Supplementary Material
Bayesian Performance Analysis for Algorithm Ranking Comparison

Jairo Rojas-Delgado[1], Josu Ceberio[2], Borja Calvo[2], and Jose A. Lozano[1,2]

[1]Basque Center for Applied Mathematics. Bilbao, Spain.
Email: {jrojasdelgado, jlozano}@bcamath.org
[2]Intelligent Systems Group. University of the Basque Country UPV/EHU. Donostia, Spain.
Email: {josu.ceberio, borja.calvo, ja.lozano}@ehu.eus

1 Numerical results

In this section we provide the numerical results for the experiments conducted in the paper. For each posterior summary, we report the mean and standard deviation between parenthesis taken from the different posterior samples.

1.1 Synthetically generated scores

Table 1: Probability of each algorithm to be the top-ranked algorithm.

	A_1	A_2	A_3	A_4
PLD	9.10E-01 (9.10E-03)	8.30E-02 (8.48E-03)	6.79E-03 (9.14E-04)	5.47E-04 (9.43E-05)
PLG	9.11E-01 (8.67E-03)	8.17E-02 (8.03E-03)	6.66E-03 (9.31E-04)	5.26E-04 (9.11E-05)
BT	9.09E-01 (8.84E-03)	8.99E-02 (8.74E-03)	9.73E-04 (2.38E-04)	7.26E-07 (3.58E-07)
MM	1.00E+00 (8.75E-06)	3.86E-05 (8.75E-06)	1.56E-09 (7.30E-10)	6.66E-14 (4.90E-14)

Table 2: Probability of each algorithm to outperform others.

		A_1	A_2	A_3	A_4
PLD	A_1		9.16E-01 (8.59E-03)	9.93E-01 (1.04E-03)	9.99E-01 (1.07E-04)
	A_2	8.36E-02 (8.59E-03)		9.24E-01 (7.72E-03)	9.93E-01 (1.00E-03)
	A_3	7.41E-03 (1.04E-03)	7.57E-02 (7.72E-03)		9.26E-01 (7.48E-03)
	A_4	6.01E-04 (1.07E-04)	6.56E-03 (1.00E-03)	7.45E-02 (7.48E-03)	
PLG	A_1		9.18E-01 (8.14E-03)	9.93E-01 (1.06E-03)	9.99E-01 (1.03E-04)
	A_2	8.22E-02 (8.14E-03)		9.25E-01 (7.83E-03)	9.94E-01 (9.73E-04)
	A_3	7.26E-03 (1.06E-03)	7.54E-02 (7.83E-03)		9.27E-01 (8.25E-03)
	A_4	5.77E-04 (1.03E-04)	6.41E-03 (9.73E-04)	7.33E-02 (8.25E-03)	
BT	A_1		9.10E-01 (8.76E-03)	9.98E-01 (4.27E-04)	1.00E+00 (3.08E-06)
	A_2	9.00E-02 (8.76E-03)		9.20E-01 (8.29E-03)	9.99E-01 (3.53E-04)
	A_3	1.88E-03 (4.27E-04)	7.99E-02 (8.29E-03)		9.26E-01 (8.38E-03)
	A_4	8.75E-06 (3.08E-06)	1.37E-03 (3.53E-04)	7.39E-02 (8.38E-03)	
MM	A_1		1.00E+00 (8.75E-06)	1.00E+00 (1.46E-09)	1.00E+00 (1.47E-13)
	A_2	3.86E-05 (8.75E-06)		1.00E+00 (8.75E-06)	1.00E+00 (1.46E-09)
	A_3	3.13E-09 (1.46E-09)	3.86E-05 (8.75E-06)		1.00E+00 (8.75E-06)
	A_4	2.00E-13 (1.47E-13)	3.13E-09 (1.46E-09)	3.86E-05 (8.75E-06)	

Table 3: Probability of an algorithm to be in the top-k ranking.

