UNIVERSITÀ DEGLI STUDI DI BERGAMO
Department of Economics

WORKING PAPERS

Strategic judgment: its game-theoretic foundations, its econometric elicitation

Emilio Zanetti Chini

November 2021 - WP N. 5 Year 2021

Working papers – Department of Economics
n. 05

Strategic judgment: its game-theoretic foundations, its econometric elicitation

Emilio Zanetti Chini

Università degli Studi di Bergamo
2021

Strategic judgment: its game-theoretic foundations, its econometric elicitation / Emilio Zanetti Chini – Bergamo: Università degli Studi di Bergamo, 2021.
Working papers of Department of Economics, n. 5
ISSN: 2974-5586
DOI: 10.13122/WPEconomics_5

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/10446/196554

Strategic judgment: its game-theoretic foundations, its econometric elicitation

Emilio Zanetti Chini

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

FIRST VERSION: October, 2019
THIS VERSION: September 2021

Abstract

We provide a new econometric methodology to detect the bias due to strategic interaction induced by subjective learning. This methodology relies on (i) a new definition of coherence based on the Likelihood Principle, specifically designed for economic forecasting; (ii) an environment named "Scoring Structure", where a Forecast User interacts with a Forecast Producer and Reality. A formal test for the null hypothesis of linearity in the Structure is introduced. Linearity implies that forecasts are strategically coherent with evaluations and viceversa. The new test has good small-sample properties and behaves consistently with theoretical requirements. Three case studies on the Federal Reserve Bank's, the Bank of England's and the Norges Bank's forecasts support the endemic nature of the strategic judgment in Macroeconomics. The economic interpretation of the results are discussed.

Keywords: Bias, Coherence, Learning, Likelihood Principle, Mis-specification, Non-linearity, Scoring Structures.
JEL: C12, C22, C44, C53.

1 Introduction

Quantifying the forecasting accuracy is one of the most important issues that any economic expert or policy institution has to deal with. Since the origins of the econometric literature, forecast errors are looked as the empirical counterpart of the unobserved economic costs that a generic representative forecaster pays in case of poor decision/forecast. Thus, in a traditional setting, the optimal forecast is the one that maximizes the sample equivalent of forecaster's own expected utility (or, equivalently, that minimizes her own expected loss)[1]. Any deviation from the optimality (or *bias*) is looked as the effect of an econometric misspecification made by forecaster. For an overview, see De Gooijer (2017); Ghysels and Marcellino (2018); Castle et al. (2019) among others.

In this paper, the forecasting process is looked as an environment where several, differently-behaving agents interact and learn on reciprocal actions. These agents can be generically identified as *forecast users* and *producers* (henceforth, FUs and FPs, respectively). The cheating among economic agents is commonly known with the generic label 'herding' and explained by the failure of rational expectation hypothesis. The rise of the forecasting bias due to herding is not a new idea in the economic literature, see Clements (2018) for a review and a stylized model that englobes some of these features.

No matter what economic motivation is postulated, the forecasting process is sensitive to subjective bias. We are interested in that peculiar subjective bias that either FPs or FUs are induced by their own non-sample information (or *judgment*, henceforth), as well in the link of this judgment with strategic behavior; see Manganelli (2009) and Ottaviani and Sorensen (2006), to which we are strictly related and complementary. According to this framework, a representative FU integrates sample information with her own pre-constituted opinions every time she takes a decision, so that the econometric treatment of these opinions is non-standard. Nevertheless, and despite the generality and flexibility of this solution, this approach mis-recognizes the reaction of a (representative) FP due to a poor evaluation of her forecast by FU. The upper panel of Figure

1 gives a visual intuition of this forecasting scenario. Henceforth, we use the term *quotation* when referring to the (potentially) biased estimate obtained by FP; specularly, we label *announcement* the (eventually) biased estimate that FU adopts after having evaluated the FP quotation. In both the cases the quotation/announcement is assumed being a combination of the output of an optimization problem and a judgmental component. Moreover, we consider the key hypothesis that quotation does not coincide with the announcement. When this hypothesis is verified, we call *bias due to strategic judgment* – or strategic judgmental bias (SJB, henceforth) – the quotation-announcement spread.

The next Section 2 provides an illustrative example on how the FP's reaction to a poor evaluation by FU can be motivated by her own learning activity, which originates a bias that cumulates and perdures in the subsequent periods. Very little is known about the theoretical conditions necessary to set an econometric analysis of such a strategic problem and its mechanics – the only paper on this topic (Manganelli, 2019) considers the FU-side only, while the literature on game-properties of forecasting is mainly theoretical or focused on bayesian calibration or claim validation (Olszewski, 2015). *Hence, how can SJB be empirically modelled and tested?*

Section 3 answers to this research question by introducing a new mathematical object named, generically, *Scoring (or Judgmental) Structure* (SS, henceforth)[2]. This object, illustrated in the lower panel of Figure 1, models a game among several players by using frequentist inference. To this aim, it englobes several elements: a utility function that both FP and FU are assumed to have; a sequence of quotations; a sequence of announcements; an entropy function that measures the amount of the uncertainty – hence, the maximum of the utility – produced by the synergy of Reality, FP and FU[3]; a divergence function, allowing to map the distance between the FP's quotation from the optimal forecast. Similarly to any game, making the SS operational requires some general rules to avoid that the final output is unbalanced in favor of FP or FU. Once these rules has been properly set, the SS makes us able to test for the hypothesis that the marginal utility of the FP coincides with the (log-)likelihood of the model adopted

by FU to reproduce the FP's quotation. When such a coincidence is verified by data, we say that the forecast is *coherent*. This coherence test is strictly related to a peculiar type of utility function, the scoring rule (SR, henceforth). This last assigns a numerical score to several competing (model-based) density forecasts. According to the statistical literature – see De Finetti (2017) and Winkler (1967) – a *proper* SR gives an incentive to FP to be honest[4]. However, this "honesty" does not ensure, *per se*, that FP will not announce a value that differs from that of the estimated model (similarly, that FU will not do an announcement different from the one she is supposed to do given the data and FP evaluation) because she uses non-sample informations in a strategic way. In other words, proper SRs are robust to judgmental bias, *not to SJB*.[5] The game-theoretic nature of the SS overcomes this problem.

Section 4 illustrates the feasibility of the proposed approach via three case studies: the U.S. survey data on GDP and unemployment rate; the professional forecasts of the U.K. inflation; and the Norwegian Output Gap growth rate forecasts. Two of these three cases suggest that the amount of strategic bias is significant. The economic and methodological implications of our findings are discussed. Finally, section 5 concludes. An Appendix provides mathematical details, while a separate Supplement conveys additional explanations and results.

2 The Strategic Bias Problem

In this section we focus on a simple forecasting exercise based on simulated data on the U.S. Industrial Production. 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.

