DEPARTMENT OF ECONOMICS AND BUSINESS ECONOMICS
AARHUS UNIVERSITY

CREATES
Center for Research in Econometric Analysis of Time Series

Forecaster’s utility and forecasts coherence

Emilio Zanetti Chini

CREATES Research Paper 2018-1

Department of Economics and Business Economics
Aarhus University
Fuglesangs Allé 4
DK-8210 Aarhus V
Denmark

Email: oekonomi@au.dk
Tel: +45 8716 5515

Forecaster's utility and forecasts coherence

Emilio Zanetti Chini*

December, 2017

Abstract

I provide general frequentist framework to elicit the forecaster's expected utility based on a Lagrange Multiplier-type test for the null of locality of the scoring rules associated to the probabilistic forecast. These are assumed to be observed transition variables in a nonlinear autoregressive model to ease the statistical inference. A simulation study reveals that the test behaves consistently with the requirements of the theoretical literature. The locality of the scoring rule is fundamental to set dating algorithms to measure and forecast probability of recession in US business cycle. An investigation of Bank of Norway's forecasts on output growth leads us to conclude that forecasts are often suboptimal with respect to some simplistic benchmark if forecaster's reward is not properly evaluated.

Keywords: Business Cycle, Evaluation, Locality Testing, Nonlinear Time Series, Predictive Density, Scoring Rules, Scoring Structures.
JEL: C12, C22, C44, C53.

*University of Pavia, Department of Economics and Management, Via San Felice 5, 27100, Pavia, Italy - *E-mail*: emilio.zanettichini@unipv.it

1 Introduction

Should the final user of the forecast of an uncertain event trust in professional forecaster's quotation? And is this quotation coherent with the prediction of an economic/statistical model or is biased by personal judgments? Being able to answer to these questions means, for any forecast user, to satisfy two strong conditions: (i) to have a full control of the forecasting process and (ii) to deal with forecaster's subjective evaluation of the uncertainty – and, as a logical consequence, with the laws of probability. These *desiderata*, despite their apparent triviality, are the fundaments of modern decision-theoretical approach to economic forecasting settled by De Finetti (1937, 2017) almost a century ago; see also Elliott and Timmermann (2008). According to this framework, the forecaster maximizes its own expected utility (or, symmetrically, minimizes the loss) when (in)correctly quoting its forecast before that event realizes, inducing a systematic bias[1]. Hence, estimating the forecaster's utility (or reward) function implies necessarily to check the coherence of his forecast.

In practice, the evaluation of the probability is done by computing the goodness-of-fit of the estimated density function (or predictive density) via the Rosenblatt (1952) probability integral transforms (PITs), which has been object of study in Econometrics and more recently has evolved in several directions[2]. Unfortunately, as the next Section 2 extensively explains, PITs do not conveys any information about the forecaster's utility underlining the quotation, and thus they might fail in recognizing a forecast drawn by a biased forecaster. Thus, some alternative approach is necessary to the pursue of this aim. One of the most successful strategy, originating from Bates and Granger (1969), is comparing and combining the loss functions estimated by alternative model(s) *via* some information criterion or encompassing testing. This literature[3], however, implicitly assumes that the loss function is known by forecast user when assessing the forecaster's quotation, whilst this is exactly the main source of the uncertainty in this game.

Scoring rules (SR) are the most immediate way to link the forecaster's reward and the mechanics of predictive density estimation. These are functions $\mathrm{R} \rightarrow \mathrm{R}$ assigning a numerical score to several competing (model) density forecasts. The score corresponds to the forecaster's utility for his correct forecasts of the event. A properly set SR incentives the forecaster to be honest[4]. Gneiting et al. (2007) (GBR, henceforth) demonstrate that the mentioned deficiency of PITs can be overlooked if appropriate scoring functions are used, where the world appropriate means to be capable to nest the Savage (1971) representation, which in turn is based on the based on the Brègman (1967) distance. Despite their eldery origins, SR have been continuously object of interest in Statistics, see Gneiting and Raftery (2007) (GR, henceforth). We assert that this literature under-evaluates the theoretical properties of SRs to be used in hypothesis testing for model selection. Patton (2016) enforces the importance of this issue by demonstrating that the forecast rankings are generally sensitive to the choice of a proper SR and concluding that forecasters should be told *ex ante* what utility functions will be used to evaluate their quotation; see also Laurent et al. (2013) for a multivariate equivalent.

The application of these theoretical results in a basic regression framework is complicated by two – statistical and operational – problems: *prima facie*, as documented in Table 1, the number of proper SRs (although not all strictly proper) explicitly built-up for density forecasts is high and the ones capable to nest a Brégman-Savage representation, despite lower, is still considerable. *In secundis*, many of them are nested; this can easy confirmed by having a look on Table 1 in Jose et al. (2008). Hence, we need a criterion to select the (unknown) reward that effectively drives the quoted forecast. Fortunately, the Bernardo (1979) theorem, according to which a SR is proper and local if and only if it is logarithmic, gives us a fundamental hint: in facts, it postulates that the locality of the SR coincides with the likelihood principle. Hence, our research question: "*Is forecaster's quotation compatible with local SR?*" Section 3 provides an answer by introducing a new framework by relying on the theory by Parry et al. (2012); Dawid et al. (2012); Ehm and Gneiting (2012), whose mathematical results have no applied counterfactuals. Namely, we assume the SR as an endogenous transition variable in a general family of nonlinear models. This make us able to rely on a well-known asymptotic theory. Such a combination of a (nonlinear) model and SRs is defined *scoring structure*. It is based on the Patton and Timmermann (2007) assumption that the forecaster utility is unknown but observable and is directly connected to the literature on equal predictive ability testing[5]. Subsequently, Section 4 define a test for the hypotheses of locality of the estimated predictive density generated by the scoring structure and Section 5 investigates the small-sample properties, as well as their relevance with respect to the (more or less implicit) requirements of theoretical literature.

Section 6 illustrates two different case studies. Namely, in the first application we introduce a modification of a standard algorithm for detection of the probability of recessions of US economy; in the second illustration, we assess the density forecasts of Norway's output gap from Bank of Norway (BoN). Our results reveal that improper scoring rules affect the dating algorithm of recessions events and the model-based forecast performances in favor of a nonlinear specification. Secondly, an equal predicitve ability test between our nonlinear scoring structure and a benchmark therein nested indicates that BoN's fan charts are severally biased when a strongly nonlinear model is assumed and, more importantly, when the SRs do not have Brègman-Savage repesentation.

Finally, Section 7 summarizes and concludes; two Appendices describe the details on the SBB algorithm and the equal predictive ability test used in our illustrations.

2 Motivating Example: the Hamill Paradox

Having a correct view of the (short term) evolution of the main macroeconomic variables is mandatory for any central bank in order to address its policy. To this aim, these institutions use to ask to a pool of professional forecasters to quote the expected probability distribution of the variable(s) under consideration. Then, these are collected and summarized via *fan charts*. Anyway, the evaluation of these forecasts is preliminary to make decisions concerning the economic policy. The following experiment shows how easy one can fail this step.

We simulate the dynamics of US Industrial Production by using the following four DGPs:

$$y_{1,t}^{(s)} = 0.9y_{1,t-1}^{(s)} - 0.795y_{1,t-2}^{(s)} + \epsilon_{1,t}^{(s)}, \qquad \epsilon_{1,t}^{(s)} \sim N(0,1);$$

$$y_{2,t}^{(s)} = 0.9y_{2,t-1}^{(s)} - 0.795y_{2,t-2}^{(s)} + (0.02 - 0.4y_{2,t-1}^{(s)} + 0.25y_{2,t-2})^{(s)}G^{(s)}(\Xi) + \epsilon_{2,t}^{(s)}, \quad \epsilon_{2,t}^{(s)} \sim N(0,1);$$

$$y_{3,t}^{(s)} = \epsilon_t^{Unfocused,(s)}, \qquad \epsilon_t^{Unfocused} \sim 0.5 \cdot [N(\mu_t, 1) + N(\mu_t + \tau_t, 1)];$$

$$y_{4,t}^{(s)} = \epsilon_t^{Hamill,(s)}, \qquad \epsilon_t^{Hamill} \sim N(\mu_t + \delta_t, \delta_t^2);$$

where: $G^{(s)}(\Xi) = \big(1 + \exp\{-\big[h(\eta_t)^{(s)} I_{(\eta_t \leq 0)}(y_{t-1}^{(s)} - \bar{y}_t^{(s)}) + h(\eta_t)^{(s)} I_{(\eta_t > 0)}(y_{t-1}^{(s)} - \bar{y}_t^{(s)})\big]\}\big)^{-1}$, 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 $\eta_t < 0$ is a function of two parameters governing the transition between the two extreme states $G = 0$ and $G = 1$; namely $h(\cdot)$ is function of $\gamma_1 = 50$, $\gamma_2 = -20$,$c = ave(y_t)$, $\eta_t = (y_{t-1}^{(s)} - \bar{y}_t^{(s)})$, $\bar{y}_t = \frac{1}{T}\sum_{t=1}^{T} y_t$ and y_{t-1} being the transition variable, $s = \{1, \ldots, S\}$, $T = 265$ and $S = 1,000$; 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 and Model 2 a GSTAR model of the same order[6]; Models 3 and 4 are the Unfocused and Hamill's forecasters, see Hamill (2001).

We then compute the PITs corresponding to these forecasts and plot them in Figure 1. Under perfect forecast, the histogram is perfectly rectangular. But in our experiment, all four histograms are almost rectangular, so that none of the alternative Forecasters is distinguishable. Such a finding – called *Hamill's Paradox* – complicates the decision-making because every model corresponds to a different policy. Is the Industrial Production following a linear (thus, cyclical phases are equally possible) or nonlinear asymmetric low of motion (and thus future downturns have low probability to happen)? Or, on the opposite side, is the professional forecasters quotation the output of a poor model (the Unfocused), so that no policy should be implemented? Moreover, we remark that density forecasts which has been generated by nonlinear models, are often multimodal, as shown in Figure 2. This is often a source of confusion with for a process driven by a mixture of distributions (*ergo*, of utilities), like the Hamill forecaster. Is it symptomatic of a change in forecaster's utility rather than a more complex scenario – like cooperative strategy? No assessment can be done according to our results.

Mitchell and Wallis (2011) severely criticize this new approach in what the data generating process used by Hamill and GBR is not robust to some basic time-series feature. This implies that traditional diagnostic tools can still be successfully applied. Their critique, however, is overcome by our example: in fact, it proves that Hamill's paradox is still an involving issue when a standard and sufficiently general time series model specification is set.

