Robustifying Score-Driven Updates by the PseudoSpherical Scoring Rule

Louis Gehringer 13322842
Supervisor: Ramon de Punder
University of Amsterdam

June 27, 2024

Abstract

Score-driven (SD) models, introduced by Creal et al. (2013) and Harvey (2013), are popular tools for modelling time-varying parameters in non-linear time series due to their responsive updating mechanism integrating the gradient of the log-likelihood function. Despite their effectiveness, these models can incorrectly update parameters in the presence of outliers, necessitating a robustification of the updating step. This paper proposes integrating the strictly proper PseudoSpherical scoring rule (PsSphS) into the SD framework to mitigate the impact of outliers. The PsSphS has been shown by Kanamori and Fujisawa (2015) to provide consistent estimates in the presence of anomalous data points, with Kleen (2023) referring to it as a robstified version of the logarithmic scoring rule found in existing SD models. We derive a Quasi-Score driven model, whereby the model maximises the PsSphS in each updating step, showing that the PsSphS introduces an exponential term that modulates the impact of large observations on parameter updates. The tuning parameter α arises as a key feature of the PsSphS, as it adjusts the model's sensitivity to outliers and can replicate the logarithmic scoring rule. This makes it a valuable tool for robust forecasting and modelling, setting the groundwork for more resilient score-driven methodologies.

Keywords: generalized autoregressive score (GAS) dynamic conditional score (DCS); Kullback Leibler; scoring rule, divergence

Statement of Originality

This document is written by Student Louis Gehringer 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. UvA Economics and Business is responsible solely for the supervision of completion of the work and submission, not for the contents.

1 Introduction

Score-driven (SD) models introduced by Creal et al. (2013) and Harvey (2013) are a popular observation-driven framework for modeling time varying parameters in non-linear time series. SD models leverage the gradient of the log-likelihood function to iteratively adjust parameters, updating parameters in the direction of steepest ascent to locally optimise the fit of the model. This yields a highly effective dynamic mechanism for capturing evolving data processes while maintaining strong computational efficiency. As a result, SD models are firmly established for modelling complex processes such as credit risk and volatility. Their flexibility and efficiency come at a cost however, as the framework is shown to incorrectly update parameters in the presence of outliers, prompting the need for a robustification of the updating step.

In this context, the application of scoring rules emerges as a pivotal concept in robustifying the SD updating mechanism. Scoring rules, as defined by Gneiting and Raftery (2007), serve as metrics for evaluating the quality of probabilistic forecasts by assigning numerical scores based on the predictive distribution and observed outcomes. Strictly proper scoring rules incentivize forecasters to provide accurate and well-calibrated predictions. Literature by Kanamori and Fujisawa (2015) and Kleen (2023) provide strong evidence that the strictly proper PseudoSpherical scoring rule (PsSphS) can deliver consistent estimates in the presence of anomalous data points. We integrate the PsSphS into the score-driven framework, demonstrating that its unique properties yield robustified parameter updating.

Building upon the foundational principles of score-driven models and scoring rules, this paper delves into the theoretical underpinnings of robustifying parameter updates using the PsSphS. We aim to enhance the model's adaptability and reliability in diverse data environments. Through a comprehensive analysis of the PsSphS's properties and its

implications for parameter estimation, this study seeks to contribute to the advancement of robust time-varying parameter modeling techniques in the realm of statistical forecasting of non-linear time series data. Section 2 explores the relevant theory surrounding the score-driven framework and strictly proper scoring rules. Section 3 defines an adaptation of the score-driven model using the PsuedoSpherical score for parameter updating. Section 4 showcases the model's performance in modelling the conditional mean and conditional variance through simulated data.

2 Theoretical Framework

The score-driven model known as the Generalized Autoregressive Score (GAS) model proposed by Creal et al. (2013) employs the score of the predictive model density to iteratively update the time-varying parameters. Suppose the observation $y_t \in \mathcal{Y}$ is drawn from the density $p(y_t|\vartheta_t, \mathcal{F}_{t-1}; \theta)$ where ϑ_t is the time varying parameter updated according to:

$$\vartheta_{t+1} = \omega + \beta\vartheta_t + \alpha S_t \nabla_t \tag{1}$$

where:

$$\nabla_t = \left[\frac{\partial \log p(y_t|\vartheta_t, \mathcal{F}_{t-1}; \theta)}{\partial \vartheta_t}\right] \tag{2}$$

corresponds to the score or gradient of the conditional log-likelihood w.r.t the time varying parameter. θ is an unknown vector containing the static parameters ω, β and α. S corresponds to a matrix function that scales the score and is typically defined as the inverse fisher information matrix (i.e $S = \mathcal{I}_t^{-1}$ where $\mathcal{I}_t = \mathbb{E}_t(\nabla_t \nabla_t')$.

The score is not to be confused with the important concept of strictly proper scoring rules discussed by Gneiting and Raftery (2007). Scoring rules assess the quality of probabilistic forecasts, assigning a numerical score based on the predictive distribution and

realized value. Scoring rules are taken to be positively oriented rewards that a forecaster seeks to maximize. Specifically, say a forecaster models an unknown distribution P according to distribution F. Let $\mathcal{P}$ denote a convex class of probability distributions. Consider a random variable $Y : \Omega \to \mathcal{Y}$, from a complete probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The scoring rule S assigns numerical values (scores) to observations $y \in \mathcal{Y}$ and distributions $F \in \mathcal{P}$ through a mapping $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R} \cup \{-\infty\} =: \bar{\mathbb{R}}$. Important is the notion of propriety and strict propriety. A scoring rule S is proper to a class of probability densities $\mathcal{P}$ if no predictive density has a higher expected score than the true unknown density p:

$$\mathbb{E}_p(S(f, Y)) \leq \mathbb{E}_p(S(p, Y)) \quad \forall f, p \in \mathcal{P} \tag{3}$$

Strict propriety extends this notion, where the inequality becomes an equality if and only if $f = p$. In other words, the expected score $\mathbb{E}_p(S(f, y))$ is maximized if $p = f$ while no other $f \in \mathcal{P}$ achieves the same expected score. Of interest is the relation with Bregman divergences, as it provides deeper insight into the properties of (strictly) proper scoring rules.

Both scoring rules and Bregman divergences are rooted in convex functions, specifically entropy functions. While scoring rules assess dissimilarity for probabilistic forecasts, Bregman divergences serve primarily in convex optimization as a generalization of the Euclidean distance. Jose et al. (2008) illustrates this relationship, remarking that if $S(\cdot)$ is a differentiable function, it satisfies:

$$\mathbb{E}_p(S(f, Y)) = S(f) + \nabla S(f) \cdot (p - f) \tag{4}$$

where $\nabla S(f)$ is the gradient of $S(\cdot)$ evaluated at f and conversely every function S that is (strictly) convex and differentiable uniquely defines a (strictly) proper scoring rule. Any

strictly convex function g defines a Bregman divergence $\mathbb{D}_g(p||f)$ as:

$$\mathbb{D}_g(p||f) = g(p) - g(f) - \Delta g(f) \cdot (p - f) \tag{5}$$

Letting $g(p) = S(p)$ it follows that for any strictly proper scoring rule the forecaster's expected loss for reporting f when the true distribution is p is a Bregman divergence, namely:

$$\mathbb{D}_S(p||f) = \mathbb{E}_p(S(p, Y) - S(f, Y)) = \mathbb{E}_p(S(p, Y)) - \mathbb{E}_p(S(f, Y)) \tag{6}$$

Which is referred to as a score divergence, as it quantifies the discrepancy between two probability distributions through a strictly proper scoring rule. A score divergence is a measure of the "information deficit" that is incurred by modelling the true distribution p using f. Bregman divergences, along with their associated strictly proper scoring rules, offer a wide variety of useful loss functions in this context. We can write the condition of propriety and strict propriety of a scoring rule in terms of its Bregman divergence Gneiting and Raftery (2007). A scoring rule $S : \mathcal{P} \times \mathcal{Y} \to \mathbb{R}$ is proper relative to $\mathcal{P}$ if $\mathbb{D}_S(p||f) \geq 0, \forall p, f \in \mathcal{P}$ and strictly proper if, additionally $\mathbb{D}_S(p||f) = 0$ if and only if $p = f, \forall p, f \in \mathcal{P}$ de Punder et al. (2023). This harmonization of scoring rules and Bregman divergences is also shown by Ovcharov (2018), characterizing strict propriety under the guise of a Bregman divergence. This demonstrates that score maximization is equivalent to the minimization of the corresponding Bregman divergence.

To further illustrate the relationship between scoring rules and Bregman divergences, consider the logarithmic scoring rule $\log \mathrm{S}(f, y) := \log f(y)$. The logarithmic scoring rule manifests itself in the GAS model through updating by the gradient of the conditional log-likelihood. In doing so, the parameter is updated in the direction of steepest ascent of the log-likelihood, corresponding to a maximization of the logarithmic score. The related

score divergence is given by (6):

$$\begin{aligned}
\mathbb{D}_{\log S}(p||f) &= \mathbb{E}_p(S(p,Y)) - \mathbb{E}_p(S(f,Y)) \\
&= \mathbb{E}_p(\log(p(y)) - \mathbb{E}_p(\log(f(y))) \\
&= \mathbb{E}_p\left(\log\frac{p(y)}{f(y)}\right) \\
&= \int_{-\infty}^{\infty} p(y)\cdot\log\frac{p(y)}{f(y)}\,dy \\
&= \mathrm{KLD}(p||f)
\end{aligned}$$

This is the familiar Kullback-Leibler divergence (KLD) that measures the dissimilarity between two probability distributions. By the strict propriety of the logarithmic scoring rule, $\mathrm{KLD}(p||f) \geq 0$ with equality if and only if $f = p$. In essence, the KLD is a measure of the additional entropy introduced by approximating the true distribution p by f. If f perfectly matches p, the KLD is zero and no excess entropy is produced. The KLD is an important measure for assessing the optimality of a GAS model.

Blasques et al. (2015) intend to show that GAS updates are information theoretically optimal by minimizing the local KLD relative to the true density. The authors reason that the sign of the score in the direction that reduces the local KLD most, such that the fit of the model is improved locally. The paper is however lacking in several respects, most notably that the local KLD measure can take on extremely large negative values, which renders it an invalid divergence measure. In addition to this, the presence of outliers can severely deteriorate the fit of the model when measured against new observations. Follow on research by Gorgi et al. (2023) acknowledges these short-comings, seeking to define an optimality condition using a global perspective of the KLD. The proposed conditional expected variation optimality deems a parameter update as optimal in expected variation if the distance between the expected updated parameter and (pseudo-)true parameter dimin-

ishes with each iteration. The idea is that the parameter update is based on the observable point y_t to update the parameter such that on average the updated parameter gets closer to the (pseudo)-true parameter. This optimality definition not only sheds light on reducing mean squared error concerning the pseudo-true parameter but also aids in selecting scaling factors for the score within the updating equation, rendering it more practical than the local-fit improvement focus of Blasques et al. (2015).

Blasques et al. (2023) introduce the Quasi score-driven (QSD) model, which disconnects the updating equation from the log-likelihood implied by the conditional density of the observations. The updating equation is now defined as $\vartheta_{t+1} = \omega + \beta\vartheta_t + \alpha \cdot \psi(y_t, X_t, \vartheta_t; \theta)$ where X_t is a vector of exogenous random variables and ψ is a differentiable measurable function defined as $\psi(y_t, \vartheta_t; \theta) = \frac{\partial \rho(y_t, X_t, \vartheta_t; \theta)}{\partial \vartheta_t} \cdot S_t$, with S_t a scaling function. $\rho(y_t, X_t, \vartheta_t) = \log p(y_t|\vartheta_t, \mathcal{F}_{t-1}; \theta)$ recovers the original GAS model by Creal et al. (2013). By relaxing the strict link between the updating equation and the conditional density, QSD models allow for freedom to design parameter updating equations guided by various statistical loss functions beyond the log-likelihood function.

Promising candidates beyond the log-likelihood are strictly proper scoring rules due to their ability to elicit accurate and well calibrated probabilistic forecasts. Of particular interest is the PseudoSpherical score (PsSphS) belonging to the semi-local PseudoSpherical family. The PsSphS is defined as a local scoring rule by Gneiting and Raftery (2007), but will depend on the density via the global norm of the density, henceforth referred to as semi-local as by de Punder et al. (2023). The use of the PsSphS is motivated by several reasons. Firstly, as a strictly proper scoring rule it defines a unique score divergence. More so, the popular logarithmic scoring rule is a limiting case of the PsSphS when α approaches 1 from above (i.e $\log \mathrm{S}(f, y) = \lim_{\alpha \downarrow 1} \mathrm{PsSphS}_\alpha(f, y)$). This motivates its exploration and hints at

its potential. Kanamori and Fujisawa (2015) found that the PseudoSpherical scoring rule produces an approximately consistent estimator of the target density, even under heavy contamination. This demonstrates that the PsSphS can provide accurate density forecasts under the influence of outliers. Substantial evidence in favor of the PsSphS and particularly the SphS (a special case of the PsSphS for $\alpha = 2$) is found by Kleen (2023), who compares the strictly proper scoring rules for density forcasting. Kleen denotes that the PsSphS has a high degree of sharpness independent of the outcome, stemming from the L^{α}-norm. A sharp forecast distribution indicates a high level of confidence or precision in the prediction. Kleen finds that the SphS is the least affected scoring rule by possibly large but infrequent measurement error when compared to other common strictly proper scoring rules, especially the logarithmic scoring rule. Furthermore, he remarks that the SphS is of interest due to its relationship to the logS, explicitly referring to the SphS as 'robustified' version of the logarithmic scoring rule. Noteworthy is that Kleen's simulations results show that if one intends to employ a scoring rule that is based on the likelihood, the SphS is preferred to the logS. This is of course the case for the score driven models which we intend to (in part) estimate using maximum likelihood estimation in this study. This provides strong theoretical evidence supporting the use of the PseudoSpherical scoring rule in robustifying the updating step of score-driven models.