Let consider the case of a set of $\{1, \ldots, \mathrm{i}, \ldots, 1000\}$ random paths of length $T = 265$

from the following DGP that represent a simulated Reality:

$$y_t^{(i)} = 0.9y_{2,t-1}^{(i)} - 0.795y_{2,t-2}^{(i)} + (0.02 - 0.4y_{2,t-1}^{(i)} + 0.25y_{2,t-2})^{(i)} G^{(i)}(\boldsymbol{\Xi}) + \epsilon_{2,t}^{(i)}, \quad \epsilon_{2,t}^{(i)} \sim N(0, 0.5), \tag{1}$$

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 $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T} y_t$.[6] 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 choice of a Smooth-Transition Autoregressive family as DGP is motivated by the fact (that will be proved in the course of the next Section) that this family of models is particularly indicated to nest the forecasting framework and coherence testing that motivates this paper. In particular, we remark that the nonlinearity of the process here created is looked as consequence of the raise of the SJB, while linearity is associated to its non-existence. Thus, the STAR-modelling have to be considered as instrumental to avoid ad-hoc refinements that would make the entire methodology less immediate to understand. Finally, the assumption that the vector parameter $\boldsymbol{\Xi}$ is known is made to allow the reader to focus on SJB as the main source of uncertainty with respect to other sources (like parameter uncertainty or measurement error).

Let now consider a forecasting environment formed by two different agents – namely FP and FU – co-exist and interact in a strategic way. This interaction is summarized in four steps:

1. At time T of the i-esim random draw, FP makes her own quotation $Q_{T+1|T}$ (say, on the conditional density) for the next period given all the information available at the current one via MonteCarlo simulation when the closed form is not available by looking at the vector of observables $\{y_t\}_{t=1}^{t}$ adopting (1) and the utility function the $U(\cdot)$ – say, the quadratic utility, $U(\boldsymbol{\Xi}, \hat{e}) = \hat{e}(\boldsymbol{\Xi})^2$, so that standard OLS inference holds. In this step, FP does not use non-sample information, so that

$Q_{T+1|T}$ is the predicted value of a direct estimation (here, an OLS regression of the dependent variable on the regressors lagged $h-$periods).

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

$$\tilde{y}_t = \hat{y}_t^{(i)} + \tilde{\pi}_1, \tag{2}$$

where $\hat{y}_t^{(i)}$ is the estimated value of the variable, $\tilde{\pi}_1$ represents the non-sample information that FU suspects in FP's quotation and is assumed being small (relatively to $\hat{y}_t^{(i)}$) and set to 0.05 to mean that the role of non-sample information is limited. This requires to compute the probability integral transform (PIT) of $\tilde{y}_t$. Under the assumptions that the sequence of density forecasts $\{f_t(\tilde{y}_t)\}_t^m$ has a nonzero Jacobian with continuous partial derivatives (Diebold et al., 1998), if $\{f_t(\tilde{y}_t)\}_t^m$ coincides with the Data Generating Process density $\{p(y)|\mathcal{I}_t\}_t^m$, then

$$z_{t+h} = \int_{-\infty}^{\tilde{Y}_{t+h}} \hat{F}_{t+h}(u|\mathcal{I}_t)du \;\sim\; i.i.d.\, U(0,1) \tag{3}$$

where $\hat{F}_{t+h}$ is the estimated cumulative density function in t+h (or predictive density) evaluated at $\tilde{Y}_{t+h}$, $\mathcal{I}_t$ the information up to time T and u a suitable measure that allow the effective computation of the integral. The PIT is computed using in the fixed-rolling windows scheme by Rossi and Sekhposyan (2011): in a set of periods $\{1, \ldots, \mathrm{T}\}$, 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 (1) 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 (3). Under perfect forecast, the histogram is perfectly rectangular. But the judgmental component $\tilde{\pi}_1$ makes the effective one-step-ahead histogram corresponding to Step 2 in upper panel of Figure 2 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 quotation, albeit the magnitude of the mis-evaluation is small. However, FP does not know where FU is wrong.

3. FP has learnt to be mis-evaluated by FU and, contemporaneously, has to make the new quotation for the next period $Q_{T+2|T+1}$, 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 $\hat{y}_t^{(i)}$ remains the same. This means that FP adds non-sample information to her original DGP to take in 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 quotation $Q_{T+2|T+1}$. Again, FU knows FP's DGP, not the amount of judgment incorporated by her in $T + 1$ 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

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

where $\tilde{\pi}_2$ represents the non-sample information 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. Notice that 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 (3) – with minor modifications to the bounds of the integral to

take in account for the new period.

Key Result 1. The lower panel of Figure 2 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 (having no estimation bias) neither to FU (because she knows FP's DGP) *if these are considered singularly*. Thus, 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.[8]

Assuming the knowledge of DGP by FU may be looked as counterintuitive for justifying the rise of SJB. One may argue that this assumption corresponds to perfect specification – ontologically antithetic to the need to add further source of information. Thus, what would be the rationale to use judgment in such a setting? Answering to this question implies to formulate an economic theory on the use of information. The inner motivations for which FU might suspect FP has added non-sample informations are compatible with several theories. In this paper we rely on the key idea that forecasting agents are imperfect maximizers facing costs to learn. According to Ilut and Valchev (2020), these learning costs are directly associated to nonlinearity in the policy function (corresponding to asymmetry in utility function) that FP has to fulfill, hence to lack of regularity in updating her beliefs. This theory is compatible with the most recent empirical literature (Manzan, 2011, 2021). Moreover, on a subjectivist perspective, we assume that FP may be influenced by indirect signals (for example, the reputation): in fact, as demonstrated by Ottaviani and Sorensen (2006), reporting the best quotation of the state is not an equilibrium if assuming a forecasting tournament with pre-specified rules. Instead, the reporting activity is a balance between two contrasting forces: (i) the proper (in the sense of De Finetti) incentive to report the honest forecast; (ii) the gain from moving away from the prior mean, deriving by the fact that the number of forecasters that correctly guess the state lower as farther is the FP's state from it. In our setting, the contest is a simple repeated game among FP with FU that may leads FP to differentiate her predictions from the true ones despite the FU activity by putting

greater weight on their private signals (that is, on her SJB) than she would in an honest report; see the Supplement.

The assumption that the DGP of FP is observable by FU is only for ease of illustration and may be relaxed. In this case, the FU has to deal with the estimation error *additionally* to the strategic bias previously considered. It can be shown that, in this case, for the FU is sufficient to know the form of the $U(\cdot)$. In facts, this allows FU to select a proper function (that is, the SR) to compute the PIT. Notice that such a function does take in account of $\tilde{\pi}$ and $\tilde{\pi}_2$ *but only one of them per time*. This means that the FU may use the best SR to perform her evaluation and still obtain a PIT similar to the one in lower panel of Figure 2. In this case, the FU may abandon the PIT and use the Autocontours, see González-Rivera et al. (2011); or may use a formal test for correct specification like the ones presented in Corradi and Swanson (2006) and their more recent development for dynamic correct specification by Rossi and Sekhposyan (2019), but the qualitative result of the above example would remain unchanged because this literature does not consider any strategic component. The next Section fills this gap.