		Top 1	Top 2	Top 3	Top 4
PLD	A_1		9.99E-01 (2.79E-04)	1.00E+00 (7.74E-07)	
	A_2		9.19E-01 (8.21E-03)	9.99E-01 (2.24E-04)	
	A_3		7.58E-02 (7.67E-03)	9.26E-01 (7.43E-03)	
	A_4		6.11E-03 (8.99E-04)	7.50E-02 (7.53E-03)	
PLG	A_1		9.99E-01 (2.73E-04)	1.00E+00 (7.20E-07)	
	A_2		9.20E-01 (8.23E-03)	9.99E-01 (2.13E-04)	
	A_3		7.56E-02 (7.80E-03)	9.27E-01 (8.21E-03)	
	A_4		5.97E-03 (8.76E-04)	7.37E-02 (8.30E-03)	
BT	A_1		9.99E-01 (2.40E-04)	1.00E+00 (4.59E-07)	
	A_2		9.20E-01 (8.33E-03)	9.99E-01 (2.04E-04)	
	A_3		8.08E-02 (8.36E-03)	9.26E-01 (8.36E-03)	
	A_4		6.68E-04 (1.87E-04)	7.46E-02 (8.46E-03)	
MM	A_1		1.00E+00 (7.30E-10)	1.00E+00 (4.91E-14)	
	A_2		1.00E+00 (8.75E-06)	1.00E+00 (7.30E-10)	
	A_3		3.86E-05 (8.75E-06)	1.00E+00 (8.75E-06)	
	A_4		1.56E-09 (7.30E-10)	3.86E-05 (8.75E-06)	

1.2 Permutation Flowshop Scheduling Problem

Table 4: Probability of each algorithm to be the top-ranked.

	GM-EDA	HGM-EDA	AGA	VNS	NVNS
PLD	1.67E-02 (1.29E-03)	3.45E-01 (1.02E-02)	4.39E-01 (1.28E-02)	1.29E-01 (5.49E-03)	7.05E-02 (3.85E-03)
PLG	1.67E-02 (1.26E-03)	3.45E-01 (1.07E-02)	4.40E-01 (1.20E-02)	1.28E-01 (5.82E-03)	7.00E-02 (3.65E-03)
BT	1.03E-03 (2.04E-04)	3.77E-01 (1.27E-02)	4.62E-01 (1.33E-02)	1.08E-01 (6.53E-03)	5.19E-02 (4.16E-03)
MM	5.49E-05 (1.76E-05)	9.13E-01 (7.06E-03)	7.96E-02 (5.82E-03)	6.99E-03 (1.08E-03)	6.18E-04 (1.46E-04)

Table 5: Probability of each algorithm to outperform others.

		GM-EDA	HGM-EDA	AGA	VNS	NVNS
PLD	GM-EDA		4.62E-02 (3.60E-03)	3.67E-02 (3.26E-03)	1.15E-01 (7.73E-03)	1.92E-01 (1.18E-02)
	HGM-EDA	9.54E-01 (3.60E-03)		4.40E-01 (1.36E-02)	7.28E-01 (1.05E-02)	8.30E-01 (8.80E-03)
	AGA	9.63E-01 (3.26E-03)	5.60E-01 (1.36E-02)		7.73E-01 (1.10E-02)	8.62E-01 (8.65E-03)
	VNS	8.85E-01 (7.73E-03)	2.72E-01 (1.05E-02)	2.27E-01 (1.10E-02)		6.46E-01 (1.34E-02)
	NVNS	8.08E-01 (1.18E-02)	1.70E-01 (8.80E-03)	1.38E-01 (8.65E-03)	3.54E-01 (1.34E-02)	
PLG	GM-EDA		4.60E-02 (3.67E-03)	3.65E-02 (3.02E-03)	1.15E-01 (8.37E-03)	1.92E-01 (1.19E-02)
	HGM-EDA	9.54E-01 (3.67E-03)		4.40E-01 (1.36E-02)	7.29E-01 (1.19E-02)	8.31E-01 (8.70E-03)
	AGA	9.63E-01 (3.02E-03)	5.60E-01 (1.36E-02)		7.74E-01 (1.07E-02)	8.63E-01 (7.97E-03)
	VNS	8.85E-01 (8.37E-03)	2.71E-01 (1.19E-02)	2.26E-01 (1.07E-02)		6.47E-01 (1.46E-02)
	NVNS	8.08E-01 (1.19E-02)	1.69E-01 (8.70E-03)	1.37E-01 (7.97E-03)	3.53E-01 (1.46E-02)	
BT	GM-EDA		2.38E-02 (2.65E-03)	1.62E-02 (1.93E-03)	1.07E-01 (7.75E-03)	1.74E-01 (9.88E-03)
	HGM-EDA	9.76E-01 (2.65E-03)		4.50E-01 (1.44E-02)	7.39E-01 (1.21E-02)	8.26E-01 (1.00E-02)
	AGA	9.84E-01 (1.93E-03)	5.50E-01 (1.44E-02)		7.80E-01 (1.11E-02)	8.59E-01 (9.15E-03)
	VNS	8.93E-01 (7.75E-03)	2.61E-01 (1.21E-02)	2.20E-01 (1.11E-02)		6.11E-01 (1.39E-02)
	NVNS	8.26E-01 (9.88E-03)	1.74E-01 (1.00E-02)	1.41E-01 (9.15E-03)	3.89E-01 (1.39E-02)	
MM	GM-EDA		2.14E-04 (6.80E-05)	1.79E-03 (4.18E-04)	1.34E-02 (2.03E-03)	8.02E-02 (5.97E-03)
	HGM-EDA	1.00E+00 (6.80E-05)		9.20E-01 (5.97E-03)	9.87E-01 (2.03E-03)	9.98E-01 (4.18E-04)
	AGA	9.98E-01 (4.18E-04)	8.02E-02 (5.97E-03)		9.20E-01 (5.97E-03)	9.87E-01 (2.03E-03)
	VNS	9.87E-01 (2.03E-03)	1.34E-02 (2.03E-03)	8.02E-02 (5.97E-03)		9.20E-01 (5.97E-03)
	NVNS	9.20E-01 (5.97E-03)	1.79E-03 (4.18E-04)	1.34E-02 (2.03E-03)	8.02E-02 (5.97E-03)	

