ISSN: 2281-1346

Department of Economics and Management

DEM Working Paper Series

Forecasters' utility and forecast coherence

Emilio Zanetti Chini
(Università di Pavia)

 # 145 (01-18)

Via San Felice, 5
I-27100 Pavia
economiaweb.unipv.it

Revised in:
August 2018

Forecasters' utility and forecast coherence

Emilio Zanetti Chini*

University of Pavia
Department of Economics and Management
Via San Felice 5 - 27100, Pavia (ITALY)
e-mail: emilio.zanettichini@unipv.it

FIRST VERSION: December, 2017
THIS VERSION: August, 2018

Abstract

We introduce a new definition of probabilistic forecasts' coherence based on the divergence between forecasters' expected utility and their own models' likelihood function. When the divergence is zero, this utility is said to be local. A new micro-founded forecasting environment, the "Scoring Structure", where the forecast users interact with forecasters, allows econometricians to build a formal test for the null hypothesis of locality. The test behaves consistently with the requirements of the theoretical literature. The locality is fundamental to set dating algorithms for the assessment of the probability of recession in U.S. business cycle and central banks' "fan" charts

Keywords: Business Cycle, Fan Charts, Locality Testing, Smooth Transition Auto-Regressions, Predictive Density, Scoring Rules and Structures.
JEL: C12, C22, C44, C53.

*This paper was initiated when the author was visiting Ph.D. student at CREATES, the Center for Research in Econometric Analysis of Time Series (DNRF78), which is funded by the Danish National Research Foundation. The hospitality and the stimulating research environment provided by Niels Haldrup are gratefully acknowledged. The author is particularly grateful to Tommaso Proietti and Timo Teräsvirta for their supervision. He also thanks Barbara Annicchiarico, Anders Bredahl Kock, Ana Beatriz Galvão, Domenico Giannone, George Kapetanios, Gary Koop, Michele Lenza, Antonio Lijoi, Alessandra Luati, Gael Martin, James Mitchell, Elisa Nicolato, Francesco Ravazzolo, Andrey Vasnev, and Francesco Violante for their feedback and discussions, as well as all seminar participants of the 34th International Symposium on Forecasting, the IAAE 2016 Annual Conference, and the NBP 2017 Workshop in Forecasting in Warsaw. The results of the empirical applications in Section 6 were obtained by developing the codes originally written by James Engel and Barbara Rossi, whom the author gratefully acknowledges. The usual disclaimers apply. This paper has circulated in MIMEO form with the title "*Testing and selecting local proper scoring rules*", and the author has been awarded a Travel Grant by the International Institute of Forecasters, for which the author is grateful. Finally, the author is in debt to the doctors, assistants, and nurses at the Department of Emathology of Policlinico "S. Matteo" of Pavia, without whose (free) care this paper would not have been completed.

1 Introduction

Should expert opinions that forecast an uncertain event be trusted? Are these opinions coherent with the prediction of an economic/statistical model or are they biased by personal judgments? Any user of such forecasts must deal with the forecasters' subjective evaluations of the uncertainty and, as a logical consequence, with the laws of probability. However, the ability to trust such forecasts requires having an effective understanding of the forecasting process. These considerations are the fundamentals of the modern decision-theoretical approach to economic forecasting that were settled by De Finetti (1937) almost a century ago. (See also Elliott and Timmermann (2008)). According to this framework, the forecaster maximizes its own expected utility (minimizes its own expected loss) when it correctly (incorrectly) provides its forecast before the event occurs, inducing a systematic bias[1]. Hence, estimating the forecaster's utility (or reward) function implies the need to check the coherence of his forecast.

According to De Finetti (2017, pp. 62-64), a forecast is coherent if it is additive and bounded. This definition is based on the forecast's coinciding with random probability (or "gain") and then being evaluated by another (but similar) entity. However, in the real world, and specially in economic institutions like central banks, forecasting is a complex procedure that involves many agents, each with its own information and/or utility-maximizing (loss-minimizing) behaviors that produce often widely diverse opinions ("quotations", henceforth) to be evaluated and combined. As a consequence, an official announcement requires a considerable amount of time and often differs from a simple average (or any other aggregated form) of each expert's opinion. In this scenario, De Finetti's definition seems weak. The non-Bayesian literature assumes that the forecasts' coherence coincides with forecasts' optimality (or rationality). In line with the strand of literature that focuses on optimal properties of asymmetric loss functions and in opposition to classical mean absolute/square error, Elliott et al. (2005) obtains by inverse-engineering the forecasters' loss from observed sequences of forecasts[2] and suggests an instrumental variable estimator to overcome this problem. However, Elliott et al. (2005) do not consider any interaction between forecasters and the forecasts' users, so they need an axiomatic approach to identify the conditions of optimality.

This paper introduces a more effective definition of coherence and provides a statistical tool that can help the decision-maker determine whether a quotation is coherent with professional forecasters' utility, in other words, whether the forecasters' judgment bias is statistically rele-

vant. Unlike Elliott et al. and the literature they have generated, we separate the coherence by optimality, even if the optimality is obtained via flexible, asymmetric utility functions. The next section, Section 2, motivates this innovation by running a simple experiment that demonstrates that the standard econometric methodology of predictive density assessments may not be able to recognize the "true" forecasting model. Then Section 3 introduces a novel theoretical framework in response to this critique that is is based on a peculiar mathematical function, the scoring rule (SR, henceforth), which assigns a numerical score to several competing (model-based) density forecasts. The score corresponds to the forecaster's utility for his or her correct forecasts of the event. The forecaster has no incentive to quote a forecast q of an observation x, which is drawn from the (estimated) distribution Q, if the reward for such quotation is bigger or equal to the reward that is associated with the forecast p under $P \neq Q$. A SR with this property is called "proper". (See Gneiting and Raftery (2007) and the aforementioned literature.) A proper SR incents the forecaster to be honest[3], but this "honesty" does not ensure that the forecast's final user will announce a value that differs from that of the estimated model for external reasons. In other words, proper SRs are not robust to judgment bias[4]. Hence, we need a different SR that is robust to judgement bias and can avoid the forecaster's identification problem. We overcome this idiosyncrasy by adopting (m-)local SRs, a peculiar class of utility functions that depend only on the realization x, its predictive density, and its first (m)-derivative(s).

The algebraic properties of local SR have been studied recently (See, in particular, Parry et al. (2012), Dawid et al. (2012) and Ehm and Gneiting (2012)). Here, the focus is on the forecasting environment and particularly on the assumption that the forecast's final user interfaces actively with professional analysts. This environment, called the "Scoring Structure" (SS, henceforth), corresponds to the game-theoretic framework introduced by Vovk and Shafer (2005). To the best of our knowledge, this study contains the first attempt to apply it. We distinguish three types of forecasters:

(i) Forecasters who maximize their (expected) utility based on an improper SR. These forecasters are improperly incented to announce their opinions or, more simply, use some "rule-of-thumb" to arrive at their predictions. For example, a forecaster who uses an asymmetric loss function would be systematically induced to be highly conservative in announcing a boost in the GDP in order to preserve his or her reputation in case of a recession. Such an improper incentive might change based on the phase of the cycle –

that is, their quotations tend to be state-dependent. As a consequence, if the forecast's user does not adopt the forecaster's (biased) opinion, there is "news" (or "surprises") that affects the market. See Gürkaynak et al. (2005) and Blanchard et al. (2013) for empirical evidence and macroeconomic theory.

(ii) Forecasters who maximize their (expected) utility based on a proper but non-local SR, such as the Continuous Rank Probability Score (CRPS). These forecasters quote their true opinions, conditional on their information sets, and it is this delimitation that becomes the real issue, as some of these forecasters – usually a huge number – might incorporate colleagues' quotations in their own information sets, resulting in herding behavior. Such herding could influence the final announcements from the forecast's user with the consequence that the same forecasters will modify their own quotations in the next period. Therefore, it is not possible to determine whether the professional forecasters drive the forecasts' users or vice versa. This scenario explains the recent empirical findings that both survey-based forecasts and market measures are informative sources with which to ascertain the future value of inflation and that the efficiency of these measures is not constant in the long run (Trehan, 2015).

(iii) Forecasters who maximize their (expected) utility based on a local SR. These forecasters ignore confidential information and "rules-of-thumb" and generate their opinions with an econometric/statistical model, which is the provided to the forecast's user. The user arrives at the same conclusion as the forecaster if the user obtains the same likelihood function, in which case the forecast is coherent. However, the user may still disregarding the forecast if, for example, the model is poor or is affected by measurement error.

We assume a forecast environment that is populated by this last type of agents, so our research question is "*Are forecast user's announcements obtained by inverse-engineering the forecasters' utility from observed sequences of forecasts that are coherent with their own quotations?*" Section 4 answers this question by providing a formal test for the null hypothesis of the locality (that is, coherence) of the estimated predictive density generated by the SS.

The small-sample properties and their relevance to the theoretical literature are investigated and discussed in Section 5. Section 6 provides two case studies. The first one introduces a modification of a standard algorithm for detecting the U.S. economy's turning points, and the second one assesses the density forecasts of Norway's output gap (OG, henceforth) published by

the Bank of Norway (BoN, henceforth). Our results reveal that improper SRs affect the dating algorithm of recession events and the BoN's forecasts, thus justifying the assumption of non-coherent forecasting environments. Finally, Section 7 summarizes and concludes. An Appendix provides details on mathematical details, while a separate Supplement gives additional results.

2 Motivating Example: the "Hamill's Paradox"

Central banks must have a correct view of the evolution of the main macroeconomic variables to set their monetary policy. These institutions ask a pool of professional forecasters to quote the expected probability distribution of the variable(s) under consideration. Then they collect and summarize these variables via "fan" charts that correspond to the offcial announcement of their future levels. The evaluation of these forecasts is preliminary and necessary to central banks' decision-making. The following experiment shows how easily one can fail in this fundamental step.

