PROCEEDINGS OF SCIENCE

Uncovering hidden new physics patterns in collider events with Bayesian probabilistic models

Darius A. Faroughy[•]

[•]*Physik-Institut, Zurich University,*
Winterthurerstrasse 190, Zurich 8057, Switzerland

E-mail: faroughy@physik.uzh.ch

Individual events at high-energy colliders like the LHC can be represented by a sequence of measurements, or 'point patterns'. Starting from this generic 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 mixed dijet samples of QCD and BSM, we demonstrate that a two-theme LDA model can learn to distinguish in (unlabelled) jet substructure data the hidden new physics patterns produced by a non-trivial BSM signature from a much larger QCD background.

40th International Conference on High Energy physics - ICHEP2020
July 28 - August 6, 2020
Prague, Czech Republic (virtual meeting)

1. A simple probabilistic model for collider events

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 $\mathcal{O}_1, \ldots, \mathcal{O}_k$, like e.g. the (p_T, η, ϕ) particle coordinates in a hadron collider. Collider events can be thought as individual realizations of a *stochastic point process* in $\mathcal{O}$, represented by an unordered distribution of points

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

where the number of measurements N, as well as their positions in $\mathcal{O}$ are random variable changing from event to event. For the interesting events, the corresponding point patterns will not be uniformly distributed in $\mathcal{O}$. 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 a generic $\mathcal{O}$, the resulting point patterns for individual events can be very sparse, or give rise to irregularly shaped patterns when averaging over many events. A complete probabilistic model for these sequence of event measurements $\mathcal{P}(e) = \mathcal{P}(o_1, o_2, \ldots, o_N)$ is very challenging, if not impossible. Our goal in this note is to show that it is possible to write a simple generative model for $\mathcal{P}(e)$ that is capable of (approximately) capturing hidden features in events and that it can be used successfully for unsupervised event classification [1, 2]. In the following we build a generative model based on three (probabilistic) model-building assumptions: (i) Measurements in an event are exchangeable, (ii) the observable space $\mathcal{O}$ is discretized, and (iii) event measurements are generated from multiple (latent) probability distributions over $\mathcal{O}$.

Exchangeability. Our first model-building assumption is that event measurements are *exchangeable*, i.e. the order in which the measurements o_i 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 π 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. 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 theorem)** *A sequence of event measurements is exchangeable iff there exists a latent variable ω over some latent space Ω, and a distribution $\mathcal{P}(\omega)$, such that*

$$\mathcal{P}(o_1, \ldots, o_N) \;=\; \int_\Omega \mathrm{d}\omega\, \mathcal{P}(\omega) \prod_{i=1}^{N} \mathcal{P}(o_i|\omega)\,. \tag{2}$$

This result implies that if event measurements in $\mathcal{O}$ are exchangeable, then these can be thought as being conditionally independent with respect to some marginalized hidden variable ω. An event is generated by first sampling some random element ω from a latent space Ω, then each measurment in the event is drawn from a distribution over $\mathcal{O}$ conditioned on the drawn ω. Looking closely at the integral representation in (2) one recognizes $\mathcal{P}(\omega)$ as a prior and $\mathcal{P}(o|\omega)$ as a likelihood, and thus justifying the use of Bayesian probabilistic modelling.

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 Ω, nor how to model the prior $\mathcal{P}(\omega)$ or the conditional distribution $\mathcal{P}(o|\omega)$ in (2). 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 in what follows will be that the distribution $\mathcal{P}(o|\omega)$ over $\mathcal{O}$ is discrete. For this to make sense, we first discretize the continuous observables spanning $\mathcal{O}$ by binning this space so that the outcome of any event measurement is now 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 $\mathcal{O}$, to finding a discrete distribution over the finite set of non-negative integers labelling the bins in $\mathcal{O}$. 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 normalized M-dimensional vector $\beta = (\beta_1, \cdots, \beta_M)$, with $\sum_m^M \beta_m = 1$ and $0 \leq \beta_m \leq 1$, where M is the total number of bins that partition $\mathcal{O}$ and the number β_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 randomly sample an ω from the prior, then, we randomly draw an index $m \in \{1, \ldots, M\}$ from the multinomial $\mathcal{P}(o|\omega, \beta)$ conditioned on ω. The resulting index points toward the bin where this measurement belongs to. The sampling of an event measurement from the multinomial can be pictured as rolling a dice with M sides and bias β, which at this level is a free parameter of our probabilistic model. In order to 'smooth' the multinomial parameter, we introduce a prior for β. The most natural prior is the *Dirichlet distribution*, a member of the exponential family that is conjugate to the multinomial distribution, defined as