Table 6: Probability of an algorithm to be in the top-k ranking.

		Top 1	Top 2	Top 3	Top 4	Top 5
PLD	GM-EDA		4.23E-02 (3.10E-03)	9.65E-02 (6.33E-03)	2.34E-01 (1.34E-02)	
	HGM-EDA		6.98E-01 (1.11E-02)	9.17E-01 (5.88E-03)	9.92E-01 (9.45E-04)	
	AGA		7.76E-01 (1.23E-02)	9.47E-01 (5.25E-03)	9.96E-01 (6.72E-04)	
	VNS		3.09E-01 (1.07E-02)	6.53E-01 (1.26E-02)	9.38E-01 (5.11E-03)	
	NVNS		1.74E-01 (8.43E-03)	3.86E-01 (1.37E-02)	8.40E-01 (1.06E-02)	
PLG	GM-EDA		4.22E-02 (3.07E-03)	9.66E-02 (6.48E-03)	2.35E-01 (1.37E-02)	
	HGM-EDA		6.99E-01 (1.20E-02)	9.18E-01 (6.14E-03)	9.92E-01 (9.72E-04)	
	AGA		7.77E-01 (1.12E-02)	9.48E-01 (4.69E-03)	9.96E-01 (5.90E-04)	
	VNS		3.08E-01 (1.18E-02)	6.53E-01 (1.45E-02)	9.38E-01 (5.76E-03)	
	NVNS		1.73E-01 (8.13E-03)	3.84E-01 (1.42E-02)	8.39E-01 (1.07E-02)	
BT	GM-EDA		8.39E-03 (1.23E-03)	5.41E-02 (5.12E-03)	2.57E-01 (1.21E-02)	
	HGM-EDA		7.19E-01 (1.22E-02)	9.09E-01 (6.33E-03)	9.87E-01 (1.46E-03)	
	AGA		7.82E-01 (1.11E-02)	9.36E-01 (5.10E-03)	9.92E-01 (9.96E-04)	
	VNS		3.12E-01 (1.22E-02)	6.47E-01 (1.27E-02)	9.18E-01 (5.86E-03)	
	NVNS		1.79E-01 (9.82E-03)	4.54E-01 (1.35E-02)	8.47E-01 (8.45E-03)	
MM	GM-EDA		6.73E-04 (1.63E-04)	7.66E-03 (1.24E-03)	8.73E-02 (7.06E-03)	
	HGM-EDA		9.92E-01 (1.24E-03)	9.99E-01 (1.63E-04)	1.00E+00 (1.76E-05)	
	AGA		9.13E-01 (6.90E-03)	9.92E-01 (1.22E-03)	9.99E-01 (1.46E-04)	
	VNS		8.60E-02 (6.75E-03)	9.14E-01 (6.75E-03)	9.93E-01 (1.08E-03)	
	NVNS		7.60E-03 (1.22E-03)	8.66E-02 (6.90E-03)	9.20E-01 (5.82E-03)	

