become arbitrarily small. We also show that, locally, optimality in conditional expected variation implies a reduction in the mean squared error with respect to the pseudo-true parameter. We discuss through several examples how these conditions provide practical insights on how to select the scaling factor that multiplies the score in the updating equation and the link function of the time-varying parameter.

The results in the paper are related to the stochastic gradient descent literature and they formalise the practical intuition given by Creal et al. (2013) to justify score-driven models, i.e.“The use of the score for updating the parameter is intuitive. It defines a steepest ascent direction for improving the model's local fit in terms of the likelihood or density at time t given the current position of the parameter. This provides the natural direction for updating the parameter”. More precisely, the score-driven update can be seen as a time-varying optimisation problem, where the sequence of objective functions is given by the conditional log-density postulated by the model. The resulting optimality properties can be viewed as a natural generalisation of the monotonicity property of the updates of a gradient descent scheme.

2. Filtering with score-driven models

Let $\{y_t\}_{t\in\mathbb{Z}}$ be a time series process with elements taking values in $\mathcal{Y} \subseteq \mathbb{R}$. Assume that the probability density function of y_t conditional on $\mathcal{F}_{t-1} = \sigma(y_{t-1}, y_{t-2}, \dots)$ is given by $\tilde{p}_t(y)$, i.e. $y_t|\mathcal{F}_{t-1} \sim \tilde{p}_t(y)$. We shall refer to $\tilde{p}_t(y)$ as the true conditional density function, which is assumed to be unknown. We specify a conditional density function to model the time series process

$$y_t|\mathcal{F}_{t-1} \sim p(y|\lambda_t), \tag{1}$$

where the $p(y|\lambda_t)$ is a probability density function and λ_t is a time-varying parameter that takes values in the set $\Lambda \subseteq \mathbb{R}$. In practice, $p(y|\lambda_t)$ may be a parametric density function that also depends on a static parameter vector to be estimated, however, this is not relevant for the optimality results discussed below. We also note that the conditional density $p(y|\lambda_t)$ may be misspecified with respect to the true conditional density, namely, there may not be a value of λ_t such that $p(y|\lambda_t) = \tilde{p}_t(y)$. Depending on the variables of interest and on the time points at which they are evaluated, the model density in equation (1) will be equivalently denoted also as $p(y_t|\lambda)$, $p(y_t|\lambda_t)$ or $p(y|\lambda)$.

The score-driven framework of Creal et al. (2013) and Harvey (2013) provides a general approach to specify observation-driven time-varying parameters. A first order score-driven model for the time-varying parameter λ_t is described by the following equation

$$\lambda_{t+1} = \omega + \beta\lambda_t + \alpha S(\lambda_t) s(y_t, \lambda_t), \tag{2}$$

where $S(\lambda_t)$ is a positive scaling factor and $s(y_t, \lambda_t)$ is the score of the predictive log-density

$$s(y_t, \lambda_t) = \frac{\partial \log p(y_t|\lambda_t)}{\partial \lambda_t}.$$

In the literature, the scaling factor $S(\lambda_t)$ is typically selected to be a transformation of the conditional Fisher information (Creal et al., 2013) and different scaling factors give rise to different model specifications, see also Ayala et al. (2023) for an empirical comparison of scaling factors in score driven models.

Blasques et al. (2015) show that score-driven parameter updates are locally optimal in reducing a local version of the KL divergence between the true conditional density and the conditional density of the model. The KL divergence between the true conditional density function $\tilde{p}_t(y)$ and the conditional density function of the model $p(y|\lambda)$ is

$$KLD_t(\lambda) = \int_{\mathbb{R}} \tilde{p}_t(y) \log \frac{\tilde{p}_t(y)}{p(y|\lambda)} dy.$$

The local KL divergence defined in Blasques et al. (2015) replaces the integration set $\mathbb{R}$ with a small interval $\mathcal{Y}_\varepsilon \subseteq \mathbb{R}$ around the observed value y_t.

The optimality results in this paper extend the work of Blasques et al. (2015) in several respects. In particular, the results in this paper are based on the global KL divergence $KLD_t(\lambda)$ and not a local version. This is a major feature since the local KL divergence in Blasques et al. (2015) is, in fact, not a divergence measure as it can be negative. As noted in Blasques et al. (2018b), the local KL divergence in Blasques et al. (2015) is positive only if $\tilde{p}_t(y_t) > p(y_t|\lambda_t)$. However, y_t is a continuous random variable and therefore, in general, we have that $\tilde{p}_t(y_t) < p(y_t|\lambda_t)$ with positive probability. This affects the interpretation of the results, as they do not actually entail that there is a divergence measure such that the assumed density $p(y|\lambda_{t+1})$ is closer to the true one $\tilde{p}_t(y)$ compared to $p(y|\lambda_t)$.

Blasques et al. (2015) consider the Newton-score parameter update, which is a special case of the score-driven update in (2) with $\beta = 1$ and $\omega = 0$,

$$\lambda_{t+1} = \lambda_t + \alpha S(\lambda_t)s(y_t, \lambda_t), \tag{3}$$

and show that the local KL divergence between $\tilde{p}_t(y)$ and $p(y|\lambda_{t+1})$ is smaller than the local KL divergence between $\tilde{p}_t(y)$ and $p(y|\lambda_t)$ for an arbitrarily small value of the score innovation $S(\lambda_t)s(y_t, \lambda_t)$.

In the following section, we derive optimality results for score-driven parameter updates with respect to the pseudo-true time-varying parameter that minimises the conditional KL divergence $KLD_t(\lambda)$. In particular, we show that the score-driven parameter update from λ_t to λ_{t+1} gets closer in expected value to the pseudo-true time-varying parameter. This provides a clear interpretation of score-driven filters as optimal approximations of a pseudo-true time-varying parameter in a misspecified framework. The criterion function used to characterise the otherwise non-unique concept of optimality is formalised in the next section.