$$\mathcal{D}(\beta|\eta) \;=\; \frac{\Gamma(\eta_1 + \cdots + \eta_M)}{\Gamma(\eta_1)\cdots\Gamma(\eta_M)} \prod_{m=1}^{M} (\beta_m)^{\eta_m - 1} \,. \tag{3}$$

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 governs the possible shapes of the Dirichlet over β space, i.e. over the $(M-1)$-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 β) of the multinomial with choosing a suitable Dirichlet distribution from which these parameters are sampled from.

Latent Dirichlet Allocation. We now need to specify the nature of the latent variable ω and the conditional dependence of the multinomial $\mathcal{P}(o|\omega, \beta)$ with ω. 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 $\beta_t = (\beta_{t1}, \cdots, \beta_{tM})$. 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 unsupervised text classification and natural language processing. The latent variable is a T-dimensional vector $\omega = (\omega_1, \ldots, \omega_T)$ describing the mixing proportions for every theme. The likelihood in De Finetti's representation takes the form of a multinomial mixture model, $\mathcal{P}(o|\omega) = \sum_{t=1}^{T} \mathcal{P}(t|\omega)\mathcal{P}(o|t,\beta_t)$, with a fixed number ($T$) of multinomials $\mathcal{P}(o|t,\beta_t)$. The discrete distributions $\mathcal{P}(t|\omega)$ are also multinomial distributions that are parametrized by the latent variable ω. These represent the probability of selecting a particular theme $\mathcal{P}(o|t,\beta_t)$ from which event measurements are extracted. Therefore, the latent space Ω is a $(T-1)$-dimensional simplex, denoted by Ω_T, spanned by the latent mixtures ω which now satisfy the convexity constraints $\sum_{t=1}^{T} \omega_t = 1, 0 \leq \omega_t \leq 1$. This implies that the most natural choice for the prior $\mathcal{P}(\omega)$ in (2) 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 $\mathcal{O}$:

$$\mathcal{P}(o_1,\ldots,o_N|\alpha,\eta) \;=\; \left(\prod_{t=1}^{T} \mathcal{D}(\beta_t|\eta_t)\right) \int_{\Omega_T} \mathrm{d}\omega\, \mathcal{D}(\omega|\alpha) \prod_{i=1}^{N} \left[\sum_{t=1}^{T} \mathcal{P}(t|\omega)\, \mathcal{P}(o_i|t,\beta_t)\right] \tag{4}$$

This model is known as Latent Dirichlet Allocation (LDA), and was first proposed as a topic model for texts[1] [4]. 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 η where the M-dimensional row η_t controls the shape of the Dirichlet for the theme multinomials 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 ω between zero and one from the beta prior, (ii) flip a coin with bias ω, (iii) if the coin lands on 'heads' select the first theme, otherwise select the second theme, (iv) from the selected theme randomly sample one measurement $o \in \mathcal{O}$. 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 mixture models. In the later, all measurements within an event are limited to come from a single theme (the mixture of theme 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 in different physical processes.

Event classification with LDA. After fixing the Dirichlet parameters α, η and the number of themes $T = 2$, we can use LDA for unsupervised event classification. The posterior distribution $\mathcal{P}(\omega,t,\beta|o_i,\alpha,\eta)$ is calculated using Bayes theorem. The idea is to learn from unlabelled collider data the theme multinomial parameters β_{tm} and use them to cluster events into two categories. We use variational inference (VI) [4] for the learning algorithm. During training, the algorithm learns the themes by identifying recurring measurement patterns, in particular, it identifies co-ocurrences between measurement bins throughout the event sample. Once the learning converges and the themes have been extracted, we build a likelihood-ratio defined by $\mathcal{L}(o_1,\ldots,o_N|\alpha) = \prod_{i=1}^{N} \frac{\mathcal{P}(o_i|\hat{\beta}_1(\alpha))}{\mathcal{P}(o_i|\hat{\beta}_2(\alpha))}$. The $\hat{\beta}_t$ are statistical estimators for the β_t's extracted from VI. The classifier is obtained by thresholding: for some suitable $c \in \mathbb{R}$, if $\mathcal{L}(o_1,\ldots,o_N|\alpha) > c$ then the event belongs to

