Detecting hidden Patterns inside jets with probabilistic models

Darius A. Faroughy

University of Zurich[UZH]

ML4Jets, Heidelberg, July 8th 2021

Overview

- Build step by step a probabilistic model for collider events

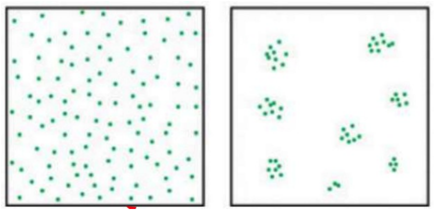
- Application: extract BSM from dijets using the lund Jet plane.

Based on:	1904.04200	- Jernej F. Kamenik
	2005.12319	- Barry Dillon
	2012.08579	- Manuel Swezc

Collider events as point patterns

- Collider event: set of observations, or measurements. $e = \{o_1, \ldots, o_n\}$

$o_i \in \mathcal{O}$ ·········· space of observables



random distribution of points $$e(o) = \sum_{i=1}^{n} \delta^{(k)}(o - o_i)$$

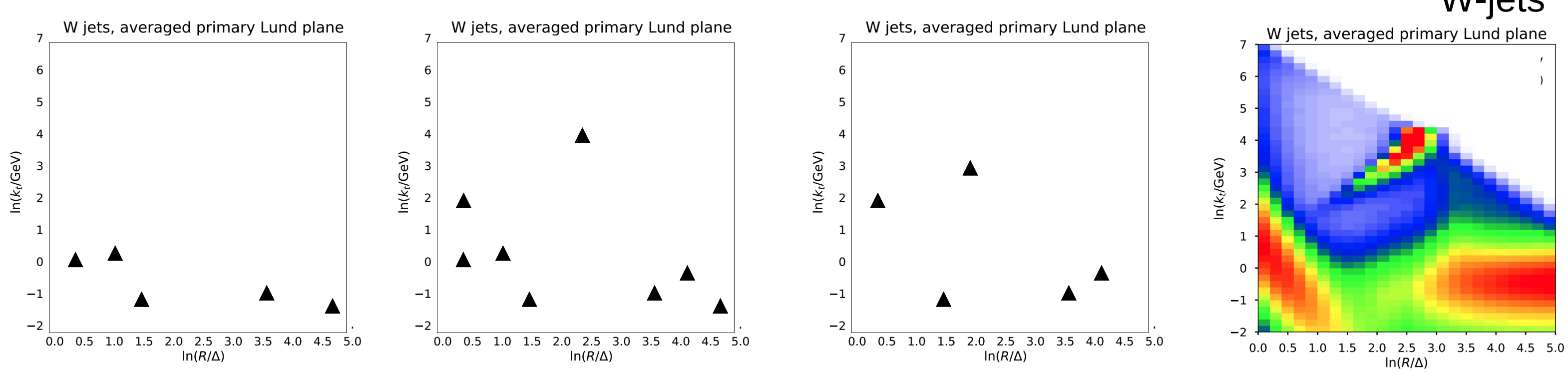
- We could model events with an underlying stochastic point process in $\mathcal{O}$ e.g. Non-homogenous Poisson process

$$N(\mathcal{R}) \equiv \#\{o \in \mathcal{R} \subset \mathcal{O}\} \quad \Longleftarrow \quad N(\mathcal{R}) \sim \mathrm{Poisson}(\lambda_{\mathcal{R}})\,, \quad \lambda_{\mathcal{R}} = \int_{\mathcal{R}} \prod^{k} \mathrm{d}\mathcal{O}\ \mu(\mathcal{O}_1, \ldots, \mathcal{O}_k)$$

- Point patterns can be sparse and give rise to irregular patterns.

example: primary Lund jet plane Dreyer et al (2018)

W-jets



Probabilistic models for events

- Modelling events with a stochastic point porcesses seems cool, but too complicated...

$\mathcal{P}(e) = \mathcal{P}(o_1, o_2, o_3, \cdots)$ How can we model this joint probability in a simple way?

felixible and tractable

Goal: build a probabilistic model for event classification (not event generators)

- Our “model-buiding” assumptions

I. Exchangeability of observations.

II. Discretization of the observable space.

III. Multiple *latent* 'classes' can contribute to each event.

I. Exchangeability

- Exchangeability of event observations (i.e. Permutation symmetry)

$$\mathcal{P}(o_1, o_2, o_3, \cdots) = \mathcal{P}(o_{\pi(1)}, o_{\pi(2)}, o_{\pi(3)}, \cdots) \qquad \pi \in \mathcal{S} \text{ permutation group}$$

De Finnetti's representation theorem (1931):

A sequence of observations is exchangeable if and only if there exists a latent variable ω *such that:*

$$\mathcal{P}(o_1, o_2, \ldots) = \int_\Omega \mathrm{d}\omega \, P(\omega) \prod_{i=1}^{n} p(o_i|\omega)$$

Latent space   Prior   Likelihood

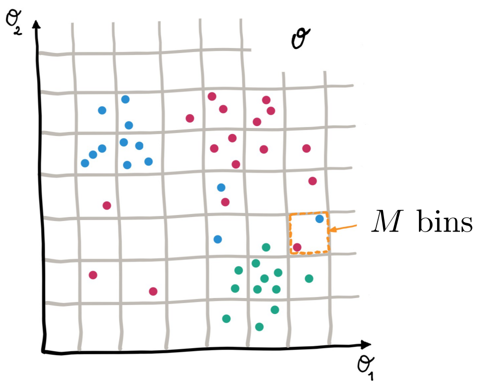
- Obervations are considered conditionally independent given a latent variable $\omega \in \Omega$
- Exchangeable not to be confused with independent and identically distributed (iid) !!
- We will need extra model-building assumptions to fix *p*, *P*, *omega*

II. Discretization

- What to take for $p(o|\omega)$?

$$\mathcal{P}(o_1, o_2, \dots) = \int_\Omega \mathrm{d}\omega\, P(\omega) \prod_{i=1}^{n} p(o_i|\omega)$$

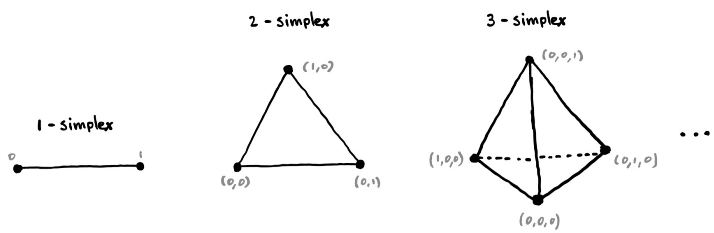
- Multinomial distributions for binned data:



$$o \sim \text{Multinomial}(\beta)$$

$$\beta = (\beta_1, \dots, \beta_M)$$

$$\begin{cases} \sum_{m=1}^{M} \beta_m = 1 \\ 0 \le \beta_m \le 1 \end{cases}$$



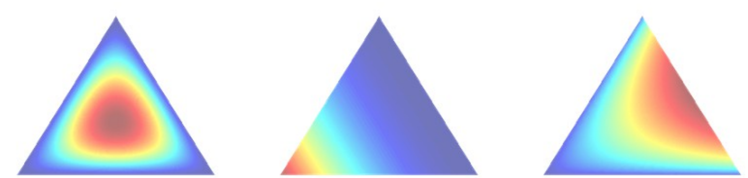
M-dimensional Simplex for $\beta = (\beta_1, \dots, \beta_M)$

- We introduce a prior for $\beta = (\beta_1, \dots, \beta_M)$

Dirichlet distribution

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

$\eta = (\eta_1, \cdots, \eta_M)$ shape parameter



III. Latent event classes

What to take for ther latent variable?

$$\mathcal{P}(o_1, o_2, \ldots) = \int_\Omega \mathrm{d}\omega \, P(\omega) \prod_{i=1}^{n} p(o_i|\omega, \beta)$$

- Event observations are generated from **multiple** latent Multinomial distributions over $\mathcal{O}$

$$p(o|\beta_t) \quad t = 1, \ldots, T$$

Mixture of multinomials:

$$p(o|\omega, \beta) = \sum_{t=1}^{T} p(t|\omega) \, p(o|\beta_t)$$

"Themes" or "Topics"

** Terminology from Natural Language Processing*

'Theme' mixing parameter $\omega = (\omega_1, \ldots, \omega_t)$

$$p(t|\omega) = \omega_t \qquad 0 \leq \omega_t \leq 1, \; \sum \omega_t = 1$$

- Latent space T-dimensional simplex (space of theme mixings)

The prior P is a Dirichlet distribution

$$D(\omega|\alpha) \qquad \alpha = (\alpha_1, \cdots, \alpha_T)$$

$$\mathcal{P}(o_1, o_2, \ldots) = \int_\Omega \mathrm{d}\omega \, P(\omega) \prod_{i=1}^{n} p(o_i|\omega, \beta)$$

Latent Dirichlet Allocation (LDA)

$$\mathcal{P}(e|\alpha) = \int_{\Omega_T} \mathrm{d}\omega\, D(\omega|\alpha) \prod_{i=1}^{n} \left(\sum_{t=1}^{T} p(t|\omega)\; p(o_i|\beta_t) \right)$$

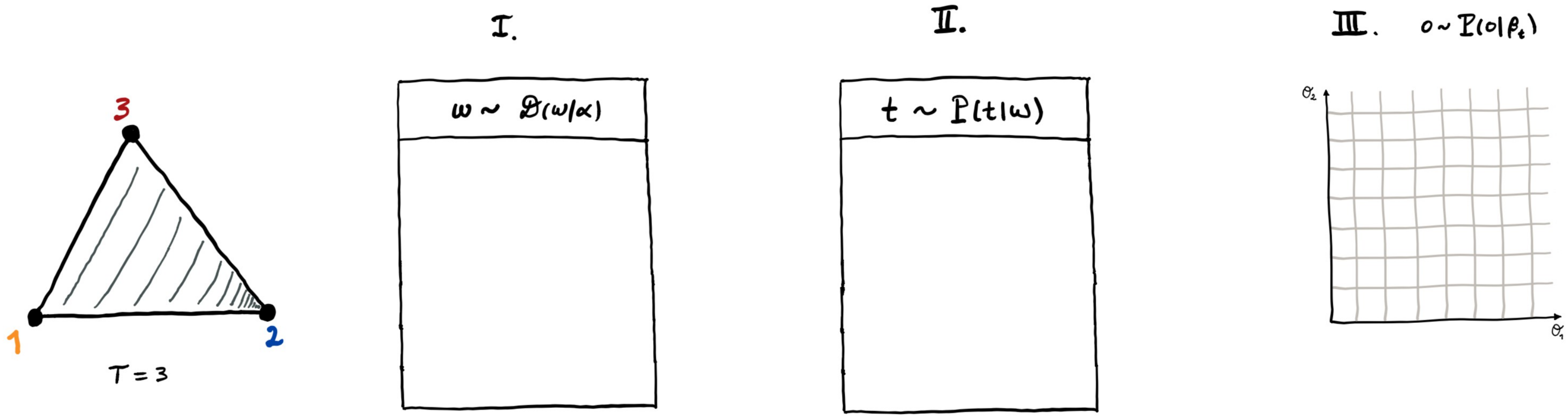
$$e = \{o_1, \cdots, o_n\}$$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan,
Journal of Machine Learning Research, 3 (2003) 993-1022.

over 30K citations!

