UNIVERSITÀ DEGLI STUDI DI BERGAMO
Department of Economics

WORKING PAPERS

The judgmental strategy of professional forecasters

Emilio Zanetti Chini

March 2025 - WP N. 31 Year 2025

Working papers – Department of Economics
n. 31

The judgmental strategy of professional forecasters

Emilio Zanetti Chini

Università degli Studi di Bergamo
2025

The judgmental strategy of professional forecasters /Emilio Zanetti Chini – Bergamo: Università degli Studi di Bergamo, 2025.
Working papers of Department of Economics, n. 31
ISSN: 2974-5586
DOI: 10.13122/WPEconomics_31

Progetto grafico: Servizi Editoriali – Università degli Studi di Bergamo
Università degli Studi di Bergamo
via Salvecchio, 19
24129 Bergamo
Cod. Fiscale 80004350163
P. IVA 01612800167

https://aisberg.unibg.it/handle/10446/297845

The judgmental strategy of professional forecasters

Emilio Zanetti Chini*

University of Bergamo
Department of Economics
Via dei Caniana, 2 - 24127, Bergamo (ITALY)
e-mail: emilio.zanettichini@unibg.it

Abstract

We introduce a new definition of forecasting coherence based on the Likelihood Principle and a "Scoring Structure" environment where a Forecast User interacts with a Forecast Producer and Reality to detect strategic interaction among economic agents' forecasting bias. This mathematical object is necessary to identify and parametrize coherence in a feasible econometric model and give it a structural interpretation. Structural coherence is evaluated through a formal test that satisfies theoretical requirements in small samples. Three case studies illustrates the evidence of strategic judgment. The economic interpretation and the consequences of our approach in Central Banking are also discussed.

Keywords: Bias, Likelihood Principle, Mis-specification, Nonlinearity, Reputation, Scoring Structures.
JEL:E50, C12, C22, C44, C53.

*The Author is particularly grateful to Raffaella Giacomini for her suggestions and support. He also thanks Barbara Annicchiarico, Federico Crudu, Matthias Hartmann, Anders B. Kock, Alessandra Luati, Jan G. de Gooijer, Michele Lenza, Simone Manganelli, Francesco Ravazzolo and Paolo Santucci De Magistris for their feedback and discussions in several (un)official meetings. A special thanks to Francesca Rossi for her feedback and support in occasion of 2024 RCEA International Conference in Economics, Econometrics, and Finance in Brunel University and in numerous unofficial meeting in Bergamo when She was in life; her contagious enthusiasm for the Econometric research and teaching was fundamental for arriving at this final version of the manuscript. Part of the results in this paper were obtained by developing the codes originally written by Barbara Rossi who is gratefully acknowledged for having shared them on her webpage. The usual disclaimers apply. The Author is also in debt with doctors, nurses and workers of Policlinico "S. Matteo" of Pavia, without whose (free) cares this paper would not have been written. Finally, this paper is dedicated to the memory of Stefano Fenoaltea, without whose guidance in the earlier phase of the Author's university career none of the ideas expressed in this paper would have been possible.

"*However, from our perspective, intuition and reasoning are not so radically different [. . .] They differ in what comes to mind.*" (Gennaioli and Shleifer, 2010).

1 Introduction

The idea that a decision-maker (DM) may employ strategies to achieve her own objectives is a long-standing concern in modern Macroeconomics. According to foundational contributions by Kydland and Prescott (1977) and Barro and Gordon (1983), the monetary authority operates in accordance with the proclaimed inflation targeting in a strategic game with private market operators who build rational expectations. However, once beliefs are formed, the Central Banker has an incentive to disregard the announced objective function and surprise the market with higher inflation to exploit the Phillips Curve effect. As a result, in the following period, operators will anticipate the incentive of the FP to disregard her announcement, thus considering it not credible. Avoiding discretion becomes the logical consequence in terms of welfare maximization.

The role of individual 'intuition' and 'reasoning'—or, in one word, *judgment*, defined by Svensson (2005) as "*information, knowledge, and views outside the scope of a particular model*"—as well as their influence on economic predictions is still an open question. Judgment is pervasive also in modern Economics: in their seminal work, N. Gennaioli and A. Schleifer—in turn, inspired by seminal works by Nobel Prize winners D. Kahneman and A. Tversky—suggest that (intuitive) judgment is a combination of data and information retrieved from human memory in order to evaluate a (testable) hypothesis. Hence, the key problem: human access to information from memory is limited and selective. This implies that intuitive judgment can work effectively in many situations but falters when the representativeness and probability of scenarios diverge.

In Econometrics, where economic forecasts based on data are sought, such a divergence is relevant. Professional economists, such as bankers or financial analysts,

use limited (or, conversely, over-abundant) data from the external environment and combine it with acquired knowledge to make decisions every day. However, like other people, they may be forced by time or data limitation/over-abundance to use some information automatically and selectively. It is not infrequent that their forecasting output is directly linked to that of other colleagues; see Gallo et al. (2002). How should a DM deal with experts' opinions/judgment? A long strand of economic, statistical, and cognitive/behavioral literature deals with this question; see Figure 1 for a stylized representation of the state-of-the-art. Here, we aim to delve deeper and split the previous research question into two questions: *(i) How do these experts' judgments impact DMs' final predictions? (ii) To what extent are experts influenced by DMs' final decisions next time and vice versa?* In other words, *are these final predictions moved by strategic behavior? If so, what is the role of incentives and reputation?*

This paper addresses these questions by examining the forecasting process as a network of two agents' unobserved judgments and their reciprocal learning on repeated actions. These agents can be categorized as forecast users and producers (henceforth, FUs and FPs, respectively).[1] FP and FU have distinct roles and access to information. While the FU is familiar with the theoretical data-generating process (DGP), her/his ability to utilize this information optimally is often constrained by time pressures and the need to align forecasts with specific policy objectives. In contrast, the FP's role is to apply statistical modeling techniques, transforming raw data into forecasts that aim to minimize expected loss without necessarily conforming to policy targets. Granger and Pesaran (2000) explore this distinction using a two-state, two-action integrated decision model. Although they do not explicitly define the roles of FU and FP, they set the main road for our framework by highlighting the importance of considering the entire distribution function rather than relying on a single point estimate, especially outside the classical quadratic loss context.[2]

[1] In our language, FU is a DM.

[2] On the importance of aligning forecasts with the specific needs and constraints of decision-

Specifically, we are interested in the distinctive bias that either FPs or FUs are subject to due to their own non-sample knowledge, as well as the relationship between judgment and strategic conduct. Henceforth, we refer to the estimate obtained by FP as a *quotation* and the estimate adopted by FU after evaluating the FP quotation as an *announcement*. In both instances, it is assumed that the quotation/announcement is a combination of an optimization problem's result and a subjective component. We consider the hypothesis that the quotation does not match the announcement, which we refer to as strategic judgment bias (SJB).

Section 2 presents two examples illustrating how learning can be related to judgmental forecasting activity, specifically the FP's response to a bad FU's appraisal. Every time a representative FU makes a decision, she combines sample knowledge with her own pre-formed opinions (or beliefs), which makes the econometric treatment of these opinions non-standard. This may incorrectly identify the reaction of a typical FP to a bad appraisal of her forecast by FU. Our concern is whether standard inference can be used in a setting characterized by SJB.

In Section 3, we address this concern by introducing a new mathematical object named '*Scoring Structure*' (SS), which models a game involving multiple players using frequentist inference; see Figure 2 for a stylized representation.[3] SS incorporates utility functions[4] by both FP and FU, a sequence of quotations, announcements, an entropy function quantifying the maximum uncertainty arising from Reality, FP, and FU synergy, and a divergence function measuring the gap between FP (FU)'s quotation (announcement) and the optimal forecast.[5] To ensure fairness, general

makers, see also Timmermann (2006).

[3]The term 'scoring' refers to the commonly known 'scoring rule,' formally defined in Appendix A.1. The choice of 'structure' emphasizes that evaluating the FP/PU requires defining additional objects for testing equilibrium optimality, all related to the same FP/FU.

[4]In what follows, we prefer to use the notion of "utility" rather than that of "loss", which is more frequently used in the econometric literature, to emphasize our connection with the Bayesian literature that, despite our frequentist approach, inspired this paper.

[5]Entropy is a statistical functional acting as a link among the distribution function and its moments. Despite the entropy adopted in this paper being one of the most general, the number of these functionals currently known is huge. The mathematical conditions ensuring their feasible estimation vary according to the specific functional adopted. The strict stationarity is one of the few assumptions commonly considered as necessary to define an estimator of entropy; see Giannerini

assumptions are imposed on the game. Using SS, we can test the hypothesis that the FP's marginal utility coincides with the likelihood of FU replicating FP's quote. When data supports this relationship, we call the forecast *coherent*. Coherence testing involves designing a utility function—or scoring rule (SR)—that encourages FP to be truthful; see De Finetti (1962). However, since FP strategically uses non-sample information, honesty does not guarantee that FP will align with the estimated model or that FU will make the expected announcement. In other words, proper SRs are robust to judgment, *not to SJB*. The game-theoretic nature of SS overcomes this challenge.

Section 4 demonstrates the viability of the suggested method with a case study of U.S. survey data, where strategic bias is substantial, and discusses the relevance of the findings.

Finally, Section 5 concludes. An Appendix contains mathematical details, while a separate Supplement provides additional results

2 The Strategic Bias Problem

This Section provides two illustrative examples that motivate the use of SS with the failure of a standard forecasting scenario. Susbection 2.1 introduces the basic notation and tools; Subsection 2.2 describes two DGPs, the forecasting protocols therein involved and the emerging key facts; finally, Subsection 2.3 discusses the results.

2.1 Set-up

We focus on a simple forecast evaluation exercise based on simulated data on the U.S. Industrial Production. We assume a forecasting environment formed by two agents—FP and FU—co-exist and interact in a strategic way with Reality. Both agents share a common objective: forecasting and assessing a (conditional) predictive density.

et al. (2015).

Henceforth, we use $\doteq$ and $\equiv$ to mean equal by definition and equivalence, respectively; a $\hat{}$ to mean estimates resulting from sample information and $\tilde{}$ for labeling non-sample information; $U(\boldsymbol{\Xi}, \hat{e}_t)$ denotes the utility of the FP, which is a function of the parameter vector $\boldsymbol{\Xi}$ to be estimated and the estimated residuals $\hat{e} \doteq \hat{y}_t - y_t$; an upper-dot means optimality of the estimate; $\{f_t(\tilde{y}_t)\}_t^m$ for sequence of density forecasts; $\hat{F}_{t+h}$ for estimated cumulative density function in t+h evaluated at $\tilde{Y}_{t+h}$; $\mathcal{I}_t$ for the information up to time T. Moreover, we rely on the Diebold et al. (1998) result that if $\{f_t(\cdot)\}_t^m$ coincides with the Data Generating Process (DGP, henceforth) density $\{p(y)|\mathcal{I}_t\}_t^m$, then

$$z_{t+h} = \int_{-\infty}^{\tilde{Y}_{t+h}} \hat{F}_{t+h}(u|\mathcal{I}_t)du \;\sim\; i.i.d.\, U(0,1) \tag{1}$$

where u is a feasible measure that allows to compute the integral.[6] The PIT is computed using the fixed-rolling windows under dynamic misspecification scheme by Rossi and Sekhposyan (2019): in a set of periods {1, ..., T}, FP produces a number P of quotations obtained by using estimates of an OLS regression. Thus, there are P out-of-sample predictions to be evaluated by FU, where the first out-of-sample prediction is based on a parameter estimated using data generated by a DGP up to time R; the second prediction is based on a parameter estimated using data up to $R+1$, and the last prediction is based on a parameter estimated using data up to $R+P-1=T$, where $R+P+h-1=T+h$ is the size of the available sample and $h=1$ being the pseudo-out-of-sample horizon, so that T=264 is the "in-sample" part.[7] Once the sequence of estimation errors has been computed, it is transformed via (1).

[6] This holds under the assumptions that the predictive density has a nonzero Jacobian with continuous partial derivatives,

[7] For ease of exposition, we assumed a quadratic loss.

2.2 Two Simulated Examples

2.2.1 Unknown constant mean N.I.D. process

Consider the case of a set of {1, ..., i, ..., 1000} random paths of length $T = 265$ from the following DGP (identified with the index "1"):

$$\text{DGP1:} \quad y_{1t}^{(i)} = \mu^{(i)} + \epsilon_{1,t}^{(i)}, \quad \epsilon_{1,t}^{(i)} \sim N(0,1), \tag{2}$$

that is, a simple stationary i.i.d. process with unknown mean. Since the error variance is inversely proportional to the forecaster's utility in many SRs, the unit variance error indicates the DGP is perfectly focused around a Normal distribution, so that no penalty is payed by FP in the case of a poor quotation.

This interaction among FP and FU, given this DGP, is summarized in four steps:

1. At time T of the i-th random draw, FP makes her own forecast $Q_{T+1|T}$ on the conditional density for the next period given all the current information.[8] When the closed form is not available, the conditional density comes from a MonteCarlo simulation by looking at the vector of observables $\{y_t\}_{t=1}^{t}$ adopting (6) and the utility function the $U(\cdot)$—say, the quadratic utility, $U(\boldsymbol{\Xi}, \hat{e}) = \hat{e}(\boldsymbol{\Xi})^2$. In this step, FP does not use judgment, so that $Q_{T+1|T}$ is the predicted value of a direct estimation.[9]

2. At time $T+1$ of the i-th random draw, FU evaluates FP's forecast using data up to T. For ease of exposition, we assume that (i) FU knows the FP's DGP corresponding to (2); (ii) FU does not know if FP has used judgment to arrive at $Q_{T+1|T}$. Thus, FU needs to test whether the realized values $Q_{T+1|T}$ are effectively generated by

$$\tilde{y}_t = \hat{y}_t^{(i)} + \tilde{\pi}_1, \tag{3}$$

[8] We use the notation Q to mean "quotation", the definition of which is not essential in this stage of the analysis.

[9] Here, an OLS regression of the dependent variable on the regressors lagged $h-$periods.

where $\hat{y}_t^{(i)}$ is the estimated value of the variable, $\tilde{\pi}_1$ represents the judgment that FU suspects in FP's forecast and is assumed being small (relatively to $\hat{y}_t^{(i)}$) and set to 0.05 to mean that the role of judgment is limited; finally this last element has to be interpreted as a shift in the mean of the distribution. We remark that $\tilde{\pi}_1$ *is not* part of the DGP because of its inherently subjective nature; thus, its exact computation is highly implausible. This requires to compute the probability integral transform (PIT) of $\tilde{y}_t$. Under optimal forecast, the histogram of the estimated PIT is perfectly rectangular. But $\tilde{\pi}_1$ makes the one-step-ahead histogram (corresponding to Step 2 in the upper panel of Figure 4) a non-perfect rectangle. Since FU knows the DGP adopted to produce $Q_{T+1|T}$, FP has now insight that FU is mis-evaluating her forecast, although the magnitude of the mis-evaluation is small. However, FP does not know where FU is wrong.

3. At the same time, FP makes forecasts for $T+2$ using data up to T. That is, FP does not update her informations and does not learn from the FU's evaluation. This implies that she limits herself collecting new random draws from true distribution and performs under a static DGP where

$$\tilde{y}_t^{(i)} = \hat{y}_t^{(i)} + \tilde{\psi}, \tag{4}$$

with $\tilde{\psi}$ small (still set to 0.05) and all other parameters of $\hat{y}_t^{(i)}$ remain the same. This means that FP adds judgment to her original DGP regardless of what FU is supposed to do in $t+2$.

4. Finally, at $T+2$, the FU evaluates FP's new forecast $Q_{T+2|T}$. Again, FU knows FP's DGP, not the amount of judgment incorporated by her in $T+1$ to arrive at $Q_{T+2|T}$. Thus, FU needs to test whether the realized $Q_{T+2|T}$ is effectively generated by

$$\tilde{y}_t^{*} = \tilde{y}_t^{(i)} + \tilde{\pi}_2, \tag{5}$$

where $\tilde{\pi}_2$ represents the judgment that FU suspects being in $Q_{T+2|T}$. Still,

$\tilde{\pi}_2 = \tilde{\pi}_1$, for ease of illustration and without loss of generality; hence, like $\tilde{\pi}_1$ and $\tilde{\psi}$, it is not part of the DGP.

Key Result 1. The lower panel of Figure 4 displays the histogram of PIT with $h = 1$ computed in Step 4 according to the same rolling-windows scheme explained above. Differently to the equivalent one in Step 2, the effect of the misspecification increases dramatically. Such a misspecification cannot be imputed to the FP—she has no estimation bias, and the error distribution is perfectly centered around a standard Normal—nor to FU—since she knows FP's DGP—*if these agents are considered singularly*. Thus, it must be imputed to π_1—originated and measured by the FP in Step 2—and to π_2—originated and measured only in Step 4. However, the lack of dynamics in the information updating and learning may be in turn looked at as a cause of misspecification. The next example will consider a more realistic DGP.

2.2.2 LSTAR(2)

Now, consider the following DGP (identified with index "2") with the same number of random paths and sample length:

$$\begin{aligned} \text{DGP2:} \quad & y_{2t}^{(i)} = 0.9y_{2,t-1}^{(i)} - 0.795y_{2,t-2}^{(i)} + (0.02 - 0.4y_{2,t-1}^{(i)} + 0.25y_{2,t-2})^{(i)} G^{(i)}(\boldsymbol{\Xi}) + \epsilon_{2,t}^{(i)}, \\ & \epsilon_{2,t}^{(i)} \sim N(0, 0.5), \end{aligned} \tag{6}$$

where: $G(\boldsymbol{\Xi}) = \left(1+\exp\left\{-\gamma(s_t - c))\right]\right\}\right)^{-1}$ is a function of the known parameter vector $\boldsymbol{\Xi} = [\gamma, c, s_t]$, formed by a slope $\gamma = 5.0$ that governs the transition between the two extreme states $G = 0$ and $G = 1$, a location parameter $c = \bar{y}_t$, and the transition variable $s_t = (y_{t-1} - \bar{y}_t)$, with $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T} y_t$.[10] Since the error variance is inversely proportional to the forecaster's utility in many SRs, the value $\sigma^2 = 0.5$ increases the potential cost of bad quotation. The model (6) is the (Logistic) Smooth-Transition Autoregressive (LSTAR) model introduced by Chan and Tong (1986). This family

[10] This DGP is similar to the one used in Teräsvirta (1994), equation (4.1).

of models is one of the most widely adopted in model production indexes. Moreover, it is particularly indicated to nest the forecasting framework and coherence testing that motivates this paper.[11]

Let us now consider a forecasting environment formed by FP and FU. This interaction is summarized in four steps:

1. The first step coincides with the one in the previous example, thus it is omitted.

2. The second step coincides with the second one in the previous example apart the notational variation in the judgmental term

$$\tilde{y}_t = \hat{y}_t^{(i)} + \tilde{\pi}_1', \tag{7}$$

   where all that concerns $\hat{y}_t^{(i)}$ and $\tilde{\pi}_1'$ is the same as in the previous example and the change in notation is just to avoid confusion.

3. FP has learnt to be mis-evaluated by FU and, at the same time, has to make the new forecast for the next period (according to our notation, $Q_{T+2|T+1}$), where

$$\tilde{y}_t^{(i)} = \hat{y}_t^{(i)} + \tilde{\psi}', \tag{8}$$

   with $\tilde{\psi}'$ small (still set to 0.05) and all other parameters of $\hat{y}_t^{(i)}$ remain the same. This means that FP adds judgment to her original DGP to account for the FU mis-evaluation and, thus, to anticipate another possible mis-evaluation also in $t+2$.

4. Finally, in time $t+2$ FU evaluates FP's new forecast $Q_{T+2|T+1}$. Again, FU knows FP's DGP, not the amount of judgment incorporated by her in $T+1$

[11] In particular, in next Section we will argument that a strong nonlinearity can be associated to significant SJB—conversely, insignificant SJB means the process is linear. Thus, the STAR-modelling has to be considered as instrumental to avoid ad-hoc refinements that would make the entire methodology less intuitive. Finally, the assumption that the vector parameter $\boldsymbol{\Xi}$ is known allows the reader to focus on the SJB as the main source of uncertainty with respect to other sources (like parameter uncertainty or measurement error).

to arrive at $Q_{T+2|T+1}$. Thus, FU needs to test whether the realized $Q_{T+2|T+1}$ is effectively generated by

$$\tilde{y}_t^* = \tilde{y}_t^{(i)} + \tilde{\pi}_2', \tag{9}$$

where $\tilde{\pi}_2'$ represents the judgment that FU suspects being in $Q_{T+2|T+1}$. Still, $\tilde{\pi}_2' = \tilde{\pi}_1'$, for ease of illustration and without loss of generality; hence, like $\tilde{\pi}_1'$ and $\tilde{\psi}'$, it is not part of the DGP. Now $\tilde{y}_t^*$ includes several judgmental components: $\tilde{\pi}_1'$, which is incorporated in $\tilde{y}_t^{(i)}$), and $\tilde{\pi}_2'$. These cumulate, and thus FU computes the one-step-ahead PIT for t+2 corresponding to (1)—with minor modifications to the bounds of the integral to take into account for the new period.