3 Theoretical framework

This Section introduces the econometric theory of SJB that generalizes the example above illustrated. Namely, Subsection 3.1 introduces the notation; Subsection 3.2 defines and characterizes the notion of coherence; finally, Subsection 3.3 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 $Y_t = \{y_1, \ldots, y_T\}$ is the vector of observed data, $\hat{Y}_t = \{\hat{y}_1, \ldots, \hat{y}_T\}'$ is the vector of optimal estimates, $\tilde{Y}_t = \{\tilde{y}_1, \ldots, \tilde{y}_T\}'$, is the vector of (potentially biased) estimates corresponding to the $\tilde{y}_t$ in (2) and $X_t = \{x_1, \ldots, x_T\}'$ is the vector of exogenous variables.[9] Moreover, it is defined on a com-

plete probability space $\{\Omega, \mathcal{F}, \mathcal{P}\}$, where Ω is the sample space; the event space $\mathcal{F}_t$ is partitioned as $\mathcal{F}_t \equiv [\Pi, \Psi] \in \mathbb{R}^k$ to denote the sub-spaces of FU and FP, respectively; $\mathcal{P} \doteq \{\mathbf{p} \in \mathcal{A} : \sum_x p_x = 1\}$ defines the set of all real vectors corresponding to strictly positive probability measures if the sample space is discrete, as assumed in this paper;[10] $\mathcal{A}$ an algebraic subset of $\tilde{\mathcal{Y}}$ representing the set of FP's actions, which, in turn, correspond to the set of all possible $\tilde{\pi}_t$ that she decides.

The set of all possible values taken by Z_t is $\mathcal{Z}_t \equiv [\mathcal{Y}_t, \hat{\mathcal{Y}}_t, \tilde{\mathcal{Y}}_t, \mathcal{X}_t]$. The Log-likelihood of FU and FP are denoted, respectively, $\mathcal{L}(\Pi)$ and $\mathcal{L}(\Psi)$, while the density of Z (or each of its partitions) is denoted $P(Z) \doteq \int p_Y(z)dz$, where $p(\cdot)$ is a continuous density function defined on $\mathcal{L}(\Omega)$ and t is omitted to ease the notation. The (one-step-ahead) distributional and density forecasts of Z_t are denoted as $P(Z_{t+1})$ and $p(Z_{t+1})$, respectively.

The FP seeks to solve a decision problem defined by the triple $\left\{\hat{Y}_t, \mathcal{A}, \mathcal{U}(\hat{p}, \mathrm{a}^*)\right\}$, where: $\hat{Y}_t$ and $\mathcal{A}$ are defined as before; and $\mathcal{U}(\hat{p}, \mathrm{a}^*)$ is a real-valued utility function that represents the reward obtained by the FP as the result of minimizing the discrepancy $|\hat{y}_{t+1} - y_{t+1}|$ – that is, the distance between density forecast for $t+1$ and data that will be observed at that time – by her own optimal judgement $a^* \in \mathcal{A}$. This last is the one that maximizes the expected utility, denoted as $EU := \int U(\hat{P}, \mathrm{a}^*)p(\hat{Y}_t)d\hat{Y}_t$, computed using the distribution P, which is believed to be the true DGP. The same holds for FU, with minor modification to the notation – that is, ψ instead of π, $\tilde{Y}$ instead of $\hat{Y}$ and $\tilde{a}^* \in \tilde{\mathcal{A}}$ instead of $a^* \in \mathcal{A}$).

3.2 Representation of the Forecasting Environment

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

Definition 1 (Structural Coherence)**.** The (h-step ahead) density forecast y_{t+h} obtained by $p(\hat{\Pi}; x_t) \in \mathcal{P}$ is *coherent relatively to the scoring structure* (or *structurally coherent*) if there is one-to-one mapping between $\mathcal{L}(\hat{\Pi}; x_t)$ and $\mathcal{L}(\hat{\Psi}; x_t)$.

This definition is an economic, empirical counterpart of the definition of coherence by

De Finetti, Ch. 3.1, pp. 62–64, that, to our best knowledge, is the only other definition available for this issue. According to this last Author, a probabilistic forecast is coherent if and only if it is finitely additive. However, despite its generality has enabled a series of mathematical as well as epistemological consequences still today 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. Our definition is (micro-)economically founded, because is based on (i) a walrasian axiomatization of the forecasting environment (that is, the existence of a market where the demand and supply of economic forecasts matches); (ii) a game-theoretic approach to forecasting (see the Supplement for details). The term "structural" emphasizes the role of the two demand/supply sides constituting the structure of any exchange-based economy. Secondly, our definition is empirical because it is based on an equivalence among two estimated objects and links the coherence to the likelihood principle. 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 that agents move away from optimal actions.

Proposition 1. *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 orders 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 2. *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:*

$$\mathbb{L}\mathfrak{s} = 0, \tag{6}$$

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 (6) is called the *Key Condition*. Two further (merely theoretical) conditions concerning the representation of $\mathfrak{s}$ via Lagrange operators are required to proof 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. □

3.3 Testing for structural coherence

To test the hypothesis that the equation (6) 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 same structure.

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:

$$\begin{aligned} y_t &= \boldsymbol{\phi}'\mathbf{w}_t + G(\gamma, \mathbf{z}_t, \mathbf{c}_k)\boldsymbol{\theta}'\mathbf{w}_t + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma^2) \\ G(\gamma, \mathbf{z}_t, \mathbf{c}_k) &= \left(1 + \exp\left\{-\gamma \prod_{k=1}^{K} (\mathbf{z}_t - \mathbf{c}_k)\right\}\right)^{-1}, \quad \gamma > 0, \quad c_1 < \cdots < c_k, < \cdots < c_K, \end{aligned} \tag{7}$$

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 as in (14) 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 (7) becomes a two-regime threshold autoregression SS (SS-TARX).

The (nonlinear) SS so defined is an algorithm that applies the Forecasting Protocol defined in Section 2 of 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 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 (7). In Supplement, Reality is assumed acting as third player, so the number of steps enlarges without loose 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 $S(y_{t+h}, x_t)$ is needed. As will appear shortly, equation (7) is necessary only as a convenient way to test the null hypothesis of structural coherence, corresponding to (6). Moreover, the Step (ii) can be seen also 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 $\tilde{y}_t$.[11]

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

Proposition 3. *Let y_t be a stochastic process generated by (7). Then:*

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

$$H_0 : \gamma = 0 \;\; vs \;\; H_1 : \gamma \neq 0 \quad in \; (7), \tag{8}$$

which can be measured by the following LM statistics:

$$S(\boldsymbol{\Xi})^{LM} = \hat{\sigma}^{-2}\hat{\mathbf{U}}'\hat{\mathbf{D}}_{\mathbf{2}}(\hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{D}}_{\mathbf{2}} - \hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{D}}_{\mathbf{1}}(\hat{\mathbf{D}}_{\mathbf{1}}'\hat{\mathbf{D}}_{\mathbf{1}})^{-1}\hat{\mathbf{D}}_{\mathbf{1}}'\hat{\mathbf{D}}_{\mathbf{2}})^{-1}\hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{U}} \sim \chi^2_n \tag{9}$$

