Bayesian Machine Learning for Scientific Research

Maximizing information from data

Ezequiel Alvarez
sequi@unsam.edu.ar
ICTP-SAIFR October 2024

Previous lectures

Lecture 1

Bayesian: assume data being sampled from a PDF, infer its parameters and learn the internal structure of the data

To learn the PDF of the data and then being able to assess, predict , generate, etc.

More scientific

Lecture 1

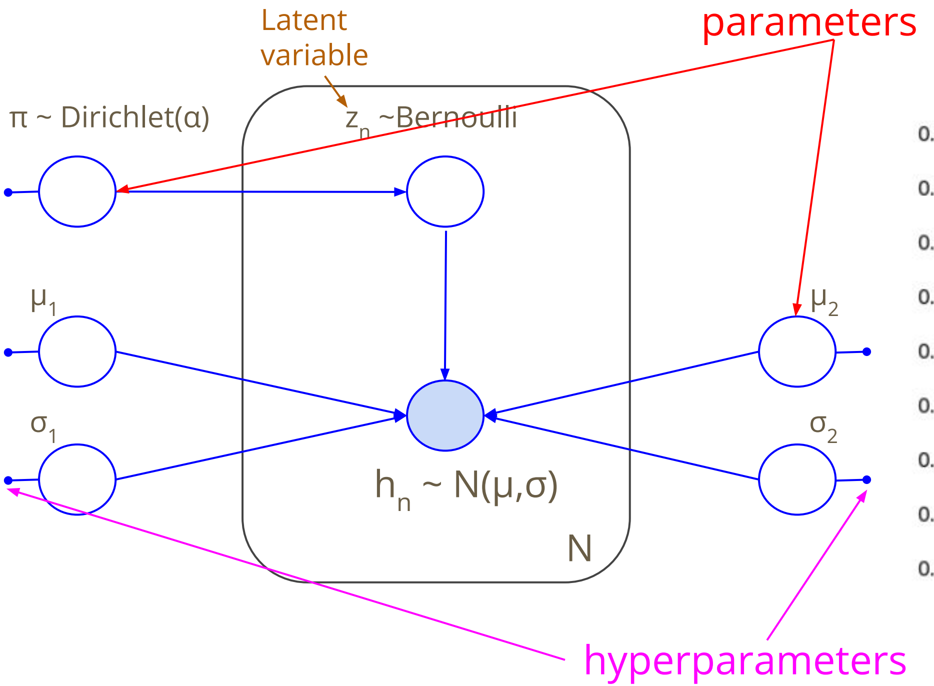
Bayes Theorem:

$$p(\theta|X) = \frac{p(X|\theta) * p(\theta)}{p(x)}$$

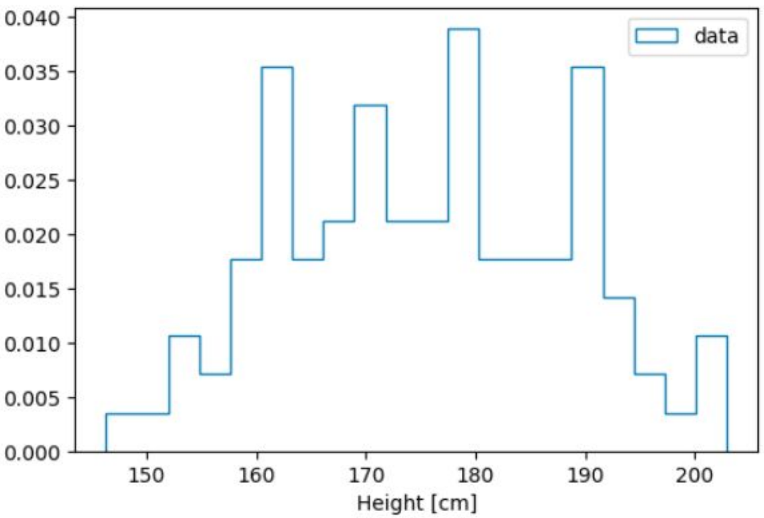
Our utility: X = data, θ=parameters

Model data as being sampled from a clever PDF with parameters θ → Infer θ once you see the data X → Connect θ to physical parameters of interest

Lecture 1

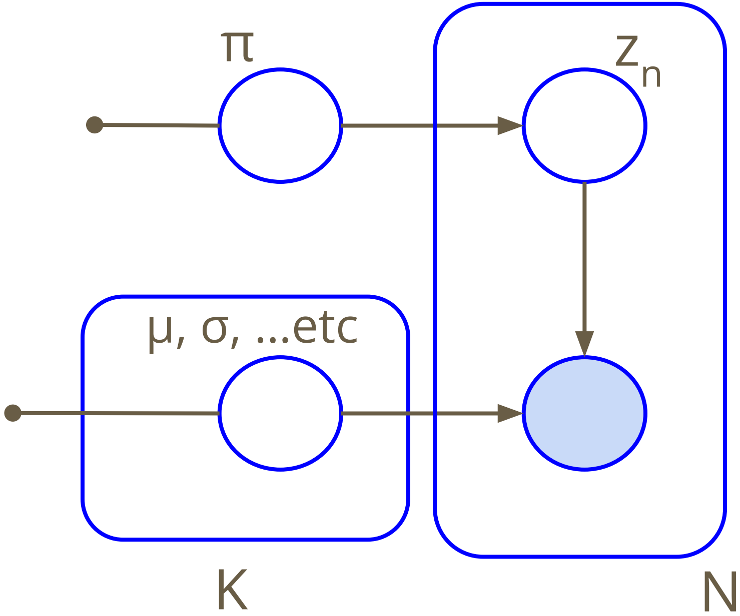


Graphical Model



1D Gaussian Mixture

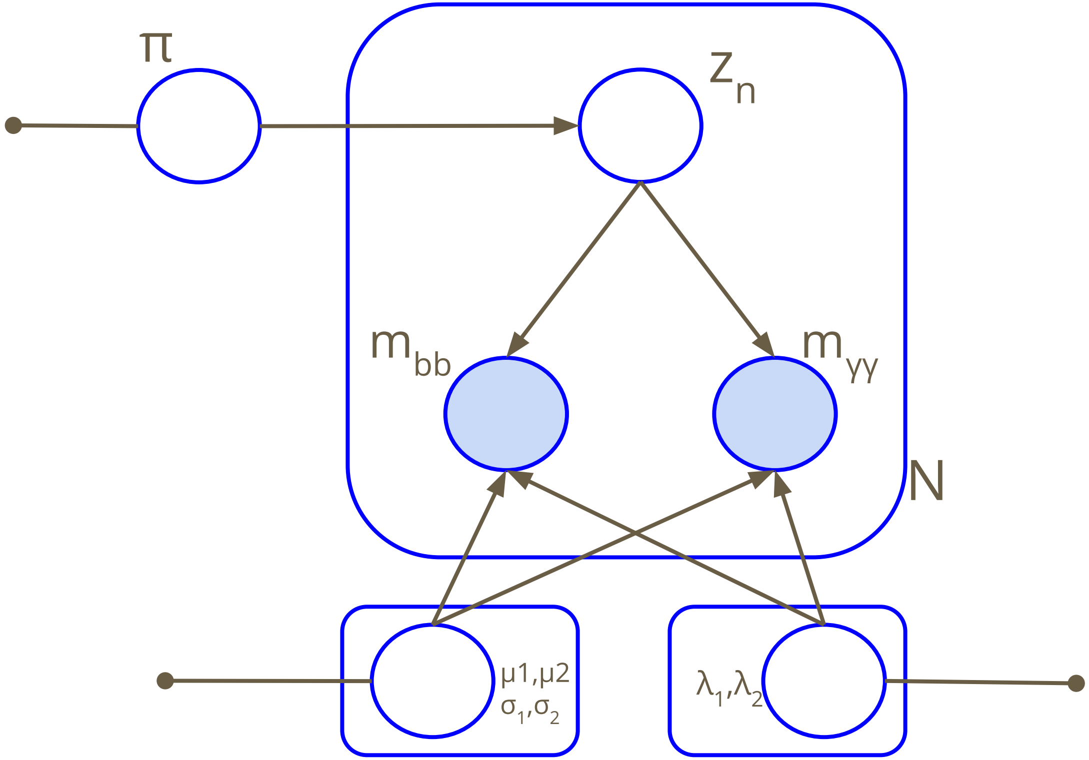
Lecture 3: Mixture Models

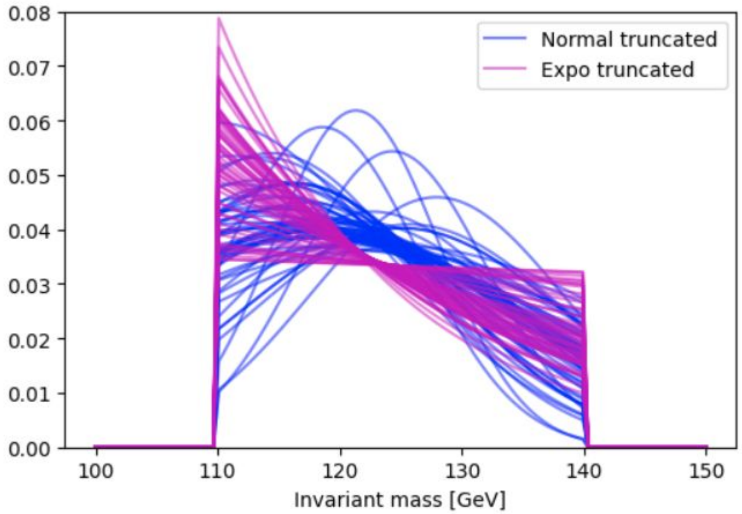


$$p(x_n|\Theta) = \sum_{k=1}^{K} \pi_k \, p(x_n|\theta_k)$$

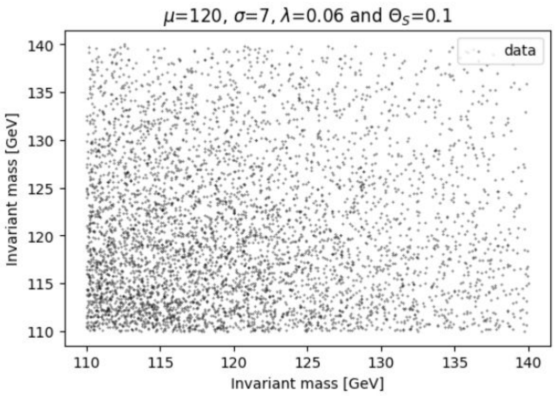
$$\log p(X|\Theta) = \sum_{n=1}^{N} \log(\sum_{k=1}^{K} \pi_k \, p(x|\theta_k))$$

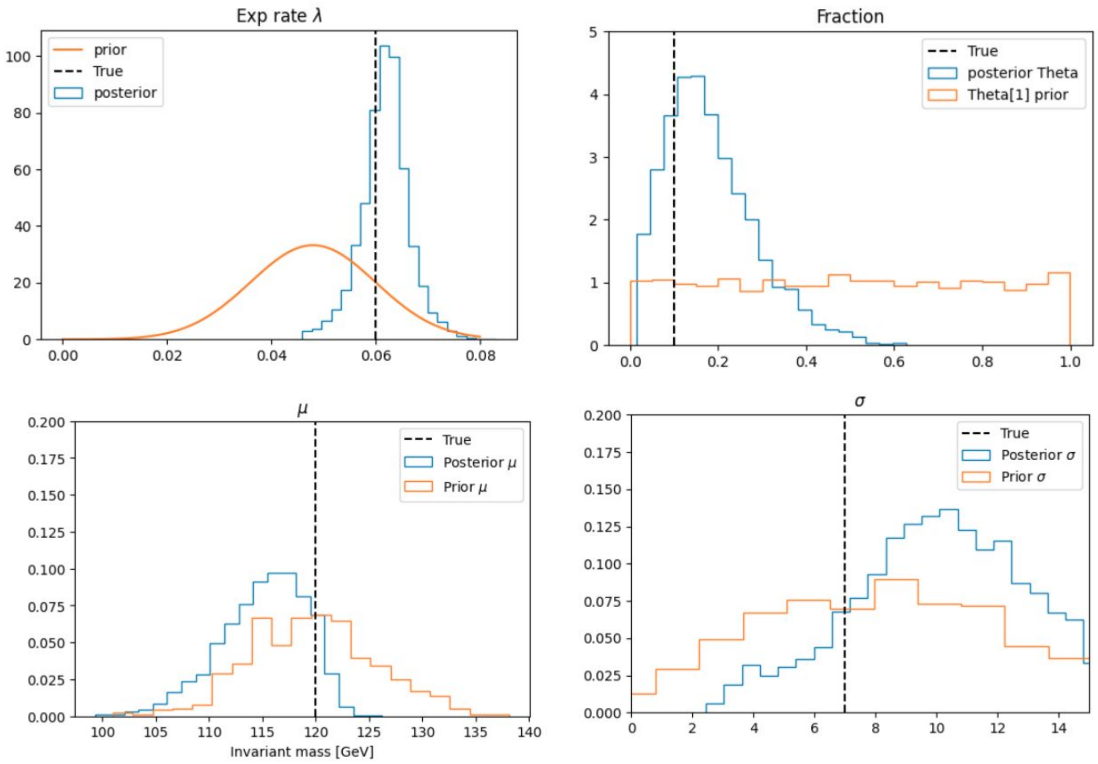
Lecture 3: 2D Mixture Model, hh → bbγγ



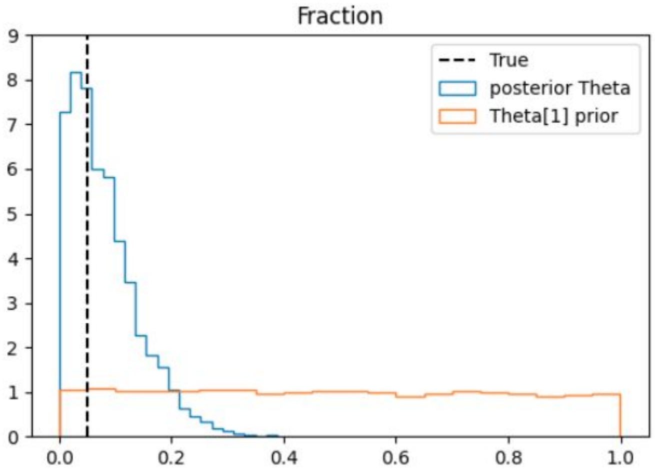


Lecture 3: The impossible.... @10%





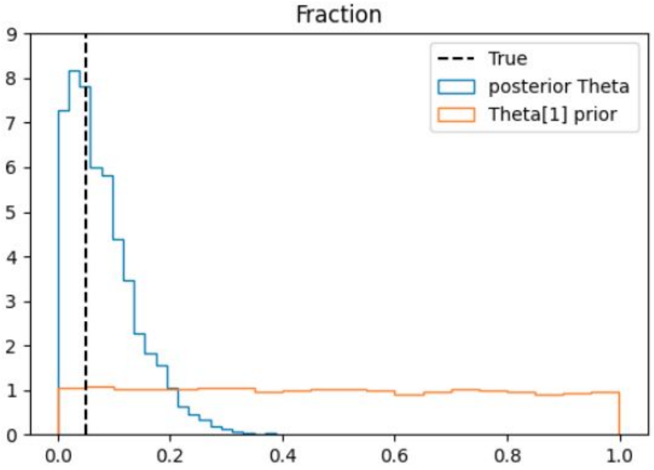
Assessment in Bayesian ML



Once you have a result from real data, how can you assess if...

- The sampling is unbiased
- The model is (fairly) correct

Assessment in Bayesian ML



Once you have a result from real data, you want to test that…

- The sampling is unbiased

Sampling diagnostics

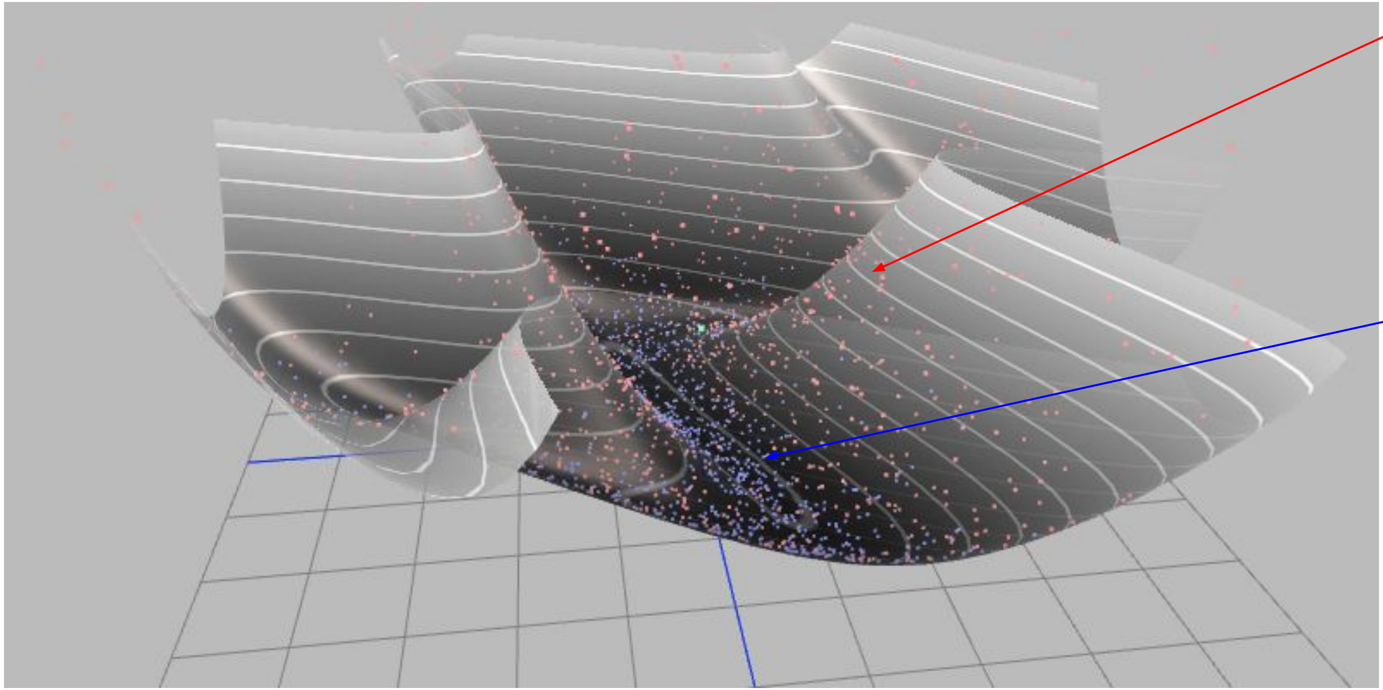
- The model works correctly

Fake data to test it

- The model is (fairly) correct for the data

Posterior Predictive Check

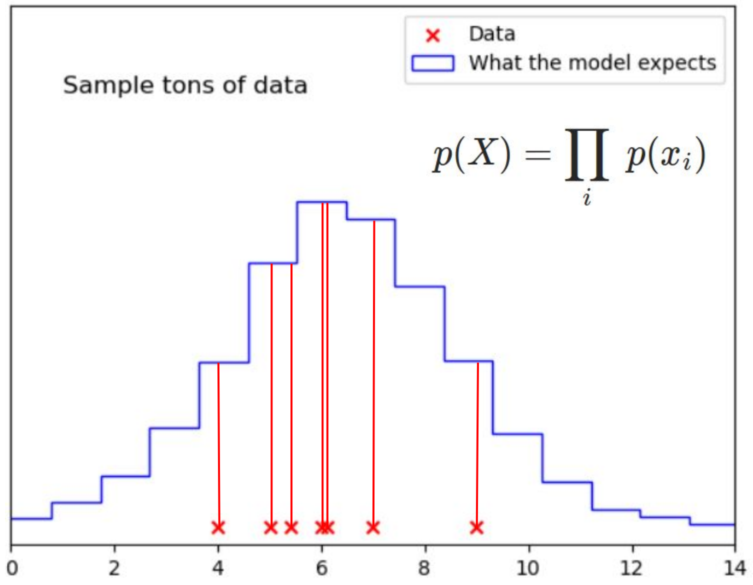
MCMC Sampling with Hamiltonian MC



Numerical integration yields some divergent trajectories (rejected)

Accepted samples

What is good and what is bad ?



$p = 10^{-6}$ … and now?

- What does it mean ?
- What do we compare it to ?