3 Methodology

3.1 Quasi-Score Driven Model

We define a Quasi score-driven model, whereby the gradient of the log-likelihood is replaced by that of the PseudoSpherical score:

$$y_t \sim p(\cdot|\vartheta_t, \mathcal{F}_{t-1}; \theta) \qquad \vartheta_{t+1} = \omega + \beta\vartheta_t + \gamma \cdot \psi_S(y_t, \vartheta_t; \theta) \tag{7}$$

where:

$$\psi_S(y_t, \vartheta_t; \theta) = \frac{\partial S(y_t, \vartheta_t; \theta)}{\partial \vartheta_t}\Big|_{\vartheta=\vartheta_t} \cdot \rho_t \qquad S(y_t, \vartheta_t; \theta) = \text{PsSphS}_\alpha(p, y) \tag{8}$$

where ρ_t denotes the scaling function with $\rho_t = \mathcal{I}_t^{-1}$ and $\mathcal{I}_t = \mathbb{E}_t(\nabla_t^2)$. This updates the parameter in the direction of steepest ascent, maximizing the PsSphS and minimizing the associated score divergence. The PseudoSpherical scoring rule is defined as:

$$\text{PsSphS}_\alpha(f, y) = \frac{f(y)^{\alpha-1}}{||f||_\alpha^{\alpha-1}}, \quad \alpha > 1 \tag{9}$$

where $f(y)$ is the predicted probability density at the observed outcome y and α is a parameter with $\alpha > 1$. $||f||_\alpha$ gives the L^α-norm i.e $||f||_\alpha = \left(\int f^\alpha \, dy\right)^{\frac{1}{\alpha}}$. The associated score divergence is given by:

$$\mathbb{D}_{\text{PsSphS}}(p||f) = ||f||_\alpha - \frac{\int (f(y) \cdot g(y))^{\alpha-1} dy}{||g||_\alpha^{\alpha-1}} \tag{10}$$

We distinguish the special case where $\alpha = 2$, which yields the Spherical score (SphS):

$$S(y_t, \vartheta_t; \theta) = \text{PsSphS}_2(p, y) = \text{SphS}(p, y) = \frac{f(y)}{||f||_2} \tag{11}$$

with divergence measure:

$$\mathbb{D}_{\text{SphS}}(p||f) = ||f||_2(1 - C(p, f)) \tag{12}$$

where $C(p,f) = \frac{\int p(y)\cdot f(y)\ dy}{\sqrt{\int f(y)^2\ dy\ \int g(y)^2\ dy}}$ is the cosine similarity between p and f. For the limiting case $\alpha \downarrow 1$ we have:

$$S(y_t, \vartheta_t; \theta) = \text{PsSphS}_{\alpha\downarrow 1}(p, y) = \text{logS}(p, y) \tag{13}$$

which motivates the comparison between the logarithmic scoring rule as a benchmark against the the PsSphS for different values of α.

3.2 Leading Examples

Consider $y_t \sim \mathcal{N}(\mu_t, 1)$ with the updating equations using logS, $\text{PsSphS}_{1.01}$ $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.5}$ and SphS:

$$y_t \sim \mathcal{N}(\mu_t, 1) \qquad \mu_{t+1} = \omega + \beta \cdot \mu_t + \gamma \cdot \psi_{\text{PsSphS}_\alpha}(y_t, \mu_t; \theta) \qquad \rho_t = 1$$

where:

$$\begin{aligned}
\psi_{\text{PsSphS}_\alpha}(y_t, \mu_t; \theta) &= \frac{\partial}{\partial \mu_t} \frac{f(y_t)^{\alpha-1}}{||f||_\alpha^{\alpha-1}} \\
&= \frac{1}{||f||_\alpha^{\alpha-1}} \frac{\partial}{\partial \mu_t} f(y_t)^{\alpha-1} \\
&= \left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha-1)(y_t - \mu_t) e^{-\frac{(\alpha-1)(y_t-\mu_t)^2}{2}}
\end{aligned}$$

with the special case for $\alpha = 2$:

$$\psi_{\text{SphS}}(y_t, \mu_t; \theta) \equiv \psi_{\text{PsSphS}_2}(y_t, \mu_t; \theta)$$

we find the L^α-norm to be independent of the parameter μ.

$$\begin{aligned}
\psi_{\text{SphS}}(y_t, \mu_t; \theta) &= \frac{1}{||f||_2} \frac{\partial}{\partial \mu_t} f(y_t) \\
&= \left(\frac{1}{\pi}\right)^{\frac{1}{4}} (y_t - \mu_t) e^{-\frac{(y_t-\mu_t)^2}{2}}
\end{aligned}$$

This allows us to observe that for the special case of the SphS, the re-parametrization $\tilde{\gamma} = \gamma \cdot ||f||_2$ yields the typically improper linear scoring rule as $||f||_2$ is constant w.r.t μ_t Gneiting and Raftery (2007). However, in this case it can be shown that due to the independence of μ_t and $||f||_2$, the linear scoring rule is in fact proper due to the fact that $||f||_2 = ||p||_2$, where p is the true density.

For the special case of the logarithmic score:

$$\begin{aligned}
\psi_{\text{logS}}(y_t, X_t, \mu_t; \theta) &\equiv \lim_{\alpha \downarrow 1} \psi_{\text{PsSphS}_\alpha}(y_t, \mu_t; \theta) \\
\psi_{\text{logS}}(y_t, X_t, \mu_t; \theta) &= \frac{\partial \log f(y_t)}{\partial \mu_t} \\
&= \frac{1}{f(y_t)} \frac{\partial f(y_t)}{\partial \mu_t} \\
&= y_t - \mu_t
\end{aligned}$$

The key result is the exponential term in the PsSphS that introduces a weighting that decays exponentially with the square of the gradient of the logarithmic score. This provides strong theoretical evidence for superior robustness to outliers, as large deviations in the data are heavily compensated by the exponential term.

We now consider the case, $y_t \sim \mathcal{N}(0, \sigma_t^2)$ with the updating equations using logS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.5}$ and SphS:

$$y_t \sim \mathcal{N}(0, \sigma_t^2) \qquad \sigma_{t+1}^2 = \omega + \beta \cdot \sigma_t^2 + \gamma \cdot \psi_{\text{PsSphS}_\alpha}(y_t, \sigma_t^2; \theta) \qquad \rho_t = 2\sigma_t^4$$

where:

$$\begin{aligned}
\psi_{\text{PsSphS}_\alpha}(y_t, \sigma_t^2; \theta) &= \frac{\partial}{\partial \sigma_t^2} \frac{f(y_t)^{\alpha-1}}{||f||_\alpha^{\alpha-1}} \cdot \rho_t \\
&= \left(\frac{\alpha}{2\pi\sigma_t^2}\right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha - 1) \frac{y_t^2}{2\sigma_t^4} e^{-\frac{(\alpha-1)y_t^2}{2\sigma_t^2}} \cdot 2\sigma_t^4 \\
&= \left(\frac{\alpha}{2\pi\sigma_t^2}\right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha - 1) y_t^2 e^{-\frac{(\alpha-1)y_t^2}{2\sigma_t^2}}
\end{aligned}$$

With the following interesting special cases:

$$\psi_{\text{SphS}}(y_t, \sigma_t^2; \theta) \equiv \psi_{\text{PsSphS}_2}(y_t, \sigma_t^2; \theta) = \left(\frac{1}{\pi\sigma_t^2}\right)^{\frac{1}{4}} y_t^2 \mathrm{e}^{-\frac{y_t^2}{2\sigma_t^2}}$$

$$\psi_{\text{logS}}(y_t, \sigma_t^2; \theta) \equiv \lim_{\alpha \downarrow 1} \psi_{\text{PsSphS}_\alpha}(y_t, \sigma_t^2; \theta) = y_t^2 - \sigma_t^2$$

Similarly as for μ_t we observe an exponential term that mitigates the effect of large data points in the data.

Our model departs the space of GAS updates, by explicitly implementing a strictly proper scoring rule in its updating equation. This makes it related to the class of Proper and Robust Autoregressive Derivative-Adaptive (PRADA) models.

3.3 Estimation by Maximum Likelihood

3.3.1 Conditional Mean

We simulate 3000 data points of a stationary AR(1) process with normally distributed error terms:

$$y_t = 0.9 y_{t-1} + \varepsilon_t \qquad \varepsilon_t \sim \mathcal{N}(0, 1)$$

Models are fit by Maximum Likelihood Estimation (MLE) using a rolling window with a window size of 500. For the first leading example $y_t \sim \mathcal{N}(\mu_t, 1)$, the log-likelihood function for the time series is given by:

$$\mathcal{L}(\omega, \beta, \gamma; \{y_t\}) = \sum_{t=1}^{n} \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}(y_t - \mu_t)^2 \right]$$

where the mean μ_t follows the updating equations described above. For each rolling window of size 500, the parameters ω, β, and γ are estimated by maximizing the log-likelihood function over that window. Specifically, for each window indexed by k (where k ranges

from 1 to $n - 499$), we solve the optimization problem:

$$(\hat{\omega}_k, \hat{\beta}_k, \hat{\gamma}_k) = \arg\max_{\omega,\beta,\gamma} \mathcal{L}(\omega, \beta, \gamma; \{y_t\}_{t=k}^{k+499})$$

where $\{y_t\}_{t=k}^{k+499}$ denotes the subset of the time series data corresponding to the k-th rolling window. The problem is considered a second time using t-distributed errors with 5 degrees of freedom (i.e $\varepsilon_t \sim t(5)$).

3.3.2 Conditional Variance

We simulate 2000 data points of a GARCH(1,1) process with normally distributed error terms:

$$y_t = \epsilon_t, \qquad \epsilon_t = \sigma_t z_t, \qquad \sigma_t^2 = 0.1 + 0.1\epsilon_{t-1}^2 + 0.7\sigma_{t-1}^2$$

with $z_t \sim \mathcal{N}(0, 1)$. Models are fit by MLE with rolling window and a window size of 1000. For $y_t \sim \mathcal{N}(0, \sigma_t^2)$, the log-likelihood is given by:

$$\mathcal{L}(\omega, \beta, \gamma; \{y_t\}) = \sum_{t=1}^{n} \left[-\frac{1}{2} \log(2\pi\sigma_t^2) - \frac{y_t^2}{2\sigma_t^2} \right]$$

where the variance σ_t^2 follows the updating equations described above. Similarly as above, for each window indexed by k (where k ranges from 1 to $n-999$), we solve the optimization problem:

$$(\hat{\omega}_k, \hat{\beta}_k, \hat{\gamma}_k) = \arg\max_{\omega,\beta,\gamma} \mathcal{L}(\omega, \beta, \gamma; \{y_t\}_{t=k}^{k+999})$$

The problem is considered once more with t-distributed errors with 5 degrees of freedom (i.e $z_t \sim t(5)$).

3.4 Estimation by the Same Scoring Rule

Adhering to Gneiting and Raftery (2007), we now use the same strictly proper scoring rule to estimate the parameters of the updating equation. We no longer derive the optimal parameters by maximizing the log-likelihood. Instead, the unknown parameter vector $\theta = (\omega, \beta, \gamma)$ is estimated by maximizing the mean score:

$$S_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} S(f, Y_i)$$

where f is the estimated density and Y_i the sample of identically distributed observations used to fit the model. If θ_0 denotes the true parameter vector, then asymptotically it holds that:

$$\text{argmax}_\theta \; S_n(\theta) \to \theta_0 \qquad \text{as } n \to \infty$$

such that $\hat{\theta}_n = \text{argmax}_\theta \; S_n(\theta)$ denotes the optimum score estimator.

Estimating by the logarithmic scoring rule is equivalent to maximizing the mean log-likelihood under maximum likelihood estimation.

Under $y_t \sim \mathcal{N}(\mu_t, 1)$, the optimal parameter vector θ optimizes the mean PsSphS for a given α:

$$\text{PsSphS}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\alpha}{2\pi} \right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)(y_i-\mu_t)^2}{2}}$$

Under $y_t \sim \mathcal{N}(0, \sigma_t^2)$ the optimal parameter vector θ optimizes the mean PsSphS for a given α:

$$\text{PsSphS}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{\alpha}{2\pi\sigma_t^2} \right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)y_i^2}{2\sigma_t^2}}$$

As for the MLE problem, the data generating process is an AR(1) and GARCH(1,1) process, with estimation using rolling window with window sizes 500 and 1000 respectively.

3.5 Evaluating Density Forecasts

We formalize the comparison between the models by first calculating the score difference between the forecasts. This involves selecting a strictly proper scoring rule, determining the score for each prediction from each respective model and taking the difference between the scores of the model predictions. Namely let $\left\{\hat{f}_t(y)\right\}$ and $\{\hat{g}_t(y)\}$ be the sequences of density forecasts we intend to compare. We select the logarithmic scoring rule to evaluate the density forecasts and then take the difference between the scores. This gives us the sequence of differences:

$$\{d_t\} = \left\{\log \mathrm{S}(\hat{f}_t(y)) - \log \mathrm{S}(\hat{g}_t(y))\right\}$$

The forecasts $\left\{\hat{f}_t(y)\right\}$ are those by the models using the PsSphS_α with $\alpha > 1$. We compare this to the baseline model with the logarithmic score (ψ_{logS}), denoted by $\{\hat{g}_t(y)\}$.

To test if the difference between the log-scores of each density forecasts are statistically significant, we apply the Diebold-Mariano (DM) test with the following hypothesis:

$$H_0 : \mathbb{E}(d_{t+1}) = 0 \qquad H_1 : \mathbb{E}(d_{t+1}) > 0$$

and corresponding test statistic:

$$t_n = \frac{\bar{d}_n}{\sqrt{\hat{\sigma}_n^2/n}}$$

where $\bar{d}_n = \frac{1}{n}\sum_{t=1}^{n} d_t$ is the average score difference between $\left\{\hat{f}_t(y)\right\}$ and $\{\hat{g}_t(y)\}$ and $\hat{\sigma}_n^2$ is a heteroskedasticity and autocorrelation consistent estimator of $\sigma_n^2 = \mathbb{V}(\sqrt{n}\bar{d}_n)$ as used in Diks et al. (2011). Furthermore, we assume it holds that $t_n \xrightarrow{d} \mathcal{N}(0,1)$. We observe that

if S is strictly proper, the following holds:

$$\begin{aligned} H_0 : \mathbb{E}(d_{t+1}) = 0 & \\ \Leftrightarrow \quad & \mathbb{E}(S(\hat{f}_t, y_t) - S(\hat{g}_t, y_t)) = 0 \\ \Leftrightarrow \quad & \mathbb{E}(S(\hat{f}_t, y_t) - S(p_t, y_t)) = \mathbb{E}(S(\hat{g}_t, y_t) - S(p_t, y_t)) \\ \Leftrightarrow \quad & -\mathrm{KLD}(\hat{f}_t) = -\mathrm{KLD}(\hat{g}_t) \end{aligned}$$

This stipulates that if S is strictly proper, the null hypothesis can be interpreted as $\hat{f}$ being equally close to the unknown p_t as $\hat{g}$. This closeness is of course measured by the Bregman Divergence corresponding to the scoring rule S. In the case of the logarithmic scoring rule, this divergence is the Kullback-Leibler Divergence. Such that the hypotheses can be interpreted as:

$$H_0 : \mathrm{KLD}(\hat{f}_t) = \mathrm{KLD}(\hat{g}_t) \qquad H_1 : \mathrm{KLD}(\hat{f}_t) < \mathrm{KLD}(\hat{g}_t)$$

Allowing us to formulate the statistical test on the quality of fit of the density forecasts directly in terms of a Bregman Divergence. Note, the sign of our alternative hypothesis flips as it is written in terms of minus the KLD, such that $\mathbb{E}(d_{t+1}) > 0 \Rightarrow \mathrm{KLD}(\hat{f}_t) < \mathrm{KLD}(\hat{g}_t)$, however we still reject for $t_n > 1.645$ (for significance $\alpha = 0.05$). Using the logarithmic scoring rule inevitably introduces a bias, since the updating equation seeks to update parameters that maximize the log-score and thus minimize the KLD, directly optimizing for this statistical test. However, given the broader role of the KLD in assessing GAS optimality, selecting the logarithmic score to assess the quality of the density forecasts is warranted despite the bias it may introduce. We would expect that if the PseudoSpherical score yields significant improvements in fit due to robustification of the updating step, that it too optimizes the KLD overall.

