Robust Quasi Score-Driven Models with an Adaptive Loss Function

Mathijs Dijkstra, 14465671
Supervisor: Ramon de Punder

June, 2025

Abstract

The score-driven framework has gained substantial popularity in recent years, with nearly 400 published papers on the topic. Despite its success, the framework is limited by its reliance on pre-specified scoring rules tied to assumed conditional densities, which may lead to model misspecification in the presence of outliers or structural breaks. This paper demonstrates that incorporating a flexible, data-adaptive loss function, specifically the Barron loss function, into the quasi score-driven (QSD) framework enables the model to determine the appropriate level of robustness from the data. The resulting QSD filter is shown to generate a strictly stationary, ergodic, and invertible sequence of time-varying parameters under mild regularity conditions. Importantly, the Barron loss function is shown to be strictly consistent for different statistical functionals depending on the value of γ, offering a flexible trade-off between bias and robustness. A Monte Carlo study on both clean and contaminated data reveals that the proposed QSD filter has superior performance over the GARCH model and the β_t GARCH$(1,1)$ model in the presence of outliers. These results highlight the advantages of data-driven robustness in time series modeling, especially when dealing with heavy-tailed or contaminated data.

This document is written by Student Mathijs Dijkstra who declares to take full responsibility for the contents of this document. I declare that the text and the work presented in this document are original and that no sources other than those mentioned in the text and its references have been used in creating it. I have not used generative AI (such as ChatGPT) to generate or rewrite text. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

1 Introduction

Time-varying parameter models are crucial in econometrics for capturing the evolving dynamics commonly observed in economic and financial time series. It is often observed that volatility typically appears in clusters, with periods of high volatility followed by high volatility and low volatility followed by low volatility (Mandelbrot 1963). To accurately model these behaviours, it is necessary to allow parameters to adjust over time in order to more accurately reflect the underlying stochastic process (Engle 1982). However, extreme observations can distort these parameter estimates. If outliers are ignored, models may fail to pick up structural breaks. On the other hand, too much influence of outliers on the model will make it vulnerable to react too much to anomalies in the data. This tension motivates the search for a framework that adaptively balances sensitivity and robustness.

Score-driven (SD) models (Creal et al. 2013; Harvey 2013) update the time-varying parameters using the scaled score of an assumed conditional density. This framework can, in some cases, naturally down weight large deviations under heavy-tailed specifications. Quasi score-driven (QSD) model (Blasques et al. 2023) generalizes the score driven framework by decoupling the update step for the time-varying parameters from the assumed conditional density, allowing for more general loss functions (e.g. Huber-Charbonnier). Yet both approaches require a pre-specified conditional density or loss function, risking misspecification.

In this paper, the loss function proposed by Barron (2019) is used in the Quasi score-driven framework. The loss function has a parameter that can be learned, letting the data decide on the optimal robustness. This added flexibility helps to avoid misspecification beforehand. In practice, this approach has implications in, for example, financial market analysis, where modelling should be accurate during both normal periods and crisis situations. Occasional extreme movements should not trigger structural breaks in the model.

1.1 Time-varying parameters

An example of an early model that allows for time-varying parameters is the Autoregressive Conditional Heteroscedasticity model (ARCH) Engle (1982) and later the generalized autoregressive conditional heteroscedasticity model (GARCH) (Bollerslev 1986). This model incorporates a time-dependent conditional variance parameter. Many variations on the GARCH model have been proposed, such as the GARCH-M model (Engle et al. 1987), the GARCH-t (Bollerslev 1987) and the GJR-GARCH model (Glosten et al. 1993).

A more recent and flexible framework is the class of score-driven models, also referred to as Generalized Autoregressive Score (GAS) models (Creal et al. 2013; Harvey 2013). These models unify and extend many existing time-varying parameter models, including GARCH, as special cases. The models update the time-varying parameters based on the scaled score function. Which is the derivative of the log-likelihood of the assumed conditional density of the data, with respect to the time-varying parameter. The scaled score function defines a direction and scale for improving the model's fit at a time t, considering the current value of the time-varying parameter. This approach is analogous to gradient ascent. The time-varying parameter ϑ_t is updated as follows

$$\vartheta_{t+1} = \omega + \sum_{i=1}^{p} A_i s_{t-i+1} + \sum_{j=1}^{q} B_j \vartheta_{t-j+1}$$

$$s_t = S_t \cdot \nabla_t, \quad \nabla_t = \frac{\partial \log(p(y_t|\vartheta_t, \mathcal{F}_{t-1}; \theta))}{\partial \vartheta_t}, \quad S_t = S(t, \vartheta_t; \theta)$$

In this formulation, θ is a vector of static parameters, A_i and B_i have appropriate dimensions and $p(\cdot|\vartheta_t, \mathcal{F}_{t-1}; \theta)$ is the probability distribution of y_t. Furthermore, Creal et al. (2013) defines a scaling function $S(t, \vartheta_t, \theta)$ which can, for example, be proportional to the variance of the score.

The GAS models have been widely used and have been published in nearly 400 articles, see www.gasmodel.com. With reviews on the topic by Artemova et al. (2022); Harvey (2022). An advantage of the GAS framework is its reliance on the full conditional density, not only first- or second-order moments of the data y_t. Under the assumption of fat-tailed distributions, the score naturally down-weights large deviating observations, preventing excessive influence from outliers. Empirical evidence for the efficiency of the score-driven approach is provided by Creal et al. (2013), while De Punder et al. (2024) shows that the score-driven updates are unique in providing an improvement under an expected Kullback and Leibler divergence measure. This further motivates the adoption of these models.

More recently, the score-driven framework has been generalized by Blasques et al. (2023). The authors decouple the updating equation of the time-varying parameters from the assumed conditional density of the innovations, yielding quasi score-driven models. This generalization retains GAS models as a subset, while encompassing alternative specifications enabled by loss functions commonly used in robust statistics. Consequently, the QSD models can reduce the effect of outliers independent of the assumed conditional distribution, effectively performing a form of winsorisation through robust loss functions such as the Huber-Charbonnier and generalized Charbonnier losses.

Various approaches have been proposed to increase robustness in models with time-varying parameters prior to QSD. As Huber (1981) notes, a robust estimator should be more sensitive to inliers than to outliers. Building on this principle, Hill (2015) applied a trimmed quasi-maximum likelihood estimator that discards the most extreme squared residuals. Blasques et al. (2024), on the other hand, models the error variable as a mixture of normal distributions, which can accurately approximate many continuous error distributions. This ensures a flexible framework for location models, allowing for robust filtering.

Despite advances, existing robustification techniques suffer from inflexibility: they often depend on pre-specified tuning constants or fixed loss shapes not learned from the data. The main problem is a trade-off between the robustness to outliers and the sensitivity of the model. Excessive down-weighting of extreme values can ignore genuine structural breaks or regime shifts, while insufficiently robust methods leave the model vulnerable to distortions from anomalous data points.

1.2 The Barron loss function

The use of robust loss functions is commonly used to reduce the sensitivity of estimators to outliers. The basis of robust estimation theory was formed by Huber (1964), with the introduction of M-estimators, which are a generalization of the maximum likelihood estimators. A widely used example is the Huber loss function (Huber 1981), which is a piecewise function that is quadratic for small residuals, but linear for large residuals. A smooth version of this loss function is known as Pseudo-Huber loss or Charbonnier loss (Charbonnier et al. 1997).

These loss functions have been effective in robustifying estimators and have been adopted in a variety of models. However, their performance is sensitive to the choice of parameter values, which are usually fixed and not learned from data. Moreover, their shapes impose fixed trade-offs between bias and robustness that may not align with the underlying data-generating process. This motivates the search for a more flexible and data-adaptive loss function, leading to the loss function proposed by Barron (2019).

The loss function proposed by Barron (2019), is a more general robust loss function (Barron loss function henceforth) that unifies many existing robust loss functions, including the Cauchy, Charbonnier, and Huber, as special cases. Through this generalisation, the loss function has more expressive power. The authors introduce a parameter γ which can be adapted to obtain different levels of robustness. Lower values of γ make the loss function less responsive to large deviations, limiting the impact outliers have.

Barron (2019) motivates that through the use of their loss function, no choice on the robustness of the loss function has to be made beforehand. The proposed loss function, with its parameter γ, can be estimated such that the data can determine the best level of robustness itself. This is achieved by framing the loss function as a negative log-likelihood of a probability distribution. They show that the maximisation of this likelihood function allows for the parameter γ to be estimated as part of the estimation of the model. The use of this form has been found effective, especially in models with multivariate output spaces. In this case, the loss function allows for independent robustness variables that can independently adapt the robustness of its loss function in each dimension. The authors proceed to show that the loss function results in improved outcomes in image synthesis and in neural networks for unsupervised molecular depth estimation.

This paper aims to make the updating step of the quasi score-driven models more robust. The restraining choice of loss function can be too restrictive and is prone to misspecification. Through the use of the loss function proposed by Barron (2019) a more flexible model is created. This flexibility enables greater freedom in the robustness of the updating equation. It is therefore hypothesized that embedding the adaptive-robust loss of Barron (2019) within the quasi score-driven framework will deliver improved performance, both in terms of parameter stability under heavy-tailed shocks and outlier resilience, compared to QSD models with pre-specified loss functions and the benchmark GARCH and β_t GARCH$(1,1)$ models.

The remainder of this paper is set up as follows. In Section 2 the formulation of the QSD filter is proposed along with some theoretical properties of the loss function (including some illustrative examples). Furthermore, this section gives theoretical details on the QSD filter. In addition, the setup of the Monte Carlo study is outlined. Section 3 discusses the results of the Monte Carlo study and provides some visual insights. Finally, Section 4 concludes.

2 A robust QSD filter

2.1 Formulation

In this paper, a model is proposed that is based on the quasi score-driven (QSD) formulation provided by Blasques et al. (2023). First, consider a time series $\{y_t\}_{t\in\mathbb{Z}}$ with a conditional density $p(\cdot|\vartheta_t,\theta)$, where $\theta \in \Theta \subseteq \mathbb{R}^q$ is a vector of constant parameters of length q and a time-varying parameter $\{\vartheta_t\}_{t\in\mathbb{Z}}$. The conditional distribution $p(\cdot|\vartheta_t,\theta)$ is often expressed by means of a time series model of the form $y_t = g(\vartheta_t, \varepsilon_t)$. In this formulation, ε_t is an i.i.d. random variable that can be interpreted as an error term. Examples are $g(\vartheta_t,\varepsilon_t) = \vartheta_t + \varepsilon_t$ for a location model and $g(\vartheta_t,\varepsilon_t) = \sqrt{\vartheta_t}\varepsilon_t$ for a volatility model. Blasques et al. (2023), proposes the following updating equation for the time-varying parameter ϑ_t

$$\vartheta_{t+1} = \omega + \alpha\psi(y_t, X_t, \vartheta_t, \theta) + \beta\vartheta_t \tag{1}$$

Here ω, α, β are real parameters and X_t is a vector of exogenous variables. Furthermore, ψ is a measurable function. The specification of the function ψ is very general and various specifications are possible. Consider, for example, the following specification,

$$\psi(y_t, \vartheta_t, \theta) := \frac{\partial \log(p(y_t|\vartheta_t, \theta))}{\partial \vartheta_t} S(\vartheta_t)$$

The resulting updating equation is now equal to that of the score-driven models of Creal et al. (2013).

In this paper, a different formulation of ψ is proposed based on the loss function derived by Barron (2019) denoted by $\rho\left(e(y_t, \vartheta_t), \gamma, \xi\right)$. Because the formulation of ϑ_t is in line with a form of gradient ascent, the scoring rule is used $(-\rho\left(e(y_t, \vartheta_t), \gamma, \xi\right))$. This results in the following specification

$$\psi(y_t, \vartheta_t, \theta) := -\frac{\partial \rho\left(e(y_t, \vartheta_t), \gamma, \xi\right)}{\partial \vartheta_t} = -\frac{\partial \rho\left(e(y_t, \vartheta_t), \gamma, \xi\right)}{\partial e(y_t, \vartheta_t)} \cdot \frac{\partial e(y_t, \vartheta_t)}{\partial \vartheta_t}$$

$e(y_t, \vartheta_t)$ is a function that measures forecast error. One could take $e(y_t, \vartheta_t) = y_t - \vartheta_t$ for location models (Blasques et al. 2023) and $e(y_t, \vartheta_t) = y_t^2 - \vartheta_t$ volatility models with mean zero (Andersen and Bollerslev 1998). The loss function, $\rho(e(y_t, \vartheta_t), \gamma, \xi)$, is given by

$$-\rho(e(y_t, \vartheta_t), \gamma, \xi) = \begin{cases} -\frac{1}{2}\Big(\frac{e(y_t, \vartheta_t)}{\xi}\Big)^2, & \gamma = 2, \\ -\log\Big(\frac{1}{2}\big(\frac{e(y_t,\vartheta_t)}{\xi}\big)^2 + 1\Big), & \gamma = 0, \\ \exp\Big(-\frac{1}{2}\big(\frac{e(y_t,\vartheta_t)}{\xi}\big)^2\Big) - 1, & \gamma = -\infty, \\ -\frac{|\gamma - 2|}{\gamma}\left(\left(\frac{e(y_t,\vartheta_t)^2}{\xi^2|\gamma-2|} + 1\right)^{\gamma/2} - 1\right), & \text{otherwise.} \end{cases} \tag{2}$$

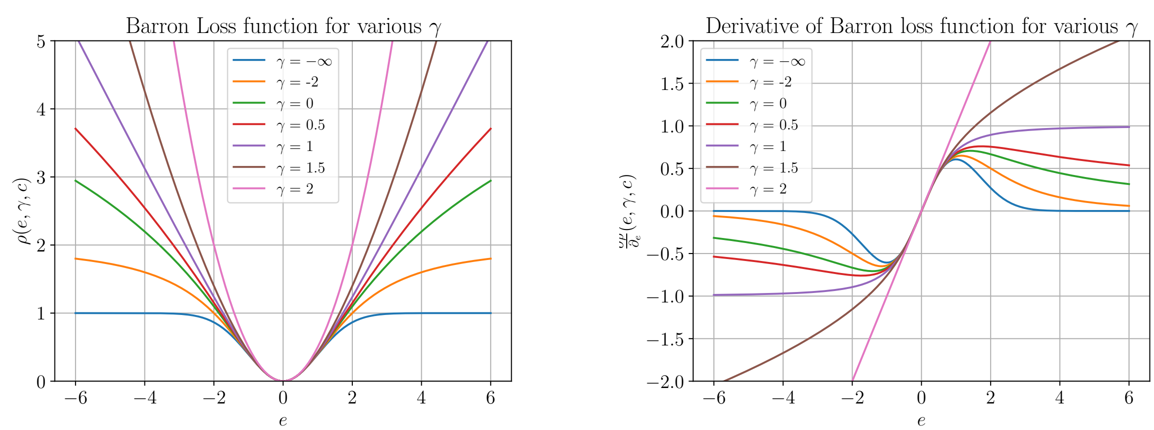
And its first order derivative with respect to the time varying parameter ϑ_t

$$\psi(y_t, \vartheta_t, \theta) = -\frac{\partial \rho(e(y_t, \vartheta_t), \gamma, \xi)}{\partial \vartheta_t} = \begin{cases} -\frac{e(y_t, \vartheta_t)}{\xi^2} \cdot \frac{\partial e(y_t, \vartheta_t)}{\partial \vartheta_t}, & \gamma = 2, \\ -\frac{2e(y_t, \vartheta_t)}{e(y_t, \vartheta_t)^2 + 2\xi^2} \cdot \frac{\partial e(y_t, \vartheta_t)}{\partial \vartheta_t}, & \gamma = 0, \\ -\frac{e(y_t, \vartheta_t)}{\xi^2} \exp\Big(-\frac{1}{2}\left(\frac{e(y_t, \vartheta_t)}{\xi}\right)^2\Big) \cdot \frac{\partial e(y_t, \vartheta_t)}{\partial \vartheta_t}, & \gamma = -\infty, \\ -\frac{e(y_t, \vartheta_t)}{\xi^2}\left(\frac{e(y_t, \vartheta_t)^2}{\xi^2 |\gamma - 2|} + 1\right)^{\frac{\gamma}{2} - 1} \cdot \frac{\partial e(y_t, \vartheta_t)}{\partial \vartheta_t}, & \text{otherwise.} \end{cases} \tag{3}$$

2.2 The loss function

To support the use of the loss function proposed by Barron (2019), properties of this loss function are described. For simplicity, write $e_t = e(y_t, \vartheta_t)$

Figure 1: The Barron Loss Function and its Gradient



(a) General loss function

(b) Derivative of general loss function

NOTE: The loss function (a) and its gradient (b) for different values of γ, the parameter ξ is fixed to 1. Notice that for $\frac{\partial e(y_t, \vartheta_t)}{\partial \vartheta_t} = -1$ (this holds for both location and volatility models), the derivative shown here is equivalent to the formulation of $\psi(y_t, \vartheta_t, \theta)$.

Figure 1 shows the Barron loss function and the first order derivative of the Barron loss function for different values of γ. These values reproduce several existing loss functions: L2 loss ($\gamma = 2$), Charbonnier loss ($\gamma = 1$), Cauchy loss ($\gamma = 0$), Geman-McClure loss ($\gamma = -2$) and Welsch loss ($\gamma = -\infty$). From Figure 1, it can be seen that lower values of γ decrease the effect that large values of e_t have. This property also holds for the derivative of the loss function, which is useful since this derivative is used in the updating equation for the time-varying parameters.

Barron (2019) further note some useful properties that will be used to show the ability of the model to generate a stationary sequence, and bounded unconditional moments (see Section 2.3). For instance, the loss function is smooth (C^∞) with respect to the parameters e_t, γ, ξ. Other useful properties are

$$\left|\frac{\partial \rho}{\partial e_t}(e_t, \gamma, \xi)\right| \leq \begin{cases} \frac{1}{\xi}\left(\frac{\gamma-2}{\gamma-1}\right)^{\left(\frac{\gamma-1}{2}\right)} \leq \frac{1}{\xi} & \text{if } \gamma \leq 1 \\ \frac{|e_t|}{\xi^2} & \text{if } \gamma \leq 2 \end{cases}$$

and

$$\frac{\partial^2 \rho}{\partial e_t^2}(e_t, \gamma, \xi) \leq \frac{1}{\xi^2}$$

This means that for values of γ less than 1, the derivatives are bounded, which prevents exploding gradients. For $\gamma \leq 2$ the gradient is bounded by a function that is linearly dependent on e_t. The loss function also has the shape parameter $\xi > 0$, which determines the shape of the quadratic bowl near $e_t = 0$.