We simulate the dynamics of U.S. Industrial Production (IP, henceforth) using four data generating processes (DGPs, henceforth):

$$
\begin{aligned}
&y_{1,t}^{(i)} = 0.9y_{1,t-1}^{(i)} - 0.795y_{1,t-2}^{(i)} + \epsilon_{1,t}^{(i)}, && \epsilon_{1,t}^{(i)} \sim N(0,1);\\
&y_{2,t}^{(i)} = 0.9y_{2,t-1}^{(i)} - 0.795y_{2,t-2}^{(i)} + (0.02 - 0.4y_{2,t-1}^{(i)} + 0.25y_{2,t-2})^{(i)}G^{(i)}(\boldsymbol{\Xi}) + \epsilon_{2,t}^{(i)}, && \epsilon_{2,t}^{(i)} \sim N(0,1);\\
&y_{3,t}^{(i)} = \epsilon_t^{Unfocused,(i)}, && \epsilon_t^{Unfocused} \sim 0.5 \cdot [N(\mu_t,1) + N(\mu_t + \tau_t, 1)];\\
&y_{4,t}^{(i)} = \epsilon_t^{Hamill,(i)}, && \epsilon_t^{Hamill} \sim N(\mu_t + \delta_t, \delta_t^2);
\end{aligned}
$$

where: $G^{(i)}(\boldsymbol{\Xi}) = \left(1 + \exp\left\{-\left[h(\eta_t)^{(i)} I_{(\eta_t \leq 0)}(y_{t-1}^{(i)} - \bar{y}_t^{(i)}) + h(\eta_t)^{(i)} I_{(\eta_t > 0)}(y_{t-1}^{(i)} - \bar{y}_t^{(i)})\right]\right\}\right)^{-1}$, and

$$
h(\eta_t) \doteq \begin{cases} \gamma_1^{-1}\exp(\gamma_1|\eta_t| - 1) & \text{if } \gamma_1 > 0, \\ 0 & \text{if } \gamma_1 = 0, \\ -\gamma_1^{-1}\log(1 - \gamma_1|\eta_t|) & \text{if } \gamma_1 < 0, \end{cases} \tag{1}
$$

for $\eta_t \geq 0$ and

$$
h(\eta_t) \doteq \begin{cases} -\gamma_2^{-1}\exp(\gamma_2|\eta_t| - 1) & \text{if } \gamma_2 > 0, \\ 0 & \text{if } \gamma_2 = 0, \\ \gamma_2^{-1}\log(1 - \gamma_2|\eta_t|) & \text{if } \gamma_2 < 0, \end{cases} \tag{2}
$$

for $\eta_t < 0$, where $G(\boldsymbol{\Xi})$ is a function of two parameters (γ_1 and γ_2) that govern the transition between the two extreme states $G = 0$ and $G = 1$, a scale parameter c and a link-function $h(\cdot)$; here, we adopt the more compact notation $\boldsymbol{\Xi} = [\gamma_1, \gamma_2, c]$; more in detail, $h(\cdot)$ is function of $\gamma_1 = 50$, $\gamma_2 = -20$, $c = ave(y_t)$, $\eta_t = (y_{t-1}^{(i)} - \bar{y}_t^{(i)})$, $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T} y_t$ and y_{t-1} being the transition variable, $T = 265$ is the length of the time series and $i = \{1, \ldots, I\}$ the number of draws, with $I = 1,000$; the sequences μ_t, τ_t and δ_t, δ_t^2 are identically distributed and mutually independent and $(\delta_t, \delta_t^2) = 0.33 \cdot (0.5, 1) + 0.33 \cdot (-0.5, 1) + 0.33 \cdot (0, 167/100)$. Model 1 is a simple autoregression of order 2; Model 2 is a GSTAR model of the same order, which is similar to the true DGP[5]; Models 3 and 4 are the "Unfocused" and "Hamill's" forecasters, respectively (Hamill, 2001).

Then we compute the PITs that correspond to these forecasts and plot them in Figure 1. Under a perfect forecast, the histogram is perfectly rectangular, but in our experiment, all four histograms are almost rectangular, so none of the alternative forecasters is distinguishable. Such a situation, called Hamill's Paradox, complicates decision-making because each model can correspond to a different policy. Does the IP follow a linear asymmetric law of motion (where cyclical phases are equally possible) or a nonlinear asymmetric one, where future downturns have low probability to happen? From another perspective, is the professional forecaster's quotation the output of a poor model (Unfocused), so no policy should be implemented? Another source of uncertainty in the evaluation of density forecasts that are generated by nonlinear models is that these forecasts are often multimodal. (See Figure 3 in Zanetti Chini (2018)). In fact, these densities might be driven by a mixture of distributions – hence, of utilities – like the "Hamill's" forecaster. Is such multi-modality symptomatic of a change in the forecaster's utility or does it indicate a more complex scenario, such as a cooperative strategy? This experiment offers no answer because, while the literature on PIT has evolved[6], nothing is known about the forecaster's utility that corresponds to the quotation.

An alternative strategy, which originates from Bates and Granger (1969), is to compare and combine the loss functions estimated by alternative model(s) by using some information criterion or encompassing test. (See Geweke and Amisano (2011) and their aforementioned literature.) The only way to overcome the Hamill's Paradox is to use the approach proposed by Gneiting et al. (2007), which uses SRs that satisfy the Savage (1971) representation, which is based on the Brègman (1967) distance. (See Laurent et al. (2013) for a multivariate equivalent.) Mitchell

and Wallis (2011) criticize Gneiting et al.'s approach because they see the DGP that Hamill uses as not robust to some basic time-series features, which implies that traditional diagnostic tools can be applied successfully. However, our example overcomes their critique by showing that Hamill's paradox is still an issue when a standard and sufficiently general specification is set. The number of functions enclosed in this family is considerable (Table 1 of the Supplement), which justifies our new theoretical framework, as explained in Section refsec:theory.

3 Theoretical framework

3.1 Set up

We assume that the probabilistic forecast of an economic event is the output of a one-period game with three players: the Forecaster, the Forecast User (or Skeptic), who has capital $\mathcal{K}$ to preserve; and Reality (or Nature). The Skeptic seeks always to refuse the Forecaster's quotations and eventually cooperates with Reality; however, no matter how the Skeptic plays, Reality acts as though the Skeptic does not win the game (called the "*excluded gambling system hypothesis*" or Cournot's Principle). These players act according to the following Forecasting Protocol:

1. $\mathcal{K}_0 := 1$;
2. Forecast User announces a bounded function $\mathcal{S} : [0, 1] \to \mathcal{R}$;
3. Forecaster announces his quotation $\mathrm{Q} \in \mathcal{R}$;
4. Reality announces $X \in \{0, 1\}$;
5. $\mathcal{K}_1 = \mathcal{K}_0 + \mathcal{S}\ (\mathrm{X} - \mathrm{Q})$,

where the binary event by which Reality materializes and the [0, 1]-support of $\mathcal{S}$ are used only for ease of illustration. The restriction on the Skeptic is that she or he must choose $\mathcal{S}$ so her or his capital remains non-negative ($\mathcal{K} \geq 0$) no matter what values the Forecaster and Reality announce for Q and X. The winner is the Skeptic if $\mathcal{K}_1 >> \mathcal{K}_0$. Otherwise, the Forecaster wins.

Remark 1. The game illustrated here is a simplified version of the "Forecasting (Sub-)game" in Vovk and Shafer (2005, p. 753). With respect to these authors, to ease the statistical treatment in Section 4, we avoid the recursion corresponding to the $n = 1, 2, \ldots$ times that the game is

re-iterated. This simplification can be removed by assuming a properly calibrated algorithm, which ensures the main restrictions and assumptions about the players for each recursion.

Remark 2. Step 2 of the Protocol is essentially an application of one of Patton (2017)'s main conclusions. He demonstrates that the forecast rankings are generally sensitive to the choice of a proper SR, and, correcting Gneiting (2011) main result, asserts that forecasters should be told ex-ante what utility functions will be used to evaluate their quotations[7].

Remark 3. Step 5 of the Protocol is a test on the forecast's coherence in terms of the Skeptic's utility, which implies that the Forecaster's reward cannot be augmented after his or her quotation. In principle, the assumption that Reality can cooperate with the Skeptic implies that, when the game is repeated n times, the sequences of outcomes S_n Q_n, X_n do not necessarily coincide with realizations of a classical stochastic process. As a consequence, classical hypothesis testing and inference are not available. Nevertheless, the excluded gambling system hypothesis allows both of them to be used. (See Shafer and Vovk (2001, Chapter 8.1).)

We are interested in forecasting the economic variable Y, which is observed by time series $Y_t = \{y_t\}_{t=1}^{T} = \{y_1, \ldots, y_t, \ldots, y_T\}^{\mathsf{T}}$, with upper index "$\mathsf{T}$" denoting the transposition, being fully represented by an information set $\mathcal{F}_t = \{y_{t-1}, y_{t-2}, \ldots\}$ and a probability density $P(Y_t)$, and having a probability density $P(Y_t)$. Let us denote the (one-step-ahead) probability forecast of Y_t as $P(Y_{t+1})$, the best Forecaster's judgment of the distributional forecast of Y_t as $Q(Y_t)$ (we omit Y_t for notational convenience), and y a draw of Q that materializes in $T+1$ and the predictive cumulative distribution function that is associated with the materialization of y as $F(y)$. Let the parameters of the Skeptic's and Forecaster's distributions be denoted as P and Q, respectively, and let $\mathcal{L}(\Theta)$ and $\mathcal{L}(\Psi)$ denote their likelihood functions and the associated parameters vectors, respectively. A hat indicates estimates. A rolling window consisting of the past m observations is used to fit a density forecast for a future observation that lies k time steps ahead. Suppose that $T = m + n$. At times $t = m, \ldots, m + n - k$ estimated density forecasts $\hat{P}(Y_{t+k})$ and $\hat{Q}(Y_{t+k})$ for Y_{t+k} are generated, each of which depends only on $Y_{t-m+1}, \ldots, Y_t$. Let $\mathcal{X}$ be a set of the possible forecaster's outcomes, $\mathcal{P}$ the family of distributions on $\mathcal{X}$ in which P belongs, and $\mathcal{A}$ an algebraic subset of $\mathcal{X}$ representing the set of actions. In particular, if the sample space is discrete (that is, dichotomous for events like the probability of a recession, or categorical for, say, the ranking position of a firm or State), P is defined by $\mathcal{P} = \{\mathbf{p} \in \mathcal{A} : \sum_x p_x = 1\}$, the set of all real vectors corresponding to strictly

positive probability measures; if the sample space is continuous (like the conditional mean of an economic time series), P is defined by $\mathcal{M}$, the set of all distributions on $\mathcal{X}$ that are absolutely continuous with respect to a $\sigma-$finite measure μ. The same applies to $\mathbf{q}$. The Forecaster seeks to solve a decision problem defined by the triple $\left\{\mathcal{X}, \mathcal{A}, \mathcal{U}(P, \mathrm{a})\right\}$, where: $\mathcal{X}$ is as previously defined; $\mathcal{A}$ is the action space; and $\mathcal{U}(P, \mathrm{a}^*)$ is a real-valued utility function that represents the reward obtained by the Forecaster as the result of minimizing the discrepancy in his or her own quotations and of action of the action $a^* \in \mathcal{A}$, which maximizes the expected utility computed using the density P, which is believed to be the true DGP, with the expected loss denoted as $EU := \int U(P, \mathrm{a}) P(Y_t) dY_t$.

