A Bayesian nonparametric procedure for comparing algorithms

Alessio Benavoli ALESSIO@IDSIA.CH
Francesca Mangili FRANCESCA@IDSIA.CH
Giorgio Corani GIORGIO@IDSIA.CH
Marco Zaffalon ZAFFALON@IDSIA.CH
IDSIA, Manno, Switzerland

Abstract

A fundamental task in machine learning is to compare the performance of multiple algorithms. This is usually performed by the frequentist Friedman test followed by multiple comparisons. This implies dealing with the well-known shortcomings of null hypothesis significance tests. We propose a Bayesian approach to overcome these problems. We provide three main contributions. First, we propose a nonparametric Bayesian version of the Friedman test using a Dirichlet process (DP) based prior. We show that, from a Bayesian perspective, the Friedman test is an inference for a multivariate mean based on an ellipsoid inclusion test. Second, we derive a *joint* procedure for the multiple comparisons which accounts for their dependencies and which is based on the posterior probability computed through the DP. The proposed approach allows verifying the null hypothesis, not only rejecting it. Third, as a practical application we show the results in our algorithm for *racing*, i.e. identifying the best algorithm among a large set of candidates sequentially assessed. Our approach consistently outperforms its frequentist counterpart.

1. Introduction

A fundamental task in machine learning is to compare multiple algorithms. The non-parametric Friedman test (Demšar, 2006) is recommended to this end. The advantages of the non-parametric approach are that it does *not* average measures taken on different data sets; it does *not* assume normality of the sample means; it is *robust* to outliers. The Friedman test is a null-hypothesis significance tests. It thus controls the Type I error, namely the probability of rejecting the null hypothesis when it is true. The framework of null hypothesis significance tests has however important drawbacks. For instance one usually sets the significance level to 0.01 or 0.05, without having the possibility of an optimal trade-off between Type I and Type II errors.

Instead, Bayesian tests of hypothesis (Kruschke, 2010) estimate the posterior probability of the null and of the alternative hypothesis, which allows to take decisions which minimize the posterior risk. See (Kruschke, 2010) for a detailed discussion of further advantages of Bayesian hypothesis testing. However there are currently no Bayesian counterparts of the Friedman test. Our first contribution is a Bayesian version of the Friedman test. We adopt the Dirichlet process (DP) (Ferguson, 1973) as a model for the prior. The DP has been already used to develop Bayesian counterparts of frequentist non-parametric estimators such as Kaplan-Meier (Susarla & Van Ryzin, 1976), the Kendall's tau (Dalal & Phadia, 1983) and the Wilcoxon signed-rank test (Benavoli et al., 2014). Our novel derivation shows that, from a Bayesian perspective, the Friedman test is an inference for a multivariate mean based on an ellipsoid inclusion test.

When the frequentist Friedman test rejects the null hypothesis, a series of multiple pairwise comparison are performed in order to assess which algorithm is significantly different from which. The traditional post-hoc analysis controls the family-wise Type I error (FWER), namely the probability of finding at least one Type I error among the null hypotheses which are rejected when performing the multiple comparisons. The traditional Bonferroni correction for multiple comparisons controls the FWER but yields too conservative inferences. More modern approaches for controlling the FWER are discussed in (Demšar, 2006; Garcia & Herrera, 2008). All such approaches simplistically treat the multiple comparisons as independent from each other. But when comparing algorithms $\{a, b, c\}$, the outcome of the comparisons (a,b), (a,c), (b,c) are *not* independent.

Proceedings of the 32^{nd} International Conference on Machine Learning, Lille, France, 2015. JMLR: W&CP volume 37. Copyright 2015 by the author(s).

Our second contribution is a *joint* procedure for the analysis of the multiple comparisons which accounts for their dependencies. We adopt the definition of Bayesian multiple comparison of (Gelman & Tuerlinckx, 2000). We analyze the posterior probability computed through the Dirichlet process, identifying statements of *joint* comparisons which have high posterior probability. The proposed procedure is a compromise between controlling the FWER and performing no correction of the significance level for the multiple comparisons. Our Bayesian procedure produces more Type I errors but less Type II errors than procedures which controls the family-wise error. In fact, it does not aim at controlling the family-wise error. On the other hand, it produces both less Type I and less Type II errors than running independent tests without correction for the multiple comparison, thanks to its ability in capturing the dependencies of the multiple comparisons.

Our overall procedure thus consists of a Bayesian Friedman test, followed by a joint analysis of the multiple comparison.

1.1. Racing

Racing addresses the problem of identifying the best algorithms among a large set of candidates. Racing is particularly useful when running the algorithms is time-consuming and thus they cannot be evaluated many times, for instance when comparing different configurations of time-consuming optimization heuristics. The idea of racing (Maron & Moore, 1997) is to test the set of algorithms in parallel and to discard early those recognized as inferior. Inferior candidates are recognized through statistical tests. The race thus progressively focuses on the better models.

There are both frequentist and Bayesian approaches for racing. A peculiar feature of Bayesian algorithms (Maron & Moore, 1997; Chien et al., 1995) is that they are able to eliminate also models whose performance is very similar with high probability (*indistinguishable models*). Detecting two indistinguishable models correspond to *accept* the null hypothesis that their performance is equivalent. It is instead impossible to accept the null hypothesis for a frequentist test. However the applicability of such Bayesian racing approaches (Maron & Moore, 1997; Chien et al., 1995) is restricted, as they assume the normality of observations.

Non-parametric frequentist are generally preferred for racing (Birattari, 2009, Chap.4.4). A popular racing algorithm is the F-race (Birattari et al., 2002). It first performs the frequentist Friedman test, followed by the multiple comparisons if the null hypothesis of the Friedman is rejected. Being frequentist, the F-race cannot eliminate indistinguishable models.

None of the racing procedures discussed so far can exactly compute the probability (or the confidence) of a series of sequential statements which are issued during the multiple comparisons, since they treat the multiple comparisons as independent.

