A Bayesian Wilcoxon signed-rank test based on the Dirichlet process

A. Benavoli, F. Mangili, G. Corani, M. Zaffalon {ALESSIO,FRANCESCA,GIORGIO,ZAFFALON}@IDSIA.CH
IPG IDSIA, Manno, Switzerland

F. Ruggeri FABRIZIO@MI.IMATI.CNR.IT
CNR IMATI, Milano, Italy

Abstract

Bayesian methods are ubiquitous in machine learning. Nevertheless, the analysis of empirical results is typically performed by frequentist tests. This implies dealing with null hypothesis significance tests and *p*-values, even though the shortcomings of such methods are well known. We propose a nonparametric Bayesian version of the Wilcoxon signed-rank test using a Dirichlet process (DP) based prior. We address in two different ways the problem of how to choose the infinite dimensional parameter that characterizes the DP. The proposed test has all the traditional strengths of the Bayesian approach; for instance, unlike the frequentist tests, it allows verifying the null hypothesis, not only rejecting it, and taking decisions which minimize the expected loss. Moreover, one of the solutions proposed to model the infinite-dimensional parameter of the DP allows isolating instances in which the traditional frequentist test is guessing at random. We show results dealing with the comparison of two classifiers using real and simulated data.

1. Introduction

The field of machine learning is constantly growing. Many novel approaches in classification, regression etc. are constantly proposed, raising the issue of assessing and comparing these new methods with the state-of-the-art. A proper way to perform such comparison is by means of statistical procedures. Tutorials on the use of parametric and nonparametric statistical tests as a methodology for comparing algorithms have been presented (Demšar, 2006; Trawiński et al., 2010; Derrac et al., 2011) on different areas of machine learning. In all these papers, the Wilcoxon signed-rank test is indicated as the nonparametric statistical procedure for the analysis of two paired samples.

Let us consider classification as a case-study. After having assessed the accuracy (or the AUC, or any other indicator) of two competing classifiers on multiple data sets, one has to formally check whether the difference among the two classifiers is significant. The Wilcoxon signed-rank is used for this purpose because, thanks to its *nonparametric* nature, it solves some major problems of the *t*-test: it does *not* assume commensurability of the measures across different data sets; it does *not* assume normality of the sample mean of the accuracy; it is *robust* w.r.t. outliers. The signed-rank test is moreover preferable also to the sign test, which is nonparametric but has much lower power.

However the Wilcoxon signed-rank test is affected by all the drawbacks which characterize the 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, 1995). 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* α, 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. Hence, decisions are simply taken by setting α =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 (Goodman, 1999; Kruschke, 2010).

Bayesian tests of hypothesis constitute an appealing alternative to the NHST analogous. They return the posterior probability of the null and the alternative hypotheses, which are thus fully characterized in terms of mean, variance, credible interval and density function. While the frequentist test can only *reject* the null hypothesis, the Bayesian one can also *accept* the null hypothesis, on the basis of the estimated posterior probability. Once

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

the costs of Type I and Type II errors are specified, a Bayesian test allows taking decisions by minimizing the expected loss; in this way, the size of the test can be better adapted to the actual need. The computation does not depend on the intention of the person who collected the data. Bayesian tests have been widely considered for clinical practice (Spiegelhalter et al., 1994) and cognitive sciences (Kruschke, 2010). They are however very rarely used in machine learning, despite the abundance of Bayesian algorithms in this area. To the best of our knowledge, no Bayesian version of the Wilcoxon signed-rank has been proposed so far. We fill this gap by proposing a nonparametric Bayesian version of the Wilcoxon signed-rank test based on the Dirichlet process (DP) (Ferguson, 1973). In his seminal paper on the Dirichlet process, Ferguson provided a Bayesian justification of many classic nonparametric estimators (Mann-Whitney statistics, median, etc.). Similar results were derived by other authors that, employing DP as prior model, were able to naturally obtain estimators related to the frequentist ones, e.g., Kaplan-Meier (Susarla & Ryzin, 1976), Kendall's tau (Dalal & Phadia, 1983). Recently there has been an increasing interest in the development of Bayesian nonparametric procedures for hypothesis testing focusing in particular on the two (or k) sample problem (Borgwardt & Ghahramani, 2009; Holmes et al., 2009; Ma & Wong, 2011; Chen & Hanson, 2014). Two sample tests deal with *unpaired* samples, while in machine learning we often work with *paired* observations, e.g., when we want to compare the accuracy of two classifiers on the same collection of datasets. Moreover, an open problem in all these procedures is how to choose the infinite dimensional parameter of the nonparametric prior in case of lack of prior information.

Here, we address the problem of how to choose the infinite dimensional parameter that characterizes the DP by means of two models corresponding to two different choices of the prior parameter: the noninformative DP prior (Dp($s = 0$)) and the prior ignorance model (IDP). Dp($s = 0$) is the nonparametric analogue of a Bayesian noninformative prior, while IDP consists of a set of DPs priors and is based on the techniques developed in *Bayesian robustness* (Berger et al., 2000; 1994; Pericchi & Walley, 1991; Walley, 1991) for modelling lack of prior information within parametric models.

By means of simulations on artificial and real world data, we use our test to decide if a certain classifier is significantly better than another. We show that the Bayesian test incurs much lower costs than the frequentist one for a wide variety of costs of Type I and Type II errors. We moreover show that the IDP test is more robust, in the sense that it acknowledges when the decision is *prior-dependent*. In other words, the IDP test suspends the judgment and becomes *indeterminate* when the option which minimizes the expected loss depends on the prior. This behavior is analogous to that observed in credal classifiers (Corani & Zaffalon, 2008), which suspend the judgment when the classification is prior-dependent, namely when the most probable class varies with the prior used to induce the classifier. The little reliability of *prior-dependent* decisions is confirmed by the fact that when the IDP test is indeterminate, the Wilcoxon signed-rank and the Dp($s = 0$) tests are virtually behaving as random guessers. Since IDP has all the positive features of a Bayesian test and it is more reliable than Wilcoxon and Dp($s = 0$), we propose IDP as a new test for comparing classifiers and other methods in machine learning. Finally, notice that the proposed test is applicable to many other fields of research where it has the potential of reducing the misleading results of NHST by avoiding the use of a significance parameter which does not represent a correct measure of the evidence provided by data (Johnson, 2013). The IDP test developed in this work can currently be used online (or downloaded as R or Matlab code) at `http://ipg.idsia.ch/software/IDP.php`.

