Portorož 2021: Physics of the Flavourful Universe

Bayesian inference for Four Tops at the LHC

Based on arxiv:2107.00668 by
E. Alvarez[a], B. M. Dillon[b], D. A. Faroughy[c], J. F. Kamenik[d], F.Lamagna[e] and MS[a]

(a) ICAS UNSAM & CONICET, Argentina (b) IFT, Heidelberg U., Germany (c) Physik-Institut UZH, Switzerland (d) JSI and Ljubljana U., Slovenia (e) CAB-IB & CONICET, Argentina

In this talk:

- A brief remainder of Four Tops at the LHC:
- Graphical models for Bayesian Inference of Four tops
  - Modelling the data
  - Numerical Inference
  - Results for a specific benchmark
- Conclusions

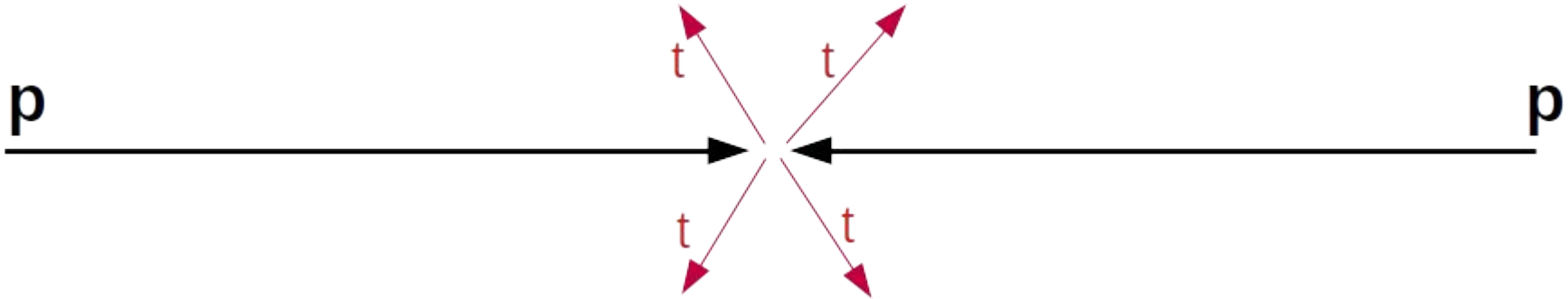
Four tops at the LHC

An increasingly sensitive SM benchmark to be explored at the LHC, with huge experimental effort and impressive results. Recent results in monolepton + 2LOS (ATLAS coll., ATLAS-CONF-2021-013, CMS coll., arxiv:1906.02805) and 2LSS + multilepton (ATLAS coll., arxiv:2007.14858. and CMS coll., arxiv:1908.06463)

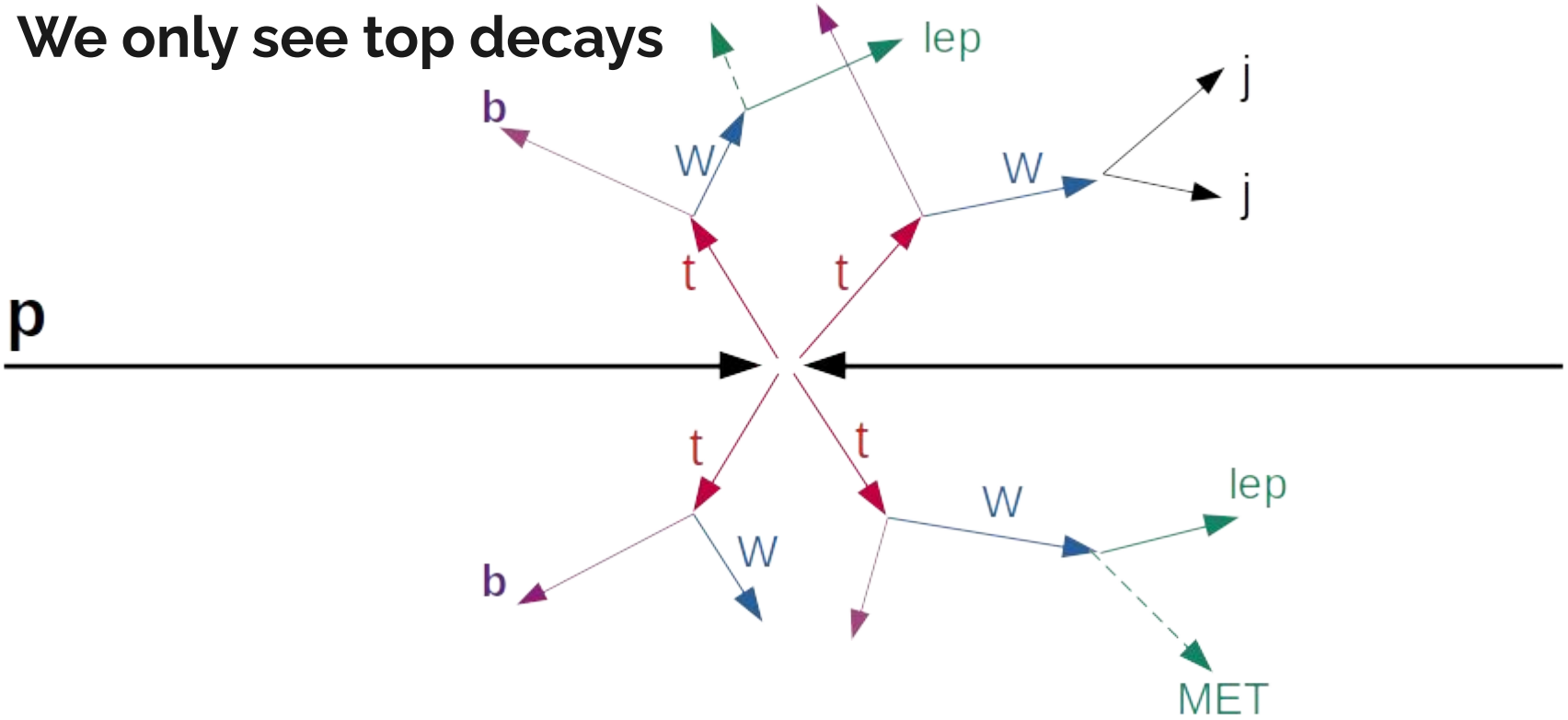
Correspondingly, state of the art calculations and improvement of SM predictions (see R. Frederix, D. Pagani, M. Zaro, arxiv:1711.02116)

A theoretically well motivated (but perhaps more importantly, still allowed) window to BSM effects (see e.g. G. Banelli, E. Salvioni, J. Serra, T. Theil, A. Weiler, arxiv:2010.05915, E. Alvarez, A. Juste, M. S., T. Vazquez Schroeder, arxiv:2011.06514 and Luc Darmé, Benjamin Fuks, Fabio Maltoni, arxiv:2104.09512).

