Bayesian Performance Analysis as an Alternative to Null Hypothesis Statistical Tests in the paper "Kernels of Mallows Models under the Hamming Distance for solving the Quadratic Assignment Problem"

Etor Arza, Aritz Pérez, Ekhiñe Irurozki, Josu Ceberio

February 26, 2020

Abstract

In this short additional document, we justify the use of Bayesian Performance Analysis instead of Null Hypothesis Statistical Tests the paper "Kernels of Mallows Models under the Hamming Distance for solving the Quadratic Assignment Problem". This document is based on the work of by Borja Calvo et. al [2, 3].

1 Introduction

Until recently, the frequentist approach has dominated the multiple comparison of algorithms, where null hypothesis statistical tests (NHST) have been the most used statistical tool to measure the uncertainty in empirical results [3]. However, many statisticians have criticized the use of NHST [7, 9], specially because of the lack of interpretability of the p-values [4, 8]. The concern about the use of NHST has increased in the last years. For example, the American Statistical Association (ASA) has recently warned about some common misinterpretations of the p-value [11]. These NHST assume H_0 to be true, thus, that the distributions from which the samples to be compared are drawn are identical, which in general, is not true (there is always a small difference in the performance of algorithms, even if it is very small) [3]. In this setting, the correct interpretation of the p-value is the probability of erroneously assuming there are differences between the distributions, when actually, there are not. Sometimes, this p-value is used to measure of the magnitude of the difference, since the p-value is sensitive to the increase of the difference in the performance of the algorithms. However, the p-value also changes with the sample size, even if the average difference between the algorithms remains unchanged. Therefore, since there will always be even a small difference between algorithms, getting a significant difference is only a matter of sample size [4]. In other words, the p-value is not able to represent the magnitude of the differences between the performance of two algorithms.

With NHST, the algorithm comparison testing is done pairwise. In addition, since usually more than 2 algorithms are being compared, the NHST has to be repeated multiple times, ($\binom{n}{2}$ to be specific, where n is the number of algorithms to be compared) and the alpha or critical p-value needs to be corrected accordingly, for example using Bonferroni correction [5]. Unlike NHST, BPA simultaneously compares several algorithms. This means there is no need to correct the alpha values nor to conduct several tests in order to compare several algorithms. Not limited to the previous, NHST have two other additional disadvantages [1]. First, NHST do not provide any information when the null hypothesis, H_0, cannot be rejected ($p > \alpha$). The correct interpretation in this case would be that there is not enough evidence to reject H_0, but it would not be correct to assume that H_0 is true. Second, NHST do not separate between the effect size and the sample size. In other words, the p-value is sensitive to both the magnitude of the difference and the sample size, which means that $p < \alpha$ almost always can be achieved with enough samples, since it is very unlikely that two different algorithms exactly the same performance, even if these differences are very small or negligible.

BPA defines a posterior distribution of the parameters that measure the differences between the compared algorithms, given the experimental data. The BPA considered in this paper uses the PL model as the sampling distribution and the Dirichlet distribution is used to model the uncertainty of the weights of the PL model. Since we are using the Plackett-Luce (PL) probability distribution [10] to model the differences, the data is required to be in the permutation space. Even if the experimental data is real valued, permutations can be obtained by ranking the average performance in each problem instance.

It is true that the obtained posterior distribution depends on the prior distribution. However, it has been shown that the prior distribution does not have a big impact on the results [3]. In order to further prove this point, we have replicated the experimentation [3] in our experimental data. The experiment consists in comparing the results of three BPA, but using favorable, uniform and deceptive priors. The favorable and deceptive priors are set considering the rankings and inverse rankings of the experimental data, respectively. As shown in Figures 1,2 and 3, even tough the priors do have an effect, they do not condition the conclusions that one may draw from the BPA.

In order to compare the results of the BPA with those of NHST, we have used Wilcoxon's signed rank test with Finner's p-value correction [6]. It is worth mentioning that this is a Post hoc procedure, and as such, Friedman's or similar test is required before it is applied. The conclusion that can be drawn from these NHST are quite similar to the ones drawn from the BPA. As shown in the critical difference Figures 4,5 and 6, there is a statistically significant difference between Hamming KMM and the rest of the methods, at a significance level of $\alpha = 0.05$. It is worth mentioning that the BPA as an advantage in terms of interpretability of the uncertainty. Consider the following example based on the experimental data of our paper. With the NHST, we conclude that Hamming KMM performs better than the rest of the methods at $\alpha = 0.05$ in both the ablation study and the comparison with other EDA methods specific to $\mathcal{S}^n$ (shown in Figures 4 and 5 respectively). On the other hand, from the BPA shown in Figures 7 and 8, one can see that the magnitude of the difference between KMM.linear and KMM.expo (Figure 7) is smaller than the difference between Caylay.KMM and Hamming.KMM (Figure 8).

