Submitted to *Journal of Statistical Theory and Practice*
Vol. 00, No. 00, Month 201X, 1–24

Imprecise Dirichlet process with application to the hypothesis test on the probability that $X \leq Y$

Alessio Benavoli[1] and Francesca Mangili[1] and Fabrizio Ruggeri[2] and Marco Zaffalon[1]

[1] *IPG IDSIA, Manno, Switzerland*
[2] *CNR IMATI, Milano, Italy*

(*Received 00 Month 201X; final version received 00 Month 201X*)

The Dirichlet process (DP) is one of the most popular Bayesian nonparametric models. An open problem with the DP is how to choose its infinite-dimensional parameter (base measure) in case of lack of prior information. In this work we present the Imprecise DP (IDP)—a prior near-ignorance DP-based model that does not require any choice of this probability measure. It consists of a class of DPs obtained by letting the normalized base measure of the DP vary in the set of all probability measures. We discuss the tight connections of this approach with Bayesian robustness and in particular prior near-ignorance modeling via sets of probabilities. We use this model to perform a Bayesian hypothesis test on the probability $P(X \leq Y)$. We study the theoretical properties of the IDP test (e.g., asymptotic consistency), and compare it with the frequentist Mann-Whitney-Wilcoxon rank test that is commonly employed as a test on $P(X \leq Y)$. In particular we will show that our method is more robust, in the sense that it is able to isolate instances in which the aforementioned test is virtually guessing at random.

AMS Subject Classification: 62G10; 62G35.

Keywords: Bayesian nonparametric test, Imprecise Dirichlet Process, Wilcoxon rank sum.

1. Introduction

The Dirichlet process (DP) is one of the most popular Bayesian nonparametric models. It was introduced by Ferguson [1] as a prior over probability distributions. In his seminal paper, Ferguson showed that the DP leads to tractable posterior inferences and can be used for Bayesian analysis of several nonparametric models, such as the estimation of a distribution function, of a mean, of quantiles, of a variance, etc. He also considered the estimation of $P(X \leq Y)$ assigning independent Dirichlet process priors to the distribution functions of X and Y. The Mann-Whitney statistic naturally arises in this case. Susarla and Van Ryzin [2] and Blum and Susarla [3] extended the results of Ferguson on estimation of the distribution function in case of right censored data obtaining a Bayesian version of the Kaplan-Meier estimator. Dalal and Phadia [4] considered the problem of estimating a measure of dependence for a bivariate distribution. The Bayes estimate is computed using a two-dimension Dirichlet prior and Kendall's tau is seen to appear naturally. A review of other similar applications of the DP can be found in [5].

The beauty of the DP is that most of these results are in closed form and that it provides a Bayesian justification of the classic nonparametric estimators. In spite of all these nice properties and of the promising initial outcomes, such a research did not result in the development of DP-based Bayesian nonparametric procedures for hypothesis testing. For instance, the most used statistical packages for DP-based Bayesian nonparametric modeling, "DPpackage" [6] and "Bayesm" [7], include procedures for density estimation, clustering and regression, but do not include any Bayesian version of the Wilcoxon rank sum, Wilcoxon sign test or other classic

nonparametric tests. It is arguable that this absence may be related to the unsettled question of how to choose the prior "parameters" of the DP in case of lack of prior information. Only very recently there has been a renewed interest in the development of Bayesian nonparametric procedures for hypothesis testing [8, 9, 10, 11, 12] – we will return on these approaches later in the paper. However, also in these cases, the choice of the prior parameters is critical, as is evident from the solution commonly chosen to address this problem, namely, the empirical Bayesian approach.

It is well known that a DP is completely characterized by its prior "parameters": the prior strength (or precision), which is a positive scalar number, and the normalized base measure. The question is, how should we choose these prior "parameters" in case of lack of prior information? The only non-empirical solution to this problem that has been proposed so far, first by Ferguson [1] and then by Rubin [13] under the name of Bayesian Bootstrap (BB), is the limiting DP obtained when the prior strength goes to zero. But the BB model has faced quite some controversy, since it is not actually noninformative and moreover it assigns zero posterior probability to any set that does not include the observations. We will discuss these two points in more detail in Section 3.

In this paper we present an alternative viewpoint to the problem of choosing the prior base measure of the DP in case of lack of prior information that overcomes the above drawbacks of the BB. The model we present generalizes to nonparametric setting earlier ideas developed in Bayesian parametric robustness, see Berger [14] and Berger et al. [15] for a review. Here lack of prior information is expressed in terms of a family $\mathscr{T}$ consisting of all prior probability measures that are compatible with the available prior information. Inferences are then carried out by considering the whole family $\mathscr{T}$. In case almost no prior information is available on the parameters of interest, $\mathscr{T}$ should be as large as possible in order to describe this state of prior ignorance. The natural candidate for $\mathscr{T}$ to represent complete ignorance is the set of all probability measures. However, it turns out that the posterior inferences obtained from this set are *vacuous* [16, Sec. 7.3.7], i.e., the posterior set coincides with the prior set. This means that there is no learning from data. Therefore, the vacuous prior model is not a practically useful way to model our prior ignorance. There is then a compromise to be made. Pericchi and Walley [17] and Walley [16] suggest, as an alternative, the use of an almost vacuous model which they call "near-ignorance" or "imprecise" model. This is a model that behaves a priori as a vacuous model for some basic inferences (e.g., prior mean, prior credible regions) but always provides non-vacuous posterior inferences.

While Bayesian robust models have already been extended to the nonparametric setting [18], that has not been the case for near-ignorance models. Note that, a nonparametric model that uses lower and upper bounds for probabilities to quantify uncertainty has been proposed by Augustin and Coolen [19], Coolen and Augustin [20]. However, this model is a purely predictive model, based on post-data assumptions, and, thus, it cannot be used straightforwardly (i.e., without bootstrap) to perform hypothesis tests. The main aim of this paper is to derive a prior near-ignorance DP, called Imprecise DP (IDP). This is the class $\mathscr{T}$ of all DPs obtained by fixing the prior strength of the DP and letting the normalized base measure vary in the set of all probability measures. We will show that the IDP behaves a priori as a vacuous model for all predictive inferences. This, together with the fact that it is a nonparametric model, allows us to start a statistical analysis with very weak assumptions about the problem of interest. However, contrarily to a full vacuous model, we will show that the IDP can learn from data.

Moreover, we will employ the IDP to develop a new Bayesian nonparametric hypothesis test on the probability that $X \leq Y$; we will call this test IDP rank-sum test, due to its similarity with the Mann-Whitney-Wilcoxon (MWW) rank-sum test. This hypothesis test is widely applied; for instance, if X and Y are health status measures in a clinical trial, $P(X \leq Y)$ is, roughly speaking, the probability that the treatment represented by Y is better (not worse) than the treatment represented by X. A Bayesian nonparametric near-ignorance model presents several advantages with respect to a traditional approach to hypothesis testing. First of all, the Bayesian approach

allows us to formulate the hypothesis test as a decision problem. This means that we can verify the evidence in favor of the null hypothesis and not only rejecting it, as well as take decisions that minimize the expected loss. Second, because of the nonparametric near-ignorance prior, the IDP rank-sum test allows us to start the hypothesis test with very weak prior assumptions, much in the direction of letting data speak for themselves. From a computational point of view, we will show that posterior inferences from the IDP can be derived by computing lower and upper bounds of expectations w.r.t. the class of DPs $\mathscr{T}$ and that, for certain inference, these lower and upper bounds can be computed in closed-form (e.g., mean and variance of $P(X \leq Y)$). When no closed form expression exists, these bounds can be computed by a simple Monte Carlo sampling from two Dirichlet distributions. This means that we do not need to use "stick breaking" Sethuraman [21] or other sampling approaches specific for DP. This computational advantage is an effect of our prior near-ignorance model.

In our view, the IDP rank-sum test appears to be a natural way to complete the work of Ferguson [1], who first showed the connection between the expectation of $P(X \leq Y)$ w.r.t. the DP and the Mann-Whitney statistic: it develops a Bayesian nonparametric near-ignorance-prior test for the probability that $X \leq Y$, which is computationally efficient and that, also for this reason, provides an effective practical alternative to the MWW test.

Note that, although the IDP test shares several similarities with a standard Bayesian approach, at the same time it embodies a significant change of paradigm when it comes to take decisions. In fact the IDP rank-sum test has the advantage of producing an *indeterminate* outcome when the decision is *prior-dependent*. In other words, the IDP test suspends the judgment (which can be translated as "I do not know whether Y is better than X") when the option that minimizes the expected loss changes depending on the DP base measure we focus on. Therefore, the IDP-based test is *robust* in the sense that it provides a determinate decision only when all the DPs, in the class the IDP represents, agree on the same decision. We will show that the number of indeterminate instances decreases as the evidence accumulates and thus that the IDP-based test is always asymptotically consistent for $P(X \leq Y)$. This is not always true for the MWW test, even though the MWW test is commonly employed as a test about $P(X \leq Y)$.

Finally, we will compare our IDP test with the MWW test and the DP-based test obtained as the prior strength goes to zero (called BB-DP test). We empirically show on several different case studies that when the IDP test is indeterminate, the MWW and BB-IDP tests are virtually behaving as random guessers. For a sample size of 20 observations, the percentage of these instances can reach almost 20%. We regard this surprising result as an important finding, with practical consequences in hypothesis testing. Assume that we are trying to compare the effects of two medical treatments ("Y is better than X") and that, given the available data, the IDP test is indeterminate. In such a situation the MWW test (or the BB-IDP test) always issues a determinate response (for instance, "I can tell that Y is better than X"), but it turns out that its response is virtually random (like if we were tossing a coin). In these cases by using MWW we would choose treatment Y, but this decision would not be reliable. In fact in these instances the MWW test could randomly return the other hypothesis ("it is not true that Y is better than X"). On the other side, the IDP test acknowledges the impossibility of making a decision in these cases. Thus, by saying "I do not know", the IDP test provides a richer information to the analyst. The analyst could for instance use this information to collect more data.

A desirable test should have a low Type I error, high power, but also high replicability. The replicability is the probability that the same conclusion is achieved in two experiments involving the same pair of treatments (i.e., the null hypothesis is accepted or rejected in both cases). Since the response of the MWW test (or of the BB-IDP test) is virtually random when the IDP is indeterminate, it is then clear that a sharp drop of replicability affects the MWW test (or the BB-IDP test) when the IDP test becomes indeterminate. Therefore, one of the advantages of the IDP test w.r.t. the MWW test (or the BB-IDP test) is the higher replicability. This has also been observed for other (imprecise) robust models, see in particular Coolen and Bin Himd [22], Benavoli et al. [23]. Finally, note that IDP tests can be used in many other applications beside the

medical one (i.e., not only for comparing two treatments), For instance, we have implemented an IDP version of the Wilcoxon signed rank sum test and used it to compare the performance of classifiers (or algorithms, more in general), see Benavoli et al. [23]. We have extended the IDP rank-sum test to account for censored data, which are common in reliability and survival analysis. R and Matlab codes of the IDP rank-sum test and these other tests are freely available at `http://ipg.idsia.ch/software/IDP.php`.

2. Dirichlet process

The Dirichlet process was developed by Ferguson [1] as a probability distribution on the space of probability distributions. Let X be a standard Borel space with Borel σ-field $\mathscr{B}_{\mathrm{X}}$ and P be the space of probability measures on $(\mathrm{X},\mathscr{B}_{\mathrm{X}})$ equipped with the weak topology and the corresponding Borel σ-field $\mathscr{B}_{\mathrm{P}}$. Let M be the class of all probability measures on $(\mathrm{P},\mathscr{B}_{\mathrm{P}})$. We call the elements $\mu \in \mathrm{M}$ nonparametric priors.