Barron (2019) explain that the first order derivative gives some intuition about how γ effects the behaviour of the loss function when it is being minimised by gradient descent. For all values of γ, the derivative is approximately linear when $|e_t| < \xi$. This means that the effect of the residual is linearly proportional to its magnitude. If $\gamma = 2$, the derivative stays linearly proportional to its magnitude. If γ is decreased to equal 1, then the derivative tends to the value of $1/\xi$ as $|e_t|$ increases. For $\gamma < 1$, the derivative is said to be redescending when $|e_t| > \xi$. The parameters γ, ξ allow for great flexibility, which can give a tailored robustness of the loss function in the QSD framework.

2.2.1 Mathematical Intuition and Examples

Example 1 and 2 give some intuition on the effect of γ on the size of the updating step of the time varying parameter.

Example 1. *Consider the simple scenario in which the parameters are fixed to $\omega = 0$ and $\beta = 1$ such that the update of ϑ_t is given by, $\vartheta_{t+1} = \alpha\psi(e(y_t, \vartheta_t), \gamma, \xi) + \vartheta_t$. For* *a location model with* $e(y_t, \vartheta_t) = y_t - \vartheta_t$*, and in the cases that* $\gamma \neq \{-\infty, 0, 2\}$*,* ψ *is given by*

$$\psi(e(y_t, \vartheta_t)) = \frac{(y_t - \vartheta_t)}{\xi^2}\left(\frac{(y_t - \vartheta_t)^2}{\xi^2|\gamma - 2|} + 1\right)^{\frac{\gamma}{2}-1}$$

For different values of γ*,* ψ *can be rewritten to give some intuition into the updating size given by* ψ *through an adaptive scaling of* $\frac{(y_t - \vartheta_t)}{\xi^2}$*. For instance, when* $\gamma = 1$*, the formulation of* ψ *reduces to*

$$\psi(e(y_t, \vartheta_t)) = \frac{(y_t - \vartheta_t)}{\xi^2}\frac{1}{\sqrt{\frac{(y_t - \vartheta_t)^2}{\xi^2} + 1}}$$

Defining the scaling term as $S_t = \frac{1}{\sqrt{\frac{(y_t - \vartheta_t)^2}{\xi^2} + 1}}$*, which satisfies* $0 < S_t \leq 1$ *as* $\sqrt{\frac{(y_t - \vartheta_t)^2}{\xi^2} + 1} \geq 1$*, this down scales* $\frac{y_t - \vartheta_t}{\xi^2}$*, where the size of* $y_t - \vartheta_t$ *determines the level of the down scaling.*

If γ *is lowered further, the size of down scaling increases even more. For example, if* $\gamma = -1$ *then the formulation of* ψ *reduces to*

$$\psi(e(y_t, \vartheta_t)) = \frac{(y_t - \vartheta_t)}{\xi^2}\frac{1}{\left(\frac{(y_t - \vartheta_t)^2}{3\xi^2} + 1\right)^{\frac{3}{2}}}$$

this results in a greater down scaling of $\frac{y_t - \vartheta_t}{\xi^2}$*.*

Conversely, increasing γ *to 2, eliminates the scaling term entirely, resulting in* $\psi = \frac{y_t - \vartheta_t}{\xi^2}$*. This formulation of* ψ *does not downscale* $\frac{y_t - \vartheta_t}{\xi^2}$ *at all. Note also that* $\psi = \frac{y_t - \vartheta_t}{\xi^2}$ *is proportionally equal to the updating equation of a GAS model that assumes that the error term is normally distributed. Revealing an interesting observation, the proposed QSD filter reduces to famous GAS models for certain values of* γ*, a detailed description can be found in Appendix G.*

These examples illustrate that smaller values of γ *lead to stronger down weighting of* $y_t - \vartheta_t$*, resulting in smaller parameter updates and hence increased robustness.*

The robustness mechanism demonstrated in the location model (Example 1) applies equally to volatility modeling, as shown in the following example.

Example 2. *The same intuition of the down scaling holds for a volatility model where* $e(y_t, \vartheta_t)$ *is taken to be* $y_t^2 - \vartheta_t$*. Consider again the simple scenario in which the parameters are fixed to* $\omega = 0$*, and* $\beta = 1$ *such that the update of* ϑ_t *is given by,* $\vartheta_{t+1} =$

$\alpha\psi(e(y_t,\vartheta_t),\gamma,\xi)+\vartheta_t$. In the cases that $\gamma \neq \{\infty, 0, 2\}$, ψ is given by

$$\psi(e(y_t,\vartheta_t)) = \frac{(y_t^2-\vartheta_t)}{\xi^2}\left(\frac{(y_t^2-\vartheta_t)^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-1}$$

For different values of γ, the size of the update of the time-varying parameter differs. For example when $\gamma = 1$, the formulation of ψ reduces to

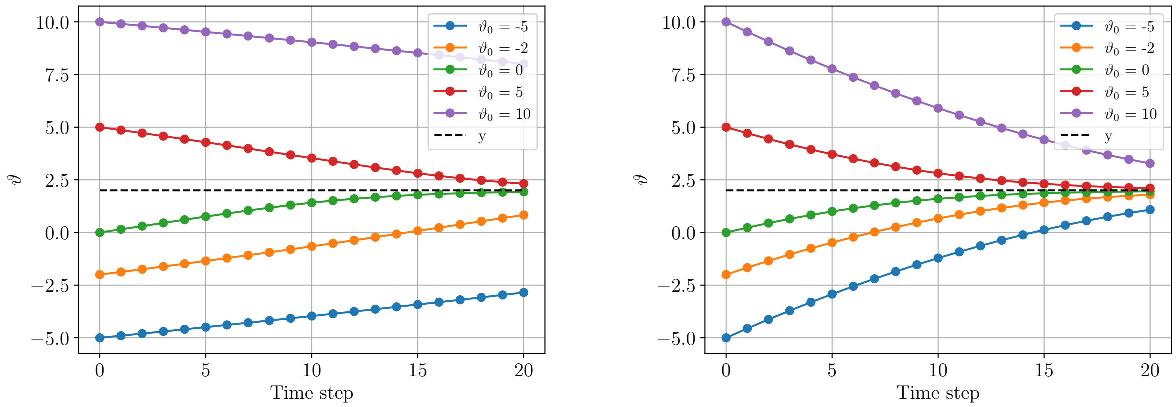
$$\psi(e(y_t,\vartheta_t)) = \frac{(y_t^2-\vartheta_t)}{\xi^2}\frac{1}{\sqrt{\frac{(y_t^2-\vartheta_t)^2}{\xi^2}+1}}$$

This is similar to Example 1 and the scaling factor $S_t = \frac{1}{\sqrt{\frac{(y_t^2-\vartheta_t)^2}{\xi^2}+1}}$ has the same effect, it down weighs $\frac{y_t^2-\vartheta_t}{\xi^2}$. Resulting in smaller updating steps compared to the case where $\gamma = 2$ in which $\psi = \frac{y_t^2-\vartheta_t}{\xi^2}$. In this case the model is proportional to the GARCH model (see Appendix G for more details). Thus, in the volatility model, the same reasoning applies: smaller values of γ lead to less responsive updates, thereby improving robustness.

Examples 1 and 2 highlight two notable properties of the proposed QSD filter. First, the filter nests well-known GAS models as special cases. Second, the updating equation can be reformulated to resemble a GAS model with a normally distributed error term, but with an adaptive scaling function. This scaling function, denoted S_t, is $\mathcal{F}_t$-measurable, whereas most scaling functions considered by Creal et al. (2013) are $\mathcal{F}_{t-1}$-measurable. For example, a typical choice in their framework is $\mathcal{I} = \mathbb{E}[\nabla_t\nabla_t' \mid \mathcal{F}_{t-1}]$, where ∇_t denotes the score (i.e., the derivative of a pre-specified log-likelihood function).

Figure 2 furthermore illustrates the trajectory of the estimate ϑ_t under the recursion (1) when the same observation $y_t = 2$ is received for 20 consecutive time steps. Assuming a location model with $e_t = y_t - \vartheta_t$, we see that ϑ_t converges to the true value $y = 2$, as expected. The effect of γ is immediately visible, a larger value of γ (2b) produces larger update steps, whereas a smaller γ yields more conservative adjustments. This highlights the role of γ in controlling the sensitivity of the loss to large residuals, thereby providing robustness against outliers.

Figure 2: Convergence Dynamics under Different Values of γ



(a) $\gamma = 0.5$ (b) $\gamma = 1.5$

NOTE: Convergence of the estimate ϑ_t under the recursion (1) when observing $y_t = 2$ for 20 timesteps. The different curves correspond to different starting values of ϑ_0. In this example, ξ is set to 1. Figure (a), $\gamma = 0.5$ and figure (b), $\gamma = 1.5$ demonstrate that larger γ, yields larger update steps and faster convergence, while smaller γ produces more conservative adjustments.

2.2.2 Consistency of the loss function

Figure 2, Example 1 and 2, illustrate how different values of γ affect the dynamics of the time-varying estimate ϑ_t. Beyond the impact on update speed and robustness, these differences also reflect a deeper statistical property of the Barron loss function, each value of γ corresponds to a different target functional. This relationship can be precisely described using the notion of consistency, as introduced by Gneiting (2011).

Proposition 1. *Let I be the interval representing the potential range of outcomes of a random variable Y, and let $\mathcal{P}$ be a class of probability distributions in the sample space of I. A scoring function $S : I \times I \to [0, \infty)$ is consistent, for a functional $T : \mathcal{P} \to \mathcal{P}_r(I)$ and $F \to T(F) \subseteq I$ if*

$$\mathbb{E}_F[S(t, Y)] \leq \mathbb{E}_F[S(x, Y)] \tag{4}$$

for all probability distributions $F \in \mathcal{P}$, all $t \in T(F)$, and all $x \in I$.

The scoring function S is strictly consistent for the functional T if it is consistent, and equality in the above expression implies that $x \in T(F)$.

The Barron loss function evaluated in the error $e(y_t, \vartheta_t) = y_t - \vartheta_t$ is a strictly consistent estimator for the mean ($T(F) = \mathbb{E}[y_t]$) if and only if $\gamma = 2$. For all other $\gamma \neq 2$, the loss function is not consistent for the mean. This intentional inconsistency ensures that the loss function is more robust to outliers by means of less variable estimates of the parameters. The same result holds for $e(y_t, \vartheta_t) = y_t^2 - \vartheta_t$. In this case, the Barron loss function is a strictly consistent scoring function for the variance if $\vartheta_t = \mathbb{E}(y_t^2)$ and $\mathbb{E}(y_t) = 0$. A verification of Proposition 1 for the Barron loss function can be found in Appendix A.

Barron (2019) also mentions a similar statement. They argue that minimising their loss in the case when $\gamma = 2$ is equivalent to estimating a mean. Similarly, for $\gamma = 1$, it is similar to estimating the median (see Appendix A.2) and for $\gamma = -\infty$ it is equivalent to local mode finding. All other values of γ can be thought of as smoothly interpolating between these three location measures.

For lower values of γ, the scoring function is consistent for more robust location measures, such as the median and the mode. This property is what gives the Barron loss function its robustness. It induces a small bias relative to the true mean, but at the same time reduces the effect of extreme errors, reducing the variance of the estimates, especially under data with extreme observations. In many realistic settings, that drop in variance more than compensates for the bias, yielding lower overall mean-squared error than the $\gamma = 2$ case.

2.3 Stationarity and invertibility of the QSD model

In this section, a verification of Lemmas 1-4 in Blasques et al. (2023) is done. These lemmas ensure that Theorem 1 in their paper holds, which ensures that the proposed QSD filter generates a stationary data-generating process, and it furthermore ensures consistency and normality of the estimated parameters. In this section, the necessary conditions for the validity of Lemmas 1-4 are shown. The full derivations can be found in Appendix B to E.

2.3.1 Stationarity

Following the notation of Blasques et al. (2023). Denote the exogenous variables, $z_t = (\varepsilon_t, X_t')' \in \mathbb{R}^d$ and note that the time-varying parameter ϑ_t is a recurrence relation of the form $\vartheta_{t+1} = \varphi(z_t, \vartheta_t)$. Conditions for the QSD model to generate a stationary sequence have been proposed by Blasques et al. (2023) in Lemma 1. It states

Proposition 2. *Assume that (z_t) is stationary and ergodic. Suppose that*

$$(i)\quad \mathbb{E}\Big[\log^{+}\big|\psi\big(g(\vartheta^0, \varepsilon_t),\, X_t,\, \vartheta^0,\, \theta\big)\big|\Big] < \infty, \quad \vartheta^0 \in F \subset \mathbb{R}; \tag{5}$$

$$(ii)\quad \mathbb{E}\big[\log \Lambda_t\big] < 0, \quad \Lambda_t = \sup_{\vartheta}\Big|\gamma\,\frac{\partial}{\partial\vartheta}\psi\big(g(\vartheta, \varepsilon_t),\, X_t,\, \vartheta,\, \theta\big) + \beta\Big| \tag{6}$$

Then there exist unique strictly stationary and ergodic solutions to $\{\vartheta_t\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$ to $y_t = g(\vartheta_t, \varepsilon_t)$ and (1).

Given that $y < \infty$, $\left|\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right| < \infty$, $\left|\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\right| < \infty$ (bounded for $\gamma \leq 2$, as noted in Barron (2019)), $\mathbb{E}\left[\log^{+}|\psi(y,\vartheta,\theta)|\right] < \infty$. So, condition (i) is satisfied.

For condition (ii) a distinction is made between two cases that require different assumptions. Using bounds from Barron (2019) and assuming bounded derivatives of $e(y,\vartheta)$

Case 1: $\gamma \leq 1$

$$\frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}\left[\left|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right|\right] + |\beta| < 1$$

Case 2: $1 < \gamma \leq 2$

$$\frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\right] + \frac{|\alpha|}{\xi^2}\mathbb{E}\left[|e(y,\vartheta)|\cdot\left|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right|\right] + |\beta| < 1$$

Under appropriate parameter constraints, Condition (ii) is satisfied and stationarity is ensured. For illustrational purposes, consider a location model: $e(y,\vartheta) = y - \vartheta$, then $\frac{\partial e}{\partial\vartheta} = -1, \quad \frac{\partial^2 e}{\partial\vartheta^2} = 0$. Which results in

$$\frac{|\alpha|}{\xi^2} + |\beta| < 1$$

Note that the same result holds for a volatility model.

2.3.2 Existence of bounded unconditional moments

Lemma 2 in Blasques et al. (2023) gives conditions for unconditional bounded moments.

Proposition 3. *Under the assumptions of Proposition 1, if the sequence $\{\Lambda_t\}$ is i.i.d. and for some $r > 0$*

$$\mathbb{E}\left[\left|\psi\big(g(\vartheta^0, \epsilon_t), X_t, \vartheta^0, \theta\big)\right|^r\right] \; < \; \infty \quad \textit{and} \quad \mathbb{E}\left[\sup_{\vartheta}\left|\frac{\partial\psi\big(g(\vartheta, \epsilon_t), X_t, \vartheta, \theta\big)}{\partial\vartheta}\right|^r\right] \; < \; \infty,$$

then ϑ_t satisfies $\mathbb{E}\,|\vartheta_t|^s < \infty$ for some $s > 0$.

These conditions hold in the proposed model if $\gamma \leq 2$, $\frac{\partial e(y,\vartheta)}{\partial\vartheta} < \infty$,$\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2} < \infty$ and $y_t < \infty$. Furthermore, it is assumed that $e(y_t, \vartheta_t)$ is bounded. Note that these assumptions are mild. For example, in the case of a volatility model where: $e(y_t, \vartheta_t) = y_t^2 - \vartheta_t$, no assumptions on γ have to be made. The same holds for a location model where $e(y_t, \vartheta_t) = y_t - \vartheta_t$.

2.3.3 Invertibility of the QSD filter

To show that the QSD filter is invertible, the following condition from Blasques et al. (2023) is shown to hold,

Proposition 4. *Let $\{y_t, X_t\}_{t\in\mathbb{Z}}$ be stationary and ergodic, and suppose that*

(i) for all $\theta \in \Theta$, there exists $\vartheta^0 \in F$ such that $\mathbb{E}\big[\log^+\big|\psi(y_t, \vartheta^0, \theta)\big|\big] < \infty, \quad \forall\, \theta \in \Theta$.

(ii) $\mathbb{E}\Big[\log \sup_{\vartheta\in F} \sup_{\theta\in\Theta}\Big|\alpha\, \dfrac{\partial}{\partial\vartheta}\psi(y_t, \vartheta, \theta) + \beta\Big|\Big] < 0$.

Then for all $\theta \in \Theta$, there exists a unique stationary and ergodic solution to $\{\vartheta_t(\theta)\}_{\theta\in\Theta}$ to $\vartheta_{t+1}(\theta) = \omega + \alpha\psi(y_t, X_t, \vartheta_t(\theta), \theta) + \beta\vartheta_t(\theta),\ t \in \mathbb{Z}$. Furthermore, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathbb{C}(\Theta, F)$, there exists $\varrho \in (0, 1)$ such that

$$\varrho^{-t} \sup_{\theta\in\Theta}\big|\hat{\vartheta}_t(\theta) - \vartheta_t(\theta)\big| \to 0 \quad \textit{a.s. as} \quad (t \to \infty).$$

The model is then said to be uniformly invertible.

Under similar assumptions as in 2.3.1 the two conditions hold. The needed assumptions are, $\gamma \leq 2$, $\frac{\partial e(y,\vartheta)}{\partial\vartheta} < \infty$,$\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2} < \infty$ and $y_t < \infty$. Furthermore, it is assumed that $e(y_t, \vartheta_t)$ is bounded.

2.3.4 Invertibility properties for the derivatives of the filter

This proposition sets the conditions for stationarity and invertibility of the derivatives of the filter.