Key Result 2. The lower panel of Figure 5 displays the histogram of PIT with $h = 1$ computed in Step 4 according to the same rolling-windows scheme explained above. Differently to the equivalent one in Step 2, the effect of the misspecification increases dramatically. Still, such a misspecification cannot be imputed to the FP—having no estimation bias—neither to FU—because she knows FP's DGP—*if these are considered singularly*. This time it must be imputed to the perduring effect of π_1', which has been originated by the FP learning in Step 2, but is measured only in Step 4.[12]

2.3 Discussion

The focus on density forecasting for both FP and FU justifies the need for a non-standard computation of SJB. Under point forecasting, where FP provides a specific forecast like $p_{t+1|t}$, if FU knows the DGP, she can compute bias π_1 (or π_1') as a simple forecast-realization difference. However, this method does not reveal the source of

[12]This example—as well as this paper—does not consider any dynamics of the learning process that FP benefits from in the transition from Step 2 to Step 3 and only aims to illustrate the SJB as a (potentially) self-exciting phenomenon that may cause a systematic failure of traditional diagnostics. Moreover, it does not aim to quantify directly $\tilde{\pi}_1$, $\tilde{\psi}$, or $\tilde{\pi}_2$. This simplistic assumption is made for ease of statistical treatment. The dynamical extrapolation of (non-strategic) judgment is discussed in Zanetti Chini (2023). Issues in the dynamic estimation of strategic judgment constitute the next step of this research.

misinterpretation. In contrast, density forecasting requires a theory to explain SJB's origin.

Assuming that FU knows the DGP may seem counterintuitive as it implies perfect specification, contradicting the need for additional information sources. So, why should FU suspect FP's output is 'wrong', and thus add judgment? Answering this question requires an economic theory of information. We consider the idea that forecasting agents are imperfect maximizers facing learning costs linked to nonlinearity in FP's policy function (related to utility function asymmetry) and irregular belief updates, as supported by recent empirical literature (Manzan, 2011, 2021; Ilut and Valchev, 2022).

In a subjectivist perspective, we also consider that FP may be influenced by indirect signals, such as reputation, as shown by Ottaviani and Sorensen (2006). According to these authors, reporting the best state quotation is not an equilibrium in a forecasting tournament with pre-specified rules, while it involves balancing two forces: (i) the incentive to report an honest forecast; (ii) the gain from deviating from the consensus. In our scenario, a simple repeated game between FP and FU may lead FP to emphasize her private signals (that is, SJB) over honest reporting, despite FU's actions; see the Supplement.

The assumption that FU can observe FP's DGP is for illustrative purposes and may be relaxed. In this case, FU must deal with estimation errors *in addition* to the previously considered strategic bias. It can be shown that, in this case, FU only needs to know the form of $U(\cdot)$ to select an appropriate SR for PIT computation. This SR takes into account either $\tilde{\pi}$ or $\tilde{\pi}_2$ (or π_2') *one at a time only*. Hence, FU may still achieve a PIT similar to the one shown in the lower panel of Figure 5 by using the best SR for evaluation. The next section aims to address this issue.

3 Theoretical framework

This Section introduces the econometric theory of SJB that generalizes the example illustrated above. Namely, Subsection 3.1 introduces the notation; the repeated game that constitutes the fundament of the new framework is explained in Section 3.2; Subsection 3.3 defines and characterizes the notion of coherence; finally, Subsection 3.4 introduces a formal test for the null hypothesis of forecasting coherence.

3.1 Notation

We are interested in the stochastic process $Z \doteq \{Z_t : \Omega \to \mathbb{R}^{k+1}, k \in \mathbb{N}, t = 1, \ldots, T\}$. This process is partitioned as $Z \equiv [Y, \hat{Y}_t, \tilde{Y}_t, X_t]$, where $Y_t : \Omega \to \mathbb{R}$ is the vector of variable of interest, $\hat{Y}_t : \Omega \to \mathbb{R}$ is the vector of optimal estimates, $\tilde{Y}_t : \Omega \to \mathbb{R}$, is the vector of (potentially biased) estimates corresponding to the $\tilde{y}_t$ in (3) and $X_t : \Omega \to \mathbb{R}^k$ is the vector of exogenous variables.[13] Let be Z_t defined on a complete probability space $\{\Omega, \mathcal{F}_t, P\}$, where Ω is the sample space; $\mathcal{F}_t = \sigma(Z_1, \ldots, Z_T)'$ is the information set to t, partitioned as $\mathcal{F}_t \equiv [\mathcal{F}_t^{\Pi}, \mathcal{F}_t^{\Psi}]$ to denote the sub-spaces of FU and FP, respectively; and P denoting a probability measure. We aim to estimate the true, but unknown, h-step-ahead predictive density for variable Z_{t+h}, with $1 \leq h < \infty$, based on $\mathcal{F}_t$ denoted as $p(Z_{t+h})$, while its distribution is $P(Z_{t+h})$, respectively.

The estimates of Z are defined as follows: $\hat{Z}_t = f_t(Z_t, Z_{t-1}, \ldots, Z_{t-m+1}; \hat{\boldsymbol{\theta}})$, and $\tilde{Z}_t = f_t(Z_t, Z_{t-1}, \ldots, Z_{t-m+1}; \tilde{\boldsymbol{\theta}})$, where the k vector $\hat{\boldsymbol{\theta}}$, collects all the estimated parameters, $\tilde{\boldsymbol{\theta}} = \alpha\hat{\boldsymbol{\theta}} + (1-\alpha)\boldsymbol{\theta}^{NS}$, and $\boldsymbol{\theta}^{NS}$ the non-sample based estimates, $\alpha \in [0, 1]$ the weight of data-driven based estimates; f and g are any measurable functions.[14] These estimates are built under the assumption that both FP and FU partition the available sample of size $T + h$ into an in-sample portion of size R and an out-of-sample portion of size P and obtained a sequence of h-step-ahead out-of-sample

[13] This notation aims to emphasize the link with previous Section 2. In some parts of this section, the difference between Y, $\hat{Y}$ and $\tilde{Y}$ is not essential—and the last two coincide under some conditions, as will be clarified in what follows. For this reason they will be denoted as Y in the course of this section. The Supplement provides an exact geometry of the equilibrium among these variables.

[14] According to this, $\hat{\theta}$ and $\tilde{\theta}$ coincides with $\alpha = 0$

density forecasts of the variable of Y_t using $\mathcal{I}_t$ and the set of all possible judgments—denoted J_t—such that $R + P - 1 + h = T + h$ and $\mathcal{I}_t^i \subseteq \mathcal{F}_t^i$, and $J_t^i \subseteq \mathcal{F}_t^i$ with $i = \{FP, FU\}$.[15] These P out-of-sample estimates of conditional predictive densities evaluated at the ex-post realizations (denoted by $p(y_{t+h}|\mathcal{I}_t)$) depends on information set via parameter $\hat{\Theta}_{t,R}$; when they depends also on judgment up to time t, they are defined as $p(y_{t+h}|\mathcal{I}_t, J_t))$ and will lead to parameter $\tilde{\Theta}_{t,R}$, with the same partition established previously for Θ, J_t and I_t holds. For each parameter sequence we assume a rolling scheme: $\hat{\Theta}$ and $\tilde{\Theta}$ are re-estimated at each $t = 1, \ldots, R$ over a window of R data spanning from $t - R + 1$ to t.

The (estimated) Log-likelihood of FU and FP are denoted, respectively, $\mathcal{L}(\hat{\Theta}^{\Pi}; z)$ and $\mathcal{L}(\hat{\Theta}^{\Psi}; z)$. Finally, let J^* the subset of optimal judgments, that is value(s) of J that maximizes the expected utility (that is, the Scoring Rule), denoted as $S := \int U(\hat{P}, \mathrm{j}^*)p(\hat{Y}_t)d\hat{Y}_t$, computed using the distribution P, which is believed to be the true DGP.

The FP seeks to solve a decision problem defined by the triple $\left\{\hat{Y}_t, J^{FU}, U(\hat{p}, \mathrm{j}^*)\right\}$, where: $\hat{Y}_t$ and J^{FU} are defined as before; and $U(\hat{p}, \mathrm{j}^*)$ is a real-valued utility function that represents the reward obtained by the FP as the result of minimizing the distance between density forecast for $t+1$ and data that will be observed at that time by her own optimal judgement $j^* \in J^{FU}$). Let denote this peculiar utility function as S, a proper SR. The same holds for FU, with minor modification to the notation. The divergence among $S(\cdot)$ and $U(\hat{p}, \mathrm{j}^*)$, that is among any scoring rule and the utility function give the optimal amount of judgment can be denoted as D. Mathematical characterization of these function are explained in the next Sub-sections.

[15] Note that this implies that both FP and FU observes a subset of the true information set but may compensate such lack with judgment; this last partition $J = [J^{\Pi}$ and $J^{\Psi}]$ to emphasize the role of FP and FU; similarly, the same partition hold for the parameter space $\Theta \equiv [\Theta^{\Pi}, \Theta^{\Psi}]$.

3.2 The Forecasting Game

We assume that the probabilistic forecast of an economic event is the output of a one-period game with three players: the FP; the FU who has capital K to preserve; and Reality. The FU suspects the FP's quotations are biased and, eventually, cooperates with Reality;[16] however, no matter how the FU plays, Reality acts as though the FU does not win the game. This rule, called "*Excluded gambling system hypothesis*"or *Cournot's Principle*", is necessary to avoid an unbalanced game in favor of FU.

Cournot principle is strictly related to the notion of internal strategy. The latter is a rule governing player's motion law on each round based on the previous moves by the other players, using only information that is internal to the game. We remark that the J is also internal to the game, so that the use of judgment as additional source of information for each player do not constitute a problem in this regard. An internal strategy for Sceptic is said *legal* if it respects the condition that FU moves so that his capital always remains non-negative, no matter how the other players move. The rationale of Cournot's principle is simply the foundation of econometric forecasting: Reality has to be the player who largely determine the output of the game and, in practice, ensures that a property E (say, a measurable function) of the sequence of events x realizing with a certain probability p denoted as $\{p_i x_i\}_{i=1}^{N}$ (for N large enough) happens almost surely if FU has an internal strategy that wins the game whenever the actual sequence fails to satisfy E. Vovk and Shafer (2001) shows that existence of such a legal internal strategy is sufficient to re-express in a game-theoretic sense some of the basic laws of Kolmogorov: an event happens almost surely sense if and only if it has probability 1 and, conversely, it do not happens if and only if it has zero probability; see the discussion in Vovk and Shafer (2005, p. 751). Thus, characterizing internal strategies is mandatory to understand the final outcome of the game. SR functions (denoted S) has exactly this objective and constitutes the link among utility and probabilistic game framework.

[16] Cooperating with reality has to be intended in the sense that there exists an internal strategy for Reality that ensures that FU wins; see Theorem 2 in Vovk and Shafer (2005, p. 750).

The two Players have different objective functions:

$$\max \mathbb{E}\left[S(Z_t; \Theta^{FP})\right] = \max_{\theta_t} \mathbb{E}\left[f(z_t, \theta_t^{FP})\right], \tag{10}$$

$$\max \mathbb{E}\left[S(Z_t; \Theta^{FU})\right] = \max_{\theta_t | \bar{\theta}_t} \mathbb{E}\left[f(z_t, \theta_t^{FU}, \bar{\theta_t}^{FU})\right], \tag{11}$$

where the notation has been previously set, except for the upper bar $\bar{}$, which denotes the target value. That is, the two agents face the same data and the same kind of model parameter, but while FP has to maximize some scoring function, FU has to do the same given a target constraint on the parameter vector θ_t. Examples of $f(\cdot)$ to maximize are a quadratic function for FP and the square of the difference between actual and target inflation and real growth for FU; see, for example, Clarida et al. (1999). In other words, we allow that $\Theta^{FP} \neq \bar{\Theta}^{FU}$. This difference is assumed to be the ultimate source of SJB. Indeed, the Players act according to the following *Forecasting Protocol*:

1. $K_0 := 1$;
2. FU announces a bounded function $S : \mathbb{R} \to \mathbb{R}$;
3. FP announces her (potentially biased) quotation $\hat{p}(X|\mathcal{I}_t, J_t) \in \mathbb{R}$;
4. Reality announces a draw from $P(Y) \in \mathbb{R}$;
5. $K_1 = K_0 + D\ \big(\mathrm{Y}, \mathrm{X}\big)$,

where S is a proper SR and $D(\cdot)$ is a divergence function among predictive density of X and predictive density of Y whose properties are specified below. FU must choose U so her capital remains non-negative ($K \geq 0$) no matter what values the FP and Reality announce for $\hat{p}(X)$ and $P(Y)$. The winner is the FU if $K_1 >> K_0$. Otherwise, the FP wins.

The game illustrated here is a modified version of the "Forecasting sub-game" by Vovk and Shafer (2005, p. 753). With respect to these authors, to ease the statistical

treatment we avoid the recursion corresponding to the $n \geq 1$ times that the game is re-iterated. This simplification can be removed by assuming an algorithm that ensures that the main restrictions and assumptions regarding the players hold for each recursion.

Step 2 of the Protocol is an application of one of Patton (2019)'s main conclusions. Namely, he demonstrates that utility-based objects like the forecast rankings are generally sensitive to the choice of a proper SR and asserts that FPs should be told ex-ante what utility functions will be used to evaluate their quotations.[17]

Step 5 of the Protocol is the basis for a test for the null hypothesis of forecasting coherence in terms of the FU's utility. The form in which the test is written implies that the FP's reward cannot be augmented after his quotation. Subsection 3.4 and Supplement shows that this can be implemented via standard LM statistics with standard χ^2 asymptotics. This coherence test is essentially based on the $D-$function, which is a Brégman divergence:

$$d(X,Y)^{Bregman} \doteq \int y\Big\{\Big(f\big[p(y)\big] + \big[p(y)-p(x)\big]f'\big[p(x)\big]\Big) - f[p(y)]\Big\}d\mu, \qquad (12)$$

where $f(\cdot)$ is a (strictly) concave function and $f'(\cdot)$ a subgradient of $f(\cdot)$.[18] This is a very general class of non-metric distance, introduced by Brègman (1967) and subsequently applied by Savage (1971) to elicit forecasters utility and capable to characterize most of the SRs described in Table 1 of the Supplement. Necessary and sufficient conditions under which $D(\cdot,\cdot)$ admits a Brègman-Savage representation are provided by Hendrickson and Buehler (1971). We are particularly interested in the special case that

$$f(z) = k(z) - \lambda \log(z), \qquad (13)$$

[17] "[...] *Specifying the target functional is generally not sufficient to elicit a forecaster's best (according to a given, consistent, loss function) prediction. Instead, forecasters should be told the single, specific loss function that will be used to evaluate their forecasts*" Patton (2019, p.3).

[18] Notice that the use of y and x here is just for ease of notation and the same definition holds for all the (couples of) partitions of Z.

where k, known in Physics as 'Bolzmann's constant', represents the marginal cost of a unit of information and is set to zero without loss of generality. Under (13) the forecasts generated by $\mathcal{M}$ are coherent with a given SS. Two peculiar values of $S(\cdot)$ are S(Y,Y) and S(X, X) (that is the SR of Y evaluated in Y), corresponding to the Entropy function of the system in X or Y. The relationship among D_B, $S(\cdot)$ and $H(\cdot)$ is well known in functional analysis by Schervish (1989) contribution and re-popularized by Gneiting and Raftery (2007) to which we refers for details. Schervish representation is a generalization of the original Brègman-Savage one and is necessary to build the formal test on the hopothesis of forecast coherence. Figure 3 summarizes these results known as "Schervish Representation".

In principle, the assumption that Reality can cooperate with the FU implies that, when the game is repeated n times, the sequences of outcomes S_n Y_n, X_n do not necessarily coincide with realizations of a stochastic process. As a consequence, classical hypothesis testing and inference is ineffective and should be substituted by another type of inference which explicitly accounts for strategic behavior, see Olszewski (2015) for a theoretical discussion of this problem. Nevertheless, Shafer and Vovk (2001, Chapter 8.1) ensure that the Cournot's Principle allows both of them to be used.

3.3 Representation of the Forecasting Environment

The De Finetti's coherence principle is a set of theorems showing that a decision maker's probability assignment $P : \mathcal{F} \to \mathbb{R}$ over a set of events $\mathcal{F}$ is a probability measure if and only if it does not expose the decision maker to Dutch books. A more formal definition is the following (Mundici, 2009):

Definition 1. Let be $E = \{Z_1, \ldots, Z_m\}$ a finite set of events, $w(Z_i)_{i=1}^m \in [0,\ 1]$ a state of world acting on E that assigns value zero-one to Z_i, so that the set of all states of world $W = (w(Z_1), \ldots, w(Z_m)) \in [0,\ 1]^{[1,\ldots,m]}$ lies in the unit m-cube; and a stake function $\sigma \in \mathbb{R}$. Then, a map $f : E \to [0,\ 1]$ is said to be W-coherent (on E)

iff for no $\sigma_1, \ldots, \sigma \in \mathbb{R}$ one has $\sum_{i=1}^{m} \sigma(f(Z_i - \upsilon(Z_i)) < 0 \; \forall \; \upsilon \in W$. Otherwise, f is W-incoherent.

Thus, according to this definition, the mathematical condition for a probabilistic forecast to be coherent is finitely additiveness. See De Finetti (2017, Ch. 3.1, pp. 62–64). However, despite its generality has enabled a series of mathematical as well as epistemological consequences still under debate (Dawid, 1982; Regazzini, 1987; Pelloni, 1996; Nau, 2001; Dawid, 2004; Predd et al., 2009), the operational definition of coherence has not been considered explicitly by the economic literature.

Consider the preliminary theory in A.1. Then, it's possible to arrive at the following

Definition 2 (Structural Coherence)**.** The (h-step ahead) density forecast y_{t+h} obtained by $p(\hat{\Theta}^{\Pi}; z_t) \in \mathcal{P}$ is *coherent relatively to the scoring structure* (or *structurally coherent*) if there is one-to-one mapping f between the estimated LogLikelihoods of FU and FP $\mathcal{L}(\hat{\Theta}^{\Pi}; z_t)$ and $\mathcal{L}(\hat{\Theta}^{\Psi}; z_t)$.

This last definition is based on (i) the existence of a market where the demand and supply of economic forecasts match; (ii) a game-theoretic approach to forecasting.[19] Secondly, our definition extends Definition 1 because it is based on an equivalence among two estimated objects and links the coherence to the likelihood principle—that is, the idea that all the relevant information in a sample is contained in the likelihood function. The non-manipulability of tests based on Likelihood-Ratio Tests in strategic forecasting is proved by Pomatto (2021). Our game-theoretic foundation avoids unfair evaluation in case agents move away from optimal actions.

The relationship among the two definitions is ensured by the following

Proposition 1. *Any structurally coherent model is also coherent in De Finetti sense.*

The proof is trivial if considering that (i) both the definitions are based on a mapping; (ii) inversion is always a mapping; (iii) the inverted map between the

[19]The term "structural" emphasizes the role of the two demand/supply sides constituting the structure of any exchange-based economy.

Log-likelihoods (that is conditional probability) of FP and FU is just an exponential rescaling of a probability; (iv) the inverse mapping among probability models $p(\Theta^{\hat{FU}})$ and $p(\hat{\Theta}^{FP})$ implies there are no configuration of the probabilities leading to certain loss. The Supplement gives a detailed proof.

Proposition 2. *The FP's reward $S(X,Y)$ is a proper SR if and only if A1 – A5 are satisfied. The same holds for the FU side if inverting the order of the variables.*

Proof. This is essentially the Theorem 1 in Gneiting and Raftery (2007). □

The next result identifies the testable hypothesis of forecasting coherence and constitutes the basis for the rest of the analysis:

Proposition 3. *Let $S(z,p)$ be an SR, possibly the Brègman-Savage representation, with q-function $\mathfrak{s}$. Then, $S(z,p)$ is local and strictly proper if and only if $\mathfrak{s}$ is such that:*

$$\mathbb{L}\mathfrak{s} = 0, \tag{14}$$

where: $\mathbb{L} := \sum_{k\geq 0}(-1)^k \mathbb{D}^k p_0 \frac{\partial}{\partial p_k}$, $\mathbb{D} := \frac{\partial}{\partial y} + \sum_{j>>0} p_{j+1}\frac{\partial}{p_j}$, $\mathbb{D}$ and $\mathbb{L}$ are total derivative and linear differential operators, respectively.

Proof. This is essentially the condition (i) in Theorem 6.4 in Parry et al. (2012). □

Equation (14) is called the *Key Condition*. Two further (merely theoretical) conditions concerning the representation of $\mathfrak{s}$ via Lagrange operators are required to prove the mentioned Parry et al.'s Theorem. Nevertheless, the Key Condition is sufficient—and, to the best of our knowledge, it is the only available—to identify a testable hypothesis of the logarithmic form of the FP' and FU's utility.

The connection between forecast coherence and the locality is ensured by the following

Theorem 1. *A density forecast $p(Z_{t+h})$ is structurally coherent if and only if $S(z,P)$ is local.*

Proof. See Appendix A.2. □

3.4 Testing for structural coherence

To test the hypothesis that equation (14) is verified by the data, we assume that $S(z;p)$ is part of a smooth-transition autoregressive scoring structure with exogenous regressors (SS-STARX, henceforth) and is treated as an observed transition variable; for a comprehensive treatment of traditional smooth-transition regression models, see van Dijk et al. (2002). This treatment is necessary to set up the null hypothesis and introduce an LM-type test using a linearization of the SS-STARX, which is equivalent to an auxiliary model with augmented regressors, the number of which depends on the type of non-linearity of the structure. Since the statistical treatment of this methodology relies on well-known principles, we leave details in the Supplement.

Our MonteCarlo simulation, also exposed in the Supplement for economy of space, shows that the test is generally well-behaving. However, several issues can be observed—to our knowledge, for the first time. One of them is the so called *Score Invariance*: the form of the SR function does not modify the empirical size and power of the test. Moreover, the dynamic misspecification tests therein discussed, differently from the original test proposed for "simple" STAR models by Eitrheim and Teräsvirta (1996) are less clear and this implies the need to disentangle the roles of SJB and simple dynamic misspecification in the forecast evaluation stage. The presence of dynamic misspecification (measured and tested separately from coherence) means that the objective function of FP may not be separable from the FU's one whereas the coherence test implicitly assumes they are.[20]

4 Empirical Application: The U.S. survey data

This section applies the SS and coherence testing to real data. Subsection 4.1 describes the data and results; Subsection 4.2 describes their relevance and effects on econometric literature; finally Subsection 4.3 provides links to Central Banking.

