

# Uncovering Hidden New Physics Patterns in Collider Events

ICHEP 2020 | PRAGUE

Darius A. Faroughy

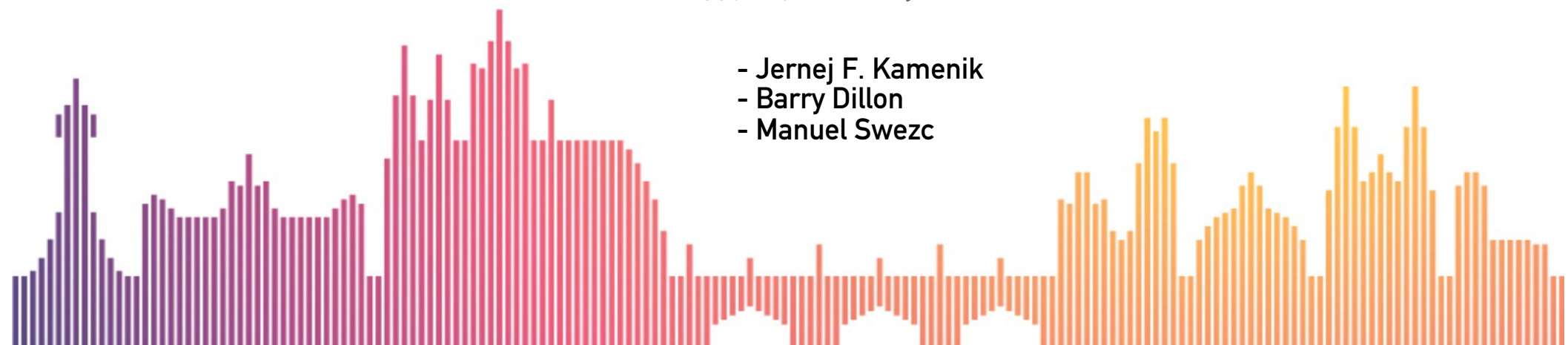


University of  
Zurich <sup>UZH</sup>

1904.04200  
2005.12319

Phys. Rev. D 100 (2019) 5, 056002

- Jernej F. Kamenik
- Barry Dillon
- Manuel Swezc



# Overview

- Last few years: several proposals for unsupervised hunting for BSM physics...

- Auto-encoders (AE) Farina et al (2018), Roy et al (2019)
- Variational AEs Cerri et al (2018)
- Cwola Metodiev et al (2018), Collins et al (2019), Amram et al (2020)
- Demix Metodiev, Thaler (2018), Komiske et al (2019), Alvarez et al (2019)
- Junipr Andreassen et al (2018, 2019)
- ...

- Unsupervised LHC searches are Challenging!

- To push these efforts even further, we probably need a deeper understanding of collider data!

Some recent developments: Geometrization of collider events (energy mover's distance)

Komiske et al (2019, 2020) , Cesarotti, Thaler (2020), Romao et al (2020)

[See Romao's Talk](#)

- This Talk:

- Is it possible to write down simple statistical models for generic collider events?
- Can we use them for unsupervised event classification tasks? e.g. Uncover BSM

(Bayesian) Probabilistic Generative Modeling (a.k.a Model-building for data)

Dillon, DAF, Kamenik (2019) Dillon, DAF, Kamenik, Scwez (2020)

# Representing collider events

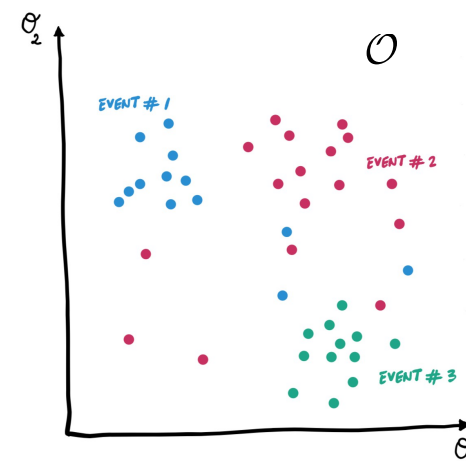
- An individual event is a sequence of multiple 'measurements'

$$\begin{cases} e_i = \{o_{i1}, o_{i2}, \dots, o_{iN_i}\} \\ o_{ij} \in \mathcal{O} \text{ 'Phase space'} \end{cases}$$

Observables may be high or low-level

- Events are **Stochastic Point Process** in phase space.

$$\mathcal{D} = \{e_1, \dots, e_N\}$$



- What is the probability distribution of a collider event?

$$P(e_i) = P(o_{i1}, o_{i2}, \dots) \quad \text{Joint probability distribution}$$

- Our broad assumptions:

- (1) Exchangeability of measurements.
- (2) Discretization of phase space.
- (3) Multiple 'latent' categories contributing to event generation.

- Disclaimer: our task is to classify events, not build a faithful event generator.

# Exchangeability

- First assumption: exchangeability of measurements

$$P(e_i) = P(o_{i1}, o_{i2}, o_{i3}, \dots) = P(o_{i\pi(1)}, o_{i\pi(2)}, o_{i\pi(3)}, \dots)$$

$\pi \in \mathcal{S}$  (permutation group)

- De Finetti's theorem (1931):

A sequence of measurements is **exchangeable** if and only if there exists a distribution functions such that

$$P(e_i) = \int_{\Omega} d\omega P(\omega) \prod_{j=1}^{N_i} P(o_{ij}|\omega)$$

Latent space

'Prior'

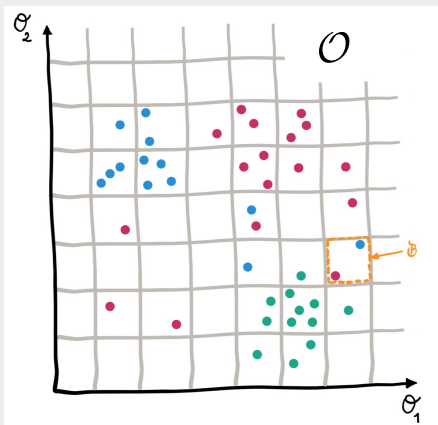
'Likelihood'

$\mathcal{D} = \{e_1, \dots, e_N\}$

- Measurements are considered **conditionally independent** given a latent variable  $\omega \in \Omega$

# Phase space discretization

- Second assumption: binned measurements



Naturally leads to **Multinomial distributions**

$$o_{ij} \sim \text{Multi}(\beta)$$

$$\left\{ \begin{array}{l} \beta = (\beta_1, \dots, \beta_M) \\ 0 \leq \beta_m \leq 1 \quad \sum_{m=1}^M \beta_m = 1 \end{array} \right.$$



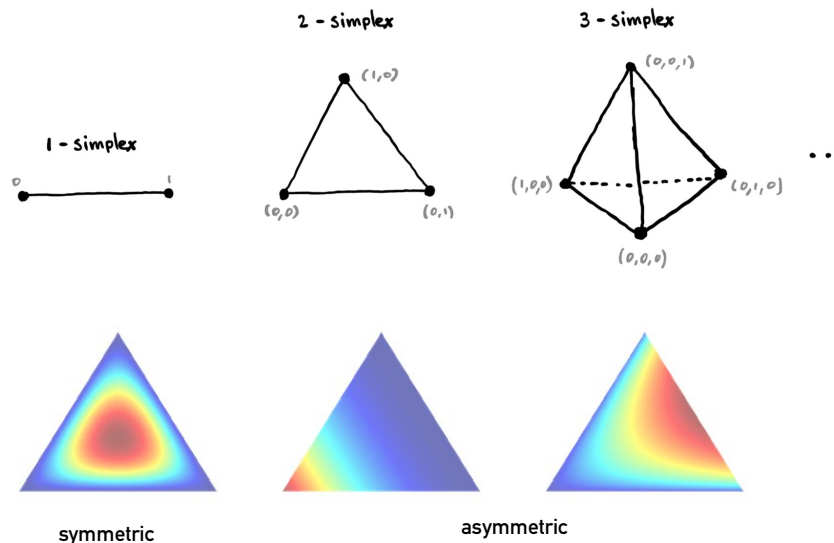
- Prior for Multinomial distribution parameters:

**Dirichlet distribution over the simplex**

$$\mathcal{D}(\beta|\alpha) = \frac{\Gamma(\alpha_1 + \dots + \alpha_M)}{\Gamma(\alpha_1) \dots \Gamma(\alpha_M)} \prod_{m=1}^M \beta_m^{\alpha_m - 1}$$

$\alpha = (\alpha_1, \dots, \alpha_M)$  concentration hyper-parameter

Dirichlet is the **conjugate** to the multinomial



# Multiple Latent Categories

- Third assumption:

- We assume collider data is generated from **multiple** Multinomial distributions over  $\mathcal{O}$

We call these latent probability distributions '**Themes**' or '**Topics**'  $P(o|\beta_t)$

(Terminology from Natural Language Processing)

Multinomial parameters  $\beta_{tm}, t = 1, \dots, T \implies T \times M$  parameters  
(# themes) x (# bins)

- Each theme encodes different information about hidden patterns associated with different physical processes or phenomena in the collider data
- Our work: we typically focused on models with two themes (T=2)

# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

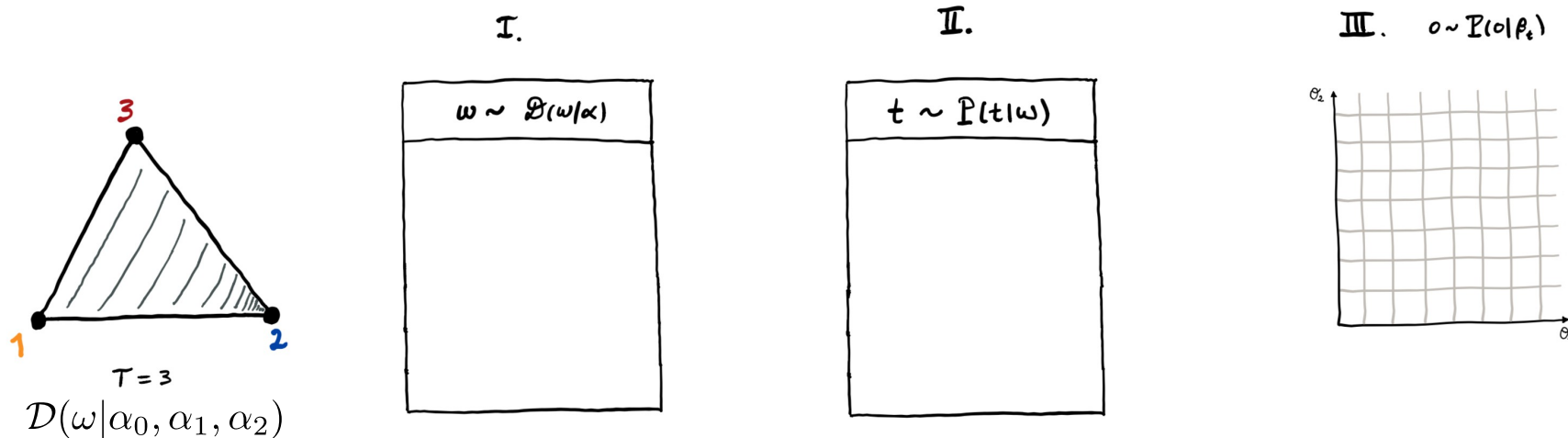
Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

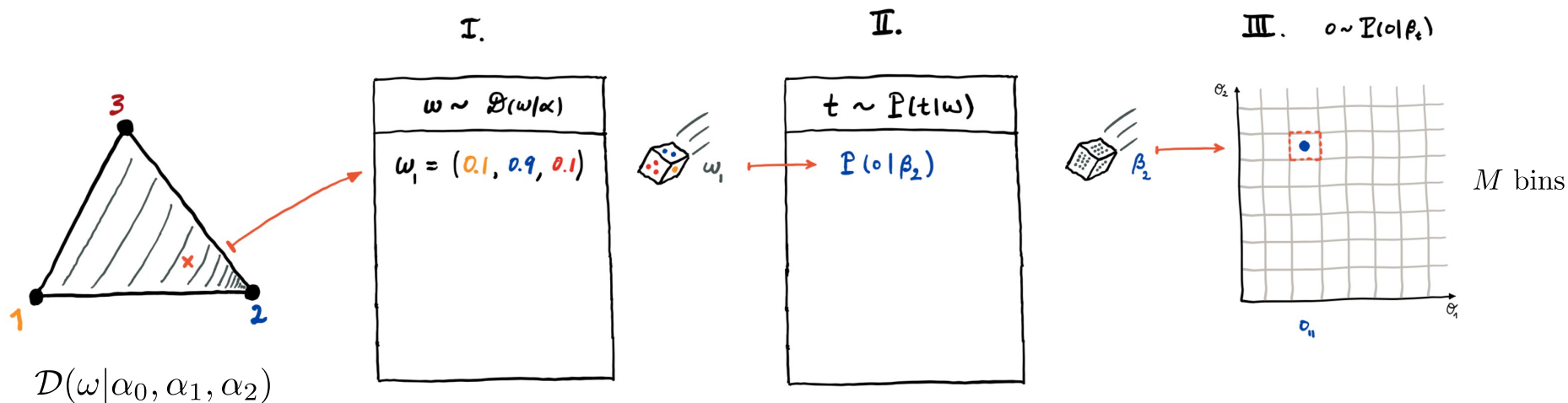
Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

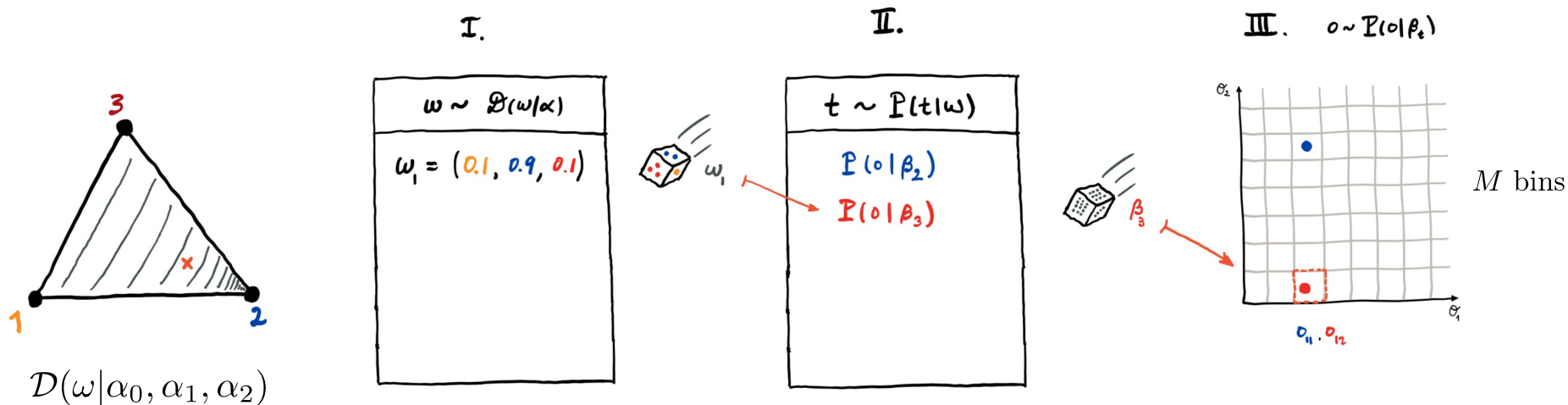
Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

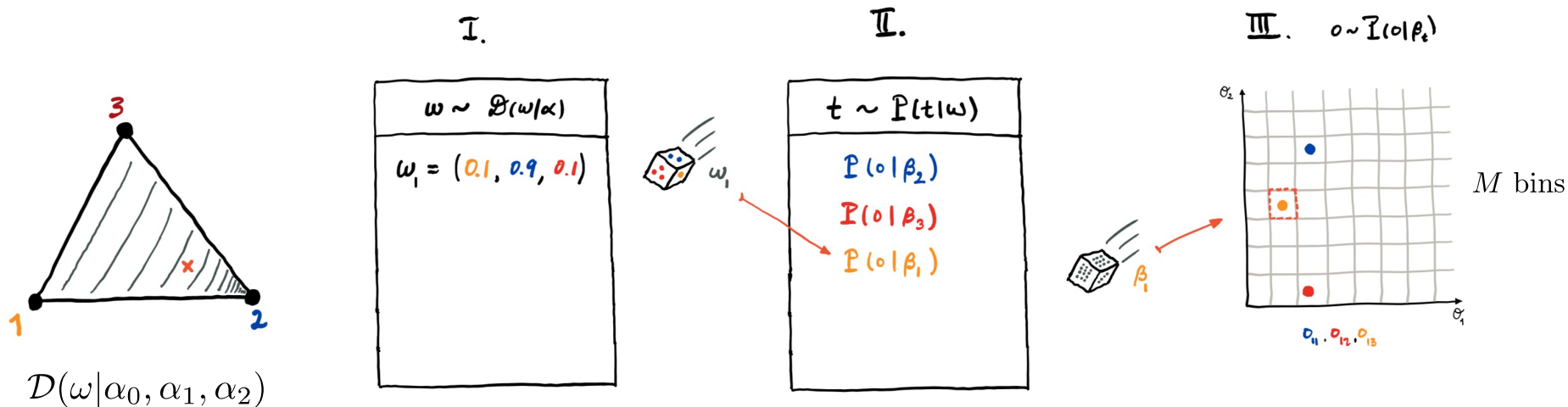
Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

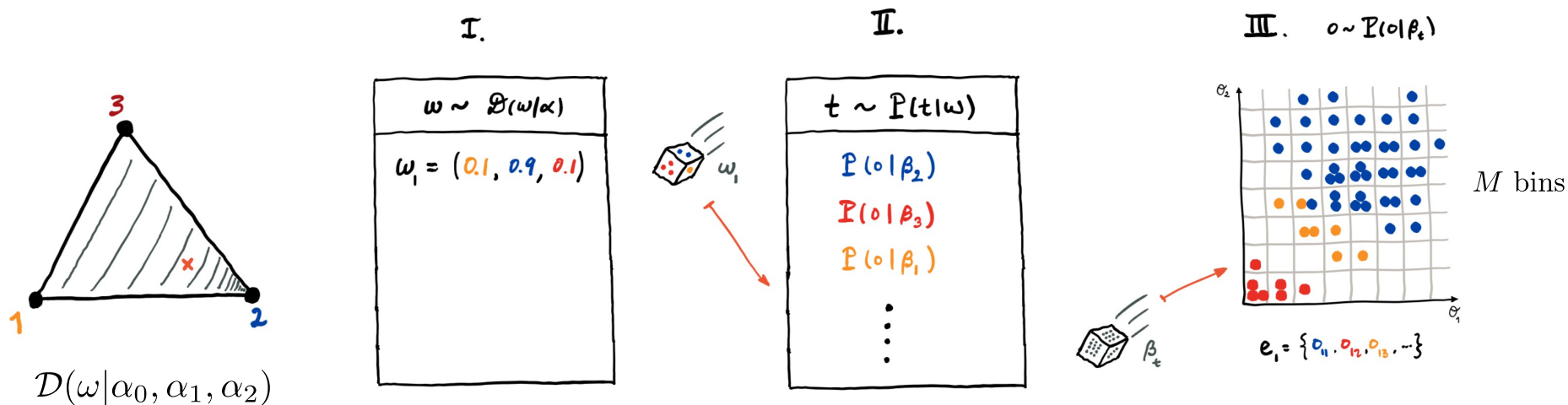
Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