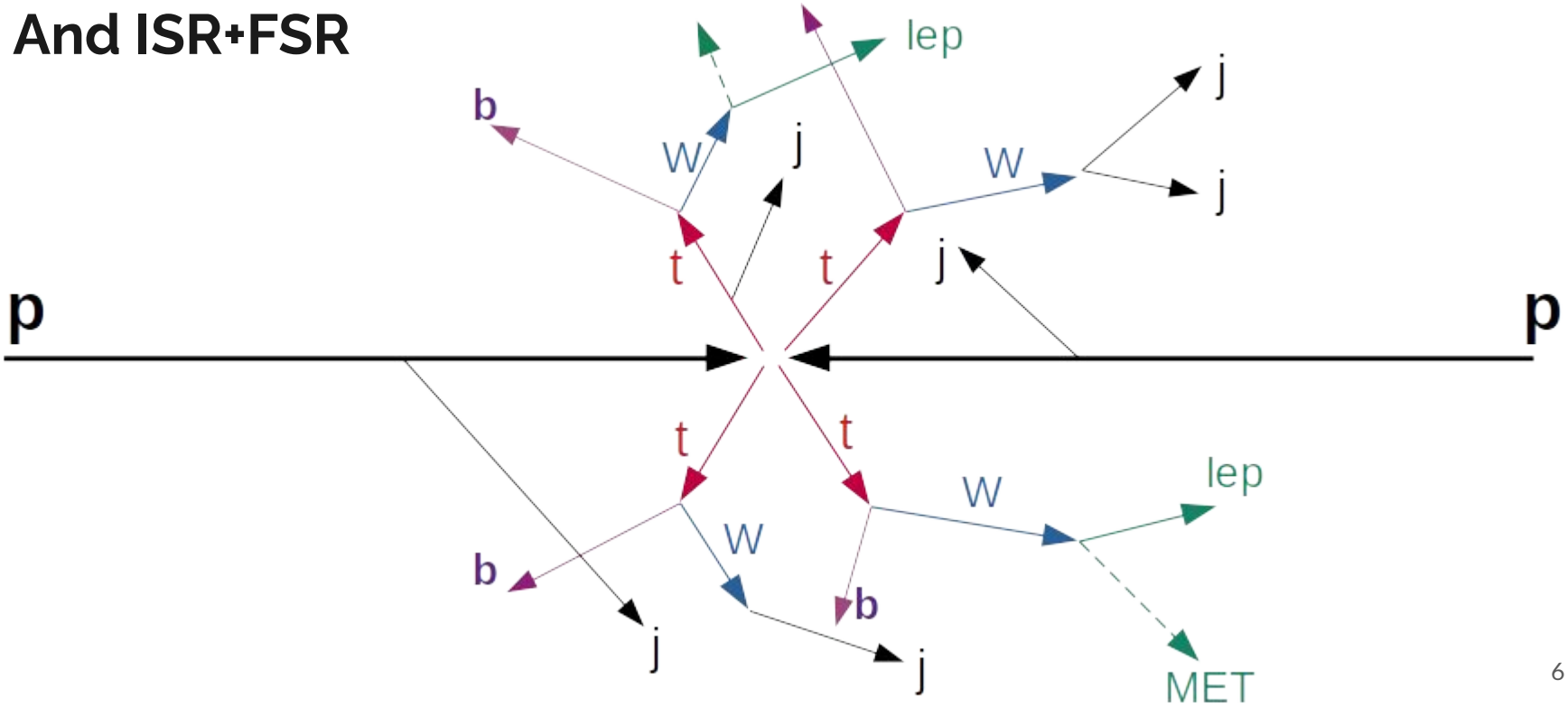
Already a very populated partonic final state



We only see top decays



And ISR+FSR



Four tops at the LHC

An increasingly sensitive SM benchmark to be explored at the LHC, with huge experimental effort and impressive results. Recent results in monolepton + 2LOS (ATLAS coll., ATLAS-CONF-2021-013, CMS coll., arxiv:1906.02805) and 2LSS + multilepton (ATLAS coll., arxiv:2007.14858. and CMS coll., arxiv:1908.06463)

Correspondingly, state of the art calculations and improvement of SM predictions (see R. Frederix, D. Pagani, M. Zaro, arxiv:1711.02116)

A theoretically well motivated (but perhaps more importantly, still allowed) window to BSM effects (see e.g. G. Banelli, E. Salvioni, J. Serra, T. Theil, A. Weiler, arxiv:2010.05915, E. Alvarez, A. Juste, M. S., T. Vazquez Schroeder, arxiv:2011.06514 and Luc Darmé, Benjamin Fuks, Fabio Maltoni, arxiv:2104.09512).

Theoretical calculations are really challenging

From R. Frederix, D. Pagani, M. Zaro, arxiv:1711.02116:

NLO corrections are large + Mixed corrections are comparable to pure QCD.

→ Theoretical calculations are expensive and necessary.

Large, accidental cancelations of (N)LO terms which involve QCD+EW couplings. Clear scale dependence of the accidental cancellations.

→ Hard to assert how BSM would change these cancellations. Need for very expensive simulations for each BSM model? (See Benjamin Fuks' talk about the subtleties of using EFTs!)

Four tops at the LHC

An increasingly sensitive SM benchmark to be explored at the LHC, with huge experimental effort and impressive results. Recent results in monolepton + 2LOS (ATLAS coll., ATLAS-CONF-2021-013, CMS coll., arxiv:1906.02805) and 2LSS + multilepton (ATLAS coll., arxiv:2007.14858. and CMS coll., arxiv:1908.06463)

Correspondingly, state of the art calculations and improvement of SM predictions (see R. Frederix, D. Pagani, M. Zaro, arxiv:1711.02116)

A theoretically well motivated (but perhaps more importantly, still allowed) window to BSM effects (among many examples, NP: 2105.03372,1910.09581,1906.09703, 1805.108351804.05598,1611.05032,1206.3064,1203.5862, 1112.3778,1107.46161101.1294, 1008.3562,hep-ph/9507411, EFT: 2104.09512, 2011.15060,2010.05915,1708.05928,1010.6304

Four tops in the 2LSS+multilepton channel

Signal and irreducible backgrounds (mainly $t\bar{t}Z$, $t\bar{t}H$ and specially $\mathbf{t\bar{t}W^{\pm}}$ + heavy-flavour) are of the same order of magnitude.

Reduced complexity of the multijet final state compared to single lepton. Added complexity of other source of MET.

Still very challenging!

Four tops in the 2LSS+multilepton channel

CMS measures something along the lines of the SM cross-section: $12.6^{+5.8}_{-5.2}$ fb

While ATLAS measures 24^{+7}_{-6} fb

It's still compatible with SM predictions but the central value is roughly twice as expected

Four tops in the 2LSS+multilepton channel

The main sources of uncertainty are:

High multiplicities final states are challenging to simulate.