Our novel procedure allows both to detect indistinguishable models and to model the dependencies between the multiple comparisons.

2. Friedman test

The comparison of multiple algorithms are organized in the following matrix:

$$\begin{array}{cc} & \textit{Performance on different cases} \\ \textit{Algorithms} & \begin{matrix} X_{11} & X_{12} & \dots & X_{1n} \\ X_{21} & X_{22} & \dots & X_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ X_{m1} & X_{m2} & \dots & X_{mn} \end{matrix} \end{array} \tag{1}$$

where X_{ij} denotes the performance of the i-th algorithm on the j-th dataset (for $i = 1, \dots, m$ and $j = 1, \dots, n$). The performance of algorithms on different cases can refer for instance to the accuracy of different classifiers on multiple data sets (Demšar, 2006) or the maximum value achieved by different solvers on different optimization problems (Birattari et al., 2002).

The performances obtained on different data sets are assumed to be independent. The algorithms are then ranked column-by-column obtaining the matrix $\mathbf{R}$ of the ranks R_{ij}, $i = 1, \dots, m$ and $j = 1, \dots, n$, with R_{ij} the rank of the i-th algorithm with respect to to the other observations in the j-th column. The row sum $\sum_{i=1}^{m} R_{ij} = m(m+1)/2$ is a constant (i.e., the sum of the first m integer), while the column sum $R_i = \sum_{j=1}^{n} R_{ij}$, $i = 1, \dots, m$, is affected by the differences between the algorithms. The null hypothesis of the Friedman test is that the different samples are drawn from populations with identical medians. Under the null hypothesis the statistic

$$Q = \frac{12}{nm(m+1)} \sum_{j=1}^{n} \left[R_j - \frac{n(m+1)}{2} \right]^2,$$

is approximately chi-square distributed with $m - 1$ degress of freedom. The approximation is reliable when $n > 7$. We denote by γ the significance of the test. The null hypothesis is rejected when the statistic exceed the critical value, namely when:

$$Q > \chi^2_{m-1,\gamma}. \tag{2}$$

For $m = 2$ the Friedman test reduces to a sign test.

3. Directional multiple comparisons

If the Friedman test rejects the null hypothesis one has to establish which are the significantly different algorithms. The commonly adopted statistic for comparing the i-th and the j-th algorithm is (Demšar, 2006; Garcia & Herrera, 2008):

$$z = (R_i - R_j)/\sqrt{\frac{m(m+1)}{6n}},$$

which is normally distributed. Comparing this statistics with the critical value of the normal distribution yields the *mean ranks test*. A shortcoming of the mean rank test is that the decisions regarding algorithms i, j depends also on the scores of all the other algorithms in the set. This is a severe issue which can have a negative impact both on Type I and the Type II errors (Miller, 1966; Gabriel, 1969; Fligner, 1984); see also (Hollander et al., 2013, Sec. 7.3).

Alternatively, one can compare algorithm i to algorithm j via either the *Wilcoxon signed-rank test* or the *sign test* (Conover & Conover, 1980). The Wilcoxon signed-rank test has more power than the sign test, but it assumes that the distribution of the data is symmetric w.r.t. its median. When this is not the case, the Wilcoxon signed-rank test is not calibrated. Conversely, the sign test does not require any assumption on the distribution of the data. Pairwise comparisons are often performed in a two-sided fashion. In this case a Type I error correspond to a claim of difference between two algorithms, while the two algorithms have instead identical performance. However, it is hardly believable that two algorithms have an actually identical performance. It is thus more interesting running one-sided comparisons. We thus perform the multiple comparisons in a directional fashion (Williams et al., 1999): for each pair i, j of algorithms we select the one-sided comparison yielding the smallest p-value. A Type S error (Gelman & Tuerlinckx, 2000) is done every time we wrongly claim that algorithm i is better than j, while the opposite is instead true. The output of this procedure will consist in the set of all directional claims for which the p-value is smaller than a suitably corrected threshold.

4. Dirichlet process

The Dirichlet process was developed by Ferguson (Ferguson, 1973) as a probability distribution on the space of probability distributions. Let $\mathbb{X}$ be a standard Borel space with Borel σ-field $\mathcal{B}_{\mathbb{X}}$ and $\mathbb{P}$ be the space of probability measures on $(\mathbb{X}, \mathcal{B}_{\mathbb{X}})$ equipped with the weak topology and the corresponding Borel σ-field $\mathcal{B}_{\mathbb{P}}$. Let $\mathbb{M}$ be the class of all probability measures on $(\mathbb{P}, \mathcal{B}_{\mathbb{P}})$. We call the elements $\mu \in \mathbb{M}$ nonparametric priors. An element of $\mathbb{M}$ is called a Dirichlet process distribution $\mathcal{D}(\alpha)$ with base measure α if for every finite measurable partition $B_1, \dots, B_m$ of $\mathbb{X}$, the vector $(P(B_1), \dots, P(B_m))$ has a Dirichlet distribution with parameters $(\alpha(B_1), \dots, \alpha(B_m))$, where $\alpha(\cdot)$ is a finite positive Borel measure on $\mathbb{X}$. Consider the partition $B_1 = A$ and $B_2 = A^c = \mathbb{X} \backslash A$ for some measurable set $A \in \mathbb{X}$, then if $P \sim \mathcal{D}(\alpha)$, let $s = \alpha(\mathbb{X})$ stand for the total mass of $\alpha(\cdot)$, from the definition of the DP we have that $(P(A), P(A^c)) \sim Dir(\alpha(A), s - \alpha(A))$, which is a Beta distribution. From the moments of the Beta distribution, we can thus derive that:

$$\mathcal{E}[P(A)] = \frac{\alpha(A)}{s}, \quad \mathcal{V}[P(A)] = \frac{\alpha(A)(s-\alpha(A))}{s^2(s+1)}, \tag{3}$$

where we have used the calligraphic letters $\mathcal{E}$ and $\mathcal{V}$ to denote expectation and variance w.r.t. the Dirichlet process. This shows that the normalized measure $\alpha^*(\cdot) = \alpha(\cdot)/s$ of the DP reflects the prior expectation of P, while the scaling parameter s controls how much P is allowed to deviate from its mean. If $P \sim \mathcal{D}(\alpha)$, we shall also describe this by saying $P \sim Dp(s, \alpha^*)$. Let $P \sim Dp(s, \alpha^*)$ and f be a real-valued bounded function defined on $(\mathbb{X}, \mathcal{B})$. Then the expectation with respect to the Dirichlet process of $E[f]$ is

$$\mathcal{E}\big[E(f)\big] = \mathcal{E}\left[\int f dP\right] = \int f d\mathcal{E}[P] = \int f d\alpha^*. \tag{4}$$