3. OPTIMALITY OF SCORE-DRIVEN UPDATES

3.1. *Optimality results in conditional expected variation*

Let us define the pseudo-true time-varying parameter λ_t^* as the value that minimises the KL divergence

$$\lambda_t^* = \arg\min_{\lambda \in \Lambda} KLD_t(\lambda).$$

We refer to Akaike (1998) and White (1982) for interpretation of KL divergence and its use in statistics and econometrics. Throughout the paper, we use the shorthand notation $E_t(\cdot)$ to denote the expectation conditional on $\mathcal{F}_t$, i.e. $E_t(\cdot) = E_t(\cdot|\mathcal{F}_t)$. We also define the function $f_t(\lambda)$ as follows

$$f_t(\lambda) = E_{t-1}[\log p(y_t|\lambda)].$$

We note that the conditional expectation $E_{t-1}[\log p(y_t|\lambda)]$ is with respect to the true conditional distribution of y_t, i.e. $\tilde{p}_t(y)$. Hence, minimising $KLD_t(\lambda)$ is the equivalent of maximising $f_t(\lambda)$ and therefore λ_t^* maximises $f_t(\lambda)$.

We classify a parameter update from λ_t to λ_{t+1} as optimal in conditional expected variation (CEV), if the distance between the expected updated parameter $E_{t-1}(\lambda_{t+1})$ and the pseudo-true λ_t^* is smaller than the distance between λ_t and λ_t^*. The interpretation is that the parameter update from λ_t to λ_{t+1} is based on the observable y_t, which is generated under the true conditional probability measure $\tilde{p}_t(y)$. A CEV optimal update is expected to process the information in y_t to update λ_t in the correct direction in such a way that on average λ_{t+1} gets closer to λ_t^*. Namely, the conditional expected variation from λ_t to λ_{t+1} is in the direction of the pseudo-true parameter λ_t^*. Note that the expectation $E_{t-1}(\lambda_{t+1})$ averages out only the impact of y_t, which is the most recent observation in the filter λ_{t+1}. A formal definition of CEV optimality is given below.

DEFINITION 1 (CONDITIONAL EXPECTED VARIATION OPTIMALITY). *A parameter update from λ_t to λ_{t+1} is optimal in conditional expected variation if*

$$\begin{cases} |\lambda_t^* - E_{t-1}(\lambda_{t+1})| < |\lambda_t^* - \lambda_t|, & \textit{if } \lambda_t \neq \lambda_t^*, \\ E_{t-1}(\lambda_{t+1}) = \lambda_t^*, & \textit{if } \lambda_t = \lambda_t^*. \end{cases}$$

We start by introducing some regularity conditions on the model conditional density.

Assumption 1. The function $f_t(\lambda)$ is twice continuously differentiable in Λ with probability one. Furthermore, the derivative and the conditional expectation of the conditional log-density p can be interchanged, i.e.

$$f_t'(\lambda) = E_{t-1} s(y_t, \lambda),$$

for any $\lambda \in \Lambda$.

Assumption 1 is a standard regularity condition that enables us to interchange integration and differentiation of the conditional log-density.

Assumption 2. The set $\Lambda \subseteq \mathbb{R}$ is open and convex. The pseudo-true time-varying parameter λ_t^* is the unique global maximum of $f_t(\lambda)$ in Λ with probability 1.

Assumption 2 ensures the existence and uniqueness of λ_t^* and imposes some smoothness conditions on $f_t(\lambda)$.

Assumption 3. The function $S(\lambda)$ is continuously differentiable. There is a constant $c > 0$ such that

$$0 < -\frac{\partial S(\lambda) f_t'(\lambda)}{\partial \lambda} \leq c \quad \textit{a.s.} \tag{4}$$

for any $\lambda \in \Lambda$.

Assumption 3 requires the expected score innovation to be a decreasing Lipschitz continuous function with respect to λ. Under these assumptions, we obtain optimality in conditional expected variation of the Newton-score parameter update.

THEOREM 1. *Let Assumptions 1-3 hold. Then, the Newton-score parameter update defined in (3) with $0 < \alpha < 2/c$ is CEV optimal.*

The proof, in Appendix, follows similar arguments as the ones that are typically used to prove the convergence of the gradient descent algorithm. Note that Assumptions 1-3 involve only the conditional expectation, with respect to the true density, of the model log-density. In practice, this expectation often reduces to some moment conditions on y_t. As we shall see in Section 4, in practice, the most restrictive assumption is the Lipschitz continuity of the expected score innovation in Assumption 3. We consider a weaker Lipschitz condition that provides an alternative to Assumption 3. This assumption can be used to deliver a local version of the optimality result in Theorem 1.

Assumption 4. For any compact subset $\Lambda_c \subset \Lambda$, there is a positive $\mathcal{F}_{t-1}$-measurable random variable c_t such that the condition in (4) with c replaced by c_t holds for any $\lambda \in \Lambda_c$.

Assumption 4 is weaker than Assumption 3 as it only requires the expected score function to be Lipschitz on compact subsets of Λ instead of the whole set Λ.

THEOREM 2. *Let Assumptions 1, 2 and 4 hold. Then, there exists $\alpha_t = a_t(\lambda_t, \lambda_t^*) > 0$, where a_t is a $\mathcal{F}_{t-1}$-measurable random function, such that the Newton-score parameter update defined in (3) with α replaced by α_t is CEV optimal.*

Theorem 2 provides a local result in the sense that it only guarantees that there is a small enough α_t such that the Newton-score update is optimal. The size of α_t depends on λ_t and λ_t^* and therefore α_t may become arbitrarily small depending on the current state of λ_t and λ_t^*. Several examples are presented in Section 4 that illustrate how the choice of the scaling function $S(\lambda)$, as well as the link function for the time-varying parameter λ_t, can affect whether the global optimality result or the local optimality result holds. It is important to remark that the concepts of local and global apply here to the parameter space Λ, not to the range of y_t. As the updating scheme in score-driven models is based on a derivative with respect to λ, it is expected that, in absence of high level assumptions on the data generating process, optimality holds when small variations of time-varying parameters are considered, see also the discussion in Van Os et al. (2022).