2 Model selection using Bayes' factors

We have several probabilistic models on permutation spaces which we could use to describe the experimental data we have collected from the comparison of several algorithms. The problem we are faced with is choosing among these models the one that fits the data the best.

A number of different approaches can be used to accomplish this goal, for example, Bayes Factors [6], Bayesian Information Criterion [7] or the Minimum Descriptive Length [4]. Each approach has a number of difficulties to consider and several decisions have to be made with careful consideration to what kind of analysis and data we are dealing with. In this section, we illustrate the use of Bayes' factors for model selection as a form of generic example while highlighting the important decisions that have to be made.

When comparing the Bayes' factors of two models, we are interested in the posterior probability of the models M_i, M_j to describe a given set of rankings of algorithms (permutations), i.e., data $S = \{\pi_1, ..., \pi_p\}$ as follows:

$$\Pr[M_i|S] = \frac{\Pr[S|M_i]\Pr[M_i]}{\Pr[S]}, \tag{1}$$

$$\Pr[M_j|S] = \frac{\Pr[S|M_j]\Pr[M_j]}{\Pr[S]} \tag{2}$$

Then, our focus is in the quotient:

$$\frac{\Pr[M_i|S]}{\Pr[M_j|S]} = \frac{\Pr[S|M_i]}{\Pr[S|M_j]}\frac{\Pr[M_i]}{\Pr[M_j]} \tag{3}$$

where $\Pr[S|M_k]$ is the marginal likelihood probability for model M_k and $\Pr[M_k]$ is the prior probability of model M_k.

The marginal likelihood probability is obtained by integrating over the parameter space of the probabilistic model, i.e.:

$$\Pr[S|M_k] = \int_{\Omega_\theta} \Pr[S|\theta, M_k]\Pr[\theta|M_k]d\theta \tag{4}$$

which is in fact the normalization constant that we get when conducting Bayesian inference of the given model parameters. The only difference with Equation (11) in our paper is merely in notation. In addition, we need to define a prior probability distribution for the models $\Pr[M_k]$. Both the computation of the marginal likelihood probability $\Pr[S|M_k]$ and the choice of the models' prior $\Pr[\theta|M_k]$ have a number of important issues that are worth to be reviewed:

- First, notice that we need to obtain the marginal likelihood probability $\Pr[S|M_k]$ that usually involves multi-dimensional integrals or sums which are inherently difficult to calculate.
- Furthermore, we need to consider that these marginal likelihood probabilities $\Pr[S|M_k]$ are really sensitive to the choice of the prior distribution of the model parameters $\Pr[\theta|M_k]$ even if the selection of such prior does not affect that much the posterior distribution itself $\Pr[\theta|S, M_k]$. The problem with this in the context of model selection is that, ideally, we don't want that the choice of such prior distribution affects the marginal likelihood distribution that we use to compare the models, i.e., $\Pr[S|M_k]$.
- In addition, we have the fact that we need to describe the prior probabilities for each of our models $\Pr[M_k]$ and, in practice, it is usually difficult to come up with sensible ways of describing those.
- Finally, there is the issue of choosing a threshold of the Bayes factor to determine whether a model is preferred over another.

2.1 Bayes' factors with Approximate Bayesian Computation

In practice, obtaining the marginal likelihood probability $\Pr[S|M_k]$ is the main challenge to come up with a closed-form expression in Bayesian analysis. This is the main reason why, in our work, we used Markov Chain Monte Carlo methods to obtain samples of the posterior distribution of the probabilistic models in permutation spaces.

To circumvent this difficulty in the context of model selection with Bayes' factors, a number of methods can be used. Again, a number of choices have to be made. We opt to exemplify how this can be accomplished by using the Approximate Bayesian Computation Rejection Algorithms [2].