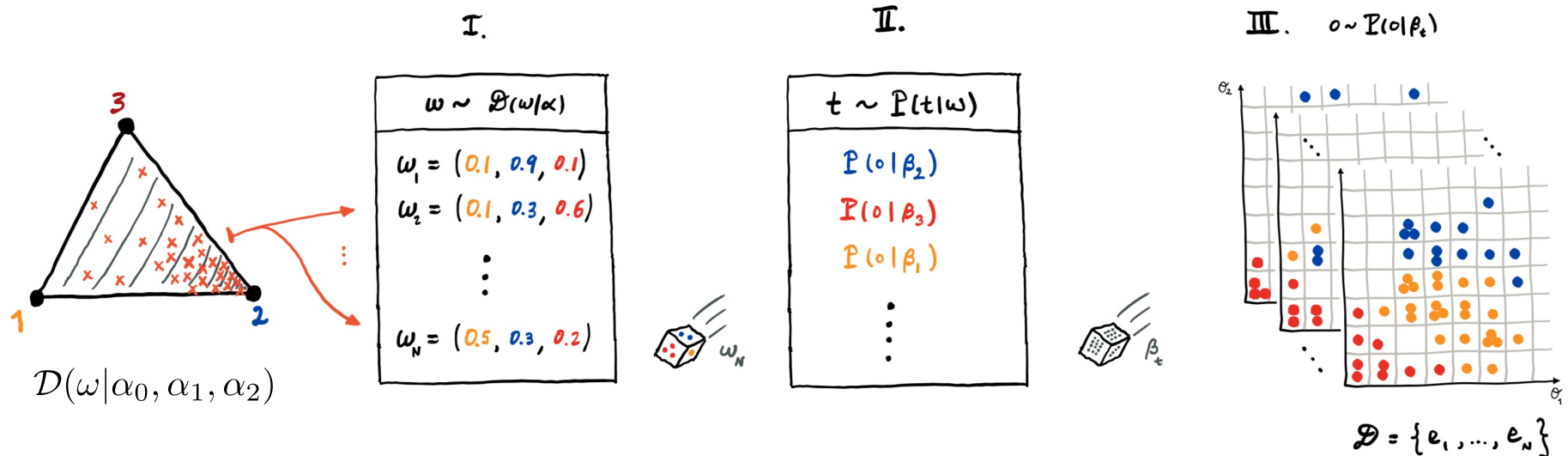
Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



# Latent Dirichlet Allocation (LDA)

Blei et al (2003)

- LDA is a **mixed-membership model**.
- Individual events are described by mixture of multiple 'latent' multinomial distributions (Themes):

$$P(e_i|\alpha) = \int_{\Delta_{T-1}} d\omega \mathcal{D}(\omega|\alpha) \prod_{j=1}^{N_i} \left[ \sum_{t=1}^T P(t|\omega) P(o_{ij}|\beta_t) \right]$$

Latent space:  
Simplex

Dirichlet  
Prior

Latent variable:  
'Theme' mixing  
proportions

Theme  
Multinomial  
distributions

$P(\cdot|\beta_t) \quad t = 1, \dots, T$

- Generative process:



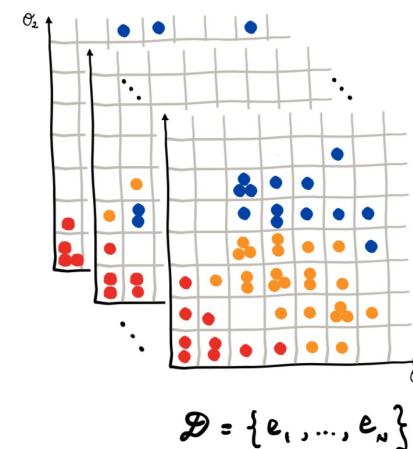
$D(\omega|\alpha_0, \alpha_1, \alpha_2)$

I.

II.

III.  $o \sim P(o|\beta_t)$

- **Mixed-Membership Models** not to be confused with **Mixture models**!
- The former allows for **shared** features between events! Hence more flexible.
- In the later all measurements in an event would come from only one theme...



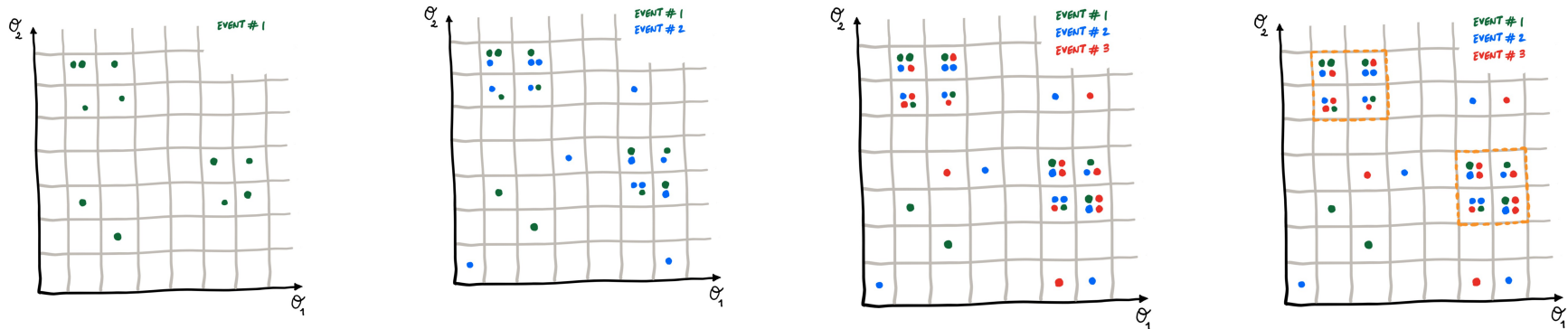
# Inference & Event classification with LDA

- Bayesian inference:  $P(\omega, \beta_t | e_i, \alpha) = \frac{P(\omega, \beta_t | \alpha)}{P(e_i | \alpha)}$  the posterior

- Solved via approximate methods: (stochastic) Variational Inference

- output: estimation of the parameters of the model  $\hat{\beta}_t, \hat{\omega}_t$

- What does LDA learn?  $\implies$  'Co-occurrence' of measurement bins in events



- LDA Likelihood Ratio Two-theme classifiers:

$T = 2$

$$\mathcal{L}(e_i | \alpha) = \prod_{j=1}^{N_i} \frac{P(o_{ij} | \hat{\beta}_1)}{P(o_{ij} | \hat{\beta}_2)} \quad \begin{cases} \mathcal{L}(e_i | \alpha) > c \implies e_i \in \mathcal{C}_1 \\ \mathcal{L}(e_i | \alpha) \leq c \implies e_i \in \mathcal{C}_2 \end{cases}$$

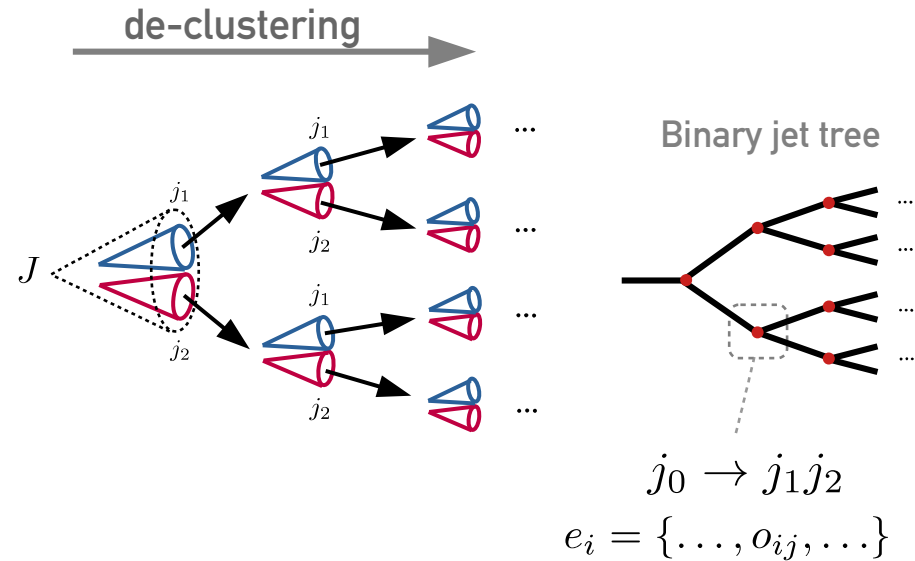
# Application: uncovering hidden patterns inside jets

