Journal of Machine Learning Research 27 (2026) 1-42 Submitted 3/26; Revised 9/26; Published 10/26

Bayesian Predictive Inference Beyond Martingales

Marco Battiston M.BATTISTON@LANCASTER.AC.UK
School of Mathematical Sciences,
Lancaster University
Lancaster, UK

Lorenzo Cappello LORENZO.CAPPELLO@UPF.EDU
Department of Economics and Business,
Universitat Pompeu Fabra,
Data Science Center, Barcelona School of Economics
Barcelona, Spain

Editor: Pierre Alquier

Abstract

There is growing interest in the so-called Bayesian Predictive Inference approach, which enables Bayesian inference without specifying a likelihood function or prior distribution and without resorting to MCMC methods. Instead, only a sequence of predictive distributions for the observations is required, and inference on the unknown estimand can be performed, cheaply in parallel, using bootstrap-type schemes. Understanding which classes of predictive distributions can be used within this framework, is still a key open question. We relax commonly used probabilistic assumptions on the observations, namely exchangeability and conditionally identical distribution, and on their predictive distributions, which are measure-valued martingales, by introducing the new class of *Almost Conditionally Identically Distributed* (a.c.i.d.) random variables. This class assumes that the predictive distributions are measure-valued almost supermartingales, and is parametrized by a sequence of parameters $(\xi_n)_{n\geq 0}$, which regulate the decay of conditional dependence among future observations. Under mild summability assumptions on $(\xi_n)_{n\geq 0}$, the resulting sequence of observations is shown to be asymptotically exchangeable, hence amenable to Bayesian Predictive Inference techniques. A.c.i.d. random variables arise naturally in recursive algorithms, and include classic approaches in Statistics and Learning Theory, such as kernel estimators, and more novel ones, such as the parametric Bayesian bootstrap.

Keywords: Almost Supermartingales, Conditional identical distribution, Kernel Methods, Martingale Posterior, Parametric Bootstrap, Predictive Resampling.

1. Introduction

Prediction lies at the foundation of Bayesian Statistics. Traditionally, much research has focused on the relationship between a specified statistical model and prior, and the resulting predictive distribution of future observations given current information (Geisser, 1993; Shmueli, 2010; Clarke and Clarke, 2018; Efron, 2020); see also Berti et al. (2025) and Fortini and Petrone (2025) for recent reviews. However, when the primary goal is inference, a (Bayesian) statistician usually specifies a statistical model and a prior, and then, given a sample, computes the corresponding posterior distribution. Predictive distributions do not play a role in this process; possibly except for the necessary checks of model fit (e.g. predictive checks, Gelman et al., 1996). A recent stream of

©2026 Marco Battiston and Lorenzo Cappello.

License: CC-BY 4.0, see https://creativecommons.org/licenses/by/4.0/. Attribution requirements are provided at http://jmlr.org/papers/v27/26-0847.html.

literature is taking a new stance, suggesting that it is possible to do Bayesian inference by simply specifying a *predictive rule* - how to update the prediction as new information becomes available - fully bypassing the need for models and priors (Fortini and Petrone, 2020; Berti et al., 2021; Fong et al., 2023; Holmes and Walker, 2023; Fortini and Petrone, 2023). The peculiarity of this approach is that prediction is done with inference in mind rather than forecasting. This emerging area, often referred to as *Bayesian predictive inference* or *modeling*, is the focus of the paper.

Several reasons support the popularity of this growing research area. First, it bypasses the need to specify a statistical model and subjective probabilities on latent parameters, which may lack interpretability or be high-dimensional. Instead, it relies on a probabilistic learning rule on the observables - an assumption that can be formally assessed, and aligns with de Finetti's view that probabilistic statements should be made on observable events. Second, it is computationally appealing: one can avoid MCMC methods, and rely on numerical methods that speed up inference (Fong et al., 2023; Fortini and Petrone, 2020). Lastly, it has the potential to bridge the two cultures —statistical inference and prediction (Breiman, 2001) — by enabling the use of machine learning algorithms for inference. However, the current framework requires stringent assumptions on the predictive rule, and many popular algorithms violate these. This paper proposes a relaxation of these assumptions to enlarge the class of algorithms admissible in predictive inference, thus paving the way for this bridge.

Let θ be an estimand and $x_{1:n}$ a sample. The key observation underlying predictive inference is that the uncertainty described by a posterior distribution $\mathbb{P}(\theta|x_{1:n})$ can be equivalently described through the predictive distribution of the yet-to-be-seen observations $\mathbb{P}(X_{n+1:\infty}|x_{1:n})$, or equivalently a sequence of one-step ahead distributions $(\mathbb{P}(X_{n+m}|x_{1:n+m-1}))_{m\geq 1}$. The scheme to move from prediction to inference is called *Predictive Resampling*, which is a generalization of the Bayesian bootstrap (Rubin, 1981), where one samples many synthetic data points from $\mathbb{P}(X_{n+1:n+M}|x_{1:n})$ for some large M, and then for each synthetic sample computes an estimate $\hat{\theta} = f(x_{1:n+M})$, where f is an appropriately chosen estimator which crucially uses both the original and the synthetic data. The resulting empirical distribution of $\hat{\theta}$ is an approximation to the so-called *Martingale Posterior* (Fong et al., 2023); see Subsection 2.2. Crucially, under suitable conditions, this posterior is fully equivalent to the one obtained with the usual application of Bayes' rule starting with a likelihood and a prior.

What conditions the predictive rule needs to satisfy to get a valid martingale posterior is one of the key research questions. Given a filtration $\mathcal{G} = (\mathcal{G}_n)_{n\geq 1}$ (for example, the natural filtration generated by $(X_n)_{n\geq 1}$), Fong et al. (2023) assume that the predictive distributions are measure-valued martingales, that is $(\mathbb{P}(X_{n+1} \in A|\mathcal{G}_n))_{n\geq 0}$ is a martingale for all Borel sets A. The same condition appears also in Fortini and Petrone (2020). The martingale assumption has many desirable implications. It represents a departure from the assumption of exchangeability on $(X_n)_{n\geq 1}$ to the more general class of *Conditionally Identically Distributed* (c.i.d.) sequences (Kallenberg, 1988; Berti et al., 2004). This enlarges the class of predictive distributions and models that can be considered, moving away from exchangeable Bayesian predictive schemes (e.g. Fortini and Petrone, 2020). Such a departure comes at no substantial cost: while the convergence of predictive distributions for exchangeable sequences is given by de Finetti's representation theorem, for c.i.d. sequences the random sequence $(\mathbb{P}(X_{n+1} \in \cdot|\mathcal{G}_n))_{n\geq 0}$ converges weakly almost surely (a.s.) by martingale arguments (see Aldous, 1985). Crucially, if $(X_n)_{n\geq 1}$ is c.i.d., it is also *Asymptotically Exchangeable*, meaning that it behaves in the limit as an exchangeable sequence (Aldous, 1985). This property

justifies the validity of the martingale posterior as an approximation for $\mathbb{P}(\theta|x_{1:n})$. What is appealing is that there is no need to specify a model and a prior or even assume exchangeability.

The martingale assumption has other desirable features; see Fong et al. (2023). Nevertheless, what matters is the almost sure weak convergence of $(\mathbb{P}(X_{n+1} \in \cdot|\mathcal{G}_n))_{n\geq 0}$ and the resulting asymptotic exchangeability of $(X_n)_{n\geq 1}$. It is becoming clear that the martingale assumption could be limiting, and the literature now includes many examples that violate this assumption; for example, the parametric Bayesian bootstrap (Holmes and Walker, 2023), an online logistic regression (Fortini and Petrone, 2025), and predictive schemes driven by sample moments (Garelli et al., 2026a). Several algorithms in learning theory and machine learning (such as kernel estimators), if considered as predictive rules, do not satisfy the martingale property either.

1.1 Preview of Contributions

In this paper, we relax the martingale assumption on predictive distributions, by assuming instead that the sequence of one-step ahead predictive distributions $(\mathbb{P}(X_{n+1} \in A|\mathcal{G}_n))_{n\geq 0}$ forms an almost supermartingale, (Robbins and Siegmund, 1971) (see Appendix A.1 for background), for every set A. Informally, it means that the sequence is not a martingale but converges to one asymptotically. The most immediate consequence of the departure from martingales is that the resulting sequence $(X_n)_{n\geq 1}$ is not necessarily c.i.d. A similar departure from the martingale property is studied in the concurrent work of Leisen et al. (2026).

In Section 3, we define *Almost Conditionally Identically Distributed* (a.c.i.d.) sequences, which generalize both c.i.d. and exchangeable sequences, and discuss a few characterizations. It is then shown in Theorem 6 that, under mild conditions, an a.c.i.d. sequence is asymptotically exchangeable. This is done by proving that any sequence of random measures that is an almost supermartingale converges weakly a.s. This result generalizes similar results for martingales (Aldous, 1985), and may be of general interest beyond the scope of this paper. An implication is that any a.c.i.d. sequence satisfying these mild conditions can be used in predictive inference.

Section 4 is devoted to illustrations showcasing the benefits of doing Bayesian predictive inference with a.c.i.d. sequences. We consider predictive rules built using old ideas in Statistics and Machine Learning, and new ones. We actually consider the former class more interesting to illustrate how Bayesian predictive inference can incorporate ideas from fields centered around prediction into a formal inference framework. For example, we show that a sequence of predictive distributions defined by kernel estimates generates an a.c.i.d. sequence. A similar result holds for an online version of the Random Forest, exploiting the link between kernel methods and Random Forest (Breiman, 2000). Lastly, another example is the Bayesian parametric bootstrap (Holmes and Walker, 2023; Fong and Yiu, 2026), which satisfies the property under certain regularity conditions.

Although the illustrations in this paper primarily serve to demonstrate examples of a.c.i.d. schemes, many of them are already of independent interest as predictive methods. For instance, kernel-based predictive schemes offer notable advantages: they do not require evaluation on a numerical grid and can be less sensitive to the ordering of data, in contrast to copula-based methods (Hahn et al., 2018; Fong et al., 2023) and predictive recursion approaches (Newton et al., 1998; Fortini and Petrone, 2020). In addition, framing kernels as a.c.i.d. procedures may be useful for interpreting these algorithms in a proper statistical framework, as done by Fortini and Petrone (2020) for Newton's predictive recursion. Informally, our result suggests that there is a statistical model underlying kernel methods, thus addressing a question posed by West (1991). In Section 5, we explore

some numerical properties of these schemes, while leaving a more comprehensive investigation of kernel methods for predictive inference to future work. Section 6 provides concluding remarks.

Notation. In the rest of the paper, all random variables (r.v.s) are assumed to be defined on a common probability space $(\Omega, \mathcal{A}, \mathbb{P})$, and $\mathcal{G} = (\mathcal{G}_n)_{n\geq 0}$ denotes a filtration on it, with $\mathcal{G}_0 = \{\emptyset, E\}$. $(X_n)_{n\geq 1}$ denotes a sequence of random variables taking values on a Polish space $(E, \mathcal{E})$. For most examples, $(E, \mathcal{E})$ can be taken as $(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$. $X_{1:n}$ is used as shorthand notation for the vector $(X_1, \ldots, X_n)$. $\overset{\text{w}}{\rightarrow}$, $\overset{\text{d}}{\rightarrow}$ and $\overset{\text{a.s.}}{\rightarrow}$ stand for weak convergence of probability measures, convergence in distribution and almost sure convergence of r.v.s, respectively. $\text{TV}(\alpha_1, \alpha_2) = \sup_{A\in\mathcal{E}} |\alpha_1(A) - \alpha_2(A)|$ denotes the total variation distance between the probability measures α_1 and α_2. Finally, $\alpha_n(\omega, A) := \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n)(\omega)$ denotes the one-step ahead predictive distribution, given $\mathcal{G}_n$. In $\alpha_n(\omega, A)$, we will often not highlight the dependence on ω, and simply write $\alpha_n(A)$, and write $(\alpha_n)_{n\geq 0}$ for the random sequence $(\alpha_n(\cdot, \cdot))_{n\geq 0}$.

2. Preliminaries

In this section, we provide some background material. Specifically, Subsection 2.1 reviews two definitions of c.i.d. r.v.s, which will be both generalized to the a.c.i.d. case in Section 3. Subsection 2.2 reviews the Predictive Resampling algorithm, whose application to a.c.i.d. r.v.s will be then illustrated in Section 5.

2.1 Conditionally Identically Distributed Sequences

Informally, a sequence of r.v.s is c.i.d. if, at any point in time, given the current history of the process $\mathcal{G}_n$, future random variables are identically distributed. This form of probabilistic dependence has been thoroughly studied in Kallenberg (1988) and Berti et al. (2004) from a theoretical point of view, and important examples of c.i.d. r.v.s were proposed in Bassetti et al. (2010), Airoldi et al. (2014), Cassese et al. (2019) and Chen et al. (2023). C.i.d. r.v.s can be formally defined as follows,

Definition 1 (C.i.d., Law) *A sequence of $\mathcal{G}$-adapted r.v.s. $(X_n)_{n\geq 1}$ is Conditionally Identically Distributed (c.i.d.) with respect to $\mathcal{G}$ if for all $k > n \geq 0$ and for all $A \in \mathcal{E}$*

$$\mathbb{P}(X_k \in A|\mathcal{G}_n) = \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n) \qquad a.s.. \tag{1}$$

In Definition 1, it is easy to check that it is actually enough to require that $\mathbb{P}(X_{n+2} \in A|\mathcal{G}_n) = \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n)$ for all $n \geq 0$ and $A \in \mathcal{E}$, that is X_{n+1} and X_{n+2} are identically distributed, conditionally on the current history $\mathcal{G}_n$. Moreover, using usual approximations with simple functions, it is possible to restate Definition 1 with condition (1) holding for expectations over all integrable functions $f : E \to \mathbb{R}$, that is, $\mathbb{E}(f(X_k)|\mathcal{G}_n) = \mathbb{E}(f(X_{n+1})|\mathcal{G}_n)$, instead of just for indicator functions, $f(x) = \mathbb{I}(x \in A)$. Another, possibly less obvious, equivalent definition of c.i.d. r.v.s is through the martingale property of their predictive distributions.

Definition 2 (C.i.d., Predictive) *A sequence of $\mathcal{G}$-adapted r.v.s. $(X_n)_{n\geq 1}$ is c.i.d. with respect to $\mathcal{G}$ if the sequence $(\mathbb{P}(X_{n+1} \in A|\mathcal{G}_n))_{n\geq 0}$ forms a $\mathcal{G}$-martingale, for each $A \in \mathcal{E}$.*

Informally, we will often write that a sequence of predictive distributions is c.i.d. to mean that the corresponding sequence $(X_n)_{n\geq 1}$ is c.i.d.

2.2 Predictive Resampling

Predictive resampling generalizes the Bayesian bootstrap (Rubin, 1981) to arbitrary predictive schemes. As in the Introduction, let θ denote an estimand of interest, which may be either finite- or infinite-dimensional, and let f be an estimator such that, given a sample $x_{1:n}$, an estimate for θ is given by $\hat{\theta}_n = f(x_{1:n})$. Both θ and f are arbitrary. Fong et al. (2023) define the finite martingale posterior as follows:

Definition 3 *(Finite Martingale Posterior) The finite martingale posterior is defined as*

$$\mathbb{P}(\hat{\theta}_{n+M} \in \cdot \mid \mathcal{G}_n) = \int \mathbb{I}(f(x_{1:n+M}) \in \cdot)\mathbb{P}(\mathrm{d}x_{n+1:n+M} \mid \mathcal{G}_n). \tag{2}$$

The martingale posterior is obtained as the limit $M \to \infty$. Under the assumption that $(X_n)_{n\geq 1}$ is exchangeable and $(\mathbb{P}(X_{n+1} \in \cdot \mid \mathcal{G}_n))_{n\geq 0}$ the corresponding sequence of predictive distributions, one can show that the martingale posterior coincides with the standard Bayesian posterior. This argument traces back to Doob (1949). If $(X_n)_{n\geq 1}$ is c.i.d., such equivalence may not hold in finite samples but only asymptotically. Informally, this happens because $(X_n)_{n\geq 1}$ behaves in the limit as an exchangeable sequence. We will return to this point in the next section when we define asymptotic exchangeability.

In Bayesian predictive inference, rather than specifying a model and prior, one directly defines the sequence $(\mathbb{P}(X_{n+1} \in \cdot \mid \mathcal{G}_n))_{n\geq 0}$ and uses it to approximate the martingale posterior. Predictive resampling is a Monte Carlo method to do that. Specifically, one draws B samples of size M, $X^{(i)}_{n+1:n+M} \sim \mathbb{P}(X_{n+1:n+M} \in \cdot \mid \mathcal{G}_n)$, for $i = 1, \ldots, B$, and computes $(\hat{\theta}^{(i)}_{n+M})_{1:B}$, where $\hat{\theta}^{(i)}_{n+M} = f(x^{(i)}_{1:n+M})$. The empirical distribution of $(\hat{\theta}^{(i)}_{n+M})_{1:B}$ then serves as an approximation of $\mathbb{P}(\hat{\theta}_{n+M} \in \cdot \mid \mathcal{G}_n)$. Sampling $X_{n+1:n+M}$ can be performed sequentially, one step at a time, using the one-step ahead predictive. If $(\mathbb{P}(X_{n+1} \in \cdot \mid \mathcal{G}_n))_{n\geq 0}$ is the empirical distribution, this scheme coincides with the Bayesian bootstrap. The key novelty is that we can move beyond the discreteness of the empirical distribution and employ continuous updates; for instance, using the Gaussian copula algorithm of Hahn et al. (2018). The same numerical scheme can be used to scrutinize the underlying model and prior; see the application to Newton's algorithm in Fortini and Petrone (2020) for details.

3. Almost Conditionally Identically Distributed Random Variables

We relax Definition 2 of c.i.d. r.v.s by allowing the sequence of predictive distributions to be an almost supermartingale (Robbins and Siegmund, 1971). We remark that a similar generalization has recently been considered in the concurrent work, Leisen et al. (2026), in which the predictive distribution is allowed to be a quasi-martingale (Fisk, 1965). Almost Conditionally Identically Distributed (a.c.i.d.) r.v.s are defined as follows,

Definition 4 (A.c.i.d., Predictive) *A sequence of $\mathcal{G}$-adapted r.v.s. $(X_n)_{n\geq 1}$ is $(\xi_n)_{n\geq 0}$-a.c.i.d. with respect to $\mathcal{G}$ if the sequence $(\mathbb{P}(X_{n+1} \in A|\mathcal{G}_n))_{n\geq 0}$ forms an almost $\mathcal{G}$-supermartingale for all $A \in \mathcal{E}$, that is, for all $n \geq 0$ and $A \in \mathcal{E}$,*

$$\mathbb{E}\left(\mathbb{P}(X_{n+2} \in A|\mathcal{G}_{n+1})|\mathcal{G}_n\right) \leq \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n) + \xi_n \qquad a.s., \tag{3}$$

where $(\xi_n)_{n\geq 0}$ is a sequence of $\mathcal{G}$-adapted non-negative r.v.s.

When $(\xi_n)_{n\geq 0}$ and/or $\mathcal{G}$ are clear from the context, we will simply write that the sequence $(X_n)_{n\geq 1}$ is a.c.i.d., instead of $(\xi_n)_{n\geq 0}$-a.c.i.d. with respect to $\mathcal{G}$. Informally, we will often write that a sequence of predictive distributions is a.c.i.d. to mean that the corresponding $(X_n)_{n\geq 1}$ is a.c.i.d.

