Testing ideal calibration for sequential predictions

Thibault Modeste, Clément Dombry, Anne-Laure Fougères

Testing ideal calibration for sequential predictions

Thibault Modeste * Clément Dombry † Anne-Laure Fougères ‡

August 28, 2024

Abstract

Forecasts and their evaluation are major tasks in statistics. In real applications, forecasts often take the form of a dynamic process evolving over time and this sequential point of view must be taken into account. A strategy for forecast evaluation is calibration theory based on the *Probability Integral Transform*. The idea is to check the conformity between the forecast and the observation. Here, ideal forecasts are characterized by conditional calibration and we present some new tests based on regression trees.

*CNRS / Université de Pau et des Pays de l'Adour / E2S UPPA Laboratoire de math´ematiques et applications IPRA, UMR 5142 B.P. 1155, 64013 Pau Cedex, France E-mail: thibault.modeste@univ-pau.fr

†Université de Franche-Comté, CNRS, LmB, F-25000 Besançon, France. E-mail: clement.dombry@univ-fcomte.fr

‡Univ. Lyon, Université Claude Bernard Lyon 1, CNRS UMR 5208, Institut Camille Jordan, F-69622 Villeurbanne, France E-mail: fougeres@math.univ-lyon1.fr

Contents

1 Introduction

Forecasting is an important task in statistics with many applications such as in meteorology (Vannitsem et al., 2021), hydrology (Tiberi-Wadier et al., 2021), health (Henzi et al., 2021), energy (Hong et al., 2016). To take account for uncertainties or statistical errors, a prediction is not just a single value. Several approaches exist to deal with these uncertainties, including interval prediction and probabilistic prediction. We will focus on the latter approach. To estimate the trend of the future, predictions then take the form of probability measures, i.e. the forecaster predicts several, even infinitely many, possible scenarios with different probabilities of success. This is the idea introduced by Epstein (1969b) where the process studied is governed by deterministic laws but the initial conditions are unknown.

Forecast verification is crucial in order to improve a forecasting method or compare different forecast strategies. Assessing forecast accuracy is done by comparing the predictive distributions and the actual observations which are real valued. But since these two quantities are not intrinsically comparable, assessing in this case is a difficult task. Two main methods exist, scoring rules (Gneiting and Raftery, 2007) and calibration theory (Gneiting et al., 2007). The idea of the first approach is to create a pseudo distance between the observations and the predictions, while the second one is in some way more qualitative. We can find an early introduction of the notion of calibration in Dawid (1984) and Diebold et al. (1998). The reliability of a probabilistic forecast is defined as its skill to be conform with the actual observation.

The evaluation methods depend on the assumptions about these measures. In some cases, probability measure is discrete and uniform over its atoms. This type of prediction is called ensemble forecast. It is a common approach in meteorology with the *Numerical Weather Prediction* (NWP) where each atom represents a different scenario starting from a different initial point. Several diagnostic tools have been introduced (Bröcker, 2009; Weigel, 2011), the rank histogram is one of them and one of the most studied (Anderson, 1996; Talagrand et al., 1997). The measures can also be assumed with a density, as is the case in financial risk management (Diebold et al., 1998).

In the general case, the fundamental notion of *Probability Integral Transform* (PIT, David and Johnson (1948)) is introduced, which plays a crucial role in calibration theory. More recently, Tsyplakov (2013) describes the use of the PIT as a diagnostic tool for calibration. The review by Gneiting and Katzfuss (2014) provides a nice discussion of these notions. Several types of calibrations have been introduced over time. Their definitions and links will be given in Section 2. For this article, we are interested in a particular type of calibration, the ideal calibration. A prediction of a phenomenon Y is said to be ideally calibrated with respect to an information $\mathcal{F}$ if the latter prediction is the conditional distribution of Y for $\mathcal{F}$. That is to say that the forecaster predicts the phenomenon perfectly with respect to the known information. Few results exist in the literature for this type of calibration. This notion of ideal calibration is closely related to the cross-calibration defined by Strähl and Ziegel (2017), which aims at predicting perfectly with the knowledge of other forecasters' information. In this context, the authors propose two tests based on the study of PIT.

In Bröcker (2022), the author also focuses on ideal calibration but in the specific case of binary events. This sub-case of probabilistic prediction is called *probability forecasting*.

The concept of his test is to write the calibration in terms of an empirical process involving the observations of the phenomenon and its forecasters. Under some conditions, this empirical process has good asymptotic behavior under the ideal calibration hypothesis. We will proceed in an analogous way.

Section 2 sets the definitions and details the classical framework, the PIT and the different notions of calibration. This section ends with Corollary 3 that is a key result from which we will be able in Section 3 to write the ideal calibration in terms of an empirical process. Under certain conditions, this process will converge to a Gaussian limit process. Since this limit has an unknown distribution, we will justify the bootstrap to approximate it. Section 4 introduces three new tests based on regression trees (CART algorithm). The purpose of this section is to show that this regression is rewritten as the functional of the empirical process that involves the PITs. The performance of these three tests are then numerically investigated in Section 5 in an auto-regressive model framework. We conclude this paper with Section 6 where we give some limitations of ideal calibration and our tests. Then we present a recent notion of weaker calibration and discuss an adaptation perspective for our tests. All the proofs are relegated to Section 7.

2 Calibration theory for the validation of dynamic forecast

2.1 Probability Integral Transform

Before introducing the different notions of calibration for probabilistic forecasts, we provide the definition of the *Probability Integral Transform* (PIT) for deterministic probability measures, see David and Johnson (1948) or Brockwell (2007).

Definition 1. *Let F be a deterministic cumulative distribution function (CDF), Y be a random variable and $V \sim \mathrm{Unif}([0,1])$ independent of the variable Y. Their PIT is defined as*

$$Z_F^Y = VF(Y^-) + (1-V)F(Y).$$

When F is continuous, we have simply $Z_F^Y = F(Y)$. In this case it is well known that $Y \sim F$ implies that $F(Y)$ is uniformly distributed on $[0,1]$; the converse is also true that $F(Y) \sim \mathrm{Unif}([0,1])$ implies $Y \sim F$. To extend this fundamental property to the general case when F is not necessarily continuous, it is necessary to introduce the randomization V in the definition of the PIT.

Lemma 1. *Let F be a deterministic CDF, Y be a random variable and Z_F^Y the associated PIT. The following statements are equivalent:*

i) $Y \sim F$;

ii) $Z_F^Y \sim \mathrm{Unif}([0,1])$.

The preceding lemma states a fundamental property of the PIT and seems to be well-known, even if we do not know a clear reference (especially for the proof of the implication $ii) \Rightarrow i)$). We postpone the proof to Section 7.

2.2 Prediction Space

We introduce the classical framework of prediction space developed in Gneiting and Ranjan (2013) and Strähl and Ziegel (2017).
Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space and $Y : (\Omega, \mathcal{G}, \mathbb{P}) \to \mathbb{R}$ a random variable representing the quantity of interest that we wish to predict. Let $\mathcal{M}_1(\mathbb{R})$ be the space of probability measures P on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ endowed with the σ-algebra generated by the maps $P \mapsto P(A)$, $A \in \mathcal{B}(\mathbb{R})$. With a slight abuse of notation, we often identify a probability measure on $\mathbb{R}$ and its CDF. A (random) probabilistic forecast is a measurable map $F\colon (\Omega, \mathcal{G}, \mathbb{P}) \to \mathcal{M}_1(\mathbb{R})$. Let $\mathcal{F} \subset \mathcal{G}$ be a sub-σ-algebra representing the information available to the forecaster. Sometimes it will be assumed that $\mathcal{F} = \sigma(X)$ where $X\colon \Omega \to \mathbb{R}^d$ is an observed covariate used to produce the prediction. The constraint that the forecaster has only access to the information encoded by $\mathcal{F}$ corresponds to the $\mathcal{F}$-measurability of F. An auxiliary random variable $V \sim \mathrm{Unif}([0,1])$, independent of $\mathcal{F}$ and Y, is introduced that will be useful to define the PIT $Z_F^Y = VF(Y^-) + (1-V)F(Y)$. Note that the PIT is measurable with respect to $\mathcal{G}$, see Proposition 20. To summarize, the one step prediction space is a triple (Y, F, V) defined on $(\Omega, \mathcal{G}, \mathbb{P})$ together with a sub-σ-algebra $\mathcal{F}$ and with F assumed $\mathcal{F}$-measurable.

A sequential prediction space consists in an extension of the preceding construction to model sequential prediction. The probability space $(\Omega, \mathcal{G}, \mathbb{P})$ is endowed with a sequence of random variables $(Y_n)_{n\in\mathbb{N}}$. At time $n \in \mathbb{N}$, we wish to predict Y_{n+T} where $T \geq 1$ is the so-called lead time. A sequence of (random) probabilistic forecasts $(F_n)_{n\in\mathbb{N}}$ is a sequence of measurable maps $F_n\colon (\Omega, \mathcal{G}, \mathbb{P}) \to \mathcal{M}_1(\mathbb{R})$. The available information evolves over time and is represented by a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$, *i.e.* a nondecreasing sequence of sub-σ-algebras $\mathcal{F}_n \subset \mathcal{G}$. The sequence of forecasts $(F_n)_{n\in\mathbb{N}}$ is assumed to be measurable with respect to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Sometimes we will assume filtration generated by a sequence of covariates $(X_n)_{n\in\mathbb{N}}$, i.e. $\mathcal{F}_n = \sigma(X_k, k \leq n)$. Finally, we consider an i.i.d. sequence $(V_n)_{n\in\mathbb{N}}$ of random variables uniformly distributed on $[0,1]$ and independent of $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and $(Y_n)_{n\in\mathbb{N}}$. To summarize, a sequential prediction space is a sequence of triples $(Y_n, F_n, V_n)_{n\in\mathbb{N}}$ defined on $(\Omega, \mathcal{G}, \mathbb{P})$ together with a filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ and with $(F_n)_{n\in\mathbb{N}}$ assumed measurable with respect to $(\mathcal{F}_n)_{n\in\mathbb{N}}$.

2.3 Ideal probabilistic forecast and calibration

The main purpose of this work is to build a statistical procedure to verify the quality of a probabilistic forecast and its ability to exploit the information available to the forecaster. This is related to the notion of *ideal forecast*.
The ideal forecast with respect to a sub-σ-algebra $\mathcal{F}$ is defined as the conditional distribution

$$F^* = \mathcal{L}(Y \mid \mathcal{F}).$$

Heuristically, the ideal forecast F^* is the best probabilistic forecast of Y that can be achieved if the forecaster has only access to the information $\mathcal{F}$.
The qualitative assessment of probabilistic forecast relies on the notion of calibration. Roughly speaking, calibration means that the observation Y and the probabilistic forecast F are in agreement and this is made precise thanks to the probability integral transform. We refer to Tsyplakov Tsyplakov (2013) and Gneiting and Katzfuss Gneiting (2014) for comprehensive reviews on probabilistic forecast and calibration.

Definition 2. *Let (Y, F, V) be a one step prediction space on $(\Omega, \mathcal{G}, \mathbb{P})$. The probabilistic forecast F is said probabilistically calibrated if $Z_F^Y \sim \text{Unif}([0,1])$.*

Gneiting and Ranjan (2013, Theorem 2.8) show that an ideal forecast F^* (with respect to any σ-field $\mathcal{F}$) is always probabilistically calibrated. A partial converse holds in the case of binary outcome: Gneiting and Ranjan (2013, Theorem 2.11) show that a probabilistically calibrated forecast F is ideal with respect to the σ-field $\mathcal{F} = \sigma(F)$ which it generates, *i.e.* it satisfies $F = \mathcal{L}(Y \mid F)$. This property is called auto-calibration. Note that the assumption of binary outcomes is crucial here, see Gneiting and Resin (2021, Example 2.4) for a counter example.
These properties show that probabilistic calibration is an important property of ideal forecast. It is is however a weak property that does not take into account the information $\mathcal{F}$. To this purpose *complete calibration* was introduced (Diebold et al., 1998; Mitchell and Wallis, 2011; Gneiting and Ranjan, 2013).

Definition 3. *Let (Y, F, V) be a one step prediction space on $(\Omega, \mathcal{G}, \mathbb{P})$. The probabilistic forecast F is said completely calibrated with respect to $\mathcal{F} \subset \mathcal{G}$ if $Z_F^Y \sim \text{Unif}([0,1])$ and Z_F^Y is independent of $\mathcal{F}$.*

This notion was introduced in Mitchell and Wallis (2011) for density forecasts. A similar notion of *cross-calibration* is introduced in Strähl and Ziegel (2017) where the information $\mathcal{F} = \sigma(F_1, \dots, F_K)$ is generated by competitive forecasts and the goal is to check whether the forecaster can efficiently exploit the information provided by alternative forecasters.
It is clear from the definition that *complete calibration* is a stronger property than *probabilistic calibration*, since independence between information and PIT is assumed in addition to the uniform distribution of the PIT. Interestingly, complete calibration is equivalent to

$$\mathcal{L}\left(Z_F^Y \mid \mathcal{F}\right) = \text{Unif}([0,1]). \tag{1}$$

The notion of complete calibration is the correct notion to verify that a forecast is ideal, as stated in the following proposition.