An element of M is called a Dirichlet process distribution $\mathscr{D}(\alpha)$ with base measure α if for every finite measurable partition $B_1,\ldots,B_m$ of X, the vector $(P(B_1),\ldots,P(B_m))$ has a Dirichlet distribution with parameters $(\alpha(B_1),\ldots,\alpha(B_m))$, where $\alpha(\cdot)$ is a finite positive Borel measure on X. Consider the partition $B_1 = A$ and $B_2 = A^c = \mathrm{X}\backslash A$ for some measurable set $A \in \mathrm{X}$, then if $P \sim \mathscr{D}(\alpha)$ from the definition of the DP we have that $(P(A),P(A^c)) \sim Dir(\alpha(A),\alpha(\mathrm{X})-\alpha(A))$, which is a Beta distribution. From the moments of the Beta distribution, we can thus derive that:

$$\mathscr{E}[P(A)] = \frac{\alpha(A)}{\alpha(\mathrm{X})},\ \mathscr{E}[(P(A)-\mathscr{E}[P(A)])^2] = \frac{\alpha(A)(\alpha(\mathrm{X})-\alpha(A))}{(\alpha(\mathrm{X})^2(\alpha(\mathrm{X})+1))}, \tag{1}$$

where we have used the calligraphic letter $\mathscr{E}$ to denote expectation w.r.t. the Dirichlet process. This shows that the normalized measure $\alpha(\cdot)/\alpha(\mathrm{X})$ of the DP reflects the prior expectation of P, while the scaling parameter $\alpha(\mathrm{X})$ controls how much P is allowed to deviate from its mean $\alpha(\cdot)/\alpha(\mathrm{X})$. Let $s = \alpha(\mathrm{X})$ stand for the total mass of $\alpha(\cdot)$ and $\alpha^*(\cdot) = \alpha(\cdot)/s$ stand for the probability measure obtained by normalizing $\alpha(\cdot)$. If $P \sim \mathscr{D}(\alpha)$, we shall also describe this by saying $P \sim Dp(s,\alpha^*)$ or, if $\mathrm{X} = \mathrm{R}$, $P \sim Dp(s,G_0)$, where G_0 stands for the cumulative distribution function of α^*.

Let $P \sim Dp(s,\alpha^*)$ and f be a real-valued bounded function defined on $(\mathrm{X},\mathscr{B})$. Then the expectation with respect to the Dirichlet process of $E[f]$ is

$$\mathscr{E}\big[E(f)\big] = \mathscr{E}\left[\int f dP\right] = \int f d\mathscr{E}[P] = \int f d\alpha^*. \tag{2}$$

One of the most remarkable properties of the DP priors is that the posterior distribution of P is again a DP. Let $X_1,\ldots,X_n$ be an independent and identically distributed sample from P and $P \sim Dp(s,\alpha^*)$, then the posterior distribution of P given the observations is

$$P|X_1,\ldots,X_n \sim Dp\left(s+n, \frac{s}{s+n}\alpha^* + \frac{1}{s+n}\sum_{i=1}^{n}\delta_{X_i}\right), \tag{3}$$

where δ_{X_i} is an atomic probability measure centered at X_i. This means that the Dirichlet process satisfies a property of conjugacy, in the sense that the posterior for P is again a Dirichlet process with updated unnormalized base measure $\alpha + \sum_{i=1}^{n}\delta_{X_i}$. From (3) and (1)–(2), we can easily derive the posterior mean and variance of $P(A)$ and, respectively, posterior expectation of f. Hereafter we list some useful properties of the DP that will be used in the sequel (see Ghosh and Ramamoorthi [24, Ch. 3]).

(a) In case $\mathrm{X} = \mathrm{R}$, since P is completely defined by its cumulative distribution function F, a-priori we say $F \sim Dp(s, G_0)$ and a posteriori we can rewrite (3) as follows:

$$F|X_1,\ldots,X_n \sim Dp\left(s+n, \frac{s}{s+n}G_0 + \frac{n}{s+n}\frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)}\right), \tag{4}$$

where I is the indicator function and $\frac{1}{n}\sum_{i=1}^{n} I_{[X_i,\infty)}$ is the empirical cumulative distribution.

(b) Consider an element $\mu \in \mathrm{M}$ which puts all its mass at the probability measure $P = \delta_x$ for some $x \in \mathrm{X}$. This can also be modeled as $Dp(s, \delta_x)$ for each $s > 0$.

(c) Assume that $P_1 \sim Dp(s_1, \alpha_1^*)$, $P_2 \sim Dp(s_2, \alpha_2^*)$, $(w_1, w_2) \sim Dir(s_1, s_2)$ and P_1, P_2, (w_1, w_2) are independent, then [24, Sec. 3.1.1]:

$$w_1P_1 + w_2P_2 \sim Dp\left(s_1+s_2, \frac{s_1}{s_1+s_2}\alpha_1^* + \frac{s_2}{s_1+s_2}\alpha_2^*\right). \tag{5}$$

(d) Let P_x have distribution $Dp(s+n, \frac{s}{s+n}\alpha^* + \frac{1}{s+n}\sum_{i=1}^{n}\delta_{X_i})$. We can write

$$P_x = w_0P + \sum_{i=1}^{n} w_i\delta_{X_i}, \tag{6}$$

where $\sum_{i=0}^{n} w_i = 1$, $(w_0, w_1, \ldots, w_n) \sim Dir(s, 1, \ldots, 1)$ and $P \sim Dp(s, \alpha^*)$. This follows from (b)-(c).

3. Prior ignorance

How should we choose the prior parameters (s, α^*) of the DP, in particular the infinite-dimensional α^*, in case of lack of prior information? To address this issue, the only prior that has been proposed so far is the limiting DP obtained for $s \to 0$, which has been introduced under the name of Bayesian Bootstrap (BB) by Rubin [13]; in fact it can be proven that the BB is asymptotically equivalent (see Lo [25] and Weng [26]) to the frequentist bootstrap introduced by Efron [27].

The BB has been criticized on diverse grounds. From an a-priori point of view, the main criticism is that taking $s \to 0$ is far from leading to a noninformative prior. Sethuraman and Tiwari [28] have shown that for $s \to 0$ a measure sampled from the DP is a degenerated (atomic) measure centered on X_0, with X_0 distributed according to α^*. As a further consequence, from an a-posteriori point of view, this choice for the prior gives zero probability to the event that a future observation is different from the previously observed data. Rubin [13] reports the following extreme example. Consider the probability that $X > C$ where C is a value larger than the largest observed value of X, i.e., $X_{(n)}$. The standard BB and bootstrap methods estimate such probability to be 0 with zero variance, which is untenable if X can assume different values from the n previously observed. Rubin also remarks that one should expect a probability that X is greater than or equal to $X_{(n)}$ of about $1/(n+1)$. This shows that a Dirichlet prior with $s \to 0$ implies definitely a very strong (and not always reasonable) information about P, and hence it cannot be considered a noninformative prior. On the other side, if we choose a DP prior with $s > 0$, the inferences provided by this model will be sensitive to the choice of the normalized measure α^*. If, for example, we decide to assign a "tail" probability of $1/(n+1)$ to $X > X_{(n)}$, in agreement with Rubin's intuition, the inferences will be different if we assume that the tail probability is concentrated on $X_{(n)}$ or if we assume that it is spread from $X_{(n)}$ to a very large value of X.

To answer to the initial question of this section, we propose the imprecise Dirichlet process (IDP). The main characteristic of the IDP is that it does not require any choice of the normalized measure α^*, it is a prior near-ignorance model and solves the issues of the BB. Before introducing the IDP, it is worth to explain what is a prior near-ignorance model with the example of a parametric model [16, Sec. 5.3.1].

Example 1. Let A be the event that a particular thumbtack lands pin-up at the next toss. Your information is that there have been m occurrences of pin up in n previous tosses. Using a Bernoulli model, the likelihood function generated by observing m successes in n trials is then proportional to $\theta^n(1-\theta)^{n-m}$ where θ is the chance of pin-up. To complete the model, we need to specify prior beliefs concerning the unknown chance θ. We can use a conjugate Beta prior $p(\theta) = Be(\theta;\alpha,\beta)$, where $\alpha,\beta > 0$ are the prior parameters of the Beta density. A-posteriori we have that $p(\theta|m,n) = Be(\theta;\alpha+m,\beta+n-m)$. Thus, the prior and posterior probabilities of A are:

$$P(A) = E[\theta] = t, \quad P(A|m,n) = E[\theta|m,n] = \frac{st+m}{s+n},$$

where $s = \alpha+\beta$ is the prior strength and $t = \alpha/(\alpha+\beta)$ the prior mean. The problem is how to choose the parameters s,t in case of lack of prior information. Walley [16, Ch. 5] proposes to use a prior near-ignorance model. A near-ignorance prior model for this example is any set of priors which generates vacuous prior probabilities for the event of interest A, i.e.,

$$\underline{P}(A) = 0, \ \ \overline{P}(A) = 1,$$

where $\underline{P},\overline{P}$ are lower and upper bounds for $P(A)$. These vacuous probabilities reflect a complete absence of prior information concerning A. For the Beta prior, since $P(A) = E[\theta] = t$, the class of priors is simply:

$$p(\theta) \in \{Be(\theta;st,s(1-t)) : \ \ 0 < t < 1\},$$

for some fixed $s > 0$, i.e., this is the set of priors obtained by considering all the Beta densities whose mean parameter t is free to span the interval $(0,1)$. Posterior inferences from this model are derived by computing lower and upper posterior bounds; in the case of event A these bounds are:

$$\underline{P}(A|m,n) = \frac{m}{s+n}, \ \ \overline{P}(A|m,n) = \frac{s+m}{s+n},$$

where the lower is obtained for $t \to 0$ and the upper for $t \to 1$. We point the reader to Walley [29] for more details about this model and to Benavoli and Zaffalon [30] for an extension of near-ignorance to one-parameter exponential families.

3.1 *Imprecise Dirichlet process*

Before introducing the IDP, we give a formal definition of (nonparametric) prior ignorance for predictive inferences. Let f be a real-valued bounded function on X, we call $E[f] = \int f dP$ a predictive inference about X and $P \in \mathrm{P}$. Let $\mu \in \mathrm{M}$ be a nonparametric prior on P and $\mathscr{E}_\mu[E(P)]$ the expectation of $E[f]$ w.r.t. μ.

Definition 1. *A class of nonparametric priors $\mathscr{T} \subset \mathrm{M}$ is called a prior ignorance model for predictive inferences about X, if for any real-valued bounded function f on X it satisfies:*

$$\underline{\mathscr{E}}[E(f)] = \inf_{\mu\in\mathscr{T}} \mathscr{E}_\mu[E(f)] = \inf f, \ \ \overline{\mathscr{E}}[E(f)] = \sup_{\mu\in\mathscr{T}} \mathscr{E}_\mu[E(f)] = \sup f, \tag{7}$$

where $\underline{\mathscr{E}}[E(f)]$ *and* $\overline{\mathscr{E}}[E(f)]$ *denote respectively the lower and upper bound of* $\mathscr{E}_\mu[E(P)]$ *calculated w.r.t. the class* $\mathscr{T}$.

From (7) it can be observed that the range of $\mathscr{E}_\mu[E(f)]$ under the class $\mathscr{T}$ is the same as the original range of f. In other words, by specifying the class $\mathscr{T}$, we are not giving any information on the value of the expectation of f. This means that the class $\mathscr{T}$ behaves as a vacuous model. We are now ready to define the IDP.

Definition 2. IDP. *We call prior imprecise DP the following class of DPs:*

$$\mathscr{T} = \{Dp(s,\alpha^*) : \ \alpha^* \in \mathrm{P}\}. \tag{8}$$

The IDP is the class of DPs obtained for a fixed $s > 0$ and by letting the normalized measure α^* to vary in the set of all probability measures P on $(\mathrm{X}, \mathscr{B}_\mathrm{X})$.