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 (8) can be measured by one of the following LM statistics:

$$\begin{aligned} LM_1 &= (SSR_0 - SSR)/\hat{\sigma_v}^2 \sim \chi^2_{3p} \;\; if \;\; K = 1 \;\; in \;\; (7) \\ LM_2 &= (SSR_0 - SSR)/\hat{\sigma}^2_{v1} \sim \chi^2_{2p} \;\; if \;\; K = 2 \;\; and \;\; c_1 = c_2 \;\; in \;\; (7) \\ LM_3 &= (SSR_0 - SSR)/\hat{\sigma}^2_{v2} \sim \chi^2_{p} \;\; if \;\; K = 2 \;\; and \;\; c_1 \neq c_2 \;\; in \;\; (7), \end{aligned} \tag{10}$$

where SSR_0 *and* SSR *are the sum of the squared residuals of SS-STARX (7) 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 (10) 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 Supplement. □

4 Empirical Applications

This Section illustrates the use of the SS and the coherence test by three case studies. Namely, Subsection 4.1 analyzes the GDP and unemployment forecasts by Survey of Professional Forecasters of Federal Reserve (SPF-FED, henceforth); Subsection 4.2 deals with the UK inflation using Bank of England' survey data; Subsection 4.3 con-

siders the Norwegian output gap (OG, henceforth) forecast by Norges Bank; finally Subsection 4.4 discusses the relevance of these applications.

4.1 The U.S. survey data

The Federal Reserve Bank has been the first institution to use professional forecasting to justify its policy decisions. The FED-SPF began in 1968 as independent study of the American Statistical Association and NBER and in late 1990 its maintainance became part of the institutional activity of the FED. 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 FED of Philadelphia at https://www.philadelphiafed.org/surveys-and-data/data-files. The release is dated November 16th, 2020. In this paper we consider two variables: the Real GDP (RGDP) and the unemployment rate (UNR), both of them for an horizon going from one to four quarters; the nowcasts are also investigated. The sample span is 1975:Q1–2020:Q4.[12]

According to Tables 1 and 2, there is a strong evidence of SJB in both 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 on eight in 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 the cases. 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 the cases.

4.2 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 reach of the Bank's institutional targets. Thus, its projections can be seen as announcements. More recently, the BoE has also published the data on the Survey of External Forecasters (SEF), which constitute the FP in our framework and are the equivalent of the U.S. SPF. It is very well-known that the BoE use 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 since 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 mean aggregate to allow their use in our SS-framework without complicating the statistical model therein nested.

The results are reported in Table 3. 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 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 (7), $\hat{\gamma} = 5.67$ 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. To this aim, 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. This last is assumed being the output of a GLSTAR(2) because the

estimated density function from that model are asymmetric, see Zanetti Chini (2018)[13]. Then the average scores

$$\bar{S}_n^f = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{f}_{t+k}, y_{t+k}), \quad \bar{S}_n^g = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{g}_{t+k}, y_{t+k}) \tag{11}$$

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

$$H_0 : \Delta^* = \bar{S}_n^f - \bar{S}_n^g = 0 \quad vs \quad H_1 : \Delta^* = \bar{S}_n^f - \bar{S}_n^g \neq 0, \tag{12}$$

which is measured by statistic

$$t_n = \sqrt{n}\frac{\Delta^*}{\hat{\sigma}_n} \sim N(0,1), \tag{13}$$

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 4, 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 case on eight, that is the weighted pseudo-spherical score (WPseudoSph) with $\alpha = 1$, corresponding to a weighted logarithmic score.

4.3 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 here used corresponds to the release 2014 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 (7). 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)[14]. The last ones 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 to the previous application on UK CPI inflation, Table 5 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[15] to be compared with the final announcement, represented by downloadable BoN fan charts. Therefore, we repeat the analysis of (11) in the previous Subsection 4.2 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 6 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.

4.4 Discussion

The lack of structural coherence defined in the Section 3 is an endemic characteristic of macroeconomic data – at least the ones published by the main Central Banks. Several consequences descends by our illustrations: first, the SJB hypothesis spurs the comparability of the model-based forecasts. In fact, whereas (as in the case of the U.K. Inflation) the SJB cannot be rejected in a third of the lags, despite at a confidence level of 10%, the equal predictive ability tests is not passed in one third on the SRs adopted; on the other side, whereas (as in the case of the Norwegian OG) the SJB hypothesis is rejected in almost all the lags, the equal predictive hypothesis is globally acceptable. This relates directly to the recent finding by Galvao et al. (2021) that judgment tends to downgrade the accuracy of pure statistically-driven ones. Moreover, the variety of SRs failing to reveal a winner is also noticeable. This further complication is an effect of the Patton (2019) results that electing *ex-ante* the form of the SR does not ensure coherence. Second, the Score Invariance principle – exposed in Supplement – makes the elicitation of the true FP's utility objectively difficult and, consequently, explains the difficulty in detecting a winner when making forecast comparisons. Thus, we should question the economic nature of such Score Invariance.

Answering to this question implies having a theory that enlighten the dynamics of the players of the Forecasting Game and, in ultimate analysis, the foundation of the SS. We find this in the recent model of costly cognitive decision making by Ilut and Valchev (2020). According to these authors, FP allocates her effort to make a costly quotation by perfectly observing all relevant objective state variables. However, the true policy function corresponding to the scoring function used by FU in her evaluation is unknown. The uncertainty on the 'true' SR is estimated by bayesian nonparametric methods focused on Gaussian Process distribution over which FP update her beliefs. The FP gradually accumulates information about the optimal quotation as function of the underlining state. Such an accumulation (i) reduces the variance of the judgment but, contemporaneously, (ii) increases the system's Entropy necessary to take the

decision; thus (iii) it explains only partially the optimal quotation at a different state realization; as a consequence, (iv) it leads to a propagation of the judgment in all the forecasting process.

The combination of (i)-(iv) helps to explain the endemic incoherence and the high variety of SRs estimates (that is the two most important findings of our empirical investigation) but not the issue of the Score Invariance. This last may be seen as a consequence of a stationary covariance function which controls the correlation among beliefs about the values of the (unknown) SR. When this correlation is imperfect, the information acquired by FP about the 'true' SR is more useful in the neighborhood of the state realization where learning occurs. Hence, in states where learning is more intense, the uncertainty over the SR is lower than in states where such learning is more rare. This non-constant reasoning by both FP and FU causes the oscillation of the SJB, hence demonstrated not identifiable by the SR – even if proper.

The relevance of these consequences requires some caveat: first of all, our SS approach is characterized by a simplistic parametrization. The forecasting environment is assumed to be populated by only two homogeneous, representative agents. This assumption have allowed us to ease the illustration and the mathematical representation of the SS. We are aware that real Economics is less simplistic, being at least two possibile source of additional complexity in both the demand and supply side: for example, in the U.S. case, (a) the FED-SPF is not used only by FED, but by any firm or investor interested in macroeconomic forecasting. Specularly, (b) the FED complements SPF with its own internal forecasting model(s).