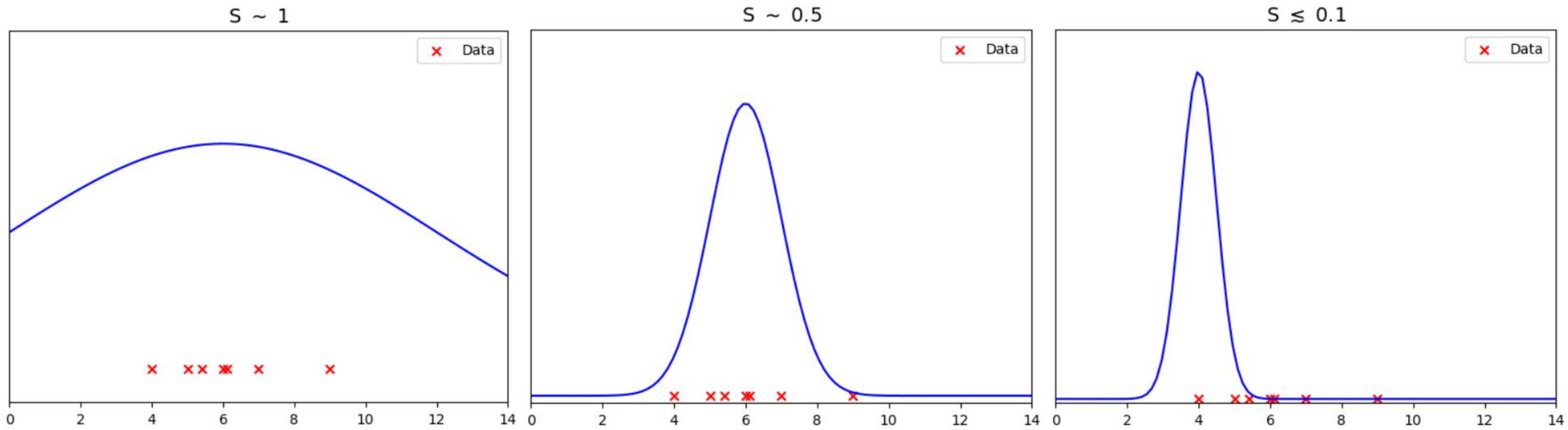
- Generate replicas of data X^{Rep}
- Compute their probability
- Compute

$$p\big(\, p(X^{rep}) < p(X) \,\big)$$

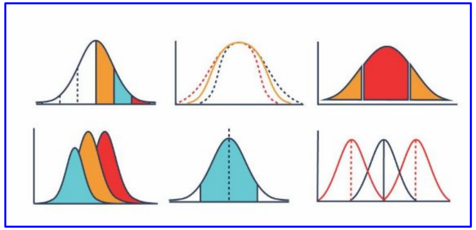
What is good and what is bad ?

$$\text{Score} = p\big(\, p(X^{rep}) < p(X) \,\big)$$

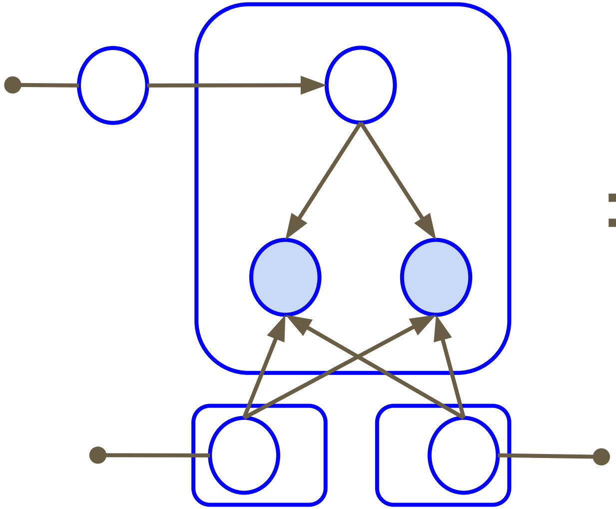
- S ~ 1 : bad model
- S ~ 0.5: good model
- S ≲ 0.1: bad model



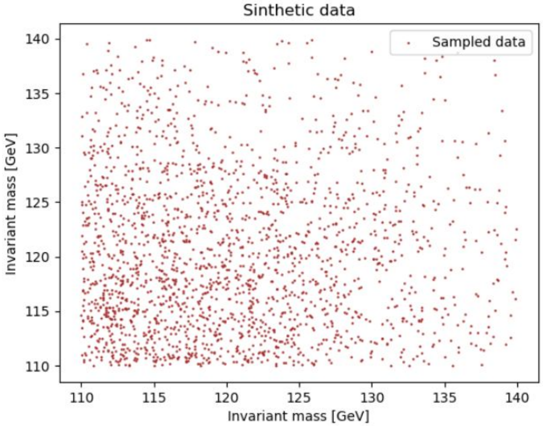
Posterior Predictive Check: pp → bbγγ



X

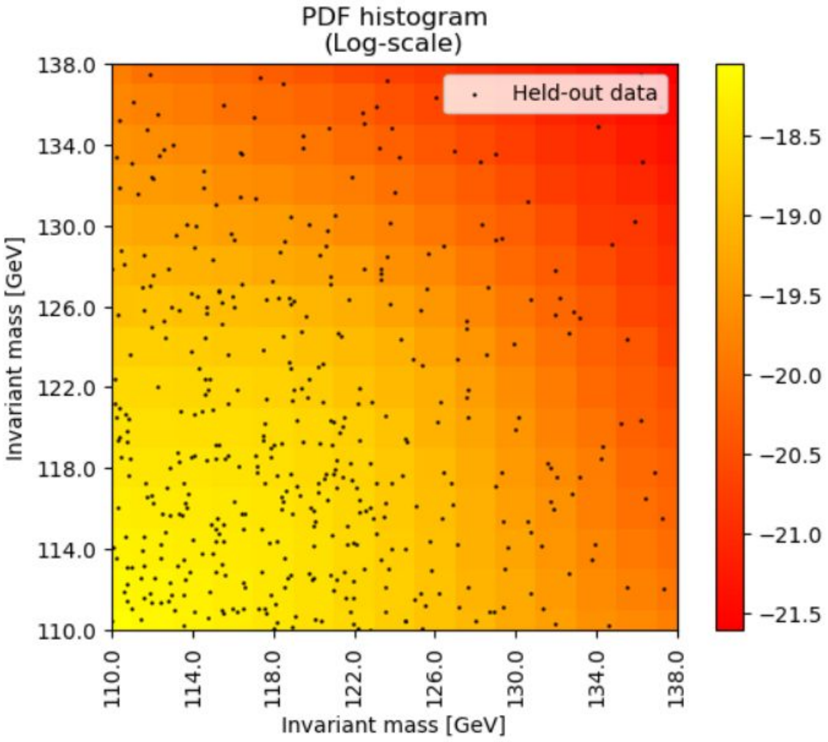


=



Posterior x model = Millions of sinthetic data !

PDF of your model given the data !

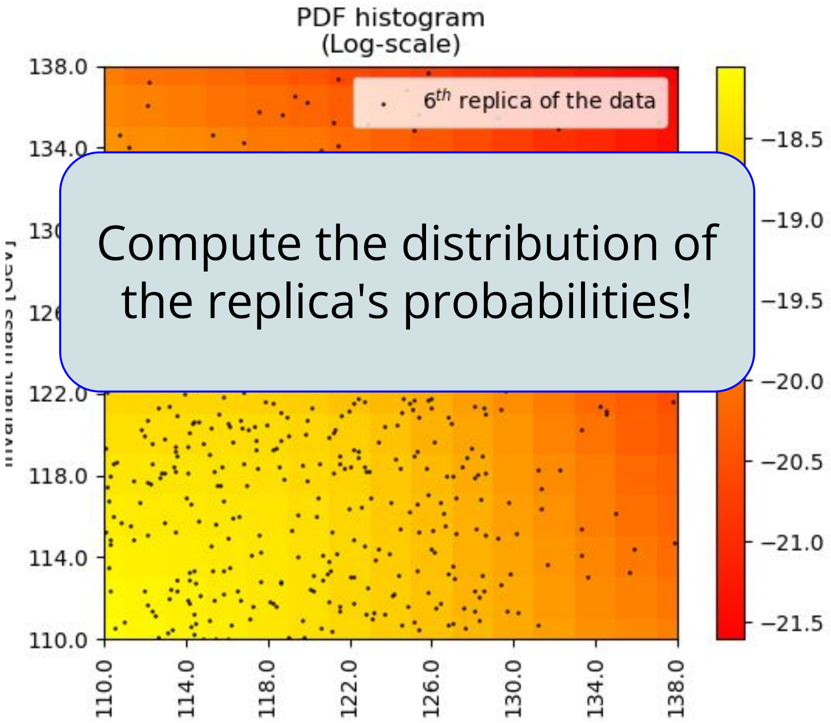


$$p(x_n|X) = \int p(x_n|z_i)\, p(z_i|X)\, dz_i$$

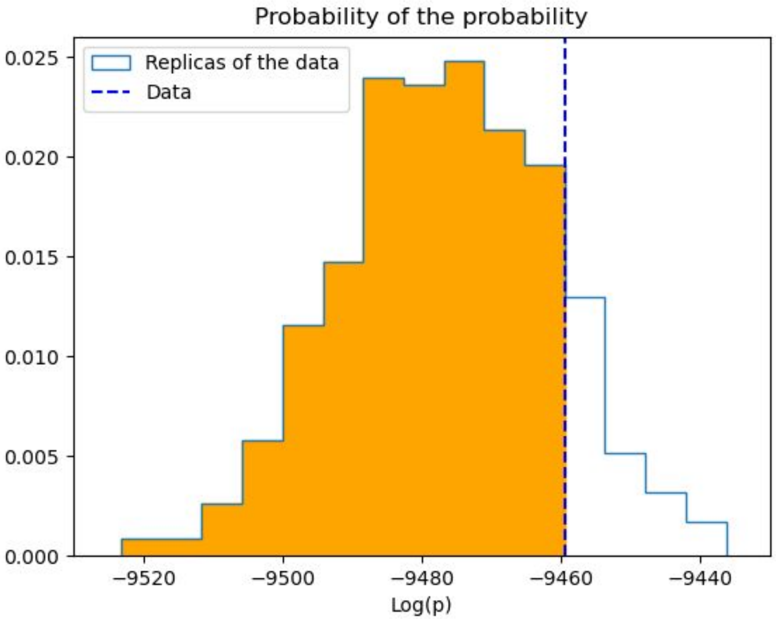
Log-scale!

Simply add the value of each bin for each held-out datapoint!

PDF of your model given the data !



Compute the distribution of the replica's probabilities!

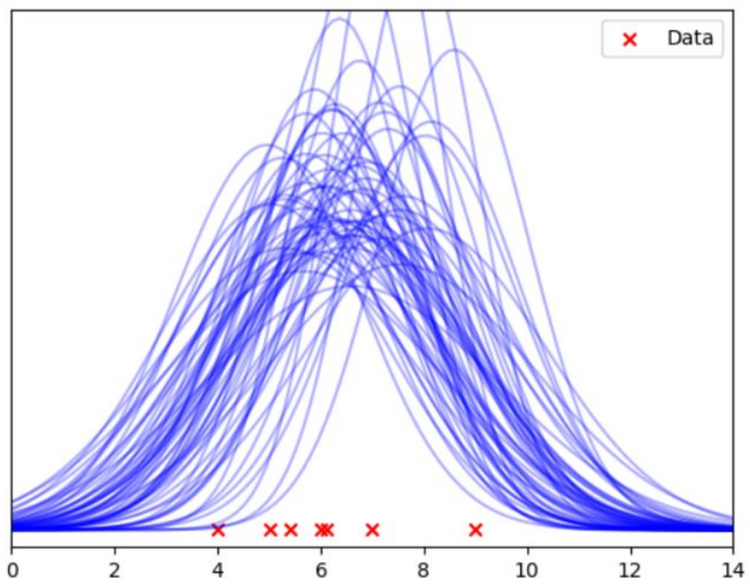


$$p\left(p(X_{held}^{rep}) < p(X_{held})\right) = 0.86$$

Inferring non-parametric distributions

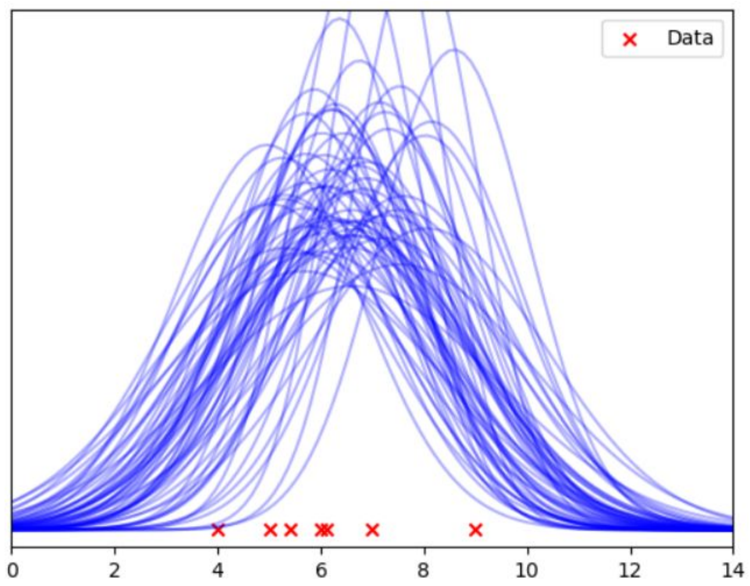
Lecture 5

Inferring non-parametric distributions



Predetermined shape:
→ then fewer parameters!

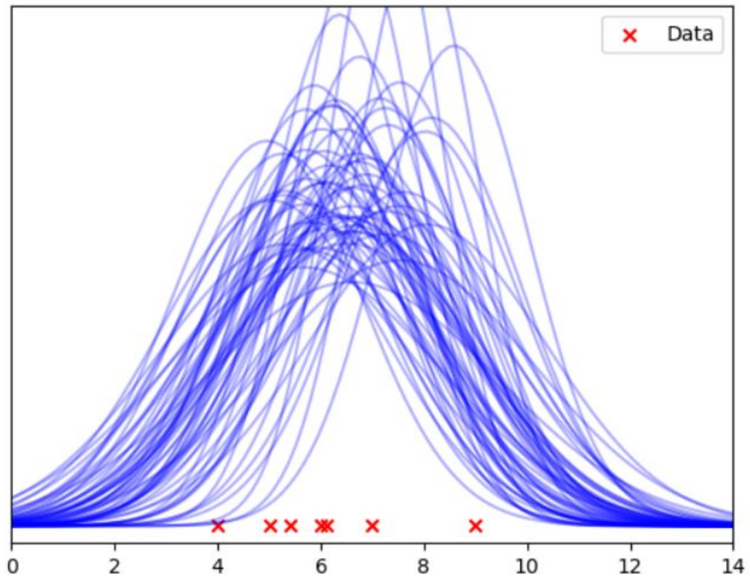
Inferring non-parametric distributions



Predetermined shape:
→ then fewer parameters!

N(μ, σ), β(a,b), γ(a,b), exp(λ), etc

Inferring non-parametric distributions

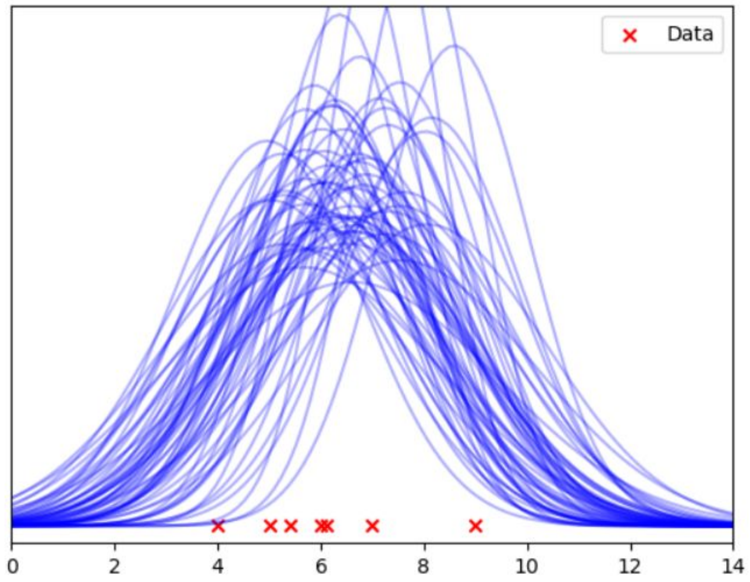


Predetermined shape:
→ then fewer parameters!

N(μ, σ), β(a,b), γ(a,b), exp(λ), etc

Real life is more sophisticated!

Inferring non-parametric distributions



Predetermined shape:
→ then fewer parameters!

N(μ, σ), β(a,b), γ(a,b), exp(λ), etc

Real life is more sophisticated!

- Smoothness
- Unimodal
- Decreasing/increasing
- etc

Inferring non-parametric distributions

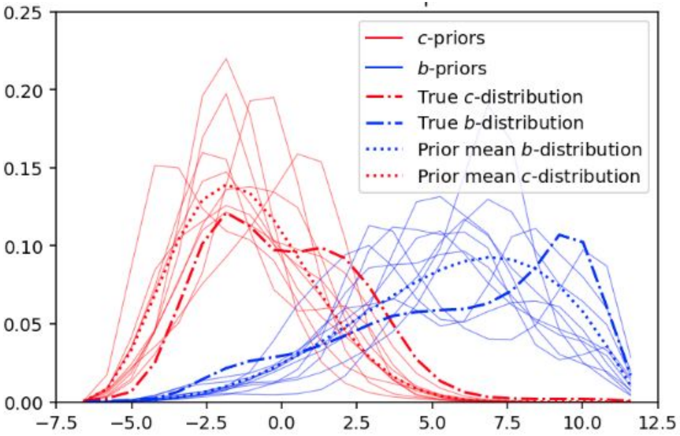
Can we sample arbitrary curves
whose only constraints are e.g.

Smoothness, unimodality, decreasing,
etc….?

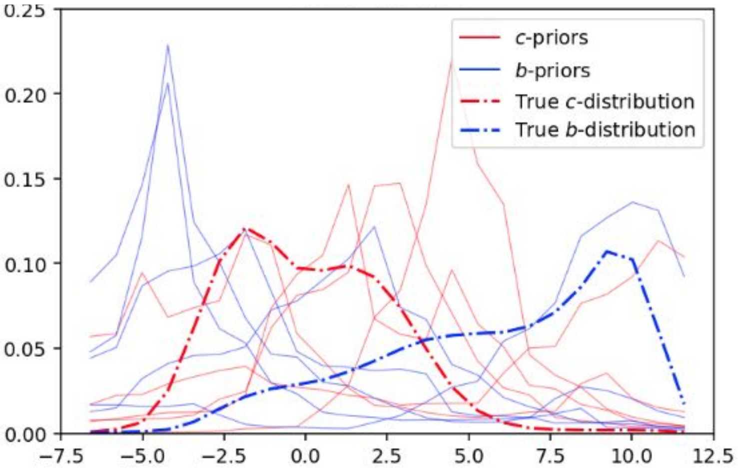
Inferring non-parametric distributions

Can we sample arbitrary curves whose only constraints are e.g.

Smoothness, unimodality, decreasing, etc….?