4 Simulation

4.1 Estimation by Maximum Likelihood

4.1.1 Conditional mean

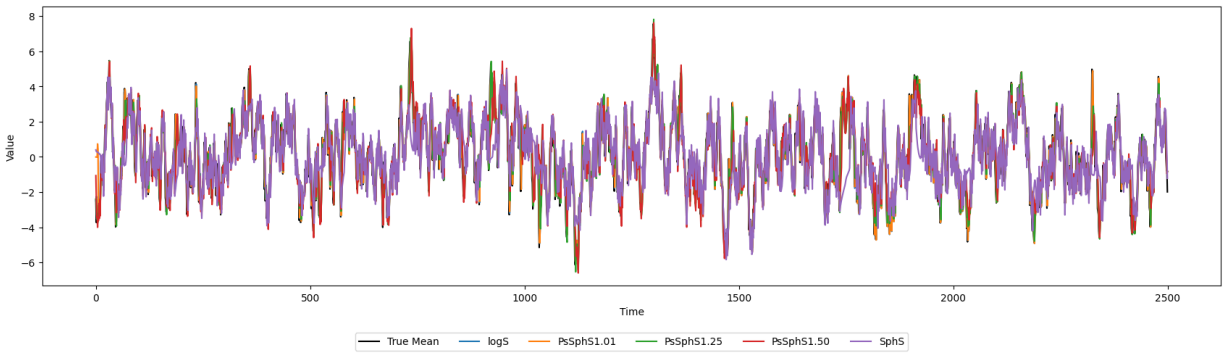

Figure 1: True conditional mean plotted against predictions by models using LogS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

We observe that the predictions made by the LogS and $\text{PsSphS}_{1.01}$ models are identical, confirming that the logarithmic scoring rule can be retrieved as a limiting case of the PsSphS in a practial setting. As the value of α increases, the predictions become increasingly conservative for spikes in the data, as would be implied by the properties of the PsSphS.

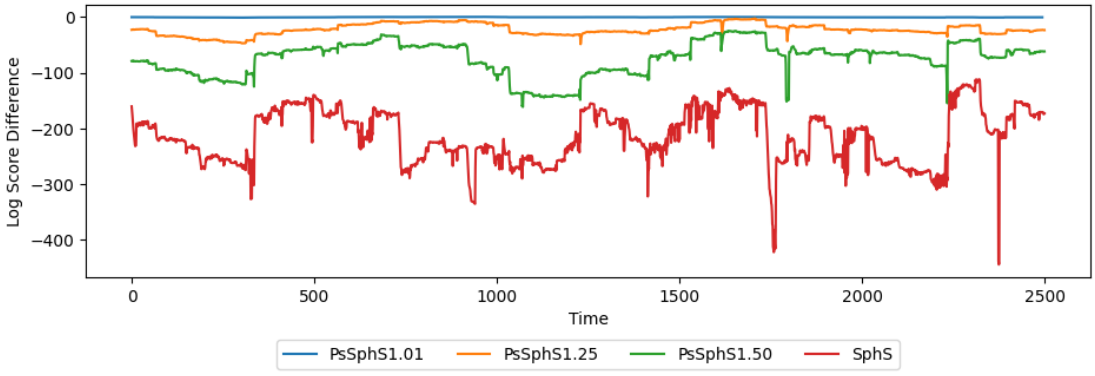

Figure 2: Log-Score difference between LogS and $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}
Test Statistic	-0.97	-2.15	-2.48	-4.43
p-value	0.83	0.98	0.99	0.99

Table 1: Diebold-Mariano Test using logS for $n = 2500$.

The log-score difference between the LogS and PsSphS models increases as α increases. The score difference is deemed statistically insignificant, but given the size of the test statistics here, the result favors the LogS model overall.

We calculate the mean squared errors (MSE) for the models and apply a grid search for values $1 < \alpha \leq 2$ to find the optimal value of α. The optimal value of α is found to be $\alpha = 1.01$. Indicating that the model with the logarithmic scoring rule is favorable in this scenario based on the MSE.

ψ_{LogS}	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}	$\psi_{\text{PsSphS}_{\text{Optimal}}}$
0.838	0.870	0.918	1.136	1.989	0.870

Table 2: Mean Squared Error of LogS, PsSphS_{α} for $\alpha > 1$ and optimal value of α.

The logarithmic scoring rule has the best overall fit, even marginally over $\text{PsSphS}_{1.01}$. In general, the MSE increases as the tuning parameter α increases, indicating that the PsSphS overcompensates deviations from the mean. The result is a poorer fit overall. Despite this, the fact that the performance of the logarithmic scoring rule can be replicated by the PsSphS with $\alpha = 1.01$ is favorable for the PsSphS, since the tuning parameter α gives an additional degree of freedom.

4.1.2 Conditional variance

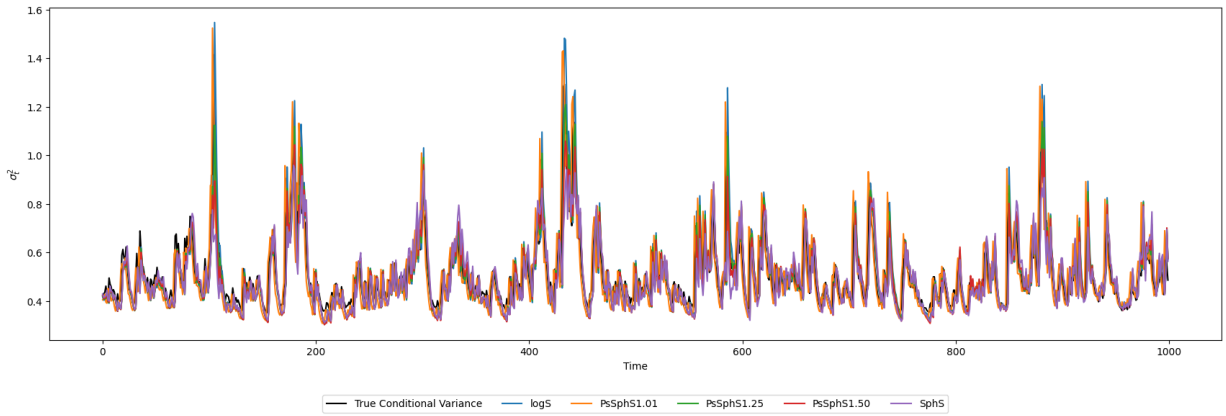

Figure 3: True conditional variance plotted against predictions by models using LogS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

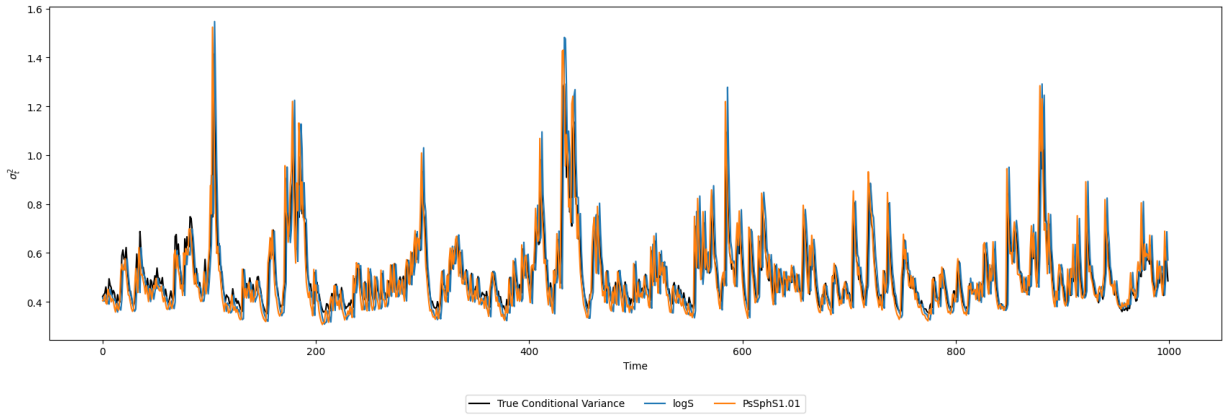

Figure 4: Predictions of LogS and $\text{PsSphS}_{1.01}$.

For the conditional variance, the PseudoSpherical scoring rule exhibits clear robustness to outliers relative to the logarithmic scoring rule. While the Logarithmic scoring rule consistently overshoots its predictions for extreme observations, the PsSphS across all values of α effectively smoothens the prediction. Noteworhty is that even the $\text{PsSphS}_{1.01}$ model does typically not overshoot as much as the LogS model. This is a strong result in favor of the PseudoSpherical scoring rule. The impact of the tuning parameter α can clearly be

observed as well, with the model becoming increasingly insensitive to changes in the data for larger values of α. In that regard the SphS does not appear to be an attractive option for modelling the conditional variance, given that it is too conservative in its predictions for more extreme observations.

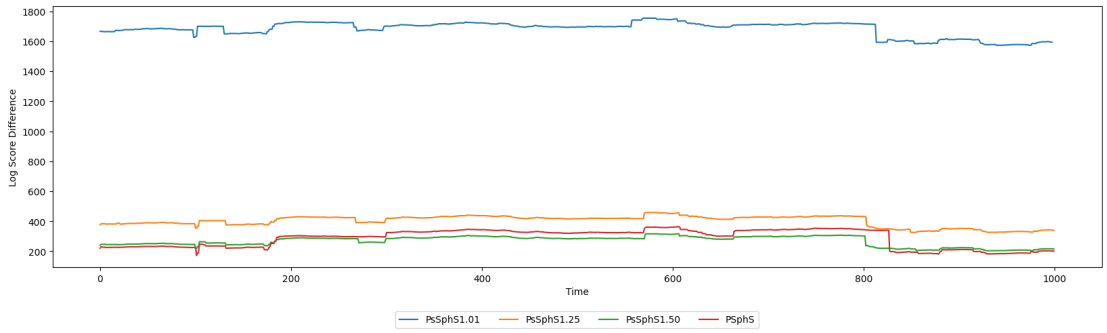

Figure 5: Log-Score difference between LogS and $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

We derive a very strong result here, namely that the PsSphS achieves a higher logarithmic score, than the logarithmic score itself. This is profound, given that the model using the logarithmic score naturally updates parameters to maximize the logarithmic scoring rule. The difference is the greatest for $\text{PsSphS}_{1.01}$, which has a substantially higher score than the LogS. The difference diminishes greatly for larger values of α, which is to be expected given that the PsSphS converges to the LogS for $\alpha \to 1^+$.

	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}
Test Statistic	35.05	11.18	8.08	4.86
p-Value	0.00	0.00	0.0	0.0

Table 3: Diebold-Mariano Test using logS for $n = 1000$.

Looking at the Diebold-Mariano test confirms that the aforementioned log-score differences are statistically significant for all variations of the PsSphS.

We calculate the mean squared errors (MSE) for the models and apply a grid search for values $1 < \alpha \leq 2$ to find the optimal value of α. The optimal value of α is found to be $\alpha = 1.55$.

ψ_{LogS}	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}	$\psi_{\text{PsSphS}_{\text{Optimal}}}$
0.0092	0.0092	0.0081	0.0081	0.0095	0.0074

Table 4: Mean Squared Error of LogS, PsSphS_α for $\alpha > 1$ and the model with the optimal value of α.

This demonstrates that the PsSphS is a superior choice to the logarithmic score for modelling the conditional variance, as all the benchmark models as well as the optimal model have smaller MSEs than the LogS model. Notable is that the MSE of the LogS and $\text{PsSphS}_{1.01}$ are identical despite Figure 4 showing that the LogS model overshoots predictions for extreme observations. It appears that the LogS model has a marginally better fit for regular observations, which brings the MSE closer to that of the robustified $\text{PsSphS}_{1.01}$. Nontheless, the plot clearly shows the preferable fit of the PsSphS models over the LogS model for anomalous data points.

4.1.3 Conditional mean under t-distribution

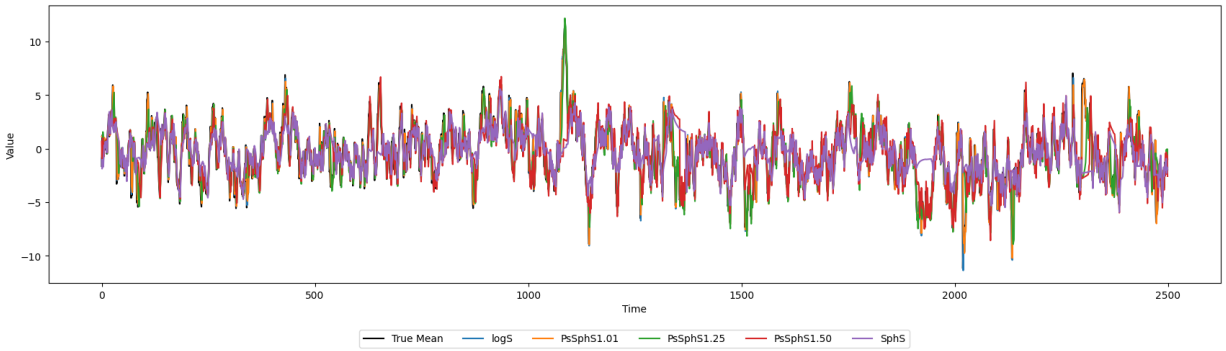

Figure 6: True conditional mean plotted against predictions by models using LogS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

By introducing t-distributed errors, we observe that only models with the PsSphS for larger values of α are unaffected by large spikes in the data.

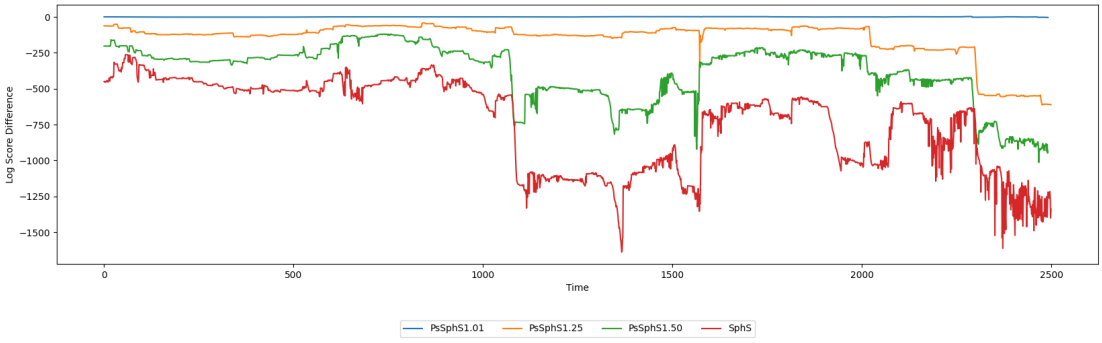

Figure 7: Log-Score difference between LogS and $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

Despite this, the log-score differences show the LogS model to have a higher log score compared to the models using the PsSphS, except for $\text{PsSphS}_{1.01}$ once more.