Misaligned charge + Fake/non-prompt leptons need to be estimated with data-driven techniques.

Charge asymmetric + High b-jet multiplicity discrepancies are addressed by using a Normalization Factor for $t\bar{t}W^{\pm}$: $NF_{ttW}=1.6\pm0.3$ for ATLAS and 1.3 ± 0.2 for CMS.

Also… not a whole lot of events (~ 300 events for 140fb^{-1})

Four tops in the 2LSS+multilepton channel

What can we do?

Improving on our knowledge of 4-top and its irreducible backgrounds would be ideal. However, obtaining the $t\bar{t}W^{\pm}$ +HF and 4-top cross-sections, and kinematical distributions, to a higher precision is a daunting task.

Maybe reduce Monte Carlo dependency by learning from the Signal Region directly using semi- or unsupervised techniques.

Learning four tops

We will focus on the 2LSS++ channel and consider only four tops and $t\bar{t}W^{\pm}$. $t\bar{t}Z$ and $t\bar{t}H$ could be included easily.

In this channel, we have roughly 1/1 events.

Let’s focus on the technique more than in the simulation dataset.

Learning four tops

We can resume our goal as the following:

Consider our imperfect simulations of the (known) physical processes as **prior knowledge** and **update** our knowledge using the **measured data** → Bayesian framework

Learning four tops

As seen in E. Alvarez, D. A. Faroughy, J. F. Kamenik, R. Morales, A. Szynkman in arxiv:1611.05032, **N_b and N_j** are the low level observables that drive the discriminatory power.

Let's focus on those! **Keep things simple.**

We consider a probabilistic mixture model → each event is generated by one of underlying the physical processes.

Because we cannot observe this assignment, it is a **latent parameter** z.

Modelling four tops

For event n, we measure $N_j = j_n$ and $N_b = b_n$.

We want to model $p(j_n, b_n)$ as a mixture of two processes: background ($t\bar{t}W^{\pm}$) and signal (four tops). This is achieved by writing the likelihood as:

Modelling four tops

For event n, we measure $N_j = j_n$ and $N_b = b_n$.

We want to model $p(j_n,b_n)$ as a mixture of two processes: background ($t\bar{t}W^{\pm}$) and signal (four tops). This is achieved by writing the likelihood as:

$$p(j_n, b_n) = \Sigma_t \, p(j_n, b_n | z_n = t) p(z_n = t)$$

Modelling four tops

For event n, we measure $N_j = j_n$ and $N_b = b_n$.

We want to model $p(j_n,b_n)$ as a mixture of two processes: background ($t\bar{t}W^{\pm}$) and signal (four tops). This is achieved by writing the likelihood as:

$$p(j_n, b_n)=p(j_n, b_n|z_n=t\bar{t}W^{\pm})p(z_n=t\bar{t}W^{\pm})+p(j_n, b_n|z_n=\text{four tops})p(z_n=\text{four tops})$$

Modelling four tops

$p(z_n = \text{four tops})$ is the probability of an event originating from a four tops hard process

$\rightarrow S/(S+B) = \pi_1$; $p(z_n = t\bar{t}W^{\pm}) = \pi_0 = 1 - \pi_1$.

$p(j_n, b_n | z_n = t)$ are the probability mass functions we would **ideally** obtain from simulation

$\rightarrow (j, b) \sim \text{Multinomial}(\gamma_{t,(j,b)})$ where $\gamma_{t,(j,b)}$ is a matrix of dimension $2 \times (d_j \times d_b - 1)$

Modelling four tops

However, we do not really know any of these parameters.

(Roughly) Frequentist approach: obtain the best parameters using a ML fit.

(Roughly) Bayesian approach: treat these parameters as random variables with prior probability distributions that need to be updated with the data as encoded in the likelihood (Bayes theorem)

Modelling four tops

Posterior = Likelihood x Prior / Evidence

Modelling four tops

$$p(\{\gamma_{t,(j,b)},\pi_t\}|j,b) = [\Sigma_m\, p(j,b|\gamma_{m,(j,b)}, z=m)\pi_m] \times p(\{\gamma_{t,(j,b)},\pi_t\}) / p(j,b)$$

Modelling four tops

But there is a problem: We have a 2x(djxdb-1)+1 parameters to infer. But we have one measurement per event → We do not have enough information to disentangle!

We can solve this by assuming conditional independence between N_j,N_b. That is:

$$p(j, b|z=t) = p(j|z=t)p(b|z=t) = \text{Multi}(\alpha_t)\text{Multi}(\beta_t)$$

Modelling four tops

We are assuming that the correlations between N_j and N_b come from the fact that there is a mixture of processes. We need to learn this mixture to learn the correlations. If supervised, it would be Naive Bayes.

This is very different from:

$$p(j, b|z=t)p(t) = p(j|z=t)p(b|z=t)p(t) \neq p(j)p(b)$$

Modelling four tops

The likelihood is now:

$$p(j,b) = \Sigma_t p(j|z=t)p(b|z=t)p(t) = \Sigma_t \pi_t \alpha_{t,j} \beta_{t,b}$$

Which implies an specific covariance matrix between j and b

$$C_{j,b} = \Sigma_{t,t'} (\pi_t \delta_{t,t'} - \pi_t \pi_{t'}) \alpha_{t,j} \beta_{t',b}$$

Modelling four tops

We go from $2x(d_j x d_b - 1)+1 \rightarrow 2x(d_j + d_b - 2)+1$ parameters

So now Bayes theorem looks like...

$$p(\{\alpha_{t,j}, \beta_{t,b}, \pi_t\}|j, b) = [\Sigma_m \pi_m \alpha_{m,j} \beta_{m,b}] \times p(\pi_{m'}) \prod_{m'} p(\alpha_{m',j}) p(\beta_{m',b}) / p(j,b)$$

Modelling four tops

We need to specify the priors. We consider the conjugate prior of the Multinomial: the Dirichlet.

$$\mathrm{Dir}(\theta|\eta) = \prod_v \theta_v^{\eta v - 1} / \mathrm{B}(\eta)$$

We map our prior knowledge to the Dirichlet hyperparameters.

The MC simulations yield estimations on the parameters $\theta = \pi$, α, and β $\rightarrow$ Expected values under the prior distribution

Modelling four tops

We can fix the expected values given by our MC by parameterizing $\eta_k = \Sigma\, p_k$ where p_k is the parameter value estimated through MC simulations and Σ a total scaling factor which encodes our confidence in the prior estimations.