- Generative process (T = 3):



$\mathcal{D}(\omega|\alpha_0, \alpha_1, \alpha_2)$

Latent Dirichlet Allocation (LDA)

$$\mathcal{P}(e|\alpha) = \int_{\Omega_T} \mathrm{d}\omega \, D(\omega|\alpha) \prod_{i=1}^{n} \left(\sum_{t=1}^{T} p(t|\omega) \; p(o_i|\beta_t) \right)$$

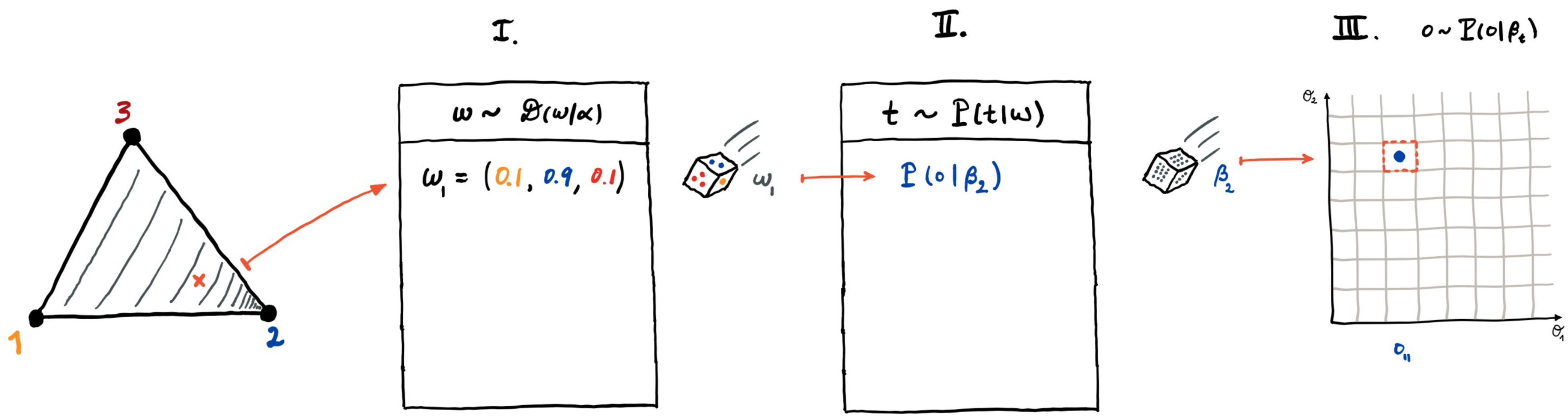
$$e = \{o_1, \cdots, o_n\}$$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan,
Journal of Machine Learning Research, 3 (2003) 993-1022.

over 30K citations!

- Generative process (T = 3):



Latent Dirichlet Allocation (LDA)

$$\mathcal{P}(e|\alpha) = \int_{\Omega_T} \mathrm{d}\omega\, D(\omega|\alpha) \prod_{i=1}^{n} \left(\sum_{t=1}^{T} p(t|\omega)\; p(o_i|\beta_t) \right)$$

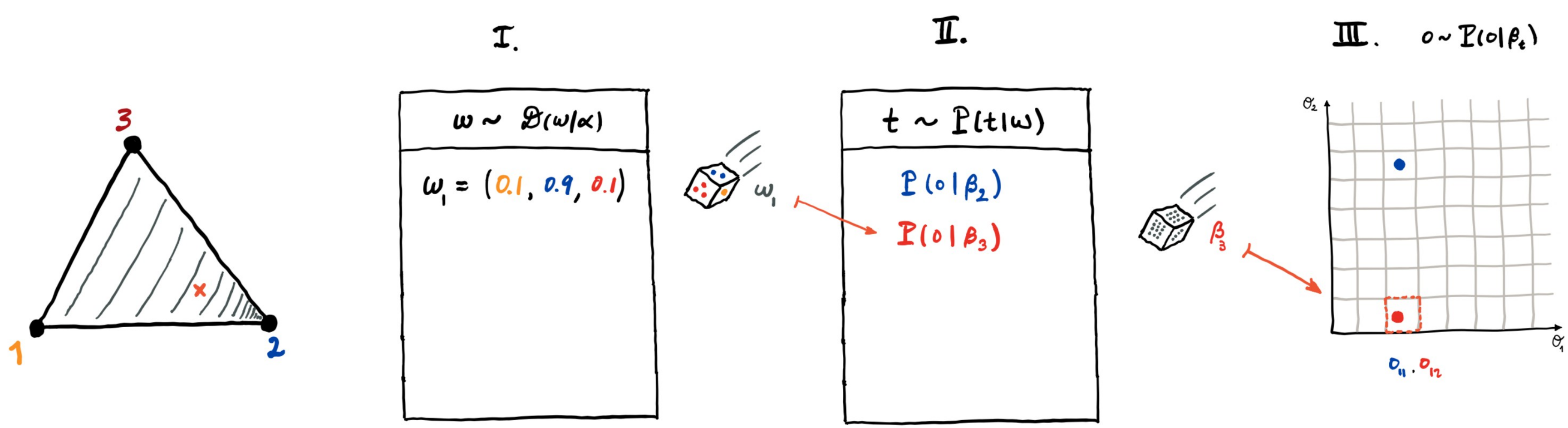
$$e = \{o_1, \cdots, o_n\}$$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan,
Journal of Machine Learning Research, 3 (2003) 993-1022.

over 30K citations!

- Generative process (T = 3):



Latent Dirichlet Allocation (LDA)

$$\mathcal{P}(e|\alpha) = \int_{\Omega_T} \mathrm{d}\omega\, D(\omega|\alpha) \prod_{i=1}^{n} \left(\sum_{t=1}^{T} p(t|\omega)\; p(o_i|\beta_t) \right)$$

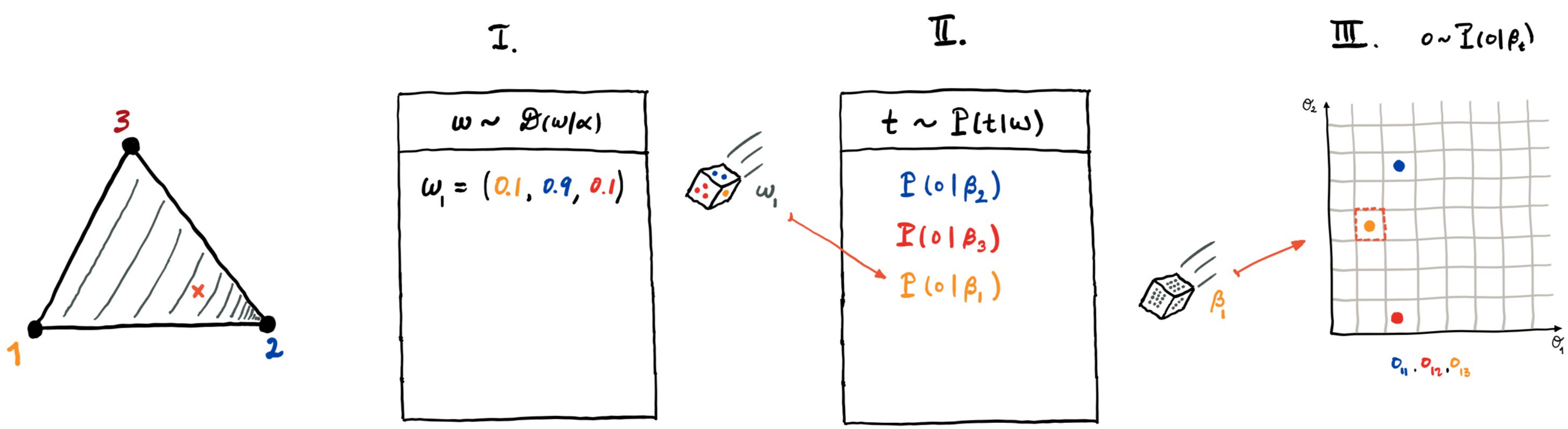
$$e = \{o_1, \cdots, o_n\}$$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan,
Journal of Machine Learning
Research, 3 (2003) 993-1022.

over 30K citations!