[20] See Komunjer and Owyang (2012) for the case of point forecasting.

4.1 Results

We consider a case study on the GDP and unemployment forecasts by the Survey of Professional Forecasters of Federal Reserve (SPF, henceforth); two other examples on U.K. and Norway's data are discussed in the Supplement.

The Federal Reserve Bank has been the first institution to use professional forecasting to justify its policy decisions. The data collection began in 1968 as an independent study of the American Statistical Association and NBER and in late 1990 its maintenance became part of the FED's institutional activity. This change in the managing institution makes the coherence of the SPF of the U.S. economy particularly interesting to test. The data can be downloaded at the Real-Time Data Research Center of the Federal Reserve of Philadelphia at https://www.philadelphiafed.org/surveys-and-data/data-files. The release is dated November 16th, 2020. In this paper we consider three variables: the Real GDP (RGDP), the GDP Deflator (GDPD) and the unemployment rate (UNR), for an horizon going from one to four quarters; the nowcasts are also investigated. The sample span is 1975:Q1–2020:Q4.[21]

According to Tables 1—3, there is a strong evidence of SJB in all the variables. However, its length is heterogenous: if the full sample is considered, the null hypothesis of no SJB in the forecast of RGDP is rejected in the first lag in all the horizons, albeit this is the only rejection case in eight nowcasts as well as in first step-ahead forecasts. The two-quarters forecasts are strategically biased also in the fourth lag. Differently, the three and four-quarters-ahead forecasts are biased in the majority of cases. The GDPD tends to reject the null homogeneously in the first 2/4 lags. On the other hand, UNR is pervaded by SJB in all the horizons. Interestingly, this finding is less evident if considering only the 'FED-managing' sample 1991:Q1–2020:Q4. Apart the first lags in nowcast and all step-ahead-forecasts, the only case in which SJB cannot be neglected is the one-year-ahead forecast. In UNR forecasts the SJB is still evident in the majority of cases.

[21] The first 24 observations have been discarded to avoid non available answers occurring occasionally in that period.

4.2 Relevance and Consequences

The lack of structural coherence described in Section 3 is endemic in macroeconomic data, at least those reported by the major central banks. Several ramifications follow from our examples: first, the SJB hypothesis promotes model-based forecast comparability. In fact, while the SJB cannot be rejected in a third of the lags, despite a confidence level of 10%, the equal predictive ability test is passed in one-third of the SRs adopted; on the other hand, while the SJB hypothesis is rejected in almost all of the lags, the equal predictive hypothesis is globally acceptable. This is directly related to the recent discovery by Galvao et al. (2021) that judgment tends to reduce the accuracy of pure statistically driven models. Furthermore, the number of SRs that fail to produce a winner is notable. This additional difficulty is a result of Patton (2019)'s findings, which reveal that selecting the form of the SR *ex-ante* does not guarantee coherence. Second, as explained in the Supplement, the score invariance principle complicates objective elicitation of the genuine FP's utility, which explains the difficulties in recognizing a winner when performing forecast comparisons. As a result, the economic nature of such score invariance should be questioned.

Answering this question requires developing a theory that describes the dynamics of the predicting game's players and, ultimately, the foundation of the SS. This is supported by Ilut and Valchev's recent model of costly cognitive decision making. According to these authors, FP devotes her time and effort to generating a pricey quotation by meticulously observing all relevant objective state factors. The true policy function corresponding to the scoring function utilized by FU in her evaluation, on the other hand, remains unknown. The uncertainty in the 'true' SR is computed using Bayesian nonparametric methods based on the Gaussian Process distribution, which FP uses to update her views. As a result of the underlining situation, the FP gradually acquires information regarding the optimal quotation. Such an accumulation (i) reduces the variance of the judgment while simultaneously (ii) increasing the system's entropy required to make the decision; as a result, (iii) it only partially

explains the optimal quotation at a different state realization; as a result, (iv) it leads to judgment propagation in all forecasting processes. The combination of (i)-(iv) explains the endemic incoherence and the large range of SRs estimates (our two most major empirical findings), but not the issue of score invariance.

This latter point can be seen as a result of a stationary covariance function that governs the correlation between thoughts about the (unknown) SR's values. When this association is imprecise, FP's knowledge of the textquoteleft true' SR is more valuable in the vicinity of the state realization, where learning occurs. As a result, the uncertainty over the SR is smaller in situations where learning is more intensive than in states where such learning is more rare. This inconsistent reasoning by both FP and FU leads the SJB to oscillate, proving that it is not recognized by the SR—even if it were correct.

The significance of these implications necessitates the following caveats: First and foremost, our SS technique is distinguished by a simple parametrization. Only two homogeneous, representative agents are expected to populate the forecasting environment. Because of this assumption, we were able to simplify the graphical and mathematical representation of the SS. We are aware that real economics is less simple, with at least two potential sources of added complexity on both the demand and supply sides: for example, in the United States, (a) the FED-SPF is used by any enterprise or investor interested in macroeconomic forecasting, not only the FED; (b) On the other side, the FED supplements SPF with its own internal forecasting model (s).

Concerning (a), increasing the complexity of our methodology entails discarding the homogeneity assumption and allowing the amount of private information to be unequal among the (potentially enormous) number of players. Unless we assume that the FED's internal forecasting model and SPF (specifically, that SPF and the collection of institutions that use it) have the same information processing dynamics, in which case the additional FP's outputs can be easily represented in an exogenous

vector $\mathbf{z_t}$ our current parametrization is insufficient to achieve this goal. A first step towards addressing this issue could be the use of networks; see, for example, Bramoullé et al. (2014). Concerning (b), a second stochastic term (say, ζ_t) with appropriate distributional assumptions should be added to equation (3), such that the PIT in the first panel of Figure 1 is drastically different—perhaps similar to the PIT in the second panel because the model is easily misspecified. As a result, these additional phrases will infect the subsequent processes. As a result, the game in Section 2 of the supplement should be correctly changed to accommodate it. In this sense, Vovk and Shafer (2005, p. 754) provide a more complex forecasting game: this final one includes a third player (four if Reality is considered one of them), the Random Generator (RG), which can be interpreted as an econometric division and is supposed to act in parallel with FP. FU then properly averages the results of FP and RG. Again, this necessitates an additional step and a more robust parametrization than the simple equation (1) of Supplement.

A second issue pertains to Step 2 of the Forecasting methodology, which assumes the SR is known to use the features of the STARX models and hence avoids using unobserved component modeling. In fact, no one has ever explored the representation of the SS in an unobserved state-space model. Furthermore, we ignored it on purpose in order to examine a dynamic in the repeated game for the economy of space. This means that we can't yet validate recent findings on the quality of SPF under- and overreactions; see, for example, Bordalo et al. (2020). Third, we hypothesized that, because the researcher will select both the scoring rule and the coherence test function, the FP's objective function will be among those examined. Testing for the hypothesis—an alternative to the strategic interaction—that there is another unidentified SR that may explain the misspecification is an ongoing nontrivial issue; for the instance of point forecasting, see Elliott et al. (2008). Finally, the difference between sub-samples appears to show that the effects of recessions and data quality (which could be highly different over decades of observation) alter the quantity of

SJB among forecasters in response to differing incentives for information processing. Using a Bayesian framework, Gaglianone et al. (2022) recently acquired empirical proof of this idea.

4.3 Connections with Central Banks' Reputation

The findings of our prior empirical investigation call into question the links between SJB and Central Bank's repute. With the central banker as the FP, market participants as the FU, and inflation targeting as the assessment criterion, our SS can be viewed as an econometric alternative to the Barro-Gordon model. Perhaps more intriguing, the location of the SR may be viewed as a link with the Central Banker's evaluation criterion—thus with the inconsistency problem: if we adopt a utilitarian paradigm, the definition of reputation means avoiding actions that may produce higher reward in the future, so any deviation from the optimal action may be considered subjective cheating that induces lower utility in the future; this is fully consistent with De Finetti's definition of SR. As a result, our location test might be viewed as a test of the Central Bank's reputation.

Clearly, the economic literature in recent years has deepened several aspects: (i) the role of committees in decisions, as well as their voting rules; (ii) the way debate during meetings has been conducted, as well as the order in which the (heterogeneously informed) members talk; and (iii) the effectiveness of transparency versus secretive policy in decisions; see, for example, Ottaviani and Sørensen (2001); Sibert (2003); Persico (2004); Prat (2005); Levy (2007).

The SS architecture, by design, cannot solve all of these issues: statistics on single central bank voters (on the FU side) are frequently unavailable. On the other hand, the availability of individual survey panel data on the FP side enables to develop more sophisticated partial equilibrium versions of our approach.

5 Conclusions

Recent advances in economic theory suggest that the rise of forecasters' judgmental bias is associated to their own subjective learning. The latter produces effects via incorporation of the non-sample information during the execution of the forecasting process. The strategic nature of this bias cannot be detected by standard econometric methods. This paper has introduced a new, micro-funded statistical approach that enables both professional forecasters and their evaluators to account their reciprocal learning in both the production and evaluation steps of the forecasting process.

This framework, named Scoring Structure, allows econometricians to build a standard LM-type test to verify the hypothesis of coherence among a representative Forecast Producer's quotations and Forecast User's subsequent announcement and vice-versa. From a general perspective, the Scoring Structure establishes a link between the likelihood principle and coherence among Forecast Producer's quotations and Forecast User's announcements. More specifically, it establishes a direct link among forecasting models linearity and coherence. This makes the proposed approach very general, flexible and easy to implement.

Our empirical investigation leads us to conclude that strategic judgmental bias is a non negligible feature also in the most well-considered institutions. This can be explained because the forecasting agents have not only bad incentives but also imperfect processing capability. Additionally, this bias represents an important issue in forecast comparison. However, the specification introduced here is prototypical. Thus, further research is necessary to make it feasible in more realistic scenarios.

References

Amisano G, Giacomini R. 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business and Economic Statistics* **25**: 177–190.

Barro RJ, Gordon DB. 1983. Rules, discretion and reputation in a model of monetary policy. *Journal of monetary economics* **12**: 101–121.

Bernardo J. 1979. Expected Information as Expected Utility. *The Annals of Statistics* **7**: 686–690.

Bordalo P, Gennaioli N, Ma Y, Shleifer A. 2020. Overreaction in macroeconomic expectations. *American Economic Review* **110**: 2748–82.

Bramoullé Y, Kranton R, D'Amours M. 2014. Strategic interaction and networks. *American Economic Review* **104**: 898–930.

Brègman L. 1967. The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming. *USSR Computational Mathematics and Mathematical Physics* **7**: 200–217.

Chan K, Tong H. 1986. On estimating thresholds in autoregressive models. *Journal of Time Series Analysis* **7**: 178–190.

Clarida R, Gali J, Gertler M. 1999. The science of monetary policy: a new keynesian perspective. *Journal of economic literature* **37**: 1661–1707.

Dawid A. 1982. The well-calibrated bayesian. *Journal of the American Statistical Association* **77**: 605–610.

Dawid A. 2007. The geometry of proper scoring rules. *The Annals of the Institute of Statistical Mathematics* **59**: 77–93.

Dawid P. 2004. Probability, causality and the empirical world: A bayes-de finetti-popper-borel synthesis. *Statistical Science* **19**: 44–57.

De Finetti B. 1962. Does it make sense to speak of "good probability appraisers"? In Good I (ed.) *The Scientist Speculates*. New York: Wiley.

De Finetti B. 2017. *Theory of probability: A critical introductory treatment*, volume 6 of *Wiley Series on Probability and Statistics*. John Wiley & Sons. Translated by Antonio Machí and Adrian Smith.

Diebold F, Gunther T, Tay A. 1998. Evaluating Density Forecasts With Applications to Financial Risk Management. *International Economic Review* **39**: 863–883.

Ehm W, Gneiting T, et al. 2012. Local proper scoring rules of order two. *The Annals of Statistics* **40**: 609–637.

Eitrheim Ø, Teräsvirta T. 1996. Testing the adequacy of smooth transition autoregressive models. *Journal of econometrics* **74**: 59–75.

Elliott G, Komunjer I, Timmermann A. 2008. Biases in macroeconomic forecasts: irrationality or asymmetric loss? *Journal of the European Economic Association* **6**: 122–157.

Gaglianone WP, Giacomini R, Issler JV, Skreta V. 2022. Incentive-driven inattention. *Journal of Econometrics* **231**: 188–212.

Gallo GM, Granger CW, Jeon Y. 2002. Copycats and common swings: the impact of the use of forecasts in information sets. *IMF staff Papers* **49**: 4–21.

Galvao A, Garratt A, Mitchell J. 2021. Does judgment improve macroeconomic density forecast? *International Journal of Forecasting* **37**: 1247–1260.

Gennaioli N, Shleifer A. 2010. What comes to mind. *The Quarterly journal of economics* **125**: 1399–1433.

Giannerini S, Maasoumi E, Dagum E. 2015. Entropy testing for nonlinear serial dependence in time series. *Biometrika* **102**: 661–675.

Gneiting T, Raftery A. 2007. Strictly Proper Scoring Rules, Prediction and Estimation. *Journal of the American Statistical Association* **102**: 359–378.

Granger CW, Pesaran MH. 2000. Economic and statistical measures of forecast accuracy. *Journal of Forecasting* **19**: 537–560.

Hendrickson A, Buehler R. 1971. Proper Scores for Probability Forecasters. *The Annals of Mathematical Statistics* **42**: 1916–1921.

Ilut C, Valchev R. 2022. Economic agents as imperfect problem solvers. *The Quarterly Journal of Economics* **forthcoming**.

Komunjer I, Owyang MT. 2012. Multivariate forecast evaluation and rationality testing. *Review of Economics and Statistics* **94**: 1066–1080.

Kydland F, Prescott E. 1977. Rules rather than discretion: The inconsistency of optimal plans. *Journal of political economy* **85**: 473–491.

Levy G. 2007. Decision making in committees: Transparency, reputation, and voting rules. *American economic review* **97**: 150–168.

Manzan S. 2011. Differential interpretation in the survey of professional forecasters. *Journal of Money, Credit and Banking* **43**: 993–1017.

Manzan S. 2021. Are professional forecasters bayesian? *Journal of Economic Dynamics and Control* **123**: 104045.

Mundici D. 2009. Interpretation of de finetti coherence criterion in łukasiewicz logic. *Annals of Pure and Applied Logic* **161**: 235–245.

Nau RF. 2001. De Finetti was right: probability does not exist. *Theory and Decision* **51**: 89–124.

Olszewski W. 2015. Calibration and Expert Testing. In Young H, Zamir S (eds.) *Handbook of Game Theory with Economic Applications*. North Holland.

Ottaviani M, Sørensen P. 2001. Information aggregation in debate: who should speak first? *Journal of Public Economics* **81**: 393–421.

Ottaviani S, Sorensen P. 2006. The strategy of professional forecasting. *Journal of Financial Economics* **81**: 441–466.

Parry M, Dawid A, Lauritzen S. 2012. Proper Local Scoring Rules. *The Annals of Statistics* **40**: 561–592.

Patton A. 2019. Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics* : 1–43.

Pelloni G. 1996. De Finetti, Friedman and the methodology of positive economics. *Journal of econometrics* **75**: 33–50.

Persico N. 2004. Committee design with endogenous information. *The Review of Economic Studies* **71**: 165–191.

Pomatto L. 2021. Testable forecasts. *Theoretical Economics* **16**: 129–60.

Prat A. 2005. The wrong kind of transparency. *American economic review* **95**: 862–877.

Predd J, Seiringer R, Elliott H, Osherson D, Poor V, Kulkarn S. 2009. Probabilistic coherence and proper scoring rules. *IEEE Transactions on Information Theory* **55**: 4786–92.

Regazzini E. 1987. De Finetti's coherence and statistical inference. *The Annals of Statistics* : 845–864.

Rossi B, Sekhposyan T. 2019. Alternative tests for correct specification of conditional predictive density. *Journal of Econometrics* **208**: 638–657.

Savage L. 1971. Elicitation of Personal Probabilities and Expectations. *Journal of American Statistical Association* **66**: 783–801.

Schervish MJ. 1989. A general method for comparing probability assessors. *The annals of statistics* **17**: 1856–1879.

Shafer G, Vovk V. 2001. *Probability and Finance. – It's only a Game.* New York: Wiley.

Sibert A. 2003. Monetary policy committees: individual and collective reputations. *The Review of Economic Studies* **70**: 649–665.

Svensson L. 2005. Monetary policy with judgment: Forecast targeting. *International Journal of Central Banking* .

Teräsvirta T. 1994. Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**: 208–218.

Timmermann A. 2006. Forecast combinations. *Handbook of economic forecasting* **1**: 135–196.

van Dijk D, Terasvirta T, Franses P. 2002. Smooth transition autoregressive models – a survey of recent developments. *Econometric Reviews* **21**: 1–47.

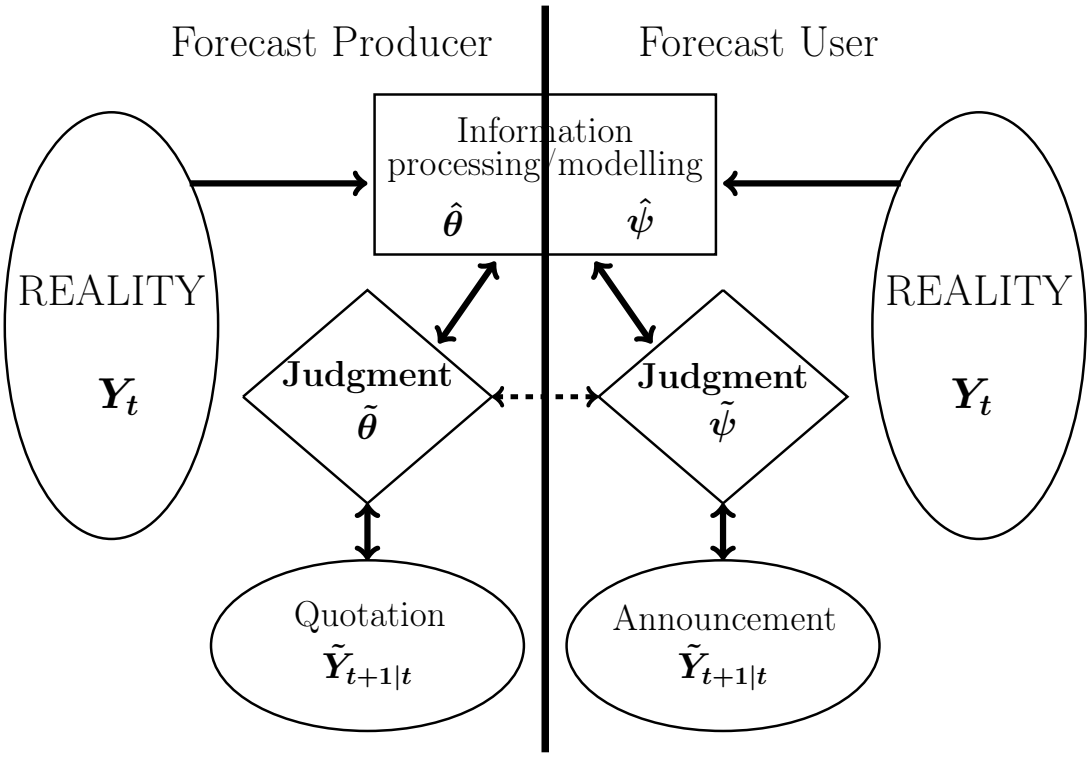
Vovk V, Shafer G. 2001. *Probability and Finance: it's Only a Game!* Wiley.

Vovk V, Shafer G. 2005. Good randomized sequential probability forecasting is always possible. *Journal of Royal Statistical Society, ser. B* **67**: 491–499.

Zanetti Chini E. 2023. Can we estimate macroforecasters' mis-behavior? *Journal of Economic Dynamics and Control* **149**: 104632.