The results in Theorems 1 and 2 rely on the assumption that the function $S(\lambda)f_t'(\lambda)$ is strictly decreasing. This ensures that $f_t(\lambda)$ does not have stationary points other than the global maximum λ_t^*, ruling out for instance the possibility of local maxima. As we shall see in Section 4, for some models, this assumption may not be satisfied or, in general, it may be difficult to check that it holds because $f_t'(\lambda)$ may not be available in closed form. In the following, we consider a weaker version of the results derived so far, which only requires the function $S(\lambda)f_t'(\lambda)$ to be strictly decreasing in a neighborhood of the global maximum of $f_t(\lambda)$.

Assumption 5. Condition in (4) holds for any λ in an open neighborhood of the global maximum of $f_t(\lambda)$ for some value $c > 0$ that may depend on λ_t^*.

Assumption 5 is a milder version of Assumption 3 that is required to hold only on a neighborhood of the global maximum of $f_t(\lambda)$.

THEOREM 3. *Let Assumptions 1, 2 and 5 hold. Then, there is an $\epsilon > 0$ such that the Newton-score parameter update defined in (3) is CEV optimal for $\lambda_t \in \{\lambda \in \Lambda : |\lambda - \lambda_t^*| < \epsilon\}$.*

Theorem 3 delivers the CEV optimality of the Newton-score update under the constraint that the parameter value λ_t is close enough to the pseudo-true λ_t^*. This condition is needed as the function $f_t(\lambda)$ is not required to be strictly concave in Λ and otherwise the parameter update may go in the direction of a local maximum.

Theorem 3 applies to a very large class of models under standard regularity conditions, provided that the conditional density function is correctly specified, i.e. $\tilde{p}_t(y) = p(y|\lambda_t^*)$. If the conditional density is correctly specified, then λ_t^* is the time-varying parameter of the true conditional density. Therefore, in this case, $f_t''(\lambda_t^*) = -I(\lambda_t^*)$, where $I(\lambda)$ is the Fisher information associated with the conditional density function $p(y|\lambda)$. Under standard regularity conditions on the density function $p(y|\lambda)$, the Fisher information $I(\lambda)$ is a continuous function and $I(\lambda) > 0$ for any $\lambda \in \Lambda$. Note that the equality $f_t''(\lambda_t) = -I(\lambda_t)$ does not hold for $\lambda_t \neq \lambda_t^*$ as the conditional expectation in $f_t''(\lambda_t)$ is taken with respect to the true parameter λ_t^*. However, under continuity of the function $f_t''(\lambda)$, we have that $f_t''(\lambda_t) \to -I(\lambda_t^*)$ as $|\lambda_t - \lambda_t^*| \to 0$. Theorem 3 imposes that λ_t is arbitrarily close to λ_t^* and thus since $I(\lambda_t^*) > 0$ we obtain that $f_t''(\lambda_t) < 0$ and is bounded from below by a function of λ_t^*. This entails that Assumption 5 holds.

Finally, we note that the CEV optimality results presented in this section also entail that the parameter update reduces the mean squared error (MSE) with respect to the pseudo-true parameter 225 λ_t^* for a small enough α.

COROLLARY 1. *Let the assumptions of either Theorem 1, Theorem 2 or Theorem 3 hold. Furthermore, assume that $E_{t-1}s(y_t,\lambda)^2 < \infty$ a.s. for any $\lambda \in \Lambda$. Then, for any (λ_t, λ_t^*), $\lambda_t \neq \lambda_t^*$, there exists a small enough $\alpha > 0$ such that the Newton-score update reduces the MSE with respect to the pseudo-true λ_t^*, i.e.*

$$E_{t-1}\{(\lambda_t^* - \lambda_{t+1})^2\} < (\lambda_t^* - \lambda_t)^2.$$

The result in Corollary 1 can only hold for a small enough α that depends on λ_t and λ_t^*. This is intuitive as we have that α must be zero in the limit case where $\lambda_t^* = \lambda_t$ to achieve $\lambda_{t+1} = \lambda_t$, which entails $E_{t-1}\{(\lambda_t^* - \lambda_{t+1})^2\} = (\lambda_t^* - \lambda_t)^2$.

In summary, Theorems 1, 2 and 3 establish CEV optimality results under the easy-to-verify 230 conditions given in Assumptions 3, 4 and 5, respectively. The result of Theorem 1 is global as it holds for any λ_t over the parameter space Λ and uniformly for a fixed value of α. On the other hand, the result of Theorem 2 is local in the sense that it holds for any λ_t over the parameter space Λ but not uniformly as α depends on λ_t and λ_t^*. Finally, the result of Theorem 3 is fully local at it holds for λ_t in a neighborhood of the pseudo-true parameter λ_t^*. We note that the 235 KL divergence is in all cases the global one, defined over the whole sample space. It is easy to see that Assumption 3 implies Assumption 4, which itself implies Assumption 5. As noted above, Assumption 5 holds for a very wide class of models under high level conditions on the true density. Assumptions 3 and 4 can be easily checked for any given density specification, including different choices of the scaling factor and of the link function. Section 4 illustrates 240 through several examples how Theorems 1, 2 and 3 apply in practice and also their implications on the range of optimality of the score coefficient α. Finally, Corollary 1 shows that, under the same assumptions of Theorems 1, 2 and 3, CEV optimality implies a reduction of the MSE with respect to the pseudo-true parameter.