The Approximate Bayesian Computation Rejection Algorithm starts by sampling hierarchically a model from the models' prior distribution $\Pr[M_k]$ and then sampling from the posterior distribution of that model $\Pr[\theta|M_k]$. After this, a simulation is conducted in which, in our context, a set of permutations $\hat{S} = \{\hat{\pi}_1, ..., \hat{\pi}_p\}$ is sampled using the the probabilistic model on permutation spaces such as $\hat{\pi} \sim \Pr[\hat{\pi}|\theta]$. An acceptance criterion $\rho(S, \hat{S}) \leq \epsilon$, e.g. a difference between datasets and a threshold parameter $\epsilon > 0$ are used to determine if the simulated set of permutations is similar to the actual data. The relative acceptance frequencies for the different models approximate the posterior distribution for these models $\Pr[M_k|S]$. Again, the choice of the distance measure, which is itself an active area of research [8, 5], the threshold parameter and the number of simulations are important choices to be made.

2.2 A practical example

As a practical example, we illustrate the use of Approximate Bayesian Computation to obtain the Bayes' factor between the Bradley-Terry model (BT) and the Plackett-Luce model with Dirichlet prior (PLD) when used to make inferences on the real case study. We set an equal prior for each model such as $\Pr[M_i]/\Pr[M_j] = 1$ where $\Pr[M_i]$ denotes the prior distribution for the BT model and $\Pr[M_j]$ denotes the prior distribution for the PLD model. We run 1000 iterations of the Approximate Bayesian Algorithm using $\epsilon = 0.7$ and we use as a distance measure between datasets the mean Kendall-tau distance between the nearest pairs of permutations in S vs. $\hat{S}$. More details are provided in our code repository[1]. As a result, we obtain the Bayes' factors presented in Figure (1).

The figure presents the Bayes' factor between the model represented in the vertical axis with respect to the model represented in the horizontal axis. As can be seen, the Bayes factors give more weight to the BT model and PL model over the MM and in general gives similar weight to the BT model and the PL model. Nevertheless, the reader should be aware of the number of choices we have made when selecting a distance measure between datasets or when setting the ϵ parameter that is not, in general, easy to justify.

[1] Available at: `https://github.com/ml-opt/BayesPermus`

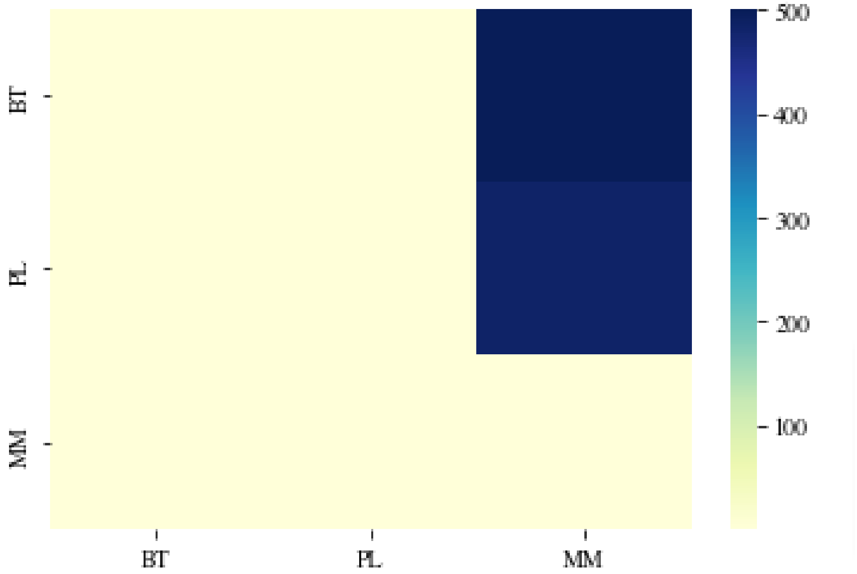

Figure 1: Bayes' factors between pairs of models. The values represented in the heat map corresponds to the Bayes factor between the model in the vertical axis and the model in the horizontal axis.

3 Sensitivity analysis of the prior distributions and their hyperparameters

A key question in Bayesian analysis is the effect of the prior on the posterior, and how we can measure this effect. This is an important question that has been widely and actively addressed in the literature [3, 1] In this section, we illustrate this type of analysis with our case study that uses real data by comparing the posterior distribution of the different probabilistic models in permutation spaces. Implementation details to reproduce the sensitivity analysis are provided in our source code repository.