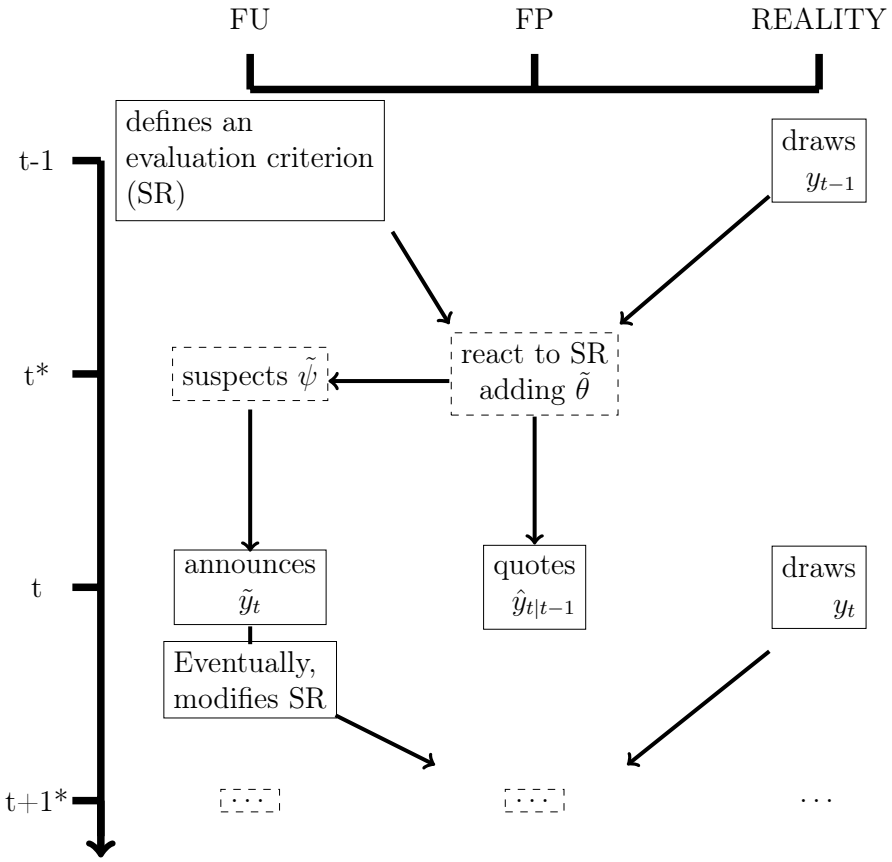
6 Tables and Figures

Figure 1: A synopsis of the classical forecasting environments



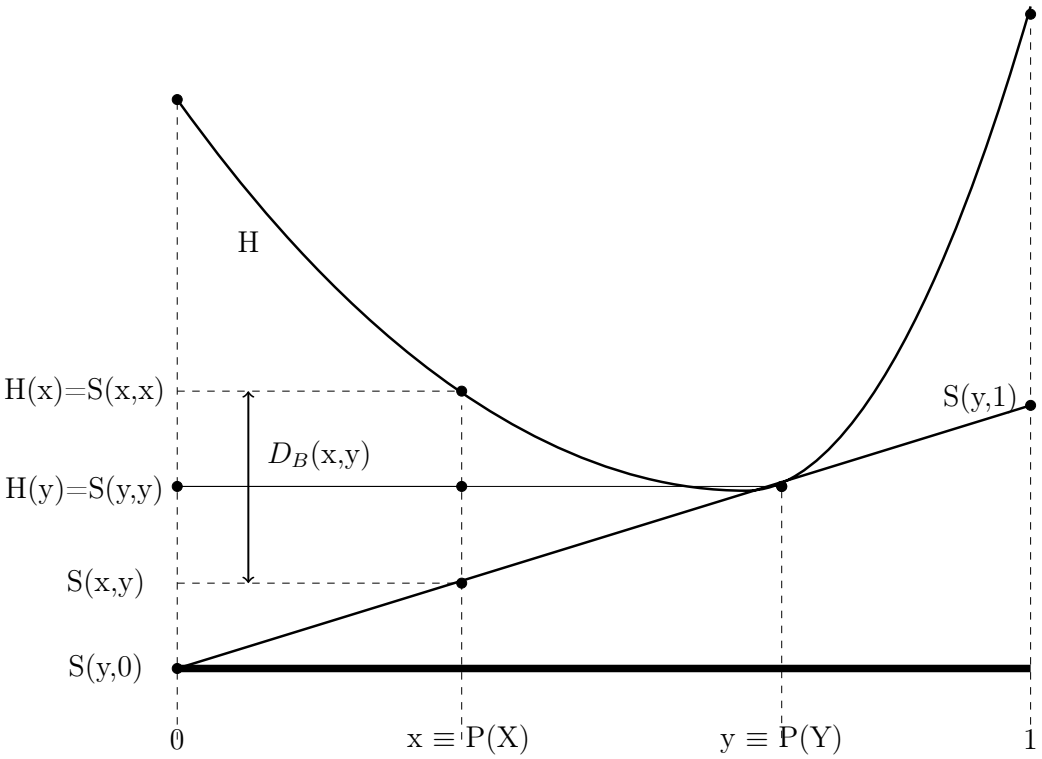
NOTE: This figure displays a visual synopsis of the classical forecasting environment, in which the FP and FU are separated entities that act according to a 'linear' scheme in which each of them (i) collects information on the variable(s) of interest from Reality, (ii) processes it via econometric modelling; (iii) eventually, adds a judgmental component (before or after the estimation/processing phase); (iv) finally, produces an announcement for the next period. This scheme is repeated for any period. In this framework, any cheating among FU and FP is assumed to be purely deterministic and thus, to influence the judgmental part only.

Figure 2: A synopsis of the proposed forecasting environments



NOTE: This figure displays the proposed forecasting environment, where each FP/FU's action may influence several, contemporaneous (and, possibly, future) decisions of the other Player in any period; dashed rectangles indicate unobserved actions; the vertical axis represent time and the horizontal axis groups the Players. The reiteration of such a multi-period, multi-agent game forms a lattice.

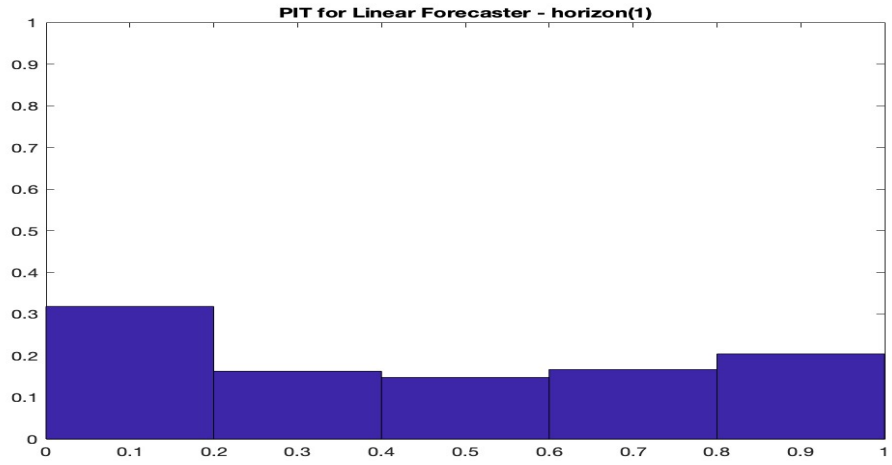
Figure 3: The Schervish representation



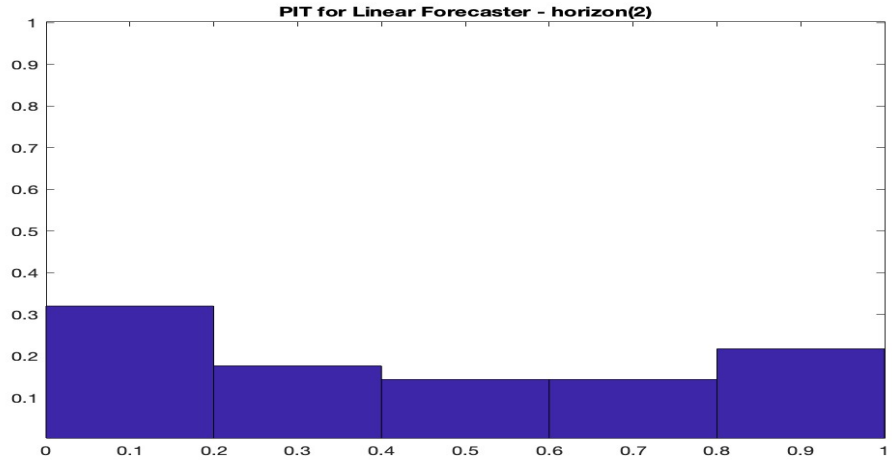
NOTE: Schematic illustration of the relationship between a (Generalized Bregman) Entropy Function H (solid convex curve) and the associated SR and Bregman Divergence d_B(x,y). Let be $x \in [0,1]$ and $y \in [0,1]$ a "true" probability forecast corresponding to P(Y) and its biased version corresponding to $P(X|\mathcal{I}, J)$ in the text). Then for any x the expected score $S(x,y) = xS(x,1) + (1-y)S(x,0)$ equals the ordinate of the tangent to H at y—that is, the solid line with slope H(y)—when evaluated at x. The Bregman divergence equals the difference between H and its tangent at y when evaluated at x. See Gneiting and Raftery (2007) (who adopts a different notation) for details.

Figure 4: The problem of Misspecification due to Strategic Judgment (Example 1)

(a)

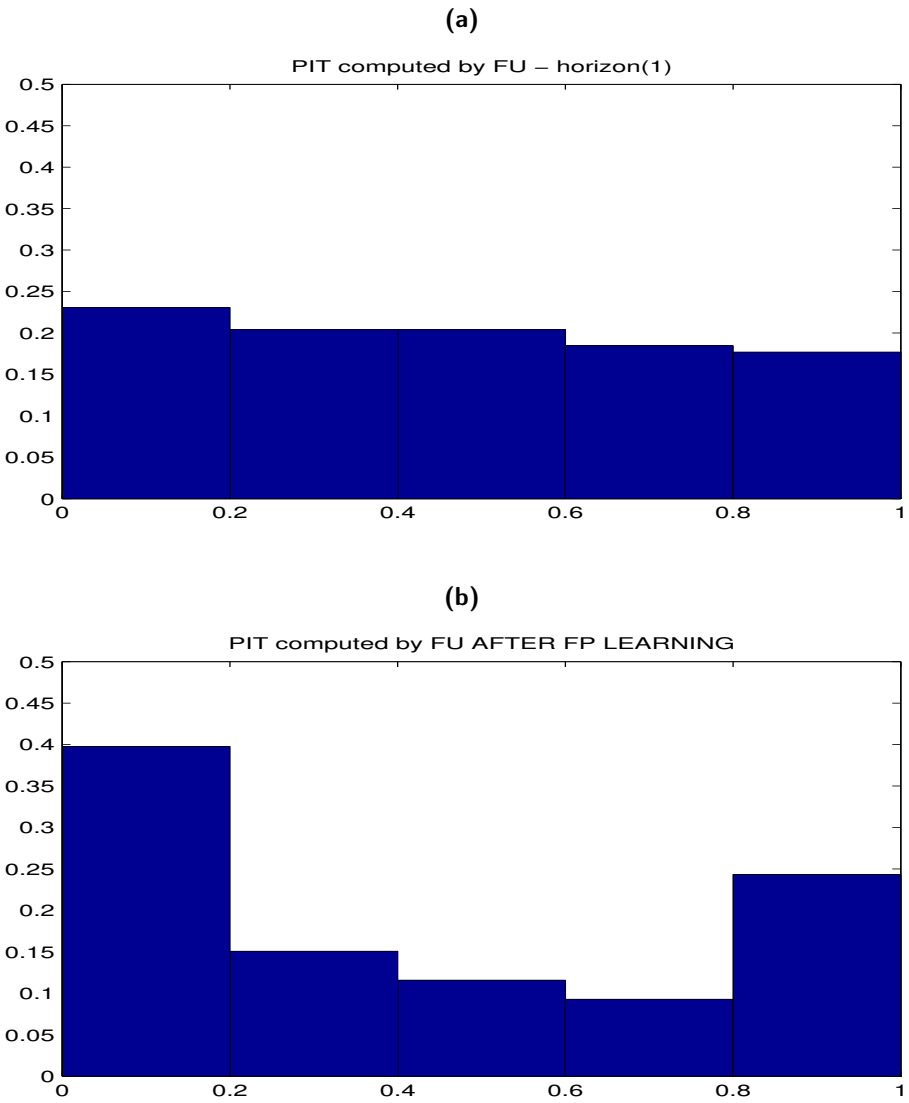


(b)



NOTE: This figure shows the problem of misspecification by FU in the illustrative i.i.d. with unknown mean example in Section 2 via the rolling-windows scheme. Namely, in a sequence of periods {1, ..., T}, FP observes a set of draws from (2) (corresponding to the data). Then, the same FP produces a number P of quotations based on direct OLS regression (assuming a quadratic loss) of the dependent variable on the regressors lagged $h-$periods. Thus, there are P out-of-sample predictions to be evaluated by FU, where the first out-of-sample prediction is based on a parameter estimated using data up to time R; the second prediction is based on a parameter estimated using data up to $R+1$, and the last prediction is based on a parameter estimated using data (generated by DGP) up to $R+P-1=T$, where $R+P+h-1=T+h$ is the size of the available sample, with $h=1$ being the pseudo-out-of-sample horizon, so that T=264 is the "in-sample" part. Once the sequence of estimation errors has been computed, the first judgmental component $\tilde{\pi}_1$ is added and the vector $\tilde{y}_t$ in (3) is collected and, successively, transformed via (1) in Step 2. Panel (b) corresponds to the same procedure replicated in the subsequent period, where the judgmental components $\tilde{\psi}$ in (4) and $\tilde{\pi}_2$ in (5) appear. In this example, R=20 and h=1. Moreover, for ease of exposition and without loss of generality, we arbitrarily set $\tilde{\pi}_1=\tilde{\pi}_2=\tilde{\psi}=0.05$.

Figure 5: The problem of Misspecification due to Strategic Judgment (Example 2)



NOTE: This figure shows the problem of misspecification by FU in the nonlinear illustrative example in Section 2 via the rolling-windows scheme. Namely, in a sequence of periods {1, ..., T}, FP observes a set of draws from (6) (corresponding to the data). Then, the same FP produces a number P of quotations based on direct OLS regression (assuming a quadratic loss) of the dependent variable on the regressors lagged $h-$periods. Thus, there are P out-of-sample predictions to be evaluated by FU, where the first out-of-sample prediction is based on a parameter estimated using data up to time R; the second prediction is based on a parameter estimated using data up to $R+1$, and the last prediction is based on a parameter estimated using data (generated by DGP) up to $R+P-1=T$, where $R+P+h-1=T+h$ is the size of the available sample, with $h=1$ being the pseudo-out-of-sample horizon, so that T=264 is the "in-sample" part. Once the sequence of estimation errors has been computed, the first judgmental component $\tilde{\pi}_1'$ is added and the vector $\tilde{y}_t$ in (7) is collected and, successively, transformed via (1) in Step 2. Panel (b) corresponds to the same procedure replicated in the subsequent period, where the judgmental components $\tilde{\psi}$ in (8) and $\tilde{\pi}_2'$ in (9) appear. In this example, R=20 and h=1. Moreover, for ease of exposition and without loss of generality, we arbitrarily set $\tilde{\pi}_1' = \tilde{\pi}_2' = \tilde{\psi}' = 0.05$.

Table 1: Structural coherence test for U.S. Real GDP

Sample: 1975:Q1 - 2020:Q4										
d	$h=0$		$h=1$		$h=2$		$h=3$		$h=4$	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**8.690**	**<0.001**	**8.708**	**<0.001**	**6.129**	**<0.001**	**4.285**	**<0.001**	**3.308**	**<0.001**
2	1.225	0.269	1.290	0.228	1.229	0.266	**2.000**	**0.027**	**1.754**	**0.059**
3	0.919	0.528	**1.628**	**0.087**	1.368	0.185	**2.232**	**0.012**	**1.916**	**0.035**
4	1.276	0.237	0.985	0.465	**2.042**	**0.023**	**1.662**	**0.079**	**2.767**	**0.002**
5	0.806	0.645	0.601	0.839	1.393	0.172	**2.190**	**0.014**	2.138	0.167
6	0.778	0.672	0.927	0.520	0.457	0.936	1.373	0.182	**2.040**	**0.023**
7	1.003	0.447	0.767	0.682	1.520	0.121	1.146	0.325	1.276	0.236
8	1.144	0.327	0.776	0.675	0.820	0.629	1.291	0.227	0.920	0.527
Sample: 1991:Q1 - 2020:Q4										
d	$h=0$		$h=1$		$h=2$		$h=3$		$h=4$	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**8.076**	**<0.001**	**8.256**	**<0.001**	**9.608**	**<0.001**	**8.473**	**<0.001**	**11.167**	**<0.001**
2	0.463	0.931	0.646	0.798	0.766	0.723	0.409	0.957	1.195	0.296
3	0.308	0.986	0.773	0.676	0.418	0.953	1.555	0.116	1.267	0.249
4	0.282	0.991	0.232	0.996	1.061	0.396	0.841	0.607	**1.626**	**0.094**
5	0.075	1.000	0.452	0.937	0.698	0.750	1.278	0.241	0.789	0.650
6	1.366	0.193	0.109	0.985	1.384	0.185	1.172	0.319	**2.017**	**0.029**
7	0.394	0.966	0.424	0.951	0.155	0.999	1.272	0.244	**1.608**	**0.100**
8	0.567	0.863	0.337	0.980	0.298	0.988	0.278	0.991	0.755	0.694

NOTE: This table reports the test statistics LM_1 (**??**) in their F-variant and corresponding p-values for data on the U.S.GDP forecasts of one, two, three and four quarters-step-ahead. Rejections are illustrated in bold. The zero-horizon in the first group of columns refers to nowcasts.

Table 2: Structural coherence test for U.S. GDP Deflator

Sample: 1975:Q1 - 2020:Q4										
d	$h=0$		$h=1$		$h=2$		$h=3$		$h=4$	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**2.233**	**0.009**	**4.121**	**<0.001**	**5.129**	**<0.001**	**4.285**	**<0.001**	**5.308**	**<0.001**
2	**2.144**	**0.012**	**3.549**	**<0.001**	**4.040**	**<0.001**	**2.000**	**0.027**	**4.654**	**<0.001**
3	1.622	0.075	**2.628**	**0.002**	**3.040**	**<0.001**	**2.232**	**0.012**	**3.916**	**<0.001**
4	1.561	0.092	1.828	0.037	**2.337**	**0.006**	**2,045**	**0.018**	**2.751**	**0.002**
5	1.481	0.121	1.659	0.066	1.655	0.066	**1.790**	**0.042**	**2.068**	**0.016**
6	1.348	0.189	1.432	0.143	1.421	0.148	1.578	0.0872	**1.837**	**0.023**
7	1.233	0.273	1.107	0.399	1.285	0.232	1.339	0.195	**1.780**	**0.043**
8	1.169	0.332	1.032	0.490	1.200	0.303	1.294	0.225	1.677	0.062
Sample: 1991:Q1 - 2020:Q4										
d	$h=0$		$h=1$		$h=2$		$h=3$		$h=4$	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**2.611**	**0.002**	**3.084**	**<0.001**	**3.221**	**<0.001**	**4.452**	**<0.001**	**4.885**	**<0.001**
2	**2.054**	**0.017**	**2,250**	**0.009**	**2.540**	**0.003**	**3.602**	**<0.001**	**3.930**	**<0.001**
3	1.708	0.055	**1.824**	**0.037**	**1.788**	**0.042**	**2.874**	**0.001**	**3.041**	**<0.001**
4	1.630	0.072	1.663	0.065	1.591	0.083	**2.188**	**0.011**	**2.090**	**0.015**
5	1.399	0.159	1.481	0.121	1.270	0.243	1.740	0.050	1.733	0.051
6	1.260	0.250	1.256	0.254	1.199	0.304	1.612	0.077	1.499	0.118
7	1.095	0.412	0.996	0.537	1.055	0.461	1.422	0.142	1.350	0.187
8	0.808	0.792	0.914	0.650	0.922	0.639	1.106	0.402	1.082	0.428

.

NOTE: This table reports the test statistics LM_1 (**??**) in their F-variant and corresponding p-values for data on the U.S.GDP Deflator forecasts of one, two, three and four quarters-step-ahead. Rejections are illustrated in bold. The zero-horizon in the first group of columns refers to nowcasts.

Table 3: Structural coherence test for U.S. UNR

Sample: 1975:Q1 - 2020:Q4										
d	$h=0$		$h=1$		$h=2$		$h=3$		$h=4$	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**185.110**	**<0.001**	**107.664**	**<0.001**	**44.650**	**<0.001**	**22.432**	**<0.001**	**24.380**	**<0.001**
2	**5.801**	**<0.001**	**5.894**	**<0.001**	**5.845**	**<0.001**	**5.548**	**<0.001**	**7.313**	**<0.001**
3	**4.190**	**<0.001**	**4.685**	**<0.001**	**4.642**	**<0.001**	**4.502**	**<0.001**	**6.151**	**<0.001**
4	**3.202**	**0.002**	**3.202**	**<0.001**	**3.424**	**<0.001**	**3.321**	**<0.001**	**4.624**	**<0.001**
5	**2.667**	**0.042**	**2.198**	**0.013**	**2.444**	**0.006**	**2.561**	**0.007**	**2.962**	**<0.001**
6	**1.856**	**0.080**	**1.871**	**0.039**	**2.055**	**0.022**	**2.145**	**0.016**	**2.570**	**0.006**
7	1.651	0.114	**1.675**	**0.074**	**1.785**	**0.053**	**1.849**	**0.042**	**1.883**	**0.039**
8	**1.922**	**0.033**	**1.884**	**0.037**	**1.870**	**0.039**	**1.896**	**0.036**	**1.777**	**0.055**
Sample: 1991:Q1 - 2020:Q4										
d	$h=0$		$h=1$		$h=2$		$h=3$		$h=4$	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**443.798**	**<0.001**	**311.933**	**<0.001**	**29.179**	**<0.001**	**29.179**	**<0.001**	**22.355**	**<0.001**
2	**5.462**	**<0.001**	**5.296**	**<0.001**	**5.166**	**<0.001**	**5.088**	**<0.001**	**5.456**	**<0.001**
3	**4.181**	**<0.001**	**3.403**	**0.005**	**4.342**	**<0.001**	**4.287**	**<0.001**	**4.143**	**<0.001**
4	**3.141**	**<0.001**	**2.586**	**0.082**	**3.458**	**<0.001**	**3.353**	**<0.001**	**3.259**	**<0.001**
5	**2.202**	**0.016**	1.675	0.492	**2.799**	**0.023**	**2.878**	**0.002**	**2.748**	**0.002**
6	1.402	0.172	0.952	0.580	**1.932**	**0.052**	**1.893**	**0.042**	**1.192**	**0.039**
7	0.803	0.641	0.868	0.588	1.092	0.374	1.179	0.306	1.204	0.209
8	0.836	0.612	0.861	0.785	0.899	0.550	0.943	0.507	0.965	0.487

NOTE: This table reports the test statistics LM_1 (**??**) in their F-variant, with corresponding p-values for data on the U.S. UNR forecasts of one, two, three and four quarters-step-ahead. Rejections are illustrated in bold. The zero-horizon in the first group of columns refers to nowcasts.