Smooth around some prior



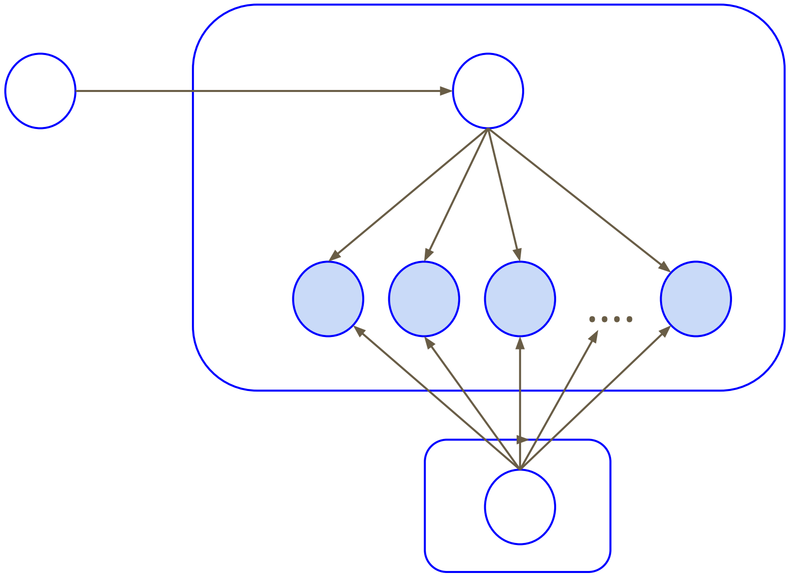
Smooth and unimodal

Inferring non-parametric distributions

Usually multi-dimension required
(otherwise ambiguities)

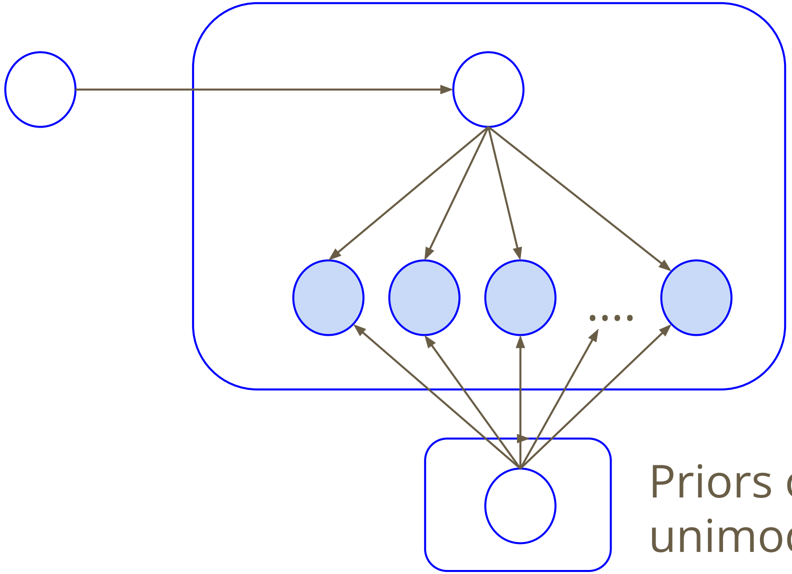
Inferring non-parametric distributions

Usually multi-dimension required
(otherwise ambiguities)



Inferring non-parametric distributions

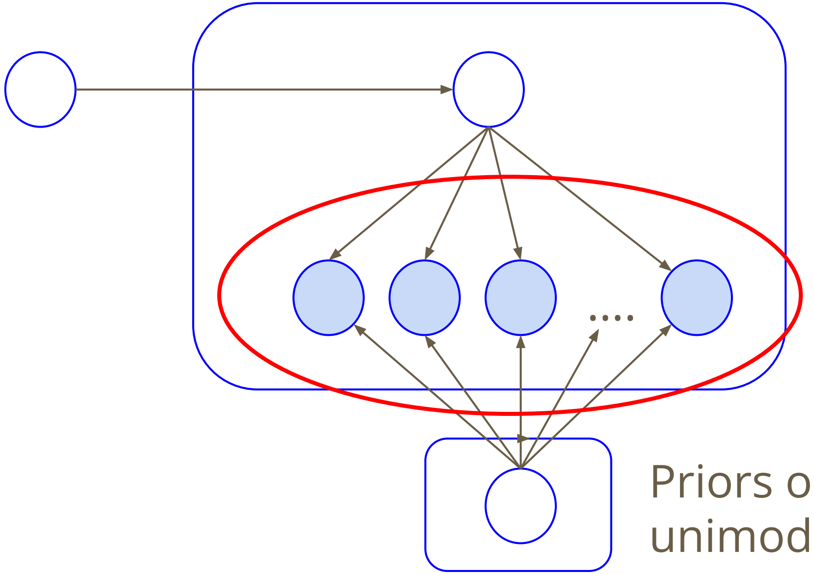
Usually multi-dimension required
(otherwise ambiguities)



Priors on these distributions!
unimodal, smooth, etc!

Inferring non-parametric distributions

Usually multi-dimension required
(otherwise ambiguities)



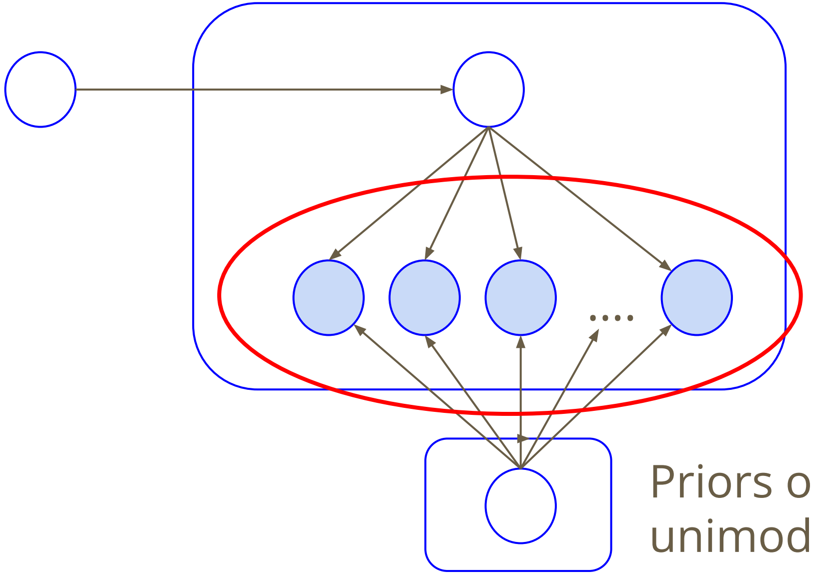
Often some prior knowledge on the structure

Priors on these distributions!
unimodal, smooth, etc!

Inferring non-parametric distributions

Usually multi-dimension required
(otherwise ambiguities)

Solve the unsolvable puzzle ?



Often some prior knowledge on the structure

Priors on these distributions!
unimodal, smooth, etc!

Mixture model for

Arbitrary, Smooth Unimodal distributions

Case: $pp \rightarrow hh \rightarrow bbbb$ (LHC physics)

Case: $pp \rightarrow hh \rightarrow bbbb$ (LHC physics)

Prior knowledge: events are either cccc, ccbb or bbbb

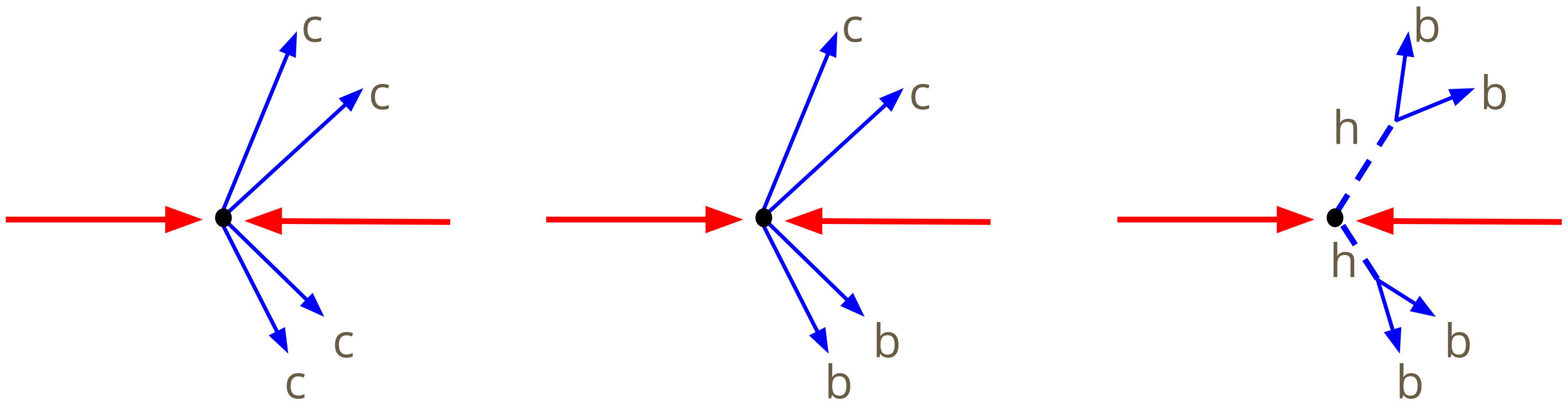
Case: pp→hh→bbbb (LHC physics)

Prior knowledge: events are either cccc, ccbb or bbbb

backgrounds   signal

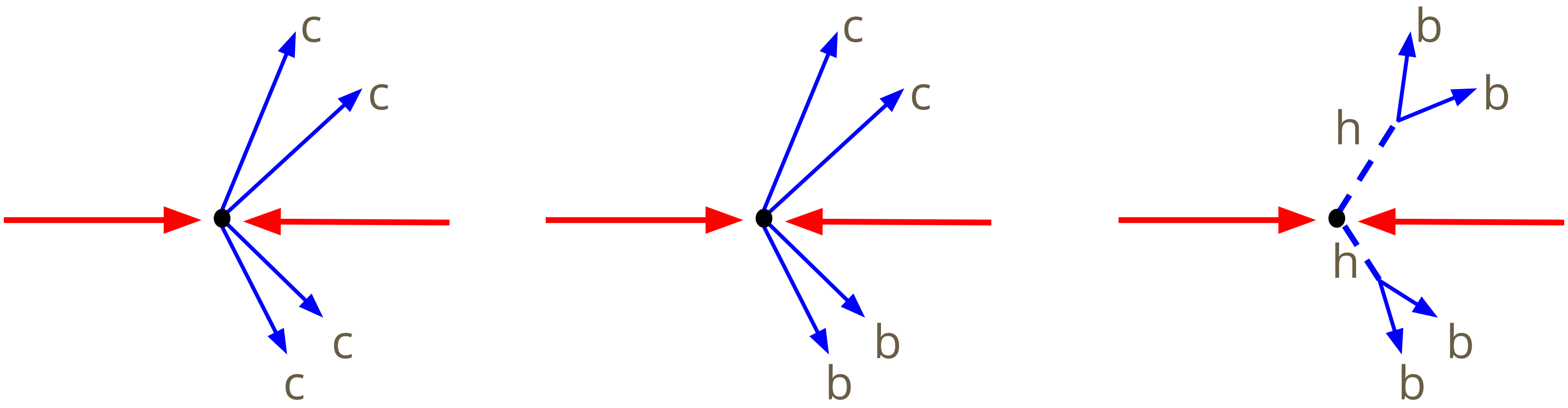
Case: pp→hh→bbbb (LHC physics)

Prior knowledge: events are either cccc, ccbb or bbbb



Case: pp→hh→bbbb (LHC physics)

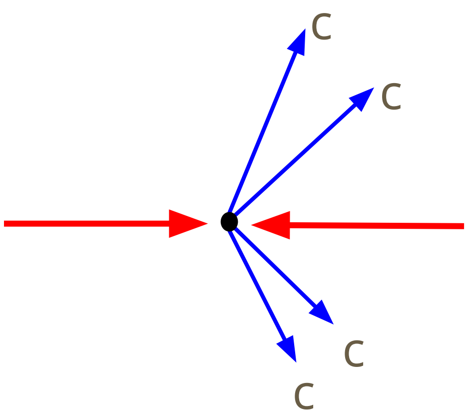
Prior knowledge: events are either cccc, ccbb or bbbb

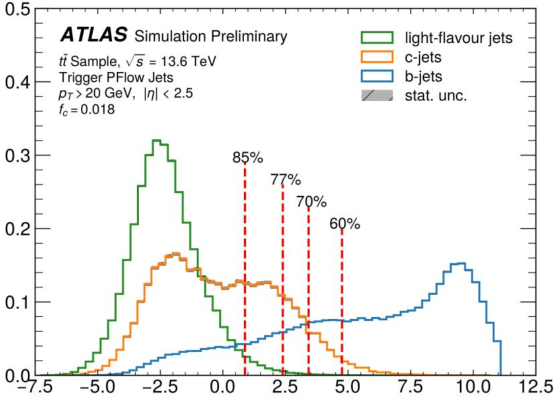


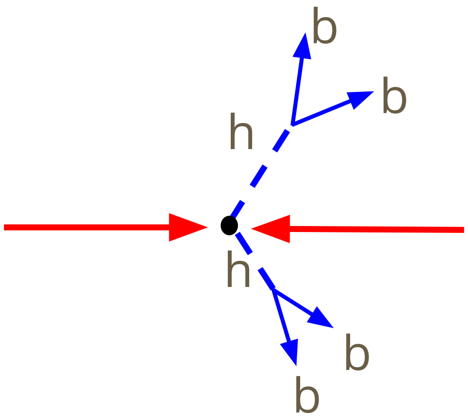
Underlying physics: they only come in pairs

Case: pp⟶hh⟶bbbb (LHC physics)

Prior knowledge: events are either cccc, ccbb or bbbb

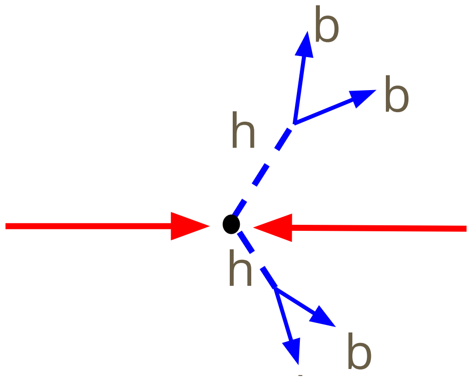






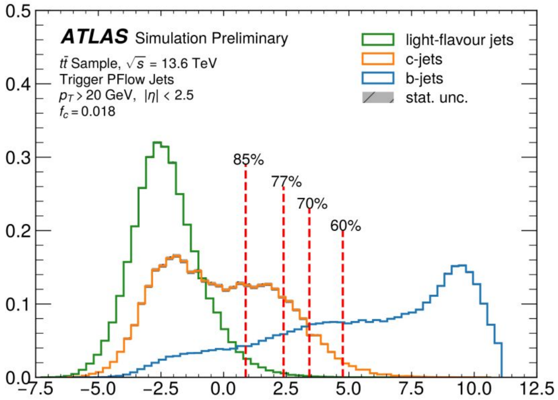
Underlying physics: they only come in pairs

Case: pp→hh→bbbb (LHC physics)

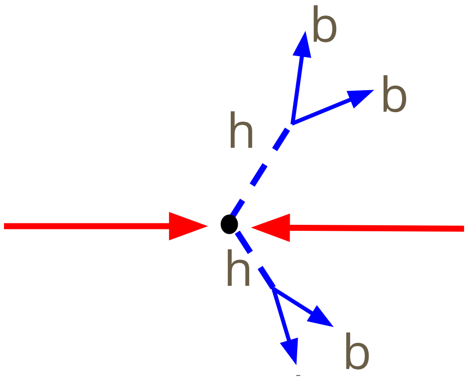


The data: N tuples of 4 numbers

X = [[3,7,2,12], [15,6,18,20].....[18,17,20,15]]



Case: pp→hh→bbbb (LHC physics)

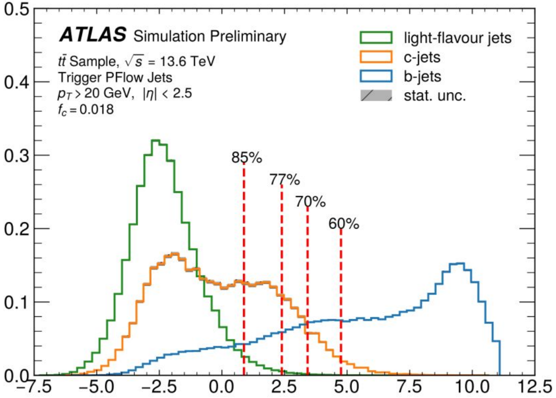


The data: N tuples of 4 numbers

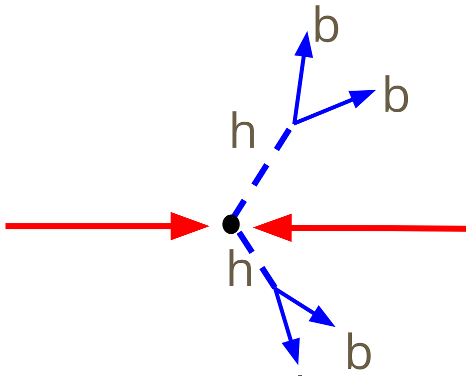
X = [[3,7,2,12], [15,6,18,20].....[18,17,20,15]]

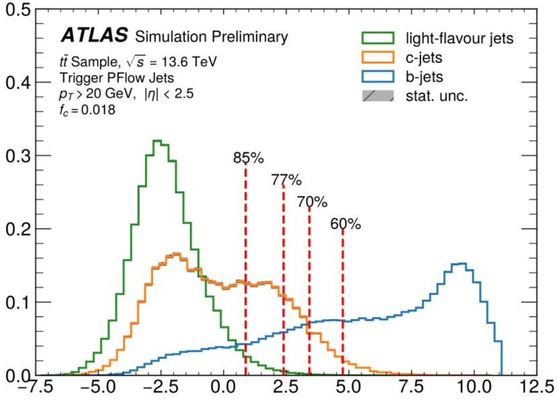
The problem is to Infer

- The shape of each individual component (c & b)
- The mixture fractions of each class
  cccc, ccbb and bbbb



Case: pp→hh→bbbb (LHC physics)



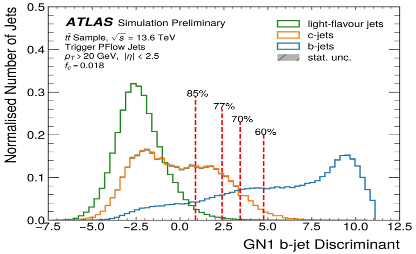


Explore 2 solutions:

1) Gaussian processes

2) Unimodal distributions

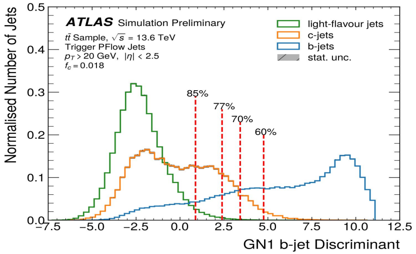
Gaussian Processes



$$f(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k * det(\boldsymbol{\Sigma})}} * e^{-\frac{1}{2} * ((\boldsymbol{x}-\boldsymbol{\mu})^T \cdot inv(\boldsymbol{\Sigma}) \cdot (\boldsymbol{x}-\boldsymbol{\mu}))}$$

We bin the score and ***x*** contains the distribution values in each bin

Gaussian Processes



We bin the score and ***x*** contains the distribution values in each bin

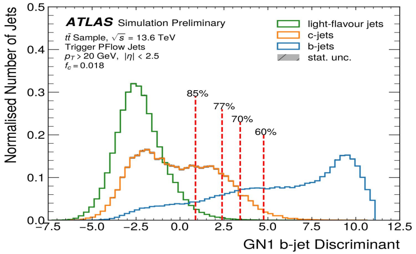
$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k * det(\boldsymbol{\Sigma})}} * e^{-\frac{1}{2} * ((\boldsymbol{x}-\boldsymbol{\mu})^T \cdot inv(\boldsymbol{\Sigma}) \cdot (\boldsymbol{x}-\boldsymbol{\mu}))}$$

Each bin is sampled around some expected μ

Define uncertainty and how related are neighbouring bins: Continuity!

$$\Sigma^{-1} = \begin{pmatrix} 2 & 1 & 0.5 & 0 & \dots \\ 1 & 2 & 1 & 0.5 & 0 & \dots \\ 0.5 & 1 & 2 & 1 & 0.5 & 0 & \dots \\ 0 & 0.5 & 1 & 2 & 1 & 0.5 & 0 \dots \\ \dots & \dots & & \dots & & \dots \end{pmatrix}$$

Gaussian Processes



$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k * det(\boldsymbol{\Sigma})}} * e^{-\frac{1}{2}*((\boldsymbol{x}-\boldsymbol{\mu})^T \cdot inv(\boldsymbol{\Sigma}) \cdot (\boldsymbol{x}-\boldsymbol{\mu}))}$$

We bin the score and ***x*** contains the distribution values in each bin

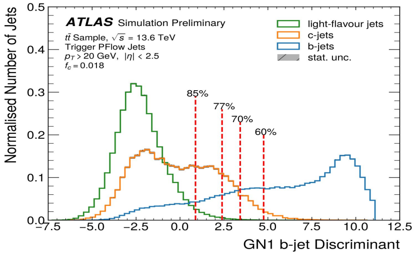
Each bin is sampled around some expected μ

Define uncertainty and how related are neighbouring bins: Continuity!

We can sample continuous curves around a central curve with very few hyperparameters

$$\Sigma^{-1} = \begin{pmatrix} 2 & 1 & 0.5 & 0 & \ldots. \\ 1 & 2 & 1 & 0.5 & 0 & \ldots \\ 0.5 & 1 & 2 & 1 & 0.5 & 0 & \ldots \\ 0 & 0.5 & 1 & 2 & 1 & 0.5 & 0 \ldots. \\ \ldots. & \ldots\ldots. & & \ldots.. & & \ldots.. \end{pmatrix}$$

Gaussian Processes



We bin the score and ***x*** contains the distribution values in each bin

$$f(\boldsymbol{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^k * det(\boldsymbol{\Sigma})}} * e^{-\frac{1}{2} * ((\boldsymbol{x}-\boldsymbol{\mu})^T \cdot inv(\boldsymbol{\Sigma}) \cdot (\boldsymbol{x}-\boldsymbol{\mu}))}$$

Each bin is sampled around some expected μ

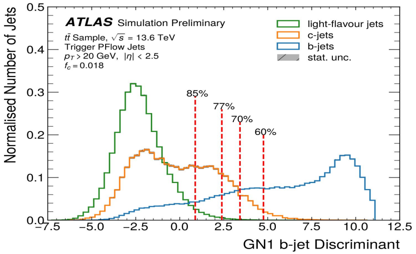
Define uncertainty and how related are neighbouring bins: Continuity!

