International Journal of Approximate Reasoning 178 (2025) 109332

Contents lists available at ScienceDirect

International Journal of Approximate Reasoning

journal homepage: www.elsevier.com/locate/ijar

Adding imprecision to hypotheses: A Bayesian framework for testing practical significance in nonparametric settings

Rodrigo F.L. Lassance [a,b,*], Rafael Izbicki [a], Rafael B. Stern [c]

[a] Department of Statistics, Federal University of São Carlos, Rodovia Washington Luis km 235, São Carlos, 13565-905, São Paulo, Brazil
[b] Institute of Mathematics and Computer Sciences, Avenida Trabalhador São-carlense 400, São Carlos, 13566-590, São Paulo, Brazil
[c] Institute of Mathematics and Statistics, Rua do Matão 1010, São Paulo, 05508-090, São Paulo, Brazil

ARTICLE INFO

Keywords:
Pragmatic hypothesis
Bayesian nonparametrics
Adherence
Goodness-of-fit
Quantile
Two-sample
Link function

ABSTRACT

Instead of testing solely a precise hypothesis, it is often useful to enlarge it with alternatives deemed to differ negligibly from it. For instance, in a bioequivalence study one might test if the concentration of an ingredient is exactly the same in two drugs. In such a context, it might be more relevant to test the enlarged hypothesis that the difference in concentration between them is of no practical significance. While this concept is not alien to Bayesian statistics, applications remain mostly confined to parametric settings and strategies that effectively harness experts' intuitions are often scarce or nonexistent. To resolve both issues, we introduce the Pragmatic Region Oriented Test (PROTEST), an accessible nonparametric testing framework based on distortion models that can seamlessly integrate with Markov Chain Monte Carlo (MCMC) methods and is available as an R package. We develop expanded versions of model adherence, goodness-of-fit, quantile and two-sample tests. To demonstrate how PROTEST operates, we use examples, simulated studies that critically evaluate features of the test and an application on neuron spikes. Furthermore, we address the crucial issue of selecting the threshold—which controls how much a hypothesis is to be expanded—even when intuitions are limited or challenging to quantify.

1. Introduction

Throughout the history of Bayesian statistics, the idea of inserting utility judgments directly into hypotheses has been often proposed, albeit remaining largely ignored in practical settings. The most pristine example of this behavior is perhaps the defense that all point null hypotheses should be reframed as composite ones [1–3]. However, this idea was either applied in very specific settings—such as switching $H_0 : \theta = \theta_0$ for $H_0 : |\theta - \theta_0| \in [\delta_L, \delta_U]$, with $\delta_L \leq 0 \leq \delta_U$ known beforehand [4,5]—or not applied at all, being described as "a lot of hard work" [6].

The appeal of using external information to enlarge hypotheses is twofold, of both theoretical and practical nature. For the former, it avoids the requirement of adding probability masses to priors—a common strategy when using Bayes factors [7–9]. As for the latter, it allows for the inclusion of objective and subjective knowledge, such as measurement errors and researcher considerations on negligible deviations respectively, ensuring that the new hypothesis is more akin to the actual interest of the researcher.

While the use of Bayesian nonparametric models for testing hypotheses has been defended over frequentist/parametric methods [10], deriving enlarged hypotheses where they can be straightforwardly applied remains an incipient subject. Considering such

* Corresponding author at: Institute of Mathematics and Computer Sciences, Avenida Trabalhador São-carlense 400, São Carlos, 13566-590, São Paulo, Brazil.
E-mail addresses: rflassance@usp.br (R.F.L. Lassance), rizbicki@ufscar.br (R. Izbicki), rstern@ime.usp.br (R.B. Stern).

https://doi.org/10.1016/j.ijar.2024.109332
Received 9 September 2024; Received in revised form 10 November 2024; Accepted 23 November 2024

Available online 28 November 2024

gap in the literature, this work brings forth a theoretical framework for hypothesis enlargement that is both capable of expanding nonparametric hypotheses based on the inputs of experts and easily applicable through currently available technologies, such as Markov Chain Monte Carlo (MCMC) methods. With this contribution, we expect researchers to be able to test complex hypotheses without having to disregard valuable information in the process.

In this work, we make use of an expanded version of pragmatic hypotheses [11,12]. For nonparametric settings, we substitute the parametric space Θ for a more abstract hypothesis space $\mathbb{H}$, whose definition depends on what is to be tested. For instance, when evaluating the characteristics of a single and unknown distribution (such as in subsections 3.2 and 3.3), we take $\mathbb{H} = \mathbb{F}$, where $\mathbb{F}$ is the space of distribution functions. However, for high-dimensional settings, $\mathbb{H}$ might instead be the set of regression functions that relate the explanatory and the response variables, such as in subsection 3.1.

Definition 1 *(Pragmatic hypothesis).* Let $\mathbb{H}$ be the hypothesis space and $H_0 \subset \mathbb{H}$ be the null hypothesis of interest. For a given dissimilarity function $d(\cdot,\cdot)$ and a threshold $\varepsilon > \min(d) \geq 0$, a pragmatic hypothesis is defined as

$$Pg(H_0, d, \varepsilon) := \bigcup_{P_0 \in H_0} \{P \in \mathbb{H} : d(P_0, P) < \varepsilon\} = \left\{ P \in \mathbb{H} : \inf_{P_0 \in H_0} d(P_0, P) < \varepsilon \right\}. \tag{1}$$

For brevity, if $d(\cdot,\cdot)$ and ε are evident, we substitute $Pg(H_0, d, \varepsilon)$ for $Pg(H_0)$.

The intuition behind Definition 1 is as follows. The purpose of the pragmatic hypothesis is to expand the null so that it contains all elements that, for all practical purposes, are similar enough to at least one element of H_0. This is the same as checking, for each $P_0 \in H_0$, which are the elements $P \in \mathbb{H}$ such that $d(P_0, P) < \varepsilon$ to then take their union, which is represented by the left side of (1). This is the same as evaluating, for each $P \in \mathbb{H}$, if the smallest difference between P and all elements of H_0 is less than ε, the right side of (1).

When $\mathbb{H} = \mathbb{F}$, pragmatic hypotheses are built from the same foundations as distortion models on probability measures [13,14], but applied to hypotheses instead of probability measures. With this link in mind, some suggestions for the dissimilarity function are:

- **L_p distance:** $d_p(G,F) = \left[\int_{x\in\Omega} |G(x) - F(x)|^p dx\right]^{1/p}$, $p \in [1, \infty)$, where Ω is the sample space of the random variable;
- **Kolmogorov or L_∞ distance:** $d_\infty(G,F) = \sup_{x\in\Omega} |G(x) - F(x)|$;
- **Total variation distance:** $d_{TV}(G,F) = \sup_{A\subseteq\Omega} |\mathbb{P}_G(A) - \mathbb{P}_F(A)|$, where $(\mathbb{P}_G, \mathbb{P}_F)$ are the probability measures associated to their respective distribution functions;
- **Kullback-Leibler divergence:** $d_{KL}(G,F) = \int_{x\in\Omega} \log\left(\frac{d\mathbb{P}_F(x)}{d\mathbb{P}_G(x)}\right) d\mathbb{P}_F(x)$;
- **ϵ-contamination or linear vacuous dissimilarity:** Adapting from [15, Theorem 5.1], $d_{LV}(G,F) = \sup_{A\subseteq\Omega:\mathbb{P}_G(A)\neq 0} \frac{\mathbb{P}_G(A) - \mathbb{P}_F(A)}{\mathbb{P}_G(A)}$;
- **Hellinger distance:** $d_H(G,F) = \sqrt{\frac{1}{2}\int |\sqrt{g} - \sqrt{f}|^2 d\mu}$, where μ is a dominating σ-finite measure and (g, f) are the densities of (G, F).

A concept closely related to pragmatic hypotheses is the Region Of Practical Equivalence [ROPE; 5], which has been adopted in fields such as Psychology [16], Medicine [17], Computer Science [18] and Economics [19]. Given a quantity θ of interest, the ROPE is the region such that any $\theta \in$ ROPE is deemed as practically indistinguishable from H_0. Of particular interest to this work are the applications of ROPE to nonparametric settings, where Bayesian versions of the signed and signed-rank tests [20, subsection 4.2.1], as well as an independence test [21], have been proposed. Assuming that $P^* \in \mathbb{H}$ is the data generating process and setting $\theta = \inf_{P_0\in H_0} d(P_0, P^*)$, we conclude that the pragmatic hypothesis and ROPE $= [\min(d), \varepsilon)$ may lead to equivalent tests, though $Pg(H_0, d, \varepsilon) \subseteq \mathbb{H}$ and ROPE $\subseteq \mathbb{R}_{\geq 0}$ are defined on different spaces.

We provide three major contributions to the identification and use of pragmatic hypotheses in practical settings:

1. Propose an intuitive testing procedure that can be seamlessly combined with MCMC methods (`PROTEST`, section 2);
2. Expand the theory of pragmatic hypotheses to nonparametric settings and explore how some hypotheses can be transformed into pragmatic ones (section 3);
3. Provide practical strategies for the choice of ε even when it is not initially clear which value it should assume (section 4). This point is particularly important since defining ε requires careful consideration of both the context and the hypothesis, as its value can heavily influence the results.

To ensure the adequacy of the procedure and demonstrate its applicability, we provide three simulated studies and an application with real data. The first simulated study (subsection 5.1) evaluates if `PROTEST` can recover the true link function of binary data generated from a generalized linear model (GLM). The second (subsection 5.2) is a comparison between `PROTEST` and the PTtest [22] as well as the Kolmogorov-Smirnov test [23]. The last simulated study evaluates the probability of `PROTEST` to commit an error as a function of the sample size. As for the application, it evaluates if data on neuron spikes resembles a Poisson process and if neurons behave differently between experiments (section 6). Lastly, we discuss the potential of these methods and link `PROTEST` to other current research areas such as three-way testing (section 7). In the appendix, we provide an introduction to the nonparametric priors used in this paper (Appendix A) and the proofs to all results presented (Appendix B). An `R` package [24] of `PROTEST` is available at rflassance/protest, also with code of some of the analyses performed.

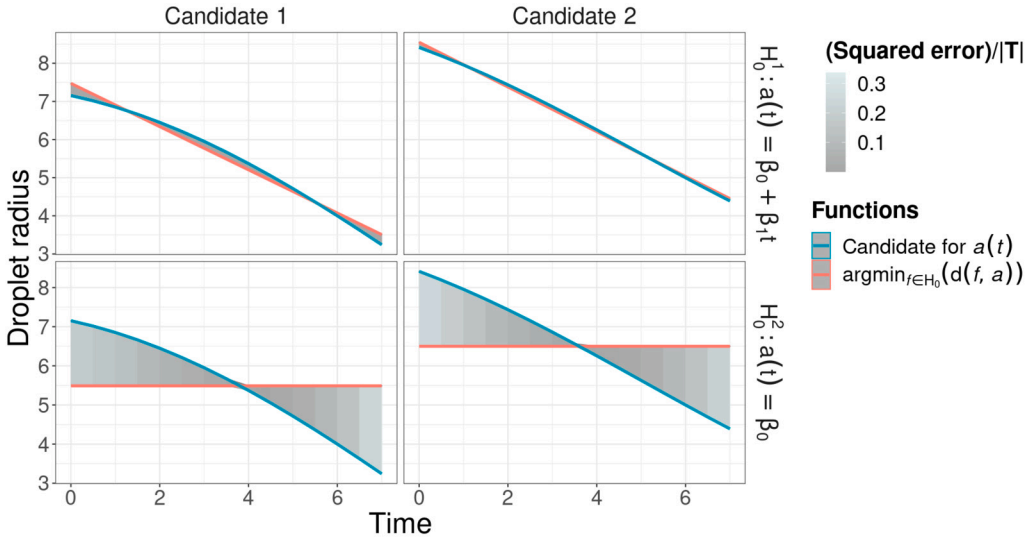


Fig. 1. Candidates (i.e., posterior draws) for $a(\cdot)$ (column) in Example 1 and their best approximations under each hypothesis (row) based on the mean squared error between functions. The scale presents the point-wise squared error divided by $|T|$.

Example 1 *(Water droplet experiment).* The free falling water droplet experiment [25] is a study that evaluates the behavior of small water droplets (ranging from 3 to 9 micrometers) as they fall through a tube in a controlled setting. One of the experiment's main objectives is to test the validity of Fick's law of diffusion, which in this case posits that the radius of the falling droplet changes linearly through time.

As the droplet falls, a camera takes pictures of it every 0.5 second and ceases recording after 7 seconds. Therefore, $T = \{0s, 0.5s, \cdots, 6.5s, 7s\}$ represents the timestamps used as the independent variable. Consequently, two hypotheses of interest are

$$\begin{cases} H_0^1 : a(t) = \beta_0 + \beta_1 t, \quad \forall t \in T, \quad (\beta_0, \beta_1) \in \mathbb{R}^2; \\ H_0^2 : a(t) = \beta_0, \quad \forall t \in T, \quad \beta_0 \in \mathbb{R}. \end{cases}$$

where $a(\cdot)$ represents the radius of the droplet at a given time. The first hypothesis represents Fick's law, while the second evaluates if time can be removed as a covariate.

In Fig. 1, two posterior draws (blue lines) of $a(\cdot)$ are compared to the linear function that best approximates them under H_0^1 and H_0^2. Since the draws are based on the data, each can be thought of as a viable candidate for the true regression function. Therefore, if such candidates for $a(\cdot)$ are close to any function of the null hypothesis, this adds credence to the possibility of it being true.

Using as dissimilarity the square root of the expected squared error between two functions, we derive the linear functions that best approximate each under H_0^1 and H_0^2 (red lines). Following Definition 1, the extent to which any of the linear functions is sufficiently similar to the original function depends on the dissimilarity being less than a threshold, which in this case is $\varepsilon \approx 0.1606$ (see Example 1.3 for the reasoning behind this choice). For both cases, the dissimilarity falls under ε on H_0^1 and over it on H_0^2, suggesting that Fick's law might be applicable for this case and that time should remain as a covariate.

2. Overview

In this section, we define the Pragmatic Region Oriented TEST (`PROTEST`, Definition 2), provide an accessible guide for performing it (`PROTEST` procedure) and introduce some properties of the procedure. A similar idea in the context of rough sets has been presented by [26, subsection 2.2], although restricted to the discrete case and not making use of dissimilarity functions.

Definition 2 *(Pragmatic region oriented test - `PROTEST`).* Let $Pg(H_0, d, \varepsilon)$ be the pragmatic hypothesis, $\mathcal{P}$ be a random function defined on $\mathbb{H}$ (e.g., a probability or a regression function) and $\alpha \in [0, 1]$. `PROTEST` is such that

- If $\mathbb{P}\left(\mathcal{P} \in Pg(H_0) | \boldsymbol{X} = \boldsymbol{x}\right) \leq \alpha$, reject the hypothesis;
- Otherwise, do not reject it.

From Definition 1, we note that

$$\mathbb{P}\left(\mathcal{P} \in Pg(H_0) | \boldsymbol{X} = \boldsymbol{x}\right) = \mathbb{P}\left(\inf_{P_0 \in H_0} d(P_0, \mathcal{P}) < \varepsilon \middle| \boldsymbol{X} = \boldsymbol{x}\right), \tag{2}$$

which implies that the test can be conducted even when the full posterior is unknown or $Pg(H_0)$ cannot be fully specified. As long as $\inf_{P_0 \in H_0} d(P_0, P)$ can be obtained for every $P \in \mathbb{H}$, estimating Equation (2) becomes a matter of sampling from $\mathcal{P}|\boldsymbol{X} = \boldsymbol{x}$ and using the

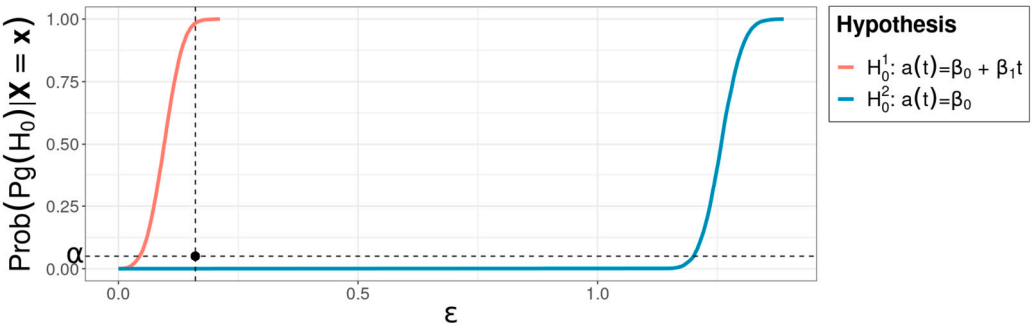


Fig. 2. Largest ε that entails rejection and the posterior probability of $Pg(H_0)$ for each hypothesis (colored curves) in Example 1.1. The black point represents the particular choice of $(\varepsilon \approx 0.1606, \alpha = 0.05)$, leading to not reject H_0^1 and reject H_0^2.

proportion of times in which $\inf_{P_0 \in H_0} d(P_0, \cdot) < \varepsilon$ as an estimate for (2). This is the motivation that leads to the PROTEST procedure, and ensures that it is fully compatible with MCMC methods and does not require knowledge of the full posterior distribution.

In the parametric setting, it is often possible to explicitly identify the pragmatic region since it is a subset of $\mathbb{R}^d$, so a posterior draw belongs to $Pg(H_0)$ if such subset contains it. This is not as straightforward when $\mathbb{H} = \mathbb{F}$, as $\mathcal{P}$ is a random object on the space of distribution functions. When dealing with hypotheses that reside in a function space, a more accessible strategy is to directly obtain $\inf_{P_0 \in H_0} d(P_0, P), \forall P \in \mathbb{H}$, which then allows for Equation (2) to be estimated through a Monte Carlo procedure.