Concerning (a), augmenting the complexity of our methodology implies to abandon the homogeneity assumption and to allow the amount of private information among the (possibly, large) number of players being asymmetric. Unless assuming that the FED's internal forecasting model and SPF (specularly, that SPF and the collection of institutions using it) have the same information processing dynamics – in which case the additional FP' outputs may easy be represented in exogenous vector $\mathbf{z_t}$ – our actual parametrization is insufficient to this aim. A first step to deal with this issue may be

the use of networks; see, among others, Bramoullé et al. (2014). Concerning (b), a second stochastic term (say, ζ_t) with opportune distributional assumptions should be added to equation (2), so that the PIT in first panel of Figure 1 would be radically different – perhaps similar to the second panel because the model would be very easy misspecified. In turn, this additional terms will contaminate the further steps. As a consequence, the game in Section 2 of Supplement should be properly modified to allow it. In this sense, a more complex forecasting game is available in Vovk and Shafer (2005, p. 754): this last 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 assumed working in parallel with FP. The outputs of FP and RG are then properly averaged by FU. Again, this requires an additional step and a richer parametrization in (7).

A second caveat concerns the Step 2 of the Forecasting Protocol (see Supplement), that assumes the SR is known to exploit the properties of the STARX models and thus avoid the use of unobserved component modelling. In fact, the representation of the SS in an unobserved state-space models has never been studied. Finally, we avoided deliberately to consider a dynamics in the repeated game for economy of space. This implies that that we are unable to validate the recent findings on the quality of SPF under/overreactions; see, among others, Bordalo et al. (2020). All these assumptions may limit the application of the new methodology.

5 Conclusions

Recent advances in economic theory suggest that the rise of the 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 econometric approach that enables both professional forecasters and their evaluators to take in account of their

reciprocal learning, either when forecasts are produced, either when these are evaluated.

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 a Forecast User' subsequent announcement and vice-versa. From a general perspective, the Scoring Structure establishes a link between likelihood principle and coherence among Forecast Producer's quotations and Forecast User's announcements. More specifically, it establishes a direct link among forecasting model's 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 by the fact that 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 here introduced is prototypical. Thus, further research is necessary to make it feasible in more realistic scenarios and in large forecast comparisons exercises.

Aknowledgments

The Author thanks Barbara Annicchiarico, Christian Brownlees, Federico Crudu, Anders B. Kock, Alessandra Luati, Jan G. de Gooijer, Michele Lenza, Antonio Lijoi, Simone Manganelli, Elisa Nicolato, Tommaso Proietti, Francesco Ravazzolo, Enrique Sentana, Timo Teräsvirta, Francesco Violante and Takashi Yamagata for their feedback and discussions in several (un)official meetings. Moreover, and for the same reason, he thanks all the participants to the 10th Nordic Econometric Meeting 2019 in Stockholm, the Rome Workshop in Time Series and Financial Econometrics at LUISS University (2019), virtual IFS2020 in Rio De Janeiro, IWEEE2020 in Venice, EC^2 in Paris, the Virtual ESEM2021 at Copenhagen. Previous versions of this paper were presented at: the NBP Workshop on Forecasting in Warsaw (2017), 2nd Annual IAAE Conference

in Milan (2016) and the 34th ISF2014 in Rotterdam, where a Travel Award Grant has been obtained by the International Institute of Forecasters, which is gratefully acknowledged. 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 web-page. 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 and support in the earlier phase of the Author's university career none of the ideas expressed in this paper would have been possible.

Notes

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

2 The term 'Scoring' is due to the function commonly known as *scoring rule*, which is formally defined introduced in Appendix A.1. The term 'Structure' has been chosen in spite of the most immediate 'Game' in order to remark that evaluating the FP/PU requires not only to define some mathematical conditions of an equilibrium, but also to pre-define other objects for measuring the optimality of such an equilibrium. These objects are related to the same FP/FU.

3 The Entropy is a statistical functional very well-known in Physics and only recently appeared in Econometrics. It acts as link among the distribution function and its moments. The amount of Entropy functionals currently known is huge. In this paper we adopt one of the most general forms. The mathematical conditions required to its estimation vary according to the specific form adopted. One of the few assumptions commonly considered necessary to define an estimator of the Entropy is the strict stationarity; see Giannerini et al. (2015) as an example.

4 "*The scoring rule is constructed according to the basic idea that the resulting device should oblige each participant to express her true feelings, because any departure from his own personal probability results in a diminution of his own average score as he sees it*" De Finetti (1962, p. 359).

5 Moreover, as documented in Table 1 of the Supplement, the number of proper SRs explicitly built

for density functions is high. Just as important, many of them are nested, so identifying the true incentive that drives the FP's forecast is not trivial.

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

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

8 This example – as well as this paper – does not consider any dynamics of the learning process that FP benefits in the transition from Step 2 to Step 3 and aims only to illustrate the SJB as a (potentially) self-exiting phenomenon that may cause a systematic failure of traditional diagnostics. This simplistic assumption is made for ease of statistical treatment. Dynamical issues are left to further research.

9 This notation aims to emphasize the link with previous Section 2. In some parts of this section, the difference of Y, $\hat{Y}$ and $\tilde{Y}$ is not essential. For this reason they will be both denoted as Y in the course of this section. The Supplement provides an exact geometry of the equilibrium among these variables.

10 If Ω is continuous, the probability space is defined on $\mathcal{M}$, the set of all distributions on $\mathcal{X}$ that are absolutely continuous respect to a $\sigma-$finite measure μ.

11 Estimation and inference are analogue to the STARX model; see the Supplement.

12 The first 24 observation have been discarded to avoid non available answer occurring occasionally in that period.

13 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 Section 3. The same reference illustrates the Monte Carlo procedure adopted for this family of models.

14 See the BoN web page at `http://www.norges-bank.no/en/Monetary-policy/` for references.

15 According to our estimates,$\hat{\gamma} = 0.89$, and the standard deviation is 0.61.

References

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

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

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.

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.

Castle J, Clements M, Hendry D. 2019. *Forecasting: An Essential Introduction*. London, UK.: Yale University Press.

Clements M. 2018. Do macroforecasters heard? *Journal of Money, Credit and Banking* **510**: 265–292.

Corradi V, Swanson N. 2006. Predictive density evaluation. In Elliott G, Granger C, Timmermann A (eds.) *Handbook of Economic Forecasting*. North Holland.

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.

De Gooijer J. 2017. *Elements of Nonlinear Time Series Analysis and Forecasting*. Series in Statistics. Berlin: Springer-Verlag.

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.

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

Ghysels E, Marcellino M. 2018. *Applied Economic Forecasting using Time Series Methods*. New York, USA: Oxford University Press.

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.

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

González-Rivera G, Senyuz Z, Yoldas E. 2011. Autocontours: dynamic specification testing. *Journal of Business & Economic Statistics* **29**: 186–200.

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

Ilut C, Valchev R. 2020. Economic Agents as imperfect problem solvers. Unpublished Manuscript, Boston College.

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

Manganelli S. 2009. Forecasting with judgment. *Journal of Business & Economic Statistics* **27**: 553–563.

Manganelli S. 2019. Deciding with Judgment. Unpublished Manuscript.

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.

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.

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 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.