Proposition 2. *Let (Y, F, V) be a one step prediction space on $(\Omega, \mathcal{G}, \mathbb{P})$ and $\mathcal{F} \subset \mathcal{G}$ be a sub-σ-field such that F is $\mathcal{F}$-measurable. The two following statements are equivalent:*

i) F is ideal with respect to $\mathcal{F}$;

ii) F is completely calibrated with respect to $\mathcal{F}$.

The link between ideal forecast and complete calibration was partially found by Diebold et al. (1998) where the first implication $i) \Rightarrow ii)$ is proved under extra regularity condition (see the first Proposition of Part 3); see also (Gneiting and Ranjan, 2013, Theorem 2.9). The equivalence $i) \Leftrightarrow ii)$ is established in (Strähl and Ziegel, 2017, Proposition 2.11) in the framework of cross-calibration. We propose here a different proof based on Fubini theorem for the conditional distribution.
For future reference, we provide an extension of Proposition 2 to the framework of a sequential prediction space.

Corollary 3. *Let $(Y_n, F_n, V_n)_{n\in\mathbb{N}}$ be a sequential prediction space on $(\Omega, \mathcal{G}, \mathbb{P})$ with F_n adapted to the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$. For a lead time $T \geq 1$, we note $Z_n = Z_{F_n}^{Y_{n+T}}$ the PIT. The following properties are equivalent:*

i) $(F_n)_{n\in\mathbb{N}}$ *is ideal with respect to* $(\mathcal{F}_n)_{n\in\mathbb{N}}$*, i.e.,*

$$F_n = \mathcal{L}(Y_{n+T} \mid \mathcal{F}_n), \quad \textit{for all } n \in \mathbb{N};$$

ii) $(F_n)_{n\in\mathbb{N}}$ *is completely calibrated with respect to* $(\mathcal{F}_n)_{n\in\mathbb{N}}$*, i.e.*

$$Z_n \sim \mathrm{Unif}([0,1]) \textit{ and independent of } \mathcal{F}_n, \quad \textit{for all } n \in \mathbb{N}.$$

In this case, assuming that $(\mathcal{F}_n)_{n\in\mathbb{N}}$ *contains the filtration generated by* $(Y_n)_{n\in\mathbb{N}}$*, the sequence* $(Z_n)_{n\in\mathbb{N}}$ *is* $(T-1)$*-dependent. In particular, when* $T = 1$*, the sequence* $(Z_n)_{n\in\mathbb{N}}$ *is i.i.d. with uniform distribution on* $[0, 1]$*.*

The assumption that $(\mathcal{F}_n)_{n\in\mathbb{N}}$ contains the natural filtration generated by the sequence $(Y_n)_{n\in\mathbb{N}}$ is realistic in many applications. It means that, at time n, the past observations $Y_n, Y_{n-1}, \ldots$ are available to the forecaster in order to produce F_n.

Remark 4. When $T = 1$ and $(\mathcal{F}_n)_{n\in\mathbb{N}}$ is the natural filtration associated with $(Y_n)_{n\in\mathbb{N}}$, *i.e.* $\mathcal{F}_n = \sigma(Y_k, k \leq n)$, the implication $i) \Rightarrow ii)$ is provided in Diebold et al. (1998) under some stronger conditions on the model and it is also observed that $(Z_n)_{n\in\mathbb{N}}$ is i.i.d. uniformly distributed. In various places of the literature, one can find that $i)$ and $ii)$ are also equivalent to

iii) $(Z_n)_{n\in\mathbb{N}}$ is i.i.d. with uniform distribution on $[0, 1]$.

Clearly $ii)$ implies $iii)$ but it appears that the converse implication does not hold. We provide a counter-example in Section 5.1.2 (Remark 18) and show that, for the unfocused forecaster, the PIT are i.i.d. uniformly distributed even if the forecast is not ideal.

3 Empirical process and calibration

The previous section stated the importance of complete calibration to verify that a forecast is ideal. Our goal is now to proceed as in Bröcker (2022) to write the calibration assumption in terms of some empirical processes.

3.1 Assumptions and notation

We will slightly modify the dynamic framework and assume that the sequences are indexed in $\mathbb{Z}$. Let us recall that, Z_n is the PIT of Y_{n+T} by F_n. The following assumptions are made :

(A1) **Model :** the filtration $(\mathcal{F}_n)_{n\in\mathbb{Z}}$ is generated by a vectorial sequence $(X_n)_{n\in\mathbb{Z}}$, contains the filtration endowed by the quantity of interest $(Y_n)_{n\in\mathbb{Z}}$ and for $n \in \mathbb{Z}$

$$\mathcal{L}(Y_{n+T} \mid \mathcal{F}_n) = \mathcal{L}(Y_{n+T} \mid X_n);$$

(A2) **Stationarity :** the sequence $(Z_n, X_n)_{n\in\mathbb{Z}}$ is stationary;

(A3) **Dependence :** For two $\sigma-$algebras $\mathcal{A}$ and $\mathcal{B}$, the $\alpha-$mixing coefficient is defined by

$$\alpha(\mathcal{A},\mathcal{B}) = \sup_{A\in\mathcal{A},B\in\mathcal{B}} |\mathbb{P}(A\cap B) - \mathbb{P}(A)\mathbb{P}(B)|.$$

The $\sigma-$algebra are independent if and only if the coefficient is null, see Rosenblatt (1961) for more details on mixing dependent coefficients. We assume that there exists $\varepsilon > 0$, such that

$$\alpha(n) := \alpha(\sigma(\ldots, X_{-1}, X_0), \sigma(X_n, X_{n+1}, \ldots)) = O(n^{-2d-\varepsilon}).$$

The consequence of the Markovian Assumption (A1) will be discussed in Remark 7. This formulation of Assumption (A2) is mathematical. It is stated in this way to make fewer assumptions. In a less general framework, for example, if the triplet $(X_n, Y_{n+1}, F_n)_{n\in\mathbb{Z}}$ is stationary then Assumption (A2) is verified. This framework means that the predicted and observed phenomena are stationary as well as the way of forecasting.
We want to test the following null hypothesis

$$(H_0) : \text{the sequence } (F_n)_{n\in\mathbb{Z}} \text{ is ideally calibrated relatively to } (\mathcal{F}_n)_{n\in\mathbb{Z}}.$$

This means that for each $n \in \mathbb{Z}$, $F_n = \mathcal{L}(Y_{n+T} \mid \mathcal{F}_n)$. With Assumption (A1) and Corollary 3, this null hypothesis can be rewritten

$$(H_0) : \text{for each } n \in \mathbb{Z},\ Z_n \sim \text{Unif}([0,1]) \text{ and } Z_n \text{ is independent of } X_n.$$

We may assume without loss of generality that the stationary sequence $(X_n)_{n\in\mathbb{Z}}$ takes its values in $[0,1]^d$. The common CDF of the X_n is denoted by F. We consider the empirical process

$$\mathbb{G}^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{Z_i\le y, X_i\le t\}} - yF(t)\right), \quad y\in[0,1], x\in[0,1]^d,$$

and denote by Γ the limit covariance function

$$\Gamma((y,t),(y',t')) = \sum_{i\in\mathbb{Z}} \text{Cov}\left(\mathbb{1}_{\{Z_0\le y, X_0\le t\}}, \mathbb{1}_{\{Z_i\le y', X_i\le t'\}}\right).$$

Here the symbol $\le$ denotes componentwise comparison of vectors so that $x \le t$ means that $x_i \le t_i$ for all $i = 1,\ldots,d$. The negation $x \not\le t$ means that $x_i > t_i$ for some $i = 1,\ldots,d$. The following Proposition is a direct application of Theorem 10.2 in Dedecker et al. (2007).

Proposition 5. *Under Assumptions (A1)-(A3) and assuming the calibration null hypothesis (H_0), the empirical process converges in distribution*

$$\sqrt{n}\mathbb{G}^{(n)} \rightsquigarrow \mathbb{G} \quad \text{in } \ell^\infty\left([0,1]\times[0,1]^d\right)$$

and the limit $\mathbb{G}$ is a centered Gaussian process with covariance function Γ.

3.2 Decomposition of the empirical process

Let us consider the decomposition of the empirical process

$$\mathbb{G}^{(n)}(y,t) = \mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) + y\mathbb{F}_3^{(n)}(y,t)$$

with

$$\begin{aligned}
\mathbb{F}_1^{(n)}(y,t) &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{Z_i \le y\}} - y\right) F(t) \\
\mathbb{F}_2^{(n)}(y,t) &= \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{Z_i \le y\}} - y\right) \left(\mathbb{1}_{\{X_i \le t\}} - F(t)\right) \\
\mathbb{F}_3^{(n)}(y,t) &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}} - F(t)\,.
\end{aligned}$$

This decomposition is motivated by the following simple yet interesting lemma.

Lemma 6. *The following properties hold true:*

1. *the process $\mathbb{F}_1^{(n)}$ is centered if and only if $Z_i \sim \mathrm{Unif}([0,1])$ for all i;*
2. *the process $\mathbb{F}_2^{(n)}$ is centered if and only if Z_i and X_i are independent for all i;*
3. *the process $\mathbb{F}_1^{(n)} + \mathbb{F}_2^{(n)}$ is centered if and only if $Z_i \sim \mathrm{Unif}([0,1])$ and Z_i and X_i are independent for all i.*

As a consequence, the ideal calibration assumption (H_0) holds if and only if $\mathbb{F}_1^{(n)}$ and $\mathbb{F}_2^{(n)}$ are both centered processes.

A short interpretation of this lemma is that the first term tests probabilistic calibration and the second term tests the independence of the PIT with the information. The last term only uses the information $(X_n)_{n\in\mathbb{Z}}$. Thus in general, to test ideal calibration, one should essentially use the first two terms $\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}$ and not the last one $\mathbb{F}_3^{(n)}$. Indeed, the sum encodes ideal calibration and has the advantage of being observable since it does not depend on the unknown CDF F, for $y, t \in [0,1] \times [0,1]^d$,

$$\mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \left(\mathbb{1}_{\{Z_i \le y\}} - y\right) \mathbb{1}_{\{X_i \le t\}}.$$

Remark 7. Assumption (A1) on the Markovian character of the conditional distribution is used only to prove the last consequence of the previous lemma. Without it, the three points are still true.

We now investigate the asymptotic behavior of the empirical processes $(\mathbb{F}_i^{(n)})_{1\le i\le 3}$. We need to introduce these two limit covariances,

$$\Gamma_1((a,t),(b,s)) = \sum_{|i|\le T-1} F(t)F(s)\mathrm{Cov}\left(\mathbb{1}_{\{Z_0\le a\}}, \mathbb{1}_{\{Z_i \le b\}}\right),$$

$$\Gamma_2((a,t),(b,s)) = \sum_{|i|\leq T-1} \mathrm{Cov}\big(\big(\mathbb{1}_{\{Z_0\leq a\}} - a\big)\big(\mathbb{1}_{\{X_0\leq t\}} - F(t)\big), \\ \big(\mathbb{1}_{\{Z_i\leq b\}} - b\big)\big(\mathbb{1}_{\{X_i\leq s\}} - F(s)\big)\big).$$

The next theorem follows from Proposition 5 and the continuous mapping theorem.

Theorem 8. *Under Assumptions (A1)-(A3) and* (H_0)*,*

$$\sqrt{n}\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \rightsquigarrow (\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3),$$

where $\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3$ *are Gaussian processes, with covariances respectively given by* Γ_1 *and* Γ_2 *for the two first processes. In the specific case where* $T = 1$*, these two processes are independent.*

We remark that the covariance structure of $\mathbb{G}_1$ and $\mathbb{G}_2$ is simple because the covariance functions involve sums with $2T-1$ terms, which is a consequence of the $(T-1)$ dependence of the sequence of PITs $(Z_i)_{i\in\mathbb{Z}}$. In particular, when $T = 1$, Γ_1 and Γ_2 involve only a single term. For the sake of brevity, we do not provide the full expression for the covariance function of $(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3)$.
For the following, we need also

$$\bar{\mathbb{G}}^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i\leq y\}}\mathbb{1}_{\{X_i\not\leq t\}} - y\bar{F}(t), \text{ where } \bar{F}(t) = 1 - F(t).$$

Similarly as before, we decompose this process as

$$\overline{\mathbb{G}}^{(n)}(y,t) = \overline{\mathbb{F}}_1^{(n)}(y,t) + \overline{\mathbb{F}}_2^{(n)}(y,t) + y\overline{\mathbb{F}}_3^{(n)}(y,t).$$

Note that for $i = 1, 2$,

$$\overline{\mathbb{F}}_i^{(n)}(y,t) = \mathbb{F}_i^{(n)}(y,\mathbf{1}) - \mathbb{F}_i^{(n)}(y,t), \ \forall y,t \in [0,1]\times[0,1]^d. \tag{2}$$

We get also a convergence result for this other empirical process.