3 Theoretical framework

We consider a time series $Y_t = \{y_t\}_{t=1}^{T} = \{y_1, \ldots, y_t, \ldots, y_T\}^\intercal$, with $\intercal$ denoting the transposition, which is fully represented by an information set $\mathcal{F}_t = \{y_{t-1}, y_{t-2}, \ldots\}$

and a probability density $P(Y_t)$. Let denote the (1-step-ahead) probability forecast of Y_t as $P(Y_{t+1})$, the best forecaster's judgement of the distributional forecast of Y_t as $Q(Y_t)$ (we will omit Y_t for notational convenience) and x a draw of Q which materializes in $T+1$ and the predictive cumulate disribution function associated to the materialization of x as $F(x)$. 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 $t = m, \dots, m+n-k$ 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 $Y_{t?m+1}, \dots, Y_t$. Then, in order to nest the theoretical framework by Vovk and Shafer (2005) and Dawid (2007), we introduce two agents: Forecaster and Nature (or Reality); the first one is assumed to have an information set at most equal to the Nature one.

Let be: $\mathcal{X}$ being a set of the possible forecaster's outcomes, $\mathcal{P}$ the family of distributions on $\mathcal{X}$ in which P is belonged and $\mathcal{A}$ a σ-algebra of subset of $\mathcal{X}$ representing the set of actions. In particular, if the sample space is discrete (that is, a dichotomous for events like the probability of a recession, or categorical like, say, the ranking position of a firm or State), P is defined by $\mathcal{P} = \{\mathbf{p} \in \mathcal{A} : \sum_x p_x = 1\}$ is the set of all real vectors corresponding to strictly positive probability measures; if it is continuous (like the conditional mean of an economic time series), P is defined by $\mathcal{M}$, the set of all distributions on $\mathcal{X}$ which are absolutely continuous with respect to a $\sigma-$finite measure μ. The same is for $\mathbf{q}$. Forecaster aims to solve a decision problem defined by the triple $\{\mathcal{X}, \mathcal{A}, \mathcal{U}(P, \mathrm{a})\}$, where: $\mathcal{X}$ is previously defined; $\mathcal{A}$ is the action space; and $\mathcal{U}(P, \mathrm{a}^*)$ is a real-valued utility function which represent the reward obtained by Forecaster as effect of minimizing the discrepancy on his own quotations, of the action $a^* \in \mathcal{A}$, which maximizes the expected utility computed using the density P believed the true DGP, with the expected loss denoted as $EU := \int U(P, \mathrm{a}) P(Y_t) dY_t$. Let 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 to any $U(P, \cdot)$. The resulting system is defined as follows:

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

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

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

iii. *Divergence function* the function $D(P, Q) := H(P) - S(P, Q)$

iv. *Scoring structure* the 5-ple $\{Y_t, \mathcal{F}_t, S(\cdot, \cdot), H(P), D(\cdot, \cdot)\}$

which, under proper conditions, is able to characterize coherent forecasts.

Remark 1. The Scoring Structure is sufficiently general, in his definition, to take in consideration both the point of view of Reality, Forecaster and Forecast User. The interactions between these three agents depends on the assumptions on each of them. These assumptions have practical relevance because, as we will see in the course of the paper, they delimit the framework of econometric assessment method to be used.

Since SRs are the main block of all this system, we formally define them as follows:

Definition 2 (m-Local, (strictly) proper scoring rule)**.** i. A scoring rule $S : \mathcal{X} \times \mathcal{P} \to \mathbb{R}$ is local of order m or m-local if it can be expressed in form of:

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

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

ii. A (local) scoring rule $S(x, Q)$ 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) $Q = P$

To make Definition 1 and Definition 2 operational, we invoke the following 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 $S()$ 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.*

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

Proof. This is essentially the Theorem 1 in GR. □

Remark 2. A1 – A3 are necessary (but not sufficient) to define the Forecaster's reward as scoring rule. In particular, A1 is encloses the three "basic assumptions" discussed in Dawid (2007)[7] and implies 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 assumption which are necessary to having a unique maximizing action. A4 characterizes the general representation of scoring rules; see Thm 1 in GR. A5 stresses the fact that Forecaster has no loss only if his DGP coincides with the Nature's one;

see Thm 1 in Bernardo (1979) and GBR. A6 is fundamental to characterize a very 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 $\mu \in \mathcal{X}$, the *Brègman score*:

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

with associated *Brègman divergence*:

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

where ψ is a (strictly) concave function. This is a very general class of non-metric distance able to characterize most of the scoring rules described in Table 1. We are particularly interested in the special case that

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

where k, also known in Physics as *Boltzmann's constant*, is commonly is set to zero without loss of generality for ease of treatment. Under (7) the forecasts generated by model $\mathcal{M}$ are coherent with a scoring structure $\{Y_t, \mathcal{F}_t, S(\cdot,\cdot), H(\cdot), D(\cdot,\cdot)\ \}$ defined on a logarithmic scoring rule, Shannon Entropy and Kullback-Liebler distance.

Remark 3. All the functions $S(\cdot,\cdot)$ in Definition 2 can be interpreted in terms of utility: $S(x, P)$ is the Forecaster's reward for the fact that the event x (truly) materializes. This is a function defined on the extended real line, that is $S(x, P) \in \overline{\mathbb{R}} = [-\infty, +\infty]$. Consequently the expected forecaster's utility, *conditionally 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 Forecaster can achieve using Nature'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. Finally, A7 is necessary only to apply the Amisano and Giacomini (2007) predictive ability test on the output of the scoring structure.

The next result constitutes the motivation for the rest of the analysis:

Proposition 2. *Let $S(x,Q)$ be a scoring rule, possibly fulfilling the Brègman-Savage representation, with q-function s. Then, $S(x,Q)$ is local and strictly proper if and only if s is such that:*

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

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 *Key Condition*. For purely theoretical reasons, the same theorem requires other two conditions concerning the representation of s via Lagrange operators. Nevertheless, the Key Condition is sufficient (and, to the best of our knowledge it the only available) to identify an empirically testable hypothesis for the assessment of the logarithmic form of the Forecaster's utility.

4 The Locality Test

This section defines a statistical hypothesis test to check if the "key equation" (8) is verified by data. In order to do this, we assume that the q-function s generating $S(x,Q)$ is an observed transition variable of a Smooth Transition Autoregressions (STAR) model introduced by Chan and Tong (1986)[8]. This is necessary to set-up the null hypothesis and introduce an LM-type test *à la* Luukkonen et al. (1988) (LST): the idea is to linearize the original nonlinear parametrization via Taylor expansion on our scoring structure to arrive an auxiliary model with augmented regressors, the number of which depends on the type of non-linearity suspected; this artificial model can be investigated by standard χ^2 asymptotics. We stress that the test does not impose any form of SR, because these are treated as unknown, observed variables. The rest of the discussion is divided in two subsections: Subsection 4.1 briefly describes the model and the null hypothesis and Subsection 4.2 the test statistic.

4.1 The Null Hypothesis

The process $\{y_t\}$ observed at $t = 1-p, 1-(p-1), \dots, -1, 0, 1, \dots, T-1, T$ is assumed to be driven by the following structure:

$$y_t = (1 - G(\gamma, \mathbf{w}_t, \mathbf{c}_k))\boldsymbol{\phi}^\intercal \mathbf{z}_t + G(\gamma, \mathbf{w}_t, \mathbf{c}_k))\boldsymbol{\theta}^\intercal \mathbf{z}_t + \epsilon_t, \ \ \epsilon_t \sim iid(0, \sigma^2) \quad (9)$$

$$G(\gamma, \mathbf{w}_t, \mathbf{c}_k) = \left(1 + \exp\left\{ -\gamma \prod_{k=1}^{K} (\mathbf{w}_t - \mathbf{c}_k)) \right\}\right)^{-1}, \ \ \gamma > 0, \ \ c_1 < \dots < c_k, < \dots < c_K, \quad (10)$$

where: $\mathbf{z}_t = (1, y_{t-1}, \dots, y_{t-p})^\intercal$ are the autoregressive covariates, $\boldsymbol{\phi} = (\phi_0, \phi_1, \dots, \phi_p)'$ are the linear part parameters, $\boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_p)^\intercal$ nonlinear part parameters, γ is the slope parameter, $\mathbf{c}_k = (c_1, \dots, c_K)$ denoting the location parameters, $\mathbf{w}_t = a^\intercal \mathbf{z}_t \odot \mathrm{s}$ is a composite transition variable, with $a = [a_1, \dots, a_p]^\intercal$, $a_i = \begin{cases} 0 \text{ if } i = d \\ 1 \text{ if } i \neq d \end{cases}$ denoting the fact that delay parameter d, which is such that $1 \leq d \leq p$, is unknown and s a generic proper SR as in Definition 2.

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 s_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 some 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 Exponential STAR (ESTAR) model. When $\gamma \to \infty$, the model (9) nests the Tong (1983) two-regimes Threshold Auto-Regressive model.

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

$$
\begin{aligned}
\boldsymbol{\Lambda}_t(\mathbf{z_t}, \Xi) &= const + \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\sigma^2 \sum_t (y_t - \boldsymbol{\phi}'\mathbf{z_t} - \boldsymbol{\theta}'\mathbf{z}_t G)^2 \\
&= const + \frac{T}{2}\ln\sigma^2 - \frac{1}{2}\sigma^2 \sum_t u_t^2(\Xi),
\end{aligned}
\tag{11}
$$

with *const* and u_t denoting a constant and the model's residual, respectively, and to the score:

$$
\boldsymbol{\Sigma}_t(\mathbf{z_t}, \Xi) = \nabla_{\Xi}\boldsymbol{\Lambda}_t(\mathbf{z_t}, \Xi) = \frac{1}{\sigma^2}\sum_t u_t(\Xi)\mathbf{d}_t, \tag{12}
$$

$$
\mathbf{d}_t = \nabla_{\Xi} u_t(\Xi) = [\mathbf{z}_t, \mathbf{z}_t G, \boldsymbol{\theta}'\mathbf{z}_t G_\gamma, \boldsymbol{\theta}'\mathbf{z}_t G_c]^\intercal, \tag{13}
$$

with $G_\gamma = \partial G/\partial\gamma$ and $G_c = \partial G/\partial c$ denoting the first derivatives of G with respect to γ and c. We want to test the hypothesis that, in (9),

$$
H_{0i} : \boldsymbol{\theta} = \mathbf{0} \ \text{ vs } \ \boldsymbol{\theta} \neq \mathbf{0}. \tag{14}
$$

The null hypothesis (14) corresponds to (8) (in particular to the null hypothesis that H_0^*: $\lambda = 0$ in the q-function s) and requires a simple LM-type test to be verified. Note that the parameters γ_i, a and c are not identified under H_0. In a similar way, we could choose $H_0' : \gamma = 0$ as our locality hypothesis, in which case neither c, a nor $\boldsymbol{\theta}$ would be identified under H_0'. This implies that the conventional maximum likelihood theory is not directly applicable to deriving test procedures for testing (14). In this case, it is still possible to proceed to built-up a null hypothesis by keeping fix the unspecified parameters, as suggested by Davies (1977).