3.2 Main Definitions and Representation of the Forecasting Environment

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

Definition 1 (SRs, entropy/divergence functions, scoring structure)**.** We define:

i. *SR* the function $S(x, Q) := U(P, \mathrm{a}_Q)$. Namely:

   (a) A SR $S : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ is local of order m (or m-local) if it can be expressed in the form of:

$$S(x, Q) = \mathfrak{s}(x, q(x), q'(x), q''(x), \ldots, q^{(m)}(x)), \tag{3}$$

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

   (b) A SR $S(x, Q)$ is (strictly) proper relative to the class of probability measures $\mathcal{P}$ if

$$S(P, P) \leq S(P, Q) \quad \forall \quad P, Q \ \in \ \mathcal{P} \tag{4}$$

   with equality if (and only if, for strict properness) $Q = P$;

ii. The *entropy function* is the function $H(P) := S(P, P) \equiv \sup_{Q \in \mathcal{P}} S(P, Q)$;

iii. The *divergence function* is the function $D(P, Q) := H(P) - S(P, Q)$;

iv. The *Scoring structure* is the 7-ple $\mathcal{SS} := \{Y_t, \mathcal{F}_t, \mathcal{X}, \mathcal{P}, S(\cdot,\cdot), H(\cdot), D(\cdot,\cdot)\}$.

Remark 4. All of the functions $S(\cdot,\cdot)$ in Definition 1 can be interpreted in terms of utility: $S(x,P)$ is the Forecaster's reward if event x (truly) materializes. Since $S(\cdot)$ is defined on the extended real line, the expected forecaster's utility, conditional to Q, can be denoted as $S(P,Q) \equiv \int_{-\infty}^{+\infty} S(P,x)dQ(x)$. $H(P)$ can be interpreted as the maximum possible of the utility that the Forecaster can achieve using Reality's true DGP to predict P. The divergence function is the difference between the maximum utility and the utility achieved by predicting the quoted predictive distribution Q, given the true distribution P. Hendrickson and Buehler (1971)provide the necessary and sufficient conditions under which $D(P,Q)$ admits a Brègman-Savage representation.

Remark 5. This definition of SS is highly general and is used only to consider Reality's, the Forecaster's, or the Skeptic's point 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. (See Appendix A.1.) If the variable of interest is a continuous function, $\mathcal{P}$ is substituted by $\mathcal{M}$.

Definition 2 (Structural Coherence)**.** The (h-step ahead) forecast y_{t+h} obtained by $P(\hat{\Theta}; x_t) \in \mathcal{P}$ (or $\mathcal{M}$) is *coherent relatively to the scoring structure* (or, more simply, is *structurally coherent*) if there is one-to-one mapping between $\mathcal{L}(\hat{\Theta}; x_t)$ and $\mathcal{L}(\hat{\Psi}; x_t)$.

Proposition 1. *The Forecaster's reward $S(P,Q)$ is a proper SR if and only if A1 – A5 are satisfied.*

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

Remark 6. A1 – A3 are necessary (but not sufficient) to define the Forecaster's reward as SR. In particular, A1 encompasses the three "basic assumptions" discussed in Dawid (2007)[8] and suggests that the reward is measurable with respect to $\mathcal{A}$ and quasi-integrable with respect to all $P \in \mathcal{P}$. 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 Forecaster has no loss only if his or her DGP coincides with that of Reality. A6 is fundamental to characterizing a general family of SRs for the case that every $Q \in \mathcal{P} = \mathcal{A} = \mathcal{M}$ has a density $q(x)$ with respect to $\mu \in \mathcal{X}$, that is the

Brègman score:

$$S(x,Q)^B \doteq \psi'[q(x)] + \int \Big\{\psi[q(x)] - q(x)\psi'[q(x)]\Big\} d\mu(x) \tag{5}$$

with associated *Brègman divergence*:

$$d(P,Q)^B \doteq \int \Big\{\Big(\psi[p(x)] + [p(x) - q(x)]\psi'[q(x)]\Big) - \psi[p(x)]\Big\} d\mu(x), \tag{6}$$

where ψ is a (strictly) concave function and ψ' a subgradient of ψ. This is a very general class of non-metric distance that can characterize most of the SRs described in Table 1 of Supplement. We are particularly interested in the special case that

$$\psi_x = k(x) - \lambda \log(x), \tag{7}$$

where k is set to zero without loss of generality. Under (7) the forecasts generated by $\mathcal{M}$ are coherent with a given SS. Finally, A7 is necessary to apply Amisano and Giacomini (2007)'s predictive ability test on the SS' outputs.

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

Proposition 2. *Let $S(x,Q)$ be an SR, possibly the Brègman-Savage representation, with q-function $\mathfrak{s}$. Then, $S(x,Q)$ is local and strictly proper if and only if $\mathfrak{s}$ is such that:*

$$\mathbb{L}\mathfrak{s} = 0, \tag{8}$$

where: $\mathbb{L} := \sum_{k\geq 0}(-1)^k \mathbb{D}^k q_0 \frac{\partial}{\partial q_k}$, $\mathbb{D} := \frac{\partial}{\partial x} + \sum_{j>>0} q_{j+1}\frac{\partial}{q_j}$, $\mathbb{D}$ 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 (8) is called the *Key Condition*. For purely theoretical reasons, the same theorem requires two other conditions concerning the representation of $\mathfrak{s}$ via Lagrange operators (Appendix A.2). Nevertheless, the Key Condition is sufficient (and, to the best of our knowledge, it is the only one available) to identify an empirically testable hypothesis for the assessment of the logarithmic form of the Forecaster's utility.

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

Theorem 1. *A density forecast Q is structurally coherent if and only if $S(x, Q)$ is local.*

Proof. See Appendix A.2. □

4 The Locality Test

To test the hypothesis that the equation (8) is verified by the data, we assume that the time-invariant q-function s generating; and $S(x; Q)$ is part of a smooth transition auto-regressive scoring structure (SS-STAR, henceforth) and is treated as observed transition variable (Chan and Tong, 1986)[9]. This treatment is necessary to set up the null hypothesis and introduce an LM-type test using a linearization of the SS-STAR, which leads to an equivalent auxiliary model with augmented regressors, the number of which depends on the type of non-linearity of the same structure. This artificial model can be investigated by standard inference.
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:

$$\begin{aligned} y_t &= \boldsymbol{\phi}^{\mathsf{T}}\mathbf{z}_t + G(\gamma, \mathbf{w}_t, \mathbf{c}_k))\boldsymbol{\theta}^{\mathsf{T}}\mathbf{z}_t + \epsilon_t, \quad \epsilon_t \sim iid(0, \sigma^2) \\ G(\gamma, \mathbf{w}_t, \mathbf{c}_k) &= \left(1 + \exp\left\{-\gamma \prod_{k=1}^{K}(\mathbf{w}_t - \mathbf{c}_k)\right\}\right)^{-1}, \quad \gamma > 0, \quad c_1 < \cdots < c_k, < \cdots < c_K, \end{aligned} \tag{9}$$

where: $\mathbf{z}_t = (1, y_{t-1}, \ldots, y_{t-p})^{\mathsf{T}}$ are the autoregressive covariates; $\boldsymbol{\phi} = (\phi_0, \phi_1, \ldots, \phi_p)^{\mathsf{T}}$ are the linear part parameters; $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_p)^{\mathsf{T}}$ are the nonlinear part parameters; γ is the slope parameter; $\mathbf{c}_k = (c_1, \ldots, c_K)$ denoting the (eventually, multiple) location parameters; $\mathbf{w}_t = a^{\mathsf{T}}\mathbf{z}_t \odot \mathbf{s}$ is a composite transition variable, with $a = [a_1, \ldots, a_p]^{\mathsf{T}}$, $a_i = \begin{cases} 0 \text{ if } i = d \\ 1 \text{ if } i \neq d \end{cases}$ indicating that delay parameter d, which is such that $1 \leq d \leq p$, is unknown; $\mathbf{s} = vec(\mathfrak{s} \bigotimes \mathbf{i})$ with $\mathfrak{s}$ is a scalar denoting a generic proper SR as in Definition 1 and $\mathbf{i}$ is a one-vector of the same dimensions of $\mathbf{z_t}$. The most common choices for K are $K = 1$, in which case the parameters $\boldsymbol{\phi} + \boldsymbol{\theta}G(\gamma, \mathbf{w}_t, \mathbf{c}_k)$ change monotonically as a function of $\mathbf{w}_t$ from $\boldsymbol{\phi}$ to $\boldsymbol{\phi} + \boldsymbol{\theta}$ and $K = 2$, in which case the parameters $\boldsymbol{\phi} + \boldsymbol{\theta}G(\gamma, \mathbf{w}_t, \mathbf{c}_k)$ 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 STAR (SS-ESTAR) model. When $\gamma \to \infty$, the equation (9) becomes a two-regime threshold auto-regressive SS (SS-TAR).

(Wee Tong (1983)).

Remark 7. The (nonlinear) SS so defined is an algorithm that applies the forecasting protocol defined in Section 3. Its use requires three steps: (i) the Skeptic specifies the form of $\mathfrak{s}$ that will be adopted to evaluate the Forecaster; (ii) the Forecaster estimates $Q(\hat{y}_{t+h})$, which is the estimated h-step-ahead forecast of conditional density of variable y and applies s to it, which is the step in which $\mathbf{w_t}$ is computed; and (iii) the evaluated quotation is compared with the realizations y_t via (9).

Remark 8. The mechanics of the forecasting exercise implemented by the Forecaster is independent of the form of the SS: no restrictions or assumptions in the forecasting model or in the methodology adopted to obtain $Q(y_{t+h})$ is needed. Equation (9) is necessary only as a convenient way to test the null hypothesis of locality.