2. Dirichlet Process

The Dirichlet process $Dp(\alpha)$ has been proposed by (Ferguson, 1973) as a probability measure on the set of probability measures on some space $\mathscr{Z}$ (for this paper we can assume $\mathscr{Z} = \mathbb{R}$). It has an infinite dimensional parameter $\alpha(\cdot)$, which is a positive finite measure over $\mathscr{Z}$, i.e., $\alpha(A) > 0$ (positive) for any (measurable) set $A \in \mathscr{Z}$ and $\alpha(\mathscr{Z}) < \infty$ (finite). Assuming that a probability measure is drawn from DP, i.e., $P \sim Dp(\alpha)$, the characteristic of DP is that, for any (measurable) partition $B_1, \ldots, B_m$ of $\mathscr{Z}$, the finite vector $(P(B_1), P(B_2), \ldots, P(B_m))$ is Dirichlet distributed $Dir(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_m))$. From the Dirichlet distribution, we can thus derive the prior mean $(\mathscr{E}[P(B_i)] = \alpha(B_i)/\alpha(\mathscr{Z}))$ and variance $(\mathscr{E}[(P(B_i) - \mathscr{E}[P(B_i)])^2] = \alpha(B_i)(\alpha(\mathscr{Z}) - \alpha(B_i))/\alpha^2(\mathscr{Z})(\alpha(\mathscr{Z}) + 1)$ of $P(B_i)$ w.r.t. the DP for any $B_i \in \mathscr{Z}$.[1] This shows that the normalized measure $\alpha(\cdot)/\alpha(\mathscr{Z})$ of DP reflects the prior expectation of P, while the scaling parameter $\alpha(\mathscr{Z})$ controls the variance of P around $\alpha(\cdot)/\alpha(\mathscr{Z})$. The normalized measure $\alpha(\cdot)/\alpha(\mathscr{Z})$ is a probability measure, therefore, when $\mathscr{Z} = \mathbb{R}$, it can be completely characterized by the cumulative distribution function $G(z) = \alpha(-\infty, z]/\alpha(\mathscr{Z})$. We can then denote the Dirichlet process by $Dp(\alpha(\mathscr{Z}), G)$.

DP is a conjugate prior in the sense that, given a sample $Z_1, \ldots, Z_n$ from $F \sim Dp(\alpha(\mathscr{Z}), G)$ of n observations which are conditionally independent given F, and fixed the prior parameters $\alpha(\mathscr{Z}) = s$ and $G = G_0$, the posterior distribution of the cumulative distribution function F of P is still

[1] We will use calligraphic letters, $\mathscr{E}, \mathscr{P}$, to denote expectation and probability w.r.t. the DP.

$Dp(\alpha_n(\mathscr{Z}), G_n)$ with

$$\alpha_n(\mathscr{Z}) = s+n, \;\; G_n = \frac{s}{s+n}G_0 + \frac{1}{s+n}\sum_{i=1}^{n} I_{[Z_i,\infty)},$$

where $I_A(z)$ is the indicator function: it is one when $z \in A$ and zero otherwise. Thus, a-posteriori we have that $\mathscr{E}[P(Z \le z)|Z_1,\ldots,Z_n] = (sG_0(z) + n_{<z})/(s+n)$, where $n_{<z} = \sum_{i=1}^{n} I_{[Z_i,\infty)}(z)$ is the number of observations Z_i falling in $(-\infty, z]$.

An issue in the use of the DP as prior measure on P is how to choose the infinite dimensional parameter G_0 in case of lack of prior information. There are two avenues that we can follow. The first assumes that prior ignorance can be modelled satisfactorily by a so-called noninformative prior. In the DP setting, the only noninformative prior that has been proposed so far is the limiting DP obtained for $s \to 0$, which has been introduced by (Ferguson, 1973) and discussed by (Rubin, 1981). The second approach suggests that lack of prior information should be expressed in terms of a set of probability distributions. This approach is known as *Bayesian robustness* (Berger et al., 2000; 1994; Pericchi & Walley, 1991; Walley, 1991; Coolen-Schrijner et al., 2009) and it has been extensively applied to model lack of prior information in parametric models. In this paper, we implement the limiting DP obtained for $s \to 0$ and we also extend the Bayesian robust approach to the nonparametric setting by considering a set of DPs obtained by fixing s to a strictly positive value and letting G_0 span the set of all distributions.

3. The Dirichlet Process-based Wilcoxon test

Let $X^n = (X_1,\ldots,X_n)$ and $Y^n = (Y_1,\ldots,Y_n)$ be two sequences of paired observations representing the accuracies of classifiers X and Y on n different data sets. Define $Z_i = Y_i - X_i$ and assume for the moment that there are no ties of type $Z_i = -Z_j$. We will discuss how to manage ties and zeros in Section 3.2. Let F be the distribution of Z and M its median. A one-sided test contrasts the null hypothesis $M \le 0$ against the alternative hypothesis $M > 0$. The easiest test about the median of a distribution is the sign test which counts the number of positive differences $Z_i > 0$; this statistic is an estimator of the probability $P(Z > 0)$. Then, a Bayesian analogous of the sign test would consider the posterior probability of $P(Z > 0)$. However, a shortcoming of the sign-test is its low power.