Corollary 9. *Under Assumptions (A1)-(A3) and* (H_0)*,*

$$\sqrt{n}\begin{pmatrix} \mathbb{F}_1^{(n)} & \overline{\mathbb{F}}_1^{(n)} \\ \mathbb{F}_2^{(n)} & \overline{\mathbb{F}}_2^{(n)} \\ \mathbb{F}_3^{(n)} & \overline{\mathbb{F}}_3^{(n)} \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G}_1 & \mathbb{G}(\cdot,\mathbf{1}) - \mathbb{G}_1 \\ \mathbb{G}_2 & -\mathbb{G}_2 \\ \mathbb{G}_3 & -\mathbb{G}_3 \end{pmatrix},$$

where $(\mathbb{G}_i)_{1\leq i\leq 3}$ *is the same as in Theorem 8.*

The tests we will propose in the next Section can be seen as functionals of these empirical processes and we will use the functional delta method to derive their asymptotic behaviour (van der Vaart and Wellner, 1996, Section 3.9).
As the ideal calibration is encoded by $(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)})$, it makes sense to ask that the asymptotics involve only the components $(\mathbb{G}_1, \mathbb{G}_2)$ as in the following theorem. Here $\ell^\infty = \ell^\infty([0,1]\times[0,1]^d)$.

Theorem 10. *Assume* $\Psi\colon \ell^\infty \times \ell^\infty \times \ell^\infty \to \ell^\infty$ *is differentiable at* $(0,0,0)$ *with*

$$\Psi(0,0,0) = 0 \text{ and } \partial_3\Psi(0,0,0) = 0.$$

Then under Assumptions (A1)-(A3) and the ideal calibration hypothesis (H_0),

$$\sqrt{n}\begin{pmatrix}\Psi\left(\mathbb{F}_1^{(n)},\mathbb{F}_2^{(n)},\mathbb{F}_3^{(n)}\right)\\ \Psi\left(\overline{\mathbb{F}}_1^{(n)},\overline{\mathbb{F}}_2^{(n)},\overline{\mathbb{F}}_3^{(n)}\right)\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathrm{d}_0\Psi(\mathbb{G}_1,\mathbb{G}_2,0)\\ \mathrm{d}_0\Psi(\mathbb{G}(\cdot,\mathbf{1})-\mathbb{G}_1,-\mathbb{G}_2,0)\end{pmatrix}.$$

It is too difficult to compute the exact law of these asymptotic empirical processes. Therefore we will approximate these limits by bootstrap. As already mentioned, the case $T = 1$ is the simplest and we will focus on this case.

3.3 Bootstrap for the lead time $T = 1$

To approximate the limiting distribution arising in Theorem 10, we will use a form of bootstrap where we replace the sequence of PITs $(Z_n)_{n\in\mathbb{Z}}$ by a new uniform sequence $(Z_n^\star)_{n\in\mathbb{Z}}$ which is i.i.d. and independent of the information $(X_n)_{n\in\mathbb{Z}}$. Note that it is not obvious that the new empirical process has the same limit, because the sequences $(X_n, Z_n)_{n\in\mathbb{Z}}$ and $(X_n, Z_n^\star)_{n\in\mathbb{Z}}$ do not have the same distributions under (H_0). Indeed, (H_0) implies that Z_n is independent of the past $(X_1,\dots,X_n)$ but in general, as Z_n depends on Y_{n+1}, there exists some dependency between Z_n and X_{n+1}, and more generally between Z_n and the future observations. We introduce the usual notation to denote the bootstrap process

$$\mathbb{G}^{(n)\star}(y,t) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{Z_i^\star\le y\}}\mathbb{1}_{\{X_i\le t\}} - yF(t),$$

and we construct the same process $(\mathbb{F}_i^{(n)\star}, \overline{\mathbb{F}}_i^{(n)\star})$ as previously, but with $Z_n^\star$ replacing Z_n.

Theorem 11. *Let* $(Z_n^\star)_{n\in\mathbb{Z}}$ *be an i.i.d. uniform random sequence, independent of* $(Z_i)_{i\in\mathbb{Z}}$ *and* $(X_i)_{i\in\mathbb{Z}}$. *Under the Assumptions (A1)-(A3) and the null hypothesis* (H_0),

$$\sqrt{n}\begin{pmatrix}\mathbb{F}_1^{(n)} & \overline{\mathbb{F}}_1^{(n)}\\ \mathbb{F}_2^{(n)} & \overline{\mathbb{F}}_2^{(n)}\\ \mathbb{F}_1^{(n)\star} & \overline{\mathbb{F}}_1^{(n)\star}\\ \mathbb{F}_2^{(n)\star} & \overline{\mathbb{F}}_2^{(n)\star}\\ \mathbb{F}_3^{(n)}-F & \overline{\mathbb{F}}_3^{(n)}-\overline{F}\end{pmatrix} \rightsquigarrow \begin{pmatrix}\mathbb{G}_1 & \mathbb{G}(\cdot,\mathbf{1})-\mathbb{G}_1\\ \mathbb{G}_2 & -\mathbb{G}_2\\ \mathbb{G}_1^\star & \mathbb{G}^\star(\cdot,\mathbf{1})-\mathbb{G}_1^\star\\ \mathbb{G}_2^\star & -\mathbb{G}_2^\star\\ \mathbb{G}_3 & -\mathbb{G}_3\end{pmatrix},$$

where

$$\begin{pmatrix}\mathbb{G}_1 & \mathbb{G}(\cdot,\mathbf{1})-\mathbb{G}_1\\ \mathbb{G}_2 & -\mathbb{G}_2\end{pmatrix} \overset{law}{=} \begin{pmatrix}\mathbb{G}_1^\star & \mathbb{G}^\star(\cdot,\mathbf{1})-\mathbb{G}_1^\star\\ \mathbb{G}_2^\star & -\mathbb{G}_2^\star\end{pmatrix},$$

and these two vector processes are independent.

Remark 12. The term bootstrap is a slight abuses of language. In fact, resampling is done with a new sample.

This result states that the bootstrapped process has the same asymptotic behavior even if the two sequences $(X_n, Z_n)_{n\in\mathbb{Z}}$ and $(X_n, Z_n^\star)_{n\in\mathbb{Z}}$ do not have the same distribution. Then thanks to Theorem 10 and Theorem 11, we get the next result that justifies the use of bootstrap to adjust our tests. As the distributions of $(\mathbb{G}_1,\mathbb{G}_2,\mathbb{G}_3)$ and $(\mathbb{G}_1^\star,\mathbb{G}_2^\star,\mathbb{G}_3)$ are not the same, it is important to assume that the functional Ψ satisfies $\partial_3\Psi(0,0,0) = 0$.

Corollary 13. *Let $(Z_n^\star)_{n\in\mathbb{Z}}$ be an i.i.d. uniform random sequence, independent of $(Z_n)_{n\in\mathbb{Z}}$ and $(X_n)_{n\in\mathbb{Z}}$. Assume $\Psi\colon \ell^\infty \times \ell^\infty \times \ell^\infty \to \ell^\infty$ is differentiable at $(0,0,0)$ with*

$$\Psi(0,0,0) = 0 \text{ and } \partial_3\Psi(0,0,0) = 0,$$

then under Assumptions (A1)-(A3) and (H_0),

$$\sqrt{n}\begin{pmatrix} \Psi\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \\ \Psi\left(\mathbb{F}_1^{(n)\star}, \mathbb{F}_2^{(n)\star}, \mathbb{F}_3^{(n)}\right) \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, 0) \\ \mathrm{d}_0\Psi(\mathbb{G}_1^\star, \mathbb{G}_2^\star, 0) \end{pmatrix}.$$

Moreover, the two limit processes have the same distribution and are independent.

In practice, the calibration of the test uses several independent sequences $(Z_n^\star)_{n\in\mathbb{Z}}$ to obtain a sample approximation of the limit distribution.

4 Testing for ideal calibration

4.1 Heuristic and strategy

The main idea driving our tests for ideal calibration relies on Corollary 3, stating that the forecast $(F_n)_{n\in\mathbb{Z}}$ is ideally calibrated if and only if

$$(H_0) : \text{for each } n \in \mathbb{Z},\ Z_n \sim \mathrm{Unif}([0,1]) \text{ and } Z_n \text{ is independent of } X_n.$$

We recall that this characterization was obtained thanks to the Markov assumption (A1), which implies that Z_n is independent from $\mathcal{F}_n$ if and only if it is independent from X_n. Note that even if the Markov assumption does not hold, our tests can still be used with a controlled level but will detect only if Z_n depends on X_n and will not be able to detect more subtle forms of non-calibration. However, it is always possible in theory to augment the dimension of the covariate space and test the dependency of Z_n from $(X_n, \dots, X_{n-m+1})$, that is consider memory of length $m \geq 1$. In practice, augmenting the dimension of the covariate space has a cost, both in terms of computational time and loss of power.

The proposed methodology for testing (H_0) relies on the simple observation that Z_n is uniformly distributed on $[0,1]$ and independent of X_n if and only if

$$\mathbb{E}\big[g(Z_n) \mid X_n \in A\big] = 0,$$

for all functions $g : [0,1] \to \mathbb{R}$ such that $\int_0^1 g(y)\ \mathrm{d}y = 0$ and measurable sets $A \subset [0,1]^p$. In practice, this integral can be estimated by

$$\frac{\frac{1}{n}\sum_{i=1}^n g(Z_i)\mathbb{1}_{\{X_i\in A\}}}{\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\in A\}}} \tag{3}$$

which must be approximately zero under (H_0). In order to detect default of calibration, one needs to find a test function g and a region A where this mean significantly deviates from 0. We will consider several natural choices for the choice of g that are related to cumulative distribution functions, moments or histograms. The search for the region A is guided by the CART algorithm as explained in the next section.

4.2 General approach for tree based tests

The CART algorithm (Breiman et al., 1984) is a popular method from statistics and machine learning used for prediction, both in classification and regression. It produces simple predictors, called trees, in the sense that they can be represented graphically by a decision tree and predict finitely many different values on finitely many different subgroups of the population. More precisely, the procedure constructs a partition of the feature space $[0,1]^d$ into regions $A_1, \dots, A_K$, called leaves, and the tree function $T\colon [0,1]^d \to \mathbb{R}$ is constant on each leaf. In regression, the predicted value on A_k is simply the sample mean of the response variable in the sub-sample of individuals with features in A_k.

For our purpose, we will use the CART algorithm to predict the PIT Z_n, or rather a transformation of it $g(Z_n)$, as a function of the covariate X_n. Under (H_0), the covariate is uninformative to predict $g(Z_n)$ but, under the alternative, the CART algorithm should detect the dependency between $g(Z_n)$ and X_n. Interestingly, the CART algorithm is completely non parametric and assumes no model between $g(Z_n)$ and X_n. Furthermore, it is known to be quite robust in high dimension, i.e. its power is not too much hindered by the curse of dimensionality.

We next briefly describe the construction of the tree in our setting. Let $(X_i, g(Z_i))_{1\le i\le n} \in [0,1]^d \times \mathbb{R}$ be the sample data. The construction of the partition $A_1, \dots, A_K$ of $[0,1]^d$ relies on recursive binary splitting, where a splitting rule is used repeatedly to form the partition. The first split forms the partition $[0,1]^d = A_1 \cup A_2$ in order to minimize the mean square error

$$\sum_{X_i \in A_1} \Big(g(Z_i) - \overline{g(A_1)}\Big)^2 + \sum_{X_i \in A_2} \Big(g(Z_i) - \overline{g(A_2)}\Big)^2, \tag{4}$$

where $\overline{g(A)}$ is the mean of the transformed PITs $g(Z_i)$ for $X_i \in A$. Not all possible partitions are used, but only the so-called admissible ones. An admissible partition is obtained by choosing a covariate index $j \in \{1, \dots, d\}$ and a threshold $u \in (0,1)$ and by letting

$$A_1 = \{x \in [0,1]^d \colon x_j \le u\} \quad \text{and} \quad A_2 = \{x \in [0,1]^d \colon x_j > u\}.$$

Finding the admissible partitions that minimize the mean square error (4) can be done very efficiently, see Breiman et al. (1984) for more details. Using this splitting rule recursively on both A_1 and A_2 we then obtain 4 leaves, renamed $A_1, \dots, A_4$. Repeating the procedure d times, we obtain the tree with depth d with 2^d leaves. For our purpose, we will consider shallow trees with depth $d = 1, 2, 3$ only.