The null hypothesis of locality can be investigated as follows:

Proposition 3. *Let y_t be a stochastic process generated by a SS-STAR represented by (9). Then:*

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

$$H_0 : \gamma = 0 \;\; vs \;\; H_1 : \gamma \neq 0 \quad in \;\; (9), \tag{10}$$

which can be measured by the following LM statistics:

$$S(\boldsymbol{\Xi})^{LM} = \hat{\sigma}^{-2}\hat{\mathbf{U}}'\hat{\mathbf{D}}_\mathbf{2}(\hat{\mathbf{D}}_\mathbf{2}'\hat{\mathbf{D}}_\mathbf{2} - \hat{\mathbf{D}}_\mathbf{2}'\hat{\mathbf{D}}_\mathbf{1}(\hat{\mathbf{D}}_\mathbf{1}'\hat{\mathbf{D}}_\mathbf{1})^{-1}\hat{\mathbf{D}}_\mathbf{1}'\hat{\mathbf{D}}_\mathbf{2})^{-1}\hat{\mathbf{D}}_\mathbf{2}'\hat{\mathbf{U}} \sim \chi^2_n \tag{11}$$

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

$$\begin{aligned} LM_1 &= (SSR_0 - SSR)/\hat{\sigma_v}^2 \sim \chi^2_{3p} \;\; if \;\; K = 1 \;\; in \;\; (9) \\ LM_2 &= (SSR_0 - SSR)/\hat{\sigma}^2_{v1} \sim \chi^2_{2p} \;\; if \;\; K = 2 \;\; and \;\; c_1 = c_2 \;\; in \;\; (9) \\ LM_3 &= (SSR_0 - SSR)/\hat{\sigma}^2_{v2} \sim \chi^2_{p} \;\; if \;\; K = 2 \;\; and \;\; c_1 \neq c_2 \;\; in \;\; (9), \end{aligned} \tag{12}$$

where SSR_0 and SSR are the sum of the squared residuals of SS-STAR (9) 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-STAR(p); p is the autoregressive order of the same SS-STAR. F-type tests equivalent to LM statistics in (12) *are preferable in small samples.*

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

Remark 9. The SR, assumed as a transition variable of a SS-STAR, is necessary to avoid the investigator's (possibly, the same Skeptic) assuming erroneously any functional form of SR when doing *ex-post* evaluation (as represented in step 5 of the Forecasting Protocol).

5 Simulation Study

5.1 Simulation Design and Results

We consider two different DGPs:

$$y_{1,t}^{(i)} = 0.4y_{1,t-1}^{(i)} - 0.25y_{1,t-2}^{(i)} + (0.01 - 0.9y_{1,t-1}^{(i)} + 0.795y_{1,t-2})^{(i)}G^{(i)}(\gamma, \mathbf{w_t}, c) + \epsilon_{1,t}^{(i)}\,, \qquad (13)$$

and

$$y_{2,t}^{(i)} = 0.8y_{2,t-1}^{(i)} - 0.7y_{2,t-2}^{(i)} + (0.01 - 0.9y_{2,t-1}^{(i)} + 0.795y_{2,t-2}^{(i)})G^{(i)}(\gamma, \mathbf{w_t}, c) + \epsilon_{2,t}^{(i)}\,, \qquad (14)$$

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 local, near-to-local, and non-local 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 1 reports the results of the Monte Carlo simulation of the locality test for the statistics F_1 and F_2 from the hypothesis system (12) discussed in Section 4. 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 locality test's empirical power. Thus, the test correctly accepts the hypothesis of locality more easily when the SS is highly nonlinear than it does in the opposite case of quasi-linear behavior, a feature we call *structure linearity bias*. This finding is counter-balanced by the functional form of the SRs' having no role in the test's empirical power. Table 2 reports the results of a simulation of the same two DGPs, where we fixed $\gamma = 10$ and all 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 locality test is invariant to the form of the SR that is assumed to drive the Forecaster's 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 $H(\cdot, \cdot)$.

5.2 Discussion

This simulation experience provides several lessons. First, the role of the SR locality hypothesis is not trivial nor easy to understand. 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 Forecaster is suspected to have when he or she produces quotations.

The score's invariance is a direct consequence of Vovk and Shafer (2005) excluded gambling system hypothesis. The forecasts (drawn by the Forecaster) 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 Forecaster's utility. According to Lindley (1982), 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 that are conditional on the same event, thus enhancing the status of the probability transform of the obtained value x[10]. In this sense, the results of our simulations are fully consistent with the De Finetti–Lindley theory. No less important, 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 (2017). 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 Skeptic deals with Forecasters, even if the Skeptic specifies (axiomatically) ex-ante the exact utility function that will be used to evaluate the Forecasters, as required by Step 2 of Forecasting Protocol, the Skeptic will never know, ex-post, if the same utility function is the one the Forecasters used. This sort of "undeterminancy" is the motivation for adopting the locality as a criterion for assessing the forecasts. The locality tells the Skeptic whether Barnard et al. (1962) likelihood principle, according to which all the evidence in a sample that is relevant to the model parameters is contained in the likelihood function, holds. In this case, the forecast must necessarily be driven by some function derived from the likelihood. Since the Skeptic is supposed to have sound knowledge about the estimation methods used to verify the Forecaster's work, any deviations from likelihood are likely represented by judgments, that is, opinions that are not justified by any statistical model.

6 Applications

6.1 Assessing Recession Probability in the U.S. Business Cycle

We evaluate the U.S. business cycle using a new version of the Bry-Boshan Quarterly (BBQ, henceforth) algorithm introduced by Harding and Pagan (2002) and Engel et al. (2005). The BBQ is an advancement of Bry and Boschan (1971) algorithm for detecting turning points in the U.S.'s GDP. (See also Artis et al. (2004)). This new version is here defined Scored BBQ (SBBQ, henceforth). Our objective is to explain the role of SS in its multiple specifications and, just as important, to identify the problems that occur when improper and/or non-local SRs and SS are assumed.

We use the IP in quarterly data that covers from 1947Q1 to 2013Q1 as a proxy variable in logarithmic transformation. We measure the effects of adopting a local SR in the BBQ algorithm using a simple AR(5) model to estimate the predictive density evaluated via logarithmic (that is, local) SR. This model corresponds to an SS in equation (9), where $\gamma = 0$. Several findings emerge from the results (reported in Table 3). First, the Logarithmic Score (LogS) conveys the minimum value between the set of SRs adopted, so it is the optimal one[11]. In addition, not surprisingly, the associated descriptive statistics coincide with the statistics in which no SR is adopted – that is, when there is no evaluation. Therefore, in this scenario the Forecaster's quotations are structurally coherent with the Skeptic's quotations, the latter of which correspond to the results obtained by running the standard BBQ algorithm. A second finding is that, according to the score invariance principle, most of the indicators' values do not vary when the functional form of the SR varies. Figure 2 shows many of these results. When log-level data are used, the estimated recessions are almost always coincident with the National Bureau of Economic Research (NBER, henceforth) recession dates, with the exception of the 1975-6 oil crisis, which is recognized only by LogS and the weighted power Score (WPwrS). The LogS is also sensitive to stagnation events like those at the of the 1960s and anticipates the Gulf-War recession in 1991-2. Growth rates convey slightly different results and, in particular, lead to over-evaluation of the recessions: the standard (unscored) BBQ overreacts in the first part of the sample, where the number of recession episodes is doubles the NBER recession dates. On the opposite side, the BBQ does not recognize the first episode in the 1980s and over-evaluates the cyclical movements after 1992. This pattern holds also for many cases of the SBB algorithm. Nevertheless, under LogS, the recessions of the early 1980s are recognized, albeit with

a large lead. Then we repeat the analysis by assuming an SS where $\gamma = 5.0$ and Forecasters use a STAR(5) model with d = 4 and three regimes based on the General-to-Specific modelling strategy in Teräsvirta et al. (2010). Table 4 shows that, under nonlinear (that is, non-local) SS, more SRs convey the same descriptive statistics as the "no score" case – namely, the LogS as in the previous analysis, the quantile score (QuantS), and the censored likelihood score (CsLS). All three of these SRs are proper. Figure 3 shows that nonlinear SSs are more compatible with the Forecaster's quotation than linear ones are: in fact, when recession probabilities are driven by a nonlinear SS, the SBB is less likely to consider years of stagnation to be recessions, as shown by the indicator's having been 1 from 1962 to 1971. In general, the LogS respects the NBER recession dates almost everywhere, apart from the end of 1989 to 1990. Non-Brègman-Savage SRs like CRPS are not sensitive to any recession since the first oil crisis (1973-74) until 1991. On the opposite side, WPwrS and WPSph over-estimate the recession episodes after 1982, as they remain at 0 until 2000. The LogS is less biased than other measures are, although the 1982 recession remains over-estimated.

These findings are supported by the formal LM locality test on the log-level series, the results of which are reported in Table 5: the locality hypothesis is strongly rejected for all specifications of nonlinear functions $G(\cdot)$. Finally, we perform a robustness check by running Amisano and Giacomini (2007) test for the null hypothesis that linear and nonlinear SS specifications for the same variable per-form equally well according to a rolling window forecasting scheme. Table 6 shows that the SS-LSTAR is preferred when Brègman-Savage SRs are adopted, although only LogS strongly rejects the null hypothesis. On the opposite side, no evidence of superior forecasting ability is found when non-Brègman-Savage families of scores are used.

6.2 Assessing the Bank of Norway's Fan Chart

The OG is one of the most important variables used in macroeconomics because it measures the percentage deviation between the actual level of GDP and the projected GDP. According to the most widely used macroeconomic models and practices, this measure is a proxy for the need for a central planner's intervention via the monetary channel. Correct assessments of the OG forecasts are part of each central bank's duties. The BoN'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. Using our framework, and differently from the previous application to the U.S. IP, we take the BoN forecasts as primitive observations, so these are y_t in equation (9). 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). (See the BoN models web page for references.) The last ones take the role of the composite transition variable $\mathbf{w_t}$, representing the Forecaster's quotations. In terms of the theoretical framework, $\gamma = 0$, the final BoN announcements correspond to the estimated fan charts; that is, the latter are perfectly coherent with internal forecasts. Similar to the previous application on IP, Table 5 rejects this hypothesis, so we must assume a non-negligible amount of bias in the BoN's fan charts. In line with this finding, we assume that the Forecaster adopts a Logistic STAR model, which has to be compared with the final announcement, represented by downloadable BoN fan charts. Therefore, we run the Amisano-Giacomini test for many SRs computed in a rolling-windows forecasting scheme, where the window has length of $m = 6$ quarters. (Among others, and despite some differences in theoretical and empirical framework, see Gneiting and Ranjan (2011).) Under LogS, the t-statistic indicates whether the distance between the BoNMPR forecasts and a forecast obtained by an econometric model is significant.