4.2 The Test Statistics

Define $\boldsymbol{\tau} = (\boldsymbol{\tau_1}, \boldsymbol{\tau_2})^{\intercal}$, where $\boldsymbol{\tau_1} = (\phi_0, \boldsymbol{\phi}^{\intercal})^{\intercal}$, $\boldsymbol{\tau_2} = \gamma$[9]. Let $\hat{\boldsymbol{\tau}}_{\mathbf{1}}$ the LS estimator of $\boldsymbol{\tau_1}$ under $H_0 : \boldsymbol{\gamma} = \mathbf{0}$, $\hat{\boldsymbol{\tau}} = (\boldsymbol{\tau}_{\mathbf{1}}', \mathbf{0}^{\intercal})^{\intercal}$. Moreover, let $\hat{\mathbf{d}}_{\mathbf{t}} = \mathbf{d}_{\mathbf{t}}(\hat{\boldsymbol{\tau}}) = (\hat{\mathbf{d}}_{\mathbf{1},\mathbf{t}}, \hat{\mathbf{d}}_{\mathbf{2},\mathbf{t}})$, where the partition conforms to that of $\boldsymbol{\tau}$. Under H_0, the test statistic is:

$$S(\Xi)^{LM} = \frac{1}{\hat{\sigma}^2}\hat{\mathbf{U}}'\hat{\mathbf{D}}_{\mathbf{2}}(\hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{D}}_{\mathbf{2}} - \hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{D}}_{\mathbf{1}}(\hat{\mathbf{D}}_{\mathbf{1}}'\hat{\mathbf{D}}_{\mathbf{1}})^{-1}\hat{\mathbf{D}}_{\mathbf{1}}'\hat{\mathbf{D}}_{\mathbf{2}})^{-1}\hat{\mathbf{D}}_{\mathbf{2}}'\hat{\mathbf{U}}, \tag{15}$$

$\mathbf{D}_i = [\mathbf{d}_{i1}, \dots, \mathbf{d}_{iT}]^{\intercal}$, $\hat{\mathbf{D}}_{\mathbf{i}} = [\hat{\mathbf{d}}_{\mathbf{i1}}, \dots, \hat{\mathbf{d}}_{\mathbf{it}}, \dots, \hat{\mathbf{d}}_{\mathbf{iT}}]^{\intercal}$, $i = \{1, 2\}$, $t = 1, \dots, T$, $\hat{\sigma}^2 = \frac{1}{T}\sum_1^T \hat{u}_t^2$ and $\hat{u}_t = y_t - \hat{\boldsymbol{\tau}}_{\mathbf{1}}^{\intercal}\mathbf{z}_{\mathbf{t}}$. When the model is an GLSTAR, $\hat{\mathbf{d}}_{\mathbf{1},\mathbf{t}} = -\mathbf{z}_{\mathbf{t}} = -(1, y_{t-1}, \dots, y_{t-p})^{\intercal}$ while $\hat{\mathbf{z}}_{\mathbf{2t}} \equiv \frac{\partial^2 u_t}{\partial\gamma\partial\gamma'}\big|_{\boldsymbol{\gamma}=\mathbf{0}} = -\frac{1}{2}\big\{\theta_{20}[y_t(y_{t-d})] - cy_t\boldsymbol{\theta}'\mathbf{z}_{\mathbf{t}} + \boldsymbol{\theta}_{\mathbf{2}}'\mathbf{z}_{\mathbf{t}}y_t y_{t-d}\big\}$ Just minor modifications are needed in notation of $\hat{\mathbf{d}}_t$ and $\mathbf{s}_t^L$ in case of LSTAR2 model due to an additional c parameter with respect to the LSTAR. The proposed test statistic depends on θ and is still unidentified unless $\theta_2 = 0$. LST prove that this problem can be circumvented by linearizing the nonlinear model via (third order) Taylor expansion. Here we adopt the same argument and argue that this linearization has a double importance because it is the only direct way to link the null hypothesis to a regression framework.

The linearized GLSTAR model

$$y_t = \phi'\mathbf{z}_{\mathbf{t}} + \theta'\mathbf{z}_{\mathbf{t}}T_3G(\,\cdot\,)\epsilon_t', \tag{16}$$

leads to the following auxiliary regression for testing linearity and symmetry:

$$\hat{\epsilon}_t' = \hat{\mathbf{z}}_{\mathbf{1t}}'\tilde{\boldsymbol{\beta}}_{\mathbf{1}} + \sum_{j=1}^{p}\beta_{2j}sy_{t-j}y_{t-d} + \sum_{j=1}^{p}\beta_{3j}sy_{t-j}y_{t-d}^2 + \sum_{j=1}^{p}\beta_{4j}sy_{t-j}y_{t-d}^3 + v_t\ , \tag{17}$$

where v_t is a $NIID(0, \sigma^2)$ process, $\tilde{\boldsymbol{\beta}}_{\mathbf{1}} = (\beta_{10}, \boldsymbol{\beta}_{\mathbf{1}}^{\intercal})^{\intercal}$, $\beta_{10} = \phi_0 - (c/4)\theta_0$, $\boldsymbol{\beta}_{\mathbf{1}} = \boldsymbol{\phi} - (c/4)\boldsymbol{\theta} + (1/4)\theta_0\mathbf{e}_d$, $\mathbf{e}_d = (0, 0, \dots, 0, 1, 0, \dots, 0)^{\intercal}$ with the d-th element equal to unit and $T_3(G) = f_1G + f_3G^3$ is the third-order Taylor expansion of $G(\Xi)$,

$f_1 = \partial G(\Xi)/\partial\Xi\big|_{\boldsymbol{\gamma}=\mathbf{0}}$ and $f_3 = (1/6)\partial^3 G(\Xi)/\partial\Xi\big|_{\boldsymbol{\gamma}=\mathbf{0}}$, $G(\Xi)$ being defined in previous section. The null hypothesis is

$$H_0 : \beta_{2j} = \beta_{3j} = \beta_{4j} = 0 \;\; j = 1, \ldots, p, \tag{18}$$

The test statistic:

$$LM_1 = (SSR_0 - SSR)/\hat{\sigma_v}^2 \; , \tag{19}$$

with SSR_0 and SSR denoting the sum of squared estimated residuals from the estimated auxiliary regression (17) and under the null and alternative, respectively and $\sigma_v^2 = (1/T)SSR$, has an asymptotic χ^2_{3p} distribution under H_0.
If the model is a GESTAR(p), then it is possible to show that the corresponding auxiliary regression is

$$\hat{\epsilon}'_t = \tilde{\boldsymbol{\beta}}_{\mathbf{1}}^{\mathsf{T}} \hat{\mathbf{z}}_{\mathbf{1}} + \boldsymbol{\beta}_{\mathbf{2}}^{\mathsf{T}} \mathbf{z_t} s y_{t-d} + \boldsymbol{\beta}_{\mathbf{3}}^{\mathsf{T}} \mathbf{z_t} s y^2_{t-d} + v'_t \; , \tag{20}$$

where v_t^{T} is a $NIID(0, \sigma^2)$ error term and $\tilde{\boldsymbol{\beta}}_{\mathbf{1}} = (\beta_{10}, \boldsymbol{\beta}'_{\mathbf{1}})'$, with $\beta_{10} = \phi_0 - c^2\theta_0$ and $\boldsymbol{\beta}_{\mathbf{1}} = \boldsymbol{\phi} - c^2\boldsymbol{\theta} + 2c\theta_0\boldsymbol{e}_d$; moreover $\boldsymbol{\beta}_{\mathbf{2}} = 2c\boldsymbol{\theta} - \theta_0\mathbf{e}_d$ and $\boldsymbol{\beta}_{\mathbf{3}} = -\boldsymbol{\theta}$. Thus the null hypothesis of linearity is

$$H'_0 : \boldsymbol{\beta}_{\mathbf{2}} = \boldsymbol{\beta}_{\mathbf{3}} = 0, \tag{21}$$

which can be tested by the test statistic:

$$LM_2 = (SSR_0 - SSR)/\hat{\sigma}^2_{v1}, \tag{22}$$

where SSR_0 and SSR are the sum of squared residuals from (20) under the null and the alternative respectively, $\hat{\sigma}^2_{v1} = (1/T)SSR$. When the null is true, the statistic (24) is asymptotically χ^2_{2p} distributed. A peculiar case of (24) is when $\boldsymbol{\beta}_2 = 0$ as

$\theta_0 = c = 0$, in which case, under the null

$$H_0'' : \boldsymbol{\beta_3} = 0, \tag{23}$$

the test has a statistic

$$LM_3 = (SSR_0 - SSR)/\hat{\sigma}^2_{v2} \tag{24}$$

distributed as a χ^2_p, with SSR_0, SSR and σ_{v2} defined in a similar way with respect the LM_2 case. The F-version of LM_1, LM_2 and LM_3, denoted as F_1, F_2 and F_3, may be preferable when testing (19) or (24) in order to preserve power in low samples. In practice the form of G is not known by the investigator. Teräsvirta (1994) proposes a battery of F-tests on the auxiliary model (17):

$$\begin{aligned} &H_{01} : \beta_4 = 0 \quad \text{vs} \quad H_{11} : \beta_4 \neq 0 \\ &H_{02} : \beta_3 = 0 | \beta_4 = 0 \quad \text{vs} \quad H_{12} : \beta_3 \neq 0 | \beta_4 = 0 \\ &H_{03} : \beta_2 = 0 | \beta_3 = 0 \ \text{ and } \ \beta_4 = 0 \quad \text{vs} \quad H_{22} : \beta_2 \neq 0 | \beta_3 = 0 \ \text{ and } \ \beta_4 = 0. \end{aligned} \tag{25}$$

and suggests an empirical rule – based on the results of a simulation experiment – to select the right transition function. For our aims, however, this is not a crucial issue, so we will do not discuss in details. The next Section 5 shows some results from this LST-derived test.