Proposition 5. *Let the conditions of Proposition 2.3.3 hold, assume that ψ admits continuous second-order derivatives with respect to its last two components, and suppose that*

$$\textit{(i) for all } \theta \in \Theta, \quad \mathbb{E}\left[\log^+ |\psi_t| + \log^+ \left\| \frac{\partial \psi_t}{\partial \theta} \right\| + \log^+ \left| \frac{\partial \psi_t}{\partial \vartheta} \right| + \log^+ |\vartheta_t(\theta)| \right] < \infty.$$

Then, for all $\theta \in \Theta$, there exists a unique strictly stationary and ergodic solution $\{\vartheta_t'(\theta)\}_{t\in\mathbb{Z}}$ to (8). *If in addition*

$$\textit{(ii)} \quad \mathbb{E}\left[\log^+ \left(\sup_f \left| \frac{\partial \psi_t}{\partial \vartheta} \right| + \sup_{\vartheta,\theta} \left\| \frac{\partial^2 \psi_t}{\partial \theta \partial \vartheta} \right\| + \sup_f \left\| \frac{\partial^2 \psi_t}{\partial \vartheta^2} \right\| + \sup_\theta \|\vartheta_t'(\theta)\| \right) \right] < \infty,$$

then, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, F)$ and $\tilde{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, \mathbb{R}^\rho)$, there exists $\varrho \in (0,1)$ such that

$$\varrho^{-t} \sup_{\theta\in\Theta} \left\| \tilde{\vartheta}_t'(\theta) - \vartheta_t'(\theta) \right\| \to 0 \quad \textit{a.s. as } t \to \infty.$$

furthermore, assuming that

$$\textit{(iii) for all } \theta \in \Theta, \quad \mathbb{E}\left[\log^+ \left\| \frac{\partial^2 \psi_t}{\partial \theta \partial \theta^\top} \right\| + \log^+ \left\| \frac{\partial^2 \psi_t}{\partial \theta \partial \vartheta} \right\| + \log^+ \left\| \frac{\partial^2 \psi_t}{\partial \vartheta^2} \right\| \right] < \infty,$$

then, for all $\theta \in \Theta$, there exists a unique strictly stationary and ergodic solution $\{\vartheta_t''(\theta)\}_{t\in\mathbb{Z}}$ to (10). Under the additional assumption

$$\textit{(iv)} \quad \mathbb{E}\left[\log^+ \left(\sup_{\vartheta,\theta_i,\theta_j} \left| \frac{\partial^3 \psi_t}{\partial \theta_i \partial \theta_j \partial \vartheta} \right| + \sup_{\vartheta,\theta_i} \left\| \frac{\partial^3 \psi_t}{\partial \theta_i \partial \vartheta^2} \right\| + \sup_f \left\| \frac{\partial^3 \psi_t}{\partial \vartheta^3} \right\| \right) \right] < \infty,$$

then, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, F)$, $\tilde{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, \mathbb{R}^\rho)$, and $\tilde{\vartheta}_1''(\cdot) \in \mathcal{C}(\Theta, \mathbb{R}^{\rho^2})$, there exists $\varrho \in (0,1)$ such that

$$\varrho^{-t} \sup_{\theta\in\Theta} \left\| \tilde{\vartheta}_t''(\theta) - \vartheta_t''(\theta) \right\| \to 0 \quad \textit{a.s. as} \quad t \to \infty.$$

In the Appendix it is shown that these conditions hold under mild assumptions. These include bounded derivatives of $e(y_t, \vartheta_t)$ with respect to ϑ, up to and including the fourth derivative and $\gamma \in \Theta \subseteq (-\infty, 2]$, $\xi > 0$ so that all partial derivatives of $\rho(e(y_t, \vartheta_t), \gamma, \xi)$ with respect to $e(y_t, \vartheta_t)$ are bounded.

2.3.5 Implications

For Propositions 2 to 5 to hold, several assumptions are required on the data-generating process. These assumptions are mild and are satisfied by a wide class of models. To summarize, the following conditions are sufficient for the QSD filter to generate a data-generating process: finite values of $y_t \in \mathbb{R}$, the function $e(y_t, \vartheta_t)$ and its derivatives w.r.t. ϑ_t up to fourth order are bounded, $\gamma \in (-\infty, 2]$, $\xi > 0$, and stability conditions like $\frac{|\alpha|}{\xi^2} + |\beta| < 1$ in the location and volatility model ensure stationarity and invertibility. Additionally, the QSD function ψ must be continuously differentiable up to third order with respect to its arguments. This ensures the existence and convergence of the derivative recursions necessary for asymptotic normality of the estimator. Barron (2019) notes that this is satisfied for all the necessary parameters, γ, ξ, e_t as the loss function is C^∞.

Under these assumptions, Propositions 2 to 5, which correspond to Lemmas 1 through 4 in Blasques et al. (2023), are satisfied. Consequently, the QSD model defines a valid and stable data-generating process. Furthermore, it implies that the conditions required for the consistency and asymptotic normality of the parameter estimators, as established in Theorem 1 of Blasques et al. (2023) are met.

Hence, the theoretical foundations for estimation using the QSD model are verified, justifying its application in empirical settings.

2.4 Estimation

Estimation of the models can be done in various ways depending on the assumptions that are made. With minimal assumptions, one could use Quasi-likelihood estimation (QLE)(Durbin 1960; Godambe 1960). In this specification, no assumption on the underlying distribution of y_t has to be made. If one is willing to assume the underlying distribution of the data-generating process, maximum likelihood estimation (MLE) is readily available. In the MLE framework, a proposal by Barron (2019) could be used. The authors define a general probability distribution based on their proposed loss function. This could in principle be used in the score-driven framework of Creal et al. (2013) and Harvey (2013). But requires a complex framework using cubic hermite splines due to non-differentiability issues in the parameter γ (see Barron (2019) for the details).

2.4.1 Quasi-likelihood estimation (QLE)

Under weak assumptions, θ can be estimated using QLE. Blasques et al. (2023) describe the QLE estimation process as follows. The authors define an unbiased estimating function $h_t = h_t(\theta) \in \mathbb{R}$ which depends on y_t and $\vartheta_t = \vartheta_t(\theta)$. Specifically, consider the case where $\vartheta_t = \mathbb{E}_{t-1}[y_t^k]$ for some $k > 0$. In this formulation, location models correspond to $k = 1$ and volatility models correspond to $k = 2$. Such that $h_t(\theta) = y_t^k - \vartheta_t(\theta)$. By construction, the following condition holds:

$$\mathbb{E}[h_t|\mathcal{F}_{t-1}] = 0$$

Where $\mathbb{E}(\cdot|\mathcal{F}_{t-1})$ is the conditional expectation given the σ-algebra, $\mathcal{F}_{t-1}$ which is generated by $\{y_u, u < t\}$ (information set up to time t). The estimator of θ can be computed by solving an estimator equation

$$\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \, ||\hat{G}_t(\theta)||$$

$$\hat{G}_t(\theta) = \frac{1}{T} \sum_{t=1}^{T} \frac{\hat{h}_t(\theta)}{\hat{\sigma}_t^2(\theta)} \cdot \frac{\partial \hat{\vartheta}_t(\theta)}{\partial \theta} \tag{7}$$

The estimator equation $\hat{G}_t(\theta)$ contains the functions $\hat{h}_t(\theta), \hat{\vartheta}_t'(\theta), \hat{\sigma}_t^2(\theta)$. These functions are approximations of the functions $h_t(\theta), \vartheta_t'(\theta), \sigma_t^2(\theta)$, which depend on the unknown values of $\{y_t; t \leq 0\}$. The recursion for $\hat{\vartheta}_t(\theta)$ is given by

$$\hat{\vartheta}_{t+1}(\theta) = \omega + \alpha\psi(y_t, \hat{\vartheta}_t(\theta), \theta) + \beta\hat{\vartheta}_t(\theta), \quad \text{for } t \geq 1$$

And $\hat{h}_t(\theta) = y_t^k - \hat{\vartheta}_t(\theta)$. An approximation for $\sigma_t^2 = \mathbb{E}[h_t^2(\theta)|\mathcal{F}_{t-1}]$ is recursively computed from $\{y_1, \ldots, y_T\}$. Blasques et al. (2023), note that the choice of $\hat{\sigma}_t^2$ is proportional to either $\hat{\vartheta}_t$ or $\hat{\vartheta}_t^2$. In this paper, $\hat{\sigma}_t^2(\theta)$ is set to $\hat{\sigma}_t^2(\theta) = \hat{\vartheta}_t(\theta)$.

The derivative of $\hat{\vartheta}_t(\theta)$ w.r.t θ is given by

$$\frac{\partial \hat{\vartheta}_{t+1}(\theta)}{\partial \theta} = \frac{\partial \omega}{\partial \theta} + \psi_t \frac{\partial \alpha}{\partial \theta} + \alpha \frac{\partial \psi_t}{\partial \theta} + \hat{\vartheta}_t(\theta) \frac{\partial \beta}{\partial \theta} + \left(\alpha \frac{\partial \psi_t}{\partial \vartheta} + \beta \right) \frac{\partial \hat{\vartheta}_t(\theta)}{\partial \theta} \tag{8}$$

Here $\psi_t = \psi_t(y_t, \hat{\vartheta}_t(\theta), \theta)$.

Under the assumptions of Propositions 2–5 Theorem 1 in Blasques et al. (2023) holds, which states that any such $\hat{\theta}$ is consistent and asymptotically normal distributed.

2.4.2 Maximum likelihood estimation (MLE)

When the conditional density is completely specified, one can use MLE to estimate the model parameters. In this case, MLE is often more efficient than the QLE estimator, but is likely inconsistent when the assumed distribution is misspecified (Blasques et al. 2023).

Suppose it is assumed that y_t is drawn from a distribution $\mathcal{D}$ with density $p(\cdot|\vartheta_t, \theta)$. The likelihood is given by $L(\theta) = \prod_{t=1}^{T} p(y_t|\hat{\vartheta}_t, \theta)$ and the log-likelihood is specified as

$$l(\theta) = \log(L(\theta)) = \sum_{t=1}^{T} \log \left(p(y_t|\hat{\vartheta}_t, \theta) \right)$$

The parameters θ are then given by the solution to

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ l(\theta)$$

Blasques et al. (2023) show in Theorems 2 and 3 that the MLE estimator is consistent and asymptotically normally distributed when the conditional distribution is correctly specified. Under misspecification, the estimator converges to the parameter that minimises the Kullback-Leibler divergence White (1996), this however need not be the unique maximiser of the log-likelihood, see Freedman and Diaconis (1982) for an example on a location model.

2.5 Monte Carlo Study

2.5.1 Monte Carlo study without outliers

A Monte Carlo experiment is done, similar to Blasques et al. (2023) and Blasques et al. (2021), to asses the finite-sample performance of the proposed Barron loss function with an estimated γ in a volatility setting. The data-generating process is specified as follows

$$y_t = \sqrt{\vartheta_t}\,\varepsilon_t, \quad \varepsilon_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1),$$

$$\vartheta_{t+1} = \omega \;+\; \alpha\,\psi\big(y_t, \vartheta_t; \theta\big) \;+\; \beta\,\vartheta_t,$$

where

$$\psi(y_t,\vartheta_t,\theta) = -\frac{\partial\rho(y_t^2-\vartheta_t,\gamma,\xi)}{\partial\vartheta_t} = \begin{cases} \dfrac{y_t^2-\vartheta_t}{\xi^2}, & \gamma = 2, \\ \dfrac{2(y_t^2-\vartheta_t)}{(y_t^2-\vartheta_t)^2+2c^2}, & \gamma = 0, \\ \dfrac{y_t^2-\vartheta_t}{\xi^2}\exp\Big(-\frac{1}{2}\left(\frac{y_t^2-\vartheta_t}{\xi}\right)^2\Big), & \gamma = -\infty, \\ \dfrac{y_t^2-\vartheta_t}{\xi^2}\left(\frac{(y_t^2-\vartheta_t)^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-1}, & \text{otherwise.} \end{cases}$$

In Appendix F, a section with all the derivatives needed for the QLE estimation process is described in Quasi-likelihood estimation (QLE).

This paper considers five special cases and two benchmarks. The five special cases are: $\gamma = 2$ (fixed L_2 loss), $\gamma = 1, \gamma = 0$ (Cauchy loss), $\gamma = -\infty$ (Welsh Loss) and estimated γ. One of the benchmarks is a GAS model (see Equation (1.1)) where ε_t is assumed to follow a standard normal distribution and a scoring function (S_t) that depends on the variance of the score. That is $S_t = \mathbb{E}[\nabla_t\nabla_t']^{-1}$. In this formulation, the GAS model is equivalent to the GARCH model proposed by Bollerslev (1986). The GARCH model is estimated using MLE. The second benchmark model is the β_t GARCH$(1,1)$, proposed by Harvey and Chakravarty (2008), which has been a popular score-driven model for volatility. This model assumes that ε_t follows a standardised Student's t distribution with $\nu > 2$ degrees of freedom, allowing for a heavy-tailed error specification. The updating equation is given by

$$\vartheta_{t+1} = \omega + \alpha \frac{\nu + 1}{\nu - 2 + \varepsilon_t^2} \varepsilon_t^2 \vartheta_t + \beta \vartheta_t \tag{9}$$

A simulation on 3000 and 4000 observations is done for each of the 1000 repetitions of the Monte Carlo simulation. $\theta = (\omega, \beta, \alpha, \gamma, \xi)$ is estimated by minimising the quasi-likelihood criterion (7) for the QSD filter models. The models are evaluated by the finite-sample bias, root mean squared error (RMSE) and the mean absolute error (MAE). Furthermore, the pathwise average RMSE and MAE of the estimated volatility $\hat{\vartheta}_t$ relative to the true ϑ_t are computed. The true parameters that generate the data are set to: $\theta_{\text{true}} = (\omega, \beta, \alpha, \gamma, \xi) = (0.07, 0.8, 0.11, 1.0, 1.2)$. This combination of parameters satisfies the conditions set in 2.3.

The pathwise RMSE and MAE are computed using

$$\widehat{\text{RMSE}} = \sqrt{\frac{1}{T} \sum_{t=1}^{T} \left(\vartheta_t - \hat{\vartheta}_t(\hat{\theta}) \right)^2}, \qquad \widehat{\text{MAE}} = \frac{1}{T} \sum_{t=1}^{T} \left| \vartheta_t - \hat{\vartheta}_t(\hat{\theta}) \right|$$

where ϑ_t denotes the true volatility parameter from the DGP and $\hat{\vartheta}_t(\hat{\theta})$ denotes the filtered volatility evaluated at the estimated parameter value $\hat{\theta}$.

2.5.2 Monte Carlo study including outliers

To evaluate the robustness of each of the models described above, a portion of the observations y_t is deliberately contaminated with extreme values. Specifically, after simulating the time series according to the previously described process, 20 observations have been replaced with extreme values (see section 3.2.1). These outliers are drawn uniformly from a range between 6 and 10 standard deviations above the volatility level of the original time series at the corresponding time step. The choice of the amount and size of the outliers is discussed in Section 3.2.

Model estimation is performed on this contaminated time series. The models are evaluated to asses their robustness to outliers. In this Monte Carlo simulation, the average pathwise RMSE and MAE between ϑ_t and $\hat{\vartheta}_t$ is of extra importance, as it gives a measure of how much the performance of the model has been affected by the added outliers. Additionally, the bias, RMSE and MAE of the parameters are given to show the effect of the outliers on the parameter estimates.

3 Results

This section presents the results of the Monte Carlo study described in Section 2.5. First, parameter estimation under clean data settings is examined, then the robustness performance of the model when the data contains outliers is assessed, and finally visual evidence of how the choice of γ affects the model's sensitivity to extreme observations and volatility clusters is shown.

3.1 Results Monte Carlo study without outliers

Table 1, 2 and 5 present the results of the Monte Carlo simulation study assessing the finite-sample performance of the QSD (Quasi Score Driven) filter using the Barron loss. The results are shown for different values of the shape parameter γ and the benchmark GARCH model and β_t GARCH$(1,1)$ model.

The results of the estimated parameters are based on 1000 simulation runs and are shown in Table 1 and 5, which describe the estimated value, bias, RMSE and MAE of each parameter averaged over all simulation runs. The latter three statistics are omitted for the GARCH and β_t GARCH(1,1) models, as their estimated parameters are not directly comparable to the true parameter values used in the QSD models.

The RMSE and MAE between the estimated volatility ($\hat{\vartheta}_t$) and the true volatility (ϑ_t) are given in Table 2. Similarly to Blasques et al. (2023), the analysis has been done for both 3000 and 4000 time steps, to assess the impact of longer time series on estimation accuracy and convergence. The table with the estimated parameters of the Monte Carlo study with 3000 time steps is given in the Appendix H.

3.1.1 Parameter estimation performance

Across both sample sizes, the bias for the parameters ω, α, and β is generally small and decreases with larger sample sizes, indicating consistency and convergence of the QSD filter. For instance, when $\gamma = 2$, the bias of ω and α are negligible (0.004 and 0.007 for $T = 3000$ and almost 0 for $T = 4000$, respectively). The estimate for β shows slightly larger bias, especially for $\gamma = 2$, but this bias tends to decrease as T grows larger.

The estimation of ξ seems more challenging, showing similar bias but larger RMSE and

Table 1: Monte Carlo study excluding outliers

		$T = 4000$					
		ω	α	β	ξ	γ	ν
GARCH	Est.	0.084	0.059	0.691	–	–	–
β_t	Est.	0.086	0.060	0.684	–	–	91.094
$\gamma = -\infty$	Est.	0.071	0.169	0.770	1.148	–	–
	Bias	**0.001**	0.059	−0.031	−0.052	–	–
	RMSE	0.033	0.108	0.122	0.289	–	–
	MAE	0.021	0.064	0.071	0.201	–	–
$\gamma = 0$	Est.	0.067	0.159	0.770	1.164	–	–
	Bias	−0.003	0.049	−0.030	**-0.036**	–	–
	RMSE	0.025	0.095	0.107	0.243	–	–
	MAE	0.017	0.052	0.065	0.169	–	–
$\gamma = 1$	Est.	0.066	0.144	0.767	1.161	–	–
	Bias	−0.004	0.034	−0.033	−0.039	–	–
	RMSE	**0.021**	0.069	0.099	0.200	–	–
	MAE	0.016	0.038	0.063	0.135	–	–
$\gamma = 2$	Est.	0.071	0.115	0.740	1.256	–	–
	Bias	**0.001**	**0.005**	−0.060	0.058	–	–
	RMSE	**0.021**	**0.047**	0.112	0.219	–	–
	MAE	0.014	**0.022**	0.075	0.157	–	–
Est. γ	Est.	0.066	0.135	0.772	1.159	0.967	–
	Bias	−0.004	0.025	**-0.028**	−0.041	−0.033	–
	RMSE	**0.021**	0.052	**0.088**	**0.178**	0.430	–
	MAE	**0.013**	0.029	**0.052**	**0.104**	0.185	–

NOTE: Monte Carlo study of 1000 repetitions on the finite sample performance of the QSD filter. The models were estimated on 4000 observations. The true parameters are $\theta_{\text{true}} = (\omega, \alpha, \beta, \gamma, \xi) = (0.07, 0.11, 0.8, 1.0, 1.2)$ and the data has not been contaminated with outliers. The QSD models were estimated using quasi likelihood estimation and the GARCH and β_t GARCH$(1,1)$ (abbreviated as β_t) were estimated using maximum likelihood estimation. Bold numbers denote the most accurate (closest to zero) value for each parameter within a given error metric.

MAE compared to other parameters. The RMSE being larger than the MAE, indicates the presence of estimation errors with heavy tails. While the overall central tendency of the parameter is near the true value of $\xi = 1.2$, occasional larger errors occur. This increase in RMSE and MAE is mostly notable when γ was fixed. This suggests that the joint estimation of ξ and γ may lead to improved fit.