- Generative process (T = 3):



Latent Dirichlet Allocation (LDA)

$$\mathcal{P}(e|\alpha) = \int_{\Omega_T} \mathrm{d}\omega \, D(\omega|\alpha) \prod_{i=1}^{n} \left(\sum_{t=1}^{T} p(t|\omega) \; p(o_i|\beta_t) \right)$$

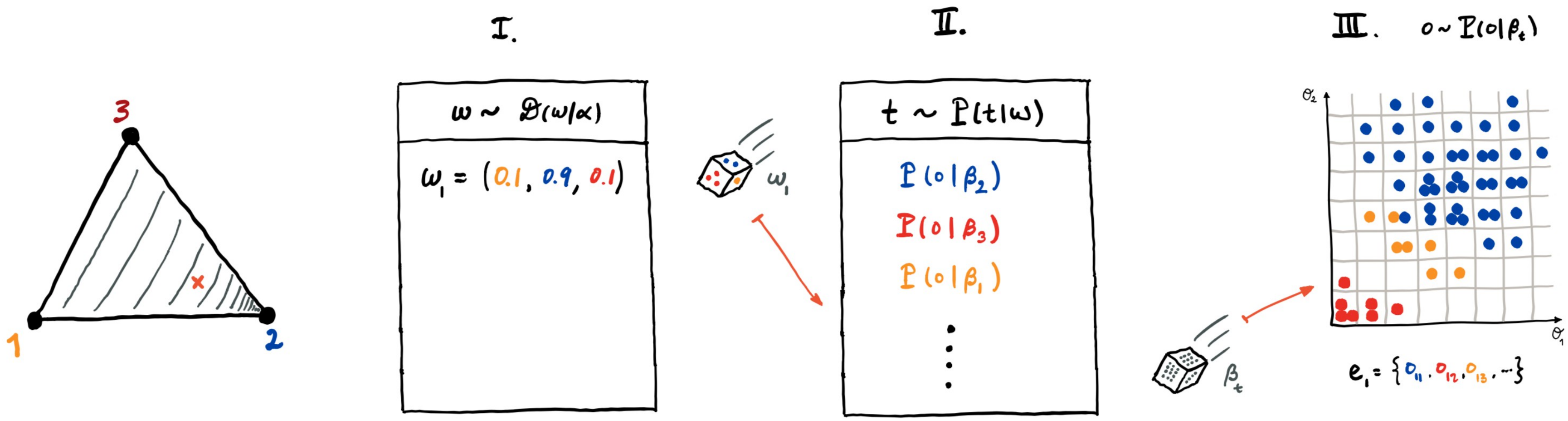
$$e = \{o_1, \cdots, o_n\}$$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan,
Journal of Machine Learning Research, 3 (2003) 993-1022.

over 30K citations!

- Generative process (T = 3):



Latent Dirichlet Allocation (LDA)

$$\mathcal{P}(e|\alpha) = \int_{\Omega_T} \mathrm{d}\omega\, D(\omega|\alpha) \prod_{i=1}^{n} \left(\sum_{t=1}^{T} p(t|\omega)\; p(o_i|\beta_t) \right)$$

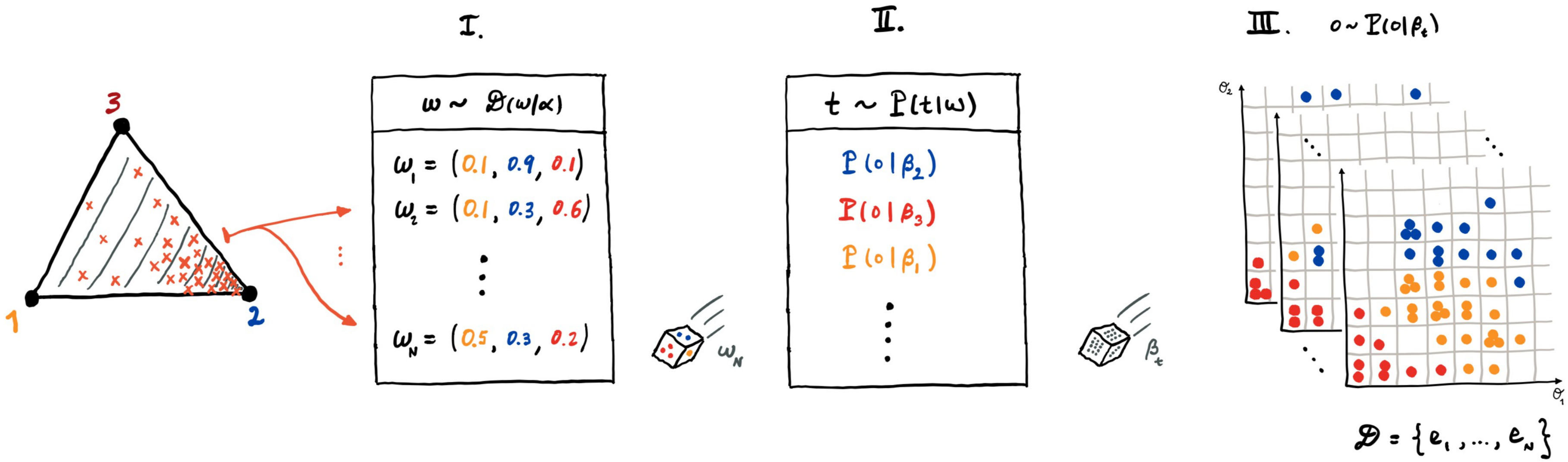
$$e = \{o_1, \cdots, o_n\}$$

Latent Dirichlet Allocation (LDA)

Blei, Ng, Jordan,
Journal of Machine Learning Research, 3 (2003) 993-1022.

over 30K citations!

- Generative process (T = 3):



- LDA is a mixed-membership model

LDA classifier

- Bayesian inference:

$$p(\omega,t,\beta|e,\alpha,\eta)=\frac{p(\omega,t,\beta,e|\alpha,\eta)}{p(e|\alpha,\eta)}$$

$$D(\omega|\alpha)\left(\prod_{t=1}^{T}D(\beta_t|\eta)\right)\left(\prod_{j=1}^{N}\prod_{i=1}^{n}p(t_i|\omega)\,p(o_i|t_i,\beta)\right)$$

$$\sum_t\int \mathrm{d}\omega\mathrm{d}\beta\,p(\omega,t,\beta,e|\alpha,\eta)$$

"evidence" intractable!

- Variational inference: inference problem $\longrightarrow$ optimization problem.

Blei, et al. Journal of the American Statistical Association 112 (Feb, 2017) 859–877

- For most applications we wish to classify events into two categories (e.g. signal & background)

We focus on Two-theme LDA models $T=2$

- LDA classifer: likelihood ratio of the 2 themes.

$$L(e|\alpha):=\prod_{o\in e}\frac{p(o\,|\hat{\beta}_2)}{p(o\,|\hat{\beta}_1)}\qquad\begin{cases}L(e|\alpha)>c & \Rightarrow\quad e\in\mathcal{C}_1\\ L(e|\alpha)\leq c & \Rightarrow\quad e\in\mathcal{C}_2\end{cases}$$

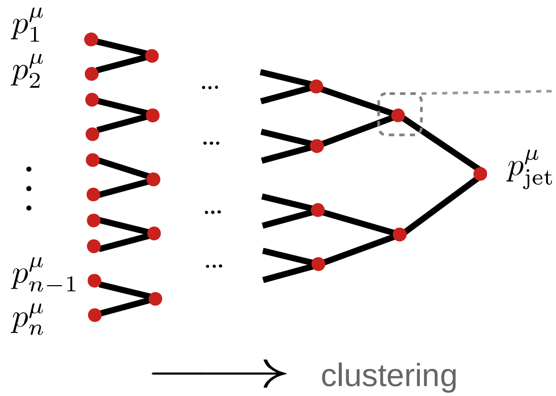
threshold

We actually have an infinite landscape of LDA classifiers...

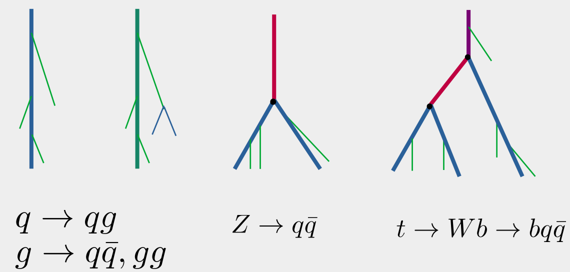
Jet clustering history

- Jet clustering history is sensitive to the underlying physics.

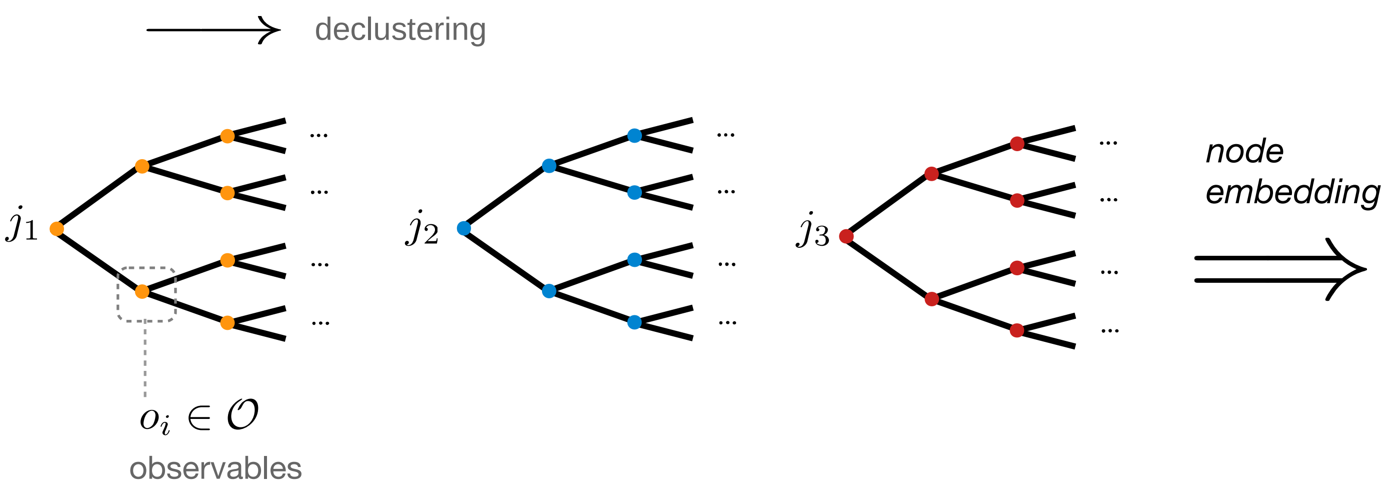
  binary tree: proxy for the radiation pattern during jet formation.