Theorem 1. *The IDP is a model of prior ignorance for all predictive inferences about X, i.e., for any real-valued bounded function* f *on* X *it satisfies:*

$$\underline{\mathscr{E}}[E(f)] = \inf f, \ \ \overline{\mathscr{E}}[E(f)] = \sup f, \tag{9}$$

where $\underline{\mathscr{E}}[E(f)]$ *and* $\overline{\mathscr{E}}[E(f)]$ *denote respectively the lower and upper bound of* $\mathscr{E}[E(f)]$ *defined in (2) calculated w.r.t. the class of DPs (8).*

The proofs of this and the next theorems are in the Appendix. To show that the IDP is a model of prior ignorance, consider for instance the indicator function $f = I_A$ for some $A \subseteq \mathrm{X}$. Since $E[I_A] = P(A)$, from (2) we have that $\mathscr{E}[P(A)] = \int I_A d\alpha^*$. Then if we choose $\alpha^* = \delta_{x_l}$ with $x_l \notin A$ and, respectively, $\alpha^* = \delta_{x_u}$ with $x_u \in A$:

$$\underline{\mathscr{E}}[P(A)] = \int I_A d\delta_{x_l} = \min I_A = 0, \ \ \overline{\mathscr{E}}[P(A)] = \int I_A d\delta_{x_u} = \max I_A = 1, \tag{10}$$

where $\underline{\mathscr{E}}[P(A)]$ and $\overline{\mathscr{E}}[P(A)]$ are the lower and upper bounds for $\mathscr{E}\big[P(A)\big]$. This is a condition of prior ignorance for $P(A)$, since we are saying that the only information about $P(A)$ is that $0 \leq P(A) \leq 1$. The lower and upper bounds are obtained from the degenerate DPs $Dp(s,\delta_{x_l})$ and $Dp(s,\delta_{x_u})$, which belong to the class (8). Note that, although the lower and upper bounds are obtained by degenerate DPs, to obtain these bounds we are considering all possible $Dp(s,\alpha^*)$ with $\alpha^* \in \mathrm{P}$ (even the ones with continuous probability measures α^*).

Theorem 2. Posterior inference. *Let* $X_1,\ldots,X_n$ *be i.i.d. samples from* P *and* $P \sim Dp(s,\alpha^*)$. *Then for any real-valued bounded function* f *on* X, *the lower and upper bounds of* $\mathscr{E}[E(f)|X_1,\ldots,X_n]$ *under the IDP model in (8) are:*

$$\begin{aligned} \underline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] &= \frac{s}{s+n}\inf f + \frac{n}{s+n}S_n(f), \\ \overline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] &= \frac{s}{s+n}\sup f + \frac{n}{s+n}S_n(f), \end{aligned} \tag{11}$$

where $S_n(f) = \frac{\sum_{i=1}^n f(X_i)}{n}$.

A-posteriori the IDP does not satisfy anymore the prior ignorance property (9). This means that learning from data takes place under the IDP. In fact let $S(f)$ be equal to $\lim_{n\to\infty} S_n(f)$, a-posteriori for $n \to \infty$ we have that:

$$\underline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big], \overline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big] \to S(f), \tag{12}$$

i.e., the lower and upper bounds of the posterior expectations converge to $S(f)$, which only

depends on data. In other words, the effect of prior ignorance vanishes asymptotically:

$$\overline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big]-\underline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big]=\frac{s}{s+n}(\sup f-\inf f)\to 0,$$

for any finite s. To define the IDP, the modeler has only to choose s. This explains the meaning of the adjective *near* in prior near-ignorance, because the IDP requires by the modeller the elicitation of a parameter. However, this is a simple elicitation problem for a nonparametric prior, since we only have to choose the value of a positive scalar (there are not infinitely dimensional parameters left). Section 4 gives some guidelines for the choice of this parameter.

Observe that IDP solves the two main drawbacks of Bayesian Bootstrap. From the a-priori point of view, we have shown in (9) that the IDP is a model of prior ignorance for predictive inferences. Moreover, the prior distributions considered can assign a non-null probability to unobserved values of X. Then, considering Rubin's example about the probability that X is greater than or equal to $X_{(n)}$, which is obtained as the expectation of $f=I_{[C,\infty)}$ with $C>X_{(n)}$, from (11) we have a-posteriori that $\underline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big]=0$ and $\overline{\mathscr{E}}\big[E(f)|X_1,\ldots,X_n\big]=\frac{s}{s+n}$. The upper expectation is greater than zero and, for $s=1$, it is equal to $1/(1+n)$. This result is obtained without specifying how the probability of $1/(1+n)$ is spread between the values $X>X_{(n)}$, and thus it is insensitive to the model specification of tail probabilities. Note that the IDP reduces to the imprecise Dirichlet model proposed by Walley [29], see also Bernard [31], de Cooman et al. [32]), when we limit ourselves to consider a finite measurable partition $B_1,\ldots,B_m$ of X. In this case, the set of priors $\{Dp(s,\alpha^*),\ \alpha^*\in \mathrm{P}\}$, reduces to a set of Dirichlet distributions with parameters $(s\alpha^*(B_1),\ldots,s\alpha^*(B_m))$.

4. An application to hypothesis testing

Hypothesis testing is an important application of nonparametric statistics. Recently there has been an increasing interest in the development of Bayesian nonparametric procedures for hypothesis testing. For instance Bayesian nonparametric approaches to the two-sample problem have been proposed using Dirichlet process mixture models or (coupling-optional) Polya trees priors by Borgwardt and Ghahramani [8], Holmes et al. [9], Ma and Wong [10], Chen and Hanson [11] and Martin and Tokdar [12]. Although prior near-ignorance may also be useful in these models and in the two-sample problem, we do not follow this avenue in the present study. Our focus is instead the hypothesis test $P(X\leq Y)\lesseqgtr P(X>Y)$ (equivalently $P(X\leq Y)\lesseqgtr 0.5$), given independent samples of sizes n_1 and n_2 from two populations. This problem arises, for example, if one wishes to compare the response X of a population with respect to the response Y of a different population in order to establish whether the two populations perform equally well or one population has generally "better" responses than the other.

The nonparametric test traditionally applied in such situations is the Mann-Whitney-Wilcoxon (MWW) rank-sum test. The null hypothesis of the MWW rank-sum test is that the two populations are equal, that is, they come from the same distribution $F_X(x)=F_Y(x)$. Let $X^{n_1}=\{X_1,\ldots,X_{n_1}\}$ and $Y^{n_2}=\{Y_1,\ldots,Y_{n_2}\}$ be two sequences of observations from the two populations. The MWW test is based on the fact that, if the two populations have the same distribution, the distribution of the linear rank statistic

$$U=\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}I_{[X_i,\infty)}(Y_j), \tag{13}$$

can be computed by considering all the possible arrangements of the observations in X^{n_1} and Y^{n_2}. At the increase of n_1 and n_2, this distribution converges to a Normal distribution with mean

and variance given by

$$E\left[\frac{U}{n_1 n_2}\right] = \frac{1}{2}, \;\; Var\left[\frac{U}{n_1 n_2}\right] = \frac{n_1 + n_2}{12 n_1 n_2}. \tag{14}$$

It is worth to stress that $F_X(x) = F_Y(x)$ implies $P(X \leq Y) = 0.5$ (i.e., it is not true that Y is better than X) but not vice versa. Thus, the MWW test cannot be used in general as a test for $P(X \leq Y)$. This limitation of the test is due to the choice of U statistic functions as the test statistic instead of estimator and the need of any frequentist method to specify the distribution of the statistic under the null hypothesis. The null hypothesis $F_X(x) = F_Y(x)$ is thus selected to be able to compute the distribution of the statistic, although, in practice, one is interested in a much weaker hypothesis to test $P(X \leq Y)$ (see Fay and Proschan [33] for a detailed discussion). To overcome this issue of the MWW test, it is often common to assume a location-shift model, which states that the two populations can only differ in locations: $F_Y(y) = F_X(y - \Delta)$. The goal is then to test the hypothesis that there is no treatment effect $\Delta = 0$ ($P(X \leq Y) = 0.5$) versus the alternative $\Delta > 0$ ($P(X \leq Y) > 0.5$) or $\Delta < 0$ ($P(X \leq Y) < 0.5$). Under this assumption, the MWW test can be interpreted as a Hodges and Lehmann [34] estimator. On the other side, the Bayesian approach provides the posterior distribution of $P(X \leq Y)$, which can be used to compute the probability of any hypothesis of interest. Therefore, we are not limited in the choice of the null hypothesis. Moreover, the MWW test is affected by all the drawbacks which characterize null hypothesis significance tests (NHST). Such tests "allow one either to reject the null hypothesis or to fail to reject it, but they do *not* provide any measure of evidence for the null hypothesis" (Raftery [35]). This prevents associating a cost to Type I and Type II errors and taking decisions by minimizing the expected loss. Instead, decision are taken on the basis of the chosen significance γ, namely the probability of rejecting the null hypothesis when it is true. In principle, one should balance significance and power of the test. Yet, a principled way of doing this is lacking (Kruschke [36]). Hence, decisions are simply taken by setting γ =0.01 or 0.05, without considering the probability of Type II errors. Moreover, the p-value and thus the outcome of the test depend on the intention of the person who has collected the data (Kruschke [36], Goodman [37]). The Bayesian approach to decision making allows basing the decisions on the value of the expected loss, whose practical meaning is much more intuitive. For example, the hypothesis test:

$$\underbrace{P(X \leq Y) \leq P(X > Y)}_{P(X \leq Y) \leq 0.5} \quad \textit{vs.} \quad \underbrace{P(X \leq Y) > P(X > Y)}_{P(X \leq Y) > 0.5}$$

can be performed in a Bayesian way in two steps. First we define a loss function

$$L(P, a) = \begin{cases} K_0 I_{\{P(X \leq Y) > 0.5\}} & \text{if } a = 0, \\ K_1 I_{\{P(X \leq Y) \leq 0.5\}} & \text{if } a = 1. \end{cases} \tag{15}$$

The first row gives the loss we incur by taking the action $a = 0$ (i.e., declaring that $P(X \leq Y) \leq 0.5$) when actually $P(X \leq Y) > 0.5$, while the second row gives the loss we incur by taking the action $a = 1$ (i.e., declaring that $P(X \leq Y) > 0.5$) when actually $P(X \leq Y) \leq 0.5$. Second, we compute the expected value of this loss:

$$\mathscr{E}[L(P, a)] = \begin{cases} K_0 \mathscr{P}[P(X \leq Y) > 0.5] & \text{if } a = 0, \\ K_1 \mathscr{P}[P(X \leq Y) \leq 0.5] & \text{if } a = 1, \end{cases} \tag{16}$$

where we have used the calligraphic letter $\mathscr{P}$ to denote the probability w.r.t. the DP priors for

F_X and F_Y. Thus, we choose $a = 1$ if

$$K_1 \mathscr{P}[P(X \leq Y) \leq 0.5] \leq K_0 \mathscr{P}[P(X \leq Y) > 0.5] \Rightarrow \mathscr{P}[P(X \leq Y) > 0.5] > \frac{K_1}{K_1 + K_0}, \quad (17)$$

or $a = 0$ otherwise. When the above inequality is satisfied, we can declare that $P(X \leq Y) > 0.5$ with probability $\frac{K_1}{K_1+K_0} = 1 - \gamma$. For the choice $\gamma = 0.05$, the MWW test and DP are closely matched. However, in the Bayesian setting, $\gamma = 0.05$ plays no special role and other choices are possible.

Finally, based on the imprecise DP model developed in this paper, we can perform a Bayesian nonparametric test that, besides overcoming the limitation of the frequentist test described above, is based on extremely weak prior assumptions, and easy to elicit, since it requires only to choose the strength s of the DP instead of its infinite-dimensional parameter α. When using the IDP set of priors, we consider for F_X and F_Y all the possible DP priors with strength lower than or equal to s (since all inferences obtained for $s' < s$ are encompassed by those obtained for s, see Walley [29]). All these priors give a posterior probability $\mathscr{P}[P(X \leq Y) > 0.5]$ included between the lower and upper bounds $\underline{\mathscr{P}}[P(X \leq Y) > 0.5]$ and $\overline{\mathscr{P}}[P(X \leq Y) > 0.5]$. Thus, according to the decision rule in (17) for some $\gamma = \frac{K_0}{K_0+K_1}$, we verify if

$$\underline{\mathscr{P}}\left[P(X \leq Y) > 0.5 | X^{n_1}, Y^{n_2}\right] > 1 - \gamma, \;\; \overline{\mathscr{P}}\left[P(X \leq Y) > 0.5 | X^{n_1}, Y^{n_2}\right] > 1 - \gamma,$$

and then proceed as follows:

(1) if both the inequalities are satisfied we can declare that $P(X \leq Y)$ is greater than 0.5 with probability larger than $1 - \gamma$;
(2) if only one of the inequality is satisfied (which has necessarily to be the one for the upper), we are in an indeterminate situation, i.e., we cannot decide;
(3) if both are not satisfied, we can declare that the probability that $P(X \leq Y)$ is greater than 0.5 is lower than the desired probability of $1 - \gamma$.