[1] Topic modelling was first used in collider studies in [3] for quark/gluon jet discrimination.

theme $t = 1$, else it belongs to theme $t = 2$. This classifier is a function of the Dirichlet parameter α, and is better thought as a continuous 'landscape' of LDA classifiers. In principle there is no robust criteria for choosing one specific set of α's over another. Preliminary results given in ref. [2] suggest that a quantity known as *perplexity* can be used to precisely select the best α.

2. Latent Dirichlet allocation for new physics in 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[2] in $\mathcal{O}$. In order for LDA to learn from measurement co-ocurrence, 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 $j_0 \to j_1 j_2$. 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 Δ, defined in [6]. We then build a multi-dimensional space $\mathcal{O}$ spanned by different combination of these observables. Moreover, we also include a 'jet label' indicating to which jet in the event the measurement belongs to. In our experiments, we used for the hidden BSM benchmark a W'–ϕ 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 $pp \to W'$ production followed by the decay chain $W' \to W\phi \to WWW$, 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 $\mathcal{O} = \{j, \log k_T, \log(1/\Delta)\}$, where the integer $j = 1, 2, \ldots$ labels the 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. For the Dirichlet prior $\mathcal{D}(\omega|\alpha)$ controlling the theme mixings we fixed it to a very asymmetric shape $\alpha_0 =$ and $\alpha_2 =$. During training, this choice forces one multinomial theme ($t = 1$) to approximate the mixed data distribution which we know (a priori)

[2] Jet substructure observables that marginalize over all particles in the event, like e.g. N-subjettiness [5], fall into this category and are therefore not useful for LDA.

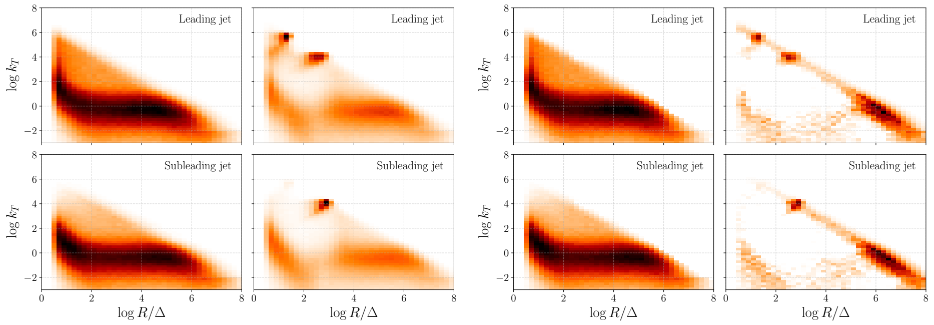

Figure 1: *Truth-level primary Lund planes for QCD background (1^{st} col.) and BSM signal (2^{nd} col.). Results for the first theme (3^{rd} col.) and second theme (4^{th} col.), learned from a two-theme LDA model trained with*100*k events composed of an unlabelled mixture of QCD and BSM at $s/b = 5\%$.*

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].

3. Conclusions

In conclusion, we have demonstrated that it is possible to build a simple 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 $\mathcal{O}$, called 'themes'. We trained a two-theme LDA model on the primary Lund plane from (unlabelled) a mixed dijet event sample with QCD background and BSM signal (from a $W' - \phi$ model) at $s/b = 5\%$. Our results indicate that LDA can successfully discover small BSM signals from unlabelled data.

References

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

[2] B.M. Dillon, D.A. Faroughy, J.F. Kamenik, M. Szewc, *JHEP* **10** (2020) 206 [2005.12319].

[3] E.M. Metodiev, J. Thaler, *Phys. Rev. Lett.* **120** (2018) 241602 [1802.00008].

[4] D.M. Blei, A.Y. Ng, M.I. Jordan, J. Lafferty, *JMLR* **3** (2003) 2003.

[5] J. Thaler, K. Van Tilburg, *JHEP* **03** (2011) 015 [1011.2268].

[6] F.A. Dreyer, G.P. Salam, G. Soyez, *JHEP* **12** (2018) 064 [1807.04758].