5 Simulation Study

This section investigates the empirical properties of the proposed locality test by a Monte Carlo experiment. We organize this section as follows: Sub-section 5.1 describes the design of the experiment; Sub-section 5.2 reports the results; Sub-section 5.3 provides a brief discussion.

5.1 Simulation Design

We consider two different data generating processes (DGP):

$$y_{1,t}^{(n)} = 0.4y_{1,t-1}^{(n)} - 0.25y_{1,t-2}^{(n)} + (0.01 - 0.9y_{1,t-1}^{(n)} + 0.795y_{1,t-2})^{(s)}G^{(n)}(\Xi) + \epsilon_{1,t}^{(n)} , \quad (26)$$

and

$$y_{2,t}^{(n)} = 0.8y_{2,t-1}^{(n)} - 0.7y_{2,t-2}^{(n)} + (0.01 - 0.9y_{2,t-1}^{(n)} + 0.795y_{2,t-2}^{(i)})G^{(n)}(\Xi) + \epsilon_{2,t}^{(n)} , \quad (27)$$

where $G^{(i)}(\Xi)$ has the same step form as in (16), $\epsilon_t^{(n)} \sim N(0,1)$, $n = \{1, \ldots, N\}$ denoting the n-esim drow of the process $\{y_t\}_{t=1}^{T}$ with $s = y_{t-1}$, $c = \frac{1}{T}y_t^{(i)}$, $N = 1,000$. $y_{2,t}^{(n)}$ (henceforth "DGP 1") is an additive nonlinear model with accentuated nonlinear behavior, due to the high autoregressive parameters driving $G(\Xi)$ that gave a high sensitivity to the size of the slope parameters; this can be the case of a macroeconomic indicator affected by an unexpected shocks affecting the whole dynamics. On the other hand, $y_{2,t}^{(n)}$ (henceforth "DGP 2") describes a mixed scenario.
In order to simulate the function $G(\cdot)$ we use a set of values of slope γ. These combinations allow us to investigate: i) the different cases of null, small, medium, extreme asymmetry respectively; ii) the effect of having different strength of nonlinearity, due to the different γ-s. Moreover, we consider three different hypotheses for T and the size α, namely $T = \{75, 150, 300\}$ – corresponding to a very small, small and medium sample sizes, respectively – and $\alpha = \{0.01, 0.05, 0.10\}$. The first 100 simulations have been discarded in order to avoid the initialization effect.

5.2 Results

Table 2 reports the results of the Monte Carlo simulation exercise of the locality test for the statistics F_1 and F_2 from the hypothesis system (25) discussed in Section 4 are reported in; the performances of F_3 statistic are poor and thus has been omitted. Several findings can be easily noticed: first, the two tests tends to be well-behaving for what concerns the empirical size. Second, and conversely, the empirical power is poor if an almost linear model is used, and in general for DGP 1. Third, 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 F_1 statistic passes from almost 0.02 when $\gamma = 0.5$ (hence almost linear model) to 0.6 when $\gamma = 500$ – hence, the increase is proportional but less than linearly – and similarly for statistic F_2. When considering DGP2 the range is still more abrupt: *ceteribus paribus*, F_1 is 0.05 when $\gamma = 0.5$ and 0.88 when $\gamma = 500$. When T increases, there role of γ became almost inflationary: 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 still more in DGP2.

Table 3 reports the results of a different exercise: we simulate the same DGPs with γ fixed at 10 and varying SRs; namely, we used all the scoring functions mentioned in Table 1 apart the logarithmic score – which has been previously investigated in Table 2). The power of the test is not affected by different SRs: the values of each F-statistic is the same for all the 19 SRs adopted (for example, the power of F_1 in DGP1 at nominal size of 5% is 0.35 with T=75, 0.57 with T=150 and 0.63 with T=300, respectively. Secondly, the DGP2 (that is the mixed scenario) delivers an extremely high power – actually, higher than the nonlinear scenario: in facts *ceteribus paribus*, the power of the F_1 statistic, is 0.77 with T=75, 0.92 with T=150 and 0.98 with T=300 and equivalent F_2 powers values are slightly lower.

5.3 Discussion

The above results convey a non-trivial picture of the role of the locality of the SRs in forecasting. In terms of statistical theory, we hold with two problems:

Structure dependence: Strong nonlinearity in the scoring structure is necessary to detect locality.

Score Invariance: if the SR is changed, the probability function of the event realization is the same.

Heuristically, we can reply to these idiosyncrasies as follows. For the model dependance problem, the SR, although treated as transition variable, is *per se* exogenous to the (G)STAR model parametrization, because it is just a functional of its distribution. In addition, we used the implicit assumption that scoring rule has an observable variable, when the converse is considerably more appropriate from a theoretical point of view. It is not so unrealistic to think that such a direct application of the LST test to this object is responsible to add some nonlinearity not captured by the AR parameters, hence leading to a spurious rejection of the null of locality symmetry when it is true.

The Score Invariance is one of the critical assumptions of the Lindley (1982) generalized theory on the admissibility of the Forecaster' utility and clearly stated as a condition for treating the scores as finitely-additive probability-behaving objects. In particular, the Lemma 4 of the same work proves the equivalence between two different scores of two quotations conditional on the same event, enhancing in a such a way the status of probability transform of the obtained value x[10]. In this sense our simulations are fully consistent with the theory and confirm the Patton (2016) theory that consistency of SRs is a non sufficient condition for coherency of the evaluation. In facts, despite some of the 19 scoring functions used in this experiment has Brègman-Savage representation, the empirical power of the test exactly identical to the ones having not, so they cannot be distinguished between these two types of utility functions. We can interpret this finding as follows: when the forecast user have to deal with different possible Forecasters, he/she can never know, *ex-post* what is the exact utility function driving them but have to specify, axiomatically, *ex-ante*. This sort of "undeterminancy" puzzle is the motivation of the adoption of the locality as criterion for SRs assessment. In fact, it tells the forecast user if the Barnard et al. (1962) likelihood principle – according to which all the evidence in a sample relevant to model parameters is contained in the likelihood function – holds or not. In this last case, the forecast is necessarily do be driven by something that exit from the likelihood. And, being the activity of any forecast user such that they need a good knowledge of estimation issues and methods, these deviations are likely represented by opinions.

6 Illustrations

6.1 Assessing Recession Probability in U.S. Business Cycle

In this application, we evaluate the US business cycle by using a new version the "Bry-Boshan Quarterly" (BBQ, henceforth) algorithm, introduced by Harding and Pagan (2002); Engel et al. (2005) (EHP, henceforth), which in turn is an advancement of the pre-existing algorithm for the detection of turning points in the US GDP. Here, the focus is twofold: (i) descriptive, by measuring the asymmetry of the business cycle by following the literature started with Burns and Mitchell (1946) and (ii) predictive, assessing of the probability of recession in line with Diebold and Ruderbusch (1989).

We use the US Index of Industrial Production in quarterly data – spanning from 1947Q1 to 2013Q1 – as proxy variable in logarithmic transformation. The source is the Federal Reserve Bank of St. Louis, Research Division. We first measure the effects of adopting the scoring rules in BBQ algorithm using a simple AR(5) to estimate the predictive density; the results are reported in Table 4 shows. Two main findings emerge immediately: first, the high duration of the expansion phase with respect to the contraction, in line with all the literature; second, the value of many indicators seems to be unaffected by different scoring rules.

We then repeat the analysis by estimating a STAR of the same order in both the linear and nonlinear part, $d = 4$ and 3 regimes, according to the General-to-Specific modelling strategy in Terasvirta and Dijk (2002). According to Table 5, the asymmetry of the cycle is increased, and in some cases, exacerbated (see the Excess of % triangle area). For example, let consider the case of duration (D): under linear model, it is 4.4 quarters in phase of contraction and 17.6 quarters in phase of expansion, and this evidence is uniform for almost all the SRs adopted with only two relevant exceptions: the LogS, which slightly under-estimate both of them (4.1 and 15.0 quarters, respectively) and Hyvaarinen Score (HS), which makes contraction still more conservative (3.6 quarters) but blows up the phase of expansion (56 quarters). When nonlinear model is used, the values of D remains the same in contraction, but, on a different side, the expansion phase (18.6 quarters on average) is augmented of more than a year. The same exceptions hold: when using a LogS, the duration of the cycle is lower both in contraction and in expansion (4.2 and 16.7 quarters, respectively); on the converse, if HS is assumed, it increase the contraction to 6.0 and reduces dramatically the expansion to 6.7 quarters – that is less than one third of the other measures. In the nonlinear case, also Weighted Power Score for the special case that power parameter is 1 (WPwrS, $\alpha = 1$) shares the same features of HS: it doubles the recession phase to 9.2 quarters and reduces the expansion to 14.3 quarters. These differences in these two scores are possibly more easy to see in the cumulated measures.

Figure 3 emphasizes the effects of SBB algorithm. When using log-level data, the estimated recessions are almost always coincident with NBER dates – with the relevant exception of 1975-6 Oil Crisis, which is recognized only by LogS and WPwrS. The LogS is sensitive to stagnation events as in end-'60s and anticipate the Gulf-War recession in 1991-2, and the same holds in both cases of linear and nonlinear specification for the DGP. However, when recessions probabilities are driven by a nonlinear model, the SBB is less prone to consider recession the years of stagnation, as proved by the fact that the indicator is permanently 1 since 1962 to 1971. In general LogS is almost everywhere right under STAR modelling apart a small deviance in end-1989-90. Differently, if data in growth rates are used, these results change: the (standard) BBQ over-react in the first part of the sample, where the number of recession episodes is the double of NBER dates and, on the opposite side, does not recognizes the first 80's episode and over-evaluate the cyclical movements after 1992. This holds also for many cases of SBB algorithm, but, at least, in under LogS recessions of first 80s are recognized – albeit with a large lead. When the same algorithm is used under STAR model, it is noticeable that for some SR, such CRPS, is not sensitive to any recession since first Oil Crisis 1973-74 to 1991. On the opposite side, WPwrS and WPSph overestimate the recession episodes after 1982, remaining at 0 until 2000. The LogS is relatively more precise than other measures although the 1982 recession remains over-estimated.