As a simple illustration, Figure 1 represents a data set in dimension $d = 2$ and the associated partition. The points $X_i \in [0,1]^2$ represent the covariates and the transformed PITs $g(Z_i)$ are represented by the color of the points. We can see a strong dependence since X_i and $g(Z_i)$ tend to be large at the same time. The CART algorithm detects this dependence and produces regions $A_1, \dots, A_5$ which are quite homogeneous.
König-Huygens Formula implies this rewriting of the splitting criterion (4)

$$\sum_{i=1}^{n} g(Z_i)^2 - \sum_{X_i \in A_1} \overline{g(A_1)}^2 - \sum_{X_i \in A_2}^{n} \overline{g(A_2)}^2. \tag{5}$$

So minimizing this variance quantity is equivalent to maximizing the following quantity

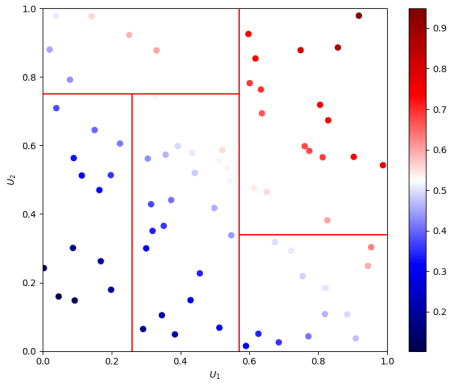

Figure 1: Created region by CART algorithm for the dummy case

$$\sum_{X_i \in A_1} \overline{g(A_1)}^2 + \sum_{X_i \in A_2}^{n} \overline{g(A_2)}^2. \tag{6}$$

In its simplest version, our proposed test for ideal calibration relies on the choice of a test function $g : [0,1] \to \mathbb{R}$ such that $\int_0^1 g(u)\mathrm{d}u = 0$ and on the choice of a depth $d \geq 1$ (typically $d = 1, 2, 3$). The test statistic takes the form

$$\Delta = \sum_{k=1}^{2^d} \sum_{X_i \in A_k} \overline{g(A_k)}^2,$$

where the partition $(A_k)_{1 \leq k \leq 2^d}$ is the one associated to the tree with depth d. Under the null hypothesis (H_0), this statistic must be close to zero. To adjust the test, we use the bootstrap (see Theorem 11) and we compute the mean square error $\Delta^\star$ obtained when fitting a regression tree with depth d to the sample $(X_i, g(Z_i^\star))_{1 \leq i \leq n}$, where $(Z_i^\star)_{1 \leq i \leq n}$ denotes an i.i.d. sample with uniform distribution on $[0,1]$ (independent of everything else). This statistic has a more explicit form coming from an optimization problem

$$\Delta = \max_{\substack{\text{admissible regions} \\ (B_1, \ldots, B_{2^d})}} \sum_{k=1}^{2^d} \sum_{X_i \in B_k} \overline{g(B_k)}^2. \tag{7}$$

4.3 Specification of test functions

In practice, using only one test function seems limited and several of them should be used. We discuss different natural choices and also how to combine different test functions in our approach.

4.3.1 Test 1: cumulative distribution function

The first test focuses on the CDF of the uniform distribution and considers the family of test functions

$$g_p(u) = \mathbb{1}_{\{u \leq p\}} - p, \quad p \in (0,1).$$

This approach is closely related to the test based on Conditional Exceedance Probability (CEP) by Strähl and Ziegel (2017), see also section 5.1.1 for more details on this test. It is natural choice because, since the CDF characterize the PIT distribution, ideal calibration is equivalent to the fact that

$$\mathbb{E}[g_p(Z_n) \mid X_n \in A] = 0, \quad \text{for all } p \in (0,1) \text{ and } A \subset [0,1]^d \ .$$

In practice we consider finitely many values $p_1 < \ldots < p_K$, (for instance $p_1 = 0.1, \ldots, p_9 = 0.9$ in the simulation study) and fit K trees with depth d. The k-th tree uses the sample $(X_i, g_{p_k}(Z_i))_{1 \le i \le n}$ and the corresponding mean squared error is noted Δ_k. Similarly, for a bootstrap sample $(Z_i^\star)_{1 \le i \le n}$ we obtain the mean squared errors $\Delta_k^\star$, $1 \le k \le K$. We use here aggregation of the K errors, that is the test statistic is $\Delta = \sum_{i=1}^k \Delta_k$ that we compare with the bootstrap distribution of $\Delta = \sum_{i=1}^k \Delta_k^\star$.

4.3.2 Test 2: moments

The second test focuses on the moments of the uniform distribution and more precisely on the first four moments. In order to check whether the PITs are uniformly distributed, we want to verify that mean, variance, skewness and kurtosis match those of the uniform distribution. In a slightly different context of calibration of ensemble forecast, this approach was used by Jolliffe and Primo (2008).
Here we consider the orthogonal polynomials

$$g_0(u) = 1, \quad g_1(u) = u - \frac{1}{2}, \quad g_2(u) = \sqrt{12}\left(u - \frac{1}{2}\right)^2, \ldots$$

that are obtained by the Gram-Schmidt orthonormalisation procedure applied to family of polynomials $(u^k)_{0 \le k \le 4}$ in the Hilbert space $L^2([0,1])$. As we will see in Proposition 16, orthogonality offers the benefit to yield asymptotically independent tests.
Thanks to this asymptotic independence, we do not aggregate the four mean square errors but rather perform four independent tests with test function $g_1, \ldots, g_4$ respectively. This strategy here seems interesting because it offers some qualitative interpretation: if a deviation to uniformity is detected, we are able to see whether it is rather in mean, variance, skewness or kurtosis.

4.3.3 Test 3: histogram and χ^2-test

The third test is related to the histogram and χ^2-test and is slightly different from the previous ones as we will see that it can be related to classification trees rather than regression trees.
For $L \ge 2$, we consider the histogram of the PITs $(Z_n)_{n \ge 1}$ based on L bins of equal size $[0, 1/L)$, ..., $[(L-1)/L, 2]$. Here we introduce the centered *vectorial* test function

$$g(u) = \left(\mathbb{1}_{[\frac{(l-1)}{L}, \frac{l}{L})}(u) - \frac{1}{L} \right)_{1 \le l \le L} .$$

On a region $A \subset [0,1]^2$, the squared (euclidean) norm

$$\|\overline{g(A)}\|^2 = \sum_{l=1}^{L} \left(\frac{\sum_{i=1}^n \mathbb{1}_{\{X_i \in A, Z_i \in [\frac{(l-1)}{L}, \frac{l}{L})\}}}{\sum_{i=1}^n \mathbb{1}_{\{X_i \in A\}}} - \frac{1}{L} \right)^2$$

corresponds, up to a multiplicative constant, to the χ^2-distance obtained when testing the uniformity of the PIT in A with the χ^2 test with L classes. This quantity is related to the Gini criterion

$$G(A) = \sum_{l=1}^{L} \left(\frac{\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A, \lceil LZ_i \rceil = l\}}}{\sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}} \right)^2$$

by the relation $\|\overline{g(A)}\|^2 = G(A) - 1/L$. Here the L classes used for the computation of the Gini criterion are the bins number $l = 1, \dots L$ and the class associated with a PIT Z_i is $\lceil LZ_i \rceil$. Hence the l-th term of the sum defining $G(A)$ corresponds to the estimated probability of the l-th class in A.

Interestingly, the Gini criterion is one of the homogeneity criteria used in the construction of classification trees. Their construction is similar as the one of regression trees and is again based on recursive binary splitting. But now a split consists in the search of the admissible partition $A_1 \cup A_2$ that maximizes the Gini criterion $G(A_1) + G(A_2)$ (instead of minimisation of the mean squared error in regression).

The methodology for our third test is the following. We fit a classification tree with depth d on the sample $(X_i, \lceil LZ_i \rceil)$ using the Gini criterion. Denoting by $(A_k)_{1 \le k \le 2^d}$ the resulting partition, the test statistic is

$$\Delta = \sum_{k=1}^{2^d} G(A_k).$$

Up to constants, this is the sum, over the different leaves, of χ^2-distances obtained when testing the uniformity of the PIT in each leaf. Adjustment of the test is again based on a bootstrap replication $\Delta^\star$ of the test statistic using the bootstrap sample $(X_i, \lceil LZ_i^\star \rceil)$.

4.4 Statistics Δ and empirical processes

In this section, we see how the CART algorithm, and especially the splitting criterion (4) and the statistic Δ, can be rewritten in terms of the empirical processes introduced in Section 3. Let us recall the form of Δ in term of optimization problem

$$\Delta = \max_{\substack{\text{admissible regions} \\ (B_1, \dots, B_{2^d})}} \sum_{k=1}^{2^d} \sum_{X_i \in B_k} \overline{g(B_k)}^2.$$

We give a proof only for the first split. For this split, the shape of the region B_1, B_2 is $\{x \in [0,1]^d \mid x \le t\}, \{x \in [0,1]^d \mid x \not\le t\}$ for an admissible $t \in [0,1]^d$, then

$$\Delta = \max_{\text{admissible } t \in [0,1]^d} \sum_{X_i \le t} \overline{g(\{x \le t\})}^2 + \sum_{X_i \not\le t} \overline{g(\{x \not\le t\})}^2 \tag{8}$$

These two terms can be rewritten in the following way to make the empirical processes appear more naturally,

$$\sum_{i=1}^{n} \mathbb{1}_{\{X_i \le t\}} \left(\frac{\sum_{k=1}^{n} g(Z_k) \mathbb{1}_{\{X_k \le t\}}}{\sum_{k=1}^{n} \mathbb{1}_{\{X_k \le t\}}} \right)^2 + \sum_{i=1}^{n} \mathbb{1}_{\{X_i \not\le t\}} \left(\frac{\sum_{k=1}^{n} g(Z_k) \mathbb{1}_{\{X_k \not\le t\}}}{\sum_{k=1}^{n} \mathbb{1}_{\{X_k \not\le\}}} \right)^2,$$

or in the same way

$$\left(\sqrt{n}\frac{\frac{1}{n}\sum_{k=1}^n g(Z_k)\mathbb{1}_{\{X_k\le t\}}}{\sqrt{\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\{X_k\le t\}}}}\right)^2+\left(\sqrt{n}\frac{\frac{1}{n}\sum_{k=1}^n g(Z_k)\mathbb{1}_{\{X_k\not\le t\}}}{\sqrt{\frac{1}{n}\sum_{k=1}^n \mathbb{1}_{\{X_k\not\le t\}}}}\right)^2. \tag{9}$$

These are the sums that can be interpreted as functionals of our processes. Indeed, let us introduce the following integral notation. Let g be a piecewise continuously differentiable function, it can be decomposed as

$$g(y)=g_0(y)+\sum_{i=1}^k w_i\mathbb{1}_{\{y\le\alpha_i\}},$$

where $g_0\in\mathcal{C}^1$ and $w_i,\alpha_i\in\mathbb{R}$. For $f\in\ell^\infty$, we define its integral with respecto to $\mathrm{d}g(y)$ as

$$\int_{\mathbb{R}} f(y)\ \mathrm{d}g(y):=\int_{\mathbb{R}} f(y)g_0'(y)\ \mathrm{d}y-\sum_{i=1}^k w_i f(\alpha_i).$$

This integral is only a notation. It is not directly related to the Stieltjes integral because our definition allows to consider functions f,g with common points of discontinuities. The minus sign in the notation is natural because the function $\mathbb{1}_{\{\cdot\le\alpha\}}$ has a negative jump. The following Lemma is useful. It is a kind of integration by parts.

Lemma 14. *Assume $g:[0,1]\to\mathbb{R}$ piecewise continuously differentiable. Then for each $t\in[0,1]^d$,*

$$\frac{1}{n}\sum_{i=1}^n\big(g(Z_i)-\int_0^1 g(u)\ \mathrm{d}u\big)\mathbb{1}_{\{X_i\le t\}}=-\int_0^1\Big(\mathbb{F}_1^{(n)}(y,t)+\mathbb{F}_2^{(n)}(y,t)\Big)\ \mathrm{d}g(y).$$

Similarly

$$\frac{1}{n}\sum_{i=1}^n\big(g(Z_i)-\int_0^1 g(u)\ \mathrm{d}u\big)\mathbb{1}_{\{X_i\not\le t\}}=-\int_0^1\Big(\overline{\mathbb{F}}_1^{(n)}(y,t)+\overline{\mathbb{F}}_2^{(n)}(y,t)\Big)\ \mathrm{d}g(y).$$

This lemma links quantity (9) with empirical processes because $\int_0^1 g(u)\ \mathrm{d}u=0$, because the quantity (9) can be rewritten

$$\left(\sqrt{n}\frac{\int_0^1\mathbb{F}_1^{(n)}(y,t)+\mathbb{F}_2^{(n)}(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t)+\mathbb{F}_3^{(n)}(1,t)}}\right)^2+\left(\sqrt{n}\frac{\int_0^1\overline{\mathbb{F}}_1^{(n)}(y,t)+\overline{\mathbb{F}}_2^{(n)}(y,t)\ \mathrm{d}g(y)}{\sqrt{\bar{F}(t)+\overline{\mathbb{F}}_3^{(n)}(1,t)}}\right)^2 \tag{10}$$

This invites us to define the functional

$$\Psi_g^F(F_1,F_2,F_3;t)=\frac{\int_0^1\big(F_1(y,t)+F_2(y,t)\big)\,\mathrm{d}g(y)}{\sqrt{F(t)+F_3(1,t)}}. \tag{11}$$

By combining (8) and (10),

$$\Delta = n \max_{\text{admissible } t \in C_\varepsilon} \Psi_g^F\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}; t\right)^2 + \Psi_g^{\overline{F}}\left(\overline{\mathbb{F}}_1^{(n)}, \overline{\mathbb{F}}_2^{(n)}, \overline{\mathbb{F}}_3^{(n)}; t\right)^2. \tag{12}$$

We had already noticed that the sum $\mathbb{F}_1^{(n)} + \mathbb{F}_2^{(n)}$ was observable because it did not depend on the CDF F. We can make the same remark about this quantity. Of course, the process $\mathbb{F}_3^{(n)}$ and the application Ψ_g^F depend on F separately but this dependence is simplified in the denominator. To use the delta method on Ψ_g^F, the condition of differentiability must be verified. This is only possible on a restriction of this set because the square root is not derivable in 0, so we need to avoid regions where F or $\overline{F}$ is too close to 0. For $\varepsilon > 0$, Let us consider the restriction on $\ell^\infty(C_\varepsilon)$ with $C_\varepsilon \subset [0,1]^d$ and

$$\forall t \in C_\varepsilon,\ \varepsilon \leq F(t) \leq 1 - \varepsilon. \tag{13}$$

Proposition 15. *Assume $g : [0,1] \to \mathbb{R}$ piecewise continuously differentiable. Let $\varepsilon \in (0,1)$, the functionals Ψ_g^F and $\Psi_g^{\overline{F}}$ are differentiable at $(0,0,0)$ when the variable t is restricted to a subset $C_\varepsilon \subset [0,1]^d$ satisfying (13) and*

$$\partial_3 \Psi_g^F(0,0,0) = \partial_3 \Psi_g^{\overline{F}}(0,0,0) = 0.$$

It remains to be seen that the application max is continuous in ℓ^∞. Then Δ converges to a certain distribution by the contiunous mapping theorem.