	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}
Test Statistic	-0.42	-1.15	-1.92	-2.39
p-value	0.66	0.88	0.97	0.99

Table 5: Diebold-Mariano Test using logS for $n = 2500$.

The log-score is statistically insignificant, but again the value of the test statistics indicate that the LogS model is favorable.

ψ_{LogS}	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}	$\psi_{\text{PsSphS}_{\text{Optimal}}}$
1.401	1.406	2.114	3.607	4.761	1.406

Table 6: Mean Squared Error of LogS, PsSphS_α for $\alpha > 1$ and optimal value of α.

We observe again that the mean squared error of the LogS and $\psi_{\text{PsSphS}_{1.01}}$ are identical. This reinforces that the PsSphS can effectively replicate the LogS model in a practical setting.

4.1.4 Conditional variance under t-distribution

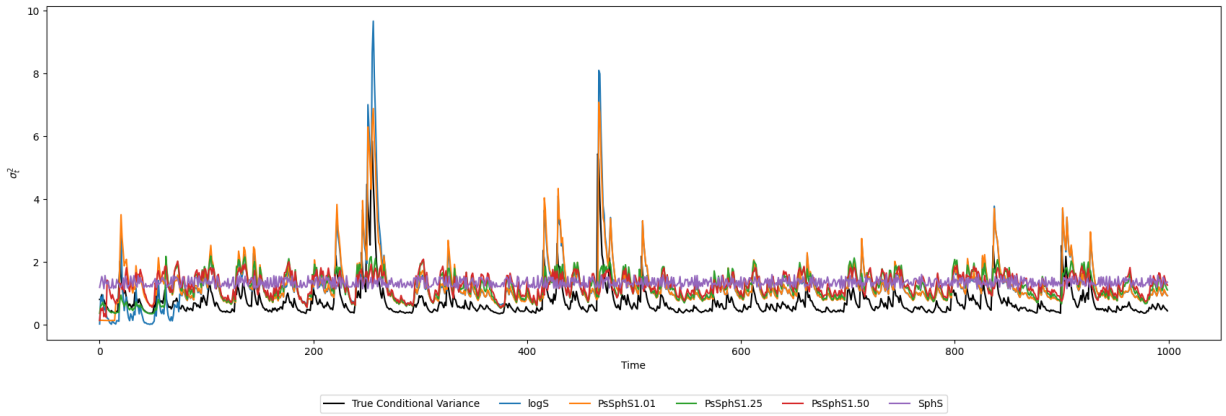

Figure 8: True conditional variance plotted against predictions by models using LogS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

We would expect the PsSphS for larger values of α to perform better here given that the outliers are more extreme, such that heavier smoothing would result in a better fit. However, the figure shows that instead the LogS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$ have the best overall fits. Notably, $\text{PsSphS}_{1.01}$ again displays improved robustness for the extreme observations compared to the LogS which significantly overshoots its predictions. Although $\text{PsSphS}_{1.25}$ generally has a good fit, it is overly conservative for larger spikes in the observations.

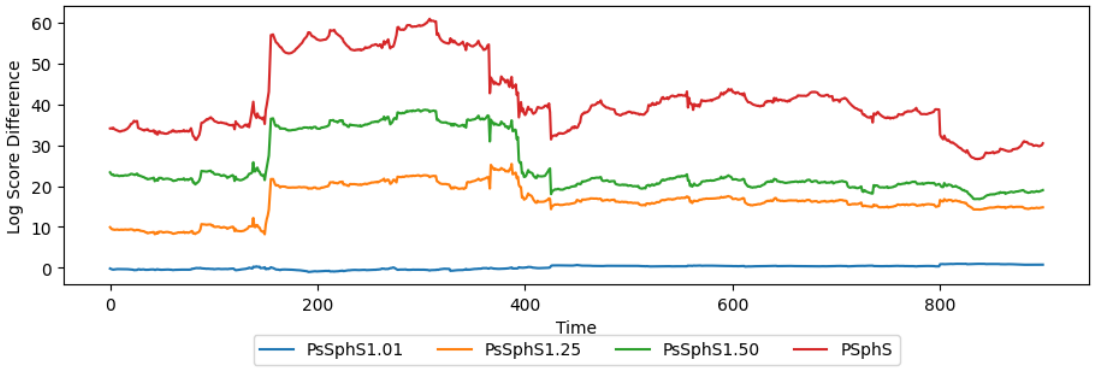

Figure 9: Log-Score difference between LogS and $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

We again find a positive log-score difference, indicating the models using the PsSphS have a higher logarithmic score than the model using the LogS model itself.

	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}
Test Statistic	-0.18	-2.40	-0.22	-0.19
p-value	0.57	0.60	0.59	0.57

Table 7: Diebold-Mariano Test using logS for $n = 1000$.

The difference is less pronounced as previously and consequently statistically insignificant. This is unlike what we would expect, given that the t-distributed errors yield more extreme outliers.

ψ_{LogS}	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}	$\psi_{\text{PsSphS}_{\text{Optimal}}}$
0.565	0.534	0.513	0.582	0.671	0.8070

Table 8: Mean Squared Error of LogS, PsSphS_α for $\alpha > 1$ and optimal value of α.

The optimal model is for $\alpha = 1.03$, although the actual MSE value is 0.8070, which is higher than the above models. This is however attributed to the high sensitivity of the optimization problem to the initial parameters. This cannot be accounted for during the grid-search for the optimal alpha value, such that the MSE value itself is only useful for relative comparisons. For this, we find $\alpha = 1.03$ to yield the smallest MSE overall. Aside from this, we observe that the model using the logS is outperformed by both the $\text{PsSphS}_{1.01}$ and $\text{PsSphS}_{1.25}$ in terms of MSE. $\text{PsSphS}_{1.01}$ appears to have the most favorable performance as it accurately predicts the spikes in the conditional variance, while the logS overshoots and the $\text{PsSphS}_{1.25}$ significantly under-predicts.

4.2 Estimation by the same scoring rule

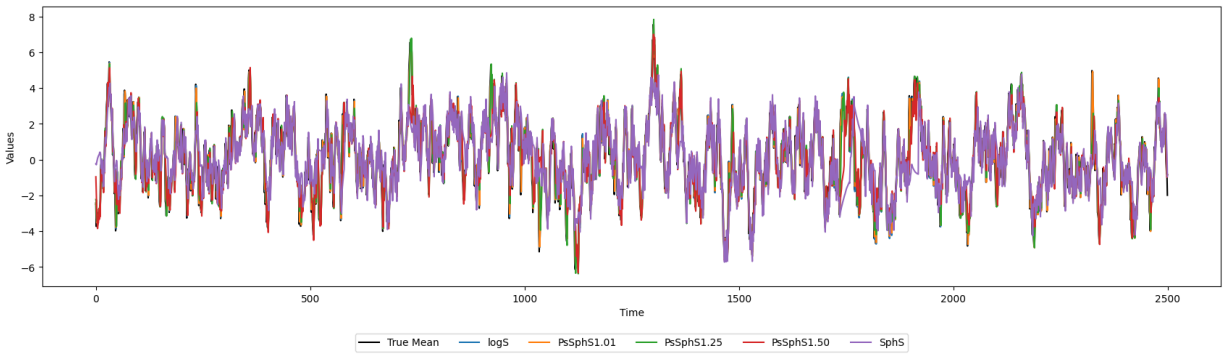

Figure 10: True conditional mean plotted against predictions by models using LogS, $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

The result for the logarithmic scoring rule is identical to that of maximum likelihood estimation. This result is intuitive, given that maximum likelihood optimizes the logarithm of the pdf, which is equivalent to maximizing the logarithmic scoring rule explicitly. A minor difference only arises from the fact that the mean of the score is optimized, which is not the case for MLE. For the PsSphS the difference is more pronounced with a heavier smoothing effect compared to MLE. This is also as expected, as direct maximization of the PsSphS is achieved through heavy penalization of deviations from the mean.

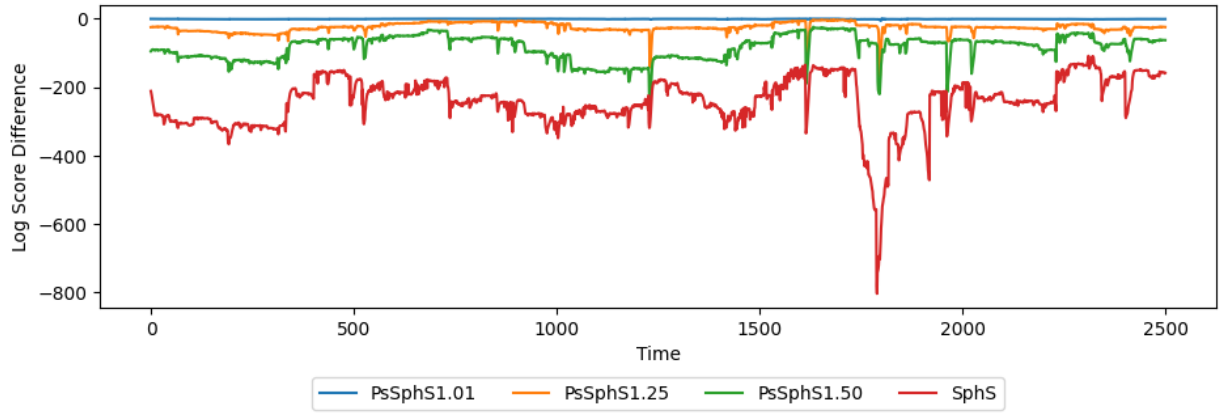

Figure 11: Log-Score difference for LogS and $\text{PsSphS}_{1.01}$, $\text{PsSphS}_{1.25}$, $\text{PsSphS}_{1.50}$ and SphS.

	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}
Test Statistic	-1.19	-1.78	-2.34	-3.23
p-value	0.88	0.96	0.99	0.99

Table 9: Diebold-Mariano Test using logS for $n = 2500$.

For the log-score difference we observe identical results to the maximum likelihood case as expected. The log-score difference is generally more consistent than for the MLE case, where it changed more frequently for the observations. The differences are statistically insignificant nontheless.

Evaluating the MSE scores across all tested values of α we find the optimal value of α to be $\alpha = 1.01$.

ψ_{LogS}	$\psi_{\text{PsSphS}_{1.01}}$	$\psi_{\text{PsSphS}_{1.25}}$	$\psi_{\text{PsSphS}_{1.50}}$	ψ_{SphS}	$\psi_{\text{PsSphS}_{\text{Optimal}}}$
0.838	0.838	0.920	1.16	2.019	0.838

Table 10: Mean Squared Error of LogS, PsSphS_{α} for $\alpha > 1$ and optimal value of α.

Generally, the MSEs are identical to when estimating by MLE, with the exception that the MSE of LogS and $\text{PsSphS}_{1.01}$ are equal.

We omit the estimation by the same scoring rule for the conditional variance, as this did not yield any usable predictions.

5 Conclusion

This paper demonstrates the efficacy of the strictly proper PseudoSpherical scoring rule in enhancing the robustness of score-driven updates. Strictly proper scoring rules, such as the logarithmic scoring rule implemented in the Generalized Autoregressive Score model, are effective in contributing to accurate and well calibrated forecasts. However, in the case of the logarithmic scoring rule, the score's direct proportionality to the size of observations makes it very sensitive to outliers, potentially causing large, incorrect parameter updates.

The PsSphS emerges as a robust alternative due to its relationship with the logarithmic scoring rule and its documented efficacy in existing literature. The PsSphS introduces an exponential term that modulates the influence of large observations for both the conditional mean and conditional variance. A key feature of the PsSphS is the tuning parameter α, which adjusts the model's sensitivity to outliers and can replicate the logarithmic scoring rule as a limiting case, showcasing its flexibility.

Simulations mostly back the theoretical findings, demonstrating the PsSphS's robustness to outliers. While this robustness for the conditional mean comes at the expense of overall fit for larger values of α, it is advantageous if conservative estimates are desired. For the conditional variance, the PsSphS displayed predictions superior in both robustness and fit compared to the logarithmic score, clearly putting the PsSphS on top overall. Implementation of the models was challenging however, as there was a high degree of sensitivity to the initial parameters used for maximum likelihood estimation. This would often lead to high computational burden as well as incoherent results.

Future research could explore a time-varying α parameter, akin to the accelerated GAS model, which sets the parameter of the score innovation to be time varying. This could also involve incorporating regime switches such that the model dynamically changes α based on

the data environment, better adapting to the data. Simulations indicate that the optimal α value varies significantly with the data environment, also suggesting that model averaging with models using the PsSphS with different α values would be an effective tool to utilize the PsSphS to the full extent.

Overall, the PsSphS provides a robust extension to the established score-driven framework. Its ability to modulate sensitivity to outliers makes it a valuable tool for robust forecasting and modeling, setting the groundwork for more resilient score-driven methodologies.

References

Blasques, F., C. Francq, and S. Laurent (2023), "Quasi score-driven models", Journal of Econometrics, 234(1), 251–275.

Blasques, F., S. J. Koopman, and A. Lucas (2015), "Information-theoretic optimality of observation-driven time series models for continuous responses", Biometrika, 102(2), 325–343.

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. F. A., C. G. H. Diks, R. J. A. Laeven, and D. J. C. Van Dijk (2023), "Localizing Strictly Proper Scoring Rules", 23(84).

Diks, C., V. Panchenko, and D. Van Dijk (2011), "Likelihood-based scoring rules for comparing density forecasts in tails", Journal of Econometrics, 163(2), 215–230.

Gneiting, T. and A. E. Raftery (2007), "Strictly Proper Scoring Rules, Prediction, and Estimation", Journal of the American Statistical Association, 102(477), 359–378.

Gorgi, P., C. S. A. Lauria, and A. Luati (2023), "On the optimality of score-driven models", Biometrika, , asad067.

Harvey, A. C. (2013). Dynamic models for volatility and heavy tails: with applications to financial time series. Econometric Society monographs. Cambridge: Cambridge Univ. Press.

Jose, V. R. R., R. F. Nau, and R. L. Winkler (2008), "Scoring Rules, Generalized Entropy, and Utility Maximization", Operations Research, 56(5), 1146–1157.

Kanamori, T. and H. Fujisawa (2015), "Robust estimation under heavy contamination using unnormalized models", Biometrika, 102(3), 559–572.

Kleen, O. (2023), "Scaling and measurement error sensitivity of scoring rules for distribution forecasts", Journal of Applied Econometrics, n/a(n/a).

Ovcharov, E. Y. (2018), "Proper scoring rules and Bregman divergence", Bernoulli, 24(1).