PROTEST procedure

1. Specify the null hypothesis H_0, the level α, the dissimilarity function $d(\cdot, \cdot)$ and the threshold ε.
2. Generate a sample $(\mathcal{P}^{(1)}, \mathcal{P}^{(2)}, \cdots, \mathcal{P}^{(N)})$, $N \in \mathbb{N}$, from the posterior distribution $\mathcal{P}|\boldsymbol{x}$. These draws can be from an MCMC chain if the posterior cannot be explicitly obtained.
3. Obtain

$$\hat{\mathbb{P}}\left(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}\right) = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\inf_{P_0 \in H_0} d(P_0, \mathcal{P}^{(i)}) < \varepsilon\right), \tag{3}$$

   where $\mathbb{I}(\cdot)$ is the indicator function.
4. Reject the hypothesis if the estimated probability in (3) is equal to or less than α.

Example 1.1 *(Water droplet experiment, continued).* Based on the PROTEST procedure, we set $\alpha = 0.05$ and determine ε based on the measurement error of the experiment (see Example 1.3 for more details) to finish the first step. For the second step, we use a Gaussian process [GP; 27] to model the regression function $a(\cdot)$ due to its widespread adoption in hypothesis testing (see [28] and [29] for some examples). More specifically, we apply a Gaussian kernel to the GP, using its posterior to draw regression functions for the test. Since the covariate set T is finite and the posterior is conjugate, we draw directly from the posterior of $a(T)$ through a multivariate normal distribution, dismissing the need for MCMC in this case. For now, we omit how to achieve step 3 (see Example 1.2 for the full discussion). The last step is evident from Fig. 2. Based on the choice of (ε, α), we assert the validity of Fick's law and keep time as a covariate.

Based on Example 1.1, we note that the model selection strategy provided by PROTEST sharply differs from that usually performed through Bayes factors [30,31]. In the latter, one picks a finite number of models and compares them to each other, with the Bayes factor indicating which models fit the data better, but without the possibility of asserting that any of them is actually true or "good enough". PROTEST performs model selection by comparing H_0 to an infinite set of other models, assuming the model implied by the null to be valid if, with high probability, the true model is negligibly different from H_0.

Aside from PROTEST, another Bayesian testing procedure that is also based on significance is the Full Bayesian Significance Test [FBST; 32], having been widely studied [33,34] and adapted to operate even in the presence of nuisance parameters [35]. Similarly to PROTEST, it does not require the researcher to alter their prior distribution so that it includes a probability mass on the precise null hypothesis of interest. However, instead of basing its decisions directly on the posterior probability of the hypothesis, the FBST uses the highest posterior density set, rejecting the hypothesis if it is not included in the set. While the original formulation of the FBST does not account for negligible deviations from a precise null hypothesis, there is an extension of it which has been applied directly to pragmatic hypotheses [12].

An alternative to both PROTEST and the FBST are tests based on the ROPE [5], which can be similar to either depending on its formulation. In nonparametric settings, there are tests based on the ROPE that use its posterior probability directly [20,21], making it more akin to PROTEST. Outside of nonparametrics, the ROPE is usually combined with the highest density interval of the parameter

Table 1
0-1-c loss function for $Pg(H_0)$.

Truth state \ Decision	Accept $Pg(H_0)$	Reject $Pg(H_0)$
$Pg(H_0)$	0	1
$Pg(H_0)^c$	c	0

being tested, which is the same as the highest posterior density set when the latter is an interval. Both testing schemes based on the ROPE have been provided with decision-theoretic foundations [36].

2.1. Properties of PROTEST

Bayes decision PROTEST is the Bayes decision for the 0-1-c loss function [37, page 215]. The Bayes decision is defined as the decision that minimizes the expected loss, and the 0-1-c loss is given by Table 1.

The expected loss if one rejects $Pg(H_0)$ is $\mathbb{P}(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x})$ and $c \times \mathbb{P}(\mathcal{P} \in Pg(H_0)^c|\boldsymbol{X} = \boldsymbol{x})$ if one accepts it. Thus, the Bayes decision is to reject $Pg(H_0)$ when

$$\frac{\mathbb{P}(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x})}{c \times \mathbb{P}(\mathcal{P} \in Pg(H_0)^c|\boldsymbol{X} = \boldsymbol{x})} \leq 1 \Longrightarrow \mathbb{P}(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}) \leq \frac{c}{1+c}.$$

Assuming that $\alpha = \frac{c}{1+c}$, we conclude that PROTEST is indeed the Bayes decision.

Monotonicity In Example 1.1, we performed two tests in sequence—H_0^1 and then H_0^2, with H_0^2 being more specific than H_0^1—while keeping α constant between tests. The following result ensures that PROTEST cannot reach the counterintuitive conclusion of rejecting H_0^1 but not H_0^2 when $H_0^1 \supseteq H_0^2$ and α is fixed. Furthermore, when $H_0^1 \not\supseteq H_0^2$ but their pragmatic versions are nested (i.e., the threshold associated with H_0^1 induces a set large enough to contain both H_0^2 and the vicinity implied by its respective threshold), then the result still holds.

Corollary 1 *(Monotonicity property of PROTEST). Let $H_0^1, H_0^2 \subset \mathbb{H}$ be such that*

$$Pg(H_0^1, d, \varepsilon_1) \supseteq Pg(H_0^2, d, \varepsilon_2)$$

and take $\alpha \in [0, 1]$. If PROTEST leads to rejecting $Pg(H_0^1)$, then it also rejects $Pg(H_0^2)$.

Posterior error control PROTEST offers some form of error control guarantee. Its posterior error rate, which is defined as the probability of $Pg(H_0)$ being true conditional on PROTEST rejecting it, is dominated by α. This result is not to be confused with the frequentist type I error.

Proposition 1. *Let $\phi(\boldsymbol{x}) = \mathbb{I}\{\mathbb{P}(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}) \leq \alpha\}$ and set the posterior error rate as $e(\boldsymbol{x}) = \mathbb{P}(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}, \phi(\boldsymbol{x}) = 1)$. Then, $e(\boldsymbol{x}) \leq \alpha$ when $\phi(\boldsymbol{x}) = 1$, and thus $\mathbb{E}[e(\boldsymbol{X})|\phi(\boldsymbol{X}) = 1] \leq \alpha$.*

Consistency Let $P \in \mathbb{H}$ be the data generating process. Under some regularity conditions, PROTEST converges asymptotically to the correct decision as long as P is not in the border of $Pg(H_0, \varepsilon)$, that is, $\inf_{P_0 \in H_0} d(P_0, P) \neq \varepsilon$. Next, we proceed to define the regularity conditions and then present the convergence result.

Assumption 1. *For every $0 < \varepsilon_1 < \varepsilon_2$, there exists $\delta(\varepsilon_1, \varepsilon_2) > 0$ such that, if $d(P_0, P_1) = \varepsilon_1$ and $\min(d(P_1, P_2), d(P_2, P_1)) < \delta(\varepsilon_1, \varepsilon_2)$, then $d(P_0, P_2) < \varepsilon_2$.*

Assumption 1 mimics the triangular inequality without being as restrictive. It admits a topological interpretation: For every point, P, in a neighborhood of P_0, N_{P_0}, there exists a neighborhood of P that is contained in N_{P_0}. When d is a distance function, this assumption follows immediately from symmetry and the triangular inequality.

Assumption 2. *Let $Pg^*(P_0, \varepsilon) = \{P \in \mathbb{H} : d^*(P_0, P) < \varepsilon\}$. There exists d^* such that,*

a) For every $P_0 \in \mathbb{H}$ and for every ε, there exists ε^* such that $Pg^*(P_0, \varepsilon^*) \subseteq Pg(P_0, \varepsilon)$.
b) For every $P_0 \in \mathbb{H}$ and ε^*, if $(X_n)_{n \in \mathbb{N}}$ are i.i.d. and $X_i \sim P$, then

$$\mathbb{P}(Pg(P_0, \varepsilon)|X_1, \ldots, X_n) \xrightarrow{n \to \infty} \mathbb{I}(P \in Pg(P_0, \varepsilon))$$

Assumption 2 requires the existence of a dissimilarity, d^*, with two properties. First, for every ε-neighborhood of H_0 according to d, there exists a ε^*-neighborhood of H_0 according to d^*, such that the latter is contained in the former. That is, for every neighborhood

according to d, there exists a smaller neighborhood according to d^*. Second, Assumption 2 requires the posterior consistency for every neighborhood according to d^*, that is, as the sample size increases, the posterior of the neighborhood goes to 1 if the data-generating mechanism belongs to the neighborhood, and goes to 0, otherwise. When d^* is the Hellinger distance, $d_H(f_1, f_2)$, [38] shows that random histograms, Pólya trees, and infinite-dimensional exponential families satisfy Assumption 2.b. Also, when d is based on the total variation distance, $d_{TV}(f_1, f_2)$ or L_2 distance, $d_{L2}(f_1, f_2)$, one can obtain Assumption 2.a by using inequalities such as $d_H(f_1, f_2) \leq d_{TV}(f_1, f_2) \leq 0.5 \times d_{L_2}(f_1, f_2)$.

Under Assumption 1 and Assumption 2, PROTEST's type I and type II error rates converge to 0, as shown below:

Theorem 1. *Under Assumption 1 and Assumption 2,*

(a) For every P such that $d(P_0, P) < \varepsilon$, if the data is generated according to P, then PROTEST's type I error rate converges to 0 as the sample size increases.

(b) For every P such that $d(P_0, P) > \varepsilon$, if the data is generated according to P, then PROTEST's type II error rate converges to 0 as the sample size increases.

Using Theorem 1, one can obtain the consistency of PROTEST in a variety of settings. For instance, under a Pólya Tree prior, if $d(f_0, f_1) = \sqrt{\int (f_0(x) - f_1(x))^2 dx}$ and d^* is the Hellinger distance, consistency follows directly from Theorem 1 and the convergence results in [38]. Similar convergence results can be found when using other nonparametric priors, such as random histograms, Dirichlet processes, and Gaussian processes [39]. For many of these nonparametric priors, there exist further results that guarantees that the error rates converge exponentially to 0 as the sample size increases. [40,41].

3. Nonparametric pragmatic hypotheses

In this section, we transform some common nonparametric hypotheses into pragmatic ones. From Definition 1, this can be achieved by finding the infimum of the dissimilarity function between H_0 and any given $P \in \mathbb{H}$. We use the data to draw specific elements from $\mathbb{H}$ and the infimum to check if they belong to $Pg(H_0)$, rejecting the hypothesis if less than $\alpha \times 100\%$ of them do. We can produce such elements by, for example, sampling from the Dirichlet or the Pólya tree processes [42–44].

In some cases, the infimum can be obtained analytically and for a wide range of dissimilarity functions (such as in subsection 3.4), while in others it requires an optimization procedure (subsection 3.2) or the choice of a specific dissimilarity (subsection 3.1, subsection 3.3). Whenever possible, the choice of the dissimilarity function should be based on how the researcher can best elicit their knowledge and interests about a problem. If that is not initially clear, we recommend the use of the classification dissimilarity due to its intuitive appeal.

Definition 3 *(Nonparametric classification dissimilarity function).* If $\mathbb{H} = \mathbb{F}$ and $F, G \in \mathbb{F}$ are distribution functions, the classification dissimilarity is given by

$$d_C(G, F) := 0.5 \left[\mathbb{P}\left(\frac{g(Z)}{f(Z)} \geq 1 \middle| Z \sim G \right) + \mathbb{P}\left(\frac{f(Z)}{g(Z)} > 1 \middle| Z \sim F \right) \right] \in [0.5, 1], \tag{4}$$

where Z is a future observation, while f and g are the respective density functions of F and G.

The idea behind Equation (4) is as follows: say that there are two possible distribution functions (F or G) that could be used to generate the future observation Z, and that there is no reason to assume one is more likely than the other, so $\mathbb{P}(Z \sim F) = \mathbb{P}(Z \sim G) = 0.5$. If the criterion for deciding from which distribution Z came from is the likelihood ratio (LR), then one should choose G when $\frac{g(Z)}{f(Z)} > 1$ and F when $\frac{f(Z)}{g(Z)} > 1$. When $\frac{g(Z)}{f(Z)} = 1$ both distributions are equally favored, so every choice is justifiable and, in particular, the choice for G. In this case, the accuracy of the decision is:

$$\begin{aligned} &\mathbb{P}\left(\left(\frac{g(Z)}{f(Z)} \geq 1 \cap Z \sim G \right) \cup \left(\frac{f(Z)}{g(Z)} > 1 \cap Z \sim F \right) \right) \\ =&0.5 \times \mathbb{P}\left(\frac{g(Z)}{f(Z)} \geq 1 | Z \sim G \right) + 0.5 \times \mathbb{P}\left(\frac{f(Z)}{g(Z)} > 1 | Z \sim F \right) \end{aligned}$$

Moreover, thanks to the Neyman-Pearson lemma [45], we conclude that the classification dissimilarity provides the highest achievable probability of correctly identifying which distribution function generated Z.

Lastly, we note that d_C obeys the conditions required for Theorem 1. Since $d_C(G, F) = 0.5 \times d_{TV}(G, F) + 0.5$ [12] and d_{TV} is a distance function, Assumption 1 follows. As for Assumption 2.a, if $0.5 \times d_H^2 + 0.5 \leq 0.5 \times d_{TV} + 0.5 = d_C < \varepsilon$, then $d_H < \sqrt{2\varepsilon - 1} =: \varepsilon^*$.

We present pragmatic versions of a model adherence test based on linear predictors (subsection 3.1), the goodness-of-fit test (subsection 3.2), the quantile test (subsection 3.3) and the two-sample test (subsection 3.4). Whenever required, we apply a subscript to $\mathbb{F}$ to avoid ambiguity on what is the random variable being referenced. For example, if $X \in \mathbb{N}$, $\mathbb{F}_X$ may contain a Poisson distribution, but not a Normal distribution.

3.1. Model adherence test

We begin with a test focused on regression models applicable to data $(\boldsymbol{y}, \boldsymbol{X})$, and therefore $\mathbb{H}$ is a space of functions of the type $g : \mathcal{X} \longrightarrow \mathbb{R}$, where $\mathcal{X}$ is the covariates' domain. Our main finding shows how to analytically obtain the pragmatic hypothesis when comparing a function to a class of linear models.

Theorem 2 *(Linear model test). Let $\mathbb{H}$ be such that*

$$g \in \mathbb{H} \iff \mathbb{E}_X(g^2) = \int_{\mathcal{X}} g(\boldsymbol{x})^2 d\mathbb{P}(\boldsymbol{x}) < \infty.$$

Let $\boldsymbol{b}(\boldsymbol{x}) = (b_1(\boldsymbol{x}), b_2(\boldsymbol{x}), \cdots, b_k(\boldsymbol{x})) \subset \mathbb{H}$ be a linearly independent set of linear functions and choose H_0 such that

$$H_0 : R(\boldsymbol{x}) = \boldsymbol{b}(\boldsymbol{x})\boldsymbol{\beta}, \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{\beta} \in \mathbb{R}^k, \tag{5}$$

where $R(\cdot)$ is the true regression function. If $d(f,g) := \sqrt{\mathbb{E}_X[(f-g)^2]}$, then $d(H_0, g) = d(\boldsymbol{b} \times \hat{\boldsymbol{\beta}}, g)$ for any $g \in \mathbb{H}$, where

$$\hat{\boldsymbol{\beta}} = A_{\boldsymbol{b}}^{-1} \times \boldsymbol{g}_{\boldsymbol{b}}, \quad A_{\boldsymbol{b}} = \begin{pmatrix} \mathbb{E}[b_1^2(\boldsymbol{X})] & \mathbb{E}[b_2(\boldsymbol{X})b_1(\boldsymbol{X})] & \cdots & \mathbb{E}[b_k(\boldsymbol{X})b_1(\boldsymbol{X})] \\ \mathbb{E}[b_1(\boldsymbol{X})b_2(\boldsymbol{X})] & \mathbb{E}[b_2^2(\boldsymbol{X})] & \cdots & \mathbb{E}[b_k(\boldsymbol{X})b_2(\boldsymbol{X})] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[b_1(\boldsymbol{X})b_k(\boldsymbol{X})] & \mathbb{E}[b_2(\boldsymbol{X})b_k(\boldsymbol{X})] & \cdots & \mathbb{E}[b_k^2(\boldsymbol{X})] \end{pmatrix},$$

$$\boldsymbol{g}_{\boldsymbol{b}}' = \Big(\mathbb{E}[g(\boldsymbol{X})b_1(\boldsymbol{X})], \quad \mathbb{E}[g(\boldsymbol{X})b_2(\boldsymbol{X})], \quad \cdots \quad , \quad \mathbb{E}[g(\boldsymbol{X})b_k(\boldsymbol{X})] \Big).$$

Some lingering aspects of Theorem 2 require additional explanations. Framing the hypothesis as the span of linearly independent functions allows for testing a diverse set of assumptions, some of them being: $\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{x}$ (standard linear regression), $\boldsymbol{b}(\boldsymbol{x}) = \boldsymbol{x}_{-i}$ (removal of the i-th entry of the vector, thus providing a variable selection procedure) and $\boldsymbol{b}(\boldsymbol{x}) = (x_1 + x_2, \boldsymbol{x}'_{-\{1,2\}})'$ (first two entries receive the same parameter β). As for the choice of the probability measure $\mathbb{P}$, if the context of the problem is not sufficient to imply one, we suggest using the empirical distribution of $\boldsymbol{b}(\boldsymbol{X})$, which leads to $\hat{\boldsymbol{\beta}} = (\boldsymbol{b}(\boldsymbol{X})'\boldsymbol{b}(\boldsymbol{X}))^{-1}\boldsymbol{b}(\boldsymbol{X})'g(\boldsymbol{X})$.

