Inferring correlated distributions: Boosted top jets

Ezequiel Alvarez[1*], Manuel Szewc[1,2†], Alejandro Szynkman[3‡],
Santiago Tanco[3°] and Tatiana Tarutina[3§]

1 International Center for Advanced Studies (ICAS) and ICIFI-CONICET, UNSAM,
25 de Mayo y Francia, CP1650, San Martín, Buenos Aires, Argentina
2 Department of Physics, University of Cincinnati, Cincinnati, Ohio 45221, USA
3 IFLP, CONICET - Dpto. de Física, Universidad Nacional de La Plata,
C.C. 67, 1900 La Plata, Argentina

★ sequi@unsam.edu.ar , † szewcml@ucmail.uc.edu , ‡ szynkman@fisica.unlp.edu.ar ,
○ santiago.tanco@fisica.unlp.edu.ar , § tarutina@fisica.unlp.edu.ar

Abstract

Improving the understanding of signal and background distributions in signal-region is a valuable key to enhance any analysis in collider physics. This is usually a difficult task because –among others– signal and backgrounds are hard to discriminate in signal-region, simulations may reach a limit of reliability if they need to model non-perturbative QCD, and distributions are multi-dimensional and many times may be correlated within each class. Bayesian density estimation is a technique that leverages prior knowledge and data correlations to effectively extract information from data in signal-region. In this work we extend previous works on data-driven mixture models for meaningful unsupervised signal extraction in collider physics to incorporate correlations between features. Using a standard dataset of top and QCD jets, we show how simulators, despite having an expected bias, can be used to inject sufficient inductive nuance into an inference model in terms of priors to then be corrected by data and estimate the true correlated distributions between features within each class. We compare the model with and without correlations to show how the signal extraction is sensitive to their inclusion and we quantify the improvement due to the inclusion of correlations using both supervised and unsupervised metrics.

1 Introduction

The Large Hadron Collider (LHC) is entering the High Luminosity era. However, in many cases the $\mathcal{O}(10)$ gain in luminosity will not translate straightforwardly to a relevant gain in precision unless it is accompanied with a more detailed understanding of the detectors and the underlying physics and distributions followed by the variables involved in each process. In fact, detector knowledge and physics modeling are intimately related, with improvements in one area allowing for improvements in the other.

In this work we are interested in developing techniques to infer underlying distributions corresponding to LHC analyses. From a statistical point of view, a given selection of events can always be modeled as a mixture of classes: each event belongs to a given class, which could be, for example, signal and the corresponding backgrounds.[1] Bayesian techniques are a powerful means of addressing these kinds of problems. By modeling events as samples from a probabilistic distribution, one can unfold the probabilistic model to latent variables. This unfolding allows for the incorporation of important prior knowledge for each one of the latent variables, thereby enhancing the posterior performance and deepening the understanding of the system. These techniques have been widely used in the Machine Learning industry (see Ref. [1] for a description of the recommended Bayesian workflow) and have been known to the LHC community for quite some time (see *e.g.* Ref. [2] for discussions about mixture models in the context of frequentists and/or Bayesian analyses) with broader advances in Bayesian computation becoming more popular in LHC applications in recent years [3–16].

To ensure that inference yields appropriate distributions, a reasonable inductive bias needs to be baked into the model. Focusing on collider applications, there have been different strategies. The more common, and most storied, tradition is to implicitly define the class-dependent distributions using simulated events and consider as free parameters only the class fractions. This is the most powerful approach provided that the simulators are a perfect model of the data. However, we know this not to be the case, and so it is standard practice to extend the model using nuisance parameters to account for different sources of mismodeling, including by the simulators themselves. This is not guaranteed to be sufficient, specially for the high-complexity, high-statistic environment of the LHC, and thus data-driven techniques have long been a staple of the community toolkit [17–19]. However, data-driven techniques also need inductive bias, mainly to restrict the possible shapes of the probability distributions involved. For continuous variables, one can achieve this by imposing a particular parametric family of

---

[1]Interference between processes implies that they need to be grouped in the same class. In this work we assume the signal does not interfere with the background and thus it is always its own separate class.

high-dimensional distributions, such as a multivariate gaussian, to which each class-dependent distribution belongs. For discrete variables, such as bin counts, this is trickier to achieve.

Another possibility that side-steps the need to fully specify a multi-dimensional distribution but still restricts the problem is to assume that all relevant observables –or the dimensions of the problem– are *conditionally independent* or uncorrelated[2] within each class. This results in a large enhancement in the inferring power, since the problem is reduced to learning sets of one-dimensional distributions and there is no need to add parameters to describe the dependence between the observables and there is no complication coming from the correlation among the characterizing observables. Examples of previous works exploiting conditional independence can be found in Refs. [11,20–24]. One should note that, even when conditional independence may not be exact, it may hold up approximately and, for finite data, can be a valid working assumption. However, the estimators will not be unbiased and will not converge to the class fractions in the limit of infinite data (as seen in *e.g.* Ref. [11]). Explicit attempts to incorporate conditional dependence rely on theoretical knowledge of factorization properties [25,26] or assume weak dependence and rely on the interpolation capabilities of deep neural networks [27–29].

In this work, we are interested in the study of systems with correlated observables where conditional independence does not hold, but we do not have any explicit parametric shape we can use to restrict the functions and the simulators are imperfect. As a first step towards a more complete inference, we assume that the correlations between the variables are well described by the simulations, but not their marginalized distributions. This choice is a compromise between approaches, where the simulator provides the relationship between observables which are used to unfold the data to learn appropriate underlying one-dimensional distributions. Thus, we avoid fully specifying the multi-dimensional structure of the data in terms of a parametric distribution without needing to assume conditional independence. In future works we envisage to also infer the correlation starting from biased priors, while maintaining some structure that allows us to avoid adding too many parameters.

To instantiate the above paradigm we study in this work top-quark related observables, in particular those from hadronically decaying boosted top quark. In general, these objects are relevant in many searches at the LHC, both for precision physics [30–32] and to search for beyond the Standard Model effects [33–35]. Boosted top quarks are observed as fat jets (which we name top jets for convenience) at LHC events, and two relevant characterizing observables of these objects are the number of clusters obtained from re-clustering the constituent particles of the jet with a smaller R [36,37] (N_{clus}) and the jet mass (Mass). The clusters within a fat jet are computed through FastJet [38] and the jet mass is computed using the four-momentum of the constituents of the fat jet. For a top jet, these two observables are correlated and therefore are a suitable set of variables to test the proposed framework. The background for hadronically decaying boosted top quarks consists of QCD jets, which originate from a quark or a gluon and have very different cluster and mass distributions. However, since QCD jets are overwhelmingly more abundant than top jets, they end up being a large background even within a pre-selection of events targeting top jets.

The objective of this work is to consider a selection of events containing top and QCD jets and infer their distributions in $\vec{x} = \{N_{\text{clus}}, \text{Mass}\}$ for each class, even if we start the inference from a biased prior. A partial goal of this set up is to work out the possibility that Monte Carlo simulations may not match the data in what respects to the marginalized $\{N_{\text{clus}}, \text{Mass}\}$ distributions. It is a partial goal because, for simplicity, along the work we assume that the correlations in the form of transfer matrices are well described by simulations. It is worth mentioning that one should expect that simulations at some level do not match the data, since

---

[2]In an abuse of language, but to align with the terminology used in the community, throughout the work we use uncorrelated for independent.

programs such as Pythia [39], Herwig [40], etc., need to use phenomenological models to describe non-perturbative QCD effects, and therefore there is an inherent bias that forbids complete match between simulation and data distributions.

This article is organized as follows. In the next section we introduce the statistical models with and without the assumption of conditional independence and discuss the details of their implementation. In section 3 we perform the inference for both models and different priors; and assess the quality of the obtained results. In section 4 we summarize the results and present our conclusions. Further details on the estimation of the transfer matrix necessary to account for correlations are given in Appendix A.

2 Mixture of multinomial with correlations

Mixture models are a crucial tool for science, since often the data in a problem consist of a collection of independently distributed elements, or data points, which are each sampled from one among a set of possible classes. For many LHC analyses, we are interested in one class in particular, which we label as signal and that we want to characterize in the presence of all other classes, labeled as contributing backgrounds. Along this work we are interested in studying top events in which a boosted top decays hadronically forming a single jet. Therefore, the dataset is a collection of jets and the classes are top (signal) or QCD (background) jets.

Mixture models allow us to construct probabilistic models of the data, by first selecting an appropriate collection of measurable observables $\vec{x}$ and defining the probability of a single jet as

$$p(\vec{x}) = \sum_{k=\{0,1\}} \pi_k \, p(\vec{x}|k) \,, \tag{1}$$

where $k = 0$ corresponds to the QCD background while $k = 1$ corresponds to a top jet, π_k is the probability of sampling a jet from jet class k and $p(\vec{x}|k)$ is the class-dependent probability of a particular observable value $\vec{x}$. Modeling the class-dependent probability distributions corresponds to capturing the relevant physics. In particular, if we consider distributions depending on tensors of parameters $\boldsymbol{\theta}^k$, $p(\vec{x}|\boldsymbol{\theta}^k)$, we can obtain the posterior probability distribution over $\zeta = \{(\pi_k, \boldsymbol{\theta}^k), k = 0, 1\}$ given N measured jets $X = \{\vec{x}_1, \dots, \vec{x}_N\}$

$$p(\zeta|X) = \frac{p(X|\zeta)p(\zeta)}{p(X)} \,. \tag{2}$$

Here $p(X|\zeta) = \prod_{n=1}^{N} p(\vec{x}_n|\zeta)$ for independent jets, $p(\zeta)$ is the prior distribution over parameters and $p(X) = \int d\zeta \, p(X|\zeta)p(\zeta)$ is the evidence or marginal likelihood. The posterior can be estimated via approximate techniques as in Ref. [14]. However, for the probabilistic models considered in this work we obtain samples from the posterior by implementing the model in the probabilistic programming language Stan [41].