6 Appendix

6.1 Derivations of Models

6.1.1 Derivation of the Partial Derivative of the Pseudospherical Score with respect to μ

Let $y \sim \mathcal{N}(\mu, 1)$ and the PDF of the normal distribution be:

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}$$

The L^α-norm of f is:

$$\begin{aligned}
\|f\|_\alpha &= \left(\int_{-\infty}^{\infty} f^\alpha(y)\, dy \right)^{\frac{1}{\alpha}} \\
&= \left(\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}} \right)^\alpha dy \right)^{\frac{1}{\alpha}} \\
&= \left(\frac{1}{(2\pi)^{\frac{\alpha}{2}}} \int_{-\infty}^{\infty} e^{-\frac{\alpha(y-\mu)^2}{2}}\, dy \right)^{\frac{1}{\alpha}}
\end{aligned}$$

The integral inside the expression is a Gaussian integral, and its solution is well-known. Namely:

$$\int_{-\infty}^{\infty} e^{-\frac{\alpha(y-\mu)^2}{2}} dy = \sqrt{\frac{2\pi}{\alpha}}$$

Substituting this result back into our expression:

$$\begin{aligned}
\|f\|_\alpha &= \left(\frac{1}{(2\pi)^{\frac{\alpha}{2}}} \sqrt{\frac{2\pi}{\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \left(\frac{1}{(2\pi)^{\frac{\alpha}{2}}} \cdot \sqrt{\frac{2\pi}{\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \left((2\pi)^{-\frac{\alpha}{2}} \cdot \sqrt{\frac{2\pi}{\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \left((2\pi)^{-\frac{\alpha+1}{2}} \cdot \alpha^{-\frac{1}{2}} \right)^{\frac{1}{\alpha}} \\
&= (2\pi)^{-\frac{\alpha+1}{2\alpha}} \alpha^{-\frac{1}{2\alpha}}
\end{aligned}$$

The pseudospherical score is:

$$\mathrm{PsSphS}_\alpha(f, y) = \frac{f(y)^{\alpha-1}}{\|f\|_\alpha^{\alpha-1}}$$

Substituting $f(y)$ and $\|f\|_\alpha$:

$$\begin{aligned}
\mathrm{PsSphS}_\alpha(f,y) &= \frac{\left(\frac{1}{\sqrt{2\pi}}e^{-\frac{(y-\mu)^2}{2}}\right)^{\alpha-1}}{\left((2\pi)^{-\frac{\alpha+1}{2\alpha}}\alpha^{-\frac{1}{2\alpha}}\right)^{\alpha-1}} \\
&= \frac{(2\pi)^{-\frac{\alpha-1}{2}}e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}}{(2\pi)^{-\frac{(\alpha+1)(\alpha-1)}{2\alpha}}\alpha^{-\frac{\alpha-1}{2\alpha}}} \\
&= (2\pi)^{-\frac{(\alpha-1)}{2\alpha}}\alpha^{\frac{\alpha-1}{2\alpha}}e^{-\frac{(\alpha-1)(y-\mu)^2}{2}} \\
&= \left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}
\end{aligned}$$

The partial derivative with respect to μ is:

$$\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\mu} = \frac{\partial}{\partial\mu}\left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}$$

Since $\left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}}$ is a constant with respect to μ, we focus on the exponential term:

$$\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\mu} = \left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} \cdot \frac{\partial}{\partial\mu}\left(e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}\right)$$

Apply the chain rule:

$$\begin{aligned}
\frac{\partial}{\partial\mu}\left(e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}\right) &= e^{-\frac{(\alpha-1)(y-\mu)^2}{2}} \cdot \frac{\partial}{\partial\mu}\left(-\frac{(\alpha-1)(y-\mu)^2}{2}\right) \\
\frac{\partial}{\partial\mu}\left(-\frac{(\alpha-1)(y-\mu)^2}{2}\right) &= (\alpha-1)(y-\mu)
\end{aligned}$$

So:

$$\begin{aligned}
\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\mu} &= \left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)(y-\mu)^2}{2}} \cdot (\alpha-1)(y-\mu) \\
&= \left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha-1)(y-\mu)e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}
\end{aligned}$$

Thus, the partial derivative of the pseudospherical score with respect to the parameter μ is:

$$\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\mu} = \left(\frac{\alpha}{2\pi}\right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha-1)(y-\mu)e^{-\frac{(\alpha-1)(y-\mu)^2}{2}}$$

6.1.2 Derivation of the Partial Derivative of Pseudospherical score w.r.t σ^2

Let $y \sim \mathcal{N}(0, \sigma^2)$ and the PDF of the normal distribution be:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$$

Similarly to above, we simplify the L^α-norm of f as:

$$\begin{aligned}
\|f\|_\alpha &= \left(\int_{-\infty}^{\infty} f^\alpha(y)\, dy \right)^{\frac{1}{\alpha}} \\
&= \left(\int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}} \right)^\alpha dy \right)^{\frac{1}{\alpha}} \\
&= \left(\frac{1}{(2\pi\sigma^2)^{\frac{\alpha}{2}}} \int_{-\infty}^{\infty} e^{-\frac{\alpha y^2}{2\sigma^2}}\, dy \right)^{\frac{1}{\alpha}}
\end{aligned}$$

Again recognizing the Gaussian integral:

$$\int_{-\infty}^{\infty} e^{-\frac{\alpha y^2}{2\sigma^2}}\, dy = \sqrt{\frac{2\pi\sigma^2}{\alpha}}$$

Simplifying our expression as:

$$\begin{aligned}
\|f\|_\alpha &= \left(\frac{1}{(2\pi\sigma^2)^{\frac{\alpha}{2}}} \sqrt{\frac{2\pi\sigma^2}{\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \left(\frac{1}{(2\pi\sigma^2)^{\frac{\alpha}{2}}} \cdot \sqrt{\frac{2\pi\sigma^2}{\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \left((2\pi\sigma^2)^{-\frac{\alpha}{2}} \cdot \sqrt{\frac{2\pi\sigma^2}{\alpha}} \right)^{\frac{1}{\alpha}} \\
&= \left((2\pi\sigma^2)^{-\frac{\alpha+1}{2}} \cdot \alpha^{-\frac{1}{2}} \right)^{\frac{1}{\alpha}} \\
&= (2\pi\sigma^2)^{-\frac{\alpha+1}{2\alpha}} \alpha^{-\frac{1}{2\alpha}}
\end{aligned}$$

Substituting $f(y)$ and $\|f\|_\alpha$ into the PsSphS:

$$\begin{aligned}
\mathrm{PsSphS}_\alpha(f,y) &= \frac{\left(\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{y^2}{2\sigma^2}}\right)^{\alpha-1}}{\left((2\pi\sigma^2)^{-\frac{\alpha+1}{2\alpha}}\alpha^{-\frac{1}{2\alpha}}\right)^{\alpha-1}} \\
&= \frac{(2\pi\sigma^2)^{-\frac{(\alpha-1)}{2}}e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}}{(2\pi\sigma^2)^{-\frac{(\alpha+1)(\alpha-1)}{2\alpha}}\alpha^{-\frac{(\alpha-1)}{2\alpha}}} \\
&= (2\pi\sigma^2)^{-\frac{(\alpha-1)}{2\alpha}}\alpha^{\frac{(\alpha-1)}{2\alpha}}e^{-\frac{(\alpha-1)y^2}{2\sigma^2}} \\
&= \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}
\end{aligned}$$

Such that the PsSphS can be written as:

$$\mathrm{PsSphS}_\alpha(f,y) = \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}$$

The partial derivative with respect to σ^2 is:

$$\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\sigma^2} = \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} \left(\alpha^{\frac{\alpha-1}{2}}e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}\right)$$

Since $\left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}}$ is a constant with respect to σ^2, we focus on the exponential term:

$$\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\sigma^2} = \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} \cdot \frac{\partial}{\partial\sigma^2}\left(e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}\right)$$

Apply the chain rule:

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}\left(e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}\right) &= e^{-\frac{(\alpha-1)y^2}{2\sigma^2}} \cdot \frac{\partial}{\partial\sigma^2}\left(-\frac{(\alpha-1)y^2}{2\sigma^2}\right) \\
\frac{\partial}{\partial\sigma^2}\left(-\frac{(\alpha-1)y^2}{2\sigma^2}\right) &= (\alpha-1)\frac{y^2}{2\sigma^4}
\end{aligned}$$

So:

$$\begin{aligned}
\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\sigma^2} &= \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} e^{-\frac{(\alpha-1)y^2}{2\sigma^2}} \cdot (\alpha-1)\frac{y^2}{2\sigma^4} \\
&= \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} \frac{(\alpha-1)y^2}{2\sigma^4}e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}
\end{aligned}$$

Thus, the partial derivative of the pseudospherical score with respect to σ^2 is:

$$\frac{\partial \mathrm{PsSphS}_\alpha(f,y)}{\partial\sigma^2} = \left(\frac{\alpha}{2\pi\sigma^2}\right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha-1)\frac{y^2}{2\sigma^4}e^{-\frac{(\alpha-1)y^2}{2\sigma^2}}$$

6.1.3 Derivation of the Partial Derivative of the Logarithmic score with respect to μ

Let $y \sim \mathcal{N}(\mu, 1)$ and the PDF of the normal distribution be:

$$f(y) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}$$

The logarithm of the PDF is:

$$\begin{aligned}
\log f(y) &= \log\left(\frac{1}{\sqrt{2\pi}} e^{-\frac{(y-\mu)^2}{2}}\right) \\
&= \log\left(\frac{1}{\sqrt{2\pi}}\right) + \log\left(e^{-\frac{(y-\mu)^2}{2}}\right) \\
&= -\frac{1}{2}\log(2\pi) - \frac{(y-\mu)^2}{2}
\end{aligned}$$

Taking the partial derivative w.r.t μ yields:

$$\begin{aligned}
\frac{\partial}{\partial \mu} \log f(y) &= \frac{\partial}{\partial \mu}\left(-\frac{1}{2}\log(2\pi) - \frac{(y-\mu)^2}{2}\right) \\
&= \frac{\partial}{\partial \mu}\left(-\frac{(y-\mu)^2}{2}\right) \\
&= -\frac{1}{2} \cdot 2(y-\mu) \cdot (-1) \\
&= y - \mu
\end{aligned}$$

Thus, the partial derivative of the logarithmic score with respect to μ is:

$$\frac{\partial}{\partial \mu} \log f(y) = y - \mu$$

6.1.4 Proof: Logarithmic scoring rule is a limiting case of PseudoSpherical Scoring rule as $\alpha \to 1^+$

We prove the following:

$$\lim_{\alpha \to 1^+} \frac{1}{\alpha - 1} \left(\frac{f(y)}{\|f\|_\alpha} \right)^{\alpha - 1} = \log f(y)$$

for the case where $y_t \sim \mathcal{N}(\mu, 1)$.

Let $g_\alpha(\mu) = \text{PsSphS}_\alpha$, then:

$$\lim_{\alpha \to 1^+} g'_\alpha(\mu) = f'(\mu), \quad \text{where } f(\mu) = \text{logS}$$

We have shown previously that:

$$\begin{aligned} g'_\alpha(\mu) &= \left(\frac{\alpha}{2\pi} \right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha - 1)(y - \mu) e^{-\frac{(\alpha-1)(y-\mu)^2}{2}} \\ f'(\mu) &= y - \mu \end{aligned}$$

To prove the expression above, we can show that:

$$\begin{aligned} \lim_{\alpha \to 1^+} \frac{1}{(\alpha - 1)} g'_\alpha(\mu) &= \lim_{\alpha \to 1^+} \frac{1}{(\alpha - 1)} \left(\frac{\alpha}{2\pi} \right)^{\frac{(\alpha-1)}{2\alpha}} (\alpha - 1)(y - \mu) e^{-\frac{(\alpha-1)(y-\mu)^2}{2}} \\ &= \lim_{\alpha \to 1^+} \left(\frac{\alpha}{2\pi} \right)^{\frac{(\alpha-1)}{2\alpha}} (y - \mu) e^{-\frac{(\alpha-1)(y-\mu)^2}{2}} \\ &= y - \mu \end{aligned}$$

Which proves that the result holds for $y_t \sim \mathcal{N}(\mu, 1)$. The result can be shown to hold for different cases in the same manner.

6.1.5 Derivation of the Partial Derivative of the Logarithmic score with respect to σ^2

Let $y \sim \mathcal{N}(0, \sigma^2)$ and the PDF of the normal distribution be:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}$$

The logarithm of the PDF is:

$$\begin{aligned}
\log f(y) &= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{y^2}{2\sigma^2}}\right) \\
&= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(e^{-\frac{y^2}{2\sigma^2}}\right) \\
&= \log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) - \frac{y^2}{2\sigma^2} \\
&= \log 1 - \log\sqrt{2\pi\sigma^2} - \frac{y^2}{2\sigma^2} \\
&= -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2} \\
&= -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{y^2}{2\sigma^2}
\end{aligned}$$

The partial derivative with respect to σ^2 is:

$$\frac{\partial}{\partial\sigma^2}\log f(y) = \frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{y^2}{2\sigma^2}\right)$$

Simplifying the derivative:

$$\frac{\partial}{\partial\sigma^2}\log f(y) = \frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2}\log\sigma^2\right) - \frac{\partial}{\partial\sigma^2}\left(\frac{y^2}{2\sigma^2}\right)$$

Computing each term separately:

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}\left(-\frac{1}{2}\log\sigma^2\right) &= -\frac{1}{2}\cdot\frac{1}{\sigma^2} = -\frac{1}{2\sigma^2} \\
\frac{\partial}{\partial\sigma^2}\left(\frac{y^2}{2\sigma^2}\right) &= \frac{y^2}{2}\cdot\frac{\partial}{\partial\sigma^2}\left(\frac{1}{\sigma^2}\right) = \frac{y^2}{2}\cdot -\frac{1}{\sigma^4} = -\frac{y^2}{2\sigma^4}
\end{aligned}$$

Combining the terms:

$$\begin{aligned}
\frac{\partial}{\partial\sigma^2}\log f(y) &= -\frac{1}{2\sigma^2} - \left(-\frac{y^2}{2\sigma^4}\right) \\
&= -\frac{1}{2\sigma^2} + \frac{y^2}{2\sigma^4} \\
&= \frac{y^2 - \sigma^2}{2\sigma^4}
\end{aligned}$$

Thus, the partial derivative of the logarithmic score with respect to σ^2 is:

$$\frac{\partial}{\partial\sigma^2}\log f(y) = \frac{y^2 - \sigma^2}{2\sigma^4}$$

6.1.6 Derivation of Fisher Information for $y \sim \mathcal{N}(\mu, 1)$

Given an observation from $y \sim \mathcal{N}(\mu, 1)$, the probability density function is:

$$f(y; \mu) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(y-\mu)^2}{2}\right)$$

The log-likelihood function for a single observation y is:

$$\log L(\mu) = \log f(y; \mu) = -\frac{1}{2}\log(2\pi) - \frac{(y-\mu)^2}{2}$$

First, we compute the first derivative of the log-likelihood with respect to μ:

$$\frac{\partial}{\partial \mu} \log L(\mu) = \frac{(y-\mu)}{1} = y - \mu$$

Next, we compute the second derivative of the log-likelihood with respect to μ:

$$\frac{\partial^2}{\partial \mu^2} \log L(\mu) = -1$$

Taking the expectation of the second derivative, we get:

$$\mathbb{E}\left[\frac{\partial^2}{\partial \mu^2} \log L(\mu)\right] = -1$$

The Fisher information $\mathcal{I}(\mu)$ is the negative of this expectation:

$$\begin{aligned}\mathcal{I}(\mu) &= -\mathbb{E}\left[\frac{\partial^2}{\partial \mu^2} \log L(\mu)\right] \\ &= -(-1) \\ &= 1\end{aligned}$$