Example 1.2 *(Water droplet experiment, continued).* As mentioned in Example 1, the covariate T is a discrete variable and all times are recorded in the experiment, therefore it is reasonable to assign a discrete uniform distribution to it. Hence

$$d(f,g) = \sqrt{\int_{\mathcal{X}} |f(\boldsymbol{x}) - g(\boldsymbol{x})|^2 d\mathbb{P}(\boldsymbol{x})} = \sqrt{\frac{1}{15}\sum_{t \in T} |f(t) - g(t)|^2},$$

which is a weighted version of the l_2 distance. From Theorem 2 and assuming T to be a column vector, $\boldsymbol{b}(\boldsymbol{X}) = (\boldsymbol{1}, T)$ for H_0^1 and $\boldsymbol{b}(\boldsymbol{X}) = \boldsymbol{1}$ for H_0^2.

Once again, we use the Gaussian process to model the data, this time applying different kernels (exponential, Gaussian and Matérn) from the default settings of the `GauPro R` package [46] for more robust results. Since T is discrete, it is only necessary to obtain draws of the regression function at its values. The remaining steps of the `PROTEST` procedure lead us to apply Theorem 2 to obtain the dissimilarity between each draw and H_0 and then take the proportion of times such dissimilarity is less than 0.1606 to reach a decision.

Fig. 3 presents the results of both tests, leading to non-rejection for H_0^1 and to rejection for H_0^2. For H_0^1, Fig. 3a shows that the significance level required to reject the hypothesis would be of at least 0.25, leading to the conclusion that the droplet radius can indeed be described as a linear function of the time. As for H_0^2, the threshold of choice leads to rejection in all cases, with a considerably higher value required for concluding otherwise. Hence, not only can the data be described by a linear model, but it also requires time to be kept as a covariate.

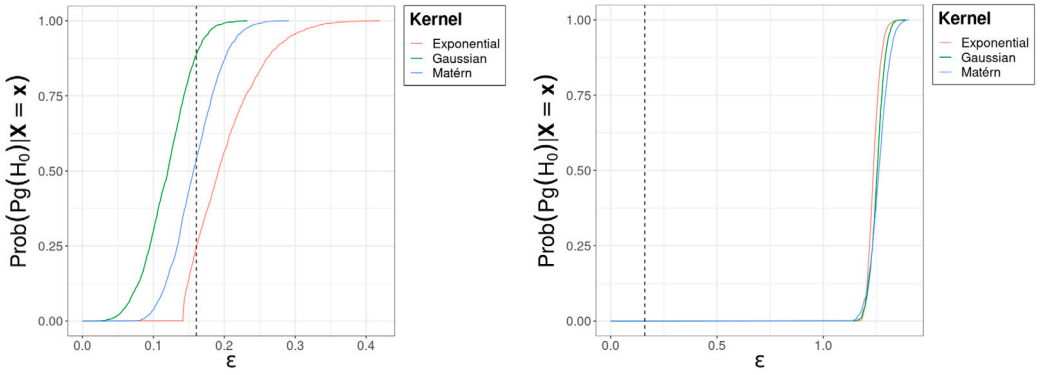
Going beyond linear regression, Theorem 2 can also be used for testing models whose regression function depends on a linear combination of $\boldsymbol{b}(\boldsymbol{x})$, such as GLMs. For a known function $h(\cdot)$, the same test can be performed by switching the hypothesis in Equation (5) for

$$H_0 : R(\boldsymbol{x}) = h(\boldsymbol{b}(\boldsymbol{x})\boldsymbol{\beta}) \iff H_0 : h^{-1}(R(\boldsymbol{x})) = \boldsymbol{b}(\boldsymbol{x})\boldsymbol{\beta}, \quad \forall \boldsymbol{x} \in \mathcal{X}, \quad \boldsymbol{\beta} \in \mathbb{R}^k, \tag{6}$$

as long as $h^{-1}(\cdot)$ can be obtained. In subsection 5.1, we use this strategy in a simulated setting to check for adherence when the response variable is binary.

3.2. Goodness-of-fit test

Let $H_0 : X \sim F$, where F is a fixed distribution function. Then, for a threshold ε and a dissimilarity function $d(\cdot,\cdot)$, the pragmatic hypothesis is given by



(a) $H_0^1 : a(t) = \beta_0 + \beta_1 t$ (b) $H_0^2 : a(t) = \beta_0$

Fig. 3. Largest ε that entails rejection and the posterior probability of $Pg(H_0)$ for each kernel of the Gaussian process in Example 1.2. The dashed line marks the threshold value ($\varepsilon \approx 0.1606$).

$$Pg(H_0) = \{P \in \mathbb{F} : d(F, P) < \varepsilon\}, \tag{7}$$

since F is the only distribution function that belongs to H_0. Thus, a goodness-of-fit test can be executed through `PROTEST`, with the problem of the dissimilarity being reduced to that of obtaining $d(F, \cdot)$.

Going beyond a single distribution function, we can also determine the pragmatic hypothesis for a parametric family. If $\theta \in \Theta$ is the parameter vector of such family, the null hypothesis is

$$H_0 : X \sim F_\theta, \quad \theta \in \Theta_0 \subseteq \Theta.$$

Hence, the pragmatic hypothesis is

$$Pg(H_0) = \left\{ P \in \mathbb{F} : \inf_{\theta \in \Theta_0} d(F_\theta, P) < \varepsilon \right\}. \tag{8}$$

This means that the process of identifying if a candidate $P \in \mathbb{F}$ belongs to $Pg(H_0)$ can be translated into an optimization procedure. For every given P, the objective is to find $\hat{\theta} \in \Theta_0$ such that $d(F_{\hat{\theta}}, P) \leq d(F_\theta, P), \forall \theta \in \Theta_0$. Then, if $\hat{\theta}$ provides a dissimilarity smaller than ε, we conclude that $P \in Pg(H_0)$.

Example 2 *($H_0 : X \sim Exp(1/\lambda), \lambda \in \mathbb{R}^+$).* In this setting, Equation (8) implies that the pragmatic hypothesis is

$$Pg(H_0) = \left\{ P \in \mathbb{F} : \inf_{\lambda \in \mathbb{R}^+} d(F_\lambda, P) < \varepsilon \right\}, \tag{9}$$

where $F_\lambda \equiv Exp(1/\lambda)$.

Choosing the L_∞ distance—the same used in the Kolmogorov-Smirnov test—for (9), it would be represented as

$$d_\infty(F_\lambda, P) = \sup_{x \in \mathbb{R}_{\geq 0}} |F_\lambda(x) - P(x)| = \sup_{x \in \mathbb{R}_{\geq 0}} |1 - \exp(-\lambda x) - P(x)|. \tag{10}$$

Therefore, $P \in Pg(H_0) \iff \exists \lambda \in \mathbb{R}^+ : d_\infty(F_\lambda, P) < \varepsilon$. Since P is fixed, such condition can be verified through an optimization procedure by finding the value for λ that minimizes $d_\infty(F_\lambda, P)$, which is achievable through general optimization routines such as the `optim` function in `R` [24]. This exact test is carried out in subsection 6.1.

We end this subsection highlighting the contribution of the ideas presented here to the field of robust statistics and distortion models more generally. Both building an imprecise probability model around a probability measure P_0 [14] and proposing a neighborhood around P_θ to then try to estimate θ [13] rely on the assumption that there is at least one probability distribution inside the neighborhood capable of adequately representing the data. Since the models are respectively represented in Equations (7) and (8), `PROTEST` can evaluate if such proposals should be used or not.

3.3. Quantile test

In this section, we propose a quantile test that only requires the data to come from a continuous distribution. Let x_0 and p_0 be such that $\mathbb{P}(X \leq x_0) = p_0$, i.e., x_0 is the p_0-quantile of X if $\mathbb{P}$ is its true probability measure. If $\mathbb{F}_X$ represents the set of distribution functions compatible with X, the hypothesis of interest for this case is

$$H_0 : F(x_0) = p_0, \quad F \in \mathbb{F}_X.$$

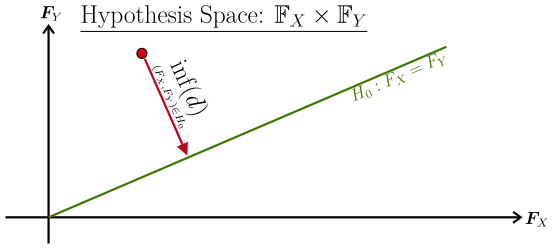


Fig. 4. Representation of the nonparametric two-sample hypothesis. The green line is the original null hypothesis, while the red dot is a pair of distribution functions.

Closed-form solutions for this hypothesis depend on the dissimilarity function of choice. Let

$$d_1(F,G) := \|F-G\|_1 = \int_{\mathbb{R}} |F(x)-G(x)|dx, \quad F,G \in \mathbb{F}. \tag{11}$$

The following theorem provides a straightforward procedure for obtaining the pragmatic hypothesis when (11) is the dissimilarity function.

Theorem 3 *(Quantile test). Let $H_0 : F(x_0) = p_0$, $F \in \mathbb{F}_X$, be the null hypothesis and take (11) as the dissimilarity function. Then, $\forall P \in \mathbb{F}_X$, if $a := \min(P^{-1}(p_0), x_0)$ and $b := \max(P^{-1}(p_0), x_0)$,*

$$\inf_{P_0 \in H_0} d(P_0, P) = \inf_{P_0 \in H_0} \int_{-\infty}^{\infty} |P_0(x) - P(x)|dx = \int_a^b |p_0 - P(x)|dx. \tag{12}$$

If for some reason (12) cannot be analytically obtained, a Monte Carlo integration procedure [47] could be used. This result is applied in subsection 6.2.

3.4. Two-sample test

In this section, we provide a pragmatic version of the nonparametric two-sample test, a test whose hypothesis originally states that the true distribution functions of two different datasets are the same. In other words, if X and Y are the random variables of interest and F_X and F_Y are their respective distribution functions, then

$$H_0 : F_X = F_Y = F, \quad F \in \mathbb{F},$$

is the hypothesis which we seek to expand.

We highlight that $\mathbb{H} = \mathbb{F} \times \mathbb{F}$, i.e., the hypothesis space is the Cartesian product of the space of distribution functions. In Fig. 4, a visualization is provided to give an idea of the peculiarities of such space. Each axis of the figure represents the distribution function of a specific population. Then, the green line represents the null hypothesis that both distributions are equal. Thus, while the red dot is an element of $\mathbb{H}$ (i.e., a given pair of distribution functions), the red arrow represents the smallest distance between such element and H_0.

The following result provides an analytical solution for the infimum that is solely based on the distance between the functions obtained from the data.

Theorem 4 *(Two-sample test). Let $H_0 : F_X = F_Y = F$, $F \in \mathbb{F}$, be the null hypothesis, and (P_X, P_Y) be a pair of distribution functions. If the dissimilarity function $d(\cdot,\cdot)$ is such that*

$$d[(F_X,F_Y),(P_X,P_Y)] = d^*(F_X,P_X) + d^*(F_Y,P_Y), \tag{13}$$

where $d^*(\cdot,\cdot)$ is a distance function, then

$$\inf_{(F_X,F_Y)\in H_0} d[(F_X,F_Y),(P_X,P_Y)] = d^*(P_X,P_Y).$$

More than simply identifying the infimum for a given dissimilarity d, Theorem 4 provides a solution that works for any distance function d^*, while keeping the intuitive appeal of reaching a decision solely based on the discrepancy between the distribution functions of X and Y, $d^*(F_X,F_Y)$. Such appeal can be observed in both classical statistical tests—such as the Kolmogorov-Smirnov test—and more recent iterations [48,49]. Moreover, our version can be seen as an enhancement of the Kolmogorov-Smirnov test, since it allows for the choice of other distance functions and takes model uncertainty into account.

Since the theorem makes no restriction on the choice of the distance function d^*, the classification dissimilarity (Definition 3) could be used in this case if we subtract it by 0.5, i.e.,

$$d_C^*(F_X, F_Y) = 0.5\left[\mathbb{P}\left(\frac{f_X(Z)}{f_Y(Z)} \geq 1 \middle| Z \sim F_X\right) + \mathbb{P}\left(\frac{f_Y(Z)}{f_X(Z)} > 1 \middle| Z \sim F_Y\right)\right] - 0.5, \tag{14}$$

where f_X and f_Y are the respective density functions of F_X and F_Y. Equation (14) is a distance function and is used in the simulated study (subsection 5.2).

4. On choosing the threshold ϵ

The current lack of standards and guidelines for establishing the threshold ϵ is the main drawback for researchers that seek to enlarge their hypotheses, so it is imperative to derive suggestions for ϵ that can be more generally applied. Although some solutions have been proposed to specific problems [11,4,50,5,51], none of them offer strategies for determining the threshold in more general settings, such as when dealing with nonparametric hypotheses.

Although we provide general suggestions on how to choose ϵ based on the type of intuition a researcher has, these suggestions serve more as a starting point for discussions. Ideally, the value of ϵ should reflect a utility judgment of the researcher, their notion of what results should be indistinguishable from the null hypothesis in practice. Considering this dependence between ϵ and the practical considerations of experts, using a threshold at all includes a new layer of subjectivity that can affect reproducibility, and thus must be carefully considered.

If, instead of making a decision, the objective is to report on the robustness of an effect, then the threshold can be used to other ends. For instance, if the significance level is fixed, one can report the largest ϵ that entails rejecting the hypothesis, so that each researcher can compare it to their own threshold and reach their conclusions. If both the significance level and the threshold can differ between researchers, the last suggestion in subsection 4.2 can convey the decision that each of them should take.

4.1. Intuitions that lead to ϵ

We begin by presenting suggestions that, if followed, are assertive enough to establish a unique value for ϵ. They consist of:

Using theory or measurement errors In this case, there is external information available to determine ϵ, coming either through theoretical assumptions, knowledge of measurement errors or both. The scope of possible dissimilarity functions for this case would then be limited to those that can use the information on ϵ to their advantage.

Example 1.3 *(Water droplet experiment, continued).* This last part of the example uses known results of Physics and more details from the original experiment [25] to determine a value for ϵ. It consists of enumerating the sources of error in the experiment, estimating the magnitude of the total error and then deriving a threshold that is compatible with the L_2 distance.

Identifying the sources of error: While the objective of the study is to evaluate the radius of droplets through time, the radius itself was not measured directly. Instead, Stoke's law was used to estimate it based on the velocity. It states that

$$V_T(t) = \frac{a(t)^2}{K_s} \Longrightarrow a(t) = \sqrt{V_T(t) \times K_s}, \tag{15}$$

where V_T is the terminal velocity and K_s is a known constant that depends on factors such as temperature and humidity (in this case, $K_s = 8.446$). However, V_T was not registered in the experiment, with the mean velocity (V_M) being used instead since it can be inferred from the pictures of the camera. Furthermore, since a square grid placed in front of the apparatus was used as a ruler to estimate the droplet's position at a given time, there is also a measurement error regarding V_M. Therefore, the experiment possesses two sources of error: one of unknown magnitude due to switching V_T for V_M in (15) and one due to measuring V_M imprecisely, whose magnitude is of at most $\delta = 0.14$.

Estimating the total error: To estimate the magnitude of the error due to using V_M instead of V_T, we first estimate V_T. Let $S(t)$ be the position of the droplet at time $t \in T$ and define

$$\Delta_i S(t) := \frac{S(t - 0.5i) - S(t + 0.5i)}{i}, \quad 1 \leq i \leq k, \quad t \in T, \tag{16}$$

as the symmetric difference quotient, where $k < |T|$ is fixed beforehand. Setting $\mathbf{\Delta} S(t) = (\Delta_1 S(t), \Delta_2 S(t), \ldots, \Delta_k S(t))'$, if

$$\hat{\boldsymbol{v}}^t = (D'WD)^{-1} D'W\mathbf{\Delta} S(t), \quad D = \left(\mathbf{1}, \frac{(1:k)^2}{|T|}\right), \quad W = diag\left(\frac{(1:k)^2}{|T|^2}\right),$$

where $\mathbf{1}$ is the unit column vector, then the first entry of $\hat{\boldsymbol{v}}^t$ is an estimate for $V_T(t)$ [52, subsection 2.2]. A similar strategy can be used when t is such that Equation (16) is not defined, but its estimates are less accurate and are disregarded for the rest of the analysis. Following the suggestion that $k < |T|/4$, we set $k = 4$.