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

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. 2011. Understanding models' forecasting performance. *Journal of Econometrics* **164**.

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.

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

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

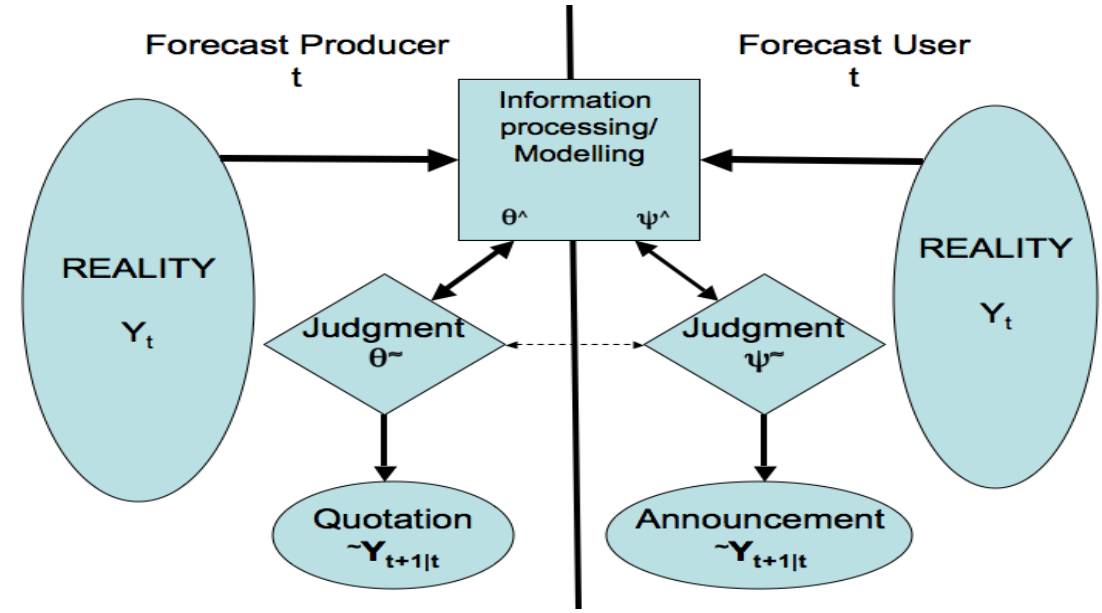
Wallis K. 2004. An assessment of Bank of England and National Institute inflation forecast uncertainties. *National Institute Economic Review* **189**: 64–71.

Winkler R. 1967. The quantification of judgment: Some methodological suggestions. *Journal of the American Statistical Association* **62**: 1105–1120.

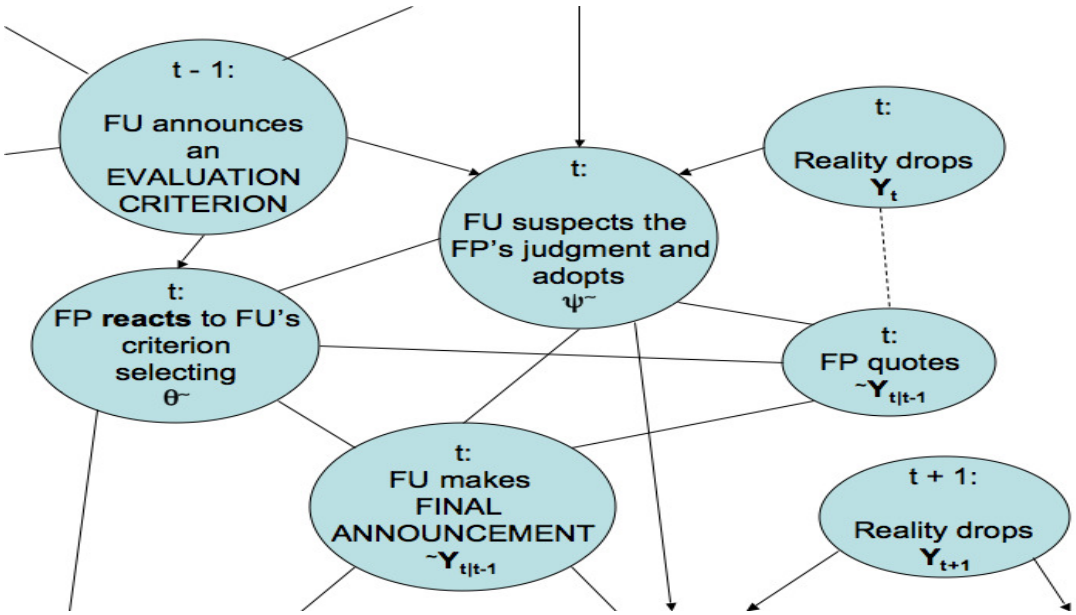
Zanetti Chini E. 2018. Forecasting dynamically asymmetric fluctuations of the US business cycle. *International Journal of Forecasting* **34**: 711–732.