This implies $\rho_t = 1$ for $y \sim \mathcal{N}(\mu_t, 1)$

6.1.7 Derivation of Fisher Information for $y \sim \mathcal{N}(0, \sigma^2)$

Given an observation from $y \sim \mathcal{N}(0, \sigma^2)$, the probability density function is:

$$f(y; \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{y^2}{2\sigma^2}\right)$$

The log-likelihood function for a single observation y is:

$$\log L(\sigma^2) = \log f(y; \sigma^2) = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{y^2}{2\sigma^2}$$

First, we compute the first derivative of the log-likelihood with respect to σ^2:

$$\frac{\partial}{\partial \sigma^2} \log L(\sigma^2) = -\frac{1}{2\sigma^2} + \frac{y^2}{2(\sigma^2)^2}$$

Next, we compute the second derivative of the log-likelihood with respect to σ^2:

$$\frac{\partial^2}{\partial \sigma^2} \log L(\sigma^2) = \frac{1}{2(\sigma^2)^2} - \frac{y^2}{(\sigma^2)^3}$$

Taking the expectation of the second derivative, and noting that $\mathbb{E}[y^2] = \sigma^2$ for $Y \sim \mathcal{N}(0, \sigma^2)$, we get:

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial^2}{\partial \sigma^2} \log L(\sigma^2)\right] &= \mathbb{E}\left[\frac{1}{2(\sigma^2)^2} - \frac{y^2}{(\sigma^2)^3}\right] \\
&= \frac{1}{2(\sigma^2)^2} - \frac{\sigma^2}{(\sigma^2)^3} \\
&= \frac{1}{2(\sigma^2)^2} - \frac{1}{2(\sigma^2)^2} \\
&= 0
\end{aligned}$$

Since the expected value of the second derivative is zero, we use the variance of the score function to compute the Fisher information. The score function $U(\sigma^2)$ is:

$$U(\sigma^2) = \frac{\partial}{\partial \sigma^2} \log L(\sigma^2) = -\frac{1}{2\sigma^2} + \frac{y^2}{2(\sigma^2)^2}$$

The Fisher information $\mathcal{I}(\sigma^2)$ is the expected value of the square of the score function:

$$\begin{aligned}
\mathcal{I}(\sigma^2) &= \mathbb{E}\left[\left(\frac{\partial}{\partial \sigma^2} \log L(\sigma^2)\right)^2\right] \\
&= \mathbb{E}\left[\left(-\frac{1}{2\sigma^2} + \frac{y^2}{2(\sigma^2)^2}\right)^2\right] \\
&= \mathbb{E}\left[\left(\frac{y^2 - \sigma^2}{2(\sigma^2)^2}\right)^2\right] \\
&= \frac{1}{4(\sigma^2)^4} \mathbb{E}\left[(y^2 - \sigma^2)^2\right]
\end{aligned}$$

Since $y^2 \sim \sigma^2 \chi_1^2$, we know that $\mathbb{E}[y^2] = \sigma^2$ and $\mathrm{Var}(y^2) = 2\sigma^4$:

$$\mathbb{E}\left[(y^2 - \sigma^2)^2\right] = \mathrm{Var}(y^2) = 2\sigma^4$$

Substituting this back into the Fisher information equation, we get:

$$\begin{aligned}
\mathcal{I}(\sigma^2) &= \frac{1}{4(\sigma^2)^4} \cdot 2\sigma^4 \\
&= \frac{1}{2\sigma^4}
\end{aligned}$$

This implies $\rho_t = \mathcal{I}(\sigma^2)^{-1} = 2\sigma_t^4$ for $y \sim \mathcal{N}(0, \sigma_t^2)$

6.1.8 Proof: Linear Scoring Rule is Strictly Proper for $y_t \sim N(\mu, 1)$ when L^2-norm is independent of μ

To prove that the linear scoring rule is strictly proper, we need to show that:

$$\mathbb{E}_p[\text{LinS}(p, y)] - \mathbb{E}_p[\text{LinS}(f, y)] \geq 0$$

with equality if and only if $p = f$.

The linear scoring rule is given by:

$$\text{LinS}(p, y) = p(y)$$

The expected score under the true density p is:

$$\mathbb{E}_p[\text{LinS}(p, y)] = \int_{-\infty}^{\infty} p(y)p(y)\, dy = \langle p, p \rangle = \|p\|_2^2$$

The expected score under the predictive density f is:

$$\mathbb{E}_p[\text{LinS}(f, y)] = \int_{-\infty}^{\infty} p(y)f(y)\, dy = \langle p, f \rangle$$

Computing the difference:

$$\mathbb{E}_p[\text{LinS}(p, y)] - \mathbb{E}_p[\text{LinS}(f, y)] = \|p\|_2^2 - \langle p, f \rangle$$

By the Cauchy-Schwarz inequality for integrals, we have:

$$\langle p, f \rangle \leq \|p\|_2 \|f\|_2$$

Equality holds if and only if $p = kf$ for some constant k. Since both p and f are probability densities, they integrate to 1, so k must be 1, implying $p = f$. We have that:

$$\|p\|_2^2 - \langle p, f \rangle \geq \|p\|_2^2 - \|p\|_2 \|f\|_2 \geq 0$$

by the Cauchy Schwarz inequality. The inequality holds strict if and only if:

$$\|p\|_2 \|f\|_2 = \langle p, f \rangle \iff p = f$$

Since the L^2-norm is independent of μ, it regardless holds that $\|p\|_2 = \|f\|_2$ and thus the linear scoring rule is strictly proper.

6.2 Simulations

6.2.1 Conditional Mean with normally distributed errors

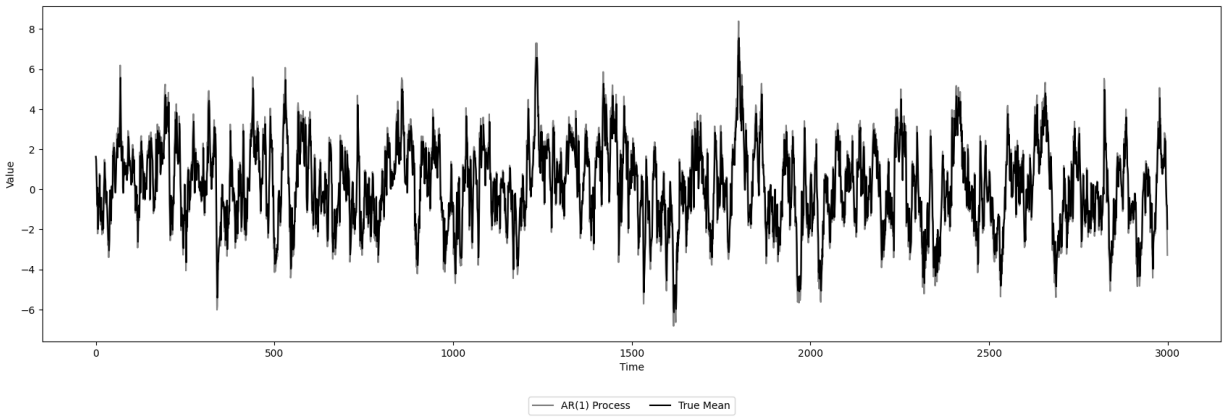

Figure 12: 3000 observations of an AR(1) process with true conditional mean.