One of the most remarkable properties of the DP priors is that the posterior distribution of P is again a DP. Let $X_1, \dots, X_n$ 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

$$P|X_1, \dots, X_n \sim Dp\left(s+n, \frac{s\alpha^* + \sum_{i=1}^n \delta_{X_i}}{s+n}\right), \tag{5}$$

where δ_{X_i} is an atomic probability measure centered at X_i. The Dirichlet process satisfies the property of conjugacy, since the posterior for P is again a Dirichlet process with updated unnormalized base measure $\alpha + \sum_{i=1}^n \delta_{X_i}$. From (3),(4) and (5) we can easily derive the posterior mean and variance of $P(A)$ and, respectively, posterior expectation of f. Some useful properties of the DP (Ghosh & Ramamoorthi (2003, Ch.3)) are the following:

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

(b) 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 (Ghosh & Ramamoorthi, 2003, Sec.3.1.1):

$$w_1 P_1 + w_2 P_2 \sim Dp\left(s_1 + s_2, \frac{s_1\alpha_1^* + s_2\alpha_2^*}{s_1 + s_2}\right). \tag{6}$$

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

$$P = w_0 P_0 + \sum_{i=1}^n w_i \delta_{X_i}, \tag{7}$$

where $(w_0, w_1, \dots, w_n) \sim Dir(s, 1, \dots, 1)$ and $P_0 \sim Dp(s, \alpha^*)$ (it follows by (a)-(b)).

5. A Bayesian Friedman test

Let us denote with $\mathbf{X}$ the vector of performances $[X_1, \dots, X_m]^T$ for m algorithms so that the records algorithms/dataset can be rewritten as

$$\mathbf{X}^n = \{\mathbf{X}_1, \dots, \mathbf{X}_n\}, \tag{8}$$

that is a set of n vector valued observations of $\mathbf{X}$, i.e., $\mathbf{X}_j$ coincides with the j-column of the matrix in (1). Let P be the unknown distribution of $\mathbf{X}$, assume that the prior distribution of P is $Dp(s, \alpha^*)$, our goal is to compute the posterior of P. From (5), we know that the posterior of P is

$$Dp\left(s + n, \alpha_n^* = \frac{s\alpha^* + \sum_{i=1}^n \delta_{\mathbf{X}_i}}{s + n}\right). \tag{9}$$

We adopt this distribution to devise a Bayesian counterpart of the Friedman's hypothesis test.

5.1. The case $m = 2$ - Bayesian sign test

Assume there are only two algorithms $\mathbf{X} = [X_1, X_2]^T$, in the next sections we will show how to assess if algorithm 2 is better than algorithm 1 (one-sided test), i.e., $X_2 > X_1$, and how to assess if there is a difference between the two algorithms (two-sided test), i.e., $X_1 \neq X_2$. The next section shows how to compare two algorithms. In particular we show how to assess the probability that a score randomly drawn for algorithm 1 is higher than a score randomly drawn for algorithm 2, $P(X_2 > X_1)$. This eventually leads to the design of a one-sided test. We then discuss how to test the hypothesis of the score of two algorithms being originated from distributions with significantly different medians (two-sided test).

5.1.1. One-sided test

In probabilistic terms, algorithm 2 is better than algorithm 1, i.e., $X_2 > X_1$, if:

$$P(X_2 > X_1) > P(X_2 < X_1) \text{ equiv. } P(X_2 > X_1) > \frac{1}{2},$$

where we have assumed that X_1 and X_2 are continuous so that $P(X_2 = X_1) = 0$. In Section 6 we will explain how to deal with the presence of ties $X_1 = X_2$. The Bayesian approach to hypothesis testing defines a loss function for each decision:

$$L(P, a) = \begin{cases} K_0 I_{\{P(X_2 > X_1) > 0.5\}} & \text{if } a = 0, \\ K_1 I_{\{P(X_2 > X_1) \leq 0.5\}} & \text{if } a = 1. \end{cases} \tag{10}$$

where the first row gives the loss we incur by taking the action $a = 0$ (i.e., declaring that $P(X_2 > X_1) \leq 0.5$) when actually $P(X_2 > X_1) > 0.5$, while the second row gives the loss we incur by taking the action $a = 1$ (i.e., declaring that $P(X_2 > X_1) > 0.5$) when actually $P(X_2 > X_1) \leq 0.5$. Then, it computes the expected value of this loss w.r.t. the posterior distribution of P:

$$\mathcal{E}\left[L(P, a)\right] = \begin{cases} K_0 \mathcal{P}\left[P(X_2 > X_1) > 0.5\right] & \text{if } a = 0, \\ K_1 \mathcal{P}\left[P(X_2 > X_1) \leq 0.5\right] & \text{if } a = 1, \end{cases} \tag{11}$$

where we have exploited the fact that $\mathcal{E}[I_{\{P(X_2 > X_1) > 0.5\}}] = \mathcal{P}[P(X_2 > X_1) > 0.5]$ (here we have used the calligraphic letter $\mathcal{P}$ to denote probability w.r.t. the DP). Thus, we choose $a = 1$ if

$$\mathcal{P}\left[P(X_2 > X_1) > 0.5\right] \quad > \quad \tfrac{K_1}{K_1 + K_0}, \tag{12}$$

and $a = 0$ otherwise. When the above inequality is satisfied, we can declare that $P(X_2 > X_1) > 0.5$ with probability $\frac{K_1}{K_0 + K_1}$. For comparison with the traditional test we will take $\frac{K_1}{K_0 + K_1} = 1 - \gamma$. However the Bayesian approach allows to set the decision rule in order to optimize the posterior risk. Optimal Bayesian decision rules for different types of risk are discussed for instance by (Müller et al., 2004).