Setting $\eta := \max_{t \in T} |V_T(t) - V_M(t)|$ as our estimate for the error of unknown magnitude, then $\exists h \in [-(\delta + \eta), (\delta + \eta)]$ such that

$$a(t) = \sqrt{K_s V_T(t)} = \sqrt{K_s (V_M(t) + h)}, \quad t \in T. \tag{17}$$

If $\hat{a}(t) = \sqrt{K_s V_M(t)}$ is the estimate of the radius at time $t \in T$, the margin of error of the radius is

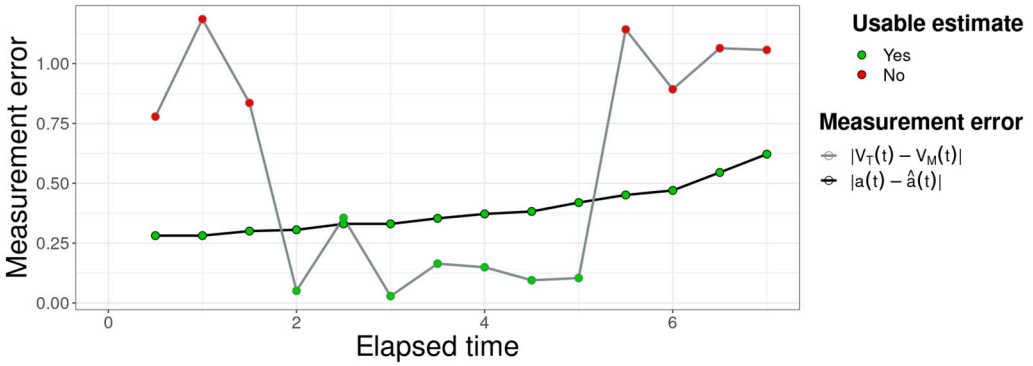


Fig. 5. Measurement errors of velocity and radius. Red dots represent poor estimates of $|V_T(t) - V_M(t)|$, green dots represent better ones. The largest green estimate of $|V_T(t) - V_M(t)|$ was the one plugged into $|a(t) - \hat{a}(t)|$.

$$\begin{aligned}\epsilon &:= \max_{t\in T, h\in[-(\delta+\eta),(\delta+\eta)]} |a(t) - \hat{a}(t)| \\ &= \max_{t\in T} \left\{ \left| \sqrt{K_s(V_M(t) - \delta - \eta)} - \hat{a}(t) \right|, \left| \sqrt{K_s(V_M(t) + \delta + \eta)} - \hat{a}(t) \right| \right\}.\end{aligned}$$

Fig. 5 shows the respective errors for both $|V_T(t) - V_M(t)|$ and $|a(t) - \hat{a}(t)|$. Excluding the first and last k estimates of V_T, $\eta = \max_{t\in T} |V_T(t) - V_M(t)| \approx 0.3555$. By plugging η in $|a(t) - \hat{a}(t)|$, we conclude that $\epsilon = \max_{t\in T} |a(t) - \hat{a}(t)| \approx 0.6218$.

Deriving a compatible threshold: The last step required for reaching the threshold ε is to adapt ϵ—which is related to the l_∞ distance—to the dissimilarity function of interest, a weighted version of the l_2 distance. Proposition 6.11 of [53] ensures that $l_2 \subset l_\infty$, therefore

$$\inf_{\beta\in\mathbb{R}^p} \sqrt{\frac{1}{n}\sum_{i=1}^{n} (\boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{x}_i))^2} \leq \sqrt{\frac{1}{n}}\epsilon \implies \inf_{\beta\in\mathbb{R}^p} \max_{i\in\{1,2,\cdots,n\}} |\boldsymbol{x}_i'\boldsymbol{\beta} - g(\boldsymbol{x}_i)| \leq \epsilon.$$

Thus, using $\varepsilon = \sqrt{\frac{1}{n}}\epsilon \approx 0.1606$ as the threshold for the l_2 distance leads to the same conclusion as using ϵ for the l_∞ distance when not rejecting the hypothesis.

Setting the threshold through the prior While specifying a probability mass to a point null hypothesis might be ill-advised if that does not represent the researcher's belief [3, page 21], attributing a prior probability to the pragmatic hypothesis itself is a valid possibility. If that is the case, we can then use this to obtain the threshold by checking which value ε should assume to match the prior. Formally, let $\mathbb{P}(\mathcal{P} \in Pg(H_0)) = \delta$ be the prior probability of $Pg(H_0)$ being true. Since

$$\mathbb{P}(\mathcal{P} \in Pg(H_0)) = \mathbb{P}\left(\inf_{P_0\in H_0} d(P_0, \mathcal{P}) < \varepsilon \right) = \delta \iff Q_\delta \left(\inf_{P_0\in H_0} d(P_0, \mathcal{P}) \right) = \varepsilon,$$

where $Q_\delta(\cdot)$ is the δ-quantile function, then ε is uniquely determined through the choice of δ and the prior over $\mathcal{P}$.

When the prior over $\mathcal{P}$ is informative but a value for δ is not clear, we can use the fact that the prior uncertainty is greater than the posterior uncertainty to our advantage. By taking $\delta = \alpha$, it is expected that $Q_\delta\left(\inf_{P_0\in H_0} d(P_0, \mathcal{P}|\boldsymbol{X} = \boldsymbol{x})\right)$ should be smaller than ε when $Pg(H_0)$ is true and greater when it is false. This suggestion is applied in subsection 5.2.

Exploring this suggestion further, there is a connection between `PROTEST` and Bayes factors that allows for practitioners to make decisions without setting neither ε nor α directly. By choosing $\mathbb{P}\left(\mathcal{P} \in Pg(H_0)\right) = 0.5$ and using this to determine ε, the Bayes factor for the pragmatic hypothesis is such that, for data $\boldsymbol{x} \in \Omega$,

$$BF = \frac{\mathbb{P}\left(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}\right)}{1 - \mathbb{P}\left(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}\right)}. \tag{18}$$

Since $\mathbb{P}\left(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}\right)$ can be obtained through `PROTEST`, one can then plug its value in (18) to obtain the Bayes factor. Since there are suggestions available on how to interpret the values of Bayes factors in terms of evidence against the null [30, subsection 3.2], this strategy is closer to what is already routinely performed by Bayesian practitioners.

Building from related studies Say that there is at least one study in the literature with positive results which can be used as reference for your own study. Since the interest here is to provide a direct comparison between their findings and yours, apply the same model and the same significance level α of your study to their data, choosing the smallest ε that leads to non-rejection. If there are multiple studies, obtain the smallest ε_i that leads to non-rejection in the i-th study, do the same for all studies and then pick $\varepsilon = \max_i \varepsilon_i$, so that none of the studies is rejected.

This approach is particularly useful for reproducibility research, since newer studies tend to have a larger sample and data with higher quality than the old one, so the same conclusion should be reached if the hypothesis is true. Other cases where this approach might be reasonable are when there has been observed an effect for a given group (geographical region, social class, species, etc.) and we wish to check if the same effect exists for a different group. A similar idea is found in Section 9.12 of [54].

Example 3 *(Worldwide gender wage gap).* The gender wage gap is a multifaceted issue that remains harming women in the workforce for the last 200 years [55], even though some advances have been made to reduce it [56]. Let X represent the difference between the wage gap of two consecutive years and let the null hypothesis be

$$H_0 : F_X(0) = 0.25, \quad F_X \in \mathbb{F}_X,$$

i.e., that only 25% of the countries have managed to reduce the wage gap between years. Using data from the "pay gap as difference in hourly wage rates" in different countries between the years of 2021 and 2020 [57], our objective is to deliberate what ε should be used in a follow-up study on the same matter.

We remove the countries with one or both entries missing, resulting in a sample of $n = 28$. Then, we use a Dirichlet process [42] with scaling parameter equal to 1 and centered on $N(0, 10)$ as the prior. Lastly, we apply Theorem 3 and choose ε as the largest value that would lead to rejecting H_0 when $\alpha = 0.05$ on `PROTEST`, leading to $\varepsilon \approx 0.0312$. Therefore, in a follow-up study, if such value of ε leads to rejection, we can safely conclude that H_0 has failed to reproduce.

4.2. Intuitions that delimit ε

When the intuitions provided by the researcher are not sufficient to provide a definitive value for ε, but can nevertheless be of use, some suggestions are:

Setting an upper bound through examples This case consists of listing the pairs of elements in the hypothesis space that the researcher assumes to be negligible from each other. Then, by obtaining the dissimilarities of those combinations and taking the largest of them, the result can be assigned as the value of ε. This represents a lower bound for the real ε of interest and, in case the test does not reject the hypothesis, provides the exact same conclusion as the "true" ε. This strategy is employed in section 6.

Using multiple candidates for ε We assume that, instead of dealing with a unique ε, there is a list or a range of values for ε which one must consider. This might happen when there are multiple professionals and each of them provides their own suggested ε, such as when there are more "liberal" or "conservative" choices available for it [50]. The idea is to simply perform `PROTEST` for each ε on the list (or to a grid based on the range of reasonable candidates) and take as final the decision that came out the most. Further still, we could weight each candidate based on some criteria (such as the importance of the professional or how much smaller a specific ε is when compared to the others) and apply the same idea. For example, since the classification dissimilarity (Definition 3) only takes values in $[0.5, 1]$, one could build a grid and use weights that decrease linearly to reach a decision, such as giving weight 1 to $\varepsilon = 0.5$ and 0 to $\varepsilon = 1$.

Direct graphical evaluation Lastly, we suggest the user to simply plot the conclusion as a function of ε and $\alpha \in [0, 1]$, and then use this graphical evaluation to decide if rejecting the hypothesis makes sense. This is the only suggestion that does not require setting neither ε nor α beforehand and should thus be used with caution. After all, this liberty could influence the analyst of the test towards making the conclusion they already agree with, biasing the results.

More than an actual suggestion for reaching conclusions, this plot acts as a tool for transparency and plurality. Since disagreements on the choice of α and ε are sure to be common, it neatly provides an indication of the decision one should take for their particular choice without the requirement of doing the whole analysis once again.

Example 3.1 *(Worldwide gender wage gap, continued).* Fig. 6 presents, for each combination of (ε, α), the decision suggested by `PROTEST`, with the red area implying rejection and the green area implying non-rejection. Based on the figure alone, we can safely conclude that researchers advocating for $\varepsilon \geq 0.1$ should not reject the hypothesis, since α would need to be set around 0.5 to lead to rejection.

5. Simulated studies

5.1. Regression on a binary response variable

This next setting uses data from a logistic regression to evaluate if the test can discriminate between link functions as the sample size grows. Let $\boldsymbol{X}$ be a 3 column matrix, with all values sampled from a $U(-3, 3)$, and Y be a binary variable such that

$$\mathbb{P}(Y_i = 1 | \boldsymbol{X}) = \frac{1}{1 + \exp(-0.5 + 1.5\boldsymbol{X}_{i,1} - 2\boldsymbol{X}_{i,2} + 0\boldsymbol{X}_{i,3})}, \quad i \in \{1, 2, \cdots, n\},$$

where n represents the sample size.

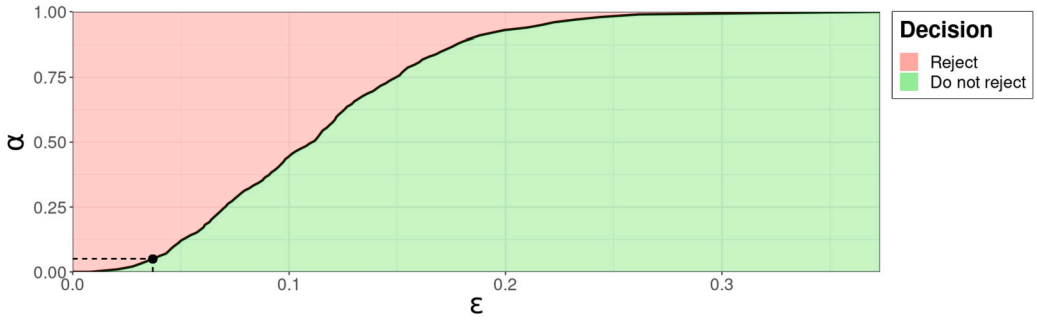


Fig. 6. Decision regions as a function of (ε, α) for the gender wage gap data (red for rejection, green for non-rejection). The black dot in the curve indicates the initial choice for ε based on $\alpha = 0.05$.

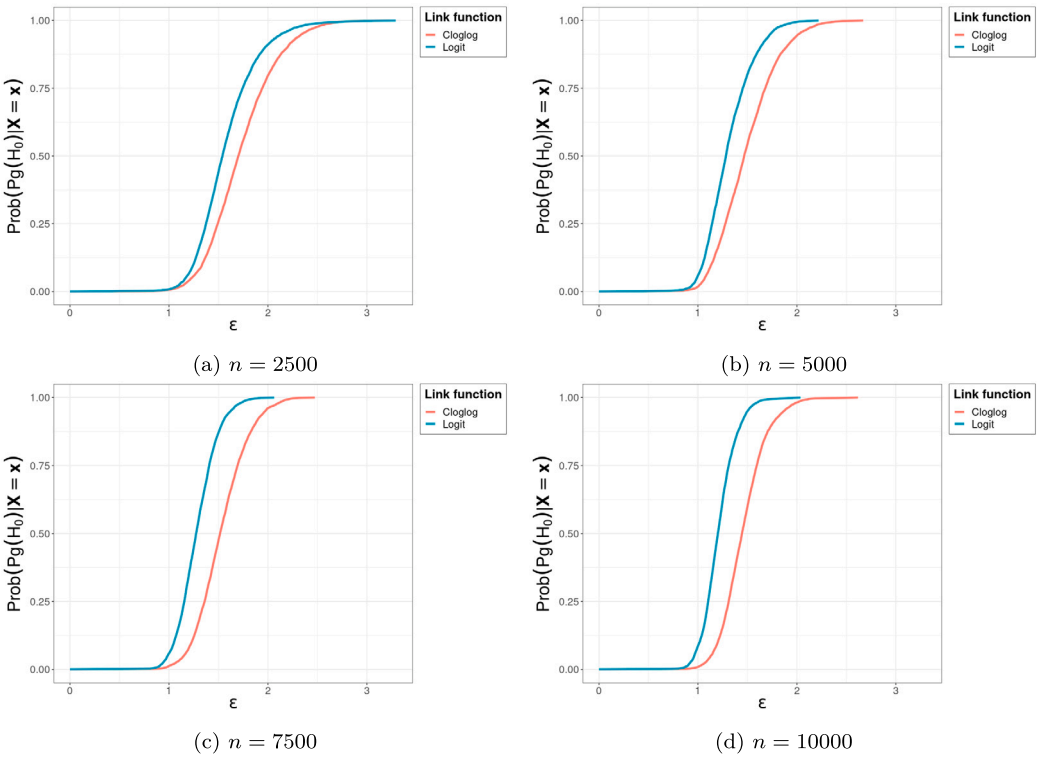


Fig. 7. Largest ε that entails rejection and the posterior probability of $Pg(H_0)$ for each link function and multiple sample sizes.

We use the nonparametric model proposed by [58] to draw estimates of $\mathbb{P}(Y = 1|\boldsymbol{X})$ and then apply the hypothesis from Equation (6) to check which link function (logit or cloglog) seems better suited for the data. The prior specification follows the second approach suggested by that paper and we set $Gamma(2, 2)$ as the prior distribution of the Dirichlet process' scaling parameter. We truncate the Dirichlet process so that it provides 30 mixture components.

Fig. 7 provides the test results for different sample sizes. We observe that, when increasing the sample size, the value of ε that would lead to rejection becomes consistently smaller for both link functions and the decision becomes less dependent on the choice of α. Still, the logit link presents a superior performance for all sample sizes, and the difference between curves becomes more apparent as well.

5.2. Similarity between normal and Student's t distributions

This simulation study compares the performance of `PROTEST` with that of other standard methods used in two-sample testing. It highlights the importance of defining a pragmatic hypothesis when information is available, as the other methods are bound to reject H_0 for large enough samples by picking up negligible but systematic differences in the data, while the results provided by `PROTEST` are robust. This same problem has been pointed out in other settings [59,60,20].

Table 2
Posterior probabilities based on $H_0 : F_X = F_Y$ from PROTEST ($\varepsilon \approx 0.0657$) and from the PTtest for different sample sizes and values of the hyperparameter c, as well as the p-value of the KStest. In all cases, one dataset was generated from a $N(0,1)$ and the other from a t_{30}.

Sample size	PROTEST				PTtest				KStest
	$c=1$	$c=4$	$c=7$	$c=10$	$c=1$	$c=4$	$c=7$	$c=10$	
10^2	0	0.4680	0.8421	0.8856	0.9509	0.8843	0.8478	0.8174	0.5806
10^3	0.0057	0.9998	1	1	1	1	1	1	0.6852
10^4	1	1	1	1	1	1	1	1	0.8232
10^5	1	1	1	1	1	1	1	1	0.1453
2×10^5	1	1	1	1	1	1	1	1	0.0215
3×10^5	1	1	1	1	1	0.9999	1	1	0.0004
4×10^5	1	1	1	1	1	0	0	1	0
5×10^5	1	1	1	1	0.0015	0	0	0	0
6×10^5	1	1	1	1	0.0978	0	0	0	0
7×10^5	1	1	1	1	0	0	0	0	0

Distribution tables have been widely present in statistical textbooks through time [61,62] and are still used nowadays for pedagogical purposes [63]. Particularly for the Student's t distribution table, a common feature is that the table becomes sparser after 30 degrees of freedom, implying that after 30 the deviations between the quantiles are deemed as negligible. Moreover, since the Student's t distribution converges to a standard Normal as its degrees of freedom tend towards infinity, some claim that using the Normal distribution as an approximation when the degrees of freedom are over 30 is good enough for most practical purposes [64]. We use this "consensus" as the basis for our simulation study, verifying how sensitive PROTEST can be to it.

Let $H_0 : F_X = F_Y$, where X represents data coming from the $N(0,1)$ and Y from the t_{30}. Table 2 presents a comparison between PROTEST, the PTtest and the Kolmogorov-Smirnov test (KStest) in such context for multiple sample sizes. In order to highlight the difference between the methods while keeping them as similar as possible, we draw from the posterior of a Pólya tree process (PT, [43,44]) for PROTEST as well. Using the same construction as [22] for the PT, we establish a centering distribution and a constant $c > 0$, which tells how much the process should rely on such distribution. Since [22, Section 5] recommends $c \in [1, 10]$, we set $c \in \{1, 4, 7, 10\}$ for the study. For both datasets, we apply a PT centered on $N(0,1)$.

