Referee 4: Report and responses

1. Well written and presented
2. Toy application of a less used Bayesian statistical method to a HEP problem

Weaknesses

1. A bit too much assumption of familiarity with statistical issues and distributions for a physics audience
2. Lack of clarity on generation and inference methodology, and on which physics information makes the problem tractable
3. Does not go beyond toy studies based on sampling from 1D curves which could have come from any topic area: the coupling to HEP is rather shallow (other than in limiting the multivariate mixtures to assume pair production of heavy flavour)

Report

A nice demonstration of Bayesian mixture methods to toy problems of tagging-score distribution inference, using four different statistical models. Convergence is shown to require additional physics information such as correlations between multiple jets in an event (this is in need of some clarification) or knowledge of e.g. smoothness and unimodality of the posterior distributions. The amount of physics in the paper is rather small, being more a context for statistical application than central to the conclusions as currently written. No physics simulation is explicitly used, instead simply sampling from provided 1D distributions, and inferring them as a closure test. On the basis that the connection between the physics content and the statistics content is rather small, I think this does not fulfil the SciPost Physics criteria of novelty and synergy between the two fields, though it is a useful demonstration of statistical methods to a physics audience: Physics Core is more appropriate on the declared conditions. It is well written and presented, and a solid contribution to the literature, but does need clearer explanation and perhaps deeper study of the additional information that makes inference possible in the 4-jet case.

We thank the Referee for reading the manuscript and their detailed report. We are grateful for all their comments which have considerably improved the quality of the work.

- In abstract and introduction, it would be helpful to clarify that "heavy flavour" in this work refers to b-jets and c-jets but not top quarks -- which are heavy flavour but kinematically distinct and cannot usually be associated to a single jet.

We have done so in the first appearance in the Introduction

- "Statically" -> "Statistically" at the end of p2

Done.

- Sub-sections 2.1.x: it would be useful to have more explanation of the distributions' features and motivations, for accessibility to physicists, and more motivation/discussion of the 4 options' pros/cons from a physics point of view

Perhaps the most intuitive way to summarize a stream of numbers is a histogram: to count the number of occurrences within each bin. This histogram is typically the “default” approximation to the density function, and can be viewed as a posterior inference with a Dirichelt prior with alpha=1, or a uniform prior (alpha being the Dirichlet distribution parameters). Starting from this purely “uninformative” prior, we designed three models to make use of the prior physical knowledge on the shape of the density. (1) The Dirichelt prior describes how concentrated the bin heights are, as alpha = infinity enforces only one bin to have mass and alpha = 0 enforces uniformity. (2) The self-normalized Gaussian process encodes the continuity of density function, as we believe

that the neighboring bins should have similar heights. This can be understood as a local smooth. (3) The unimodality prior enforces the density function to be unimodal, or the bin height only has one local mode. The continuity and unimodality assumptions will probably be even more relevant when modeling physically meaningful quantities. From our experiments, we find these structured priors (2 and 3) are able to extract more information from noisy data, and hence are generally recommended.

- Sec 3 start: "In this Section" -> "In this section" (not used as a proper noun here)

Done.

- Sec 3.1: it would be good to note that the identification of a hadronic jet with a specific parton flavour is not a complete picture -- as that is equating a colour-singlet object with a non-singlet and the QCD radiation structure of jets emerges from colour dipoles between partons -- but that it has served well as an approximation, particularly for heavy-quark tagging and to a much lesser extent for quark/gluon discrimination

We have changed *originated* by *initiated by* a b-quark, and so on, to remark what is pointed out by the Referee. Given the context, we considered that it was not worthwhile to go into further detail.

- Sec 3.1: LHC collaborations are also increasingly using continuous or pseudo-continuous b-tagging, in which scores are used directly or in differential binnings rather than below/above thresholds for fixing working points. This introduces challenges of calibration, but of course provides more nuanced information for analyses.

The Referee is correct about this, we have added in Section 3.1 some sentences about pseudo-continuous b-tagging and citing some papers in which it is used. We also comment that this is not the same as the presented work, but instead is about using a few calibrated working points. In any case, we agree with the Referee that this should be mentioned here.

- p6: "label switching along the inference process" needs some clarification: this whole sentence seems a bit cryptic to a typical physicist

We have added a sentence shortly explaining what is label-switching and a reference to get further information about it.

- Sec 3.2: an explicit statement of the computing setup used for these modelling studies would be useful. Stan has been implied but is not explicitly stated, with references to e.g. the ordered vector use appearing in "computing font" but without enough context to understand it. Equally, samples sizes, etc. would be useful to have stated. (I see the computing is specified in Appendix A. Either bring it forward into the main text -- I think think would be best -- or explicitly instruct the reader to see App A for details and avoid having the main text refer to features that require the computing context.

We agree with the Referee and we have added the requested sentence about the Stan language. We have also briefly described what is an ordered-vector in statistical programming and added the corresponding reference. We have added a reference to the Appendix for the computational details.