Let us now compute $\mathcal{P}\left[P(X_2 > X_1) > 0.5\right]$ for the DP in (9). From (7) with $P \sim Dp\left(s + n, \alpha_n^*\right)$, it follows that:

$$P(X_2 > X_1) = w_0 P_0(X_2 > X_1) + \sum_{i=1}^n w_i I_{\{X_{2i} > X_{1i}\}},$$

where $(w_0, w_1, \dots, w_n) \sim Dir(s, 1, \dots, 1)$ and $P_0 \sim Dp(s, \alpha^*)$. The sampling of P_0 from $Dp(s, \alpha^*)$ should be performed via stick breaking. However, if we take $\alpha^* = \delta_{\mathbf{X}_0}$, we know from property (a) in Section 4 that $P_0 = \delta_{\mathbf{X}_0}$ and thus we have that

$$\mathcal{P}\left[P(X_2 > X_1) > 0.5\right] = P_{Dir}\left[\sum_{i=0}^n w_i I_{X_{2i} > X_{1i}} > \frac{1}{2}\right], \tag{13}$$

where P_{Dir} is the probability w.r.t. the Dirichlet distribution $Dir(s, 1, \dots, 1)$. In other words, as the prior base measure is discrete, also the posterior base measure $\alpha_n = s\delta_{\mathbf{X}_0} + \sum_{i=1}^n \delta_{\mathbf{X}_i}$ is discrete with finite support $\{\mathbf{X}_0, \mathbf{X}_1, \dots, \mathbf{X}_n\}$. Sampling from such DP reduces to sampling the probability w_i of each element $\mathbf{X}_i$ in the support from a Dirichlet distribution with parameters $(\alpha(\mathbf{X}_0), \alpha(\mathbf{X}_1), \dots, \alpha(\mathbf{X}_n)) = (s, 1, \dots, 1)$.

Hereafter, for simplicity, we will therefore assume that $\alpha^* = \delta_{\mathbf{x}}$. In Section 7, we will give more detailed justifications for this choice. Notice, however, that the results of the next sections can easily be extended to general α^*.

5.1.2. Two-sided test

In the previous section, we have derived a Bayesian version of the one-sided hypothesis test, by calculating the posterior probability of the hypotheses to be compared. This approach cannot be used to test the two-sided hypothesis $P(X_2 > X_1) = 0.5$ ($a = 0$) against $P(X_2 > X_1) \neq 0.5$ ($a = 1$) since $P(X_2 > X_1) = 0.5$ is a point null hypothesis and, thus, its posterior probability is zero. A way to approach two-sided tests in Bayesian analysis is to use the $(1-\gamma)\%$ symmetric (or equal-tail) credible interval (SCI) (Gelman et al., 2013; Kruschke, 2010). If 0.5 lies outside the SCI of $P(X_2 > X_1)$, we take the decision $a = 1$, otherwise $a = 0$:

$$\begin{aligned} &a = 0 \text{ if } 0.5 \in (1-\gamma)\% \text{ SCI}(P(X_2 > X_1)),\\ &a = 1 \text{ if } 0.5 \notin (1-\gamma)\% \text{ SCI}(P(X_2 > X_1)). \end{aligned}$$

Since $P(X_2 > X_1)$ is univariate distributed, its SCI can be computed as follows: $\text{SCI}(P(X_2 > X_1)) = [c, d]$, with c, d are respectively the $\gamma/2\%$ and $(1-\gamma/2)\%$ percentiles of the distribution of $P(X_2 > X_1)$.

5.2. The case $m \geq 3$ - Bayesian Friedman test

The aim of this section is to derive a DP based Bayesian version of the Friedman test. To obtain this new test, we generalize the approach derived in the previous section for the two-sided hypothesis test. Consider the following function:

$$R(X_i) = \sum_{i \neq k=1}^{m} I_{\{X_i > X_k\}} + 1, \tag{14}$$

that is the sum of the indicators of the events $X_i > X_k$, i.e., the algorithm i is better than algorithm k. Therefore, $R(X_i)$ gives the rank of the algorithm i among the m algorithms we are considering. The constant 1 has been added to have $1 \leq R(X_i) \leq m$, i.e., minimum rank one and maximum rank m. Our goal is to test the point null hypothesis that $E[R(X_1), \dots, R(X_m)]^T$ is equal to $[(m+1)/2, \dots, (m+1)/2]^T$, where $(m+1)/2$ is the mean rank of a algorithm under the hypothesis that they have the same performance. To test this point null hypothesis, we can check whether:

$$\begin{aligned} &[(m+1)/2, \dots, (m+1)/2]^T \\ &\qquad \in (1-\gamma)\% \text{ SCR}(E[R(X_1), \dots, R(X_m)]^T), \end{aligned}$$

where SCR is the symmetric credible region for $E[R(X_1), \dots, R(X_m)]^T$. When the inclusion does not hold, we declare with probability $1-\gamma$ that there is a difference between the algorithms. To compute SCR, we exploit the following results.