Table 2: RMSE and MAE of $\hat{\vartheta}_t$ for two sample sizes

	$T = 3000$		$T = 4000$	
	RMSE	MAE	RMSE	MAE
GARCH	**0.034**	**0.024**	**0.033**	**0.024**
β_t	**0.034**	0.025	**0.033**	**0.024**
$\gamma = -\infty$	0.049	0.050	0.046	0.046
$\gamma = 0$	0.052	0.049	0.048	0.045
$\gamma = 1$	0.052	0.049	0.048	0.045
$\gamma = 2$	0.054	0.050	0.049	0.046
Est. γ	**0.046**	**0.044**	**0.045**	**0.043**

NOTE: This table reports the RMSE and MAE between the estimated volatility path $\hat{\vartheta}_t$ and the true volatility in Monte Carlo experiments with $T = 3000$ and $T = 4000$ time steps, excluding outliers. Parameter estimation results for the $T = 3000$ case are provided in Appendix H. Bold values in the top two rows indicate the lowest error per metric within each column. The bold values in the final row emphasize that the QSD model with estimated γ outperforms all fixed-γ variants.

The robustness parameter γ demonstrates clear sensitivity to sample size. With 3000 observations, the estimated γ exhibits a noticeably larger RMSE and MAE compared to the other parameters. However, this RMSE substantially reduces (from 0.602 to 0.430) when the time series increases to 4000 observations, while the MAE also reduces (from 0.238 to 0.185). This gives an indication that the estimation of γ needs a sufficient amount of data to distinguish between different levels of robustness.

3.1.2 Benchmark models performance

The GARCH and β_t GARCH$(1,1)$ model, which use different filtering methods, yield the lowest RMSE and MAE of the estimated volatility $\hat{\vartheta}_t$ across both sample sizes (0.034 for $T = 3000$ and 0.033 for $T = 4000$). The parameter estimates differ from the estimated parameters of the QSD model. This was however to be expected, as there is a difference in the dynamics of the GARCH and β_t GARCH(1,1) model and the QSD filter. The QSD filter allows the updating function ψ to be negative, enabling faster downward adjustment of the time-varying parameter. This is not possible in the standard GARCH and β_t GARCH(1,1) model, where the updating term is proportional to y_t^2 and therefore always non-negative. This ability of the QSD filter allows for more aggressive downward updating, which ensures new information is processed quickly. Since the data was generated using the QSD model, which incorporates this ability to react more quickly, the estimated parameters of the benchmark models can be expected to obtain different values to compensate for this difference in model dynamics.

The β_t GARCH$(1,1)$ model's estimation of the degrees of freedom parameter ν provides additional insight into model behaviour under clean data conditions. It was estimated at approximately 89-90, by which the student's t distribution is approaching normality, suggesting the model does not need heavy tails. This is a verification of the data being generated by a Gaussian error distribution, which does not need extra robustness.

3.1.3 Tests

Welch's t-tests were conducted to evaluate whether the RMSE of estimated volatility $\hat{\vartheta}_t$ differs significantly between fixed and estimated values of γ. For example, in case of a sample size of 3000, the resulting p-value the test between $\gamma = 0$ and the estimated γ model is $2.016 \cdot 10^{-7}$. this value falls well below the conventional 5% significance level, indicating statistically significant differences in RMSE.

Similarly, for the sample size of 4000, the corresponding p-values are 0.001488 for $\gamma = 0$ and 0.009145 for $\gamma = 2$ against the estimated γ model. Again, both p-values are below the 5% threshold, reinforcing the conclusion that estimating γ leads to significantly different, and in this case, lower, RMSE values. These results suggest that allowing γ to be estimated rather than fixed improves the accuracy of volatility estimation.

3.2 Results Monte Carlo with outliers

To evaluate the robustness of the proposed QSD filter, a Monte Carlo simulation is conducted with data containing outlier contamination as described in Section 2.5. The results are presented in Table 3 and 4. This analysis examines how the introduction of extreme observations affects parameter estimation accuracy across different specifications.

The presence of outliers significantly inflates both bias, RMSE and MAE for most estimated parameters compared to the clean data Monte Carlo study. The robustness parameters γ and ξ are most affected by outlier contamination, which is expected given their role in controlling the QSD filter's resistance to extreme observations.

Table 3: Monte Carlo study including outliers

		ω	α	β	ξ	γ	ν
$T = 4000$							
GARCH	Est.	0.176	0.036	0.570	–	–	–
β_t	Est.	0.103	0.071	0.680	–	–	5.631
$\gamma = -\infty$	Est.	0.114	0.128	0.759	1.825	–	–
	Bias	0.044	0.018	**-0.041**	0.627	–	–
	RMSE	0.096	0.138	0.204	2.378	–	–
	MAE	0.054	0.071	0.128	1.051	–	–
$\gamma = 0$	Est.	0.114	0.141	0.748	1.692	–	–
	Bias	0.044	0.031	−0.052	0.492	–	–
	RMSE	0.096	0.158	0.210	2.163	–	–
	MAE	0.055	0.082	0.130	0.936	–	–
$\gamma = 1$	Est.	0.109	0.180	0.712	1.608	–	–
	Bias	0.039	0.070	−0.088	0.408	–	–
	RMSE	0.084	0.180	0.218	1.968	–	–
	MAE	0.050	0.109	0.143	0.834	–	–
$\gamma = 2$	Est.	0.118	0.034	0.699	2.087	–	–
	Bias	0.048	−0.076	−0.101	0.887	–	–
	RMSE	0.101	0.102	0.244	2.045	–	–
	MAE	0.058	0.090	0.148	1.022	–	–
Est. γ	Est.	0.098	0.127	0.759	1.424	0.812	–
	Bias	**0.028**	**0.017**	**-0.041**	**0.224**	−0.188	–
	RMSE	**0.067**	**0.110**	**0.164**	**1.516**	1.291	–
	MAE	**0.036**	**0.065**	**0.100**	**0.613**	0.502	–

NOTE: Monte Carlo study of 1000 repetitions on the finite sample performance of the QSD filter and the benchmark models (GARCH model and β_t GARCH(1,1) (abbreviated as β_t)). The models were estimated on 4000 observations which included outliers. The true parameters are $\theta_{\text{true}} = (\omega, \alpha, \beta, \gamma, \xi) = (0.07, 0.11, 0.8, 1.0, 1.2)$. In addition to the generated data, outliers 20 have been added, which contaminated the data. The QSD models were estimated using quasi likelihood estimation and the two benchmark models were estimated using maximum likelihood estimation. Bold numbers denote the most accurate (closest to zero) value for each parameter within a given error metric.

The estimated value of γ decreases substantially under outlier contamination, averaging 0.812 compared to the true value of 1.0. This reduction indicates that the estimation procedure correctly identifies the need for increased robustness in the presence of outliers. When ξ is estimated jointly with other parameters, the model achieves robustness primarily

Table 4: RMSE and MAE of $\hat{\vartheta}_t$ under contaminated setting ($T = 4000$)

	RMSE	MAE
GARCH	0.155	0.112
β_t	0.086	0.077
$\gamma = -\infty$	0.090	0.100
$\gamma = 0$	0.085	0.078
$\gamma = 1$	0.079	0.078
$\gamma = 2$	0.096	0.076
Est. γ	**0.074**	**0.065**

NOTE: This table reports the RMSE and MAE between the estimated volatility path $\hat{\vartheta}_t$ and the true volatility for a Monte Carlo study with $T = 4000$ time steps under a contaminated setting (i.e., with outliers). Bold numbers indicate the lowest error within each column.

through this downward adjustment of γ.

The parameter ξ, which controls the curvature of the quadratic bowl around zero in the loss function, shows consistent overestimation across all QSD model variants. This overestimation represents an attempt to achieve robustness by flattening the loss function around zero. The effect is most pronounced when γ is fixed at 2, where the high γ value makes the function highly reactive to outliers, necessitating larger ξ values for compensation. Notably, when γ and ξ are estimated jointly, the overestimation of ξ is minimized, as robustness is achieved more efficiently through the reduction in γ.

The QSD filter variants demonstrate superior robustness compared to benchmark models. Among all models, the QSD model with estimated γ achieves the lowest pathwise RMSE and MAE (Table 4), confirming the benefit of allowing the robustness parameter to adapt to data conditions.

The GARCH model exhibits poor performance under outlier contamination, with the largest RMSE of $\hat{\vartheta}_t$ of all the models tested. A Welch's t-test against the worst performing QSD model ($\gamma = 2$) gives a p-value of $2.2 \cdot 10^{-16}$. This gives great evidence that the QSD models are outperforming the GARCH model.

The β_t GARCH$(1,1)$ model demonstrates better robustness against outliers, with an average pathwise RSME of 0.086 and MAE of 0.077. While this an improvement over the GARCH model, it is still inferior to the estimated γ model. A Welch's t-test between the two RMSE gives a p-value of $2.2 \cdot 10^{-16}$, giving sufficient statistical evidence for the improvement of the estimated γ QSD filter. Interestingly, the results show that the estimated value of ν decreased substantially from around 90 in the clean setting, to 5.631 in the contaminated setting. This demonstrated the model's need to accommodate for the heavy-tailed nature of the data with outliers.

3.2.1 Monte Carlo study design considerations

The experimental design requires careful calibration of outlier characteristics to ensure meaningful robustness evaluation. Both the number of outliers and the size of the outliers must be balanced. Excessive contamination transforms outliers from anomalies into systematic features of the data-generating process, while insufficiently extreme values fail to challenge model robustness. Similarly, if the outliers are not extreme enough, they may fail to meaningfully distort the estimates of the parameters. In this setup, the size and number of outliers have been selected such that the difference between robust and non-robust models becomes evident.

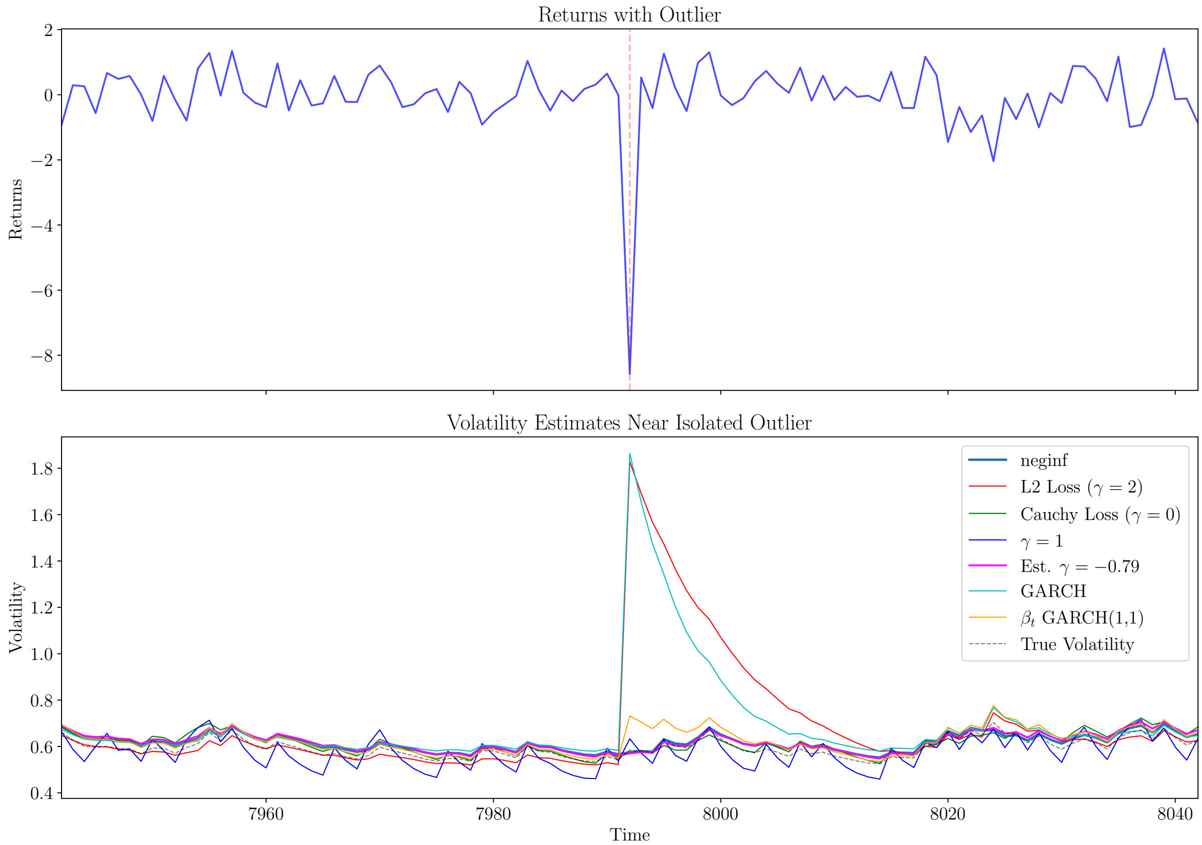
The results show that the use of the Barron loss function in a QSD setting allows the QSD filter to down weigh the influence of outliers. This robustness mechanism translates into substantial performance gains, with the QSD filter achieving lower average pathwise RMSE of $\hat{\vartheta}_t$ relative to both benchmark models.

3.3 The effect of outliers and clustered volatility

To visualise how the choice of γ influences outlier sensitivity, plots have been created that show the estimated volatility under the different models. For this purpose, data has been generated for 15000 time steps according to the volatility specification described in the Monte Carlo study (Section 2.5, where $\gamma_{\text{true}} = 1$), and $\varepsilon_t \sim \mathcal{N}(0, 1)$. Similar to the Monte Carlo study, 10 outliers have been randomly added to the generated data y_t (not the volatility process), which contaminates the data. These outliers are selected such that they lie 10 to 15 standard deviations from the true volatility at that time. Also included is a cluster of 30 subsequent extreme observations, which were 5 to 8 standard deviations from the volatility generated under $\gamma = 1$.

3.3.1 Effect of outliers

Figure 3: Returns with single outlier



NOTE: The effect of outliers on the estimated volatility. Above, the return includes an outlier. Below, the corresponding estimated volatilities and true volatility.

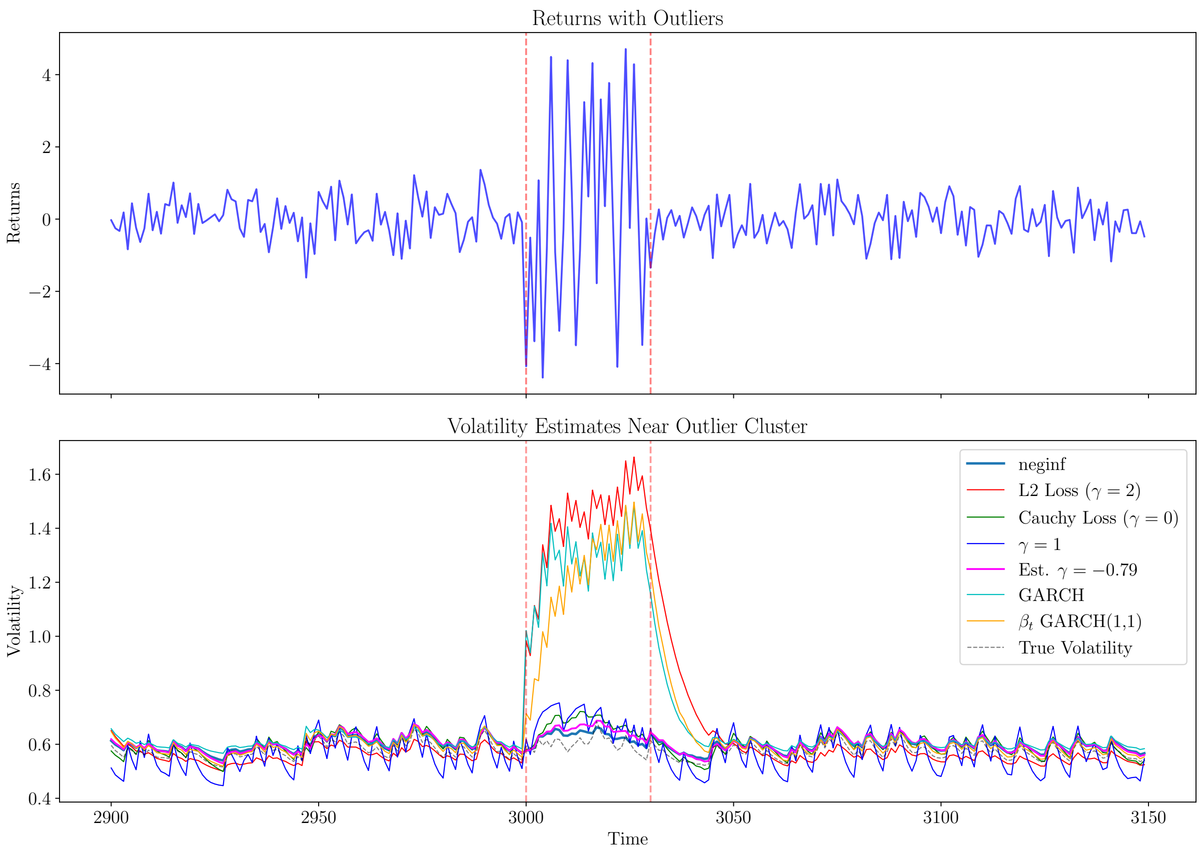
Figure 3 demonstrates the different responses of various model specifications to a single extreme observation. The results reveal a clear relationship between the γ parameter and outlier sensitivity.

When $\gamma = 2$ (equivalent to L2 loss), the estimated volatility is substantially elevated for approximately 25 time periods after the outlier has occurred. The GARCH model displays similar behaviour, maintaining overestimated volatility for a comparable duration. This persistent elevation represents a significant distortion that could substantially impact volatility forecasting.

The β_t GARCH$(1, 1)$ model shows intermediate sensitivity, overestimating volatility for approximately 15 periods. While less severe than the $\gamma = 2$ or the GARCH model, this is still a meaningful distortion of the volatility forecast.

In contrast, lower values of γ show resilience against the effect of outliers on the estimated volatility. When γ was estimated from the contaminated data, it was estimated to be -0.79, which is lower than the true value of $\gamma = 1$. But, due to the effect of the outliers, the true volatility is better estimated by $\gamma = -0.79$. That is, the model without the outliers was correctly specified by $\gamma = 1$, but due to the added outliers the model with $\gamma = 1$ is now misspecified relative to the contaminated data.

Figure 4: Returns with multiple subsequent outliers



NOTE: The effect of multiple outliers on the estimated volatility. Above, the return including multiple outliers. Below, the corresponding estimated volatilities.

Figure 4 shows the cluster of outliers and the estimated volatility under the different models. This situation tests model performance during extended periods of anomalous behaviour.