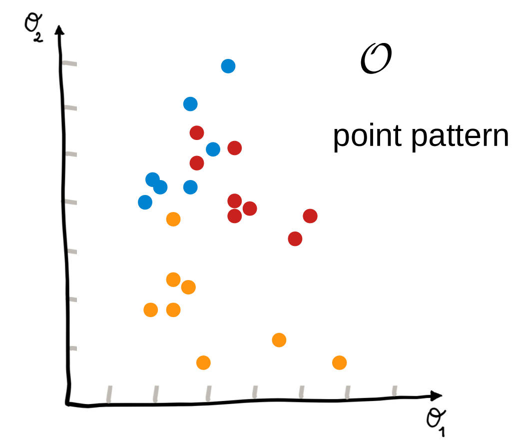


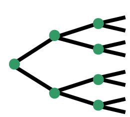
Proxy for $1 \to 2$ splitting or massive decay



- “De Finnetti” representation for jets:





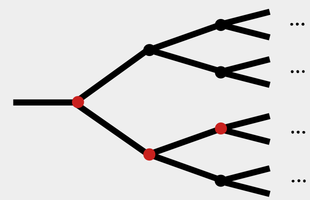
$\mathcal{P}($  $) \simeq \int_\Omega d\omega\, \mathcal{P}(\omega) \prod_{\bullet \in j} \mathcal{P}(\bullet\,|\omega)$ De Finnetti

Jet observables & data samples

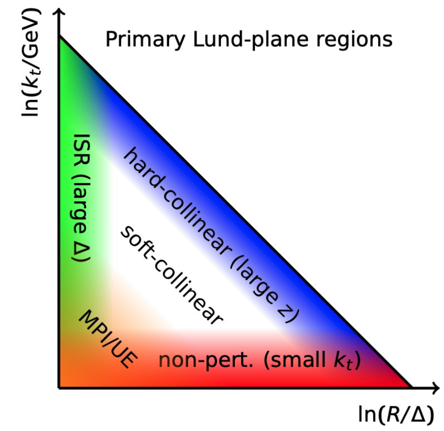
- Train LDA on full events with Lund observables:

$$\mathcal{O}_{\text{Lund}} = \left\{ \ell, \log(k_t), \log\left(\frac{1}{\Delta}\right) \right\}$$

Primary Lund plane
Dreyer et al (2018)



Label indicating to which jet the measurement belongs too, mass-ordered jets.

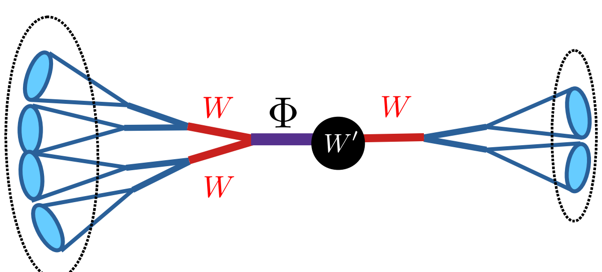


- Dijet data: unlabeled mixture of QCD (b) + BSM (s)

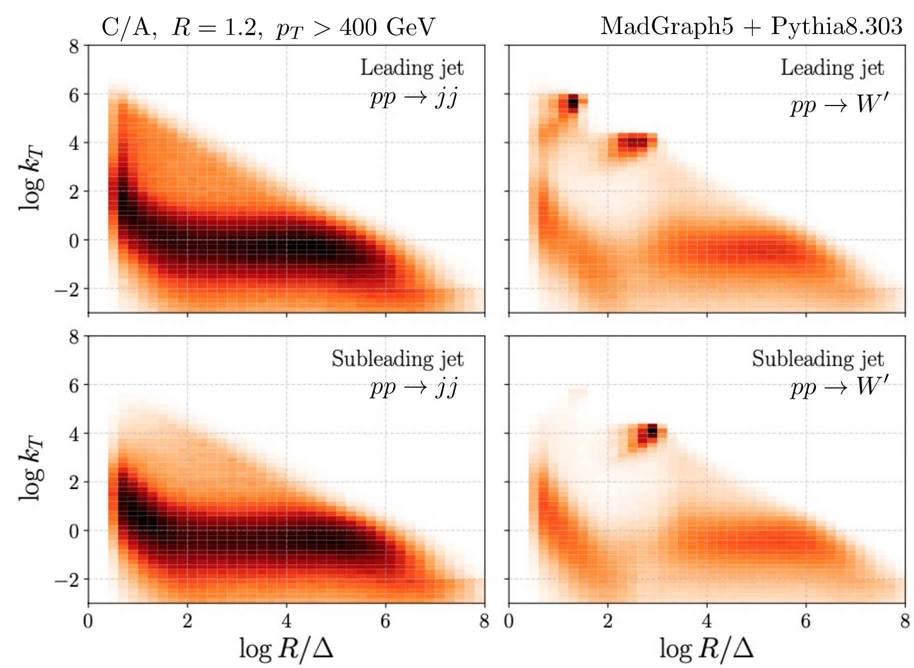
$$s/b \ll 1$$

$$pp \to W' \to \Phi W^{\pm}$$
$$\Phi \to W^{\pm} W^{\mp}$$

$$m_{W'} = 3\,\text{TeV}$$
$$m_{\Phi} = 400\,\text{GeV}$$



Agashe et al (2017)



truth-level

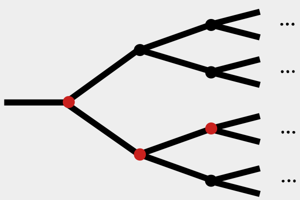
- Training performed with `Gensim` package (python)

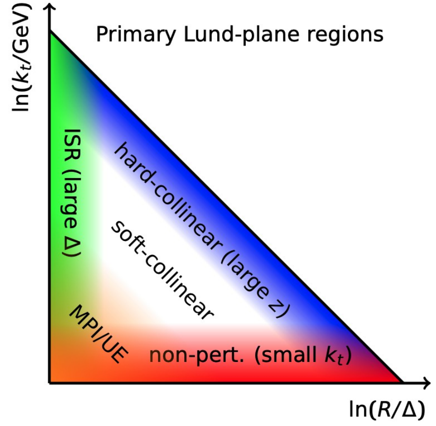
Jet observables & data samples

- Train LDA on full events with Lund observables:

$$\mathcal{O}_{\text{Lund}} = \left\{ \ell, \log(k_t), \log\left(\frac{1}{\Delta}\right) \right\}$$

Primary Lund plane
Dreyer et al (2018)





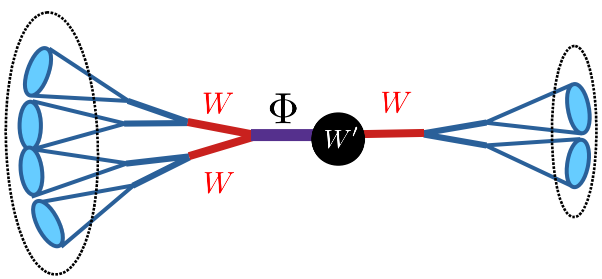
Label indicating to which jet the measurement belongs too, mass-ordered jets.

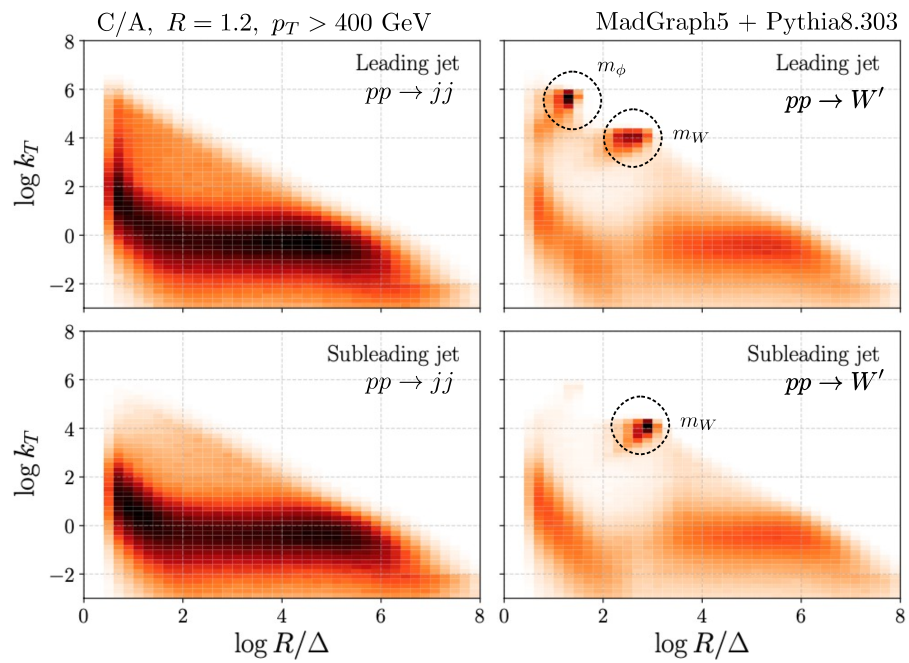
- Dijet data: unlabeled mixture of QCD (b) + BSM (s)

$$s/b \ll 1$$

$$pp \to W' \to \Phi W^{\pm}$$
$$\Phi \to W^{\pm} W^{\mp}$$

$$m_{W'} = 3\,\text{TeV}$$
$$m_{\Phi} = 400\,\text{GeV}$$





truth-level

- Training performed with `Gensim` package (python)

Extracting rare signals with LDA

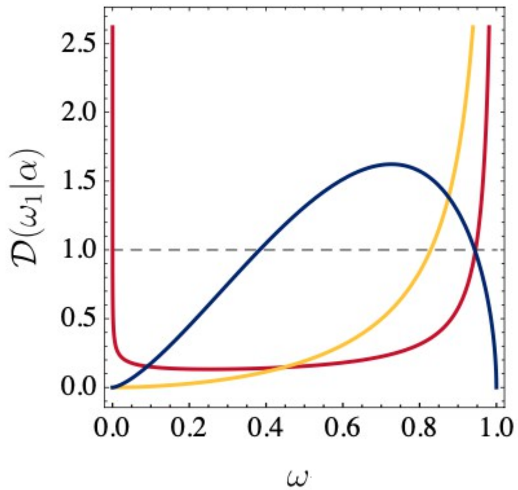
- if LDA works well:
  - $p(o|\beta_1)$ 1st theme: should include most QCD features
  - $p(o|\beta_2)$ 2nd theme: should include most signal features (e.g. BSM)

- Which Dirichlet prior for the theme mixture? $\omega \sim D(\omega|\alpha_1, \alpha_2)$