The Corollary 13 states that $\Delta^\star$ approximates the asymptotic distribution of Δ. Sometimes, the ideal calibration of a forecaster $(F_n)_{n\in\mathbb{Z}}$ will be tested with several functions g. When these functions are pairwise orthogonal, see Equation (14), the aggregation of the multiple tests will be exact.

Proposition 16. *Assume $g, f : [0,1] \to \mathbb{R}$ **centred** piecewise continuously differentiable. Let $\varepsilon \in (0,1)$, under Assumptions (A1)-(A3) and (H_0), if*

$$\int_0^1 f(u)g(u)\ \mathrm{d}u = 0, \tag{14}$$

then $\sqrt{n}\begin{pmatrix} \Psi_g^F\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \\ \Psi_g^{\overline{F}}\left(\overline{\mathbb{F}}_1^{(n)}, \overline{\mathbb{F}}_2^{(n)}, \overline{\mathbb{F}}_3^{(n)}\right) \end{pmatrix}$ and $\sqrt{n}\begin{pmatrix} \Psi_f^F\left(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)}, \mathbb{F}_3^{(n)}\right) \\ \Psi_f^{\overline{F}}\left(\overline{\mathbb{F}}_1^{(n)}, \overline{\mathbb{F}}_2^{(n)}, \overline{\mathbb{F}}_3^{(n)}\right) \end{pmatrix}$ are asymptotically independent in $\ell^\infty(C_\varepsilon)$.

5 Numerical illustrations

The aim of this section is to illustrate the performances of the three tests defined in Section 4.3. We first propose a simulation study and then an application on real data related to weather forecast.

5.1 Simulation study

5.1.1 Competing testing procedures

Regression and classification tree. We consider the three tests introduced in Section 4.3. A brief preliminary exploration led us to limit the depth of the trees to two splits.

The other arbitrarily fixed parameters are the partition of the interval [0,1] of Test 1, the number of polynomials of Test 2 and the number of classes of Test 3. For the first test, we choose the partition into ten classes $((i-1)/10, i)/10)$, $i = 1, \dots, 10$. For the second test, polynomials up to degree 4 are used. For the third test, 7 classes are considered. Adjustment of the tests is done through $B = 600$ bootstrap replications. In the rest of the Section, these tests will be respectively denoted T_1, T_2 and T_3.

Conditional Exceedence Probability : The testing procedure by Strähl and Ziegel's (Strähl and Ziegel, 2017) consists in writing the characterization of the ideal calibration, namely that the sequence of PIT $(Z_n)_{n\in\mathbb{N}}$ are uniformly distributed and independent of the information $(X_n)_{n\in\mathbb{N}}$, in terms of logistic regression. In our framework, this translates into

$$\text{logit}(\mathbb{P}(Z_n \leq z \mid X_n)) = \beta_{0,z} + \sum_{i=1}^{d} \beta_{i,z} X_n^{(i)}, \quad n \geq 0, \tag{15}$$

with $z \in (0,1)$, $\text{logit}(z) = \log(z/(1-z))$ and $X_n^{(i)}$ the i^{th} coordinate of X_n. The null hypothesis (H_0) corresponds to

$$\beta_{0,z} = \text{logit}(z) \text{ and } \beta_{i,z} = 0 \text{ for each } z \in (0,1) \text{ and } i \in \{1, \dots, d\}. \tag{16}$$

The logistic regression is run for different values of z to test if Equation (16) is satisfied. To combine the different tests, multiple test adjustment are used (Cox and Lee, 2008). For more details, see Strähl and Ziegel (2017, Section 6.1) . In the following, this test will be referred as CEP.

Remark 17. Other tests related to this problem have been introduced in the literature. For example, in Strähl and Ziegel (2017) or Held et al. (2010), the authors work on a gaussian scale and apply a linear model to detect a dependence between the transformed PIT and the covariates. We decide to focus here on nonparametric models and therefore do not include these latter tests in the benchmark. Besides, one could also think of considering the weaker hypothesis of probabilistic calibration (see Berkowitz (2001)). However, most of these procedures test the i.i.d. of PITs, which is not equivalent to complete calibration, see Remark 18. In our context, we consider serial dependence so that these tests might lead to rejection most of the time for "bad reasons".

We propose two deta generating processes to test the performances of the competitors described in the previous section. The first one is the simple autoregressive model $AR(1)$. The second one is a specifically designed to challenge the CEP test.

5.1.2 Numerical experiment in the autoregressive model

Data generating process. For $\rho \in (-1,1), \alpha^2 > 0$, define

$$Y_0 \sim \mathcal{N}(0, \alpha^2), \quad Y_{n+1} = \rho Y_n + \varepsilon_n, \quad n \geq 0,$$

where $(\varepsilon_n)_{n\in\mathbb{N}}$ is an i.i.d. sequence with gaussian distribution $\mathcal{N}(0, \sigma^2)$, $\sigma^2 > 0$. The initial variance α^2 is chosen such that the sequence $(Y_n)_{n\in\mathbb{N}}$ is stationary, i.e. $\alpha^2 = \sigma^2/(1-\rho^2)$. The correlation ρ represents the strength of the dependence across time. For $\rho = 0$, the sequence is i.i.d. We assume that the covariate and observations are equal, i.e. $X_n = Y_n$

for all $n \geq 0$. The known information is thus $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$. The autoregressive equation implies that, at time $n \geq 0$, the ideal forecast F_n^* is $\mathcal{N}(\rho X_n, \sigma^2)$.

Different forecasters. In this autoregressive model, four different forecasts are considered for comparison, denoted respectively by $F_n^{(1)}, \ldots, F_n^{(4)}$. Note that the three first alternatives have been considered in Gneiting and Ranjan (2013); Strähl and Ziegel (2017):

- Climatological forecaster: $F_n^{(1)} = \mathcal{N}(0, \alpha^2)$;
- Unfocused forecaster: $F_n^{(2)} = \frac{1}{2}\mathcal{N}(\rho X_n, \sigma^2) + \frac{1}{2}\mathcal{N}(\rho X_n + \tau_n, \sigma^2)$ with $\tau_n = \pm 1$ with probability $1/2$ independently of $(X_n, \varepsilon_n)_{n\in\mathbb{N}}$;
- Sign-reversed forecaster: $F_n^{(3)} = \mathcal{N}(-\rho X_n, \sigma^2)$;
- Corrupted observation forecaster: $F_n^{(4)} = \mathcal{N}(\rho(X_n + \delta_n), \sigma^2)$, where $\delta_n \sim \mathcal{N}(0, 1)$ independently of $(X_n, \varepsilon_n)_{n\in\mathbb{N}}$.

Remark 18. For the first two forecasters, probabilistic calibration holds, meaning that the associated PITs are uniformly distributed (Gneiting and Ranjan, 2013). The unfocused forecaster also has the interesting additional property that the PITs are i.i.d. Indeed, for $n \in \mathbb{N}$, and $\sigma^2 = 1$,

$$Z_n = \frac{1}{2}\Phi(Y_{n+1} - \rho Y_n) + \frac{1}{2}\Phi(Y_{n+1} - \rho Y_n - \tau_n),$$

where Φ is the CDF of standard Gaussian distribution, so that $Z_n = \frac{1}{2}\Phi(\varepsilon_n) + \frac{1}{2}\Phi(\varepsilon_n - \tau_n)$, implying the indepndance of the PITs. Note that the PITs of the unfocused forecaster are i.i.d. uniformly distributed without ideal calibration because the PIT depends on $\mathcal{F}_n$.

Results. Table 1 summarizes the results obtained for the four tests under the different alternatives. The empirical powers reported therein are calculated from 1000 replications, with a test level chosen at $\alpha = 0.05$ for the four tests. The variance parameter σ^2 is 1 for all simulations. All the test parameters (depth, number of classes, level, ...) are fixed as detailed in Section 5.1.1.

Climatological Forecaster $F_n^{(1)}$

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	0.08	0.07	0.06	0.07
$(0.3, 50)$	**0.35**	0.23	0.17	0.27
$(0.5, 50)$	**0.81**	0.63	0.48	0.75
$(0.8, 50)$	**0.99**	**0.97**	**0.97**	**0.99**
$(0.1, 100)$	0.10	0.07	0.06	0.07
$(0.3, 100)$	**0.62**	0.46	0.31	0.54
$(0.5, 100)$	**0.97**	0.91	0.78	**0.98**
$(0.8, 100)$	1	1	1	1
$(0.1, 500)$	0.39	0.24	0.15	0.32
$(0.3, 500)$	1	1	1	1
$(0.5, 500)$	1	1	1	1
$(0.8, 500)$	1	1	1	1

Unfocused Forecaster $F_n^{(2)}$

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	0.04	0.06	0.04	0.03
$(0.3, 50)$	0.05	0.05	0.05	0.03
$(0.5, 50)$	0.05	0.06	0.05	0.05
$(0.8, 50)$	0.06	0.05	0.05	0.05
$(0.1, 100)$	0.04	0.06	0.04	0.03
$(0.3, 100)$	0.05	0.05	0.05	0.03
$(0.5, 100)$	0.05	0.06	0.05	0.05
$(0.8, 100)$	0.06	0.06	0.05	0.05
$(0.1, 500)$	0.06	0.06	0.05	0.04
$(0.3, 500)$	0.07	0.06	0.06	0.06
$(0.5, 500)$	0.07	0.05	0.06	0.06
$(0.8, 500)$	0.05	0.07	0.05	0.06

Sign-reversed Forecaster $F_n^{(3)}$

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	**0.17**	0.13	0.10	0.14
$(0.3, 50)$	**0.92**	0.81	0.63	0.87
$(0.5, 50)$	1	1	1	1
$(0.8, 50)$	1	1	1	1
$(0.1, 100)$	**0.32**	0.19	0.14	0.25
$(0.3, 100)$	**0.99**	**0.98**	0.90	**0.99**
$(0.5, 100)$	1	1	1	1
$(0.8, 100)$	1	1	1	1
$(0.1, 500)$	0.94	0.85	0.66	0.92
$(0.3, 500)$	1	1	1	1
$(0.5, 500)$	1	1	1	1
$(0.8, 500)$	1	1	1	1

Corrupted observation Forecaster $F_n^{(4)}$

(ρ, N)	T1	T2	T3	CEP
$(0.1, 50)$	0.05	0.05	0.05	0.05
$(0.3, 50)$	0.06	0.08	0.05	0.05
$(0.5, 50)$	0.08	**0.16**	0.09	0.10
$(0.8, 50)$	0.16	**0.46**	0.17	0.23
$(0.1, 100)$	0.05	0.05	0.05	0.05
$(0.3, 100)$	0.06	0.08	0.05	0.05
$(0.5, 100)$	0.08	**0.19**	0.09	0.10
$(0.8, 100)$	0.20	**0.65**	0.31	0.36
$(0.1, 500)$	0.05	0.04	0.05	0.05
$(0.3, 500)$	0.06	0.10	0.09	0.08
$(0.5, 500)$	0.15	**0.61**	0.35	0.34
$(0.8, 500)$	0.80	**1**	**0.98**	**0.99**

Table 1: *Empirical power of the four competing tests for Alternatives 1 to 4 with different numbers of realizations N and values of parameter ρ, based on 1000 replications, with theoretical level test 0.05. The closer this rate is to 1, the better the test performs. Boldface highlights the best results obtained.*

Several comments can be formulated on the basis of Table 1:

- As expected, the power of the tests increases with the sample size.
- Alternative 2 fails to be detected by any test. This fact has already been noted in Strähl and Ziegel (2017). This is still very surprising as the PITs are only uniformly distributed and independent but not of the information (see Remark 18).
- Tests based on regression trees (T_1, T_2, T_3) perform better or are at least comparably to CEP test. Unfortunately, no single test is best in all cases.
- Due to the shape of the alternatives, the greater the time dependency, the more powerful the tests are.

5.1.3 Non linear model

Data generating process. Let us now consider a framework that might be specifically challenging for the CEP test. We consider an i.i.d. sequence $(\mu_n, \varepsilon_n, \delta_n)_{n\in\mathbb{N}}$ the marginals are also independent and $\varepsilon_n \sim \text{Unif}(\{-1, 1\})$, $\mu_n \sim \mathcal{N}(0, 1)$ and $\delta_n \sim \mathcal{N}(0, 1)$. We define the sequence

$$Y_{n+1} = \varepsilon_n \mu_n + \delta_n, \quad n \geq 1.$$

The covariate used for prediction is $X_n = \mu_n$? Note that that the laws of Y_{n+1} given $X_n = \pm x$ are equal. This symetry suggests that the coefficent associated to X_n in the logistic regression will vanish so that the dependence will not be detected by the CEP.