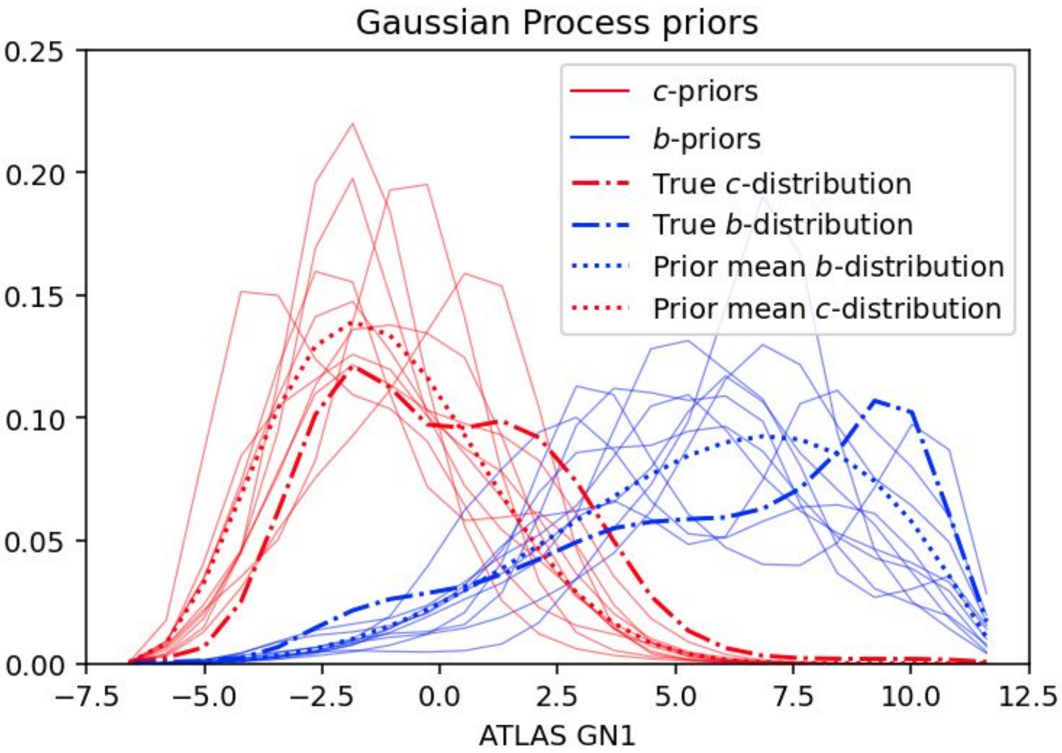
Prior information

We can sample continuous curves around a central curve with very few hyperparameters

$$\Sigma^{-1} = \begin{pmatrix} 2 & 1 & 0.5 & 0 & \ldots. & & & \\ 1 & 2 & 1 & 0.5 & 0 & \ldots & & \\ 0.5 & 1 & 2 & 1 & 0.5 & 0 & \ldots & \\ 0 & 0.5 & 1 & 2 & 1 & 0.5 & 0 & \ldots. \\ \ldots. & \ldots\ldots. & & \ldots.. & & \ldots.. & & \end{pmatrix}$$

Gaussian Processes

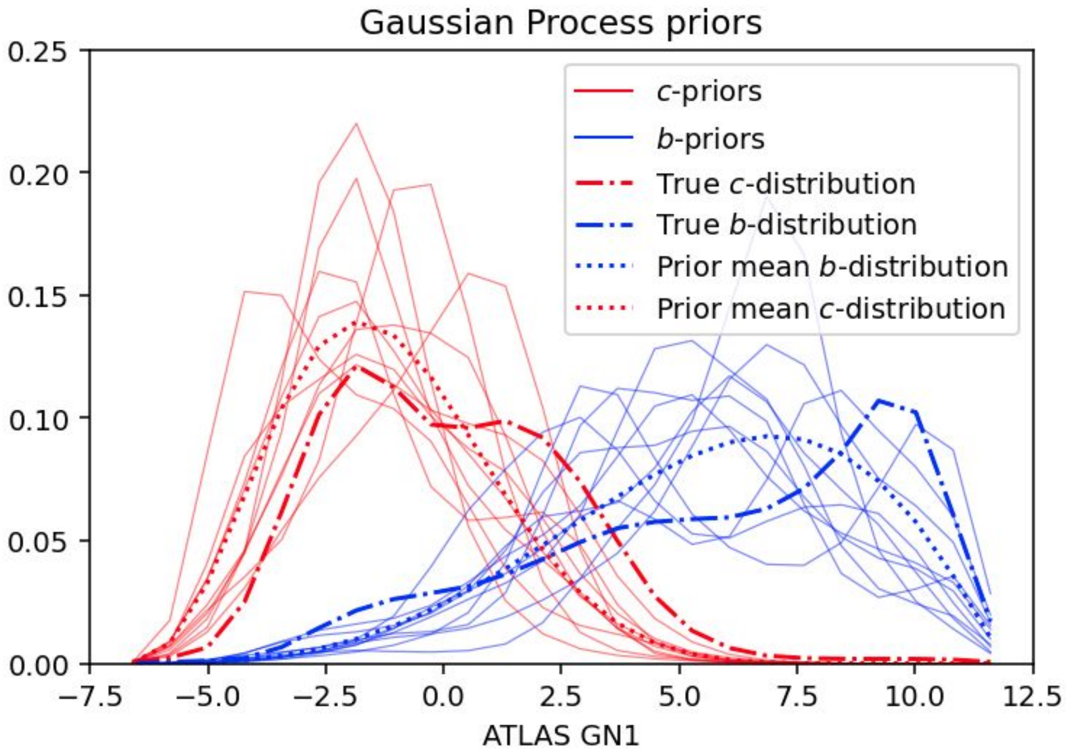




Gaussian Processes

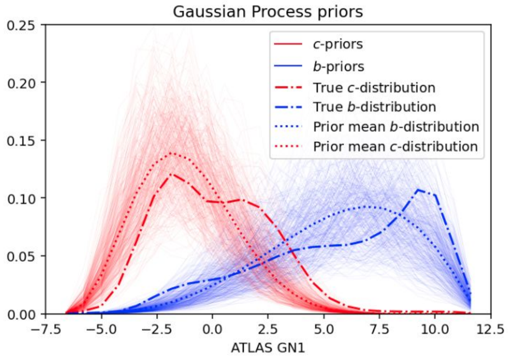
The game:

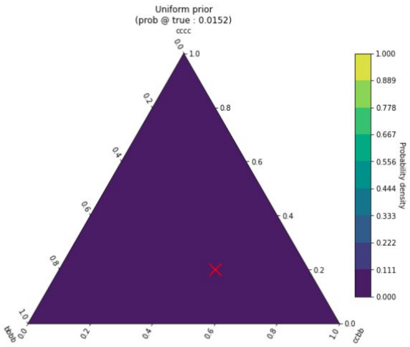
- Starts with biased prior
- The data will shift the posterior to the most likely distribution, which should be the true
- Leverage:
  - Multidimensionality
  - Continuity
  - bbbb, ccbb, cccc



Gaussian Processes: Results

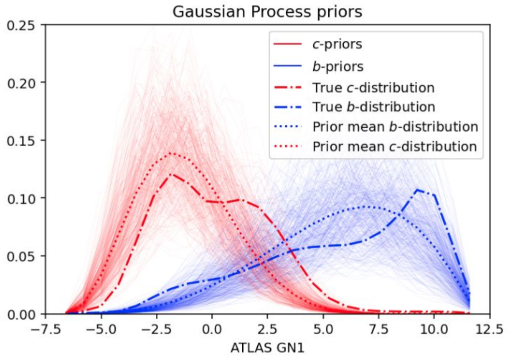
This is how we start

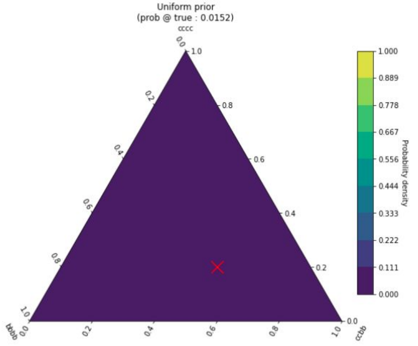




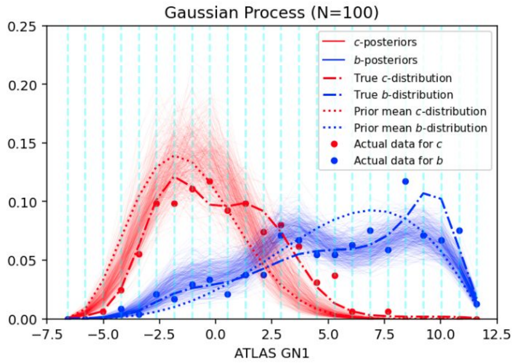
Gaussian Processes: Results

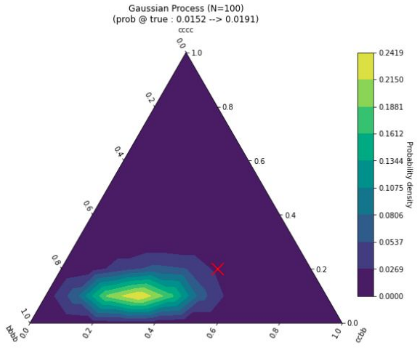
This is how we start





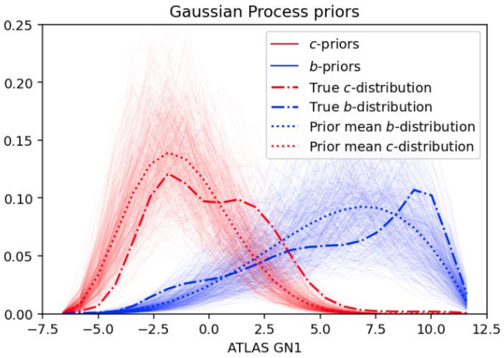
After seeing 100 events

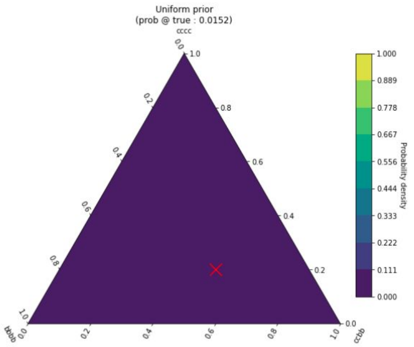




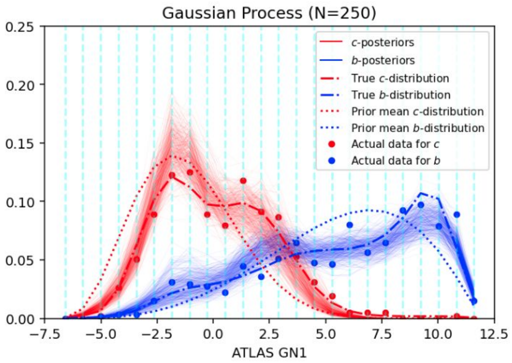
Gaussian Processes: Results

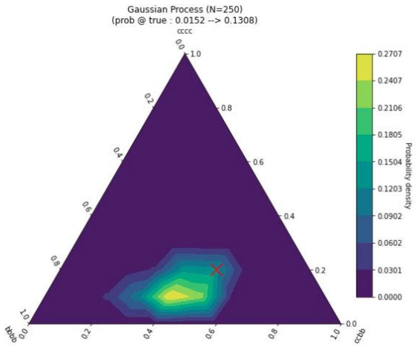
This is how we start





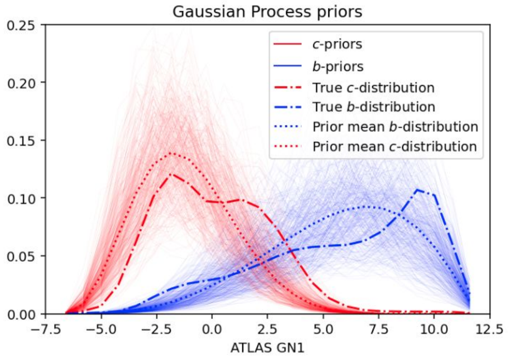
After seeing 250 events

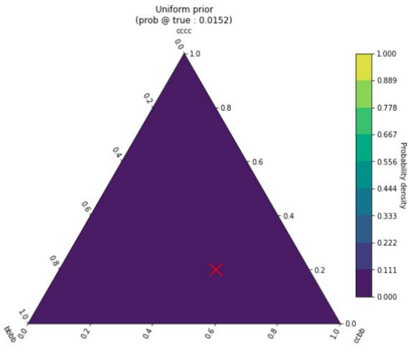




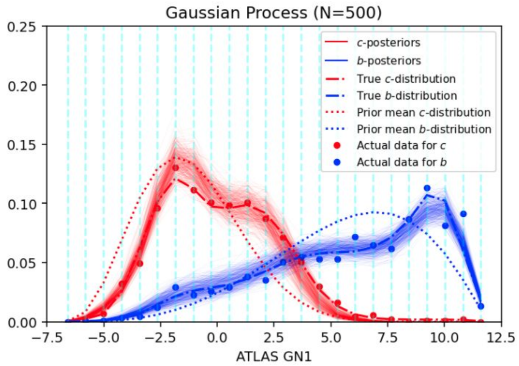
Gaussian Processes: Results

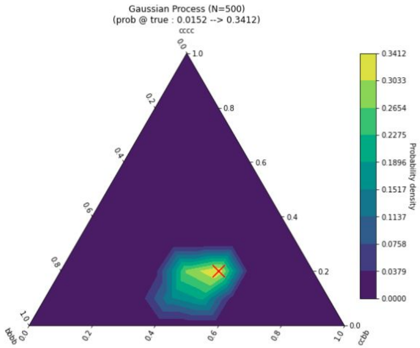
This is how we start





After seeing 500 events

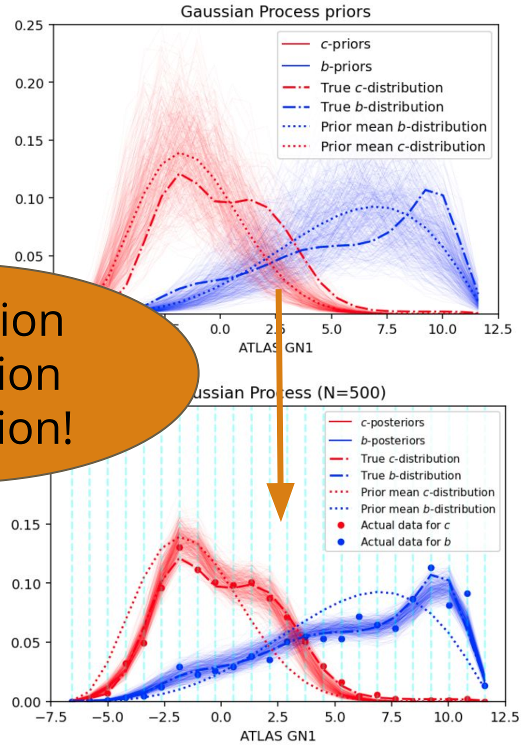


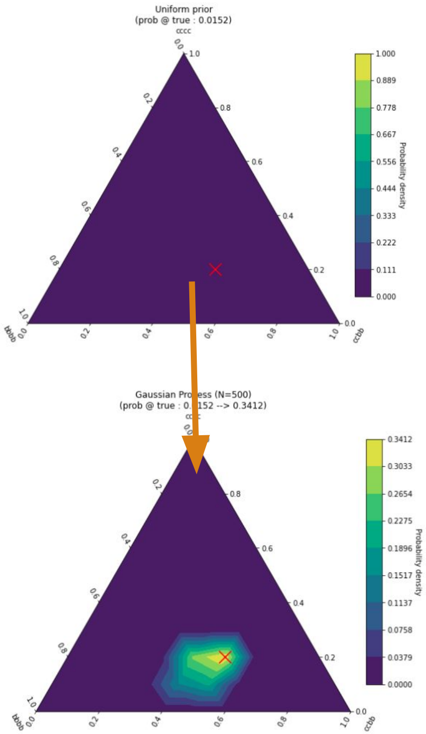


Gaussian Processes: Results

This is how we start

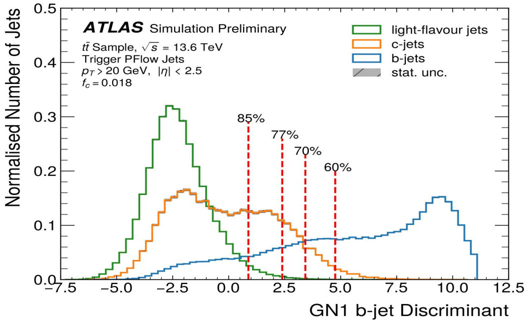
After seeing 500 events





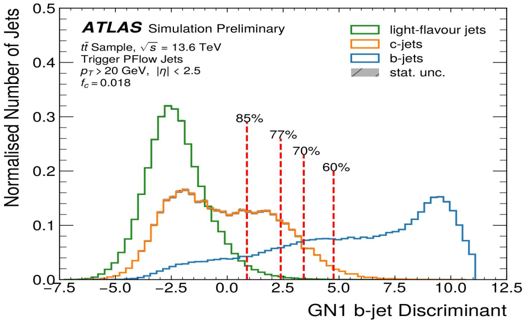
Correlation
correlation
correlation!

Unimodal model



How to sample unimodal arbitrary continuous curves ?

Unimodal model

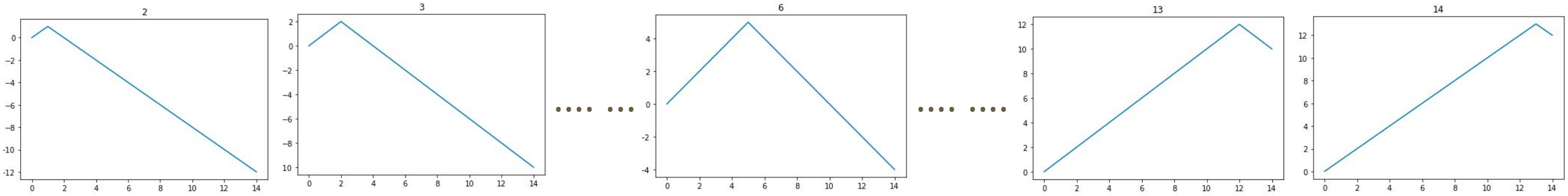


Prior information!

How to sample unimodal arbitrary continuous curves ?