- Jet de-clustering history

$$e = \{J_1, J_2, \dots\} \quad p_T(J_1) > p_T(J_2) > \dots$$

Clustered with Cambridge/Aachen algorithm

$$R \sim 1 \text{ 'fat jets'}$$

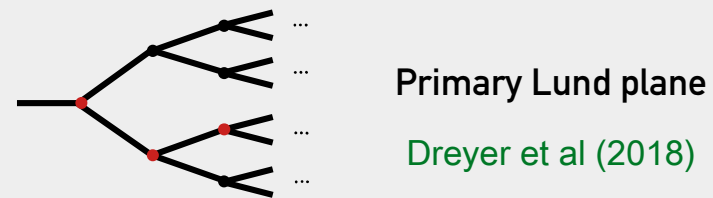


- Two data representations:

$$\mathcal{O}_{\text{Mass}} = \left\{ \underset{\text{Jet label}}{\ell}, m_{j_0}, \underset{\text{mass-drop}}{\frac{m_{j_1}}{m_{j_0}}} \right\}$$



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



Both Jet substructure observables are good discriminating massive hadronic resonances from QCD

# Resonant (B)SM in jet substructure

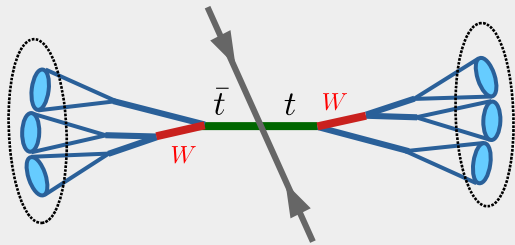
- Trained **two-theme LDA** ( $T = 2$ ) on:

- **Unlabeled** di-jet samples with QCD + 'signal'  $s/b \ll 1$

- Two benchmark signal processes:

i) Boosted tops

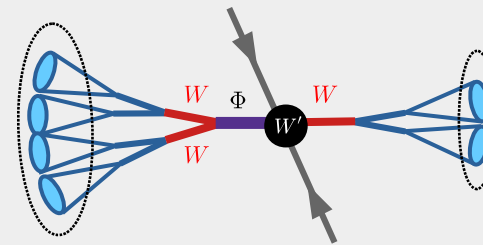
$$pp \rightarrow t\bar{t}, \quad t \rightarrow bW^\pm$$



ii) BSM cascade decay

Collins et al (2018)

$$pp \rightarrow W' \rightarrow \Phi W^\pm, \quad \Phi \rightarrow W^\pm W^\mp$$



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

- Trained using asymmetric Dirichlet prior

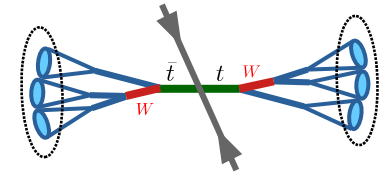
$$\mathcal{D}(\omega | \alpha_0, \alpha_1) \quad \alpha_0 \ll \alpha_1$$

$$\begin{cases} P(o|\beta_1) \implies \text{'QCD-like' theme} \\ P(o|\beta_2) \implies \text{'Rare' theme} \end{cases}$$

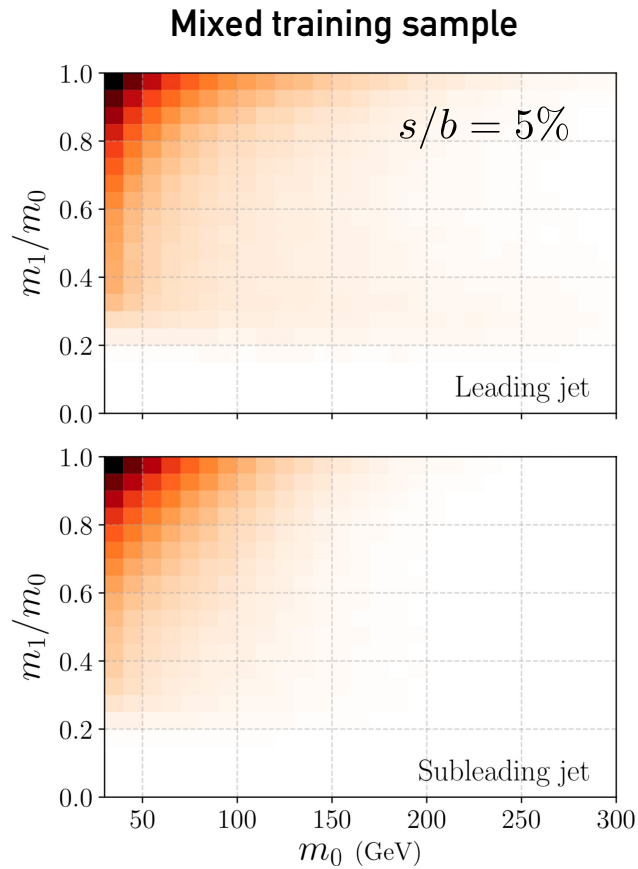
e.g. hopefully describing BSM

# 'Re-discovering' the Top-quark

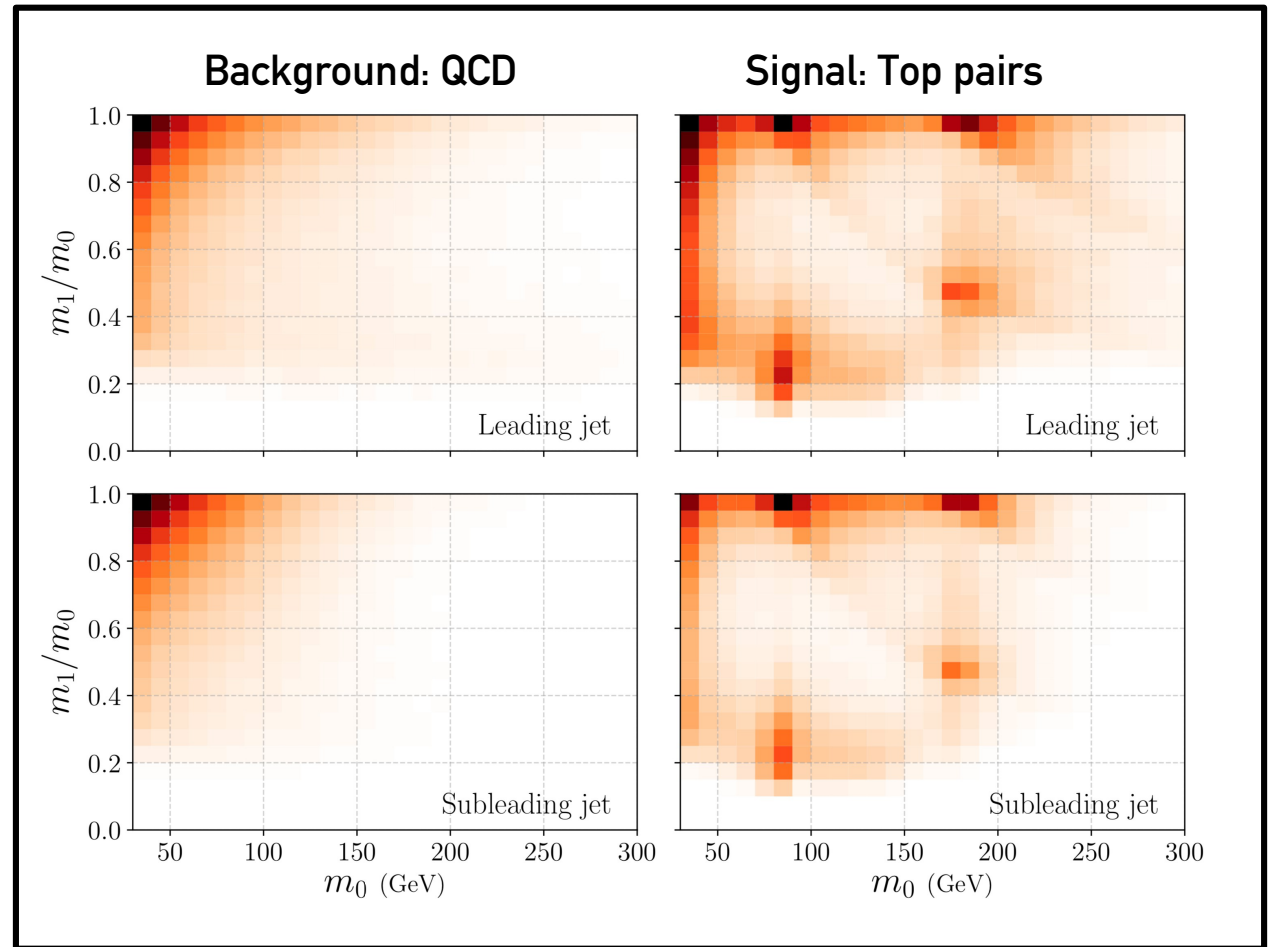
- Boosted tops  $\mathcal{O}_{\text{Mass}} = \left\{ \ell, m_{j_0}, \frac{m_{j_1}}{m_{j_2}} \right\}$