Notice that, because $\mathcal{G}_n \subset \mathcal{G}_{n+1}$, formula (3) reduces to $\mathbb{P}(X_{n+2} \in A|\mathcal{G}_n) \leq \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n)+\xi_n$. Since condition (3) must hold for all $A \in \mathcal{E}$, by applying it to A^c, we obtain that it must also hold that $\mathbb{P}(X_{n+2} \in A|\mathcal{G}_n) \geq \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n) - \xi_n$ for all $A \in \mathcal{E}$. Finally, using condition (3) on both A and A^c, it follows that, for $A \in \mathcal{E}$, $|\mathbb{P}(X_{n+2} \in A|\mathcal{G}_n) - \mathbb{P}(X_{n+1} \in A|\mathcal{G}_n)| \leq \xi_n$, which then provides a natural generalization of Definition 1 to the a.c.i.d. case, in terms of total variation distance between $\mathbb{P}(X_{n+2} \in \cdot|\mathcal{G}_n)$ and $\mathbb{P}(X_{n+1} \in \cdot|\mathcal{G}_n)$.

Definition 5 (A.c.i.d., Law) *A sequence of $\mathcal{G}$-adapted r.v.s. $(X_n)_{n\geq 1}$ is $(\xi_n)_{n\geq 0}$-a.c.i.d. with respect to $\mathcal{G}$ if, for all $n \geq 0$, a.s.*

$$TV\left(\mathbb{P}(X_{n+2} \in \cdot|\mathcal{G}_n), \mathbb{P}(X_{n+1} \in \cdot|\mathcal{G}_n)\right) \leq \xi_n.$$

The same argument that led to Definition 5 justifies working with almost supermartingales: requiring the predictive distributions to form either a supermartingale or a submartingale cannot generalize the definition of c.i.d. r.v.s, because in that case they would be both, and thus a martingale. Definition 5 also highlights the difference with the condition of Leisen et al. (2026), where the upper bound can depend on a given Borel set.

As already mentioned in the Introduction, the main requirement for a sequence of predictive distributions to apply the posterior resampling approach of Fong et al. (2023) is *asymptotic exchangeability*. Theorem 6 provides sufficient conditions on $(\xi_n)_{n\geq 0}$ that ensure an a.c.i.d. sequence is asymptotically exchangeable. This is achieved by showing that any sequence of random measures forming an almost supermartingale converges weakly a.s., extending analogous results for martingales (e.g. see Aldous, 1985; Berti et al., 2004). Given a.s. weak convergence, asymptotic exchangeability follows from Lemma 8.2 of Aldous (1985). The proof is in Section B of the Appendix.

Theorem 6 *Let $(X_n)_{n\geq 1}$ be a $(\xi_n)_{n\geq 0}$-a.c.i.d. sequence and $(\alpha_n(\cdot,\cdot))_{n\geq 0}$ the corresponding sequence of 1-step ahead predictive distributions. If the sequence $(\xi_n)_{n\geq 0}$ satisfies*

$$\mathbb{E}\left(\sum_{n=0}^{\infty} \xi_n\right) < \infty \tag{4}$$

then there exists a random probability measure $\alpha : \Omega \times \mathcal{A} \to [0,1]$ such that

$$\mathbb{P}(\omega \in \Omega : \alpha_n(\omega,\cdot) \xrightarrow{w} \alpha(\omega,\cdot)) = 1.$$

Moreover, $(X_n)_{n\geq 1}$ is asymptotically exchangeable, that is, as $n \to \infty$,

$$(X_n, X_{n+1}, \ldots) \xrightarrow{d} (Z_1, Z_2, \ldots),$$

where $Z = (Z_1, Z_2, \ldots)$ is an exchangeable sequence directed by α.

In most applications, the sequence $(\xi_n)_{n\geq 0}$ is deterministic, in which case condition (4) reduces to the summability condition $\sum_{n=0}^{\infty} \xi_n < \infty$. Informally, Theorem 6 states that the distribution of

$(X_n)_{n\geq N}$ can be approximated for large N by that of an exchangeable sequence whose predictive distributions converge to the same random limit $\alpha(\cdot)$. From a Bayesian perspective, the random measure $\alpha(\cdot)$ represents the unknown statistical model, and its law corresponds to the prior distribution. Given a sample, whose information is summarized by the sigma-algebra $\mathcal{G}_n$, $\mathbb{P}(\alpha(\cdot) \in \cdot|\mathcal{G}_n)$ is the posterior distribution, and $\mathbb{E}[\alpha(\cdot)|\mathcal{G}_n]$ the predictive distribution. Theorem 6 does not provide a characterization of the asymptotic model and its properties, such as whether it is absolutely continuous. Such characterizations are available for martingale random measures through the tools developed in Berti et al. (2013), who also link the absolute continuity of the random limit to convergence in total variation, but remain unavailable for almost supermartingale random measures. We return to this point in the discussion.

If the sequence $(\alpha_n)_{n\geq 0}$ forms a martingale (Definition 2), standard properties of martingales imply $\alpha_n(\cdot) = \mathbb{E}[\alpha(\cdot)|\mathcal{G}_n]$. That is, the predictive distribution is an unbiased estimator of the limiting measure; in this context, the statistical model. Notably, this result depends solely on the martingale property and does not require exchangeability; see the notion of predictive coherence in Fong et al. (2023). However, under the a.c.i.d. condition (Definition 4), the sequence $(\alpha_n)_{n\geq 0}$ fails to be a martingale, and in general, $\alpha_n(\cdot) \neq \mathbb{E}[\alpha(\cdot)|\mathcal{G}_n]$. This naturally raises the question of whether this bias vanishes or can be reduced. The following corollary to Theorem 6 provides a partial answer.

Corollary 7 *Let $(X_n)_{n\geq 1}$ be an (a.c.i.d.) asymptotically exchangeable sequence with directing measure $\alpha(\cdot)$. Then, for $M \to \infty$.*

$$\mathbb{P}\left(X_{n+M} \in \cdot|\mathcal{G}_n\right) \xrightarrow{w} \mathbb{E}\left(\alpha(\cdot)|\mathcal{G}_n\right) \quad a.s.. \tag{5}$$

In words, Corollary 7 states that, if we are predicting future observations using the sequence of predictive distributions of an asymptotically exchangeable sequence, the M-step ahead predictive distribution will become closer (in a distance metrizing weak convergence) to corresponding predictive distribution obtained using the Bayesian posterior distribution as M increases.

4. Examples of a.c.i.d. Sequences

Throughout the section, $\mathcal{G}$ will be the natural filtration, $(E, \mathcal{E}) = (\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p))$, and $(\eta_n)_{n\geq 1}$ an appropriately chosen deterministic sequence such that $\eta_n \to 0$. When possible, the predictive distributions $(\alpha_n)_{n\geq 0}$ will be specified recursively. This is not a requirement, but is convenient for implementing predictive resampling. Examples in Sections 4.1 and 4.2 are considered by Holmes and Walker (2023) and related ones are studied mathematically by Garelli et al. (2026a) and Fong and Yiu (2026). Here, the goal is to illustrate the benefit of working with our new technical condition. Examples in Sections 4.3 and 4.4 are new, and will define usable new predictive resampling schemes.

4.1 Gaussian Sample Mean Driven Sequence

Let $\mathrm{N}(\mu, \Sigma)$ denote a p-variate Gaussian distribution with mean μ and covariance Σ, and $\phi(x; \mu, \Sigma)$ the corresponding density function. We can specify the generative process of a sequence $(X_n)_{n\geq 1}$ as follows: choose $\hat{\theta}_0 \in \mathbb{R}^p$ and Σ a valid variance-covariance matrix, and sample inductively $X_{n+1}|\mathcal{G}_n \sim \mathrm{N}(\hat{\theta}_n, \Sigma)$, where $\hat{\theta}_n$ is updated recursively as $\hat{\theta}_n = (1-\eta_n)\hat{\theta}_{n-1} + \eta_n X_n$. $(\mathrm{N}(\hat{\theta}_n, \Sigma))_{n\geq 0}$ is a sequence of parametric predictive random measures, where the randomness is driven by $\hat{\theta}_n$. When $\eta_n = \frac{1}{n}$, $\hat{\theta}_n$ corresponds to the sample mean (the MLE).

It is easy to show that $(X_n)_{n\geq 1}$ is in general not c.i.d.; instead, it is an a.c.i.d. sequence. Indeed, since $\alpha_n(A) = \int_A \phi(x; \hat{\theta}_n, \Sigma)\mathrm{d}x$, it holds that

$$\mathbb{E}\left(\alpha_{n+1}(A)|\mathcal{G}_n\right) = \int_{\mathbb{R}^p}\int_A \phi(x; \hat{\theta}_{n+1}, \Sigma)\mathrm{d}x\phi(x_{n+1}; \hat{\theta}_n, \Sigma)\mathrm{d}x_{n+1} = \int_A \phi(x; \hat{\theta}_n, (1+\eta_{n+1}^2)\Sigma)\mathrm{d}x,$$

where the second equality follows from the recursion formula for $\hat{\theta}_{n+1}$, and solving a convolution of Gaussian variables. To check Definition 5, we note that

$$\mathrm{TV}\left(\mathbb{P}(X_{n+2}\in\cdot|\mathcal{G}_n), \mathbb{P}(X_{n+1}\in\cdot|\mathcal{G}_n)\right) = \mathrm{TV}\left(\mathrm{N}(\hat{\theta}_n, (1+\eta_{n+1}^2)\Sigma), \mathrm{N}(\hat{\theta}_n, \Sigma)\right) \leq \frac{3}{2}p\eta_{n+1}^2,$$

where the inequality follows from Theorem 1.1 of Devroye et al. (2018) (see Appendix Section A.2). Then, $(X_n)_{n\geq 1}$ is a.c.i.d. with parameter $\xi_n = \frac{3}{2}p\eta_{n+1}^2$. The fact that $\mathbb{P}(X_{n+2}\in\cdot|\mathcal{G}_n) = \mathrm{N}(\hat{\theta}_n, (1+\eta_{n+1}^2)\Sigma)$ also implies that there is no η_{n+1} that makes the sequence c.i.d., except for the degenerate case where $\eta_{n+1} = 0$, which corresponds to no parameter update.

From Theorem 6, it suffices to assume that $\sum_{n=1}^{\infty}\eta_n^2 < \infty$ in order to prove that $(\mathrm{N}(\hat{\theta}_n, \Sigma))_{n\geq 0}$ converges weakly a.s. and that $(X_n)_{n\geq 1}$ is asymptotically exchangeable. The proof is straightforward, relying only on convolutions of Gaussian distributions and a standard bound on the total variation distance. It would be of interest to see if our argument applies to the more general sequence of Garelli et al. (2026a), who study sequences driven by the sample mean and the sample variance and prove almost sure convergence in total variation, or even their follow-up work (Garelli et al., 2026b), where they consider a mixture of such sequences.

4.2 Bayesian Parametric Bootstrap

Let P_θ denote a probability distribution parametrized by an unknown parameter $\theta\in\Theta$; $p(x;\theta)$ denote the corresponding density function (or probability mass function if the random variable is discrete), $p'(x;\theta)$ and $p''(x;\theta)$ the first and second derivatives with respect to θ, $s(x;\theta) := \frac{\partial}{\partial\theta}\log p(x;\theta)$ the score function, and $\mathcal{I}(\theta) = \mathbb{E}(s^2(x;\theta))$ the Fisher information. Let $\hat{\theta}_0$ be an initialization, we inductively sample for $n\geq 1$

$$X_n|\mathcal{G}_{n-1} \sim P_{\hat{\theta}_{n-1}}, \tag{6}$$

where $\hat{\theta}_n$ is updated recursively with Stochastic Gradient Descent (SGD)

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \eta_n Z(x_n, \hat{\theta}_{n-1}), \tag{7}$$

where $Z(x_n, \hat{\theta}_{n-1}) := \mathcal{I}(\hat{\theta}_{n-1})^{-1}s(x_n, \hat{\theta}_{n-1})$ is the natural gradient update, and η_n the step size. The use of the natural gradient follows the description of the parametric Bayesian bootstrap given by Fong and Yiu (2026).

To state the main result, we make the following assumptions,

Assumption 1 *Consider the data-generating process* (6)-(7) *and assume*

a) *p is twice differentiable with respect to θ.*

b) *$\mathcal{I}(\theta)$ is continuous and there exists ϵ such that $\epsilon < \mathcal{I}(\theta) < \infty$ for all $\theta\in\Theta$.*

c) *The sequence $(Z(X_n, \hat{\theta}_{n-1}))_{n\geq 0}$ is uniformly integrable.*

d) There exists $C < \infty$ such that $\int p''(x;\theta)\mathrm{d}x < C$ for all $\theta \in \Theta$.

Assumptions 1 a)–c) match those in Fong and Yiu (2026). a) and b) are standard. Assumption c) allows us to use Doob's martingale convergence theorem on $(\hat{\theta}_n)_{n\geq 1}$. This is a fairly strong assumption, which may be difficult to verify in practice. However, Fong and Yiu (2026) provide a sufficient condition that is easier to check in examples (see Appendix Proposition 19). We discuss alternatives below. Assumption 1 d) is used in a Taylor expansion of the density $p(x;\hat{\theta}_n)$, which we employ to bound $\mathrm{TV}(\mathbb{E}(P_{\hat{\theta}_n}|\mathcal{G}_{n-1}), P_{\hat{\theta}_{n-1}})$. It is reminiscent of classical conditions for asymptotic normality of maximum likelihood estimators, except that it is a condition on derivatives of the density and not of the log density as more commonly done (as in Van der Vaart (2000) Section 5.6). Nevertheless, we will show that it is verifiable for some parametric models.

Proposition 8 *Let $(X_n)_{n\geq 1}$ be sampled inductively as described in* (6)-(7)*. Under Assumption 1 and $\sum_{n=1}^{\infty}\eta_n^2 < \infty$, it holds that*

- *$(X_n)_{n\geq 1}$ is a.c.i.d., with $\xi_n := C\frac{\eta_{n+1}^2}{2\epsilon}$, where C and ϵ are from Assumption 1 d) and b), respectively;*
- *$(X_n)_{n\geq 1}$ is asymptotically exchangeable.*

The sequence $(P_{\hat{\theta}_n})_{n\geq 0}$ is not a martingale, which means that the resulting sequence $(X_n)_{n\geq 1}$ is not c.i.d. However, Proposition 8 shows that $(X_n)_{n\geq 1}$ is a.c.i.d. The result is stated for a univariate parameter but it can be generalized to the multivariate setting. This result complements the work of Fong and Yiu (2026) because they study asymptotic properties for the sequence $(\hat{\theta}_n)_{n\geq 0}$ while we characterize the law of the observables.

Example 1 *(Gaussian, unknown mean, known variance) Let $P_\theta = N(\theta, \sigma^2)$. We have that $\mathcal{I}(\theta) = \frac{1}{\sigma^2}$ and $s(x;\theta) = \frac{(x-\theta)}{\sigma^2}$, and so $Z(x,\theta) = (x-\theta)$. In this case, Assumptions 1 a)-c) trivially hold. The second derivative is*

$$p''(x;\theta) = \phi(x;\theta,\sigma^2)\left[\frac{(x-\theta)^2}{\sigma^2} - \frac{1}{\sigma^2}\right],$$

which implies that $\int p''(x;\theta)\mathrm{d}x = \frac{\sigma^2-1}{\sigma^2} < \infty$. This shows Assumption 1 d) holds.

Example 2 *(Student-t distribution) Let P_θ be a location-scale Student-t distribution with location θ and known scale τ and degrees of freedom ν.Setting $\tau = 1$, the update is given by Fong and Yiu (2026):*

$$Z(\theta, Y) = \frac{(\nu+3)(Y-\theta)}{\nu + (Y-\theta)^2},$$

who also prove that Assumptions 1 a)-c) hold. The case is interesting because both $Z(\theta,Y)$ and $\mathcal{I}(\theta)$ are unbounded. In Appendix Section B.2, we show that $\int p''(x;\theta)\mathrm{d}x < C$, for some finite constant C, which implies that Assumption 1 d) also holds.

Remark 9 *An alternative update for the parametric bootstrap is proposed in Holmes and Walker (2023). This version uses the standard gradient $Z'(x,\theta) = s(x,\theta)$, instead of the natural gradient. We remark that Proposition 8 continues to hold with this update, provided Assumption 1(c) is replaced by the assumption that the sequence $(\mathcal{I}(\hat{\theta}_n))_{n\geq 0}$ is uniformly integrable. An alternative assumption is also discussed below*

Assuming that $\Theta = [\underline{\theta}, \bar{\theta}]$, for arbitrary finite $\underline{\theta}, \bar{\theta}$, allows one to drop the uniform integrability assumption completely, both in the case of $Z(x, \theta)$ and $Z'(x, \theta)$. Define a constrained SGD sequence as follows

$$\tilde{\theta}_n = \begin{cases} \underline{\theta} & \text{if } \hat{\theta}_n < \underline{\theta} \\ \hat{\theta}_n & \text{if } \underline{\theta} \leq \hat{\theta}_n \leq \bar{\theta} \\ \bar{\theta} & \text{if } \hat{\theta}_n > \bar{\theta}, \end{cases} \tag{8}$$

$(\tilde{\theta}_n)_{n\geq 0}$ is not a martingale but it is still possible to prove that $(X_n)_{n\geq 1}$ is a.c.i.d. and drop Assumption 1 c).

Proposition 10 *Let $(X_n)_{n\geq 1}$ be sampled inductively as described in* (6)-(8)*. Under Assumption 1 a), b), and d), and $\sum_{n=1}^{\infty} \eta_n^2 < \infty$, the statement of Proposition 8 holds with $\xi_n := C\frac{\eta_{n+1}^2}{2\epsilon}$, where ϵ and C are from Assumption 1 b) and d).*

4.3 Kernel Estimators

Kernel density estimators are popular nonparametric methods to estimate smooth density functions. Let $K : \mathbb{R} \to \mathbb{R}$ be an integrable function satisfying $\int_{\mathbb{R}} K(u)\mathrm{d}u = 1$ called *kernel*, and $h = (h_n)_{n\geq 0}$ be a sequence of parameters called *bandwidth*. The pair (K, h) is generally chosen by the user, with K bounded, and h such that $h_n \to 0$ at a prespecified rate; see Silverman (2018) for an introduction, and Wolverton and Wagner (1969) for the sequential version of kernel estimators.

We use kernels to construct sequences of predictive distributions: choose an initial distribution α_0, and a pair (K, h), define recursively $\alpha_n(A)$

$$\alpha_n(A) = \frac{n-1}{n}\alpha_{n-1}(A) + \frac{1}{nh_n}\int_A K\left(\frac{x - X_n}{h_n}\right)\mathrm{d}x, \tag{9}$$

and $X_{n+1} \sim \alpha_n$. Since $\int \frac{1}{h_n}K(\frac{x}{h_n})\mathrm{d}x = 1$, we will denote the normalized kernel $K_n(x) := \frac{1}{h_n}K(\frac{x}{h_n})$ and let $\mu_n(A) := \int_A K_n(x)\mathrm{d}x$ be the corresponding measure. We remark that the predictive scheme (9) can also be seen as a specific case of the general class of recursive scheme proposed in Example 3.1 of Leisen et al. (2026).

The next two propositions show that the sequence of predictive distributions $(\alpha_n)_{n\geq 0}$ defined in (9) is never c.i.d., except for the degenerate case of no smoothing, but it can be a.c.i.d. under suitable conditions on the kernel and bandwidth.

More specifically, Proposition 11 establishes that the only instance in which the sequence defined by (9) is c.i.d. occurs when the kernel is degenerate and equal to the Dirac delta function. This scenario entails no smoothing and corresponds to adopting the empirical distribution function as the predictive distribution. Informally, it means that the martingale condition would rule out any sequential kernel estimator.