Figure 1: A synopsis of the classical and proposed forecasting environments

(a) The classical environment

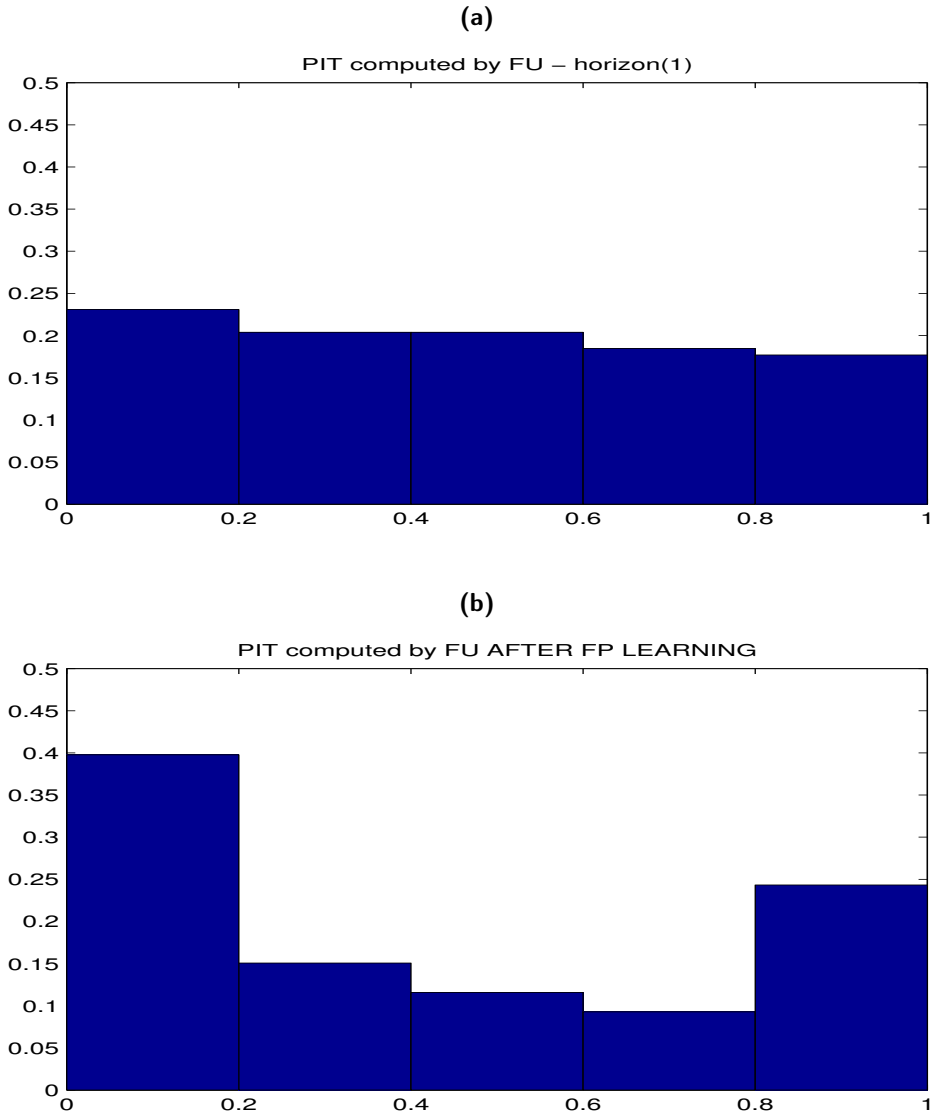


(b) The proposed environment



NOTE: The panel (a) 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 the Reality, (ii) processes it via econometric modelling; (iii) eventually, adds a judgmental component on them (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 a purely deterministic and thus, to influence only on the judgmental part. Panel (b) displays the proposed forecasting environment, where a the same players act in a more complex scenario, where each FP/FU's action may influence on several, contemporaneous and future decisions of the other Player in any period; notice that, in this scenario, Reality acts periodically and unidirectionally (that is, it is not influenced by any other Player).

Figure 2: The problem of Misspecification due to Strategic Judgment



NOTE: This figure shows the problem of misspecification by FU in the illustrative example in Section 2 via the rolling-windows scheme by Rossi and Sekhposyan (2011). Namely, in a sequence of periods {1, ..., T}, FP observes a set of draws from (1) (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 (2) is collected and, successively, transformed via (3) in Step 2. The panel (b) corresponds to the same procedure replicated in the subsequent period, where the 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 set arbitrarily $\tilde{\pi}_1=\tilde{\pi}_2=\tilde{\psi}=0.05$.

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 (10) 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. 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 (10) 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.

Table 3: 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 (10) 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 4: 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 5: 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 $F_{34,29}$ describer in (10) 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 6: 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.

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 2 (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:

$$S(z, p) = \mathfrak{s}(z, p(z), p'(z), p''(z), \ldots, p^{(m)}(z)), \tag{14}$$

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 $\mathcal{SS} := \left\{\mathcal{Z}_t, \mathcal{F}_t, \mathcal{P}, S(\cdot, \cdot), H(\cdot), D(\cdot, \cdot)\right\}$.

The Definition 1 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* FPs 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 order of the variables in the functions – that is $S(Y,X)$, $H(Y)$ and $D(Y,X)$.

This definition of SS is highly general and is used only to consider both FPs 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 functions 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)[16] 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 $P \in \mathcal{P} = \mathcal{A}$ ($\mathcal{M} = \mathcal{A}$ 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*:

$$S(y,p)^{Bregman} \doteq f\big[p(x)\big] + \int y\Big\{f\big[p(y)\big] - p(y)f'\big[p(y)\big]\Big\}d\mu \tag{15}$$

with associated *Brègman divergence*:

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

where $f(\cdot)$ is a (strictly) concave function and $f'(\cdot)$ a subgradient of $f(\cdot)$.[17] 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 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

$$f(z) = k(z) - \lambda \log(z), \tag{17}$$

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 (17) the forecasts generated by $\mathcal{M}$ are coherent with a given SS. 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 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 conditions (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 aim, it is sufficient to notice that Key Equation (6) 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 one; 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
"*Strategic judgment: its game-theoretic foundations, its econometric elicitation*"
(FOR ONLINE PUBBLICATION ONLY)

Emilio Zanetti Chini

September, 2021

1 Introduction

This Supplement gives additional analysis and results which cannot be put in the Main Document for space reasons. Namely, the next Section 2 describes the game assumed in Section 3 of Main Document; Section 3 provides details of the Monte Carlo Simulation of the SS-STARX model introduced in Section 4 of Main Document; 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; finally Section 5 reports the proof of Proposition 3 of Section 4 of Main Document.

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; however, no matter how the FU plays, Reality acts as though the FU does not win the game. This rule, called "*Cournot's Principle*", is necessary to avoid that the game is unbalanced in favor of FU. These 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 his/her (potentially biased) quotation $\hat{p}(X) \in \mathbb{R}$;
4. Reality announces a draw from $P(Y) \in \mathbb{R}$;
5. $K_1 = K_0 + D\ \big(\mathrm{Y},\ \mathrm{X}\big)$,

FU must choose S so his 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 in Section 4 of Main Document, 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 about the players hold for each recursion.

The 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[1].

The Step 5 of the Protocol is 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. This coherence test is essentially based on the $D-$function, which will be better defined soon. 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 who 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 Simulation Study

3.1 Simulation Design and Results

We consider two different DGPs:

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

and

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

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 gave high sensitivity to the size of the slope parameters. Such 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 $\alpha = \{0.01, 0.05, 0.10\}$.

Table 2 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 Main

Document. The performances of the F_3 statistic are poor, so it is 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 still 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 3 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 1 of the Supplement, 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 FPs quotation. This second feature is called '*Score Invariance*' and, as theoretically demonstrated by Paragraph 11.2 in Parry et al. (2012), it holds also for $D(\cdot, \cdot)$ and $H(\cdot, \cdot)$.

3.2 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 this last, 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[1]. 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 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 motivation for adopting the locality (which means coherence) as 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

[1] 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 $P(x) = p$.*"

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.

4 Families of Scoring Structures

In Section 3 of the Main Document we have introduced the general theory of Scoring Structure (SS). In our subsequent applications of Sections 5, we have considered a large number of SRs, and each one can be adopted in the same SS. All these SRs are reported in Table 1 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 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:

$$\mathbf{\Lambda}_t(\mathbf{z_t}, \mathbf{\Xi}) = const + \frac{T}{2} \ln \sigma^2 - \frac{1}{2}\sigma^2 \sum_t u_t^2(\mathbf{\Xi}), \tag{3}$$

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

$$\begin{aligned} \mathbf{\Sigma}_t(\mathbf{w_t}, \mathbf{\Xi}) &= \nabla_{\mathbf{\Xi}} \mathbf{\Lambda}_t(\mathbf{w_t}, \mathbf{\Xi}) = \frac{1}{\sigma^2} \sum_t u_t(\mathbf{\Xi}) \mathbf{d}_t, \\ \mathbf{d}_t &= \nabla_{\mathbf{\Xi}} u_t(\mathbf{\Xi}) = [\mathbf{w}_t, \mathbf{w}_t G, \boldsymbol{\theta}'\mathbf{w}_t G_\gamma, \boldsymbol{\theta}'\mathbf{w}_t G_c]^\mathsf{T}, \end{aligned} \tag{4}$$