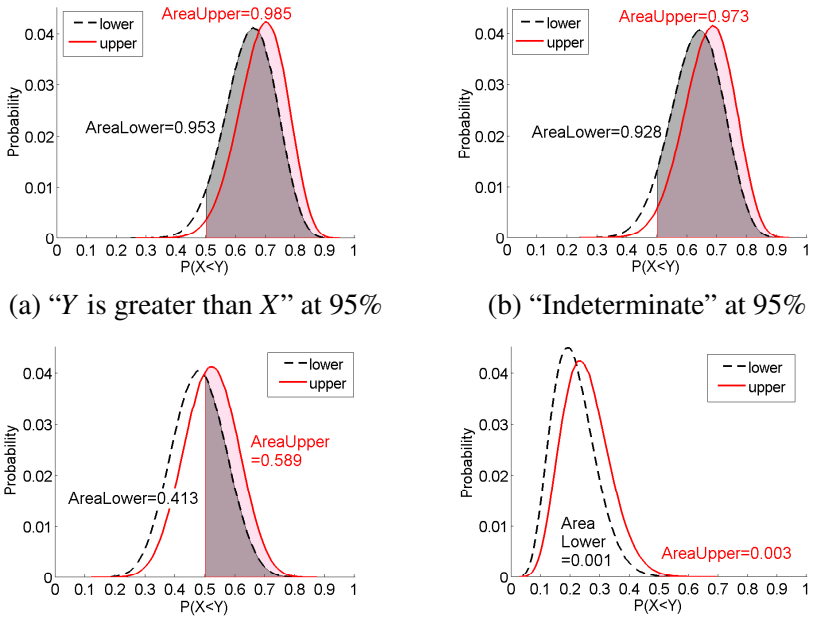
When our model of prior ignorance returns an indeterminate decision, it means that the evidence from the observations is not enough to declare either that the probability of the hypothesis being true is larger or smaller than the desired value $1 - \gamma$; more measurements are necessary to take a decision.

The three cases are respectively depicted in Figure 1. Observe that the posterior distributions of $P(X \leq Y)$, from which the lower and upper probabilities above are derived, give us much more information than the simple result of the hypothesis test. In particular we can derive the posterior lower and upper probabilities of $P(X \leq Y) < 0.5$. For instance, from both Figure 1 (c) and (d) we can see that Y is not greater than X at 95%, but only in Figure (d) it is evident that X is greater than Y at 95%. While in the case shown in Figure 1 (b), we can say neither that Y is greater than X nor that X is greater than Y. (To distinguish these two cases it would be more appropriate to perform a "two-sided" hypothesis test.)

In the next section we prove that the IDP is a model of prior ignorance for $P(X \leq Y)$ and derive the posterior results which are necessary to evaluate $\mathscr{P}[P(X \leq Y) > 0.5]$ and perform the test. Note that, for the moment, we assume that there are no ties between X and Y; we will discuss how to account for the presence of ties in Section 4.3.

4.1 *IDP model for* $P(X \leq Y)$

Let the samples X^{n_1} and Y^{n_2} be drawn, respectively, from F_X and F_Y. As prior for (F_X, F_Y), we assume that $F_X \sim Dp(s_1, G_1)$ and $F_Y \sim Dp(s_2, G_2)$, where $s_1, s_2 \in \mathbb{R}$ and G_1, G_2 are two cumulative distribution functions. Hereafter, to simplify the presentation, we take $s_1 = s_2 = s$.



(a) "Y is greater than X" at 95%

(b) "Indeterminate" at 95%

(c) "Y is not greater than X" at 95%

(d) "Y is not greater than X" at 95%

Figure 1. Four possible results of the hypothesis test. The dark and light filled areas correspond respectively to the lower and upper probabilities of the event $P(X \leq Y) > 0.5$. The numerical values of these lower and upper probabilities are also reported in the figures.

F_X and F_Y are assumed to be independent. The probability $P(X \leq Y)$ is given by $P(X \leq Y) = E[I_{[X,\infty)}(Y)] = \int F_X(y) dF_Y(y)$. As derived by Ferguson [1], by the properties of the Dirichlet process, it follows that a-priori $\mathscr{E}[P(X \leq Y)] = \int G_1(y) dG_2(y)$. It can be shown that the set of priors $\mathscr{T}$ in (8) satisfies the condition of prior ignorance also for $P(X \leq Y)$. In fact, since $\mathscr{E}[P(X \leq Y)] = \int G_1(y) dG_2(y)$, if $G_i \in \mathrm{P}$, we have that

$$\underline{\mathscr{E}}[P(X \leq Y)] = 0, \quad \overline{\mathscr{E}}[P(X \leq Y)] = 1,$$

where the lower (upper) bound is obtained for $dG_1 = \delta_{X_0}$ and $dG_2 = \delta_{Y_0}$ with $X_0 > Y_0$ ($X_0 < Y_0$). Thus, prior ignorance about the mean of $P(X \leq Y)$ is satisfied. Furthermore, let us consider the probability of $P(X \leq Y) < 0.5$ with respect to the Dirichlet process. A-priori, for $dG_1 = \delta_{X_0}$ and $dG_2 = \delta_{Y_0}$ we have that

$$\begin{array}{l} \text{if } X_0 < Y_0, \text{ then } \mathscr{P}[P(X \leq Y) = 1] = 1 \text{ and thus } \underline{\mathscr{P}}[P(X \leq Y) \leq 0.5] = 0 \\ \text{if } X_0 > Y_0, \text{ then } \mathscr{P}[P(X \leq Y) = 0] = 1 \text{ and thus } \overline{\mathscr{P}}[P(X \leq Y) \leq 0.5] = 1. \end{array}$$

A similar reasoning leads to $\underline{\mathscr{P}}[P(X \leq Y) > 0.5] = 0$, $\overline{\mathscr{P}}[P(X \leq Y) > 0.5] = 1$, thus, prior ignorance about the hypothesis $P(X \leq Y) > 0.5$ is also satisfied. Given the two sequences of measurements, a-posteriori one has:

$$\mathscr{E}[P(X \leq Y) | X^{n_1}, Y^{n_2}] = \int G^*_{n_1}(y) dG^*_{n_2}(y),$$

with $G^*_{n_i} = \frac{s}{s+n_i} G_i + \frac{1}{s+n_i} \sum_{j=1}^{n_i} I_{[Z_j,\infty)}$, where $Z_j = X_j$ for $i = 1$ and $Z_j = Y_j$ for $i = 2$. It follows that:

$$\begin{aligned} \mathscr{E}[P(X \leq Y)|X^{n_1}, Y^{n_2}] = \frac{s}{s+n_1}\frac{s}{s+n_2}\int G_1(y)dG_2(y) + \frac{n_1}{s+n_1}\frac{s}{s+n_2}\frac{1}{n_1}\sum_{j=1}^{n_1}(1-G_2(X_j^-)) \\ + \frac{s}{s+n_1}\frac{n_2}{s+n_2}\frac{1}{n_2}\sum_{j=1}^{n_2} G_1(Y_j) + \frac{n_1}{s+n_1}\frac{n_2}{s+n_2}\frac{U}{n_1 n_2}, \end{aligned} \tag{18}$$

where $1 - G_2(X^-) = \int I_{[X,\infty)} dG_2$. Then, the lower and upper posterior bounds of the posterior expectations of $P(X \leq Y)$ given the set of priors $\mathscr{T}$ are:

$$\begin{aligned} \underline{\mathscr{E}}[P(X \leq Y)|X^{n_1}, Y^{n_2}] &= \frac{U}{(s+n_1)(s+n_2)}, \\ \overline{\mathscr{E}}[P(X \leq Y)|X^{n_1}, Y^{n_2}] &= \frac{U}{(s+n_1)(s+n_2)} + \frac{s(s+n_1+n_2)}{(s+n_1)(s+n_2)}, \end{aligned} \tag{19}$$

obtained in correspondence of the extreme distributions $dG_1 \to \delta_{X_0}$, $dG_2 \to \delta_{Y_0}$, with $X_0 > \max(\{Y_0, \ldots, Y_{n_1}\})$, $Y_0 < \min(\{X_0, \ldots, X_{n_2}\})$ (lower) and $X_0 < \min(\{Y_0, \ldots, Y_{n_1}\})$, $Y_0 > \max(\{X_0, \ldots, X_{n_2}\})$ (upper) and U is given in (13). The posterior probability distribution of $P(X \leq Y)$ w.r.t. the Dirichlet processes, which is used to perform the Bayesian test of the difference between the two populations, is, in general, computed numerically (Monte Carlo sampling) by using the stick-breaking construction of the Dirichlet process. We will show in the remaining part of this section that, in correspondence to the discrete priors that give the upper and lower bounds of the posterior distributions of $P(X \leq Y)$, a more efficient procedure can be devised. Consider the limiting posteriors that give the posterior lower and upper expectations in (19):

$$G_{n_i}(y) = \frac{s}{s+n_i} I_{[Z_0,\infty)} + \frac{1}{s+n_i}\sum_{j=1}^{n_i} I_{[Z_j,\infty)}, \tag{20}$$

where the lower bound is obtained with $Z_0 = X_0 > \max(\{Y_0, \ldots, Y_{n_1}\})$ for $i = 1$ and $Z_0 = Y_0 < \min(\{X_0, \ldots, X_{n_2}\})$ for $i = 2$, and the upper bound with $Z_0 = X_0 > \max(\{Y_0, \ldots, Y_{n_1}\})$ for $i = 1$, and $Z_0 = Y_0 < \min(\{X_0, \ldots, X_{n_2}\})$ for $i = 2$.

Lemma 1. *A cumulative distribution function F_{n_i} sampled from the Dirichlet process $Dp(s + n_i, G_{n_i})$ with base probability distribution G_{n_i} as that defined in (20) is given by:*

$$F_{n_i} = w_{i0} I_{[Z_0,\infty)} + \sum_{j=1}^{n_i} w_{ij} I_{[Z_j,\infty)}, \tag{21}$$

where $w_{i\cdot} = (w_{i0}, w_{i1}, \ldots, w_{in_i}) \sim Dir(s, \overbrace{1, \ldots, 1}^{n_i})$.

Lemma 1 states that any distribution F_{n_i} sampled from $DP(s + n_i, G_{n_i})$ has the form (21). Since the probability density function relative to F_{n_i}, i.e., $w_{i0}\delta_{Z_0} + \sum_{j=1}^{n_i} w_{ij}\delta_{Z_j}$, has a discrete support, we do not need stick-breaking to sample from a Dirichlet Process when its base measure is discrete; we only need to sample the weights $(w_{i0}, w_{i1}, \ldots, w_{in_i})$ in (21) from the Dirichlet distribution with parameters $(s, 1, \ldots, 1)$. Moreover, if the distributions of X and Y are DPs with discrete base measures G_{n_1} and G_{n_2}, each predictive inference $E[f(X,Y)]$ can be written as a function of the weights $(w_{i0}, w_{i1}, \ldots, w_{in_i})$ only, and the relative distribution can be derived from the (Dirichlet) distribution of these weights. Using this result and the fact that the posteriors that give lower and upper bounds for $\mathscr{P}\left[P(X \leq Y) > c|X^{n_1}, Y^{n_2}\right]$ have the discrete base measures

(20), we can obtain the following result.