Proposition 11 *Let $(\alpha_n(\cdot))_{n\geq 0}$ be a sequence of the predictive distributions defined by* (9)*. Then, the corresponding probability law is c.i.d. if and only if $K(x) = \delta(x)$ (Dirac delta function) almost everywhere. This corresponds to the empirical measure $\alpha_n(\cdot) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \in \cdot)$.*

Proposition 11 corresponds to Theorem 1 of West (1991), who proved a similar result for the classical (non-recursive) kernel density estimator where, for a fixed sample size n, the bandwidths $h_{1:n}$ are all equal.

Proposition 12 shows that $(X_n)_{n\geq 1}$ can be a.c.i.d. and asymptotically exchangeable. Consequently, kernel estimators can be used for predictive inference if one works with our weaker technical condition. For $i \leq n$, define the kernel convolution

$$(K_{n+1} * K_i)(x; X_i) := \int \frac{1}{h_{n+1}} K\left(\frac{x - x_{n+1}}{h_{n+1}}\right) \frac{1}{h_i} K\left(\frac{x_{n+1} - X_i}{h_i}\right) \mathrm{d}x_{n+1}, \tag{10}$$

and let $\mu_{n+1*i}(\cdot; X_i)$ be the corresponding probability measure.

Proposition 12 *Let $(\alpha_n(\cdot))_{n\geq 0}$ be a sequence of the predictive distributions defined by* (9)*. Suppose that there exist $d > 0$ and $C > 0$ such that $K(\cdot)$ satisfies for all $n \geq i \geq 1$*

$$TV(\mu_{n+1*i}(\cdot, X_i), \mu_i(\cdot, X_i)) \leq C \frac{h_{n+1}^d}{h_i^d} \quad a.s.. \tag{11}$$

Then, the corresponding probability law is a.c.i.d. with parameter

$$\xi_n := \frac{1}{n+1} \sum_{i=1}^{n} \frac{1}{n} C \frac{h_{n+1}^d}{h_i^d}. \tag{12}$$

Moreover, a sufficient condition for $(X_n)_{n\geq 1}$ to be asymptotically exchangeable is that h satisfies

$$\frac{h_{n+1}^d}{n+1} \sum_{i=1}^{n} \frac{1}{h_i^d} = o\left(\log^{-1} n\right). \tag{13}$$

Proposition 12 provides an answer to a question raised by West (1991) regarding the existence of a statistical model underlying kernel-based predictives. The result suggests that, although kernel estimators may not correspond to a proper statistical model at finite sample sizes, they do in the limit, as they define a sequence that is asymptotically exchangeable under suitable conditions. This interpretation is analogous to results of Fortini and Petrone (2020) for the predictive recursion of Newton et al. (1998).

At present, we lack theoretical results for a.c.i.d. sequences that would allow us to determine whether the limiting distribution is absolutely continuous. Such a result is available for c.i.d. sequences in Berti et al. (2012) and has been used by Fortini and Petrone (2020) and Fong et al. (2023). However, that argument of Berti et al. (2012) relies on the martingale structure of the associated measure-valued process and does not apply to non-martingale settings. Moreover, it does not appear to be readily extendable to our framework. All we know is that $(P_n)_{n\geq 0}$ is absolutely continuous for all n. In the rest of the section, we discuss kernels and bandwidths that satisfy the conditions in Proposition 12.

Choice of the kernel. The condition (11) on K is not particularly restrictive. As examples, it is shown to hold for Gaussian, Laplace and Uniform kernels. Details on the calculations for the uniform kernel can be found in the Appendix.

Example 3 *(Gaussian kernel) Let $K(x) = \phi(x; 0, 1)$ for $x \in \mathbb{R}$, which implies $K_n(x) = \phi(x; X_n, h_n^2)$. The convolution $(K_{n+1} * K_i)(x; X_i) = \phi(x; X_i, h_i^2 + h_{n+1}^2)$, following standard calculation of convolution of Gaussians. Condition* (11) *holds: $TV(\mu_{n+1*i}(\cdot, X_i), \mu_i(\cdot, X_i)) \leq C \frac{h_{n+1}^2}{h_i^2}$, where the inequality follows from Theorem 1.3 of Devroye et al. (2018) (see Appendix Section A.2).*

Example 4 *(Laplace kernel) Let $K(x) = Laplace(x; \mu, b)$, a Laplace kernel with location μ and scale b, so that $K_n(x) = Laplace(x; X_n, h_n)$. Then, $(K_{n+1} * K_i)(x; X_i) = Laplace(x; X_i, h_{n+1} + h_i)$ and Proposition 18(a) in the Appendix shows that condition* (11) *holds with an upper bound on the total variation given by $C\frac{h_{n+1}}{h_i}$ for some $C > 0$.*

Choice of the bandwidth. The choice of the bandwidth is more critical. Condition (13) is satisfied by sequences h of the type $h_n = e^{-n}$. This is because $\sum_{i=1}^{n} \frac{1}{h_i^d} = e\frac{e^{\epsilon(n+1)}-1}{e^d-1}$, which implies that

$$\frac{h_{n+1}^d}{n+1}\sum_{i=1}^{n}\frac{1}{h_i^d} = \frac{1}{(n+1)e^{d(n+1)}}e\frac{e^{d(n+1)}-1}{e^d-1} = O\left(\frac{1}{n}\right).$$

This suggests that the resulting sequence $(\alpha_n)_{n\geq 0}$ needs to approach the empirical distribution rapidly, as n increases, in order for Proposition 12 to hold. This happens by virtue of the kernel "shrinking" as n increases due to the fast decay of the bandwidth.

A bandwidth that decays at a polynomial rate ($h_n = n^{-b}$ with $b > 0$) does not satisfy (13). This may at first seem very restrictive, since a standard condition for kernel density estimation consistency is $h_n n \to \infty$. However, note that (13) is a condition on the sequence of predictive distributions, not on the sequence of estimators. This example highlights a crucial aspect: we are prescribing a rule for the bandwidth in generating synthetic data for predictive resampling, not for the bandwidth of the estimator used in predictive resampling. The latter can be chosen according to standard results on the frequentist consistency of kernel estimators. This is in line with what done by Fong et al. (2023). In addition, we do not claim that such a decay is necessary for asymptotic exchangeability. It is sufficient for the observables to belong to the a.c.i.d. class.

A practical method to select the bandwidth is described in Appendix Section C.3.

Remark 13 *(Multivariate kernels) As is often the case in the kernel literature, results for the univariate setting generalize to the multivariate one. Specifically, let $K : \mathbb{R}^p \to \mathbb{R}$ be an integrable function satisfying $\int_{\mathbb{R}^p} K(u)\mathrm{d}u = 1$, and consider the case of the isotropic bandwidth $h = (h_n)_{n\in\mathbb{N}}$, meaning that each coordinate is scaled by the same h_n. In this setting, we write $K_n(x) = \frac{1}{h_n^p}K(\frac{x}{h_n})$ with $x \in \mathbb{R}^p$. Similarly, we will have that $\int_{\mathbb{R}^p} K_n(x)\mathrm{d}x = 1$. Proposition 12 will hold with identical conditions.*

4.4 Kernel Regression

Let $(Y_n, X_n)_{n\geq 1}$ be composed of a response variable $Y_n \in \mathbb{R}$, and a vector of covariates, $X_n \in \mathbb{R}^p$. Similar to the previous section, given an initial distribution α_0, kernels $K_y : \mathbb{R} \to \mathbb{R}$ and $K_x : \mathbb{R}^p \to \mathbb{R}$, and a sequence h, for $A \in \mathcal{B}(\mathbb{R}^{p+1})$, define $\alpha_n(A) := \mathbb{P}((Y_{n+1}, X_{n+1}) \in A|\mathcal{G}_n)$ recursively as follows

$$\alpha_n(A) = \frac{n-1}{n}\alpha_{n-1}(A) + \frac{1}{n}\int_A \frac{1}{h_n}K_y\left(\frac{y-Y_n}{h_n}\right)\frac{1}{h_n^p}K_x\left(\frac{x-X_n}{h_n}\right)\mathrm{d}x\mathrm{d}y, \tag{14}$$

The predictive distribution above implies that the one-step ahead conditional mean

$$\mathbb{E}(Y_{n+1}|x, \mathcal{G}_n) = \frac{\sum_{i=1}^{n} Y_i\, K_x(\frac{x-X_i}{h_i})}{\sum_{i=1}^{n} K_x(\frac{x-X_i}{h_i})} \tag{15}$$

corresponds to a recursive version of the Nadaraya-Watson kernel regression estimator; see Györfi et al. (2006) for more details and examples.

Let $K(y,x) = K_y(y)K_x(x)$. If $K(y,x)$ is such that the corresponding measure and kernel convolution measure satisfies (11), and h satisfies (13), $(Y_n, X_n)_{n\geq 1}$ is a.c.i.d. and asymptotically exchangeable by Proposition 12 (see Remark 13). Below are two examples where this holds.

Example 5 *(Gaussian isotropic kernel) Choose both K_y and K_x to be standard univariate Gaussian kernels. The Gaussian isotropic kernel is such that $\frac{1}{h_n^p}K_x\left(\frac{||x-X_n||_2}{h_n}\right) = \phi_p\left(x; X_n, h_n^2 I_p\right)$, where ϕ_p denotes the multivariate normal density and I_p a diagonal matrix of size $p \times p$. Similarly,*

$$K_n(y,x) = \frac{1}{h_n}K_y\left(\frac{y-Y_n}{h_n}\right)\frac{1}{h_n^p}K_x\left(\frac{||x-X_n||_2}{h_n}\right) = \phi_{p+1}(y,x;(y_n,x_n),h_n^2 I_{p+1}).$$

The kernel convolution is $(K_{n+1} * K_i)(y,x;(Y_i,X_i)) = \phi_{p+1}(y,x;(Y_i,X_i),(h_i^2+h_{n+1}^2)I_{p+1})$. The TV bound (11) *is $TV(\mu_{n+1*i}(\cdot,(Y_i,X_i)),\mu_i(\cdot,(Y_i,X_i))) \leq \frac{3}{2}(p+1)\frac{h_{n+1}^2}{h_i^2}$, by Theorem 1.1 of Devroye et al. (2018) (see Appendix Section A.2). Hence, if h satisfies* (13)*, $(Y_n, X_n)_{n\geq 1}$ is a.c.i.d.*

Example 6 *(Random Forests) Random Forests (RFs) are a popular family of learning algorithms for classification and regression. Most theoretical results for the RF exploit the connection between kernel methods and random forest; see Biau and Scornet (2016) for a review.*

Under several assumptions - namely, the regressors $X \in [0,1]^p$ and are uniformly distributed on the p-dimensional unit cube, regression trees have a fixed number of leaves l, the partition of each tree is such that at each stage the feature to be used in the split is chosen at random, and the split point is a draw from a uniform random variable - Breiman (2000) showed that, as the number of trees and the sample size go to infinity, the mean regression function of a RF $f_{RF}(x) \approx f_{RF}^{ker}(x)$, where $f_{RF}^{ker}(x)$

$$f_{RF}^{ker}(x) = \frac{\sum_{i=1}^n Y_i \exp(-\lambda \,||\, x - X_i \,||_1)}{\sum_{i=1}^n \exp(-\lambda \,||\, x - X_i \,||_1)}, \tag{16}$$

where $\lambda = \log l/p$. $f_{RF}(x)$ matches the regression function of a Nadaraya-Watson kernel regression estimator with a Laplace isotropic kernel. This approximation is referred to as kernel-based RF (Scornet, 2016), where the approximate nature is clear since f_{RF} with a finite number of trees is piecewise constant while (16) *is not.*

Let us select $K_y = Laplace(y;\mu_y,b)$ and $K_x(x) = \prod_{j=1}^p Laplace(x_j;\mu_{x_j},b)$, and $K_n(y,x) = Laplace(y;\mu_y,h_n)\prod_{j=1}^p Laplace(x_j;\mu_{x_j},h_n)$. Define a sequence of predictive distributions $(\alpha_n)_{n\geq 0}$ via the online scheme (14)*. The sequence $(Y_n,X_n)_{n\geq 1}$ sampled via $(\alpha_n)_{n\geq 0}$ has a mean regression function equal to*

$$\mathbb{E}(Y_{n+1}|x,\mathcal{G}_n) = \frac{\sum_{i=1}^n Y_i \exp(-h_i \,||\, x - X_i \,||_1)}{\sum_{i=1}^n \exp(-h_i \,||\, x - X_i \,||_1)},$$

a recursive version of (16) *with bandwidths indexed by i. This corresponds to an online version of the kernel-based RF. Following the same calculations as in Example 4, the kernel convolution is*

$$(K_{n+1} * K_i)(y,x;(Y_i,X_i)) = Laplace(y;\mu_y,h_n+h_i)\prod_{j=1}^p Laplace(x_j;\mu_{x_j},h_n+h_i),$$

and an upper bound for (11) *is $C(p+1)\frac{h_{n+1}}{h_i}$ (see Proposition 18 b)). Thus, the sequence $(Y_n,X_n)_{n\geq 1}$ is a.c.i.d.*

5. Numerical Illustrations

In this section, we illustrate predictive resampling using the a.c.i.d. predictive schemes presented in the previous section. We focus on the nonparametric schemes based on kernels since illustrations for the parametric examples can be found in Holmes and Walker (2023); Garelli et al. (2026a); Fong and Yiu (2026). The code for reproducibility is available online. The goal of the section is not to advocate in favor of one bootstrap scheme versus another, or predictive resampling versus standard ways of computing the posterior. Rather, we want to provide numerical evidence that the examples in Section 4 provide viable alternatives to existing proposals. As argued by Garelli et al. (2026a), optimal predictive schemes will vary based on the problem and data at hand. To our knowledge, there does not yet exist a theoretical framework in this recent literature to make such a comparison, and numerical studies are currently the only available means of comparison.

The martingale posterior can be defined for any parameter of interest. Garelli et al. (2026b) consider selected population functionals, such as mean, variance, kurtosis, etc. Here, we consider infinite-dimensional ones. As discussed in previous sections, we do not have theoretical guarantees on whether KerP's limiting model is absolutely continuous. For this reason, we consider two functionals θ that do not require such guarantees: in Subsection 5.1, a probability distribution; in Subsection 5.2, a regression function. Subsection 5.3 includes applications to real data. Kernel-based Predictive (KerP) schemes are implemented in `R`. We consider three types of kernels: Gaussian, Uniform, and Laplace. Methods to select the bandwidth h are described in Appendix Section C.3: they differ depending on whether we consider univariate or multivariate data, but all of them rely on standard selection rules implemented in `R`. More details on the implementation and pseudo-code can be found in Appendix Section C.3. In Appendix C.4, we replicate the simulation study to assess the sensitivity of the results for KerP to different bandwidth selection rules. Although bandwidth choice is important for any kernel-based procedure, we find that our results are largely insensitive to the alternatives considered. This robustness may reflect the fact that all of the bandwidth selection rules are based on widely used criteria. Nevertheless, the proposed rule is heuristic and warrants further investigation. In the remainder of the paper, we therefore report results based on one of the most standard configurations.

To estimate a probability distribution, we compare KerP to predictive resampling based on the Gaussian Copula (GauC) algorithm of Hahn et al. (2018) (implemented in Fong et al. (2023), code available at `https://github.com/edfong/MP`), Bayesian inference based on Gaussian Dirichlet Process Mixtures (DPM) (implemented in the `dirichletprocess` `R` package of Ross and Markwick (2018)), and the mixture of Gaussians driven by sample means and variances (GLPR) of Garelli et al. (2026b) (code made available by the authors). In regression, we compare KerP to regression function estimates of Gaussian Processes (GP) (implemented in the `laGP` `R` package of Gramacy (2016)), Gaussian DPMs, and GLPR. GPs directly target the mean regression function, while for DPMs and kernel-based methods, we estimate the joint density of (Y, X) and compute numerically $f(x) := \mathbb{E}(Y|x)$. For GLPR, we use the regression function defined in Garelli et al. (2026b). All predictive methods use the same number of synthetic samples and number of bootstrap replicates. Both GauC and GLPR require hyperparameter tuning: we rely on the built-in procedure for GauC and greedy search for GLPR. Further details on alternative methods are in Appendix Section C.2.

Given an estimand θ, let $\hat{\theta}$ denote the posterior mean, and $\hat{\theta}_{2.5}$ and $\hat{\theta}_{97.5}$ respectively the posterior .025 and .975 quantiles. Given a grid $\{x_i\}_{1:N}$, we use the following statistics to compare the methods: $env = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}(\hat{\theta}_{2.5}(x_i) \leq \theta(x_i) \leq \hat{\theta}_{97.5}(x_i))$, which measures if θ is covered by the 95% credible

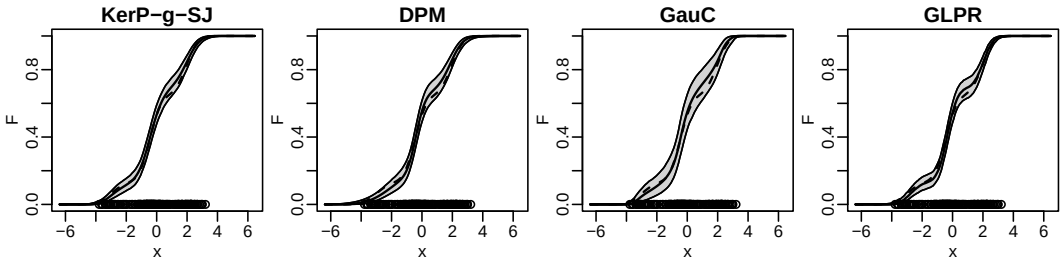


Figure 1: Simulations: probability distribution (one example). Estimates for one of the random P ($n = 200$). In each panel: true p (dashed line), posterior mean $\hat{p}$ (solid line), and 95% credible regions (gray shaded area). Dots depict the observations. The four panels differ by the method: KerP with Gaussian kernel and bandwidth via the method of Sheather and Jones (1991) (left), DPM (middle-left), GauC (middle-right), and GLPR (right).

region along the grid (a proxy for coverage), $dev = \frac{1}{N}\sum_{i=1}^{N}|\hat{\theta}(x_i) - \theta(x_i)|$, which measures the bias in L1 norm, and $awd = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_{97.5}(x_i) - \hat{\theta}_{2.5}(x_i))$, which measures the average width of the interval. $rdev$ and rwd will denote the corresponding metrics normalized by the true values. In the regression data sets, we will also compute the mean squared error $MSE = \frac{1}{n_{test}}\sum_{i=1}^{n_{test}}(y_i - \hat{f}(x_i))^2$. We will also report computing times as obtained on an Intel Xeon E5-2630[1]. The different methods are implemented in different programming languages - GauC in `Python`, DPM and KerP in `R`, and GLPR in `C++/Rcpp` - which complicates a direct comparison of their computational efficiency.

5.1 Probability Distributions

Let $Y_n|P \overset{iid}{\sim} P$, where P is a mixture model with a finite number of components and density p. P is random: for each new data set we sample the number of components, the weights of the mixture, the type of kernel (Gaussian or Student-t ($\nu = 10$)), and the location and scale parameters of each kernel; see Appendix Section C.1 for details. DPMs and GauC are specifically designed to estimate this class of mixture models, whereas KerP is not. DPMs will be considered as a benchmark, since inference is exact and does not rely on predictive resampling. We consider 100 simulated data sets sampled from 100 distinct distributions for two sample sizes ($n = 50, 200$). GauC requires a grid to numerically approximate a density (DPM and KerP do not). We fix it to $\{x_i\}_{1:N} := \{-4, \ldots, 3.95\} \times \hat{\sigma} + \bar{X}$, where $\bar{X}, \hat{\sigma}$ are respectively the sample mean and standard deviation, and $N = 100$. The same grid is also used to compute env, dev, and awd. GLPR requires the number of mixture components as a hyperparameter: we fix it to the maximum number of components of P in the simulation study. Figure 1 depicts estimates (solid lines), 95% credible regions (gray shaded area), true density (dashed line), and data (points) for one possible realization P and data for $n = 200$. In this example, all methods perform similarly.