Now, let us retrace all steps of the PROTEST procedure, but skipping the choice of α and step 4 altogether, since we are only interested in the posterior probabilities.

1. The null hypothesis is $H_0 : F_X = F_Y$. We use (14) as the dissimilarity function and follow the prior thresholding guideline presented in subsection 4.1 for establishing ε. For each $c \in \{1, 4, 7, 10\}$, we obtain ε such that $\mathbb{P}[Pg(H_0)] = 0.5$ and choose the most restrictive of them, which in this case resulted in $\varepsilon \approx 0.0657$.
2. Since the number of parameters of the PT is infinite, we draw from a partially specified PT [44] instead. Following [65], we set $\log_2 n \approx 20$ as the number of layers, n being the largest sample size of Table 2.
3. From Theorem 4, we conclude that, for any (P_X, P_Y) obtained from the data,

$$\inf_{(F_X,F_Y)\in H_0} d\left[(F_X,F_Y),(P_X,P_Y)\right] = d_C^*(P_X,P_Y).$$

Now, let Ω be the sample space of both datasets and $(\Omega_i)_{i\in\{1,\cdots,I\}}$ be the sets obtained from the partition of the last layer of the PT. Then, if F and G come from partially specified PTs centered on the same distribution function,

$$\begin{aligned}
\mathbb{P}\left(\frac{f(Z)}{g(Z)} > 1 \middle| Z\sim F\right) &= \sum_{i=1}^{I} \mathbb{P}\left(\frac{f(Z)}{g(Z)} > 1 \middle| \{Z\sim F\}\cap\{Z\in\Omega_i\}\right) F\left(Z\in\Omega_i\right) \\
&= \sum_{i=1}^{I} \mathbb{P}\left(\frac{F\left(Z\in\Omega_i\right)}{G\left(Z\in\Omega_i\right)} > 1 \middle| \{Z\sim F\}\cap\{Z\in\Omega_i\}\right) F\left(Z\in\Omega_i\right) \\
&= \sum_{i=1}^{I} \mathbb{I}\left(\frac{F\left(Z\in\Omega_i\right)}{G\left(Z\in\Omega_i\right)} > 1\right) F\left(Z\in\Omega_i\right),
\end{aligned}$$

and thus (14) can be obtained analytically, easing the calculation of (3).

From Table 2, we see that the PTtest and the KStest provide the desired outcome for smaller samples, but reject when the sample size is large enough. Of course, rejecting the hypothesis is no fault of such tests since H_0 is false, but it is an indication that the test may be too rigorous on negligible differences that are perfectly compatible with real-world data when the sample size is large.

Unlike the other tests, PROTEST remains consistent for all cases as the sample size grows, and this is not to be confused with the method being permissive. Compared to the PTtest, its probability was generally lower for small sample sizes, but this is largely a consequence of choosing the more conservative ε. Moreover, the true dissimilarity between $N(0,1)$ and t_{30} is around 0.005 and, when using this value for ε instead, for no sample size did PROTEST reach a probability other than 0.

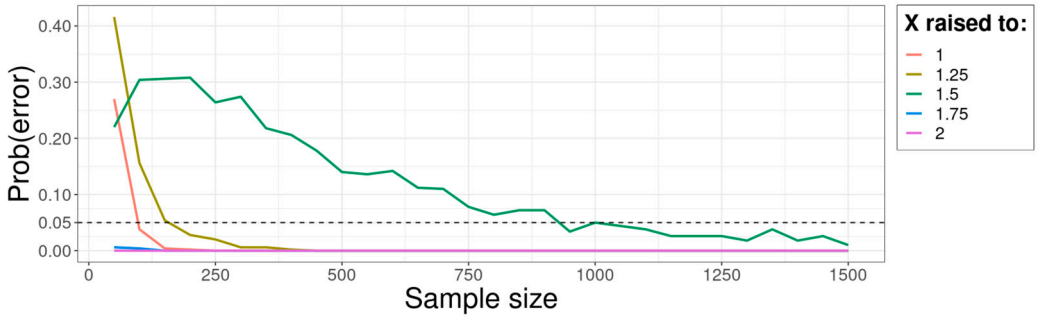


Fig. 8. Probability of PROTEST to commit an error as a function of the sample size and of the value that X was raised to. In this setting, $\alpha = 0.05$ and $\varepsilon = d(H_0, x^{1.4}) \approx 0.1639$.

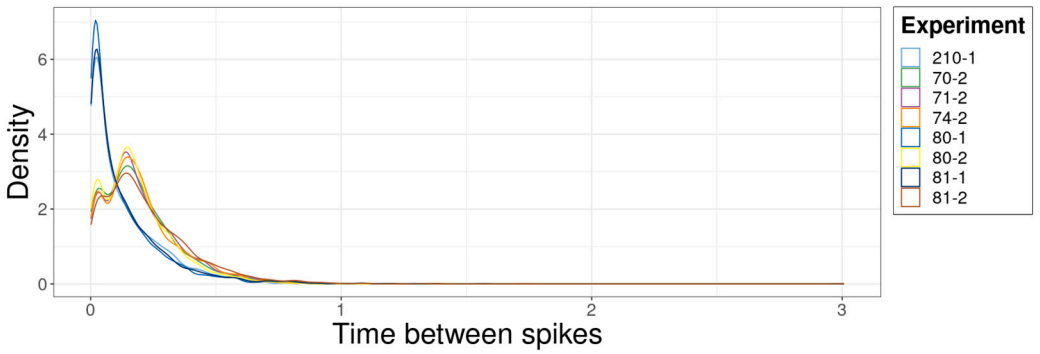


Fig. 9. Smoothed sample density of the time between spikes of neuron "2494" for each experiment.

5.3. PROTEST decision error

In this study, we evaluate the effect of the sample size in the probability of PROTEST to commit a decision error. More specifically, we apply the linearity test of Theorem 2 to data $Y = X^k + \epsilon$, where $X \sim U(0,3)$, $\epsilon \sim N(0,2)$ and $k \in \{1, 1.25, 1.5, 1.75, 2\}$. We set $\alpha = 0.05$ and $\varepsilon = d(H_0, x^{1.4}) \approx 0.1639$. To model the regression function, we set $R(.) \sim GP(0, K(\cdot,\cdot))$, where $K(\boldsymbol{x}, \boldsymbol{x'}) = \exp(-2||\boldsymbol{x} - \boldsymbol{x'}||_2)$.

Fig. 8 shows the estimated posterior error probability as a function of the sample size for each k. For a given sample size and a specific k, these error estimates are the proportion of the 500 samples drawn where PROTEST leads to the wrong decision. For almost all cases, the probability of PROTEST to be wrong quickly goes to zero. As for when $k = 1.5$, the data generating process provides dissimilarities closer to the borders set by ε, and thus PROTEST requires more data to provide assertive decisions.

6. Application: neuron spike analysis

In this section, we apply PROTEST to data on the time between neuron spikes (in microseconds) of an epilepsy patient exposed to visual stimuli (pictures in varied contexts, each context represents an experiment). The first test evaluates if a Poisson process [66] can describe the data, while the second uses the median to verify if the neuron activity is similar across experiments. In both tests, we use a Dirichlet process (DP, [42]) with a centering distribution gamma and scaling parameter of 1. To stipulate the hyperparameters of the gamma distribution, we remove one of the experiments from the data and use its maximum likelihood estimates (MLE).

The original dataset [67] is composed of 42 patients and the brain activity of their amygdala and hippocampus as they were subjected to the stimuli. The authors identified clusters of activity, which were assumed to represent individual neurons, and registered a total of 1576 individual neurons. We restricted the analysis to the neuron "2494" due to it having a high number of experiments applied (8 in total) and a reasonably high sample size in each experiment (minimum of 693, maximum of 2691). As for the experiments, we use the notation "a-b" to represent session b of experiment a, since the same type of visual stimuli might be presented at different times.

Fig. 9 presents the smoothed sample densities for each experiment of neuron "2494". This plot alone already puts the assumption of a Poisson process into question, since some cases exhibit a bimodal behavior with peaks not that close to 0. As for the median, the densities of the experiments seem to be roughly divided in two groups, so the intragroup median might be similar enough.

For both tests, we use available information on how neurons work to set an upper bound for ε through examples, a procedure described in subsection 4.2. Since a neuron spike typically lasts for 1 millisecond [68, Section 1.1.1], it would be physically impossible for another spike to be observed in such interval. This is also corroborated by the fact that the smallest time observed between spikes is 0.0016 second, i.e., 1.6 milliseconds. Therefore, if the difference between two distribution functions could be attributed to the 1 millisecond threshold, they should be deemed as practically equivalent.

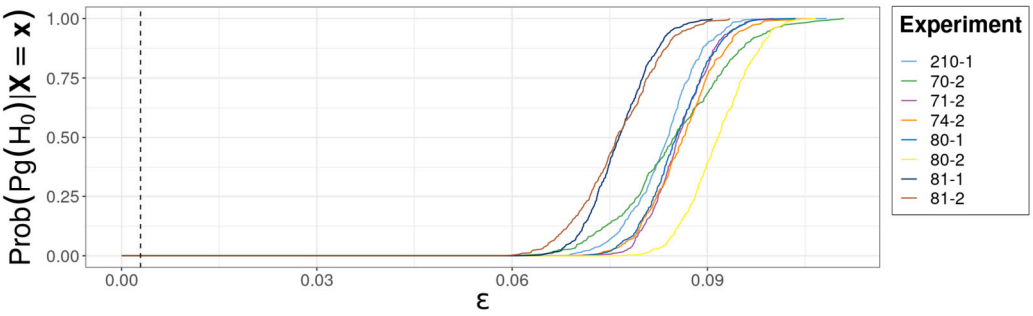


Fig. 10. Largest ε that entails rejection and the posterior probability of $Pg(H_0)$ for each experiment. The pragmatic hypothesis is expanded from $H_0 : T_i|\lambda \overset{ind.}{\sim} Exp(1/\lambda), \lambda \in \mathbb{R}_{\geq 0}, i \in \{1, \cdots, n\}$. The dashed line is the value of ε.

We turn once again to the experiment excluded from the analysis to derive a distribution function of reference and to establish ε from it. First, we model the excluded experiment through a gamma distribution due to its flexibility. Then, taking $tol = \pm 0.001$ as the criteria for practical equivalence, we use it to establish equivalent distribution functions and then obtain their dissimilarity.

Let $A \sim Gamma(\hat{\alpha}, \hat{\beta})$, where $\hat{\alpha} \approx 1.2172$ and $\hat{\beta} \approx 5.0581$ are the MLE based on the removed experiment. If $B \sim Gamma(\tilde{\alpha}, \tilde{\beta})$ is practically equivalent to A, then $\mathbb{E}[B] = \mathbb{E}[A] + tol$, i.e., the means differ by at most 1 millisecond. Furthermore, if $\mathbb{V}[B] = \mathbb{V}[A]$ (the variance is the same for both variables), then

$$\mathbb{V}[B] = \frac{\tilde{\alpha}}{\tilde{\beta}^2} = \frac{\mathbb{E}[B]}{\tilde{\beta}} = \frac{\frac{\hat{\alpha}}{\hat{\beta}} + tol}{\tilde{\beta}} = \mathbb{V}[A] \Longrightarrow \tilde{\beta} = \hat{\beta} + \frac{tol}{\mathbb{V}[A]} \approx \{5.0372, 5.0792\};$$

$$\mathbb{E}[B] = \frac{\tilde{\alpha}}{\tilde{\beta}} = \frac{\hat{\alpha}}{\hat{\beta}} + tol \Longrightarrow \tilde{\alpha} = \tilde{\beta}\left(\frac{\hat{\alpha}}{\hat{\beta}} + tol\right) \approx \{1.2071, 1.2273\}.$$

Lastly, if F_A, F_{B^-} and F_{B^+} are the respective distribution functions of A and B with the negative/positive tolerance, we set $\varepsilon = \max\{d(F_A, F_{B^-}), d(F_A, F_{B^+})\}$.

6.1. First test: Poisson process

The first test verifies if the spiking behavior observed in the data can be adequately described as a Poisson process [66] $N(t)$, a counting process such that $N(t) \sim Poisson(\lambda t), \forall t \in \mathbb{R}_{\geq 0}$. If $(X_1, \cdots, X_n)$ is the moment in time where each spike has occurred, $T_1 := X_1$ and $T_i := X_i - X_{i-1}$, $i \in \{2, \cdots, n\}$, then

$$H_0 : N(t) \text{ is a Poisson process} \Longleftrightarrow H_0 : T_i|\lambda \overset{ind.}{\sim} Exp(1/\lambda), \lambda \in \mathbb{R}^+, i \in \{1, \cdots, n\},$$

therefore the problem can be reframed as the goodness-of-fit test introduced in subsection 3.2, particularly that of Example 2.

By using the L_∞ distance from Equation (10) as the dissimilarity function, we conclude that $\varepsilon \approx 0.0029$. Hence, we should expect a difference of at most 0.0029 between a distribution function drawn from the DP and the exponential distribution that is closest to it for any $x \in (0, \infty)$.

Fig. 10 provides the largest ε that leads to rejecting the hypothesis for each value of α in each experiment. From it, it is clear that taking $\varepsilon \approx 0.0029$ leads to rejection for all experiments, since $\mathbb{P}[\mathcal{P} \in Pg(H_0)|\boldsymbol{T}]$ becomes greater than 0 only when $\varepsilon \geq 0.06$. This result means that either the hypothesis should be rejected or that the choice of ε was too strict. Considering that the values of ε that would lead to non-rejection are considerably far from the initial estimate, we reject the hypothesis for all experiments.

6.2. Second test: median of time between spikes

This second test is a particular case of the quantile test (subsection 3.3, $p_0 = 0.5$), an instance of PROTEST that has already been demonstrated in Example 3. In this case, the remaining steps required for performing PROTEST are to assume a value for x_0 and to derive ε for this case. For the former, we use the experiment that was removed from the original data, which provides a sample median of around 0.1787 second between spikes, implying that the null hypothesis can be expressed as

$$H_0 : F(0.1787) = 0.5, \quad F \in \mathbb{F}.$$

As for the latter, we once again turn to the scheme based on the 1 millisecond threshold, which when applied for the distance in Equation (11) results in $\varepsilon \approx 0.001$.

Table 3 presents the results of PROTEST for this case, as well as information on the sample size and the sample median of each experiment. We observe that the test provides assertive decisions in all cases, requiring either a considerably high significance level α to reject or not requiring it at all. Following our intuition, the experiments whose sample medians are closer to 0.1787 are the ones that lead to non-rejection.

Table 3
Comparison between experiments based on the sample median and the smallest value of α that would lead to the rejection of H_0 ($\varepsilon \approx 0.001$).

Experiment	Sample size	Sample median	α for rejecting H_0
70-2	693	0.1651	0.970
71-2	2388	0.1668	1
74-2	1834	0.1718	1
80-1	2487	0.0693	0
80-2	1919	0.1601	0.975
81-1	2691	0.0785	0
81-2	1547	0.1793	1
210-1	2279	0.0795	0

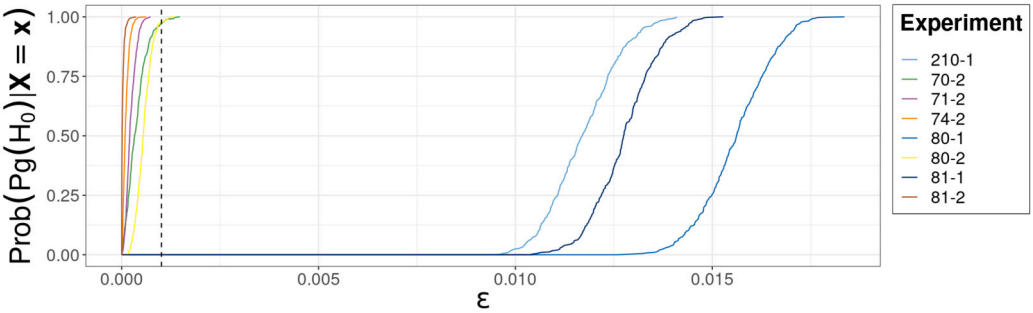


Fig. 11. Largest ε that entails rejection and the posterior probability of $Pg(H_0)$ for each experiment. The pragmatic hypothesis is expanded from $H_0 : F(0.1787) = 0.5$, $F \in \mathbb{F}$. The dashed line is the value of ε.

Fig. 11 provides more nuanced results, clearly contrasting between the experiments that were rejected and the ones that were not. While the conclusion of not rejecting the hypothesis for experiments whose curves reach their peak early is hardly contestable, the rejection for the other cases will depend on how strict is the choice of ε. Still, the clear divide between the curves is more evidence of the robustness of our decision.

We end this subsection noting that there are alternative formulations for this test that rely less heavily on the excluded experiment. For instance, if F_i and F_j are the distribution of the i-th and j-th experiment,

$$H_0 : F_i^{-1}(0.5) = F_j^{-1}(0.5) = q, \quad q \in \mathbb{R}^+, \tag{19}$$

is also a hypothesis that evaluates the similarity of the median for different experiments. While this setting is not contemplated by any of the pragmatic hypotheses explored in this work, it is a generalization of the two-sample test (subsection 3.4). Therefore, the monotonicity property (Corollary 1) guarantees that, if the pragmatic two-sample hypothesis is not rejected, then the pragmatic version of Equation (19) is not rejected as well.