Table 7 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. The values and p-values show that 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 statistical coherence of the quotation. On the other side, it does not reject the null hypothesis if any of several other improper functionals, such as the WPwrS, most Weighted Pseudo-Spherical (WPseudoSph) scores, and Log-Cosh (LCS) scores, are used. The Supplement provides details on each of these functionals.

6.3 Discussion

What do we learn from these applications? Concerning the evaluation of business cycle downturns, our results show that the dating algorithms – and, thus, the probability of recession – are sensitive to the locality hypothesis, as different SRs convey the same cycle indicator. This result seems counterintuitive: since most of these estimated SRs differ (and, in some cases, they are consistently different), we would expect that each of these functionals corresponds to a unique value. On the other side, the score invariance principle that is empirically demonstrated in Section 4 supports Patton (2017)'s conclusion that an ex-ante selection of SRs for evaluation according to their axiomatic properties is not sufficient to obtain a coherent forecast. In fact, all the three of the SRs that are associated with the "coherent" descriptive statistics (Table 4) and forecast (Table 6) of the U.S. business cycle have Brègman-Savage representation. This apparent contradiction with our original goal of obtaining a unique and coherent evaluation of economic forecasts is a basis on which to motivate the use of conditional predictive ability tests when the null hypothesis of locality of forecasting environment is rejected: there would no need to compare two density forecasts if one of them were known to be coherent.

Concerning the assessment of the BoN's fan charts, we have to assume that the measure of the Forecaster's coherence can be done by simply averaging a time series of each quotation. More general combination schemes have recently become available: Kapetanios et al. (2015) suggest a sieve estimator of weighted means, where weights are allowed to be nonlinear combinations of density functions, and Billio et al. (2013) develop a combination method that assigns time-varying weights to competing densities via a Markov-Switching state-space model. In principle, all these methods can be nested in our framework.

The proposed methodology is a stylized way to address the elicitation of the Forecaster's true utility. For example, we assumed the SR is known (because of the ex-ante elicitation in Step 2 of the Forecasting Protocol) and observed that it exploits the properties of the STAR family of models. Nevertheless, in the real world, the Skeptic can never be sure that the Forecaster is considering the same utility function, which requires that the observability assumption be relaxed. In this sense, a state-space representation of the SS could be a useful development. Moreover, the game-theoretic framework that constitutes the skeleton of the SS adopted in this paper is a simple version of more general games in that we assumed that there is no uncertainty in the Forecaster's activity, whereas the Forecasting Game II in Vovk and Shafer (2005) seminal

paper assumed it acts as a fourth player. (See Vovk and Shafer (2005, p. 754).) Moreover, no dynamics in the game underline the assumed Forecasting Protocol, suggesting from a theoretical perspective that the Forecaster cannot use any further information he or she receives, nor can the Skeptic account for the Forecaster's past behavior. This implication limits many of the applied contexts because the Skeptic would be able to test whether a Forecaster's quotation is biased but would not know in what direction the bias acts. Finally, we assumed that the Forecaster is a single entity (or, equivalently, a set of entities with perfectly homogeneous behavior that can be represented by a single agent) so, when different types of SR were adopted in the course of our applications, there were no interactions between them. Hence, further research is necessary to generalize the SS approach in these (and, possibly, other) directions.

7 Conclusions

We introduced a novel frequentist framework named the Scoring Structure that assumes interactions between a Forecaster, a Forecast User, and Reality to make coherent evaluations of econometric forecasts. This framework allows econometricians to build an LM-type test to verify the hypothesis of locality of the Forecaster's expected utility. The test's empirical power properties are consistent with the fundamental requirements stated in the literature on decision theory. Because of its generality and flexibility, the Scoring Structure is a fundamental tool with which to elicit the Forecaster's utility in several economic applications. The specification used in this paper is simple with respect to real-world needs, so we suggest using the new approach in companion with standard predictive ability tests and encourage further research efforts to refine this methodology by relaxing the Scoring Structure's theoretical assumptions and related operational methods.

Notes

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

2 See also Christoffersen and Diebold (1996, 1997); Patton and Timmermann (2007a), among others. Patton and Timmermann (2007b) consider the case in which the utility functions are unknown, but this last strand of literature considers only point forecasting.

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

4 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 Forecaster's forecast is not easy; see also Table 1 in Jose et al. (2008).

5 The true DGP is Model 2, but with $\gamma_1 = 20$ and $\gamma_2 = -5$ and the autoregressive parameter of the nonlinear part augmented to 0.6 and 0.45 (instead of 0.4 and 0.25). For details on this peculiar generalization of the smooth transition autoregressive family of models, we refer the reader to Canepa and Zanetti Chini (2016)) and notice that both DGP2 and the true DGP are similar to those used in Teräsvirta (1994), equation (4.1). A large strand of literature demonstrates that this variable is typically nonlinear. See Anderson and Teräsvirta (1992), among others.

6 See Rossi and Sekhposyan (2013) for statistical inference with unstable forecast environments and González-Rivera and Sun (2017) and the aforementioned literature for a generalization of PITs.

7 According to Patton (2017, p. 3), "[. . .] *Specifying the target functional is generally not sufficient to elicit a forecaster's best (according to a given, consistent, loss function) prediction. Instead, forecasters should be told the single, specific loss function that will be used to evaluate their forecasts.*"

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

9 This test was built in previous versions of this paper to assume the generalized version of the STAR model presented in Canepa and Zanetti Chini (2016). However, this complication does not improve the results presented here. The equivalent tables with the result of the simulations with STAR model can be provided under request.

10 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 $P(x) = p$.*"

11 In this peculiar example, the SRs with negative value are not meaningful because they are associated with the probability of recession events and do not allow the application of the algorithmic procedure.

References

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

Anderson H, Teräsvirta T. 1992. Characterizing nonlinearities in business cycles using smooth transition autoregressive models. *Journal of Applied Econometrics* **7**: 119–136.

Artis M, Marcellino M, Proietti T. 2004. Dating Business Cycles: A Methodological Contribution with an Application to the Euro Area. *Oxford Bullettin of Economics and Statistics* **66**: 537–565.

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

Bates J, Granger C. 1969. The combination of forecasts. *Operations Research Quarterly* **20**: 451–468.

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

Billio M, Casarin R, Ravazzolo F, van Dijk H. 2013. Time-varying Combinations of Predictive Densities Using Nonlinear Filtering. *Journal of Econometrics* **177**: 213–232.

Blanchard O, L' Huillier J, Lorenzoni G. 2013. News, Noise and Fluctuations: An Empirical Expoloration. *The American Economic Review* **103**: 3045–3070.

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.

Bry G, Boschan C. 1971. *Cyclical analysis of time series: Selected procedures and computer programs.* New York: NBER.

Canepa A, Zanetti Chini E. 2016. Dynamic asymmetries in house price cycles: A generalized smooth transition model. *Journal of Empirical Finance* **37**: 91–103.

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

Christoffersen P, Diebold F. 1996. Further Results on Forecasting and Model Selection under Asymmetric Loss. *Journal of Applied Econometrics* **13**: 808–817.

Christoffersen P, Diebold F. 1997. Optimal prediction under asymmetric loss. *Econometric Theory* **11**: 561–571.

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

Dawid A, Lauritzen S, Parry M. 2012. Proper Local Scoring Rules on Discrete Sample Space. *The Annals of Statistics* **40**: 593–608.

De Finetti B. 1937. La prévision: ses lois logiques, ses sources subjectives. *Annales de l' Institut Henri Poincarè* **7**: 1–68.

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.

Ehm W, Gneiting T. 2012. Proper Local Scoring Rules on Discrete Sample Space. *The Annals of Statistics* **40**: 609–637.

Elliott G, Komunjer I, Timmermann A. 2005. Estimation and Testing of Forecast Rationality under Flexible Loss. *Review of Economic Studies* **72**: 1107–1125.

Elliott G, Timmermann A. 2008. Economic Forecasting. *Journal of Economic Literature* **46**: 3–56.

Engel J, Haugh D, Pagan A. 2005. Some methods for assessing the need for non-linear models in business cycle analysis. *International Journal of Forecasting* **21**: 651–662.

Geweke J, Amisano G. 2011. Optimal prediction pools. *Journal of Econometrics* **164**: 130–141.

Gneiting T. 2011. Making and evaluating point forecasts. *Journal of American Statistical Association* **106**: 746–762.

Gneiting T, Balabdaoui F, Raftery A. 2007. Probabilistic forecasts, calibration and sharpness. *Journal of Royal Statistical Society* **69**: 243–268.

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, Sun Y. 2017. Density forecast evaluation in unstable environments. *International Journal of Forecasting* **33**: 416–432.

Gürkaynak R, Sack B, Swanson E. 2005. The Sensitivity of Long-Term Interest Rates to Economic News: Evidence and Implications. *The American Economic Review* **95**: 425–436.

Hamill T. 2001. Interpretation of rank histograms for verifying ensemble forecasts. *Monthly Weather Review* **129**: 550–560.

Harding D, Pagan A. 2002. Dissecting the cycle: a methodological investigation. *Journal of Monetary Economics* **49**: 365–381.

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

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

Kapetanios G, Mitchell J, Price S, Fawcett N. 2015. Generalised Density Forecast Combinations. *Journal of Econometrics* **188**: 150–165.

Laurent S, Rombouts J, Violante F. 2013. On loss functions and ranking forecasting performances of multivariate volatility models. *Journal of Econometrics* **173**: 1–10.

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.

Mitchell J, Wallis K. 2011. Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. *Journal of the Applied Econometrics* **26**: 1023–1040.

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

Patton A. 2017. Comparing Possibly Misspecified Forecasts. Duke University Working Paper.

Patton A, Timmermann A. 2007a. Properties of optimal forecasts under asymmetric loss and nonlinearity. *Journal of Econometrics* **140**: 884–918.

Patton A, Timmermann A. 2007b. Testing Forecast Optimality Under Unknown Loss. *Journal of the American Statistical Association* **102**: 1172–1184.

Rossi B, Sekhposyan T. 2013. Conditional predictive density evaluation in the presence of instabilities. *Journal of Econometrics* **177**: 199–212.

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

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

Tong H. 1983. *Threshold Models in Non-Linear Time Series Analysis.* Number 21 in Lecture Notes in Statistics. New York: Springer-Verlag.