A Mathematical details

A.1 Prelminary theory

Let $\overline{\mathbb{R}} = [-\infty, +\infty]$ denote the extended real line and the functions $H(Y) : \mathcal{M} \to \overline{\mathbb{R}}$ and $D(X, Y) : \mathcal{M} \times \mathcal{M} \to \overline{\mathbb{R}}$ be associated with any $U(P, \cdot)$. The resulting objects are defined as follows:

Definition 3 (SRs, entropy/divergence functions, scoring structure)**.** We define:

i. *(Local) Scoring Rule (SR)* the function $S(y, p) := U(p, \mathrm{a}_p)$ where $S : \mathcal{Y} \times \mathcal{P} \to \mathbb{R}$ is local of order m (or m-local) if it can be expressed in the form of:

$$S(z, p) = \mathfrak{s}(z, p(z), p'(z), p''(z), \ldots, p^{(m)}(z)), \tag{15}$$

where $\mathfrak{s} = \mathcal{Z} \times \mathcal{P}_m \to \mathbb{R}$ is the scoring function (or "p-function") of $S(z, p)$, $\mathcal{P}_m := \mathbb{R}^+ \times \mathbb{R}^m$ is a real-valued, infinitely differentiable function, m is a finite integer, and the prime (') denotes the differentiation with respect to z.

ii. The *Entropy* is the function $H(X) := S(X, X) \equiv \sup_{X \in \mathcal{X}} S(Y, X)$, where the notation X and Y omits $P(\cdot)$ and the partition of Y for convenience;

iii. The *Divergence* is the function $D(X, Y) := H(X) - S(X, Y)$, where the same omission holds;

iv. The *Scoring Structure* is the 6-ple $\mathcal{SS} := \left\{\mathcal{Z}_t, \mathcal{F}_t, \mathcal{P}, S(\cdot, \cdot), H(\cdot), D(\cdot, \cdot)\right\}$.

Definition 3 makes all the functions interpretable in terms of utility: $S(z, P)$ is the FP's reward if event z (truly) materializes. Since $S(\cdot)$ is defined on the extended real line, the *expected* FP's utility, conditional to X, can be denoted as $S(X, Y) \equiv \int_{-\infty}^{+\infty} S(X, x) dY(x)$. $H(X)$ can be interpreted as the maximum possible of the utility that the FP can achieve using Reality's true DGP to predict X. The divergence function is the difference between the maximum utility and the utility achieved by predicting the quoted predictive distribution $P(X)$, given the true distribution $P(Y)$

or their density equivalent. Notice that the same interpretations hold for FU while inverting the variables' order in the functions – that is $S(Y, X)$, $H(Y)$ and $D(Y, X)$.

This definition of SS is highly general and is only used to consider both FP's and FU's points of view. Several assumptions about each of them must be made to define the type of interaction that occurs among these three players and to delineate the econometric methodology to be used:

A 1. *$\mathcal{P}$ (or $\mathcal{M}$, if sample space is continuous) is assumed such that EU exists for all $a \in \mathcal{A}$, $P \in \mathcal{P}$.*

A 2. *$\mathcal{A}$ is compact.*

A 3. *$U(\hat{p}, a)$ is strictly convex in* a.

A 4. *$S(X, Y)$ is strictly convex and minimized in X. Equivalently, the strictly convex $S(Y, X)$ is minimized in Y.*

A 5. *$D(X, Y) - D(X_0, Y)$ is affine in X, and $D(X, Y) \geq 0$, with equality achieved at $X = Y$. The same property holds if inverting the variables' order.*

A 6. *$\hat{p}_{t+k}(Z)$ is a measurable function of the data in a rolling estimation window.*

A1 – A3 are necessary (but not sufficient) to define the FP's reward as SR. In particular, A1 encompasses the three "basic assumptions" discussed in Dawid (2007)[22] and suggests that the reward is measurable with respect to $\mathcal{A}$ and quasi-integrable with respect to all $p \in \mathcal{P}$ (or, $\mathcal{M}$, if continuous). A2 and A3 are convenience assumptions that are necessary only to have a unique maximizing action. A4 characterizes the general representation of SRs. A5, justified by Theorem 1 in Bernardo (1979), stresses that the FP has no loss only if his DGP coincides with that of Reality. A6 is fundamental to characterizing a general family of SRs for the case that every

[22] In our simplified notation: a) there exists exactly one $\boldsymbol{p} \in \mathcal{A}$ for any $p \in \mathcal{P}$; b) distinct distributions in P have distinct actions in $\mathcal{A}$; c) Every a $\in \mathcal{A}$ is a Bayes act for some $P \in \mathcal{P}$; see Dawid (2007), p. 80.

$P \in \mathcal{P} = \mathcal{A}$ ($\mathcal{M} = \mathcal{A}$ for the continuous case) has a density, for example, $p(Y)$ empirically realized as $p(y)$, with respect to $\mu \in \mathcal{Y}$, that is the *Brègman score*:

$$S(y,p)^{Bregman} \doteq f\big[p(x)\big] + \int y\Big\{f\big[p(y)\big] - p(y)f'\big[p(y)\big]\Big\}d\mu \tag{16}$$

with the associated *Brègman divergence*. Finally, A6 is necessary to apply Amisano and Giacomini (2007)'s predictive ability test on the SS' outputs.

A.2 Proof of Theorem 1

Let $\mathbb{L}$ and $\mathbb{D}$ be the same operators defined in Section 3, $\Lambda = \sum_{k\geq 0}(-1)^k D^k \partial/\partial p_k$ the Lagrange operator defined in equation (25) of Parry et al. (2012), and $\mathbb{I}$ the identity operator.

To prove the "if" part of the statement we need to show that, if (i) $\mathbb{L}\mathfrak{s} = 0$, (ii) $\mathfrak{s} = (\mathbb{I} - \mathbb{L})s$, s being a generic 0-homogeneous q-function, and (iii) $s = \Lambda\phi$, where ϕ is a generic 1-homogeneous p-function, then $\mathcal{L}(\Pi) \equiv \mathcal{L}(\Psi)$. The Key Condition (i) is a consequence of the fact that (assuming a comparison among X and Y as defined in the Main Text just as example), in $p(x) = p(y)$ $S(\cdot)$ is a stationary point under an infinitesimal variation $\delta p(\cdot)$ of $p(\cdot)$ (if assuming that $p(\cdot) + \delta p(\cdot)$ is still a density function); in turn, this leads to use classical variational analysis arguments by Parry et al. (2012), pages 569–71. (ii) is a consequence of Corollary 6.3 by Parry et al. (2012). (iii) is a consequence of Theorem 5.3 and Corollary 6.3 by the same authors. Since each single condition (i)–(iii) holds, Proposition 1 can be applied. Now, we need show only that if two $S(\cdot)$ are key local, their likelihood functions coincide; to this end, it is sufficient to notice that Key Equation (14) is the only binding condition because it must be satisfied for any $S(\cdot)$ function, even if (ii) and (iii) are not satisfied. Now $\mathcal{L}(\cdot)$ is, by definition, a simple linear (product) transform of $log\big(p(\cdot)\big)$ – that is, the same LogS; see Ehm et al. (2012). The operators $\mathbb{D}$ and $\mathbb{L}$ here adopted are linearly invariant by Corollary 11.3 and Theorem 11.4 of Parry et al. (2012). Hence

the statement.

To prove the "only if" part of the statement we need to show that if $\mathcal{L}(\Pi) \equiv \mathcal{L}(\Psi)$, then $S(\cdot)$ is key local. This is trivial when $p(y) \equiv p(x)$, in which case there is no evaluation. In the non-trivial case that $p(y) \neq p(x)$, the forecast is coherent when the expected score of FU coincides with FP's; in turn, this condition is ensured by Theorem 1 in Bernardo (1979), where the Expected Information of FU (that is, the "distance" between changing its opinion from $p_{\Pi}(\cdot)$ to $p_{\Pi}(\cdot|x)$ after that data materializes and maintaining $p_{\Pi}(\cdot)$ without any regard to data) can be written as a Kullback-Liebler divergence. By definition, the expected information is zero only when this expected utility of having such insight for FU coincides with FP – that is, the difference between expected information for FU and FP is zero. Hence the statement.

SUPPLEMENT
to
"*The judgmental strategy of professional forecasters*"
(FOR ONLINE PUBBLICATION ONLY)

June 2024

Abstract

This Supplement provides additional analysis and results which cannot be put in the Main Document for reasons of space. Namely, Section 1 illustrates the model and inference; Section 2 provides details of the Monte Carlo Simulation of the SS-STARX model introduced in Section 4 of the Main Document; Section 3 provides additional empirical examples; Section 4 gives a taxonomy of the utility functions characterizing the Scoring Structures to better understand the motivation leading to the use of Bregman-Savage-type of SR; Section 5 provides a formal proof of Proposition 1 in Main Document; Section 6 reports the proof of Proposition 3 in Section 4 of the Main Document; finally, Section 7 discusses the estimation; .

1 The SS-STARX

The process $\{y_t\}$ observed at $t = 1-p, 1-(p-1), \ldots, -1, 0, 1, \ldots, T-1, T$ is assumed to have the following parametrization:

$$\begin{aligned} y_t &= \boldsymbol{\phi}'\mathbf{w}_t + G(\gamma, \mathbf{z}_t, \mathbf{c}_k)\boldsymbol{\theta}'\mathbf{w}_t + \epsilon_t, \;\; \epsilon_t \sim iid(0, \sigma^2) \\ G(\gamma, \mathbf{z}_t, \mathbf{c}_k) &= \left(1 + \exp\left\{ -\gamma \prod_{k=1}^{K} (\mathbf{z}_t - \mathbf{c}_k) \right\}\right)^{-1}, \;\; \gamma > 0, \;\; c_1 < \cdots < c_k, < \cdots < c_K, \end{aligned} \tag{1}$$

where: $\mathbf{w}_t = (1, y_{t-1}, \ldots, y_{t-p})'$ are the autoregressive covariates; $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)'$ are the linear part parameters; $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)'$ are the nonlinear part parameters; γ is the slope parameter; $\mathbf{c}_k = (c_1, \ldots, c_K)$ denoting the (eventually, multiple) location parameters; $\mathbf{z}_t = a'\mathbf{x}_t \odot \mathbf{s}$ is a composite transition variable, with $a = [a_1, \ldots, a_p]'$, $a_i = 0$ if $i = d$ and 1 if $i \neq d$ indicating that delay parameter d, which is such that $1 \leq d \leq p$, is unknown and $\mathbf{x_t}$ the vector of FP's quotation; $\mathbf{s} = vec(\mathfrak{s} \otimes \mathbf{i})$ with $\mathfrak{s}$ is a scalar denoting a generic (strictly) proper SR and $\mathbf{i}$ is a one-vector of the same dimensions of $\mathbf{x_t}$. The most common choices for K are $K = 1$, in which case the parameters $\boldsymbol{\phi} + \boldsymbol{\theta} G(\gamma, \mathbf{z}_t, \mathbf{c}_k)$ change monotonically as a function of $\mathbf{z}_t$ from $\boldsymbol{\phi}$ to $\boldsymbol{\phi} + \boldsymbol{\theta}$ and $K = 2$, in which case the parameters $\boldsymbol{\phi} + \boldsymbol{\theta} G(\gamma, \mathbf{z}_t, \mathbf{c}_k)$ change symmetrically at the point where the function reaches its own minimum. A peculiar form of this latter case is when $K = 2$ and $c_1 = c_2$ and the transition function defines the SS-Exponential STARX (SS-ESTARX) model. When $\gamma \to \infty$, the equation (1) becomes a two-regime threshold autoregression SS (SS-TARX).

The (nonlinear) SS so defined is an algorithm that applies the Forecasting Protocol described in Section 2 of the Supplement. Its use requires three steps: (i) the FU specifies the form of $\mathfrak{s}$ that will be adopted to evaluate the FP; (ii) the FP estimates $\hat{p}(y_{t+h})$ and applies $S(\tilde{y}_t, x_t)$ to it – that is, the FU makes sure that FP's bias $\tilde{\psi}_t$ appears in x_t – so that $\mathbf{z_t}$ can be computed; and (iii) the quotation is compared with the realizations y_t via (1). In the Supplement, Reality is assumed to act as a third player, so the number of steps grows without loss of generality.

The mechanics of the forecasting exercise executed by FP is independent of the form of the SS: no restrictions or assumptions, neither in the forecasting model nor in the methodology adopted to obtain $S(y_{t+h}, x_t)$ are needed. As it will appear shortly, equation (1) is necessary only as a convenient way to test the null hypothesis of structural coherence, corresponding to equation (14) in Main Document. Moreover, Step (ii) can also be seen from the FP's side: that is, she presumes that FU will play against her quotation and thus, after having estimated $\hat{p}(x_{t+h})$, she incorporates the FU's bias $\tilde{\pi}_t$

in $\tilde{y}_t$.[1]

The null hypothesis of structural coherence can be investigated as follows:

Proposition 1. *Let y_t be a stochastic process generated by (1). Then:*

(i) The locality can be tested via the hypothesis system

$$H_0 : \gamma = 0 \;\; vs \;\; H_1 : \gamma \neq 0 \quad in \; (1), \tag{2}$$

which can be measured by the following LM statistics:

$$S(\boldsymbol{\Xi})^{LM} = \hat{\sigma}^{-2}\hat{\mathbf{U}}'\hat{\mathbf{D}}_{\mathbf{2}}(\hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{D}}_{\mathbf{2}} - \hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{D}}_{\mathbf{1}}(\hat{\mathbf{D}}_{\mathbf{1}}'\hat{\mathbf{D}}_{\mathbf{1}})^{-1}\hat{\mathbf{D}}_{\mathbf{1}}'\hat{\mathbf{D}}_{\mathbf{2}})^{-1}\hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{U}} \sim \chi^2_n \tag{3}$$

where $\hat{\mathbf{U}}$, $\hat{\mathbf{D}}_{\mathbf{1}}$, $\hat{\mathbf{D}}_{\mathbf{2}}$ denote properly defined matrices; $\hat{\sigma}^{-2}$ is an estimator of the unconditional variance of SS; n is the length of the vector of nonlinear parameters.

(ii) Alternatively, the system (2) can be measured by one of the following LM statistics:

$$\begin{aligned} LM_1 &= (SSR_0 - SSR)/\hat{\sigma_v}^2 \sim \chi^2_{3p} \;\; if \;\; K = 1 \;\; in \;\; (1) \\ LM_2 &= (SSR_0 - SSR)/\hat{\sigma}^2_{v1} \sim \chi^2_{2p} \;\; if \;\; K = 2 \;\; and \;\; c_1 = c_2 \;\; in \;\; (1) \\ LM_3 &= (SSR_0 - SSR)/\hat{\sigma}^2_{v2} \sim \chi^2_{p} \;\; if \;\; K = 2 \;\; and \;\; c_1 \neq c_2 \;\; in \;\; (1), \end{aligned} \tag{4}$$

where SSR_0 and SSR are the sum of the squared residuals of SS-STARX (1) linearized via the Taylor expansion, $\hat{\sigma_v}^2$, $\hat{\sigma}^2_{v1}$, and $\hat{\sigma}^2_{v2}$ are estimators of unconditional variance of the same linearized SS-STARX(p); p is the autoregressive order of the same SS-STARX. F-type tests equivalent to LM statistics in (4) are preferable in small samples.

Proof. It is a re-proposition of the existing results by Luukkonen et al. (1988) and Teräsvirta (1994), and thus it is shown in the Supplement. □

[1] Estimation and inference are analogue to the STARX model; see the Supplement.

2 Simulation Study

This section reports the MonteCarlo simulation experiment that constitutes the foundation of some intuitions explained in the Main Text. Namely, Subsection 2.1 describes the Data Generating Process (DGP, heceforth); Subsection 2.2 displays the results; finally Subsection 2.3 provides a discussion.

2.1 Simulation Design

We consider two different DGPs:

$$y_{1,t}^{(i)} = 0.4y_{1,t-1}^{(i)} - 0.25y_{1,t-2}^{(i)} + (0.01 - 0.9y_{1,t-1}^{(i)} + 0.795y_{1,t-2})^{(i)}G^{(i)}(\gamma, \mathbf{w_t}, c) + \epsilon_{1,t}^{(i)}\,, \quad (5)$$

and

$$y_{2,t}^{(i)} = 0.8y_{2,t-1}^{(i)} - 0.7y_{2,t-2}^{(i)} + (0.01 - 0.9y_{2,t-1}^{(i)} + 0.795y_{2,t-2}^{(i)})G^{(i)}(\gamma, \mathbf{w_t}, c) + \epsilon_{2,t}^{(i)}\,, \quad (6)$$

where $G^{(i)}(\gamma, \mathbf{w_t}, c) = (1 + \exp\{-\gamma(\mathbf{w_t} - c)\})^{-1}$, $\epsilon_t^{(i)} \sim N(0,1)$, $i = \{1, \ldots, I\}$ denoting the i-th draw of the process $\{y_t\}_{t=1}^{T}$ with $c = \frac{1}{T}y_t^{(i)}$, $I = 1,000$.

$y_{1,t}^{(i)}$ (henceforth "DGP 1") is an additive nonlinear model with accentuated nonlinear behavior because of the high autoregressive parameters that drive $G(\cdot)$, which give high sensitivity to the size of the slope parameters. This can be the case of a macroeconomic indicator that is affected by an unexpected shock that pervades the time series dynamics. On the other hand, $y_{2,t}^{(i)}$ (henceforth "DGP 2") describes a mixed scenario. To simulate the function $G(\cdot)$, we use a set of values to investigate the cases of null, small, and high nonlinearity in the SS, corresponding to a coherent, near-to-coherent, and non-coherent forecast scenario, respectively. We also consider three hypotheses for T and three sample sizes – $T = \{75, 150, 300\}$ for very small, small, and medium-sized samples, respectively – and $\alpha = \{0.01, 0.05, 0.10\}$.

We also investigate the (dynamic) adequacy of the SS via three diagnostic tests, originally introduced by Eitrheim and Teräsvirta (1996). These are (i) a test for no error

autocorrelation, where we assumed the errors of the generating process followed an AR(1) process $u_t = \rho u_{t-1} + \epsilon$, $\epsilon \sim NID(0,1)$ and $\rho = \{0.2, 0.4\}$. (ii) A test for no additive nonlinearity, where we added to the previously described DGP a logistic function $G_2(\gamma^{(2)}, s_t, y_{t-1}))$ with AR-coefficients $\pi_0 = 0.01$, $\pi_2 = -1.8$, $\pi_3 = 1.6$,$\gamma^{(2)} = \{5, 10, 100\}$ and $G_1(\cdot)$ fixed to $\gamma = 20$; fixing G_1 ensures that the behavior of the additive component remains isolated from the additional component, so that no interaction among FP and FU can be observed. (iii) A test for parameter constancy, where the AR-coefficients have been simulated according to a logistic smooth change $\lambda_1 = (0, 0.4, -0.25)'$ and $\lambda_2 = (0.2, -0.9, -0.795)'$.

All these tests aim to give the investigator information regarding the existence of pure mis-specifications in the functional form *regardless* the internal setting of the Forecasting Protocol. This implies that they are not able to capture further strategic interaction, which can only be inferred via the locality test presented in the Main Text.

2.2 Results

Table 1 reports the results of the Monte Carlo simulation of the coherence test for the statistics F_1 and F_2 from the hypothesis system (17) discussed in Section 4 of the Main Document. The performances of the F_3 statistic are poor, so they are omitted. The two test statistics behave well for what concerns the empirical size. Conversely, the empirical power is poor if an almost-linear specification of the SS is used, and in general for DGP 1. Moreover, the empirical power is highly sensitive to the values of the slope. For example, under DGP1 and T=75 and $\alpha = 0.10$, the power of the F_1 statistic passes from almost 0.02 when $\gamma = 0.5$ (hence, an almost linear model) to 0.6 when $\gamma = 500$. Therefore, the increase is proportional but less than linear, as is similar for statistic F_2. When DGP2 is considered, the range is more abrupt: *ceteribus paribus*, F_1 is 0.05 when $\gamma = 0.5$ and 0.88 when $\gamma = 500$. The role of γ becomes almost inflationary as T increases. For example, when $T = 300$, and $\alpha = 0.05$, the range of the power of F_1 in DGP1 is [0.001 – 0.892] and is even more in DGP2. Therefore, there is strong evidence of a relationship between the SS's degree of nonlinearity and the test's

empirical power. Thus, the test correctly accepts the hypothesis of coherence more easily when the SS is highly nonlinear than it does in the opposite case of quasi-linear behavior, a feature we call *structure linearity bias*. This finding is counter-balanced by the functional form of the SRs having no role in the test's empirical power. Table 2 reports the results of a simulation of the same two DGPs, where we fixed $\gamma = 10$ and most of the scoring functions mentioned in Table 9 apart from the logarithmic score previously investigated. The value of each F-statistic is the same for all nineteen SRs adopted (e.g., the power of F_1 in DGP1 at the nominal size of 5% is 0.35 with T=75, 0.57 with T=150, and 0.63 with T=300). The empirical power of the test under DGP2 (i.e., the mixed scenario) is high in general – in particular, higher than the nonlinear scenario. *Ceteribus paribus*, the power of the F_1 statistic is 0.67 with $T = 75$, 0.72 with $T = 150$, and 0.85 with $T = 300$, and the equivalent F_2 power values are slightly lower – at least in case of middle sample dimensions. In other words, when the SS parameters are fixed, the empirical power of the test is invariant to the form of the SR that is assumed to drive the FP's quotation. This second feature is called '*Score Invariance*' and, as theoretically demonstrated in Paragraph 11.2 in Parry et al. (2012), it holds also for $D(\cdot, \cdot)$ and $H(\cdot, \cdot)$.