7. Discussion

`PROTEST` offers a new paradigm for hypothesis testing, one that is theoretically sound, easy to apply and highly adaptable to practical settings. Moreover, although the pragmatic versions covered here represent enhancements over nonparametric hypotheses routinely evaluated, there are still many other hypotheses left to be expanded, such as the one at the end of subsection 6.2. Cases that deal with multivariate or high-dimensional settings are probably the ones where greater attention should be directed, given how common these models have become. We have developed an `R` package that implements `PROTEST` and that reproduces some of the analyses in this paper. Its development version can be found in rflassance/protest.

The `PROTEST` procedure can be extended to the context of three-way testing—which can accept, reject or remain undecided towards a hypothesis—linking it more closely to the work of [5]. This can be done by switching `PROTEST` for its three-way version, which also retains the monotonicity property.

Definition 4 *(Three-way `PROTEST`).* Let $Pg(H_0, d, \varepsilon)$ be the pragmatic hypothesis and $\mathcal{P}$ be a random object over $\mathbb{H}$. The three-way `PROTEST` is such that, for $0 \leq \alpha_1 \leq \alpha_2 \leq 1$,

- If $\mathbb{P}\left(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}\right) \leq \alpha_1$, reject the hypothesis;
- If $\alpha_1 < \mathbb{P}\left(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x}\right) < \alpha_2$, remain undecided;
- Otherwise, accept the hypothesis.

The exact decision rules in Definition 4 have been previously introduced in the context of rough sets [26, subsection 2.3] and of decision theory with an imprecise loss function [69, subsection 3.3], although neither make use of pragmatic hypotheses. Through the logic of the latter, when there is uncertainty on the constant c of the 0-1-c loss such that $c \in \left[\frac{\alpha_2}{1-\alpha_2}, \frac{\alpha_1}{1-\alpha_1}\right]$, then the three-way PROTEST corresponds to the Bayes decision.

Other three-way testing procedures are the GFBST [70] and coherent agnostic tests in general [71]. We note that all of these procedures heavily rely on pragmatic hypotheses, so the contributions in section 3 can be of use even if one uses a procedure other than PROTEST. Further still, if PROTEST is adapted to evaluate a credibility region instead of $\mathbb{P}(\mathcal{P} \in Pg(H_0)|\boldsymbol{X} = \boldsymbol{x})$, it will acquire new properties that make it a fully coherent procedure in the sense presented by [72, Definition 2.6].

Abbreviations

DP	Dirichlet Process
GFBST	Generalized Full Bayesian Significance Test
GLM	Generalized Linear Model
GP	Gaussian Process
KStest	Kolmogorov-Smirnov test
MCMC	Markov Chain Monte Carlo
PROTEST	Pragmatic Region Oriented Test
PT	Pólya Tree Process
PTtest	Pólya Tree test

CRediT authorship contribution statement

Rodrigo F.L. Lassance: Writing – original draft, Validation, Software, Methodology, Formal analysis, Conceptualization. **Rafael Izbicki:** Writing – review & editing, Conceptualization. **Rafael B. Stern:** Writing – review & editing, Supervision, Conceptualization.

Funding sources

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001; Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) [grant numbers 2013/07699-0, 2019/11321-9, 2023/07068-1]; and Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) [grant numbers 309607/2020-5, 422705/2021-7, 305065/2023-8].

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Dani Gamerman, Julio M. Stern, Luben M. C. Cabezas and Luis G. Esteves for the fruitful conversations and suggestions regarding PROTEST.

Appendix A. Prior processes used in this work

A.1. Dirichlet process

The Dirichlet process [DP; 42] is a prior process applicable to a variety of different settings, thus justifying its popularity in the literature. In this section, we restrict our attention to cases that use it to model distribution functions directly, which require a centering distribution P_C and a scaling parameter α.

Definition 5. Let Ω be the sample space. A distribution function $\mathcal{P}$ follows a Dirichlet process $DP(P_C, \alpha)$ if, for every finite partition $(B_i)_{1 \leq i \leq d}$, $d \in \mathbb{N}$, of Ω,

$$\big(\mathcal{P}(B_1), \mathcal{P}(B_2), \cdots, \mathcal{P}(B_k)\big) \sim Dirichlet\big[\alpha\big(P_C(B_1), P_C(B_2), \cdots, P_C(B_k)\big)\big]. \tag{A.1}$$

As a particular case of the Equation (A.1), one can take the partition $B_1 = (-\infty, x]$ and $B_2 = (x, \infty)$, which leads to

$$\big(\mathcal{P}((-\infty, x]), \mathcal{P}((x, \infty))\big) \sim Dirichlet\big[\alpha P_C((-\infty, x]), \alpha P_C((x, \infty))\big].$$

Since, by definition, $\mathcal{P}(X \leq x) = \mathcal{P}((-\infty, x]) = 1 - \mathcal{P}((x, \infty))$, we can restrict the analysis to $\mathcal{P}(X \leq x)$, leading to the conclusion that, for all $x \in \Omega$,

$$\mathcal{P}(X \leq x) \sim Beta(\alpha P_C(X \leq x), \alpha(1 - P_C(X \leq x)));$$

$$\mathbb{E}[\mathcal{P}(X \leq x)] = \frac{\alpha P_C(X \leq x)}{\alpha P_C(X \leq x) + \alpha - \alpha P_C(X \leq x)} = P_C(X \leq x);$$

$$\mathbb{V}[\mathcal{P}(X \leq x)] = \frac{\alpha^2[P_C(X \leq x)(1 - P_C(X \leq x))]}{\alpha^2(\alpha + 1)} = \frac{P_C(X \leq x)(1 - P_C(X \leq x))}{1 + \alpha}.$$

Therefore, P_C is the expectation of the DP and α indicates how much the process is concentrated around P_C.

One of the advantages of the DP is that its posterior is conjugate [42]. For a sample $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$, let $\delta_{x_i} := \mathbb{I}(x_i \in A)$, with $i \in \{1, 2, \cdots, n\}$ and $A \subseteq \Omega$. Then,

$$\mathcal{P}|\boldsymbol{X} = \boldsymbol{x} \sim DP\left(\frac{\alpha P_C + \sum_{i=1}^n \delta_{x_i}}{\alpha + n}, \alpha + n\right).$$

This implies that the expectation of the posterior DP is a mixture model, with probability $\frac{\alpha}{\alpha+n}$ of drawing the next observation from P_C and $\frac{n}{\alpha+n}$ of drawing from the discrete distribution $\frac{1}{n}\sum_{i=1}^n \delta_{x_i}$.

Lastly, we use a strategy proposed by [73] for approximately sampling from the DP with an arbitrary degree of precision. Let $(Y_i)_{i\in\mathbb{N}}$ and $(\theta_i)_{i\in\mathbb{N}}$ be independent random sequences such that, $\forall i \in \mathbb{N}$, $Y_i \sim P_C$ and $\theta_i \sim \beta(1, \alpha)$. Then

$$p_i = \theta_i \prod_{j<i}(1 - \theta_j), \quad \mathcal{P} = \sum_{i=1}^{\infty} p_i \delta_{Y_i} \Longrightarrow \mathcal{P} \sim DP(P_C, \alpha). \tag{A.2}$$

From this construction, we observe that p_i becomes increasingly closer to 0, therefore the latter terms of the sample $(Y_i)_{i\in\mathbb{N}}$ are increasingly less probable to be drawn, thus truncating the process at a specific $n \in \mathbb{N}$ can provide a distribution function that is as close to the true distribution as desired. Moreover, Equation (A.2) shows that, with probability 1, the DP provides a discrete distribution.

A.2. Gaussian process

Assuming a regression setting, let $Y = R(\boldsymbol{x}) + \epsilon$ represent the conditional distribution $Y|R(\boldsymbol{x})$, where $\epsilon \sim N(0, \sigma^2)$ and $R(\cdot)$ is a random function. We say that $R(\cdot)$ is a Gaussian process (GP) if all of its finite-dimensional representations behave according to a multivariate normal distribution, that is,

$$R(\boldsymbol{X}) \sim N(m(\boldsymbol{X}), K(\boldsymbol{X}, \boldsymbol{X})), \quad \forall \boldsymbol{X} \subset \mathcal{X}, \tag{A.3}$$

where $\mathcal{X}$ is the covariate space, $m(\cdot)$ is a mean function and $K(\cdot, \cdot)$ is the covariate function. From (A.3) and the conjugacy property of the normal distribution, we conclude that, for $\boldsymbol{X}^* \subset \mathcal{X}$,

$$R(\boldsymbol{X}^*)|\boldsymbol{y}, \boldsymbol{X}, \sigma^2 \sim N(\mu(\boldsymbol{X}^*), \Sigma(\boldsymbol{X}^*, \boldsymbol{X}^*)),$$

$$\mu(\boldsymbol{X}^*) := m(\boldsymbol{X}^*) + K(\boldsymbol{X}, \boldsymbol{X}^*)(K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2\mathbb{I})^{-1}(\boldsymbol{y} - m(\boldsymbol{X})),$$

$$\Sigma(\boldsymbol{X}^*, \boldsymbol{X}^*) := K(\boldsymbol{X}^*, \boldsymbol{X}^*) - K(\boldsymbol{X}, \boldsymbol{X}^*)(K(\boldsymbol{X}, \boldsymbol{X}) + \sigma^2\mathbb{I})^{-1}K(\boldsymbol{X}^*, \boldsymbol{X}).$$

The covariate function $K(\cdot, \cdot)$ is built from a kernel function. Some common choices for the kernel function are:

- **Gaussian:** $K_G(\boldsymbol{x}, \boldsymbol{x}') = \exp(-s||\boldsymbol{x} - \boldsymbol{x}'||_2)$, $s > 0$;
- **Matérn:** $K_M(\boldsymbol{x}, \boldsymbol{x}') = t\left(1 + s||\boldsymbol{x} - \boldsymbol{x}'||_2 + \frac{(s||\boldsymbol{x}-\boldsymbol{x}'||_2)^2}{3}\right) K_G(\boldsymbol{x}, \boldsymbol{x}')$, $s > 0$, $t > 0$.

A.3. Pólya tree process

To define the Pólya Tree process (PT), we first introduce some notation. Let Π be a collection of separable binary trees of partitions of Ω, the sample space. At the m-th layer of the partition, the collection of sets $\{B_i^{(m)}; i = 0, \cdots, 2^m - 1\}$ is such that

$$\bigcup_{i=0}^{2^m-1} B_i^{(m)} = \Omega; \quad B_i^{(m)} \cap B_j^{(m)} = \emptyset, \forall i \neq j; \quad B_i^{(m)} = B_{2i}^{(m+1)} \cup B_{2i+1}^{(m+1)}.$$

It is possible to drop the superscript of $B_i^{(m)}$ by representing the subsets of Ω in base 2. For instance, B_0 represents the first set at the first layer, B_{11} the fourth and last set of the second layer and B_{010} the third set of the third layer. Let $B_\emptyset = \Omega$ and take

$$E = \{0, 1\}, \quad E^m = \overbrace{E \times \cdots \times E}^{m \text{ times}} = \{0, 1\}^m, \quad E^0 = \emptyset, \quad E^* = \bigcup_{m=0}^{\infty} E^m.$$

Definition 6. [43]. Let $\Theta = \{\theta_\epsilon : \epsilon \in E^*\}$ be the probabilities of a variable to belong to each partition and $\mathcal{A} = \{\boldsymbol{\alpha}_\epsilon : \epsilon \in E^*\}$ be the parameters related to such probabilities. Set $\boldsymbol{\epsilon}_m = (\epsilon_1, \cdots, \epsilon_m)$ as the variable's trajectory up to the m-th layer, where, for each layer $i \in \{1, 2, \cdots, m\}$, $\epsilon_i = 0$ if the variable is on the left set and $\epsilon_i = 1$ if it is on the right one. $\mathcal{P}$ follows a Pólya Tree process $PT(\Pi, \mathcal{A})$ if the following conditions are satisfied:

1. All random variables in Θ are mutually independent;
2. For every $m \in \{1, 2, \cdots\}$ and every $\boldsymbol{\epsilon}_m \in E^m$, $\theta_{\boldsymbol{\epsilon}_m} \sim Beta(\alpha_{(\boldsymbol{\epsilon}_m,0)}, \alpha_{(\boldsymbol{\epsilon}_m,1)})$;
3. For every $m \in \{1, 2, \cdots\}$ and every $\boldsymbol{\epsilon}_m \in E^m$,

$$P(B_{\boldsymbol{\epsilon}_m}|\Theta) = \prod_{i=1}^{m} (\theta_{\boldsymbol{\epsilon}_{i-1}})^{\epsilon_i} (1 - \theta_{\boldsymbol{\epsilon}_{i-1}})^{1-\epsilon_i}.$$

From Definition 6, we note that the posterior is conjugate. For $\boldsymbol{x} = (x_1, \cdots, x_n)$, let $n_{\boldsymbol{\epsilon}_m} = \sum_{j=1}^{n} \mathbb{I}(x_j \in B_{\boldsymbol{\epsilon}_m})$, where $\mathbb{I}(.)$ is the indicator function. Then,

$$\mathcal{P}|\boldsymbol{X} = \boldsymbol{x} \sim PT(\Pi, \mathcal{A}^*); \quad \mathcal{A}^* = \{\alpha^*_\epsilon : \epsilon \in E^*\}; \quad \alpha^*_{\boldsymbol{\epsilon}_m} = \alpha_{\boldsymbol{\epsilon}_m} + n_{\boldsymbol{\epsilon}_m}.$$

This implies that, if any $\alpha_{\boldsymbol{\epsilon}_m}$ is updated by an observation, then at the next layer it suffices to check if $\alpha_{(\boldsymbol{\epsilon}_m,0)}$ or $\alpha_{(\boldsymbol{\epsilon}_m,1)}$ should be updated as well.

Another feature of the PT is that it can be applied to discrete, continuous and singular continuous random variables [74]. For $c > 0$, the following choice for the parameters on $\mathcal{A}$ guarantees that, with probability one, the distribution function is of the same type as that of the random variable.

$$\alpha_{(\boldsymbol{\epsilon}_m,0)} = \alpha_{(\boldsymbol{\epsilon}_m,1)} = \begin{cases} \frac{c}{2^m}, & \text{if } X \text{ is discrete;} \\ c, & \text{if } X \text{ is continuous singular;} \\ c \times m^2, & \text{if } X \text{ is absolutely continuous.} \end{cases}, \quad \forall m \in \mathbb{N}.$$

The PT can be built such that its expectation is equal to a centering distribution P_C of choice. To do that, we first look at the conditional distribution $B_{\boldsymbol{\epsilon}_m}|B_{\boldsymbol{\epsilon}_{m-1}}$. For a fixed Θ, $\mathcal{P}$ is known and the conditional distribution is given by

$$\mathcal{P}(B_{\boldsymbol{\epsilon}_m}|B_{\boldsymbol{\epsilon}_{m-1}}, \Theta) = (\theta_{\boldsymbol{\epsilon}_{m-1}})^{\epsilon_m} (1 - \theta_{\boldsymbol{\epsilon}_{m-1}})^{1-\epsilon_m}. \tag{A.4}$$

Thus, the expectation of the process is

$$\begin{aligned} \mathbb{E}_{\mathcal{P}}[\mathcal{P}(B_{\boldsymbol{\epsilon}_m}|B_{\boldsymbol{\epsilon}_{m-1}})] &= \mathbb{E}_\Theta\{\mathbb{E}_{\mathcal{P}}[\mathcal{P}(B_{\boldsymbol{\epsilon}_m}|B_{\boldsymbol{\epsilon}_{m-1}}, \Theta)]\} = \mathbb{E}_\Theta\left[(\theta_{\boldsymbol{\epsilon}_{m-1}})^{\epsilon_m}(1 - \theta_{\boldsymbol{\epsilon}_{m-1}})^{1-\epsilon_m}\right] \\ &= \left(\frac{\alpha_{(\boldsymbol{\epsilon}_{m-1},0)}}{\alpha_{(\boldsymbol{\epsilon}_{m-1},0)} + \alpha_{(\boldsymbol{\epsilon}_{m-1},1)}}\right)^{\epsilon_m} \left(\frac{\alpha_{(\boldsymbol{\epsilon}_{m-1},1)}}{\alpha_{(\boldsymbol{\epsilon}_{m-1},0)} + \alpha_{(\boldsymbol{\epsilon}_{m-1},1)}}\right)^{1-\epsilon_m} = \frac{1}{2}, \end{aligned} \tag{A.5}$$

since $\alpha_{(\boldsymbol{\epsilon}_m,0)} = \alpha_{(\boldsymbol{\epsilon}_m,1)}$. From (A.5), one can set the centering distribution P_C through its quantile function $P_C^{-1}(.)$ and choose the partitions such that $P_C(B_0) = P_C(B_1) = \frac{1}{2}$ and, $\forall m \in \mathbb{N}$, $P_C(B_{(\boldsymbol{\epsilon}_m,0)}|B_{\boldsymbol{\epsilon}_m}) = P_C(B_{(\boldsymbol{\epsilon}_m,1)}|B_{\boldsymbol{\epsilon}_m}) = \frac{1}{2}$, ensuring that $\mathbb{E}_{\mathcal{P}}(\mathcal{P}) = P_C$.

Since the Pólya tree process possesses an infinite number of parameters, it is impossible to update them all in practice. To avoid this problem, [44] introduces a Pólya tree that is specified only up to a layer M, a partially specified PT. If, after the layer M, one tries to obtain a probability, it can be done by setting $\mathcal{P}(B_{\boldsymbol{\epsilon}_{M+1}}|B_{\boldsymbol{\epsilon}_M}) = P_C(B_{\boldsymbol{\epsilon}_{M+1}}|B_{\boldsymbol{\epsilon}_M})$. In this case, $\mathbb{E}_{\mathcal{P}}(\mathcal{P}) = P_C$ remains valid.