Theorem 1. *When* $\alpha^* = \delta_{\mathbf{x}}$ *the mean of* $[R(X_1), \dots, R(X_m)]^T$ *is:*

$$E[R(X_1), \dots, R(X_m)]^T = w_0 \mathbf{R}_0 + \mathbf{R}\mathbf{w}, \tag{15}$$

where $(w_0, \mathbf{w}^T) = (w_0, w_1, \dots, w_n) \sim Dir(s, 1, 1, \dots, 1)$, $\mathbf{R}$ *is the matrix of ranks and* $\mathbf{R}_0 = \int [R(X_1), \dots, R(X_m)]^T d\alpha^*(\mathbf{X})$. *The mean and covariance of* $w_0\mathbf{R}_0 + \mathbf{R}\mathbf{w}$ *are:*

$$\boldsymbol{\mu} = \mathcal{E}\left[w_0 \mathbf{R}_0 + \mathbf{R}\mathbf{w}\right] = \tfrac{s\mathbf{R}_0}{s+n} + \tfrac{\mathbf{R}\mathbf{1}}{s+n}, \tag{16}$$

$$\begin{aligned} \boldsymbol{\Sigma} &= Cov\left[w_0 \mathbf{R}_0 + \mathbf{R}\mathbf{w}\right] \\ &= E[w_0^2]\mathbf{R}_0\mathbf{R}_0^T + \mathbf{R}E[\mathbf{w}\mathbf{w}^T]\mathbf{R}^T \\ &+ \mathbf{R}_0 E[w_0 \mathbf{w}^T]\mathbf{R}^T + \mathbf{R}E[\mathbf{w}w_0]\mathbf{R}_0^T \\ &- E\left[w_0\mathbf{R}_0 + \mathbf{R}\mathbf{w}\right] E\left[w_0\mathbf{R}_0 + \mathbf{R}\mathbf{w}\right]^T, \end{aligned} \tag{17}$$

where $\mathbf{1}$ *is a* n*-dimensional vector of ones and the expectations on the r.h.s. are taken w.r.t.* $(w_0, \mathbf{w}^T) \sim Dir(s, 1, \dots, 1)$. ■

Its proof and that of the next theorems can be found in the appendix (supplementary material). For a large n, we have that

$$\begin{aligned} \boldsymbol{\mu} &= E\left[w_0\mathbf{R}_0 + \mathbf{R}\mathbf{w}\right] \approx \tfrac{1}{n}\mathbf{R}\mathbf{1}, \\ \boldsymbol{\Sigma} &= Cov\left[w_0\mathbf{R}_0 + \mathbf{R}\mathbf{w}\right] \\ &\approx \mathbf{R}E[\mathbf{w}\mathbf{w}^T]\mathbf{R}^T - \mathbf{R}E[\mathbf{w}]E[\mathbf{w}^T]\mathbf{R}^T, \end{aligned} \tag{18}$$

which tends respectively to the sample mean and sample covariance of the vectors of ranks. Hence, for a large n we can approximately assume that the null hypothesis mean ranks vector μ_0 is included in the $(1-\gamma)\%$ SCR if

$$\left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)^T \boldsymbol{\Sigma}^{-1} \left(\boldsymbol{\mu} - \boldsymbol{\mu}_0\right)\Big|_{m-1} \leq \rho, \tag{19}$$

where $\Big|_{m-1}$ means that we must take only $m-1$ components of the vectors and the covariance matrix,[1] $\boldsymbol{\mu}_0 = [(m+1)/2, \dots, (m+1)/2]^T$ and

$$\rho = \text{Finv}(1-\gamma, m-1, n-m+1)\frac{(n-1)(m-1)}{n-m+1},$$

where $Finv$ is the inverse of the F-distribution. Therefore, from a Bayesian perspective, the Friedman test is an inference for a multivariate mean based on an ellipsoid inclusion test. Note that, for small n we should compute the SCR by Monte Carlo sampling probability measures from the posterior DP (9) $P = w_0 P_0 + \sum_{j=1}^{n} w_j \delta_{\mathbf{X}_j}$. We leave the calculation of the exact SCR for future work.

The expression in (16) for the posterior expectation of $E[R(X_1), \dots, R(X_m)]^T$ w.r.t. the DP in (9) remains valid for a generic α^* since

$$\begin{aligned} \mathcal{E}[R_{m0}] &= \mathcal{E}\left[\int \sum_{k=1}^{m-1} I_{\{X_m > X_k\}}(\mathbf{X}) dP_0(\mathbf{X})\right] \\ &= \int \sum_{k=1}^{m-1} I_{\{X_m > X_k\}}(\mathbf{X}) d\mathcal{E}\left[P_0(\mathbf{X})\right] \\ &= \int \sum_{k=1}^{m-1} I_{\{X_m > X_k\}}(\mathbf{X}) d\alpha^*(\mathbf{X}) \end{aligned} \tag{20}$$

[1] Note in fact that $\boldsymbol{\mu}^T\mathbf{1} = m(m+1)/2$ (a constant), therefore there are only $m-1$ degrees of freedom.

It can be observed that (16) is the sum of two terms. The second term is proportional to the sampling mean rank of algorithm X_m, where the mean is taken over the n datasets. The first term is instead proportional to the prior mean rank of the algorithm X_m. Notice that for $s \to 0$ the posterior expectation of $E[R(X_1), \dots, R(X_m)]$ reduces to:

$$\mathcal{E}\left[E[R(X_1), \dots, R(X_m)]^T | \mathbf{X}^n\right] = \tfrac{1}{n}\left[R_1, \dots, R_m\right]^T. \tag{21}$$

Therefore, for $s \to 0$ we obtain the mean ranks used in the Friedman test.

5.3. A Bayesian multiple comparisons procedure

When the Bayesian Friedman test rejects the null hypothesis that all algorithms under comparison perform equally well, that is when the inequality in (19) is not satisfied, our interest is to identify which algorithms have significantly different performance. Following the directional multiple comparisons procedure introduced in section 3, we first need to consider, for each of the $k = m(m-1)/2$ pairs i, j of algorithms, the posterior probability of the alternative hypotheses $P(X_i > X_j) > 0.5$ and $P(X_j > X_i) > 0.5$ and select the statement with the largest posterior probability. Then, we need to perform a multiple comparisons procedure testing all the $k = m(m-1)/2$ selected statements, say $X_j > X_i$. For this, we follow the approach proposed in (Gelman & Tuerlinckx, 2000) that starts from the statement having the highest posterior probability and accepts as many statements as possible stopping when the joint posterior probability of all them being true is less than $1-\gamma$. The multiple comparison proceeds as follows.