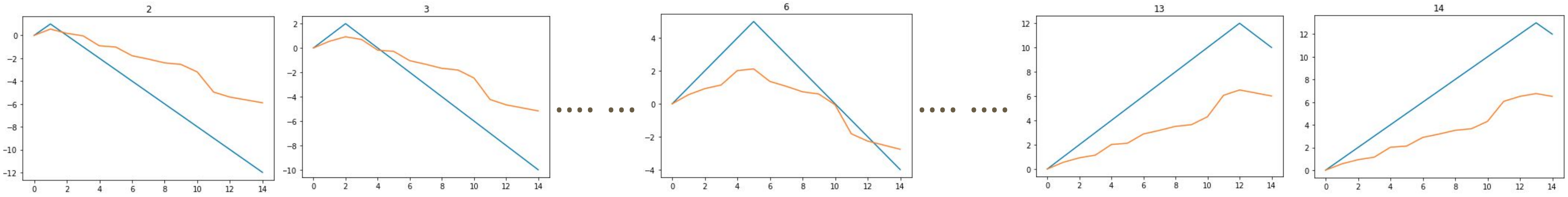
Unimodal model

Construct strict linear unimodal, one for each bin



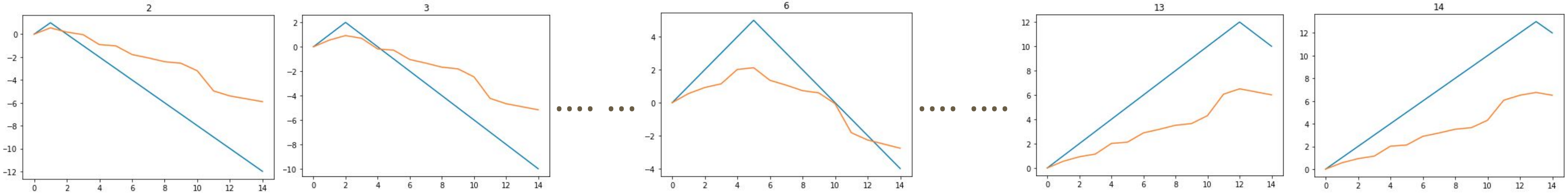
Unimodal model

Allow for randomness with a half normal $|N(0,0.5)|$ at each step

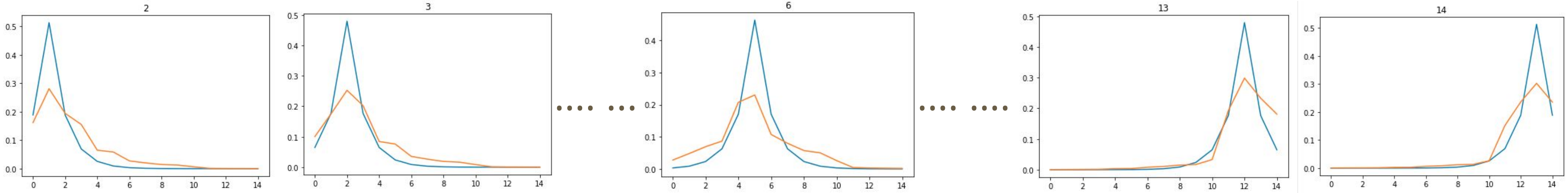


Unimodal model

Allow for randomness with a half normal |*N*(0,0.5)| at each step

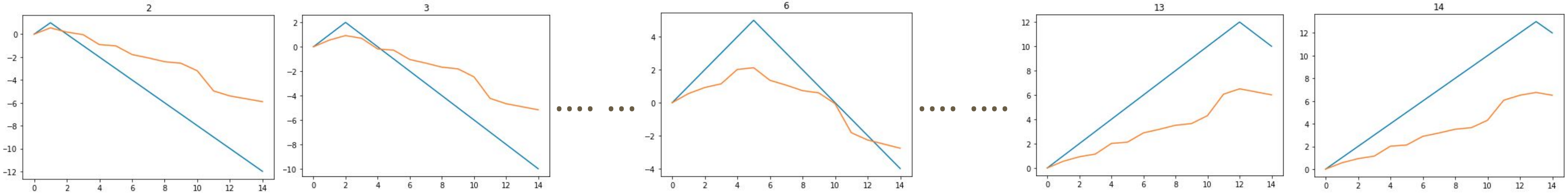


Apply *softmax()* to make them integrate to unity

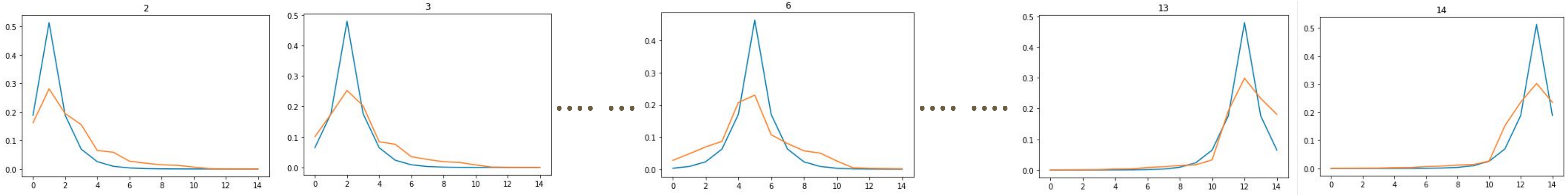


Unimodal model

Allow for randomness with a half normal $|N(0,0.5)|$ at each step



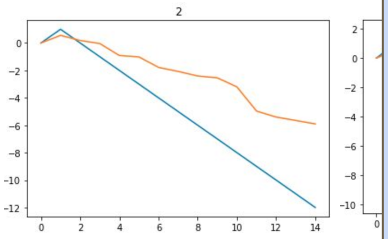
Apply *softmax()* to make them integrate to unity



How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

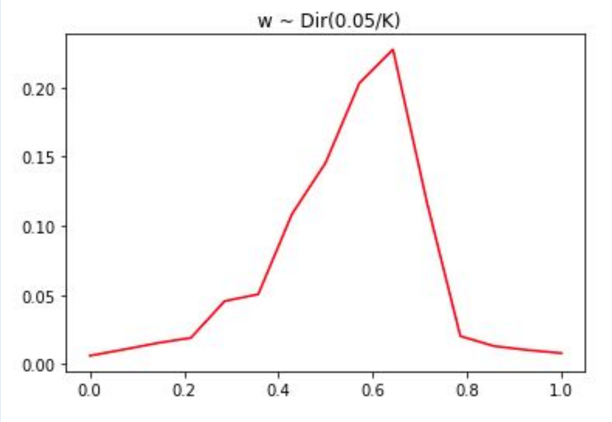
Allow for random

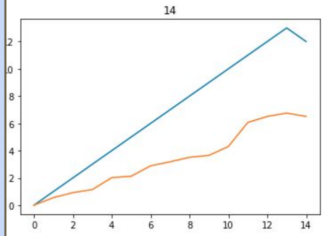


Sum orange curves weighted by

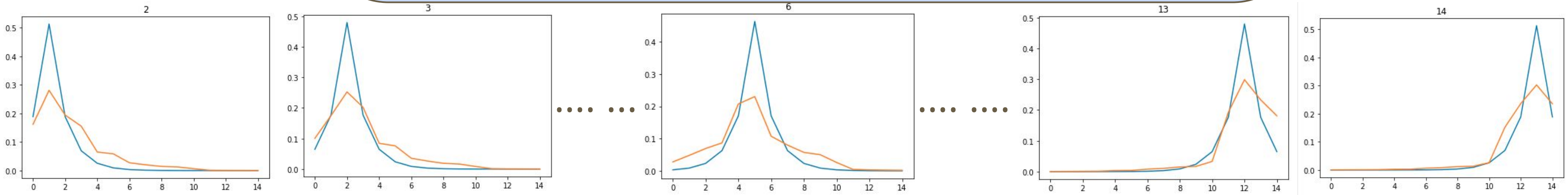
w ~Dirichlet(α)

With α small





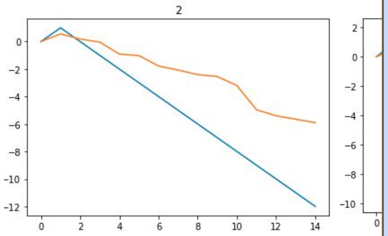
Apply *softmax()* to

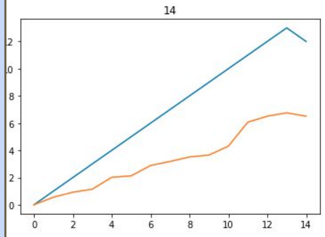


How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random

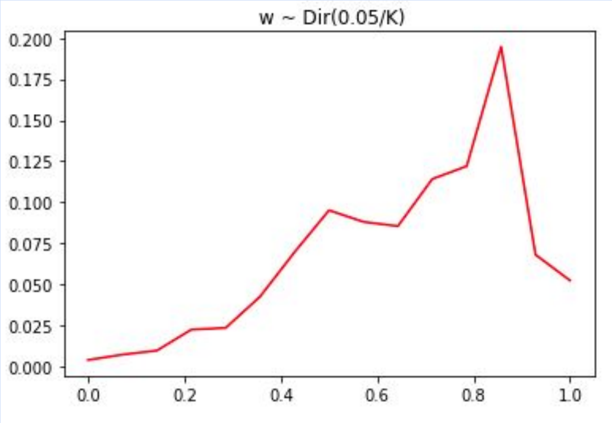




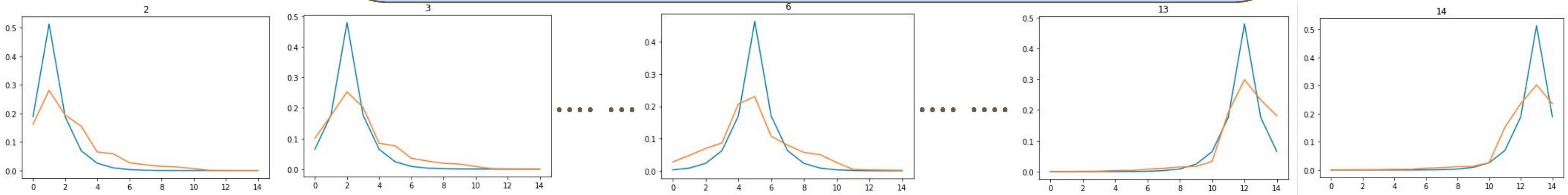
Sum orange curves weighted by

w ~Dirichlet(α)

With α small



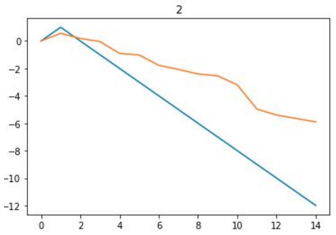
Apply *softmax()* to

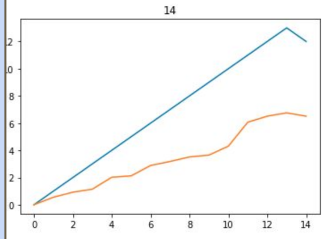


How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random

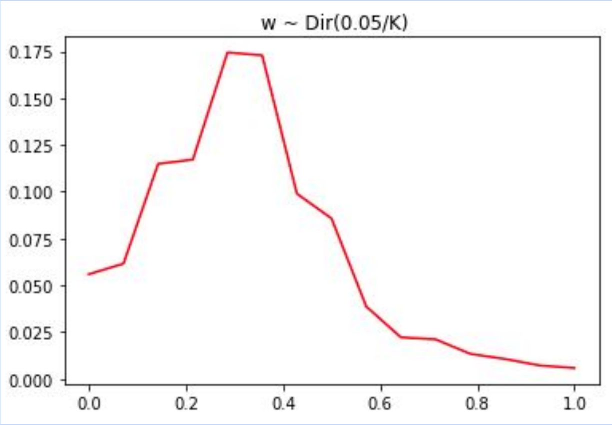




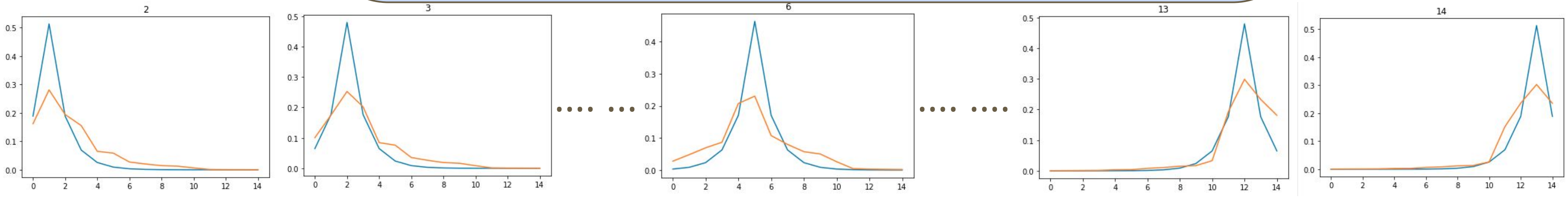
Sum orange curves weighted by

w ~Dirichlet(α)

With α small



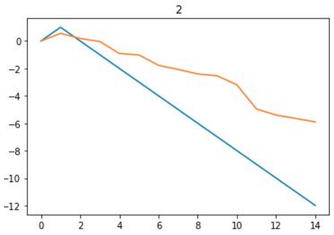
Apply *softmax()* to

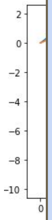


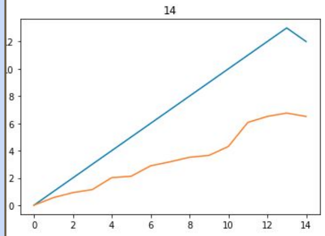
How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random



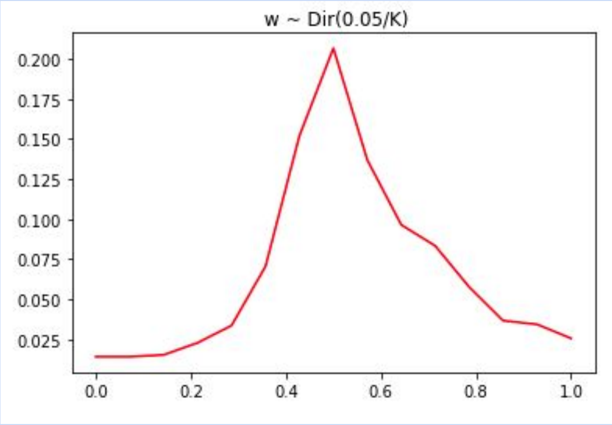




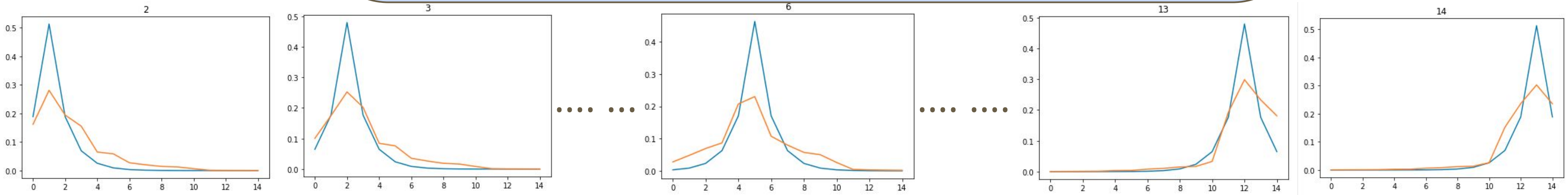
Sum orange curves weighted by

w ~Dirichlet(α)

With α small



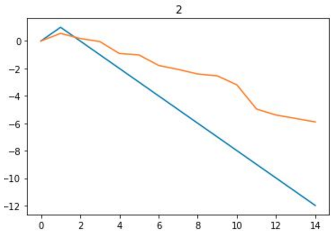
Apply *softmax()* to

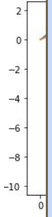


How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random

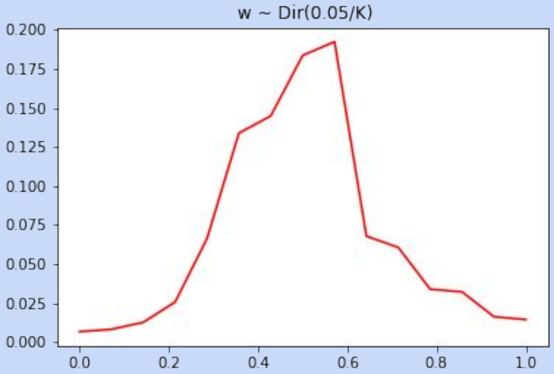


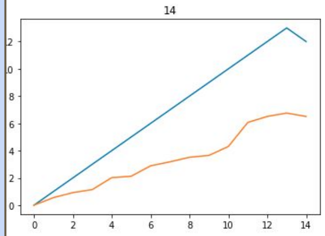


Sum orange curves weighted by

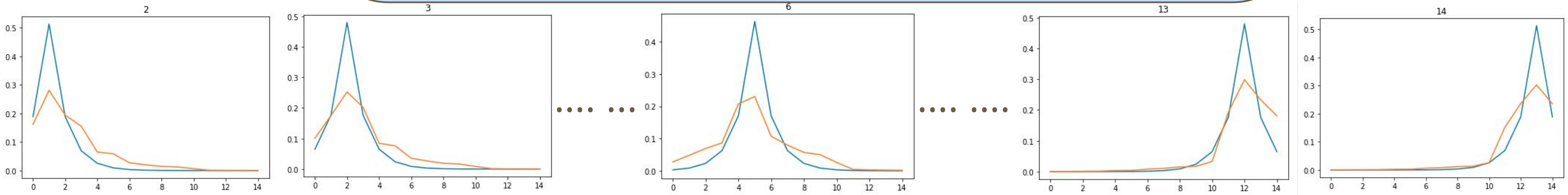
w ~Dirichlet(α)

With α small





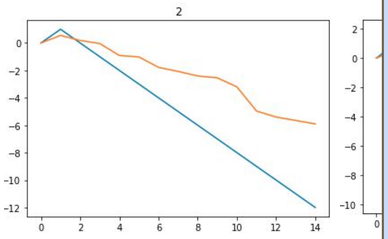
Apply *softmax()* to

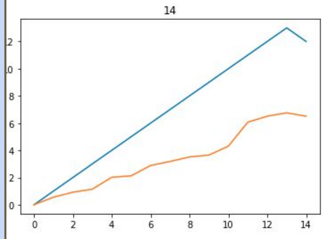


How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random

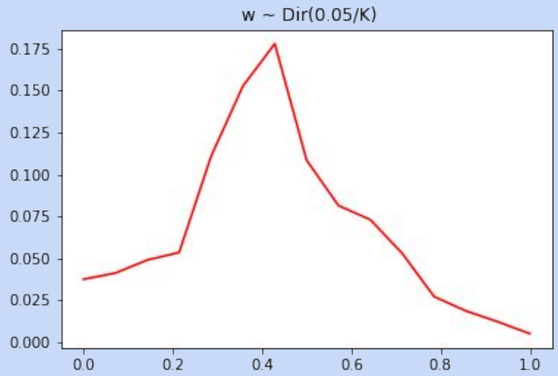




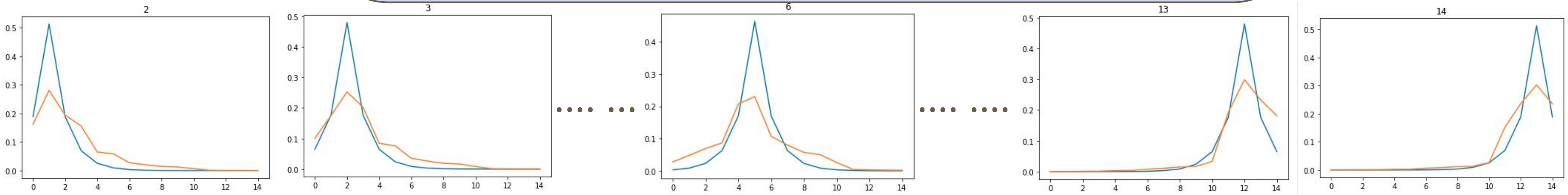
Sum orange curves weighted by

w ~Dirichlet(α)

With α small



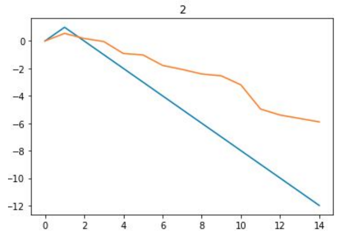
Apply *softmax()* to

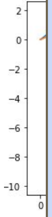


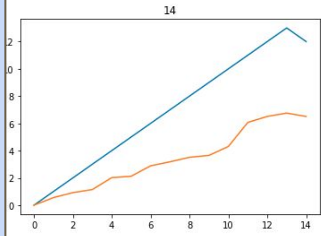
How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random



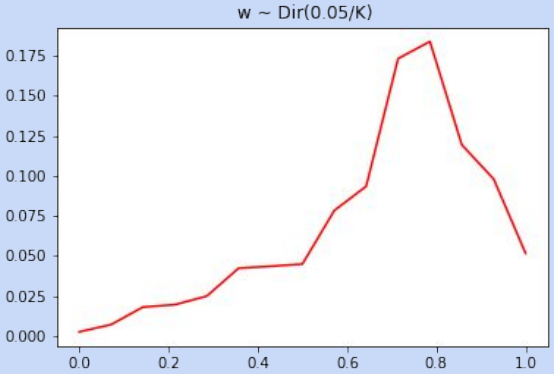




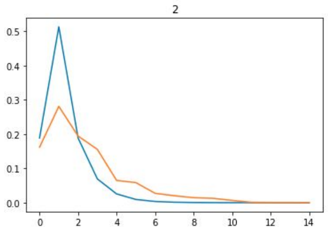
Sum orange curves weighted by

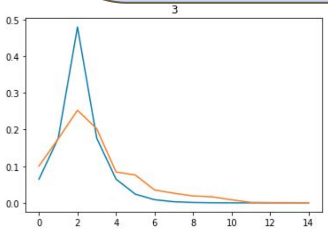
w ~Dirichlet(α)

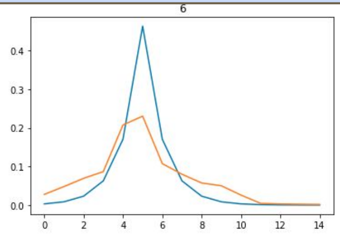
With α small

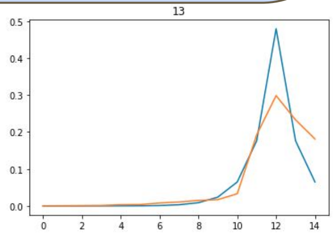


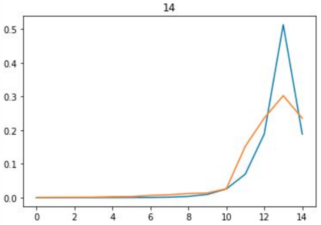
Apply *softmax()* to







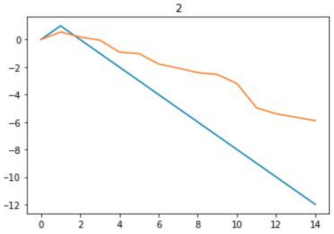


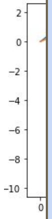


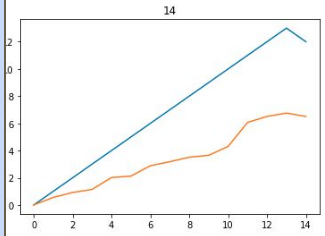
How to have unimodal at any bin and with some freedom of shape in other bins?