Appendix B. Proofs and technical findings

Proposition 2 *(Monotonicity property of the three-way* `PROTEST`*). Let* $H_0^1, H_0^2 \subset \mathbb{H}$ *be such that* $Pg(H_0^1, d, \varepsilon_1) \supseteq Pg(H_0^2, d, \varepsilon_2)$ *and* $0 \leq \alpha_1 \leq \alpha_2 \leq 1$. *Then, the three-way* `PROTEST` *(Definition 4) has the monotonicity property, i.e.,*

- *If the test rejects* $Pg(H_0^1)$*, then it rejects* $Pg(H_0^2)$ *as well;*
- *If the test remains undecided on* $Pg(H_0^1)$*, it does not accept* $Pg(H_0^2)$.

Proof. Let $\mathcal{P}$ be a random object on $\mathbb{H}$. Since $Pg(H_0^1) \supseteq Pg(H_0^2)$,

$$\mathbb{P}(\mathcal{P} \in Pg(H_0^1)) \geq \mathbb{P}(\mathcal{P} \in Pg(H_0^2)). \tag{B.1}$$

If $Pg(H_0^1)$ is rejected,

$$\mathbb{P}(\mathcal{P} \in Pg(H_0^1)) \leq \alpha_1 \overset{(B.1)}{\Longrightarrow} \mathbb{P}(\mathcal{P} \in Pg(H_0^2)) \leq \alpha_1.$$

If $Pg(H_0^1)$ remains undecided,

$$\alpha_1 < \mathbb{P}(\mathcal{P} \in Pg(H_0^1)) < \alpha_2 \overset{(B.1)}{\Longrightarrow} \mathbb{P}(\mathcal{P} \in Pg(H_0^2)) < \alpha_2. \quad \square$$

Proof of Corollary 1. The result follows by taking $\alpha_1 = \alpha_2$ in Proposition 2. $\square$

Proof of Proposition 1. For a fixed $\boldsymbol{x}$, the value of $\phi(\boldsymbol{x})$ is known. Therefore,

$$\mathbb{P}(\mathcal{P} \in Pg(H_0), \phi(\boldsymbol{x}) = 1 | \boldsymbol{X} = \boldsymbol{x}) = \begin{cases} \mathbb{P}(\mathcal{P} \in Pg(H_0) | \boldsymbol{X} = \boldsymbol{x}), & \text{if } \phi(\boldsymbol{x}) = 1; \\ 0, & \text{otherwise.} \end{cases} \tag{B.2}$$

By definition, $\mathbb{P}(\mathcal{P} \in Pg(H_0) | \boldsymbol{X} = \boldsymbol{x}) \leq \alpha$ if $\phi(\boldsymbol{x}) = 1$, therefore (B.2) is less than or equal to α as well. Assuming that $\{\boldsymbol{X} = \boldsymbol{x}\}$ and $\{\phi(\boldsymbol{x}) = 1\}$ are not contradictory statements,

$$\begin{aligned} e(\boldsymbol{x}) &= \mathbb{P}(\mathcal{P} \in Pg(H_0) | \boldsymbol{X} = \boldsymbol{x}, \phi(\boldsymbol{x}) = 1) = \frac{\mathbb{P}(\mathcal{P} \in Pg(H_0), \phi(\boldsymbol{x}) = 1 | \boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(\phi(\boldsymbol{x}) = 1 | \boldsymbol{X} = \boldsymbol{x})} \\ &= \mathbb{P}(\mathcal{P} \in Pg(H_0), \phi(\boldsymbol{x}) = 1 | \boldsymbol{X} = \boldsymbol{x}) \leq \alpha, \end{aligned}$$

and hence $\mathbb{E}[e(\boldsymbol{X}) | \phi(\boldsymbol{X}) = 1] \leq \alpha$. $\square$

Lemma 1. *If d satisfies Assumption 1, then*

a) For every $P_1 \in Pg(P_0, \varepsilon_2)$, there exists ε^* such that $Pg(P_1, \varepsilon^*) \subseteq Pg(P_0, \varepsilon_2)$.
b) For every P_1 such that $d(P_0, P_1) > \varepsilon_1$, there exists ε^* such that $Pg(P_1, \varepsilon^*) \subseteq (Pg(P_0, \varepsilon_1))^c$.

Proof. a) Let $P_1 \in Pg(P_0, \varepsilon_2)$. By Definition 1, there exists $0 < \varepsilon_1 < \varepsilon_2$ such that $d(P_0, P_1) = \varepsilon_1$. Hence, it follows from Assumption 1 that, for every P_2 such that $d(P_1, P_2) < \delta(\varepsilon_1, \varepsilon_2)$, one obtains $d(P_0, P_2) < \varepsilon$. It follows from Definition 1 that, by taking $\varepsilon^* = \delta(\varepsilon_1, \varepsilon_2)$, one obtains $Pg(P_1, \varepsilon^*) \subseteq Pg(P_0, \varepsilon_2)$.

b) Let $\varepsilon_2 = d(P_0, P_1)$, $\varepsilon^* = \delta(\varepsilon_1, \varepsilon_2)$, and take an arbitrary $P_2 \in Pg(P_1, \varepsilon^*)$. By construction, $d(P_1, P_2) < \varepsilon^*$. If $d(P_0, P_2) < \varepsilon_1$, then it would follow from Assumption 1 that $d(P_0, P_1) < \varepsilon_2$. Since $d(P_0, P_1) = \varepsilon_2$, conclude that $d(P_0, P_2) > \varepsilon_1$, that is, $P_2 \notin Pg(P_0, \varepsilon_1)$. Since P_2 is an arbitrary element of $Pg(P_1, \varepsilon^*)$, conclude that $Pg(P_1, \varepsilon^*) \subseteq (Pg(P_0, \varepsilon_1))^c$. $\square$

Proof of Theorem 1. Let ϕ be PROTEST, that is, $\phi(x) = 1$ when $Pg(H_0, \varepsilon)$ is rejected, and otherwise $\phi(x) = 0$.

(a) Since $P \in Pg(H_0, \varepsilon)$, it follows from Lemma 1 that there exists ε_1^* such that $Pg(P, \varepsilon^*) \subseteq Pg(H_0, \varepsilon)$. Also, it follows from Assumption 2.a that there exists ε_2^* such that $Pg^*(P, \varepsilon_2^*) \subseteq Pg(P, \varepsilon_1^*)$. Hence, $Pg^*(P, \varepsilon_2^*) \subseteq Pg(H_0, \varepsilon)$. Using Assumption 2.b, $\mathbb{P}(\{X : \mathbb{P}(Pg^*(P, \varepsilon_2^*) | X) > \alpha\} | P) \overset{n \to \infty}{\longrightarrow} 1$. Since $Pg^*(P, \varepsilon_2^*) \subseteq Pg(H_0, \varepsilon)$, $\mathbb{P}(\{X : \mathbb{P}(Pg(H_0, \varepsilon) | X) \geq \alpha\} | P) \overset{n \to \infty}{\longrightarrow} 1$. The proof follows from observing that the type I error rate is $\mathbb{P}(\{X : \mathbb{P}(Pg(H_0, \varepsilon) | X) < \alpha\} | P)$.

(b) Since $d(P_0, P) > \varepsilon$), it follows from Lemma 1 that there exists ε_1^* such that $Pg(P, \varepsilon^*) \subseteq (Pg(H_0, \varepsilon))^c$. Also, it follows from Assumption 2.a that there exists ε_2^* such that $Pg^*(P, \varepsilon_2^*) \subseteq Pg(P, \varepsilon_1^*)$. Hence, $Pg^*(P, \varepsilon_2^*) \subseteq (Pg(H_0, \varepsilon))^c$. Using Assumption 2.b, $\mathbb{P}(\{X : \mathbb{P}(Pg^*(P, \varepsilon_2^*) | X) > \alpha\} | P) \overset{n \to \infty}{\longrightarrow} 1$. Since $Pg^*(P, \varepsilon_2^*) \subseteq (Pg(H_0, \varepsilon))^c$, $\mathbb{P}(\{X : (\mathbb{P}(Pg(H_0, \varepsilon))^c | X) \geq \alpha\} | P) \overset{n \to \infty}{\longrightarrow} 1$. The proof follows from observing that the type II error rate is $\mathbb{P}(\{X : \mathbb{P}(Pg(H_0, \varepsilon) | X) \geq \alpha\} | P)$. $\square$

Lemma 2 *(Infimum on a Hilbert space from a subspace of linear functionals). Let $\mathcal{H}$ be a Hilbert space and $\boldsymbol{b} = (b_1, b_2, \cdots, b_k)$ be a basis of linear functionals that constitutes the subspace $H \subset \mathcal{H}$. If $d(\cdot, \cdot)$ and $(\cdot, \cdot)$ are the distance function and the scalar product induced by the norm of $\mathcal{H}$ and $f_{\boldsymbol{\beta}} := \sum_{i=1}^k \beta_i \times b_i$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \cdots, \beta_k) \in \mathbb{R}^k$, then $\inf_{h \in H} d(h, g) = \inf_{\beta \in \mathbb{R}^k} d(f_{\boldsymbol{\beta}}, g) = d(f_{\hat{\boldsymbol{\beta}}}, g)$ for $g \in \mathcal{H}$, where*

$$\hat{\boldsymbol{\beta}} = A_{\boldsymbol{b}}^{-1} \times \boldsymbol{g}_{\boldsymbol{b}}, \quad A_{\boldsymbol{b}} = \begin{pmatrix} (b_1, b_1) & (b_2, b_1) & \cdots & (b_k, b_1) \\ (b_1, b_2) & (b_2, b_2) & \cdots & (b_k, b_2) \\ \vdots & \vdots & \ddots & \vdots \\ (b_1, b_k) & (b_2, b_k) & \cdots & (b_k, b_k) \end{pmatrix}, \quad \boldsymbol{g}_{\boldsymbol{b}} = \begin{pmatrix} (g, b_1) \\ (g, b_2) \\ \vdots \\ (g, b_k) \end{pmatrix}.$$

Proof. By construction, H is a closed linear subspace. From corollary 5.4 of [75], for each $g \in \mathcal{H}$, $f_{\hat{\beta}}$ is characterized by

$$(g - f_{\hat{\boldsymbol{\beta}}}, f_{\boldsymbol{\beta}}) = \sum_{j=1}^k \beta_j (g - f_{\hat{\boldsymbol{\beta}}}, b_j) = 0, \quad \forall \boldsymbol{\beta} \in \mathbb{R}^k \Longrightarrow (g - f_{\hat{\boldsymbol{\beta}}}, b_j) = 0, \quad \forall j \in \{1, 2, \cdots, k\}.$$

Therefore,

$$(g - f_{\hat{\boldsymbol{\beta}}}, b_j) = (g, b_j) - \sum_{i=1}^k \hat{\beta}_i (b_i, b_j) = 0, \quad \forall j \in \{1, 2, \cdots, k\},$$

thus leading to the linear system

$$\begin{cases} \sum_{i=1}^{k} \hat{\beta}_i(b_i, b_1) = (g, b_1) \\ \sum_{i=1}^{k} \hat{\beta}_i(b_i, b_2) = (g, b_2) \\ \vdots \\ \sum_{i=1}^{k} \hat{\beta}_i(b_i, b_k) = (g, b_k) \end{cases} \implies A_{\boldsymbol{b}} \times \hat{\boldsymbol{\beta}} = \boldsymbol{g}_{\boldsymbol{b}} \implies \hat{\boldsymbol{\beta}} = A_{\boldsymbol{b}}^{-1} \times \boldsymbol{g}_{\boldsymbol{b}}. \quad \square$$

Proof of Theorem 2. We note that $\mathbb{H} \equiv L_2(\mathcal{X}, \sigma(\mathcal{X}), \mathbb{P})$ is a Hilbert space and that $\text{span}\{b_1, b_2, \cdots, b_k\} = H_0$, therefore Lemma 2 follows by switching H for H_0. Moreover,

$$(b_i, b_j) = \int_{\mathcal{X}} b_i(\boldsymbol{x}) b_j(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) = \mathbb{E}[b_i(\boldsymbol{X}) b_j(\boldsymbol{X})], \quad \forall i, j \in \{1, 2, \cdots, k\};$$

$$(g, b_i) = \int_{\mathcal{X}} g(\boldsymbol{x}) b_i(\boldsymbol{x}) d\mathbb{P}(\boldsymbol{x}) = \mathbb{E}[g(\boldsymbol{X}) b_i(\boldsymbol{X})], \quad \forall i \in \{1, 2, \cdots, k\}. \quad \square$$

Proof of Theorem 3. The proof is done in parts.

- If $P(x_0) = p_0$, then $\inf_{P_0 \in H_0} d(P_0, P) = \int_a^b |p_0 - P(x)| dx = \int_{x_0}^{x_0} |p_0 - P(x)| dx = 0$.
  Subproof. If $P(x_0) = p_0$, then $P \in H_0$. If that is the case,

$$\inf_{P_0 \in H_0} \int_{-\infty}^{\infty} |P_0(x) - P(x)| dx = \int_{-\infty}^{\infty} |P(x) - P(x)| dx = \int_{-\infty}^{\infty} 0 dx = 0. \quad \blacksquare$$

- If $P(x_0) < p_0$, then $\inf_{P_0 \in H_0} d(P_0, P) = \int_a^b |p_0 - P(x)| dx$.
  Subproof. $P(x_0) < p_0 \implies a = x_0$ and $b = P^{-1}(p_0)$. Let $P^*(\cdot)$ be such that

$$P^*(x) := \begin{cases} p_0, & \text{if } x \in [x_0, b]; \\ P(x), & \text{otherwise.} \end{cases}$$

  Thus, proving the result is equivalent to showing that

$$\inf_{P_0 \in H_0} d(P_0, P) = d(P^*, P) = \int_a^b |p_0 - P(x)| dx.$$

  Suppose by contradiction that $\exists P' \in H_0 : d(P^*, P) > d(P', P)$. Hence,

$$\int_a^b |p_0 - P(x)| dx > \int_{-\infty}^{\infty} |P'(x) - P(x)| dx \geq \int_a^b |P'(x) - P(x)| dx. \tag{B.3}$$

  For $x \in [a, b]$, $P(x) \leq p_0 \leq P'(x) \implies P(x) - p_0 \leq 0 \leq P'(x) - p_0$. Since $[P'(x) - p_0] - [P(x) - p_0] = |P'(x) - p_0| + |P(x) - p_0|$,

$$\begin{aligned} \int_a^b |P'(x) - P(x)| dx &= \int_a^b |[P'(x) - p_0] - [P(x) - p_0]| dx \\ &= \int_a^b |P'(x) - p_0| + |P(x) - p_0| dx \geq \int_a^b |p_0 - P(x)| dx, \end{aligned}$$

  which contradicts (B.3), therefore $\inf_{P_0 \in H_0} d(P_0, P) = \int_a^b |p_0 - P(x)| dx$. $\blacksquare$

- If $P(x_0) > p_0$, then $\inf_{P_0 \in H_0} d(P_0, P) = \int_a^b |p_0 - P(x)| dx$.
  Subproof. $P(x_0) > p_0 \implies b = x_0$. Let $(P_n^*)_{n \geq 1}$ be a sequence of distribution functions such that

$$P_n^*(x) := \begin{cases} p_0, & \text{if } x \in \left[a, x_0 + \frac{1}{n}\right); \\ P(x), & \text{otherwise.} \end{cases}$$

  By construction, $P_n^* \in H_0, \forall n \geq 1$, and

$$d(P_n^*, P) = \int_a^{x_0 + \frac{1}{n}} |p_0 - P(x)| dx = \int_a^{x_0} |p_0 - P(x)| dx + \int_{x_0}^{x_0 + \frac{1}{n}} |p_0 - P(x)| dx,$$

which converges decreasingly to $\int_a^{x_0} |p_0 - P(x)|dx$ as $n \to \infty$.

The proof follows by contradiction. Suppose $\exists P' \in H_0 : \inf_{P_0 \in H_0} d(P_0, P) = d(P', P) \neq \int_a^b |p_0 - P(x)|dx$. Similarly to the previous subproof,

$$\begin{aligned}\int_a^b |P'(x) - P(x)|dx &= \int_a^b |[P(x) - p_0] - [P'(x) - p_0]|dx \\ &= \int_a^b |P(x) - p_0|dx + \int_a^b |P'(x) - p_0|dx \\ &\geq \int_a^b |p_0 - P(x)|dx,\end{aligned}$$

and thus $d(P', P) > \int_a^b |p_0 - P(x)|dx$. But since $d(P_n^*, P) \xrightarrow{n\to\infty} \int_a^b |p_0 - P(x)|dx$, then $\exists n_0 \in \mathbb{N}$ such that, $\forall n \geq n_0$,

$$d(P_n^*, P) < d(P', P) \Longrightarrow \inf_{P_0 \in H_0} d(P_0, P) \neq d(P', P),$$

therefore $\inf_{P_0 \in H_0} d(P_0, P) = \int_a^b |p_0 - P(x)|dx$. ■

Based on each of the subproofs presented, we can safely conclude that

$$\inf_{P_0 \in H_0} d(P_0, P) = \int_a^b |p_0 - P(x)|dx. \quad \square$$

Proof of Theorem 4. Without loss of generality, we assume that $\mathbb{H} = \mathbb{F}_X \times \mathbb{F}_Y = \mathbb{F}_X \times \mathbb{F}_X = \mathbb{F}_Y \times \mathbb{F}_Y$. After all, if X and Y were defined on different distribution spaces, we could simply take $\mathbb{F} := \mathbb{F}_X \cup \mathbb{F}_Y$ and use this space instead.