Theorem 3. *For any $c \in [0,1]$, it holds that*

$$\underline{\mathscr{P}}\left[P(X \leq Y) > c | X^{n_1}, Y^{n_2}\right] = P_D\left[g(w_{1\cdot}, w_{2\cdot}, X^{n_1}, Y^{n_2}) > c\right], \tag{22}$$

with

$$g(w_{1\cdot}, w_{2\cdot}, X^{n_1}, Y^{n_2}) = \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} w_{1j} w_{2k} I_{(X_j,\infty)}(Y_k)$$

where $(w_{i0}, w_{i1}, \ldots, w_{in_i}) \sim Dir(s, \overbrace{1, \ldots, 1}^{n_i})$ for $i = 1,2$ and the probability P_D is computed w.r.t. the Dirichlet distributions of $w_{1\cdot}$ and $w_{2\cdot}$. The mean and variance of $g(w_{1\cdot}, w_{2\cdot}, X^{n_1}, Y^{n_2})$ are:

$$\mu = E_W[W]^T A E_V[V], \quad \sigma^2 = trace[A^T E_W[WW^T] A E_V[VV^T]] - \mu^2, \tag{23}$$

where $W = [w_{11}, \ldots, w_{1n_1}]^T$, $V = [w_{21}, \ldots, w_{2n_2}]^T$ and their expectations E_W, E_V are taken w.r.t. the Dirichlet distributions of $w_{1\cdot}$ and $w_{2\cdot}$, $E[WW^T]$ and $E[VV^T]$ are $n_i \times n_i$ square-matrix of elements $e_{jk} = (s+n_i)^{-1}(s+n_i+1)^{-1}(1+I_{\{j\}}(k))$ ($i = 1$ and 2, respectively), and A is an $n_1 \times n_2$ matrix with elements $a_{jk} = I_{(X_j,\infty)}(Y_k)$.

Corollary 1. *For any $c \in [0,1]$, it holds that*

$$\overline{\mathscr{P}}\left[P(X \leq Y) > c | X^{n_1}, Y^{n_2}\right] = P_D\left[g(w_{1\cdot}, w_{2\cdot}, X^{n_1}, Y^{n_2}) > c\right], \tag{24}$$

with

$$g(w_{1\cdot}, w_{2\cdot}, X^{n_1}, Y^{n_2}) = w_{10}w_{20} + w_{10}\sum_{j=1}^{n_2} w_{2j} + w_{20}\sum_{j=1}^{n_1} w_{1j} + \sum_{j=1}^{n_1}\sum_{k=1}^{n_2} w_{1j}w_{2k} I_{(X_j,\infty)}(Y_k),$$

where $(w_{i0}, w_{i1}, \ldots, w_{in_i}) \sim Dir(s, \overbrace{1, \ldots, 1}^{n_i})$ for $i = 1,2$. Consider the augmented vectors $W = [w_{10}, w_{11}, \ldots, w_{1n_1}]^T$, $V = [w_{20}, w_{21}, \ldots, w_{2n_2}]^T$, and the matrix A with elements $a_{jk} = I_{(X_{j-1},\infty)}(Y_{k-1})$ for all $j,k \neq 1$ and $a_{jk} = 1$ if $j = 1$ or $k = 1$. The mean and variance of $g(w_{1\cdot}, w_{2\cdot}, X^{n_1}, Y^{n_2})$ can be computed using the same formulas as in (23), where, this time, $E[WW^T]$ and $E[VV^T]$ are $(n_i+1) \times (n_i+1)$ square-matrices ($i = 1$ and 2, respectively) of elements $e_{jk} = (s+n_i)^{-1}(s+n_i+1)^{-1}\tilde{e}_{jk}$ with $\tilde{e}_{jk} = (1+I_{\{j\}}(k))$ for all $j,k \neq 1$ and $\tilde{e}_{jk} = s(1+sI_{\{j\}}(k))$ if $j = 1$ or $k = 1$.

Theorem 3 and Corollary 1 show that the lower and upper bounds of $\underline{\mathscr{P}}\left[P(X \leq Y) > c | X^{n_1}, Y^{n_2}\right]$ can be computed by Monte Carlo sampling from the Dirichlet distributions of the weight vectors $w_{1\cdot}, w_{2\cdot}$ and, thus, no stick-breaking is necessary.

To perform the hypothesis test, we select $c = 1/2$ and, according to the decision rule (17) for some K_0, K_1, we check if

$$\underline{\mathscr{P}}\left[P(X \leq Y) > \tfrac{1}{2} | X^{n_1}, Y^{n_2}\right] > 1-\gamma, \;\; \overline{\mathscr{P}}\left[P(X \leq Y) > \tfrac{1}{2} | X^{n_1}, Y^{n_2}\right] > 1-\gamma,$$

where $\gamma = \frac{K_0}{K_0+K_1} \in (0,1)$ (e.g., $1-\gamma = 0.95$).

4.2 *Choice of the prior strength s*

The value of s determines how quickly lower and upper posterior expectations converge at the increase of the number of observations. A way to select a value of s is by imposing that the degree of robustness (indeterminacy) $\overline{\mathscr{E}}[P(X \leq Y)|X^{n_1}, Y^{n_2}] - \underline{\mathscr{E}}[P(X \leq Y)|X^{n_1}, Y^{n_2}]$ is reduced to a fraction of its prior value ($\overline{\mathscr{E}}[P(X \leq Y)] - \underline{\mathscr{E}}[P(X \leq Y)] = 1$) after one observation (X_1, Y_1). Imposing a degree of imprecision close to 1 after the first observation increases the probability of an indeterminate outcome of the test, whereas, a value close to 0 makes the test less reliable (in fact the limiting value of 0 corresponds to the BB which will be shown in Section 6 to be less reliable than the IDP). Then, the intermediate value of $1/2$ is a frequent choice in prior-ignorance modeling [17, 29]. Although this is a subjective way to choose the degree of conservativeness (indeterminacy), we will show in Section 6 that it represents a reasonable trade-off between the reliability and indeterminacy of the decision. From (19) for $n_1 = n_2 = 1$, it follows that

$$\overline{\mathscr{E}}[P(X \leq Y)|X_1, Y_1] - \underline{\mathscr{E}}[P(X \leq Y)|X_1, Y_1] = \tfrac{s^2+2s}{(s+1)^2}.$$

Thus, by imposing that,

$$\frac{s^2+2s}{(s+1)^2} = \frac{1}{2},$$

we obtain $s = \sqrt{2}-1$. Observe that the lower and upper probabilities produced by a value of s are always contained in the probability intervals produced by the larger value of s. Then, whenever we are undecided for s_1 we are also for $s_2 > s_1$. Nonetheless as, for large n the distance between the upper and lower probabilities goes to 0, also the indeterminateness goes to zero.

4.3 *Managing ties*

To account for the presence of ties between samples from the two populations ($X_i = Y_j$), the common approach is to test the hypothesis $[P(X < Y) + \frac{1}{2}P(X = Y)] \leq 0.5$ against $[P(X < Y) + \frac{1}{2}P(X = Y)] > 0.5$. Since

$$P(X < Y) + \frac{1}{2}P(X = Y) = E\left[I_{(X,\infty)}(Y) + \tfrac{1}{2}I_{\{X\}}(Y)\right] = E[H(Y - X)],$$

where $H(\cdot)$ denotes the Heaviside step function, i.e., $H(z) = 1$ for $z > 0$, $H(z) = 0.5$ for $z = 0$ and $H(z) = 0$ for $z < 0$, in case of ties the U statistic becomes

$$U = \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} H(Y_j - X_i), \tag{25}$$

and it represents the number of pairs (X_i, Y_j) for which $X_i < Y_j$ plus half of the number of pairs (X_i, Y_j) for which $X_i = Y_j$. The results presented in Section 4 are still valid if we substitute $I_{(X_j,\infty)}(Y_k)$ with $H(Y_k - X_j)$ in matrix A.

5. Asymptotic consistency

From the expression of the lower and upper means in (19), it can be verified that for $n_1, n_2 \to \infty$:

$$\underline{\mathscr{E}}[P(X \leq Y)|X^{n_1}, Y^{n_2}], \overline{\mathscr{E}}[P(X \leq Y)|X^{n_1}, Y^{n_2}] \simeq \mathscr{E}[P(X \leq Y)|X^{n_1}, Y^{n_2}] \simeq \frac{U}{n_1 n_2}.$$

Notice that in this section the symbol $\simeq$ will be used to indicate asymptotic equivalence. The imprecision (degree of robustness) goes to zero for $n_1, n_2 \to \infty$ and the expectation $\mathscr{E}[P(X \leq Y)|X^{n_1}, Y^{n_2}]$ is asymptotically equivalent to the Mann-Whitney statistic [1]. The consistency of the IDP rank-sum test can be verified by considering the asymptotic behavior of the posterior lower and upper distributions of $P(X \leq Y)$ and compare it to the asymptotic distribution of the statistic U/n_1n_2. For ease of presentation, we limit ourselves to the case $n_1 = n_2 = n$. In Lehmann and D'Abrera [38, Appendix A.5] it is proved that U/n_1n_2 converges for $n_1, n_2 \to \infty$ to a Normal distribution with mean $E[U_{ij}] = P(X \leq Y)$, for all $i = 1, \ldots, n_1$ and $j = 1, \ldots, n_2$, and variance

$$\frac{1}{n} Cov[U_{ij}, U_{i,k\neq j}] + \frac{1}{n} Cov[U_{ij}, U_{k\neq i,j}], \tag{26}$$

where $U_{rt} = I_{(X_r,\infty)}(Y_t)$. In the following theorem an equivalent result is proved for the lower distribution of $P(X \leq Y)$ in the IDP rank-sum test

Theorem 4. *Assuming that $n_1 = n_2 = n$, for $n \to \infty$ the IDP rank-sum test lower distribution converges to a Normal distribution with mean $E[U_{ij}] = P(X \leq Y)$ and variance given by Equation (26).*

The above proof can be easily generalized to the upper distribution (the terms due to w_{10} and w_{20} vanish asymptotically) and to the case $n_1 \neq n_2$ (following the same procedure as in Lehmann and D'Abrera [38, Th. 9]). Theorem 4 proves that the (upper and lower) distribution of the IDP rank-sum test is asymptotically equivalent to the distribution of the statistic U/n_1n_2 and, thus, the IDP rank-sum test is consistent as a test for $P(X \leq Y)$. Conversely, the MWW test is only consistent in the case $P(X \leq Y) = 0.5$ and $F_X = F_Y$ or $P(X \leq Y) \neq 0.5$ and $F_X \neq F_Y$, while it is not consistent for $P(X \leq Y) = 0.5$ and $F_X \neq F_Y$. For instance if $X \sim N(0,1)$ and $Y \sim N(0, \sigma^2)$ with $\sigma^2 > 1$, two Normal distributions with different variance, then $P(X \leq Y) = 0.5$ but the distributions are different. In this case, if we apply MWW test with a significance level $\gamma = 0.05$, MWW will return the alternative hypothesis in approximatively 8.7% of the cases (for a large σ^2), see DasGupta [39, Sec. 25.5]. This means that MWW is not calibrated as a test for $P(X \leq Y) = 0.5$ and it is not powerful as a test for $F_X(x) \neq F_Y(x)$. Conversely, because of Theorem 4, our IDP test with $\gamma = 0.05$ will return the alternative hypothesis (asymptotically) in 5% of the cases, which is correct since $P(X \leq Y) = 0.5$.

6. Numerical simulations

Consider a Monte Carlo experiment in which $n_1,\ n_2$ observations X, Y are generated based on

$$X \sim N(0,1),\ \ Y \sim N(\Delta, 1),$$

with Δ ranging on a grid from -1.5 to 1.5. To facilitate the comparison of IDP tests with more traditional tests (which never issue indeterminate outcomes) we introduce a new test (called "50/50 when indeterminate") which returns the same response as the IDP when this is determinate, and issues a random answer (with 50/50 chance) otherwise. We want to stress that the test "50/50 when indeterminate" has been introduced only for the sake of comparison. We are not suggesting that when the IDP is indeterminate we should toss a coin to take the decision. On the contrary we claim that the indeterminacy of the IDP is an additional useful information that our approach gives to the analyst. In these cases she/he knows that (i) her/his posterior decisions would depend on the choice of the prior G_0; (ii) deciding between the two hypotheses under test is a difficult problem as shown by the comparison with the Bayesian Bootstrap DP (BB-DP) rank-sum test ($s = 0$) and MWW tests. Based on this additional information, the analyst can for example decide to collect additional measurements to eliminate the indeterminacy (in fact we have seen that when the number of observations goes to infinity the indeterminacy goes to zero).

