Looking for di-jet signals with LDA

Barry M. Dillon

in collaboration with:
D. Faroughy, J. Kamenik, & M. Swezc

ML4Jets, 16/01/2020
Goal: find rare signals in boosted di-jet events at colliders

Method: Latent Dirichlet Allocation (LDA)  
• LDA is an unsupervised machine learning technique
• Describes data using a probabilistic model
• Latent features in the data are described by probability distributions
• Inference algorithms are used to extract these latent distributions
The basic workflow

1. **Experiment (or a simulator)**
2. **Event data**
3. **Statistical model describing di-jet events**
4. **Inference: extract latent distributions**
5. **Classifier**
6. **Enriched sample for analysis**
Jet features

A single jet = a list of splittings, i.e. \( e_j = \{ f_1, f_2, \ldots, f_n \} \)

How do we represent the splittings?
Jet features

e_j = \{f_1, f_2, \ldots, f_n\}

Mass representation

Each splitting represents: \( j_0 \rightarrow j_1, j_2 \) \( m_{j_1} > m_{j_2} \)

Represent splittings as:

\[ f_i = [m_{j_0}, \frac{m_{j_1}}{m_{j_0}}] \]

[subjet mass, mass drop]

Jozef Stefan Institute, Ljubljana  
Barry M. Dillon
Jet features

\[ e_j = \{ f_1, f_2, \ldots, f_n \} \]

\[ R \]

De-clustering representation

Each splitting represents:

\[ j_0 \rightarrow j_1, j_2 \quad m_{j_1} > m_{j_2} \]

Represent splittings as:

\[ f_i = [R_{12}, k_T] \]

[angular splitting, kT splitting]

kT representation

Observables are binned. No concept of distance between splittings.
Jet features

The latent probability distributions are defined over this feature space.

QCD jets:
Jet features

The latent probability distributions are defined over this feature space.

Averaging over many QCD jets:
Jet features

The latent probability distributions are defined over this feature space.

Top jets:
Jet features

The latent probability distributions are defined over this feature space.

Averaging over many top jets:
A probabilistic model for jets

Probabilistic model: quantifies how likely it is that a certain event was generated, given some number of latent distributions, and hyper-parameters.

LDA is our probabilistic model, it is a mixed membership probabilistic model.

Mixed membership:
- Each event is generated by sampling from multiple latent distributions.
- The prevalence of each latent distribution in the sample is parameterised by hyper-parameters.

LDA for di-jets:
- For simplicity, two latent distributions (signal and background).
- Hyper-parameters incorporate class imbalance expected for rare signals.
LDA for di-jets

Mixed membership ⇒ an event can contain both background and signal features

Hyper-parameters enter through a Dirichlet prior, describing how likely it is to generate events with certain proportions of signal and background features.

A generative picture for LDA:

- Dirichlet prior
- Latent proportions
- Latent component
- Jet feature
- Latent distributions

Repeat for each feature
Repeat for each event
LDA for di-jets

1 - Draw a multinomial randomly from the prior distribution
2 - Draw a latent component from the multinomial
3 - Draw jet feature randomly from the appropriate latent distribution
4 - Repeat for each jet feature

A generative picture for LDA:
LDA for di-jets

\[ P(e | \bar{\alpha}, t_z) = \int d\theta \ \pi(\theta | \bar{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^{K} \theta(z) P(f_i | t_z) \]

A generative picture for LDA:

\[ \pi(\theta) \rightarrow \theta(z) \rightarrow z \rightarrow f_i \rightarrow P(f_i | t_z) \]
Inference

Data + LDA model \rightarrow \text{Extract latent distributions}

Technique: Variational Inference

Aims to maximise the probability that the LDA model produced the events.

\[ P(\mathcal{D}|\beta) \]

\[ P(\beta|\mathcal{D}) = \frac{P(\mathcal{D}|\beta)P(\beta)}{P(\mathcal{D})} \]

Uses: co-occurring features within the event.