3.2. *Discussion on mean reversion* 245

The conditional expected variation optimality discussed in Section 3.1 concerns the pseudo-true time-varying parameter at time t, i.e. λ_t^*. In practice, the updated time-varying parameter λ_{t+1} is useful to approximate λ_{t+1}^* and not λ_t^*. However, the parameter update from λ_t to λ_{t+1} relies on the observable variable y_t at time t. Therefore, assumptions on how y_{t+1} relates to y_t, or, equivalently, on how λ_{t+1}^* relates to λ_t^*, are required in order to make any claim on the 250 optimality of the score-driven parameter update with respect to λ_{t+1}^*. In the rest of the section, we discuss how CEV optimality is retained with respect to λ_{t+1}^* under some conditions. The discussion below also motivates the use of the mean reverting score-driven specification in (2), which is often considered in the empirical applications instead of the Newton-score specification.

First, we note that a form of CEV optimality also holds with respect to λ_{t+1}^* when the pseudo- 255 true parameter is a martingale process such that $E_{t-1}(\lambda_{t+1}^*) = \lambda_t^*$. Under this condition, and together with Assumptions 2-3, we obtain that if $\lambda_t \neq \lambda_t^*$, then $|E_{t-1}(\lambda_{t+1}^* - \lambda_{t+1})| < |\lambda_t^* - \lambda_t|$. This result implies that the distance between the expected λ_{t+1}^* and λ_{t+1} is smaller than the distance between λ_t^* and λ_t. We note that the result follows immediately from Theorem 1 as we are assuming here that $E_{t-1}(\lambda_{t+1}^*) = \lambda_t^*$. 260

Second, if we assume that λ_t^* is a mean reverting process with conditional expectation given by $E_{t-1}(\lambda_{t+1}^*) = \omega + \beta\lambda_t^*$, and $\beta \in (0, 1]$, we obtain an optimality property for the general case of a score-driven parameter update in (2). In particular, following the same arguments as in the proof of Theorem 1, it is straightforward to prove that for $0 < \alpha < 2\beta/c$ the score-driven update in (2) satisfies $|E_{t-1}(\lambda_{t+1}^* - \lambda_{t+1})| < \beta|\lambda_t^* - \lambda_t|$, if $\lambda_t^* \neq \lambda_t$. The implication is that the distance between the expected λ_{t+1}^* and λ_{t+1} is smaller than the distance between λ_t^* and λ_t multiplied by the autoregressive coefficient β. The interpretation is that, due to its mean reverting behavior, λ_{t+1}^* can be predicted to revert towards its unconditional mean and the score-driven parameter update reduces the expected deviation between λ_{t+1}^* and λ_{t+1} more than the reduction due to the predictability implied by mean reversion. We also note that the intercept and autoregressive parameters ω and β are assumed to be the same for λ_t and λ_t^*. In practice, ω and β in (2) are not known and they have to be estimated. Due to the parametric nature of score-driven models, the static parameters are usually estimated by the method of maximum likelihood, see Blasques et al. (2022), under conditions of filter invertibility, see Blasques et al. (2018a).

4. EXAMPLES

4.1. *Beta-t-EGARCH model*

Consider the Student-t scale model with exponential link function

$$y_t = \exp(\lambda_t/2)\varepsilon_t,$$

where ε_t has a Student-t distribution with zero mean, unit variance, and degrees of freedom parameter $2 < \nu < \infty$. For this model, the conditional Fisher information of the parameter of interest, λ_t, is a constant. Using a unit scaling, i.e. $S(\lambda_t) = 1$, leads to the following score innovation

$$S(\lambda_t)s(y_t, \lambda_t) = \frac{(\nu+1)y_t^2}{(\nu-2)\exp(\lambda_t) + y_t^2} - 1.$$

The corresponding model is the Beta-t-EGARCH originally proposed by Harvey & Chakravarthy (2008), see also Harvey (2013), as a model for the scale parameter of a Student-t distribution with $\nu > 0$.

In this example, the conditions of Theorem 1 are satisfied. The conditional log-density of the model, up to additive constants, is

$$\log p(y_t|\lambda_t) = -\frac{\lambda_t}{2} - \frac{\nu+1}{2}\log\left\{1 + \frac{y_t^2}{(\nu-2)\exp(\lambda_t)}\right\}.$$

Assumption 1 can be shown to hold by the dominated convergence theorem, provided that $E_{t-1}(y_t^2) < \infty$ with probability 1. Furthermore, $E_{t-1}(y_t^2) < \infty$ entails that Assumption 2 holds, as the Student-t log-likelihood has a unique maximum with respect to the variance parameter, see Fan et al. (2014). As concerns Assumption 3, we have that

$$\frac{\partial S(\lambda)f_t'(\lambda)}{\partial\lambda} = -E_{t-1}\frac{(\nu-2)(\nu+1)y_t^2\exp(\lambda)}{\{(\nu-2)\exp(\lambda) + y_t^2\}^2},$$

which is strictly negative and uniformly bounded from below by a constant that depends on the degrees of freedom parameter $\nu \in [2, \infty)$.