Table 5.1 summarizes the median metrics across 100 repetitions. Performances are comparable across methodologies. As expected, DPM achieves the lowest bias (dev) and highest coverage (env) but can have the widest credible bands (awd) when $n = 200$. KerP based on the Gaussian and Uniform kernel achieves similar performances in terms of dev and awd, even if the coverage is lower. It

1. Available in the computer cluster Correfoc at the Universitat Pompeu Fabra.

	n = 50				n = 200			
Method	*env*	*dev*	*awd*	*time*	*env*	*dev*	*awd*	*time*
GauC	0.500	0.022	0.053	–	0.541	0.013	0.040	–
DPM	0.994	0.018	0.090	10.685	0.931	0.009	0.046	73.647
GLPR	0.572	0.022	0.087	0.338	0.509	0.011	0.043	0.366
KerP-g-cv	0.831	0.020	0.092	10.916	0.769	0.011	0.044	11.612
KerP-l-cv	0.744	0.026	0.095	8.842	0.719	0.014	0.045	9.396
KerP-u-cv	0.700	0.019	0.091	7.608	0.694	0.009	0.044	8.084

Table 1: Simulations: probability distribution. Median across 100 data sets for different sample sizes (n). *env* is a measure of coverage (the higher the better), *dev* is an L1 bias (the lower the better), and *awd* is the average width of the credible bands (the lower the better). In KerP, g stands for Gaussian kernel, l for Laplace, u for uniform, and SJ for the bandwidth selection of Sheather and Jones (1991).

remains superior to the other predictive resampling methods across all accuracy metrics. Comparable numerical performance is achieved about seven times faster than MCMC-based methods. GauC's time is not reported because some runs' times were unreliable (due to our implementation on the cluster). Nevertheless, GLPR is the fastest method. It is unclear why it should be faster than kernel-based methods. A possible explanation is that it is implemented in a more efficient programming language (C++ as opposed to the others that are on R). The resampling methods can be significantly accelerated when run on a computing cluster in parallel, see Fong et al. (2023).

5.2 Regression Function

Let $Y_n = f(X_n) + \epsilon_n$ with $\epsilon_n \overset{iid}{\sim} P$ and $X_n \in \mathbb{R}$. For each regression function f, we sample $n = 200$ training data with $X_n \overset{iid}{\sim} \mathcal{U}(0, 5)$, while test points are defined on a regular grid $\{x_i\}_{i:N} := \{-4, \ldots, 3.95\}$ with $N = 100$ evenly spaced values. To simulate data, we consider three Data-Generating Mechanisms (DGM) where we vary how we sample f and the noise distribution P. DGMs 1 and 2 have $f \sim GP$ with mean 0 and a Gaussian kernel. In DGM1, P is Gaussian zero mean and variance one, while in DGM2, P is Student-t zero mean and $\nu = 5$. DGM3 has $f \sim GP$ with mean 0 and a kernel obtained by summing a Gaussian and an Exponential one, P is Gaussian zero mean and variance one. Since all DGMs are GPs, GP performance can be interpreted as an oracle, because the model is correctly specified. More details and the exact parametrization are in Appendix Section C.1. For each combination, we simulate 100 data sets. We compute env, dev, awd, and MSE on the test data for KerPs, GPs, DPMs, and GLPR (the hyperparameter "number of clusters" is optimized on a grid $(3, 5, 6, 7, 9)$, we report the best-performing version of the algorithm). Figure 2 depicts estimates (solid lines), 95% credible regions (gray shaded area), true regression function (dashed line), and data (points) for one possible realization of DGM1. In this example, KerP (Laplace kernel and CV bandwidth) GLPR and GP recover f, with GP and GLPR having smoother bands. This largely follows from the choice of the kernel. DPM oversmooths and fails to capture some local features of f, and has credible bands comparable to KerP. This provides further evidence that predictive resampling offers an alternative to computationally intensive MCMC-based methods.

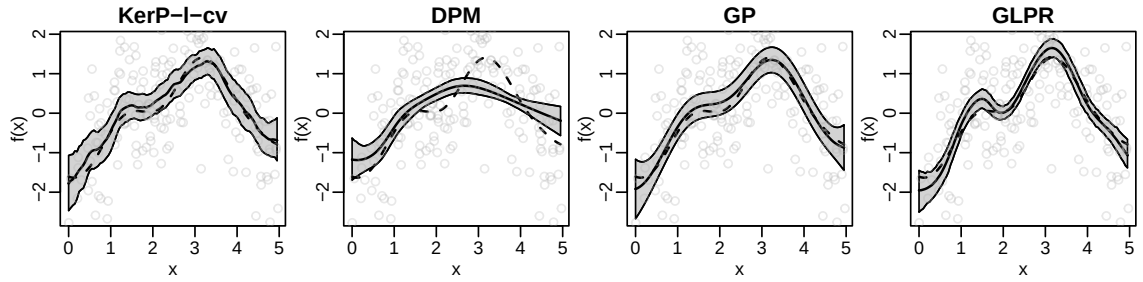


Figure 2: Simulations: regression function (one example). Estimates for one of the random f ($n = 200$, DMG1). In each panel: true p (dashed line), posterior mean $\hat{p}$ (solid line), and 95% credible regions (gray shaded area). Dots depict the observations. Four panels differ by the method: KerP with Laplace kernel and bandwidth via cross-validation (left), DPM (middle-left), GP (middle-right), and GLPR (right).

Table 5.2 summarizes the median of the 100 repetitions of the statistics considered for the three DGMs. GPs achieve the best (oracle) performance across the considered metrics. Surprisingly, KerP with Laplace and Uniform kernels is comparable to GPs' performance - which is a baseline oracle performance. Interestingly, KerP with Laplace kernel achieves the lowest test MSE, even in DGM1 when GP is the true DGM. Recall that this algorithm is linked to an infinite limit of the RF, a connection that may explain the good performance. The good performance of KerP with a uniform kernel may be explained by the uniform sampling of the covariates. KerP with the Gaussian kernel and Gaussian DPMs oversmooth the regression function. GLPR has bias comparable to that of KerP with Gaussian kernel; however, its credible bands have lower coverage, and are substantially narrower. In terms of computational times (for DGM3), GP is by far the fastest method. While GLPR remains faster than KerP, the difference becomes less substantial compared to Table 5.1. Moreover, if hyperparameter optimization was included in the computational time, there would be practically no differences between the two methods, despite GLPR's more efficient implementation in C++. However, as mentioned above, a rigorous comparison of computational efficiency requires additional work.

5.3 Data Analysis

In this subsection, although there are no guarantees that the limiting model associated with the sequence induced by KerP is absolutely continuous, we plot density functions instead of cumulative distribution functions for ease of visualization. Recall that, due to the finite truncation involved in predictive resampling, the predictive distribution P_{n+M} defined by KerP is absolutely continuous for finite M. Determining whether the full sequence converges to an absolutely continuous model is left for future work.

Galaxy data. This data set is a standard benchmark for Bayesian nonparametric estimation and clustering (Roeder (1990), publicly available in R). It contains $n = 82$ galaxies from six conic sections of an unfilled survey of the Corona Borealis region. Figure 3 includes replicates of the analysis of Fong et al. (2023) for GauC and DPM (middle and right panels) and the predictive resampling posterior obtained with KerP with Gaussian kernel and bandwidth sequence selected via the method

	DGM1: Gaussian ϵ				DGM2: Student-t ϵ				DGM3: Gauss-Exp Kernel				
Method	*env*	*dev*	*awd*	*MSE*	*env*	*dev*	*awd*	*MSE*	*env*	*dev*	*awd*	*MSE*	*time*
DPM	0.34	0.34	0.38	1.15	0.36	0.31	0.37	1.13	0.43	0.33	0.47	1.71	1404.52
GLPR	0.67	0.21	0.51	1.06	0.70	0.19	0.51	1.07	0.72	0.22	0.60	1.62	91.74
GP	0.94	0.15	0.73	1.05	0.94	0.16	0.72	1.04	0.94	0.19	0.89	1.58	0.06
KerP-g-cv	0.78	0.21	0.68	1.06	0.79	0.20	0.66	1.06	0.84	0.24	0.78	1.62	365.86
KerP-l-cv	0.90	0.17	0.75	1.04	0.90	0.17	0.73	1.04	0.92	0.20	0.89	1.58	235.68
KerP-u-cv	0.94	0.17	0.79	1.04	0.94	0.16	0.76	1.03	0.92	0.21	0.93	1.60	228.71

Table 2: Simulations: regression function. Median across 100 data sets for three different data generating mechanisms (n_{train} = 200,n_{test} = 100). *env* is a measure of coverage (the higher the better), *dev* is an L1 bias (the lower the better), and awd is the average width of the credible bands (the lower the better), MSE is the sample mean squared error on test data (the lower the better). In KerP, g stands for Gaussian kernel, l for Laplace, u for uniform, cv for the bandwidth selection with smoothed cross-validation.

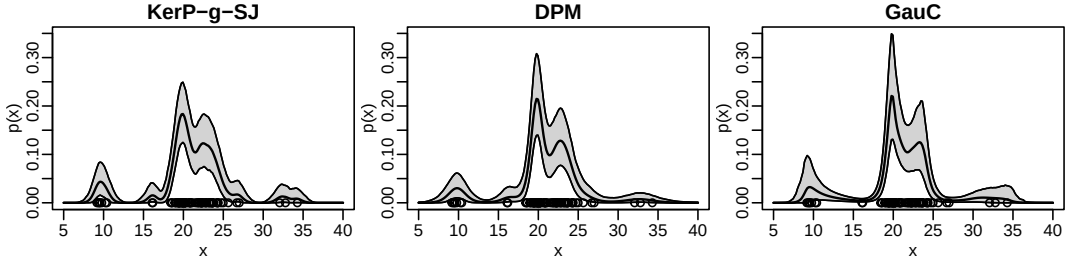


Figure 3: Galaxy data. In each panel: posterior mean $\hat{p}$ (solid line), and 95% credible regions (gray shaded area). Dots depict the observations. Three panels differ by the method: KerP with Gaussian kernel and bandwidth via Sheather and Jones (1991) (left), DPM (middle), and GauC (right).

of Sheather and Jones (1991) (B = 2000, M = 3000). KerP estimates are more closely aligned with DPM than GauC, with a mode around 17 km/sec not captured by GauC.

Air quality data. In this example, we estimate a bivariate density of the radiation and ozone variables from the `airquality` data set in `R`. They are daily measurements (from 01-05-1973 to 30-09-1973) collected in New York of solar radiation (in Langleys) and mean ozone (in parts per billion). There are 42 missing observations which we remove, resulting in n = 111. Figure 4 depicts contour plots of the mean density estimates obtained via predictive resampling of GauC (same parametrization as in Fong et al. (2023)) and KerP with Gaussian kernel with bandwidth selected via the multivariate plug-in method of Wand and Jones (1994) (M = 2000, B = 500). This results in two distinct bandwidth sequences for the two dimensions, similar to what is done for GauC where each dimension has its correlation parameter. Contour plots of the posterior standard deviation for the two methods are in the Appendix (Figure 6). Compared to GauC, KerP's estimate is more diffuse and the mode in the upper corner of the panel appears to be more skewed. KerP's predictive resampling posterior has wider credible regions around the two modes (Figure 6).

LIDAR data. It consists of 221 observations from a light detection and ranging (LIDAR) experiment. It includes a regressor, the distance traveled before the light is reflected back to its source (x, range),

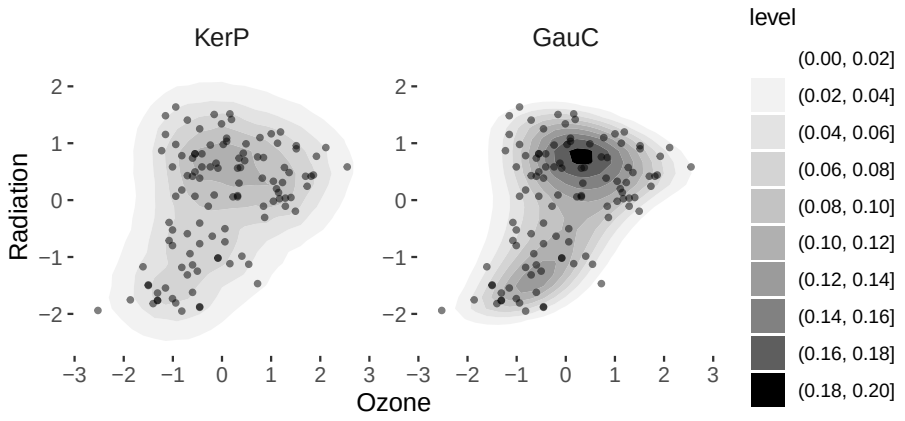


Figure 4: Air quality data: posterior mean density. In each panel: contour plot of the posterior mean. Dots depict the observations. The two panels differ by the method: KerP with Gaussian kernel and bandwidth via the multivariate plug-in method of Wand and Jones (1994) (left), and GauC with parameters as in Fong et al. (2023) (right).

and a dependent variable, the logarithm of the ratio of received light from two laser sources (y, log intensity). We estimate the regression function with the full data sets via KerP and DPM (Figure 5) and the conditional densities at $x = 0$ and $x = -3$ (Appendix Figures 7 and 8). In addition, using holdout data sets (10 random splits, 80% training, 20% test), we do a prediction exercise using the regression function estimated on the training data. For KerP, we used a Gaussian kernel with bandwidth selected via the multivariate plug-in method of Wand and Jones (1994) ($M = 1000$, $B = 600$).

In Figure 5, KerP regression function appears less smooth than DPM and with wider credible regions. The two estimates differ slightly for positive values of range, with DPM having a linear behavior, while the KerP estimate has a slightly positive curvature. KerP MSE is 0.1003 and DPM MSE is 0.1085, approximately 8% higher. Appendix Figure 7 depicts a conditional density at a value $x = 0$, which is in the middle of the data set. The DPM estimate appears remarkably smoother and with much narrower credible bands than KerP. As in Fong et al. (2023), we estimated the conditional density at $x = -3$, which is well outside the interval of observed values (Appendix Figure 8). KerP does not perform well, with wide credible bands and bumpy estimates. This illustrates a known limitation of kernel methods: with low bandwidth and no nearby observations, there is no borrowing of information. Additionally, the Gaussian kernel may be ill-suited in this scenario, because the regressor (range) forms an almost regular grid.

6. Discussion

We defined a new class of predictive schemes suitable for bootstrap methods like predictive resampling. It was known that, if one only requires the sequence of random variables to be asymptotically exchangeable, then the predictive rule does not need to be a martingale. The main contribution of this paper is the definition of a new property - almost conditionally identically distributed variables - which provides a practical criterion for determining whether an algorithm can be used for predictive

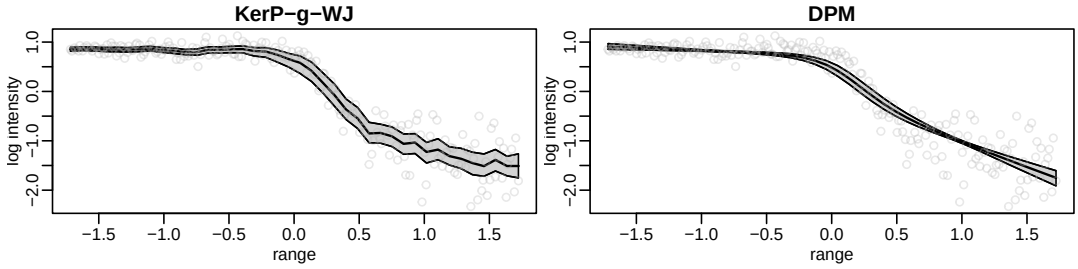


Figure 5: LIDAR: regression function. In each panel: posterior mean $\hat{f}$ (solid line), and 95% credible regions (gray shaded area). Dots depict the observations. The two panels differ by the method: KerP with Gaussian kernel and bandwidth via the multivariate plug-in method of Wand and Jones (1994) (left), and DPM (right).

resampling. We demonstrated that existing and new predictive rules satisfy the a.c.i.d. property: notably, kernel estimators, including an approximation to the random forests, and the parametric Bayesian bootstrap.

The departure from the martingale property may at first appear non-innocuous: the predictive distribution is no longer an unbiased estimator of the limiting measure, a property often associated with coherence in the Bayesian literature (see also predictive coherence, Fong et al. (2023)). However, modern statistics provides numerous examples in which introducing bias into an estimator need not be detrimental. Our numerical results provide a further example in the context of Bayesian predictive inference: finite-sample accuracy is not compromised and, in some cases, is even improved.

To provide a possible explanation for this observation, recall that predictive distributions associated with exchangeable sequences are characterized by two properties (Fortini et al., 2000): (i) they are measure-valued martingales, and (ii) they are symmetric functions of past observations. C.i.d. r.v.s generalize exchangeability to asymptotic exchangeability by relaxing the second property while preserving the first, thereby retaining the no-bias property of the predictive distribution. A.c.i.d. r.v.s, in contrast, relax both properties. Interestingly, some of the a.c.i.d. schemes considered here, despite allowing for bias in the predictive distribution, appear to be less sensitive to the ordering of the observations than their c.i.d. (martingale) counterparts. In some respects, it might be possible that they could be "closer" to exchangeability than the corresponding c.i.d. schemes. This raises the question of whether reduced sensitivity to the ordering of past observations may itself be a desirable feature and, if so, whether a suitable notion of "closeness to exchangeability" can be formalized, one that accounts simultaneously and balances for deviations from both predictive properties characterizing exchangeability.

To establish the a.c.i.d. property, we relied on total variation bounds. This approach can be substantially more tractable than verifying almost sure weak convergence of a predictive sequence, although it is not always feasible or straightforward. Developing conditions that imply the a.c.i.d. property while working with a smaller class of sets would be valuable. One way would be to use algebras that generate the Borel sigma-algebra and monotone class-type arguments. An alternative route may involve bounding other metrics (for example, the Wasserstein distance) that enable control over the total variation distance. For instance, Chae and Walker (2020) demonstrate that, under

sufficient smoothness of the underlying densities, the total variation distance can be bounded by a power of the Wasserstein distance.

More technical tools for almost-supermartingale random measures are necessary. In the martingale case, there are central limit theorems (Berti et al., 2004) and sufficient conditions for the limiting model to be absolutely continuous (Berti et al., 2013). These results are of interest to statisticians: for example, Fortini and Petrone (2020) use CLTs to build asymptotic credible intervals; Fortini and Petrone (2020) also characterize the limit model of Newton's algorithm, and Fong et al. (2023) that of the Gaussian copula. Similar results would enable analogous developments in this context for specific procedures, such as kernel estimators.

From a methodological standpoint, it would be of interest to examine the a.c.i.d. property for other algorithms in machine learning and statistical learning. We did not investigate wavelet-based or orthonormal basis estimators, and our analysis of random forests was based on a simplified approximation: there are alternatives to this such as the proposals in Scornet (2016). Among the methods we considered, kernel-based predictive resampling stands out as especially promising. Compared to the Bayesian bootstrap, kernel methods are better suited for estimating continuous parameters; compared to copula-based approaches, they do not require a discretization grid, and benefit from an extensive literature and widespread use. However, it remains crucial to develop principled procedures for selecting key parameters, such as the bandwidth, when using kernel methods in this setting.