We define three different prior distributions, an uninformed prior, an informed prior and a deceptive prior. The uninformed prior gives equal contribution to each algorithm such as no assumption is made on which algorithm should perform better than another. For the informed prior, we give more weight to the best performing algorithm a smaller weight to the second-best performing algorithm and so on until we give the smallest weight to the worst-performing algorithm. For the deceptive prior, we proceed as before but reverse the way we set the weights.

Figure 2 shows the probability of each algorithm being the top-ranked algorithm using the different prior distributions. The first column represents the results for the BT model, the second column represents the results for the PLD model and the third column represents the result for the PLG model. The horizontal axis represents the kind of prior used to make the inferences in which we use the name of the model with an initial capital letter that stands for the uninformed prior (U), the informed prior (I) and the deceptive prior (D).

In general, when we use the empirical data to bias the prior distribution towards the best algorithms, i.e., when using the informed prior, we see that the posterior distribution is slightly more in agreement with the ground truth when compared with the deceptive prior. However, the differences are practically negligible in this case and definitely do not alter the general conclusions. This same trend is observed when we compare other posterior summaries, such as the probability of each algorithm outperforming another.

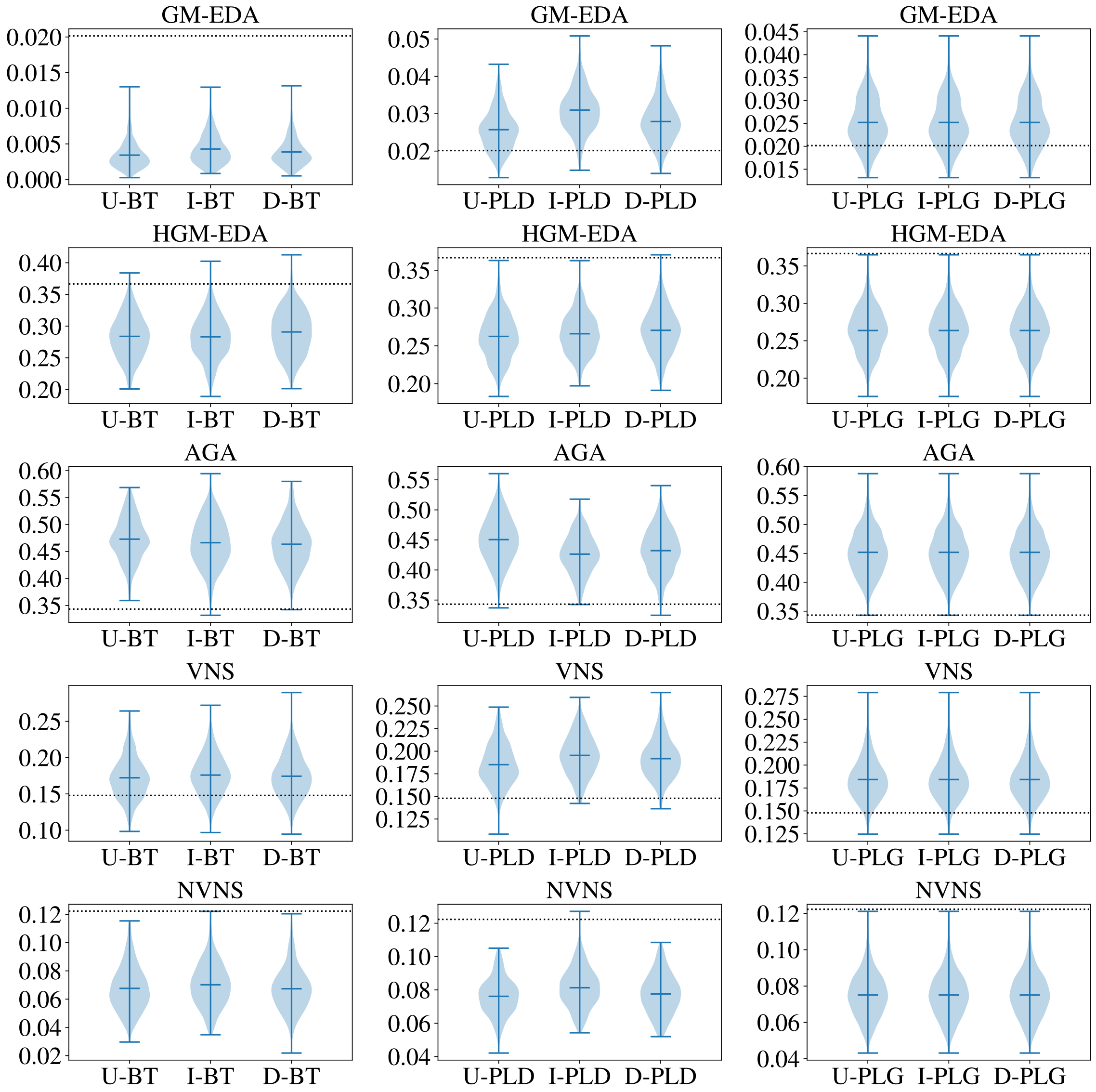

Figure 2: Probability of each algorithm being the top-ranked algorithm. The first column represents the results for the BT model, the second column represents the results for the PLD model and the third column represents the result for the PLG model. The horizontal axis represents the kind of prior used to make the inferences in which we use the name of the model with an initial capital letter that stands for the uninformed prior (U), the informed prior (I) and the deceptive prior (D).

References