The results of the MonteCarlo simulation for diagnostic tests are displayed in Tables 3 and 4. Some *caveat* are required to interpret the empirical power properties: first and noticeably, in several occasions, the diagnostic tests tend to be weakly oversized—in particular in DGP 1; this sounds different from the results by Eitrheim and Teräsvirta (1996) for a traditional STAR econometric modelling. Second, under DGP 1, all the tests have good power, in particular the serial correlation test; the tests of no additive asymmetry and parameter constancy are characterized by a duality: when the slope is high, that is $\gamma_1 = 20$ the power is extreme when γ_2 is high, while it drops for low-medium values of the same (additive) parameter (around 0.50-0.75 vs almost 1.00 at $\alpha = 5\%$, T=100 in no additive nonlinearity test, around 0.50 vs 0.88 for LM_2 statistic at the same nominal and sample size for parameter constancy). On the other hand, under DGP 2, the change in scale of the power is globally less pronounced; interestingly,

the test for no serial correlation is less powerful.

2.3 Discussion

This simulation provides several lessons. First, the Structure's linearity bias means the power of the locality test depends on the type of model that the SS assumes. In this sense, the magnitude of γ is proportional to the degree of bias that the FP is suspected to have when he or she produces quotations.

Second, the Score Invariance is a direct consequence of the Cournot's Principle. The forecasts (drawn by the FP) and the observations (drawn by Reality) are not correlated if the functional form of the SR is elicited before the event occurs, which is one of the most critical assumptions of Lindley (1982)'s generalized theory on the admissibility of the FP's utility. In fact, according to the latter, the Score Invariance is a necessary and sufficient condition for treating the scores as finitely additive, probability-behaving objects – that is, for being coherent in the sense of De Finetti (2017). In particular, Lindley's Lemma 4 demonstrates the equivalence between two scores that correspond to two quotations when these are conditional on the same event, thus enhancing the status of the probability transform of the obtained value x[2]. In this sense, the results of our simulations are fully consistent with the De Finetti–Lindley theory.

Third, our simulations confirm that the SRs' consistency – so axiomatically determined – is a non-sufficient condition for the coherence of the forecasts' evaluation, as suggested by Patton (2019). Although some of the nineteen scoring functions used in this experiment have Brègman–Savage representation, the test's empirical power coincides with that of the test statistics corresponding to SRs. Therefore, when the FU deals with FP, even if the FU specifies (axiomatically) ex-ante the exact utility function that will be used to evaluate the FP, as required by Step 2 of the Forecasting Protocol, the FU will never know, ex-post, if the same utility function is the one the FPs used. This sort of "undeterminancy" is the rason for adopting the locality (which means coherence) as

[2] According to Lindley (1982, p. 4) "*It follows that a person could proceed by choosing his probability p in advance of knowing what score function was to be used and then, when it was announced, providing x satisfying $P(x) = p$.*"

a criterion for assessing the forecasts. In fact, locality tells the FP whether Barnard et al. (1962)' likelihood principle, according to which all the evidence in a sample that is relevant to the model parameters is contained in the likelihood function, holds. In this case, the forecast must necessarily be driven by some function derived from the likelihood. Since the FU is supposed to have sound knowledge about the estimation methods used to verify the FP's work, any deviations from likelihood are likely represented by judgments.

Finally, the SJB may complicate the traditional econometric diagnostics. The difficulty to reach full power for many tests, also in some of the most extreme cases, is symptomatic of how easily one can confuse the endogenous accumulation of errors that characterizes SS with purely exogenous mis-specifications.

3 Further Illustrations

This Section provides two further case studies in addition to the one in the Main Text; namely, Subsection 3.1 reports the case of Bank of Engalnd's "Fan Charts" for the U.K. inflation and Subsection 3.2 discusses the equivalent ones for the Norwegian output growth projections by Norges Bank.

3.1 The U.K. inflation

The Bank of England (BoE, henceforth) adopts and publishes probabilistic forecasts in the form of 'Fan Charts' on several key macroeconomic indicators in support of its policy decisions since 1996. Perhaps the most famous example of these indicators is the inflation rate; see, among others, Wallis (2004); Mitchell and Hall (2005). The Monetary Policy Committee (MPC, henceforth) provides monthly projections on the CPI inflation; it acts as FU and is fully responsible for the achievement of the Bank's institutional targets. Thus, its projections can be seen as announcements. More recently, the BoE has also published data on the Survey of External Forecasters (SEF), which constitute the FP in our framework and are equivalent to the U.S. SPF. It is

very well-known that the BoE uses an asymmetric two-pieces Normal autoregressive process to produce its density forecasts; see, among others, Boero et al. (2008, 2011). In this paper we consider data from January 2014 to December 2019, corresponding to the release of August 2019. These can be downloaded at: www.bankofengland.co.uk/inflation-report/2019/august-2019. The data on the SEF has been considered as a mean aggregate to allow their use in our SS-framework without complicating the statistical model nested therein.

The results are reported in Table 5. The UK inflation forecasts are strategically non-coherent, despite the heterogeneity of the results according to the type of data: the Core CPI is affected by strategic judgment in all lags and by the Contribution of Energy to the CPI Inflation in the majority of the lags. Hence, the resulting SS-STARX is characterized by a significantly high slope parameter—in equation (15) of A2 in the Main Document, $\hat{\gamma} = 5.67$ with standard deviation 0.97. However, since the CPI Inflation is non-coherent only in a minority of the lags, we need further investigation to have a definitive assessment of these forecasts. For this purpose, we re-adapt the empirical example in Gneiting and Ranjan (2011) to compare the BoE's announcements with the equivalent professional forecasts. We consider two f and g two predictive densities, where f is the BoE announcements and g is the SEF quotation. The latter is assumed being the output of a GLSTAR(2) because the estimated density function from that model is asymmetric, see Zanetti Chini (2018).[3] The same reference illustrates the Monte Carlo procedure adopted for this family of models. Then the average scores

$$\bar{S}_n^f = \frac{1}{n-k-1} \sum_{t=m}^{m+n-k} S(\hat{f}_{t+k}, y_{t+k}), \quad \bar{S}_n^g = \frac{1}{n-k-1} \sum_{t=m}^{m+n-k} S(\hat{g}_{t+k}, y_{t+k}) \tag{7}$$

are computed by aggregating the sequences of forecasts generated by the pseudo-out-of-sample forecasting experiment where the sample is formed by $n = 72$ observations and the forecast horizon is $k = 4$ for uniformity with the evidence in Tab. 3. The null

[3]Since this model is potentially able to nest the more traditional STAR model, its use in this experiment does not constitute a loss of generality of the SS-STAR framework explained in Main Document.

hypothesis is that the two average scores are equal, so that the hypothesis system is

$$H_0 : \Delta^* = \bar{S}_n^f - \bar{S}_n^g = 0 \quad vs \quad H_1 : \Delta^* = \bar{S}_n^f - \bar{S}_n^g \neq 0, \tag{8}$$

which is measured by statistic

$$t_n = \sqrt{n}\frac{\Delta^*}{\hat{\sigma}_n} \sim N(0,1), \tag{9}$$

where $\hat{\sigma}_n^2 = \frac{1}{n-k+1}\sum_{j=-(k-1)}^{k-1}\sum_{t=m}^{m+n-k-|j|}\Delta_{t,k}\Delta_{t+|j|,k}$, and $\Delta_{t,k} = S_n^f - S_n^g$. In this exercise, the LogS is in negative orientation, so f is preferable to g if and only if $S^f < S^g$.

According to Table 6, there is not a clear superiority of the BoE forecasts with respect to the GLSTAR. In 2/3 of the cases the null hypothesis cannot be rejected. In the remaining cases, the BoE is superior to the nonlinear asymmetric model in only one in eight cases, that is the weighted pseudo-spherical score (WPseudoSph) with $\alpha = 1$, corresponding to a weighted logarithmic score.

3.2 The Norway Output Gap

The Bank of Norway's Monetary Policy Report (BoNMPR) issued probabilistic forecasts of OG from March 2008 to December 2017, using fan charts to visualize the deciles of the predictive distributions. The time series of quarterly OG investigated here is stated in percentage changes over twelve months; the first quarter extends from March 31 to May 30, while the second quarter extends from July 1 to September 30, and so on. The data corresponds to the 2014 release and can be downloaded at http://www.norges-bank.no/en/about/published/publications/monetary-policy-report/.

Also in this case, we take the BoN forecasts as primitive observations, so these are y_t in equation (15) of Appendix A2 of the Main Document. On the other side, the BoNMPR forecasts are the product of the bank's internal econometric model, such

as the System Averaging Model (SAM) or the Norway Economic Model (NEMO)[4]. The latter take the role of the composite transition variable $\mathbf{z_t}$. According to our SS-framework, when $\gamma = 0$, the final BoN announcements correspond to the estimated fan charts (that is, the latter are perfectly coherent with internal forecasts). Differently from the previous application on UK CPI inflation, Table 7 rejects this hypothesis in a minority of the lags here considered, so we conclude that the amount of SJB in the BoN's fan charts is negligible. In line with this finding, we assume that the FP adopts a Logistic STAR(1) model with small slope[5] to be compared with the final announcement, represented by downloadable BoN fan charts. Therefore, we repeat the analysis of (7) in the previous Subsection 3.1 with a window of length of $m = 6$ quarters. Under LogS, the t-statistic indicates whether the distance between the BoNMPR forecasts and a forecast obtained by an econometric model is significant.

Table 8 reports the results of this approach for a prediction horizon of $k = 1$ quarters ahead and a test period ranging from the first quarter of 2008 to the first quarter of 2017, for a total of $n = 34$ density forecast cases. According to the p-values, the superiority of the BoN approach is not unambiguously clear. Under LogS and other proper functional forms, such as Quantum (qS), Conditional Likelihood (CLS), and Interval Scores (IntS), the test rejects the null hypothesis of no equal predictive ability of SS versus the benchmark model, thus confirming the structural coherence of the quotation. On the other side, it does not reject the null hypothesis if any of several other non-proper functionals, such as the Weighted Power (WPwrS), most Weighted Pseudo-Spherical (WPseudoSph), and Log-Cosh (LCS) scores, are used; see the Supplement for the details of each SR here adopted.

4 Families of Scoring Structures

In Section 3 of the Main Document we introduced the general theory of Scoring Structure (SS). In our subsequent applications in Sections 5, we considered a large number

[4]See the BoN web page at http://www.norges-bank.no/en/Monetary-policy/ for references.
[5]According to our estimates,$\hat{\gamma} = 0.89$, and the standard deviation is 0.61.

of SRs, and each one can be adopted in the same SS. All these SRs are reported in Table 9 jointly with their associated entropy and divergence functions, their probabilistic measure and the bibliographic reference.

We remark that many of these SRs are not included in the Brègman family of functions that recent econometric theory requires in order to have coherent forecasts.

5 Extendend Proof of Proposition 1

Assume that the h-step ahead density forecasts $p(y)_{t+h}$ obtained from $p(\hat{\Theta}^{\Pi}; z_t)$ are structurally coherent. Hence, by Definition 2 there exists a one-to-one mapping f such that:

$$\mathcal{L}(\hat{\Theta}^{\Pi}; z_t) = f(\mathcal{L}(\hat{\Theta}^{\Psi}; z_t))$$

The log-likelihood $\mathcal{L}(\hat{\Theta}; z_t)$ of a model given observed data z_t is defined as:

$$\mathcal{L}(\hat{\Theta}; z_t) = \log p(z_t \mid \hat{\Theta})$$

For structural coherence, there must exist a one-to-one function f such that:

$$\log p(z_t \mid \hat{\Theta}^{\Pi}) = f(\log p(z_t \mid \hat{\Theta}^{\Psi}))$$

The function f maps the log-likelihoods of one model to another. To find a direct relationship between the probabilities, we need to invert the logarithmic transformation. Thus, if $\mathcal{L}(\hat{\Theta}; z_t) = \log p(z_t \mid \hat{\Theta})$, we can write:

$$p(z_t \mid \hat{\Theta}^{\Pi}) = \exp(f(\log p(z_t \mid \hat{\Theta}^{\Psi})))$$

For convenience, we define a new function g as:

$$g = \exp f$$

This new function g represents a one-to-one transformation directly between the probabilities of the two models:

$$p(z_t \mid \hat{\Theta}^{\Pi}) = g(p(z_t \mid \hat{\Theta}^{\Psi}))$$

By Definition 1, a set of assigned probabilities is coherent if it does not lead to certain loss through betting. Consider the (unconditional) probabilities assigned to the models, π and ψ:

$$P(\hat{\Theta}^{\Pi}) = \pi \quad \text{and} \quad P(\hat{\Theta}^{\Psi}) = \psi$$

The one-to-one mapping g implies that there are no configurations of the probabilities $P(\hat{\Theta}^{\Pi})$ and $P(\hat{\Theta}^{\Psi})$ that lead certain loss. That is, if:

$$\pi = g(\psi)$$

then the probabilities assigned to the models must be such that they do not allow for certain loss through betting, thus satisfying De Finetti's definition of coherence. Since structural coherence ensures that the evaluations of forecast quality are consistent via the one-to-one mapping f, and since this coherence translates directly into a one-to-one mapping between the probabilities via $g = \exp f$, the assigned probabilities cannot lead to certain loss. Hence the statement.

6 Proof of Proposition 3

(i) Let denote the log-likelihood function of the T observations by $\mathbf{\Lambda}_t(\mathbf{w_t}, \mathbf{\Xi})$ with $\mathbf{\Xi} = [\boldsymbol{\phi}, \boldsymbol{\theta}, \gamma, c]$ and the score vector by $\mathbf{\Sigma}_t(\mathbf{w_t}, \mathbf{\Xi})$ evaluated at $(\boldsymbol{\theta_0}, \boldsymbol{\phi_0}, 0, c_0)$. Then, standard results lead to the following log-likelihood function:

$$\mathbf{\Lambda}_t(\mathbf{z_t}, \mathbf{\Xi}) = const + \frac{T}{2} \ln \sigma^2 - \frac{1}{2}\sigma^2 \sum_t u_t^2(\mathbf{\Xi}), \tag{10}$$

with *const* and $u_t(\boldsymbol{\Xi}) = (y_t - \boldsymbol{\phi}'\mathbf{w_t} - \boldsymbol{\theta}'\mathbf{w}_t G)$ denoting a constant and the model's residual, respectively, and to the score:

$$\begin{aligned} \boldsymbol{\Sigma}_t(\mathbf{w_t}, \boldsymbol{\Xi}) &= \nabla_{\boldsymbol{\Xi}} \boldsymbol{\Lambda}_t(\mathbf{w_t}, \boldsymbol{\Xi}) = \frac{1}{\sigma^2} \sum_t u_t(\boldsymbol{\Xi}) \mathbf{d}_t, \\ \mathbf{d}_t &= \nabla_{\boldsymbol{\Xi}} u_t(\boldsymbol{\Xi}) = [\mathbf{w}_t, \mathbf{w}_t G, \boldsymbol{\theta}'\mathbf{w}_t G_\gamma, \boldsymbol{\theta}'\mathbf{w}_t G_c]^\mathsf{T}, \end{aligned} \tag{11}$$

with $G_\gamma = \partial G/\partial\gamma$ and $G_c = \partial G/\partial c$ denoting the first derivatives of G with respect to γ and c.

Moreover, let define: $\boldsymbol{\tau} = (\boldsymbol{\tau_1}, \tau_2)^\mathsf{T}$, where $\boldsymbol{\tau_1} = (\phi_0, \boldsymbol{\phi}^\mathsf{T})^\mathsf{T}$, $\tau_2 = \gamma$[1], $\hat{\boldsymbol{\tau}}_\mathbf{1}$ the LS estimator of $\boldsymbol{\tau_1}$ under $H_0 : \gamma = 0$, $\hat{\boldsymbol{\tau}} = (\boldsymbol{\tau}_\mathbf{1}', 0^\mathsf{T})^\mathsf{T}$ and $\hat{\mathbf{d}}_\mathbf{t} = \mathbf{d_t}(\hat{\boldsymbol{\tau}}) = (\hat{\mathbf{d}}_{\mathbf{1,t}}, \hat{d}_{2,t})$, where the partition conforms to that of $\boldsymbol{\tau}$, $\hat{\mathbf{D}}_\mathbf{i} = [\hat{\mathbf{d}}_{\mathbf{i1}}, \ldots, \hat{\mathbf{d}}_{\mathbf{it}}, \ldots, \hat{\mathbf{d}}_{\mathbf{iT}}]^\mathsf{T}$, $i = \{1, 2\}$, $t = 1, \ldots, T$, $\hat{\sigma}^2 = \frac{1}{T}\sum_1^T \hat{u}_t^2$ and $\hat{u}_t = y_t - \hat{\boldsymbol{\tau}}_\mathbf{1}^\mathsf{T}\mathbf{w_t}$. Then by standard Breusch and Pagan (1980) arguments, under H_0, the test statistic is the equation (15) in the Main Text.

When the nonlinear function $G(\cdot)$ is a logistic, $\hat{\mathbf{d}}_{\mathbf{1,t}} = -\mathbf{w_t} = -(1, y_{t-1}, \ldots, y_{t-p})^\mathsf{T}$ while $\hat{\mathbf{w}}_\mathbf{2t} \equiv \frac{\partial^2 u_t}{\partial\gamma\partial\gamma'}\big|_{\gamma=\mathbf{0}} = -\frac{1}{2}\big\{\theta_{20}[y_t(y_{t-d})] - c y_t \boldsymbol{\theta}'\mathbf{w_t} + \boldsymbol{\theta}_\mathbf{2}'\mathbf{w_t} y_t y_{t-d}\big\}$. Just minor modifications are needed in notation of $\hat{\mathbf{d}}_t$ and $\mathbf{s}_t^L$ in case of exponential or second-order-logistic model due to an additional c parameter with respect to the logistic model. The proposed test statistic depends on θ and is still unidentified unless $\theta_2 = 0$. This problem has been originally identified by Davies (1977).

(ii) Luukkonen et al. (1988) prove that the Davies' problem can be circumvented by linearizing the nonlinear model via (third order) Taylor expansion. Namely, let denote T_3 a third-order Taylor expansion operator. Then, the linearized LSTAR-SS

$$y_t = \phi'\mathbf{w_t} + \theta'\mathbf{w_t} T_3 G\big(\,\cdot\,\big)\epsilon_t', \tag{12}$$

leads to the following auxiliary regression for testing linearity:

$$\hat{\epsilon}'_t = \hat{\mathbf{w}}'_{\mathbf{1t}}\tilde{\boldsymbol{\beta}}_{\mathbf{1}} + \sum_{j=1}^{p}\beta_{2j}sy_{t-j}y_{t-d} + \sum_{j=1}^{p}\beta_{3j}sy_{t-j}y_{t-d}^2 + \sum_{j=1}^{p}\beta_{4j}sy_{t-j}y_{t-d}^3 + v_t, \;\; v_t \sim NIID(0,\sigma^2), \tag{13}$$

where: $\tilde{\boldsymbol{\beta}}_{\mathbf{1}} = (\beta_{10}, \boldsymbol{\beta}_{\mathbf{1}}^{\mathsf{T}})^{\mathsf{T}}$, $\beta_{10} = \phi_0 - (c/4)\theta_0$, $\boldsymbol{\beta}_{\mathbf{1}} = \boldsymbol{\phi} - (c/4)\boldsymbol{\theta} + (1/4)\theta_0\mathbf{e}_d$, $\mathbf{e}_d = (0,0,\ldots,0,1,0,\ldots,0)^{\mathsf{T}}$ with the d-th element equal to unit and $T_3(G) = f_1G + f_3G^3$ is the third-order Taylor expansion of $G(\boldsymbol{\Xi})$, $f_1 = \partial G(\boldsymbol{\Xi})/\partial\boldsymbol{\Xi}\big|_{\gamma=0}$ and $f_3 = (1/6)\partial^3 G(\boldsymbol{\Xi})/\partial\boldsymbol{\Xi}\big|_{\gamma=0}$, $\boldsymbol{\Xi}$ being defined above. Hence, the null hypothesis of locality becomes testable by following hypothesis system:

$$H_0 : \beta_{2j} = \beta_{3j} = \beta_{4j} = 0 \;\; j = 1,\ldots,p, \quad \text{vs} \quad H_1 : \beta_{2j} = \beta_{3j} = \beta_{4j} \neq 0 \tag{14}$$

corresponding to statistic LM_1 in equation (17) of the Main Text, with SSR_0 and SSR denoting the sum of squared estimated residuals from the estimated auxiliary regression (13) and under the null and alternative, respectively and $\sigma_v^2 = (1/T)SSR$, has an asymptotic χ^2_{3p} distribution under H_0.