- For each comparison perform a Bayesian sign test and derive the posterior probability $\mathcal{P}(P(X_j > X_i) > 0.5)$ that the hypothesis $P(X_j > X_i) > 0.5$ is true (that is, algorithm j is better i) and vice versa $\mathcal{P}(P(X_i > X_j) > 0.5)$. Select the direction with higher posterior probability for each pair i, j.

- Sort the posterior probabilities obtained in the previous step for the various pairwise comparisons in decreasing order. Let $\mathcal{P}_1, \dots, \mathcal{P}_k$ be the sorted posterior probabilities, $S_1, \dots, S_k$ the corresponding statements $X_j > X_i$ and $H_1, \dots, H_k$ the corresponding hypotheses $P(S_1) > 0.5, \dots, P(S_k) > 0.5$.

- Accept all the statements S_i with $i \leq \ell$, where ℓ is the greatest integer s.t.: $\mathcal{P}(H_1 \wedge H_2 \wedge \cdots \wedge H_\ell) > 1-\gamma$.

Note that, if none of the hypotheses has at least $1-\gamma$ posterior probability of being true, then we make no statement. The joint posterior probability of multiple hypotheses $\mathcal{P}(H_1 \wedge H_2 \wedge \cdots \wedge H_\ell)$ can be computed numerically by Monte Carlo sampling the probability measures $P = \sum_{j=0}^{n} w_j \delta_{\mathbf{X}_j}$ with $(w_0, w_1, \dots, w_n) \sim Dir(s, 1, \dots, 1)$ and evaluating the fraction of times all the ℓ conditions (one for each statement under consideration)

$$P(S_j) = \sum_{l=0}^{n} w_l I_{\{S_j\}} > 0.5, \quad j = 1, \dots, \ell$$

are verified simultaneously. This way we assure that the posterior probability $1 - \mathcal{P}(H_1 \wedge H_2 \wedge \cdots \wedge H_\ell)$ that there is an error in the list of accepted statements is lower than γ. Thus the Bayesian does not assume independence between the different hypotheses, like the frequentist; instead it considers their joint distribution. Assume, for example, that X_1 and X_2 are very highly correlated (take for simplicity $X_1 = X_2$); when using the Bayesian test, if we accept the statement $X_1 > X_3$ we will automatically accept also $X_2 > X_3$, whereas this is not true for the frequentist test with either the Bonferroni or the Holm's correction (or other sequential procedures). On the other side, if the p-value of two statements is 0.05, the frequentist approach without correction will accept both of them, whereas this is not true for our approach, since the joint probability of two hypotheses having marginal posterior probability of 0.95, can very easily be lower than 0.95.

6. Managing ties

To account for the presence of ties between the performances of two algorithms ($X_i = X_j$), the common approach is to replace the ranking $I_{X_j > X_i}$ (which assigns 1 if $X_j > X_i$ and 0 otherwise) with $I_{\{X_j > X_i\}} + 0.5 I_{\{X_j = X_i\}}$ (which assigns 1 if $X_j > X_i$, 0.5 if $X_j = X_i$ and 0 otherwise). Consider for instance the one-sided test in Section 5.1.1. To account for the presence of ties, we will to test the hypothesis $[P(X_i < X_j) + \frac{1}{2}P(X_i = X_j)] \leq 0.5$ against $[P(X_i < X_j) + \frac{1}{2}P(X_i = X_j)] > 0.5$. Since

$$\begin{aligned}&P(X_i < X_j) + \tfrac{1}{2}P(X_i = X_j) = \\ &E\left[I_{\{X_j > X_i\}} + \tfrac{1}{2} I_{\{X_j = X_i\}}\right] = E[H(X_j - X_i)],\end{aligned}$$

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 $z < 0$. The results presented in the previous sections are still valid if we substitute $I_{\{X_j > X_i\}}$ with $H(X_j - X_i)$ and $P_0(X_j > X_i)$ with $P_0(X_j > X_i) + 0.5 P_0(X_j = X_i)$.

7. Choosing the prior parameters

The DP is completely characterized by its prior parameters: the prior strength (or precision) s and the normalized base measure α^*. According to the Bayesian paradigm, we should select these parameters based on the available prior information. When no prior information is available, there are several alternatives to define a noninformative DP. The first solution to this problem has been proposed first by (Ferguson, 1973) and then by (Rubin, 1981) under the name of Bayesian Bootstrap (BB). It is the limiting DP obtained when the prior strength s goes to zero. In this case the choice of α^* is irrelevant, since the posterior inferences only depend on data for $s \to 0$. Ferguson has shown in several examples that the posterior expectations derived using the limiting DP coincide with the corresponding frequentist statistics (e.g., for the sign test statistic, for the Mann-Whitney statistic etc.). In (16), we have shown that this is also true for the Friedman statistic. However, 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 (Rubin, 1981). This latter issue can also give rise to numerical problems. For instance, in the Bayesian Friedman the covariance matrix obtained in (18) under the limiting DP is often ill-conditioned, and thus the matrix inversion in (19) can be numerically unstable. Instead, using a "non-null" prior introduces regularization terms in the covariance matrix, as can be seen from (18). For this reason, we have assumed $s > 0$ and $\alpha^* = \delta_{X_1=X_2=\cdots=X_m}$, so that a priori $\mathcal{E}[E[H(X_j - X_i)]] = 1/2$ for each i, j and $\mathcal{E}[E[R(X_i)]] = (m+1)/2$. This allows us to overcome the effects of ill-conditioning, although this prior is not "noninformative": a-priori we are assuming that all the algorithms have the same performance. However, it has the nice feature that the decisions of the Bayesian sign test implemented with this prior does not depend on the choice of s as it is shown in the next theorem.