The $\gamma = 2$, GARCH and β_t GARCH$(1,1)$ model reacted most dramatically to the contaminated period, appropriately capturing the elevated volatility during the outlier cluster but subsequently overestimating volatility for approximately 25 periods after the contamination ends. This is beneficial in this short period of high volatility, but when the high volatility has passed, the model overestimated the volatility for approximately 25 time steps.

Lower-γ specifications respond more conservatively to the outlier cluster. This is not beneficial during the period of high volatility, but it does ensure that the regular trajectory is followed steadily.

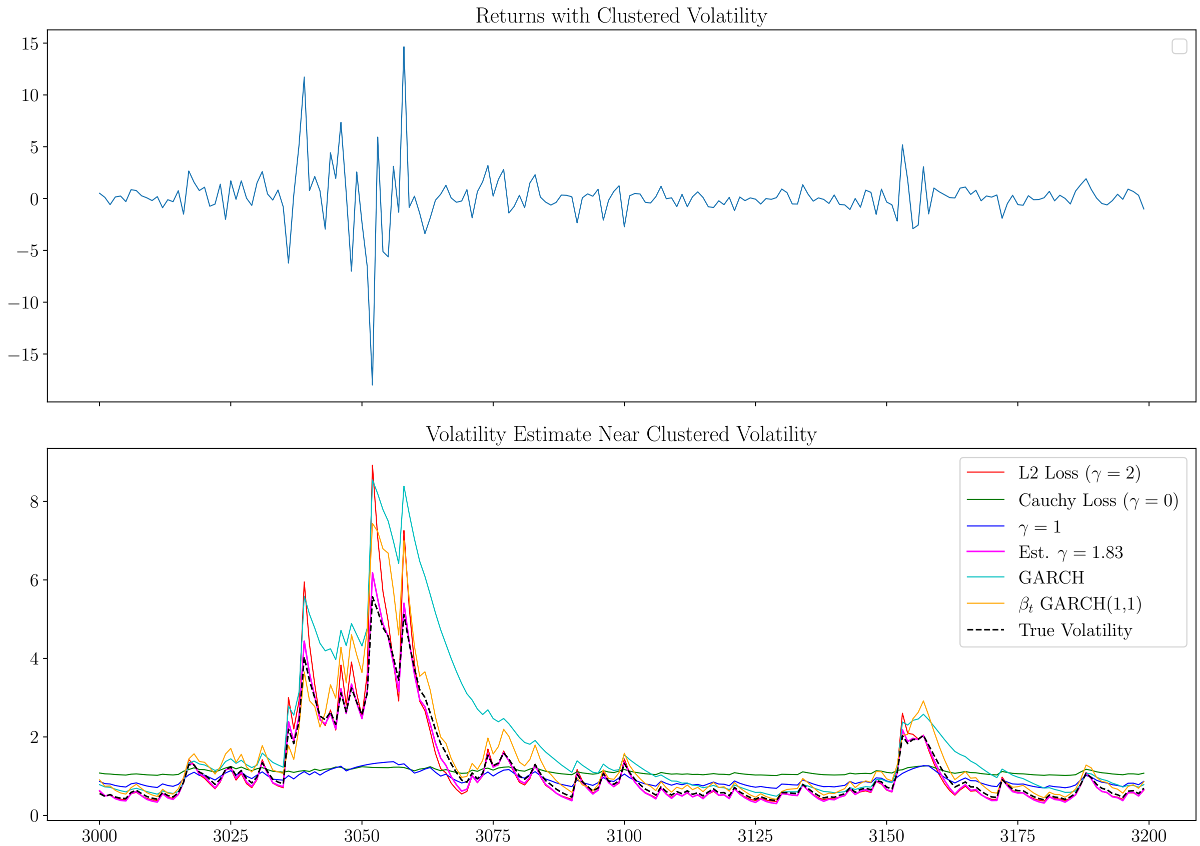
The good performance of the QSD model with the estimated γ parameter is evident in the RMSE of the pathwise estimated and true volatility. The model achieves the lowest RMSE of 0.031, substantially outperforming fixed specifications: $\gamma = 2$ (0.130), $\gamma = 1$ (0.052), $\gamma = 0$ (0.035), as well as the benchmark GARCH model (0.115) and β_t GARCH(1, 1) model (0.070).

3.3.2 Clustered volatility

While the ability of the proposed QSD filter to be robust against outliers is a key feature of the model, it is equally important that the model remains responsive to periods of clustered volatility. This clustering, often observed in economic or financial data (Mandelbrot 1963), is evaluated though the use of a synthetic dataset of 4000 time steps. The sample was divided into four regimes, each with distinct distributional characteristics to simulate clustered volatility. The first 1000 time steps and between time steps 2000 and 3000, the error term was standard normally distributed. In the period 1000 to 2000 the error term was standard-t distributed with 5 degrees of freedom. In the last 1000 time steps the data was standard-t distributed with 3 degrees of freedom. The true parameters of the data-generating process are $\theta = (0.05, 0.18, 0.85, 1.8, 0.8)$. The resulting time series contains clustering and Figure 5 shows a snapshot of a period with higher and clustering volatility. A figure that does not show the fixed γ models and a figure that shows multiple subsequent clusters can be found in Appendix I.

The figure shows the QSD filter's capability to react to the increased and clustering volatility. The estimated γ model manages to find the closest fit to the true volatility, which becomes evident from the lowest RMSE of the fitted volatility to the true volatility. The estimated γ model has an RMSE of 0.235 compared to $\gamma = 1$ (1.558) , $\gamma = 0$ (1.700), β_t GARCH(1, 1) (1.790), $\gamma = 2$ (1.293), GARCH (3.129).

Figure 5: Clustering volatilities



NOTE: Estimated volatilities under data with clustered regimes of increasing tail heaviness of some fixed γ models, an estimated γ model and the GARCH and β_t GARCH$(1,1)$ model.

The GARCH and β_t GARCH$(1,1)$ show some over-fitting of the true volatility. Especially, the GARCH model overestimates the volatility when a volatility cluster has passed. The fixed γ models also struggled, especially the models which had γ fixed to low values ($\gamma = \{0,1\}$), were not able to react to the increased volatility. The L2 Loss model ($\gamma = 2$) reacted to the increased volatility much better, but overestimated it near the peaks. This further confirms that the estimated value of $\gamma = 1.83$ was better, since it was slightly less reactive to the larger observations.

This analysis uncovers important aspects of the choice of the robustness parameter γ, on the estimation of the time-varying parameter. The ability to estimate γ allows the model to learn the optimal robustness to outliers. Fixed parameter choices, while potentially optimal under specific conditions, lack this adaptive capability and may result in either excessive sensitivity to outliers or insufficient responsiveness to genuine volatility changes.

3.4 Considerations and further research

This paper's main contribution is the use of an adaptive loss function in an application of the Quasi score-driven framework. The use of this loss function shows great flexibility, covering well-known loss functions such as the L2 loss ($\gamma = 2$), Charbonnier loss ($\gamma = 1$), Chauchy loss ($\gamma = 0$), Geman-McClure loss ($\gamma = -2$) and Welsch loss ($\gamma = -\infty$). Allowing it to adjust to the level of robustness characterised by the data. However, this flexibility comes at the cost of increased computational complexity, which limits the scale of Monte Carlo simulations and the need for a sufficient number of data points to differentiate between different levels of robustness.

The proposed QSD filter has not yet been applied to empirical data. Evaluating its performance in real-world scenarios with occasional anomalous observations presents a promising topic for future research. Another potential extension is to allow the robustness parameter γ to vary over time, enabling the model to dynamically adjust its sensitivity to outliers rather than assuming a fixed level of robustness throughout. This is similar to an approach in a paper by D'Innocenzo et al. (2023). This could be beneficial when the data exhibits periods where the model should be highly reactive to price changes and other periods where a more robust measure is needed. Furthermore, the proposed QSD framework should be extended to handle multivariate models.

4 Conclusion

The score-driven models (Creal et al. 2013; Harvey 2013) have gained significant popularity over the last decade, with almost 400 published papers using the framework. Recently, Blasques et al. (2023) introduced the quasi score-driven (QSD) framework through a generalisation of the score-driven framework which decouples the updating equation of the time-varying parameter from the assumed distribution the innovation terms are following. This generalisation allows for the use of a broad class of loss functions in the updating equation.

This paper introduces an implementation of the quasi score-driven framework by incorporating the general adaptive loss function proposed by Barron (2019) into the parameter updating mechanism. This resulting QSD filter allows the robustness of the updating step to be learned from the data, rather than a fixed, pre-specified loss function.

Under mild regularity conditions, it is shown that the proposed filter generates a strictly stationary, ergodic, and invertible sequence of time-varying parameters and a valid data-generating process. Furthermore, the existence of bounded unconditional moments and differentiability of the filter recursion is established, ensuring that the resulting quasi-likelihood estimator is both consistent and asymptotically normal. In addition, it is shown that for different values of γ, the loss function is a consistent estimator, in the sense of Gneiting (2011), for different functionals. For example, for $\gamma = 2$ the loss function is consistent for the mean and for $\gamma = 1$, the loss function is consistent for the median under the condition that the error is much larger than ξ. This flexibility ensures the robustness of the QSD filter by allowing smooth interpolation between different functionals, trading off bias and variance depending on the data.

Monte Carlo simulations on clean data without outliers demonstrated convergence of the parameter estimates to their true values as the sample size increased, validating the derived asymptotic theory. Additionally, a Monte Carlo study on data to which outliers were added showed that joint estimation of γ alongside the other model parameters leads to lower RMSE and MAE of the estimated time-varying parameter ϑ_t compared, to both fixed-γ QSD filters and benchmark GARCH and β_t GARCH$(1, 1)$ models. Robustness is achieved through data-driven down weighting of large residuals, without requiring strong distributional assumptions on the innovations or manual tuning of loss functions.

Taken together, the proposed QSD filter offers a robust and adaptable alternative to traditional volatility models. By learning the robustness parameter γ directly from the data, the QSD filter handles outliers without requiring strong distributional assumptions or manual tuning. This flexibility makes it a valuable tool for applications in time series analysis.

References

Andersen, T. G. and T. Bollerslev (1998), "Answering the Skeptics: Yes, Standard Volatility Models do Provide Accurate Forecasts", *International Economic Review*, *39*(4), 885.

Artemova, M., F. Blasques, J. Van Brummelen, and S. J. Koopman (2022). "Score-Driven Models: Methodology and Theory".

Barron, J. T. (2019). "A General and Adaptive Robust Loss Function". In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4326–4334. IEEE.

Blasques, F., L. André, and A. C. van Vlodrop (2021), "Finite Sample Optimality of Score-Driven Volatility Models: Some Monte Carlo Evidence", *Econometrics and Statistics*, *19*, 47–57.

Blasques, F., F. Christian, and L. Sébastien (2023), "Quasi score-driven models", *Journal of Econometrics*, *234*(1), 251–275.

Blasques, F., J. van Brummelen, P. Gorgi, and S. J. Koopman (2024), "Maximum Likelihood Estimation for Non-Stationary Location Models with Mixture of Normal Distributions", *Journal of Econometrics*, *238*(1), 105575.

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.

Charbonnier, P., L. Blanc-Feraud, G. Aubert, and M. Barlaud (1997), "Deterministic edge-preserving regularization in computed imaging", *IEEE Transactions on Image Processing*, *6*(2), 298–311.

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

De Punder, R., T. Dimitriadis, and R.-J. Lange (2024). "Kullback-Leibler-based characterizations of score-driven updates". Version Number: 2.

D'Innocenzo, E., B. Schwaab, A. Lucas, and X. Zhang (2023), "Modeling Extreme Events: Time-Varying Extreme Tail Shape", *Journal of Business and Economic Statistics*, *42*, 903–917.

Durbin, J. (1960), "Estimation of Parameters in Time-Series Regression Models", *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *22*(1), 139–153.

Engle, R. F. (1982), "Autoregressive Conditional Heteroscedasticity with Estimates of the Variance of United Kingdom Inflation", *Econometrica*, *50*(4), 987.

Engle, R. F., D. M. Lilien, and R. P. Robins (1987), "Estimating Time Varying Risk Premia in the Term Structure: The Arch-M Model", *Econometrica*, *55*(2), 391.

Freedman, D. A. and P. Diaconis (1982), "On Inconsistent M-Estimators", *The Annals of Statistics*, *10*(2), 454 – 461.

Glosten, L. R., R. Jagannathan, and D. E. Runkle (1993), "On the Relation between the Expected Value and the Volatility of the Nominal Excess Return on Stocks", *The Journal of Finance*, *48*(5), 1779–1801.

Gneiting, T. (2011), "Making and Evaluating Point Forecasts", *Journal of the American Statistical Association*, *106*(494), 746–762.

Godambe, V. P. (1960), "An Optimum Property of Regular Maximum Likelihood Estimation", *The Annals of Mathematical Statistics*, *31*(4), 1208–1211.

Harvey, A. C. (2013). *Dynamic Models for Volatility and Heavy Tails*. Cambridge University Press.

Harvey, A. C. (2022), "Score-Driven Time Series Models", *Annual Review of Statistics and Its Application*, *9*(1), 321–342.

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

Hill, J. B. (2015), "Robust estimation and inference for heavy tailed GARCH", *Bernoulli*, *21*(3), 1629 – 1669.

Huber, P. J. (1964), "Robust Estimation of a Location Parameter", *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Huber, P. J. (1981). *Robust statistics.* Wiley series in probability and mathematical statistics. J. Wiley & sons.

Mandelbrot, B. (1963), "The Variation of Certain Speculative Prices", *The Journal of Business*, *36*(4), 394.

White, H. (1996). *Estimation, Inference and Specification Analysis.* Cambridge University Press: Cambridge University Press.

5 Appendix

A Verification of proposition 1, consistency of the loss function

Proposition: Let I be the interval representing the potential range of outcomes of a random variable Y, and let $\mathcal{P}$ be a class of probability distributions in the sample space of I. A scoring function $S : I \times I \to [0, \infty)$ is consistent, for a functional $T : \mathcal{P} \to \mathcal{P}_r(I)$ and $F \to T(F) \subseteq I$ if

$$\mathbb{E}_F[S(t, Y)] \leq \mathbb{E}_F[S(x, Y)] \tag{10}$$

for all probability distributions $F \in \mathcal{P}$, all $t \in T(F)$, and all $x \in I$.

The scoring function S is strictly consistent for the functional T if it is consistent, and equality in the above expression implies that $x \in T(F)$.

Verification:

A.1 The mean

It is shown that if and only if $\gamma = 2$, the Barron loss function is a consistent scoring function for the mean if $e(y_t, \vartheta_t) = y_t - \vartheta_t$ and $\vartheta_t = \mathbb{E}[Y_t]$.

Consider the case for $\gamma = 2$. In this case, the Barron loss function is defined as $\rho(e(y_t, \vartheta_t), \gamma, \xi) = \frac{1}{2}\left(\frac{y_t - \vartheta_t}{\xi}\right)^2$. By differentiating the expectation of ρ it is shown that for $\vartheta_t = \mathbb{E}[y_t]$ the function is uniquely minimised.

$$\frac{\partial}{\partial \vartheta_t}\mathbb{E}_F\left[\frac{1}{2}\Big(\frac{y_t - \vartheta_t}{\xi}\Big)^2\right] = \mathbb{E}_F\left[-\frac{(y_t - \vartheta_t)}{\xi^2}\right] = 0$$

This solves for $\vartheta_t = \mathbb{E}[y_t]$. Since the function is convex, the solution uniquely minimises the expected value and therefore $t = \mathbb{E}[y_t] \in T(F)$ has the lowest score in expected value. Therefore, for $\gamma = 2$, the score function is consistent.

Now consider in the most general case, in the case that $\gamma \neq \{2, 0, -\infty\}$, the loss function is given by

$$\rho(e(y_t, \vartheta_t), \gamma, \xi) = \frac{|\gamma - 2|}{\gamma}\left(\left(\tfrac{e(y_t,\vartheta_t)^2}{\xi^2|\gamma-2|} + 1\right)^{\gamma/2} - 1\right)$$

Differentiating this function in expectation with respect to ϑ_t results in

$$\frac{\partial}{\partial \vartheta_t}\mathbb{E}_F\left[\frac{|\gamma - 2|}{\gamma}\left(\left(\tfrac{e(y_t,\vartheta_t)^2}{\xi^2|\gamma-2|} + 1\right)^{\gamma/2} - 1\right)\right] = \mathbb{E}_F\left[-\frac{(y_t - \vartheta_t)}{\xi^2}\left(\tfrac{(y_t-\vartheta_t)^2}{\xi^2|\gamma-2|} + 1\right)^{\gamma/2-1}\right] = 0$$

This function, however, generally does not minimise at $\vartheta_t = E[y_t]$. That is, there are functionals $x \neq t \in I$ which receive a lower score in expectation. This violates proposition 1, and the scoring function is therefore said to be inconsistent for the functional $\mathbb{E}[y_t]$ for $\gamma \neq \{2, 0 - \infty\}$.

In the case that $\gamma = 0$, the loss function is given by,

$$\rho(e(y_t, \vartheta_t), \gamma, \xi) = \log\Big(\tfrac{1}{2}\big(\tfrac{e(y_t,\vartheta_t)}{\xi}\big)^2 + 1\Big)$$

Differentiating the function (in expectation) with respect to ϑ_t results in:

$$\frac{\partial}{\partial \vartheta_t}\mathbb{E}_F\left[\log\Big(\tfrac{1}{2}\big(\tfrac{e(y_t,\vartheta_t)}{\xi}\big)^2 + 1\Big)\right] = \mathbb{E}_F\left[-\frac{2(y_t - \vartheta_t)}{(y_t - \vartheta_t)^2 + 2\xi^2}\right] = 0$$

This function is also generally not minimised at $\vartheta_t = \mathbb{E}[y_t]$, therefore, the scoring function is inconsistent for that functional.

The same arguments hold for the case where $\gamma = -\infty$, in that case, the scoring function is also not minimised at $\vartheta_t = \mathbb{E}[y_t]$. Barron (2019) note that this function instead is minimised at the mode.

A.2 The median

Barron (2019) note that for $\gamma = 1$ the loss function is similar to the estimation of a median. It is not equivalent because the loss function does not reduce to L1 loss for $\gamma = 1$, but rather a smoothed version of L1 (Charbonnier loss). The loss function reduces to

$$\rho((y_t, \vartheta_t), 1, \xi) = \sqrt{\left(\frac{(y_t - \vartheta_t)}{\xi}\right)^2 + 1} - 1$$

Barron (2019) note that for values of ξ that are much smaller than $y_t - \vartheta_t$, the loss function is approximately a L1 loss function

$$\rho((y_t - \vartheta_t), 1, \xi) \approx \frac{|y_t - \vartheta_t|}{\xi} - 1$$

To see that this function in expectation minimises at $\vartheta_t = \mathbb{E}[y_t]$, first expectation is taken.

$$\mathbb{E}_F\left[\frac{|y_t - \vartheta_t|}{\xi} - 1\right] = \frac{1}{\xi}\mathbb{E}_F\left[|y_t - \vartheta_t|\right] - 1$$

Since the location of the minimum of this function does not depend on the -1 term, it can be omitted.