4.2. *Beta-t-GARCH model*

Consider the following scale model with Student-t innovations

$$y_t = \sqrt{\lambda_t}\varepsilon_t,$$

where ε_t has a standardised Student-t distribution with $\nu > 2$ degrees of freedom. Selecting the inverse of the conditional Fisher information as scaling function, $S(\lambda_t) = 2\lambda_t^2$, yields the following score innovation for the time-varying parameter

$$S(\lambda_t)s(y_t, \lambda_t) = \frac{(\nu+1)y_t^2}{(\nu-2) + y_t^2/\lambda_t} - \lambda_t.$$

The resulting score-driven model is the Beta-t-GARCH model discussed in Creal et al. (2013) and Harvey (2013). 285

For this model, the assumptions of Theorems 1 and 2 are not satisfied, but the assumptions of Theorem 3 hold. The conditional log-density of the model, up to additive constants, is

$$\log p(y_t|\lambda_t) = -\frac{1}{2}\log\lambda_t - \frac{\nu+1}{2}\log\left\{1 + \frac{y_t^2}{(\nu-2)\lambda_t}\right\}.$$

Assumptions 1 and 2 hold as discussed in the previous paragraph for the Beta-t-EGARCH model, while neither Assumption 3 nor Assumption 4 can be directly verified, as

$$\frac{\partial S(\lambda) f_t'(\lambda)}{\partial\lambda} = E_{t-1}\frac{(\nu+1)y_t^4}{\{(\nu-2)\lambda + y_t^2\}^2} - 1.$$

On the other hand, Assumption 5 can hold depending on the shape of the true conditional density. For instance, Assumption 5 is immediately satisfied if the conditional density is correctly specified. We also note that in the limit case $\nu \to \infty$, the Student-t distribution converges to the normal and the score innovation becomes

$$S(\lambda_t)s(y_t, \lambda_t) = y_t^2 - \lambda_t.$$

As discussed in Creal et al. (2013), the resulting score-driven model corresponds to the integrated GARCH(1,1) model by Bollerslev (1986). Theorem 1 holds for this example as Assumption 3 is satisfied, given that

$$\frac{\partial S(\lambda) f_t'(\lambda)}{\partial\lambda} = -1.$$

By Theorem 1, as $c = 1$, the global optimality condition holds for $0 < \alpha < 2$.

Finally, we note that choosing different scaling functions in the score innovation, such as the identity or the square root of the inverse of the conditional Fisher information, leads to different results. For instance, using the latter up to a proportionality constant, i.e. using $S(\lambda_t) = 2\lambda_t$, the score innovation is

$$S(\lambda_t)s(y_t, \lambda_t) = \frac{(\nu+1)y_t^2}{(\nu-2)\lambda_t + y_t^2} - 1$$

and, consequently,

$$\frac{\partial S(\lambda) f_t'(\lambda)}{\partial\lambda} = -E_{t-1}\frac{(\nu+1)(\nu-2)y_t^2}{\{(\nu-2)\lambda + y_t^2\}^2}.$$

By taking the limit $\lambda \to 0$, we can see that $\partial S(\lambda) f_t'(\lambda)/\partial\lambda$ is not bounded from below by a constant, and therefore Assumption 3 does not hold. On the other hand, Assumption 4 is satisfied and the local optimality result of Theorem 2 applies.

4.3. *Exponential Poisson autoregression*

Consider the following Poisson time series model with exponential link function

$$y_t|\lambda_t \sim Po\{\exp(\lambda_t)\},$$

where $Po(\mu)$ denotes a Poisson distribution with mean μ. Selecting the inverse of the conditional Fisher information as scaling function, $S(\lambda_t) = \exp(-\lambda_t)$, leads to the following score innovation for the time-varying parameter λ_t,

$$S(\lambda_t)s(y_t, \lambda_t) = \frac{y_t}{\exp(\lambda_t)} - 1.$$

The resulting score-driven model is equivalent to the Po-EINGARCH model in Gorgi (2018) and it is a special case of the class of Poisson observation-driven models of Davis et al. (2003). See also Blazsek & Escribano (2016) for a generalization of this model to panel data.

For this model, the assumptions of Theorem 1 are not satisfied but, instead, the assumptions of Theorem 2 hold. In particular, the conditional log-density of the model is

$$\log p(y_t|\lambda_t) = y_t\lambda_t - \exp(\lambda_t) - \log(y_t!).$$

Assumption 1 holds provided that $E_{t-1}(y_t) < \infty$ with probability 1. Furthermore, $E_{t-1}(y_t) < \infty$ ensures that the pseudo-true parameter $\lambda_t^* = \log\{E_{t-1}(y_t)\}$ is the unique maximiser of $f_t(\lambda) = E_{t-1}\{\log p(y_t|\lambda)\}$ with probability 1 and therefore Assumption 2 is satisfied. Finally, we notice that

$$\frac{\partial S(\lambda) f_t'(\lambda)}{\partial\lambda} = -\frac{E_{t-1}(y_t)}{\exp(\lambda)}.$$

Therefore, given the open set $\Lambda = \mathbb{R}$, Assumption 3 does not hold as $1/\exp(\lambda) \to \infty$ as $\lambda \to -\infty$ and, furthermore, $E_{t-1}(y_t)$ may not be necessarily bounded by a constant, depending on the true conditional density. Instead, Assumption 4 holds as, for any compact subset $\Lambda_c \subset \mathbb{R}$, we can define the $\mathcal{F}_{t-1}$ measurable random variable $c_t = E_{t-1}(y_t)/\inf_{\lambda\in\Lambda_c}\exp(\lambda)$ that satisfies $-\partial S(\lambda) f_t'(\lambda)/\partial\lambda \leq c_t$ for any $\lambda \in \Lambda_c$.