These results prove that the dating algorithms are sensitive to the the time series model, *ergo* of the probability of recession, and, in particular that is a non negligible difference between linear and non-linear parametrization. The role of nonlinearity is evidenced in Figure 4, where the original BB and LogS-Scored BB algorithms are compared with the transition functions of (Multiple Regime) STAR models. While indicator functions tends to over-evaluate Oil Crises, the transition functions Gs lie on the range [0 – 0.5] in the case of G_1 and G_3) and [0.5 – 1.0] in the case of G_2. In particular, Panel (b) plots the same transition function by ordering the observation in ascending order in a such a way that it is possible to appreciate the smoothness (or, on the opposite side, the abruptness) of the transition from extreme quiet state to extreme event 1. In the first transition function G_1, the large majority of the observations are in the upper tail of the sigmoid – thus implying an over-evaluation of recessions; the second transition function G_2 the observations are quite equally distributed; G_3 instead tends to stay in the lower half of the range, thus is more conservative in treating recession events. The difference of all these transitions with BB-derived Indicators is in the simplistic zero-one representation of the events of these last ones; not surprisingly, the observations in recession state of BB-type indicators are comparable to the over-fitting transition G_1 and G_2.

6.2 Assessing Bank of Norway's Fan Chart

The output gap (OG) measures the percentage deviation between GDP and projected potential GDP. Since the OG is one of the most important variables used in applied Macroeconomics, a correct assessment of its forecasts is part of every Central Bank.

The Bank of Norway's Monetary Policy Report (BoNMPR) has issued probabilistic forecasts of OG since March 2008 to December 2017, by using fan charts to visualize the deciles of the predictive distributions. The quarterly OG here investigated are in percentage changes over 12 months; the first quarter extends from March 31st to May 30, the second quarter from July 1st to September 30th, and so on[11]. Our empirical analysis uses the test by Amisano and Giacomini (2007) to assess the predictive performance of BoNMPR and, following the example of Gneiting and Ranjan (2011), compares the BoNMPR density forecasts to those derived from a simplistic AR(1) model that uses a rolling estimation window of length $m = 6$ quarters; see Appendix 2 for details.

According to our theoretical framework, the benchmark model nests a scoring structure that recover the Key Condition. In fact, it corresponds to Scoring Structure (9) with $\gamma = 0$, which in turn is the null hypothesis of local SRs, that is the forecast corresponds exactly to BoN quotation. On the other side, BoNMPR forecasts are the product of the bank's internal econometric model like System Averaging Model or Norway Economic MOdel; see the BoN models web page for references. In order to treat them with our method, we assume that BoN forecasts, whenever rejecting the locality hypothesis, correspond to the nonlinear Scoring Structure (9) with $\gamma > 0$. This is compatible with the empirical finding that nonlinear models are often associated with multimodal distribution as shown in Section 2. Hence, the Amisano-Giacomini t-statistic, if logarithmic SR (LogS) is assumed, tells us the distance between the BoNMPR forecasts and a forecast coming from a "purely" statistical model.

Table 6 compares the two methods at a prediction horizon of $k = 1$ quarters ahead, for 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 superiority of the Bank of Norway is not unambiguously clear, how shown by the different values τ and p-values. Under LogS, and other functional forms (like Quantum, Conditional Likelihood or Interval Scores) the test do rejects the null hypothesis of no equal predictive ability of Scoring Structure versus benchmark model, thus confirming the statistical coherency of the quotation. On the other side, it does not reject the null in several other occasions such as the WPwrS, most of Weighted Pseudo-Spherical (WPseudoSph) and Log-Cosh (LCS) Scores.

6.3 Discussion

For what concerns the SBBQ, a couple of *caveat* should be noted: first, the fact that, in many situations, different SRs convey the same cycle indicator is apparently counter-intuitive, since, being of them mutually different (consistently, in some cases), we would expect that each functional corresponds a unique value. On the other side, the Score Invariance principle empirically proved in the previous section seems to us a way to justify such an evidence. Moreover, this is an effect of the fact that many of the selected SRs are nested. It is not surprising that the measures of cycle in Tables 1 and Table 2 shows the same qualitative features in LogS and WPwrS with power equal to 1: in facts, according to their theoretical definitions, these two measures coincide. Concerning the peculiarity of HS, we have to notice that it is another case of local SR. In fact, it is based on a (non-trivial) differential transform of the logarithmic score, which increases its sensitivity.

Second, the EHP methodology here adopted and modified is not the only available for our aim. Artis et al. (2004) provides an alternative, more elaborated algorithm based on a Markov-Switching AR model. Third, and differently to Diebold and Ruderbusch (1989), we do not assume a bayesian strategy for the replication of probabilities of downturn. The results of the test should be looked in light of Teräsvirta's modelling strategy.

Also in our second illustration – the assessment of BoN Fan Charts – we have to make several considerations: first, in our exercise, we treated the BoNMPR forecasts as primitive data to be compared by a benchmark. Hence, according to Vovk and Shafer (2005) game-theoretical theory, BoN is Forecaster and we are the Forecast User (or Skeptic); moreover, since Reality has already played his role at the date of writing this paper, its role in building the (wrong) incentives of forecaster is annihilated. These particular circumstances are the only way to use simple predictive ability tests. In real world, a Central Bank is a Skeptic who have to deal of a (large) pool of professional forecasters and must deal with their sub-game against Reality and with cooperation between them in every time. Moreover, in several Central Banks, the phase of econometric analysis of Forecasters quotations and the consequent final assessment is separated by the final decisions of the Board of Governors. This generates, potentially, a huge decision bias.

Second, we assumed that the measure of Forecaster coherency can be done via simple average of a time series of each quotation. However, more general combination schemes are nowadays available: for example, Kapetanios et al. (2015) suggest a sieve estimator of weighted means where weights are allowed to be nonlinear combinations of density functions. On a different perspective, Billio et al. (2013) develop a combination method which assign time-varying weights to competing densities via Markov-Switching state-space model. All these methods can be nested in our scoring stucture framework.

Finally, and consequently, the proposed methodology is a very stylized way to address the elicitation of the true Forecaster's incentive. We ought that further research will be pursued in order to generalize it in several directions. For example, the scoring rule has been here assumed unknown but observed to exploit the properties of STAR family of models. Relaxing this last implicit assumption via unobserved component models could be an interesting development. Moreover, the formal testing is not the only strategy to do assessment of forecasts; for example, a graphical check via Murphy's diagram according to Ehm et al. (2016) might contribute to cover some of the counterintuitive results of our applications.

7 Conclusions

We introduced a novel frequentist framework based on a scoring structure which assumes (simplistic) interactions between Forecaster and Reality in order to make coherent evaluations of econometric forecasts. This structure allows the econometrician to build an LM-type test in order to verify the hypothesis of locality of the reward of forecaster's expected utility. The nonlinear model assumed to generate the predictive density forecasts is set to consider the scoring rule as a transition variable and, consequently, to nest the statistical inference of an established literature. The empirical power properties of the test are consistent with the fundamental requirements stated in literature on Decision Theory.

The scoring structure is a fundamental tool to elicit the forecaster's utility in several economic applications. In business cycle, the decision rules for a standard dating algorithm tends to exacerbate the probability of a recession. On the other side, our example on Bank of Norway's fan charts of the output gap reveals that relative coherency assessment requires a careful selection of the scoring functions between the many available. Thus, the refinement of this methodology via relaxation of scoring structure's operational and theoretical assumptions is a primary interest for econometric forecasting.

Notes

1 In what follows, we prefer to use the notion of "*utility*" rather than "*loss*" – which is more

frequently used in the econometric literature. This allows us to have a direct link the literature in Bayesian Statistics which, despite our frequentist approach, somewhat inspired this paper.

2 We identify two strands of literature: the first one (Rossi and Sekhposyan, 2013) generalizes the the statistical inference below the PITs to hold with unstable forecast environments. In the second, the statitical object is modified and generalized to improve the empirical power of the test statistic to assess the hypothesis of correct specification; see, Gonzáles-Rivera et al. (2011); Gonzáles-Rivera and Sun (2017) and the aforementioned literature.

3 See, *inter alia*, Mitchell and Hall (2005); Bao et al. (2007); Hall and Mitchell (2007); Amisano and Giacomini (2007); Clements and Harvey (2010); Kascha and Ravazzolo (2010); Geweke and Amisano (2011); Clements and Harvey (2012).

4 "*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*" (De Finetti, 1962, *cit.*, p. 359).

5 See, *inter alia*, Diebold and Mariano (1995); West (1996); Clark and McCracken (2001); Giacomini and White (2006); Giacomini and Rossi (2010)

6 For details on this peculiar generalization we refer the reader to Canepa and Zanetti Chini (2016) and notice that this DGP is similar to the one used in Teräsvirta (1994), equation (4.1). A large strand of literature demonstrates that this peculiar variable is typically nonlinear, see Anderson and Teräsvirta (1992), for example.

7 That is, 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.

8 In preliminary versions of this paper, this test has been built-up assuming the generalized version of the STAR model presented in Canepa and Zanetti Chini (2016). However, this complication does not improve the results here presented. Anyway, the equivalent tables with the result of the simulations with GSTAR model can be provided under request.

9 In the most simple case, $\boldsymbol{\tau}$ is a scalar. Anyway, in the most general case of multiple regime STAR model, it became a vector of length k; thus we prefer to maintain the vector notation.

10 "*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 saysfying* $P(x) = p$" (Lindley, 1982, *cit.*, p. 4.)

11 Data can be downloaded from the Bank of Norway's web page. We used the release of 2014.

Acknowledgments

This paper has been initialized when the author was visiting PhD student at CREATES - Center for Research in Econometric Analysis of Time Series (DNRF78), 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. This paper has circulated in form of MIMEO with title "*Testing and selecting local proper scoring rules*" and has been awarded of a Travel Grant by the International Institute of Forecaster, which is particularly acknowledged. The Author thanks Barbara Annicchiarico, Anders Bredahl Kock, Ana Beatriz Galvão, Domenico Giannone, Gary Koop, Michele Lenza, Antonio Lijoi, Alessandra Luati, Gael Martin, James Mitchell, Elisa Nicolato, Francesco Ravazzolo, Andrey Vasnev and Francesco Violante for their useful feedback and discussions, as well as all seminar participants to the "Introduction to Risk Management, ALM and Derivative Prices" Conference held in Aarhus University (January 2013), CREATES Lunch Seminar (May 2014) in the same University, the 34^{th} International Symposium on Forecasting in Rotterdam (July 2014), the IAAE 2016 Annual Conference in the University of Milano Bicocca (June 2016) and the NBP 2017 Workshop in Forecasting in Warsaw (November 2017). All the computations in this paper have been performed by using MATLAB R2009b. The codes for replication of the figures and tables are available from the author upon request. In particular, the results on the empirical applications in Section 6 have been obtained developing the codes by James Engel and Barbara Rossi who are gratefully acknowledged. The usual disclaimer applies. Finally, the Author is in debt with the doctors, assistants and nurses of the Department of Emathology of Policlinico "S. Matteo" of Pavia, without whose (free) cares this paper would possibly remain unfinished.

