Uncovering hidden patterns in collider events with Bayesian probabilistic models

Darius A. Faroughy\textsuperscript{1,}\textsuperscript{*}

\textsuperscript{1}Physik-Institut, University of Zürich
Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

Abstract  Individual events at high-energy colliders like the LHC can be represented by a sequence of measurements, or ‘point patterns’ in an observable space. Starting from this data representation, we build a simple Bayesian probabilistic model for event measurements useful for unsupervised event classification in beyond the standard model (BSM) studies. In order to arrive to this model we assume that the event measurements are exchangeable (and apply De Finetti’s representation theorem), the data is discrete, and measurements are generated from multiple ‘latent’ distributions, called themes. The resulting probabilistic model for collider events is a mixed-membership model known as Latent Dirichlet Allocation (LDA), a model extensively used in natural language processing applications. By training on point patterns in the primary Lund plane, we demonstrate that a two-theme LDA model can learn to distinguish in unlabelled dijet events the hidden new physics patterns produced by a BSM signature from a much larger QCD background. This note is based on refs. [1, 2].

I. INTRODUCTION

A collider event $e$ can be represented by a sequence $\{o_1, o_2, \ldots, o_N\}$ of measurements, or observations $o_i$, taking values in some space $\mathcal{O}$ spanned by a set of observables $\mathcal{O}_1, \ldots, \mathcal{O}_k$. For example, the $(p_T, \eta, \phi)$ particle coordinates at a hadron collider. Collider events can be thought as individual realizations of a stochastic point process in $\mathcal{O}$. Each event can be represented by a distribution of points

$$e(o) = \sum_{i=1}^{N} \delta^{(k)}(o - o_i),$$

in $\mathcal{O}$, where the number $N$ of event measurements can be a random variables changing from event to event, or a deterministic quantity. For most collider events, the corresponding point patterns will not be uniformly distributed over $\mathcal{O}$. For instance, at hadron colliders a substantial amount of the energy from the high-energy $pp$-collision is emitted in the form of collimated sprays of hadrons. These hadronic sprays, known as jets, lead to clustered point patterns in the space $\mathcal{O} = (\eta, \phi)$. For high-level observables spanning $\mathcal{O}$, the resulting point patterns for each event can be quite sparse, or give rise to irregularly shaped patterns when averaging over many events. For example, event point patterns in the Lund planes \cite{3} are both sparse and irregular in shape. Building a completely general probabilistic model for event measurements $\mathcal{P}(e) = \mathcal{P}(o_1, o_2, \ldots, o_N)$ for an arbitrary $\mathcal{O}$ is therefore very challenging.

II. A SIMPLE PROBABILISTIC MODEL FOR COLLIDER EVENTS

In this note we show that it is possible to write down a simple Bayesian probabilistic model for $\mathcal{P}(e)$ that is capable of describing to a good approximation the generative process for event measurements. Moreover, following refs. [1, 2], we demonstrate that the model can be used for unsupervised event classification. The probabilisitic model is based on three model-building

\textsuperscript{*} Proceedings for the 40th International Conference on High Energy Physics - ICHEP 2020 Prague, Czechia
Email: faroughy@physik.uzh.ch
assumptions: (i) Measurements in an event are exchangeable, (ii) the observable space $O$ is discretized, and (iii) event measurements are generated from multiple (latent) probability distributions over $O$.

**a. Exchangeability.** The first of the these assumptions requires that all event measurements are exchangeable, i.e. the order in which the measurements $o_i$ of an event are extracted is irrelevant. This implies permutation invariance:

$$\mathcal{P}(o_1, \ldots, o_N) = \mathcal{P}(o_{\pi(1)}, \ldots, o_{\pi(N)}),$$

where $\pi$ is any element of the permutation group of $N$ indices. Exchangeability must not be confused with independent and identically distributed (iid). For iid measurements, the probability distribution would be completely factorizable and indeed exchangeable, but the converse wouldn’t necessarily be true, not all exchangeable sequences are iid. Exchangeability actually implies a weaker notion of statistical independence called ‘conditional independence’. Both concepts are related through De Finetti’s representation theorem:

**De Finetti’s representation theorem:** A sequence of event measurements is exchangeable iff there exists a latent variable $\omega$ over some latent space $\Omega$, and a distribution $\mathcal{P}(\omega)$, such that

$$\mathcal{P}(o_1, \ldots, o_N) = \int_\Omega d\omega \mathcal{P}(\omega) \prod_{i=1}^N \mathcal{P}(o_i|\omega).$$

This result implies that if event measurements in $O$ are exchangeable, then these can be thought as being conditionally independent with respect to some marginalized hidden variable $\omega$. An event is generated by first sampling some random element $\omega$ from a latent space $\Omega$, then each measurement in the event is drawn from a distribution over $O$ conditioned on the drawn $\omega$. Looking closely at the integral representation in (3) one recognizes $\mathcal{P}(\omega)$ as a prior and $\mathcal{P}(o|\omega)$ as a likelihood, thus justifying the use of Bayesian probabilistic modelling for exchangeable data.

**b. Measurement Discretization.** Permutation invariance leads to a very simple conditional structure for $\mathcal{P}(o_1, \ldots, o_N)$, but De Finetti’s theorem does not specify how to select the latent space $\Omega$, nor how to model the prior $\mathcal{P}(\omega)$ or the conditional distribution $\mathcal{P}(o|\omega)$ in (3). For this, we need additional assumptions. One possibility, which makes parameter inference much simpler, is to choose the prior and likelihood to be conjugate distributions, for instance, these can belong to the exponential family. Our second model-building assumption is that the distribution $\mathcal{P}(o|\omega)$ over $O$ is a discrete distribution. For this to make sense, we discretize the continuous observables spanning $O$ by binning this space so that the outcome of any event measurement is a discrete unit, or token, represented by the bin it populates. Notice that this ‘tokenization’ of event measurements reduces the problem of finding a continuous distribution $\mathcal{P}(o|\omega)$ over a multidimensional space $O$, to finding a discrete distribution over the finite set of non-negative integers labelling the bins in $O$. From all the discrete distributions in the exponential family, the most natural choice for $\mathcal{P}(o|\omega)$ is the multinomial distribution (a multivariate generalization of the binomial distribution). This distribution is parametrized by a $M$-dimensional vector $\beta = (\beta_1, \cdots, \beta_M)$, satisfying

$$\sum_{m=1}^M \beta_m = 1 \quad \text{and} \quad 0 \leq \beta_m \leq 1,$$

where $M$ is the total number of bins that partition $O$ and the number $\beta_m$ represents the probability that a measurement $o_i$ populates the $m$th bin. In order to generate an individual (tokenized) event measurement $o_i$, we first draw $\omega$ from the prior, then, we randomly draw an index $m \in \{1, \ldots, M\}$ from the multinomial $\mathcal{P}(o|\omega, \beta)$ conditioned on $\omega$. The resulting index points toward the bin in $O$ that the measurement belongs to. The sampling of an event measurement from the multinomial can be pictured as rolling a dice with $M$ sides and bias $\beta$, which at this level is a free parameter of our probabilistic model.
In order to ‘smooth’ the multinomial parameter, we introduce a prior for $\beta$. The most natural prior is the Dirichlet distribution, a member of the exponential family that is conjugate to the multinomial distribution, defined as

$$
\mathcal{D}(\beta|\eta) = \frac{\Gamma(\eta_1 + \cdots + \eta_M)}{\Gamma(\eta_1) \cdots \Gamma(\eta_M)} \prod_{m=1}^{M} (\beta_m)^{\eta_m - 1}.
$$

The Dirichlet $\mathcal{D}(\cdot|\eta)$ is a family of distributions with concentration parameter $\eta = (\eta_1, \ldots, \eta_M)$, $\eta_m > 0$, and $\Gamma(x)$ denotes the Gamma function. The concentration parameter controls the shape of the Dirichlet distribution over $\beta$ space. This space is an $(M - 1)$-dimensional simplex. Notice that introducing this prior makes our model fully Bayesian, since we have replaced the task of fixing a large set of parameters (the probabilities $\beta$) of the multinomial with choosing a suitable Dirichlet distribution from which these parameters are sampled from.

