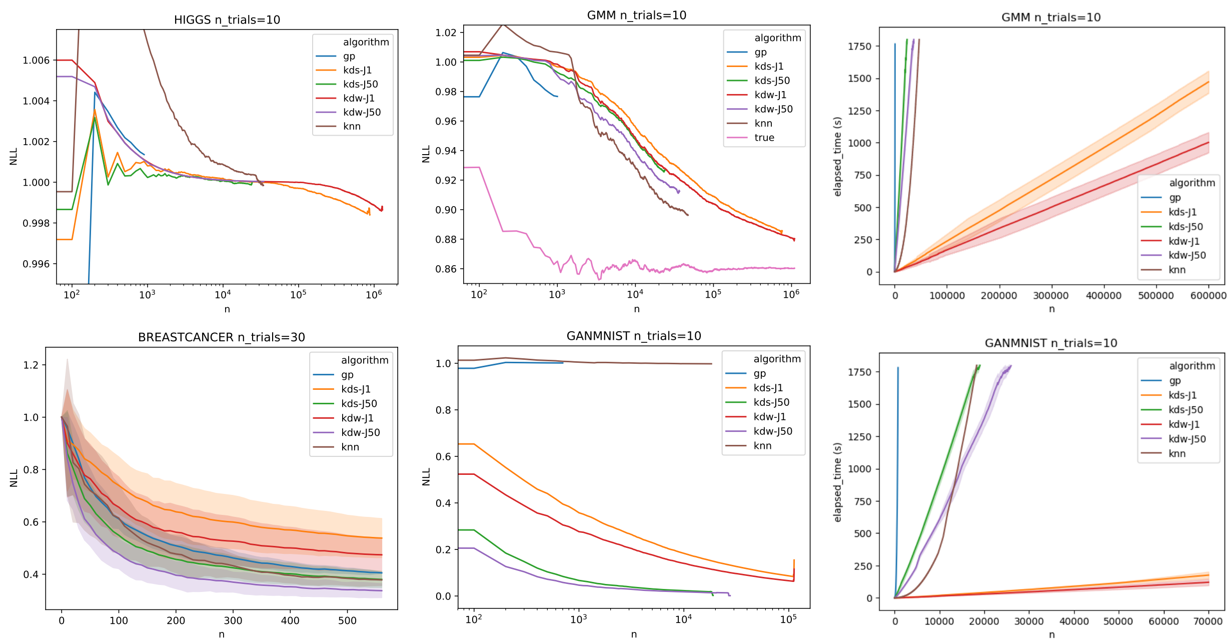
General. We thank the reviewers for their detailed comments and suggestions. In order to address the remarks, Figure 2 will be replaced with the following one that uses a log scale on the x-axis for large n cases to improve the readability. The zoom on GMM is no longer needed. We replace the plot of NLL vs elapsed-time for GMM by a comparison of n vs elapsed time (in seconds) for GMM ($d = 2$) and GANMNIST ($d = 784$), as suggested by reviewer #2.

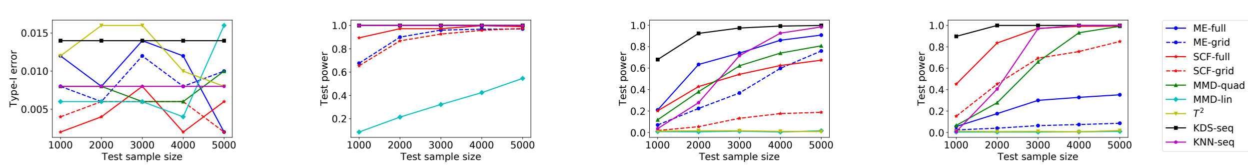


Review #1 1) We forgot to protect the capital letters in bibtex files. It will be corrected. 2) “Given the different computational times ... they are of difficult interpretation": we hope that the log scale on x-axis and the new figures relating n and elapsed time will help the interpretation.

Review #2 Model Switching is intuitively explained in the introduction (lines 39-42). We will add the term “switch": “*The resulting* **switch** *distribution still has a Bayesian form...*". Context Tree Switching builds on this idea and all the required definitions are included in this paper. As remarked by reviewer #4, Context Tree Weighting is a well-known technique in information theory and its applications and the necessary definitions are included in our paper. “I found the description of the problem in 6.2 hard to understand and so it was hard to interpret the results." We will add a reference to “Casella & Berger, Statistical Inference, Ch. 8".

Review #3 See response to Review #2.

Review #4 1) "... more intuition regarding the use of the random data rotations.": random rotations are only used for the experiments on two sample testing to avoid unfair comparisons: without rotating the data, kd-trees easily find the direction featuring the differences, resulting in much higher performances – a fact illustrated by the following novel figure that will be included in the appendix.



(a) SG. $d = 50$. (b) GMD. $d = 100$. (c) GVD. $d = 50$. (d) Blobs. $d = 2$.

2) Regarding the switching rate, the following remark will be added: *In [29], the authors observe better empirical performance with $\alpha_m^\gamma = n^{-1}$ for any cell γ, where n is the number of samples observed at the root partition Ω when the m-th sample is observed in γ. With this switching rate they were able to provide a good redundancy bound for bounded depth trees. In our unbounded case, we observed a better empirical performance with $\alpha_m^\gamma = m^{-1}$.*

3) We will cite the suggested papers.