If $I = \mathbb{R}$ and y_t follows a distribution with pdf $p(\cdot)$ then the expectation can be written out as

$$\frac{1}{\xi}\mathbb{E}_F\left[|y_t - \vartheta_t|\right] = \frac{1}{\xi}\left(\int_{-\infty}^{\vartheta_t}(y_t - \vartheta_t)p(y_t)dy_t - \int_{\vartheta_t}^{\infty}(y_t - \vartheta_t)p(y_t)dy_t\right)$$

Now, taking the derivative results in

$$
\begin{aligned}
&\frac{\partial}{\partial \vartheta_t}\frac{1}{\xi}\left(\int_{-\infty}^{\vartheta_t}(y_t-\vartheta_t)p(y_t)dy_t-\int_{\vartheta_t}^{\infty}(y_t-\vartheta_t)p(y_t)dy_t\right)\\
&\frac{1}{\xi}\left(\frac{\partial}{\partial \vartheta_t}\int_{-\infty}^{\vartheta_t}(y_t-\vartheta_t)p(y_t)dy_t-\frac{\partial}{\partial \vartheta_t}\int_{\vartheta_t}^{\infty}(y_t-\vartheta_t)p(y_t)dy_t\right)\\
&\frac{1}{\xi}\left(\left[\int_{-\infty}^{\vartheta_t}\frac{\partial}{\partial \vartheta_t}(y_t-\vartheta_t)p(y_t)dy_t+(\vartheta_t-\vartheta_t)p(\vartheta_t)\right]-\left[-(\vartheta_t-\vartheta_t)p(\vartheta_t)+\frac{\partial}{\partial \vartheta_t}\int_{\vartheta_t}^{\infty}(y_t-\vartheta_t)p(y_t)dy_t\right]\right)\\
&\frac{1}{\xi}\left(\int_{-\infty}^{\vartheta_t}-p(y_t)dy_t+\int_{\vartheta_t}^{\infty}p(y_t)dy_t\right)\\
&\frac{1}{\xi}(+1-2P(\vartheta_t))
\end{aligned}
$$

In the third line Leibniz's rule has been applied because the integration bound depends on the variable being differentiated over. Setting the last line to zero to find the minimum reveals that $F(\vartheta_t)=0.5$, meaning that ϑ_t is the median. Therefore the function is consistent for the median if $\gamma=1$ and $\xi << e(y_t,\vartheta_t)=y_t-\vartheta_t$

A.3 The variance

It is shown that the Barron loss function is a strictly consistent scoring function for the variance functional $T(F)=\mathrm{Var}(Y_t)$ if and only if $\gamma=2$, under the assumption that $\mathbb{E}[Y_t]=0$.

The residual is defined as:

$$e(y_t,\vartheta_t)=y_t^2-\vartheta_t,$$

so that the forecast ϑ_t aims to match the second moment $\mathbb{E}[y_t^2]=\mathrm{Var}(y_t)$.

Consider the case where $\gamma=2$, in this case the Barron loss simplifies to:

$$\rho(e(y_t,\vartheta_t),\gamma=2,\xi)=\frac{1}{2}\left(\frac{y_t^2-\vartheta_t}{\xi}\right)^2.$$

Taking the expected loss:

$$\mathbb{E}_F\left[\frac{1}{2}\left(\frac{y_t^2-\vartheta_t}{\xi}\right)^2\right],$$

differentiating with respect to ϑ_t:

$$\frac{\partial}{\partial\vartheta_t}\mathbb{E}_F\left[\frac{1}{2}\left(\frac{y_t^2-\vartheta_t}{\xi}\right)^2\right]=\mathbb{E}_F\left[-\frac{(y_t^2-\vartheta_t)}{\xi^2}\right]=0.$$

This yields $\vartheta_t=\mathbb{E}[y_t^2]$, which equals $\mathrm{Var}(y_t)$ under the assumption $\mathbb{E}[y_t]=0$.

Since the loss is convex in ϑ_t, this critical point is the unique global minimum. Hence, for $\gamma=2$, the Barron loss is strictly consistent for the variance functional.

Now consider the more general case where $\gamma\neq\{2,0,-\infty\}$ in this case, the general Barron loss becomes

$$\rho(e(y_t,\vartheta_t),\gamma,\xi)=-\frac{|\gamma-2|}{\gamma}\left(\left(\frac{e(y_t,\vartheta_t)^2}{\xi^2|\gamma-2|}+1\right)^{\gamma/2}-1\right),$$

where $e(y_t,\vartheta_t)=y_t^2-\vartheta_t$.

Differentiating the expected loss with respect to ϑ_t gives:

$$\frac{\partial}{\partial\vartheta_t}\mathbb{E}_F\left[-\frac{|\gamma-2|}{\gamma}\left(\left(\frac{(y_t^2-\vartheta_t)^2}{\xi^2|\gamma-2|}+1\right)^{\gamma/2}-1\right)\right]=\mathbb{E}_F\left[-\frac{(y_t^2-\vartheta_t)}{\xi^2}\left(\frac{(y_t^2-\vartheta_t)^2}{\xi^2|\gamma-2|}+1\right)^{\gamma/2-1}\right]=0.$$

This function, however, generally does not minimise at $vartheta_t=\mathbb{E}[y_t^2]$. Therefore, the solution to this estimating equation is generally not at $\vartheta_t=\mathbb{E}[y_t^2]$, but may lie elsewhere, depending on the distribution of Y_t.

As a result, for $\gamma\neq\{2,0,-\infty\}$, the expected score is not uniquely minimised at the true second moment. Hence, the scoring function is not strictly consistent for the variance.

In the case that $\gamma = 0$, the loss function is given by,

$$\rho(e(y_t, \vartheta_t), \gamma, \xi) = \log\Big(\tfrac{1}{2}\big(\tfrac{e(y_t,\vartheta_t)}{\xi}\big)^2 + 1\Big)$$

Differentiating the function (in expectation) with respect to ϑ_t results in:

$$\frac{\partial}{\partial \vartheta_t}\mathbb{E}_F\left[\log\Big(\tfrac{1}{2}\big(\tfrac{e(y_t,\vartheta_t)}{\xi}\big)^2 + 1\Big)\right] = \mathbb{E}_F\left[-\frac{2(y_t^2 - \vartheta_t)}{(y_t - \vartheta_t)^2 + 2c^2}\right] = 0$$

This function is also generally not minimised at $\vartheta_t = \mathbb{E}[y_t^2]$, therefore, the scoring function is inconsistent.

The same arguments hold for the case where $\gamma = -\infty$, in that case the scoring function is also not minimised at $\vartheta_t = \mathbb{E}[y_t^2]$ and therefore not consistent.

The same special case hold for $\gamma = 1$ if $\xi << e(y_t, \vartheta_t) = y_t^2 - \vartheta_t$. Then the loss function is consistent for the median.

B Verification of proposition 2, stationarity

Verification of Stationarity

Proposition: Assume that (z_t) is stationary and ergodic. Suppose that

$$\text{(i)} \quad \mathbb{E}\Big[\log^+\big|\psi\big(g(\vartheta^0, \varepsilon_t),\, \vartheta^0,\, \theta\big)\big|\Big] < \infty, \quad \vartheta^0 \in F \subset \mathbb{R}; \tag{11}$$

$$\text{(ii)} \quad \mathbb{E}\big[\log \Lambda_t\big] < 0, \quad \Lambda_t = \sup_{\vartheta}\Big|\gamma \frac{\partial}{\partial \vartheta}\psi\big(g(\vartheta, \varepsilon_t),\, \vartheta,\, \theta\big) + \beta\Big| \tag{12}$$

Then there exist unique strictly stationary and ergodic solutions to $\{\vartheta_t\}_{t\in\mathbb{Z}}$ and $\{y_t\}_{t\in\mathbb{Z}}$ to $y_t = g(\vartheta_t, \varepsilon_t)$ and (1).

Verification:

Condition (i)

$$\mathbb{E}\Big[\log^{+}\big|\psi\big(g(\vartheta^{0},\varepsilon_{t}),\ \vartheta^{0},\ \theta\big)\big|\Big] < \infty, \quad \vartheta^{0} \in F \subset \mathbb{R};$$

For the proposed model, we have that $y_t = g(\vartheta_t, \varepsilon_t)$ and the updating equation of $\vartheta_{t+1} = \varphi(z_t, \vartheta_t) = \omega + \alpha\psi(y_t, \vartheta_t, \theta) + \beta\vartheta_t$. In this proposition ϑ^0 is to be held constant. ψ is given by

$$\psi(y, \vartheta^0, \theta) = -\frac{\partial\rho\big(e(y,\vartheta^0)\big)}{\partial e(y,\vartheta)}\,\frac{\partial e(y,\vartheta^0)}{\partial\vartheta}$$

It is assumed that $y < \infty$.

$$\begin{aligned}
|\psi(y,\vartheta^0,\theta)| &= \left|-\frac{\partial\rho\big(e(y,\vartheta^0)\big)}{\partial e(y,\vartheta)}\,\frac{\partial e(y,\vartheta^0)}{\partial\vartheta}\right| \\
&= \left|\frac{\partial\rho\big(e(y,\vartheta^0)\big)}{\partial e(y,\vartheta)}\,\frac{\partial e(y,\vartheta^0)}{\partial\vartheta}\right| \\
&= \left|\frac{\partial\rho\big(e(y,\vartheta^0)\big)}{\partial e(y,\vartheta)}\right| \cdot \left|\frac{\partial e(y,\vartheta^0)}{\partial\vartheta}\right|
\end{aligned}$$

it is assumed that $\left|\frac{\partial e(y,\vartheta^0)}{\partial\vartheta}\right| < \infty, \forall y, \vartheta^0$.

For the conditions on $\left|\frac{\partial\rho\left(e(y,\vartheta^0)\right)}{\partial e(y,\vartheta)}\right|$ each of the cases of γ are checked. For $\gamma = 2$ the derivative grows with $e(y,\vartheta^0)$, but it is assumed that $y < \infty$. In the paper of Barron (2019), the authors note that $\frac{\partial}{\partial x}\rho(x,\gamma,\xi) \leq \frac{|x|}{\xi^2} < \infty$, for $\gamma \leq 2$ with $\xi > 0$. So it is now concluded that both derivatives are finite. This means that

$$|\psi(y,\vartheta^0,\theta)| < \infty, \quad \log^{+}|\psi(y,\vartheta^0,\theta)| < \infty.$$

Taking expectations gives the desired results.

Condition (ii)

To verify

$$\mathbb{E}\big[\log \Lambda_t\big] < 0, \quad \Lambda_t = \sup_{\vartheta}\Big|\alpha \frac{\partial}{\partial \vartheta}\psi\big(g(\vartheta, \epsilon_t),\, \vartheta,\, \theta\big) + \beta\Big| \tag{13}$$

First, consider the second-order derivative

$$\left|\frac{\partial \psi(y, \vartheta, \theta)}{\partial \vartheta}\right| = \left|\frac{\partial^2 \rho(e(y, \vartheta))}{\partial e(y, \vartheta)^2} \cdot \left(\frac{\partial e(y, \vartheta)}{\partial \vartheta}\right)^2 + \frac{\partial \rho(e(y, \vartheta))}{\partial e(y, \vartheta)} \cdot \frac{\partial^2 e(y, \vartheta)}{\partial \vartheta^2}\right|$$

The following expression can be simplified

$$\sup_{\vartheta}\left|\alpha \frac{\partial \psi(y, \vartheta, \theta)}{\partial \vartheta} + \beta\right| = \left|\alpha \frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta} + \beta\right|$$

where ϑ^* is the ϑ that maximizes the expression. By the triangle inequality, we have that

$$\left|\alpha \frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta} + \beta\right| \leq |\alpha| \left|\frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta}\right| + |\beta|$$

For stationarity to hold and by Jensens inequality (because $\log(\cdot)$ is concave)

$$\mathbb{E}\left[\log\left(|\alpha| \left|\frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta}\right| + |\beta|\right)\right] \leq \log\left(\mathbb{E}\left[|\alpha| \left|\frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta}\right| + |\beta|\right]\right) < 0$$

Which is equivalent to

$$\mathbb{E}\left[|\alpha| \left|\frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta}\right| + |\beta|\right] = |\alpha| \mathbb{E}\left[\left|\frac{\partial \psi(y, \vartheta^*, \theta)}{\partial \vartheta}\right|\right] + |\beta| < 1$$

Barron (2019) derive the following conditions

$$\left|\frac{\partial^2 \rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\right| \leq \frac{1}{\xi^2}$$

$$\left|\frac{\partial \rho(e(y,\vartheta)}{\partial e(y,\vartheta)}\right| \leq \begin{cases} \frac{1}{\xi} & \text{if } \gamma \leq 1 \\ \frac{|e(y,\vartheta)|}{\xi^2} & \text{if } \gamma \leq 2 \end{cases} \tag{14}$$

Which results in the following bounds

$$\left|\frac{\partial \psi(y,\vartheta,\theta)}{\partial \vartheta}\right| \leq \frac{1}{\xi^2}\left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{|e(y,\vartheta)|}{\xi^2}\left|\left(\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\right)\right| \quad \text{if } \gamma \leq 2$$

and

$$\left|\frac{\partial \psi(y,\vartheta,\theta)}{\partial \vartheta}\right| \leq \frac{1}{\xi^2}\left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{1}{\xi}\left|\left(\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\right)\right| \quad \text{if } \gamma \leq 1$$

Case 1: $\gamma \leq 1$

Using the bound for $\gamma \leq 1$:

$$\begin{aligned} |\alpha|\mathbb{E}\left[\left|\frac{\partial \psi(y,\vartheta^*,\theta)}{\partial \vartheta}\right|\right] &\leq |\alpha|\mathbb{E}\left[\frac{1}{\xi^2}\left(\frac{\partial e(y,\vartheta^*)}{\partial \vartheta}\right)^2 + \frac{1}{\xi}\left|\frac{\partial^2 e(y,\vartheta^*)}{\partial \vartheta^2}\right|\right] \\ &= \frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y,\vartheta^*)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}\left[\left|\frac{\partial^2 e(y,\vartheta^*)}{\partial \vartheta^2}\right|\right] \end{aligned}$$

For stationarity, we need:

$$\frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y,\vartheta^*)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}\left[\left|\frac{\partial^2 e(y,\vartheta^*)}{\partial \vartheta^2}\right|\right] + |\beta| < 1$$

Case 2: $1 < \gamma \leq 2$

Using the bound for $\gamma \leq 2$:

$$|\alpha|\mathbb{E}\left[\left|\frac{\partial\psi(y,\vartheta^*,\theta)}{\partial\vartheta}\right|\right] \leq |\alpha|\mathbb{E}\left[\frac{1}{\xi^2}\left(\frac{\partial e(y,\vartheta^*)}{\partial\vartheta}\right)^2 + \frac{|e(y,\vartheta^*)|}{\xi^2}\left|\frac{\partial^2 e(y,\vartheta^*)}{\partial\vartheta^2}\right|\right]$$
$$= \frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y,\vartheta^*)}{\partial\vartheta}\right)^2\right] + \frac{|\alpha|}{\xi^2}\mathbb{E}\left[|e(y,\vartheta^*)|\cdot\left|\frac{\partial^2 e(y,\vartheta^*)}{\partial\vartheta^2}\right|\right]$$

For stationarity, we need

$$\frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y,\vartheta^*)}{\partial\vartheta}\right)^2\right] + \frac{|\alpha|}{\xi^2}\mathbb{E}\left[|e(y,\vartheta^*)|\cdot\left|\frac{\partial^2 e(y,\vartheta^*)}{\partial\vartheta^2}\right|\right] + |\beta| < 1$$

Special Case: Location and Volatility Error Model

If we assume that $e(y,\vartheta) = y_t - \vartheta_t$ for a location model and $e(y_t,\vartheta_t) = y_t^2 - \vartheta_t$ for a volatility model then

$$\frac{\partial e(y,\vartheta)}{\partial\vartheta} = -1$$
$$\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2} = 0$$

This simplifies our conditions to

For both $\gamma \leq 1$ and $1 < \gamma \leq 2$:

$$\frac{|\alpha|}{\xi^2} + |\beta| < 1$$

C Verification of proposition 3, existence of bounded unconditional moments

Verification of Existence of bounded unconditional moments

Proposition:

Under the assumptions of Proposition 1, if the sequence $\{\Lambda_t\}$ is i.i.d. and for some $r > 0$

$$\mathbb{E}\left[\left|\psi\big(g(\vartheta^0, \epsilon_t), X_t, \vartheta^0, \theta\big)\right|^r\right] \;<\; \infty \quad \text{and} \quad \mathbb{E}\left[\sup_{\vartheta}\left|\frac{\partial\psi\big(g(\vartheta, \epsilon_t), X_t, \vartheta, \theta\big)}{\partial\vartheta}\right|^r\right] \;<\; \infty,$$

then ϑ_t satisfies $\mathbb{E}\,|\vartheta_t|^s < \infty$ for some $s > 0$.

Verification:

C.1 Condition (i)

If one assumes that ε_t is i.i.d, then Λ_t (13) is i.i.d.

The first part then reduces to

$$\mathbb{E}\left[\left|\psi\big(g(\vartheta^0, \epsilon_t), \vartheta^0, \theta\big)\right|\right] \;<\; \infty,$$

If we set $r = 1$.

In B it is shown that when $\gamma \leq 2$, and the other assumptions mentioned there, then $\left|\psi\big(g(\vartheta^0, \epsilon_t), X_t, \vartheta^0, \theta\big)\right| \;<\; \infty$. Taking expectation gives the desired result and shows that the first condition holds.