Theorem 2. *When $\alpha^* = \delta_{X_1=X_2}$, we have that*

$$\begin{aligned}&\mathcal{P}\left[P(X_2 > X_1) + \tfrac{1}{2}P(X_2 = X_1) > \tfrac{1}{2}\right]\\&= \int_{0.5}^{1} Beta(z; n_g, n - n_t - n_g)dz\\&= 1 - B_{1/2}(n_g, n - n_t - n_g),\end{aligned} \tag{22}$$

where $n_g = \sum_{i=1}^{n} I_{\{X_2 > X_1\}}$, $n_t = \sum_{i=1}^{n} I_{\{X_2 = X_1\}}$ and $B_{1/2}(\cdot,\cdot)$ is the regularized incomplete beta function computed at $1/2$. ■

An alternative solution to the problem of choosing the parameters of the DP in case of lack of prior information, which is "noninformative" and solves ill-conditioning, is represented by the prior near-ignorance DP (IDP) (Benavoli et al., 2014). It consists of a class of DP priors obtained by fixing $s > 0$ (e.g., $s = 1$) and by letting the normalized base measure α^* vary in the set of all probability measures. Since $s > 0$, IDP introduces regularization terms in the covariance matrix. Moreover, it is a model of prior ignorance, since a priori it assumes that all the relative ranks of the algorithms are possible. Posterior inferences are therefore derived considering all the possible prior ranks, which results in lower and upper bounds for the inferences (calculated considering the least favorable and the most favorable prior rank). The application of IDP to multivariate inference problems, as in (19), can be computationally quite involving.

S_i	p-value Sign test	$1-\mathcal{P}(H_i)$ Bayesian test	$1-\mathcal{P}(H_1 \wedge \cdots \wedge H_i)$ Bayesian test
$X_1 < X_3$	0.0000	0.0000	0.0000
$X_1 < X_4$	0.0000	0.0000	0.0000
$X_1 < X_2$	0.0000	0.0000	0.0000
$X_2 < X_3$	0.0494	0.0307	0.0307
$X_2 < X_4$	0.0494	0.0307	0.0595
$X_3 < X_4$	0.5722	0.5000	0.5263

Table 1. P-values and posterior probabilities of the sign and Bayesian tests.

Example 1. Consider $m = 4$ algorithms X_1, X_2, X_3 and X_4 tested on $n = 30$ datasets and assume that the mean ranks of the algorithms are $R_1 = 1.3$, $R_2 = 2.5$, $R_3 = 2.90$ and $R_4 = 3.30$ (the observations considered in this example can be found in the supplementary material). This gives a p-value of 10^{-9} for the Friedman test and, thus, we can reject the null hypothesis. We can then start the multiple comparisons procedure to find which algorithms are better (if any). Each pair of algorithms i, j is compared in the direction $X_j > X_i$ that gives a number of times the condition $X_j > X_i$ is verified in the n observations larger than $n/2 = 15$. This way we guarantee that the p-value for the comparison in the selected direction is more significant than in the opposite direction. We apply the Bayesian multiple comparison procedure to the four simulated algorithms and compare it to the traditional procedure using the sign test. Table 1 compares the p-values obtained for the sign-test with the posterior probabilities $1 - \mathcal{P}(H_i)$ of the hypothesis $P(S_i) > 0.5$ being false given by the Bayesian procedure. Note that the p-value of the sign-test is always lower that the posterior probability obtained with the prior $\alpha^* = \delta_{X_1=X_2=X_3=X_4}$ showing that the sign-test somehow favors the null hypothesis. The third column of Table 1 shows the posterior probabilities $1 - \mathcal{P}(H1 \wedge \cdots \wedge H_i)$ that at least one of the hypotheses $P(S_1) > 0.5, \ldots, P(S_i) > 0.5$ is false. All statements for which this probability is smaller than $\gamma = 0.05$ are retained as significant. Thus, while only three p-values of the sign-test fall below the conservative threshold $\gamma/(k - i + 1)$ of the Holm's procedure, and five p-value falls below the unadjusted threshold γ, the Bayesian test shows that up to four statements the probability of error remains below the threshold $\gamma = 0.05$.

7.1. Racing experiments

We experimentally compare our procedure (Bayesian Friedman test with joint multiple comparisons) with the well-established F-race. The setting are as follows. We perform both the frequentist and the Friedman with significance α=0.05. For the Bayesian multiple comparison, we accept statements of joint comparison whose posterior probability is larger than 0.95. For the frequentist multiple comparison we consider two options: using the sign test (S) or the mean-ranks (MR) test. This yield the following algorithms: Bayesian race, F-Race$_S$ and F-Race$_{MR}$. Within the F-race we perform the multiple comparison keeping the significance level at 0.05, thus without controlling the FWER error. This is the approach described by (Birattari et al., 2002). By controlling the FWER the procedure would loose too power making less effective the racing. This remain true even if more modern procedures than the Bonferroni correction are adopted. This also indirectly shows that controlling the FWER is not always the best option. Within the Bayesian race, we define two algorithms as indistinguishable if $\frac{1}{2} - \epsilon < P(X_2 > X_1) < \frac{1}{2} + \epsilon$, where $\epsilon = 0.05$.

We consider q candidates in each race. We sample the results of the j-th candidate from a normal with mean μ_i and variance σ_i^2. Before each race, the means $\mu_1, \dots, \mu_q$ are uniformly sampled from the interval $[0, 1]$; the variances $\sigma_1^2 = \dots = \sigma_q^2 = \rho^2$. The best algorithm is thus the one with the highest mean. We fix the overall number of maximum allowed assessments to $M = 300$. For each assessment of an algorithm we decrease M of one unit. In this experimental setting, everytime we assess the i-th algorithm we increase the number of random generated observations (from the normal with mean μ_i and variance σ_i^2) of five new observations. If multiple candidates are still present when the number of maximum assessments is achieved, we choose the candidate which has so far the best average performance. We perform 200 repetitions for each setting. In each experiment we track the following indicators: the absolute distance between the rank of the candidate eventually selected and the rank of the best algorithm (mean absolute error, denoted by *MAE*) and the fraction of the number of required iterations w.r.t. the maximum allowed M (denoted by *ITER*).