The choice of parametric distributions $p(\vec{x}|\boldsymbol{\theta}^k)$ depends on the choice of measured variables. In this work we consider the case where the characterizing observable $\vec{x}$ corresponds to the number of clusters in the jet[3] and the mass of the jet, $\vec{x} = \{N_{\text{clus}}, \text{Mass}\}$. We bin the mass of the jet in 15 bins between $m_{\min} = 150$ GeV and $m_{\max} = 225$ GeV, while the N_{clus} is a naturally discrete variable that in our framework takes 5 possible values. Because the most general distribution of a two-dimensional discrete variable is a multinomial, one might be tempted to describe the class-dependent likelihoods as

$$p(N_{\text{clus}} \in \text{bin}_i, \text{Mass} \in \text{bin}_j|k) \equiv p(i,j|k) = \theta_{ij}^k \,, \tag{3}$$

---

[3]The clustering is performed with the k_T algorithm [42] in FastJet setting $R = 0.3$ and a cut $p_T > 3$ GeV over the constituents of the fat jet.

with

$$\sum_{i=1}^{D_1}\sum_{j=1}^{D_2}\theta_{ij}^k = 1, \tag{4}$$

where D_1 and D_2 are the number of bins for each variable and θ_{ij}^k the parameters of class k. This runs into the problem of attempting to describe a two-dimensional distribution with $D_1 \times D_2 - 1$ parameters $p(i,j)$ with a mixture of K distributions of the same number of parameters $p(i,j|k)$, which is ill-posed and leads to mode collapse. To avoid over-parameterizing the data, we need to include additional structure in our probabilistic model. This additional structure or model bias allows us to extract useful classes. The *arts et métiers* of Bayesian inference is to select the appropriate model for the task at hand, with its compromise between bias and performance. As we show below, a multi-dimensional mixture model is easily addressed under a hypothesis of conditionally independent classes, since the number of parameters in each class gets drastically reduced (in this two dimensional case) from $D_1 \times D_2 - 1$ to $D_1 + D_2 - 2$. However, if conditional independence does not hold in the problem then correlation matters and the problem becomes considerably more difficult. Addressing the latter case is the objective of this work. We describe both cases in the following paragraphs, focusing on the latter, which is the case of interest.

2.1 Conditionally independent model

The simplest model that we can think of is the one that constrains θ_{ij}^k the most by assuming conditional independence. Following Ref. [11], we impose per-class conditional independence

$$p(i,j|k) = p(i|k)p(j|k), \tag{5}$$

or in other words

$$\theta_{ij}^k = \alpha_i^k \beta_j^k, \tag{6}$$

where $\alpha_i^k = p(i|k)$ and $\beta_j^k = p(j|k)$, satisfying $\sum_{i=1}^{D_1}\alpha_i^k = 1$ and $\sum_{j=1}^{D_2}\beta_j^k = 1$ respectively. Here the α_i^k and β_j^k are the parameters of the probabilistic multinomial model assumed for the system. The resulting graphical model is shown in Fig. 1 and the full likelihood can be
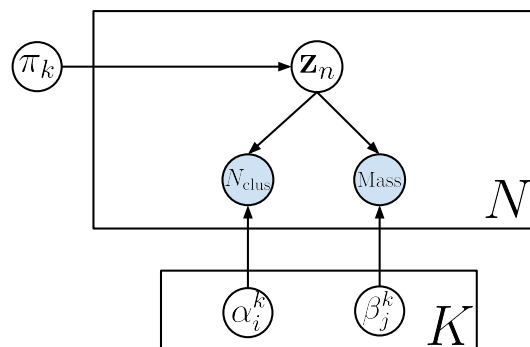
Figure 1: Graphical model for the conditionally independent model. Solid white (blue) circles represent latent (observed) variables. Boxes represent the repeated sampling of variables.

written as

$$p(X|\zeta) = \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \alpha_{i_n}^k \beta_{j_n}^k \right), \qquad (7)$$

which can be usefully rewritten as

$$p(X|\zeta) = \prod_{i=1}^{D_1} \prod_{j=1}^{D_2} \left(\sum_{k=1}^{K} \pi_k \alpha_i^k \beta_j^k \right)^{N_{ij}}, \qquad (8)$$

where N_{ij} is the total number of events populating bin (i, j) and $\sum_{i=1}^{D_1} \sum_{j=1}^{D_2} N_{ij} = N$.

2.2 A simulation-assisted model for the lack of conditional independence

As shown in Ref. [11], if the assumption of conditional independence is incorrect, for a sufficiently large number of events the bias of the model introduced in section 2.1 will shift the posterior from the true values, because the corresponding hypotheses do not hold. Since in this work we consider a large dataset, this bias is present as shown in section 3. To incorporate correlations, we introduce two additional transfer matrices that convolute the truly factorized per-class distributions

$$\theta_{ij}^k = \sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} C_{ij;i'j'}^k \alpha_{i'}^k \beta_{j'}^k, \qquad (9)$$

with $\sum_{i=1}^{D_1} \sum_{j=1}^{D_2} C_{ij;i'j'}^k = 1$ for all i', j', k. This corresponds to writing

$$p(ij|k) = \sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} p(i, j, i', j'|k) = \sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} p(i, j|i', j', k) p(i'|k) p(j'|k), \qquad (10)$$

and identifying $p(i'|k) = \alpha_{i'}^k$, $p(j'|k) = \beta_{j'}^k$ and $C_{ij;i'j'}^k = p(i, j|i', j', k)$.

The introduction of latent, unobservable, variables with conditional independence and transfer matrices that relate them to the observable values of $\vec{x}$ is at this point merely bookkeeping, with the transfer matrices acting as a generalization of covariance matrices that capture more complex relations between the two variables. Thus, if we do not impose any constraints on the transfer matrix the problem again becomes ill-posed. However, this changes if we assume in the following that the transfer matrices are fixed parameters, which we determine from Monte Carlo simulations as detailed in section 2.2.1, and focus on inferring the posterior distribution on $\alpha_{i'}^k$ and $\beta_{j'}^k$. This assumption treats any errors in the description of the transfer matrices as subleading and allows for meaningful inference. Future efforts will be devoted to improving this approximation. The resulting graphical model is depicted in Fig. 2 and the full likelihood can be written as

$$p(X|\zeta) = \prod_{n=1}^{N} \left(\sum_{k=1}^{K} \pi_k \left\{ \sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} C_{i_n j_n; i'j'}^k \alpha_{i'}^k \beta_{j'}^k \right\} \right), \qquad (11)$$

which can be usefully rewritten

$$p(X|\zeta) = \prod_{i=1}^{D_1} \prod_{j=1}^{D_2} \left(\sum_{k=1}^{K} \pi_k \left\{ \sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} C_{ij; i'j'}^k \alpha_{i'}^k \beta_{j'}^k \right\} \right)^{N_{ij}}, \qquad (12)$$

where again N_{ij} is the total number of events populating bin (i, j).
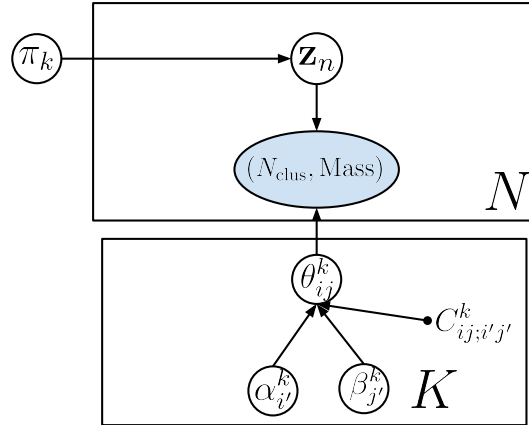
Figure 2: Graphical model for the simulation-assisted model with the addition of the transfer matrices, which are hyperparameters at this level.

2.2.1 Transfer matrices

To obtain posterior samples for each considered model we utilize a Hamiltonian Monte Carlo (HMC) sampler [43] implemented in the Stan [41] statistical software package. HMC is a particular variant of Markov Chain Monte Carlo that makes use of the derivatives of the posterior with respect to the parameters of interest to explore the relatively high-dimensional parameter space and it has been chosen due to its combination of sampling efficiency and speed. We have ensured that the sampled posterior chains pass the standard diagnostics of convergence and autocorrelation.

For the simulation-assisted model, we estimate the transfer matrix on an additional dataset, different from data. In this "training" or "Monte Carlo" dataset, we have to label QCD and top events and estimate the transfer matrix via an expectation-maximization (EM) algorithm that maximizes the per-class likelihood. To do so, we introduce auxiliary latent variables

$$\gamma_{ij;i'j'} = p(i',j'|i,j), \tag{13}$$

that encode the probability of an event populating the bin (i',j') in the sample space defined by the marginal distributions (which we denote as "marginal bin") given that it populates the observable bin (i,j). These auxiliary variables are estimated during the E-step via the identity

$$\gamma_{ij;i'j'} = p(i'j'|ij) = \frac{p(i,j,i',j')}{p(i,j)}, \tag{14}$$

and are needed to estimate the marginal likelihoods $p(i'|k) = \alpha_{i'}^k$ and $p(j'|k) = \beta_{j'}^k$ and the transfer matrix $p(i,j|i',j',k) = C_{ij;i'j'}^k$ during the M-step. To update the parameters of interest during the M-step, we consider the likelihood in marginal space

$$p(i',j'|k) = p(i'|k)p(j'|k) = \alpha_{i'}^k \beta_{j'}^k, \tag{15}$$

and assume that we have $N_{i',j'}^{\mathrm{MC},k}$ events per (i',j') bin given by

$$N_{i'j'}^{\mathrm{MC},k} = \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} \gamma_{ij;i'j'} N_{ij}^{\mathrm{MC},k}, \tag{16}$$

where $N_{ij}^{\mathrm{MC},k}$ are the total number of events belonging to the Monte Carlo simulation of class k that populate bin (i, j). Thus the likelihood over the labeled dataset is

$$\mathcal{L} = \prod_{i'=1}^{D_1} \prod_{j'=1}^{D_2} \left(\alpha_{i'}^k \beta_{j'}^k \right)^{N_{i'j'}^{\mathrm{MC},k}} , \tag{17}$$

which can be maximized with respect to $\alpha_{i'}^k$ and $\beta_{j'}^k$ yielding the usual estimates

$$\alpha_{i'}^k = \frac{\sum_{j'} N_{i'j'}^{\mathrm{MC},k}}{\sum_{i'} \sum_{j'} N_{i'j'}^{\mathrm{MC},k}} = \frac{N_{i'}^{\mathrm{MC},k}}{N^{\mathrm{MC},k}} ,$$

$$\beta_{j'}^k = \frac{\sum_{i'} N_{i'j'}^{\mathrm{MC},k}}{\sum_{i'} \sum_{j'} N_{i'j'}^{\mathrm{MC},k}} = \frac{N_{j'}^{\mathrm{MC},k}}{N^{\mathrm{MC},k}} , \tag{18}$$

where $N^{\mathrm{MC},k}$ is the total number of events for the Monte Carlo sample of class k. The transfer matrix is then estimated as

$$\begin{aligned}
C_{ij;i'j'}^k &= p(i, j | i', j', k) \\
&= \frac{p(i', j' | i, j, k) p(i, j | k)}{p(i', j' | k)} \\
&= \gamma_{ij;i'j'}^k \frac{N_{ij}^{\mathrm{MC},k}}{N_{i'j'}^{\mathrm{MC},k}} .
\end{aligned} \tag{19}$$

The EM algorithm is shown as a pseudo-code in appendix A. After training, the transfer matrix can be used when inferring the conditionally independent distributions in "marginal space" from data. In this sense, we are treating the imperfections of the simulator as concentrated in the marginal likelihoods per-class, with the transfer matrix being correct up to subleading effects.