The Wilcoxon signed-rank test is more powerful than the sign test, as it accounts not only for the sign but also for the size of the differences Z_i. It does so by ranking the absolute value of the differences and then comparing the ranks of the positive and negative differences (Demšar, 2006). The Wilcoxon signed-rank test assumes the symmetry of F w.r.t. its median M and computes the statistic:

$$T^+ = \sum_{\{i:\, Z_i \ge 0\}} r_i(|Z_i|) = \sum_{1\le i\le j\le n} T^+_{ij}, \qquad (1)$$

where $r_i(|Z_i|)$ is the rank of $|Z_i|$ and

$$T^+_{ij} = \begin{cases} 1 & \text{if } Z_i \ge -Z_j, \\ 0 & \text{otherwise.} \end{cases}$$

The decision is taken by comparing the observed value of the statistic $R^+ = \frac{2T^+}{n(n+1)}$ against its critical value, which depends on the significance α. For a large number of data, the distribution of R^+ under the null hypothesis is approximately normal with mean $1/2$. Then, in practice, considering for example a one-sided test evaluating $M \le 0$ against $M > 0$, the null hypothesis is rejected when the observed value of R^+ is significantly larger (according to the significance level α) than $1/2$. Based on the definition of T^+_{ij}, one can interpret the statistic R^+ as an estimator of the probability that, given two independent observations Z and Z' from F, $Z \ge -Z'$. This probability can be written as

$$\begin{aligned} &P(Z \ge -Z') = P(Z \le 0,\, Z' > 0,\, |Z'| \ge |Z|) \\ &+ P(Z > 0,\, Z' \le 0,\, |Z| \ge |Z'|) + P(Z > 0,\, Z' > 0), \end{aligned}$$

from which it can be noticed that $P(Z \ge -Z')$ considers at the same time the probability that Z is positive and the probability that negative differences are smaller than the positive ones; for this reason the Wilcoxon statistic is more sensitive than the sign test statistic to the presence of a bias (positive if $P(Z \ge -Z') > 1/2$ or negative if $P(Z \ge -Z') < 1/2$) in the differences Z_i. In analogy with the Wilcoxon signed-rank test, we propose a Bayesian test based on the probability

$$P(Z \ge -Z') = \iint I_{[-z',\infty)}(z)\, d(F(z)F(z')) = E[I_{[-Z',\infty)}(Z)].$$

The test compares the hypothesis $P(Z \ge -Z') \le 1/2$ against $P(Z \ge -Z') > 1/2$. Notice that the Wilcoxon signed-rank test needs to assume the symmetry of F to be able to specify the distribution of R^+ under the null hypothesis. In this context, another advantage of the Bayesian approach is that it does not require the symmetry assumption since all inferences are derived from the posterior distribution of $P(Z \ge -Z') > 1/2$ which follows directly from the prior distribution for F and the sequence of observations $Z^n = (Z_1,\ldots,Z_n)$. We propose the Dirichlet process as prior for F.

Theorem 1 *If* $F \sim Dp(\alpha(\mathscr{Z}), G)$, *then*

$$\begin{aligned} \mathscr{E}\left[P(Z \ge -Z')\right] &= \iint I_{[-z',\infty)}(z)\, d\mathscr{E}[F(z)F(z')] \qquad (2) \\ &= \iint I_{[-z',\infty)}(z)\, \frac{d\left[G(\min(z,z')) + \alpha(\mathscr{Z})\, G(z)G(z')\right]}{\alpha(\mathscr{Z})+1}. \end{aligned}$$

This result is similar to that in Lemma 2.1 of (Dalal & Phadia, 1983) for the Kendall's tau. Its proof and that of the next theorems can be found in the appendix (supplementary material). To use the DP for evaluating the posterior probability of $P(Z \geq -Z') > 1/2$, we must choose the base CDF G_0. We focus on the case where we have no prior information about the functional form of F (which would justify the use of a nonparametric test), and about the value of $P(Z \geq -Z')$, and propose a model which is capable of modeling a situation of complete prior ignorance about the expectation of $P(Z \geq -Z')$. For any choice of s and G_0 the prior and posterior expectation of $P(Z \geq -Z')$ can be derived from Theorem 1, by taking, respectively, $\alpha(\mathscr{Z}) = s$ and $G = G_0$ for the prior and $\alpha(\mathscr{Z}) = s+n$ and $G = G_n$ for the posterior. Since the form of G_0 does not affect the posterior for $s \to 0$, this is a frequent choice for modeling a noninformative prior. This prior has been introduced under the name of Bayesian Bootstrap by (Rubin, 1981). The prior and posterior expectations, in this case, are given by the following theorem.

Theorem 2 *Given that $F \sim Dp(s, G_0)$, for $s \to 0$ one has*

$$\mathscr{E}\left[P(Z \geq -Z')\right] = \int I_{[0,\infty)}(z) dG_0(z), \tag{3}$$

$$\mathscr{E}[P(Z \geq -Z')|Z^n] = \frac{1}{n(n+1)}\left[\sum_{i=1}^{n}\sum_{j=1}^{n} I_{[-Z_i,\infty)}(Z_j) + \sum_{j=1}^{n} I_{[0,\infty)}(Z_j)\right]. \tag{4}$$

It can be easily seen that the posterior expectation obtained for $s \to 0$ is equal to R^+. This shows that $\mathscr{E}[P(Z \geq -Z')]$ is closely related to the Wilcoxon signed-rank statistic R^+. This result extends to the Wilcoxon signed-rank a similar result obtained by (Ferguson, 1973) concerning the relationship between $\mathscr{E}[P(X \leq Y)]$ (with X and Y representing independent unpaired samples) and the Mann-Whitney U statistics. Note that, although in (2) the posterior means of $P(Z \geq -Z')$ and R^+ appear to be closely related, the posterior distribution of $P(Z \geq -Z')$ is, in general, different from that assumed for R^+ under the null hypothesis (although, if the null hypothesis is true they converge to the same distribution for large n), and thus one should not expect the frequentist and Bayesian tests to make the same decisions even for $s \to 0$. The prior expectation for $s \to 0$ depends on the choice of the prior base measure G_0. For example, by choosing G_0 symmetric around zero, we obtain a prior expectation of $1/2$. However, in a situation of complete prior ignorance, we have no reason to assign, a priori, any specific value to the probability $P(Z \geq -Z')$. Moreover, Rubin has highlighted a second critical point: the Bayesian bootstrap assigns zero posterior probability to any set that does not include the observations, since for $s \to 0$, $\mathscr{E}[P(A)|Z^n] = (\alpha(A) + n_A)/(s+n) \to 0$ when $n_A = 0$, i.e., whenever there are not observations in the set A (Rubin, 1981). This is not suitable for a Bayesian model that can be used for predictive inferences. In case of lack of prior information, a more natural way to model prior ignorance may be to consider the set of all distributions G_0 (Walley, 1991), (Walley, 1996). In other words, we keep s fixed and assume that $G_0 \in \Gamma = \{\text{all distributions}\}$, and then compute the lower and upper expectations for all the functions of interest in the statistical analysis. We call this model prior near-ignorance DP (IDP).