Reparametrization:

$$(\alpha_1, \alpha_2) \rightarrow (\rho, \Sigma) \qquad \begin{cases} \Sigma = \alpha_1 + \alpha_2 \\ \rho = \dfrac{\alpha_2}{\alpha_1} \end{cases}$$

Controls the shape asymmetry



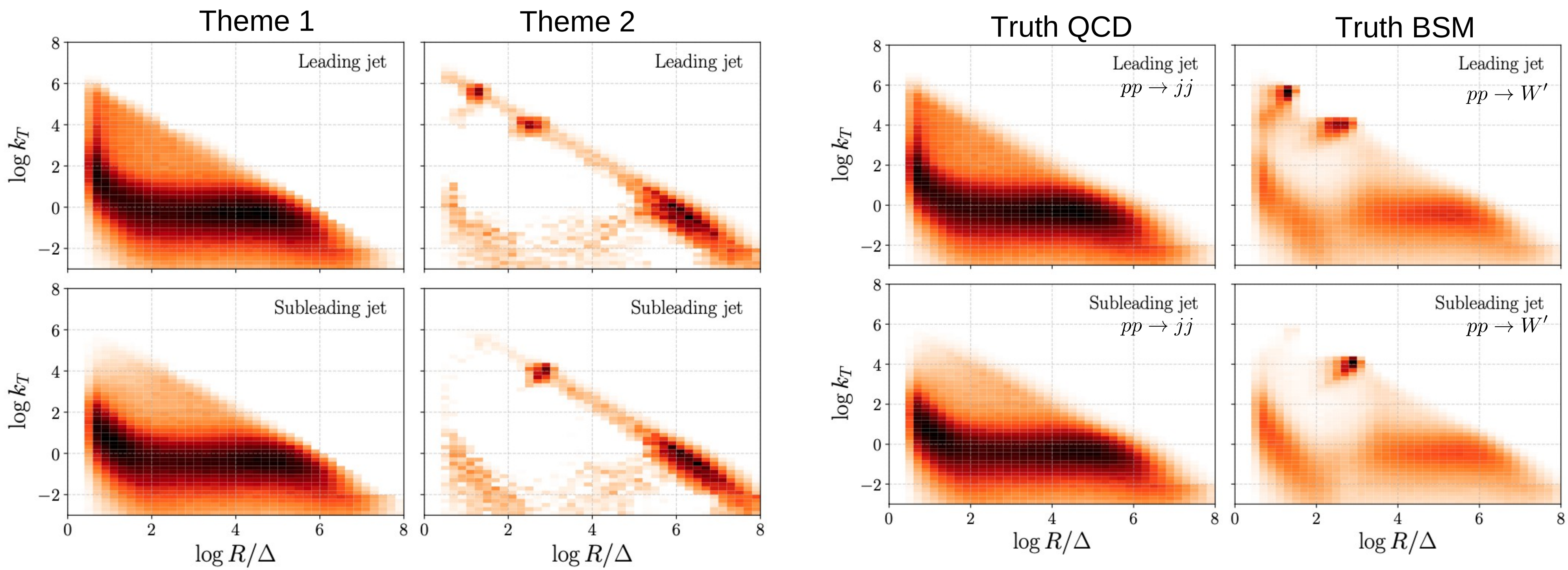
Prior with asymmetric shape $\Longleftarrow$ We want to discover *rare* signal in a data sample dominated by QCD $b \gg s$

$\rho \sim 0.1$ usually works...

Uncovering BSM physics from the Lund plane

$$pp \to W' \to \Phi W^{\pm},\ \Phi \to W^{\pm}W^{\mp}$$

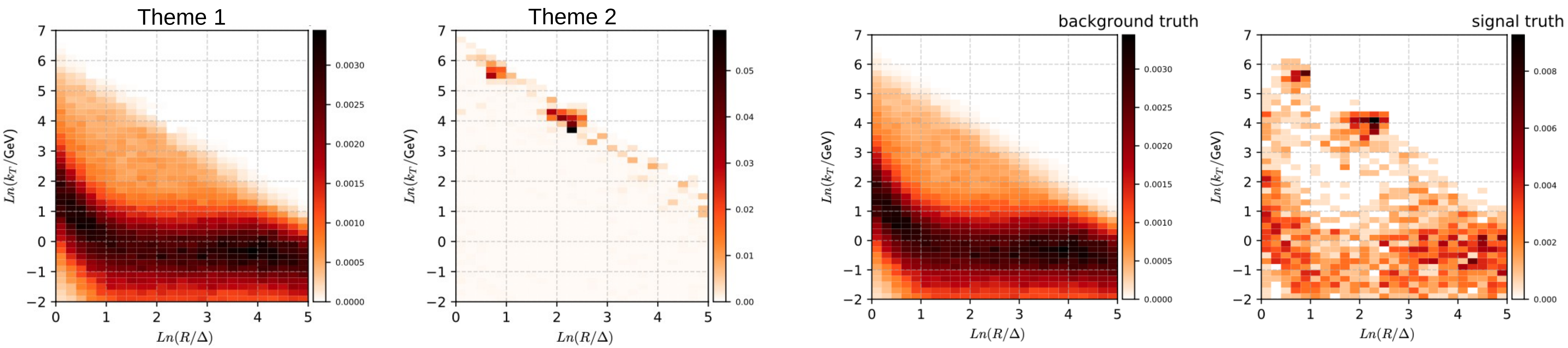
~ 100k events $s/b = 0.01$ $\mathcal{O}_{\mathrm{Lund}} = \left\{\ell, \log(k_t), \log\left(\frac{R}{\Delta R}\right)\right\}$ $(\rho, \Sigma) = (0.1, 1)$



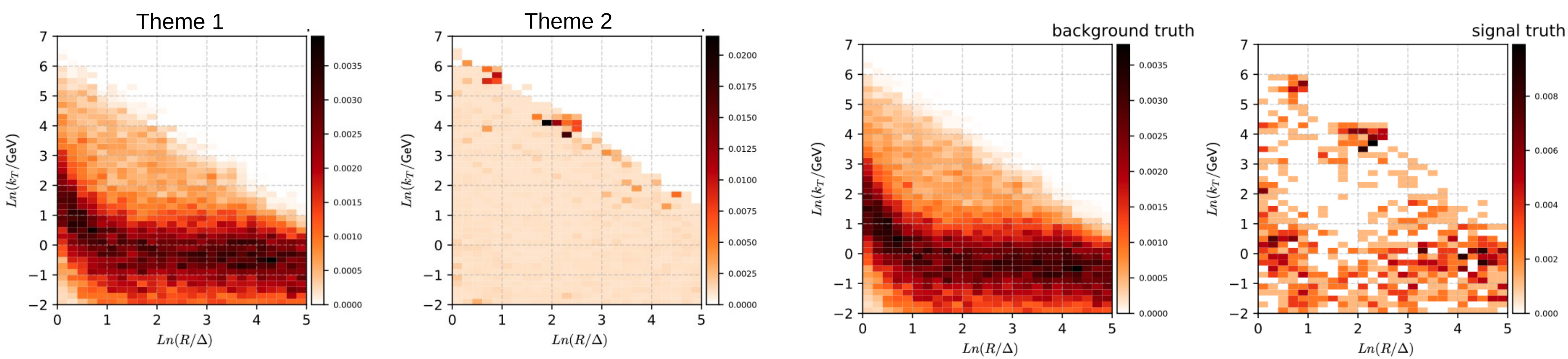
LDA discovers the hard/colinear splittings of the massive resonance decays in the Primary lund plane.

- What if we train on much less events?

$\left\{ \begin{array}{l} \text{10k QCD events} \\ \text{100 signal events} \end{array} \right.$ $s/b = 0.01$ $\mathcal{O}_{\text{Lund}} = \left\{ \log(k_t), \log\left(\frac{R}{\Delta R}\right) \right\}$ $(\rho, \Sigma) = (0.0009, 5.2)$



$\left\{ \begin{array}{l} \text{1600 QCD events} \\ \text{40 signal events} \end{array} \right.$ $s/b = 0.025$ $\mathcal{O}_{\text{Lund}} = \left\{ \log(k_t), \log\left(\frac{R}{\Delta R}\right) \right\}$ $(\rho, \Sigma) = (0.09, 4.0)$



- LDA works well with small data samples!

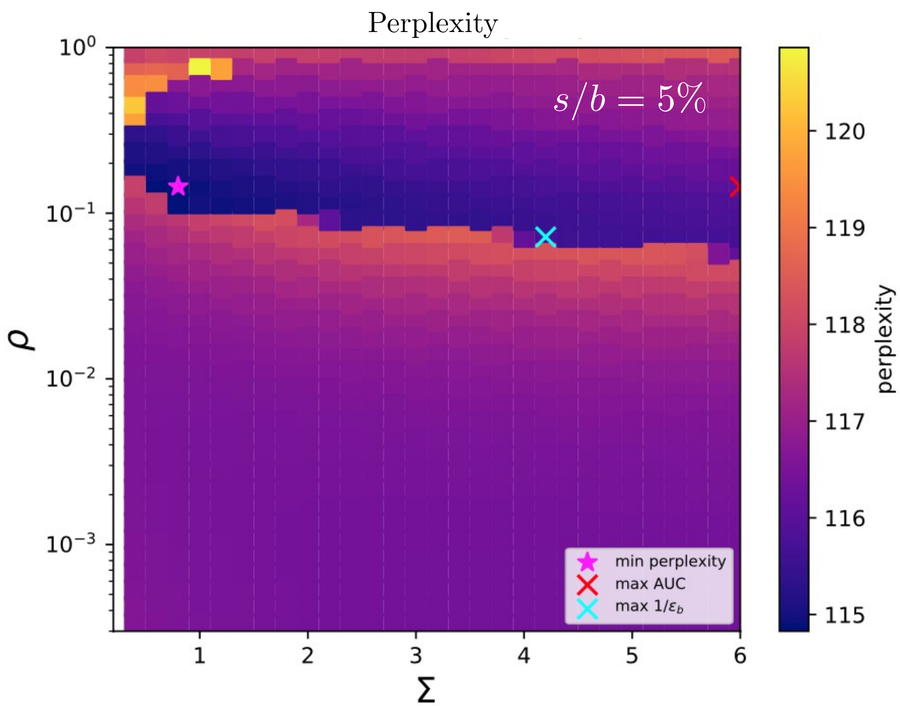
Perplexity & the landscape of LDA classifiers