Looking at the mean and variance of a given possible outcome θ_k:

$$E[\theta_v]_\eta = p_v \qquad Var[\theta_v]_\eta = p_v(1-p_v)/\Sigma$$

Putting it all together: Generative process

We can understand it using a **plate diagram** which encodes the **generative process** of the data:

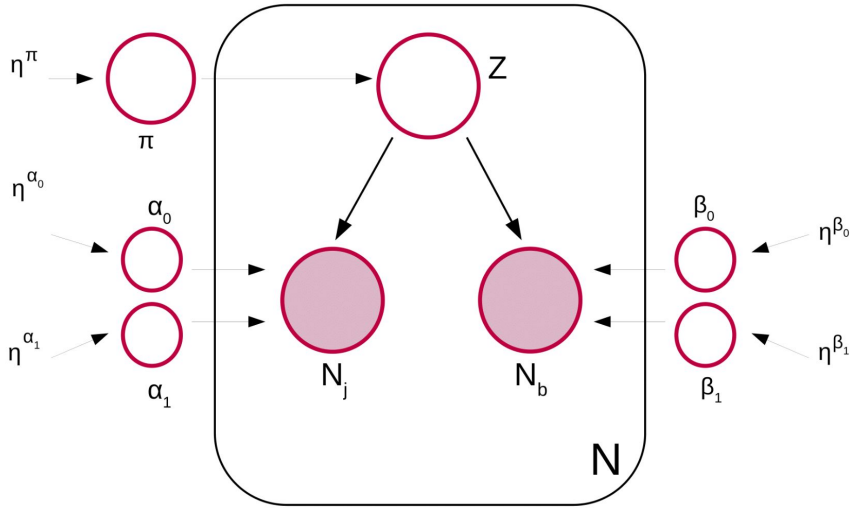
Sample fractions $\pi_0, \pi_1 \sim \text{Dir}(\eta^{\pi})$

For t=1,2:

- Sample light jet multinomials $\alpha_t \sim \text{Dir}(\eta^{\alpha t})$
- Sample b-jet $\beta_t \sim \text{Dir}(\eta^{\beta t})$

For event n=1,..,N:

- Sample event assignment $z_n \sim \text{Multi}(\pi_0, \pi_1)$
- Sample $j_n \sim \text{Multi}(\alpha_{zn})$
- Sample $b_n \sim \text{Multi}(\beta_{zn})$



Learning four tops

The thing is: we cannot do the inference procedure exactly. The evidence is intractable: a lot of possible assignments and thus updates of the prior.

Learning four tops

The thing is: we cannot do the inference procedure exactly. The evidence is intractable: a lot of possible assignments and thus updates of the prior.

Luckily, there is a vast literature on the subject and a lot of techniques. EM+priors for finding the MAP, Variational Inference for approximate, fast and analytical inference and Markov Chain Monte Carlos for “exact” numerical inference.

Learning four tops

The thing is: we cannot do the inference procedure exactly. The evidence is intractable: a lot of possible assignments and thus updates of the prior.

Luckily, there is a vast literature on the subject and a lot of techniques. EM+priors for finding the MAP, Variational Inference for approximate, fast and analytical inference and Markov Chain Monte Carlos for “exact” numerical inference.

Our model is so simple (everything is either multinomial or dirichlet!) that we are able to write the latter in python without the need to resort to dedicated software (such as pymc, emcee or pyro).

A simple benchmark

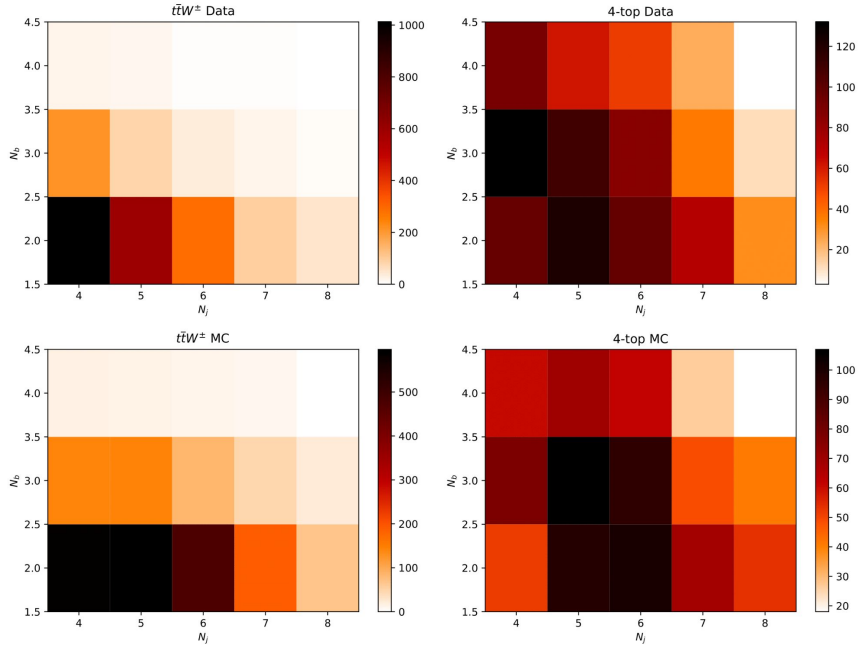
We consider the case with N=500, f_1=0.30 (roughly a Luminosity of 800 fb^{-1}).

To study our algorithm we use MC samples as “true data” and a smeared version as our “MC” which plays the role of prior.

We make an exception for the signal fractions, for which we assume no prior knowledge.