Theorem 3 *Given the DP prior $Dp(s, G_0)$, with $G_0 \in \Gamma$, the prior lower and upper expectations are obtained for $dG_0 = \delta_{Z_0}$ with $Z_0 < 0$ (lower) and $Z_0 > 0$ (upper), and are*

$$\underline{\mathscr{E}}[P(Z \geq -Z')] = 0, \quad \overline{\mathscr{E}}[P(Z \geq -Z')] = 1; \tag{5}$$

the posterior lower expectation of $P(Z \geq -Z')$ is obtained for $dG_0(z) = \delta_{Z_0}$ with $Z_0 < -\max|Z_i|$ and is

$$\underline{\mathscr{E}}[P(Z \geq -Z')|Z^n] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I_{[-Z_i,\infty)}(Z_j)}{(s+n)(s+n+1)} + \frac{\sum_{j=1}^{n} I_{[0,\infty)}(Z_j)}{(s+n)(s+n+1)}; \tag{6}$$

the posterior upper expectation is obtained for $dG_0(x) = \delta_{Z_0}$ with $Z_0 > \max|Z_i|$ and is

$$\overline{\mathscr{E}}[P(Z \geq -Z')|Z^n] = \frac{\sum_{i=1}^{n}\sum_{j=1}^{n} I_{[-Z_i,\infty)}(Z_j)}{(s+n)(s+n+1)} + \frac{\sum_{j=1}^{n} I_{[0,\infty)}(Z_j)}{(s+n)(s+n+1)} + \frac{s^2+2ns+s}{(s+n)(s+n+1)}. \tag{7}$$

From Theorem 3 it follows that, given the prior $Dp(\alpha(\mathscr{Z}), G_0)$, with $\alpha(\mathscr{Z}) \leq s$, the posterior expectation of $P(Z \geq -Z')$ will be bounded by $\underline{\mathscr{E}}[P(Z \geq -Z')|Z^n]$ and $\overline{\mathscr{E}}[P(Z \geq -Z')|Z^n]$ whatever is the choice of G_0. Thus, one should not worry about the fact the upper and lower expectations are obtained by extreme priors ($Dp(s, \delta_{Z_0})$), since they are only used to identify the range of values where the expectation provided by any other (smoother) prior will fall. The imprecision, defined as the difference between the upper and lower expectations, can be derived from Theorem 3 as $\overline{\mathscr{E}}[P(Z \geq -Z')|Z^n] - \underline{\mathscr{E}}[P(Z \geq -Z')|Z^n] = \frac{s^2+2ns+s}{(s+n)(s+n+1)}$, and goes to zero for large n. Thus, $\overline{\mathscr{E}}[P(Z \geq -Z')|Z^n], \underline{\mathscr{E}}[P(Z \geq -Z')|Z^n]$ tend to the asymptotic limit of R^+ for $n \to \infty$. To perform the hypothesis test, we need to know the posterior probability of $P(Z \geq -Z') > 1/2$. The next theorem gives an important result which can be used to efficiently approximate it by Monte Carlo sampling, in correspondence of the atomic priors that give the upper and lower distributions.

Theorem 4 *Consider one of the limiting priors that give the posterior lower and upper expectations in (6)–(7). Let dF_n be sampled from its posterior; then $dF_n = w_0 \delta_{Z_0} + \sum_{j=1}^{n} w_j \delta_{Z_j}$, where $(w_0, w_1, \ldots, w_n) \sim Dir(s, 1, \ldots, 1)$ and, for any $a \in [0,1]$, it holds that*

$$\begin{aligned} \underline{\mathscr{P}}\left[P(Z \geq -Z') > a|Z^n\right] &= P\left[\underline{g}(w_\cdot, Z^n) > a\right], \\ \overline{\mathscr{P}}\left[P(Z \geq -Z') > a|Z^n\right] &= P\left[\overline{g}(w_\cdot, Z^n) > a\right], \end{aligned} \tag{8}$$

with $\underline{g}(w.,Z^n) = \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j I_{[-Z_i,\infty)}(Z_j)$ *and*

$$\overline{g}(w.,X^n) = w_0(2-w_0) + \sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j I_{[-Z_i,\infty)}(Z_j),$$

and P is computed w.r.t. this Dirichlet distribution.

Based on this theorem, we can numerically approximate $\overline{\mathscr{P}}$ and $\underline{\mathscr{P}}$ by Monte Carlo sampling the vector of weights $(w_0, w_1, \ldots, w_n)$ from the Dirichlet distribution. This means that we do not need stick-breaking or other sampling techniques specific for DP.

Let a_1 be the decision of preferring classifier Y to X and a_0 its opposite; we can formulate the Bayesian test in terms of a loss function which assigns loss l_1 to an error of type I (taking the decision a_1 when Y is not better than X) and loss l_0 to an error of type II (taking the decision a_0 when Y is actually better than X). To minimize the expected loss, the decision a_0 should be preferred if

$$l_0 \mathscr{P}[P(Z \geq -Z') > \tfrac{1}{2}|Z^n] \leq l_1 \mathscr{P}[P(Z \geq -Z') \leq \tfrac{1}{2}|Z^n],$$

that is

$$\mathscr{P}[P(Z \geq -Z') > \tfrac{1}{2}|Z^n] \leq \frac{l_1}{l_0 + l_1}, \tag{9}$$

where the probability $\mathscr{P}[P(Z \geq -Z') > 1/2|Z^n]$ is evaluated from the posterior distribution of F. Notice that this decision problem is the Bayesian analogous of a frequentist one-sided test with $\alpha = l_0/(l_0+l_1)$ where we have used the posterior probability of $P(Z \geq -Z') > 1/2$ given the data in place of the likelihood of the data given the null hypothesis. In the IDP model, both the lower and upper probabilities of $P(Z \geq -Z') > 1/2$ are compared with the threshold $l_1/(l_0+l_1)$ and the decision is made based on the following rules: (i) if $\underline{\mathscr{P}} > l_1/(l_0+l_1)$ we prefer classifier Y; (ii) if $\overline{\mathscr{P}} < l_1/(l_0+l_1)$ we prefer classifier X; (iii) if $\overline{\mathscr{P}} > l_1/(l_0+l_1)$ and $\underline{\mathscr{P}} < l_1/(l_0+l_1)$ we are not able to make a decision which yields minimum expected loss for any choice of the prior measure G_0.