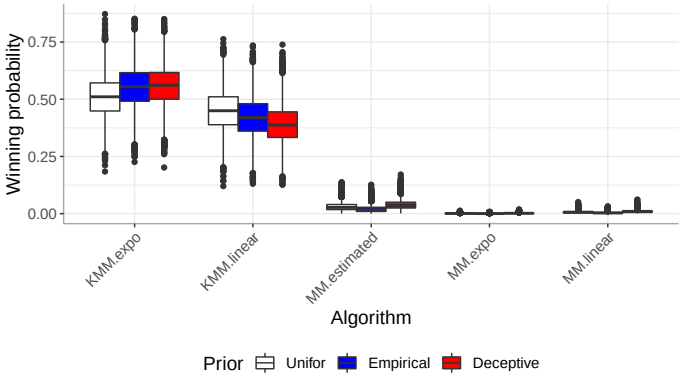

Figure 1: BPA with different priors of the ablation study.

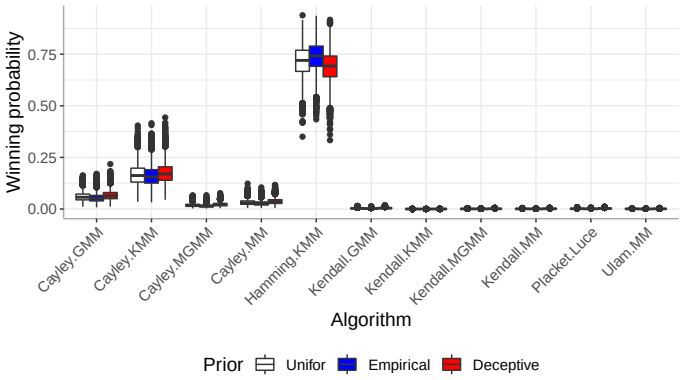

Figure 2: BPA with different priors of the comparison with other EDAs specific to $\mathcal{S}^n$.

The above example shows how the uncertainty on the BPA is represented in two ways. The first kind of uncertainty, the uncertainty associated with the sample size used in the BPA, is represented by the credibility intervals. And the other kind of uncertainty, associated with the similarity of performance of the algorithms, is associated with a similar probability of being the best algorithm or the overlapping between two credibility intervals.
In conclusion, we believe that BPA is a promising alternative to NHST.

References

[1] A. Benavoli, G. Corani, J. Demšar, and M. Zaffalon. Time for a change: A tutorial for comparing multiple classifiers through Bayesian analysis. *The Journal of Machine Learning Research*, 18(1):2653–2688, 2017.

[2] B. Calvo and G. Santafé Rodrigo. scmamp: Statistical comparison of multiple algorithms in multiple problems. *The R Journal, Vol. 8/1, Aug. 2016*, 2016.

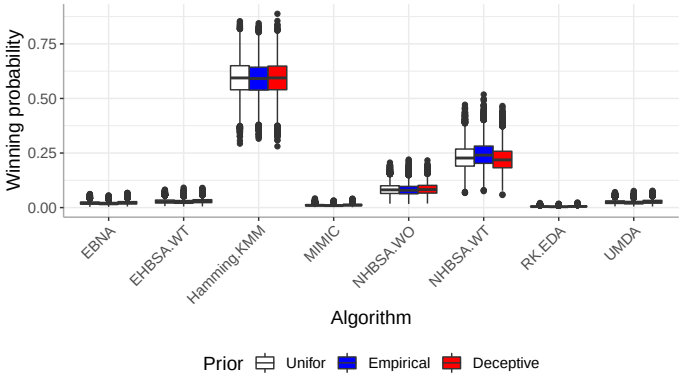

Figure 3: BPA with different priors of the comparison with other non-specific EDAs.

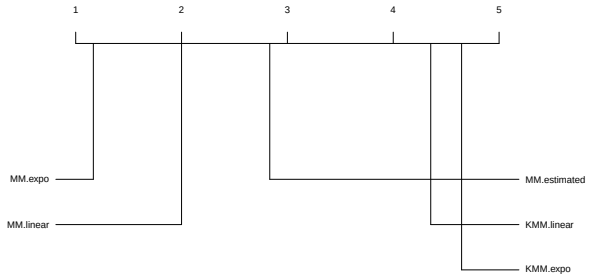

Figure 4: NHST of the ablation study.

[3] B. Calvo, O. M. Shir, J. Ceberio, C. Doerr, H. Wang, T. Bäck, and J. A. Lozano. Bayesian performance analysis for black-box optimization benchmarking. In *Proceedings of the Genetic and Evolutionary Computation Conference Companion*, GECCO '19, pages 1789–1797, New York, NY, USA, 2019. ACM.

[4] J. Cohen. The earth is round (p<. 05). In *What if there were no significance tests?*, pages 69–82. Routledge, 2016.