4.4. *Poisson autoregression*

Consider the Poisson time series model with identity link function

$$y_t|\lambda_t \sim Po(\lambda_t).$$

Choosing the inverse of the conditional Fisher information as scaling function, $S(\lambda_t) = \lambda_t$, leads to the following score innovation

$$S(\lambda_t)s(y_t, \lambda_t) = y_t - \lambda_t.$$

The resulting model is the Poisson autoregression by Fokianos et al. (2009). For this model, the assumptions of Theorem 1 are satisfied. The conditional log-density of the model is

$$\log p(y_t|\lambda_t) = y_t \log(\lambda_t) - \lambda_t - \log(y_t!).$$

Assumption 1 trivially holds provided that $E_{t-1}(y_t) < \infty$ with probability 1. Furthermore, $E_{t-1}(y_t) < \infty$ implies that $\lambda_t^* = E_{t-1}(y_t)$ is the unique maximiser of $f_t(\lambda) =$

$E_{t-1}\{\log p(y_t|\lambda)\}$ with probability 1 and therefore Assumption 2 also holds. Finally, Assumption 3 holds immediately as

$$\frac{\partial S(\lambda) f_t'(\lambda)}{\partial \lambda} = -1.$$

Therefore, Theorem 1 applies with $0 < \alpha < 2$ as $c = 1$.

Contrary to the Beta-t-(E)GARCH examples, where the exponential link function leads to global optimality, in (exponential) Poisson autoregressions, choosing the identity link function leads to global optimality, while the exponential link function leads to local optimality.

4.5. *Student-t location model* 305

Consider the location model with Student-t innovations

$$y_t = \lambda_t + \varepsilon_t,$$

where ε_t has a Student-t distribution with zero mean, ν degrees of freedom and scale σ. Selecting the constant $S(\lambda_t) = \nu/(\nu + 1)$ as a scaling function yields the following score innovation for the time-varying location λ_t,

$$S(\lambda_t)s(y_t, \lambda_t) = \frac{(y_t - \lambda_t)/\sigma}{1 + (y_t - \lambda_t)^2/\sigma^2\nu}.$$

The resulting model is the score-driven location model given by Harvey & Luati (2014). The conditional log-density of the model, up to additive constants, is

$$\log p(y_t|\lambda_t) = -\frac{\nu + 1}{2} \log \left\{ 1 + \left(\frac{y_t - \lambda_t}{\sqrt{\nu}\sigma} \right)^2 \right\}.$$

Similarly to the case of the Beta-t-GARCH model, Assumptions 1 and 2 hold as long as $E_{t-1}|y_t| < \infty$ with probability 1. In addition, we have that

$$\frac{\partial S(\lambda) f_t'(\lambda)}{\partial \lambda} = E_{t-1} \frac{(y_t - \lambda)^2/\sigma^3\nu - 1/\sigma}{\{1 + (y_t - \lambda)^2/\sigma^2\nu\}^2}, \tag{5}$$

from which we note that Assumptions 3 and 4 do not hold and therefore Theorems 1 and 2 do not apply. Instead, Assumption 5 holds under some conditions on the true conditional distribution. For instance, if $p_t(y)$ is symmetric, the pseudo-true parameter λ_t^* is the conditional expectation 310 of the true conditional distribution. Then, if the parameter σ^2 is close to the variance of y_t and ν is relatively large, Assumption 5 holds and Theorem 3 applies. Another case where Theorem 3 immediately applies is when the Student-t density is correctly specified, as for λ_t close to λ_t^*, $f_t''(\lambda_t)$ in equation (5) is negative and bounded from below by the negative Fisher information evaluated at λ_t^*, i.e. $c = I(\lambda_t^*) = (\nu + 1)(\nu + 3)/\nu^2$, see Harvey & Luati (2014, equation 11). 315 Therefore, the correctly specified Student-t score-driven location model is locally CEV optimal as stated by Theorem 3 for $0 < \alpha < 2\nu^2/\{(\nu + 1)(\nu + 3)\}$. This result is novel in the score-driven literature as no arguments on the range of α are available.

The examples considered in this section show how the specification of a different link function for the time-varying parameter and the selection of the scaling factor give rise to different 320 models and affect the optimality properties of the score-driven parameter update. The examples also illustrate that Assumptions 3-5 form a practical framework to determine if global or local optimality results hold for a specific model. We also note that the results immediately extend to the stationary case as discussed in Section 3.2. For example, the correctly specified stationary Student-t score-driven model for the location parameter is locally CEV optimal as stated by 325

Theorem 3 for $0 < \alpha < 2\beta\nu^2/\{(\nu+1)(\nu+3)\}$, where β is the autoregressive parameter of the stationary score-driven filter.

5. CONCLUDING REMARKS

Time-varying parameters of observation-driven models are often interpreted as misspecified filtering recursions. Theoretical analyses of these models are mainly concerned with their stochastic properties, such as stationarity, ergodicity and invertibility; see Blasques et al. (2018a) for a discussion on the latter. In contrast, for linear models, filtering recursions are often studied according to optimality properties that they possess with respect to criteria such as minimum mean square distance or mean absolute deviation, with the literature that dates back to Kalman (1960). This paper contributes to the literature that aims to investigate optimality properties of non-linear models, according to specific criteria that characterise the non-unique concept of optimality, with focus on the class of score-driven models. The paper provides optimality properties for score-driven filters in the direction of the pseudo-true parameter. The parallel drawn with the stochastic gradient descent method constitutes the basis for possible extensions to the case of multivariate score-driven filters, where the choice of the scaling matrix of the score is an open topic of debate.