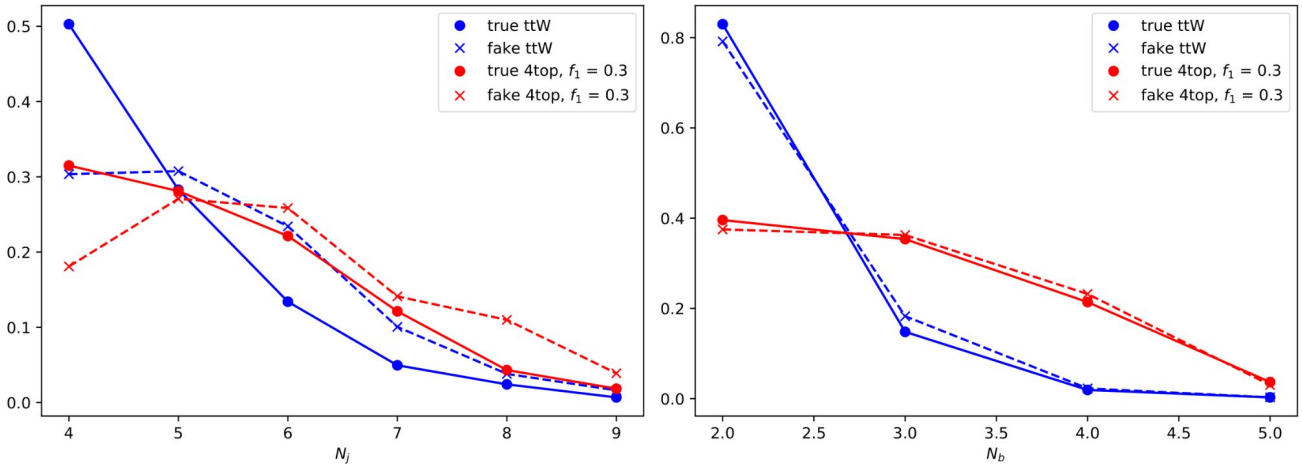
A simple benchmark

This is the two dimensional distributions



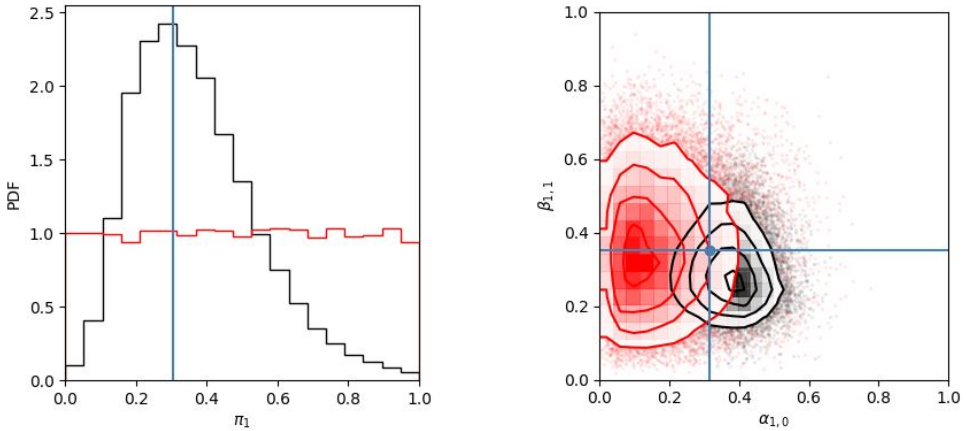
A simple benchmark

And here are the 1d projections we use.



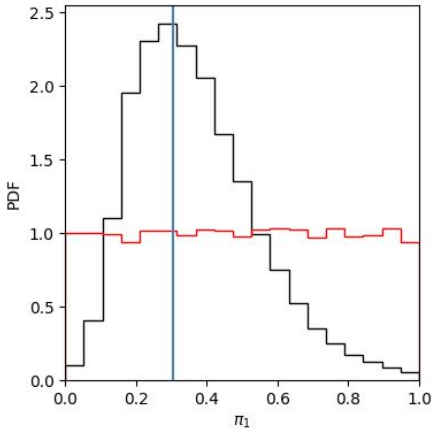
Corner plots

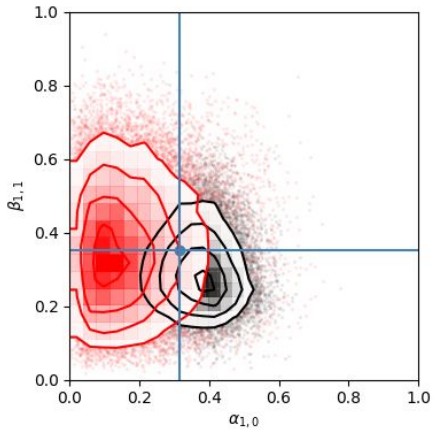
For each parameter, we show its 1d histogram and its 2d correlations with every other parameter



Corner plots

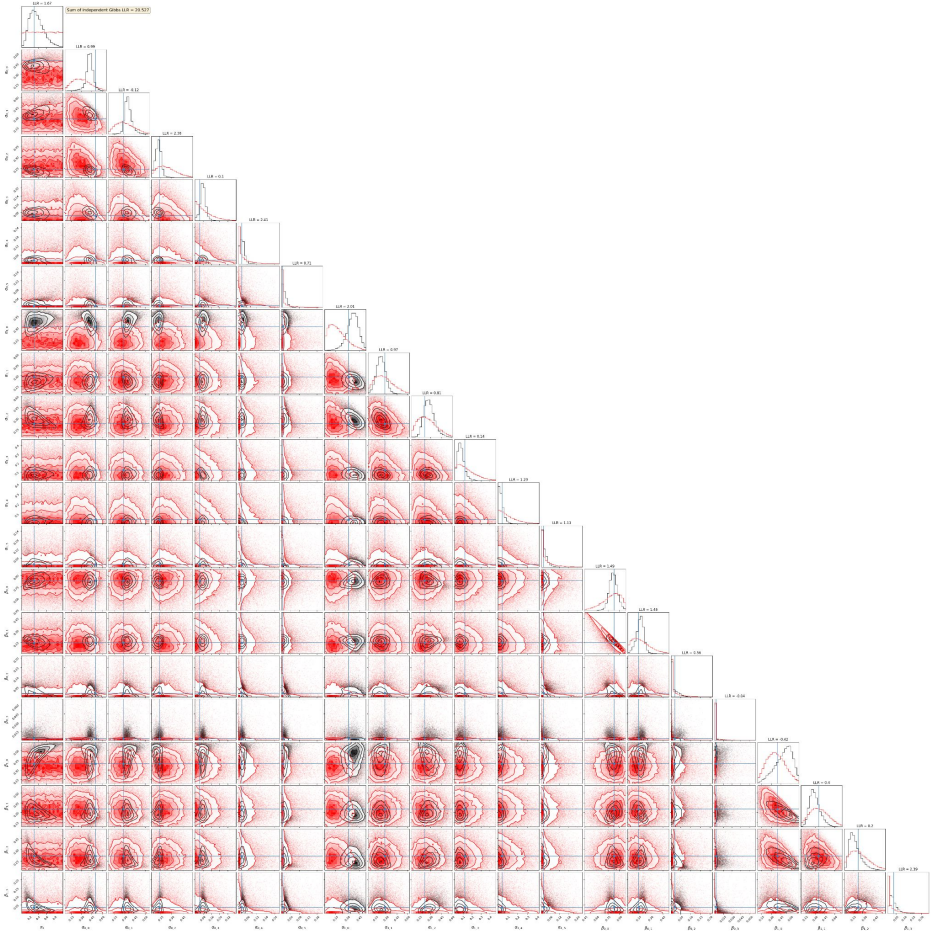
For each parameter, we show its 1d histogram and its 2d correlations with every other parameter