Setting	Bayesian		F-Race$_S$		F-Race$_{MR}$	
q, ρ	# ITER	MAE	# ITER	MAE	# ITER	MAE
$30, 1$	0.63	0.70	0.67	0.80	0.58	0.77
$50, 1$	0.72	0.92	0.80	1.22	0.69	1.15
$100, 1$	0.75	1.84	0.84	2.31	0.70	2.05
$100, 0.5$	0.67	1.15	0.74	1.19	0.62	1.26
$100, 0.1$	0.36	0.28	0.36	0.30	0.35	0.30
$200, 0.1$	0.50	0.46	0.51	0.63	0.50	0.58

Table 2. Experimental results of racing.

The simulation results are shown in Table 2 for different values of q and ρ. From Table 2, it is evident that our method is the best in terms of MAE. F-Race$_{MR}$ is in some case faster than our method, but it has always a higher MAE. Instead, F-Race$_S$ is always outperformed by our method both in terms of MAE and ITER. F-Race$_S$ and the Bayesian tests perform the pairwise multiple comparisons with a sign test and, respectively, a Bayesian version of the sign test, so they are quite similar. The lower ITER of the Bayesian method is also due to the fact that it can declare that two algorithms are indistinguishable. This allows to remove many algorithms and so to speed up the decision process. Table 3 reports the average number of algorithms removed because indistinguishable in the above listed case-studies. Note that, the pairs declared as indistinguishable from the Bayesian test were always be truly indistinguishable according to the adopted criterion described above. Therefore, the Bayesian test is very accurate on detecting when two algorithms are indistinguishable.

Setting	Bayesian
q, ρ	# indistinguishable
$30, 1$	0.9
$50, 1$	1.7
$100, 1$	4.5
$100, 0.5$	3.1
$100, 0.1$	2.4
$200, 0.1$	2.7

Table 3. Average number of algorithms declared to be indistinguishable.

Although the Bayesian method removes more algorithms, it has lower MAE than F-Race$_S$, because with the joint test is able to reduce the number of Type-I errors, but without penalizing the power too much.

8. Conclusions

We have proposed a novel Bayesian method based on the Dirichlet Processes (DP) for performing the Friedman test, together with a *joint* procedure for the analysis of the multiple comparisons which accounts for their dependencies and which is based on the posterior probability computed through the DP. We have then employed this new test for performing algorithms racing. Experimental results have shown that our approach is competitive both in terms of accuracy and speed in detecting the best algorithm. We plan to extend this work in two directions. The first is to implement Bayesian versions of other multiple nonparametric tests such as for instance the Kruskal-Wallis test. Second, we plan to derive new tests for algorithms racing which are able to compare the algorithms using more than one metric at the same time.

Acknowledgments

This work was partly supported by the Swiss NSF grant nos. 200021_146606 / 1.

References

Benavoli, Alessio, Mangili, Francesca, Corani, Giorgio, Zaffalon, Marco, and Ruggeri, Fabrizio. A bayesian wilcoxon signed-rank test based on the dirichlet process. In *Proc. of the 31st International Conference on Machine Learning (ICML-14)*, pp. 1026–1034, 2014.

Birattari, Mauro. *Tuning Metaheuristics: A Machine Learning Perspective*, volume 197. Springer Science & Business Media, 2009.

Birattari, Mauro, Stutzle, Thomas, Paquete, Luis L, Varrentrapp, K, and Langdon, WB. A racing algorithm for configuring metaheuristics. In *GECCO 2002 Proceedings of the Genetic and Evolutionary Computation Conference*, pp. 11–18. Morgan Kaufmann, 2002.

Chien, Steve, Gratch, Jonathan, and Burl, Michael. On the efficient allocation of resources for hypothesis evaluation: A statistical approach. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17(7):652–665, 1995.

Conover, William Jay and Conover, WJ. Practical nonparametric statistics. 1980.

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. *Journal of Machine Learning Research*, 7:1–30, 2006.

Ferguson, Thomas S. A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, 1(2):pp. 209–230, 1973.

Fligner, Michael A. A note on two-sided distribution-free treatment versus control multiple comparisons. *Journal of the American Statistical Association*, 79(385):pp. 208–211, 1984.

Gabriel, K Ruben. Simultaneous test procedures–some theory of multiple comparisons. *The Annals of Mathematical Statistics*, pp. 224–250, 1969.

Garcia, Salvador and Herrera, Francisco. An extension on Statistical Comparisons of Classifiers over Multiple Data Sets for all pairwise comparisons. *Journal of Machine Learning Research*, 9(12), 2008.

Gelman, Andrew and Tuerlinckx, Francis. Type S error rates for classical and Bayesian single and multiple comparison procedures. *Computational Statistics*, 3, 2000.

Gelman, Andrew, Carlin, John B, Stern, Hal S, Dunson, David B, Vehtari, Aki, and Rubin, Donald B. *Bayesian data analysis*. CRC press, 2013.

Ghosh, Jayanta K and Ramamoorthi, RV. *Bayesian nonparametrics*. Springer (NY), 2003.

Hollander, Myles, Wolfe, Douglas A, and Chicken, Eric. *Nonparametric statistical methods*, volume 751. John Wiley & Sons, 2013.

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

Maron, Oden and Moore, Andrew W. The racing algorithm: model selection for lazy learners. *Artificial Intelligence Review*, 11(1):193–225, 1997.

Miller, Rupert G. *Simultaneous statistical inference*. Springer, 1966.

Müller, Peter, Parmigiani, Giovanni, Robert, Christian, and Rousseau, Judith. Optimal sample size for multiple testing: the case of gene expression microarrays. *Journal of the American Statistical Association*, 99(468):990–1001, 2004.

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

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

Williams, Valerie SL, Jones, Lyle V, and Tukey, John W. Controlling error in multiple comparisons, with examples from state-to-state differences in educational achievement. *Journal of Educational and Behavioral Statistics*, 24(1):42–69, 1999.