Truth level distributions

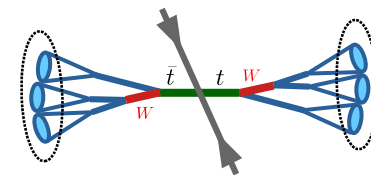


~50k events

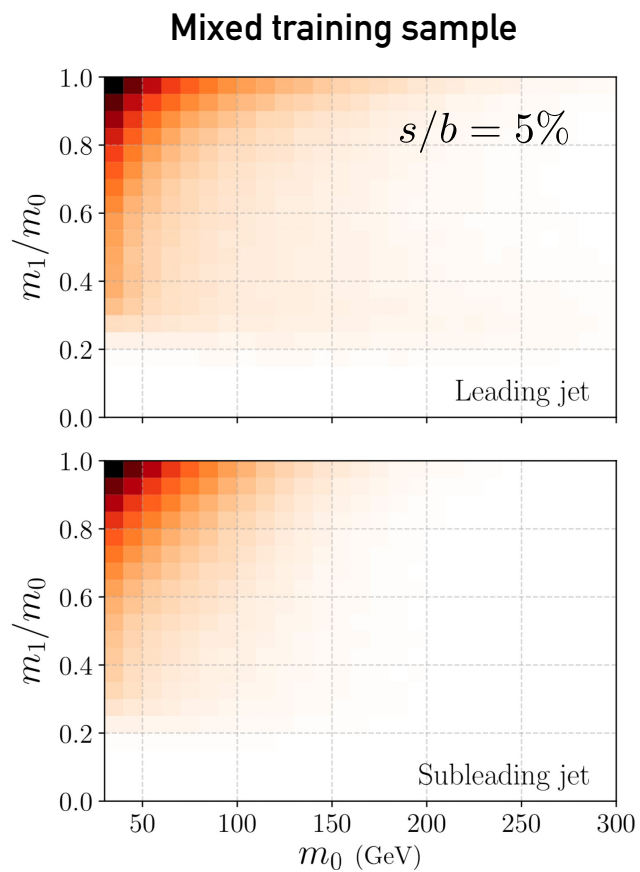


# 'Re-discovering' the Top-quark

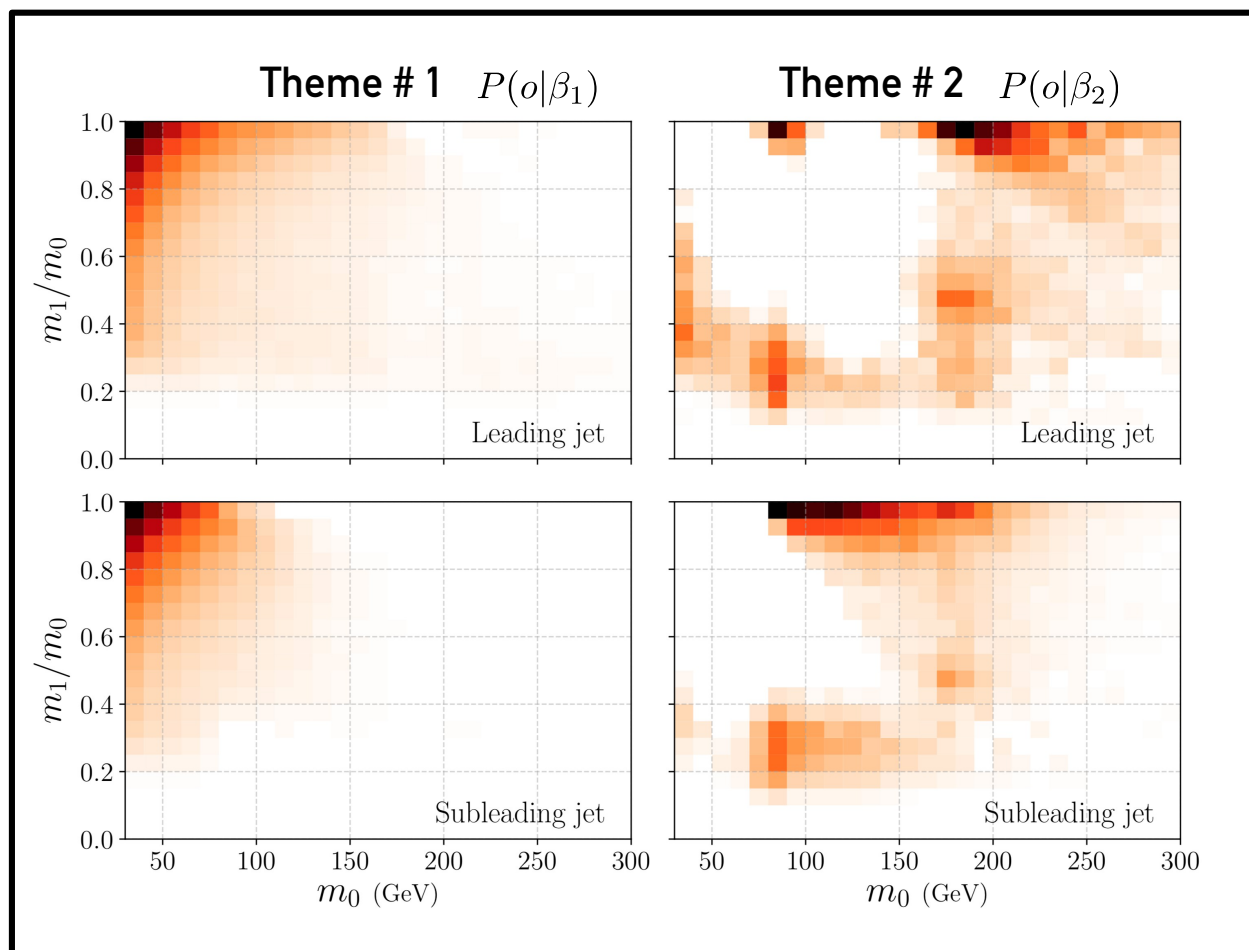
- Boosted tops  $\mathcal{O}_{\text{Mass}} = \left\{ \ell, m_{j_0}, \frac{m_{j_1}}{m_{j_2}} \right\}$



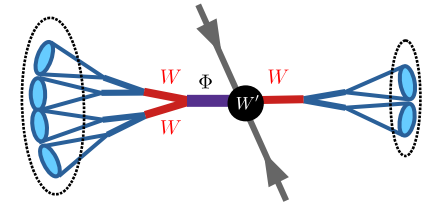
What did LDA discovered?