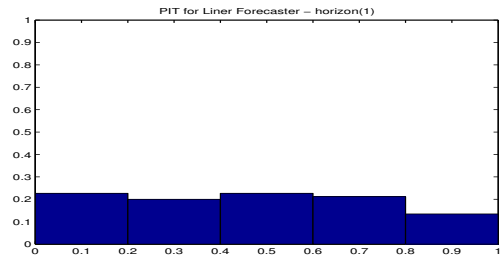
Trehan B. 2015. Survey Measures of Expected Inflation and the Inflation Process. *Journal of Money, Credit and Banking* **47**: 207–222.

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

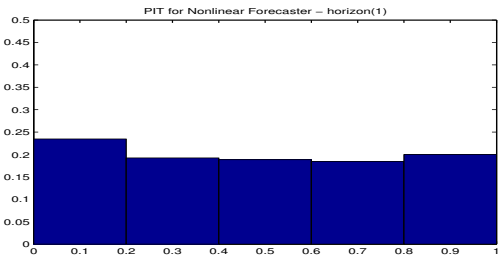
Zanetti Chini E. 2018. Forecasting dynamically asymmetric fluctuations of the U.S. business cycle. *International Journal of Forecasting* **34**: 711–732.

Figure 1: The Hamill's paradox

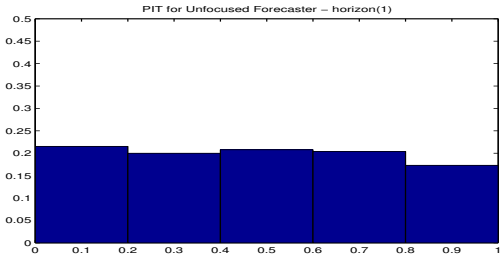
(a) PIT under linear forecast



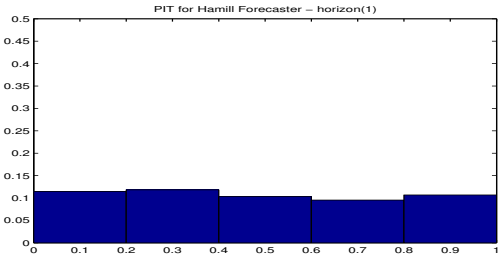
(b) PIT under GSTAR forecast



(c) PIT under unfocused forecast



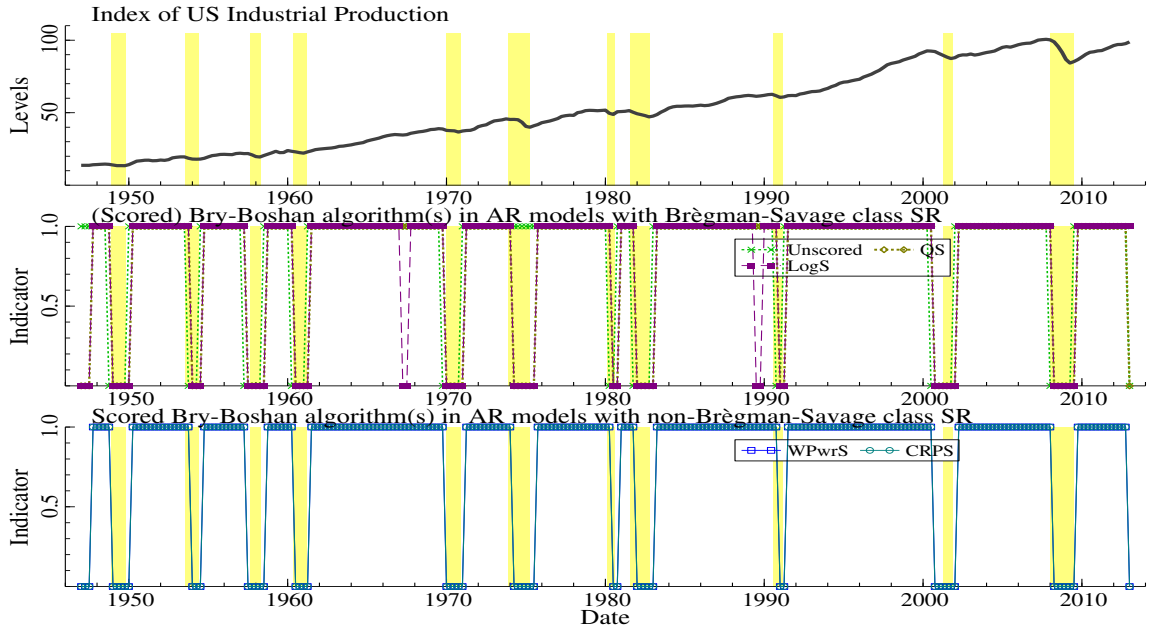
(d) PIT under Hamill's forecast



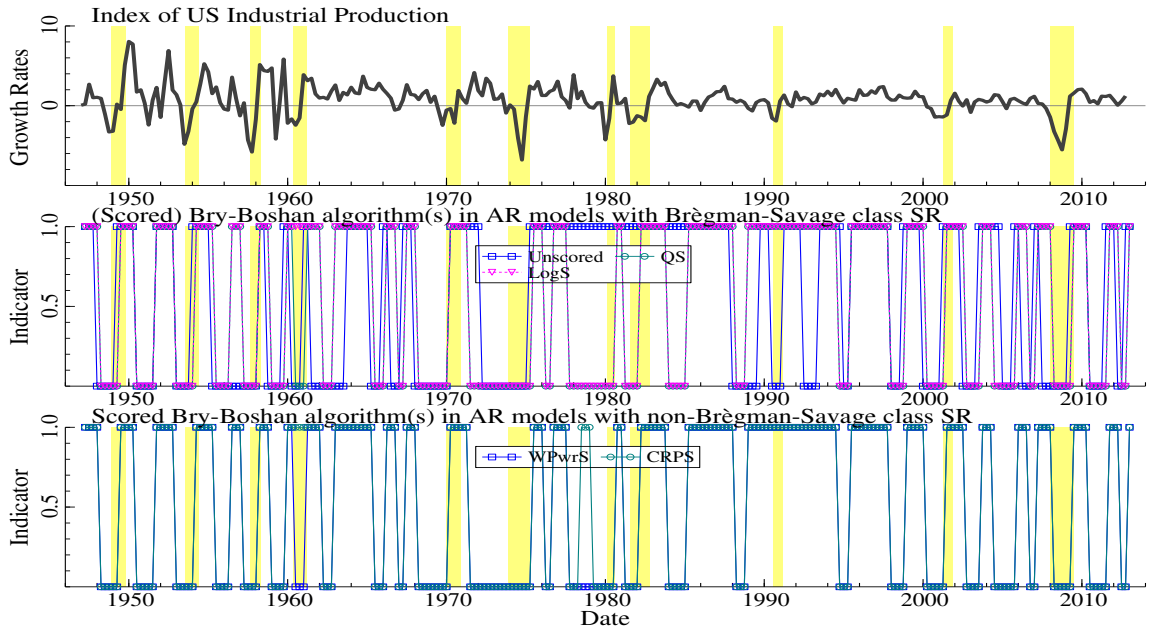
NOTE: This figure displays the output of the motivating example described in Section 2. Panel plots the PIT that corresponds to a forecast obtained by a linear AR(5) model; panel (b) plots the PIT that corresponds to a forecast obtained by Zanetti Chini's 2018 GSTAR model of the same order; panels (c) and (d) plot the two Hamill (2001) specifications: the Unfocused and Hamill Forecasters, respectively. The bars represent the histogram that corresponds to each Forecaster's quotation. If the Forecaster has the same information that Reality has (that is, under perfect forecast), the histogram would be perfectly rectangular.

Figure 2: The effects of different SR in BBQ algorithm under linear (local) Scoring Structure

(a) Analysis of data in (log-)levels



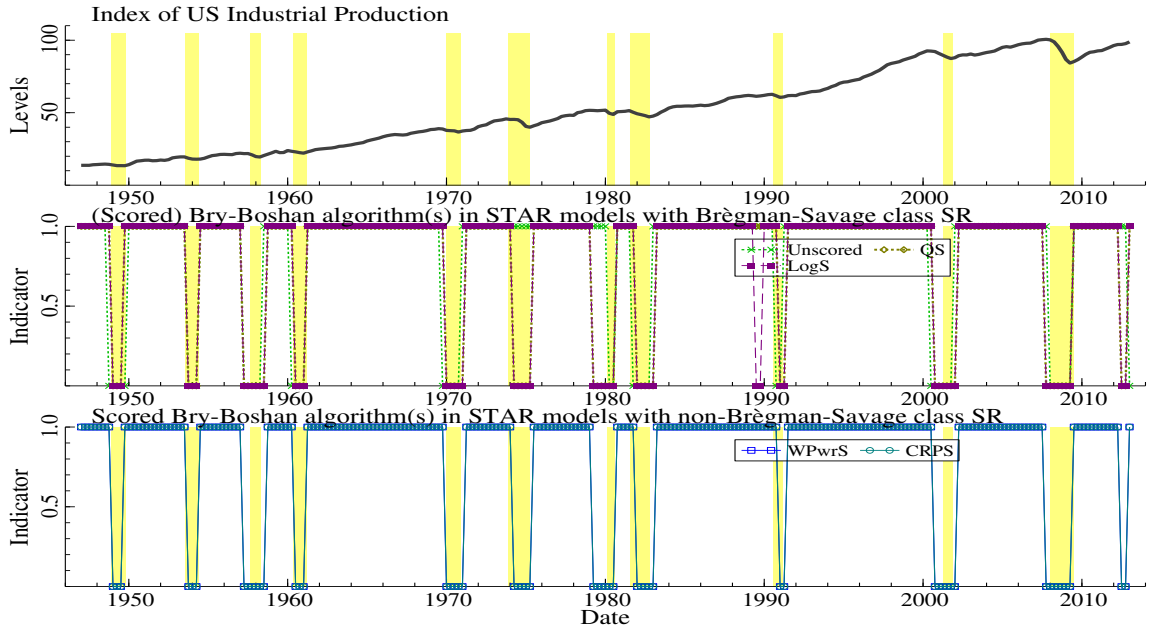
(b) Analysis of data in quarterly growth rates