Software: \textit{gensim} (Radim Řehůřek and Petr Sojka, 2010)

We have a GitHub: https://github.com/barrydillon89/LDA-jet-substructure

Jozef Stefan Institute, Ljubljana

Barry M. Dillon
Classification

With the extracted latent distributions, we then construct a classifier to filter and enrich the events in the sample.

Two methods:

Use LDA to infer the proportions of the different processes in each event:

\[ \hat{\theta}(\tilde{z}) = \arg\max_{\theta} P(e, \theta(\tilde{z})|\bar{\alpha}, t_z) \]

Construct alternative test-statistic, the likelihood ratio:

\[ L(e) = L(f_1, \ldots, f_{n_f}) = \prod_{i=1}^{n_f} \frac{P(f_i|t_S)}{P(f_i|t_B)} \]
The Dirichlet prior

\[ P(e|\vec{\alpha}, t_z) = \int d\theta \, \pi(\theta|\vec{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^{K} \theta(z) P(f_i|t_z) \]

\[ \rho = \alpha_1/\alpha_0 \]

\[ \Sigma = \alpha_0 + \alpha_1 \]

\[ \int_0^1 d\theta \, \pi(\theta|\alpha_0, \alpha_1) \left[ \theta P(f_i|t_B) + (1 - \theta) P(f_i|t_S) \right] = \frac{1}{1 + \rho} \left( P(f_i|t_B) + \rho P(f_i|t_S) \right) \]

\[ \rho \text{ represents the ratio of signal to background features in the sample} \]

\[ \Sigma \text{ controls the distribution of signal and background processes in each event} \]

\[ \Sigma \ll 1 \quad \rightarrow \quad \text{events mostly composed of a single process} \]

\[ \Sigma \gg 1 \quad \rightarrow \quad \text{events composed of a large mixture of processes} \]
How to fix the hyper-parameters?

Choose those that maximise the probability for the data to be generated. With unsupervised learning, we cannot use truth labels to choose the best model.

In practice, minimise:

$$\text{perplexity} = e^{-\frac{\log(P(\text{events}|\rho, \Sigma))}{N}}$$

Good model := hyper-parameters that minimise perplexity.

\[ P(e|\tilde{\alpha}, t_z) = \int d\theta \pi(\theta|\tilde{\alpha}) \prod_{i=1}^{n_f} \sum_{z=1}^{K} \theta(z)P(f_i|t_z) \quad \rho = \frac{\alpha_1}{\alpha_0} \]
\[ \Sigma = \alpha_0 + \alpha_1 \]
A test case: $pp \rightarrow W' \rightarrow \phi W \rightarrow WWW$ (Agashe et al, 2017&2018)

$m_{W'} = 3$ TeV $m_\phi = 400$ GeV $m_{jj} \in [2730, 3190]$ GeV

Background: QCD di-jets

Data: mixed unlabelled sample, $S/B = 1\%$
BSM jets

A test case: \( pp \rightarrow W' \rightarrow \phi W \rightarrow WWW \) \hspace{1em} (Agashe et al, 2017&2018)

\( m_{W'} = 3 \text{ TeV} \quad m_{\phi} = 400 \text{ GeV} \quad m_{jj} \in [2730, 3190] \text{ GeV} \)

**Background:** QCD di-jets \hspace{1em} **Data:** mixed unlabelled sample, \( S/B = 1\% \)

---

Jozef Stefan Institute, Ljubljana

Barry M. Dillon
A test case:

\[ pp \rightarrow W' \rightarrow \phi W \rightarrow WWW \]

\[ m_{W'} = 3 \text{ TeV} \]

\[ m_{\phi} = 400 \text{ GeV} \]

\[ m_{jj} \in [2730, 3190] \text{ GeV} \]

Background: QCD di-jets

Data: mixed unlabelled sample, \( S/B = 1\% \)
A test case:

\[ pp \rightarrow W' \rightarrow \phi W \rightarrow WWW \]

\[ m_{W'} = 3 \text{ TeV} \]

\[ m_{\phi} = 400 \text{ GeV} \]

\[ m_{jj} \in [2730, 3190] \text{ GeV} \]

**Background:** QCD di-jets

**Data:** mixed unlabelled sample, \( S/B = 1\% \)

**Potential drawback:** does not naturally incorporate a bump hunt

Jozef Stefan Institute, Ljubljana

Barry M. Dillon
Fixing the hyper-parameters

We scan over the hyper-parameters:

- **lots of local minima,** close to models with best AUC and best rejection rate at fixed mis-tag.
- **global minimum at vanishing rho.**
Fixing the hyper-parameters

High-performance regions match those of (local) minimum perplexity.