~50k events

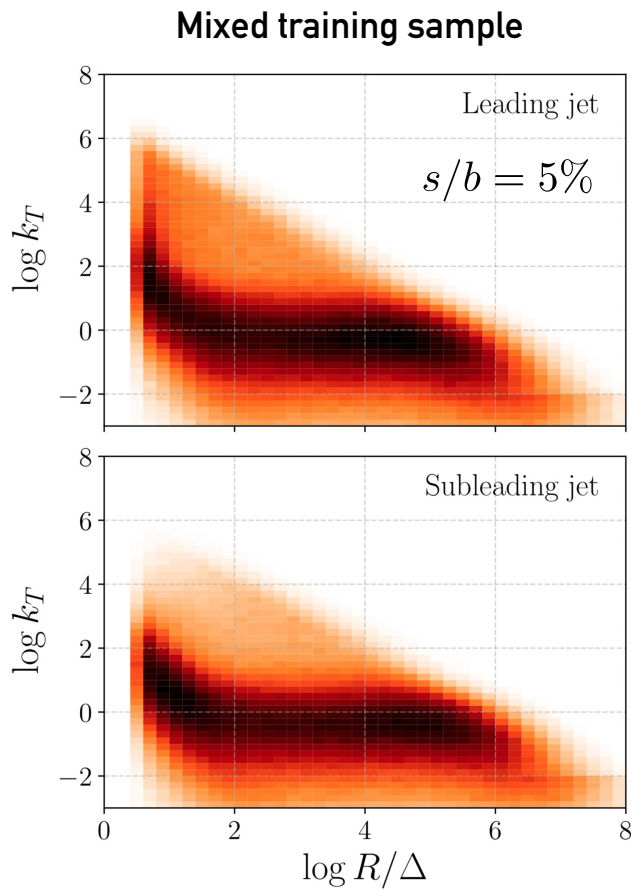


# Unsupervised BSM search

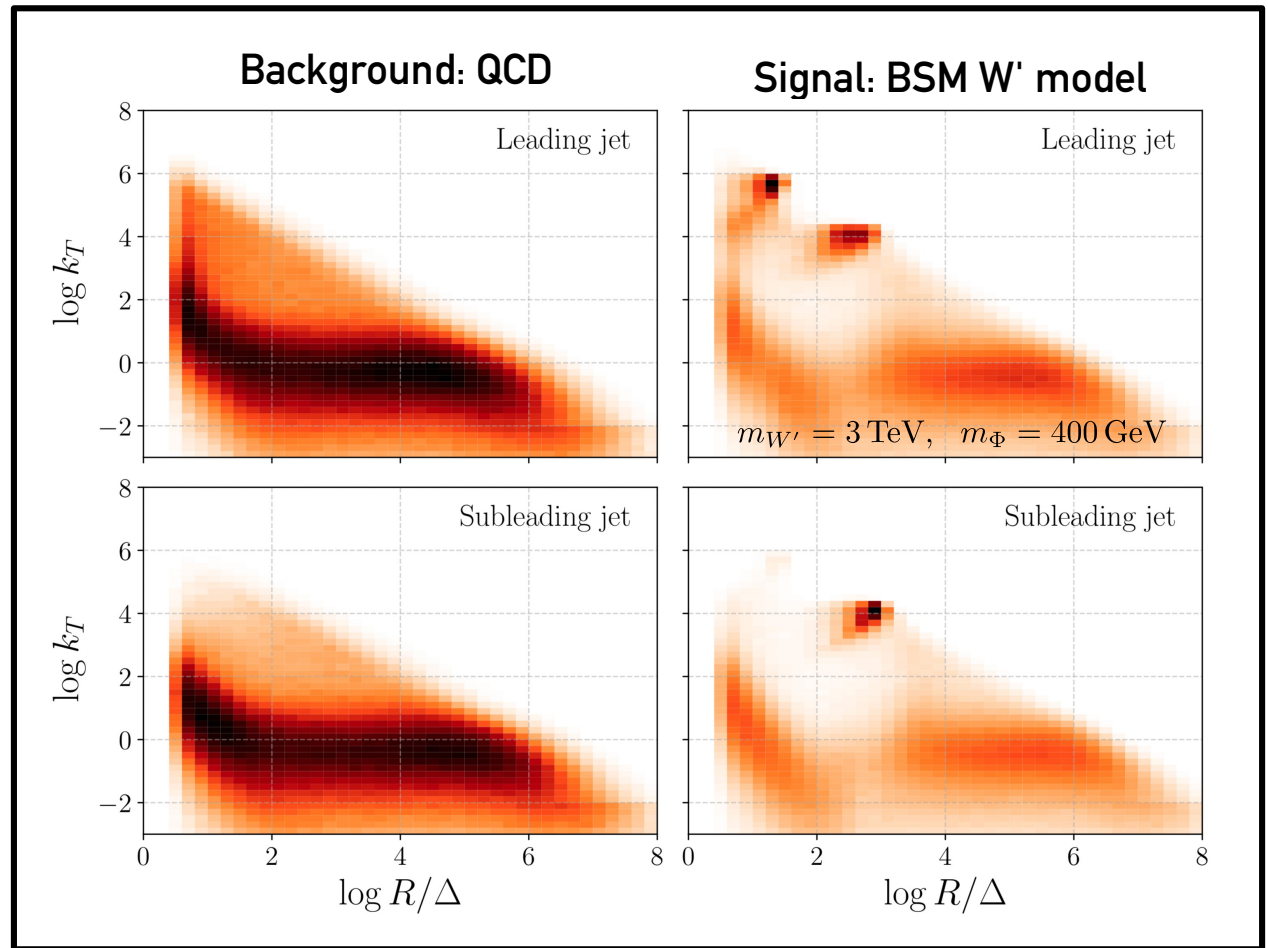


- BSM heavy cascade decay  $\mathcal{O}_{\text{Lund}} = \left\{ \ell, \log(k_t), \log\left(\frac{1}{\Delta}\right) \right\}$

## Truth level distributions

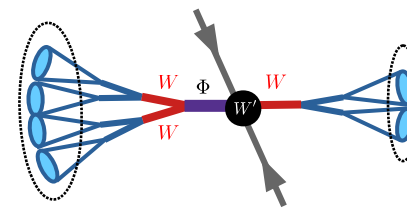


~50k events



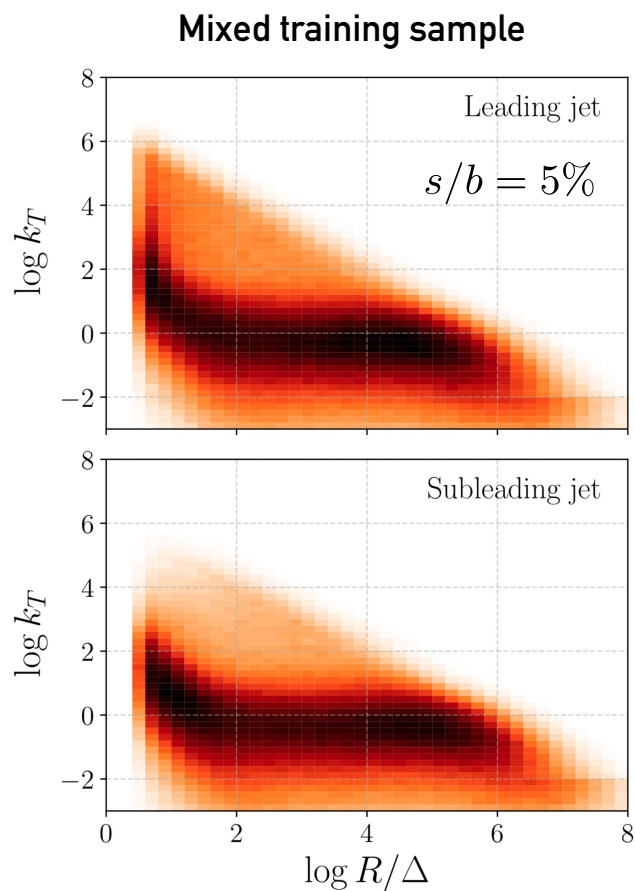
The Primary Lund Triangle

# Unsupervised BSM search

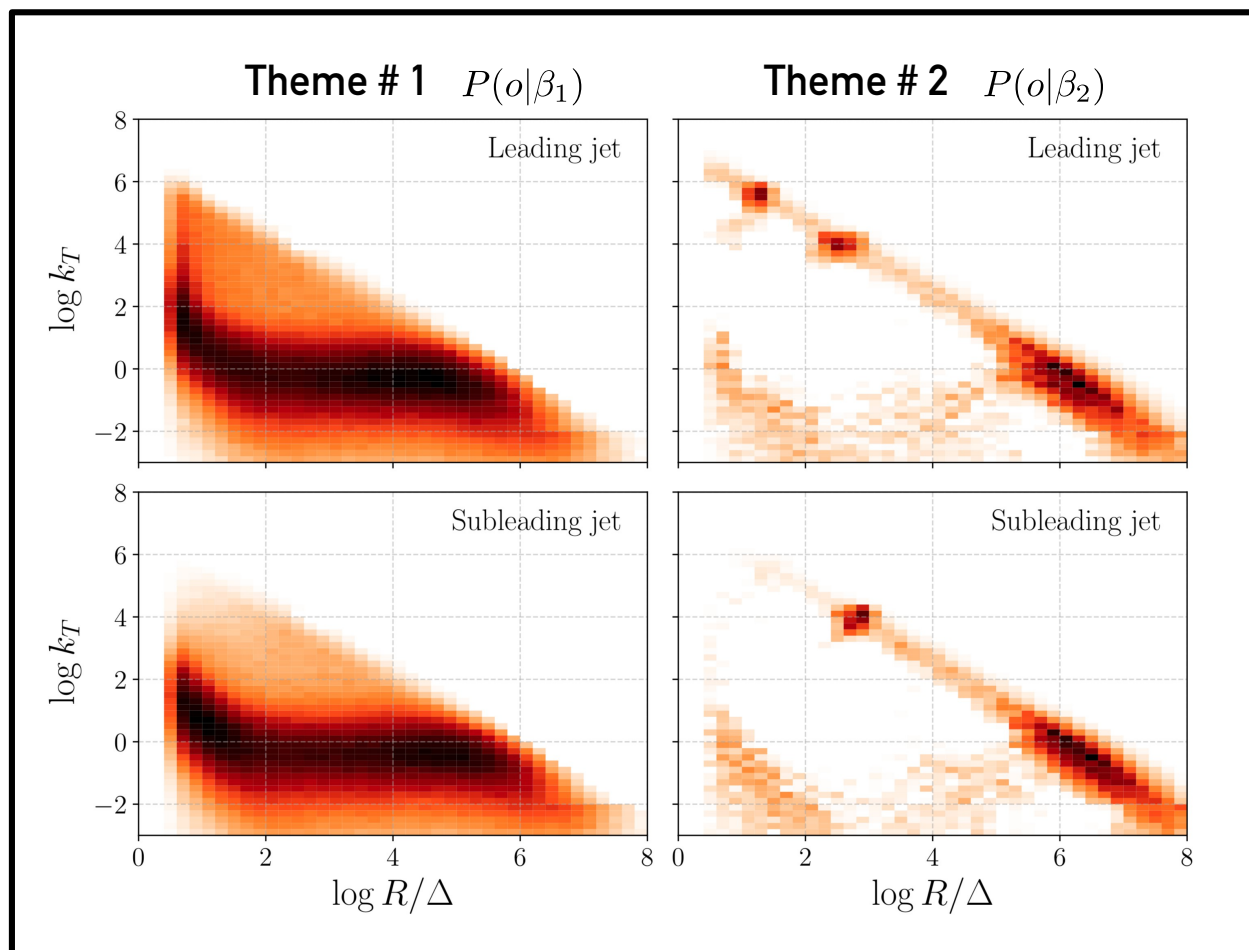


- BSM heavy cascade decay  $\mathcal{O}_{\text{Lund}} = \left\{ \ell, \log(k_t), \log\left(\frac{1}{\Delta}\right) \right\}$

What did LDA discovered?