c. **Latent Dirichlet Allocation.** We now need to specify the nature of the latent variable $\omega$ and the conditional dependence of the multinomial $\mathcal{P}(o|\omega, \beta)$ with $\omega$. This brings us to our third model-building assumption which is that the measurements $o_i$ in an event are assumed to arise from multiple multinomial distributions $\mathcal{P}(o|t, \beta_t)$, labeled by a finite index $t \in \{1, \ldots, T\}$ and parametrized by $\beta_t = (\beta_{t1}, \ldots, \beta_{tM})$. Each multinomial distribution represents an underlying event category, or theme, potentially describing features from multiple underlying physical processes or phenomena. “Themes” is a terminology borrowed from the machine learning community, specifically from topic modelling and natural language processing. The latent variable is a $T$-dimensional vector $\omega = (\omega_1, \ldots, \omega_T)$ describing the relative proportion of every theme in the event. The likelihood in De Finetti’s representation takes the form of a multinomial mixture model

$$
\mathcal{P}(o|\omega) = \sum_{t=1}^{T} \mathcal{P}(t|\omega) \mathcal{P}(o|t, \beta_t).
$$

The discrete distributions $\mathcal{P}(t|\omega)$ are also multinomial distributions that are parametrized by the latent variable $\omega$. These represent the probability of selecting a particular theme $\mathcal{P}(o|t, \beta_t)$ from which event measurements are extracted. The latent space $\Omega$ is a now a $(T - 1)$-dimensional simplex, denoted by $\Omega_T$, spanned by the latent mixtures $\omega$ which now satisfy the convexity constraints as in (4).\footnote{The simplex $\Omega_T$ must not be confused with the simplices for the multinomial theme parameters $\beta_t$.} This implies that the most natural choice for the prior $\mathcal{P}(\omega)$ in (3) is the Dirichlet distribution over such simplex. With these model-building assumptions, we finally arrive to a fairly simple generative model for collider events over $\mathcal{O}$:

$$
\mathcal{P}(o_1, \ldots, o_N|\alpha, \eta) = \left( \prod_{t=1}^{T} \mathcal{D}(\beta_t|\eta_t) \right) \int_{\Omega_T} d\omega \mathcal{D}(\omega|\alpha) \prod_{i=1}^{N} \left[ \sum_{t=1}^{T} \mathcal{P}(t|\omega) \mathcal{P}(o_i|t, \beta_t) \right]
$$

This model is known as Latent Dirichlet Allocation (LDA), and was first proposed as a topic model for texts\footnote{Other topic models have been previously used for collider studies in [4] for quark/gluon jet discrimination.} [5]. The model has two (multidimensional) model-building parameters governing the shapes of the Dirichlet distributions: the $T$-dimensional vector $\alpha = (\alpha_1, \ldots, \alpha_T)$ for the theme mixing proportions and a $T \times M$ matrix $\eta$ where the $M$-dimensional row $\eta_t$ controls the shape of the Dirichlet for the theme multinormals over $\mathcal{O}$. The number of themes $T$ is also a model building parameter to be fixed before training these models with data. The simplest possible model is the two-theme LDA model. When $T = 2$, the Dirichlet prior $\mathcal{D}(\omega|\alpha_1, \alpha_2)$ becomes a beta distribution over the unit interval, and $\mathcal{P}(t|\omega)$ is a binomial distribution over $t \in \{1, 2\}$. After fixing the priors, the generative process for a single collider event goes as follows:

- (i) Draw a random mixing $\omega$ between zero and one from the beta prior.
- (ii) Flip a coin with bias $\omega$. 

1. Other topic models have been previously used for collider studies in [4] for quark/gluon jet discrimination.
• (iii) If the coin lands on ‘heads’ select the first theme ($t = 1$), otherwise select the second theme ($t = 2$).

• (iv) Randomly sample one event measurement $o \in \mathcal{O}$ from the selected theme multinomial by rolling an $M$-sided dice with bias $\beta_t$.

• (v) Repeat steps (ii-iv) until all measurements $o_1, \ldots, o_N$ in the event have been generated.