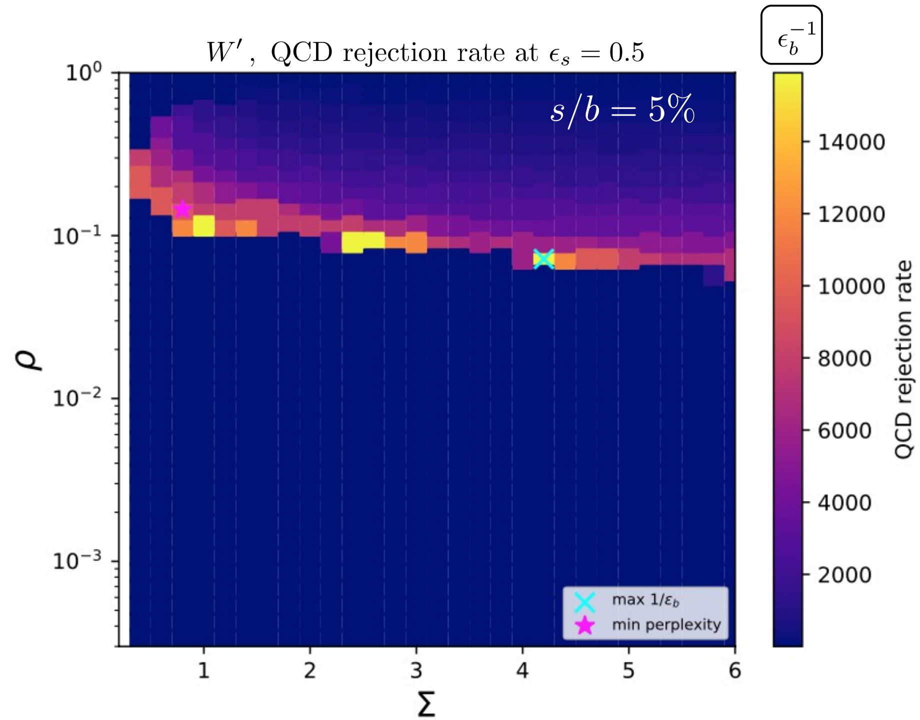
- We need a criteria for selecting from all possible LDA models the one with the “best” performance.

  i.e. a statistical goodness-of-f t test for the model.

- Perplexity measures how well the probabilsitic model f ts the data sample. Good models have a lower perplexity score

- Trained ~1000 LDA models in the (ρ, Σ) - plane:





Summary

- We showed that simple generative probabilistic models can be used to describe collider events represented as point patterns.

- Under very broad assumptions we arrived to the Latent Dirichlet Allocation (LDA) model.

- We demonstrated that LDA trained on the Lund plane can be used to uncover heavy resonances in dijet samples in a fully unsupervised way.

- LDA is just one of many possible probabilistic model...

- It can used as a building block for more complex models which could be useful for jet physics.

- Much more to explore! e.g. Bump hunt using LDA trained on the Lund jet planes or other observables.

IArxiv beta

About us

IArxiv beta

Developed by:

Ezequiel Alvarez (ICAS)
Daniel de Florian (ICAS)
Federico Lamagna (CAB CNEA)
Cesar Miquel (Easytech)
Manuel Szewc (ICAS)

Powered by:

icas.unsam.edu.ar
easytechgreen.com
unsam.edu.ar

iarxiv.org

iarxiv.org

Uses LDA to sort daily arxiv papers by learning your topic preferences

Thank You!

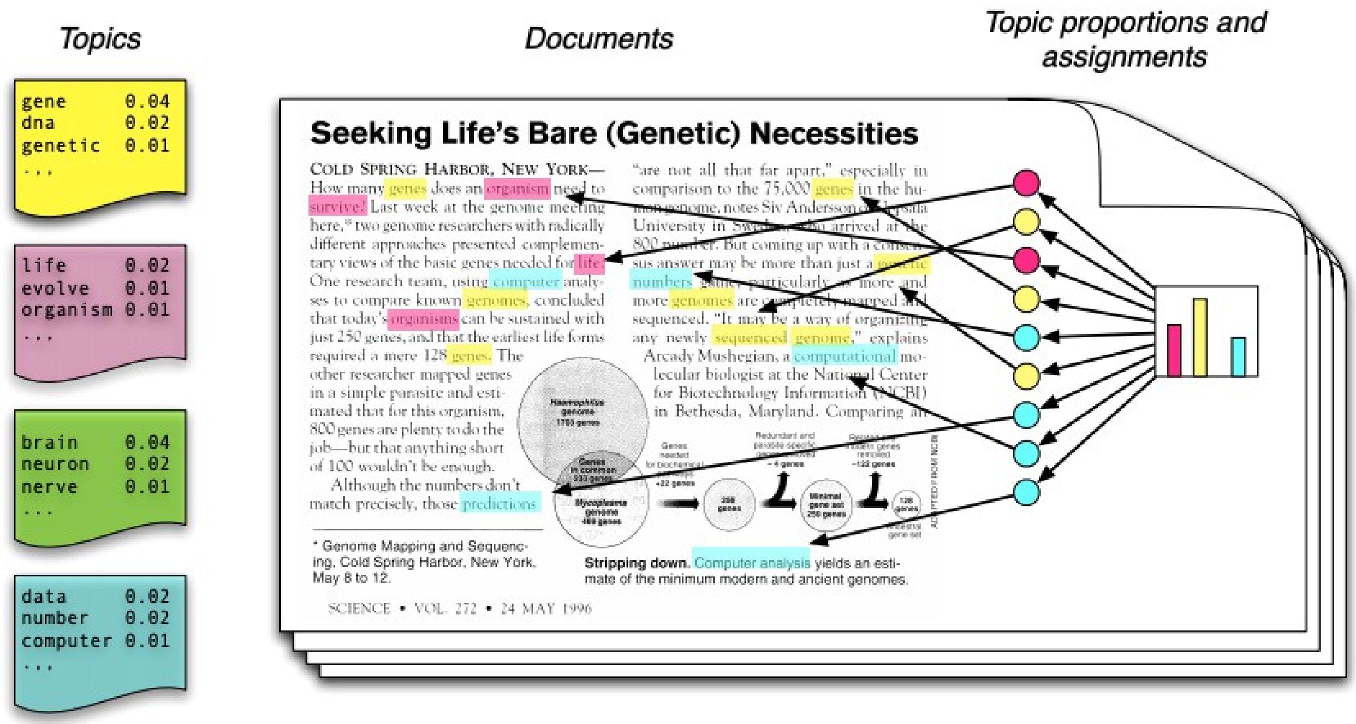
Back-up material

Topic Models for texts

- LDA conceived for Natural Language Processing

Blei, Ng, Jordan,
Journal of Machine Learning
Research, 3 (2003) 993-1022.

over 30K citations!



Fully Unsupervised ML

- LDA uncovers the hidden topics in a collection of documents
- Documents: unstructured collection of words

Bag-of-words (Bow)

Topics: distributions over vocabulary

- Text / Collider Physics correspondance:

corpus -------------- event samples
document ---------- event
vocabulary --------- space of observables
word ----------------- bin
topic ----------------- histogram

Learning the latent variables

- The posterior for an event:

$$p(\omega, t, \beta | e, \alpha, \eta) = \frac{p(\omega, t, \beta, e | \alpha, \eta)}{p(e|\alpha, \eta)}$$

$$D(\omega|\alpha)\left(\prod_{t=1}^{T} D(\beta_t|\eta)\right)\left(\prod_{j=1}^{N}\prod_{i=1}^{n} p(t_i|\omega)\, p(o_i|t_i, \beta)\right)$$

$$\sum_t \int \mathrm{d}\omega \mathrm{d}\beta\, p(\omega, t, \beta, e|\alpha, \eta)$$ "evidence"

Intractable integral!

- Variational inference: inference problem $\longrightarrow$ optimization problem

Propose a simple family of distributions $\mathcal{Q}$

$$q^* = \underset{q \,\in\, \mathcal{Q}}{\operatorname{argmin}}\, d_{\mathrm{KL}}[q, p]$$

posterior

Kullback-Liebler divergence $d_{\mathrm{KL}}[q,p] = \langle \log q \rangle - \langle \log p \rangle + \underbrace{\log p(e)}$

Log-evidence..... still intractable

Instead we maximize evidence lower-bound (ELBO):

$$q^* = \underset{q \,\in\, \mathcal{Q}}{\operatorname{argmax}}\, \mathcal{L}[q]$$

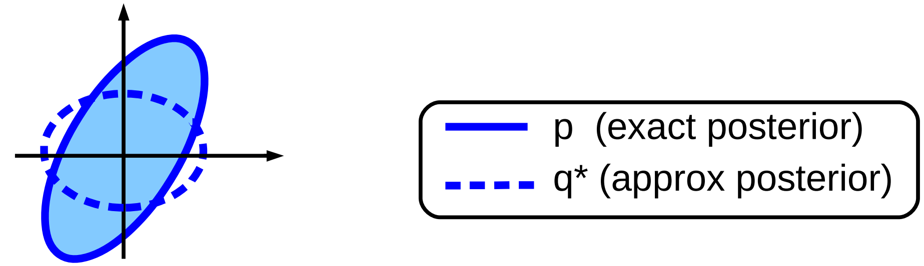
$$\mathcal{L}[q] := \langle \log p \rangle - \langle \log q \rangle$$

$$\log p(e) = d_{\mathrm{KL}}[q,p] + \mathcal{L}[q] \quad \Longrightarrow \quad \log p(e) \geq \mathcal{L}[q]$$

Choose $\mathcal{Q}$ flexible enough to approximate posterior... but simple enough for efficient optimization.

