Referee 2: Report and responses

Strengths

1. The paper presents a novel statistical approach for going beyond fixed working-point jet flavour taggers when analyzing multi-component and multi-jet event datasets.

2. The ideas presented in this work have multiple potential generalizations and applications to other similar problems in high energy particle phenomenology.

Weaknesses

1. The methods are tested on toy examples and without direct comparison to existing established approaches. Consequently, their performance and viability in realistic scenarios is difficult to evaluate.

Report

In the submitted manuscript, the authors explore a bayesian statistical method to infer mixtures of processes producing events with multiple quark-flavoured hadronic jets without relying on fixed working point jet-taggers. The method aims to extract the weights of the process components as well as the jet discrimant performances in-situ. It leverages the co-occurrances of different jet flavour combinations, the knowledge of possible truth-level flavour multiplocities predicted by individual component processes, as well as general assumptions on the smoothness and unmodality of physical jet flavour discriminants. The later are implemented through a choice of different Bayesian priors leading to several distinct statistical models with varying prediction and computational performances.
The presented research is topical and potentially highly relevant for experimental collaborations at the LHC. The approach also has potential to be developed further and applied to other similar datasets using continuous discriminants as part of object/event selection.
However in its present form, the paper fails to convincingly demonstrate the full potential of the method as well as pinpoint all of its potential weaknesses. Thus I would ask the authors to fully address the issues listed below before making my recommendation.

We thank the Referee for reading the manuscript and their report. The questions and proposals have considerably improved the work.

Requested changes

1. The authors should quantify the inference power of their method by comparing it to a more traditional cut-based approach.

We thank the Referee for this proposal, we have added a new Section (3.4) which addresses this point. In this section we compare the classification power of the proposed method to the usual cut-based method still used to select events in some analyses. In order to do such a comparison we use a ROC curve and we find that the proposed Bayesian framework has a larger Area Under the Curve (AUC) than the usual cut-based method. We also study the potential bias using both methods, also finding that the Bayesian method is likely to be less biased. The calculations, details and discussions can be found in the new version of the manuscript.

2. In the toy example the "truth-level" fractions of the two process admixtures are both reasonably large. In practice however, one often deals with signal to background ratios which are orders of magnitude below unity (before applying jet flavour discrimination). It is not clear how the proposed method would perform in this kind of scenarios.

The Referee is correct also in this point: many times the signal fraction is much smaller than the backgrounds. We have simulated samples with 10% and 1% signal fraction (assuming signal as bbbb) and run over these samples the same inference as in the others. Interestingly we find a very good extraction of the parameters even for the 1% case. The reason for this is that the framework utilizes the class ccbb to learn the individual b-component, while simultaneously using this learning to learn the very small bbbb fraction. Results can be found in the new Fig. 8. Overall, we believe that using Bayesian techniques including a carefully designed prior is even more useful in these imbalanced data scenarios.

To study smaller fractions we would need to simulate an order of magnitude of more events (5k), in such a way that 0.1% has an absolute value of at least ~5 events in expectation. We consider that a further exploration of the sensitivity for extremely small signal fraction should be done elsewhere.

3. From the existing discussion it is also not clear how the different models' computational and discriminative performance scales with the dimensionality of the problem, both in terms of the number of jets, number of events and number of admixtures. Thus it is difficult to evaluate its applicability to other, real world scenarios.

This is a good point that we want the method to achieve both a high computational scalability, and a high finite sample convergence rate. Our Figs. 10 and 11 demonstrate how our model fitting improves with the number of events in the sample. (Fig. 12 repeats these figures for different seeds to assess the stability upon different datasets.) This fitting performance is quantified both in the log predictive density and in root mean squared error.

In addition to varying event size, we also demonstrate the performance for 1D events and 4D events. And in the new version also we further test our method and benchmarks along a variety of signal fractions (20%, 10% and 1%). Our proposed method exhibits a robust advantage against various benchmarks in these extensive comparisons.

Of course our simulated study has not exhausted all varying quantities. We consider that having the scripts in the Github can facilitate the task for any reader to perform any tailored simulation..

4. More specifically on the last point: From figures 3 & 4 it seems that for low number of events (N=100), the posterior distributions of admixture fractions are significantly inconsistent with their true values, thus exhibiting bias and potentially putting the method's robustness in question.

What the Referee suggests is true as it is in the figure. However, the intention in putting these plots here is to showcase how the convergence (as N becomes large) drifts from a fixed prior to the truths in the posteriors. If we would know that only counts on 100 events in the dataset, then we probably would have (1) chosen less bins to reduce the number of unknowns in comparison to the number of datapoints, as well as to avoid fluctuations, and (2) chosen a better designed prior that reflected the physical knowledge.

5. In their current approach, the authors choose to bin the otherwise continuous discriminant distribution to build tractable statistical models. It is not clear how essential this simplification is and if there are ways to model the full continuous discriminant response (perhaps using techniques similar to PDF extraction by the NNPDF collaboration).

This is a good point that the present paper only considers the discretely binned b scores. Currently it is a model choice, and some of our model techniques are specific to this discrete setting. Extension to continuous scores is possible, for example we may replace the discrete Dirichelt prior by a Dirichelt process prior. We think this extension is beyond the scope of this paper, and we leave it for future investigation. We also note that the discrete binning is not purely a simplification; rather, it makes our tool more robust against moderate measurement errors. Perhaps this binning is especially relevant for b-scores, as the observed b-scores are point estimates from some neural nets, whose uncertainty is typically ignored.

6. The labeling of true and prior c-fractions in Figure 2 are confusing. Why is the latter drawn as a horizontal line? Shouldn't it be a (flat?) distribution? Then I would suggest a different visualization, to make this more obvious.

In fact, the horizontal line *is* a flat distribution. There was a mistake in the previous version that may have confused the Referee: the horizontal line was going beyond its (0,1) range on the x-axis. We have fixed this in the new version, and also added a sentence in the caption clarifying that only the posterior is filled. Whereas the dotted horizontal line is a flat distribution, but not filled to avoid superpositions of filled areas.

7. The text contains several typographical and grammatical errors as well some possibly incorrect statements. Some examples to be revised a listed below:

a) Second sentence on page 2 is over the top. To be removed or revised.

Done.

b) Page 2, line 9: expression "mutual information" is not clear. It may or may not refer to the formal, technical term. To be revised.

Yes it is the technical term. The intention is to not use the word *correlation* which is not strictly correct.

c) Page 2, paragraph 3, first line: Acronym HEP is never used. To be removed.

Done.

d) Page 9, 2nd paragraph of Sec. 3.3, line 3: the term "hack" is inappropriate in this context. To be revised.

Done.

d) Page 9, 3rd paragraph of Sec. 3.3: the last sentence is grammatically incorrect and should be revised.

We have changed it and added another sentence for clarification.

e) Page 10, last paragraph above Sec. 4: the expression "light statistical analysis" is vague. It should be revised and made more precise.

Done.

f) Page 10, last paragraph above Sec. 4. The meaning of the next to last sentence starting with "We find that the Unimodal..." is unclear. It should be revised.

We agree with the Referee that it was confusing. We have rephrased it to make it clearer.