LDA is a mixed-membership model because each measurements $o_i$ within an event can arise from multiple themes (e.g. a ‘head’ or a ‘tail’ theme when $T = 2$), and each event within a sample exhibits these themes with different proportions. Mixed-membership models are not to be confused with classical mixture models. In the later, all measurements within an event are limited to come from a single theme (the mixture of themes is manifest at the event sample level, and not at the event level), while the former are more flexible probabilistic model that are capable of capturing common features between different underlying physical processes.

d. Event classification with LDA. After fixing the Dirichlet parameters $\alpha, \eta$ and the number of themes $T = 2$, we can use LDA for unsupervised event classification. The posterior distribution $P(\omega, t, \beta|o, \alpha, \eta)$ is calculated using Bayes theorem. The idea is to learn from unlabelled collider data the theme multinomial parameters $\beta_m$ and use them to cluster events into two categories. We use variational inference (VI) [5] for the learning algorithm. During training, the algorithm learns the themes by identifying recurring measurement patterns, in particular, it identifies co-occurrences between measurement bins throughout the event sample. Once the learning converges and the themes have been extracted, we build a likelihood-ratio defined by

$$L(o_1, \ldots, o_N|\alpha) = \prod_{i=1}^{N} \frac{P(o_i|1, \hat{\beta}_1(\alpha))}{P(o_i|2, \hat{\beta}_2(\alpha))}. \quad (8)$$

The $\hat{\beta}_t$ are statistical estimators for the $\beta_t$’s extracted from VI. The classifier is obtained by thresholding: for some suitable $c \in \mathbb{R}$, if $L(o_1, \ldots, o_N|\alpha) > c$ then the event belongs to theme $t = 1$, else it belongs to theme $t = 2$. This classifier is a function of the Dirichlet parameter $\alpha$, and is better thought as a continuous ‘landscape’ of LDA classifiers. In principle there is no robust criteria for choosing one specific set of $\alpha$’s over another. Preliminary results given in ref. [2] suggest that a quantity known as perplexity can be used to precisely select the best $\alpha$.

III. LATENT DIRICHLET ALLOCATION FOR JET SUBSTRUCTURE

We now demonstrate how a two-theme LDA model can be used to uncover Beyond the SM (BSM) physics hiding in multi-jet events. First, we choose a set of jet observables for $\mathcal{O}$. Observables that associate only one measurement to each event are not suitable for our method because this would produce for each event a single measurement\(^3\) in $\mathcal{O}$. In order for LDA to learn from measurement co-occurrence, we need observables that produce for every event a pattern of points in $\mathcal{O}$. One possibility is to use observables extracted from the de-clustering history of jets. The jet de-clustering procedure generates a binary tree where each node corresponds to a splitting of a mother subjet into two subsequent daughter subjets $j_0 \rightarrow j_1j_2$. During each splitting, a set of measurements $o$ is registered, generating a sequence of points in $\mathcal{O}$ for the whole de-clustering tree. Assuming the de-clustering history to be exchangeable (i.e. ignoring the conditional dependence between measurements) is a good enough approximation for event classification purposes. For the splitting observables we choose quantities that are sensitive to generic decay configurations of massive resonances, like the subjets invariant mass $m_0$, mass drop $m_1/m_0$, and Lund plane observables, $k_T$ and $\Delta$, defined in [3]. We then build a multi-dimensional space $\mathcal{O}$ spanned by different combination of these observables. Moreover, we