We start by comparing the performance of the BB-DP and IDP tests. To evaluate the performance of the tests, we have used the loss function defined in (15). In particular, for each value of Δ we have performed 20000 Monte Carlo runs by generating in each run $n_1 = n_2 = 20$ observations for X, Y. The average loss for the cases (i) $K_1 = K_2 = 1$ (i.e., $\gamma = 0.5$) (ii) $K_1 = 1$ and $K_2 = 9$ (i.e., $\gamma = 0.1$) and (iii) $K_1 = 1$ and $K_2 = 19$ (i.e., $\gamma = 0.05$) is shown in Figure 2 as a function of Δ. In particular, we report (i) the loss of the BB-DP test ($s = 0$); (ii) the loss of the

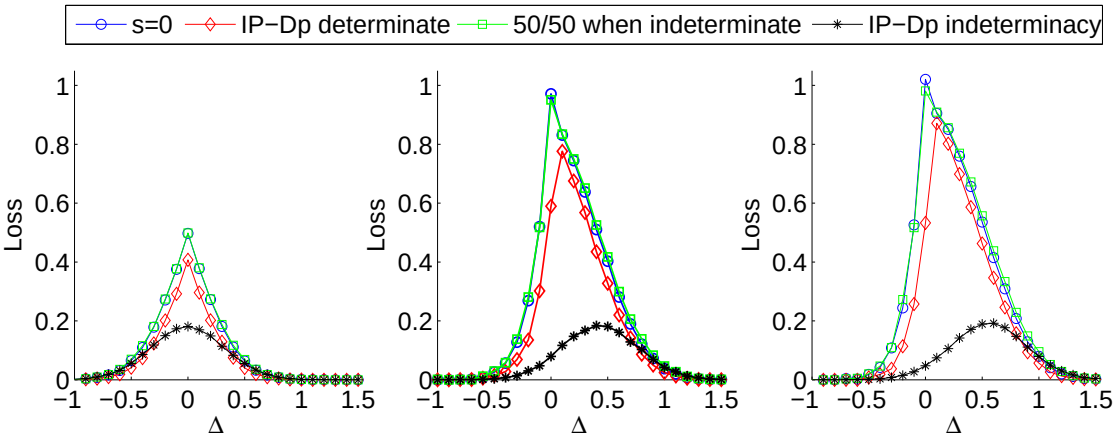


Figure 2. Loss as a function of Δ for the case $K_0 = K_1 = 1$ (left), $K_0 = 1, K_1 = 9$ (center) and $K_0 = 1, K_1 = 19$ (right).

IDP test when it is determinate; (iii) the indeterminacy of the IDP test, i.e., the number of times it returns an indeterminate response divided by the total number of Monte Carlo runs; (iv) the loss of the "50/50 when indeterminate" test.

From Figure 2, it is evident that the performance of the BB-DP and 50/50 tests practically coincide. Furthermore, since we noticed from experimental evidence that in all cases in which IDP is determinate, BB-DP returns the same response as IDP, the difference between the two tests is only in the runs where the IDP is indeterminate. In these runs, BB-DP is clearly guessing at random, since overall it has the same loss as the 50/50 test. Therefore, the IDP is able to isolate several instances in which BB-DP is guessing at random, thus providing useful information to the analyst. Assume, for instance, that we are trying to compare the effects of two medical treatments ("Y is better than X") and that, given the available data, the IDP is indeterminate. In such situation the BB-DP test always issues a determinate response (I can tell if "Y is better than X"), but it turns out that its response is virtually random (like if we were tossing a coin). On the other side, the IDP acknowledges the impossibility of making a decision and thus, although BB-DP and the IDP (more precisely the "50/50 when indeterminate" test) have the same loss, the IDP provides more information. Note that, for all the three loss functions, the maximum percentage of runs in which the IDP is indeterminate is about 18%; this means that for some value of Δ, BB-DP is issuing a random answer in 18% of the cases, which is a large percentage. For large $|\Delta|$, i.e. when the hypothesis test is easy, there are no indeterminate instances and both the BB-DP and the IDP tests have zero loss. It is interesting to note that, for the cases $K_1 = 1$ and $K_2 = 9$ (or $K_2 = 19$) (Figure 2 center and right) it is more risky (we may incur a greater loss) taking the action $a = 1$ than $a = 0$, and thus the indeterminacy curve is shifted to the $\Delta > 0$ quadrant.

We have also compared the IDP test and the one-sided MWW NHST implemented according to the conventional decision criterion, $p < 0.05$. It is well known that the decision process in NHST is flawed. It is based on asking what is the probability of the data statistic if the null hypothesis were true. This means that NHST can only reject the null hypothesis ($\Delta \leq 0$), contrarily to a Bayesian analysis that can also accept this hypothesis. Furthermore, in a Bayesian analysis we have a principled way to determine γ (i.e., by means of a loss function) which is

lost when putting decisions in the format $p < 0.05$ (or the more vague $p < 0.1$). Because of these differences, it is difficult to compare the Bayesian with the NHST approach, where we do not have a clear interpretation of the significance level. However, we believe a relatively fair comparison can be carried out by setting γ equal to the significance level of the NHST test, so that the decision criteria adopted by the two test are as similar as possible. Figure 3 shows the power for the case $\gamma = 0.05$, $n_1 = n_2 = 10$ and $n_1 = n_2 = 20$. In case $n_1 = n_2 = 20$ (Figure 3,

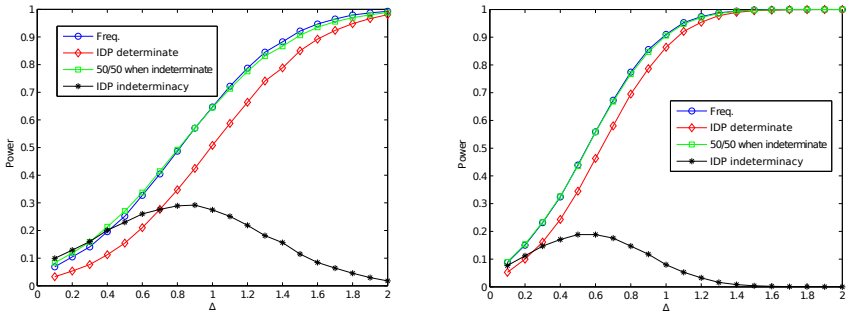


Figure 3. Power as a function of the difference of the medians Δ for the case $n_1 = n_2 = 10$ (left) and $n_1 = n_2 = 20$ (right) with $\gamma = 0.05$. Here "Freq" denotes the MWW test.

right) it is evident that the performance of the MWW and 50/50 tests practically coincide. Since it can be verified experimentally that when the IDP is determinate the two tests return the same results, this again suggests that when the IDP is indeterminate we have equal probability that $p < 0.05$ or $p > 0.05$, as it is shown in Figure 4. The IDP test is able to isolate some instances in which also the MWW test is issuing a random answer. Note that, for $\Delta = 0.5$, the maximum percentage of runs in which the IDP test is indeterminate is large, about 18%; this means that MWW is issuing a random answer in 18% of the cases.

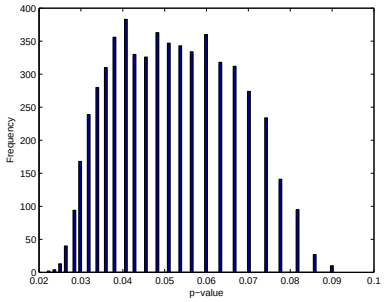


Figure 4. Distribution of MWW p-values in the IDP indeterminate cases for $n_1 = n_2 = 20$, $\gamma = 0.05$ and $\Delta = 0.5$.

The results for the case $n_1 = n_2 = 10$ (Figure 3, left) lead to similar conclusions. The performance of the MWW and 50/50 tests (almost) coincide. The 50/50 test is slightly better for $\Delta \leq 0.9$ and slightly worse for $\Delta > 0.9$. $\Delta = 0.9$ is the value which corresponds to the maximum indeterminacy of the IDP, i.e. 30%. Thus, for $\Delta = 0.9$, MWW is guessing at random in 30% of the runs.

It is worth analyzing also the case $\Delta = 0$. We know that in this case the frequentist test is calibrated, i.e., when $\gamma = 0.05$ the percentage of correct answers is 95% (although it can be noticeably larger for small values of n_1, n_2 since the discreteness of the MWW statistic originates a gap between the chosen γ and the actual significance of the MWW test). Table 1 shows the accuracy (percentage of correct answers) for $\Delta = 0$. The performance of the MWW and 50/50 tests are similar also in this case. The difference is about 1% (for $n_1 = n_2 = 10$) and 0.5% (for $n_1 = n_2 = 20$).

	Accuracy $n_1 = n_2 = 10$	Accuracy $n_1 = n_2 = 20$
MWW	0.955	0.952
50/50 test	0.945	0.947
IDP when determinate	0.911	0.924
Indeterminacy	0.068	0.045

Table 1. Accuracy for $\Delta = 0$ and $\gamma = 0.05$.

	Accuracy $\gamma = 0.1$	Accuracy $\gamma = 0.25$
MWW	0.8995	0.7552
50/50 test	0.8993	0.7482
IDP when determinate	0.8568	0.6777
IDP indeterminacy	0.081	0.142

Table 2. Accuracy in case $\Delta = 0$ for $n_1 = n_2 = 20$

Also in this case, when the IDP is determinate, it returns the same responses as MWW. This result holds independently of the choice of γ, as shown by Figure 5 and Table 2 where we have repeated the above experiment for $n_1 = n_2 = 20$ with, this time, $\gamma = 0.1$ and $\gamma = 0.25$.

Finally, Figure 6 shows the error (one minus the accuracy) of the IDP test as a function of s, when $\gamma = 0.1$, $n_1 = n_2 = 20$ and $\Delta = 0$. Clearly, the error of the MWW test is constantly equal to $\gamma = 0.1$ (we are under the null hypothesis of MWW). The error of the IDP test when determinate decreases with s, because of the increase of the indeterminacy. The error of the 50/50 test has a convex trend, clearly decreasing for $s < 0.2$ and increasing for $s > 0.5$. This (together with the other results of this section) may be seen as an empirical confirmation that the choice of $s = \sqrt{2} - 1$ is appropriate, since it guarantees a good trade-off between robustness and indeterminacy.

Finally , note that all the above differences/similarities between the three tests appear also in the case where we consider location-shift models with distributions different from Gaussians (e.g., Student-t distribution with one or two degrees of freedom). These results have been omitted for shortness.

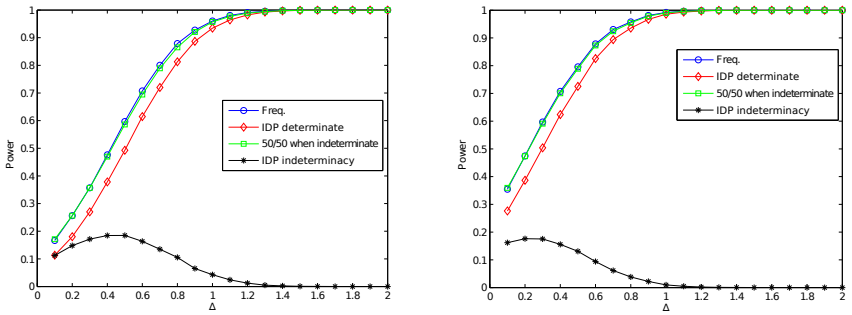


Figure 5. Power as a function of the difference of the medians Δ for $n_1 = n_2 = 20$, $\gamma = 0.1$ (left) and $\gamma = 0.25$ (right). Here "Freq" denotes the MWW test.