Acknowledgments

L.C. acknowledges the support of Ayudas Fundacion BBVA a Proyectos de Investigacion Cientifica 2021, the Spanish Ministry of Economy and Competitiveness grant PID2022-138268NB-I00, financed by MCIN/AEI/0.13039/501100011033, FSE+MTM2015-67304-P, and FEDER, EU; the grant Ramon y Cajal 2022 RYC2022-038467-I, financed by MCIN/AEI/10.13039/501100011033 and FSE+, and the Severo Ochoa Programme for Centres of Excellence in R&D (Barcelona School of Economics CEX2019-000915-S), funded by MCIN/AEI/10.13039/50110001103. M.B. gratefully acknowledges support for this work from the London Mathematical Society, Research in Pairs grant [grant number 42418]. We also thank the staff at the Scientific computing core facility at Universitat Pompeu Fabra for their help. L.C. and M.B. thank the Associate Editor and three Reviewers for helpful suggestions that significantly improved the paper.

Appendix A. Background and Auxiliary results

A.1 Almost Supermartingales

For ease of reading, we recall some results about almost supermartingale from Robbins and Siegmund (1971). We present these results in a slightly simplified formulation, which is still sufficient for the proof of Theorem 6. Let $(Z_n)_{n\geq 0}$ and $(\xi_n)_{n\geq 0}$ be two sequences of random variables adapted to a filtration $(\mathcal{G}_n)_{n\geq 0}$. Also, let ξ_n be non-negative for all $n \geq 0$. Then, $(Z_n)_{n\geq 0}$ is a $(\xi_n)_{n\geq 0}$-*almost supermartingale* if for all $n \geq 0$,

$$\mathbb{E}(Z_{n+1}|\mathcal{G}_n) \leq Z_n + \xi_n \quad \text{a.s..}$$

Robbins and Siegmund (1971) have shown that almost supermartingales satisfy counterparts of the Doob convergence theorem and maximal inequality, valid for supermartingales. Specifically,

Theorem 14 (Theorem 1, Robbins and Siegmund (1971)) *If $(Z_n)_{n\geq 0}$ is a $(\xi_n)_{n\geq 0}$-almost supermartingale, then $\lim_{n\to\infty} Z_n$ exists and is finite on the set $\{\omega \in \Omega : \sum_{n=0}^{\infty} \xi(\omega) < \infty\}$.*

Proposition 15 (Proposition 2, Robbins and Siegmund (1971)) *If $(Z_n)_{n\geq 0}$ is a $(\xi_n)_{n\geq 0}$-almost supermartingale, then for any $a > 0$, $m \in \mathbb{N}$, and $n \geq m$*

$$\mathbb{P}\left(\max_{m\leq k\leq n} Z_k \geq a|\mathcal{G}_m\right) \leq \frac{1}{a}\left[Z_m + \mathbb{E}\left(\sum_{k=m}^{n-1} \xi_k|\mathcal{G}_m\right)\right].$$

A.2 Bounds on the Total Variation Distance

TV bounds between Gaussians. The following results are given in Devroye et al. (2018). Theorem 16 provides an upper and lower bounds between univariate Gaussians with arbitrary mean and variance. Theorem 17 is the corresponding multivariate result under the assumption of equal mean vectors. Alternative bounds in total variation between high-dimensional Gaussian distributions were derived in Barabesi and Pratelli (2024).

Theorem 16 *(Theorem 1.3, Devroye et al. (2018)) Let $\mu_1, \mu_2 \in \mathbb{R}$ and $\sigma_1^2, \sigma_2^2 \in \mathbb{R}^+$, the following upper and lower bounds hold*

$$\frac{1}{200}\min\left\{1, \max\left\{\frac{|\sigma_1^2 - \sigma_2^2|}{\sigma_1^2}, \frac{40|\mu_1 - \mu_2|}{\sigma_1}\right\}\right\} \leq$$
$$\leq TV(N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2) \leq \frac{3|\sigma_1^2 - \sigma_2^2|}{2\sigma_1^2} + \frac{|\mu_1 - \mu_2|}{2\sigma_1}.$$

Theorem 17 *(Theorem 1.1, Devroye et al. (2018)) Let $\mu \in \mathbb{R}^p$, Σ_1 and Σ_2 be positive definite matrices, and $\lambda_1, \ldots, \lambda_p$ denote the eigenvalues of $\Sigma_1^{-1}\Sigma_2 - I_d$*

$$\frac{1}{100} \leq \frac{TV(\Phi(\cdot;\mu,\Sigma_1), \Phi(\cdot;\mu,\Sigma_2)}{\min\left\{1, \sqrt{\sum_{i=1}^p \lambda_i^2}\right\}} \leq \frac{3}{2}.$$

TV bounds between Laplace. We include a proof for the following result as we did not find a reference for it. In the main manuscript, we denoted with Laplace$(x;\mu,c)$, a Laplace kernel with location μ and scale c. In this section, we use Laplace(μ,c) (or Laplace$(\cdot;\mu,c)$)to refer to the corresponding probability measure.

Proposition 18 *(a) Let $X \sim Laplace(\mu, c)$ and $Y \sim Laplace(\mu, c + a_n)$ for $c \geq a_n > 0$ and $a_n \to 0$. For some $C > 0$, the following holds*

$$TV(Laplace(\mu,c), Laplace(\mu, c + a_n)) = \left(\frac{c}{c+a_n}\right)^{\frac{c}{a_n}} - \left(\frac{c}{c+a_n}\right)^{\frac{c+a_n}{a_n}} \leq C\frac{a_n}{c}.$$

(b) Let $X = (X_1, \dots, X_p) \sim P_X$ *be such that* $X_i \overset{iid}{\sim} Laplace(\mu, c)$ *for* $i = 1, \dots, p$ *and* $Y = (Y_1, \dots, Y_p) \sim P_Y$ *be such that* $Y_i \overset{iid}{\sim} Laplace(\mu, c + a_n)$ *for* $i = 1, \dots, p$*, then*

$$TV(P_X, P_Y) \leq Cp\frac{a_n}{c}.$$

Proof For part *(a)* we have that

$$\begin{aligned}
&\mathrm{TV}(\mathrm{Laplace}(\mu, c), \mathrm{Laplace}(\mu, c + a_n)) = \\
&= \int \frac{1}{2} \left| \frac{1}{2(c + a_n)} \exp\left(-\frac{|x|}{c + a_n}\right) - \frac{1}{2c} \exp\left(-\frac{|x|}{c}\right) \right| \mathrm{d}x \\
&= \int_0^\infty \frac{1}{2} \left| \frac{1}{c + a_n} \exp\left(-\frac{x}{c + a_n}\right) - \frac{1}{c} \exp\left(-\frac{x}{c}\right) \right| \mathrm{d}x \\
&= \int_{x^*}^\infty \frac{1}{c + a_n} \exp\left(-\frac{x}{c + a_n}\right) - \frac{1}{c} \exp\left(-\frac{x}{c}\right) \mathrm{d}x \\
&= \exp\left(-\frac{x^*}{c + a_n}\right) - \exp\left(-\frac{x^*}{c}\right),
\end{aligned}$$

where $x^* = \log \frac{c+a_n}{c}(c + a_n)c/a_n$ is the point where the two kernels intersect. This proves the equality

$$\mathrm{TV}(\mathrm{Laplace}(\mu, c), \mathrm{Laplace}(\mu, c + a_n)) = \left(\frac{c}{c + a_n}\right)^{\frac{c}{a_n}} - \left(\frac{c}{c + a_n}\right)^{\frac{c+a_n}{a_n}}.$$

To see the inequality, rewrite the TV as

$$\left(\frac{c}{c + a_n}\right)^{\frac{c}{a_n}} - \left(\frac{c}{c + a_n}\right)^{\frac{c+a_n}{a_n}} = \left(\frac{c}{c + a_n}\right)^{\frac{c}{a_n}} \left[1 - \frac{c}{c + a_n}\right] \leq \left(\frac{c}{c + a_n}\right)^{\frac{c}{a_n}} \frac{a_n}{c},$$

and note that the first term is positive, bounded, and strictly decreasing.

For part *(b)*, consider for simplicity the case $p = 2$. The proof for general p will follow from an identical argument.

$$\begin{aligned}
\mathrm{TV}(P_X, P_Y) &= \int \int \frac{1}{2} \left| p_{x_1}(x_1) p_{x_2}(x_2) - p_{y_1}(x_1) p_{y_2}(x_2) \right| \mathrm{d}x_1 \mathrm{d}x_2 \\
&= \int \int \frac{1}{2} \left| (p_{x_1}(x_1) - p_{y_1}(x_1)) p_{x_2}(x_2) + (p_{x_2}(x_2) - p_{y_2}(x_2)) p_{y_1}(x_1) \right| \mathrm{d}x_1 \mathrm{d}x_2 \\
&\leq \int \int \frac{1}{2} \left| (p_{x_1}(x_1) - p_{y_1}(x_1)) \right| \mathrm{d}x_1 p_{x_2}(x_2) \mathrm{d}x_2 + \\
&\qquad + \int \frac{1}{2} \left| (p_{x_2}(x_2) - p_{y_2}(x_2)) \right| \int p_{y_1}(x_1) \mathrm{d}x_1 \mathrm{d}x_2 \\
&\leq 2C\frac{a_n}{c},
\end{aligned}$$

where the first inequality is the triangle inequality, the second one is part *(a)*. ■

A.3 Auxiliary Results on Parametric Bootstrap

We recall Proposition 1 of Fong and Yiu (2026). This proposition provides a condition that implies Assumption 1 c), from Section 4.2 in the main manuscript.

Proposition 19 *(Fong and Yiu (2026)) For $X \sim P_\theta$, suppose there exists non-negative constants $B, C < \infty$ such that the following holds for all $\theta \in \Theta \subseteq \mathbb{R}$,*

$$\mathbb{E}(Z(X,\theta)^4) \leq B + C\theta^4,$$

then Assumption 1 c) is satisfied.

Appendix B. Proofs

B.1 Proof of Section 3

Proof [Proof of Theorem 6] The state space E is assumed to be a Polish space, that is a separable topological space, metrizable by a complete metric, and $\mathcal{E}$ the corresponding Borel σ-algebra. Therefore, there exists a countable subclass $\mathcal{A} \subseteq \mathcal{E}$ that generates $\mathcal{E}$, that is, $\sigma(\mathcal{A}) = \mathcal{E}$, and is probability determining, which means that if two probability measures agree on $\mathcal{A}$, then they must also agree on $\mathcal{E}$. Indeed, let us consider the class of finite intersections of open balls with rational radius (E is completely metrizable) centered at the points of a countable dense subset (E is separable). We denote this class by $\mathcal{A}$, which is countable. Because E is separable, every open set is a countable union of open balls, hence also of elements of $\mathcal{A}$. Therefore, being $\mathcal{E}$ the smallest σ-algebra generated by the open sets, it is also generated by $\mathcal{A}$. Finally, the class $\mathcal{A}$ is closed under intersections by construction, hence it is a probability determining class from Dynkyn's Lemma.

For any $A \in \mathcal{A}$, $(\alpha_n(\cdot, A))_{n\geq 0} = (\mathbb{P}(X_n \in A|\mathcal{G}_{n-1})))_{n\geq 0}$ forms an almost $\mathcal{G}$-supermartingale. Therefore, it converges almost surely by Theorem 14 and the assumption that $\sum_{n=0}^{\infty} \xi_n < \infty$ a.s. Being the class $\mathcal{A}$ countable, the exceptional null sets can be grouped into a unique null set, denoted N_1. Therefore, the sequence of probability measures $(\alpha_n(\omega, \cdot))_{n\geq 0}$ converges over all elements $A \in \mathcal{A}$ of a determining class, for all $\omega \in N_1^c$. If it can also be shown that that the sequence $(\alpha_n(\cdot, \cdot))_{n\geq 0}$ is a.s. tight, which means that there exist a null set $N_2 \in \mathcal{E}$ such that for all $\omega \in N_2^c$ the sequence $(\alpha_n(\omega, \cdot))_{n\geq 0}$ is tight. Then, for all $\omega \in (N_1 \cup N_2)^c$, the sequence of probability measures $(\alpha_n(\omega, \cdot))_{n\geq 0}$ converges for all A in a probability determining class and is tight, hence converges weakly to a probability measure $\alpha(\omega, \cdot)$, from Theorem 5.1 and subsequent Corollary of Billingsley (1999), hence the result of Theorem 6 follows.

Therefore, it remains to show that $(\alpha_n(\cdot, \cdot))_{n\geq 0}$ is a.s. tight. Fix $\epsilon > 0$ arbitrary, and, for every $j \in \mathbb{N}$, set $\delta_j := \epsilon 2^{-2j}$. Moreover, for every $m \in \mathbb{N}$, let $\bar{\alpha}_m$ denote the mean measure of α_m, defined as

$$\bar{\alpha}_m(A) := \int_\Omega \alpha_m(\omega, A)\mathbb{P}(d\omega).$$

Because E is a Polish space, $\bar{\alpha}_m$ is a tight probability measure for every $m \in \mathbb{N}$, from Theorem 1.3 of Billingsley (1999). Therefore, for every $m \in \mathbb{N}$, there exists a sequence of compact subsets $(K_{m,j})_{j\in\mathbb{N}}$ of E, s.t. $\bar{\alpha}_m(K_{m,j}^c) \leq \delta_j = \epsilon 2^{-2j}$, for all m and $j \in \mathbb{N}$. For each $m \in \mathbb{N}$, by denoting with $\mathcal{P}(\mathcal{E})$ the set of probability measures on $(E, \mathcal{E})$, we can define the sequence of sets

$$\Theta_m := \{\theta \in \mathcal{P}(\mathcal{E}) : \theta(K_{m,j}^c) \leq 2^{-j}, \forall j \geq 1\}.$$

For each m, Θ_m is a subset of probability measures on E, whose elements are tight measures (using the same approximating compact sets as $\bar{\alpha}_m$ for every level of approximation j). Also, let us notice that if, for a specific $\omega \in \Omega$, there exists $m \in \mathbb{N}$ (which may depend on ω) s.t. $\forall k \geq m,\ \alpha_k(\omega, \cdot) \in \Theta_m$, then the sequence $(\alpha_n(\omega, \cdot))_{n\in\mathbb{N}}$ is tight. Indeed, choose $\epsilon' > 0$ arbitrary, and pick j large enough s.t. $2^{-j} < \epsilon'$. Then, for all $k \geq m$, $\alpha_k(\omega, \cdot) \in \Theta_m$, so $\alpha_k(\omega, K_{m,j}) \geq 1 - 2^{-j} > 1 - \epsilon'$. For each $k < m$, $\alpha_k(\omega, \cdot)$ is a tight measure (E Polish), so there is a compact K_k (dependent on ω) s.t. $\alpha_k(\omega, K_k) > 1 - \epsilon'$. Finally, $K = K_{m,j} \cup (\cup_{k=1}^{n} K_k)$ is a compact set (being finite union of compact sets), s.t. $\alpha_n(\omega, K) > 1 - \epsilon'$ for all $n \in \mathbb{N}$. Therefore, in order to show that $(\alpha(\cdot, \cdot))_{n\geq 0}$ is a.s. tight, it is enough to show that

$$\mathbb{P}\left(\{\omega \in \Omega : \exists m \in \mathbb{N} \text{ s.t. } \forall k \geq m,\ \alpha_k(\omega, \cdot) \in \Theta_m\}\right) = 1. \tag{17}$$

The left hand side of equation (17) can be lower bounded as follows,

$$\begin{aligned}
&\mathbb{P}\left(\{\omega \in \Omega : \exists m \in \mathbb{N} \text{ s.t. } \forall k \geq m,\ \alpha_k(\omega, \cdot) \in \Theta_m\}\right) \\
&= \mathbb{P}\left(\cup_{m\in\mathbb{N}} \cap_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, \cdot) \in \Theta_m\}\right) \\
&= \mathbb{P}\left(\cup_{m\in\mathbb{N}} \cap_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \leq 2^{-j}, \forall j \in \mathbb{N}\}\right) \\
&= \mathbb{P}\left(\cap_{j\in\mathbb{N}} \cup_{m\in\mathbb{N}} \cap_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \leq 2^{-j}\}\right) \\
&= 1 - \mathbb{P}\left(\cup_{j\in\mathbb{N}} \cap_{m\in\mathbb{N}} \cup_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right) \\
&\geq 1 - \sum_{j\in\mathbb{N}} \mathbb{P}\left(\cap_{m\in\mathbb{N}} \cup_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right),
\end{aligned} \tag{18}$$

where the second equality follows from the definition of Θ_m. The j-th term in the last sum can be upper bounded as follows,

$$\begin{aligned}
&\mathbb{P}\left(\cap_{m\in\mathbb{N}} \cup_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right) \\
&= \lim_{m\to\infty} \mathbb{P}\left(\cup_{k\geq m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right) \\
&= \lim_{m\to\infty} \mathbb{P}\left(\lim_{n_m\to\infty} \cup_{k=m}^{n_m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right) \\
&= \lim_{m\to\infty} \lim_{n_m\to\infty} \mathbb{P}\left(\cup_{k=m}^{n_m} \{\omega \in \Omega :\ \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right) \\
&= \lim_{m\to\infty} \lim_{n_m\to\infty} \mathbb{P}\left(\{\omega \in \Omega :\ \max_{m\leq k\leq n_m} \alpha_k(\omega, K_{m,j}^c) \geq 2^{-j}\}\right) \\
&\leq \lim_{m\to\infty} \lim_{n_m\to\infty} 2^j \left(\mathbb{E}(\alpha_m(\cdot, K_{m,j}^c)) + \mathbb{E}(\sum_{k=m}^{n_m} \xi_k)\right) \\
&= \lim_{m\to\infty} 2^j \bar{\alpha}_m(K_{m,j}^c) + 2^j \lim_{m\to\infty} \sum_{k=m}^{\infty} \mathbb{E}(\xi_k) \\
&= \lim_{m\to\infty} 2^j \bar{\alpha}_m(K_{m,j}^c) \\
&\leq \lim_{m\to\infty} 2^j \epsilon 2^{-2j} = \epsilon 2^{-j},
\end{aligned}$$

where the first inequality follows from the maximal inequality for almost supermartingales, Proposition 15, and unconditioning over $\mathcal{G}_m$, which means integrating both sides of the inequality over E. The second to last equality follows from the assumption that $\mathbb{E}(\sum_{k=1}^{\infty} \xi_k) < \infty$, which in turns implies that $\sum_{k=1}^{\infty} \mathbb{E}(\xi_k) < \infty$, due to monotone convergence and ξ_n being non-negative random variables. Finally, the last inequality follows by the specific choice of $K_{m,j}$.