- "Mean-field" variational family:

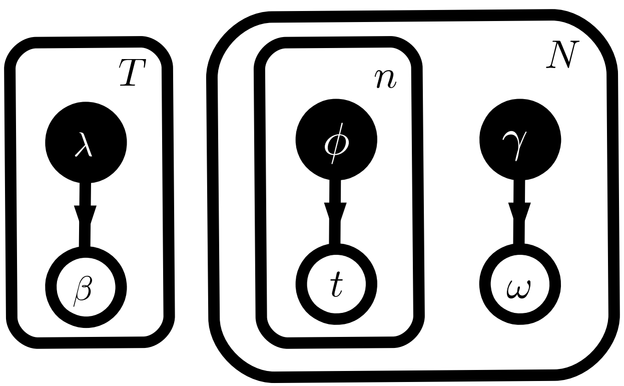
$$q(\theta|\mu) = \prod_i q(\theta_i|\mu_i)$$



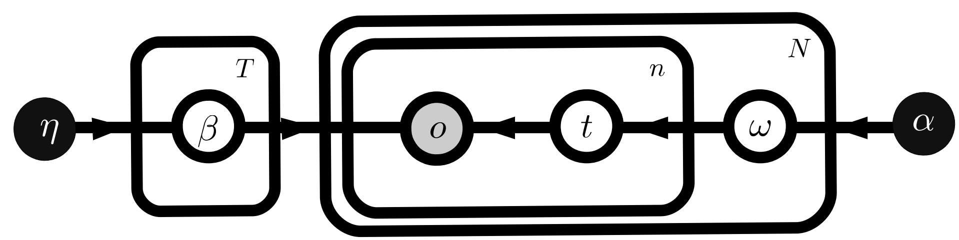
- LDA variational inference:

$$q(\omega, t, \beta|\lambda, \phi, \gamma) = q(\omega|\gamma)\, q(t|\phi)\, q(\beta|\lambda)$$

LDA mean-field approximation



LDA



$$q(\omega) = \text{Dirichlet}(\omega|\gamma)$$
$$q(t) = \text{Multinomial}(\phi)$$
$$q(\beta) = \text{Dirichlet}(\beta|\lambda)$$

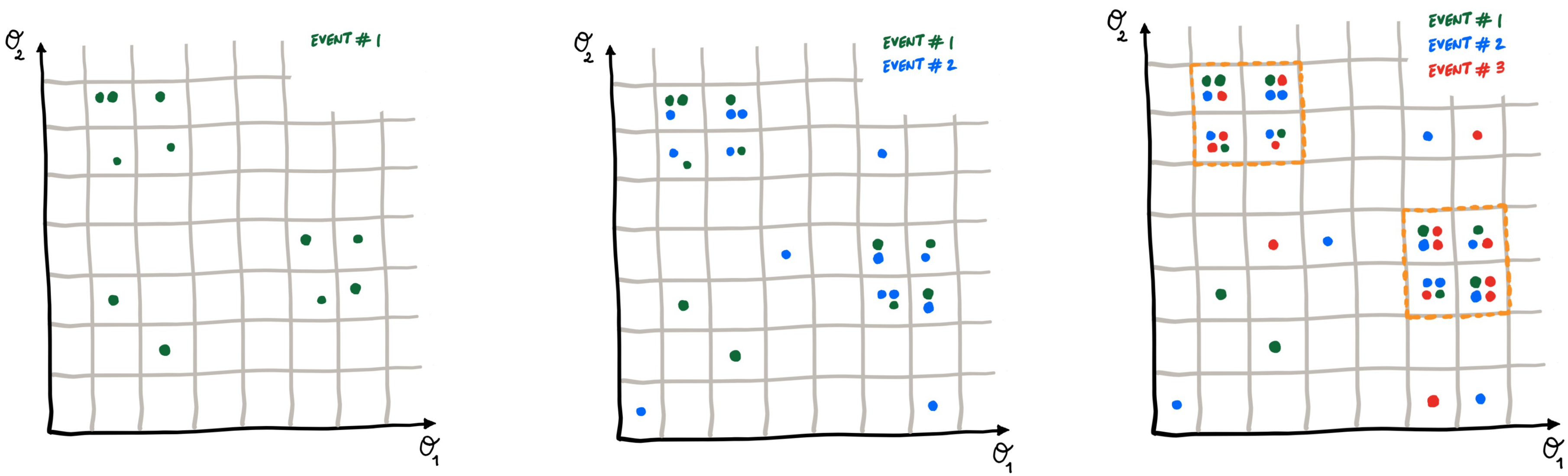
$$(\lambda^*, \phi^*, \gamma^*) = \underset{(\lambda,\phi,\gamma)}{\text{argmax}}\, \mathcal{L}[q(\omega, t, \beta)|\lambda, \phi, \gamma)]$$

Co-ocurrences

- What does LDA learn?

- LDA learns by identifying recurring measurement patterns

  Captures the statistical dependencies between event measurements in the event ensemble



Finds Co-ocurrences between event measurement throughout the event sample.

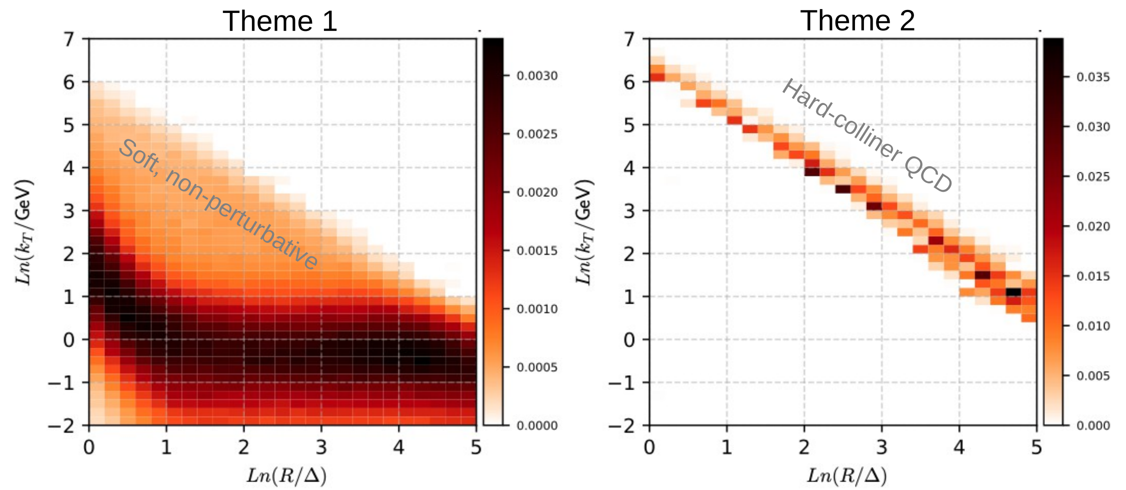
(LDA clusters in the same themes measurments that tend to co-occur together)

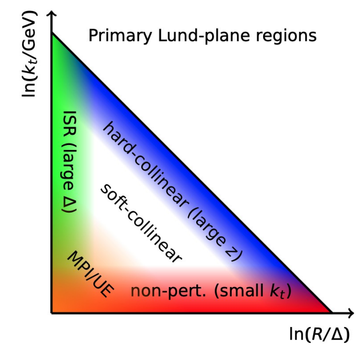
- What if there is NO signal?