The prior/hyper-parameters allow for a focusing on small S/B.
BSM jets in the Lund plane

A test case: \( pp \rightarrow W' \rightarrow \phi W \rightarrow WWW \)

\( m_{W'} = 3 \text{ TeV} \quad m_\phi = 400 \text{ GeV} \quad m_{jj} \in [2730, 3190] \text{ GeV} \)

Background: QCD di-jets

Data: mixed unlabelled sample, \( S/B = 1\% \)
BSM jets in the Lund plane

A test case: \( pp \rightarrow W' \rightarrow \phi W \rightarrow WWW \)
\( m_{W'} = 3 \text{ TeV} \quad m_\phi = 400 \text{ GeV} \quad m_{jj} \in [2730, 3190] \text{ GeV} \)

Background: QCD di-jets

Data: mixed unlabelled sample, \( S/B = 1\% \)

Jozef Stefan Institute, Ljubljana

Barry M. Dillon
BSM jets in the Lund plane

A test case: \( pp \rightarrow W' \rightarrow \phi W \rightarrow WWW \)

\( m_{W'} = 3 \text{ TeV} \quad m_\phi = 400 \text{ GeV} \quad m_{jj} \in [2730, 3190] \text{ GeV} \)

Background: QCD di-jets  
Data: mixed unlabelled sample, \( S/B = 1\% \)

LDA separates soft and hard physics!
Conclusions

LDA can extract rare signals from data, with no a-priori knowledge of the signal

Key component: the Dirichlet prior
- probabilistic model naturally describes class imbalance in the data

LDA could form part of a more complete analysis
- requires additional analysis to quantify the significance of the uncovered signals

Currently, the technique is only suitable for di-jet searches.
( extending this is a work in progress..)

More generally, these results obtained using LDA point towards possible uses of probabilistic models for unsupervised searches at colliders.
LHC Olympics: our attempt..

Procedure:

- split the dataset into overlapping invariant mass bins
- train multiple LDA models on each bin, extracting underlying latent distributions
- select candidate signal hypothesis, i.e. a latent distribution
- perform a bump hunt on the data after cutting using LDA classifier

Background estimation:

- use the uncut invariant mass distribution as a background template
- fix the total number of background events using a control region
- systematic errors estimated using the background sample
LHC Olympics: our attempt..

Dataset 1:
- our method assumes a di-jet resonance, including only the two leading jets in the analysis

The extracted latent distributions in invariant mass bins 2-3 TeV and 2.5-3.5 TeV:

Background

\[ j_1 \quad m_{j_1} > m_{j_2} \quad j_2 \]
LHC Olympics: our attempt.

Dataset 1:

- our method assumes a di-jet resonance, including only the two leading jets in the analysis

The extracted latent distributions in invariant mass bins 2-3 TeV and 2.5-3.5 TeV:

Signal

\[ j_1 \quad m_{j_1} > m_{j_2} \quad j_2 \]
LHC Olympics: our attempt..

Dataset 1:

- our method assumes a di-jet resonance, including only the two leading jets in the analysis

Bumphunter with this hypothesis on dataset1, and on the background sample:

Dataset 1 (3.8σ)  Dataset BG (1.8σ)

Jozef Stefan Institute, Ljubljana  Barry M. Dillon
LHC Olympics: our attempt...

Dataset 1:

- our method assumes a di-jet resonance, including only the two leading jets in the analysis

Bumphunter with this hypothesis on dataset1, and on the background sample:

Dataset 1 (3.8$\sigma$)  
Dataset BG (1.8$\sigma$)