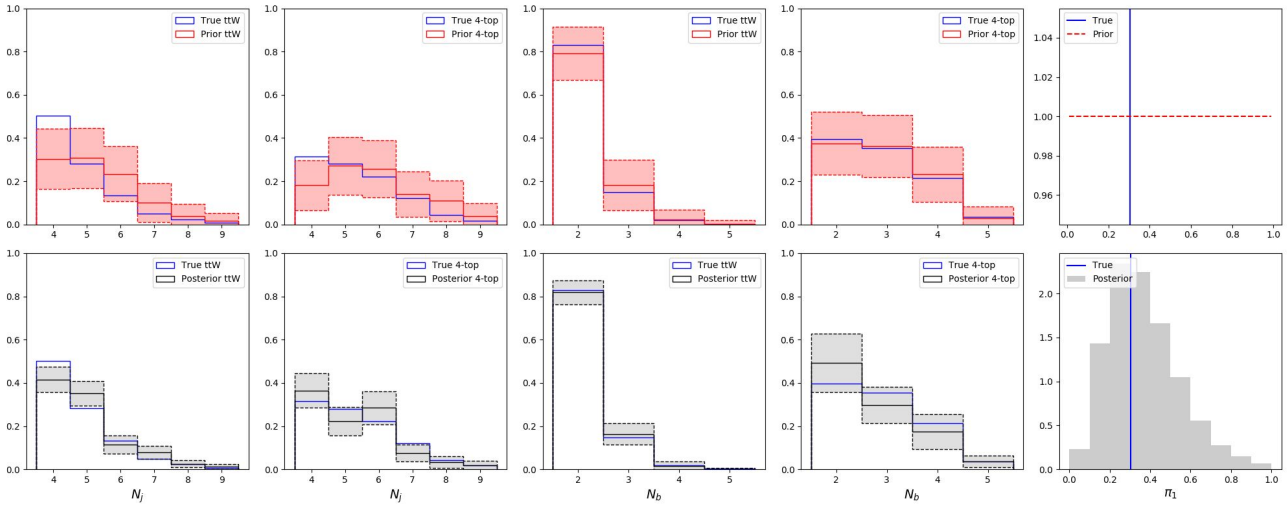
This is a particularly good example where we capture correlations

There are a lot of them...



Grouping parameters together

Now we can compare prior vs posterior



General observations

From this and other benchmarks:

Good convergence + uncertainty reduction!

N_j is easier to fix. N_b is harder and π is the hardest.

N_j has a lot of very populated bins. This is not the case for N_b.

The limitations in π probably reflect the limitations of our modelling (and the use of a non-informative prior). However, once we have N_j and N_b simulations we can trust, we can obtain π in the usual manner.

As the main problem is obtaining N_j, this algorithm seems relevant.

Evaluating our results

To analyse the results of the inference, we consider the log-likelihood ratio of the correct parameters given the prior and the posterior probabilities.

In a real application, we do not know the true values. However, this is where the Bayesian framework is useful. There is a vast literature on model selection techniques!

Because we have a generative model, sanity checks and interpretability are straightforward.

Conclusions

We are able to model the data with a generative process. Conditional independence (which could be broken by systematics!) is a key modelling assumption

This allows us to improve on our Monte Carlo estimations by treating them as prior knowledge which is to be updated through event measurements.

We are able to correct the N_j distributions properly

What could we do next?

Tune our Monte Carlo generators on signal regions

Measure the signal cross-section with reduced systematics

Test for NP effects

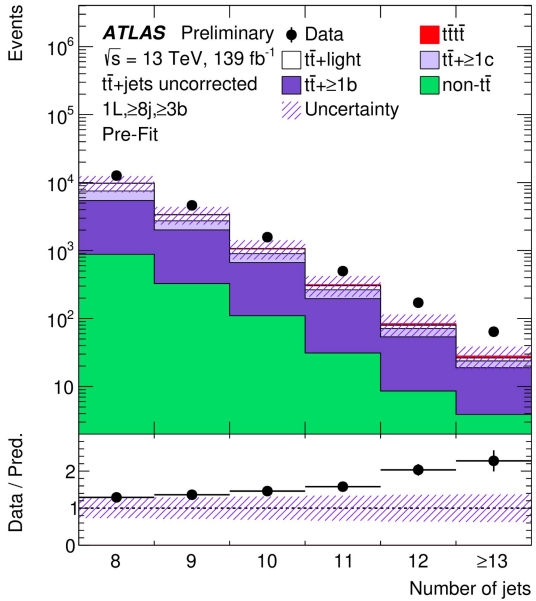
Adapt to other channels

Hvala!

Back-up slides

From ATLAS 1L+2LOS search

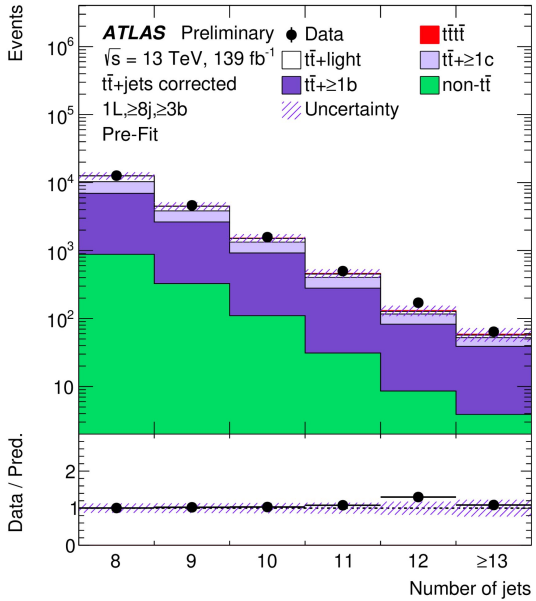
Uncertainties related to the background simulations at large N_j



From ATLAS 1L+2LOS search

Corrected through various clever techniques that need to trust that the MC extrapolates between different regions

They introduce additional systematics



Four tops in the 2LSS+multilepton channel

Let's look under the hood!