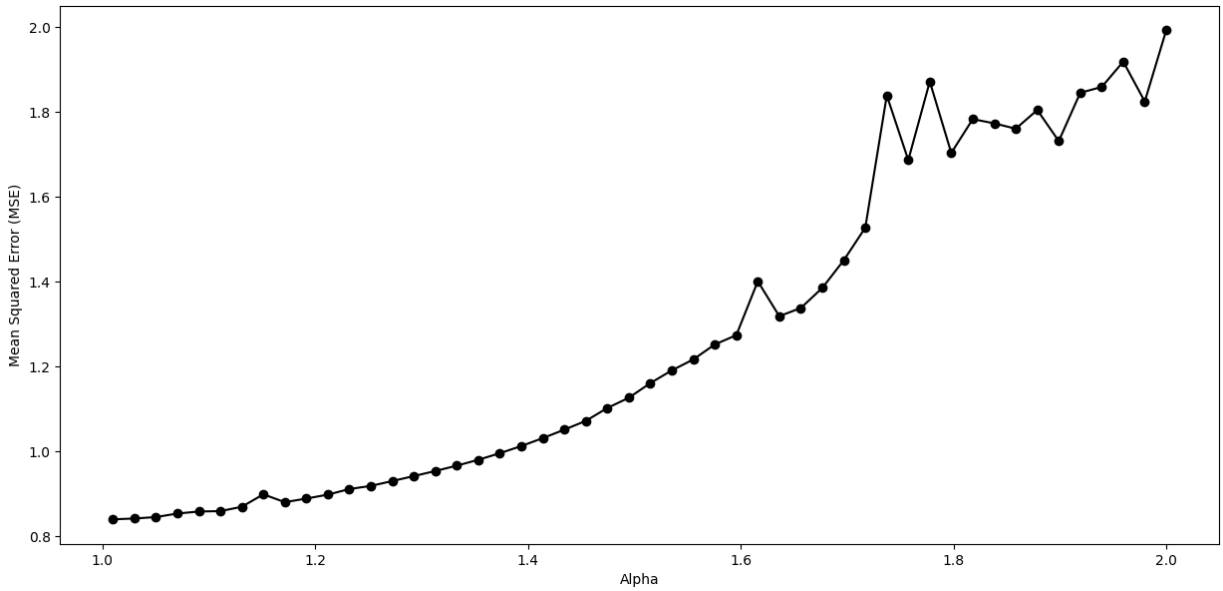

Figure 13: MSE of PsSphS_α for $1.01 \leq \alpha \leq 2$

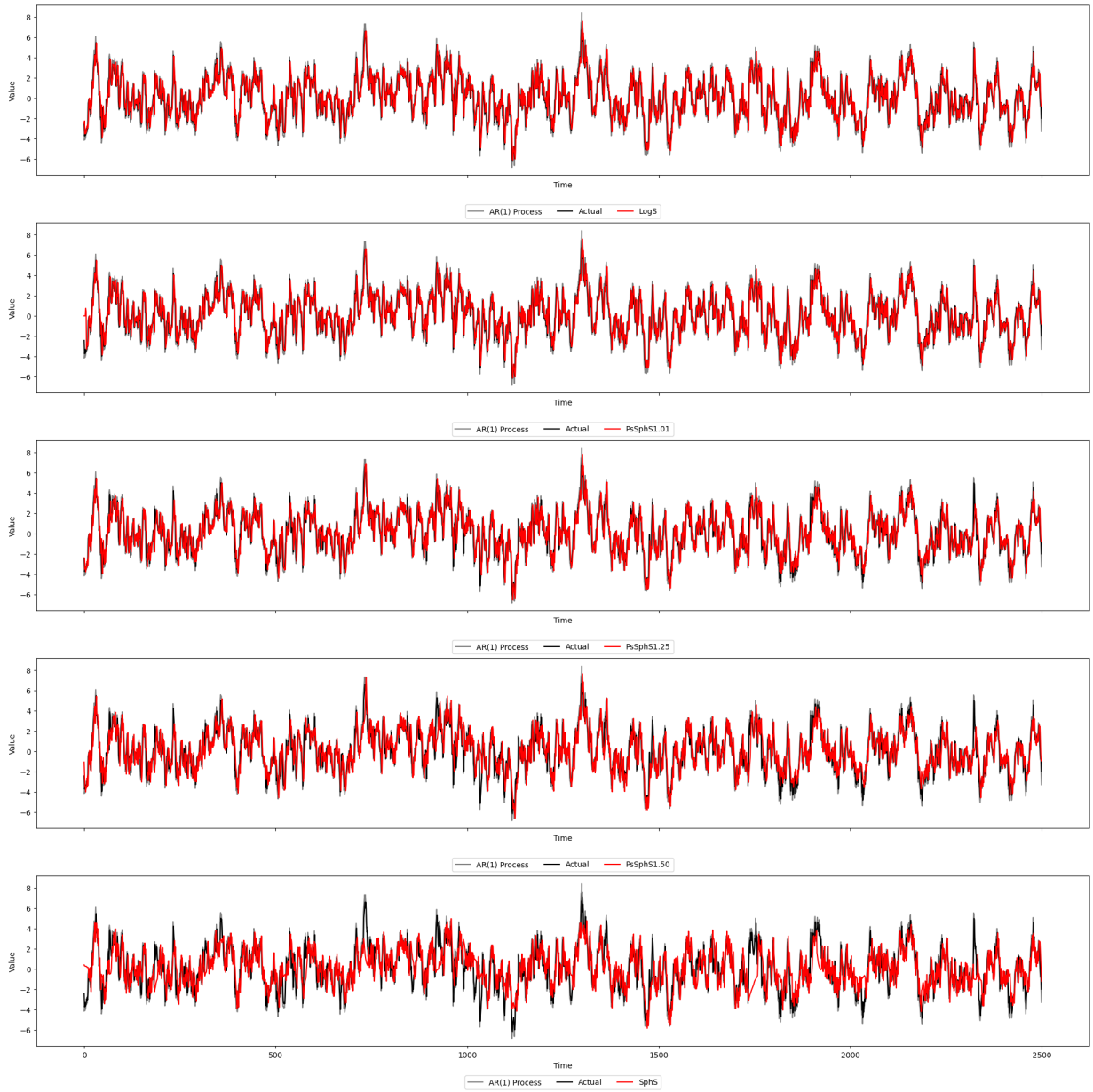

Figure 14: Model predictions for true conditional mean

6.2.2 Conditional Variance with normally distributed errors

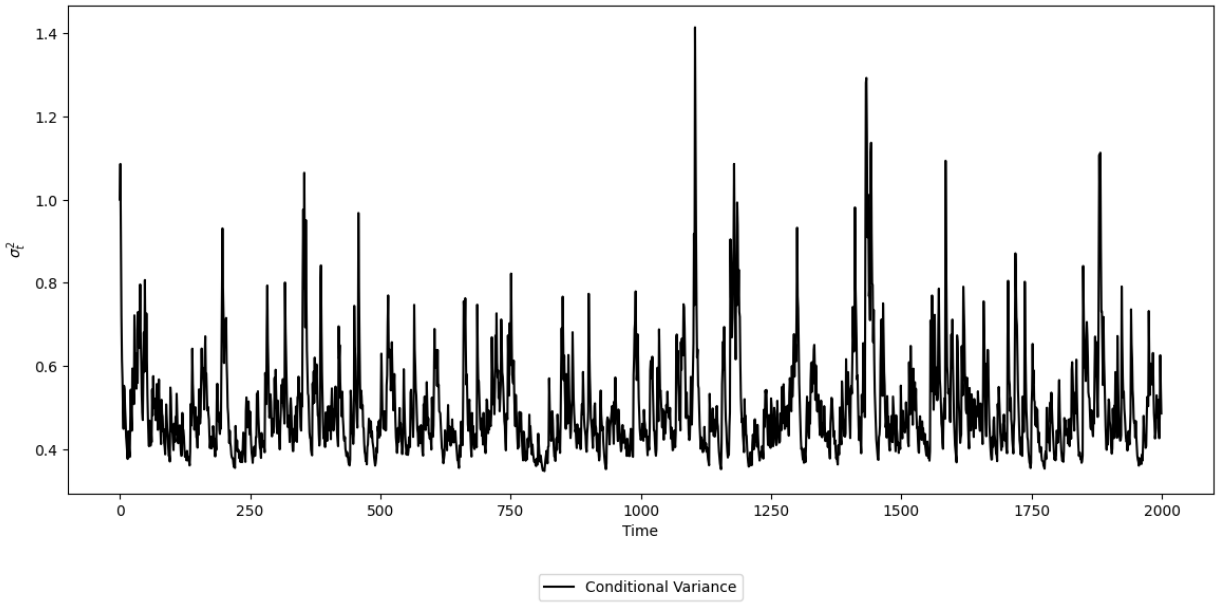

Figure 15: 2000 observations of conditional variance of GARCH(1,1) process

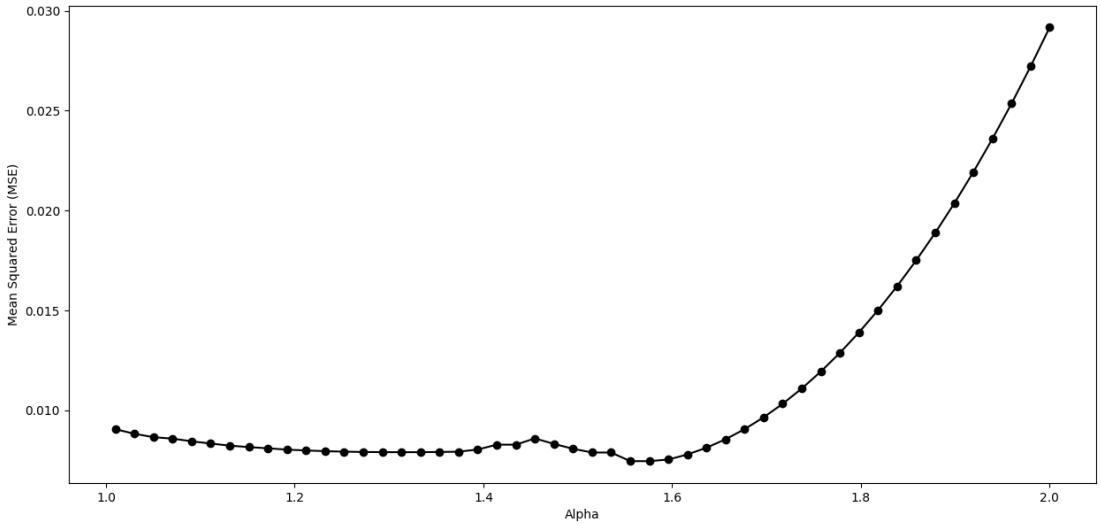

Figure 16: MSE of PsSphS_α for $1.01 \leq \alpha \leq 2$

Model Predictions:

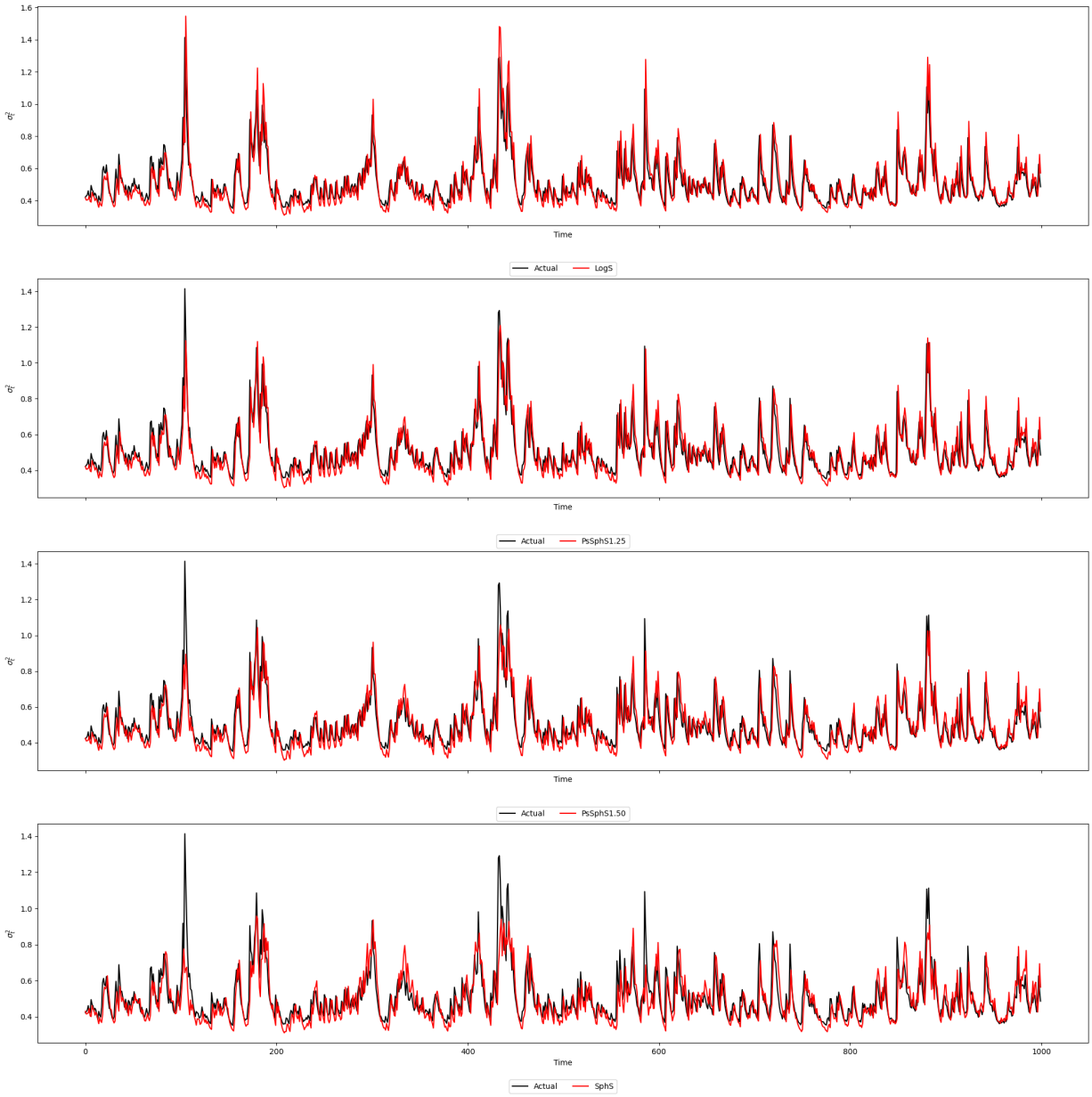

Figure 17: Model predictions for conditional variance

6.2.3 Conditional mean with t-distributed errors

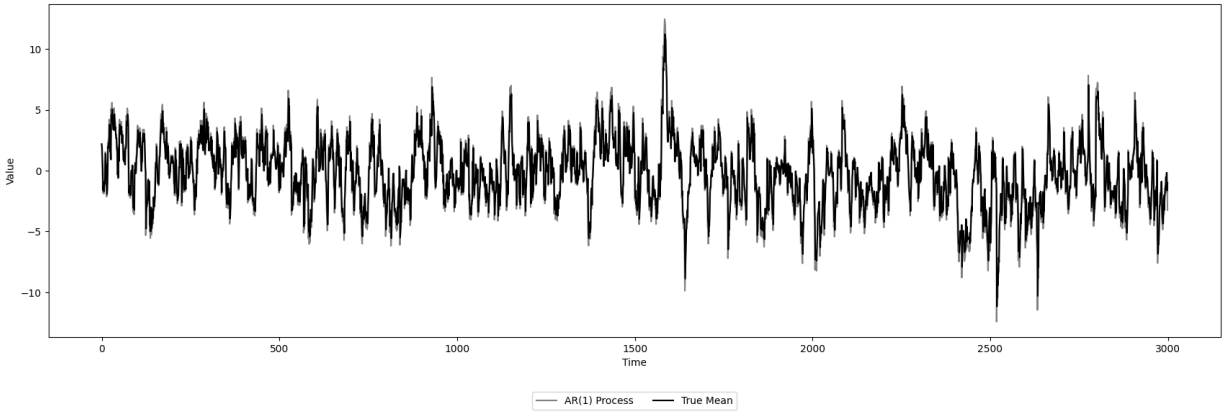

Figure 18: 3000 observations of an AR(1) process with true conditional mean.