C.2 Condition (ii)

To show that the second condition holds,

$$\begin{aligned}
&\mathbb{E}\left[\sup_{\vartheta}\Big|\frac{\partial\psi\big(g(\vartheta,\epsilon_t),\vartheta,\theta\big)}{\partial\vartheta}\Big|^r\right] \\
=&\mathbb{E}\left[\sup_{\vartheta}\Big|\frac{\partial^2\rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\cdot\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2+\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\cdot\left(\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right)\Big|^r\right]<\infty
\end{aligned}$$

We take again the case where $r=1$. Notice that

$$\begin{aligned}
&\sup_{\vartheta}\Big|\frac{\partial^2\rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\cdot\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2+\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\cdot\left(\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right)\Big| \\
&\leq\sup_{\vartheta}\Big|\frac{\partial^2\rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\cdot\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big|+\sup_{\vartheta}\Big|\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\cdot\left(\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right)\Big| \\
&=\sup_{\vartheta}\Big|\frac{\partial^2\rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\Big|\cdot\Big|\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big|+\sup_{f}\Big|\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\Big|\cdot\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big| \\
&\leq\frac{1}{\xi^2}\Big|\sup_{\vartheta}\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big|+\sup_{f}\Big|\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\Big|\cdot\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big| \\
&\leq\frac{1}{\xi^2}\cdot\sup_{\vartheta}\Big|\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big|+\sup_{f}\frac{|e(y,\vartheta)|}{\xi^2}\cdot\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big|
\end{aligned}$$

Where in the last step we use that $\gamma\leq 2$. As before, we now assume that $\frac{\partial e(y,\vartheta)}{\partial\vartheta}<\infty$,$\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}<\infty$ and $y_t<\infty$. If we then assume that $e(y,\vartheta)$ is bounded, then the following condition holds. (If $\gamma\leq 1$ then $\frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\leq\frac{1}{\xi^2}$ and we don't have to assume the latter.)

$$\frac{1}{\xi^2}\cdot\sup_{\vartheta}\Big|\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big|+\sup_{f}\frac{|e(y,\vartheta)|}{\xi^2}\cdot\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big|<\infty$$

Such that by taking expectations.

$$\mathbb{E}\left[\frac{1}{\xi^2}\cdot\sup_{\vartheta}\Big|\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big|+\sup_{f}\frac{|e(y,\vartheta)|}{\xi^2}\cdot\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big|\right]<\infty$$

Note that these assumptions are mild, if we consider a location model with $e(y_t, \vartheta_t) = y_t - \vartheta_t$ or a volatility model with $e(y_t, \vartheta_t) = y_t^2 - \vartheta_t$, then these conditions are satisfied without any assumptions on γ. In more general, if $\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2} = 0$, then no requirement on γ is needed.

Since both moment conditions from Lemma 2 are satisfied, there exists some $s \in (0, 1)$ such that $\mathbb{E}[|\vartheta_t|^s] < \infty$, ensuring that there exist bounded finite moments.

D Verification of proposition 4, invertibility of the QSD filter

Verification of Invertibility of the QSD filter

Proposition:

Let $\{y_t, X_t\}_{t\in\mathbb{Z}}$ be stationary and ergodic, and suppose that

(i) for all $\theta \in \Theta$, there exists $\vartheta^0 \in F$ such that $\mathbb{E}\big[\log^+\big|\psi(y_t, \vartheta^0, \theta)\big|\big] < \infty$,

(ii) $\mathbb{E}\Big[\log \sup_{\vartheta\in F} \sup_{\theta\in\Theta} \Big|\alpha \dfrac{\partial}{\partial \vartheta} \psi(y_t, \vartheta, \theta) + \beta\Big|\Big] < 0.$

Then for all $\theta \in \Theta$, there exists a unique stationary and ergodic solution to $\{\vartheta_t(\theta)\}_{\theta\in\Theta}$ to $\vartheta_{t+1}(\theta) = \omega + \alpha\psi(y_t, X_t, \vartheta_t(\theta), \theta) + \beta\vartheta_t(\theta),\ t \in \mathbb{Z}$. Furthermore, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathbb{C}(\Theta, F)$, there exists $\varrho \in (0, 1)$ such that

$$\varrho^{-t} \sup_{\theta\in\Theta} \big|\hat{\vartheta}_t(\theta) - \vartheta_t(\theta)\big| \to 0 \quad \text{a.s. as} \quad (t \to \infty).$$

The model is then said to be uniformly invertible.

Verification:

D.1 Condition (i)

This condition is similar to Condition (i) in Lemma 1. The difference is that for all $\theta \in \Theta$ there exists a $\vartheta^0 \in F$ such that the given expectation is finite. Again have a look at

$$\begin{aligned}
|\psi(y, \vartheta^0, \theta)| &= \left| -\frac{\partial \rho\big(e(y, \vartheta^0)\big)}{\partial e(y, \vartheta)} \frac{\partial e(y, \vartheta^0)}{\partial \vartheta} \right| \\
&= \left| \frac{\partial \rho\big(e(y, \vartheta^0)\big)}{\partial e(y, \vartheta)} \frac{\partial e(y, \vartheta^0)}{\partial \vartheta} \right| \\
&= \left| \frac{\partial \rho\big(e(y, \vartheta^0)\big)}{\partial e(y, \vartheta)} \right| \cdot \left| \frac{\partial e(y, \vartheta^0)}{\partial \vartheta} \right|
\end{aligned}$$

If it is assumed $y < \infty$ and that $\left|\frac{\partial e(y,\vartheta^0)}{\partial \vartheta}\right| < \infty, \forall y, \vartheta^0$, then since $\left|\frac{\partial \rho\left(e(y,\vartheta^0)\right)}{\partial e(y,\vartheta)}\right| \leq \frac{|e(y,\vartheta^0)|}{\xi^2}$ for $\gamma \leq 2$, for all values of $\theta \in \Theta$ there must exist a ϑ^0 for which ψ is finite (since it is assumed that for all ϑ^0 it is finite). Taking the $\log^+$ and expectation retains its boundedness.

D.2 Condition (ii)

Condition (ii) is very similar to Condition (ii). The steps taken are the same, but we now take that

$$\sup_{\vartheta \in F} \sup_{\theta \in \Theta} \left| \alpha \frac{\partial \psi(y, \vartheta, \theta)}{\partial \vartheta} + \beta \right| = \sup_{\vartheta \in F\, \theta \in \Theta} \left| \alpha \frac{\partial \psi(y, \vartheta, \theta)}{\partial \vartheta} + \beta \right| = \left| \alpha \frac{\partial \psi(y, \vartheta^*, \theta^*)}{\partial \vartheta} + \beta \right|$$

Where ϑ^*, θ^* is the combination($(\vartheta^*, \theta^*) \in F \times \Theta$) that maximizes the function. By the triangle inequality and taking expectations, we then obtain the following result

$$|\alpha| \mathbb{E}\left[\left|\frac{\partial \psi(y, \vartheta^*, \theta^*)}{\partial \vartheta}\right|\right] + |\beta| < 1$$

Using the bounds of the derivatives of the Barron loss function, we obtain

Case 1: $\gamma \leq 1$

Using the bound for $\gamma \leq 1$:

$$
\begin{aligned}
|\alpha|\mathbb{E}\left[\left|\frac{\partial \psi(y, \vartheta^*, \theta^*)}{\partial \vartheta}\right|\right] &\leq |\alpha|\mathbb{E}\left[\frac{1}{\xi^2}\left(\frac{\partial e(y, \vartheta^*)}{\partial \vartheta}\right)^2 + \frac{1}{\xi}\left|\frac{\partial^2 e(y, \vartheta^*)}{\partial \vartheta^2}\right|\right] \\
&= \frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y, \vartheta^*)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}\left[\left|\frac{\partial^2 e(y, \vartheta^*)}{\partial \vartheta^2}\right|\right]
\end{aligned}
$$

For stationarity, we need

$$
\frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y, \vartheta^*)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi}\mathbb{E}\left[\left|\frac{\partial^2 e(y, \vartheta^*)}{\partial \vartheta^2}\right|\right] + |\beta| < 1
$$

Case 2: $1 < \gamma \leq 2$

Using the bound for $\gamma \leq 2$:

$$
\begin{aligned}
|\alpha|\mathbb{E}\left[\left|\frac{\partial \psi(y, \vartheta^*, \theta^*)}{\partial \vartheta}\right|\right] &\leq |\alpha|\mathbb{E}\left[\frac{1}{\xi^2}\left(\frac{\partial e(y, \vartheta^*)}{\partial \vartheta}\right)^2 + \frac{|e(y, \vartheta^*)|}{\xi^2}\left|\frac{\partial^2 e(y, \vartheta^*)}{\partial \vartheta^2}\right|\right] \\
&= \frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y, \vartheta^*)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi^2}\mathbb{E}\left[|e(y, \vartheta^*)| \cdot \left|\frac{\partial^2 e(y, \vartheta^*)}{\partial \vartheta^2}\right|\right]
\end{aligned}
$$

For stationarity, we need

$$
\frac{|\alpha|}{\xi^2}\mathbb{E}\left[\left(\frac{\partial e(y, \vartheta^*)}{\partial \vartheta}\right)^2\right] + \frac{|\alpha|}{\xi^2}\mathbb{E}\left[|e(y, \vartheta^*)| \cdot \left|\frac{\partial^2 e(y, \vartheta^*)}{\partial \vartheta^2}\right|\right] + |\beta| < 1
$$

We now have confirmed that the DGP is invertible, since lemma 3 from Blasques et al. (2023) holds.

E Verification of proposition 5, invertibility properties for the derivatives of the filter

Proposition:

Let the conditions of Proposition 3 hold, assume that ψ admits continuous second-order derivatives with respect to its last two components, and suppose that

$$\text{(i) for all } \theta \in \Theta, \quad \mathbb{E}\left[\log^+ |\psi_t| + \log^+ \left\|\frac{\partial \psi_t}{\partial \theta}\right\| + \log^+ \left|\frac{\partial \psi_t}{\partial \vartheta}\right| + \log^+ |\vartheta_t(\theta)|\right] < \infty.$$

Then, for all $\theta \in \Theta$, there exists a unique strictly stationary and ergodic solution $\{\vartheta'_t(\theta)\}_{t\in\mathbb{Z}}$ to (8). If in addition

$$\text{(ii)} \quad \mathbb{E}\left[\log^+ \left(\sup_f \left|\frac{\partial \psi_t}{\partial \vartheta}\right| + \sup_{\vartheta,\theta} \left\|\frac{\partial^2 \psi_t}{\partial \theta \partial \vartheta}\right\| + \sup_f \left\|\frac{\partial^2 \psi_t}{\partial \vartheta^2}\right\| + \sup_\theta \|\vartheta'_t(\theta)\|\right)\right] < \infty,$$

then, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, F)$ and $\tilde{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, \mathbb{R}^\rho)$, there exists $\varrho \in (0,1)$ such that

$$\varrho^{-t} \sup_{\theta \in \Theta} \left\|\tilde{\vartheta}'_t(\theta) - \vartheta'_t(\theta)\right\| \to 0 \quad \text{a.s. as } t \to \infty.$$

furthermore, assuming that

$$\text{(iii) for all } \theta \in \Theta, \quad \mathbb{E}\left[\log^+ \left\|\frac{\partial^2 \psi_t}{\partial \theta \partial \theta^\top}\right\| + \log^+ \left\|\frac{\partial^2 \psi_t}{\partial \theta \partial \vartheta}\right\| + \log^+ \left\|\frac{\partial^2 \psi_t}{\partial \vartheta^2}\right\|\right] < \infty,$$

then, for all $\theta \in \Theta$, there exists a unique strictly stationary and ergodic solution $\{\vartheta''_t(\theta)\}_{t\in\mathbb{Z}}$ to (10). Under the additional assumption

$$\text{(iv)} \quad \mathbb{E}\left[\log^+ \left(\sup_{\vartheta,\theta_i,\theta_j} \left|\frac{\partial^3 \psi_t}{\partial \theta_i \partial \theta_j \partial \vartheta}\right| + \sup_{\vartheta,\theta_i} \left\|\frac{\partial^3 \psi_t}{\partial \theta_i \partial \vartheta^2}\right\| + \sup_f \left\|\frac{\partial^3 \psi_t}{\partial \vartheta^3}\right\|\right)\right] < \infty,$$

then, for all starting functions $\hat{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, F)$, $\tilde{\vartheta}_1(\cdot) \in \mathcal{C}(\Theta, \mathbb{R}^\rho)$, and $\tilde{\vartheta}_1''(\cdot) \in \mathcal{C}(\Theta, \mathbb{R}^{\rho^2})$, there exists $\varrho \in (0,1)$ such that

$$\varrho^{-t} \sup_{\theta \in \Theta} \left\| \tilde{\vartheta}_t''(\theta) - \vartheta_t''(\theta) \right\| \to 0 \quad \text{a.s. as } t \to \infty.$$

Verification:

E.1 Condition (i)

To check condition one, we split the expectation into the four separate parts and evaluate them separately.

$$\begin{aligned} &\mathbb{E}\left[\log^+ |\psi_t| + \log^+ \left\| \frac{\partial \psi_t}{\partial \theta} \right\| + \log^+ \left| \frac{\partial \psi_t}{\partial \vartheta} \right| + \log^+ |\vartheta_t(\theta)| \right] \\ = &\ \mathbb{E}\left[\log^+ |\psi_t|\right] + \mathbb{E}\left[\log^+ \left\| \frac{\partial \psi_t}{\partial \theta} \right\|\right] + \mathbb{E}\left[\log^+ \left| \frac{\partial \psi_t}{\partial \vartheta} \right|\right] + \mathbb{E}\left[\log^+ |\vartheta_t(\theta)|\right] \end{aligned}$$

From left to right, the first term $\mathbb{E}\left[\log^+ |\psi_t|\right] < \infty$ by the same arguments as in Lemma 1, Condition (i).

The second term is $\mathbb{E}\left[\log^+ \left\| \frac{\partial \psi_t}{\partial \theta} \right\|\right]$, this derivative only contains the variables γ and ξ, all other parameters in the vector θ result in a derivative equal to 0. For notational simplicity, those derivatives are omitted. The derivative is then given by:

$$\frac{\partial \psi_t}{\partial \theta} = \begin{bmatrix} \frac{e}{\xi^2}\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-1}\left[\frac{1}{2}\ln\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right) - \frac{e^2\left(\frac{\gamma}{2}-1\right)}{\xi^2\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right)|\gamma-2|(\gamma-2)}\right] \\ \left[-\frac{2e}{\xi^3} + \frac{e}{\xi^2}\left(\frac{\gamma}{2}-1\right)\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-2} \cdot (-2)\frac{e^2}{\xi^3|\gamma-2|}\right]\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right)^{\frac{\gamma}{2}-1} \end{bmatrix} \cdot \frac{\partial e}{\partial \vartheta}$$

Where the first row is the derivative with respect to γ and the second with respect to ξ. The derivative is given for the values of γ which are not $2, 0, -\infty$ and ($e = e(y_t, \vartheta_t)$). The derivative if finite for the values of $\gamma \neq 2, 0, -\infty$, if e is bounded. Moreover, the following condition should hold $\xi > 0$, this condition was already enforced by Barron (2019).

The third term is $\mathbb{E}\left[\log^+ \left|\frac{\partial \psi_t}{\partial \vartheta}\right|\right]$, this term is also finite. To show this, we can use a similar method as in Verification of proposition 3, existence of bounded unconditional moments.

$$
\begin{aligned}
\frac{\partial \psi}{\partial \vartheta} =& \Big|\frac{\partial^2 \rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2} \cdot \left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2 + \frac{\partial \rho(e(y,\vartheta))}{\partial e(y,\vartheta)} \cdot \left(\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\right)\Big| \\
&\leq \Big|\frac{\partial^2 \rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2} \cdot \left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2\Big| + \Big|\frac{\partial \rho(e(y,\vartheta))}{\partial e(y,\vartheta)} \cdot \left(\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\right)\Big| \\
&= \Big|\frac{\partial^2 \rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\Big| \cdot \Big|\left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2\Big| + \Big|\frac{\partial \rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\Big| \cdot \Big|\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\Big| \\
&\leq \frac{1}{\xi^2}\Big|\left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2\Big| + \Big|\frac{\partial \rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\Big| \cdot \Big|\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\Big| \\
&\leq \frac{1}{\xi^2} \cdot \Big|\left(\frac{\partial e(y,\vartheta)}{\partial \vartheta}\right)^2\Big| + \frac{|e(y,\vartheta)|}{\xi^2} \cdot \Big|\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2}\Big|
\end{aligned}
$$

In the last step, we use that $\gamma \leq 2$. As before, we now assume that $\frac{\partial e(y,\vartheta)}{\partial \vartheta} < \infty$,$\frac{\partial^2 e(y,\vartheta)}{\partial \vartheta^2} < \infty$ and $y_t < \infty$. If we then assume that $e(y,\vartheta)$ is bounded, then the condition is finite. Taking the $\log^+$ and expectations gives the desired bound.

The final term is $\mathbb{E}[\log^+ |\vartheta_t(\theta)|]$. To show boundedness of this, we use the result of C. We see that since $\mathbb{E}[|\vartheta_t|^s] < \infty$ for some $s > 0$, therefore $\mathbb{E}[\log^+ |\vartheta_t|] < \infty$ (Blasques et al. 2023).

Now we have concluded that all the elements of the sum are bounded, therefore the sum is bounded and condition (i) is satisfied.

E.2 Condition (ii)

We split the proof into four parts

First term: By the chain-rule

$$\frac{\partial\psi}{\partial\vartheta} = \Big|\frac{\partial^2\rho(e(y,\vartheta))}{\partial e(y,\vartheta)^2}\cdot\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2 + \frac{\partial\rho(e(y,\vartheta))}{\partial e(y,\vartheta)}\cdot\left(\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right)\Big|$$

following the same arguments in Verification of proposition 3, existence of bounded unconditional moments to end up in

$$\frac{\partial\psi}{\partial\vartheta} \leq \frac{1}{\xi^2}\cdot\sup_{\vartheta}\Big|\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)^2\Big| + \sup_{f}\frac{|e(y,\vartheta)|}{\xi^2}\cdot\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big|$$

Where in the last step we use that $\gamma \leq 2$. As before, we now assume that $\frac{\partial e(y,\vartheta)}{\partial\vartheta} < \infty$,$\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2} < \infty$ and $y_t < \infty$. If we then assume that $e(y,\vartheta)$ is bounded, then $\frac{\partial\psi}{\partial\vartheta}$ is bounded.