The null hypothesis asserts that, as long as $F_X = F_Y$, the distribution function of both random variables can be any element of $\mathbb{F}$. Thus, if Ω is the sample space,

$$H_0 : (F_X, F_Y) \in \mathbb{F} \times \mathbb{F} : F_X(z) = F_Y(z), \forall z \in \Omega.$$

Therefore,

$$\begin{aligned}Pg(H_0) &= \left\{ (P_X, P_Y) \in \mathbb{F} \times \mathbb{F} : \inf_{(F_X, F_Y) \in H_0} d[(F_X, F_Y), (P_X, P_Y)] < \varepsilon \right\} \\ &= \left\{ (P_X, P_Y) \in \mathbb{F} \times \mathbb{F} : \inf_{P_0 \in \mathbb{F}} d[(P_0, P_0), (P_X, P_Y)] < \varepsilon \right\}.\end{aligned}$$

From (13),

$$\inf_{P_0 \in \mathbb{F}} d[(P_0, P_0), (P_X, P_Y)] = \inf_{P_0 \in \mathbb{F}} [d^*(P_0, P_X) + d^*(P_0, P_Y)]. \tag{B.4}$$

Now, since d^* is a distance function, the properties of symmetry and triangle inequality [76] imply that

$$\inf_{P_0 \in \mathbb{F}} [d^*(P_0, P_X) + d^*(P_0, P_Y)] = \inf_{P_0 \in \mathbb{F}} [d^*(P_X, P_0) + d^*(P_0, P_Y)] \geq d^*(P_X, P_Y). \tag{B.5}$$

Since $P_X \in \mathbb{F}$, the equality in (B.5) is guaranteed if $P_0 = P_X$. $\square$

Data availability

The code used for this work is publicly available on GitHub and all data has been obtained from sources adequately cited throughout this work.

References

[1] W. Edwards, H. Lindman, L.J. Savage, Bayesian statistical inference for psychological research, Psychol. Rev. 70 (3) (1963) 193, https://doi.org/10.1037/h0044139.
[2] I.J. Good, Some logic and history of hypothesis testing, in: Dover Books on Mathematics, Dover Publications, 2009, pp. 129–148.
[3] J.O. Berger, Statistical Decision Theory and Bayesian Analysis, Springer-Verlag, New York, 1985.
[4] B.P. Hobbs, B.P. Carlin, Practical Bayesian design and analysis for drug and device clinical trials, J. Biopharm. Stat. 18 (1) (2007) 54–80, https://doi.org/10.1080/10543400701668266.

[5] J.K. Kruschke, Rejecting or accepting parameter values in Bayesian estimation, Adv. Methods Pract. Psychol. Sci. 1 (2) (2018) 270–280, https://doi.org/10.1177/2515245918771304.
[6] E.E. Leamer, 3 things that bother me, Econ. Rec. 64 (4) (1988) 331–335, https://doi.org/10.1111/j.1475-4932.1988.tb02072.x.
[7] H. Jeffreys, Theory of Probability, 3rd edition, Oxford, Oxford, England, 1998.
[8] R.E. Kass, Bayes factors in practice, J. R. Stat. Soc., Ser. D, Stat. 42 (5) (1993) 551–560, https://doi.org/10.2307/2348679.
[9] H. Migon, D. Gamerman, F. Louzada, Statistical Inference: An Integrated Approach, second edition, Chapman & Hall/CRC Texts in Statistical Science, CRC Press, 2014.
[10] G. Corani, A. Benavoli, F. Mangili, M. Zaffalon, Bayesian hypothesis testing in machine learning, in: A. Bifet, M. May, B. Zadrozny, R. Gavalda, D. Pedreschi, F. Bonchi, J. Cardoso, M. Spiliopoulou (Eds.), Machine Learning and Knowledge Discovery in Databases, Springer International Publishing, Cham, 2015, pp. 199–202.
[11] J.L. Hodges, E.L. Lehmann, Testing the approximate validity of statistical hypotheses, J. R. Stat. Soc., Ser. B, Methodol. 16 (2) (1954) 261–268, https://doi.org/10.1111/j.2517-6161.1954.tb00169.x.
[12] L.G. Esteves, R. Izbicki, J.M. Stern, R.B. Stern, Pragmatic hypotheses in the evolution of science, Entropy 21 (9) (2019), https://doi.org/10.3390/e21090883.
[13] T. Augustin, R. Hable, On the impact of robust statistics on imprecise probability models: a review, Struct. Saf. 32 (6) (2010) 358–365, https://doi.org/10.1016/j.strusafe.2010.06.002, modeling and Analysis of Rare and Imprecise Information.
[14] S. Destercke, I. Montes, E. Miranda, Processing distortion models: a comparative study, Int. J. Approx. Reason. 145 (2022) 91–120, https://doi.org/10.1016/j.ijar.2022.03.007.
[15] I. Montes, E. Miranda, S. Destercke, Unifying neighbourhood and distortion models: part I – new results on old models, Int. J. Gen. Syst. 49 (6) (2020) 602–635, https://doi.org/10.1080/03081079.2020.1778682.
[16] R.D. Morey, J.N. Rouder, Bayes factor approaches for testing interval null hypotheses, Psychol. Methods 16 (4) (2011) 406–419, https://doi.org/10.1037/a0024377.
[17] D. van Ravenzwaaij, R. Monden, J.N. Tendeiro, J.P.A. Ioannidis, Bayes factors for superiority, non-inferiority, and equivalence designs, BMC Med. Res. Methodol. 19 (1) (Mar. 2019), https://doi.org/10.1186/s12874-019-0699-7.
[18] G. Corani, A. Benavoli, J. Demšar, F. Mangili, M. Zaffalon, Statistical comparison of classifiers through Bayesian hierarchical modelling, Mach. Learn. 106 (11) (2017) 1817–1837, https://doi.org/10.1007/s10994-017-5641-9.
[19] J. Fitzgerald, The Need for Equivalence Testing in Economics, I4R Discussion Paper Series 125, Institute for Replication (I4R), 2024, s.l., https://hdl.handle.net/10419/296190.
[20] A. Benavoli, G. Corani, J. Demšar, M. Zaffalon, Time for a change: a tutorial for comparing multiple classifiers through Bayesian analysis, J. Mach. Learn. Res. 18 (77) (2017) 1–36, http://jmlr.org/papers/v18/16-305.html.
[21] A. Benavoli, C. de Campos, Bayesian independence test with mixed-type variables, in: 2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA), 2021, pp. 1–13.
[22] C.C. Holmes, F. Caron, J.E. Griffin, D.A. Stephens, Two-sample Bayesian nonparametric hypothesis testing, Bayesian Anal. 10 (2) (2015) 297–320, https://doi.org/10.1214/14-BA914.
[23] A.N. Kolmogorov, 15. On the Empirical Determination of a Distribution Law, Springer Netherlands, Dordrecht, 1992, pp. 139–146.
[24] R Core Team R, A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2022, https://www.R-project.org/.
[25] H.A. Duguid, A study of the evaporation rates of small freely falling water droplets, Master's thesis, Missouri University of Science and Technology, Rolla, Missouri, USA, 1969.
[26] Y. Yao, S. Wong, A decision theoretic framework for approximating concepts, Int. J. Man-Mach. Stud. 37 (6) (1992) 793–809, https://doi.org/10.1016/0020-7373(92)90069-W.
[27] C.E. Rasmussen, C.K.I. Williams, Gaussian Processes for Machine Learning, The MIT Press, 2005.
[28] A. Benavoli, F. Mangili, Gaussian processes for Bayesian hypothesis tests on regression functions, in: G. Lebanon, S.V.N. Vishwanathan (Eds.), Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics, PMLR, San Diego, California, USA, in: Proceedings of Machine Learning Research, vol. 38, 2015, pp. 74–82.
[29] J. Liu, B. Coull, Robust hypothesis test for nonlinear effect with Gaussian processes, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, vol. 30, Curran Associates, Inc., 2017, https://proceedings.neurips.cc/paper_files/paper/2017/file/2bb232c0b13c774965ef8558f0fbd615-Paper.pdf.
[30] R.E. Kass, A.E. Raftery, Bayes factors, J. Am. Stat. Assoc. 90 (430) (1995) 773–795, https://doi.org/10.1080/01621459.1995.10476572.
[31] J.K. Nielsen, M.G. Christensen, A.T. Cemgil, S.H. Jensen, Bayesian model comparison with the g-prior, IEEE Trans. Signal Process. 62 (1) (2014) 225–238, https://doi.org/10.1109/TSP.2013.2286776.
[32] C.A.d.B. Pereira, J.M. Stern, Evidence and credibility: full Bayesian significance test for precise hypotheses, Entropy 1 (4) (1999) 99–110, https://doi.org/10.3390/e1040099.
[33] C.A.B. Pereira, J.M. Stern, The e-value: a fully Bayesian significance measure for precise statistical hypotheses and its research program, São Paulo J. Math. Sci. 16 (1) (2020) 566–584, https://doi.org/10.1007/s40863-020-00171-7.
[34] R. Kelter, On the measure-theoretic premises of Bayes factor and full Bayesian significance tests: a critical reevaluation: commentary to Ly and wagenmakers, Comput. Brain Behav. 5 (4) (2021) 572–582, https://doi.org/10.1007/s42113-021-00110-5.
[35] S. Cabras, W. Racugno, L. Ventura, Higher order asymptotic computation of Bayesian significance tests for precise null hypotheses in the presence of nuisance parameters, J. Stat. Comput. Simul. 85 (15) (2014) 2989–3001, https://doi.org/10.1080/00949655.2014.947288.
[36] P. Schwaferts, T. Augustin, Bayesian decisions using regions of practical equivalence (rope): Foundations, Tech. Rep., Universitätsbibliothek der Ludwig-Maximilians-Universität München, 2020, https://doi.org/10.5282/UBM/EPUB.74222, https://epub.ub.uni-muenchen.de/id/eprint/74222.
[37] M. Schervish, Theory of Statistics, Springer Series in Statistics, Springer, New York, 1995.
[38] A. Barron, M.J. Schervish, L. Wasserman, The consistency of posterior distributions in nonparametric problems, Ann. Stat. 27 (2) (Apr. 1999), https://doi.org/10.1214/aos/1018031206.
[39] I. Castillo, A semiparametric Bernstein–von Mises theorem for Gaussian process priors, Probab. Theory Relat. Fields 152 (1–2) (2012) 53–99, https://doi.org/10.1007/s00440-010-0316-5.
[40] S. Ghosal, J.K. Ghosh, A.W. van der Vaart, Convergence rates of posterior distributions, Ann. Stat. 28 (2) (2000) 500–531, https://doi.org/10.1214/aos/1016218228.
[41] X. Shen, L. Wasserman, Rates of convergence of posterior distributions, Ann. Stat. 29 (3) (2001) 687–714, https://doi.org/10.1214/aos/1009210686.
[42] T.S. Ferguson, A Bayesian analysis of some nonparametric problems, Ann. Stat. 1 (2) (1973) 209–230, https://doi.org/10.1214/aos/1176342360.
[43] M. Lavine, Some aspects of Polya tree distributions for statistical modelling, Ann. Stat. 20 (3) (1992) 1222–1235, https://doi.org/10.1214/aos/1176348767.
[44] M. Lavine, More aspects of Polya tree distributions for statistical modelling, Ann. Stat. 22 (3) (1994) 1161–1176, https://doi.org/10.1214/aos/1176325623.
[45] J. Neyman, E.S. Pearson, On the problem of the most efficient tests of statistical hypotheses, Philos. Trans. R. Soc. Lond., Ser. A, Contain. Pap. Math. Phys. Character 231 (1933) 289–337, https://doi.org/10.1098/rsta.1933.0009.
[46] C. Erickson, GauPro: Gaussian process fitting, r package version 0.2.13.9000, https://github.com/collinerickson/gaupro, 2024.

[47] C.P. Robert, G. Casella, Monte Carlo Statistical Methods, 2nd edition, Springer Texts in Statistics, Springer, Berlin, 2005.
[48] M.H. de Almeida Inácio, R. Izbicki, L.E. Salasar, Comparing two populations using Bayesian Fourier series density estimation, Commun. Stat., Simul. Comput. 49 (1) (2018) 261–282, https://doi.org/10.1080/03610918.2018.1484480.
[49] R. de Carvalho Ceregatti, R. Izbicki, L.E.B. Salasar, Wiks: a general Bayesian nonparametric index for quantifying differences between two populations, Test 30 (2021) 274–291, https://doi.org/10.1007/s11749-020-00718-y.
[50] J.H. Gross, Testing what matters (if you must test at all): a context-driven approach to substantive and statistical significance, Am. J. Polit. Sci. 59 (3) (2014) 775–788, https://doi.org/10.1111/ajps.12149.
[51] D. Lakens, A.M. Scheel, P.M. Isager, Equivalence testing for psychological research: a tutorial, Adv. Methods Pract. Psychol. Sci. 1 (2) (2018) 259–269, https://doi.org/10.1177/2515245918770963.
[52] W. Wang, L. Lin, Derivative estimation based on difference sequence via locally weighted least squares regression, J. Mach. Learn. Res. 16 (81) (2015) 2617–2641, https://doi.org/10.5555/2789272.2912083.
[53] G. Folland, Real Analysis: Modern Techniques and Their Applications, Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts, Wiley, 2013.
[54] D. Lakens, Improving your statistical inferences, https://doi.org/10.5281/ZENODO.6409077, https://zenodo.org/record/6409077, 2022.
[55] C. Goldin, Understanding the Gender Gap: An Economic History of American Women, American Studies Collection, Oxford University Press, 1990.
[56] F.D. Blau, L.M. Kahn, The gender wage gap: extent, trends, and explanations, J. Econ. Lit. 55 (3) (2017) 789–865, https://doi.org/10.1257/jel.20160995.
[57] UNECE, UNECE statistical database, https://w3.unece.org/PXWeb2015/pxweb/en/STAT/, 2023. (Accessed 23 September 2023).
[58] M. DeYoreo, A. Kottas, A fully nonparametric modeling approach to binary regression, Bayesian Anal. 10 (4) (2015) 821–847, https://doi.org/10.1214/15-BA963SI.
[59] J. Cohen, The Earth is round ($p < 05$), Am. Psychol. 49 (12) (1994) 997–1003, https://doi.org/10.1037/0003-066x.49.12.997.
[60] J. Faber, L.M. Fonseca, How sample size influences research outcomes, Dent. Press J. Orthod. 19 (4) (2014) 27–29, https://doi.org/10.1590/2176-9451.19.4.027-029.ebo.
[61] R.A. Fisher, F. Yates, Statistical Tables for Biological, Agricultural and Medical Research, sixth edition, 6th edition, Hafner Publishing Company, 1963.
[62] G. Casella, R. Berger, Statistical Inference, Chapman and Hall/CRC, 2024.
[63] P. Mitchell, Teaching statistical appreciation in quantitative methods, MSOR Connections 16 (2) (2018) 37, https://doi.org/10.21100/msor.v16i2.554.
[64] M.A. Pett, Nonparametric Statistics for Health Care Research: Statistics for Small Samples and Unusual Distributions, SAGE Publications, Inc, 2016.
[65] T. Hanson, W.O. Johnson, Modeling regression error with a mixture of Polya trees, J. Am. Stat. Assoc. 97 (460) (2002) 1020–1033, https://doi.org/10.1198/016214502388618843.
[66] S.M. Ross, A First Course in Probability, 8th edition, Pearson, 2012.
[67] M.C. Faraut, A.A. Carlson, S. Sullivan, O. Tudusciuc, I. Ross, C.M. Reed, J.M. Chung, A.N. Mamelak, U. Rutishauser, Dataset of human medial temporal lobe single neuron activity during declarative memory encoding and recognition, Sci. Data 5 (1) (2018), https://doi.org/10.1038/sdata.2018.10.
[68] W. Gerstner, W.M. Kistler, Spiking Neuron Models, Cambridge University Press, 2002.
[69] P.M. Schwaferts, T. Augustin, Imprecise hypothesis-based Bayesian decision making with composite hypotheses, in: A. Cano, J. De Bock, E. Miranda, S. Moral (Eds.), Proceedings of the Twelveth International Symposium on Imprecise Probability: Theories and Applications, in: Proceedings of Machine Learning Research, vol. 147, PMLR, 2021, pp. 280–288.
[70] J.M. Stern, R. Izbicki, L.G. Esteves, R.B. Stern, Logically-consistent hypothesis testing and the hexagon of oppositions, Log. J. IGPL 25 (5) (2017) 741–757, https://doi.org/10.1093/jigpal/jzx024.
[71] L.G. Esteves, R. Izbicki, J.M. Stern, R.B. Stern, The logical consistency of simultaneous agnostic hypothesis tests, Entropy 18 (7) (2016), https://doi.org/10.3390/e18070256.
[72] L.G. Esteves, R. Izbicki, J.M. Stern, R.B. Stern, Logical coherence in Bayesian simultaneous three-way hypothesis tests, Int. J. Approx. Reason. 152 (2023) 297–309, https://doi.org/10.1016/j.ijar.2022.10.019.
[73] J. Sethuraman, A constructive definition of Dirichlet priors, Stat. Sin. 4 (2) (1994) 639–650, http://www.jstor.org/stable/24305538.
[74] E.G. Phadia, Prior Processes and Their Applications: Nonparametric Bayesian Estimation, 2nd edition, Springer International Publishing, 2016.
[75] H. Brezis, Functional Analysis, Sobolev Spaces and Partial Differential Equations, Springer, New York, 2011.
[76] E. Kreyszig, Introductory Functional Analysis with Applications, Wiley Classics Library, Wiley, 1978.