~50k events



# Conclusions

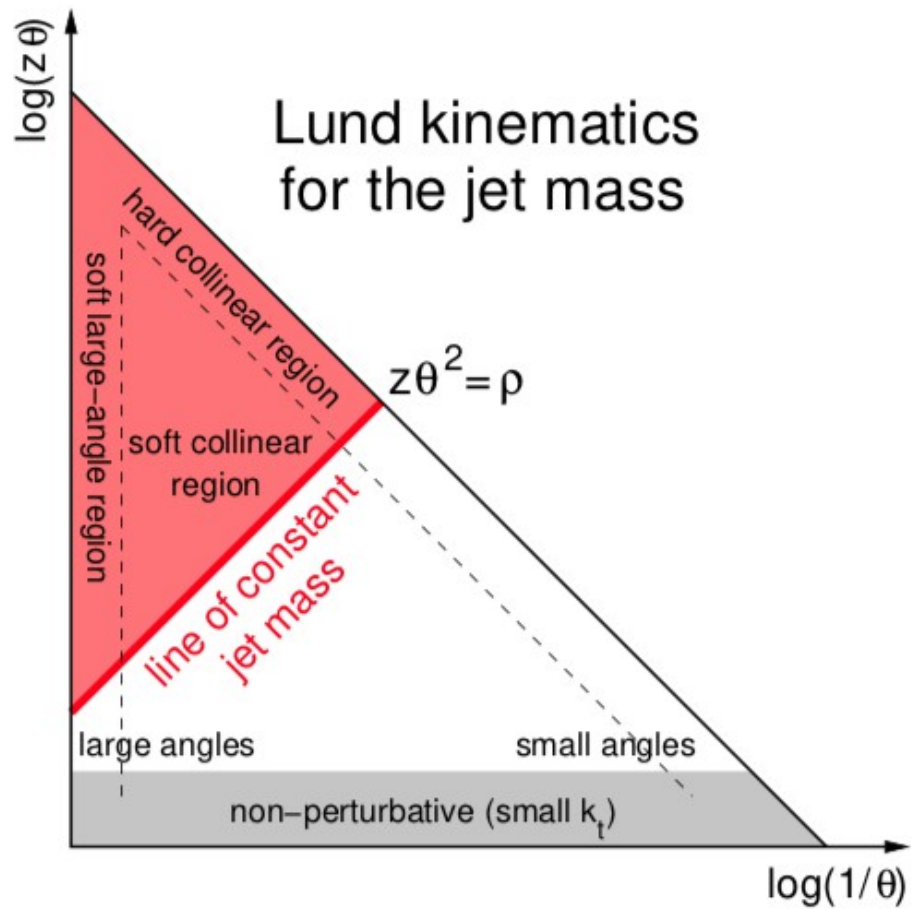
- We showed that it is possible to model collider events with simple probability distributions capable of capturing hidden patterns in the data
  - Can be used for unlabeled event classification tasks without labels.
- Method is based on Bayesian probabilistic modeling using 3 broad assumptions:
  - Collider event measurements are exchangeable.
  - Probability distributions over phase space are Multinomials with Dirichlet priors (discretization).
  - Multiple hidden categories (themes/topics) reflecting different underlying physical processes.
- We focused on one collider model: [Latent Dirichlet Allocation \(LDA\)](#)
  - Here individual events are mixtures of 'latent' theme distributions (i.e. mixed-membership model).
  - More model-building options could be explored in the future...
- We used a 'two-theme' LDA to discover the top quark and BSM decay cascades in dijets.
  - Data representation: high-level jet substructure observables (e.g. Lund Plane) extracted from jet de-clustering trees.

# Conclusions

- We showed that it is possible to model collider events with simple probability distributions capable of capturing hidden patterns in the data
  - Can be used for unlabeled event classification tasks without labels.
- Method is based on Bayesian probabilistic modeling using 3 broad assumptions:
  - Collider event measurements are exchangeable.
  - Probability distributions over phase space are Multinomials with Dirichlet priors (discretization).
  - Multiple hidden categories (themes/topics) reflecting different underlying physical processes.
- We focused on one collider model: [Latent Dirichlet Allocation \(LDA\)](#)
  - Here individual events are mixtures of 'latent' theme distributions (i.e. mixed-membership model).
  - More model-building options could be explored in the future...
- We used a 'two-theme' LDA to discover the top quark and BSM decay cascades in dijets.
  - Data representation: high-level jet substructure observables (e.g. Lund Plane) extracted from jet de-clustering trees.

Thanks!

**Back-up material**



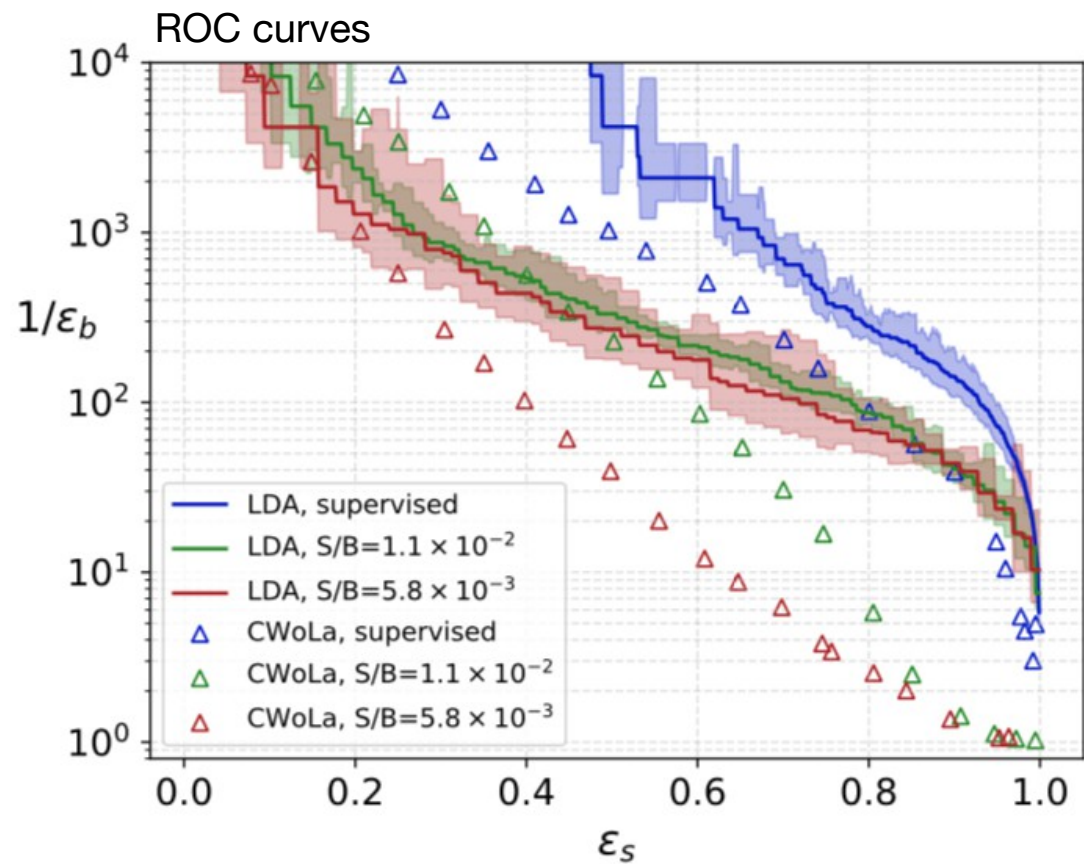
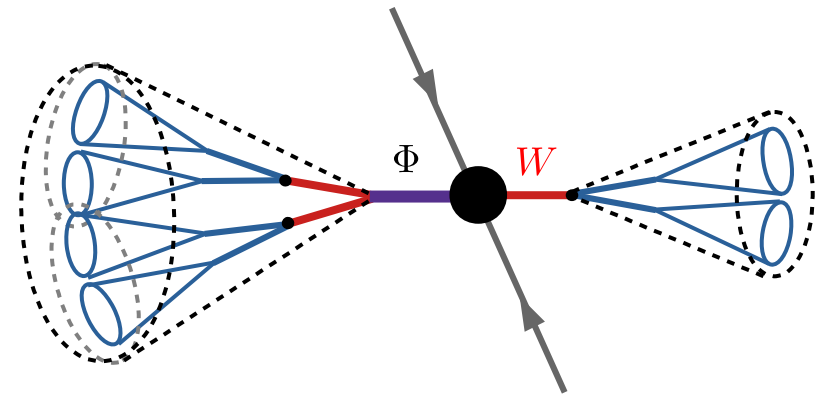
# Uncovering unknown BSM