Thus, the IDP test can return a determinate decision only in the first two cases, whereas in the third case we are in an indeterminate situation where it is not possible to reach a decision. Notice that the fact of preferring classifier X does not imply that X is better than Y, but only that one can expect a smaller loss by choosing X. Indeed, if $l_1 > l_0$, one may prefer X even when Y is likely to be better, if the evidence in favor of Y is not sufficiently large to compensate the larger cost in case of error. As an illustrative example, Figure 1 shows the posterior upper and lower distributions of $P(Z \geq -Z')$ obtained from $n = 20$ observations Z_i sampled from a standard normal distribution. Based on these posterior estimates, the test will decide in favor of classifier Y if $l_1/(l_0+l_1) < 0.76$, and conversely in favor of classifier X if $l_1/(l_0+l_1) > 0.87$. It will be instead indeterminate if $0.76 \leq l_1/(l_0+l_1) \leq 0.87$.

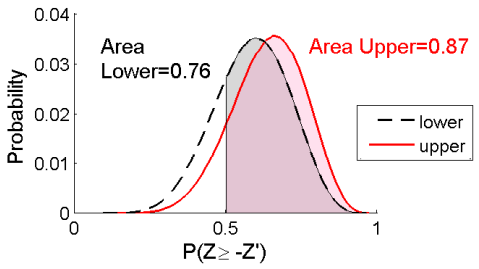

Figure 1. Posterior distributions of $P(Z \geq -Z')$. Area lower (upper) gives the value of $\underline{\mathscr{P}}[P(Z \geq -Z') > 1/2|Z^n]$ ($\overline{\mathscr{P}}[P(Z \geq -Z') > 1/2|Z^n]$), i.e. the integral of the lower (upper) distribution from $1/2$ to ∞.

Finally, by exploiting the results in (Janssen, 1994), it is possible to show that the above test is asymptotically consistent as a test for $P(Z \geq -Z')$, in the sense that the posterior lower and upper distributions of $P(Z \geq -Z')$ converge to the asymptotic distribution of the statistic R^+ (a Normal distribution). Conversely, the Wilcoxon-signed rank test is asymptotically consistent only under the assumption that the distribution F is symmetric; when this is not the case the test is not calibrated and, thus, asymptotically inconsistent. Consider this example: $P(Z \geq -Z') = 0.5$ but the median m is positive and, thus F is asymmetric (for instance $F = w\mathrm{U}[1,5] + (1-w)\mathrm{U}[-12,5]$, with $w = 0.46514$, $m = 1$ and $\mathrm{U}[a,b]$ is the uniform distribution on the interval $[a,b]$), which can happen if a classifier is slightly better than the other except in a few cases where it is worse. In this case, is one classifier better than the other? Note that the frequentist test would neither be calibrated if the answer is yes, nor powerful if it is not (the probability of rejecting H_0 converges to 6.5% for frequentist test). In this case our test coherently declares that no classifier can be preferred (it is asymptotically calibrated).

3.1. How to choose s in IDP

The value of s determines how quickly lower and upper posterior expectations converge as the number of observations increases. A way to select a value of s is by imposing that the degree of imprecision $\overline{\mathscr{E}}[P(Z \geq -Z')|Z_1] - \underline{\mathscr{E}}[P(Z \geq -Z')|Z_1]$ is reduced to a fraction of its prior value ($\overline{\mathscr{E}}[P(Z \geq -Z')] - \underline{\mathscr{E}}[P(Z \geq -Z')] = 1$) after the first observation $Z_1 = Y_1 - X_1$. 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 Bayesian bootstrap which will be shown in the next section to be less reliable than IDP). Then the intermediate value of $1/2$ is a frequent choice in prior-ignorance modeling (Pericchi & Walley, 1991; Walley, 1996). Although this is a subjective way to choose the

degree of conservativeness (indeterminacy), it represents a reasonable trade-off between the reliability and indeterminacy of the decision. From (6)–(7) it follows that $\overline{\mathscr{E}}[P(Z \geq -Z')|Z_1] - \underline{\mathscr{E}}[P(Z \geq -Z')|Z_1] = \frac{s^2+3s}{(s+1)(s+2)}$. Thus, by imposing that $\frac{s^2+3s}{(s+1)(s+2)} = \frac{1}{2}$, we obtain $s = (\sqrt{17}-3)/2$. Observe that the lower and upper probabilities produced by a value of s are always contained in the probability intervals produced by larger values of s. Then whenever we are undecided for s_1 we are also for $s_2 > s_1$. Nonetheless, for large n the distance between the upper and lower probabilities goes to zero, then also the indeterminacy goes to zero.

3.2. Managing ties

So far, it has been assumed that there is zero probability of ties ($Z_i = -Z_j$) and zeros ($Z_i = 0$). Notice that the zeros can be interpreted as ties, since $Z_i = 0 = -Z_i$. If ties are possible, the common approach to account for them is to consider the probability $[P(Z \geq -Z') + \frac{1}{2}P(Z = -Z')]$ (Sidak et al., 1999). Note that $P(Z \geq -Z') + \frac{1}{2}P(Z = -Z')$ is equal to $E[I_{(-z',\infty)}(z) + \frac{1}{2}I_{\{-z'\}}(z)]$ which in turns is equal to $E[H(z+z')]$, where $H(\cdot)$ denotes the Heaviside step function, i.e., $H(z) = 1$ for $z > 0$, $H(z) = 1/2$ for $z = 0$ and $H(z) = 0$ for $z < 0$. The procedure presented in this section is easily extended to the case of ties by substituting $I_{[-Z_i,\infty)}(Z_j)$ with $H(Z_i + Z_j)$ in the computation of $\underline{\mathscr{P}}\big[P(Z \geq -Z') > 1/2|Z^n\big]$ and $\overline{\mathscr{P}}\big[P(Z \geq -Z') > 1/2|Z^n\big]$.