- Sec 3.2: why are these choices of GP hyperparams the appropriate/natural ones?

As we describe in text, we have tested inferring σ and ρ with some setup, but the improvement is not noticeable, whereas their inferred values range around the currently used values.

- Sec 3.2, p8: how is it _known_ that the reason for non-improvement in the GP case is "non-identifiability of the problem"? And what exactly does this jargon mean... that the absence of other variables with which to correlate means there is not enough information to unmix the distributions?

The Referee is correct, this is what it means in this case. There is an ambiguity in the model to describe the data in 1D, and therefore the model does not have enough information to extract the correct original distributions. This can be realized by counting data and unknowns: the full data is a 1D histogram, and this should be explained as the sum of two unknown distributions, henceforth there is a degeneracy since there is more than one way to obtain the data. In this discretized problem the unknowns would be the value of each distribution in each bin, and the mixing fraction coefficient. Whereas the data is the counting in each bin of all the dataset.

- p8: the b distribution seems to have semi-converged, but in the opposite direction of movement from the prior than what was required. (The Dirichlet case also has a bit of this -- less obviously converged, but with similar features.) What's the reason for that? Can this be sort of mismatch be predicted from the shapes of mixture functions that are being closure-tested?

The idea of this Fig. 2 in Sect. 3.2 is to show in a simple scenario the main idea of the inference process. The data does not have enough complexity to correctly provide the inner structure of the ongoing processes. Although some statements could eventually be worked out with the available results and some more that one may explore (more data, different priors, etc), we do not consider expanding more on these results in this work. Nevertheless, it could be interesting to explore in more detail the 1D case, specially with the Unimodal case (which in some cases it may control the aforementioned ambiguity) to study its potential scope, as discussed in the Discussion section.

- p8: unimodality does not seem fully guaranteed: both the b and c distributions in Fig 1 have inflections on their tails that come close to, or marginally drift into, multi-modality. In general is there a necessity that tagging score distributions will be unimodal when transported to a process other than that where they were trained/calibrated? For example, calibrations from ttbar events might well generate additional structures when applied to jets in events with more tops.

The Referee is correct and very sharp with these observations on which we agree. We do discuss this fact in the Discussion section (3rd from last paragraph). Although we do not go deep in the discussion, this is a high point for our follow-ups in this line of research. We will explore in next works using Secondary Vertex displacements and others, which yes can be expected to be unimodal.

- p8: it would be good to float Fig 2 to the [b]ottom of the page, to avoid the bit of text on model 3 getting "trapped" there, and easily missed by the reader

We thank the Referee for this suggestion, which we have suitably incorporated.

- p8/9: it's claimed that the unimodal mixture model posterior "approaches the true values" by comparison to the priors (which it does not use). It does seem closer than the prior-guided models, but still shows an excess near the c peak and a deficit on its RHS, similar to the priors -- which is presumably a coincidence. As the priors aren't shown on the central plot, it's hard to exactly compare these, but I think a bit more discussion of the "approach" is needed to soften the conclusion, as it may be doing better but seems to be converging to something significantly far from the truth in several bins near the peak. The blue b-jet distribution is much more poorly converged and described and this should be mentioned. The point-estimate model has _very_ similar features on the c distribution, which deserves some discussion as it's too close to be coincidence, and now the b-jet distribution is better converged (on to one of the two implied fits from the unimodal posteriors) but is not accurate other than on the position of the peak. Basically, there's a lot more going on here, with interesting grouping of models, and it's too glib just to say that the last two models have "a fair approach to the two curves" -- I think that's an overstatement of the level of improvement, particularly for the b-tag distribution.

Our text was not clear enough: we refer to the priors within the same Unimodal model, and not the other models priors. We have clarified this in the new version of the text. It *does* use priors, these are "arbitrary

smooth unimodal curves with the c-distribution above the b-distribution in the first bins". These are shown in Fig. 2, at column 1 and row 3. We also added a sentence of why the c-distribution has better agreement.

(Also notice that the dashed lines in Dirichlet and GP models are the mean of the priors, not the priors itself.)

- Sec 3.3: the generation of the 4-jet data is not made clear. I can understand that physics process types could induce interesting correlations containing extra information, but here I _think_ the toy model just involves sampling again from the two GN1 curves: is this correct? So what structure (other than normalisation) is there to be exploited by the statistical model that doesn't just factorise into 4 x 1D problems? This may seem obvious to the authors, but I suspect not for many readers. Is there any coherent ordering of the jets, i.e. which scores go in which tuple entries, e.g. corresponding to pT ordering of the jets? Or is that random, in which correlations would presumably be eliminated?