Using this upper bound in (18),

$$\mathbb{P}\left(\{\omega \in \Omega : \exists m \in \mathbb{N} \text{ s.t. } \forall k \geq m,\ \alpha_k(\omega, \cdot) \in \Theta_m\}\right) \geq 1 - \sum_{j \in \mathbb{N}} \epsilon 2^{-j} = 1 - \epsilon,$$

and, because ϵ is arbitrary, this implies (17), hence it concludes the proof. ■

Proof [Proof of Corollary 7] From Theorem 6, for all ω in a set of probability one, we have that for all $f : E \to \mathbb{R}$ continuous and bounded,

$$\lim_{n \to \infty} \mathbb{E}(f(X_{n+1})|\mathcal{G}_n)(\omega) \to \int_{\mathbb{R}} f(x)\alpha(\omega, dx), \tag{19}$$

where $\mathbb{E}(f(X_{n+1})|\mathcal{G}_n)(\omega) = \int_{\mathbb{R}} f(x_{n+1})\mathbb{P}(X_{n+1} \in dx_{n+1}|\mathcal{G}_n)(\omega)$. On the same set of probability one, we have that for all $f : E \to \mathbb{R}$ continuous and bounded,

$$\begin{aligned}
\lim_{M \to \infty} \mathbb{E}\left(f(X_{n+M})|\mathcal{G}_n\right) &= \lim_{M \to \infty} \mathbb{E}\left[\mathbb{E}\left[f(X_{n+M})|\mathcal{G}_{n+M-1}\right] \middle| \mathcal{G}_n\right] \\
&= \mathbb{E}\left[\lim_{M \to \infty} \mathbb{E}\left[f(X_{n+M})|\mathcal{G}_{n+M-1}\right] \middle| \mathcal{G}_n\right] \\
&= \mathbb{E}\left[\int_{\mathbb{R}} f(x)\alpha(\cdot, dx) \middle| \mathcal{G}_n\right],
\end{aligned}$$

where the first equality follows from the fact that $\mathcal{G}_n \subset \mathcal{G}_{n+M-1}$, the second one from the conditional bounded convergence theorem, and the last one from (19). Hence, $\mathbb{P}\left(X_{n+M} \in \cdot|\mathcal{G}_n\right) \xrightarrow{w} \mathbb{E}[\alpha(\cdot)|\mathcal{G}_n]$ almost surely. ■

B.2 Proofs of Section 4.2

Proposition 8 proof is done in two steps. First we show that $(\hat{\theta})_{n \geq 1}$ converge $a.s.$; then we prove that Definition 4 holds. The following Lemma is trivial and surely known but we did not find a reference for it.

Lemma 20 *Let $(Z_n)_{n \geq 1}$ be a sequence of a.s. finite r.v.s. such that $Z_n \xrightarrow{a.s.} Z$, for some a.s. finite r.v. Z. Let $(\eta_n)_{n \geq 1}$ be a deterministic positive summable sequence $\sum_{n=1}^{\infty} \eta_n < \infty$. Then, it holds that $\sum_{n=1}^{\infty} \eta_n Z_n < \infty$ a.s.*

Proof Without loss of generality, we can assume that all Z_n are a.s. non-negative. Choose $\omega' \in \{\omega \in \Omega : Z_n(\omega) \to Z(\omega), Z(\omega) < \infty\}$. Then, for every $\epsilon > 0$, there exists $N_{\omega'}$ such that for

all $n > N_{\omega'}$, $Z_n(\omega') < Z(\omega') + \epsilon$. Therefore,

$$\sum_{n=N_{\omega'}}^{\infty} \eta_n Z_n(\omega) < (Z(\omega') + \epsilon) \sum_{n=N_{\omega'}}^{\infty} \eta_n < \infty,$$

because ϵ is arbitrary, $(\eta_n)_{n\geq 1}$ summable, and $Z(\omega') < \infty$, the statement follows. ■

Proof [Proof of Proposition 8] First note that $(\hat{\theta})_{n\geq 1}$ is a martingale

$$\mathbb{E}(\hat{\theta}_n|\mathcal{G}_{n-1}) = \hat{\theta}_{n-1} + \eta_n \mathcal{I}(\hat{\theta}_{n-1})^{-1}\mathbb{E}(s(x_n, \hat{\theta}_{n-1})|\mathcal{G}_{n-1}) = \hat{\theta}_{n-1},$$

where the second equality follows from standard properties of the score function. In addition,

$$\begin{aligned}\lim_{n\to\infty} \mathbb{E}(\hat{\theta}_n^2) &= \lim_{n\to\infty} \sum_{i=1}^{n} \mathbb{E}(\mathbb{E}((\hat{\theta}_i - \hat{\theta}_{i-1})^2|\mathcal{G}_{i-1})) = \sum_{i=1}^{\infty} \eta_i^2 \mathbb{E}(\mathbb{E}(Z(x_i, \hat{\theta}_{i-1})^2|\mathcal{G}_{i-1})) \\ &= \sum_{i=1}^{\infty} \eta_i^2 \mathbb{E}(Z(x_i, \hat{\theta}_{i-1})^2) < \infty,\end{aligned}$$

where the first equality follows by martingale properties, the last inequality by Assumption 1 c). Hence, $(\hat{\theta})_{n\geq 1}$ converges a.s. by Doob's martingale convergence theorem.
Now, we show that $(X_n)_{n\geq 1}$ is a.c.i.d. Recall that $\alpha_n(A) = \int_A p(x; \hat{\theta}_n)\mathrm{d}x$.

$$\begin{aligned}\mathbb{E}(\alpha_{n+1}(A) - \alpha_n(A)|\mathcal{G}_n) =& \mathbb{E}\left(\int_A p(x; \hat{\theta}_{n+1}) - p(x; \hat{\theta}_n)\mathrm{d}x|\mathcal{G}_n\right) \\ =& \mathbb{E}\left(\int_A p(x; \hat{\theta}_n)\left(\frac{p(x; \hat{\theta}_{n+1})}{p(x; \hat{\theta}_n)} - 1\right)\mathrm{d}x|\mathcal{G}_n\right). \end{aligned} \tag{20}$$

Let us look at the integrand. A Taylor expansion of $p(x; \theta)$ around $\hat{\theta}_n$ is

$$p(x; \theta) = p(x; \hat{\theta}_n) + p'(x; \hat{\theta}_n)(\theta - \hat{\theta}_n) + \frac{p''(x; \nu_n)}{2}(\theta - \hat{\theta}_n)^2,$$

where ν_n is a point between θ and $\hat{\theta}_n$. We can rewrite the integrand as

$$\begin{aligned}\frac{p(x; \hat{\theta}_{n+1})}{p(x; \hat{\theta}_n)} - 1 &= (\hat{\theta}_{n+1} - \hat{\theta}_n)\frac{p'(x; \hat{\theta}_n)}{p(x; \hat{\theta}_n)} + \frac{(\hat{\theta}_{n+1} - \hat{\theta}_n)^2}{2}\frac{p''(x; \nu_n)}{p(x; \hat{\theta}_n)} \\ &= (\hat{\theta}_{n+1} - \hat{\theta}_n)s(x; \hat{\theta}_n) + \frac{(\hat{\theta}_{n+1} - \hat{\theta}_n)^2}{2}\frac{p''(x; \nu_n)}{p(x; \hat{\theta}_n)}.\end{aligned}$$

Plugging in the above in (20) results in two separate integrals. Recalling that $(\hat{\theta}_{n+1} - \hat{\theta}_n) = \eta_{n+1}Z(x_{n+1}; \hat{\theta}_n)$, the first term can be rewritten as

$$\begin{aligned}&\mathbb{E}\Big(\int_A p(x; \hat{\theta}_n)\eta_{n+1}Z(x_{n+1}; \hat{\theta}_n)s(x; \hat{\theta}_n)\mathrm{d}x|\mathcal{G}_n\Big) = \\ &= \eta_{n+1}\int_{\mathbb{R}} p(x_{n+1}; \hat{\theta}_n)\int_A p(x; \hat{\theta}_n)\mathcal{I}(\hat{\theta}_n)^{-1}s(x_{n+1}; \hat{\theta}_n)s(x; \hat{\theta}_n)\mathrm{d}x\mathrm{d}x_{n+1} \\ &= \eta_{n+1}\mathcal{I}(\hat{\theta}_n)^{-1}\int_A p(x; \hat{\theta}_n)s(x; \hat{\theta}_n)\int_{\mathbb{R}} p(x_{n+1}; \hat{\theta}_n)s(x_{n+1}; \hat{\theta}_n)\mathrm{d}x_{n+1}\mathrm{d}x = 0,\end{aligned}$$

where the last equality follows from $\int_{\mathbb{R}} p(x_{n+1};\hat{\theta}_n)s(x_{n+1};\hat{\theta}_n)\mathrm{d}x_{n+1} = 0$. The second integral can be written as

$$\begin{aligned}
&\mathbb{E}\Big(\int_A p(x;\hat{\theta}_n)\frac{(\hat{\theta}_{n+1}-\hat{\theta}_n)^2}{2}\frac{p''(x;\nu_n)}{p(x;\hat{\theta}_n)}\mathrm{d}x|\mathcal{G}_n\Big) = \\
&= \int_{\mathbb{R}} p(x_{n+1};\hat{\theta}_n)\int_A p(x;\hat{\theta}_n)\frac{(\hat{\theta}_{n+1}-\hat{\theta}_n)^2}{2}\frac{p''(x;\nu_n)}{p(x;\hat{\theta}_n)}\mathrm{d}x\,\mathrm{d}x_{n+1} \\
&= \int_{\mathbb{R}} p(x_{n+1};\hat{\theta}_n)\frac{(\hat{\theta}_{n+1}-\hat{\theta}_n)^2}{2}\int_A p(x;\hat{\theta}_n)\frac{p''(x;\nu_n)}{p(x;\hat{\theta}_n)}\mathrm{d}x\,\mathrm{d}x_{n+1} \\
&\leq C\int_{\mathbb{R}} p(x_{n+1};\hat{\theta}_n)\frac{(\hat{\theta}_{n+1}-\hat{\theta}_n)^2}{2}\mathrm{d}x_{n+1},
\end{aligned}$$

where $C < \infty$ is a constant that exists by Assumption 1 d).
$(X_n)_{n\geq 1}$ is a.c.i.d. with ξ_n given by

$$C\frac{\eta_{n+1}^2}{2}\int_{\mathbb{R}} p(x_{n+1};\hat{\theta}_n)Z^2(x_{n+1};\hat{\theta}_n)\mathrm{d}x_{n+1} = C\frac{\eta_{n+1}^2}{2}\mathcal{I}(\hat{\theta}_n)^{-1} \leq C\frac{\eta_{n+1}^2}{2\epsilon} := \xi_n \tag{21}$$

where ϵ is from Assumption 1 b). Because η_n^2 is summable, so is ξ_n. ■

Proof (Proof of Proposition 10). It is not necessary to show that $(\tilde{\theta}_n)_{n\geq 1}$ converges almost surely. We follow the same steps as in the proof of Proposition 8, up to the definition of ξ_n in (21). At this point, the information function $\mathcal{I}(\theta)$ is assumed to be continuous and greater than ϵ (Assumption 1 b)), so it achieves a minimum, greater than ϵ, on the compact set $\Theta = [\underline{\theta}, \bar{\theta}]$. Therefore, we can upper bound equation (21) by $C\frac{\eta_{n+1}^2}{2\epsilon}$, which is summable under the assumptions of Proposition 10. ■

Proof (Proof of Example 2) (Student-t distribution) Let P_θ be a location-scale Student-t distribution with location θ and known scale τ and degrees of freedom ν, with density $p(x;\theta)$ with density

$$p(x;\theta) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\tau\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-\theta)^2}{\nu\tau^2}\right)^{-\frac{\nu+1}{2}}.$$

We need to show that $\int p''(x;\theta)\mathrm{d}x < \infty$ to show Assumption 1 d). The first derivative $p'(x;\theta)$ of $p(x;\theta)$ is

$$\frac{dp(x;\theta)}{d\theta} = p(x;\theta)\left(\frac{\nu+1}{\nu\tau^2}\right)\left(1+\frac{(x-\theta)^2}{\nu\tau^2}\right)^{-1}(x-\theta).$$

Differentiating again and simplifying, the second derivative $p''(x;\theta)$ is

$$\begin{aligned}
\frac{d^2p(x;\theta)}{d^2\theta} = p(x;\theta)\left(\frac{(\nu+1)(\nu+3)}{(\nu\tau^2)^2}\right)\left(1+\frac{(x-\theta)^2}{\nu\tau^2}\right)^{-2}(x-\theta)^2 + \\
- p(x;\theta)\left(\frac{\nu+1}{\nu\tau^2}\right)\left(1+\frac{(x-\theta)^2}{\nu\tau^2}\right)^{-1}.
\end{aligned}$$

In order to show that $\int p''(x;\theta)\mathrm{d}x < C < \infty$, for some constant C, we can just show that, after dropping constants and changing variable to $y = x - \theta$, the following two integrals are both bounded,

$$\begin{aligned}\int_{-\infty}^{\infty} p(y;\theta)\left(1+\frac{y^2}{\nu\tau^2}\right)^{-2} y^2 dy &\propto \int_{-\infty}^{\infty} p(y;\theta)\frac{y^2}{(\nu\tau^2+y^2)^2}\mathrm{d}y \\ &\leq \frac{1}{4\nu\tau^2}\int_{-\infty}^{\infty} p(y;\theta)\mathrm{d}y = \frac{1}{4\nu\tau^2},\end{aligned}$$

where the inequality follows from the fact that the function $f(y) = \frac{y^2}{(\nu\tau^2+y^2)^2}$ is positive and bounded, with maxima at $y = \pm\sqrt{\nu\tau^2}$. The second integral,

$$\int_{-\infty}^{\infty} p(y;\theta)\left(1+\frac{y^2}{\nu\tau^2}\right)^{-1}\mathrm{d}y \propto \int_{-\infty}^{\infty} p(y;\theta)\frac{1}{y^2+\nu\tau^2}\mathrm{d}y \leq \int_{-\infty}^{\infty} p(y;\theta)\frac{1}{\nu\tau^2}\mathrm{d}y = \frac{1}{\nu\tau^2}.$$

■

B.3 Proofs of Section 4.3

To prove Proposition 11, we first consider the sequence of densities $(p_n)_{n\geq 0}$ corresponding to the sequence of measures $(\alpha_n)_{n\geq 0}$ defined in (9) such that $\alpha_n(A) = \int_A p_n(x)\mathrm{d}x$ holds for all $A \in \mathcal{B}(\mathbb{R})$ and $n \geq 0$. Existence of such densities is guaranteed by construction. $p_n(x)$ can be similarly defined recursively:

$$p_n(x) = \frac{1}{n}\sum_{i=1}^{n-1} K_i(x - X_i) = \frac{n-1}{n}p_{n-1}(x) + \frac{1}{n}K_n(x - X_n), \tag{22}$$

where recall that $K_n(x) = \frac{1}{h_n}K(\frac{x}{h_n})$.

Lemma 21 *Consider the sequence $(p_n)_{n\geq 0}$ defined by* (22)*. Suppose that $\mathbb{E}(p_{n+1}(x)|\mathcal{G}_n) = p_n(x)$ holds a.s. for all $x \in \mathbb{R}$, then it must be that $K(x) = \delta(x)$, the Dirac delta function.*

Proof Fix $x \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}\left(p_{n+1}(x)|\mathcal{G}_n\right) &= \mathbb{E}\left(\frac{n}{n+1}p_n(x) + \frac{1}{n+1}K_{n+1}\left(x - X_{n+1}\right)|\mathcal{G}_n\right) \\ &= p_n(x) + \frac{1}{n+1}\left[\mathbb{E}\left(K_{n+1}\left(x - X_{n+1}\right)|\mathcal{G}_n\right) - p_n(x)\right].\end{aligned}$$

For $\mathbb{E}(p_{n+1}(x)|\mathcal{G}_n) = p_n(x)$ to hold, it must be that $\mathbb{E}\left(K_{n+1}\left(x - X_{n+1}\right)|\mathcal{G}_n\right) = p_n(x)$. In other words,

$$\begin{aligned}&\mathbb{E}\left(K_{n+1}\left(x - X_{n+1}\right)|\mathcal{G}_n\right) - p_n(x) = \\ &= \int K_{n+1}\left(x - x_{n+1}\right)\left(\frac{1}{n}\sum_{i=1}^{n} K_i(x_{n+1} - X_i)\right)\mathrm{d}x_{n+1} - p_n(x) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\int K_{n+1}\left(x - x_{n+1}\right)K_i(x_{n+1} - X_i)\mathrm{d}x_{n+1} - K_i(x - X_i)\right].\end{aligned}$$

By the same argument in the proof of West (1991) Theorem 1, the term above equals zero if and only if all the terms in the square bracket equal zero. The reason is that it must hold for all n a.s., so for all samples $x_{1:n}$ except for some measure zero set. Letting subsets of the data go to infinity, the corresponding kernel density goes to 0, which implies that all the summands must be equal to zero. We need to show that

$$\int K_{n+1}\left(x - x_{n+1}\right) K_i(x_{n+1} - X_i)\mathrm{d}x_{n+1} - K_i(z) = \int K_{n+1}(z-y)K_i(y)\mathrm{d}y - K_i(z) = 0,$$

where the first equality follows by substitution. The convolution of densities above has as characteristic function the product of characteristic function. The equation can be rewritten in terms of characteristic functions corresponding to the densities above

$$\gamma_{n+1}(t)\gamma_i(t) - \gamma_i(t) = 0.$$

$\gamma_i(t)$ cannot be equal to 0 for all t, which implies that $\gamma_{n+1}(t) = 1$. This completes the proof. ■

Proof [Proof of Proposition 11] In order to prove that the sequence is c.i.d., we need to check that $(\alpha_n(A))_{n\in\mathbb{N}}$ is a martingale for all $A \in \mathcal{E}$. It is actually sufficient to check the martingale property for sets of the form $A = (-\infty, t]$ for all $t \in \mathbb{R}$. Writing $\alpha_{n+1}((-\infty, t]) = F_{n+1}(t)$, we have that

$$\begin{aligned}
\mathbb{E}\left(F_{n+1}(t)|\mathcal{G}_n\right) &= \mathbb{E}\left(\frac{n}{n+1}F_n(t) + \frac{1}{(n+1)h_{n+1}}\int_{-\infty}^{t} K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x|\mathcal{G}_n\right) \\
&= \frac{n}{n+1}F_n(t) + \frac{1}{(n+1)h_{n+1}}\mathbb{E}\left(\int_{-\infty}^{t} K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x|\mathcal{G}_n\right) \\
&= F_n(t) + \frac{1}{(n+1)}\left[\mathbb{E}\left(\int_{-\infty}^{t}\frac{1}{h_{n+1}}K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x|\mathcal{G}_n\right) - F_n(t)\right].
\end{aligned}$$

In order to be a martingale, the second term needs to be equal to zero for all $n \in \mathbb{N}$. Therefore,

$$\begin{aligned}
\mathbb{E}\left(\int_{-\infty}^{t}\frac{1}{h_{n+1}}K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x|\mathcal{G}_n\right) &= F_n(t) = \mathbb{P}(X_{n+1} \leq t|\mathcal{G}_n) \\
&= \mathbb{E}(\mathbb{I}(X_{n+1} \leq t)|\mathcal{G}_n).
\end{aligned}$$

The integral on the left hand side is

$$\int_{-\infty}^{t}\frac{1}{h_{n+1}}K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x = \int_{-\infty}^{\frac{t-X_{n+1}}{h_{n+1}}} K\left(u\right)\mathrm{d}u.$$

Taking the limit $h_{n+1} \to 0^+$, we see that if $t - X_{n+1} > 0$, $\lim_{h_{n+1}\to 0^+}\int_{-\infty}^{\frac{t-X_{n+1}}{h_{n+1}}} K\left(u\right)\mathrm{d}u = \int_{-\infty}^{+\infty} K\left(u\right)\mathrm{d}u = 1$. Instead, if $t - X_{n+1} < 0$ $\lim_{h_{n+1}\to 0^+}\int_{-\infty}^{\frac{t-X_{n+1}}{h_{n+1}}} K\left(u\right)\mathrm{d}u = \int_{-\infty}^{-\infty} K\left(u\right)\mathrm{d}u = 0$. Therefore,

$$\lim_{h_{n+1}\to 0^+}\int_{-\infty}^{\frac{t-X_{n+1}}{h_{n+1}}} K\left(u\right)\mathrm{d}u = \mathbb{I}(X_{n+1} \leq t).$$

a.s. with respect to $\mathbb{P}(\cdot|\mathcal{G}_n)$ (given that the event $\{X_{n+1} = t\}$ has probability zero under this conditional distribution). Because this needs to hold for all $n \in \mathbb{N}$, the sequence of r.v.s is c.i.d. if for all $n \in \mathbb{N}$ $h_n \to 0^+$, which corresponds to the case of empirical distribution function, $F_n(t) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{I}(X_i \leq t)$.