The code can be found at `GitHub` [44].

2.3 A quantifier for the goodness of the inference

To assess the quality of the inference procedure we introduce the distances D that quantify how far the mean values $(\overline{\pi_k}, \overline{\alpha^k}$ and $\overline{\beta^k})$ of the inferred parameters $(\pi_k, \alpha^k$ and $\beta^k)$ are from their true values

$$D(\pi_k) = |\overline{\pi_k} - \pi_k^{true}| , \tag{20}$$

$$D(\alpha_i^k) = |\overline{\alpha_i^k} - \alpha_{true\,i}^k| , \tag{21}$$

$$D(\beta_i^k) = |\overline{\beta_i^k} - \beta_{true\,i}^k| . \tag{22}$$

We also compute their fluctuations such that the uncertainties of the distances D are given by the standard deviations of the distributions of the inferred parameters.

Additionally, we compute the Kullback-Leibler (KL) divergence between the posterior and the data distributions both for each class and for the complete case where no labels are used. The latter case in particular allows us to get a broader picture which does not rely on knowledge of "true" parameters. In all cases, a smaller divergence signals a better agreement.

Although the KL divergences provide useful metrics, they are not rigorous statistical comparisons between models. Depending on the access to the true parameters, one could use an array of more involved but powerful techniques. If the true parameters are known, posterior credible intervals could be computed and a coverage study could be performed by bootstrapping the dataset. If one does not wish to rely on access to the true parameters, a posterior predictive check could be performed, which also involves additional samplings under the learned posterior model. Thus, more rigorous tests will in general involve additional, expensive simulations. In this work, we find that the cheaper metrics provide enough information to justify the need for correlations and validate the specific modeling proposed, without the necessity of more expensive computations. However, future implementations may and probably will need to rely on these well-established suite of metrics to ensure that the learned probabilistic model is appropriate.

2.4 Dataset

For the purposes of the following sections we resort to simulations in order to benchmark our study. With the aim to facilitate a connection with previous analysis, we employ the standard Top Quark Tagging Reference Dataset [45]. This dataset is generated with `Pythia8` [39] at c.m. energy of 14 TeV and uses the default `Delphes` [46] simulation of the ATLAS detector. Furthermore, it contains 1.2M events for training, 400K events for testing and 400K events available for validation. The reconstruction of fat jets is performed with the anti-k_T algorithm [47] in `FastJet` [38] setting $R = 0.8$ and demanding $p_T \in (550, 650)$ GeV and $|\eta| < 2$. For the present analysis, we select those events in the training sample which satisfy an initial selection cut in the invariant mass of the jets m_j defined by 150 GeV $< m_j <$ 225 GeV. The choice of this particular range is dictated, on the one hand, by the requirement to take into consideration all the details of the top mass peak and, on the other hand, by the necessity to restrict the amount of parameters to be optimal for the successful inference process.

3 Results

We next discuss the results for inferring the true top distributions $\vec{x} = \{N_{\text{clus}}, \text{Mass}\}$ for the frameworks assuming the mixture model with and without the transfer matrices. The first goal, presented in Section 3.1, is to carry out an initial comparative analysis of the two models' performance via studying the inferred posteriors as a function of the number of events in a given test sample and the choice of prior distributions. In Section 3.2 we assess the quality of both models with the metric defined in Section 2.3. As detailed in section 2.2.1, all the posterior samples are obtained via the `Stan` [41] statistical software package.

3.1 Why correlations matter

As presented in Section 2, to include existing correlations between the mass and the number of clusters of the jet we extend the conditionally independent mixture model with transfer matrices derived from simulations. For the selection criteria and the chosen jet mass range defined in Section 2.4 (150-225 GeV), we plot in the right (left) panel of Fig. 3 the top (QCD) jet mass and the number of clusters in the jet[4] as 2D-histograms where the color scale indicates the number of events in a given bin, going to darker tones for larger amounts. Merely a visual inspection suggests the existence of correlations within each class. We quantify their

---

[4]We have checked that discarding the bin with just one cluster in the jet is beneficial for the inference process, therefore the number of clusters will cover the range from two to six throughout the following analysis.
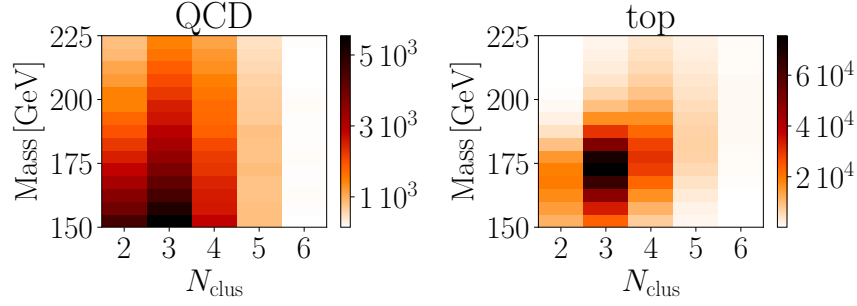
Figure 3: True two-dimensional distributions for QCD and top jets. We observe a non-negligible correlation in the top distribution.

Table 1: Mutual information MI and Pearson correlation coefficient r of the mass and number of clusters per jet for QCD and top jets.

MI(QCD)	MI(top)	MI(top)/MI(QCD)	r(QCD)	r(top)	r(top)/r(QCD)
0.008	0.069	8.6	0.12	0.36	3

strengths by calculating the mutual information (MI) [48] and the Pearson correlation coefficient (r) [49] presenting the results in Table 1. We see there that both variables are in fact correlated, although the correlation is smaller for QCD jets. We also observe how MI is more sensitive than r, pointing towards the importance of non-linear dependencies between variables.

As a first measure of the impact of the inclusion of correlations on the inference performance, we study the inferred probability over the mixture fraction, namely, the probability π_1 of top jets for a given sample as a function of the number of events contained in it. To compute the posterior, we benchmark our models with the following prior distributions for the parameters ζ: i) a uniform distribution for the probability of top jets (π_1) in the sample, ii) marginal QCD and top distributions for the number of clusters, iii) a decreasing linear approximation for the QCD jet mass distribution, and iv) a deformation of the marginal top jet mass distribution. The rationale behind this choice is to study whether a combination of agnostic and biased priors can still lead to an acceptable posterior once the algorithm sees the data. In particular, the deformation of the marginal top jet mass distribution consists in cutting off the top mass peak. This puts under stress a major discriminating feature between top and QCD jets and allows to test the robustness of the inference process.

The prior values for the fractions π_k are generated with a uniform Dirichlet distribution

$$\pi_{\mathbf{0,1}} \sim \text{Dirichlet}(1,1). \tag{23}$$

The prior values for the parameters $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$ are obtained by sampling other Dirichlet distributions

$$\boldsymbol{\alpha}^k \sim \text{Dirichlet}(\boldsymbol{\eta}_\alpha^k), \tag{24}$$

$$\boldsymbol{\beta}^k \sim \text{Dirichlet}(\boldsymbol{\eta}_\beta^k), \tag{25}$$

where the vectors $\boldsymbol{\eta}_{\alpha,\beta}^k$ are given by

$$\boldsymbol{\eta}_{\alpha,\beta}^k = \boldsymbol{p}_{\alpha,\beta}^k \, \Sigma, \tag{26}$$

with $\boldsymbol{p}_\alpha^k$ and $\boldsymbol{p}_\beta^k$ standing for the input parameters of number of clusters and mass probability distributions, respectively, and Σ is a common normalization. As the parameters $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$ are generated by the Dirichlet distribution, their expectation values are given by the input parameters of the multinomial distributions $\boldsymbol{p}_{\alpha,\beta}^k$

$$\mathrm{E}[\boldsymbol{\alpha}^k] = \boldsymbol{p}_\alpha^k, \tag{27}$$

$$\mathrm{E}[\boldsymbol{\beta}^k] = \boldsymbol{p}_\beta^k, \tag{28}$$

and the variance results

$$\mathrm{Var}[\boldsymbol{\alpha}^k] = \frac{\boldsymbol{p}_\alpha^k(1-\boldsymbol{p}_\alpha^k)}{\Sigma+1}, \tag{29}$$

$$\mathrm{Var}[\boldsymbol{\beta}^k] = \frac{\boldsymbol{p}_\beta^k(1-\boldsymbol{p}_\beta^k)}{\Sigma+1}. \tag{30}$$

The input parameters of the multinomial distributions $\boldsymbol{p}_{\alpha,\beta}^k$ for the model with and without correlations are taken to be the parameters of the marginal distributions for the case of the number of clusters for both top and QCD, the marginal distribution with a deformation of the peak of the top jet mass distribution, and a decreasing linear approximation for the QCD mass distribution. Notice that the normalization factor Σ controls the variation of the prior values for the parameters $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$. Large Σ implies $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$ concentrated on their mean $\boldsymbol{p}_{\alpha,\beta}^k$, and vice versa. Although the choice of Σ is subjective, it is not arbitrary: the prior needs to encode the trust in the Monte Carlo simulations while providing an adequate range of allowed discrepancies. Because in this work we have access to the true parameter values, we can validate the choice of Σ simply by observing that the resulting posterior can be likelihood-driven and centered around these true values for large enough number of events. When applying this method to data where the true values are unknown, a data-driven prior validation should be implemented. One possibility is to perform prior predictive checks [50] to ensure that the chosen Σ produces realistic datasets. Another possibility is to build a hierarchical model where Σ is promoted to a random variable and marginalized over (see Ref. [51], Chapter 5), reducing the prior dependence at the expense of increased model uncertainty. Since the goal of this work is to validate the introduction of correlations, we leave a more detailed prior exploration for future work.

Since the transfer matrix that encodes correlations is obtained from Monte Carlo simulations, we seek to reduce the possibility of overfitting by computing the transfer matrix on a different sample from the one used for posterior inference. We also use this different sample to compute two additional sets of distributions: the distributions in marginal space $\{\alpha_{i'}^k, \beta_{j'}^k\}$ that are obtained in addition to the transfer matrix, and the marginal distributions. The latter are obtained by marginalizing over one of the variables ($\{\alpha_i^{k,\mathrm{marg.}} = \sum_{j=1}^{D_2} p(ij|k), \beta_j^{k,\mathrm{marg.}} = \sum_{i=1}^{D_1} p(ij|k)\}$) and match the distributions in marginal space only when there is no correlation (or in other words when the transfer matrix is the identity).

As detailed in previous paragraphs, the marginal distributions are used to define the prior distributions. For the model with no correlations, the marginal distributions are also the target or "true" distributions against which we compare both the prior and the posterior distributions. For the model with correlations, the "true" distributions consist of the distributions in marginal space obtained along with the transfer matrix.