It is correct that data is sampled from two GN1 curves, but with the important relationship being that: of the four numbers, they can only come in pairs from each curve (i.e. it cannot be e.g. cccb). This is translated into the likelihood in a precise way. Hence, it is not a 4 x 1D problem, because at the event-by-event level samples can only be of the 3 classes cccc, ccbb and bbbb. There is no ordering of the jets, and this is included in the likelihood expression, which takes the ccbb class as any of its 6 possible combinations.

Another leverage for the inference is that not only the 4 jet b-scores are conditionally independent, but also that they are all constructed from only 2 individual components, and therefore considerably reducing the number of unknowns. We have added a paragraph remarking this in the Discussion.

- It's also not clear to me what are the relative normalisations of the two models (b and c) in either of the toy studies: depending on the signal channel of interest, and its main backgrounds in the signal region, the b and c rates could be enormously different... and this would presumably inject some physics into the mixture inference?

We may be missing/not understanding what is being asked here, but maybe is is already clarified with the previous answer. b and c have distributions in GN1 -which integrate to 1-, and the setup (in the 4D) is such that events can only be cccc, ccbb and bbbb. The relative fractions of these classes in the dataset have different values, as we state in the text.

- Sec 3.3: Figs 3-6 _need_ to be brought forward to p10 onwards -- it is extremely difficult to read and understand when the text discusses figures located 6-8 pages away.

We are thankful for this observation. We have performed a modification in this direction and now we believe the draft is easier to read.

- p10: The far-away Figs 7-10 also need to be moved closer. The text here discusses them in turn, so maybe they would work best inline, between each paragraph of discussion text.

Idem here. We have created an appendix for the summary figures, and just refer to them as Appendix figures. They are important, but the main interest at this level is to showcase one case in detail, and these figures are here to show that the method works well for many other seeds.

- Discussion: As earlier, it's unclear where the extra information is coming from in the 4-jet case. Either there are correlations in the toy-data sampling which need more explanation, or is the extra information the restriction to even numbers of b-jets? This would be a very good thing to be clear about, as it's key to understanding what information can be leveraged to make the mixture inference tractable. This is also key in the final paragraph of the discussion, which seems to be proposing a mechanism for inferring true distributions of observables -- which of course is the whole aim of differential measurements and unfolding -- but where the

"multidimensionality" discussed is, I think, of a different character from that used here. (I am not convinced by this paragraph, since of course extraction of distributions this way has a lot of unconsidered prior art, but it is possible that Bayesian mixtures could add something in concert with established methods.)

The Referee is correct, the restriction to even number of jets is an important ingredient to make the inference tractable.

The Referee is also correct in the subtlety that the multidimensional discussed here is in some sense different to the one in the main multijet problem. We thank the Referee for this observation, and we have added a paragraph in Discussion to clarify this important point. We have also clarified in the last sentence that the expected advantage in that proposed problem is the conditional independence.

- I expected there to be a non-toy study, e.g. using MC generation for a 4t model and seeing how the methods perform on that. I suppose obtaining some approximation to a b-tag score for real MC was too difficult, but it would have been interesting. Without physics input, the paper seems to "just" be a demonstration of Bayesian mixture-inference closure on two distributions that happen to have come from b-tagging, with a little physics content in the even-Nb restriction of the mixtures in the 4-jet case. I think this is fine, but it's good to be clear about the extent to which there is physics content in this application of inference.

The Referee is correct in their observations. We should mention that the whole enterprise of finally including these techniques in observables is quite large and we find it more suitable to divide it into building blocks. Moreover, since each building block has its complications and novelties we find that this is also a better idea from the point of view of better understanding each advancement in depth.

In particular, using physics MC would need a b-tagger with some detail. However (as suggested by the Referee), Delphes has a very simple tagger that would have not fulfilled the requirements. Using true GN1 was beyond the purposes of this work, and using another tagger, such as a Vertex Displacement, would have been another project. As a matter of fact the latter is what we are currently considering, since it is a physical distribution. There will be follow-ups about this and previous projects, and we expect to be able to correctly assess its advantages in a realistic scenario, which would satisfy the Referee expectations indicated in this point.

Available results, in any case, already allow us to do an apple-vs-apple comparison with current methods by comparing the classifying power for a multi-jet signal, in this case bbbb. With this purpose we have added Section 3.4 in which we compare the method to usual cut-based methods. We compare both methods through their corresponding ROC curve, finding that the current Bayesian framework is a better classifier.

- There are quite a few formatting issues in the references, e.g.:
- [1], [2], [4], [7], [18], [23]... capitalisation issues in ATLAS, LHC, QCD, whole title, ABCD, etc.
- [9] Different author-name formats from the rest
- [23] T. A. Collaboration, [24] S. D. Team, ... should be proper collaboration names
- [20], [21], [22]: missing proper experimental collaboration names
- [21] s -> \sqrt{s} etc. and generally fix title formatting

We are grateful that the paper has passed through such a keen eye. We have adopted most of the suggestions, a few of them come as they are in the chosen Latex style, and we find it suitable that the Journal homogenizes this to its style.