What BSM we plugged in?  $W' + \text{scalar}$

$$pp \rightarrow W' \rightarrow \Phi W^\pm, \quad \Phi \rightarrow W^\pm W^\mp$$

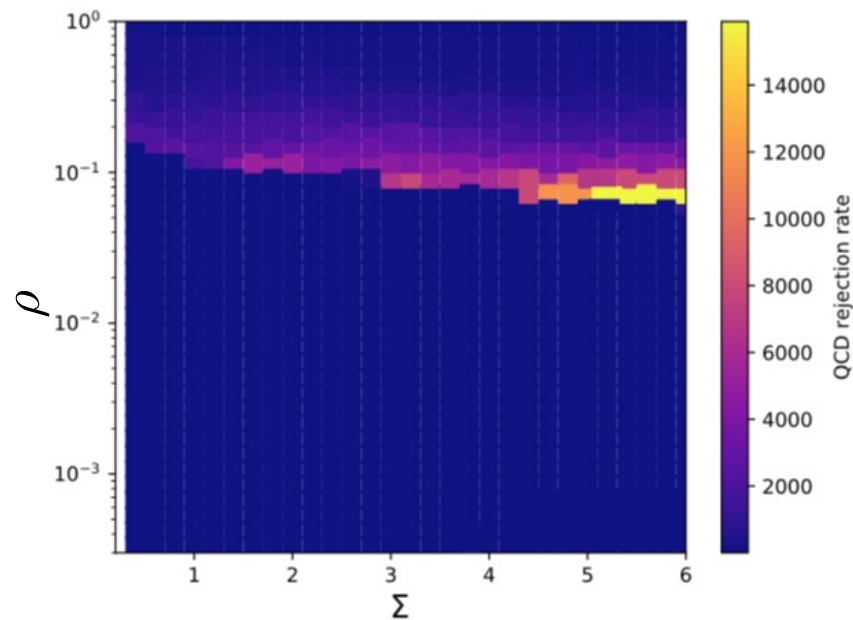
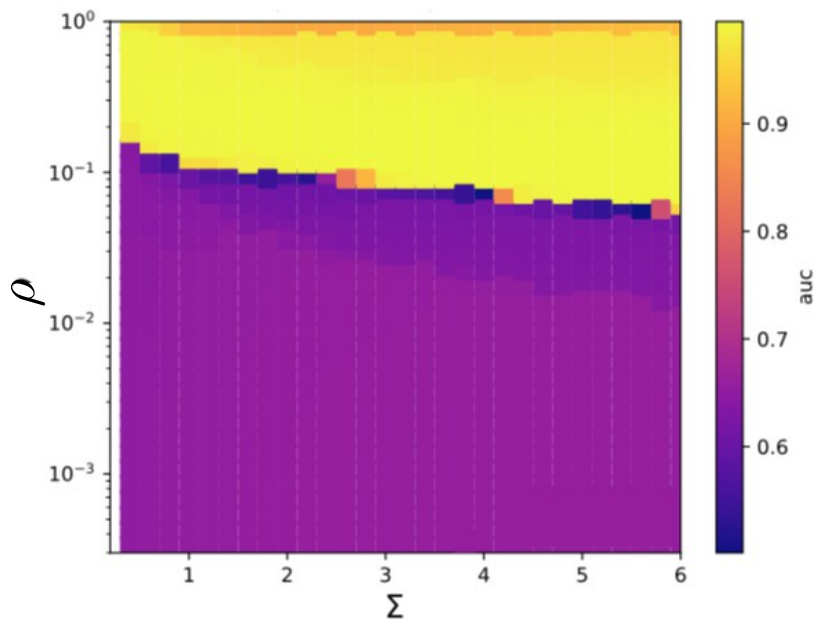
$$m_{W'} = 3 \text{ TeV}, \quad m_\Phi = 400 \text{ GeV}$$

$$S/B = 1\%, \quad 0.5\%$$

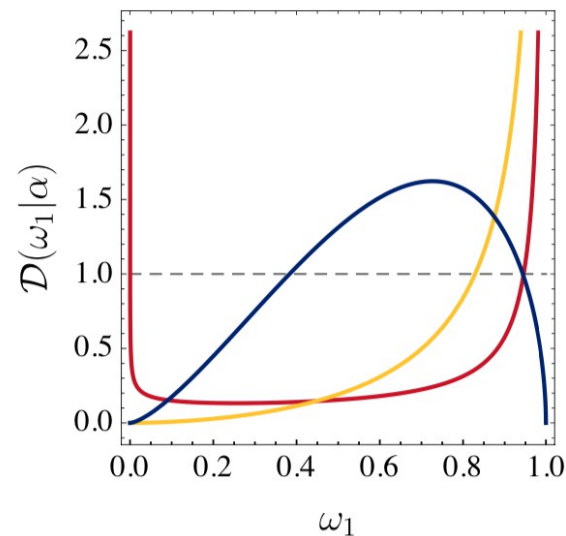


# Landscape of LDA classifiers

How to choose the prior? We can scan over the Dirichlet prior hyperparameters:

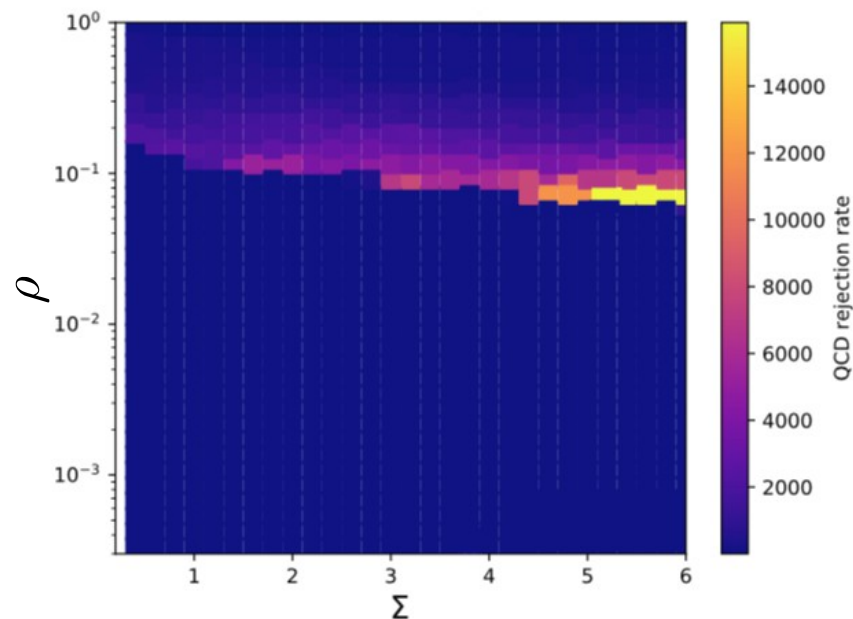
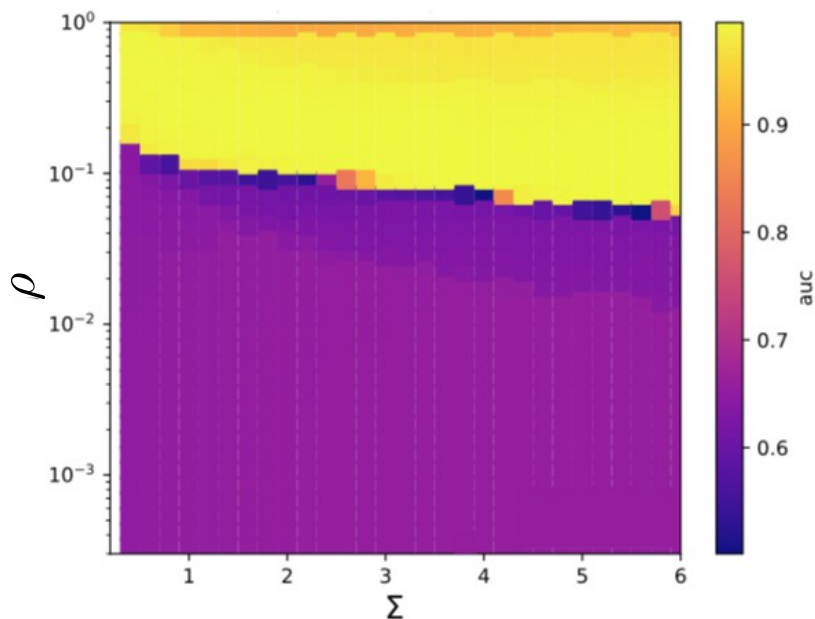


$$\mathcal{D}(\omega|\alpha_0, \alpha_1)$$
$$\rho = \alpha_0/\alpha_1$$
$$\Sigma = \alpha_0 + \alpha_1$$



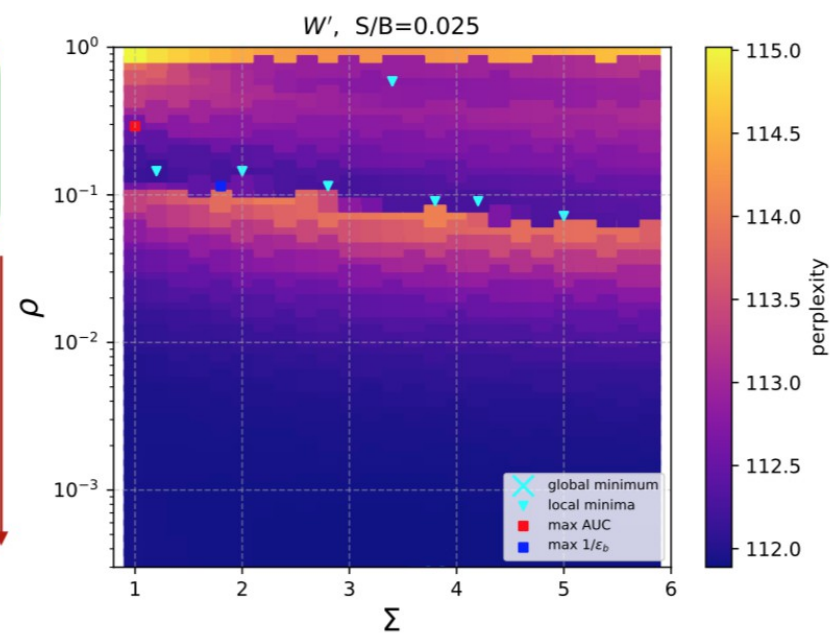
# Landscape of LDA jet taggers

How to choose the prior? We can scan over the Dirichlet prior hyperparameters:



lots of local minima, close to models with best AUC and best rejection rate at fixed mis-tag.

global minimum at vanishing rho, but this is a trivial solution.



Perplexity  $\sim$  goodness of fit!