APPENDIX

Proof of Theorem 1. First, we show that Assumption 3 implies the so-called co-coercivity of the function $S(\lambda)f_t'(\lambda)$, namely,

$$\{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\}(\lambda_1 - \lambda_2) \leq -\frac{1}{c}\{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\}^2, \quad \forall \lambda_1, \lambda_2 \in \Lambda.$$

In particular, by the mean value theorem we obtain that

$$d_t(\bar{\lambda})(\lambda_1 - \lambda_2) = \{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\},$$

where $d_t(\lambda) = \partial S(\lambda)f_t'(\lambda)/\partial\lambda$ and $\bar{\lambda}$ is a point between λ_1 and λ_2. Assumption 3 imposes that $-c < d_t(\lambda) < 0$ a.s. for any $\lambda \in \Lambda$. Therefore, we immediately obtain the desired result

$$\begin{aligned}\{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\}(\lambda_1 - \lambda_2) &= \frac{1}{d_t(\bar{\lambda})}\{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\}^2 \\ &\leq \sup_{\lambda\in\Lambda}\frac{1}{d_t(\lambda)}\{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\}^2 \\ &\leq -\frac{1}{c}\{S(\lambda_1)f_t'(\lambda_1) - S(\lambda_2)f_t'(\lambda_2)\}^2.\end{aligned}$$

Next, by Assumption 1, we have that

$$\begin{aligned}\{E_{t-1}(\lambda_{t+1}) - \lambda_t^*\}^2 =& \{\lambda_t + \alpha S(\lambda_t)f_t'(\lambda_t) - \lambda_t^*\}^2 \\ =& (\lambda_t - \lambda_t^*)^2 + 2\alpha S(\lambda_t)f_t'(\lambda_t)(\lambda_t - \lambda_t^*) + \alpha^2 S(\lambda_t)^2 f_t'(\lambda_t)^2.\end{aligned}$$

Therefore, from the co-coercivity of $S(\lambda)f_t'(\lambda)$ and accounting that $f_t'(\lambda_t^*) = 0$ a.s. since by Assumption 2 the function f_t is continuously differentiable and λ_t^* is its unique maximiser in the open set Λ, we obtain that

$$\begin{aligned}\{E_{t-1}(\lambda_{t+1})-\lambda_t^*\}^2 \leq & (\lambda_t-\lambda_t^*)^2-\frac{2}{c}\alpha S(\lambda_t)^2 f_t'(\lambda_t)^2+\alpha^2 S(\lambda_t)^2 f_t'(\lambda_t)^2 \\ \leq & (\lambda_t-\lambda_t^*)^2-\alpha\left(\frac{2}{c}-\alpha\right) S(\lambda_t)^2 f_t'(\lambda_t)^2. \qquad (6)\end{aligned}$$

Finally, we notice that $0<\alpha<2/c$ by assumption. Hence, we have that $|E_{t-1}(\lambda_{t+1})-\lambda_t^*|<|\lambda_t-\lambda_t^*|$ if $S(\lambda_t)f_t'(\lambda_t)\neq 0$. Assumption 3 entails that the function $S(\lambda)f_t'(\lambda)$ is strictly decreasing. This, together with $f_t'(\lambda_t^*)=0$, implies that $S(\lambda_t)f_t'(\lambda_t)=0$ if and only if $\lambda_t=\lambda_t^*$. Therefore, we conclude that $|E_{t-1}(\lambda_{t+1})-\lambda_t^*|<|\lambda_t-\lambda_t^*|$ if $\lambda_t\neq\lambda_t^*$ and $|E_{t-1}(\lambda_{t+1})-\lambda_t^*|=0$ if $\lambda_t=\lambda_t^*$. □

Proof of Theorem 2. The proof is equivalent to the proof of Theorem 1 with the difference that the co-coercivity of $S(\lambda)f_t'(\lambda)$ only holds on compact subsets of Λ and the Lipschitz constant c_t is a $\mathcal{F}_{t-1}$-measurable random variable. For any $\lambda_1,\lambda_2\in\Lambda$, we define the compact set $\Lambda(\lambda_1,\lambda_2)=[\min(\lambda_1,\lambda_2),\max(\lambda_1,\lambda_2)]$. Note that $\Lambda(\lambda_1,\lambda_2)\subset\Lambda$ since Λ is convex. Next, by Assumption 4, we obtain that

$$\begin{aligned}\{S(\lambda_1)f_t'(\lambda_1)-S(\lambda_2)f_t'(\lambda_2)\}(\lambda_1-\lambda_2) &= \frac{1}{d_t(\bar\lambda)}\{S(\lambda_1)f_t'(\lambda_1)-S(\lambda_2)f_t'(\lambda_2)\}^2 \\ &\leq -\frac{1}{\tilde c_t(\lambda_1,\lambda_2)}\{S(\lambda_1)f_t'(\lambda_1)-S(\lambda_2)f_t'(\lambda_2)\}^2,\end{aligned}$$

where $\tilde c_t(\lambda_1,\lambda_2)=-\sup_{\lambda\in\Lambda(\lambda_1,\lambda_2)} d_t(\lambda)$. The proof then follows the same argument as in the proof of Theorem 1 by replacing α in the updating equation in (3) with $\alpha_t=\delta/\tilde c_t(\lambda_t,\lambda_t^*)$ for some $\delta\in(0,2)$, which together with

$$\{E_{t-1}(\lambda_{t+1})-\lambda_t^*\}^2\leq(\lambda_t-\lambda_t^*)^2-\alpha_t\left\{\frac{2}{\tilde c_t(\lambda_t,\lambda_t^*)}-\alpha_t\right\}S(\lambda_t)^2 f_t'(\lambda_t)^2$$

entails the desired result. □

Proof of Theorem 3. The proof follows the same argument as the proof of Theorem 1 with the only difference that the result holds in the subset $\Lambda_t^*=\{\lambda\in\Lambda:|\lambda-\lambda_t^*|<\epsilon\}$ instead of the whole parameter set Λ. In particular, we have that $\lambda_t\in\Lambda_t^*$ by assumption and $\lambda_t^*\in\Lambda_t^*$ by the definition of the set Λ_t^*. Assumption 5 implies that there is a small enough ϵ such that Assumption 3 holds for the set Λ_t^* instead of Λ. Finally, Assumptions 1 and 2 hold also for the set Λ_t^* as they hold for the set Λ and Λ_t^* is a subset of Λ. □