Mixed derivative: Differentiating once more in θ shows likewise that $\sup_{\vartheta,\theta}\left\|\frac{\partial^2\psi_t}{\partial\theta\partial\vartheta}\right\| < \infty$

Starting from the already-derived

$$\frac{\partial\psi(y,\vartheta,\theta)}{\partial\vartheta} = \frac{\partial^2\rho(e,\gamma,\xi)}{\partial e^2}\left(\frac{\partial e}{\partial\vartheta}\right)^2 + \frac{\partial\rho(e,\gamma,\xi)}{\partial e}\frac{\partial^2 e}{\partial\vartheta^2},$$

we differentiate once more in θ

$$\frac{\partial^2\psi(y,\vartheta,\theta)}{\partial\theta\,\partial\vartheta} = \frac{\partial}{\partial\theta}\Big(\frac{\partial^2\rho(e,\gamma,\xi)}{\partial e^2}\Big)\left(\frac{\partial e}{\partial\vartheta}\right)^2 + \frac{\partial}{\partial\theta}\Big(\frac{\partial\rho(e,\gamma,\xi)}{\partial e}\Big)\frac{\partial^2 e}{\partial\vartheta^2}$$

In each of the four cases $\gamma \in \{2, 0, -\infty, \text{other}\}$ one checks from (2)–(3) that $\frac{\partial^2\rho}{\partial e^2}$, $\frac{\partial\rho}{\partial e}$ and their γ-derivatives are bounded by constants depending only on (γ,ξ). Under the mild assumptions,

$$\sup_{\vartheta,\theta}\Big|\frac{\partial e(y,\vartheta)}{\partial\vartheta}\Big|,\ \sup_{\vartheta,\theta}\Big|\frac{\partial}{\partial\theta}\left(\frac{\partial e(y,\vartheta)}{\partial\vartheta}\right)\Big|,\ \sup_{\vartheta,\theta}\Big|\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\Big|,\ \sup_{\vartheta,\theta}\Big|\frac{\partial}{\partial\theta}\left(\frac{\partial^2 e(y,\vartheta)}{\partial\vartheta^2}\right)\Big| < \infty,$$

each term in the above display is uniformly bounded. Hence

$$\sup_{\vartheta,\theta}\left\|\frac{\partial^2\psi}{\partial\theta\,\partial\vartheta}\right\| < \infty$$

Second ϑ-derivative: A further ϑ-derivative introduces third derivatives of $\rho(e)$

$$\frac{\partial^2\psi(y,\vartheta,\theta)}{\partial\vartheta^2} = \frac{\partial^3\rho(e,\gamma,\xi)}{\partial e^3}\left(\frac{\partial e}{\partial\vartheta}\right)^3 + 3\,\frac{\partial^2\rho(e,\gamma,\xi)}{\partial e^2}\,\frac{\partial e}{\partial\vartheta}\,\frac{\partial^2 e}{\partial\vartheta^2} + \frac{\partial\rho(e,\gamma,\xi)}{\partial e}\,\frac{\partial^3 e}{\partial\vartheta^3}.$$

(all bounded when $\gamma \leq 2$, or by a constant otherwise) times factors of $\frac{\partial e}{\partial\vartheta}$, $\frac{\partial^2 e}{\partial\vartheta^2}$, $\frac{\partial^3 e}{\partial\vartheta^3}$. Assuming that those suprema with respect to ϑ are finite results in $\sup_f \|\frac{\partial^2\psi(y,\vartheta,\theta)}{\partial\vartheta^2}\| < \infty$.

Filter-derivative. If we then furthermore assume that $\mathbb{E}[\sup_\theta |\vartheta_t(\theta)) < \infty]$, then combining all the previous finite terms and taking the $\log^+$ and after that the expectation shows that condition two also holds.

E.3 Condition (iii)

Two of the three terms in the expectation in this condition were already shown to be finite under certain assumptions in the previous condition, taking the $\log^+$ maintains its boundedness. The only term that remains is

$$\log^+\left\|\frac{\partial^2\psi_t}{\partial\theta\partial\theta^\top}\right\|$$

To show boundedness of this, take the derivative with respect to γ and ξ again. This again results in boundedness for $\xi > 0$ and $\gamma \neq 2$

Therefore all terms in the expectation are finite and condition three is satisfied.

E.4 Condition (iiii)

The last condition uses third derivatives of ψ with respect to θ and ϑ (equivalently 4 order derivatives of ρ). These derivatives will result in factors of $\frac{\partial e}{\partial \vartheta}$, $\frac{\partial^2 e}{\partial \vartheta^2}$, $\frac{\partial^3 e}{\partial \vartheta^3}$, $\frac{\partial^4 e}{\partial \vartheta^4}$. Assuming that taking the supremum are finite for those terms and further more $\left\| \frac{\partial^4 \rho(e,\gamma,\xi)}{\partial e^4} \right\|$ is finite.

Shows that this last condition is also verified. And therefore lemma 4 in Blasques et al. (2023) holds.

F QLE volatility model

The Monte Carlo study on a volatility model was estimated by QLE as described in Quasi-likelihood estimation (QLE). For this the following results are needed to compute $\frac{\partial \hat{\vartheta}_{t+1}(\theta)}{\partial \theta}$ where $\theta = (\omega, \alpha, \beta, \gamma, \xi)$

The derivative of $\hat{\vartheta}_t(\theta)$ w.r.t θ is given by

$$\frac{\partial \hat{\vartheta}_{t+1}(\theta)}{\partial \theta} = \frac{\partial \omega}{\partial \theta} + \psi_t \frac{\partial \alpha}{\partial \theta} + \alpha \frac{\partial \psi_t}{\partial \theta} + \hat{\vartheta}_t(\theta) \frac{\partial \beta}{\partial \theta} + \left(\alpha \frac{\partial \psi_t}{\partial \vartheta} + \beta \right) \frac{\partial \hat{\vartheta}_t(\theta)}{\partial \theta}$$

Here $\frac{\partial \omega}{\partial \theta} = (1, 0, 0, 0, 0)'$, $\psi_t \frac{\partial \alpha}{\partial \theta} = (0, \psi_t, 0, 0, 0)'$, $\alpha \frac{\partial \psi_t}{\partial \theta} = (0, 0, 0, \alpha \frac{\partial \psi}{\partial \gamma}, \alpha \frac{\partial \psi}{\partial \xi})'$, $\hat{\vartheta}_t(\theta) \frac{\partial \beta}{\partial \theta} = (0, 0, \hat{\vartheta}_t, 0, 0)'$. When we set $e = e(y_t, \vartheta_t) = y_t^2 - \vartheta_t$ then the second order derivative needed are

$$\frac{\partial \psi}{\partial \vartheta} = \frac{\partial^2 \rho}{\partial \vartheta^2} = \frac{\partial^2 \rho}{\partial e^2} = \begin{cases} \dfrac{1}{\xi^2}, & \gamma = 2, \\ \dfrac{1}{\xi^2} \dfrac{1 - \frac{1}{2}(e/\xi)^2}{\left(1 + \frac{1}{2}(e/\xi)^2\right)^2}, & \gamma = 0, \\ \dfrac{1}{\xi^2}\Big(1 - \frac{e^2}{\xi^2}\Big) \exp\Big(-\frac{1}{2}(e/\xi)^2\Big), & \gamma = -\infty, \\ \dfrac{1}{\xi^2}\left(\frac{(e/\xi)^2}{|\gamma-2|} + 1\right)^{\frac{\gamma}{2}-2}\Big[\frac{(e/\xi)^2}{|\gamma-2|} + 1 \ + \ \left(\frac{\gamma}{2} - 1\right) \frac{2e^2}{\xi^2\,|\gamma-2|}\Big], & \text{otherwise.} \end{cases}$$

$$\frac{\partial \psi}{\partial \gamma} = \frac{e\left(\frac{e^2}{\xi^2|\gamma-2|} + 1\right)^{\frac{\gamma}{2}-1}\left(\frac{\ln\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right)}{2} - \frac{e^2\left(\frac{\gamma}{2}-1\right)}{\xi^2\left(\frac{e^2}{\xi^2|\gamma-2|}+1\right)|\gamma-2|(\gamma-2)}\right)}{\xi^2} \quad , \gamma \neq 0, 2, -\infty$$

$$\frac{\partial \psi}{\partial \xi} = \begin{cases} -\dfrac{2e}{\xi^3}, & \gamma = 2, \\ -\dfrac{8e\xi}{\left(e^2 + 2c^2\right)^2}, & \gamma = 0, \\ \Big[-\dfrac{2e}{\xi^3} + \dfrac{e^3}{\xi^5}\Big] \exp\Big(-\frac{1}{2}\left(\frac{e}{\xi}\right)^2\Big), & \gamma = -\infty, \\ \Big[-\dfrac{2e}{\xi^3} \ + \ \dfrac{e}{\xi^2}\Big(\frac{\gamma}{2} - 1\Big)\Big(\frac{e^2}{\xi^2|\gamma-2|} + 1\Big)^{\frac{\gamma}{2}-2} \cdot (-2)\frac{e^2}{\xi^3|\gamma-2|}\Big]\Big(\frac{e^2}{\xi^2|\gamma-2|} + 1\Big)^{\frac{\gamma}{2}-1}, & \text{otherwise.} \end{cases}$$

G Reduction to known GAS models

The proposed framework flexibility allows for reductions to know GAS models. This section shows two examples. When we consider a location model, with $e(y_t, \vartheta_t) = y_t - \vartheta_t$, then when $\gamma = 2$, the updating equation of the time-varying parameter is given by: $\vartheta_{t+1} = \omega + \alpha\frac{y_t - \vartheta_t}{\xi^2} + \beta\vartheta_t$. This therefore reduces to a location-GAS model assuming a Gaussian error distribution. To see this, consider the log-likelihood function

$$\log(y_t|\theta_t, \vartheta_t) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y_t - \vartheta_t)^2}{2\sigma^2}$$

Taking the derivative w.r.t. the time-varying parameter results in

$$\frac{\partial \log(y_t|\theta_t, \vartheta_t)}{\partial \vartheta_t} = \frac{y_t - \vartheta_t}{\sigma^2}$$

Which can be scaled by some scaling matrix S_t (e.g I or $\mathcal{I}_t = \mathbb{E}\left[\left(\frac{\partial \log(y_t|\theta_t,\vartheta_t)}{\partial \vartheta_t}\right)^2 \middle| \mathcal{F}_{t-1}\right]$). The resulting equation is given by

$$\vartheta_{t+1} = \omega + \alpha\frac{y_t - \vartheta_t}{\sigma^2} + \beta\vartheta_t$$

If $\xi = \sigma$ then this expression is equivalent to the proposed QSD model under $\gamma = 2$, otherwise they proportionally equivalent.

A similar result holds for a reduction to the GARCH model in the case that $\gamma = 2$ and we have a volatility model with $(e(y_t, \vartheta_t) = y_t^2 - \vartheta_t)$. We see that the structure becomes similar to the GARCH model. The QSD model with $\gamma = 2$ is exactly equivalent to the GARCH model after a rescaling by ξ^2 and some relabelling of the parameters.

These examples show that the proposed QSD filter nests well know GAS models that have been used in many applications. But due to the flexibility and data-driven nature of the loss function it does not limit it self to any prefixed models. Instead, it adapts flexibly to the characteristics of the data, allowing for estimation of the most appropriate model structure.

H Monte Carlo study with 3000 time steps excluding outliers

This table shows the result of the Monte Carlo study done on a 3000 time steps using a 1000 repetitions, as set up in 2.5. The results are discussed in Section 3.1.

Table 5: Monte Carlo study excluding outliers

		$T = 3000$					
		ω	α	β	ξ	γ	ν
GARCH	Est.	0.087	0.059	0.682	–	–	–
β_t	Est.	0.088	0.061	0.678	–	–	89.060
$\gamma = -\infty$	Est.	0.075	0.174	0.754	1.172	–	–
	Bias	0.005	0.064	-0.046	-0.028	–	–
	RMSE	0.039	0.120	0.140	0.579	–	–
	MAE	0.024	0.072	0.085	0.266	–	–
$\gamma = 0$	Est.	0.070	0.171	0.754	1.169	–	–
	Bias	**0.000**	0.061	-0.046	**-0.034**	–	–
	RMSE	0.026	0.118	0.125	**0.258**	–	–
	MAE	0.019	0.065	0.079	0.184	–	–
$\gamma = 1$	Est.	0.070	0.151	0.750	1.150	–	–
	Bias	**0.000**	0.041	-0.060	-0.051	–	–
	RMSE	0.025	0.088	0.123	0.238	–	–
	MAE	0.017	0.046	0.078	0.160	–	–
$\gamma = 2$	Est.	0.075	0.117	0.721	1.170	–	–
	Bias	0.005	**0.007**	-0.079	0.058	–	–
	RMSE	0.026	**0.057**	0.140	0.320	–	–
	MAE	0.017	**0.030**	0.095	0.184	–	–
Est. γ	Est.	0.067	0.137	0.765	1.231	0.953	–
	Bias	-0.003	0.027	**-0.035**	-0.038	-0.047	–
	RMSE	**0.021**	0.060	**0.095**	0.298	0.602	–
	MAE	**0.014**	0.033	**0.058**	**0.124**	0.238	–

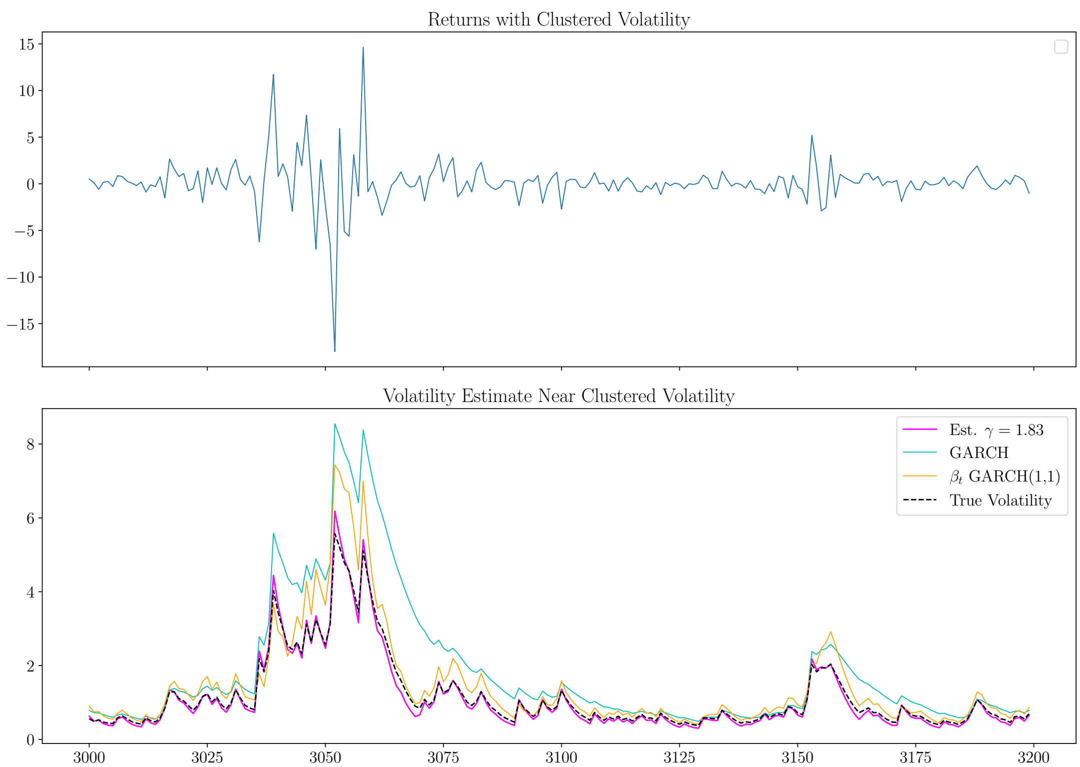
NOTE: Monte Carlo study of 1000 repetitions on the finite sample performance of the QSD filter. The models were estimated on 3000 observations. The true parameters are $\theta_{\text{true}} = (\omega, \beta, \alpha, \gamma, \xi) = (0.07, 0.8, 0.11, 1.0, 1.2)$ and the data has not been contaminated with outliers. The QSD models were estimated using Quasi Likelihood estimation and the GARCH and β_t GARCH(1,1) were estimated using maximum likelihood. Bold numbers denote the most accurate (closest to zero) value for each parameter within a given error metric.

I Extra plots Clustering Volatility

This section shows some extra plots that are part of the analysis done in section 3.3.2. Figure 6 shows the same plot as 5 but without the fixed γ models. This allows for a better comparison between the QSD model and the GARCH and β_t GARCH(1,1) model.

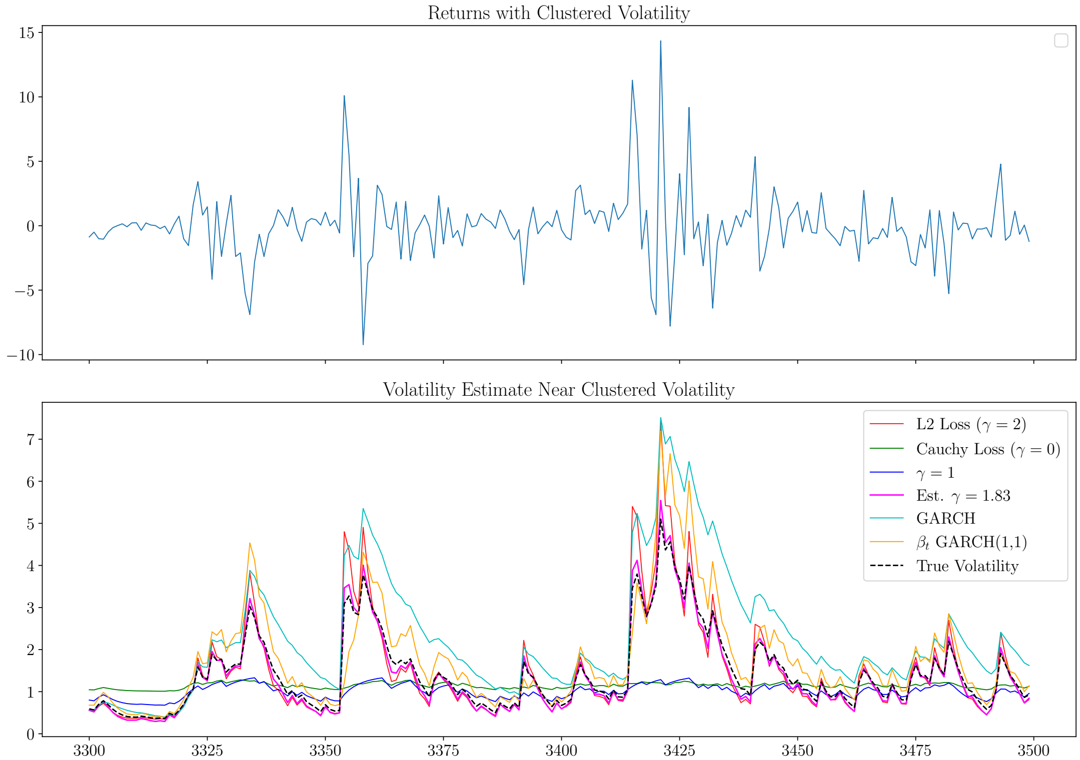
In Figure 7 a different snapshot of the same time series analysed in Section 3.3.2 is shown, this depicts a more densely clustered region.

Figure 6: Clustering Volatilities showing the most important models



NOTE: The same plot as in 3.3.2, but with only the GARCH model, β_t GARCH$(1, 1)$ and the estimated γ model. This plot allows for a better comparison between the estimated γ model and the benchmark models.

Figure 7: Multiple subsequent clusters of volatility



NOTE: Same time series as 3.3.2 but focussed on an area with more subsequent clusters. This shows that the model is able to capture multiple subsequent volatility clusters.