References

Amisano, G., Giacomini, R., 2007. Comparing Density Forecasts via Weighted Likelihood Ratio Tests. Journal of Business and Economic Statistics 25 (2), 177–190.

Anderson, H., Teräsvirta, T., 1992. Characterizing nonlinearities in business cycles using smooth transition autoregressive models. Journal of Applied Econometrics 7 (S1), S119–S136.

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 (4), 537–565.

Bao, Y., Lee, T., Saltoglu, B., 2007. Comparing Density Forecast Models. Journal of Forecasting 26 (3), 203–225.

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.

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.

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

Burns, A., Mitchell, W., 1946. Measuring Business Cycles. National Bureau of Economic Research.

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

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

Chan, K., Tong, H., 1986. On estimating thresholds in autoregressive models. Journal of Time Series Analysis 7 (3), 178–190.

Clark, T., McCracken, M., 2001. Test for equal forecast ability and encompassing for nested models. Journal of Econometrics 105, 85–110.

Clements, M., Harvey, D., 2010. Forecast Encompassing Tests and Probability Forecasts. Journal of Applied Econometrics 25, 1028–1062.

Clements, M., Harvey, D., 2012. Combining probability forecasts. International Journal of Forecasting 27, 208–223.

Davies, R., 1977. Hypothesis testing when a nuisance parameter is present only under the alternative. Biometrika 64 (1), 247–254.

Dawid, A., 2007. The geometry of proper scoring rules. The Annals of the Institute of Statistica Mathematics 59, 77–93.

Dawid, A., Lauritzen, S., Parry, M., 2012. Proper Local Scoring Rules on Discrete Sample Space. The Annals of Statistics 40 (1), 593–608.

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

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. Wiley, New York.

De Finetti, B., 2017. Theory of probability: A critical introductory treatment. Vol. 6. John Wiley & Sons.

Diebold, F., Mariano, R., 1995. Comparing Predictive Accuracy. Journal of Business & Economic Statistics 20 (1), 134–144.

Diebold, F., Ruderbusch, G., 1989. Scoring the Leading Indicators. The Journal of Business 62, 369–391.

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

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

Ehm, W., Gneiting, T., 2012. Proper Local Scoring Rules on Discrete Sample Space. The Annals of Statistics 40 (1), 609–637.

Ehm, W., Gneiting, T., Jordan, A., Krueger, F., 2016. Of quantiles and expectiles: consistent scoring functions, Choquet representations and forecast rankings. Journal of Royal Statistical Society, ser. B 78, 505–562.

Elliott, G., Timmermann, A., 2008. Economic Forecasting. Journal of Economic Literature 46 (1), 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.

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

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

Giacomini, A., White, H., 2006. Tests for Conditional Predictive Ability. Econometrica 74 (6), 1545–1578.

Giacomini, R., Rossi, B., 2010. Forecast Comparisons in Unstable Environments. Journal of Applied Econometrics 25, 595–620.

Gneiting, T., Balabdaoui, F., Raftery, A. E., 2007. Probabilistic forecasts, calibration and sharpness. Journal of Royal Statistical Society 69 (2), 243–268.

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

Gneiting, T., Ranjan, R., 2011. Comparing Density Forecasts Using Threshold- and Quantile-Weighted Scoring Rules. Journal of Business & Economic Statistics 29 (3), 411–422.

Gonzáles-Rivera, G., Senyuz, Z., Yoldas, E., 2011. Autocontours: Dynamic Specification Testing. Journal of Business Economics and Statistics 29, 186–200.

Gonzáles-Rivera, G., Sun, Y., 2017. Density forecast evaluation in unstable environments. International Journal of Forecasting 33, 416–432.

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

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

Hall, S., Mitchell, J., 2007. Combining Density Forecasts. International Journal of Forecasting 23 (1), 1–13.

Hamill, T., 2001. Interpretation of rank histograms for verifying ensemble forecasts. Monthly Weather Review 125, 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 (6), 1916–1921.

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

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

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

Kascha, C., Ravazzolo, F., 2010. Combining Inflation Density Forecasts. Journal of Forecasting 29 (1–2), 231–250.

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 (3), 491–499.

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

Mitchell, J., Hall, S., 2005. Evaluating, Comparing and Combining Density Forecasts Using the KLIC with an Application to the Bank of England and NIESR "Fan" Charts of Inflation. Oxford Bulletin of Economics and Statistics 67 (S1), 995–1033.

Mitchell, J., Wallis, K., 2011. Evaluating density forecasts: Forecast combinations, model mixtures, calibration and sharpness. Journal of the Applied Econometrics 26 (6), 1023–1040.

Parry, M., Dawid, A., Lauritzen, S., 2012. Proper Local Scoring Rules. The Annals of Statistics 40 (1), 561–592.

Patton, A., 2016. Comparing Possibly Misspecified Forecasts.

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

Rosenblatt, M., 1952. Remarks on a Multivariate Transformation. Annals of Mathematical Statistics 23 (3), 470–472.

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.

Teräsvirta, T., 1994. Specification, estimation and evaluation of smooth transition autoregressive models. Journal of the American Statistical Association 89 (425), 208–218.

Terasvirta, T., Dijk, D. V., 2002. Smooth transition autoregressive models – a survey of recent developments. Econometric Reviews 21, 1–47.

Tong, H., 1983. Threshold Models in Non-Linear Time Series Analysis. No. 21 in Lecture Notes in Statistics. Springer-Verlag, New York.

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

Vovk, V., Shafer, G., 2005. Good randomized sequential probability forecasting is always possible. Journal of Royal Statistical Society, ser. B 67 (747-763), 491–499.

West, K., 1996. Asymptotic inference about predictive ability. Econometrica 64, 1067–1084.

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

Zanetti Chini, E., 2017. Generalizing Smooth Transition Autoregressions. DEM Woking Paper no. 138.

Table 1: Scoring Rules for density of continuous variables and their features

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

NOTE: $A_t \subseteq \mathcal{X}$, $t \in \mathcal{T} = \mathcal{X}$ so that $\{A_t\} \equiv \{t\}$; $\mathcal{P}$: Borel probability measure; L: Lebesgue probability measure, μ: σ-finite measure

Table 2: Empirical Size and Power of LM test for Locality for different slope parameters

Empirical Size

T	γ	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	0.0532	0.0539	0.0543	0.0006	0.0015	0.0022	0.0208	0.0210	0.0215
150		0.0002	0.0034	0.0099	0.0385	0.0387	0.0392	0.0001	0.0001	0.0001	0.0133	0.0133	0.0133
300		0.0000	0.0000	0.0001	0.0544	0.0544	0.0544	0.0001	0.0001	0.0001	0.0172	0.0172	0.0172

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	0.0211	0.021
	0.5	0.0009	0.0066	0.0192	0.0492	0.0497	0.0501	0.0082	0.0291	0.0498	0.0360	0.0368	0.0388
	1	0.0029	0.0132	0.0217	0.0982	0.0098	0.0103	0.1553	0.3588	0.4658	0.0256	0.0354	0.0422
	5	0.1184	0.2163	0.2958	0.0990	0.2026	0.2879	0.5436	0.6207	0.6593	0.4495	0.5254	0.5538
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	0.7261	0.7302	0.7324
300	10	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
	50	0.7458	0.8588	0.8894	0.7487	0.8460	0.8753	0.9980	0.9981	0.9981	0.7686	0.7707	0.7715
	100	0.7752	0.8740	0.9035	0.7830	0.8621	0.8892	0.9997	0.9998	0.9999	0.7676	0.7703	0.7709
	500	0.7890	0.8916	0.9214	0.7998	0.8796	0.9035	0.9978	1.0000	1.0000	0.7658	0.7675	0.7683

Table 3: Empirical Power of LM test for Locality for different Scoring functions 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	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPwrS (General)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPwrS ($\beta = -1$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPwrS ($\beta = 0$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPwrS ($\beta = 1/2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPwrS ($\beta = 2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
PsdSphS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPsdSph	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPsdSphS ($\beta = -1$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPsdSphS ($\beta = 0$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPsdSphS ($\beta = 1/2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
WPsdSphS ($\beta = 2$)	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
TsallisS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
ES	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
GES	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
PSpctr	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
CRPS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
QuantS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
HS	0.1993	0.3542	0.4459	0.1710	0.3189	0.4163	0.7026	0.7755	0.8080	0.5978	0.6845	0.7210
$T = 150$												
QSR	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPwrS (General)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPwrS ($\beta = -1$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPwrS ($\beta = 0$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPwrS ($\beta = 1/2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPwrS ($\beta = 2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
PsdSphS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPsdSph	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPsdSphS ($\beta = -1$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPsdSphS ($\beta = 0$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPsdSphS ($\beta = 1/2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
WPsdSphS ($\beta = 2$)	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
TsallisS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
ES	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
GES	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
PSpctr	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
CRPS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
QuantS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
HS	0.2831	0.4536	0.5721	0.2788	0.4553	0.5625	0.9162	0.9250	0.9287	0.8994	0.9039	0.9090
$T = 300$												
QSR	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPwrS (General)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPwrS ($\beta = -1$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPwrS ($\beta = 0$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPwrS ($\beta = 1/2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPwrS ($\beta = 2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
PsdSphS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPsdSph	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPsdSphS ($\beta = -1$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPsdSphS ($\beta = 0$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPsdSphS ($\beta = 1/2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
WPsdSphS ($\beta = 2$)	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
TsallisS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
ES	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
GES	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
PSpctr	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
CRPS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
QuantS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694
HS	0.4903	0.6347	0.7035	0.4818	0.5933	0.6553	0.9876	0.9877	0.9881	0.7662	0.7691	0.7694

Table 4: BC effects via SRs estimated from AR(5)

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

Table 5: BC effects via SRs estimated from MR-STAR(5,5,4)