Proof of Corollary 1. First, we obtain that the conditional variance of λ_{t+1} given $\mathcal{F}_{t-1}$ is

$$Var_{t-1}(\lambda_{t+1})=\alpha^2 S(\lambda_t)^2 g_t(\lambda_t),$$

where $g_t(\lambda) = Var_{t-1}\{s(y_t, \lambda)\}$ and $g_t(\lambda) < \infty$ a.s. by assumption. Next, we show that the result holds under the conditions of Theorem 1. From the inequality in equation (6), we obtain

$$\begin{aligned} E_{t-1}\{(\lambda_t^* - \lambda_{t+1})^2\} &= Var_{t-1}(\lambda_{t+1}) + \{E_{t-1}(\lambda_{t+1}) - \lambda_t^*\}^2 \\ &\le Var_{t-1}(\lambda_{t+1}) + (\lambda_t - \lambda_t^*)^2 - \alpha\left(\frac{2}{c} - \alpha\right) S(\lambda_t)^2 f_t'(\lambda_t)^2 \\ &\le (\lambda_t - \lambda_t^*)^2 + \alpha S(\lambda_t)^2 \left[\alpha\{g_t(\lambda_t) + f_t'(\lambda_t)^2\} - \frac{2}{c}\right]. \end{aligned}$$

Therefore, it follows that $E_{t-1}\{(\lambda_t^* - \lambda_{t+1})^2\} < (\lambda_t - \lambda_t^*)^2$ when $\alpha < 2/[c\{g_t(\lambda_t) + f_t'(\lambda_t)^2\}]$ and $\lambda_t \neq \lambda_t^*$. Finally, we note that the result also holds under the conditions of either Theorem 2 or Theorem 3 instead of Theorem 1 based on an equivalent argument. □

REFERENCES

AKAIKE, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle*. New York, NY: Springer New York, pp. 199–213.

AYALA, A. L., BLAZSEK, S. & LICHT, A. (2023). Score function scaling for qar plus beta-t-egarch: an empirical application to the s&p 500. *Applied Economics* , 1–14.

BLASQUES, F., GORGI, P., KOOPMAN, S. J., WINTENBERGER, O. et al. (2018a). Feasible invertibility conditions and maximum likelihood estimation for observation-driven models. *Electronic Journal of Statistics* **12**, 1019–1052.

BLASQUES, F., KOOPMAN, S. J. & LUCAS, A. (2015). Information-theoretic optimality of observation-driven time series models for continuous responses. *Biometrika* **102**, 325–343.

BLASQUES, F., KOOPMAN, S. J. & LUCAS, A. (2018b). Amendments and corrections: 'information-theoretic optimality of observation-driven time series models for continuous responses'. *Biometrika* **105**, 753–753.

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**, 211–223.

BLASQUES, F., VAN BRUMMELEN, J., KOOPMAN, S. J. & LUCAS, A. (2022). Maximum likelihood estimation for score-driven models. *Journal of Econometrics* **227**, 325–346.

BLAZSEK, S. & ESCRIBANO, A. (2016). Score-driven dynamic patent count panel data models. *Economics Letters* **149**, 116–119.

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

CATANIA, L. (2021). Dynamic adaptive mixture models with an application to volatility and risk. *Journal of Financial Econometrics* **19**, 531–564.

CATANIA, L. & BILLÉ, A. G. (2017). Dynamic spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Applied Econometrics* **32**, 1178–1196.

CATANIA, L. & LUATI, A. (2022). Semiparametric modeling of multiple quantiles. *Journal of Econometrics (forthcoming)* .

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

CREAL, D., KOOPMAN, S. J. & LUCAS, A. (2013). Generalized autoregressive score models with applications. *Journal of Applied Econometrics* **28**, 777–795.

DAVIS, R. A., DUNSMUIR, W. T. & STREET, S. B. (2003). Observation-driven models for poisson counts. *Biometrika* **90**, 777–790.

DURBIN, J. & KOOPMAN, S. J. (2012). *Time series analysis by state space methods*. Oxford University Press.

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

FAN, J., QI, L. & XIU, D. (2014). Quasi-maximum likelihood estimation of garch models with heavy-tailed likelihoods. *Journal of Business & Economic Statistics* **32**, 178–191.

FOKIANOS, K., RAHBEK, A. & TJØSTHEIM, D. (2009). Poisson autoregression. *Journal of the American Statistical Association* **104**, 1430–1439.

GASPERONI, F., LUATI, A., PACI, L. & D'INNOCENZO, E. (2023). Score-driven modeling of spatio-temporal data. *Journal of the American Statistical Association* **118**, 1066–1077.

GORGI, P. (2018). Integer-valued autoregressive models with survival probability driven by a stochastic recurrence equation. *Journal of Time Series Analysis* **39**, 150–171.

GORGI, P. (2020). Beta–negative binomial auto-regressions for modelling integer-valued time series with extreme observations. *Journal of the Royal Statistical Society, series B* **82**, 1325–1347.

HARVEY, A. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge: Cambridge University Press.

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

HARVEY, A. & CHAKRAVARTHY, T. (2008). Beta-t-(e)garch. *Working paper, University of Cambridge* .

HARVEY, A. & LUATI, A. (2014). Filtering with heavy tails. *Journal of the American Statistical Association* **109**, 1112–1122.

KALMAN, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering* **82**, 35–45.

KOOPMAN, S. J., LUCAS, A. & SCHARTH, M. (2016). Predicting time-varying parameters with parameter-driven and observation-driven models. *The Review of Economics and Statistics* **98**, 97–110.

PATTON, A. J., ZIEGEL, J. F. & CHEN, R. (2019). Dynamic semiparametric models for expected shortfall (and value-at-risk). *Journal of econometrics* **211**, 388–413.

VAN OS, B., LANGE, R.-J. & VAN DIJK, D. (2022). Robust observation-driven models using proximal-parameter updates. *TI discussion paper* .

WHITE, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica* **50**, 1–25.