Unimodal model

Allow for random



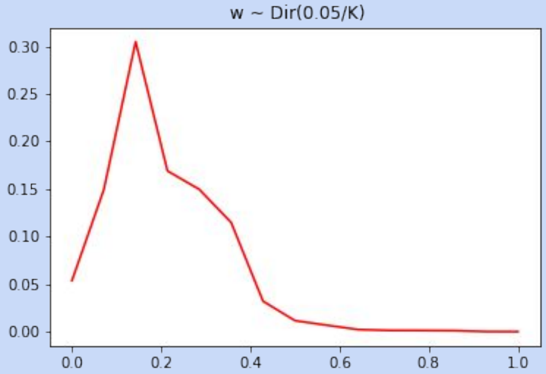




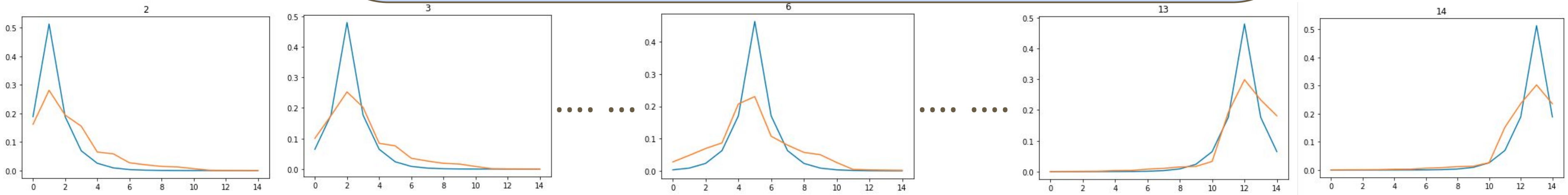
Sum orange curves weighted by

w ~Dirichlet(α)

With α small

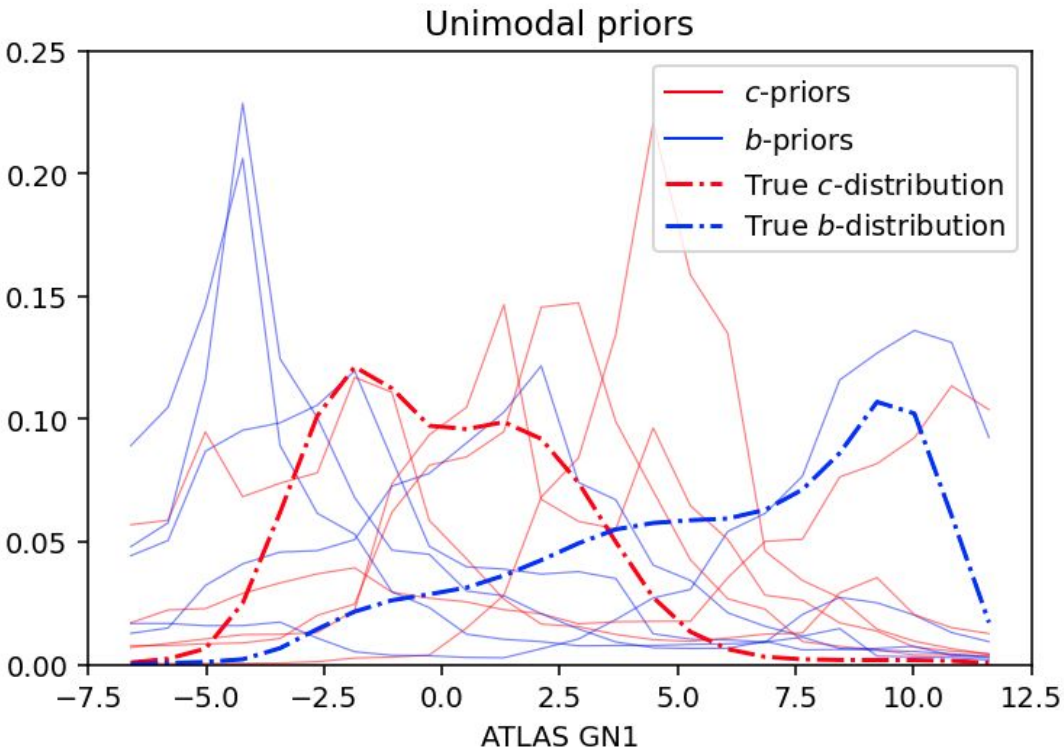


Apply *softmax()* to



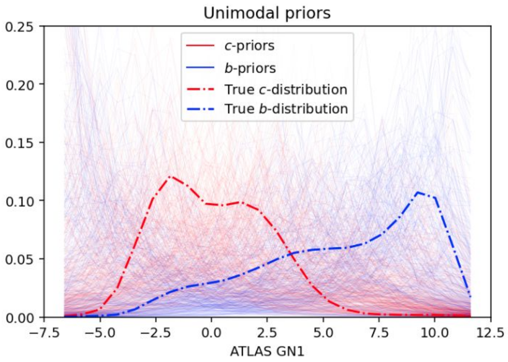
How to have unimodal at any bin and with some freedom of shape in other bins?

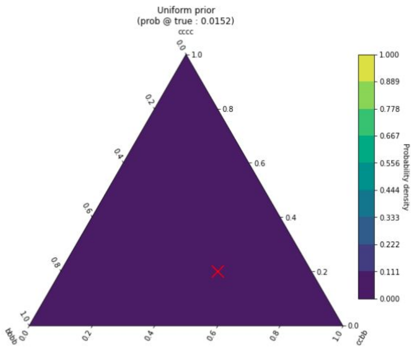
Unimodal model



Unimodal model: Results

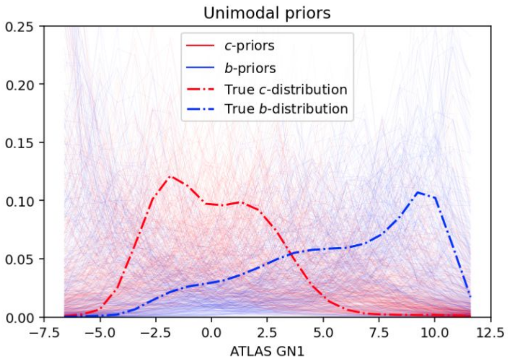
This is how we start

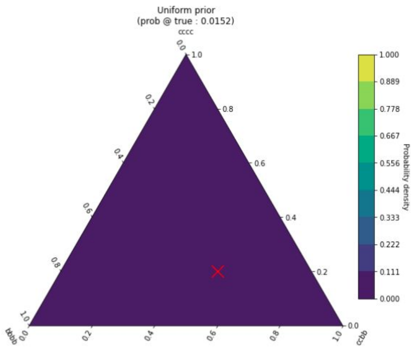




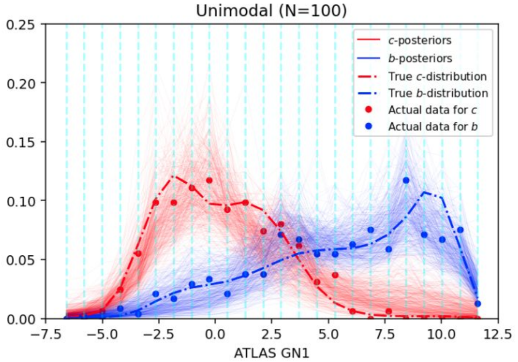
Unimodal model: Results

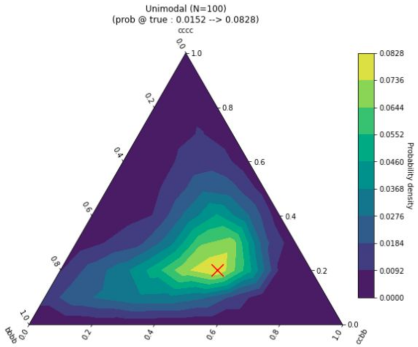
This is how we start





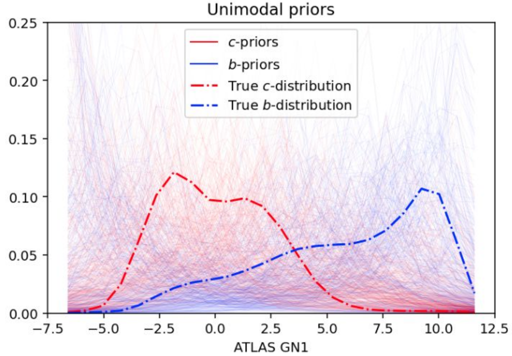
After seeing 100 events

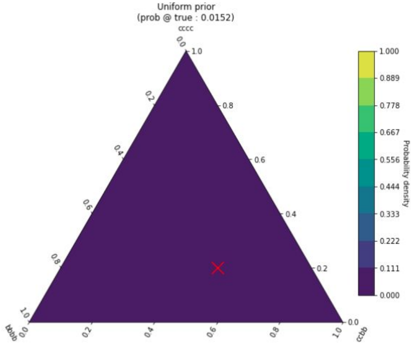




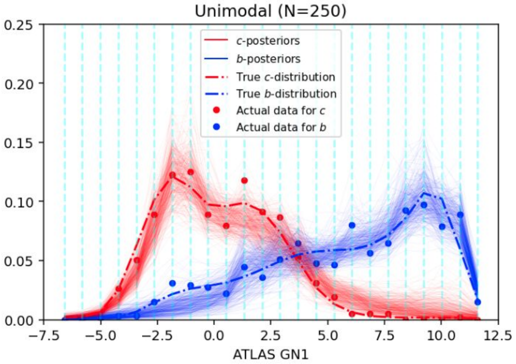
Unimodal model: Results

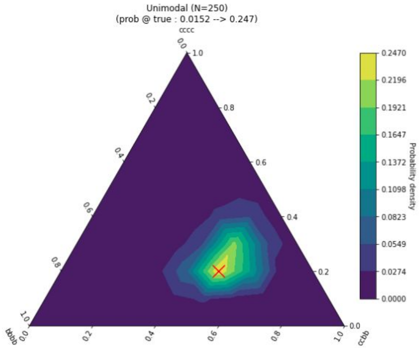
This is how we start





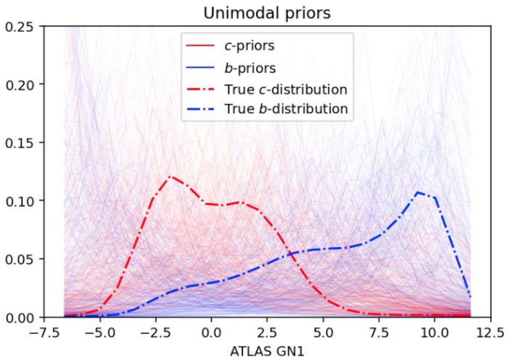
After seeing 250 events

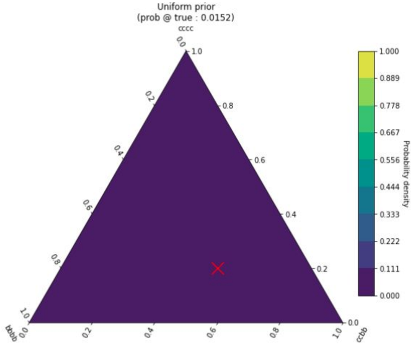




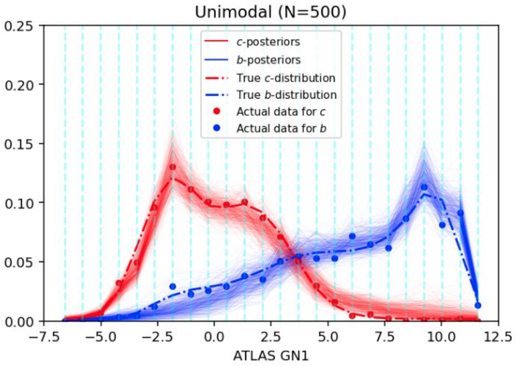
Unimodal model: Results

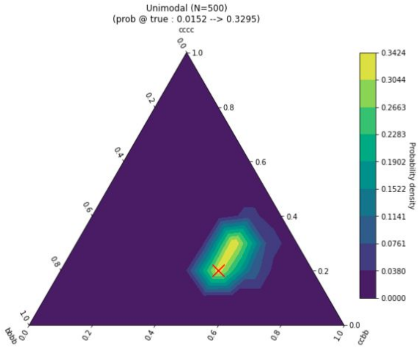
This is how we start





After seeing 500 events



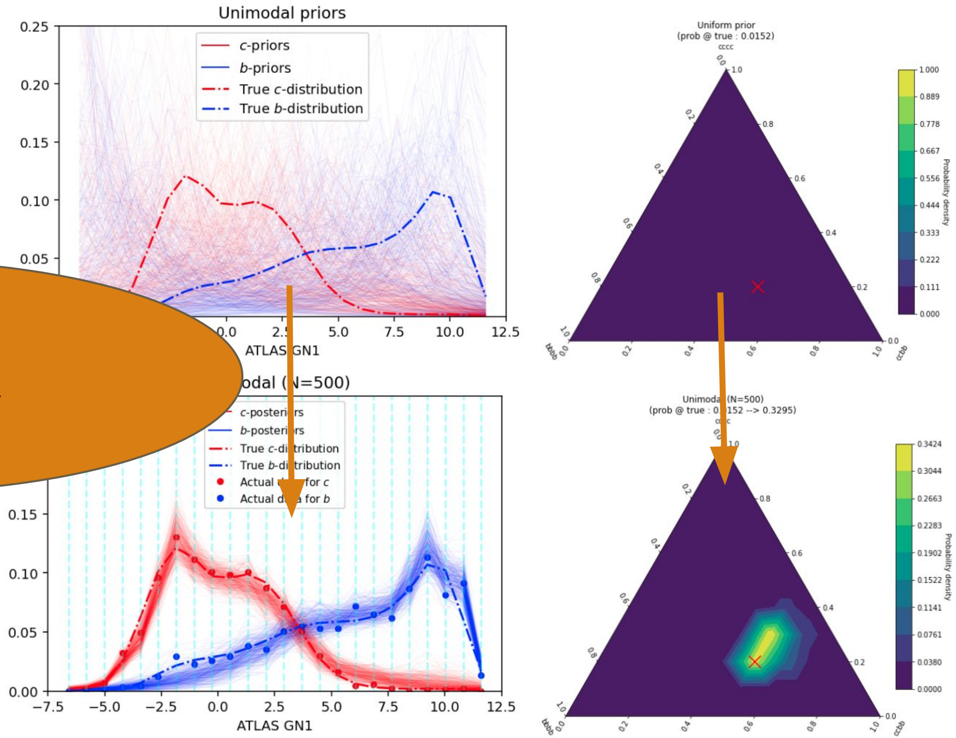


Unimodal model: Results

This is how we start

Correlation
+
unimodality
knowledge

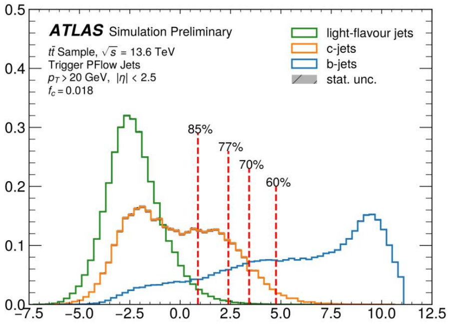
After seeing 500 events



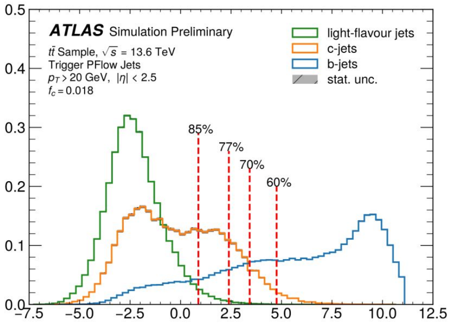
Summarizing

Summarizing

- Multidimensional Mixture of arbitrary continuous unimodal distributions

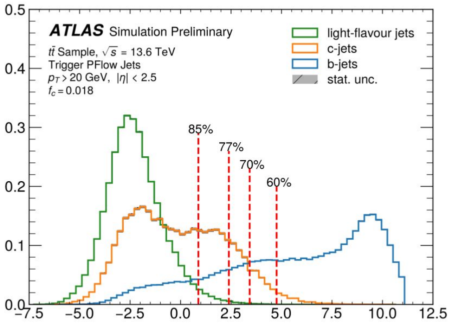


Summarizing

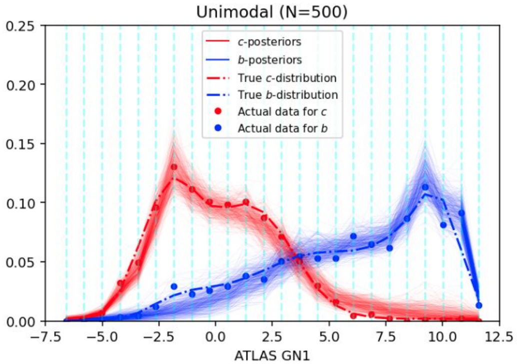


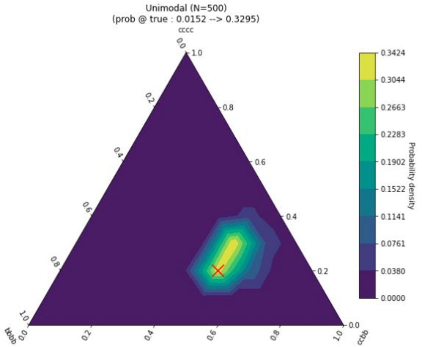
- Multidimensional Mixture of arbitrary continuous unimodal distributions
- Leverage on multidimensionality, continuity, unimodality and prior knowledge on allowed classes

Summarizing

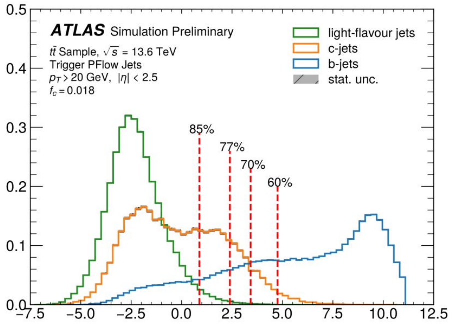


- Multidimensional Mixture of arbitrary continuous unimodal distributions
- Leverage on multidimensionality, continuity, unimodality and prior knowledge on allowed classes
- Infer everything through structured priors!

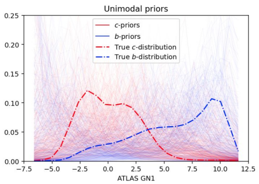


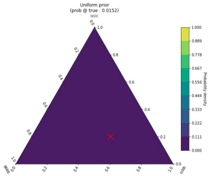


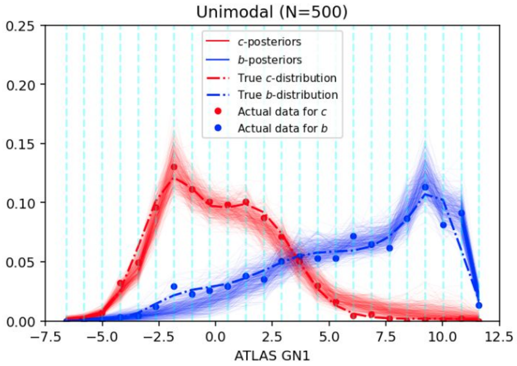
Summarizing

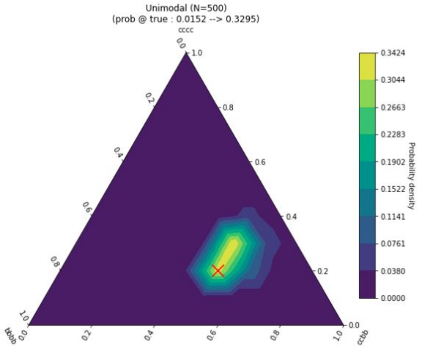


- Multidimensional Mixture of arbitrary continuous unimodal distributions
- Leverage on multidimensionality, continuity, unimodality and prior knowledge on allowed classes
- Infer everything through structured priors!