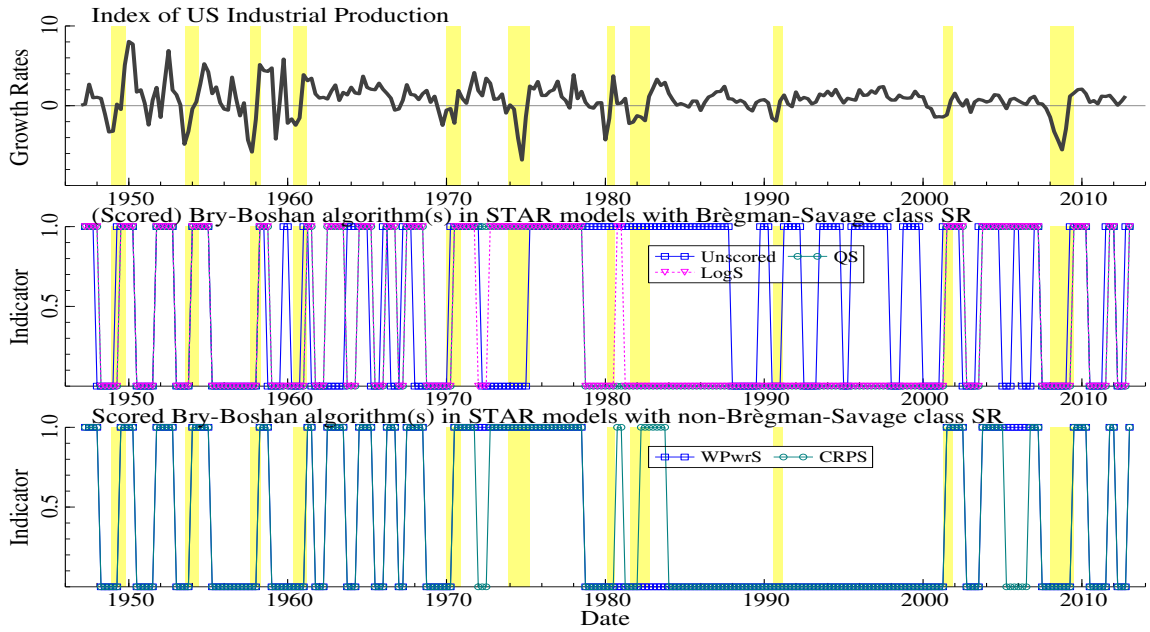
NOTE: This figure shows the output of the (S)BB algorithm introduced in Section 6.1 for data in (log-)levels (panel (a)) and growth rates (panel (b)) when the SS is assumed to satisfy the "Key Equation" or, equivalently, $\gamma = 0$ in equation (9). In both panels, the upper part of the panel shows the series in levels, while the output of the algorithmic procedures in the remaining ones is obtained using a logarithmic transform. According to our theoretical framework, the data are Reality, and the Indicators represent five Forecasters (i.e., the indicator labeled "Unscored" is the original Bry-Boshan algorithm, which simulates an unbiased evaluation, and the other four indicators are the SBBs whose evaluations of recession probability are via quadratic (QS), logarithmic (LogS), weighted power score (WPwrS), and continuous-ranking probability score (CRPS). The official NBER recession dates (the bands in yellow) represent the Forecast User's final evaluation. The Forecasters are grouped based on whether they meet the criterion of being Brègman divergence-minimizers (central sub-panel of both panels (a) and (b)) or not (lower sub-panel of the same). Software used for computation: MATLAB R2009b; image editing: OxMetrics.

Figure 3: The effects of different SR in BBQ algorithm under nonlinear Scoring Structure

(a) Analysis of data in (log-)levels



(b) Analysis of data in quarterly growth rates



NOTE: This figure plots the output of the (S)BB algorithm introduced in Section 6.1 for data in (log-)levels (panel (a)) and growth rates (panel (b)) when the SS is assumed not to satisfy the "Key Equation", or, equivalently, $\gamma \neq 0$ in equation (9). In both of the panels, the upper sub-panel shows the series in levels, while the output of the algorithmic procedures in the remaining ones is obtained using a logarithmic transform. According to our theoretical framework, the data are Reality, and the Indicators represent five Forecasters (i.e., the indicator labeled "Unscored" is the original Bry-Boshan algorithm, which simulates an unbiased evaluation, and the other four indicators are the SBBs that evaluate recession probability via quadratic (QS), logarithmic (LogS), weighted power score (WPwrS), and continuous-ranking probability score (CRPS). The official NBER recession dates (the bands in yellow) represent the Forecast User's final evaluation. The Forecasters are grouped according to whether they meet the criterion of being Brègman divergence-minimizers (central sub-panel of both panels (a) and (b)) or not (lower sub-panel of the same). Software used for computation: MATLAB R2009b; image editing: OxMetrics.

Table 1: Empirical Size and Power of LM test for Locality 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 5, where in equation (9) 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 (12) that we adopted to test the hypothesis system (10). In this experiment, the first 100 simulations were discarded to avoid the initialization effect. Software used: MATLAB R2009b.

Table 2: Empirical Power of LM-locality test 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 5 , where in equation (9) 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 (12) that we adopted to test the hypothesis system (10). In this experiment, the first 100 simulations were discarded to avoid the initialization effect. Software used: MATLAB R2009b.

Table 3: Dissecting Business Cycle: effects of adopting the SBBQ algorithm with SR elicited by linear model

S(Q, y)	$S(\cdot,\cdot)^{AR}$	Duration		Amplitude		Cumulated Value		Excess Mov. as % of Triangular Area		Cum.Val. of Duration		Cum.Val. of amplitde		Cum.Value of Excess	
		C	E	C	E	C	E	C	E	C	E	C	E	C	E
No SR	**0.0000**	**4.0769**	**15.0000**	**-0.0402**	**0.1806**	**-0.1879**	**2.2817**	**79.7578**	**60.0237**	**0.4063**	**0.6700**	**-1.2077**	**0.7509**	**2.1099**	**3.3342**
QSR	0.8127	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
WPowerS	-2.6311	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha=-1$)	32.9164	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
" ($\alpha=0$)	34.3658	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
" ($\alpha=1/2$)	-63.1365	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha=1$)	-2.2902	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha=2$)	-17.6323	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PseudoSph	1.7631	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
" ($\alpha=-1$)	1.9145	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
" ($\alpha=0$)	0.5000	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
" ($\alpha=1/2$)	1.0000	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
" ($\alpha=1$)	-19.2982	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha=2$)	-0.9825	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LogS	**0.0021**	**4.0769**	**15.0000**	**-0.0402**	**0.1806**	**-0.1879**	**2.2817**	**79.7578**	**60.0237**	**0.4063**	**0.6700**	**-1.2077**	**0.7509**	**2.1099**	**3.3342**
IntS	3.5000	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
TsallisS	1.2455	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
ES	-1.1195	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GES	-0.8426	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PseudoSpectrumS	-16.1862	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRPS	-8.1440	-	-	-	-	-	-	-	-	-	-	-	-	-	-
QuantS	1.2735	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
CLS	2.6982	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059
CsLS	0.0066	4.2500	16.1538	-0.0433	0.1908	-0.2028	2.4879	68.9083	11.1129	0.3771	0.6277	-1.1379	0.6588	2.4808	7.0334
HS	0.0381	3.6667	56.0000	-0.0047	0.5954	-0.0235	25.2830	190.4135	19.7733	0.1575	0.6819	-1.6418	0.7247	1.1166	0.0357
LCS	0.1529	4.4545	17.6667	-0.0475	0.2070	-0.2212	3.0766	93.1130	10.2991	0.3384	0.6446	-1.0399	0.7133	1.6780	7.9059

NOTE: This table reports the descriptive statistics that result from the application of the SBBQ algorithm introduced in Section 6.1 on IP in log-levels. The SRs adopted here correspond to an equivalent number of Forecasters whose behavior is evaluated by the Skeptic under linear (local) SSÐthat is, equation (9) with $\gamma = 0$ denoting a perfectly coherent forecasting scenario. According to Teräsvirta et al.'s 2010 General-to-Specific modelling strategy, Forecasters select an AR(5) model to estimate the one-step-ahead predictive density of IP. The first column indicates the type of scoring function; the second column indicates its estimate given the AR(5) specification; and the remaining columns indicate Harding and Harding and Pagan's 2002 "dissection" statistics introduced for each of the two phases (i.e., 'C' indicates contractions and 'E' indicates expansions. See the Supplement for details. 'No SR' corresponding to $S(\cdot,\cdot)^{AR} = 0.0000$ indicates that the Forecaster's information set coincides with Reality (so no evaluation is needed). The LogS (in bold) is the nearest to Reality, and the corresponding statistics coincide with the ones obtained by the standard (non-scored) BBQ algorithm. '-' indicates that the algorithm aborts the computation because of the SR's negative value. Data source: Federal Reserve Bank of St. Louis, Research Division. Software used: MATLAB R2009b.

Table 4: Dissecting Business Cycle: effects of adopting the SBBQ algorithm with SR elicited by non-linear model

S(Q, y)	$S(\cdot,\cdot)^{STAR}$	Duration		Amplitude		Cumulated Value		Excess Mov. as % of Triangular Area		Cum.Val. of Duration		Cum.Val. of amplitde		Cum.Value of Excess	
		C	E	C	E	C	E	C	E	C	E	C	E	C	E
No SR	**0.0000**	**4.2308**	**16.7500**	**-0.0604**	**0.2250**	**-0.1824**	**2.9900**	**31.5354**	**81.6785**	**0.4223**	**0.6531**	**-0.8133**	**0.7162**	**2.2606**	**2.3361**
QSR	0.8127	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
WPowerS	1.4812	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
" ($\alpha = -1$)	54.7040	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
" ($\alpha = 0$)	-106.5745	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha = 1/2$)	-0.2337	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha = 1$)	-0.8191	9.2727	14.3000	-1.7576	1.5429	-7.4182	37.0634	-181.0128	131.2204	1.0861	1.0388	-1.2176	1.1698	-4.1934	1.6266
" ($\alpha = 2$)	-27.4562	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PseudoSph	1.7631	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
" ($\alpha = -1$)	1.9689	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
" ($\alpha = 0$)	0.4999	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
" ($\alpha = 1/2$)	1.0000	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
" ($\alpha = 1$)	-94.9719	-	-	-	-	-	-	-	-	-	-	-	-	-	-
" ($\alpha = 2$)	-0.8928	-	-	-	-	-	-	-	-	-	-	-	-	-	-
LogS	**0.0025**	**4.2308**	**16.7500**	**-0.0604**	**0.2250**	**-0.1824**	**2.9900**	**31.5354**	**81.6785**	**0.4223**	**0.6531**	**-0.8133**	**0.7162**	**2.2606**	**2.3361**
IntS	3.5000	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
TsallisS	1.2455	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
ES	-1.1383	-	-	-	-	-	-	-	-	-	-	-	-	-	-
GES	-16.7098	-	-	-	-	-	-	-	-	-	-	-	-	-	-
PseudoSpectrumS	-8.1440	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CRPS	0.9012	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
QuantS	**0.0076**	**4.2308**	**16.7500**	**-0.0604**	**0.2250**	**-0.1824**	**2.9900**	**31.5354**	**81.6785**	**0.4223**	**0.6531**	**-0.8133**	**0.7162**	**2.2606**	**2.3361**
CLS	0.7807	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529
CsLS	**0.0072**	**4.2308**	**16.7500**	**-0.0604**	**0.2250**	**-0.1824**	**2.9900**	**31.5354**	**81.6785**	**0.4223**	**0.6531**	**-0.8133**	**0.7162**	**2.2606**	**2.3361**
HS	1.7303	5.9524	6.7000	-2.7365	2.8518	-11.3085	13.7579	56.5168	77.8627	0.6297	0.8032	-0.7702	1.1889	3.2995	2.0767
LCS	0.1529	4.4167	18.4545	-0.0652	0.2452	-0.1968	3.4011	16.6674	29.2511	0.3917	0.6075	-0.7358	0.6208	2.9447	1.4529