Data generating process. Now consider the Climatological Forecaster $F = \mathcal{N}(0, 2)$. Recall that this forecaster is not ideally calibrated for the information $(\mathcal{F}_n)_{n\in\mathbb{N}}$. Thus the sequence of PIT $(Z_F^{Y_{n+1}})_{n\in\mathbb{N}}$ is not simultaneously independent of the information and uniformly distributed. More precisely, the PITs will be uniformly distributed – since F is the distribution of $(Y_n)_{n\in\mathbb{N}}$ – and the PITs will also be i.i.d. The only difference with

bootstrap samples $(Z_n^\star)_{n\in\mathbb{N}}$ will be their independence with respect to the information $(\mathcal{F}_n)_{n\in\mathbb{N}}$.
Table 2 summarizes the results obtained for the four tests for different values of N. The empirical powers reported therein are calculated from 1000 replications, with a test level chosen as $\alpha = 0.05$ for the four tests. All the test parameters are fixed as detailed in Section 5.1.1.

(d, N)	T1	T2	T3	CEP
$(1, 50)$	0.14	**0.29**	0.14	0.09
$(2, 50)$	0.08	**0.26**	0.11	0.09
$(3, 50)$	0.08	**0.20**	0.11	0.09
$(1, 100)$	0.25	**0.56**	0.24	0.11
$(2, 100)$	0.20	**0.64**	0.23	0.11
$(3, 100)$	0.12	**0.53**	0.19	0.11
$(1, 200)$	0.44	**0.94**	0.63	0.13
$(2, 200)$	0.42	**0.96**	0.65	0.13
$(3, 200)$	0.29	**0.94**	0.5	0.13

Table 2: *Empirical power of the four competing tests with different numbers of realizations N and values of depth d, based on 1000 replications, with theoretical level test 0.05. Since CEP test does not depend on d, its power is repeated three times. The closer this rate is to 1, the better the test performs. Boldface highlights the best results obtained.*

Several comments can be formulated on the basis of Table 2:

- Naturally, the power increases with the size of the sample.
- Test 2 outperforms uniformly all its competitors, whatever the sample size or the depth are.
- Increasing the depth d, i.e. making too many splits in the regression tree, might decrease the power.
- The CEP test has a weak power, even for larger sample size $N/$

5.2 Real Data related to Weather Forecasting

We next consider an illustration with real data related to weather forecasting . It consists at 2-meter temperature forecast at the surface of the station of Airport Lyon-Bron (France) between 01/01/2011 and 31/12/2014. These forecasts take the form of ensemble forecasts. In other words, the forecast is made up of 35 equiprobable scenarios, obtained from numerical simulations (NWP). Obviously, such forecast cannot be ideally calibrated. In practice, these simulations are only the first step of forecasting. Statistical post-processing is done before they can be used, allowing them to be partially debiased and their under-dispersion to be fixed (Hamill and Colucci, 1997; Richardson, 2001).
We use here the simple method of statistical postprocessing called *Ensemble Model Output Statistics* (EMOS) and introduced in Gneiting et al. (2005). Let $\mathbf{x}_n \in \mathbb{R}^{35}$ the ensemble forecast produced the day n. In the simplest EMOS model, the predictive distribution F_n

given the information provided by the ensemble forecast is a normal distribution of the form

$$F_n = \mathcal{N}(a\overline{\mathbf{x}}_n + b, (c\sigma(\mathbf{x}_n) + d)^2),$$

where $(\overline{\mathbf{x}}_n, \sigma(\mathbf{x}_n))$ are respectively the empirical mean and standard deviation of the ensemble and $a, b, c, d \in \mathbb{R}$ are model parameters. The parameters are determined adaptively, meaning that they will change over time to best adapt to the seasonality of the weather. In addition, the parameters are chosen so as to minimize an empirical risk

$$a_n, b_n, c_n, d_n = \underset{a,b,c,d}{\operatorname{argmin}} \frac{1}{T} \sum_{t=1}^{T} \operatorname{loss}(F_{n-t}, y_{n+1-t}),$$

with t is the delay time of adaptation and y_n is the realization of Y_n. Without going into detail, the loss function used in this minimisation is the *Continuous Ranked Probability Score* (CRPS), that is widely used in practice Epstein (1969a); Hersbach (2000); Bröcker (2012).

We compare the 4 tests for ideal calibration presented above to check whether the post-processing uses "perfectly" the ensemble forecast $\mathbf{x}_n$, i.e. the information from the numerical weather predictions. The experiment will be repeated 4 times, one repetition per yearr. Delay time is $T = 30$ so that the number of realizations is $N = 365 - 30 = 335$. The following table shows the p-values of the different tests with this dataset. Remember that the test rejects the hypothesis that the post-processing method is ideally calibrated if the p-value is less than 0.05. Test 2 returns four p-values, one for each moment, that are independent of each other.

	Test 1	Test 2	Test 3	CEP
Year 2011	0.42	(0.83,0.09,0.01,0)	0.05	0.24
Year 2012	0.05	(0.10,0.53,0,0.18)	0.46	0.47
Year 2013	0.94	(0.705,0,0,0)	0.27	0.59
Year 2014	0.32	(0.79, 0.07, 0, 0.02)	0.51	0.22

On the whole, the post-processing method is not rejected, if the 3rd and 4th order moment is not taken into account. It is justified not to take them into account, as this method only estimates the first two moments. The CEP test has the advantage of being stable over the four tests, unlike our tests based on the CART algorithm.

6 Discussion

6.1 Testing cross-calibration

In the simulations, we compared our tests to the test CEP in Strähl and Ziegel (2017). The framework of this article is the *cross-calibration*, which is different from our setting of *ideal calibration*. We have adapterd their test to our framework but it is worth noting that the reverse is quite possible, meaning to take our tree based tests to test for cross-calibration. We present here some further details on cross-calibration. Let $(F_{1,n})_{n\in\mathbb{N}}, \ldots, (F_{k,n})_{n\in\mathbb{N}}$ be k different dynamical forecasters. The dynamical forecaster $(F_n)_{n\in\mathbb{N}}$ is said *cross-calibrated* with respect to $(F_{1,n})_{n\in\mathbb{N}}, \ldots, (F_{k,n})_{n\in\mathbb{N}}$ if

$$\forall n \in \mathbb{N},\ Z_{F_n}^{Y_{n+1}} \sim \text{Unif}([0,1]) \text{ and } Z_{F_n}^{Y_{n+1}} \perp\!\!\!\perp (F_{1,n}, \ldots, F_{k,n}, \mathcal{G}_n$$

where $\mathcal{G}_n = \sigma(Y_k, k \leq n, F_n, F_{1,n}, \ldots, F_{k,n})$. Ideal calibration compares the forecast F_n with the ideal forecast

$$F_n^* = \mathcal{L}(Y_{n+1} \mid Y_k, k \leq n, F_{1,n}, \ldots, F_{k,n}).$$

This means that the forecaster perfectly uses the information from the past observations and from the other forecaster, and hence performs better than his competitors. Our tree based test can be adapted with very little modification to the framework of cross-calibration. Similarly as for the CEP test for cross-calibration, we can consider $0 < p_1 < \cdots < p_K) < 1$ and consider the explanatory variable $X_n = (Y_n, F_{i,n}^{-1}(p_j),\ 1 \leq i \leq k,\ 1 \leq j \leq K)$. Then we can use our procedures to test whether the PITs $F_n^{Y_{n+1}}$, $n \geq 0$, are independent of X_n, $n \geq 0$.

6.2 Weaker calibration

In applications, for instance in meteorology, it is impossible to retrieve exactly the "true" conditional distribution, so that testing for ideal calibration might be too optimistic. The null hypothesis is very restrictive and the test may lead to many rejections. We have nevertheless proposed several statistics to test whether a forecast is ideal or not. The ideas behind the construction of these tests could be useful and moreover the proof of the asymptotic normality and asymptotic behaviour of the boostrap is interesting and hold under general assumptions (e.g. including several dependence). .

A recent kind of calibration, Gneiting and Resin (2021), is the $\mathcal{T}$-calibration where $\mathcal{T}$ is a functional of the probability measure such as the median, the mean, the variance...

Definition 4. *A random forecast F is $\mathcal{T}$-calibrated if*

$$\mathcal{T}(\mathcal{L}(Y \mid \mathcal{T}(F))) = \mathcal{T}(F) \ a.s.$$

A functional $\mathcal{T}$ is said to be identifiable if there exists $V \colon \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ such that $V(\cdot, y)$ is increasing, left-continuous for all $y \in \mathbb{R}$ and satisfying for each $F \in \mathcal{P}$,

$$\begin{cases} \forall x < \mathcal{T}(F), \ \int_{\mathbb{R}} V(x,y)\ F(\mathrm{d}y) < 0 \\ \forall x > \mathcal{T}(F), \ \int_{\mathbb{R}} V(x,y)\ F(\mathrm{d}y) > 0 \\ \int_{\mathbb{R}} V(\mathcal{T}(F),y)\ F(\mathrm{d}y) = 0 \end{cases}.$$

In a similar way to their proof of Theorem 2.10, the $\mathcal{T}$-calibration can be written in terms of a conditional expectation.

Proposition 19. *Let $\mathcal{T}$ be an identifiable functional, and let F be a random forecast. The forecast F is $\mathcal{T}$-calibrated if and only if*

$$\mathbb{E}[V(\mathcal{T}(F), Y) \mid \mathcal{T}(F)] = 0.$$

Proof. We have

$$\mathbb{E}[V(\mathcal{T}(F), Y) \mid \mathcal{T}(F)] = \int_{\mathbb{R}} V(\mathcal{T}(F), y)\ F_{\mathcal{T}}(\mathrm{d}y), \text{ where } F_{\mathcal{T}} = \mathcal{L}(Y \mid \mathcal{T}(F)).$$

Then it is equal to 0 if and only if $\mathcal{T}(F) = \mathcal{T}(F_{\mathcal{T}})$. □

This writing in terms of conditional expectation makes it possible to use the same method based on empirical process to analyse $\mathcal{T}-$calibration. More precisely, similarly as in Lemma 6, one can show that the process

$$\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\mathcal{T}(F_i)\leq x\}}V(\mathcal{T}(F_i),Y_i)\right)_{x\in\mathbb{R}}$$

is centered if and only if the dynamic forecast $(F_n)_{n\in\mathbb{Z}}$ is $\mathcal{T}$-calibrated . This preliminary result suggests to follow the techniques presented in this article to propose tests for $\mathcal{T}$-calibration.

7 Proofs

7.1 Proofs of Section 2

Proof of Lemma 1. The implication $i) \Rightarrow ii)$ is stated in (Brockwell, 2007, Lemma 2). We prove here the converse implication $ii) \Rightarrow i)$ and assume that $Z_F^Y \sim \text{Unif}([0,1])$. We denote by G the CDF of Y and we want to prove that $G = F$.
The fact that F is nonincreasing together with the definition

$$Z_F^Y = VF(Y^-) + (1-V)F(Y)$$

imply the following inclusions: for all $x \in \mathbb{R}$,

$$\{Y \leq x\} \subset \{Z_F^Y \leq F(x)\},$$
$$\{Y > x\} \subset \{Z_F^Y \geq F(x)\}.$$

Taking probabilities, we deduce

$$G(x) \leq \mathbb{P}\left(Z_F^Y \leq F(x)\right) = F(x),$$
$$1 - G(x) \leq \mathbb{P}(\{Z_F^Y \geq F(x)\}) = 1 - F(x).$$

As a consequence, $F(x) = G(x)$ and, $x \in \mathbb{R}$ being arbitrary, $F = G$. □

For future reference, we note that Lemma 1 can be rewritten in the following equivalent form. The implication $i) \Rightarrow ii)$ states that

$$\forall z \in [0,1], \quad \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{vF(y^-)+(1-v)F(y)\leq z\}} F(\mathrm{d}y)\mathrm{d}v = z, \tag{17}$$

where the left hand side is an integral form for $\mathbb{P}(Z_F^Y \leq z)$ valid when $Y \sim F$. The implication $ii) \Rightarrow i)$ states that

$$\left(\forall z \in [0,1], \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{vF(y^-)+(1-v)F(y)\leq z\}} G(\mathrm{d}y)\mathrm{d}v = z\right) \Rightarrow (G = F) \tag{18}$$

Equality for all $z \in [0,1]$ can be restricted to equality for all z in a dense subset.
The following technical proposition justifies the measurability of the PIT for random probabilistic forecast and may possibly be skipped at first reading.