To prove necessity, note that

$$\begin{aligned}\mathbb{E}\left(\alpha_{n+1}(A)|\mathcal{G}_n\right) &= \alpha_n(A) + \frac{1}{(n+1)}\left[\mathbb{E}\left(\int_A K_{n+1}\left(x - X_{n+1}\right)\mathrm{d}x|\mathcal{G}_n\right) - \alpha_n(A)\right] \\ &= \alpha_n(A) + \frac{1}{(n+1)}\left[\int_A \mathbb{E}\left(K_{n+1}\left(x - X_{n+1}\right)|\mathcal{G}_n\right)\mathrm{d}x - \int_A p_n(x)\mathrm{d}x\right].\end{aligned}$$

In order for the term in the squared bracket to be equal to zero for all $A \in \mathcal{B}(\mathbb{R})$,

$$\mathbb{E}\left(K_{n+1}\left(x - X_{n+1}\right)|\mathcal{G}_n\right) = p_n(x)$$

must holds almost surely. The result follows by Lemma 21. ■

Proof [Proof of Proposition 12] First notice that the predictive distribution in (9) can be written non-recursively as

$$\alpha_n(A) = \frac{1}{nh_i}\sum_{i=1}^{n}\int_A K\left(\frac{x - X_i}{h_i}\right)dx = \frac{1}{n}\sum_{i=1}^{n}\mu_i(A; X_i). \tag{23}$$

Moreover, for any $A \in \mathcal{B}(\mathbb{R})$, by (9), we have that

$$\mathbb{E}(\alpha_{n+1}(A) - \alpha_n(A)|\mathcal{G}_n) = \frac{1}{n+1}\left[\mathbb{E}\left(\int_A \frac{1}{h_{n+1}}K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x\middle|\mathcal{G}_n\right) - \alpha_n(A)\right].$$

Because $X_{n+1}|\mathcal{G}_n \sim \alpha_n$ and then plugging (23), the expectation on the right hand side equals to

$$\begin{aligned}&\mathbb{E}\left(\int_A \frac{1}{h_{n+1}}K\left(\frac{x - X_{n+1}}{h_{n+1}}\right)\mathrm{d}x\middle|\mathcal{G}_n\right) = \\ &\quad= \int\int_A \frac{1}{h_{n+1}}K\left(\frac{x - x_{n+1}}{h_{n+1}}\right)\mathrm{d}x\,\alpha_n(\mathrm{d}x_{n+1}) \\ &\quad= \int\int_A \frac{1}{h_{n+1}}K\left(\frac{x - x_{n+1}}{h_{n+1}}\right)\frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_i}K\left(\frac{x_{n+1} - X_i}{h_i}\right)\mathrm{d}x\,\mathrm{d}x_{n+1} \\ &\quad= \frac{1}{n}\sum_{i=1}^{n}\int_A\int \frac{1}{h_{n+1}}K\left(\frac{x - x_{n+1}}{h_{n+1}}\right)\frac{1}{h_i}K\left(\frac{x_{n+1} - X_i}{h_i}\right)\mathrm{d}x_{n+1}\,\mathrm{d}x \\ &\quad= \frac{1}{n}\sum_{i=1}^{n}\mu_{n+1*i}(A; X_i).\end{aligned}$$

Hence,

$$\begin{aligned}\mathbb{E}(\alpha_{n+1}(A)|\mathcal{G}_n) - \alpha_n(A) =& \frac{1}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n}\mu_{n+1*i}(A; X_i) - \frac{1}{n}\sum_{i=1}^{n}\mu_i(A; X_i)\right) \\ =& \frac{1}{n+1}\frac{1}{n}\sum_{i=1}^{n}\left(\mu_{n+1*i}(A; X_i) - \mu_i(A; X_i)\right) \\ \leq& \frac{1}{n+1}\frac{1}{n}\sum_{i=1}^{n}C\frac{h_{n+1}^d}{h_i^d} = \xi_n,\end{aligned}$$

where the last inequality follows by (11). This shows that $(X_n)_{n\geq 1}$ is $\mathcal{G}$-a.c.i.d. with ξ_n given in the last line above.
The summability of $(\xi_n)_{n\geq 1}$ follows from Theorem 3.29 and Theorem 3.25 in Rudin et al. (1964), which implies asymptotically exchangeability of $(X_n)_{n\geq 1}$ by Theorem 6. ■

In addition, we include the details of an additional example (uniform kernel), which we briefly mentioned in the main manuscript.

Example 7 *(Uniform kernel) Let $K(x) = \frac{1}{2}$ for $|x| \leq 1$, which implies $K_n(x) = \mathcal{U}(x; X_n - h_n, X_n + h_n)$, a uniform density with lower and upper bounds $(X_n - h_n, X_n + h_n)$. The convolution of uniform kernels $(K_{n+1} * K_i)(x; X_i, h)$ can be computed analytically. For $h_{n+1} < h_i$,*

$$(K_{n+1} * K_i)(x; X_i, h) = \begin{cases} \frac{x - X_i + h_{n+1} + h_i}{4h_{n+1}h_i} & X_i - (h_i + h_{n+1}) \leq x < X_i - (h_i - h_{n+1}) \\ \frac{1}{2h_i} & X_i - (h_i - h_{n+1}) \leq x < X_i + (h_i - h_{n+1}) \\ \frac{X_i - x + h_{n+1} + h_i}{4h_{n+1}h_i} & X_i + (h_i - h_{n+1}) \leq x < X_i + (h_i + h_{n+1}). \end{cases}$$

and 0 if $x < X_i - (h_i + h_{n+1})$ or $x \geq X_i + (h_i + h_{n+1})$. This is the density of a trapezoidal distribution centered at X_i.
The total variation can be computed exactly, as it is given by a simple geometric argument, by integrating over the region in which the trapezoidal density is over the uniform one, or viceversa. Therefore, by symmetry of both densities around X_i,

$$TV(\mu_{n+1*i}(\cdot, X_i), \mu_i(\cdot, X_i)) = 2 \int_{X_i - (h_i + h_{n+1})}^{X_i - h_i} \frac{x - X_i + h_{n+1} + h_i}{4h_{n+1}h_i} \mathrm{d}x = \frac{1}{4} \frac{h_{n+1}}{h_i}.$$

Appendix C. Simulations

All code to reproduce the simulation study is available online at `https://github.com/lorenzocapp/acid_paper`. We include here some details for completeness.

C.1 Simulations: Data Generating Mechanism

Section 5.1. We repeat 100 times the following steps: i) sample a random distribution P (as described in the next paragraph), ii) given P, sample n $i.i.d.$ values from P.

In Step i), we generate a mixture as follows: a) sample number of components $n_{comp} \sim \mathcal{UD}\{2, 3\}$ ($\mathcal{UD}$ stands for uniform discrete), b) sample the means of the kernel components, $(\mu_1, \ldots, \mu_{n_{comp}})$, each with a proposal distribution $\mathcal{U}(-3, 3)$ and acceptance rule $|\mu_i - \mu_j| \geq 1.5$ for all $i \neq j$, c) sample the scale parameter of each mixture component $\sigma_i^2 \sim \mathcal{U}(0.5, 2)$ for $i = 1, \ldots, n_{comp}$, d) sample the kernel type $K_i \sim \mathcal{UD}\{\text{Gaussian, Student-t (df=5)}\}$, for $i = 1, \ldots, n_{comp}$, e) sample $u_i \sim \mathcal{U}(0, 1)$ for $i = 1, \ldots, n_{comp}$ and define weights of the mixture $(w_1, \ldots, w_{n_{comp}}) = (u_1/\sum u_i, \ldots, u_{n_{comp}}/\sum u_i)$.
Section 5.2. There are three data generating mechanisms. For all of them, we repeat 100 times the following: i) generate a regression function f, ii) generate covariates in training data $i.i.d.$ $X_i \overset{iid}{\sim} \mathcal{U}(0, 5)$, $i = 1, \ldots, n_{train}$, define covariates in test data on a regular grid $[0, 4.95]$ of 100 points, iii) given f, $Y_i = f(x_i) + \epsilon_i$, for $i = 1, \ldots, n_{train} + n_{test}$.

Data generating mechanism 1. $f \sim \mathcal{GP}(0, K)$, where K has entries $k(x, y) = \exp\{||x - y||_2\}$ when $x \neq y$ and $k(x, x) = 1 + 10^{-8}$ (10^{-8} is the nugget). $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, 1)$.

Data generating mechanism 2. $f \sim \mathcal{GP}(0, K)$, where K has entries $k(x, y) = \exp\{||x - y||_2\}$ when $x \neq y$ and $k(x, x) = 1 + 10^{-8}$ (10^{-8} is the nugget). $\epsilon_i \overset{iid}{\sim}$ Student-t (df=5).

Data generating mechanism 3. $f \sim \mathcal{GP}(0, K)$, where K has entries $k(x, y) = 0.8 \exp\{||x - y||_2\} + 0.2 \exp\{||x - y||_1\}$ when $x \neq y$ and $k(x, x) = 1 + 10^{-8}$ (10^{-8} is the nugget). $\epsilon_i \overset{iid}{\sim} \mathrm{N}(0, 1)$.

C.2 Implementation: Alternative Methodologies

Gaussian Copula (GauC) (Hahn et al., 2018). We use the predictive resampling implementation of Fong et al. (2023), available at `https://github.com/edfong/MP`, which includes a dedicated Python module (we refer explicitly to the relevant files below). In Section 5.1, the implementation is based on the file `1_univariate_copula.py`, with a minor modification: we impose a lower bound on the copula correlation parameter ρ, which in the authors' code is set to a value ρ_{auto} following an automatic selection procedure. This change addresses the cases in which the automatic selection procedure yields $\rho_{auto} \approx 0$, effectively reducing the model to the independence copula, which results in no updates. We considered a grid of lower bounds $\rho_{lower} \in \{0.4, 0.5, 0.6\}$, set $\rho_{\text{selected}} = \max\{\rho_{\text{auto}}, \rho_{lower}\}$, and adopted the setting with the best empirical performance: 0.4. Specifically, we set $\rho_{\text{selected}} = \max\{\rho_{\text{auto}}, 0.4\}$. This simple adjustment leads to substantial improvements over the default choice ρ_{auto}. In Section 5.3, we used `1_univariate_copula.py` for the galaxy data, and `2_bivariate_copula.py` for the air quality data (no modifications). For further implementation details, we refer to Fong et al. (2023).

Dirichlet Process Mixture (DPM) (Lo, 1984). We use the `R` package `dirichletprocess` of Ross and Markwick (2018), following the code provided at `https://github.com/edfong/MP`, which was designed to compare DPM with GauC (we refer to the exact scripts below). In Section 5.1, we use `1_gmm_dpmm.R` for density estimation with a Gaussian location-scale Dirichlet Process mixture. For the Gaussian-Inverse Gamma base measure, we set the hyperparameters to $c(0, 1, 0.1, 0.1)$. In Section 5.2, we estimate a joint bivariate density using a multivariate Gaussian location-scale DPM. We set the hyperparameters of the multivariate Gaussian-Wishart base measure to $c(\mathbf{0}, \mathbf{I}, \mathbf{I}, \mathbf{I})$. From the posterior, we numerically approximate the regression function. In Section 5.3, we use `1_galaxy_dpmm.R` for the galaxy data (Gaussian DPM with unknown mean and variance) and `4_lidar_DPMM.R` for the LIDAR data, applying the same regression function used in Section 5.2.

Gaussian Processes. We use the `R` package `laGP` (Gramacy, 2016). In Section 5.2, we define an anisotropic GP object via `newGPsep`, optimize kernel parameters using `mleGPsep`, and predict on the test set using the mean from `predGPsep`.

C.3 Implementation: Predictive Resampling for Kernels

Algorithm 1 outlines the implementation of predictive resampling with kernel methods (KerP) used in Section 5. Given a statistical estimator f, the predictive resampling component of the algorithm (lines 15–21) follows a standard bootstrap-like scheme. What distinguishes KerP is its specific considerations regarding parameter selection. Apart from the choice of kernel (a comparison is included in the main manuscript), the only parameter to be specified is the bandwidth sequence $(h_n)_{n\geq 1}$. While bandwidth selection has been widely studied in the contexts of kernel density estimation and kernel regression (see Jones et al. (1996) for a review), its application within a

bootstrap (predictive resampling) framework is novel. We leave the development of a dedicated method for bandwidth selection tailored to predictive resampling to future work. Here, we propose a heuristic approach (lines 1–14) that combines existing bandwidth selection methods with the theoretical results presented in this paper. We describe the rationale for this proposal below.

First, observe that it is not necessary to choose an initial distribution $\alpha_0(\cdot)$ or to choose a full sequence $h_{1:n+M}$ in advance. Given a sample $x_{1:n}$, the predictive resampling procedure does not require the full sequence $(\alpha_i)_{0:n-1}$. We decompose the procedure into two stages: (1) estimate the predictive distribution $\alpha_n(\cdot)$, and (2) define a sequence $h_{n:n+M}$ to iteratively update $\alpha_n(\cdot)$ via (9) and perform sampling.

In the first stage, we estimate the density $p_n(x)$ associated with $\alpha_n(\cdot)$ using a "standard" kernel density estimator, where all observations are assigned the same bandwidth. This design choice allows us to leverage existing bandwidth selection methods. Specifically, choose a method g that given a sample outputs a bandwidth (see next subsection for a discussion of the methods g considered): $h_n = g(x_{1:n})$, where the subscript emphasizes the sample size dependency—common in most methods. Compute $h_n = g(x_{1:n})$ and then set $h_i = h_n$ for all $i \leq n$. This is a natural choice, as it corresponds to a standard kernel density estimate and ensures that all observed data points have equal resampling weights. As opposed to the "synthetic" data points generated in predictive resampling which will have decaying weights.

To construct the sequence $h_{n:n+M}$, we refer to Proposition 12, which requires that $(h_n)_{n\geq 1}$ satisfies condition (13). One way to ensure this is to use an exponentially decaying sequence; see the discussion in Section 4.3. We therefore define $h_{n+i} = b_1 e^{-b_2 i}$, where b_1 and b_2 are chosen such that the sequence interpolates between h_n and h_{n+M}. How to obtain h_n was described in the previous paragraph. To determine h_{n+M}, we draw a synthetic data set of size M via the Bayesian bootstrap and evaluate $h_{n+M} = g(x_{1:n+M})$. To reduce the variance of h_{n+M} (it is a function of the synthetic data set generated through the Bayesian bootstrap), we repeat this step ten times and average the results. We then set $b_1 = h_n$ and compute $b_2 = \log(h_n/h_{n+M})/M$. In numerical studies, we noticed that the bandwidth sequence can benefit from a scaling to avoid trajectories that are too smooth or too bumpy. We added a global scaling c for this. Results in the simulation study are with $c = 1$ (except in the data analysis section).

C.4 Sensitivity to Bandwidth Selection Method

There are many bandwidth selection functions g available; see Jones et al. (1996). In this context, there is no difference between density estimation and regression. Recall that even when we use KerP for regression, we need a multivariate density to sample the synthetic data (see Section 4.4). However, the bandwidth selection methods differ depending on whether the problem is univariate or multivariate. We explore here the sensitivity of KerP to several bandwidth selection methods available in standard `R` packages.

In the univariate case, we consider methods implemented in the `density` function from the `stats` package, which serves as the base kernel estimator in `R`: Silverman's "rule of thumb" (`nrd0`)(Silverman, 2018), Scott's variation of Silverman's "rule of thumb" (`nrd`) Scott (2015), the plug-in method of Sheather and Jones (1991) (`SJ`) and unbiased cross-validation (`ucv`). For multivariate settings, we use the `ks` package (Duong et al., 2024), using the plug-in method of Wand and Jones (1994) (`pi`) and smoothed cross-validation (`scv`). The corresponding `R` function names are indicated in parentheses.

Algorithm 1 KerP predictive resampling (with bandwidth selection)

Inputs: $x_{1:n}$, K (kernel), g (bandwidth selection method), c (scaling factor bandwidth), f (estimator of interest)

1: Compute $h_n = f(x_{1:n})$. ▷ Initialize bandwidth
2: Set $h_{1:n-1} = h_n$.
3: **for** $j = 1$ **to** 10 **do** ▷ Determine endpoint of the bandwidth sequence
4: Sample $X_{n+1:M}$ given $x_{1:n}$ with the Bayesian Bootstrap.
5: Compute $h^{(j)}_{n+M} = g(x_{1:n+M})$.
6: **end for**
7: $h_{n+M} = \frac{1}{10}\sum_{j=1}^{10} h^{(j)}_{n+M}$
8: Set $b_1 = h_n$ and $b_2 = \log(h_n/h_{n+M})/M$ ▷ Define exponentially decaying bandwidth sequence
9: **for** $i = 1$ **to** $M - 1$ **do**
10: Compute $h_{n+i} = b_1 e^{-b_2 i}$
11: **end for**
12: Scale $h_{1:n+M} = c\, h_{1:n+M}$
13: Estimate $p_n(x)$ via kernel density estimate with K and $h_{1:n}$ (such as (9)). ▷ Starting distribution for resampling
14: Compute $\alpha_n(\cdot)$.
15: **for** $b = 1$ **to** B **do** ▷ Predictive resampling
16: **for** $i = 0$ **to** $M - 1$ **do**
17: Sample $X_{n+i} \sim \alpha_n$.
18: Update α_{n+i} via (9) with x_{n+i} and h_{n+i}.
19: **end for**
20: Compute $\hat{\theta}^{(b)}_{n+M} = f(x_{1:n+M})$.
21: **end for**

return Martingale Posterior : $\hat{\theta}^{(1)}_{n+M}, \ldots, \hat{\theta}^{(B)}_{n+M}$.

To compare univariate methods, we replicate the numerical study from Section 5.1. Table C.4 summarizes the median metrics from that same simulation study (univariate cumulative distribution function estimation). While the main manuscript discusses differences between kernel choices, here we focus on comparing bandwidth selection strategies. Overall, performance across SJ and UCV is fairly similar, with the plug-in approach of Sheather and Jones (1991) slightly outperforming cross-validation. Notably, scaling the bandwidth sequence using the parameter c (see line 14 of Algorithm 1) can improve performance. For example, the Gaussian kernel with $c = 0.9$ outperforms $c = 1$—the default shown in the manuscript—by yielding lower bias, higher coverage, and similar width. However, this improvement is inconsistent across kernels: the uniform kernel performs better with $c = 1.1$. In this context, Silverman's rule of thumb does not seem to perform well, nor does Scott's variation of Silverman's rule of thumb.

To compare multivariate methods, we replicated the simulation study from Section 5.2. Table C.4 summarizes the median metrics considering the three data-generating mechanisms (regression function estimation). Overall, the performance is comparable across bandwidth selection methods (`pi`, `cv`) and scaling values ($c \in \{0.9, 1, 1.1\}$). A possible exception is the case of the Gaussian kernel with $c = 0.9$, which exhibits lower bias. Although the performance of the Gaussian kernel is inferior to the other kernel types.