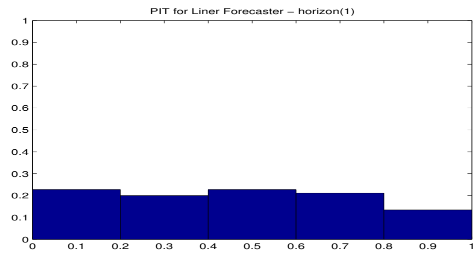
S(Q, y)	$S(\cdot,\cdot)^{STAR}$	Duration		Amplitude		Cumulated Value		Excess Mov. as % of Triangular Area		Cum.Val. of Duration		Cum.Val. of amplitde		Cum.Value of Excess	
		C	E	C	E	C	E	C	E	C	E	C	E	C	E
No SR	0.0000	4.2308	16.7500	-0.0604	0.2250	-0.1824	2.9900	31.5354	81.6785	0.4223	0.6531	-0.8133	0.7162	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	0.0021	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

Table 6: Amisano and Giacomini (2007) test for Norges Bank's Fan Chart of Output Growth at a prediction horizon of 12 month for different scoring rules with Bank of Norway taking the role of $\bar{S}^f$ and the benchmark AR density forecast the role of $\bar{S}^g$

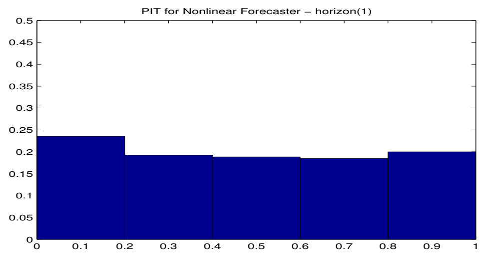
S(Q, y)	$\bar{S}^f$	$\bar{S}^g$	σ	t	P-value
QSR	0.5582	0.5582	0.0000	-	0.0000
WPowerS	2.8147	2.9534	0.1011	-3.3575	0.9991
" ($\alpha = -1$)	527.6771	2.9534	1.08e05	-0.0032	0.5012
" ($\alpha = 0$)	527.5193	670.8714	1.08e05	-0.0032	0.5012
" ($\alpha = 1/2$)	-1,062.301	-1,352.7636	4.43e05	0.0016	0.4993
" ($\alpha = 1$)	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
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
" ($\alpha = 1/2$)	-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
LogS	0.0273	0.0273	0.0000	-	0.0000
IntS	3.5000	3.5000	0.0000	-	0.0000
TsallisS	1.2229	1.2229	0.0000	-	0.0000
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
CRPS	0.0132	0.0120	7.2746e-06	395.9622	0.0000
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	0.0000
LCS	0.0552	0.0552	0.0000	-	0.0000

Figure 1: The Hamill paradox

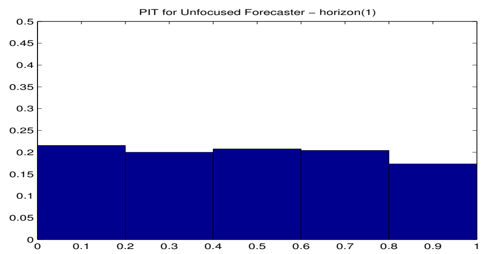
(a) PIT under linear forecast



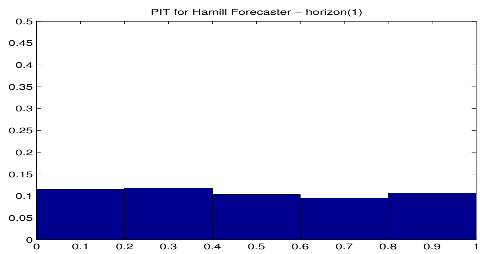
(b) PIT under GSTAR forecast



(c) PIT under unfocused forecast

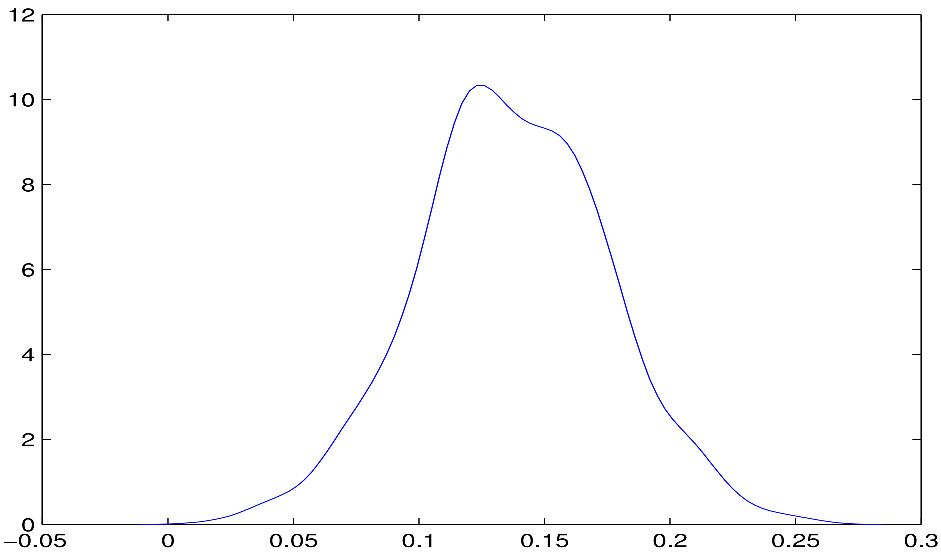


(d) PIT under Hamill's forecast



NOTE: This figure plot the output of the motivating example described in Section 2.

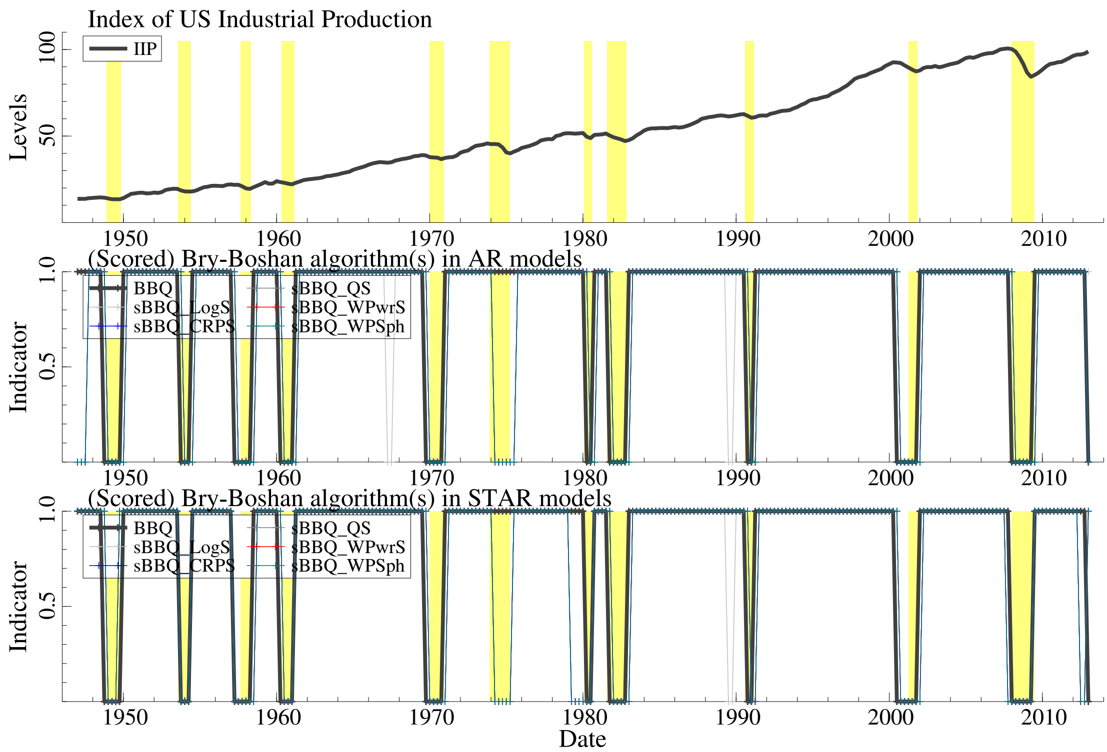
Figure 2: Simulated predictive density of a GSTAR model



NOTE: This figure shows the density of a GSTAR model with parameter values: $\phi_0 = 0.0$, $\phi_1 = 0.4$, $\phi_2 = -0.25$, $\theta_0 = 0.01$, $\theta_1 = -0.9$, $\theta_2 = 0.795$, $\gamma_1 = -200$, $\gamma_2 = 100$, $c = \bar{y}_t$; we adopt $S = 1,000$ Monte-Carlo draws and $B = 10.000$ bootstrap replications; see Zanetti Chini (2017) for details.