4. Numerical Simulations

Consider a Monte Carlo experiment in which paired values of accuracies X_i, Y_i are generated for $n = 30$ multiple data sets based on the Gaussian models:

$$\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ \Delta \end{bmatrix}, \begin{bmatrix} \sigma^2 & \rho\sigma^2 \\ \rho\sigma^2 & \sigma^2 \end{bmatrix}\right), \qquad (10)$$

for $i = 1,\ldots,n$, with Δ (difference in accuracy) ranging from -0.07 to 0.07 and $\sigma = 0.12$. We have selected these values on the basis of extensive classification experiments performed using WEKA (Witten et al., 2011). Hereafter, due to the limited space, we only report the results for the case $\rho = 0$; the results obtained with correlation (e.g., $\rho = 0.95$) lead to similar conclusions. The aim of this section is to compare three methods to evaluate if the classifier Y is better than classifier X (i.e., $\Delta > 0$): (i) one-sided Wilcoxon signed-ranks test; (ii) Bayesian Bootstrap Dp($s = 0$); (iii) prior ignorance Dirichlet process (IDP) model. The one-sided Wilcoxon test has been implemented according to the conventional decision criterion: p-value less than $\alpha = 0.05$. To evaluate the performance of the tests, we have considered the average loss produced by each method (i.e. the proportion of wrong decisions multiplied by the corresponding loss) with different values of (l_0, l_1). Fig. 2 reports as a function of Δ and for two different values of (l_0, l_1): (i) the loss of the Dp($s = 0$) test; (ii) the loss of the Wilcoxon test; (iii) the loss of the IDP test when it is determinate; (iv) the indeterminacy of the IDP test, i.e., the number of runs it returns an indeterminate response divided by the total number of Monte Carlo runs. Let us start comparing Wilcoxon versus Dp($s = 0$). From

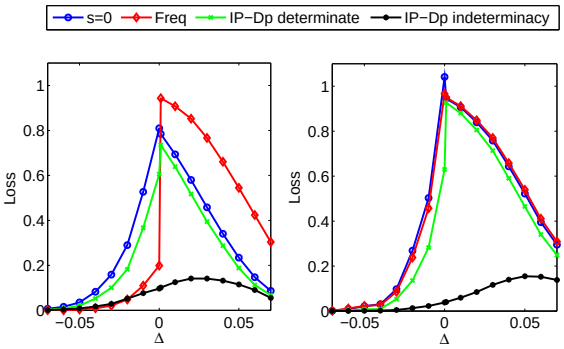

Figure 2. Loss for $l_1 = 4$ (left) and $l_1 = 19$ (right).

the plot relative to the loss $(l_0 = 1, l_1 = 4)$, it can be seen that Wilcoxon is too conservative towards the null hypothesis. When $\Delta > 0$ it has a greater loss than Dp($s = 0$) and, thus, a lower power. This conservativeness can be quantified by computing the areas under the curves in Fig. 2 for the Wilcoxon and Dp($s = 0$) tests, i.e., the average loss (averaged over the Monte Carlo runs and the values of Δ). The results are shown in Table 1 for different values of (l_0, l_1). It is evident that Dp($s = 0$) has always lower average loss than Wilcoxon, (see in particular $(l_0, l_1) = (1,4)$ which corresponds to the plot in Fig. 2 (left)). The only exception is for $(l_0, l_1) = (1,19)$ (Fig. 2 (right)). In this case, the DP-based tests declares that classifier Y is better than classifier X when the posterior probability of the hypothesis $\Delta > 0$ is greater than $1 - \alpha = 0.95$. Thus for this choice of the loss, Wilcoxon and DP are closely matched. However, the choice $(l_0, l_1) = (1,19)$ is extremely conservative, implying that the cost associated to a Type I error is 19-times greater than the cost associated to a Type II error. Consider, for instance, that clinical trials (Spiegelhalter et al., 1994) are usually designed to have significance of 5% and power 80 or 90% (the computation of power requires doing a number of assumptions). We can roughly infer that in these cases a Type II error is regarded about two or four times worse than a Type I error. The advantage of the Bayesian test is that one can make decisions minimizing the expected loss. Conversely, in the frequentist Wilcoxon test, one always makes the myopic choice $\alpha = 0.05$ regardless of the loss. Table 1 shows that for a wide variety of values of cost configurations, including $(l_0, l_1) = (1,2)$ and $(l_0, l_1) = (1,4)$, the Bayesian test incurs much lower loss than the Wilcoxon signed-rank. To

Loss, $l_0=1$	$l_1=1$	$l_1=2$	$l_1=4$	$l_1=9$	$l_1=19$
Dp($s=0$)	.025	.034	.044	.053	.061
Wilcoxon	.048	.049	.050	.054	.061

Table 1. Total average loss.

Loss, $l_0=1$	$l_1=1$	$l_1=2$	$l_1=4$	$l_1=9$	$l_1=19$
IDP	.023	.031	.040	.049	.057
Dp($s=0$)	.023	.031	.040	.049	.057
Wilcoxon	.047	.047	.048	.051	.057

Table 2. Average loss in the IDP determinate cases.

% H_0/H_1	$l_1=1$	$l_1=2$	$l_1=4$	$l_1=9$	$l_1=19$
Dp($s=0$)	45 / 45	43 / 48	43/ 50	43/ 54	42/ 55
Wilcoxon	100 / 0	100/ 0	100/ 0	100/ 0	50/ 50

Table 3. % of H_0/H_1 decisions in the IDP indeterminate instances averaged over Δ for $l_0=1$ and different values of l_1.