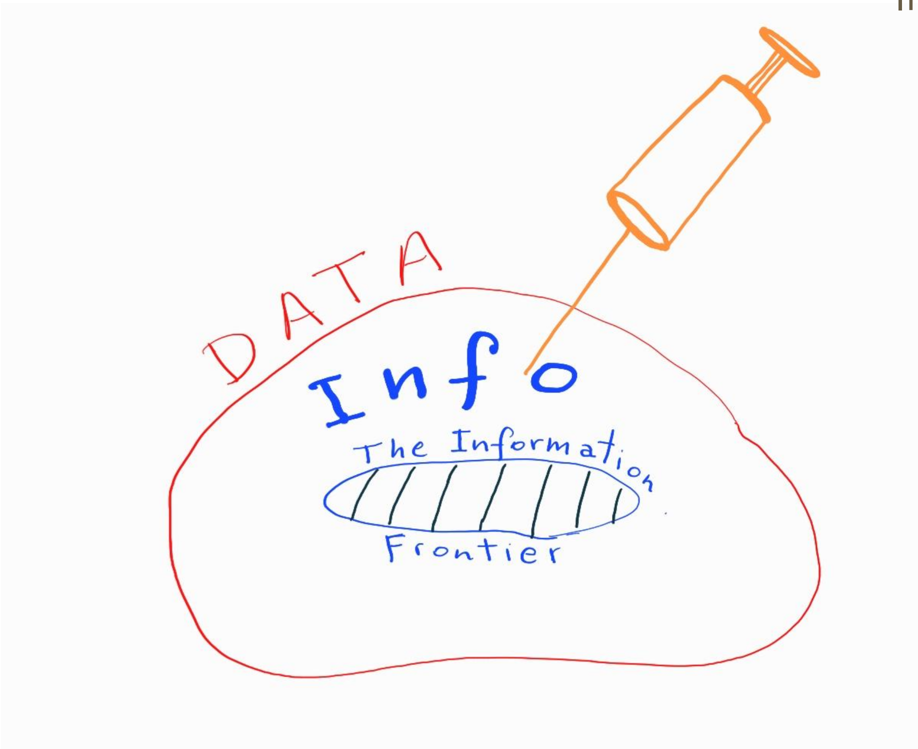


Learning Summary

Lecture 1

Intro to Bayesian ML

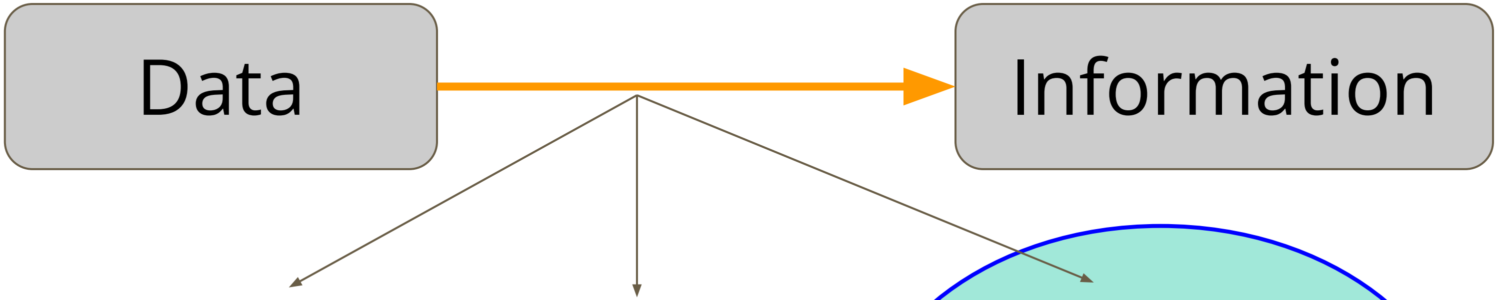
The Information Frontier



Inject *catalysts*:

- Modeling
- Tools & techniques
- Prior info

Typical problem in science



Cut based / control-treated

Misses to take benefit of correlations

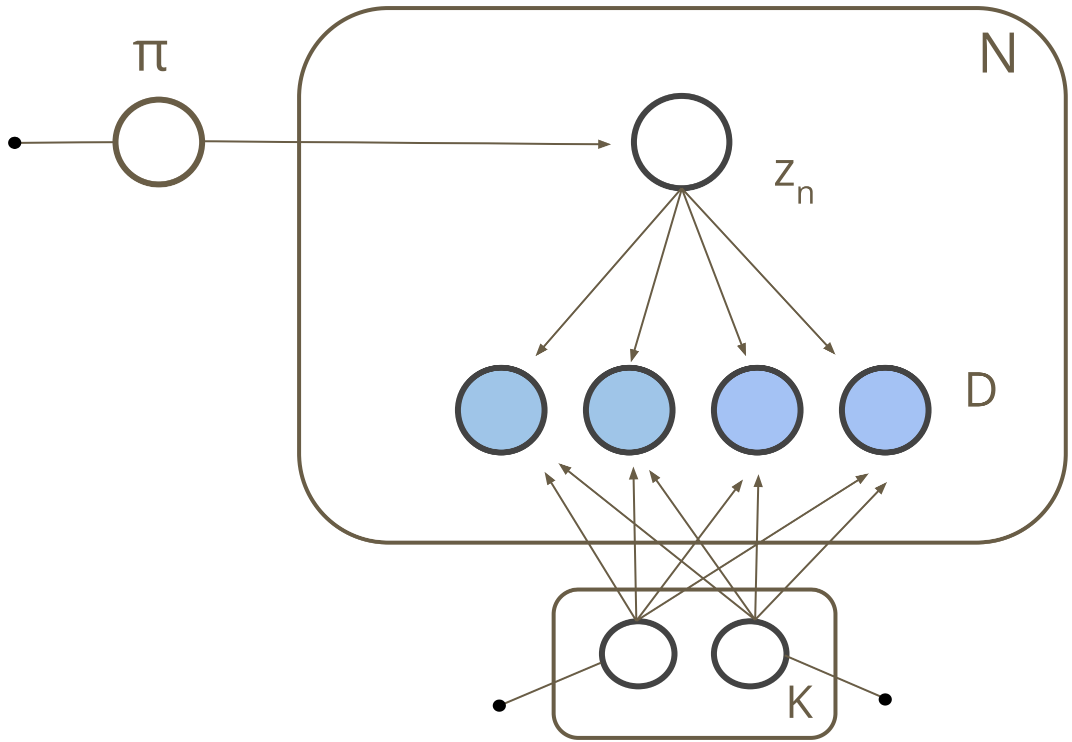
Simulation-based Neural Networks

Extracts correlations, but learns "too much" from simulations

Bayesian tools

This minicourse! Reduce simulation's impact at the price of increasing modelling impact

Bayesian Inference



- No hard cuts
- Soft assignments
- No signal/control regions
- K classes & D observables
- Deployment of data internal structure
- Controlled injection of prior knowledge

Lecture 2

Meet the tool

The box factory

Box OK: *p*
Box wrong: *1-p*

Tomorrow: k' boxes ok out of n'

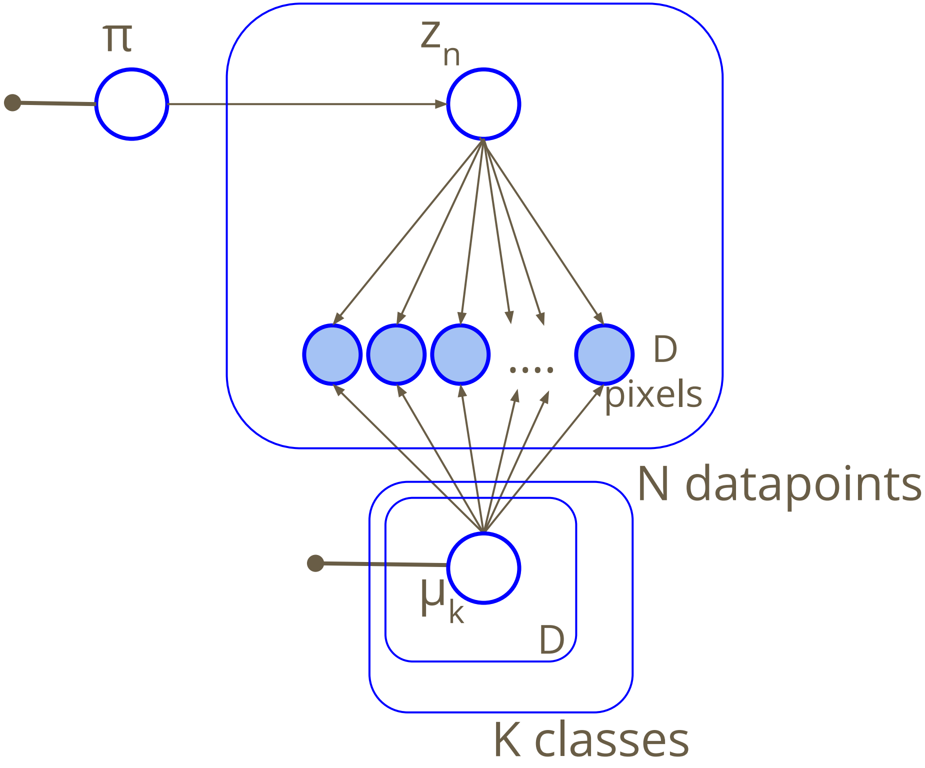
X = k', n'
Θ = p

$$p(\theta|X) = \frac{1}{B(n'+n, k'+k)} p^{k'+k}(1-p)^{n'+n-k'-k}$$

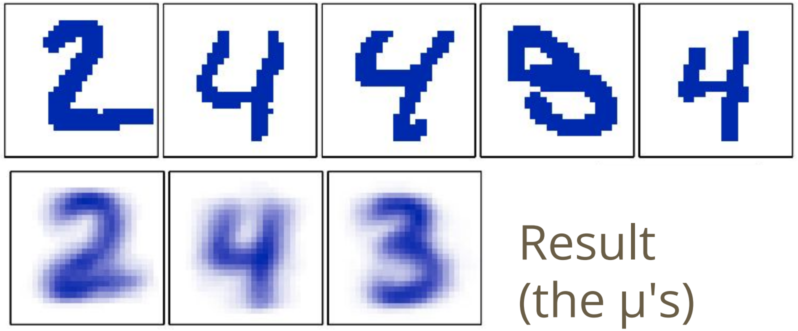
Today's posterior is tomorrow's prior!

$$p(\theta|X) = \frac{1}{B(n,k)} p^{k} (1-p)^{n-k}$$

Mixture of Bernoulli



Input



Result
(the μ's)

Scientifically: z_n is the probability of each class

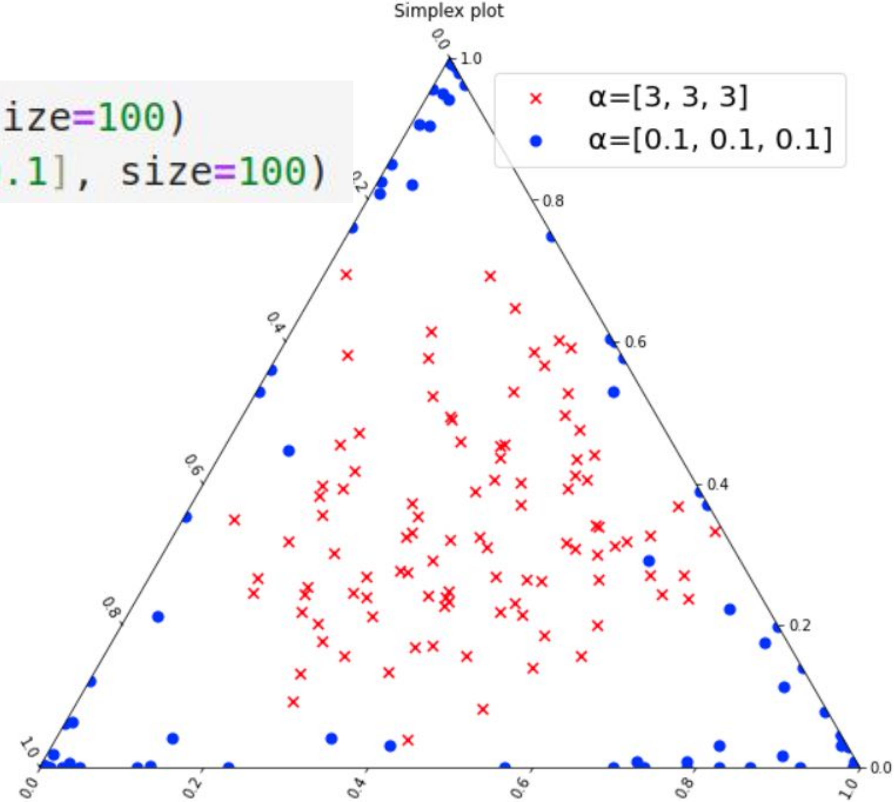
Lecture 3

Mixture Models

Dirichlet Distribution

```
sample1 = np.random.dirichlet([3,3,3], size=100)
sample2 = np.random.dirichlet([0.1,0.1,0.1], size=100)
```

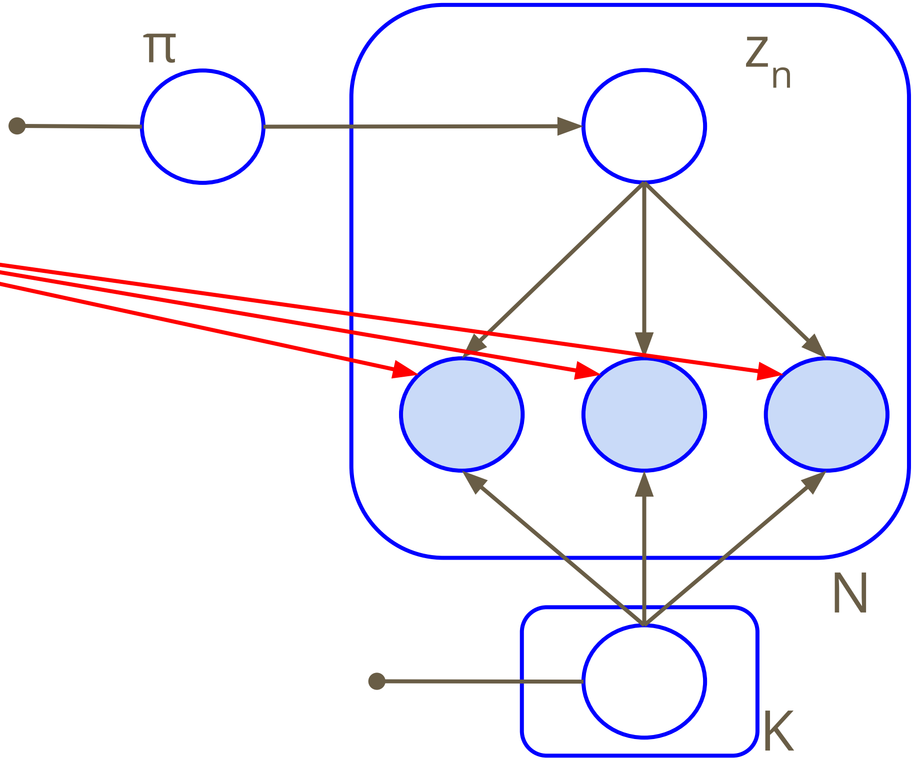
p(X) = *Dir*(X; α) = $\frac{1}{B(\boldsymbol{\alpha})} \prod_{i=1}^{K} x_i^{\alpha_i - 1}$



Mixture Models: where is the hack ?

Conditionally independence

Once the class is defined,
they are independent

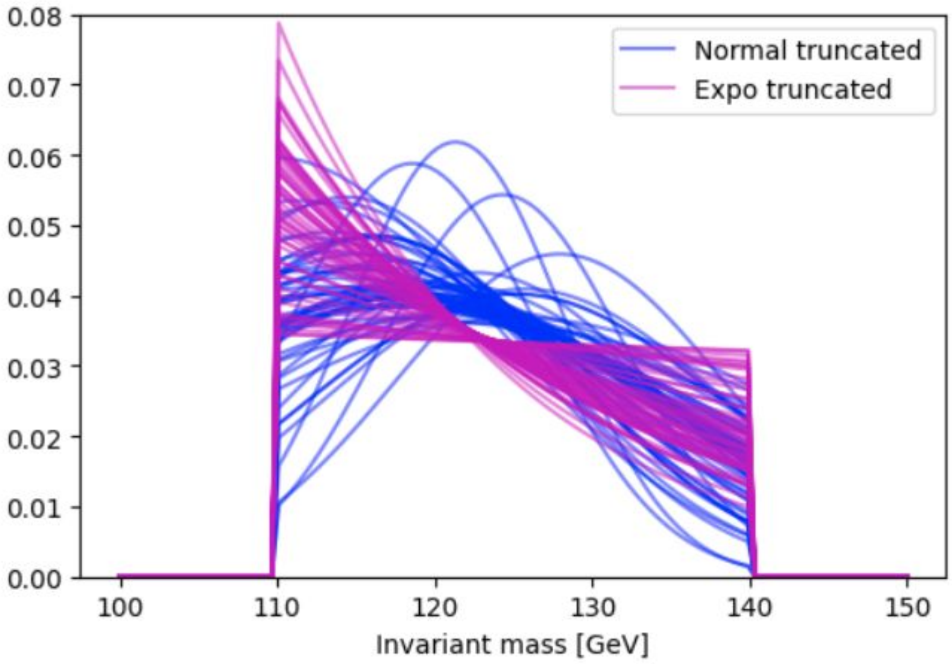


Huge reduction in number of parameters to infer!

better grip!
(B^D vs $B*D$)

2D Mixture Model: hh → bbγγ

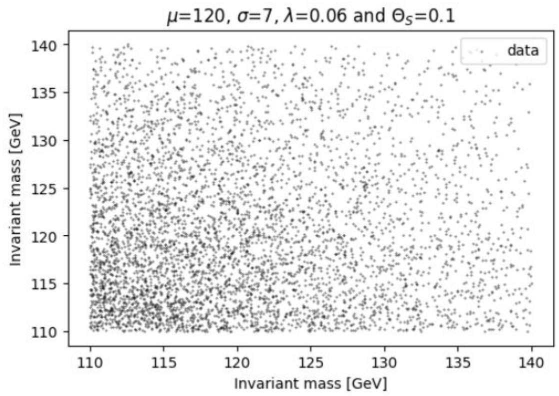
Truncated distributions!

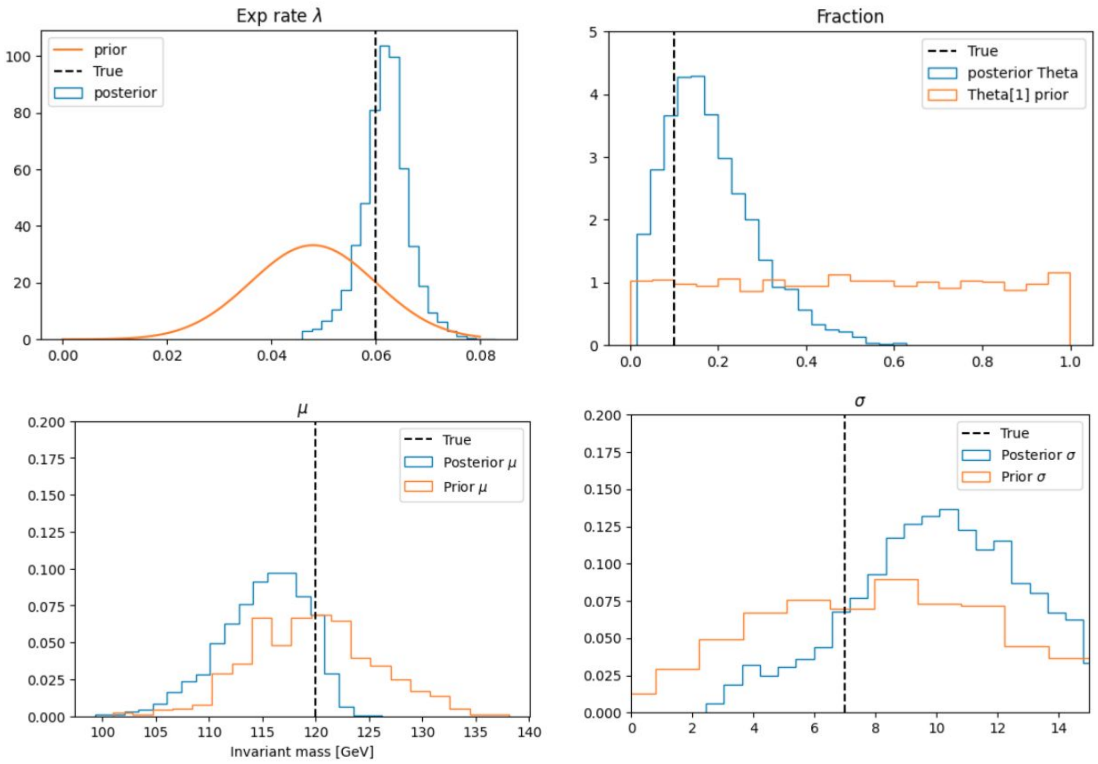


Hey.... but they are quite different!