Furthermore, in order to study the degree of dependence of the inferred parameters on the priors, we evaluate the inference outputs for two different allowed variations ($\Sigma = 100, 1400$) of the prior distributions around the corresponding mean values, which we term "loose" and
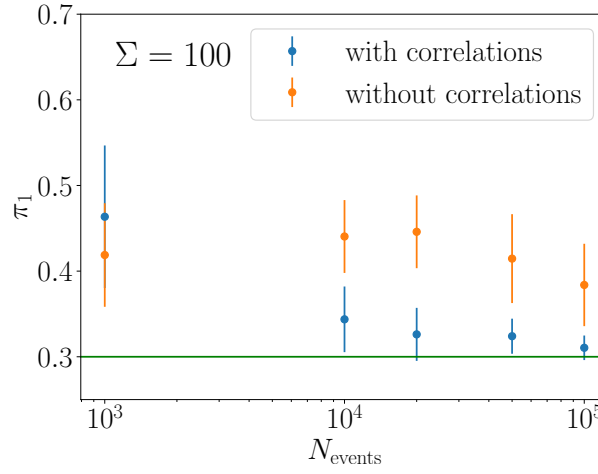
Figure 4: Inferred probability of top jets π_1 as a function of the number of events in the test samples for $\Sigma = 100$. The blue(orange) points stand for this probability for the model with(without) correlations.

"tight" accounting for the restrictions that they impose on the posterior. In addition, we aim at studying the inference performance in unbalanced samples of top and QCD jets. Therefore, we set the fraction of top jets $\pi_1 = 0.3$, independently of the number of events in the sample. Since we observed, as in principle expected, that a larger fraction turned out to improve the distinction between classes, we consider this relatively low top jets fraction as a conservative value to determine the goodness of the inference procedure. Much lower fractions will negatively affect the inference performance, although we have not explicitly explored the lowest sensitivity of the tagger since we expect fairly enriched samples to be available from $t\bar{t}$ production.

3.1.1 Loose priors

We show in Fig. 4 the inferred probability of top jets as a function of the number of events in the test samples for $\Sigma = 100$. The blue (orange) points represent this probability for the mixture model with (without) correlations. The vertical error bars stand for the standard deviation over a set of 1500 samples of the posterior, and the horizontal green line corresponds to the true value of the probability of top jets ($\pi_1 = 0.3$). We see that for 10^3 events both the uncorrelated and correlated models give compatible probabilities within the uncertainties but both lie above the true value. As the number of events increases, the orange points show large fluctuations and they still keep away from $\pi_1 = 0.3$. Instead, the mean values of the blue points display a monotonously decreasing behaviour towards the true value. From $2 \cdot 10^4$ events these points are compatible within $\sim 1\sigma$ uncertainty with $\pi_1 = 0.3$. Moreover, the size of the blue error bars decrease as a function of the number of events. This tendency is not verified by the orange error bars which remain roughly of the same size.

It is worth stressing that the priors used in the inference problem do not match the true distributions in the data, and therefore the inferential process corrects the priors and approaches the trues as it sees the data.

These features suggest that, as expected, the model with correlations is more suitable to reproduce the data. As the amount of events increase, the intra-class correlations have a greater and greater impact on the inference process and the model without correlations is not capable to correctly reproduce the unlabelled data. Even so, the probability of top jets
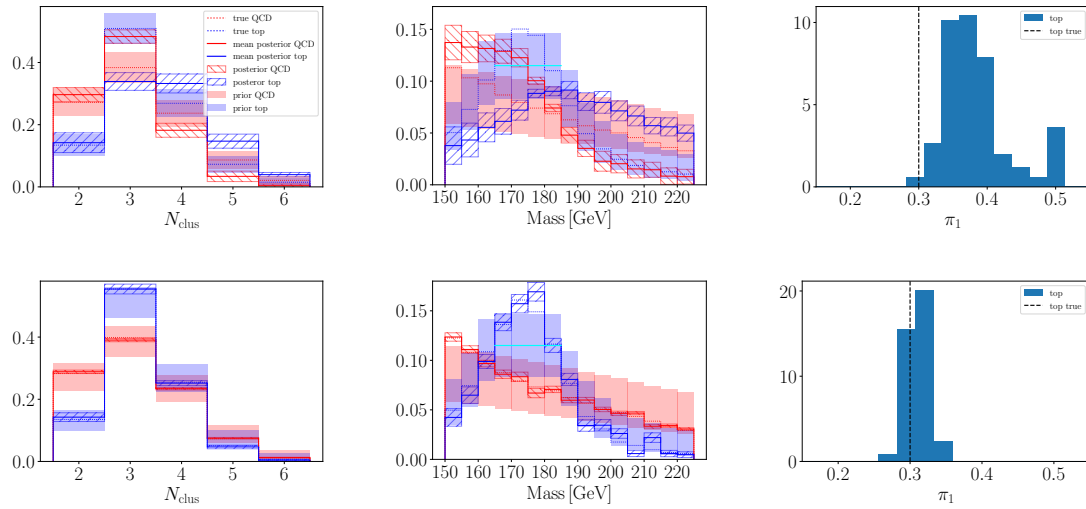
Figure 5: Number of clusters, mass and the probability of top jets distributions for the model with (bottom row) and without (upper row) correlations with $\Sigma = 100$. See text for the details.

is just one parameter in the mixture models and we are equally interested in retrieving the correct distributions for the posteriors corresponding to the number of clusters and the mass of the jet. As we show in the next paragraphs, these other distributions also exhibit a good agreement with the data for the model which includes correlations and fail in the case of the model without correlations.

For the model with(without) correlations, we then show in the bottom(upper) row, left and middle panels, of Fig. 5 the true values (dotted lines) together with the 1σ prior (shaded regions) and the marginalized posterior (solid lines for the mean values and hatched regions for the standard deviations) of the number of clusters and the jet mass distributions. This figure is made for a sample of 10^5 events in which case the posterior probability of top jets for the model with correlations reaches a closer value to the truth when compared with smaller samples. In the right panels we display the probability of top jets with the true value (vertical dashed line) and the posterior distribution in blue for both models (the prior is uniform in this case and then not shown). In agreement with Fig. 4, we see here that the probability of top jets misses the true value for the model with no correlations, whereas it matches the truth for the model with correlations.

The plot for the number of clusters (left panel) shows that the model without correlations (upper row) is deficient in reproducing the main feature of these distributions: a higher proportion of events with three clusters for top jets associated to the decay products of the top (a b quark and a W with a subsequent decay into two quarks). We see that the mean values of the posteriors for top (solid blue lines) and QCD (solid red lines) jets appear inverted in that bin and there are also relatively smaller inversions in the last two bins. In contrast with this, the model with correlations reproduces well all the bins keeping particularly the correct proportion in the bin with three clusters (left panel, bottom row).

Moreover, we observe in the middle panel that the mean values of the top (solid blue lines) and QCD (solid red lines) jet mass posterior distributions closely follow the true values for the model with correlations (bottom row) but they fail in the case of the model without correlations (upper row) diminishing the top mass peak and inverting the distributions in the higher bins region. The most remarkable point here is that for the model with correlations the top mass posterior almost recovers the peak shape in spite of the deformation of the peak introduced by

the prior (the mean value of the prior in the four bins surrounding the top mass is highlighted with a cyan line). Finally, the dispersion associated to the posterior of the probability of top jets, as well as the spread (hatched regions) corresponding to the posteriors of the number of clusters and the jet mass, are smaller for the model with correlations than for the model without correlations. This is to be expected, since the model without correlations does not correspond to the data and therefore can find many unphysical combinations of parameters that explain the data fairly well without a huge drop-off in the likelihood, *i.e.* there is no privileged true parameter choice from which any parameter variations result in a tremendous likelihood drop. This enhancement in good parameter samples increases the uncertainties over parameter space.

The same abundance of unphysical parameter choices for the model without correlations causes the inference process to find two disconnected regions in parameter space which are related by an inversion in the class labels. That is, some posterior samples show a "label-switching" where the two inferred classes are inverted, disregarding the prior distribution. To account for this unphysical phenomenon, the reported posterior samples are the result of a post-processing with a criterion based on prior knowledge of the problem: every time we get $\pi_1 > 0.5$, we redefine $\pi_1 \to 1 - \pi_1$ and re-label the learned per-class distributions, with the final $\pi_1 \in [0, 0.5]$. This kind of workaround to avoid label-switching is a common practice for Mixture Models [52].

In conclusion, when the mixture model includes correlations the posteriors not only better approach the truth but also end up being more precise due to a lower variance of the posterior.[5]

3.1.2 Tight priors

We analyze now the effects of considering more restricted priors in the inference process. We study the case in which the multiplicative factor in the Dirichlet arguments in Eq. (26) is large, namely $\Sigma = 1400$. This large factor yields that the draws from the Dirichlet are more concentrated around its mean. In Fig. 6 we show the inferred probability of top jets as a function of the number of events in the test samples. As in Fig. 4 the blue(orange) points depict the probability of top jets for the mixture model with(without) correlations. The vertical error bars denote the standard deviation over a set of 1500 values of the posterior, and the horizontal green line corresponds to the true value of the probability of top jets ($\pi_1 = 0.3$).
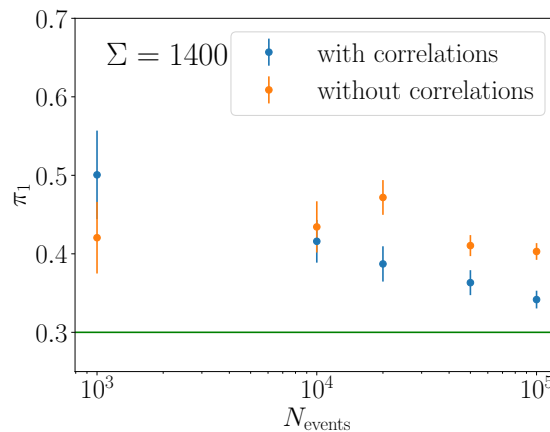
Figure 6: The inferred probability of top jets π_1 as a function of the number of events in the test samples for $\Sigma = 1400$. The blue(orange) points stand for this probability for the model with(without) correlations.