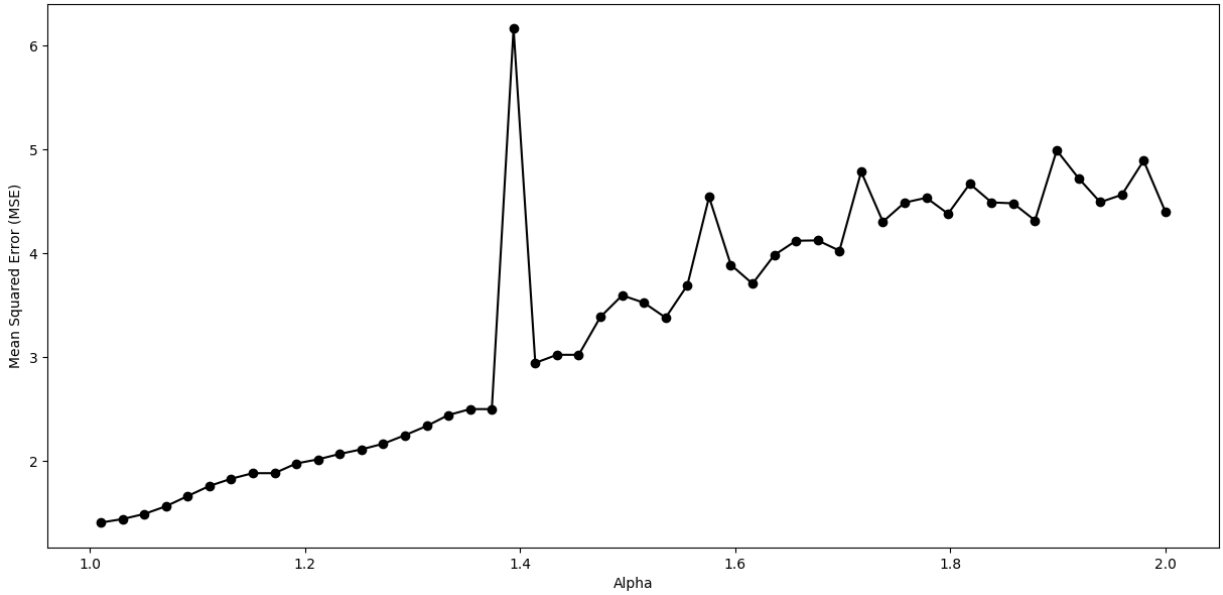

Figure 19: MSE of PsSphS_α for $1.01 \leq \alpha \leq 2$

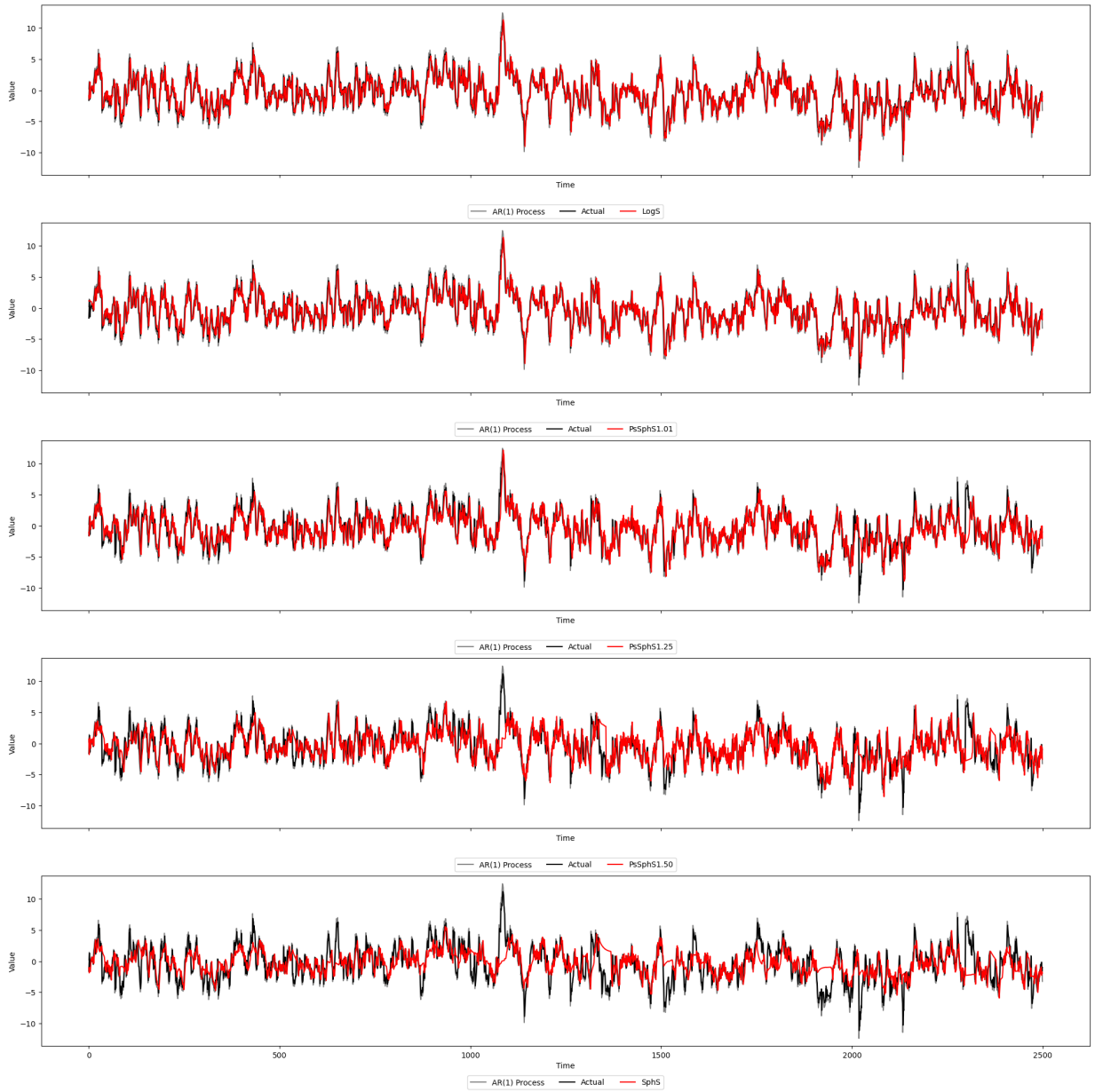

Figure 20: Model predictions for true conditional mean

6.2.4 Conditional variance with t-distributed errors

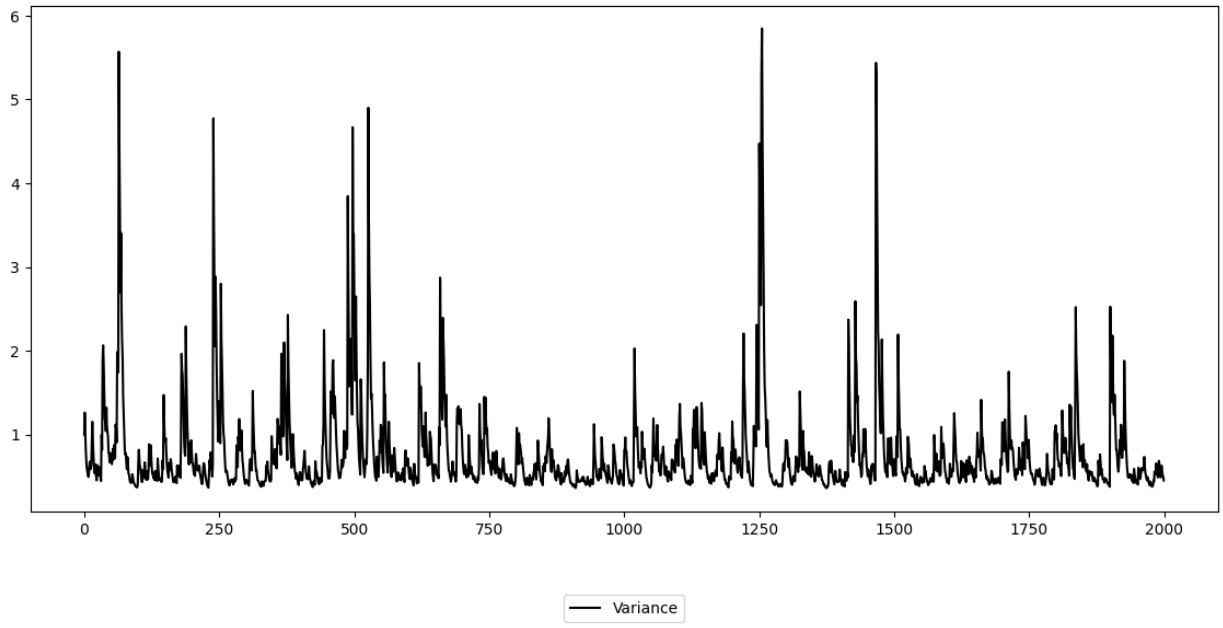

Figure 21: 2000 observations of the conditional variance of GARCH(1,1) process.

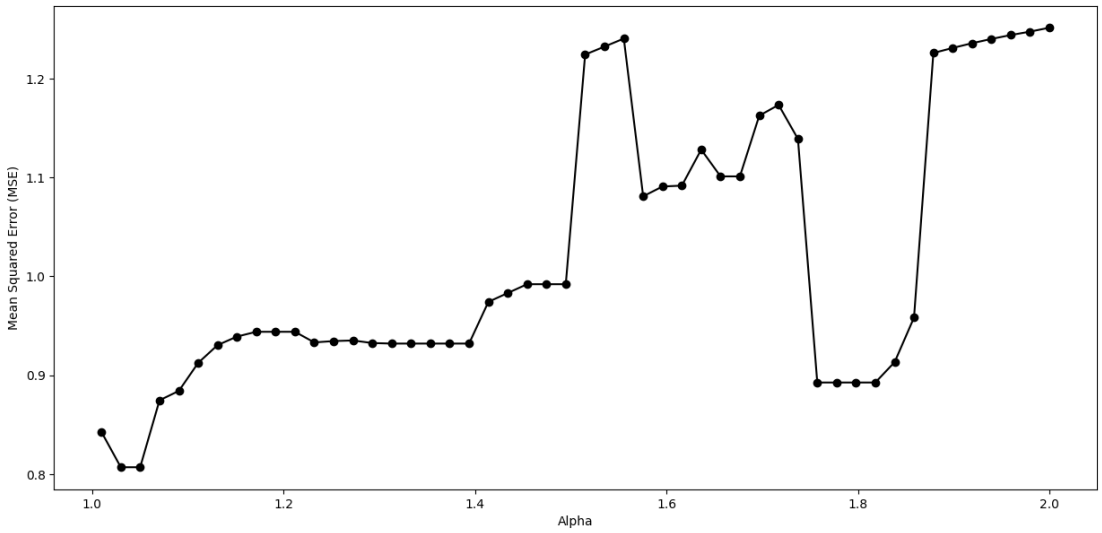

Figure 22: MSE of PsSphS_α for $1.01 \leq \alpha \leq 2$

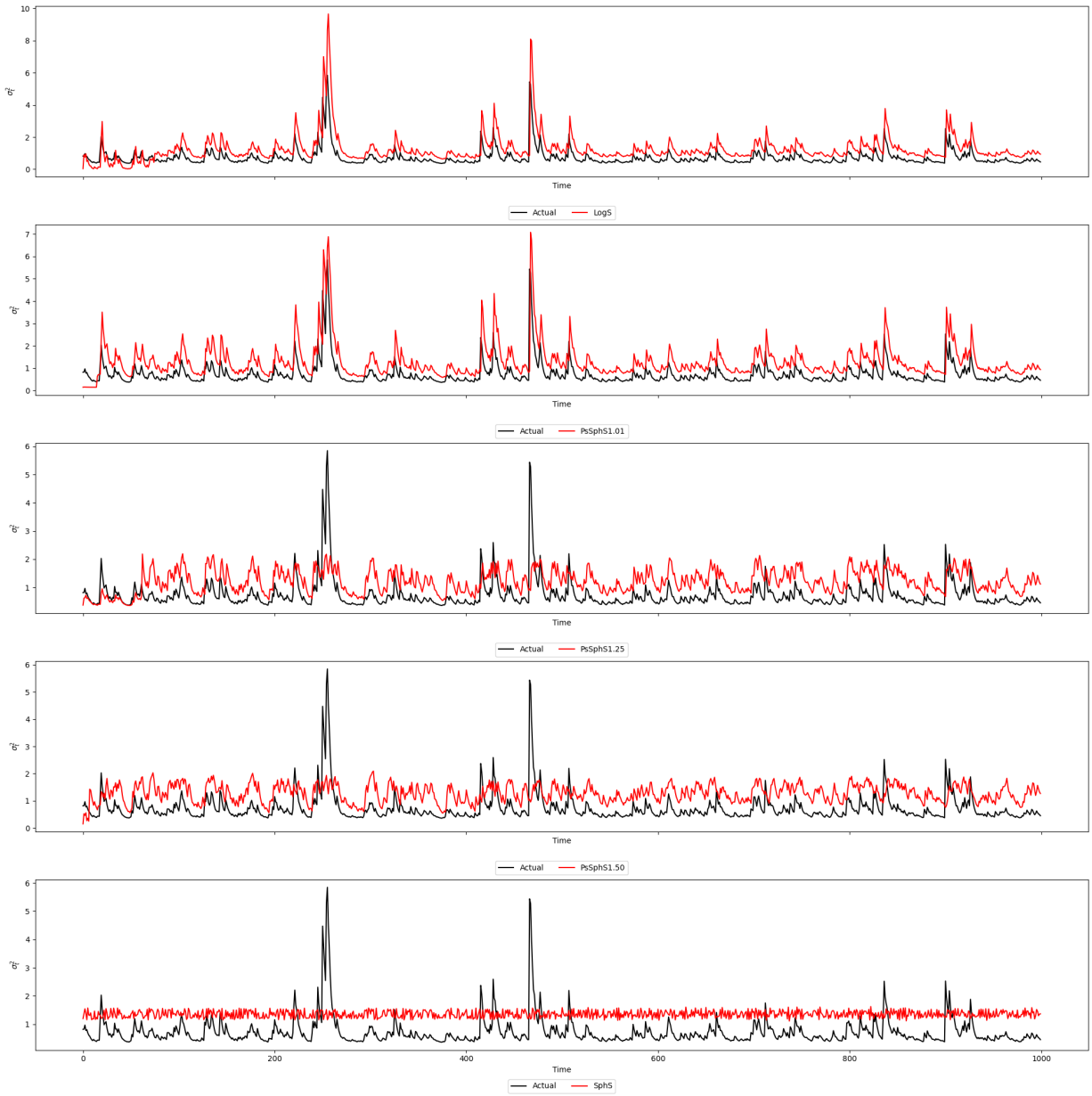

Figure 23: Model predictions for the conditional variance

6.2.5 Conditional mean by Same Scoring Rule

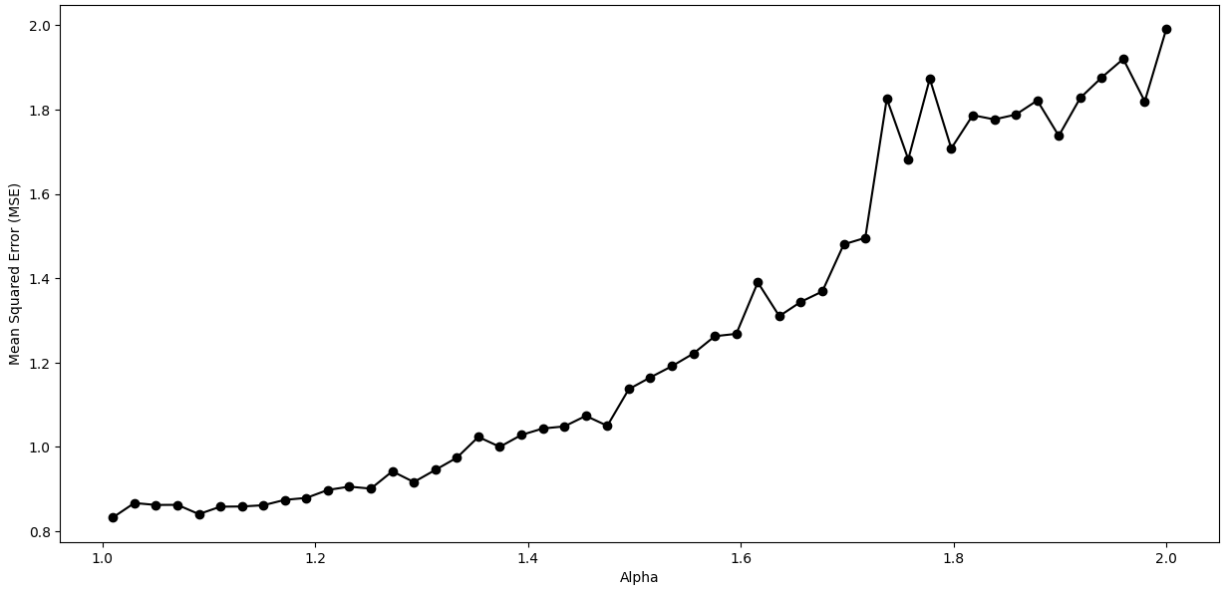

Figure 24: MSE of PsSphS_α for $1.01 \leq \alpha \leq 2$

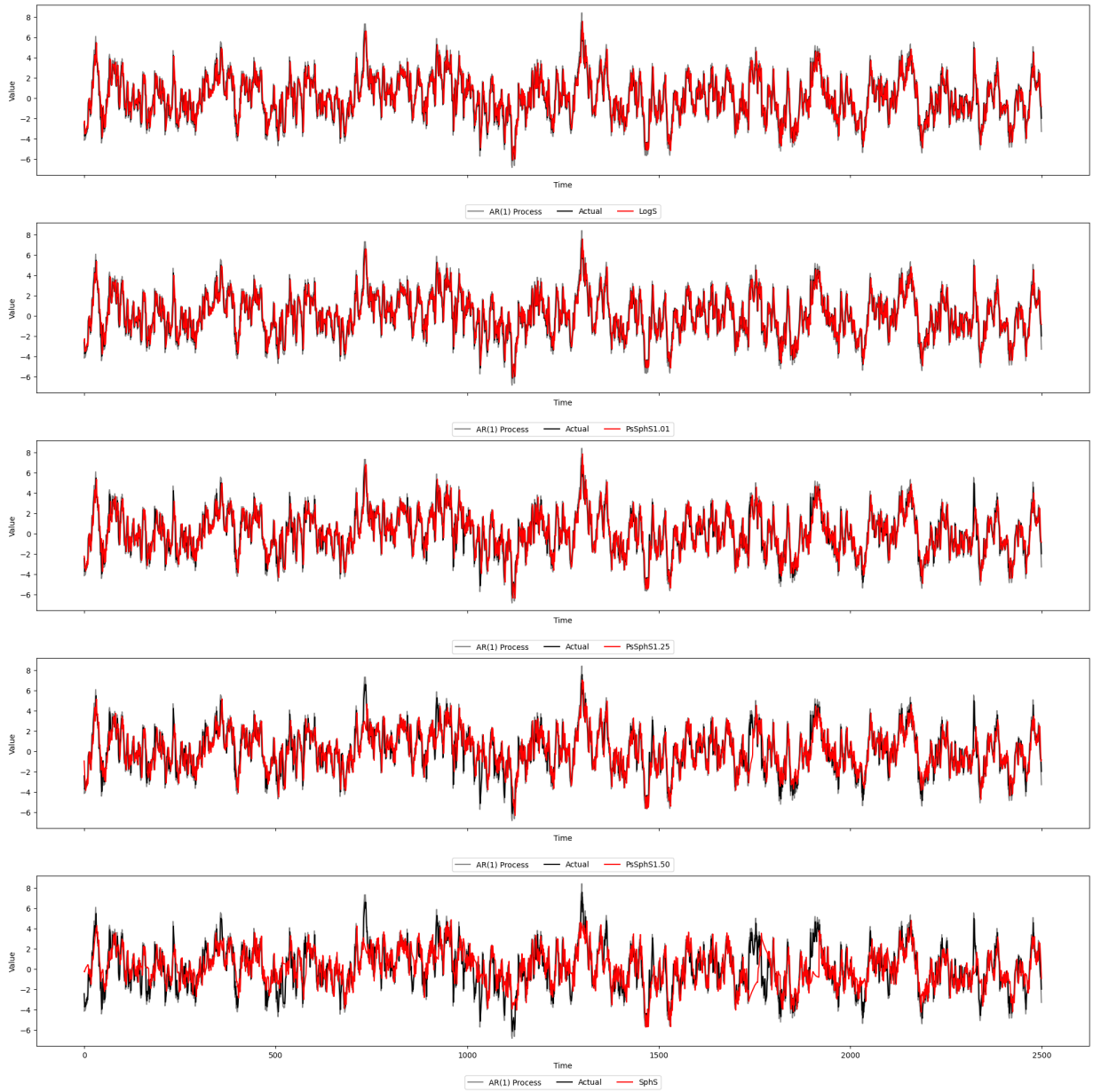

Figure 25: Model predictions using same scoring rule for conditional mean