7. Conclusions

In this paper we have proposed a model of prior ignorance for nonparametric inference based on the Dirichlet process (DP), by extending the approach proposed by Pericchi and Walley [17], Walley [29] and based on the use of sets of prior distributions. We developed a prior near-ignorance DP model (IDP) for inference about a variable X by fixing the prior strength of the DP and letting the normalized probability measure vary in the set of all distributions. We have proved that the IDP is consistent and a-priori vacuous for all predictive inferences that can be

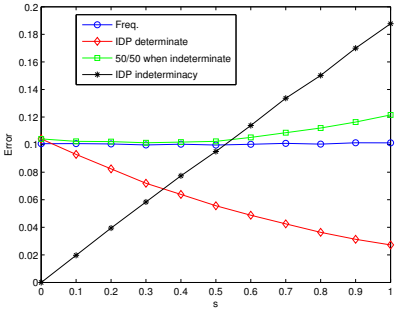


Figure 6. Error as a function of s for $n_1 = n_2 = 20$, $\gamma = 0.1$ and $\Delta = 0$. Here "Freq" denotes the MWW test.

defined as the expectation of a real-valued bounded function of X. The proposed IDP model has two main merits. First, it removes the need for specifying the infinite-dimensional parameter of the DP (only an upper bound for the strength s of the DP must be assumed a-priori), thus making the elicitation of the prior very easy. Second, it allows computing the posterior inferences for which no closed form expression exists, by simple Monte Carlo sampling from the Dirichlet distribution, thus avoiding more demanding sampling approaches typically used for the DP (e.g., stick breaking). Based on this new prior near-ignorance model, we have proposed a general, simple and conservative approach to Bayesian nonparametric tests, and in particular we have developed a robust Bayesian alternative to the Mann-Whitney-Wilcoxon test: the IDP rank-sum test. We have shown that our test is asymptotically consistent, while this is not always the case for the Mann-Whitney-Wilcoxon test. Finally, by means of numerical simulations, we have compared the IDP rank-sum test to the Mann-Whitney-Wilcoxon test and the Bayesian test obtained from the DP when the prior strength goes to zero. Results have shown that the IDP test is more robust, in the sense that it is able to isolate instances in which these tests are practically guessing at random. Given these interesting results, as future work we plan to use this approach to implement Bayesian versions of the most used frequentist nonparametric tests. In the long run, our aim is to build a statistical package for Bayesian nonparametric tests.

8. Appendix

Proof of Theorem 1: From (2) assuming that $P \sim Dp(s,\alpha^*)$ one has that $\mathscr{E}[E(f)] = \int f d\alpha^*$. Define $x_l = \arg\inf_{x\in \mathrm{X}} f(x)$ and $x_u = \arg\sup_{x\in \mathrm{X}} f(x)$, then (9) follows by:

$$\underline{\mathscr{E}}[E(f)] = \inf_{\alpha^*\in \mathrm{P}} \int f d\alpha^* = \int f d\delta_{x_l} = f(x_l), \quad \overline{\mathscr{E}}[E(f)] = \sup_{\alpha^*\in \mathrm{P}} \int f d\alpha^* = \int f d\delta_{x_u} = f(x_u),$$

which are the infimum and supremum of f by definition. The lower and upper bounds are thus obtained by the following degenerate DPs $Dp(s,\delta_{x_l})$ and $Dp(s,\delta_{x_u})$, which belong to the class (8). In case x_l is equal to ∞ (or $-\infty$), with $f(x_l)$ we mean $\lim_{x_l\to\infty} f(x_l)$, similar for the upper.
Proof of Theorem 2: By exploiting the fact that $\mathscr{E}[E(f)|X_1,\ldots,X_n] = \int f \;\; d(\frac{s}{s+n}\alpha^* + \frac{n}{s+n}\frac{1}{n}\sum_{i=1}^n \delta_{X_i})$, the proof is similar to that of Theorem 1 (the lower and upper bounds are again obtained by degenerate DPs $Dp(s,\delta_{x_l})$ and $Dp(s,\delta_{x_u})$).
Proof of Lemma 1: It follows from the properties (a) and (c) of the DP in Section 2.
Proof of Theorem 3: Based on the stick-breaking construction, a sample F_0 from the generic DP $Dp(s,G_0)$ can be written as $F_0(x) = \sum_{k=1}^{\infty} \pi_k \delta_{\tilde{X}_k}$ where $\pi_k = \beta_k \prod_{i-1}^{k-1}(1-\beta_i)$, $\beta_k \sim Beta(1,s)$, and

$X_k \sim G_0$. Then, using (6), we have that

$$F_n(x) = \sum_{i=1}^{n} w_i \delta_{X_i} + w_0 \sum_{k=1}^{\infty} \pi_k \delta_{\tilde{X}_k}, \tag{27}$$

where $(w_0, w_1, \ldots, w_n) \sim Dir(s, \overbrace{1, \ldots, 1}^{n_i})$. Consider the two samples $F_X(x)$ and $F_Y(y)$ from the posterior distributions of X and Y given the generic DP priors $Dp(s, G_{10})$ and $Dp(s, G_{20})$. The probability of $P(X \leq Y) > c$ is $\mathscr{P}\left[P(X \leq Y) > c\right] = \mathscr{P}[\int F_{n_1}(y) dF_{n_2}(y) > c]$. Then, the posterior lower probability of $P(X \leq Y) > c$ is obtained by minimizing $\int F_{n_1}(y) dF_{n_2}(y)$, which, by (27), is equal to

$$\begin{aligned}
&\int \left(\sum_{i=1}^{n_1} w_{1i} I_{(X_i,\infty)}(y) + w_{10} \sum_{k=1}^{\infty} \pi_{1k} I_{(\tilde{X}_k,\infty)}(y) \right) \left(\sum_{j=1}^{n_2} w_{2j} \delta_{Y_j}(y) + w_{20} \sum_{l=1}^{\infty} \pi_{2l} \delta_{\tilde{Y}_l}(y) \right) dy \\
&= \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{1i} w_{2j} I_{(X_i,\infty)}(Y_j) + w_{20} \sum_{i=1}^{n_1} \sum_{l=1}^{\infty} w_{1i} \pi_{2l} I_{(X_i,\infty)}(\tilde{Y}_l) \\
&+ w_{10} \sum_{k=1}^{\infty} \sum_{j=1}^{n_2} \pi_{1k} w_{2j} I_{(\tilde{X}_k,\infty)}(Y_j) + w_{10} w_{20} \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \pi_{1k} \pi_{2l} I_{(\tilde{X}_k,\infty)}(\tilde{Y}_l)
\end{aligned} \tag{28}$$

The minimum of $\int F_{n_1}(y) dF_{n_2}(y)$ is always found in correspondence of prior DPs such that the posterior probability of sampling $\tilde{X}_k < Y_j, \tilde{Y}_l$ or $\tilde{Y}_l > X_i$ is zero, so that only the term $\sum_{i=1}^{n_1} \sum_{j=1}^{n_2} w_{1i} w_{2j} I_{(X_i,\infty)}(Y_j)$ remains in (28). Priors of such kind are, for example, the extreme DP priors that give the posterior lower mean in (19) and the posterior Dirichlet process F_{n_i} with base probability G_{n_i} given by (20). From the property of the Dirichlet distribution, we know that $E[w_{ij}] = 1/(s+n_i)$ and, thus, we can rewrite the lower expectation given in the first equation of (19) as

$$\mu = \sum_{j=1}^{n_1} \sum_{k=1}^{n_2} \frac{1}{s+n_1} \frac{1}{s+n_2} I_{(X_j,\infty)}(Y_k) = E_W[W]^T A E_V[V],$$

For the variance, we have that $\sigma^2 = E[(\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} w_{1j} w_{2k} I_{(X_j,\infty)}(Y_k))^2] - \mu^2$. Thus, by exploiting the equality

$$\left(\sum_{j=1}^{n_1} \sum_{k=1}^{n_2} w_{1j} w_{2k} I_{(X_j,\infty)}(Y_k) \right)^2 = W^T A V W^T A V = V^T A^T W W^T A V,$$

the linearity of expectation and the independence of W, V, one obtains

$$E[V^T A^T W W^T A V] = E_V[V^T A^T E_W[W W^T] A V] = E_V[V^T A^T E_W[W W^T] A V].$$

Since the result of this product is a scalar, it is equal to its trace and thus we can use the cyclic property $trace[E_V[V^T A^T E_W[W W^T] A V]] = trace[A^T E_W[W W^T] A E_V[V V^T]]$, and finally obtain $\sigma^2 = trace[A^T E_W[W W^T] A E_V[V V^T]] - \mu^2$. The proof is easily completed by deriving $E_W[W W^T]$ and $E_V[V V^T]$ from the fact that w_{ij}, w_{kl} are independent and $E_W[w_{ij}^2] = \frac{2}{(s+n_i)(s+n_i+1)}$, $E_W[w_{ij} w_{il}] = \frac{1}{(s+n_i)(s+n_i+1)}$.

Proof of Corollary 1: First, observe that the posterior upper probability of $P(X \leq Y) > c$ is obtained in correspondence of the extreme DP prior that gives the posterior upper mean in (19) and

has base probability dG_{n_i} given by (20). The probability of $X \leq Y$ for a given realization F_{n_1} of $Dp(s,G_{n_1})$, and F_{n_2} of $Dp(s,G_{n_2})$ is:

$$\mathscr{P}\left[P(X \leq Y) > c | X^{n_1}, Y^{n_2}\right] = \mathscr{P}\left[\int F_{n_1}(y) dF_{n_2}(y) > c\right]$$
$$= \mathscr{P}\left[\int \left(w_{10} I_{(X_0,\infty)}(y) + \sum_{j=1}^{n_1} w_{1j} I_{(X_j,\infty)}(y)\right)\left(w_{20}\delta_{Y_0}(y) + \sum_{j=1}^{n_2} w_{2j}\delta_{Y_j}(y)\right) dy > c\right]$$
$$= \mathscr{P}\left[w_{10}w_{20} + w_{10}\sum_{j=1}^{n_2} w_{2j} + w_{20}\sum_{j=1}^{n_1} w_{1j} + \sum_{j=1}^{n_1}\sum_{k=1}^{n_2} w_{1j}w_{2k} I_{(X_j,\infty)}(Y_k) > c\right].$$

The computations are similar to those in Theorem 3, but in this case we must also consider the expectations $E_W[w_{i0}^2] = s(s+1)/(s+n_i)(s+n_i+1)$, $E_W[w_{i0}w_{ij}] = s/(s+n_i)(s+n_i+1)$ for $j > 0$.

Proof of Theorem 4: Our goal is to prove the convergence to a normal distribution of the Bayesian bootstrapped two-sample statistic $U_{DP} = \sum_{i,j} w_{1i}w_{2j} I_{[X_i,\infty)}(Y_j)$, which implies the asymptotic normality of the DP rank sum test lower distribution, since the contribution of the prior G_0 vanishes asymptotically. The asymptotic normality of U_{DP} can be proved by means of Lemma 6.1.3. of Lehmann [40], which states that given a sequence of random variables T_n, the distributions of which tend to a limit distribution L, the distribution of another sequence T_n^* satisfying $E[(T_n^* - T_n)^2] \to 0$ also tends to L. Said $h(x,y) = I_{[x,\infty)}(y)$, $h_1(x) = E_Y[h(x,Y)]$ and $h_2(y) = E_X[h(X,y)]$, the theorem will be proved by applying the lemma to

$$T_n = \sqrt{n}\left[\frac{1}{n}\left(\sum_{i=1}^{n} h_1(X_i) - \theta\right) + \frac{1}{n}\left(\sum_{j=1}^{n} h_2(Y_j) - \theta\right)\right]$$

and $T_n^* = \sqrt{n}(U_{DP} - \theta)$ where $\theta = E[U_{ij}] = E[h_1(X)] = E[h_2(Y)]$. T_n is a sum of independent terms and thus, from the central limit theorem, it converges to a Gaussian distribution with mean 0 and variance $\sigma^2 = \sigma_1^2 + \sigma_2^2$, where $\sigma_1^2 = Var[h_1(X)]$ and $\sigma_2^2 = Var[h_2(Y)]$). Note that

$$\sigma_1^2 = Cov[h(X,Y), h(X,Y')] = Cov[U_{ij}, U_{i,k\neq j}],$$
$$\sigma_2^2 = Cov[h(X,Y), h(X',Y)] = Cov[U_{ij}, U_{k\neq i,j}].$$