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 $\boldsymbol{\tau_1} = (\phi_0, \boldsymbol{\phi}^\mathsf{T})^\mathsf{T}$, $\tau_2 = \gamma^2$, $\hat{\boldsymbol{\tau}}_\mathbf{1}$ 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 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 $\hat{\mathbf{w}}_\mathbf{2t} \equiv \frac{\partial^2 u_t}{\partial\gamma\partial\gamma'}\big|_{\gamma=\mathbf{0}} = -\frac{1}{2}\big\{\theta_{20}[y_t(y_{t-d})] - cy_t\boldsymbol{\theta}'\mathbf{w_t} + \boldsymbol{\theta}_\mathbf{2}'\mathbf{w_t}y_t y_{t-d}\big\}$. 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

$$y_t = \phi'\mathbf{w_t} + \theta'\mathbf{w_t}T_3G\big(\,\cdot\,\big)\epsilon_t', \tag{5}$$

leads to the following auxiliary regression for testing linearity:

$$\hat{\epsilon}_t' = \hat{\mathbf{w}}_\mathbf{1t}'\tilde{\boldsymbol{\beta}}_\mathbf{1} + \sum_{j=1}^{p}\beta_{2j}sy_{t-j}y_{t-d} + \sum_{j=1}^{p}\beta_{3j}sy_{t-j}y_{t-d}^2 + \sum_{j=1}^{p}\beta_{4j}sy_{t-j}y_{t-d}^3 + v_t, \;\; v_t \sim NIID(0, \sigma^2), \tag{6}$$

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. Hene null hypothesis of

locality become testable by following hypothesis system:

$$H_0 : \beta_{2j} = \beta_{3j} = \beta_{4j} = 0 \;\; j = 1, \ldots, p, \quad \text{vs} \quad H_1 : \beta_{2j} = \beta_{3j} = \beta_{4j} \neq 0 \tag{7}$$

corresponding to statistic LM_1 in equation (17) of Main Text, with SSR_0 and SSR denoting the sum of squared estimated residuals from the estimated auxiliary regression (6) 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

$$\hat{\epsilon}'_t = \tilde{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathsf{T}} \hat{\mathbf{w}}_{\mathbf{1}} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathsf{T}} \mathbf{w}_{\mathbf{t}} s y_{t-d} + \boldsymbol{\beta}_{\mathbf{3}}^{\mathsf{T}} \mathbf{w}_{\mathbf{t}} s y_{t-d}^2 + v'_t, \;\; v'_t \sim NIID(0, \sigma^2), \tag{8}$$

where $\tilde{\boldsymbol{\beta}}_{\mathbf{1}} = (\beta_{10}, \boldsymbol{\beta}'_{\mathbf{1}})'$, with $\beta_{10} = \phi_0 - c^2\theta_0$ and $\boldsymbol{\beta}_{\mathbf{1}} = \boldsymbol{\phi} - c^2\boldsymbol{\theta} + 2c\theta_0 \mathbf{e}_d$; 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

$$H'_0 : \boldsymbol{\beta}_{\mathbf{2}} = \boldsymbol{\beta}_{\mathbf{3}} = 0 \quad \text{vs} \quad H'_1 : \boldsymbol{\beta}_{\mathbf{2}} = \boldsymbol{\beta}_{\mathbf{3}} \neq 0 \tag{9}$$

which can be tested by the test statistic LM_2 in equation (17) of Main Text, where SSR_0 and SSR are the sum of squared residuals from (8) under the null and the alternative respectively, $\hat{\sigma}^2_{v1} = (1/T)SSR$. A peculiar case of (9) is when $\boldsymbol{\beta}_2 = 0$ as $\theta_0 = c = 0$, in which case, the null becomes

$$H''_0 : \boldsymbol{\beta}_{\mathbf{3}} = 0 \quad \text{vs} \quad H''_1 : \boldsymbol{\beta}_{\mathbf{3}} \neq 0 \tag{10}$$

which test statistic corresponds to statistic LM_3 in equation (17) of Main Text, with SSR_0, SSR and σ_{v2} defined in a similar way with respect 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 (7) or (9) or (10) 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 (6):

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

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 do not discuss in details. This ends the proof.

6 Estimation

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

$$SSR = \sum_{t=1}^{T} \left(y_t - \hat{\boldsymbol{\psi}}' \xi_{\mathbf{t}}' \right)^2, \tag{12}$$

where:

$$\hat{\boldsymbol{\psi}} = [\hat{\boldsymbol{\phi}}, \hat{\boldsymbol{\theta}}] = \left(\sum_{t=1}^{T} \mathbf{x}_{\mathbf{t}}'(\gamma, c) \xi_{\mathbf{t}}(\gamma, c) \right)^{-1} \left(\sum_{t=1}^{T} \xi_{\mathbf{t}}'(\gamma, c) y_t \right), \tag{13}$$

and

$$\xi_{\mathbf{t}}(\hat{\gamma}, \hat{c}) = \left[\mathbf{w}_{\mathbf{t}}, \mathbf{w}_{\mathbf{t}}' G(\cdot) \right]. \tag{14}$$

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 easy computed. In a such a way, the nonlinear least square minimization problem, otherwise necessary, computationally more demanding and not available in closed-form, reduces to a minimization on two parameters, and is solved via a grid search over γ, c.[3]

7 Tables and Graphs

Table 1: 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	$K(Q_V) + tr\{\boldsymbol{V}_P - \boldsymbol{V}_Q\boldsymbol{\Gamma}_Q\} - (\mathbf{x}-\boldsymbol{\mu}_P)'\boldsymbol{\Gamma}_P^{-1}(\mathbf{x}-\boldsymbol{\mu}_P)$	$-\log\det\boldsymbol{\Gamma}_P - mK$	$\mathcal{P}$	$tr(\boldsymbol{\Gamma}_P^{-1}\boldsymbol{\Gamma}_Q) - \log\det(\boldsymbol{\Gamma}_P - 1\boldsymbol{\Gamma}_Q) + + (\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q)'\boldsymbol{\Gamma}_P^{-1}(\boldsymbol{\mu}_P-\boldsymbol{\mu}_Q) - K$	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	$\sum_{i=1}^M w_i\Big\{\sum_{t=1}^T -1y_{(-t)}\big[(\frac{y_{i,(-t)}}{x_{i,(-t)}\rho-1}\big]\Big/\rho(\rho-1)\Big\}$	Yes	Gospodinov and Maasoumi (2020)

NOTE: This table summarizes the best of our knowledge of the literature on the geometry of Scoring Rules. 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 inform 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.

Table 2: 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 3, where in equations (1) and (2) 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 Main equation. In this experiment, the first 100 simulations were discarded to avoid the initialization effect. Software used: MATLAB R2009b.

Table 3: 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 3 , where in equations (1) and (2) 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 Main Document. In this experiment, the first 100 simulations were discarded to avoid the initialization effect. Software used: MATLAB R2009b.

References

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

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.

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.

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

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.

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

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.

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

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