[1] Borja Calvo et al. “Bayesian Performance Analysis for Black-Box Optimization Benchmarking”. In: *Proceedings of the Genetic and Evolutionary Computation Conference Companion*. GECCO ’19. Prague, Czech Republic: Association for Computing Machinery, 2019, pp. 1789–1797. ISBN: 9781450367486. DOI: 10.1145/3319619.3326888. URL: https://doi.org/10.1145/3319619.3326888.

[2] Katalin Csilléry et al. “Approximate Bayesian Computation (ABC) in practice”. In: *Trends in Ecology and Evolution* 25.7 (2010), pp. 410–418. ISSN: 0169-5347. DOI: https://doi.org/10.1016/j.tree.2010.04.001. URL: https://www.sciencedirect.com/science/article/pii/S0169534710000662.

[3] Fatemeh Ghaderinezhad and Christophe Ley. “On the Impact of the Choice of the Prior in Bayesian Statistics”. In: *Bayesian Inference on Complicated Data*. Ed. by Niansheng Tang. Rijeka: IntechOpen, 2019. Chap. 1. DOI: 10.5772/intechopen.88994. URL: https://doi.org/10.5772/intechopen.88994.

[4] Peter Grünwald and Teemu Roos. “Minimum description length revisited”. In: *International Journal of Mathematics for Industry* 11.01 (2019), p. 1930001. DOI: 10.1142/S2661335219300018.

[5] Facundo Mémoli. “Distances Between Datasets”. In: *Modern Approaches to Discrete Curvature*. Ed. by Laurent Najman and Pascal Romon. Cham: Springer International Publishing, 2017, pp. 115–132. ISBN: 978-3-319-58002-9. DOI: `10.1007/978-3-319-58002-9_3`. URL: `https://doi.org/10.1007/978-3-319-58002-9_3`.

[6] Richard D. Morey, Jan-Willem Romeijn, and Jeffrey N. Rouder. “The philosophy of Bayes factors and the quantification of statistical evidence”. In: *Journal of Mathematical Psychology* 72 (2016). Bayes Factors for Testing Hypotheses in Psychological Research: Practical Relevance and New Developments, pp. 6–18. ISSN: 0022-2496. DOI: `https://doi.org/10.1016/j.jmp.2015.11.001`. URL: `https://www.sciencedirect.com/science/article/pii/S0022249615000723`.

[7] P. Stoica and Y. Selen. “Model-order selection: a review of information criterion rules”. In: *IEEE Signal Processing Magazine* 21.4 (2004), pp. 36–47. DOI: `10.1109/MSP.2004.1311138`.

[8] Nikolaj Tatti. “Distances between Data Sets Based on Summary Statistics”. In: *Journal of Machine Learning Research* 8.5 (2007), pp. 131–154. URL: `http://jmlr.org/papers/v8/tatti07a.html`.