From E. Alvarez, A. Juste, M.S. and T. Vazquez Schroeder arxiv:2011.06514

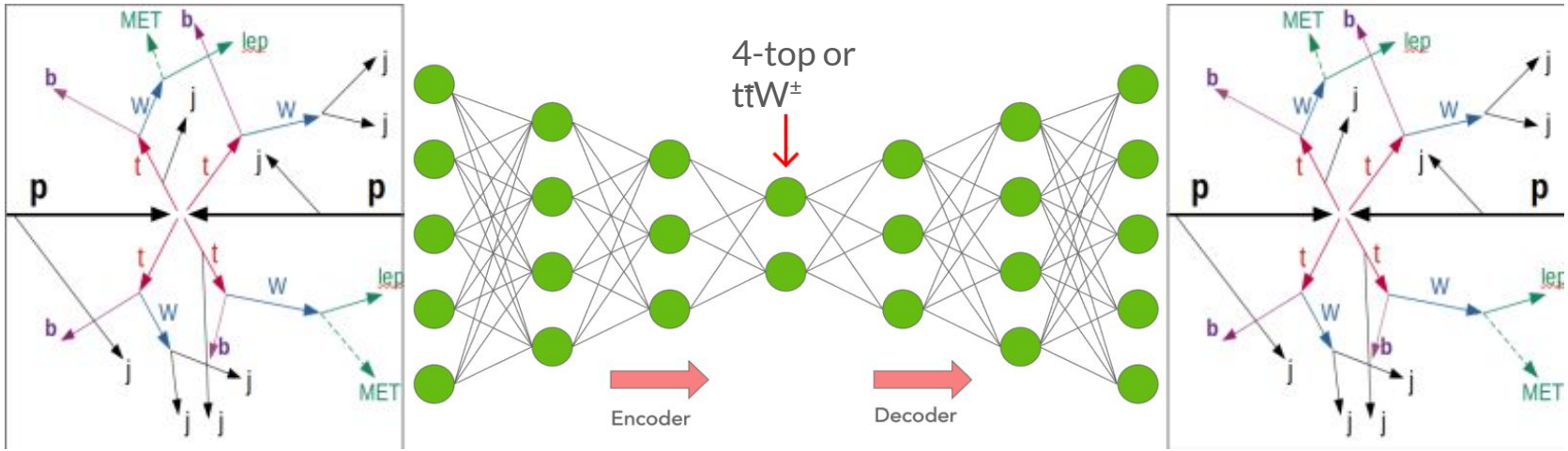
Very similar strategies... but not exactly the same.

four-top-quarks	ATLAS		CMS	
	2LSS	$\geq$ 3L	2LSS	$\geq$ 3L
Total lepton charge	$\pm$2	-	$\pm$2	-
Lepton p_T [GeV]	28 (all ℓ)		25/20	25/20/20(/20)
Number of jets and b-jets	$\geq$ 6j $\geq$ 2bj (77% eff.)		$\geq$ 6j $\geq$ 2bj OR 5j $\geq$ 3bj (55-70% eff.)	$\geq$ 5j $\geq$ 2bj OR 4j $\geq$ 3bj (55-70% eff.)
H_T [GeV]	$>$ 500		$>$ 300	
$\lvert m_{e^\pm e^\pm}\rvert$ (2LSS) or $\lvert m_{OSSF}\rvert$ (3L) [GeV]	$>$ 15	-	$>$ 12	
$\lvert m_{e^\pm e^\pm} - m_Z\rvert$ (2LSS) or $\lvert m_{OSSF} - m_Z\rvert$ (3L) [GeV]	$>$ 10		-	$>$ 15
Other	-		Missing transverse momentum cuts	

Table 3: Comparison of event selections between the ATLAS [5] and CMS [6] four-top-quarks analyses. H_T is the scalar p_T sum of jets, leptons and b-jets.

Learning 4-top

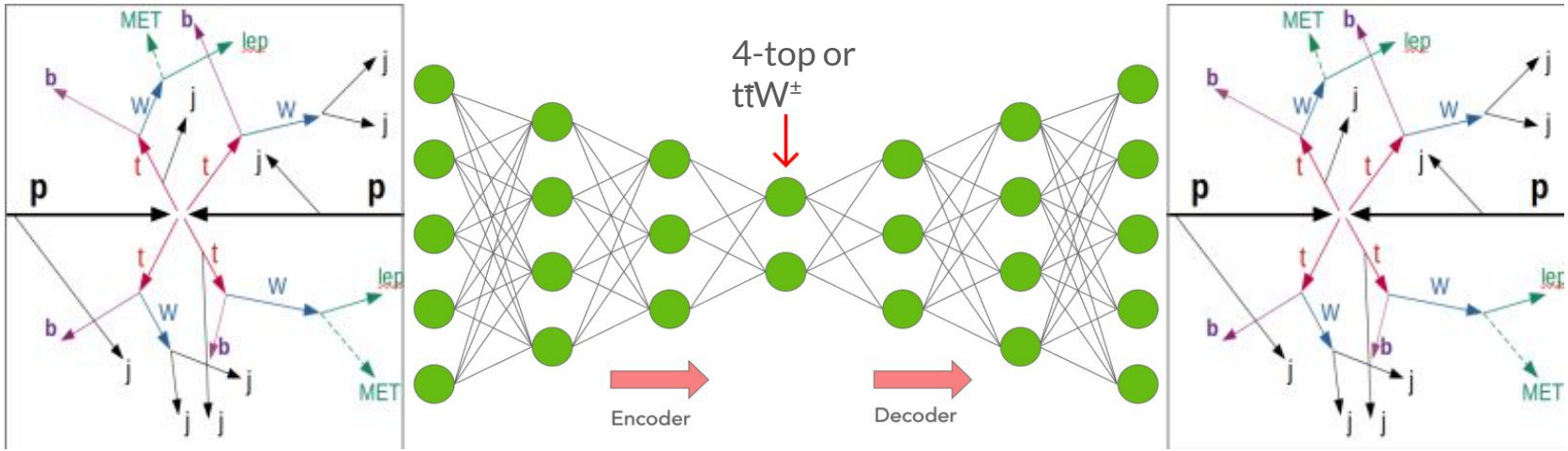
We tried various techniques. LDA, **AE**, VAEs



Learning 4-top

We used a lot of kinematical information, mixing N_b, N_j with p_T, ΔR and E

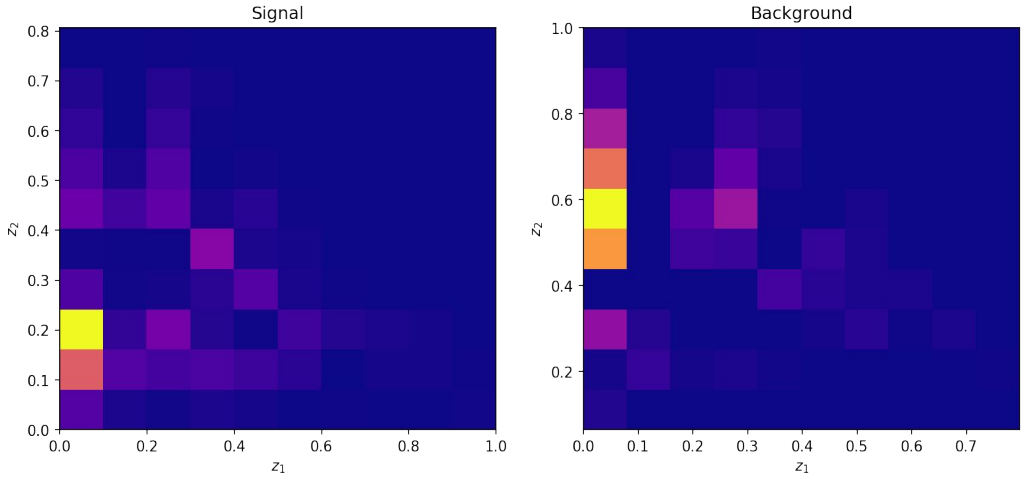
We tried various techniques. LDA, **AE**, VAEs