---

[5]It may be interesting to notice that this conclusion is along the lines of what occurs in Ref. [53], where the goal of reducing the variance is linked to the best modeling of the systematics in the data.
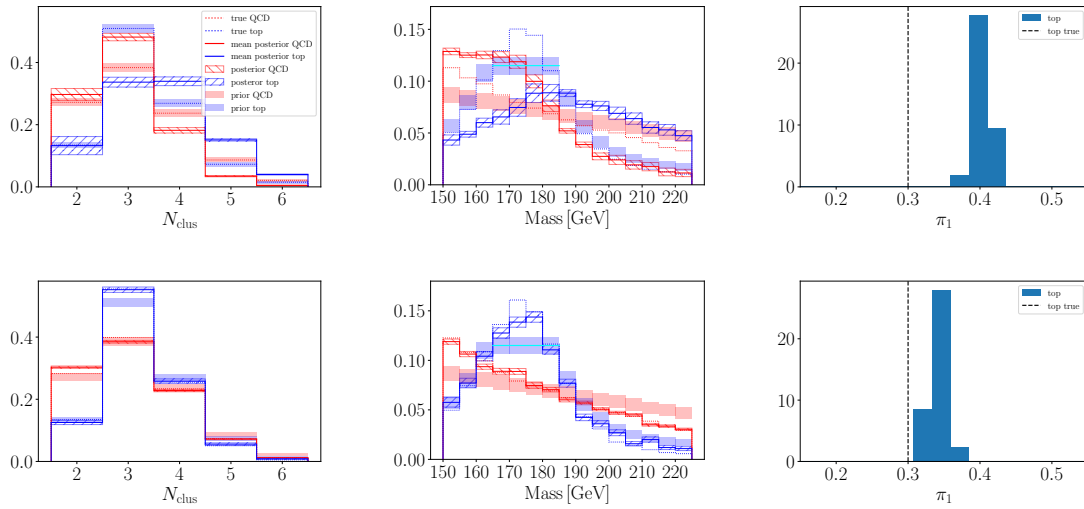
Figure 7: Number of clusters, mass and the probability of top jets distributions for the model with (bottom row) and without (upper row) correlations with $\Sigma = 1400$. See text for the details.

As in the case of $\Sigma = 100$, we observe that for 10^3 events the uncorrelated and correlated models give compatible probabilities within the uncertainties but both lie above the true value. In contrast with the loose prior, however, the mean value for the model without correlations is closer to the true value. In spite of this, from 10^4 events on, the mean values of the blue points show again a decreasing behaviour towards the true value whereas the orange points exhibit a less pronounced slope and slightly larger fluctuations. Even considering this decreasing behaviour, the blue points are not compatible within the uncertainty with $\pi_1 = 0.3$, contrary to what we found for $\Sigma = 100$, and this still occurs for 10^5 events. The uncertainties also diminish as the number of events increases and the model with correlations ends up being relatively precise but not too accurate. Given that we expect more dependence on the priors compared to the loose prior case even for a relatively large number of events, it is not altogether surprising that the model with correlations is close but misses the truth because the transfer matrix is approximate and the allowed variation of the priors is more restrictive than in the case of $\Sigma = 100$. Despite that, we will see in the following that the posterior distributions corresponding to the number of clusters and the mass of the jet agree with data visibly better for the model which includes correlations than for model without correlations.

Fig. 7 corresponds to $\Sigma = 1400$ and it is similar to Fig. 5. For the model with(without) correlations, in the bottom(upper) row, left and middle panels, we show the true values (dotted lines) together with the prior (shaded regions) and the marginalized posterior (solid lines for the mean values and hatched regions for the standard deviations) of the number of clusters and the jet mass distributions. Likewise with loose priors, Section 3.1.1, this figure is built with a sample of 10^5 events in which case the value of the posterior probability of top jets for the model with correlations is closer to the truth when compared to smaller samples. In the right panel we display the probability of top jets with the true value (vertical dashed line) and the posterior distribution in blue for both models (the prior is again uniform in this case and not shown in the figure). In accordance with Fig. 6, we observe that the probability of top jets misses the true value for the model without correlations, whereas it approaches the truth for the model with correlations.

As was the case for the loose priors, for the number of clusters (left panel) the model without correlations (upper row) fails to recreate the bin with three clusters in the jet. The posteriors for top (solid blue lines) and QCD (solid red lines) jets appear inverted there and with smaller inversions in the last two bins. The model with correlations does reproduce closely all the bins (left panel, bottom row). Additionally, the behavior of the mean values of the top (solid blue lines) and QCD (solid red lines) jet mass posterior distributions is similar to the $\Sigma = 100$ case. They nearly approximate the true values for the model with correlations (bottom row) but they are unsuccessful in the case of the model without correlations (upper row) where the top mass peak is reduced and the distributions in the higher bins region are inverted. In contrast to $\Sigma = 100$, however, the dispersion associated to the posterior of the probability of top jets, as well as the spread (hatched regions) corresponding to the posteriors of the number of clusters and the jet mass, is comparable for the models with and without correlations. The tighter prior is regularizing the inference so that many of the unphysical parameter choices become more disfavored and thus do not contribute as significantly to the uncertainty. For the same reason, we observe an absence of switching between the probability of top and QCD jets for $\Sigma = 1400$. Here again, for the model with correlations, the remarkable signature is that the top mass posterior manages to approach the peak shape even if the model is much more constrained by the large deformation of the peak (the mean value of the prior in the four bins surrounding the top mass is highlighted again with a cyan line).

In summary, the use of tight priors has shown, as expected, that the model without correlations keeps failing the true distributions. More importantly, for the model with correlations, it has been helpful to test the robustness of the inference process when biased priors are kept away from the truth.

3.1.3   Prior comparison

As a direct comparison between the loose ($\Sigma = 100$) and tight ($\Sigma = 1400$) priors for the model with correlations, we show in Fig. 8 the inferred probability of top jets as a function of events in a given sample. We also add a third possibility considering an intermediate case with $\Sigma = 500$ to browse the stability of different allowed variations of the priors. For all cases, we observe how if the number of events is too small the prior dominates and biases the inferred probability. The number of events needed to overcome the prior bias depends on the choice of Σ.
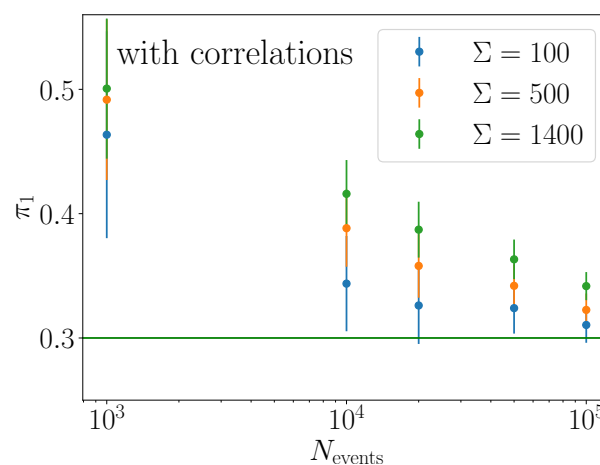
Figure 8: The inferred probability of top jets π_1 in the model with correlations as a function of the number of events in the test samples for different allowed variations of the priors.
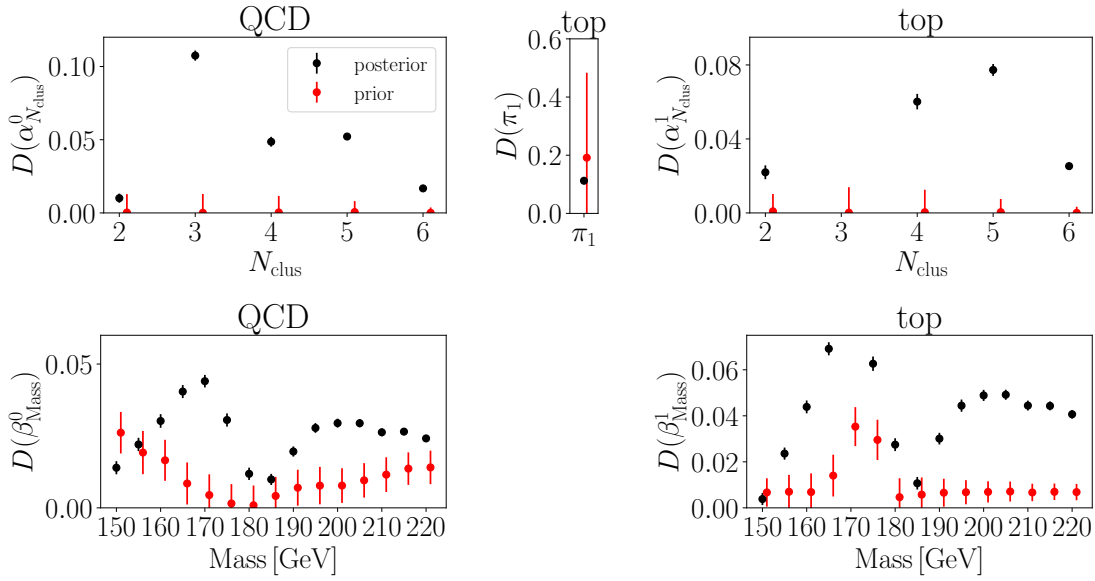
Figure 9: Distances of the mean values of the inferred (black circles) and prior (red circles) parameters π_k, $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$ from their true values for the model without correlation and $\Sigma = 1400$.

In the previous sections we came to the conclusion that the model with correlations adequately captures the data distributions for both the loose and tight priors. However, one should note that the tight prior still biases all distributions even for 10^5 events while the loose prior systematically allows the posterior probabilities to better fit the data once the amount of events is enough to dominate the inference process. The prior with $\Sigma = 500$ displays a smooth progression between the other two values.

3.2 Model quality evaluation

This section aims at evaluating the quality of both uncorrelated and correlated models through the metric defined in Eqs. 20-22. We compute the distance (D) between the values of the inferred parameters (the probability of top jets in the sample, the per-class probability of a certain number of clusters in the jet and the per-class probability of a given bin of jet mass) and their true values. For the model without correlations, we consider a single prior normalization, while for the model with correlations we study the impact of changing the normalization and consider both loose and tight priors as defined in section 3.1. In all the cases we quantify the performance of the models employing samples of 10^5 events.