Lastly, note that we discussed the possibility of using one sequence of bandwidths for generating synthetic data and another for estimation. The latter can be set according to the standard rates given by consistency results. We considered both options.

		n = 50			n = 200		
Ker	Bw	*env*	*dev*	*awd*	*env*	*dev*	*awd*
Gaussian	0.9nrd	0.784	0.025	0.090	0.725	0.013	0.043
	0.9nrd0	0.828	0.022	0.089	0.772	0.011	0.043
	0.9SJ	0.831	0.020	0.092	0.772	0.011	0.044
	0.9ucv	0.841	0.020	0.093	0.756	0.010	0.046
	1nrd	0.700	0.029	0.091	0.669	0.016	0.044
	1nrd0	0.800	0.024	0.089	0.762	0.013	0.043
	1SJ	0.825	0.021	0.093	0.769	0.011	0.044
	1ucv	0.812	0.022	0.094	0.759	0.011	0.046
	1.1nrd	0.588	0.034	0.091	0.600	0.018	0.043
	1.1nrd0	0.756	0.027	0.090	0.716	0.014	0.043
	1.1SJ	0.819	0.023	0.093	0.731	0.012	0.044
	1.1ucv	0.781	0.025	0.095	0.769	0.012	0.046
Laplace	0.9nrd	0.359	0.039	0.091	0.425	0.022	0.044
	0.9nrd0	0.575	0.032	0.092	0.600	0.017	0.044
	0.9SJ	0.753	0.026	0.095	0.712	0.014	0.045
	0.9ucv	0.700	0.029	0.097	0.747	0.013	0.046
	1nrd	0.269	0.044	0.091	0.303	0.025	0.045
	1nrd0	0.441	0.036	0.090	0.481	0.020	0.044
	1SJ	0.691	0.029	0.095	0.666	0.016	0.045
	1ucv	0.637	0.032	0.099	0.669	0.014	0.047
	1.1nrd	0.231	0.050	0.090	0.216	0.028	0.044
	1.1nrd0	0.334	0.041	0.091	0.359	0.023	0.044
	1.1SJ	0.637	0.032	0.095	0.547	0.018	0.045
	1.1ucv	0.562	0.036	0.098	0.653	0.016	0.047
Uniform	0.9nrd	0.781	0.019	0.088	0.719	0.010	0.043
	0.9nrd0	0.728	0.018	0.088	0.706	0.010	0.044
	0.9SJ	0.697	0.018	0.090	0.684	0.009	0.044
	0.9ucv	0.713	0.018	0.091	0.678	0.009	0.045
	1nrd	0.800	0.020	0.088	0.744	0.011	0.044
	1nrd0	0.762	0.020	0.088	0.716	0.010	0.044
	1SJ	0.725	0.018	0.090	0.706	0.010	0.044
	1ucv	0.741	0.018	0.091	0.688	0.009	0.045
	1.1nrd	0.831	0.020	0.088	0.747	0.011	0.043
	1.1nrd0	0.803	0.019	0.088	0.725	0.010	0.043
	1.1SJ	0.750	0.018	0.090	0.725	0.010	0.044
	1.1ucv	0.772	0.019	0.092	0.713	0.010	0.045

Table 3: Simulations: sensitivity to bandwidth in cumulative distribution function estimation. Median across 100 data sets for different sample sizes (n). *env* is a measure of coverage (the higher the better), *dev* is an L1 bias (the lower the better), and awd is the average width of the credible bands (the lower the better). In KerP, column Ker refers to the kernel, column bw to the bandwidth: nrd0 stands for Silverman (2018) rule of thumb, nrd for Scott (2015) modification of Silverman's rule of thumb, SJ stands for the bandwidth selection of Sheather and Jones (1991), and ucv for unbiased cross-validation. The number in front of the bw method refers to the scaling factor c.

		DGM1: Gaussian ϵ				DGM2: Student-t ϵ				DGM3: Gauss-Exp Kernel			
Ker	Bw	*env*	*dev*	*awd*	*MSE*	*env*	*dev*	*awd*	*MSE*	*env*	*dev*	*awd*	*MSE*
Gaussian	0.9pi	0.845	0.183	0.677	1.045	0.850	0.173	0.662	1.044	0.880	0.202	0.806	1.593
	0.9scv	0.830	0.184	0.661	1.045	0.830	0.174	0.647	1.047	0.860	0.210	0.773	1.597
	1pi	0.810	0.186	0.662	1.047	0.820	0.180	0.646	1.048	0.860	0.209	0.781	1.596
	1scv	0.800	0.191	0.640	1.049	0.810	0.181	0.630	1.049	0.830	0.216	0.743	1.599
	1.1pi	0.795	0.194	0.647	1.054	0.800	0.181	0.631	1.048	0.850	0.216	0.765	1.601
	1.1scv	0.770	0.195	0.631	1.053	0.790	0.187	0.618	1.052	0.815	0.226	0.729	1.602
Laplace	0.9pi	0.910	0.166	0.736	1.033	0.910	0.164	0.722	1.026	0.920	0.205	0.882	1.586
	0.9scv	0.900	0.163	0.712	1.028	0.895	0.161	0.698	1.026	0.910	0.201	0.847	1.579
	1pi	0.900	0.167	0.712	1.030	0.900	0.164	0.701	1.029	0.920	0.198	0.861	1.580
	1scv	0.890	0.164	0.694	1.033	0.890	0.161	0.682	1.032	0.910	0.200	0.831	1.579
	1.1pi	0.890	0.165	0.689	1.030	0.880	0.161	0.680	1.034	0.910	0.195	0.833	1.582
	1.1scv	0.880	0.168	0.682	1.035	0.870	0.164	0.665	1.033	0.900	0.204	0.800	1.584
Uniform	0.9pi	0.890	0.185	0.769	1.043	0.900	0.180	0.756	1.040	0.870	0.234	0.943	1.606
	0.9scv	0.905	0.175	0.741	1.042	0.895	0.173	0.725	1.041	0.870	0.224	0.897	1.603
	1pi	0.910	0.177	0.737	1.042	0.900	0.174	0.726	1.039	0.865	0.222	0.914	1.606
	1scv	0.910	0.169	0.713	1.036	0.900	0.166	0.701	1.041	0.885	0.211	0.867	1.594
	1.1pi	0.910	0.170	0.719	1.035	0.900	0.166	0.698	1.036	0.870	0.218	0.880	1.596
	1.1scv	0.900	0.165	0.684	1.035	0.900	0.160	0.671	1.035	0.880	0.204	0.836	1.593

Table 4: Simulations: sensitivity to bandwidth in regression function estimation. Median across 100 data sets for three different data generating mechanisms (n_{train} = 200,n_{test} = 100). *env* is a measure of coverage (where higher is the better), *dev* is an L1 bias (the lower the better), and awd is the average width of the credible bands (the lower the better), MSE is the sample mean squared error on test data (the lower the better). In KerP, column Ker refers to the kernel, column bw to the bandwidth: pi stands for the bandwidth selection of Wand and Jones (1994), and scv for smoothed cross-validation. The number in front of the bw method refers to the scaling factor c.

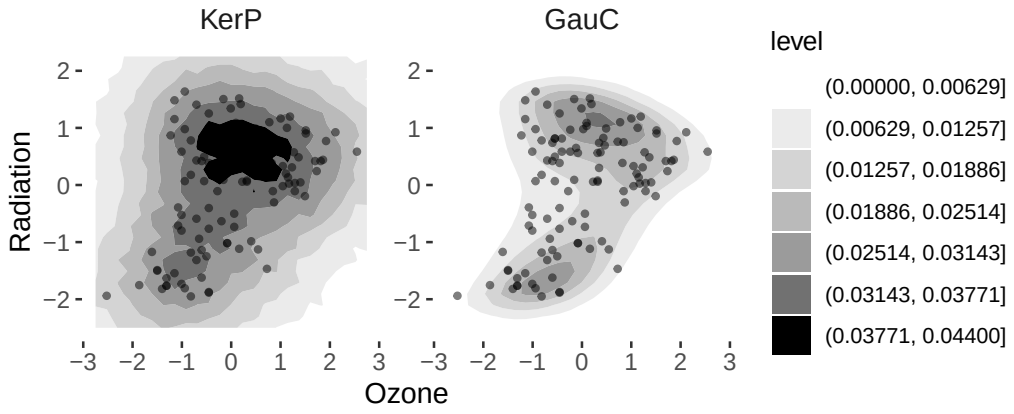


Figure 6: Air quality data: posterior standard deviations. In each panel: contour plot of the posterior standard deviations. Dots depict the observations. The two panels differ by the method: KerP with Gaussian kernel and bandwidth via the multivariate plug-in method of Wand and Jones (1994) (left), and Gaussian copula with the same parameters as in Fong et al. (2023)(right)

Appendix D. Data analysis

D.1 Air Quality Data

Figure 6 depicts the posterior standard deviation of the density estimates obtained with KerP and GauC. KerP is run with a Gaussian kernel with $(h_n)_{n\geq 1}$ chosen with the bandwidth multivariate method of Wand and Jones (1994), $M = 2000$, $B = 500$, and $c = 1.2$.

D.2 LIDAR

Figures 7 and 8 depict the condition density $p(y|x = 0)$ and $p(y|x = -3)$ of the LIDAR regression data set. KerP is run with a Gaussian kernel with $(h_n)_{n\geq 1}$ chosen with the bandwidth multivariate method of Wand and Jones (1994), $M = 1000$, $B = 600$, and $c = 5$. The regression function in the manuscript is obtained with the same parameterization except for $c = 2$. The same setup is used for the out-of-sample prediction exercise.

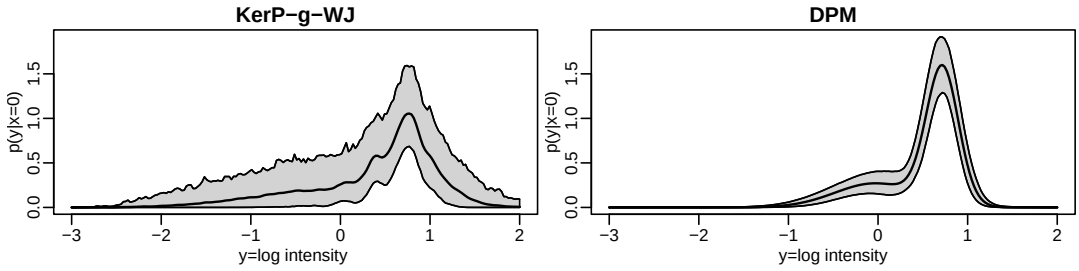


Figure 7: LIDAR: conditional density $p(y|x = 0)$. In each panel: posterior mean $\hat{p}$ (solid line), and 95% credible regions (gray shaded area). The two panels differ by the method: KerP with Gaussian kernel and bandwidth via the multivariate plug-in method of Wand and Jones (1994) (left), and DPM (right).

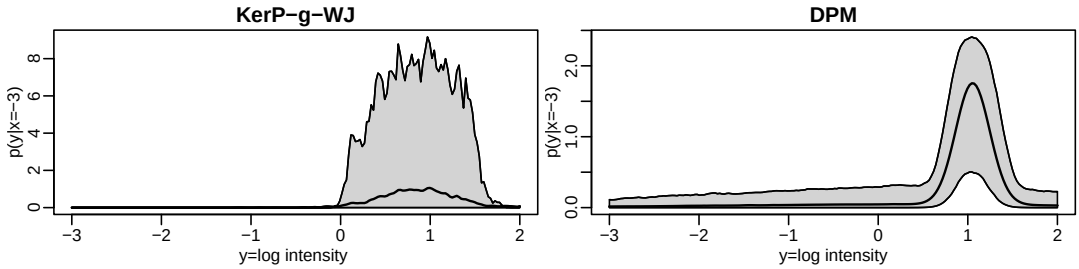


Figure 8: LIDAR: conditional density $p(y|x = -3)$. In each panel: posterior mean $\hat{p}$ (solid line), and 95% credible regions (gray shaded area). The two panels differ by the method: KerP with Gaussian kernel and bandwidth via the multivariate plug-in method of Wand and Jones (1994) (left), and DPM (right). Note: the y-axis in the two panels is on a different scale.

References

E. M. Airoldi, T. Costa, F. Bassetti, F. Leisen, and M. Guindani. Generalized species sampling priors with latent beta reinforcements. *Journal of the American Statistical Association*, 109:1466–1480, 2014.

D. J. Aldous. Exchangeability and related topics. In *Ecole d'Ete de Probabilites de Saint-Flour*, pages 1–198. Springer, 1985.

L. Barabesi and L. Pratelli. An inequality for the total variation distance between high-dimensional centered gaussian laws. *Statistics and Probability Letters*, 211:110–148, 2024.

F. Bassetti, I. Crimaldi, and F. Leisen. Conditionally identically distributed species sampling sequences. *Advances in Applied Probability*, 42:433–459, 2010.

P. Berti, L. Pratelli, and P. Rigo. Limit theorems for a class of identically distributed random variables. *Annals of Probability*, 32(3):2029–2052, 2004.

P. Berti, L. Pratelli, and Rigo P. Limit theorems for empirical processes based on dependent data. *Electronic Journal of Probability*, 17(9):1–18, 2012.

P. Berti, L. Pratelli, and P. Rigo. Exchangeable sequences driven by an absolutely continuous random measure. *Annals of Probability*, 41(3b):2090–2102, 2013.

P. Berti, E. Dreassi, L. Pratelli, and P. Rigo. A class of models for Bayesian predictive inference. *Bernoulli*, 27(1):702–726, 2021.

P. Berti, E. Dreassi, F. Leisen, L. Pratelli, and P. Rigo. A probabilistic view on predictive constructions for Bayesian learning. *Statistical Science*, 40(1):25–39, 2025.

G. Biau and E. Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

P. Billingsley. *Convergence of Probability Measures*. Wiley, Chicago, 1999.

L. Breiman. Some infinity theory for predictor ensembles. Technical Report 577, University of California Berkeley, 2000.

L. Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, 16(3):199–231, 2001.

A. Cassese, W. Zhu, M. Guindani, and M. Vannucci. A bayesian nonparametric spiked process prior for dynamic model selection. *Bayesian Analysis*, 14:553–572, 2019.

M. Chae and S. G. Walker. Wasserstein upper bounds of the total variation for smooth densities. *Statistics and Probability Letters*, 163:108–771, 2020.

K. Chen, W. Shen, and W. Zhu. Covariate dependent beta-gos process. *Computational Statistics and Data Analysis*, 180:107–662, 2023.

B. Clarke and J. Clarke. *Predictive statistics: Analysis and inference beyond models*. Cambridge University Press, Cambridge, 2018.

L. Devroye, A. Mehrabian, and T. Reddad. The total variation distance between high-dimensional gaussians with the same mean. *arXiv preprint arXiv:1810.08693*, 2018.

J. L. Doob. Application of the theory of martingales. In *Le calcul des probabilites et ses applications*, pages 23–27. Éditions du Centre National de la Recherche Scientifique (C.N.R.S.), Paris (Colloques Internationaux du C.N.R.S. XIII), 1949.

T. Duong, M. Wand, J. Chacon, and A. Gramacki. Package 'ks'. *CRAN*, 2024.

B. Efron. Prediction, estimation, and attribution. *Journal of the American Statistical Association*, 115:636–655, 2020.

D. L. Fisk. Quasi-martingales. *Transactions of the American Mathematical Society*, 120(3):369–389, 1965.

E. Fong and A. Yiu. Asymptotics for parametric martingale posteriors. *Biometrika*, 113(2), 2026.

E. Fong, C. Holmes, and S. G. Walker. Martingale posterior distributions. *Journal of the Royal Statistical Society: Series B*, 85(5):1357–1391, 2023.

S. Fortini and S. Petrone. Quasi-Bayes properties of a procedure for sequential learning in mixture models. *Journal of the Royal Statistical Society: Series B*, 82(4):1087–1114, 2020.

S. Fortini and S. Petrone. Prediction-based uncertainty quantification for exchangeable sequences. *Philosophical Transactions of the Royal Society A*, 381(2247):20220142, 2023.

S. Fortini and S. Petrone. Exchangeability, prediction and predictive modeling in bayesian statistics. *Statistical Science*, 40(1):40–67, 2025.

S. Fortini, L. Ladelli, and E. Regazzini. Exchangeability, predictive distributions and parametric models. *Sankhya: The Indian Journal of Statistics, Series A*, 62:86–109, 2000.

S. Garelli, F. Leisen, L. Pratelli, and P. Rigo. Asymptotics of predictive distributions driven by sample means and variances. *Journal of Applied Probability (to appear)*, 2026a.

S. Garelli, F. Leisen, L. Pratelli, and P. Rigo. Flexible learning with adaptive predictive distributions. *Bayesian Analysis (to appear)*, 2026b.

S. Geisser. *Predictive inference: An introduction*. Chapman and Hall, New York, 1993.

A. Gelman, X. L. Meng, and H. Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica*, 6:733–760, 1996.

R. B. Gramacy. lagp: large-scale spatial modeling via local approximate Gaussian processes in r. *Journal of Statistical Software*, 72:1–46, 2016.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

P. R. Hahn, R. Martin, and S. G. Walker. On recursive Bayesian predictive distributions. *Journal of the American Statistical Association*, 113(523):1085–1093, 2018.

C. C. Holmes and S. G. Walker. Statistical inference with exchangeability and martingales. *Philosophical Transactions of the Royal Society A*, 381(2247):20220143, 2023.

M. C. Jones, J. S. Marron, and S. J. Sheather. A brief survey of bandwidth selection for density estimation. *Journal of the American Statistical Association*, 91(433):401–407, 1996.

O. Kallenberg. Spreading and predictable sampling in exchangeable sequences and processes. *Annals of Probability*, 16(2):508–534, 1988.

F. Leisen, L. Pratelli, and Rigo P. Weak convergence in probability of predictive distributions. *Electronic Communications in Probability*, 31:1–15, 2026.

A. Y. Lo. On a class of Bayesian nonparametric estimates: I. Density estimates. *Annals of Statistics*, 12(1):351–357, 1984.

M. A. Newton, F. A. Quintana, and Y. Zhang. Nonparametric Bayes methods using predictive updating. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, volume 133 (D. Dey, P. Muller and D. Sinha, eds.), pages 45–61. Springer, 1998.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971.

K. Roeder. Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. *Journal of the American Statistical Association*, 85(411):617–624, 1990.

G. J. Ross and D. Markwick. *dirichletprocess: An R package for fitting complex Bayesian nonparametric models*, 2018.

D. B. Rubin. The Bayesian bootstrap. *Annals of Statistics*, 9(1):130–134, 1981.

W. Rudin et al. *Principles of mathematical analysis*. McGraw-hill New York, 1964.

E. Scornet. Random forests and kernel methods. *IEEE Transactions on Information Theory*, 62(3): 1485–1500, 2016.

D. W. Scott. *Multivariate density estimation: theory, practice, and visualization*. John Wiley & Sons, 2015.

S. J. Sheather and M. C. Jones. A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B*, 53(3):683–690, 1991.

G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

B. W. Silverman. *Density estimation for statistics and data analysis*. Routledge, 2018.

A. W. Van der Vaart. *Asymptotic statistics*. Cambridge university press, 2000.

M. Wand and M. C. Jones. Multivariate plug-in bandwidth selection. *Computational Statistics*, 9: 97–116, 1994.

M. West. Kernel density estimation and marginalization consistency. *Biometrika*, 78(2):421–425, 1991.

C. Wolverton and T. Wagner. Asymptotically optimal discriminant functions for pattern classification. *IEEE Transactions on Information Theory*, 15(2):258–265, 1969.