Proposition 20. *Let (Y, F, V) be a one step prediction space on $(\Omega, \mathcal{G}, \mathbb{P})$. Then the PIT Z_F^Y is measurable on $(\Omega, \mathcal{G}, \mathbb{P})$.*

Proof. We denote by μ the probability kernel associated with the random forecast F (seen as a random CDF). Define, for $w, y \in \Omega \times \mathbb{R}$,

$$g_1(\omega, y) = F(\omega, y^-) = \mu(\omega,]-\infty, y[),$$
$$g_2(\omega, y) = F(\omega, y) = \mu(\omega,]-\infty, y]).$$

We prove the joint measurability of g_1 and g_2 in both variables (ω, y). The measurability properties of μ imply that, for all fixed $y \in \mathbb{R}$, $g_1(\cdot, y)$ and $g_2(\cdot, y)$ are measurable. For fixed $\omega \in \Omega$, $g_1(\omega, \cdot)$ is left continuous and $g_2(\omega, \cdot)$ is right continuous. Consider, for $n \geq 1$, the approximations

$$g_1^n(\omega, y) = g_1\left(\omega, \frac{\lfloor ny \rfloor}{n}\right) \quad \text{and} \quad g_2^n(\omega, y) = g_2\left(\omega, \frac{\lceil ny \rceil}{n}\right).$$

Note that we use the floor operator (closest lowest integer) in g_1 and the ceiling operator (closest largest integer) in g_2. The measurability of g_1^n and g_2^n is easily checked as well as their pointwise convergence as $n \to \infty$ to g_1 and g_2 respectively. The measurability of g_1 and g_2 follows.
Finally, the measurability of the PIT is a consequence of the equality

$$Z_F^Y(\omega) = V(\omega) g_1(\omega, Y(\omega)) + (1 - V(\omega)) g_2(\omega, Y(\omega))$$

and from basic properties of measurability (composition, product and sum of measureable maps). □

Proof of Proposition 2. For the direct implication, let $A \in \mathcal{F}$ and $z \in [0, 1]$, as $F = \mathcal{L}(Y \mid \mathcal{F})$,

$$\mathbb{P}(A \cap \{Z_F^Y \leq z\}) = \int_A \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{vF(\omega,]-\infty,y[)+(1-v)F(\omega,]-\infty,y]) \leq z\}} F(\omega, \mathrm{d}y) \mathrm{d}v \mathbb{P}(\mathrm{d}\omega),$$

by Fubini Theorem for conditional distribution and V is independent of $\mathcal{F}$. Then with the Equation (17) for $\omega \in A$ fixed,

$$\int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{uF(\omega,]-\infty,y[)+(1-u)F(\omega,]-\infty,y]) \leq z\}} F(\omega, \mathrm{d}y) \mathrm{d}u = z$$

Hence $\mathbb{P}(A \cap \{Z_F^Y \leq z\}) = \mathbb{P}(A) z$, so $Z_F^Y \sim \mathrm{Unif}([0, 1])$ and is independent of $\mathcal{F}$.

For the reciprocal implication, the $\mathcal{F}-$measurability of F allows us to apply Fubini Theorem for conditional distribution. Let $z \in [0, 1] \cap \mathbb{Q}$ and $A \in \mathcal{F}$, the independence and uniform distribution on $[0, 1]$ imply,

$$\mathbb{E}\left[\mathbb{1}_A \mathbb{1}_{\{Z_F^Y \leq z\}}\right] - \mathbb{P}(A) z = \int_A \underbrace{\int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{vF(\omega,]-\infty,y[)+(1-v)F(\omega,]-\infty,y]) \leq z\}} \mu(\omega, \mathrm{d}y) \mathrm{d}v - z}_{\mathcal{F}\text{-measurable}} \mathbb{P}(\mathrm{d}\omega),$$

where μ is $\mathcal{L}(Y \mid \mathcal{F})$. The $\mathcal{F}$-measurability of the integrand is a consequence of the Fubini Theorem. As this integral is null for $A \in \mathcal{F}$,

$$a.s., \ \forall z \in [0, 1] \cap \mathbb{Q}, \int_0^1 \int_{\mathbb{R}} \mathbb{1}_{\{uF(\omega,]-\infty,y[)+(1-u)F(\omega,]-\infty,y]) \leq z\}} \mu(\omega, \mathrm{d}y) \mathrm{d}u = z.$$

Then by the Equation (18),

$$a.s.,\ F = \mu = \mathcal{L}(Y \mid \mathcal{F}).$$

□

Proof of Corollary 3. It is a direct consequence of the Proposition 2. To get the last point, it suffices to remark that if $(\mathcal{F}_n)_{n\in\mathbb{N}}$ contains the filtration endowed by $(Y_n)_{n\in\mathbb{N}}$ then Z_n is $\sigma(\mathcal{F}_{n+T}, V_n)$ measurable, and Z_{n+T} is independent of $\mathcal{F}_{n+T}$ and V_n. □

7.2 Proofs of Section 3

Proof of Proposition 5. Under (H_0) with Assumptions (A1)-(A3), Theorem 10.2 of Dedecker et al. (2007) is applicable. In this book, the convergence is in the Skorokhod space $\mathcal{D}([0,1]^d)$. But the authors prove the tighness in $\ell^\infty([0,1]^d)$. Then by their Proposition 4.2, the convergence is also in ℓ^∞. □

Proof of Lemma 6. 1. This equivalence is the definition of $Z \sim \text{Unif}([0,1])$;

2. Recall that $(Z_n, X_n)_{n\in\mathbb{N}}$ is stationary

$$\begin{aligned}\mathbb{F}_2^{(n)} \text{ is centred} &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \text{Cov}\left(\mathbb{1}_{\{Z_1\le y\}}, \mathbb{1}_{\{X_1\le t\}}\right) = 0\\ &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathbb{P}(Z_1\le y, X_1\le t) = \mathbb{P}(Z_1\le y)\mathbb{P}(X_1\le t)\\ &\Leftrightarrow (Z_1,X_1) \text{ are independent.}\end{aligned}$$

3. For $y,t \in [0,1]\times[0,1]^d$,

$$\mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}_{\{Z_i\le y\}} - y\right)\mathbb{1}_{\{X_i\le t\}}.$$

The stationarity implies

$$\begin{aligned}\mathbb{F}_1^{(n)} + \mathbb{F}_2^{(n)} \text{ is centred} &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathbb{E}\left[\left(\mathbb{1}_{\{Z_1\le y\}} - y\right)\mathbb{1}_{\{X_1\le t\}}\right] = 0\\ &\Leftrightarrow \forall y,t \in [0,1]^{d+1},\ \mathbb{P}(Z_1\le y, X_1\le t) = yF(t)\\ &\Leftrightarrow (Z_1,X_1) \text{ are independent and } Z_1 \sim \text{Unif}([0,1]).\end{aligned}$$

□

Proof of Theorem 8. As the evaluation is continuous, the mapping theorem yields

$$\begin{aligned}\sqrt{n}\begin{pmatrix}\mathbb{F}_1^{(n)}\\ \mathbb{F}_2^{(n)}\\ \mathbb{F}_3^{(n)}\end{pmatrix} &= \begin{pmatrix}\Phi_1\\ \Phi_2\\ \Phi_3\end{pmatrix}\left(\sqrt{n}\mathbb{G}^{(n)}\right)\\ &\rightsquigarrow \begin{pmatrix}\Phi_1(\mathbb{G})\\ \Phi_2(\mathbb{G})\\ \Phi_3(\mathbb{G})\end{pmatrix} = \begin{pmatrix}\mathbb{G}_1\\ \mathbb{G}_2\\ \mathbb{G}_3\end{pmatrix},\end{aligned}$$

where

$$\begin{cases}\Phi_1(G)(y,t) = G(y,\mathbf{1})F(t)\\ \Phi_2(G)(y,t) = G(y,t) - G(y,\mathbf{1})F(t) - yG(1,t)\\ \Phi_3(G)(y,t) = G(1,t)\end{cases}.$$

The covariance functions of the two first processes are, for $a, b \in [0,1]$ and $s, t \in [0,1]^d$,

$$\begin{cases} \Gamma_1((a,t),(b,s)) = \sum_{i\in\mathbb{Z}} \mathrm{Cov}\left(F(t)\mathbb{1}_{\{Z_0\leq a\}}, F(s)\mathbb{1}_{\{Z_i\leq b\}}\right) \\ \Gamma_2((a,t),(b,s)) = \sum_{i\in\mathbb{Z}} \mathrm{Cov}\left(H_0(a,t), H_i(b,s)\right), \\ \text{where } H_i(a,t) = \left(\mathbb{1}_{\{Z_i\leq a\}} - a\right)\left(\mathbb{1}_{\{X_i\leq t\}} - F(t)\right) \end{cases}.$$

The first function is directly simplified because the sequence of PITs $(Z_i)_{i\in\mathbb{Z}}$ is $(T-1)$ dependent. However, this is not true for the sequence $(X_i, Z_i)_{i\in\mathbb{Z}}$. Nevertheless, the simplification is possible. Let $|i| \geq T$, then Z_i is independent of (X_0, Z_0, X_i),

$$\mathrm{Cov}\left(H_0(a,t), H_i(b,s)\right) = \mathbb{E}[\mathbb{1}_{\{Z_i\leq b\}} - b]\mathbb{E}[\left(\mathbb{1}_{\{X_i\leq s\}} - F(s)\right)H_0(a,t)] = 0.$$

In the case where $T = 1$, the sequence $(Z_i)_{i\in\mathbb{Z}}$ is independent. The independence of $\mathbb{G}_1$ and $\mathbb{G}_2$ is a consequence of the decorrelation between $\mathbb{F}_1^{(n)}$ and $\mathbb{F}_2^{(n)}$. For $i \neq j$, Z_i is independent of (X_i, Z_j) then

$$\mathrm{Cov}(H_i(a,t), \mathbb{1}_{\{Z_j\leq b\}} - b) = \mathbb{E}[\mathbb{1}_{\{Z_i\leq a\}} - a]\mathbb{E}[\left(\mathbb{1}_{\{X_i\leq t\}} - F(t)\right)(\mathbb{1}_{\{Z_j\leq b\}} - b)] = 0.$$

For $i = j$, the PIT Z_i is independent of X_i then

$$\mathrm{Cov}(H_i(a,t), \mathbb{1}_{\{Z_i\leq b\}} - b) = \mathbb{E}[\mathbb{1}_{\{X_i\leq t\}} - t]\mathbb{E}[\left(\mathbb{1}_{\{Z_i\leq a\}} - a\right)(\mathbb{1}_{\{Z_i\leq b\}} - b)] = 0.$$

□

Proof of Theorem 10. It is a direct consequence of the δ-method in van der Vaart and Wellner (1996, Theorem 3.9.4) and the fact that as $\partial_3\Psi(0,0,0) = 0$,

$$\mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, \mathbb{G}_3) = \mathrm{d}_0\Psi(\mathbb{G}_1, \mathbb{G}_2, 0)$$

. □

Proof of Theorem 11. The random sequence $(Z_i, Z_i^\star, X_i)_{i\in\mathbb{Z}}$ still checks Assumptions (A1)-(A3) then this convergence is a consequence of Theorem 10.2 in Dedecker et al. (2007),

$$\sqrt{n}\begin{pmatrix} \mathbb{G}^{(n)} \\ \mathbb{G}^{(n)\star} \end{pmatrix} \rightsquigarrow \begin{pmatrix} \mathbb{G} \\ \mathbb{G}^\star \end{pmatrix}.$$

The application of the Mapping Theorem is done in the same way as in the proof of Theorem 8. The essential part of this theorem is the equality between the classical limit and the bootstrapped limit and the asymptotic independence. This part will be shown in several steps. As many of these steps are identical, we will not show them all. Moreover, since they are Gaussian vectors, the pairwise independence of the components implies the independence of the vectors.

1. $\mathbb{G}_1$ and $\mathbb{G}_1^\star$ have the same distribution;
2. $\mathbb{G}_2$ and $\mathbb{G}_2^\star$ have the same distribution;
3. $\mathbb{G}_1$ is independent of $\mathbb{G}_1^\star$;
4. $\mathbb{G}_2$ is independent of $\mathbb{G}_2^\star$;

5. $\mathbb{G}_2$ is independent of $\mathbb{G}_1^\star$;

6. $\mathbb{G}_1$ is independent of $\mathbb{G}_2^\star$.