compare Wilcoxon and Dp($s=0$) with IDP, we distinguish two cases: (i) the instances in which IDP is determinate; (ii) the instances in which IDP is indeterminate. The loss (averaged w.r.t. Δ and the Monte Carlo runs) for the first case is shown in Table 2, while for the second case Table 3 reports the percentage of times the Wilcoxon and Dp($s=0$) tests have returned a wrong decision in the two cases where the truth is H_0 or, respectively, H_1. From Tables 2–3, it can respectively be seen that: (i) in the IDP determinate instances the loss of Dp($s=0$) coincides with that of IDP; (ii) in the IDP indeterminate instances Dp($s=0$) is almost a random guesser. For $l_1<19$, Wilcoxon test has always greater loss than that of Dp($s=0$) and IDP in the determinate instances and it always returns H_0 in the indeterminate instances. The only exception is the case $l_1=19$ where the losses coincide in the determinate instances while, in the indeterminate ones, Wilcoxon is a perfect random guesser. From Fig. 2 (right) it can be seen that the percentage of runs in which IDP is indeterminate is high (e.g., about 16% for $\Delta=0.05$); this means that Dp($s=0$) and Wilcoxon are issuing an almost random answer in 16% of the cases, which is a large percentage (a similar comment can be done for Dp($s=0$) in the case $l_1\leq 19$, see in particular $l_1=4$ in Fig. 2 (left)). For $(l_0,l_1)=(1,19)$, since in the determinate instances Wilcoxon and IDP have the same loss and in the indeterminate ones Wilcoxon is a random guesser, we could paradoxically design a new test that coincides with IDP in the IDP determinate instances and issues a random answer in the indeterminate ones that overall has the same loss of Wilcoxon. This shows that IDP is more reliable than Wilcoxon. In fact, assume that one is trying to compare the accuracy of two classifiers to determine if "Y is better than X" and that, given the available data, IDP is indeterminate. In such a situation the Wilcoxon test always issues a determinate response (pretending to be able to conclude whether "Y is better than X" or not), but its response is simply random (like tossing a coin). On the other side, the IDP acknowledges the impossibility of making a decision (I do not know whether "Y is better than X"). In such cases one knows that (i) her/his posterior decisions would depend on the choice of G_0; (ii) reaching a decision given the observed data is difficult, and in fact the Wilcoxon behaves like a random guesser. Based on the indeterminate outcome of the IDP test, one can for example decide to run the classifiers on additional datasets to eliminate the indeterminacy (in fact when the number of observations goes to infinity the indeterminacy goes to zero).

4.1. Practical case studies

We consider three different classifiers: naive Bayes (NB) and two variants of the tree-augmented naive Bayes (TAN) (Friedman et al., 1997) which differ as for the score used for learning the TAN structure. We denote such two variants of TAN as TAN_{BDeu} and TAN_{mdl}. We run the WEKA implementation (Witten et al., 2011) of such classifiers on 70 data sets from the UCI repository: 54 classification data sets and 16 regression data sets, which we use for classification having discretized into 4 bins the target variable. We evaluate via 10 folds cross-validation the accuracy of each classifier on each data set. Then we compare pairs of classifiers via the Wilcoxon signed-rank test and its two novel Bayesian variants (Table 4.1). We run the tests in a one-sided fashion. When comparing NBC with a TAN the null hypothesis is that the median accuracy of NBC is no smaller than that of TAN; the alternative hypothesis is that the median accuracy of TAN is instead greater than that of NBC. The three tests consistently identify both TANs as significantly more accurate than NBC. Indeed, TAN is well-known to perform better than naive Bayes (Friedman et al., 1997). For both TANs, the DP(s=0) returns probability 1 for the alternative hypothesis. Given the large sample size (n=70), the upper and lower posterior probability of the alternative hypothesis computed by IDP collapse on a single point, namely 1. On the other hand, the three tests consistently report no significant difference between the two TANs. The lower and upper posteriors for this last case are shown in Fig. 3.

Pair of classifiers	Wilcoxon p-value	DP(s=0) $P(H_1\|D)$	IDP $[\underline{P}(H_1\|D), \overline{P}(H_1\|D)]$
NBC-TAN_{mdl}	1e-06	1	[1, 1]
NBC-TAN_{BDeu}	1e-07	1	[1, 1]
TAN_{BDeu}-TAN_{mdl}	.79	.26	[.23. .30]

Table 4. Statistical comparison of pair of classifiers. We report the posterior probability of H_1 for the test DP(s=0) and the interval of the posterior probability of H_1 for IDP.

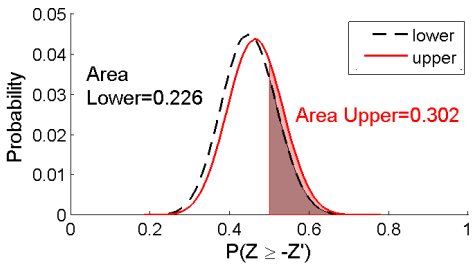

Figure 3. Posterior probability for TAN_{BDeu}-TAN_{mdl}.

4.2. Replicability analysis

According to (Bouckaert, 2004), a desirable test has low Type I error, high power and high *replicability*. The replicability is the probability that the same conclusion is achieved in two experiments involving the same pair of classifiers (i.e., the null hypothesis is accepted or rejected in both cases). We follow the experimental setup of (Demšar, 2006). We randomly draw 15 data sets among the 70 available. We repeat the drawing 1000 times. Every time we run the statistical tests to compare the accuracy of classifiers on the drawn data sets. We consider two pairs of classifiers: NB–TAN_{mdl} and NB–TAN_{BDeu}. This yields 1000 experiments for each pair of classifiers. In the following we describe how to measure replicability of a hypothesis test. The outcome e_i of the hypothesis test in the i-th experiment is 0 or 1 depending on whether the null hypothesis is accepted or rejected. The replicability R is defined as:

$$R = 1 - 2\frac{\sum_i (e_i - \overline{e})^2}{n-1},$$

where $\overline{e}$ is the mean outcome of the hypothesis test over the n=1000 repetitions. Thus R ranges between 0.5 (random decisions) and 1 (perfectly repeatable decisions). To allow a fair measure of repeatability we focus on the loss $(l_0, l_1) = (1, 19)$, in which case the decisions of the frequentist (α=0.05) and that of the Bayesian tests are closely matched, as already discussed. The IDP test suspends the judgment becoming *indeterminate* when the decision is *prior-dependent*, namely when the loss is minimized by returning either H_0 or H_1 depending on the prior. We separately evaluate the replicability of the decisions made by the tests when the IDP test is determinate and indeterminate, as reported in Tab. 5. Strikingly, a sharp drop of replicability affects both the Wilcoxon and the DP(s=0) test when the IDP test becomes indeterminate. For both tests, when assessing both pairs of classifiers, the replicability drops from about 90% to about 50%. In practice both the Wilcoxon and the DP(s=0) test behave as random guessers when the IDP is indeterminate.

Pair	$\%H_1$	%Ind	Replicability			
			Wilcoxon		DP (s=0)	
			Det	Ind	Det	Ind
$\text{NB-TAN}_{\text{mdl}}$	87	14	.88	.50	.88	.56
TAN_{mdl} - TAN_{BDeu}	6	8	.94	.51	.95	.50

Table 5. Replicability results. We denote by $\%H_1$ the proportion of times in which the Wilcoxon test rejects the null hypothesis and by %Ind the proportion of times in which the IDP test becomes indeterminate.