We show the performance of the model without correlations and tight priors ($\Sigma = 1400$) in Fig. 9, where the black points stand for the distance between the mean of the posterior parameters and their true values. The left(right) top panel depicts the distance for the QCD(top) number of clusters. Along with the number of clusters of top jets, we display the distance for the probability of top jets in the sample. The left(right) bottom panel contains the distance corresponding to the QCD(top) mass distribution. Additionally, the red points represent the distances between the mean of the prior parameters and their true values. Even considering standard deviations, this figure shows for all the variables that the distances for the posterior parameters are consistently larger than the corresponding ones for the prior parameters.[6] This

---

[6]This is not true for all bins, with a couple of bins in the mass distributions, nor for the probability of top jets showing better agreement at the posterior level. In the latter case, the spread in the prior values of this probability is relatively large due to the fact that they are generated with a uniform distribution.
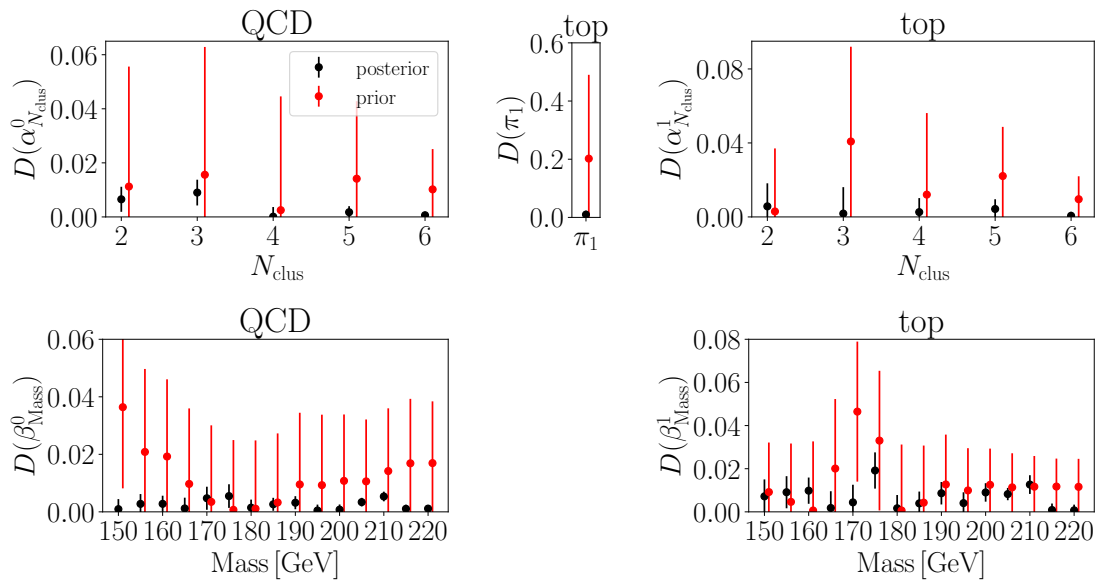
Figure 10: Distances of the mean values of the inferred (black circles) and prior (red circles) parameters π_k, $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$ from their true values for the model with correlation and $\Sigma = 100$.

is in complete agreement with the results of the previous section where we conclude that the lack of correlations in the model prevents the inference from obtaining posterior distributions compatible with the truth. This conclusion remains unchanged when $\Sigma = 100$ is employed in the inference, the only difference being an expected larger dispersion of the distances for the prior parameters and also a significant fluctuation in the distances of the posterior parameters due to the increased abundance of unphysical solutions for the uncorrelated model.

Figs. 10 and 11 correspond to the model with correlations for loose ($\Sigma = 100$) and tight ($\Sigma = 1400$) priors, respectively. Like Fig. 9, they depict in black(red) points the distance between the mean of the posterior(prior) parameters and their true values. As a general pattern, we see in Fig. 10 that the distances for the mean values of the posterior parameters are smaller than the corresponding ones for the prior parameters with the exception of some few bins in the cluster and mass distributions. Additionally, the black points have in all cases much smaller fluctuations than the red points. The metric quantifies in this way not only the improvement of the posteriors with respect to the priors but also the performance observed in Section 3.1.1 for the model with correlations where we find that the inferred parameters approach the truth with relatively small uncertainties. Additionally, by comparing Figs. 10 and 11 with Fig. 9 we also confirm the better performance of the model with correlations.[7]

Fig. 11 shows that in most of the bins the mean values of the posteriors (black points) are closer to the truth than the ones of the priors (red points). Besides, even if the uncertainties in the priors are reduced with respect to $\Sigma = 100$, they are still larger than the uncertainties in the posteriors. In particular, we see in the right panel (bottom row) that the distances of the posteriors in the bins corresponding to the peak of the top jet mass distribution are considerably closer than the priors in terms of significance illustrating how the inference process corrects the bias introduced by the priors. The posterior parameters result to be relatively precise but, as we have already discussed in Section 3.1.2, they still miss the truth, and in some bins at more than 2σ, because the transfer matrix is approximate and the allowed exploration of parameter space is strongly constrained for tight priors.

---

[7]Notice that the uncertainties of the red points seem to be smaller in Fig. 9 than in Fig. 11 but this is just an optical effect due to the use of different scales in the vertical axis.
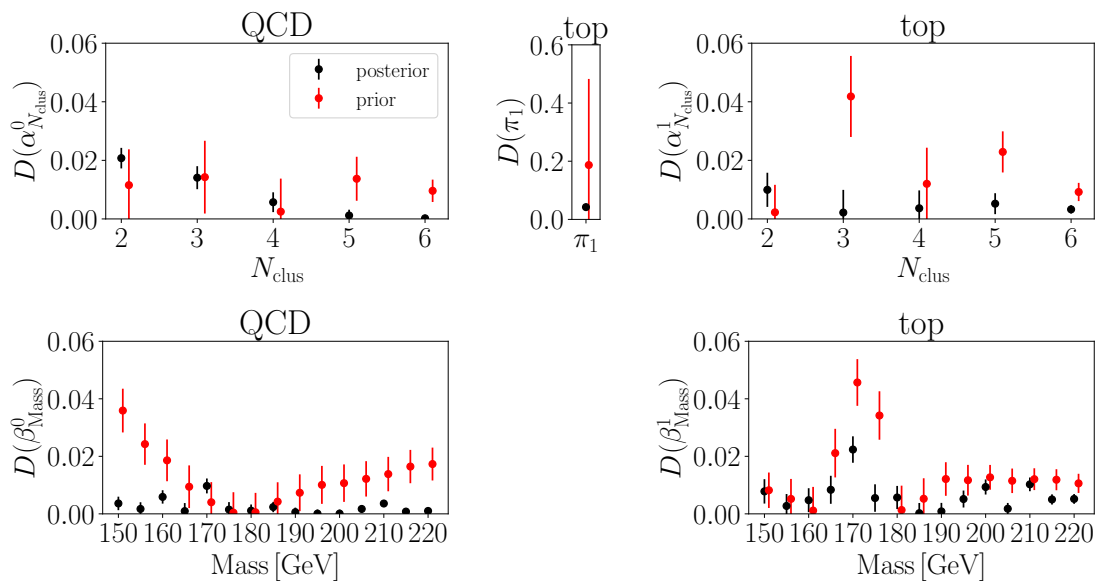
Figure 11: Distances of the mean values of the inferred (black circles) and prior (red circles) parameters π_k, $\boldsymbol{\alpha}^k$ and $\boldsymbol{\beta}^k$ from their true values for the model with correlation and $\Sigma = 1400$.

A remarkable example where the tight restriction on the priors leads the posterior to fail to hit the true value occurs in the fifth bins of the jet mass distributions (that is, for the mass in the interval $[170, 175]$ GeV). We see that the maximal separation between the prior and the truth in the top jet mass distribution takes place precisely in that bin. Since that separation is quite large in terms of the allowed variation of the prior, the way that the inference manages to deal with this difficulty is to compensate the insufficient growth of the top jet mass posterior by increasing the posterior of the QCD jet mass. This is in complete correspondence with Fig. 7 and exemplifies, in addition, the correlations of the distances D between classes. It is worth observing that the mentioned compensation is rather milder in the case of $\Sigma = 100$ as is shown both in Fig. 5 and Fig. 10. Since we see from Figs. 5 and 7 that the uncertainties in the posteriors are generally larger for $\Sigma = 100$ than for $\Sigma = 1400$, this result suggests that it might be preferable to work with looser priors at the expense of having larger uncertainties in the posteriors, or in the other way around, posteriors corresponding to tighter priors may be relatively more precise but in values more distant from the truth.

Finally, we compare the outcome of the different frameworks in the full two-dimensional distributions on the characterizing observables $\vec{x} = \{N_{\text{clus}}, \text{Mass}\}$. We show in Fig. 12 the Maximum a Posteriori (MAP) for the posterior $p(\vec{x})$ of the model with correlations, for different cases. To make a quantitative comparison of the similarity between the posteriors in different frameworks, the priors, and the true PDFs, it is suitable to use the Kullback–Leibler (KL) divergence [48], which is a measure of the *statistical difference* between a model PDF and the true PDF. We summarize all the relevant KL-divergence distances to their corresponding true PDF in Table 2. In this table we compare KL-divergence distances between the prior distributions and true distributions (KL(prior,true) in the fourth column) to the distances between posterior and the true distributions (KL(posterior,true) in the fifth column) for the model with and without correlations ("yes" and "no" in the third column of the table stand for the former and the latter, respectively). As before, we consider two scenarios for prior spread here: loose ($\Sigma = 100$) and tight ($\Sigma = 1400$). We calculate the distances separately for three categories of events: QCD events, top events and both QCD and top events.
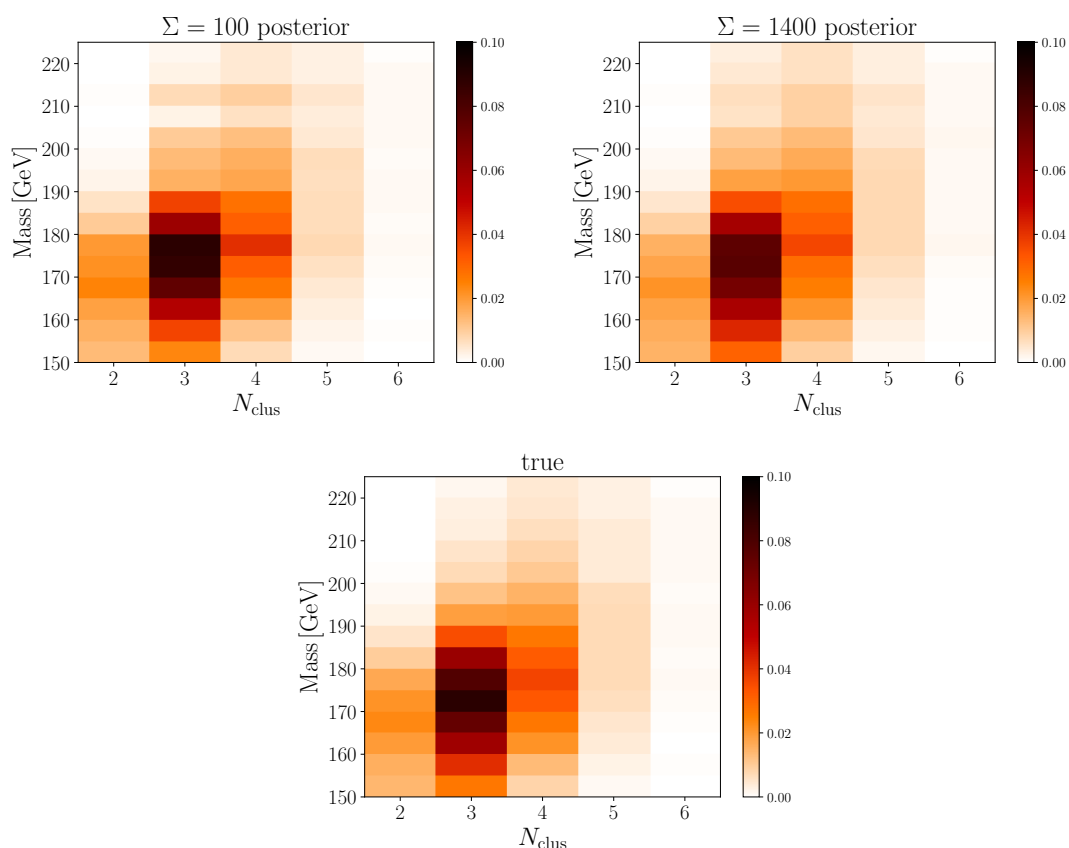
Figure 12: MAP for the posterior $p(\vec{x})$ of the model with correlations for the case of loose priors (top left panel), tight priors (top right panel) compared to truth PDFs (bottom panel).