Firstly, for $T = 1$, the PITs $(Z_i)_{i\in\mathbb{Z}}$ and the bootstrapped PITs $(Z_i^\star)_{i\in\mathbb{Z}}$ are independent and have the same distribution. Then it is the same for the processes $(\mathbb{F}_1^{(n)}, \mathbb{F}_2^{(n)})$ and their limits. So the Point 1 and 3 are shown. Point 2 comes from the fact that for the lead time $T = 1$, the covariance of $\mathbb{G}_2$ simplifies to

$$\Gamma_2((a,t),(b,s)) = \mathrm{Cov}(\left(\mathbb{1}_{\{Z_0\le a\}} - a\right)\left(\mathbb{1}_{\{X_0\le t\}} - F(t)\right), \\ \left(\mathbb{1}_{\{Z_0\le b\}} - b\right)\left(\mathbb{1}_{\{X_0\le s\}} - F(s)\right)),$$

which is the same that $\mathbb{G}_2^\star$. The last three points are shown by studying the correlation of the processes. As the proofs are identical, we only detail Point 4. For $i, j \in \mathbb{Z}$,

$$\begin{aligned}
&\mathrm{Cov}\left(\left(\mathbb{1}_{\{Z_i\le a\}} - a\right)\left(\mathbb{1}_{\{X_i\le t\}} - F(t)\right), \left(\mathbb{1}_{\{Z_j^\star\le b\}} - b\right)\left(\mathbb{1}_{\{X_j\le s\}} - F(s)\right)\right) \\
&\quad = \mathbb{E}\left[\left(\left(\mathbb{1}_{\{Z_i\le a\}} - a\right)\left(\mathbb{1}_{\{X_i\le t\}} - F(t)\right)\left(\mathbb{1}_{\{Z_j^\star\le b\}} - b\right)\left(\mathbb{1}_{\{X_j\le s\}} - F(s)\right)\right)\right] \\
&\quad = \mathbb{E}[\mathbb{1}_{\{Z_j^\star\le b\}} - b]\mathbb{E}\left[\left(\left(\mathbb{1}_{\{Z_i\le a\}} - a\right)\left(\mathbb{1}_{\{X_i\le t\}} - F(t)\right)\left(\mathbb{1}_{\{X_j\le s\}} - F(s)\right)\right)\right] \\
&\quad = 0
\end{aligned}$$

□

Proof of Lemma 14. Let us recall that for $y, t \in [0,1]\times[0,1]^d$,

$$\mathbb{F}^{(n)}(y,t) := \mathbb{F}_1^{(n)}(y,t) + \mathbb{F}_2^{(n)}(y,t) = \frac{1}{n}\sum_{i=1}^n \left(\mathbb{1}_{\{Z_i\le y\}} - y\right)\mathbb{1}_{\{X_i\le t\}}.$$

Let us develop the integral

$$\begin{aligned}
\int_0^1 \mathbb{F}^{(n)}(y,t)\ \mathrm{d}g(y) &= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\le t\}}\left(\int_0^1 g_0'(y)(\mathbb{1}_{\{Z_i\le y\}} - y)\ \mathrm{d}y - \sum_{j=1}^k w_j(\mathbb{1}_{\{Z_i\le\alpha_j\}} - \alpha_j)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\le t\}}\left(g_0(1) - g_0(Z_i) - [g_0(y)y]_{y=0}^1 + \int_0^1 g_0(y)\ \mathrm{d}y - \sum_{j=1}^k w_j(\mathbb{1}_{\{Z_i\le\alpha_j\}} - \alpha_j)\right) \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{X_i\le t\}}\left(\int_0^1 g(u)\ \mathrm{d}u - g(Z_i)\right).
\end{aligned}$$

□

Proof of Proposition 15. Let $h_1, h_2, h_3 \in \ell^\infty$, the function Ψ_g^F is null at $(0,0,0)$,

$$\begin{aligned}
\Psi_g^F(h_1^{(n)}, h_2^{(n)}, h_3^{(n)}; t) &= -\frac{\int_0^1 h_1(y,t) + h_2(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t) + h_3(1,t)}} \\
&= -\frac{\int_0^1 h_1(y,t) + h_2(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t)}} \times \left(1 - \frac{h_3(1,t)}{2F(t)} + o(\|h_3\|/F)\right).
\end{aligned}$$

The Taylor expansion is uniform in t because $\varepsilon \leq F(t) \leq 1-\varepsilon$. We therefore define the linear application

$$H(h_1,h_2,h_3;t) = -\frac{\displaystyle\int_0^1 h_1(y,t)+h_2(y,t)\ \mathrm{d}g(y)}{\sqrt{F(t)}} \tag{19}$$

The continuity of this linear application is a direct consequence of the following inequality

$$\left|\int_0^1 h_1(y,t)+h_2(y,t)\ \mathrm{d}g(y)\right| \leq \|h_1+h_2\|_\infty \left(\|g_0'\|_\infty + \sum_{j=1}^k |w_j|\right). \tag{20}$$

This inequality also proves that there exists $C>0$ such that

$$\left\| \Psi_g^F(h_1^{(n)},h_2^{(n)},h_3^{(n)};\cdot) + \frac{\displaystyle\int_0^1 h_1(y,\cdot)+h_2(y,\cdot)\ \mathrm{d}g(y)}{\sqrt{F(\cdot)}} \right\|_\infty \leq C(\|h_1\|+\|h_2\|)\|h_3\|,$$

and this bound is $o(\|h_1\|+\|h_2\|+\|h_3\|)$. This concludes that Ψ_g^F is differantiable at $(0,0,0)$ and $\mathrm{d}_0\Psi_g^F$ is the linear application H. □

Proof of Proposition 16. Let f,g be centred piecewise continuously differentiable. For sake of simplification, we assume that they are just continuously differentiable. Let us prove that for all $t,s \in C_\varepsilon$,

$$-\int_0^1 (\mathbb{G}_1(y,t)+\mathbb{G}_2(y,t))\ \mathrm{d}g(y) \perp\!\!\!\perp -\int_0^1 (\mathbb{G}_1(y,s)+\mathbb{G}_2(y,s))\ \mathrm{d}f(y).$$

By the Mapping theorem as the integral is continuous and Lemma 14,

$$\begin{aligned}\int_0^1 -g'(y)\,(\mathbb{G}_1(y,t)+\mathbb{G}_2(y,t))\ \mathrm{d}y &= \lim\sqrt{n}\int_0^1 -g'(y)\left(\mathbb{F}_1^{(n)}(y,t)+\mathbb{F}_2^{(n)}(y,t)\right)\ \mathrm{d}y \\ &= \lim\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n g(Z_i)\mathbb{1}_{\{X_i\leq t\}}\right).\end{aligned}$$

By a dependent Central Limit Theorem,

$$\sqrt{n}\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n g(Z_i)\mathbb{1}_{\{X_i\leq t\}}\\ \frac{1}{n}\sum_{i=1}^n f(Z_i)\mathbb{1}_{\{X_i\leq s\}}\end{pmatrix} \rightsquigarrow \mathcal{N}(\mathbf{0},\Sigma),$$

where

$$2\Sigma_{1,2} = \sum_{i\in\mathbb{Z}} \mathrm{cov}\left(g(Z_0)\mathbb{1}_{\{X_0\leq t\}}, f(Z_i)\mathbb{1}_{\{X_i\leq s\}}\right)$$

If $i\neq 0$ by the Assumption (A1) and the Corollary 3, Z_i is independent of (Z_0,X_0,X_i), for $i>0$, or Z_0 is independent of (Z_i,X_i,X_0), for $i<0$, then

$$\mathrm{cov}\left([g(Z_0)-m(g)]\mathbb{1}_{\{X_0\leq t\}}, [f(Z_i)-m(f)]\mathbb{1}_{\{X_i\leq s\}}\right) = 0,$$

and for $i = 0$,

$$\begin{aligned}\operatorname{cov}\left(g(Z_0)\mathbb{1}_{\{X_0 \le t\}}, f(Z_0)\mathbb{1}_{\{X_0 \le s\}}\right) &= \mathbb{E}\left(g(Z_0)f(Z_0)\mathbb{1}_{\{X_0 \le t\}}\right) \\ &= F(t) \times \mathbb{E}\left(f(Z_0)g(Z_0)\right) \\ &= F(t)\int_0^1 f(u)g(u)\,\mathrm{d}u = 0.\end{aligned}$$

As the limit is Gaussian, this decorrelation implies the independence of the marginals. □

References

Anderson, J. L. (1996). A Method for Producing and Evaluating Probabilistic Forecasts from Ensemble Model Integrations. *Journal of Climate*, 9(7):1518–1530.

Berkowitz, J. (2001). Testing density forecasts, with applications to risk management. *Journal of Business & Economic Statistics*, 19(4):465–474.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall, New York, NY.

Bröcker, J. (2012). Evaluating raw ensembles with the continuous ranked probability score. *Quarterly Journal of the Royal Meteorological Society*, 138(667):1611–1617.

Bröcker, J. (2022). Uniform calibration tests for forecasting systems with small lead time. *Statistics and Computing*, 32(6).

Brockwell, A. (2007). Universal residuals: A multivariate transformation. *Statistics and Probability Letters*, 77(14):1473 – 1478.

Bröcker, J. (2009). Reliability, sufficiency, and the decomposition of proper scores. *Quarterly Journal of the Royal Meteorological Society*, 135(643):1512–1519.

Cox, D. D. and Lee, J. S. (2008). Pointwise testing with functional data using the westfall-young randomization method. *Biometrika*, 95(3):621–634.

David, F. N. and Johnson, N. L. (1948). The probability integral transformation when parameters are estimated from the sample. *Biometrika*, 35(1/2):182–190.

Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, 147(2):278–292.

Dedecker, J., Doukhan, P., Lang, G., Leon, J. R., Louhichi, S., and Prieur, C. (2007). *Weak dependence: with examples and applications*, volume 190 of *Lectures Notes in Statistics*. Springer.

Diebold, F. X., Gunther, T. A., and Tay, A. S. (1998). Evaluating density forecasts with applications to financial risk management. *International Economic Review*, 39(4):863–883.

Epstein, E. S. (1969a). A scoring system for probability forecasts of ranked categories. *Journal of Applied Meteorology*, 8(6):985–987.

Epstein, E. S. (1969b). Stochastic dynamic prediction1. *Tellus*, 21(6):739–759.

Gneiting, T. (2014). *Calibration of medium-range weather forecasts*. European Centre for Medium-Range Weather Forecasts.

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

Gneiting, T. and Katzfuss, M. (2014). Probabilistic forecasting. *Annual Review of Statistics and Its Application*, 1:125–151.

Gneiting, T. and Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378.

Gneiting, T., Raftery, A. E., Westveld, A. H., and Goldman, T. (2005). Calibrated Probabilistic Forecasting Using Ensemble Model Output Statistics and Minimum CRPS Estimation. *Monthly Weather Review*, 133(5):1098 – 1118.

Gneiting, T. and Ranjan, R. (2013). Combining predictive distributions. *Electronic Journal of Statistics*, 7:1747–1782.

Gneiting, T. and Resin, J. (2021). Regression diagnostics meets forecast evaluation: Conditional calibration, reliability diagrams, and coefficient of determination.

Hamill, T. M. and Colucci, S. J. (1997). Verification of eta-rsm short-range ensemble forecasts. *Monthly Weather Review*, 125(6):1312–1327.

Held, L., Rufibach, K., and Balabdaoui, F. (2010). A score regression approach to assess calibration of continuous probabilistic predictions. *Biometrics*, 66.

Henzi, A., Kleger, G.-R., Hilty, M. P., Wendel Garcia, P. D., Ziegel, J. F., and for Switzerland, R.-.-I. I. (2021). Probabilistic analysis of covid-19 patients' individual length of stay in swiss intensive care units. *PloS one*, 16(2):e0247265.

Hersbach, H. (2000). Decomposition of the continuous ranked probability score for ensemble prediction systems. *Weather and Forecasting*, 15(5):559–570.

Hong, T., Pinson, P., Fan, S., Zareipour, H., Troccoli, A., and Hyndman, R. (2016). Probabilistic energy forecasting: Global energy forecasting competition 2014 and beyond. *International Journal of Forecasting*, 32(3):896–913.

Jolliffe, I. T. and Primo, C. (2008). Evaluating rank histograms using decompositions of the chi-square test statistic. *Monthly Weather Review*, 136(6):2133 – 2139.

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

Richardson, D. S. (2001). Measures of skill and value of ensemble prediction systems, their interrelationship and the effect of ensemble size. *Quarterly Journal of the Royal Meteorological Society*, 127(577):2473–2489.

Rosenblatt, M. (1961). Independence and dependence. In Neyman, J., editor, *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 2, pages 431–443. University of California Press, Berkeley, CA. (Berkeley, CA, 20–30 July 1960). Zbl:0105.11802. MR:133863.

Strähl, C. and Ziegel, J. (2017). Cross-calibration of probabilistic forecasts. *Electron. J. Statist.*, 11(1):608–639.

Talagrand, O., Vautard, R., and Strauss, B. (1997). Evaluation of probabilistic prediction systems. *Workshop on Predictability, 20-22 October 1997*, pages 1–26.

Tiberi-Wadier, A.-L., Goutal, N., Ricci, S., Sergent, P., Taillardat, M., Bouttier, F., and Monteil, C. (2021). Strategies for hydrologic ensemble generation and calibration: On the merits of using model-based predictors. *Journal of Hydrology*, 599:126233.

Tsyplakov, A. (2013). Evaluation of probabilistic forecasts: Proper scoring rules and moments. *SSRN Electronic Journal*.

van der Vaart, A. and Wellner, J. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer.

Vannitsem, S., Bremnes, J., Demaeyer, J., Evans, G., Flowerdew, J., Hemri, S., Lerch, S., Roberts, N., Theis, S., Atencia, A., Ben Bouallègue, Z., Bhend, J., Dabernig, M., De Cruz, L., Hieta, L., Mestre, O., Moret, L., Odak Plenković, I., Schmeits, M., Taillardat, M., Van den Bergh, J., Van Schaeybroeck, B., Whan, K., and Ylhaisi, J. (2021). Statistical postprocessing for weather forecasts—review, challenges and avenues in a big data world. *Bulletin of the American Meteorological Society*, 102(3):681–699.

Weigel, A. P. (2011). *Ensemble Forecasts*, chapter 8, pages 141–166. John Wiley & Sons, Ltd.