The behavior of the IDP test *cannot* be mimicked by a reject option, which would suspend the judgment whenever the p-value of the frequentist test is close to 0.05. The IDP test checks whether the decision to be taken is prior-dependent, yielding a more complex behavior than a reject option. On the one hand the IDP test does *not always* gets indeterminate when the p-value is close to 0.05; on the other hand, in some cases it does get indeterminate when the p-value is quite far from 0.05. This means that the indeterminacy of IDP does not only depend on the p-value , but on the observations (not only the statistic). However, the median p-value of the cases in which IDP suspends the decision is close to 0.05. Moreover, the p-values of the cases in which IDP suspends the judgment are (almost) symmetrically distributed around 0.05. This explains why the replicability of the Wilcoxon test drops down to 50% in the IDP indeterminate instances.

5. Conclusions

We have proposed a novel Bayesian method based on the Dirichlet Processes (DP) for performing the Wilcoxon signed-rank test. We have developed two tests: one based on a noninformative prior and one based on a conservative model of prior ignorance (IDP). The Bayesian approach is more flexible than the frequentist one, as it allows for taking decision which minimize the expected loss. Experimental results show that the prior ignorance method is more reliable than both the frequentist test and the noninformative Bayesian one, being able to isolate instances in which these tests are almost guessing at random. We plan to extend this approach to implement Bayesian versions of multiple nonparametric tests such as for instance the Friedman test. In the long run, our aim is to build a statistical package for Bayesian nonparametric tests.

Acknowledgments

This work was partly supported by the Swiss NSF grants nos. 200021_146606 / 1 and 200020_137680 / 1.

References

Berger, J. O., Moreno, E., Pericchi, L. R., Bayarri, M. J., Bernardo, et al. An overview of robust Bayesian analysis. *Test*, 3(1):5–124, 1994.

Berger, James O., Rios Insua, David, and Ruggeri, Fabrizio. Bayesian robustness. In *Robust Bayesian Analysis*, volume 152 of *Lecture Notes in Statistics*, pp. 1–32. Springer New York, 2000.

Borgwardt, Karsten M and Ghahramani, Zoubin. Bayesian two-sample tests. *arXiv preprint arXiv:0906.4032*, 2009.

Bouckaert, Remco R. Estimating replicability of classifier learning experiments. In *Proc. of the twenty-first International Conference on Machine Learning*, pp. 15–22, 2004.

Chen, Yuhui and Hanson, Timothy E. Bayesian nonparametric k-sample tests for censored and uncensored data. *Computational Statistics and Data Analysis*, 71(C):335–346, 2014.

Coolen-Schrijner, Pauline, Coolen, Frank PA, Troffaes, Matthias CM, and Augustin, Thomas. Imprecision in statistical theory and practice. *Journal of Statistical Theory and Practice*, 3(1):1–9, 2009.

Corani, Giorgio and Zaffalon, Marco. Learning reliable classifiers from small or incomplete data sets: the naive credal classifier 2. *The Journal of Machine Learning Research*, 9:581–621, 2008.

Dalal, S.R. and Phadia, E.G. Nonparametric Bayes inference for concordance in bivariate distributions. *Communications in Statistics-Theory and Methods*, 12(8):947–963, 1983.

Demšar, Janez. Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7:1–30, 2006.

Derrac, Joaquín, García, Salvador, Molina, Daniel, and Herrera, Francisco. A practical tutorial on the use of nonparametric statistical tests as a methodology for comparing evolutionary and swarm intelligence algorithms. *Swarm and Evolutionary Computation*, 1(1):3–18, 2011.

Ferguson, Thomas S. A Bayesian Analysis of Some Nonparametric Problems. *The Annals of Statistics*, 1(2):pp. 209–230, 1973. ISSN 00905364.

Friedman, Nir, Geiger, Dan, and Goldszmidt, Moises. Bayesian network classifiers. *Machine learning*, 29(2-3):131–163, 1997.

Goodman, Steven N. Toward evidence-based medical statistics. 1: The p–value fallacy. *Annals of internal medicine*, 130(12):995–1004, 1999.

Holmes, C.C, Caron, F., Griffin, J.E., and Stephens, D.A. Two-sample Bayesian nonparametric hypothesis testing. *arXiv preprint arXiv:0910.5060*, 2009.

Janssen, Paul. Weighted bootstrapping of U–statistics. *Journal of statistical planning and inference*, 38(1):31–41, 1994.

Johnson, Valen E. Revised standards for statistical evidence. *Proceedings of the National Academy of Sciences*, 110(48):19313–19317, 2013.

Kruschke, John K. Bayesian data analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, 1(5):658–676, 2010.

Ma, Li and Wong, Wing Hung. Coupling optional Pólya trees and the two sample problem. *Journal of the American Statistical Association*, 106(496), 2011.

Pericchi, L. R. and Walley, P. Robust Bayesian credible intervals and prior ignorance. *International Statistical Review*, pp. 1–23, 1991.

Raftery, Adrian E. Bayesian model selection in social research. *Sociological methodology*, 25:111–164, 1995.

Rubin, Donald B. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):pp. 130–134, 1981. ISSN 00905364.

Sidak, Z., Sen, P.K., and Hajek, J. *Theory of Rank Tests*. Probability and Mathematical Statistics. Elsevier Science, 1999. ISBN 9780080519104.

Spiegelhalter, David J, Freedman, Laurence S, and Parmar, Mahesh KB. Bayesian approaches to randomized trials. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, pp. 357–416, 1994.

Susarla, V. and Ryzin, J. Van. Nonparametric bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association*, 71(356):pp. 897–902, 1976. ISSN 01621459.

Trawiński, B., Graczyk, M., Telec, Z, and Lasota, T. Nonparametric statistical analysis of machine learning algorithms for regression problems. In *Knowledge-Based and Intelligent Information and Engineering Systems*, pp. 111–120. Springer, 2010.

Walley, P. *Statistical Reasoning with Imprecise Probabilities*. Chapman and Hall, New York, 1991.

Walley, P. Inferences from multinomial data: learning about a bag of marbles. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):3–57, 1996.

Witten, Ian H, Frank, Eibe, and Hall, Mark A. *Data Mining: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.