Difficult to disentangle this mixture if parameters are unknown!

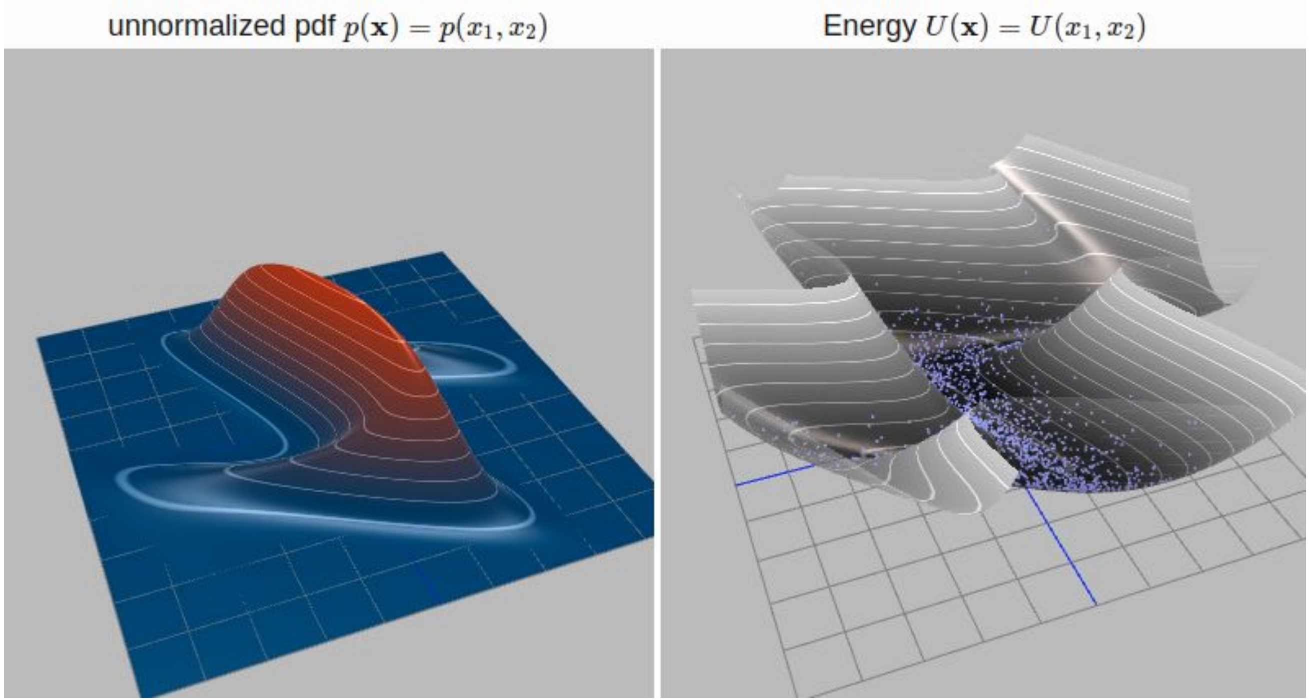
The impossible.... @10%





Lecture 4
Diagnoses and assessment

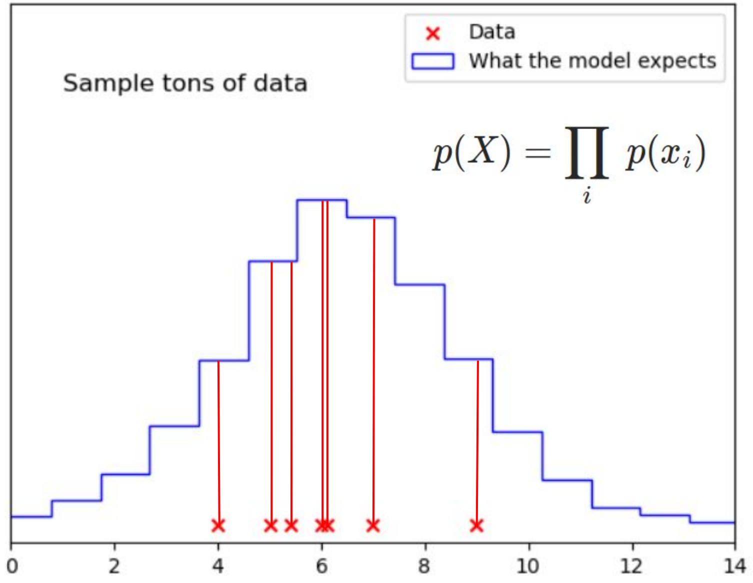
MCMC Sampling with Hamiltonian MC



$$
\begin{aligned}
H(\rho, \theta) &= -\log p(\rho, \theta) \\
&= -\log p(\rho|\theta) - \log p(\theta \\
&= T(\rho|\theta) + V(\theta),
\end{aligned}
$$

Simulate trajectory of fictitious particle

What is good and what is bad ?



$p = 10^{-6}$ ... and now?

- What does it mean ?
- What do we compare it to ?

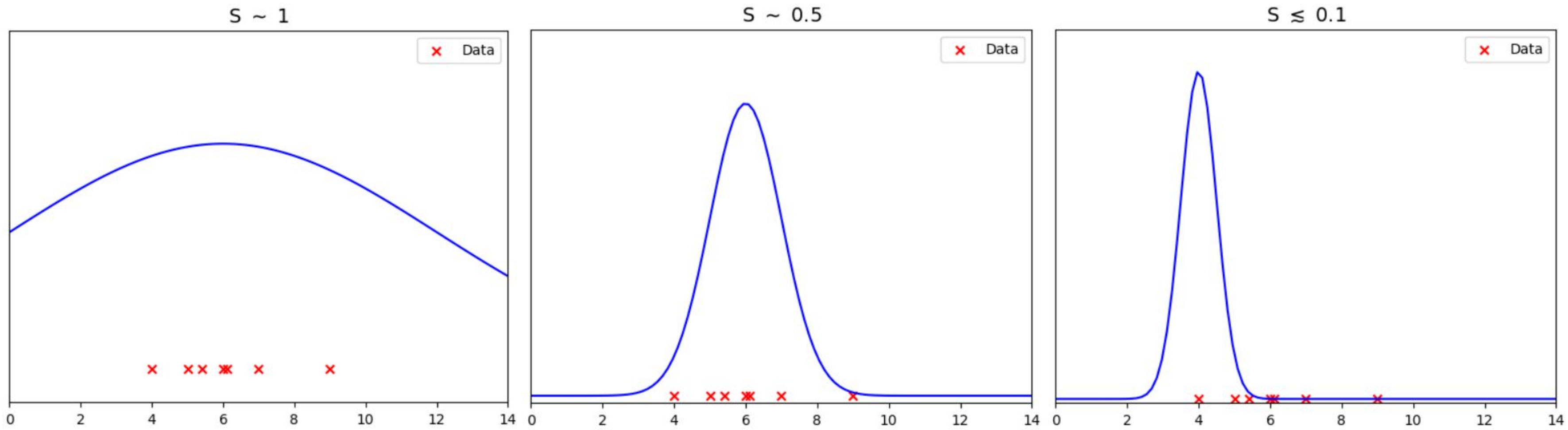
- Generate replicas of data X^{Rep}
- Compute their probability
- Compute

$$p\left(p(X^{rep}) < p(X) \right)$$

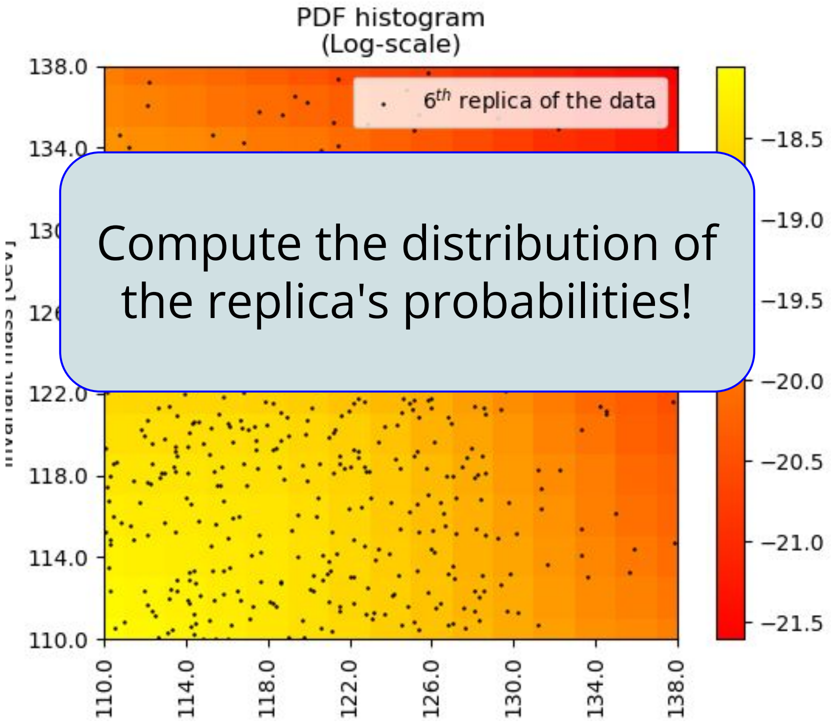
What is good and what is bad ?

$$\text{Score} = p\big(\, p(X^{rep}) < p(X) \,\big)$$

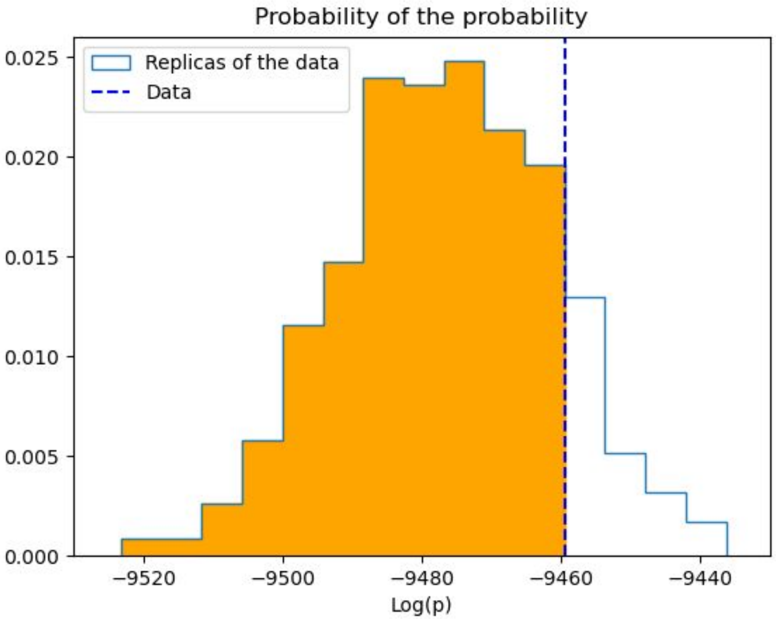
- S ~ 1 : bad model
- S ~ 0.5: good model
- S ≲ 0.1: bad model



PDF of your model given the data !



Compute the distribution of the replica's probabilities!



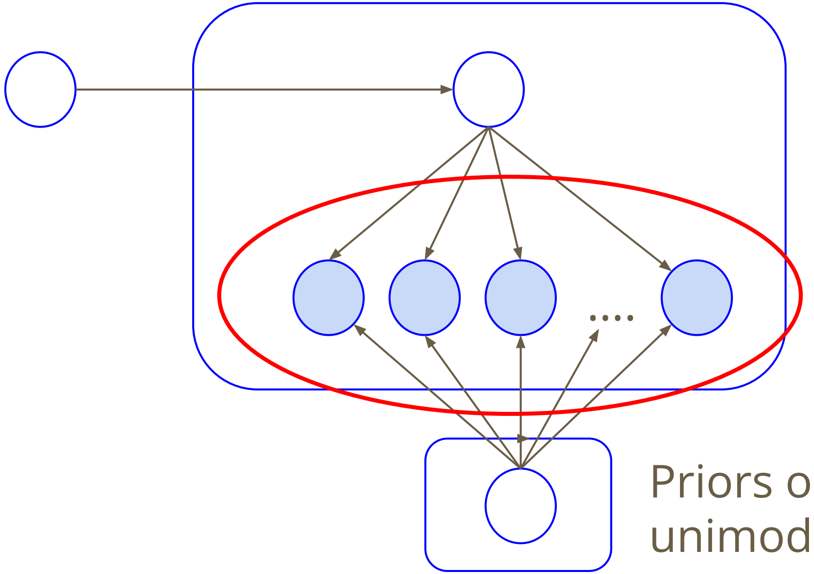
$$p\left(p(X_{held}^{rep}) < p(X_{held})\right) = 0.86$$

Lecture 5
Inferring non-parametric distributions

Inferring non-parametric distributions

Usually multi-dimension required
(otherwise ambiguities)

Solve the unsolvable puzzle ?



Often some prior knowledge on the structure

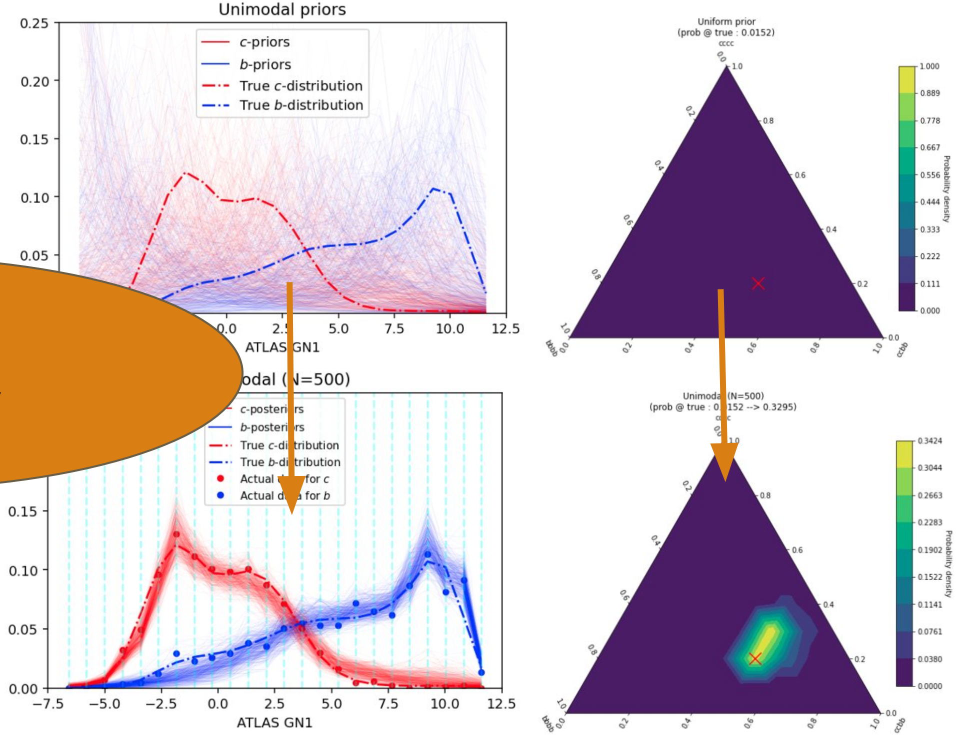
Priors on these distributions!
unimodal, smooth, etc!

Unimodal model: Results

This is how we start

Correlation
+
unimodality
knowledge

After seeing 500 events



Lecture 6+
Open questions

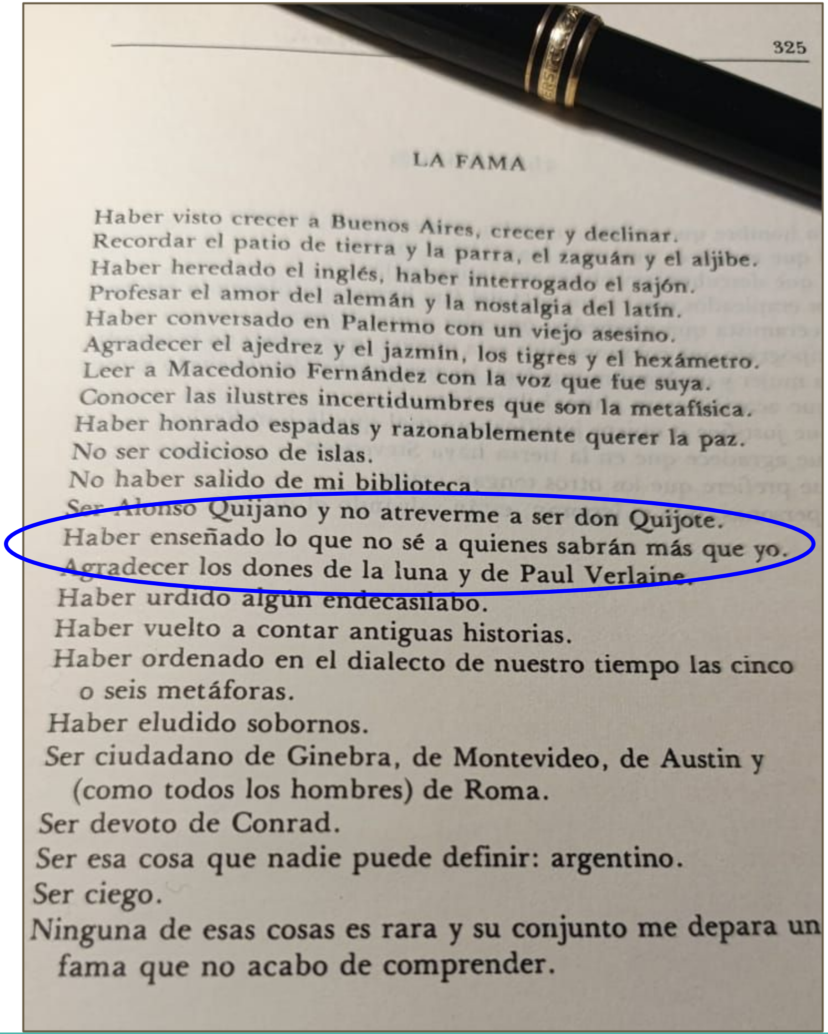
Open questions

- Correlated observables
- A Universe beyond Mixture Models
- ....

Disclaimer

I'm not a statistician...I'm learning!

J.L.Borges: *La fama* (1981)



325

LA FAMA

Haber visto crecer a Buenos Aires, crecer y declinar.
Recordar el patio de tierra y la parra, el zaguán y el aljibe.
Haber heredado el inglés, haber interrogado el sajón.
Profesar el amor del alemán y la nostalgia del latín.
Haber conversado en Palermo con un viejo asesino.
Agradecer el ajedrez y el jazmín, los tigres y el hexámetro.
Leer a Macedonio Fernández con la voz que fue suya.
Conocer las ilustres incertidumbres que son la metafísica.
Haber honrado espadas y razonablemente querer la paz.
No ser codicioso de islas.
No haber salido de mi biblioteca.
Ser Alonso Quijano y no atreverme a ser don Quijote.
Haber enseñado lo que no sé a quienes sabrán más que yo.
Agradecer los dones de la luna y de Paul Verlaine.
Haber urdido algún endecasílabo.
Haber vuelto a contar antiguas historias.
Haber ordenado en el dialecto de nuestro tiempo las cinco
o seis metáforas.
Haber eludido sobornos.
Ser ciudadano de Ginebra, de Montevideo, de Austin y
(como todos los hombres) de Roma.
Ser devoto de Conrad.
Ser esa cosa que nadie puede definir: argentino.
Ser ciego.
Ninguna de esas cosas es rara y su conjunto me depara una
fama que no acabo de comprender.

Thank you very much!

Thanks to

- Organizing committee
- Nathan Berkovits
- Rogério Rosenfeld
- Great staff
- Fantastic audience!!!

sequi@unsam.edu.ar