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

March 30, 2020

Abstract

In this short additional document, we justify the use of Bayesian Performance Analysis instead of Null Hypothesis Statistical Tests in the paper "Kernels of Mallows Models under the Hamming Distance for solving the Quadratic Assignment Problem".

1 Introduction

Bayesian Performance Analysis (BPA) is also a suitable method for the statistical assessment of experimental comparison of multiple optimization algorithms. In the following paragraphs, we will address the different concerns that a reader new to BPA may have, and will try to show that i) The overall conclusions are similar with either BPA or Mann-Whitney, ii) BPA provides additional information on the algorithm performance comparison, iii) the probability distribution assumed (the Plackett-Luce probability distribution) is suitable for multiple algorithm performance comparison and vi) by using an uninformative prior, the results are not biased.

i) The overall conclusions are similar with either BPA or Mann-Whitney

In order to compare the results of the BPA with those of null hypothesis statistical tests (NHST), we have used Wilcoxon's signed-rank test (also known as the Mann-Whitney test when used in non-paired data) with Finner's p-value correction [5]. It is worth mentioning that this is a post hoc procedure, and as such, Friedman's (or equivalent) test is required before multiple pairwise tests are performed. Figures 1, 2 and 3 show the Wilcoxon's signed-rank test and the BPA side by side, for the ablation study, the comparison with EDAs specific to the space of permutations of size n, and comparison with other EDAs adapted to work in the space of permutations of size n, respectively. The results of the NHST and BPA are quite similar. As shown in the critical difference

Figures 1a,2a and 3a, there is a statistically significant difference between Hamming KMM and the rest of the methods, at a significance level of $\alpha = 0.05$. As shown in Figures 1b,2b and 3b, BPA also shows a significant difference between Hamming KMM EDA and the rest of the methods. The conclusions obtained by NHST are also supported by the evidence shown by BPA.

ii) BPA provides additional information on the algorithm performance comparison.

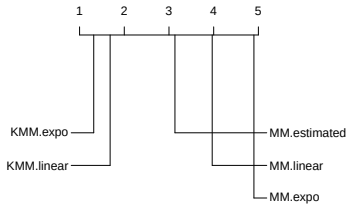
As mentioned in the previous point, the overall conclusions are the same with the two methods. However, BPA provides additional information beyond the statistical significance of the results. Specifically, the weights of the posterior distribution of the Plackett-Luce probability model can be interpreted as the probability of an algorithm being the highest ranked one. In this sense, the BPA can be interpreted by considering two types of uncertainties: The first kind of uncertainty, the uncertainty related to the sample size used in the BPA, is represented by the credible intervals and specifically, the overlapping of two credible intervals. In other words, the lack of data will increase the length of the credible intervals. The other kind of uncertainty, related to the similarity of performance of the algorithms, is associated with a similar probability of being the best algorithm.

iii) The Plackett-Luce probability distribution is suitable for algorithm performance comparison.

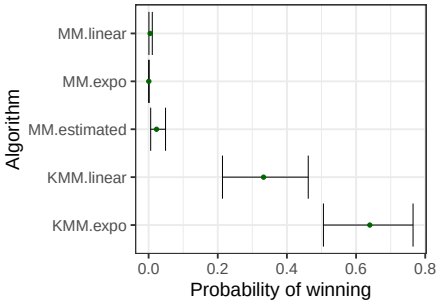
Both NHST and BPA make assumptions on the data. For NHST, and specifically for the Mann-Whitney test, it is assumed that, under H_0, the average performances of the two algorithms being compared are identical [1, 9]. On the other hand, Bayesian analysis assumes the data has a known distribution. In the case of BPA, it is assumed that the algorithm performance ranking data follows a Plackett-Luce (PL) [7, 8] distribution. The Plackett-Luce distribution is known to be suitable to model preferences [4], and specifically, algorithm performance analysis [3]. In a nutshell, we believe that assuming a PL distribution in the algorithm ranking data is reasonable in the context of algorithm performance comparison.

iv) By using an uninformative prior, the results are not biased.

Regarding the concern of the reviewer about the dependence of the results on the prior distribution, an uninformative prior has been used in the BPA. Specifically, a Dirichlet [6] prior with $\alpha = 1$ has been used to estimate the parameters of the PL from the experimental results. This is an uninformative prior because it gives the same probability to all the distributions that can be represented by a PL. This is the most common used approach in Bayesian statistics when there is no prior knowledge about the phenomenon that is being modeled [2].

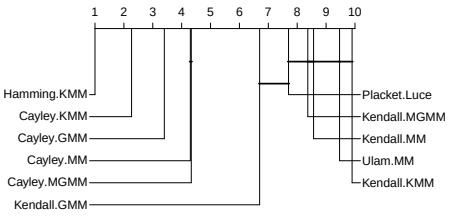


(a) NHST

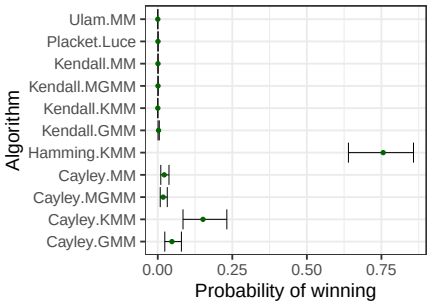


(b) BPA

Figure 1: Ablation study, where Hamming KMM EDA with an exponential decrease of the expectation is compared with the other simpler versions of Hamming EDAs.

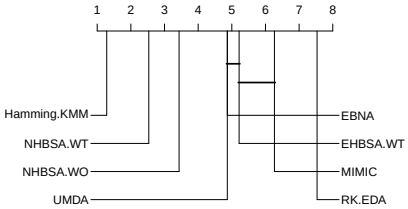


(a) NHST

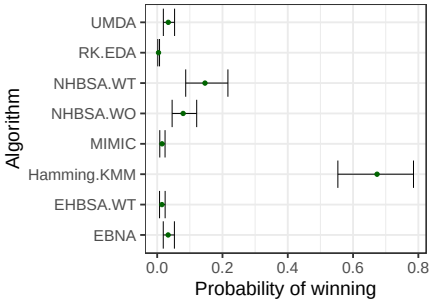


(b) BPA

Figure 2: Comparison with other EDAs specific to the space of permutations of size n, including, but not limited to, other Mallows model based EDAs.



(a) NHST



(b) BPA

Figure 3: Comparison with other non-specific EDAs, or EDAs that have been adapted to work in the space of permutations.

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] J. M. Bernardo and A. F. Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[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] D. E. Critchlow, M. A. Fligner, and J. S. Verducci. Probability Models on Rankings. *Journal of Mathematical Psychology*, 35:294–318, 1991.

[5] 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.

[6] S. Kotz, N. L. Johnson, N. Balakrishnan, and N. L. Johnson. *Continuous Multivariate Distributions*. Wiley Series in Probability and Statistics. Wiley, New York, 2nd ed edition, 2000.

[7] R. D. Luce. Individual choice behavior, a theoretical analysis. *Bull. Amer. Math. Soc*, 66(1960):259–260, 1960.

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

[9] J. W. Pratt. Robustness of some procedures for the two-sample location problem. *Journal of the American Statistical Association*, 59(307):665–680, 1964.