[5] O. J. Dunn. Estimation of the Medians for Dependent Variables. *The Annals of Mathematical Statistics*, 30(1):192–197, Mar. 1959.

[6] S. García, A. Fernández, J. Luengo, and F. Herrera. Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, 180(10):2044–2064, 2010.

[7] G. Gigerenzer and J. N. Marewski. Surrogate science: The idol of a universal method for scientific inference. *Journal of Management*, 41(2):421–440, 2015.

[8] S. Greenland, S. J. Senn, K. J. Rothman, J. B. Carlin, C. Poole, S. N.

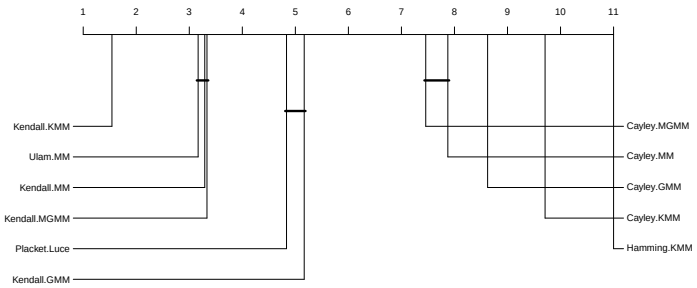

Figure 5: NHST of the comparison with other EDAs specific to $\mathcal{S}^n$.

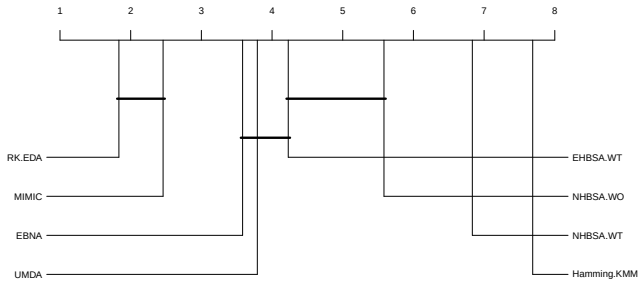

Figure 6: NHST of the comparison with other non-specific EDAs.

Goodman, and D. G. Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European journal of epidemiology*, 31(4):337–350, 2016.

[9] D. E. Morrison and R. E. Henkel. *The significance test controversy: A reader*. Transaction Publishers, 2006.

[10] R. L. Plackett. The analysis of permutations. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 24(2):193–202, 1975.

[11] R. L. Wasserstein and N. A. Lazar. The asa statement on p-values: context, process, and purpose, 2016.

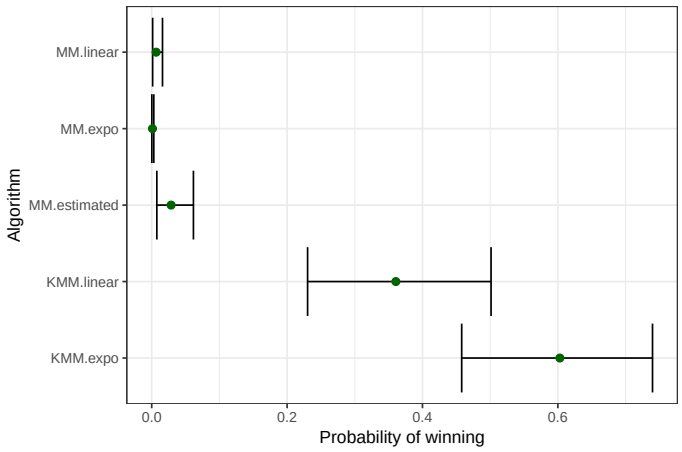

Figure 7: The BPA of the ablation study considered in our paper.

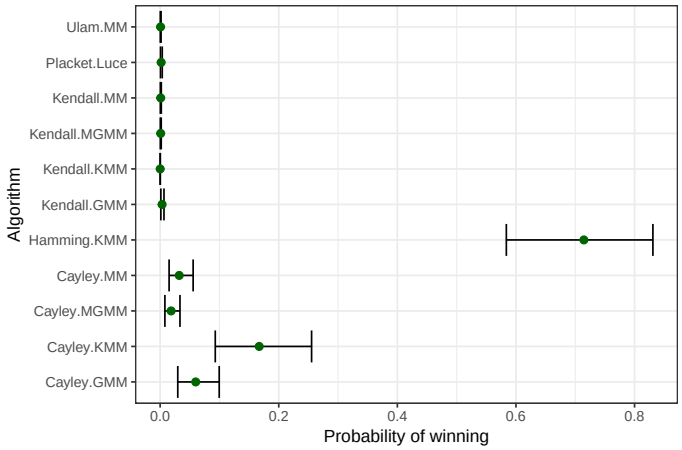

Figure 8: The BPA of the EDAs specific to $\mathcal{S}^n$ considered in our paper.