From Theorem 3, the mean of the lower distribution of U_{DP} is $\mu_l = E_W[W]^T A E_V[V] = \frac{U}{(s+n)^2}$, and thus, for large n, it is asymptotic to U/n^2 which converges, in turn, to $E[U_{ij}] = \theta$.Then, also $E[T_n^*] = 0$ so that

$$E[(T_n^* - T_n)^2] = Var[T_n^*] + Var[T_n] - 2Cov[T_n^*, T_n].$$

The proof will be completed by showing that $Var[T_n^*] \to \sigma^2$ and $Cov[T_n^*, T_n] \to \sigma^2$. For the variance of U_{DP} (23), first note that we can rewrite $E_W[WW^T] = E_V[VV^T] = (D + J_n)\frac{1}{(s+n)(s+n+1)}$ where D is the diagonal matrix of ones (identity matrix) and J_n is the $n \times n$ matrix of ones. Thus, we have that

$$A^T E_W[WW^T] A E_V[VV^T] = A^T(D+J_n)A(D+J_n)\frac{1}{(s+n)^2(s+n+1)^2},$$

and, for large n,

$$\frac{trace(A^T(D+J_n)A(D+J_n))}{(s+n)^2(s+n+1)^2} \rightarrow \frac{trace(A^TA)+trace(A^TAJ_n)+trace(A^TJ_nA)+trace(A^TJ_nAJ_n)}{n^2(n+1)^2}.$$

The above sum has four terms at the numerator:

$$\begin{aligned} trace(A^TA) &= \sum_{i,j} a_{ij}^2 = \sum_{i,j} I_{(X_i,\infty)}(Y_j), \\ trace(A^TAJ_n) &= \sum_{i,j} a_{ij} \sum_k a_{ik} = \sum_{i,j} I_{(X_i,\infty)}(Y_j) + \sum_{i,j\neq k} I_{(X_i,\infty)}(Y_j) I_{(X_i,\infty)}(Y_k), \\ trace(A^TJ_nA) &= \sum_{i,j} a_{ij} \sum_k a_{kj} = \sum_{i,j} I_{(X_i,\infty)}(Y_j) + \sum_{i\neq k,j} I_{(X_i,\infty)}(Y_j) I_{(X_k,\infty)}(Y_j), \end{aligned}$$

and $trace(A^TJ_nAJ_n) = trace(A^T\mathbb{1}\mathbb{1}^TA\mathbb{1}\mathbb{1}^T) = trace(\mathbb{1}^TA\mathbb{1}\mathbb{1}^TA^T\mathbb{1}) = U^2$, where $\mathbb{1}$ is the unit vector. Then we have that

$$\begin{aligned} \sigma_l^2 = &\frac{3\sum_{i,j} I^2_{(X_i,\infty)}(Y_j) - 3n^2\mu_l^2}{n^2(n+1)^2} + \frac{\sum_{i,j\neq k} I_{(X_i,\infty)}(Y_j) I_{(X_i,\infty)}(Y_k) - n^2(n-1)\mu_l^2}{n^2(n+1)^2} \\ &+ \frac{\sum_{i\neq k,j} I_{(X_i,\infty)}(Y_j) I_{(X_k,\infty)}(Y_j) - n^2(n-1)\mu_l^2}{n^2(n+1)^2} + \frac{3n^2+2n^2(n-1)+n^4}{n^2(n+1)^2}\mu_l^2 - \mu_l^2. \end{aligned}$$

Note that $(\frac{3n^2+2n^2(n-1)+n^4}{n^2(n+1)^2} - 1)\mu_l^2 = 0$ and, since the first term in σ_l^2 goes to zero as $1/n^2$, for large n,

$$\sigma_l^2 \rightarrow \frac{\sum_{i,j\neq k} I_{(X_i,\infty)}(Y_j) I_{(X_i,\infty)}(Y_k) - n^2(n-1)\mu_l^2}{n^2(n+1)^2} + \frac{\sum_{i\neq k,j} I_{(X_i,\infty)}(Y_j) I_{(X_k,\infty)}(Y_j) - n^2(n-1)\mu_l^2}{n^2(n+1)^2}.$$

For large n, it can be shown Lehmann and D'Abrera [38, Th. 9] that the right-hand side of the above equations tends to $\frac{1}{n}Cov[U_{ij}, U_{i,k\neq j}] + \frac{1}{n}Cov[U_{ij}, U_{k\neq i,j}] = \frac{1}{n}\sigma^2$, and thus $Var[T_n] = Var[\sqrt{n}U_{DP}] \rightarrow \sigma^2$. For the covariance we have

$$\begin{aligned} Cov[T_n, T_n^*] &= \left(E[U_{DP} \sum_{i=1}^n h_1(X_i)] + E[U_{DP} \sum_{j=1}^n h_2(Y_j)] - 2\theta \right) \\ &= \left(E[\sum_{i,j} w_{1i}w_{2j} E_{Y_j}[h(X_i,Y_j)] \sum_{i=1}^n h_1(X_i)] + E[\sum_{i,j} w_{1i}w_{2j} E_{X_i}[h(X_i,Y_j)] \sum_{j=1}^n h_2(Y_j)] - 2\theta \right) \\ &= \left(E_X[\sum_{i=1}^n E[w_{1i}] h_1(X_i) \sum_{i=1}^n h_1(X_i)] + E_Y[\sum_{j=1}^n E[w_{2j}] h_2(Y_j) \sum_{j=1}^n h_2(Y_j)] - 2\theta \right) \\ &= \frac{1}{n} \left((\sum_{i=1}^n E_X[h_1(X_i))^2] + (\sum_{j=1}^n E_Y[h_2(Y_j))^2] - 2\theta \right) = Var[h_1(X)] + Var[h_2(Y)] = \sigma^2. \end{aligned}$$

Acknowledgements

This work has been partially supported by the Swiss NSF grants nos. 200020_137680 / 1 and 200021_146606 / 1.

References

[1] T. S. Ferguson, A Bayesian Analysis of Some Nonparametric Problems, The Annals of Statistics 1 (2) (1973) 209–230, ISSN 00905364.

[2] V. Susarla, J. Van Ryzin, Nonparametric Bayesian Estimation of Survival Curves from Incomplete Observations, Journal of the American Statistical Association 71 (356) (1976) 897–902, ISSN 01621459.

[3] J. Blum, V. Susarla, On the posterior distribution of a dirichlet process given randomly right censored observations, Stochastic Processes and their Applications 5 (3) (1977) 207–211, ISSN 0304-4149, doi:http://dx.doi.org/10.1016/0304-4149(77)90030-8.
[4] S. Dalal, E. Phadia, Nonparametric Bayes inference for concordance in bivariate distributions, Commun. in Statistics-Theory and Methods 12 (8) (1983) 947–963.
[5] E. G. Phadia, Inference Based on Complete Data, in: Prior Processes and Their Applications, Springer, New York, 109–153, 2013.
[6] A. Jara, T. Hanson, F. Quintana, P. Müller, G. Rosner, DPpackage: Bayesian Semi- and Nonparametric Modeling in R, Journal of Statistical Software 40 (5) (2011) 1–30.
[7] P. Rossi, R. McCulloch, Bayesm: Bayesian inference for marketing/micro-econometrics, R package version (2010) 2–2.
[8] K. M. Borgwardt, Z. Ghahramani, Bayesian two-sample tests, arXiv preprint arXiv:0906.4032 .
[9] C. Holmes, F. Caron, J. Griffin, D. A. Stephens, Two-sample Bayesian nonparametric hypothesis testing, arXiv preprint arXiv:0910.5060 .
[10] L. Ma, W. H. Wong, Coupling optional Pólya trees and the two sample problem, Journal of the American Statistical Association 106 (496).
[11] Y. Chen, T. E. Hanson, Bayesian nonparametric k-sample tests for censored and uncensored data, Computational Statistics and Data Analysis 71 (C) (2014) 335–346.
[12] R. Martin, S. Tokdar, A nonparametric empirical Bayes framework for large-scale multiple testing, Biostatistics 13 (3) (2012) 427–439.
[13] D. B. Rubin, Bayesian Bootstrap, The Annals of Statistics 9 (1) (1981) 130–134, ISSN 00905364.
[14] J. O. Berger, An overview of robust Bayesian analysis with discussion, Test 3 (1) (1994) 5–124.
[15] J. O. Berger, D. Rios Insua, F. Ruggeri, Bayesian Robustness, in: D. Rios Insua, F. Ruggeri (Eds.), Robust Bayesian Analysis, vol. 152 of *Lecture Notes in Statistics*, Springer New York, 1–32, 2000.
[16] P. Walley, Statistical Reasoning with Imprecise Probabilities, Chapman and Hall, New York, 1991.
[17] L. R. Pericchi, P. Walley, Robust Bayesian credible intervals and prior ignorance, International Statistical Review (1991) 1–23.
[18] F. Ruggeri, Nonparametric Bayesian robustness, Chilean Journal of Statistics 2 (2010) 51–68.
[19] T. Augustin, F. Coolen, Nonparametric predictive inference and interval probability, Journal of Statistical Planning and Inference 124 (2) (2004) 251–272.
[20] F. Coolen, T. Augustin, A nonparametric predictive alternative to the Imprecise Dirichlet Model: The case of a known number of categories, International Journal of Approximate Reasoning 50 (2) (2009) 217–230.
[21] J. Sethuraman, A constructive definition of Dirichlet priors, Statistica Sinica 4 (2) (1994) 639–650.
[22] F. P. Coolen, S. Bin Himd, Nonparametric predictive inference for reproducibility of basic nonparametric tests, Journal of Statistical Theory and Practice (ahead-of-print) (2014) 1–28.
[23] A. Benavoli, F. Mangili, G. Corani, M. Zaffalon, F. Ruggeri, A Bayesian Wilcoxon signed-rank test based on the Dirichlet process, in: Proceedings of The 32th International Conference on Machine Learning (ICML), 1026–1034, 2014.
[24] J. K. Ghosh, R. Ramamoorthi, Bayesian nonparametrics, Springer (NY), 2003.
[25] A. Y. Lo, A Large Sample Study of the Bayesian Bootstrap, The Annals of Statistics 15 (1).
[26] C.-S. Weng, On a Second-Order Asymptotic Property of the Bayesian Bootstrap Mean, The Annals of Statistics 17 (2) (1989) 705–710, ISSN 00905364.
[27] B. Efron, Bootstrap Methods: Another Look at the Jackknife, The Annals of Statistics 7 (1) (1979) 1–26, ISSN 00905364.
[28] J. Sethuraman, R. C. Tiwari, Convergence of Dirichlet measures and the interpretation of their parameter, Defense Technical Information Center, 1981.
[29] P. Walley, Inferences from multinomial data: learning about a bag of marbles, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1) (1996) 3–57.
[30] A. Benavoli, M. Zaffalon, A model of prior ignorance for inferences in the one-parameter exponential family, Journal of Statistical Planning and Inference 142 (7) (2012) 1960–1979.
[31] J.-M. Bernard, An introduction to the imprecise Dirichlet model for multinomial data, International Journal of Approximate Reasoning 39 (23) (2005) 123–150, imprecise Probabilities and Their Applications.
[32] G. de Cooman, E. Miranda, E. Quaeghebeur, Representation insensitivity in immediate prediction under exchangeability, International Journal of Approximate Reasoning 50 (2) (2009) 204 – 216.

[33] M. P. Fay, M. A. Proschan, Wilcoxon-Mann-Whitney or t-test? On assumptions for hypothesis tests and multiple interpretations of decision rules, Statistics Surveys 4 (2010) 1–39.
[34] J. L. Hodges, E. L. Lehmann, Estimates of Location Based on Rank Tests, The Annals of Mathematical Statistics 34 (2) (1963) 598–611.
[35] A. E. Raftery, Bayesian model selection in social research, Sociological methodology 25 (1995) 111–164.
[36] J. K. Kruschke, Bayesian data analysis, Wiley Interdisciplinary Reviews: Cognitive Science 1 (5) (2010) 658–676.
[37] S. N. Goodman, Toward evidence-based medical statistics. 1: The P–value fallacy, Annals of internal medicine 130 (12) (1999) 995–1004.
[38] E. Lehmann, H. J. M. D'Abrera, Nonparametrics: statistical methods based on ranks, McGraw-Hill,San Francisco, 1975.
[39] A. DasGupta, Asymptotic theory of statistics and probability, Springer (NY), 2008.
[40] E. Lehmann, Elements of Large-Sample Theory, Springer-Verlag New York, 1998.