If the model is a ESTAR(p), then it is possible to show that the corresponding auxiliary regression is

$$\hat{\epsilon}'_t = \tilde{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathsf{T}}\hat{\mathbf{w}}_{\mathbf{1}} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathsf{T}}\mathbf{w}_{\mathbf{t}}sy_{t-d} + \boldsymbol{\beta}_{\mathbf{3}}^{\mathsf{T}}\mathbf{w}_{\mathbf{t}}sy_{t-d}^2 + v'_t, \;\; v'_t \sim NIID(0,\sigma^2), \tag{15}$$

where $\tilde{\boldsymbol{\beta}}_{\mathbf{1}} = (\beta_{10}, \boldsymbol{\beta}'_{\mathbf{1}})'$, with $\beta_{10} = \phi_0 - c^2\theta_0$ and $\boldsymbol{\beta}_{\mathbf{1}} = \boldsymbol{\phi} - c^2\boldsymbol{\theta} + 2c\theta_0\boldsymbol{e}_d$; moreover $\boldsymbol{\beta}_{\mathbf{2}} = 2c\boldsymbol{\theta} - \theta_0\mathbf{e}_d$ and $\boldsymbol{\beta}_{\mathbf{3}} = -\boldsymbol{\theta}$. Thus the null hypothesis of linearity is

$$H'_0 : \boldsymbol{\beta}_{\mathbf{2}} = \boldsymbol{\beta}_{\mathbf{3}} = 0 \quad \text{vs} \quad H'_1 : \boldsymbol{\beta}_{\mathbf{2}} = \boldsymbol{\beta}_{\mathbf{3}} \neq 0 \tag{16}$$

which can be tested by the test statistic LM_2 in equation (17) of the Main Text, where SSR_0 and SSR are the sum of squared residuals from (15) under the null and the alternative respectively, $\hat{\sigma}_{v1}^2 = (1/T)SSR$. A peculiar case of (16) is when

$\boldsymbol{\beta}_2 = 0$ as $\theta_0 = c = 0$, in which case, the null becomes

$$H_0'' : \boldsymbol{\beta_3} = 0 \quad \text{vs} \quad H_1'' : \boldsymbol{\beta_3} \neq 0 \tag{17}$$

which test statistic corresponds to statistic LM_3 in equation (17) of the Main Text, with SSR_0, SSR and σ_{v2} defined in a similar way with respect to the LM_2 case. As well known in the literature, F-version of LM_1, LM_2 and LM_3, denoted as F_1, F_2 and F_3, may be preferable when testing (14) or (16) or (17) in order to preserve power in low samples; in this case the F-statistics has n and $T - p - n$ degrees of freedom. In practice the form of G is not known by the investigator; see CH 5.3 in Teräsvirta et al. (2010) among others. Teräsvirta (1994) proposes a battery of F-tests on the auxiliary model (13):

$$\begin{aligned} &H_{01} : \beta_4 = 0 \quad \text{vs} \quad H_{11} : \beta_4 \neq 0 \\ &H_{02} : \beta_3 = 0 | \beta_4 = 0 \quad \text{vs} \quad H_{12} : \beta_3 \neq 0 | \beta_4 = 0 \\ &H_{03} : \beta_2 = 0 | \beta_3 = 0 \text{ and } \beta_4 = 0 \quad \text{vs} \quad H_{22} : \beta_2 \neq 0 | \beta_3 = 0 \text{ and } \beta_4 = 0. \end{aligned} \tag{18}$$

and suggests an empirical rule – based on the results of a simulation experiment – to select the right transition function. For our aims, however, this is not a crucial issue, so we will not discuss further details. This ends the proof.

7 Estimation

In the line of Teräsvirta (1994) the estimation of (11) in the Main Text is done via conditional least squares (CLS) by concentrating the sum of the square residuals function with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, that is minimizing:

$$SSR = \sum_{t=1}^{T} \left(y_t - \hat{\boldsymbol{\psi}}' \xi_{\mathbf{t}}' \right)^2, \tag{19}$$

where:

$$\hat{\boldsymbol{\psi}} = [\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}] = \left(\sum_{t=1}^{T} \mathbf{x}'_{\mathbf{t}}(\gamma, c)\xi_{\mathbf{t}}(\gamma, c) \right)^{-1} \left(\sum_{t=1}^{T} \xi'_{\mathbf{t}}(\gamma, c) y_t \right), \tag{20}$$

and

$$\xi_{\mathbf{t}}(\hat{\gamma}, \hat{c}) = \Big[\mathbf{w_t}, \mathbf{w}'_{\mathbf{t}} G(\cdot) \Big]. \tag{21}$$

This is possible because if $\boldsymbol{\gamma}$ and c are known and fixed, the GSTAR model is linear in $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, which can be easily computed. In such a way, the nonlinear least square minimization problem, otherwise necessary, computationally more demanding and not available in closed-form, is reduced to a minimization on two parameters, and is solved via a grid search over γ, c.[2]

8 Tables and Graphs

Table 1: Empirical Size and Power of LM test for Coherence for different slope parameters

Empirical Size

T		DGP 1						DGP 2					
		F_1			F_2			F_1			F_2		
		$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
75		0.0015	0.0078	0.0207	0.0032	0.0239	0.0643	0.0093	0.0415	0.0625	0.0108	0.0461	0.0715
150		0.017	0.0234	0.0399	0.0085	0.0387	0.0692	0.0110	0.0469	0.0748	0.0133	0.0550	0.0810
300		0.0020	0.0340	0.0591	0.0106	0.0444	0.0744	0.0150	0.0525	0.0917	0.0172	0.0577	0.0902

Empirical Power

T	γ	F_1			F_2			F_1			F_2		
		$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
	0.1	0.0009	0.0083	0.0185	0.0505	0.0511	0.0525	0.0004	0.0004	0.0012	0.0208	0.0411	0.0616
	0.5	0.0009	0.0066	0.0192	0.0492	0.0497	0.0501	0.0082	0.0291	0.0498	0.0360	0.0551	0.0910
	1	0.0029	0.0132	0.0217	0.0982	0.0098	0.0103	0.1553	0.3588	0.4658	0.1566	0.2354	0.3432
	5	0.1184	0.2163	0.2958	0.0990	0.2026	0.2879	0.5436	0.6207	0.6593	0.4495	0.5254	0.6538
75	10	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
	50	0.2775	0.4630	0.5737	0.2338	0.4104	0.5286	0.7713	0.8395	0.8677	0.6608	0.7588	0.8067
	100	0.2871	0.4738	0.5992	0.2385	0.4261	0.5457	0.7721	0.8441	0.8699	0.6702	0.7648	0.8063
	500	0.3060	0.4836	0.6104	0.2531	0.4386	0.5560	0.7844	0.8484	0.8789	0.6727	0.7787	0.8179
	0.1	0.0001	0.0037	0.0081	0.0424	0.0429	0.0429	0.0001	0.0001	0.0001	0.0222	0.0223	0.0223
	0.5	0.0006	0.0019	0.0046	0.0253	0.0253	0.0254	0.0062	0.0162	0.0291	0.0504	0.0535	0.0548
	1	0.0002	0.0029	0.0043	0.0092	0.0110	0.0116	0.2933	0.5335	0.6510	0.0326	0.0365	0.0392
	5	0.1571	0.2714	0.3496	0.1489	0.2527	0.3103	0.7577	0.7747	0.7814	0.7230	0.7324	0.7345
150	10	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
	50	0.4340	0.6340	0.7414	0.4360	0.6377	0.7410	0.9840	0.9863	0.9872	0.9709	0.9788	0.9798
	100	0.4391	0.6437	0.7502	0.4393	0.6578	0.7605	0.9836	0.9867	0.9874	0.9761	0.9817	0.9820
	500	0.4577	0.6679	0.7728	0.4617	0.6744	0.7755	0.9859	0.9891	0.9899	0.9773	0.9831	0.9844
	0.1	0.0000	0.0001	0.0005	0.0461	0.0462	0.0463	0.0001	0.0001	0.0001	0.0188	0.0188	0.0190
	0.5	0.0000	0.0001	0.0001	0.0374	0.0375	0.0377	0.0029	0.0051	0.0083	0.0685	0.0691	0.0695
	1	0.0000	0.0001	0.0001	0.0213	0.0216	0.0221	0.5055	0.7431	0.8299	0.0794	0.0804	0.0822
	5	0.1909	0.3085	0.3694	0.1680	0.2425	0.2858	0.9139	0.9152	0.9170	0.6261	0.6742	0.7244
300	10	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.6626	0.7691	0.7694
	50	0.7458	0.8588	0.8894	0.7487	0.8460	0.8753	0.9980	0.9981	0.9981	0.6812	0.7700	0.7879
	100	0.7752	0.8740	0.9035	0.7830	0.8621	0.8892	0.9997	0.9998	0.9999	0.7094	0.7703	0.7901
	500	0.7890	0.8916	0.9214	0.7998	0.8796	0.9035	0.9978	1.0000	1.0000	0.7248	0.7675	0.8048

NOTE: This table reports the results of the Monte Carlo simulation experiment described in Section 2, where in equations (5) and (6) the parameter = 10 is fixed and the functional form of the SR varies. F_1 and F_2 are the F-type statistics that correspond to LM_1 and LM_2 in equation (17) of the Main Text. In this experiment, the first 100 simulations were discarded to avoid the initialization effect. Software used: MATLAB R2009b.

Table 2: Empirical Power of LM test for the null hypothesis of coherence for different scoring rules and $\gamma = 10$

S(p, x)	DGP 1						DGP 2					
	F_1			F_2			F_1			F_2		
$T = 75$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$	$\alpha = .01$	$\alpha = .05$	$\alpha = .10$
QSR	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPwrS (General)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPwrS ($\beta = -1$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPwrS ($\beta = 0$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPwrS ($\beta = 1/2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPwrS ($\beta = 2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
PsdSphS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPsdSph	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPsdSphS ($\beta = -1$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPsdSphS ($\beta = 0$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPsdSphS ($\beta = 1/2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
WPsdSphS ($\beta = 2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
TsallisS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
ES	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
GES	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
PSpctr	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
CRPS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
QuantS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
HS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.5326	0.6755	0.7580	0.5978	0.6845	0.7210
$T = 150$												
QSR	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6152	0.7250	0.7987	0.6810	0.7792	0.8259
WPwrS (General)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPwrS ($\beta = -1$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPwrS ($\beta = 0$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPwrS ($\beta = 1/2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPwrS ($\beta = 2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
PsdSphS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPsdSph	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPsdSphS ($\beta = -1$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPsdSphS ($\beta = 0$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPsdSphS ($\beta = 1/2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
WPsdSphS ($\beta = 2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
TsallisS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
ES	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
GES	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
PSpctr	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
CRPS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
QuantS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
HS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.6162	0.7250	0.9287	0.6810	0.7792	0.8259
$T = 300$												
QSR	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPwrS (General)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPwrS ($\beta = -1$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPwrS ($\beta = 0$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPwrS ($\beta = 1/2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPwrS ($\beta = 2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
PsdSphS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPsdSph	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPsdSphS ($\beta = -1$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPsdSphS ($\beta = 0$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPsdSphS ($\beta = 1/2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
WPsdSphS ($\beta = 2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
TsallisS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
ES	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
GES	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
PSpctr	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
CRPS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
QuantS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045
HS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.7849	0.8570	0.9226	0.7662	0.8450	0.9045

NOTE: This table reports the results of the Monte Carlo simulation experiment described in Section 2 , where in equations (5) and (6) the parameter $\gamma = 10$ is fixed and the functional form of the SR varies. F_1 and F_2 are the F-type statistics that correspond to LM_1 and LM_2 in equation (17) of the Main Document. In this experiment, the first 100 simulations were discarded to avoid the initialization effect. Software used: MATLAB R2009b.

Table 3: Empirical Size and Empirical Power of tests for serial correlation, no additive incoherence and parameter constancy under DGP 1.

Empirical Size

T	γ_1	γ_2	Nominal size	No error autocorrelation $\rho = 0$				No additional incoherence	Parameter constancy		
				q=1	q=2	q=4	q=10	H_0	LM_1	LM_2	LM_3
			$\alpha = 0.01$	0.0043	0.0037	0.0072	0.0053	0.0030	0.0012	0.0005	0.0010
	2		$\alpha = 0.05$	0.0628	0.0482	0.0354	0.0435	0.0085	0.0011	0.0016	0.0021
			$\alpha = 0.10$	0.0991	0.0908	0.0901	0.0895	0.0189	0.0022	0.0031	0.0029
			$\alpha = 0.01$	0.0045	0.0084	0.0070	0.0040	0.0064	0.0020	0.0021	0.0025
100	20		$\alpha = 0.05$	0.0526	0.0392	0.0186	0.0167	0.0107	0.0035	0.0034	0.0033
			$\alpha = 0.10$	0.0970	0.0953	0.0515	0.0393	0.0520	0.0059	0.0058	0.0082
			$\alpha = 0.01$	0.0094	0.0109	0.0108	0.0097	0.0089	0.0067	0.0073	0.0071
	200		$\alpha = 0.05$	0.0479	0.0489	0.0529	0.0518	0.0493	0.0089	0.0143	0.0112
			$\alpha = 0.10$	0.0981	0.0994	0.1108	0.0123	0.0941	0.0199	0.0212	0.0203
			$\alpha = 0.01$	0.0116	0.0093	0.0112	0.0057	0.0064	0.0063	0.0056	0.0051
	2		$\alpha = 0.05$	0.0502	0.0457	0.0416	0.0436	0.0095	0.0453	0.0398	0.0233
			$\alpha = 0.10$	0.1028	0.0991	0.1054	0.0983	0.0134	0.0707	0.0730	0.0681
			$\alpha = 0.01$	0.0081	0.0084	0.0043	0.0074	0.0088	0.0076	0.0065	0.0071
300	20		$\alpha = 0.05$	0.0507	0.0563	0.0287	0.0312	0.0258	0.0434	0.0376	0.0297
			$\alpha = 0.10$	0.1045	0.0910	0.0608	0.0441	0.0849	0.0904	0.0923	0.0907
			$\alpha = 0.01$	0.0250	0.0137	0.0149	0.0000	0.0142	0.0034	0.0034	0.0056
	200		$\alpha = 0.05$	0.0621	0.0434	0.0568	0.0208	0.0657	0.0486	0.0493	0.0644
			$\alpha = 0.10$	0.1321	0.1125	0.0995	0.0356	0.1033	0.1064	0.1057	0.1002

Empirical Power

T	γ_1	γ_2	Nominal size	No error autocorrelation $\rho = 0.2$				No error autocorrelation $\rho = 0.4$				No additional incoherence	Parameter constancy		
				q=1	q=2	q=4	q=10	q=1	q=2	q=4	q=10	H_1	LM_1	LM_2	LM_3
			$\alpha = 0.01$	0.0158	0.0574	0.1420	0.5645	0.0576	0.0795	0.8193	0.9016	0.4052	0.2563	0.3721	0.5676
	2	5	$\alpha = 0.05$	0.0950	0.1120	0.3688	0.6425	0.0840	0.1411	0.9535	0.9425	0.5026	0.2844	0.4577	0.6941
			$\alpha = 0.10$	0.1111	0.1552	0.5023	0.7119	0.1830	0.2073	0.9356	0.9994	0.6730	0.4055	0.5698	0.7211
			$\alpha = 0.01$	0.1342	0.1538	0.2834	0.5888	0.1945	0.0228	0.7557	0.8940	0.6047	0.2735	0.3440	0.5860
100	20	10	. $\alpha = 0.05$	0.1834	0.2621	0.4024	0.7090	0.2777	0.3044	0.7053	0.9122	0.7361	0.3808	0.5439	0.7338
			$\alpha = 0.10$	0.2055	0.2127	0.4628	0.7900	0.3371	0.4883	0.8444	0.9956	0.8096	0.4330	0.5399	0.7632
			$\alpha = 0.01$	0.0185	0.0488	0.1234	0.5540	0.1259	0.1340	0.6363	0.8821	0.8851	0.7540	0.8024	0.8643
	200	100	$\alpha = 0.05$	0.0599	0.2113	0.2953	0.7341	0.1780	0.9180	0.7035	0.9333	0.9200	0.8021	0.8770	0.9100
			$\alpha = 0.10$	0.1425	0.1146	0.5334	0.8176	0.1946	0.9567	0.8490	0.9949	0.9762	0.8761	0.8934	0.9116
			$\alpha = 0.01$	0.0290	0.0312	0.6014	0.7642	0.1047	0.1235	1.0000	1.0000	0.6118	0.0215	0.3353	0.7728
	2	1.	$\alpha = 0.05$	0.0940	0.0953	0.7970	0.8430	0.2974	0.2724	1.0000	1.0000	0.6535	0.0835	0.6957	0.9442
			$\alpha = 0.10$	0.1859	0.1914	0.8667	0.9045	0.4006	0.4229	1.0000	1.0000	0.7149	0.1804	0.8281	0.9917
			$\alpha = 0.01$	0.0987	0.0810	0.7414	0.9968	0.1190	0.1331	1.0000	1.0000	0.7747	0.0215	0.3353	0.7728
300	20	10	$\alpha = 0.05$	0.1866	0.1652	0.8827	0.9980	0.2230	0.2448	1.0000	1.0000	0.8161	0.0835	0.6957	0.9442
			$\alpha = 0.10$	0.2372	0.2323	0.9069	1.0000	0.3036	0.3172	1.0000	1.0000	0.9496	0.1804	0.8281	0.9940
			$\alpha = 0.01$	0.1369	0.1492	0.7963	0.9042	0.0860	0.1422	1.0000	1.0000	1.0000	0.8994	0.8941	0.8778
	200	100	$\alpha = 0.05$	0.2135	0.3108	0.8420	0.9641	0.2156	0.2980	1.0000	1.0000	1.0000	0.9305	0.9550	0.9910
			$\alpha = 0.10$	0.4001	0.3650	0.6823	0.9911	0.3349	0.4364	1.0000	1.0000	1.0000	0.9883	0.9973	1.0000

Table 4: Empirical Size and Empirical Power of tests for serial correlation, no additive nonlinearity and parameter constancy under DGP 2.

Empirical Size

T	γ_1	γ_2	Nominal size	No error autocorrelation $\rho = 0$				No additional nonlinearity	Parameter constancy		
				q=1	q=2	q=4	q=10	H_0	LM_1	LM_2	LM_3
100	2		$\alpha = 0.01$	0.0093	0.0091	0.0047	0.0065	0.0081	0.0023	0.0031	0.0025
			$\alpha = 0.05$	0.0571	0.0520	0.0433	0.0413	0.0434	0.0062	0.0050	0.0065
			$\alpha = 0.10$	0.1230	0.1130	0.0993	0.0900	0.0963	0.0150	0.0133	0.0111
	20		$\alpha = 0.01$	0.0122	0.0134	0.0070	0.0040	0.0050	0.0053	0.0045	0.0051
			$\alpha = 0.05$	0.0536	0.0498	0.0422	0.0167	0.0357	0.0085	0.0083	0.0091
			$\alpha = 0.10$	0.0630	0.0782	0.0830	0.0393	0.1020	0.0137	0.0124	0.0144
	200		$\alpha = 0.01$	0.0110	0.0109	0.0088	0.0069	0.0094	0.0098	0.0094	0.0096
			$\alpha = 0.05$	0.0502	0.0578	0.0510	0.0445	0.0330	0.0277	0.0253	0.0220
			$\alpha = 0.10$	0.1012	0.1044	0.0878	0.0799	0.0405	0.0625	0.0592	0.0595
300	2		$\alpha = 0.01$	0.0126	0.0120	0.0092	0.0060	0.0092	0.0083	0.0080	0.0094
			$\alpha = 0.05$	0.0504	0.0585	0.0526	0.0412	0.0455	0.0502	0.0451	0.0323
			$\alpha = 0.10$	0.1084	0.1139	0.0955	0.0883	0.1202	0.0920	0.0891	0.0981
	20		$\alpha = 0.01$	0.0093	0.0094	0.0043	0.0024	0.0118	0.0045	0.0030	0.0071
			$\alpha = 0.05$	0.0530	0.0510	0.0422	0.0390	0.0550	0.0399	0.0401	0.0432
			$\alpha = 0.10$	0.1105	0.1210	0.0841	0.0774	0.1099	0.0799	0.0723	0.0707
	200		$\alpha = 0.01$	0.0250	0.0137	0.0149	0.0000	0.0142	0.0040	0.0025	0.0055
			$\alpha = 0.05$	0.0621	0.0543	0.0488	0.0208	0.0557	0.0511	0.0388	0.0400
			$\alpha = 0.10$	0.1221	0.1124	0.0955	0.0356	0.1083	0.1045	0.0957	0.0946

Empirical Power

T	γ_1	γ_2	Nominal size	No error autocorrelation $\rho = 0.2$				No error autocorrelation $\rho = 0.4$				No additional nonlinearity	Parameter constancy		
				q=1	q=2	q=4	q=10	q=1	q=2	q=4	q=10	H_1	LM_1	LM_2	LM_3
100	2	5	$\alpha = 0.01$	0.0165	0.0774	0.2261	0.4420	0.1024	0.1233	0.5560	0.6023	0.1807	0.3500	0.3274	0.3560
			$\alpha = 0.05$	0.1283	0.1551	0.3738	0.6182	0.1555	0.2747	0.6334	0.7224	0.2526	0.2457	0.3978	0.4810
			$\alpha = 0.10$	0.1561	0.1860	0.4940	0.7721	0.2545	0.4222	0.8033	0.8331	0.3630	0.3593	0.4554	0.6118
	20	10	$\alpha = 0.01$	0.0335	0.0944	0.1363	0.3825	0.1047	0.2230	0.4500	0.6922	0.2059	0.4676	0.4777	0.4226
			$\alpha = 0.05$	0.0922	0.1031	0.1544	0.5022	0.0777	0.3034	0.5233	0.7559	0.3192	0.6250	0.5450	0.5957
			$\alpha = 0.10$	0.1406	0.1437	0.1932	0.6222	0.1371	0.3852	0.6115	0.7993	0.4342	0.7480	0.7621	0.7517
	200	100	$\alpha = 0.01$	0.0675	0.1213	0.2934	0.6540	0.1312	0.2290	0.7112	0.8225	0.7884	0.6271	0.6432	0.6933
			$\alpha = 0.05$	0.1439	0.1835	0.4145	0.7341	0.3425	0.6635	0.8440	0.9640	0.8542	0.7439	0.7330	0.6233
			$\alpha = 0.10$	0.2227	0.3346	0.5084	0.8176	0.6462	0.8747	0.9225	1.0000	0.9442	0.8250	0.8034	0.8119
300	2	1.	$\alpha = 0.01$	0.0290	0.0312	0.5014	0.6022	0.1047	0.1235	0.4020	0.4224	0.5872	0.5607	0.5427	0.5744
			$\alpha = 0.05$	0.0940	0.0953	0.5670	0.7151	0.2974	0.2724	0.5652	0.6431	0.6743	0.6541	0.6407	0.6142
			$\alpha = 0.10$	0.1459	0.1614	0.6267	0.8232	0.4006	0.4229	0.6972	0.7323	0.7028	0.7030	0.7331	0.7244
	20	10	$\alpha = 0.01$	0.0987	0.0810	0.4414	0.8768	0.0890	0.1331	0.6842	0.8736	0.6225	0.7565	0.7353	0.7458
			$\alpha = 0.05$	0.1866	0.1652	0.5827	0.8680	0.2430	0.2848	0.7644	0.9042	0.6992	0.8021	0.7954	0.7885
			$\alpha = 0.10$	0.2372	0.2323	0.7069	0.8940	0.3436	0.4272	0.8240	0.9300	0.7736	0.8825	0.8778	0.8844
	200	100	$\alpha = 0.01$	0.0746	0.1045	0.5735	0.8341	0.0845	0.1560	0.7994	0.9310	0.8454	0.9031	0.8967	0.8856
			$\alpha = 0.05$	0.1015	0.1778	0.6421	0.8663	0.2156	0.2980	0.8350	0.9667	0.9040	0.9661	0.9467	0.9360
			$\alpha = 0.10$	0.1795	0.2341	0.7024	0.9115	0.2449	0.3364	0.9130	0.9956	0.9539	0.9901	0.9567	0.9612

Table 5: Structural Coherence test for U.K. Inflation

d	CPI Inflation		Core CPI Inflation		Contribution of Energy	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**1.500**	**0.049**	**2.190**	**0.001**	**3.526**	**<0.001**
2	1.351	0.109	**2.075**	**0.015**	**2.822**	**<0.001**
3	**1.374**	**0.097**	**1.763**	**0.011**	**2.550**	**<0.001**
4	**1.386**	**0.091**	**1.524**	**0.042**	**1.968**	**0.003**
5	1.230	0.198	**1.477**	**0.055**	**1.604**	**0.028**
6	1.199	0.229	**1.498**	**0.049**	**1.427**	**0.073**
7	1.105	0.342	**1.402**	**0.083**	1.250	0.181
8	1.083	0.372	**1.566**	**0.034**	1.035	0.449

NOTE: This table reports the test statistics LM_1 in equation (18) of the Main Document in their $F_{72,65}$,-variant, with corresponding p-values for data on the U.K. CPI Inflation forecasts according to the MPC Inflation Report of August 2019. Rejections are illustrated in bold.