It can be seen from Table 2 that: 1) the KL-divergence decreases if the correlations are included in the model compared to the model without correlations, 2) KL(posterior,true) < KL(prior,true) in the case of the model with correlations is taken into account, 3) KL-divergence is slightly smaller for loose priors compared to tight priors in the case of the model with correlations taken into account(which is in accordance with our earlier results), 4) KL-divergence is smaller for QCD events compared to top events (because the QCD priors are closer to the true values), 5) as expected, KL-divergence is usually smaller for the dataset that includes both QCD and top events. These results summarize the performance of our proposed objective of determining the underlying distributions in top events with correlation among the characterizing observables.

4 Discussion and conclusions

In this manuscript, we have proposed a completely Bayesian strategy to study mixture models with correlated discrete variables, where the correlation can be learned from simulations, and we have applied said strategy to learn the density of QCD and top jets from measured jets at the LHC. We have defined a set of interesting variables which are either discrete or discretized and learned the relevant class-dependent multinomial distributions. We have shown how the quality of the model depends on the choice of prior and, more importantly, on the ability of the model to incorporate multidimensional information, which we do by estimating the per-class dependence with existing simulators.

Table 2: Posterior to true and prior to true KL-divergence distances in the different frameworks. We have boldfaced the results corresponding to learning each class true distributions when exploiting the correlations, which are among the main motivations for this work.

Class	Prior spread	Correlated model	KL(prior, true)	KL(posterior, true)
QCD	tight	yes	0.022	**0.00175**
Top	tight	yes	0.038	**0.011**
both	tight	yes	0.022	**0.0108**
QCD	loose	yes	0.022	**0.00174**
Top	loose	yes	0.039	**0.0096**
both	loose	yes	0.022	**0.0105**
QCD	tight	no	0.025	0.13
Top	tight	no	0.11	0.45
both	tight	no	0.039	0.016
QCD	loose	no	0.024	0.16
Top	loose	no	0.12	0.48
both	loose	no	0.038	0.01

We have shown how, although the model without correlations is a good model of the full data, the specific classes are better modeled when correlations are taken into account. We have proposed specific metrics that quantify the improvement of the posterior with respect to the prior and allow us to compare different models. In all cases, we have observed that the top distribution suffers from the largest impact of correlations and thus is the most improved when the model incorporates them. We speculate that the methodology introduced to mix data-driven one-dimensional estimators and simulator-driven correlations is useful beyond this specific example or application. In particular, physics analysis where the backgrounds are data-driven could use this approach to correct the inference on the data in the signal region with simulations. The usefulness of this will depend on how trustworthy the multidimensional data-driven template is, which needs to be assessed in a case-by-case basis. However, we highlight the fact that because correlations are defined implicitly through simulators, we can incorporate highly non-trivial dependencies.

In the future, this model could be extended by incorporating more classes (such as W jets), and also by attempting to infer the correlations from the data itself. Of course, there is no free lunch and incorporating data-driven correlations should be compensated by additional model nuances.

Acknowledgments

Funding information   This work was partially supported by CONICET. M.S. acknowledges the support in part by the DOE grant de-sc0011784 and NSF OAC-2103889, OAC-2411215, and OAC-2417682.

A   Expectation-maximization algorithm

In this appendix, we present in a schematic way the technical details of the expectation-maximization algorithm used to calculate the transfer matrices previously introduced in sub-section 2.2.1. The reader is referred to this section for a brief introduction to the EM algorithm and the explanation of the notation.

Algorithm 1 The EM algorithm for per-class transfer matrix estimation.

$N_{ij}^{\text{MC},k} \leftarrow$ Number of events in Monte Carlo sample for class k in bin (i,j).

$N^{\text{MC},k} \leftarrow \sum_i \sum_j N_{ij}^{\text{MC},k}$

Initialize $C_{ij;i'j'}^{k,0}$

Initialize $\alpha_{i'}^{k,0}$

Initialize $\beta_{j'}^{k,0}$

$\mathcal{L}_0 \leftarrow \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} N_{ij}^{\text{MC},k} \ln\left(\sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} C_{ij;i'j'}^{k,0} \alpha_{i'}^{k,0} \beta_{j'}^{k,0}\right)$

for t=1,...,Max Epochs **do**

    E-Step

    $\gamma_{ij;i'j'}^{k,t} = p_t(i'j'|ij) = \dfrac{p_t(ij,i'j')}{p_t(ij)} \leftarrow \dfrac{C_{ij;i'j'}^{k,t-1} \alpha_{i'}^{k,t-1} \beta_{j'}^{k,t-1}}{\sum_{i'} \sum_{j'} C_{ij;i'j'}^{k,t-1} \alpha_{i'}^{k,t-1} \beta_{j'}^{k,t-1}}$

                                     ▷ Probability of event in marginal bin (i',j') if measured in bin (i,j).

    M-Step

    $N_{i'j'}^{\text{MC},k,t} \leftarrow \sum_i \sum_j N_{ij}^{\text{MC},k} \gamma_{ij;i'j'}^{k,t}$

                                     ▷ Estimated number of events in marginal bin (i',j').

    $N_{i'}^{\text{MC},k,t} \leftarrow \sum_{j'} N_{i'j'}^{\text{MC},k,t}$

    $N_{j'}^{\text{MC},k,t} \leftarrow \sum_{i'} N_{i'j'}^{\text{MC},k,t}$

    Update α

    $\alpha_{i'}^{k,t} = p_t(i') \leftarrow \dfrac{N_{i'}^{\text{MC},k,t}}{N^{\text{MC},k}}$

    Update β

    $\beta_{j'}^{k,t} = p_t(j') \leftarrow \dfrac{N_{j'}^{\text{MC},k,t}}{N^{\text{MC},k}}$

    Update C

    $C_{ij;i'j'}^{k,t} = p_t(ij|i'j') = \dfrac{p_t(ij,i'j')}{p_t(i'j')} = p_t(i'j'|i,j)\dfrac{p_t(i,j)}{p_t(i'j')} \leftarrow \gamma_{ij;i'j'}^{k,t} \dfrac{N_{ij}^{\text{MC},k}}{N_{i'j'}^{\text{MC},k,t}}$

    Update likelihood value

    $\mathcal{L}_t \leftarrow \sum_{i=1}^{D_1} \sum_{j=1}^{D_2} N_{ij}^{\text{MC},k} \ln\left(\sum_{i'=1}^{D_1} \sum_{j'=1}^{D_2} C_{ij;i'j'}^{k,t} \alpha_{i'}^{k,t} \beta_{j'}^{k,t}\right)$

    if $\mathcal{L}_t \leq \mathcal{L}_{t-1}$ **then**

        Break for loop

    end if

end for

Return $C_{ij;i'j'}^{k,t}$

References

[1] A. Gelman et al., *Bayesian workflow*, (arXiv preprint) doi:10.48550/arXiv.2011.01808.

[2] K. Cranmer, *Practical statistics for the LHC*, in *2011 European school of high-energy physics*, CERN, Geneva, Switzerland, ISBN 9789290834014, doi:10.5170/CERN-2014-003 [preprint doi:10.48550/arXiv.1503.07622].

[3] B. M. Dillon, D. A. Faroughy and J. F. Kamenik, *Uncovering latent jet substructure*, Phys. Rev. D **100**, 056002 (2019), doi:10.1103/PhysRevD.100.056002 [preprint doi:10.48550/arXiv.1904.04200].

[4] B. M. Dillon, D. A. Faroughy, J. F. Kamenik and M. Szewc, *Learning the latent structure of collider events*, J. High Energy Phys. **10**, 206 (2020), doi:10.1007/JHEP10(2020)206 [preprint doi:10.48550/arXiv.2005.12319].

[5] B. M. Dillon, D. A. Faroughy, J. F. Kamenik and M. Szewc, *Learning latent jet structure*, Symmetry **13**, 1167 (2021), doi:10.3390/sym13071167.

[6] I. Brivio, S. Bruggisser, N. Elmer, E. Geoffray, M. Luchmann and T. Plehn, *To profile or to marginalize – A SMEFT case study*, SciPost Phys. **16**, 035 (2024), doi:10.21468/SciPostPhys.16.1.035 [preprint doi:10.48550/arXiv.2208.08454].

[7] D. Yallup and W. Handley, *Hunting for bumps in the margins*, J. Instrum. **18**, P05014 (2023), doi:10.1088/1748-0221/18/05/P05014 [preprint doi:10.48550/arXiv.2211.10391].

[8] A. Fowlie, *The Bayes factor surface for searches for new physics*, Eur. Phys. J. C **84**, 426 (2024), doi:10.1140/epjc/s10052-024-12792-9 [preprint doi:10.48550/arXiv.2401.11710].

[9] J. Albert, C. Balázs, A. Fowlie, W. Handley, N. Hunt-Smith, R. Ruiz de Austri and M. White, *A comparison of Bayesian sampling algorithms for high-dimensional particle physics and cosmology applications*, Comput. Phys. Commun. **315**, 109756 (2025), doi:10.1016/j.cpc.2025.109756 [preprint doi:10.48550/arXiv.2409.18464].

[10] A. Fowlie, *stanhf: HistFactory models in the probabilistic programming language Stan*, Eur. Phys. J. C **85**, 923 (2025), doi:10.1140/epjc/s10052-025-14495-1 [preprint doi:10.48550/arXiv.2503.22188].

[11] E. Alvarez, B. M. Dillon, D. A. Faroughy, J. F. Kamenik, F. Lamagna and M. Szewc, *Bayesian probabilistic modeling for four-top production at the LHC*, Phys. Rev. D **105**, 092001 (2022), doi:10.1103/PhysRevD.105.092001 [preprint doi:10.48550/arXiv.2107.00668].

[12] E. Alvarez, M. Spannowsky and M. Szewc, *Unsupervised quark/gluon jet tagging with Poissonian mixture models*, Front. Artif. Intell. **5**, 852970 (2022), doi:10.3389/frai.2022.852970 [preprint doi:10.48550/arXiv.2112.11352].

[13] E. Alvarez, *Bayesian inference to study a signal with two or more decaying particles in a non-resonant background*, (arXiv preprint) doi:10.48550/arXiv.2210.07358.

[14] E. Alvarez, M. Szewc, A. Szynkman, S. Tanco and T. Tarutina, *Exploring unsupervised top tagging using Bayesian inference*, SciPost Phys. Core **6**, 046 (2023), doi:10.21468/SciPostPhysCore.6.2.046 [preprint doi:10.48550/arXiv.2212.13583].