NOTE: This table reports the descriptive statistics that result from the application of the SBBQ algorithm introduced in Section 6.1 on IP in log-levels. The SRs adopted here correspond to an equivalent number of Forecasters whose behavior has to be evaluated by the Skeptic under nonlinear (non-local) SS – that is, equation (9) with $\gamma \neq 0$, denoting a non-coherent forecasting scenario. According to Teräsvirta et al.'s 2010 General-to-Specific modelling strategy, Forecasters select a STAR(5) model to estimate the one-step-ahead predictive density of IP; in particular, the STAR specification has $d = 4$ and three regimes corresponding to two $G(\cdot)$ with associated nonlinear and autoregressive parameters. These estimates are not reported for space reasons. The first column indicates the type of scoring function; the second column indicates its estimate given the STAR(5) specification; and the remaining columns indicate Harding and Harding and Pagan's 2002 "dissection" statistics for each of the two phases (I.e., 'C' indicates contractions and 'E' indicates expansions). 'No SR' corresponding to $S(\cdot,\cdot)^{STAR} = 0.0000$ indicates that the Forecaster's information set coincides with Reality (so there is no need for evaluation). LogS is the minimum score between the set of positive estimates, but its statistics, which still correspond to those obtained by standard BBQ algorithm, coincide with those evaluated according to either the quantum (QuantS) or censored likelihood scores (CsLS, in bold). '-' indicates that the algorithm aborts the computation because of the SR's negative value. Data Source: Federal Reserve Bank of St. Louis, Research Division. Software used: MATLAB R2009b.

Table 5: LM locality test for real data

IP					
LM_1		LM_2		LM_3	
F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
220.5872	<0.01	220.0645	<0.01	623.50	<0.01
OG					
LM_1		LM_2		LM_3	
F-statistic	P-value	F-statistic	P-value	F-statistic	P-value
99.3900	<0.01	97.5790	<0.01	251.32	<0.01

NOTE: This table reports the test statistics (12) in their F-variant, with corresponding p-values for data on the U.S. Industrial Production Index and the Norges Bank Output Growth is estimated according to the fan charts published in the release of January 2014. Data Source: Federal Reserve Bank of St. Louis, Research Division (IP) and Bank of Norway's web page (OG).

Table 6: Relative predictive ability of Forecaster's quotations for IP.

S(Q, y)	$\bar{S}^f$	$\bar{S}^g$	σ	t	P-value
Brègman-Savage type					
LogS	1.0073	0.0269	0.0040	**19.7831**	<0.01
QuantS	0.5530	0.4082	0.0887	**1.6594**	0.0485
CsLS	0.0961	0.0340	0.0345	**1.8841**	0.0297
Non Brègman-Savage type					
WPowerS	0.8147	0.9534	0.6110	-0.1575	0.5625
WPseudoSph	1.0849	1.0857	0.0251	-0.0329	0.5131
CRPS	0.1320	0.1431	0.0273	-0.0216	0.5086

NOTE: This table reports the result of the Amisano and Giacomini's 2007 test for the U.S. index of industrial production at a prediction horizon of 12 months for different SRs corresponding to the predictive density functions f and g, respectively, where f is the Forecaster's nonlinear specification (corresponding to the same STAR model used in Table 4) given a non-local SS, and g is a linear specification given a non-local SS. The average scores $\bar{S}_n^f = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{f}_{t+k}, y_{t+k})$ and $\bar{S}_n^g = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{g}_{t+k}, y_{t+k})$ was computed by aggregating the sequences of forecasts generated by the pseudo-out-of-sample forecasting experiment described via the Monte Carlo method in Zanetti Chini (2018, pp. 718). The null hypothesis that $\Delta^* = \bar{S}_n^f - \bar{S}_n^g = 0$ is measured by statistic $t_n = \sqrt{n}\frac{\Delta^*}{\hat{\sigma}_n} \sim N(0,1)$, 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 in positive orientation is adopted, so f is preferable to g if and only if $S^f > S^g$. Statistics that lead to rejection of the null hypothesis are reported in bold.

Table 7: Relative predictive ability of Forecaster's quotations for OG

S(Q, y)	$\bar{S}^f$	$\bar{S}^g$	σ	t	P-value
LogS	0.0273	0.0275	0.0070	**-2.5601**	0.9949
QSR	0.5582	0.5390	0.0961	-0.3251	0.3820
WPowerS	2.8147	2.9534	0.1011	-3.3575	0.9991
" ($\alpha = -1$)	527.6771	2.9534	1.08e05	-0.0032	0.5012
" ($\alpha = 0$)	527.5193	670.8714	1.08e05	-0.0032	0.5012
" ($\alpha = 1/2$)	-1,062.301	-1,352.7636	4.43e05	0.0016	0.4993
" ($\alpha = 1$)	1.1986	1.4481	0.3275	-1.8662	0.9653
" ($\alpha = 2$)	-262.6019	-333.8081	2.66e03	0.0065	0.4974
PseudoSph	2.9817	2.9817	0.0000	0.0000	0.5000
WPseudoSph	1.9916	1.9931	1.2709e-05	-299.5681	1.0000
" ($\alpha = -1$)	0.4999	0.5000	7.8768e-12	-3.8e05	1.0000
" ($\alpha = 0$)	1.0000	1.0000	0.0000	0.0000	0.5000
" ($\alpha = 1/2$)	-1,927.5277	1.0000	3.8e06	0.0005	0.4997
" ($\alpha = 1$)	1.1986	1.4481	0.3275	-1.8662	0.9653
" ($\alpha = 2$)	-0.8559	-0.8361	0.0020	-23.4613	1.0000
IntS	3.5000	3.5000	0.0000	0.0000	0.5000
TsallisS	1.2229	1.2229	0.0000	0.0000	0.5000
ES	-0.1237	-0.0834	0.0085	-11.5709	1.0000
GES	1.1626	1.2485	0.0388	-5.4196	1.0000
PseudoSpectrumS	-7.8530	-7.8530	0.0000	0.0000	0.5000
CRPS	0.0132	0.0120	7.2746e-06	**395.9622**	$<$0.01
QuantS	-0.1909	-0.1835	0.0002	-63.5174	1.0000
CLS	-0.1467	-0.4232	0.4021	1.6841	0.0499
CsLS	0.0088	0.0076	8.0784e-06	**375.7488**	$<$0.01
LCS	0.0552	0.0569	0.0105	-0.2100	0.5831

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 announcements and g is a non-linear specification (corresponding to the same STAR(5) of Table 3) given a non-local SS. The average scores $\bar{S}_n^f = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{f}_{t+k}, y_{t+k})$ and $\bar{S}_n^g = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{g}_{t+k}, y_{t+k})$ was computed by aggregating the sequences of forecasts generated by the pseudo-out-of-sample forecasting experiment via Monte Carlo procedure described in Zanetti Chini (2018, pp. 718). The null hypothesis that $\Delta^* = \bar{S}_n^f - \bar{S}_n^g = 0$ is measured by statistic $t_n = \sqrt{n}\frac{\Delta^*}{\hat{\sigma}_n} \sim N(0,1)$, 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 $S^f < S^g$. Counterintuitive results are marked in bold. Data source: Bank of Norway's web page, release of 2014.

A Appendix

A.1 Assumptions

A 1. *$\mathcal{P}$ is assumed such that EU exists for all $a \in \mathcal{A}$, $P \in \mathcal{P}$.*

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

A 3. *$U(P, a_Q)$ is strictly convex in* a.

A 4. *The entropy $H(P)$ associated to $S(\cdot)$ is (strictly) convex in P, integrable with respect to $P \in \mathcal{P}$ and quasi-integrable with respect to all $Q \in \mathcal{P}$ and such that H^* is a sub-tangent of H at point P.*

A 5. *$S(P, Q)$ is affine, real-valued for all $P, Q \in \mathcal{P}$ and minimized in Q at $Q = P$.*

A 6. *$D(P, Q) - D(P, Q_0)$ is affine in P, and $D(P, Q) \geq 0$, with equality achieved at $Q = P$*

A 7. *$\hat{P}_{t+k}$ and $\hat{Q}_{t+k}$ are measurable functions of the data in a rolling estimation window.*

A.2 Proof of Theorem 1

Let $\mathbb{L}$ and $\mathbb{D}$ be the same operators defined in Section 3, $\Lambda = \sum_{k\geq 0}(-1)^k D^k \partial/\partial q_k$ 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 q-function, then $\mathcal{L}(\Theta) \equiv \mathcal{L}(\Psi)$. The Key Condition (i) is a consequence of the fact that, in $P = Q$, $S(\cdot)$ is a stationary point under an infinitesimal variation $\delta q(\cdot)$ of $q(\cdot)$ (if assuming that $q(\cdot) + \delta q(\cdot)$ 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 (8) 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 $log(p(\cdot)$ – that is, the same LogS; see Ehm and Gneiting (2012). The operators $\mathbb{D}$ and $\mathbb{L}$ here adopted are linearly invariant by Theorem 11.3 and 11.4 of Ehm and Gneiting (2012). Hence the statement.

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