\(^3\) Jet substructure observables that marginalize over all particles in the event, like e.g. $N$-subjettiness [6], fall into this category and are therefore not useful for LDA.
In our studies, we used for the hidden BSM benchmark a $W'\to \phi$ model [1, 2] with a boson mass $M_{W'} = 3$ TeV and scalar mass $M_{\phi} = 0.4$ TeV. For the signal process we considered $pp \to W'\to W\phi \to WWWW$, with $W$ bosons decaying hadronically. For the background we considered QCD dijet production. We generated 100k background and signal events and performed jet clustering using the $C/A$ algorithm with $R = 1$. For the splitting observables we used the primary Lund plane $\mathcal{O}$ spanned by $\{\log k_T, \log R/\Delta\}$, and also included labels $j = 1, 2, \ldots$ indicating to which de-clustered jet in the event the measurement belongs (leading jet, subleading jet, etc…, ordered by invariant mass). The truth-level distributions for the primary Lund plane are given in figure 1, for the QCD background (first column) and signal (second column), for the leading jet (top row) and subleading jet (bottom row). The region near the hypothenuse of the Lund triangles describe the hard and collinear splittings. This region exhibits discriminating features between signal and background: for the signal we find two (one) dark clusters for the leading (subleading) jet, corresponding to the massive decay $\phi \to WW \to jjjj (W \to jj)$, while for the QCD background we expect a uniform pattern along the hypothenuse. We also can see non-perturbative features discriminating between background and signal along the $\log k_T \sim 0$ axis.

We produced an unlabelled mixed sample of 100k events with $s/b = 5\%$ and used it to train a two-theme LDA on the primary Lund plane with the Gensim python package [7]. For the Dirichlet prior $\mathcal{D}(\omega|\alpha)$ controlling the theme mixings we fixed it to a very asymmetric shape $\alpha_0 \approx 0.9$ and $\alpha_2 \approx 0.1$. During training, this choice forces one multinomial theme ($t = 1$) to approximate the mixed data distribution which we know (a priori) to be QCD-dominated because $s \ll b$. On the other hand, the other theme ($t = 2$) is expected to learn non-QCD patterns in the Lund plane, with the hope that it picks up signal features. The outcome of the learned themes are shown in figure 1: the first theme (third column) matches very well with the QCD truth level distribution (first column), while the second theme (fourth column) contains the new physics signal features present in the truth level signal (second column). This result demonstrates that the two-theme LDA model can extract small BSM signals from a large background in a completely unsupervised manner. For more details see ref. [2].

IV. CONCLUSIONS

In conclusion, we have demonstrated that it is possible to build a simple generative probabilistic model for collider events. This model can be used for unsupervised event classification, e.g. for extracting BSM physics from jet substructure. The method presented here is based on a Bayesian probabilistic model called Latent Dirichlet Allocation. We arrived to this model starting from three main assumptions: (i) collider event measurements are to a good approximation
‘exchangeable’, leading to De Finetti’s integral representation for $\mathcal{P}(o_1, o_2, \ldots)$, (ii) individual measurements are discrete (i.e. tokenized), and (iii) measurements arise from a multiplicity of latent (multinomial) distributions over $O$, called ‘themes’. We trained a two-theme LDA model on the primary Lund plane from a mixed dijet events sample with QCD background and BSM signal (a $W' - \phi$ model) at $s/b = 5\%$. Our results show that LDA can successfully discover small BSM signals in unlabelled data.

[1] B.M. Dillon, D.A. Faroughy and J.F. Kamenik, *Uncovering latent jet substructure*, Phys. Rev. D **100** (2019) 056002 [1904.04200].

[2] B. Dillon, D. Faroughy, J. Kamenik and M. Szewc, *Learning the latent structure of collider events*, JHEP **10** (2020) 206 [2005.12319].

[3] F.A. Dreyer, G.P. Salam and G. Soyez, *The Lund Jet Plane*, JHEP **12** (2018) 064 [1807.04758].

[4] E.M. Metodiev and J. Thaler, *Jet Topics: Disentangling Quarks and Gluons at Colliders*, Phys. Rev. Lett. **120** (2018) 241602 [1802.00008].

[5] D.M. Blei, A.Y. Ng, M.I. Jordan and J. Lafferty, *Latent dirichlet allocation*, Journal of Machine Learning Research **3** (2003) 2003.

[6] J. Thaler and K. Van Tilburg, *Identifying Boosted Objects with N-subjettiness*, JHEP **03** (2011) 015 [1011.2268].

[7] R. Rehurek and P. Sojka, *Software framework for topic modelling with large corpora*, in *LREC 2010 WORKSHOP ON NEW CHALLENGES FOR NLP FRAMEWORKS*, pp. 45–50, 2010.