Learning 4-top

We used a lot of kinematical information, mixing N_b, N_j with p_T, ΔR and E

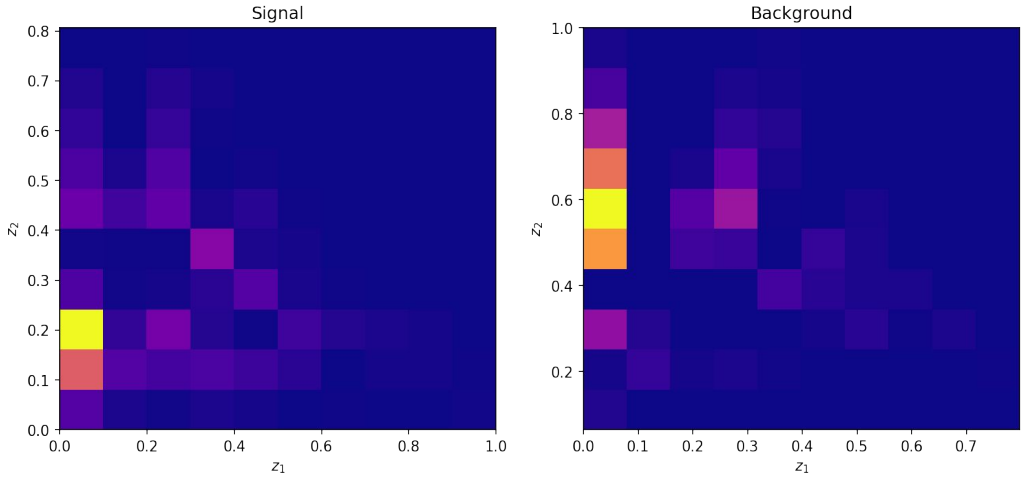
We find a latent space with islands: some clustering



Learning 4-top

We find a latent space with islands: some clustering

But the discriminatory power is mainly in N_b and N_j as in E. Alvarez, D. A. Faroughy, J. F. Kamenik, R. Morales, A. Szynkman in arxiv:1611.05032



Gibbs sampler

We obtain T “independent” samples of parameters drawn from the posterior.

Any expected value $E_{z,\pi,\alpha,\beta}[f(z,\pi,\alpha,\beta)]$ can be approximated by the mean over the samples $\Sigma_i f(z_i,\pi_i,\alpha_i,\beta_i)/T$

We can plot the marginalized distributions simply by drawing histograms on the relevant parameters.

We need the conditional distributions of each parameter conditioned on the others $p(\theta_v|\theta_{\setminus v}) \rightarrow$ Really simple because we have only Multinomials and Dirichlets.

Gibbs sampler

To start a given iteration t, we only need the sufficient statistics $N^{(t)}_{kjb}$.

$N^{(t)}_{k,j,b}$ is the number of measurements of j and b assigned to class k. It can be obtained merely by having the event class assignments.

From there we can obtain $N = \Sigma_k N_k = \Sigma_{k,j} N^{(t)}_{k,j} = \Sigma_{k,b} N^{(t)}_{k,b} = \Sigma_{k,j,b} N^{(t)}_{k,j,b}$

So we start the Gibbs sampler with an initial random event class assignment $Z^{(0)}$ which we'll then forget about later

Gibbs sampler

So, for iteration t:

$$\pi^{(t)} \sim \mathrm{Dir}(\{\eta_k^{\pi} + N_k^{(t-1)}, k = 1, .., K\})$$

for $k = 1, \ldots, K$

$$\alpha_k^{(t)} \sim \mathrm{Dir}(\{\eta_j^{\alpha_k} + N_{kj}^{(t-1)}, j = 1, .., d_j\})$$

$$\beta_k^{(t)} \sim \mathrm{Dir}(\{\eta_b^{\beta_k} + N_{kb}^{(t-1)}, b = 1, .., d_b\})$$

and then for n=1,...,N

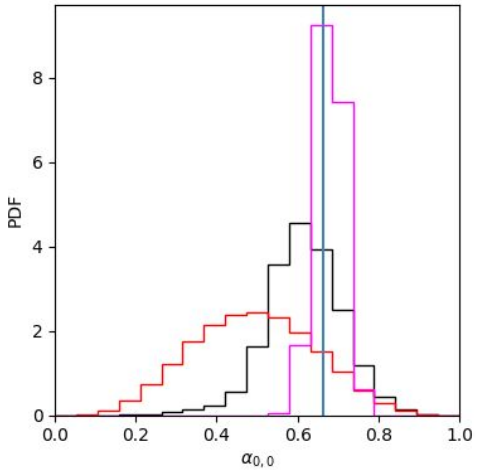
$$z_n^{(t)} \sim \mathrm{Multinomial}(\{\frac{\pi_k^{(t)} \alpha_{kj_n}^{(t)} \beta_{kb_n}^{(t)}}{\sum_{l=1}^{K} \pi_l^{(t)} \alpha_{lj_n}^{(t)} \beta_{lb_n}^{(t)}}, k = 1, \ldots, K\})$$

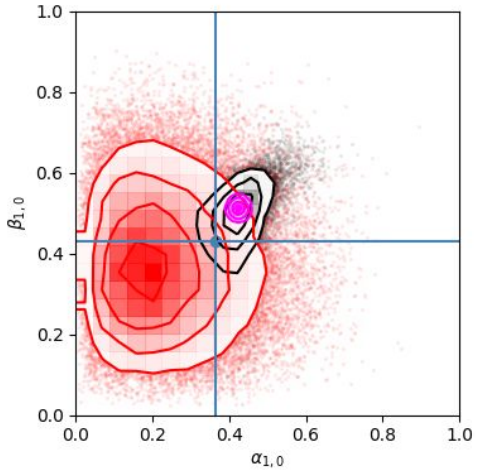
In practice you need to burn-in and thin the samples to get more or less independent samples.

Afterwards, you can get a lot of information: corrected distributions p(π,α,β|X), an event-by-event probabilistic tagger p(z_n|x_n) where we marginalize over π,α,β, etc.

Variational Inference

Variational Inference is an approximated inference technique which assumes certain factorizations





It is inherently limited to find what we need it to find! Distributions are too narrow

Machine Learning at the LHC

Although ML algorithms have been a staple of HEP analyses for a long time, there is a current boom of semi- and unsupervised algorithms applied to LHC physics (see e.g. M. Feickert, B. Nachman's Living Review, arxiv:2102.02770, the LHC Olympics 2020, arxiv:2101.08320 and The Dark Machines Anomaly Score Challenge, arxiv:2105.14027)

Among the many possible applications, these algorithms can enhance LHC analyses by reducing systematics of MC simulations and, for BSM searches, reducing the necessary model hypotheses.