Table 6: Relative predictive ability of Forecaster's quotations for UK CPI

S(Q, y)	$\bar{S}^f$	$\bar{S}^g$	σ	t	P-value
LogS	0.068	0.075	0.002	**302.50**	**<0.001**
QSR	0.167	0.113	0.356	**468.783**	**<0.001**
WPowerS	0.356	0.344	0.270	**44,444**	**<0.001**
" ($\alpha = -1$)	3.456	3.780	1.405	-0.231	0.591
" ($\alpha = 0$)	0.0124	0.131	0.002	**302.50**	**<0.001**
" ($\alpha = 1/2$)	3.889	3.888	1.000	0	0.500
" ($\alpha = 1$)	120.050	2.855	3.569	**32.891**	**<0.001**
" ($\alpha = 2$)	2.240	2.256	0.154	-0.101	0.540
PseudoSph	3.884	4.000	0.092	-1.260	0.890
WPseudoSph	2.775	2.200	1.159	0.496	0.311
" ($\alpha = -1$)	1.050	0.996	0.004	**13.500**	**<0.001**
" ($\alpha = 0$)	1.000	1.000	1.000	0.000	0.500
" ($\alpha = 1/2$)	-3,504.67	235.666	5750.999	0.568	0.286
" ($\alpha = 1$)	0.855	0.875	0.005	**-4.000**	**<0.001**
" ($\alpha = 2$)	306.440	357.900	900.550	-0.057	0.714
IntS	0.990	2.500	1.060	-1.424	0.920
TsallisS	2.046	2.046	1.000	0.000	0.500
ES	-4.780	2.900	18.251	0.000	0.499
GES	16.757	17.700	25.099	-0.037	0.515
PseudoSpectrumS	13.330	1.560	8.950	1.315	0.807
CRPS	0.889	0.046	0.009	**93.666**	**<0.001**
QuantS	-0.660	0.350	0.315	-3.206	0.998
CLS	0.485	1.290	1.355	-0.594	0.722
CsLS	1.080	2.560	3.069	-0.484	0.685
LCS	0.445	0.543	0.037	-2.648	0.995

NOTE: This table reports the result of the Amisano and Giacomini's 2007 test for the BoN's fan chart of OG at a prediction horizon of one year for different SRs of two density forecasts, f and g, respectively, where f is the BoE announcement and g is a non-linear specification (corresponding to a STAR(2)) given a non-local SS. Rejections are marked in bold.

Table 7: Structural Coherence test for Norway's Output Gap

d	LM_1		LM_2		LM_3	
	F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
1	**2.391**	**0.009**	**3.582**	**<0.001**	1.326	0.221
2	**2.510**	**0.007**	**2.335**	**0.011**	**1.890**	**0.041**
3	1.544	0.118	**1.991**	**0.030**	1.006	0.497
4	1.032	0.469	1.409	0.174	1.294	0.241
5	1.205	0.694	1.634	0.090	1.128	0.372
6	1.078	0.421	1.083	0.416	1.080	0.419
7	0.688	0.853	0.777	0.761	0.944	0.567
8	0.530	0.961	0.796	0.740	1.045	0.455

NOTE: This table reports the test statistics $F_{34,29}$ described in eq. (18) of the Main Text with corresponding p-values for data on the Norges Bank Output Growth is estimated according to the fan charts published in the release of January 2014. Rejections are illustrated in bold.

Table 8: Relative predictive ability of Forecaster's quotations for OG

S(Q, y)	$\bar{S}^f$	$\bar{S}^g$	σ	t	P-value
LogS	0.028	0.027	0.007	**2.560**	**<0.001**
QSR	0.558	0.539	0.096	-0.325	0.382
WPowerS	2.815	2.953	0.101	-3.358	0.999
" ($\alpha = -1$)	527.677	2.953	1.08e05	-0.003	0.501
" ($\alpha = 0$)	527.519	670.871	1.08e05	-0.003	0.501
" ($\alpha = 1/2$)	-1,062.30	-1,352.763	4.43e05	0.002	0.499
" ($\alpha = 1$)	1.199	1.448	0.328	-1.866	0.965
" ($\alpha = 2$)	-262.602	-333.808	2.6e03	0.007	0.497
PseudoSph	2.982	2.982	1.000	0.000	0.500
WPseudoSph	1.992	1.993	1.27e-05	-299.568	1.000
" ($\alpha = -1$)	0.499	0.500	7.87e-12	-3.8e05	1.000
" ($\alpha = 0$)	1.000	1.000	1.000	0.000	0.500
" ($\alpha = 1/2$)	-1,927.528	1.000	3.8e06	0.001	0.499
" ($\alpha = 1$)	1.199	1.448	0.328	-1.866	0.965
" ($\alpha = 2$)	-0.856	-0.836	0.002	-23.461	1.000
IntS	3.599	3.500	0.005	**135.250**	**<0.001**
TsallisS	1.223	1.223	1.000	0.000	0.500
ES	-0.124	-0.083	0.009	-11.571	1.000
GES	1.163	1.249	0.039	-5.420	1.000
PseudoSpectrumS	-7.8530	-7.853	1.000	0.000	0.500
CRPS	0.0132	0.012	7.276e-06	**395.962**	**<0.001**
QuantS	-0.184	-0.192	1e04	**63.517**	**<0.001**
CLS	-0.147	-0.423	0.402	1.684	0.049
CsLS	0.009	0.008	8.08e-06	**375.748**	**<0.001**
LCS	0.055	0.057	0.011	-0.210	0.583

NOTE: This table reports the result of the Amisano and Giacomini's 2007 test for the BoN's fan chart of OG at a prediction horizon of 12 months for different SRs of two density forecasts, f and g, respectively, where f is the BoN announcement and g is a non-linear specification (corresponding to the same STAR(1)) given a non-local SS. Rejections are illustrated in bold.

Table 9: Scoring Rules for density of continuous variables and their features

Score	S(P,x)	H(p,x)	Measure	$d(P,Q)$	Brègman-Savage type	Reference
QS	$2p(x) - \|\|p\|\|_2^2$	$\|\|p(x)\|\|_2^2$	L_2	$\|\|p-q\|\|_2^2$	Yes	Brier (1950)
LogS	$k \log p(x)$	$\sum_{j=1}^m p \log p$	L_2	$\sum_j q_j \ln(\frac{q}{p})$	Yes	Good (1952)
RPS	$\int \{\|Q(A_t) - 1_{A_t}(x))\}^2 d\mu(t)$	$\int P(A_t)\{1-P(A_t)\}d\mu(t)$	μ	$\int \{P(A_t)Q(A_t)\}^2 d\mu(t)$	No	Epstein (1969)
PseudoSph	$\frac{p(x)^{\alpha-1}}{\|\|p\|\|_\alpha^{1-\alpha}}$	$\|\|p\|\|_\alpha$	L_α	$\|\|p\|\|_\alpha$	No	Good (1971)
IntS	$(u-l) + \frac{2}{\alpha}(l-x)I_{(x<l)} + + \frac{2}{\alpha}(x-u)I_{(x>u)}$	$\int S_\alpha^{int} dp(x)$	$\mathcal{P}$	$\|\|p\|\|_\alpha$	No	Winkler (1972)
CRPS	$\frac{1}{2}E_F\|\|X-X'\| - E_F\|\|X-x\|\|$	$\frac{1}{2}E_F\|\|X-X'\|\|$	$\mathcal{P}_1$	$\int_{-\infty}^{+\infty}(F(x)-G(x))$	No	Matheson and Winkler (1972)
TsallisS	$\frac{k}{d(x)-1}\sum_{t=1}^W p_i(x)(1-p_i(x)^{d-1})$	$-\sum p(x)^d$	L	$\sum p(x)q(x)^{(d-1)} - -(d-1)H(Q) - H(P)$	Yes	Tsallis (1988)
PseudoSpectrum	$-\|\phi_P(\mathbf{y}) - e^{i<\mathbf{x},\mathbf{y}>}\|^2$	$-\|\phi_P(\mathbf{y})\|$	$\mathcal{P}$	$\int_u \|\|\alpha-\beta\|\|^2$	No	Eaton et al. (1996)
DispersionS	$K(Q_V) + tr\{\boldsymbol{V}_P\boldsymbol{\Gamma}_P^{-1} - \boldsymbol{V}_Q\boldsymbol{\Gamma}_Q\} -(\mathbf{x}-\boldsymbol{\mu}_P)'\boldsymbol{\Gamma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)$	$-\log\det\boldsymbol{\Gamma}_P - mK$	$\mathcal{P}$	$tr(\boldsymbol{\Gamma}_P^{-1}\boldsymbol{\Gamma}_Q) - \log\det(\boldsymbol{\Gamma}_P - 1\boldsymbol{\Gamma}_Q) + + (\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q)'\boldsymbol{\Gamma}_P^{-1}(\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q) - K$	Yes	Dawid and Sebastiani (1999)
GMR	$\sum_{i=1}^M w_{i,t}p_{i,t}$	$\int(\frac{1}{\eta(\eta+1)}(x^{\eta+1}-1))(\frac{dp}{dq})dq$	L	$\int(1-(p/q)^\eta)qd\nu$	Yes	Granger et al. (2004)
Hyvaärinen	$((\ln q)'(x)^2 + 2(\ln q)''(y)$	$E_P p(x) \bigtriangledown \ln p(x)$	L	$\frac{1}{2}\int p(x)\| \bigtriangledown \ln p(x) - \bigtriangledown \ln q(x)\|dx$	Yes	Hyvärinen (2005)
ES	$\frac{1}{2}E_F\|\|\mathbf{X}-\mathbf{X}'\|\|^\beta - E_F\|\|\mathbf{X}-\mathbf{x}\|\|^\beta$	$\frac{1}{2}E_F\|\|\mathbf{X}-\mathbf{X}'\|\|$	$\mathcal{P}_\beta$	$\int_{-\infty}^{+\infty}(F(\mathbf{x})-G(\mathbf{x}))$	No	Gneiting and Raftery (2007)
GES	$\frac{1}{2}E_F\|\|\mathbf{X}-\mathbf{X}'\|\|_\alpha^\beta - E_F\|\|\mathbf{X}-\mathbf{x}\|\|_\alpha^\beta$	$\frac{1}{2}E_F\|\|\mathbf{X}-\mathbf{X}'\|\|_\alpha^\beta$	$\mathcal{P}$	$\int_{-\infty}^{+\infty}(F(\mathbf{x})-G(\mathbf{x}))$	No	Gneiting and Raftery (2007)
WPower	$\frac{(p_i/q_i)^{\beta-1}-1}{\beta-1} - \frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta}$	$\frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta}$	L_β	$\frac{(E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}])^{1/\beta}-1}{\beta-1}$	Depends	Jose et al. (2008)
WPseudoSph	$\frac{1}{\beta-1}((\frac{p_i/q_i}{(E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}^{\beta-1}])^{1/\beta}} - 1)$	$\frac{p_i/q_i}{(E_{\mathbf{p}}[\mathbf{p}/\mathbf{q}^{\beta-1}])^{1/\beta}}$	L_β	$\frac{E_{\mathbf{p}}[(\mathbf{p}/\mathbf{q})^{\beta-1}]-1}{\beta(\beta-1)}$	Depends	Jose et al. (2008)
QuantS	$2(I_{[x\le F^{-1}(\alpha)]} - \alpha)(F^{-1}(\alpha) - y)$	$\int S(\alpha;x)dp(x)$	$\mathcal{P}$	$\|\|p-q\|\|_2^2$	No	Cervera and Munoz (1996)
CLS	$I_{(y_t+1\in A_t)}\log(\frac{\hat{f}_t(y_{t+1})}{\int_{A_t}\hat{f}_t(s)ds})$	$\int_A p\log p$	L_2	$\int_t p_t(x)\ln(\frac{q(x)}{p(x)})dx$	No	Diks et al. (2011)
CsLS	$I_{(y_t+1\in A_t)}\log\hat{f}_t(x_{t+1}) + + I_{(y_{t+1}\in A_t^c)}\log(\int_{A_t^c}\hat{f}_t(s)ds)$	-	L_2	-	No	Diks et al. (2011)
TW-CRPS	$\frac{1}{2}w(z)E_F\|\|X-X'\| - E_F\|\|X-x\|\|$	$\frac{1}{2}E_F\|\|X-X'\|\|$	$\mathcal{P}_1$	$\int_{-\infty}^{+\infty}(F(x)-G(x))$	No	Gneiting and Ranjan (2011)
QW-CRPS	$2(I_{[x\le F^{-1}(\alpha)]} - \alpha)(F^{-1}(\alpha) - y)w(\alpha)d\alpha$	$\frac{1}{2}E_F\|\|X-X'\|\|$	$\mathcal{P}_1$	$\int_{-\infty}^{+\infty}(F(x)-G(x))$	No	Gneiting and Ranjan (2011)
Log-coshS	$-\ln\cosh\frac{q'(x)}{q(x)} + \frac{q'(x)}{q(x)}\tanh\frac{q'(x)}{q(x)} + + (\frac{q''(x)}{q(x)} - \frac{q(x)'}{q(x)})(1-\tanh\frac{q'(x)}{q(x)})$				Yes	Ehm et al. (2012)
GM	$\sum_{i=1}^M w_{i,t}p_{i,t}$	$(\sum_{i=1}^M w_i p_{i,t}^{-\rho})^{1/\rho}$	L	$\sum_{i=1}^M w_i\Big\{\sum_{t=1}^T -1y_{(-t)}[(\frac{y_{i,(-t)}}{x_{i,(-t)}\rho-1}]\Big/\rho(\rho-1)\Big\}$	Yes	Gospodinov and Maasoumi (2020)

NOTE: This table summarizes the literature on the geometry of Scoring Rules, to the best of our knowledge. In particular, the first column displays the names of each functional; the second column displays the formula of the S-functional; the third displays the formula of the associated Entropy function; the fourth displays the probability measure necessary to build the same functions; the fifth the associated Divergence function; the sixth column informs the reader if the mentioned SR has Brgman-Savage representation (in peculiar cases, there is not a unique answer, because it depends on the weights' definition); finally, the seventh column indicates the reference corresponding to the S-function. Finally, note that $A_t \subseteq \mathcal{X}$, $t \in \mathcal{T} = \mathcal{X}$ so that $\{A_t\} \equiv \{t\}$; $\mathcal{P}$ indicates a Borel probability measure; L a Lebesgue probability measure; μ a σ-finite measure.

References

Amisano G, Giacomini R. 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. *Journal of Business and Economic Statistics* **25**: 177–190.

Barnard G, Jenkins G, Winsten C. 1962. Likelihood Inference and Time Series. *Journal of Royal Statistical Society, ser. A* **125**: 321–372.

Boero G, Smith J, Wallis K. 2008. Uncertainty and Disagreement in Economic Prediction: The Bank of England Survey of External Forecasters. *The Economic Journal* **118**: 1107–1127.

Boero G, Smith J, Wallis K. 2011. Scoring rules and survey density forecasts. *International Journal of Forecasting* **27**: 379–393.

Breusch T, Pagan A. 1980. The Lagrange Multiplier Test and its Applications to Model Specification in Econometrics. *Review of Economic Studies* **67**: 239–253.

Brier G. 1950. Verification of the forecasts Expressed in Terms of Probability. *Monthly Weather Review* **78**: 1–3.

Cervera J, Munoz J. 1996. Proper Scoring Rules for Fractiles. In Bernardo J, Berger J, Dawid A, Smith A (eds.) *Bayesian Statistics 5*. Oxford, UK: Oxford University Press, 513–519.

Davies R. 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika* **64**: 247–254.

Dawid P, Sebastiani P. 1999. Coherent Dispersion Criteria for Optimal Experimental Design. *The Annals of Statistics* **27**: 65–81.

De Finetti B. 2017. *Theory of probability: A critical introductory treatment*, volume 6 of *Wiley Series on Probability and Statistics*. John Wiley & Sons. Translated by Antonio Machí and Adrian Smith.

Diks C, Panchenko V, van Dijk D. 2011. Likelihood-based scoring rules for comparing density forecasts in tails. *Journal of Econometrics* **163**: 215–230.

Eaton M, Giovagnoli A, Sebastiani P. 1996. A Predictive Approach to the Bayesian Design Problem with Application to Normal Regression Models. *Biometrika* **83**: 111–125.

Ehm W, Gneiting T, et al. 2012. Local proper scoring rules of order two. *The Annals of Statistics* **40**: 609–637.

Eitrheim Ø, Teräsvirta T. 1996. Testing the adequacy of smooth transition autoregressive models. *Journal of econometrics* **74**: 59–75.

Epstein E. 1969. A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology* **8**: 985–987.

Gneiting T, Raftery A. 2007. Strictly Proper Scoring Rules, Prediction and Estimation. *Journal of the American Statistical Association* **102**: 359–378.

Gneiting T, Ranjan R. 2011. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. *Journal of Business & Economic Statistics* **29**: 411–422.

Good I. 1952. Rational Decisions. *Journal of Royal Statistical Society, Ser. B* **14**: 107–114.

Good I. 1971. Comment on "Measuring Information and Uncertainty "by R. J. Buheler. In Godambe V, Sprott A (eds.) *Foundations of Statistical Inference*. Toronto: Holt, Rinehart and Winston, 337–339.

Gospodinov N, Maasoumi E. 2020. Generalized aggregation of misspecified models: With an application to asset pricing. *Journal of Econometrics* **forthcoming**.

Granger C, Maasoumi E, Racine J. 2004. A dependence metric for possibly nonlinear processes. *Journal of Time Series Analysis* **25**: 649–669.

Hyvärinen A. 2005. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research* **6**: 695–709.

Jose V, Nau R, Winkler R. 2008. Scoring Rules, Generalized Entropy, and Utility Maximization. *Operation Research* **56**: 1146–1157.

Lindley D. 1982. Scoring Rules and the Inevitability of Probability. *Revue Internationale de Statistique* **50**: 1–11.

Luukkonen R, Saikkonen P, Teräsvirta T. 1988. Testing linearity against smooth transition autoregressive models. *Biometrika* **75**: 491–499.

Matheson J, Winkler R. 1972. Scoring Rules for Continuous Probability Distributions. *Management Science* **22**: 1087–1096.

Mitchell J, Hall S. 2005. Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR "Fan"Charts of Inflation. *Oxford Bulletin of Economics and Statistics* **67**: 995–1033.

Parry M, Dawid A, Lauritzen S. 2012. Proper Local Scoring Rules. *The Annals of Statistics* **40**: 561–592.

Patton A. 2019. Comparing possibly misspecified forecasts. *Journal of Business & Economic Statistics* : 1–43.

Teräsvirta T. 1994. Specification, estimation and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89**: 208–218.

Teräsvirta T, Tjostheim D, Granger C. 2010. *Modelling Nonlinear Economic Time Series*. Advenced Text in Econometrics. Oxford, UK.: Oxford University Press.

Tsallis C. 1988. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics* **52**: 479–487.

Wallis K. 2004. An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review* **189**: 64–71.

Winkler R. 1972. A Decision-Theoretic Approach to Interval Estimation. *Journal of American Statistical Association* **67**: 187–191.

Zanetti Chini E. 2018. Forecasting dynamically asymmetric fluctuations of the US business cycle. *International Journal of Forecasting* **34**: 711–732.