Figure 3: The effects of different SR in BBQ algorithm

(a) Analysis of data in logarithms



(b) Analysis of data in quarterly growth rates

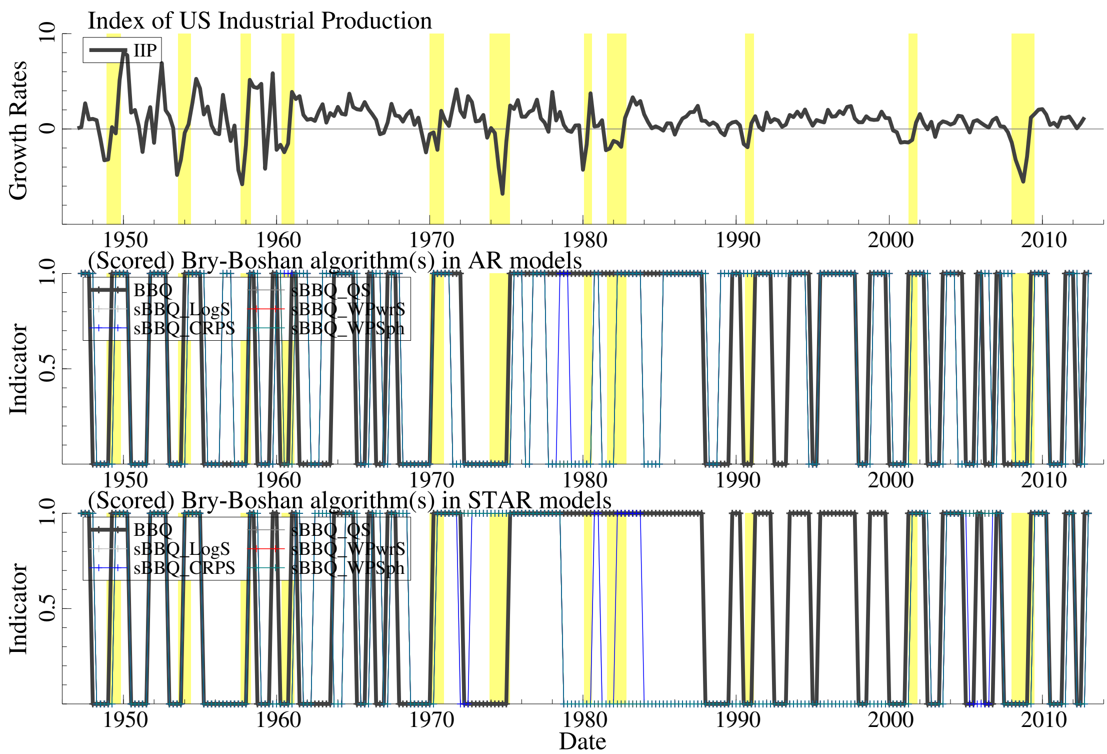
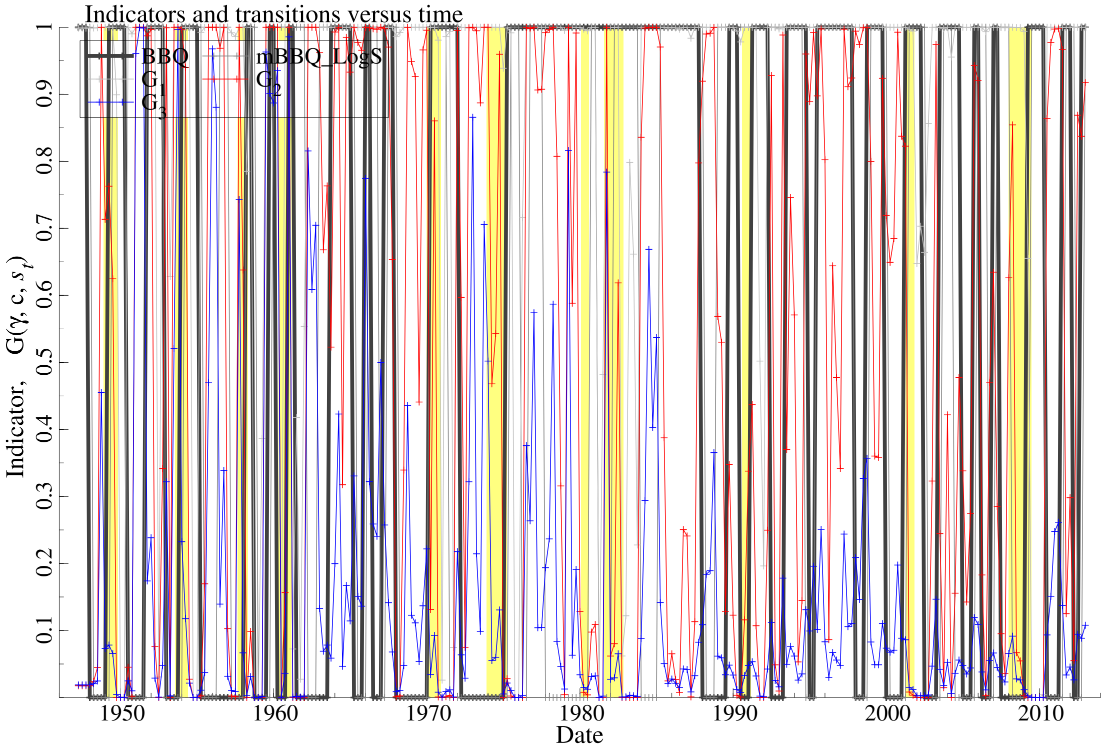
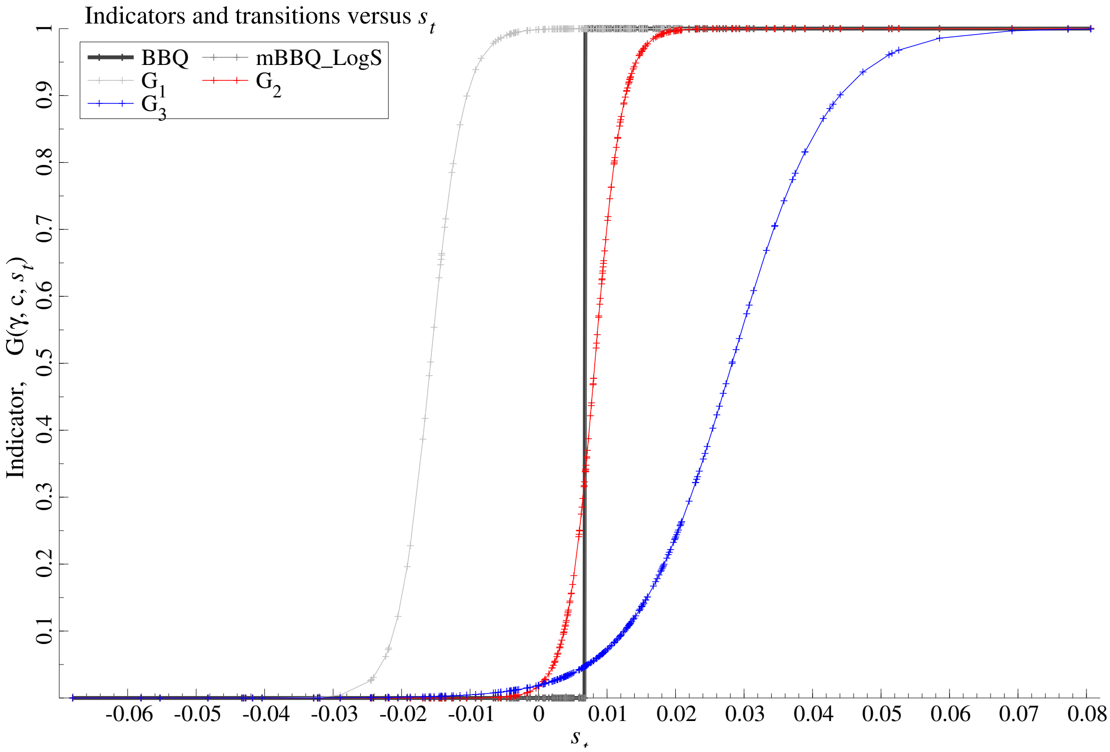


Figure 4: The SBBQ algorithm and nonlinear modelling.

(a) Indicators from (S)BBQ algorithm and STAR transition function vs time



(b) Indicators from (S)BBQ algorithm and STAR transition function vs transition variable



A The Scored Bry-Boshan Algorithm

Understanding the business cycle means to realize how the transition probabilities are likely to be affected by the nature of Δy_t, which is the scope of the EHP dating algorithm, in which the phases are defined as $S_t = 1(\Delta y_t > 0)$, i.e. when there is a positive growth rate at time t, the economy is in a state of expansion (E), while a negative one refers to a contraction (C). Under random walk hypothesis, the probabilities of a change in phase at time t is

$$\begin{aligned} Pr(EC) &: Pr(S_{t+1} = 0 | S_t = 1) = Pr(\Delta y_{t+1} < 0 | \Delta y_t > 0) \\ Pr(CE) &: Pr(S_{t+1} = 1 | S_t = 0) = Pr(\Delta y_{t+1} > 0 | \Delta y_t < 0) \end{aligned} \tag{28}$$

Our innovation consists in understanding how the transition probabilities are likely to be affected by the nature of Δy_t *given* the estimated scoring rule(s).
Let the turning point dates produce K phases of expansions and contractions, with an index $(i = 1, ..., K)$ denoting the i-th expansion or contractions. Then the definition of peaks and troughs can be stated follows:

$$\begin{aligned} \text{peak at } t &= \{(\Delta_2 y_t, \Delta y_t) > 0, (\Delta y_{t+1}, \Delta_2 y_{t+2}) < 0\} \\ \text{trough at } t &= \{(\Delta_2 y_t, \Delta y_t) < 0, (\Delta y_{t+1}, \Delta_2 y_{t+2}) > 0\} \end{aligned} \tag{29}$$

with $\Delta_2 y_t = y_t - y_{t-2}$. Additionally we are interested the following indicators of cycle:

- Duration: D_i
- Amplitude: A_i
- Cumulative Movements: $0.5 \cdot (D_i \cdot A_i)$
- Excess cumulated movements: $E_i = (C_{Ti} - C_i + 0.5 \cdot A_i)/D_i$
- Coefficient Variation: $CV_i = \frac{\sqrt{(1/K \sum_{i=1}^{K} (D_j^i - \bar{D}^i))}}{1/K \sum_{j=1}^{K} D_j^i}$

A set of restrictions have to be imposed in order to define properly what movement can be defined as cycle and what is contrarily treated as outlier. The literature suggests: $K = 2$ for definition of peak, which produces the expression (29); the minimal duration to have a complete cycle is 5 quarters; the turning phase (the number of quarters before the peak) has been set at 1.

The indicators stated above can be thought as two states of a Markov process, whose transition is the object of investigation. Thus, a peak at time t demarcates a state of expansion at time t, $S_t = 1$, from a state of contraction at time t +1, $S_{t+1} = 0$. The states S_t are then a binary Markov process, which might be summarized with the transition probabilities $Pr(S_{t+1}|S_t)$.

Model selection has been conducted via Bayesian Information Criterion (BIC); this suggests an autoregressive order 5. We first estimated a linear AR(5) model; the estimated model are the basis for computing the h-step ahead density forecast via Monte Carlo simulation. A set of 25 different scoring rules – additionally to the case that no score is computed on (29) is then achieved and used to proceed with our modified version of EHP algorithm, which we call Scored Bry-Boshan (SBB) algorithm.

B The Amisano-Giacomini test.

Let be $\bar{S}^f_n = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{f}_{t+k}, y_{t+k})$ and $\bar{S}^g_n = \frac{1}{n-k-1}\sum_{t=m}^{m+n-k} S(\hat{g}_{t+k}, y_{t+k})$ the average scores of two density forecast, f and g, respectively; then, the test statistic for the null hypothesis that $\Delta^* = \bar{S}^f_n - \bar{S}^g_n = 0$ is

$$t_n = \sqrt{n}\frac{\Delta^*}{\hat{\sigma}_n}, \quad \hat{\sigma}^2_n = \frac{1}{n-k+1}\sum_{j=-(k-1)}^{k-1}\sum_{t=m}^{m+n-k-|j|}\Delta_{t,k}\Delta_{t+|j|,k} \tag{30}$$

and $\Delta_{t,k} = S^f_n - S^g_n$. In this kind of analysis, f is preferable to g if and only if $S^f < S^g$.

Research Papers 2018

CREATES
Center for Research in Econometric Analysis of Time Series