Train ~ 100k QCD events

Asymmetric Dirichlet prior

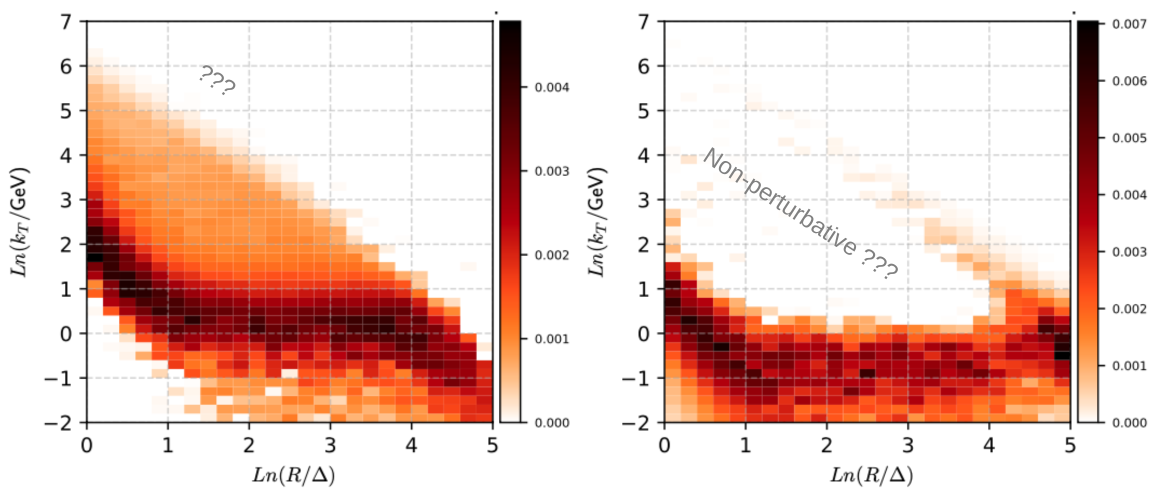
$$(\rho, \Sigma) = (0.1, 1)$$





Symmetric Dirichlet prior

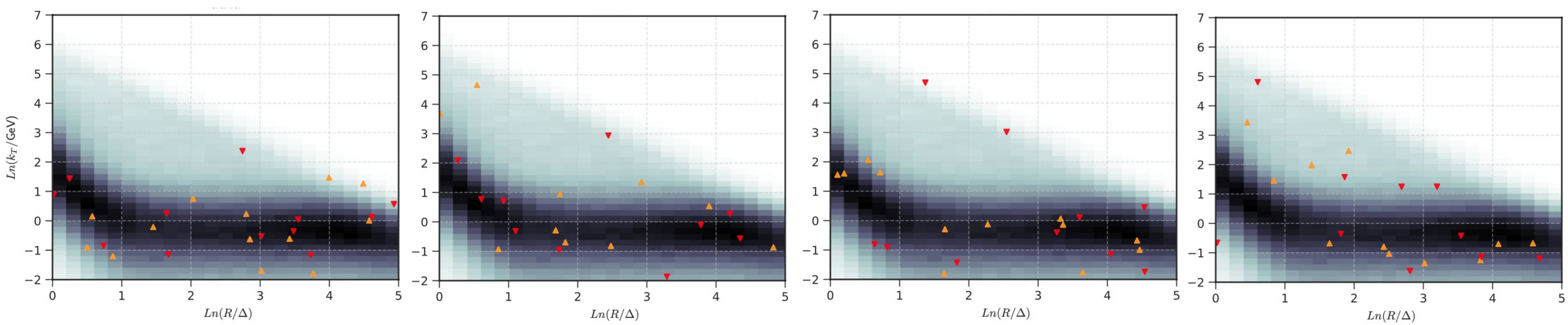
$$(\rho, \Sigma) = (0.75, 1.8)$$



Point pattern co-ocurrences in the Lund plane

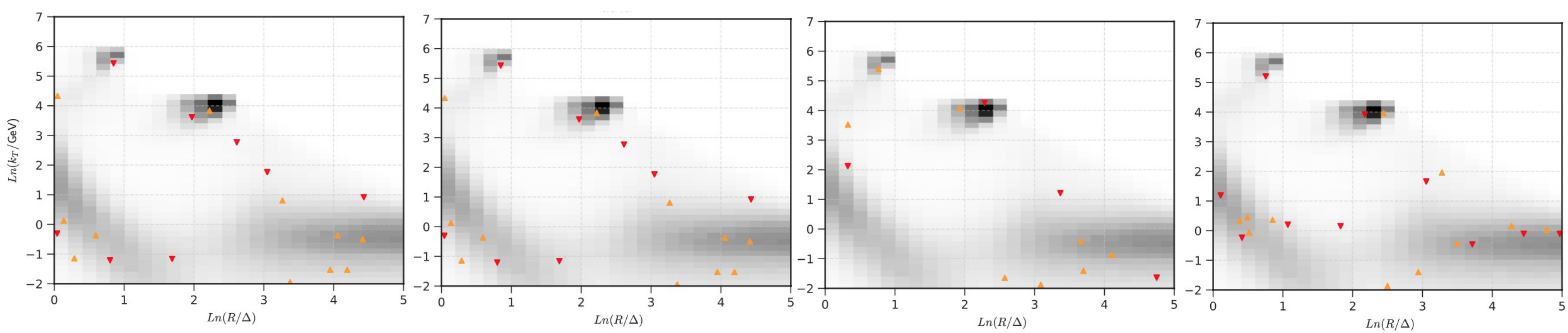
- 4 QCD events:

leading jet
sub-leading jet



- 4 signal events:

leading jet
sub-leading jet

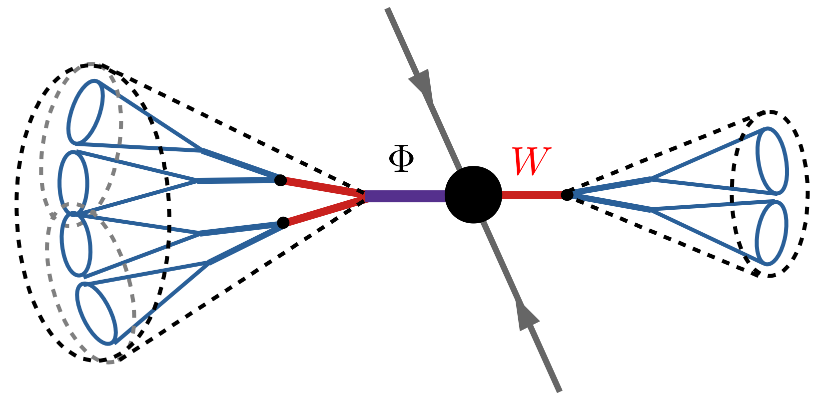


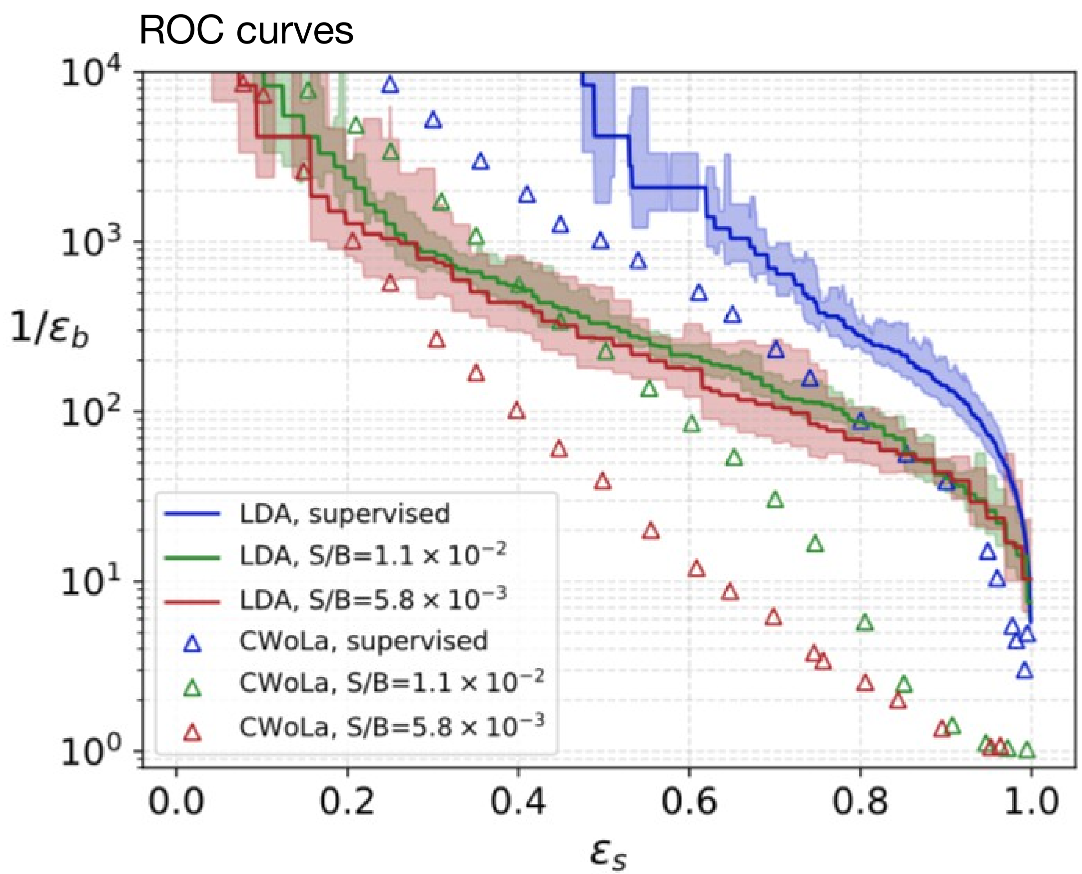
Uncovering unkown BSM

What BSM we plugged in? W’ + scalar

$$pp \to W' \to \Phi W^{\pm}, \; \Phi \to W^{\pm} W^{\mp}$$

$$m_{W'} = 3\,\text{TeV}, \quad m_{\Phi} = 400\,\text{GeV}$$





Perplexity

For an event sample $\mathcal{D} = \{e_1, \ldots, e_N\}$

$$\text{perplexity}(\mathcal{D}) := 2^{-b} \qquad b = \frac{1}{n_{\text{tot}}} \sum_{j=1}^{N} \log p(e_j) \approx \frac{1}{n_{\text{tot}}} \sum_{j=1}^{N} \mathcal{L}(e_j)$$

Total number of measurements

ELBO

- Perplexity is the measure of how well a generative model fits the data sample.

Good models have a lower perplexity score, i.e. a greater probability it generated the observed data.

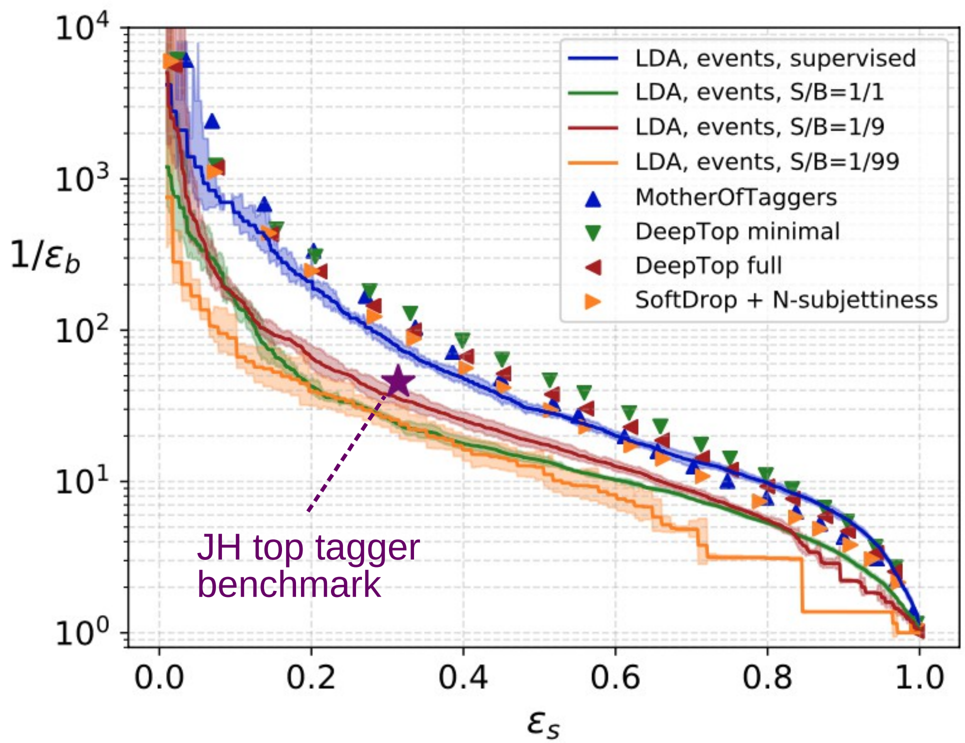
Back to 1995: 're-discovering' Top-quarks

- Train two-theme LDA on mixed (unlabelled) QCD + tops sample ~ **50k events**

Small signal: $s/b = 0.05$

$$\mathcal{O}_{\text{Mass}} = \left\{ \ell, m_{j_0}, \frac{m_{j_1}}{m_{j_0}} \right\} \qquad (\rho, \Sigma) = (0.1, 1.5)$$

LDA classifier performance:



Moderate performance for unsupervised LDA classifiers