[15] E. Alvarez and Y. Yao, *Inferring flavor mixtures in multijet events*, SciPost Phys. Core **7**, 076 (2024), doi:10.21468/SciPostPhysCore.7.4.076 [preprint doi:10.48550/arXiv.2404.01387].

[16] E. Alvarez, L. Da Rold, M. Szewc, A. Szynkman, S. A. Tanco and T. Tarutina, *Improvement and generalization of ABCD method with Bayesian inference*, SciPost Phys. Core **7**, 043 (2024), doi:10.21468/SciPostPhysCore.7.3.043 [preprint doi:10.48550/arXiv.2402.08001].

[17] S. Adler et al., *Search for the decay* $K^+ \to \pi^+ \nu \bar{\nu}$, Phys. Rev. Lett. **76**, 1421 (1996), doi:10.1103/PhysRevLett.76.1421 [preprint doi:10.48550/arXiv.hep-ex/9510006] (An approach to the ABCD idea might be found here).

[18] S. Adler et al., *Evidence for the decay* $K^+ \to \pi^+ \nu \bar{\nu}$, Phys. Rev. Lett. **79**, 2204 (1997), doi:10.1103/PhysRevLett.79.2204 [preprint doi:10.48550/arXiv.hep-ex/9708031] (An approach to the ABCD idea might be found here).

[19] J. Bardhan, T. Mandal, S. Mitra, C. Neeraj and M. Patra, *Unsupervised and lightly supervised learning in particle physics*, Eur. Phys. J. Spec. Top. **233**, 2559 (2024), doi:10.1140/epjs/s11734-024-01235-x [preprint doi:10.48550/arXiv.2403.13676].

[20] E. M. Metodiev, B. Nachman and J. Thaler, *Classification without labels: Learning from mixed samples in high energy physics*, J. High Energy Phys. **10**, 174 (2017), doi:10.1007/JHEP10(2017)174 [preprint doi:10.48550/arXiv.1708.02949].

[21] A. Andreassen, B. Nachman and D. Shih, *Simulation assisted likelihood-free anomaly detection*, Phys. Rev. D **101**, 095004 (2020), doi:10.1103/PhysRevD.101.095004 [preprint doi:10.48550/arXiv.2001.05001].

[22] K. Benkendorfer, L. Le Pottier and B. Nachman, *Simulation-assisted decorrelation for resonant anomaly detection*, Phys. Rev. D **104**, 035003 (2021), doi:10.1103/PhysRevD.104.035003 [preprint doi:10.48550/arXiv.2009.02205].

[23] G. Kasieczka, B. Nachman, M. D. Schwartz and D. Shih, *Automating the ABCD method with machine learning*, Phys. Rev. D **103**, 035021 (2021), doi:10.1103/PhysRevD.103.035021 [preprint doi:10.48550/arXiv.2007.14400].

[24] S. Klein, M. Leigh, S. Mulligan and T. Golling, *Strong CWoLa: Binary classification without background simulation*, (arXiv preprint) doi:10.48550/arXiv.2503.14876.

[25] E. M. Metodiev, J. Thaler and R. Wynne, *Anomaly detection in collider physics via factorized observables*, Phys. Rev. D **110**, 055012 (2024), doi:10.1103/PhysRevD.110.055012 [preprint doi:10.48550/arXiv.2312.00119].

[26] K. Desai, B. Nachman and J. Thaler, *Moment extraction using an unfolding protocol without binning*, Phys. Rev. D **110**, 116013 (2024), doi:10.1103/PhysRevD.110.116013 [preprint doi:10.48550/arXiv.2407.11284].

[27] A. Hallin, J. Isaacson, G. Kasieczka, C. Krause, B. Nachman, T. Quadfasel, M. Schlaffer, D. Shih and M. Sommerhalder, *Classifying anomalies through outer density estimation*, Phys. Rev. D **106**, 055006 (2022), doi:10.1103/PhysRevD.106.055006 [preprint doi:10.48550/arXiv.2109.00546].

[28] A. Hallin, G. Kasieczka, T. Quadfasel, D. Shih and M. Sommerhalder, *Resonant anomaly detection without background sculpting*, Phys. Rev. D **107**, 114012 (2023), doi:10.1103/PhysRevD.107.114012 [preprint doi:10.48550/arXiv.2210.14924].

[29] J. A. Raine, S. Klein, D. Sengupta and T. Golling, *CURTAINs for your sliding window: Constructing unobserved regions by transforming adjacent intervals*, Front. Big Data **6**, 899345 (2023), doi:10.3389/fdata.2023.899345 [preprint doi:10.48550/arXiv.2203.09470].

[30] ATLAS collaboration: M. Aaboud et al., *Measurement of jet-substructure observables in top quark, W boson and light jet production in proton-proton collisions at* $\sqrt{s} = 13$ *TeV with the ATLAS detector*, J. High Energy Phys. **08**, 033 (2019), doi:10.1007/JHEP08(2019)033 [preprint doi:10.48550/arXiv.1903.02942].

[31] CMS collaboration: A. M. Sirunyan et al., *Measurement of differential $t\bar{t}$ production cross sections using top quarks at large transverse momenta in pp collisions at $\sqrt{s} = 13$ TeV*, Phys. Rev. D **103**, 052008 (2021), doi:10.1103/PhysRevD.103.052008 [preprint doi:10.48550/arXiv.2008.07860].

[32] ATLAS collaboration: G. Aad et al., *Measurement of the energy asymmetry in $t\bar{t}j$ production at 13 TeV with the ATLAS experiment and interpretation in the SMEFT framework*, Eur. Phys. J. C **82**, 374 (2022), doi:10.1140/epjc/s10052-022-10101-w [preprint doi:10.48550/arXiv.2110.05453].

[33] CMS collaboration: A. M. Sirunyan, *Search for electroweak production of a vector-like quark decaying to a top quark and a Higgs boson using boosted topologies in fully hadronic final states*, J. High Energy Phys. **04**, 136 (2017), doi:10.1007/JHEP04(2017)136 [preprint doi:10.48550/arXiv.1612.05336].

[34] ATLAS collaboration: G. Aad et al., *Search for single production of a vectorlike T quark decaying into a Higgs boson and top quark with fully hadronic final states using the ATLAS detector*, Phys. Rev. D **105**, 092012 (2022), doi:10.1103/PhysRevD.105.092012 [preprint doi:10.48550/arXiv.2201.07045].

[35] J. Y. Araz, A. Buckley and B. Fuks, *Searches for new physics with boosted top quarks in the MadAnalysis 5 and Rivet frameworks*, Eur. Phys. J. C **83**, 664 (2023), doi:10.1140/epjc/s10052-023-11779-2 [preprint doi:10.48550/arXiv.2303.03427].

[36] J. R. Forshaw and M. H. Seymour, *Subjet rates in hadron collider jets*, J. High Energy Phys. **09**, 009 (1999), doi:10.1088/1126-6708/1999/09/009 [preprint doi:10.48550/arXiv.hep-ph/9908307].

[37] S. Catani, Y. L. Dokshitzer and B. R. Webber, *Average number of jets in deep inelastic scattering*, Phys. Lett. B **322**, 263 (1994), doi:10.1016/0370-2693(94)91118-5.

[38] M. Cacciari, G. P. Salam and G. Soyez, *FastJet user manual*, Eur. Phys. J. C **72**, 1896 (2012), doi:10.1140/epjc/s10052-012-1896-2 [preprint doi:10.48550/arXiv.1111.6097].

[39] C. Bierlich et al., *A comprehensive guide to the physics and usage of PYTHIA 8.3*, SciPost Phys. Codebases 8 (2022), doi:10.21468/SciPostPhysCodeb.8 [preprint doi:10.48550/arXiv.2203.11601].

C. Bierlich et al., *Codebase release 8.3 for PYTHIA*, SciPost Phys. Codebases 8-r8.3 (2022), doi:10.21468/SciPostPhysCodeb.8-r8.3.

[40] M. Bähr et al., *Herwig++ physics and manual*, Eur. Phys. J. C **58**, 639 (2008), doi:10.1140/epjc/s10052-008-0798-9 [preprint doi:10.48550/arXiv.0803.0883].

[41] Stan development team, *Stan modeling language users guide and reference manual* (2017), https://hero.epa.gov/hero/index.cfm/reference/details/reference_id/4235802.

[42] S. D. Ellis and D. E. Soper, *Successive combination jet algorithm for hadron collisions*, Phys. Rev. D **48**, 3160 (1993), doi:10.1103/PhysRevD.48.3160 [preprint doi:10.48550/arXiv.hep-ph/9305266].

[43] M. Betancourt, *A conceptual introduction to Hamiltonian Monte Carlo*, (arXiv preprint) doi:10.48550/arXiv.1701.02434.

[44] M. Szewc and T. Tarutina, *Inferring correlated distributions: Boosted top jets* (2025), https://github.com/ttarutina/BoostedTopJets.

[45] G. Kasieczka, T. Plehn, J. Thompson and M. Russel, *Top quark tagging reference dataset*, Zenodo (2019), doi:10.5281/zenodo.2603256.

[46] J. de Favereau, C. Delaere, P. Demin, A. Giammanco, V. Lemaître, A. Mertens and M. Selvaggi, *DELPHES 3: A modular framework for fast simulation of a generic collider experiment*, J. High Energy Phys. **02**, 057 (2014), doi:10.1007/JHEP02(2014)057 [preprint doi:10.48550/arXiv.1307.6346].

[47] M. Cacciari, G. P. Salam and G. Soyez, *The anti-k_t jet clustering algorithm*, J. High Energy Phys. **04**, 063 (2008), doi:10.1088/1126-6708/2008/04/063 [preprint doi:10.48550/arXiv.0802.1189].

[48] C. M. Bishop, *Pattern recognition and machine learning*, Springer, New York, USA, ISBN 9780387310732 (2006).

[49] L. Lyons, *Statistics for nuclear and particle physicists*, Cambridge University Press, Cambridge, UK, ISBN 9780521379342 (1986), doi:10.1017/CBO9781139167710.

[50] S. T. Radev, U. K. Mertens, A. Voss, L. Ardizzone and U. Köthe, *Bayesflow: Learning complex stochastic models with invertible neural networks*, IEEE Trans. Neural Netw. Learn. Syst. **33**, 1452 (2020), doi:10.1109/TNNLS.2020.3042395.

[51] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin, *Bayesian data analysis*, CRC Press, Boca Raton, USA, ISBN 9781439840955 (2013).

[52] M. Stephens, *Dealing with label switching in mixture models*, J. R. Stat. Soc. B: Stat. Methodol. **62**, 795 (2000), doi:10.1111/1467-9868.00265.

[53] CMS collaboration, *Development of systematic uncertainty-aware neural network trainings for binned-likelihood analyses at the LHC*, (arXiv preprint) doi:10.48550/arXiv.2502.13047.