Dirichlet Process Mixture Models

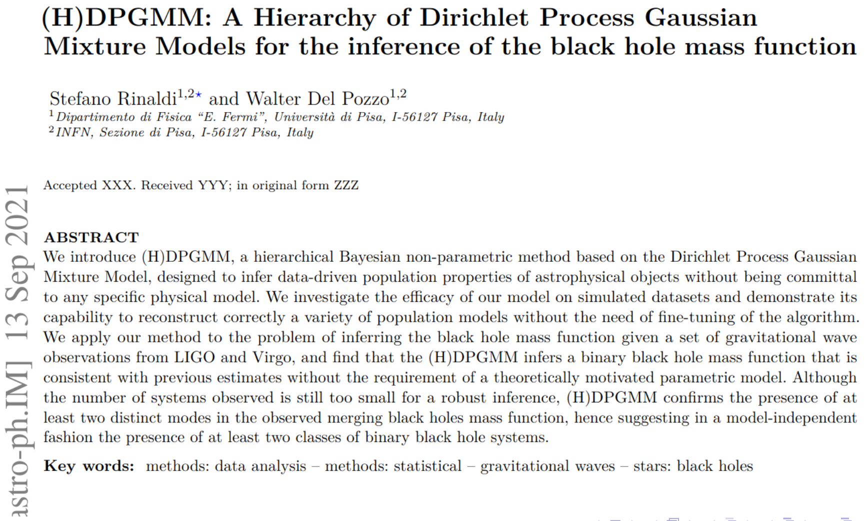
Bayesian Nonparametric Modelling

Mateusz Kapusta

SJC

8th May 2024

Motivation



(H)DPGMM: A Hierarchy of Dirichlet Process Gaussian Mixture Models for the inference of the black hole mass function

Stefano Rinaldi[1,2]⋆ and Walter Del Pozzo[1,2]
[1] *Dipartimento di Fisica "E. Fermi", Università di Pisa, I-56127 Pisa, Italy*
[2] *INFN, Sezione di Pisa, I-56127 Pisa, Italy*

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT
We introduce (H)DPGMM, a hierarchical Bayesian non-parametric method based on the Dirichlet Process Gaussian Mixture Model, designed to infer data-driven population properties of astrophysical objects without being committal to any specific physical model. We investigate the efficacy of our model on simulated datasets and demonstrate its capability to reconstruct correctly a variety of population models without the need of fine-tuning of the algorithm. We apply our method to the problem of inferring the black hole mass function given a set of gravitational wave observations from LIGO and Virgo, and find that the (H)DPGMM infers a binary black hole mass function that is consistent with previous estimates without the requirement of a theoretically motivated parametric model. Although the number of systems observed is still too small for a robust inference, (H)DPGMM confirms the presence of at least two distinct modes in the observed merging black holes mass function, hence suggesting in a model-independent fashion the presence of at least two classes of binary black hole systems.

Key words: methods: data analysis – methods: statistical – gravitational waves – stars: black holes

What is Dirichlet Process?

One of the key methods used in the field of **nonparametric Bayesian modeling**. The basic definition is quite hard and varies. Here simplified version will be given. We need to specify two basic parameters:

- concentration parameter α,
- continous distribution over some probabilistic space H.

One can construct many realizations of the Dirichlet Process, here two most popular versions will be given: **Chinese restaurant process** and **Stick-breaking process**.

Chinese restaurant process

Let's imagine following the construction of X_n: For $n = 1$ we sample $X_0 \sim H$, for $n > 1$

- with probability $\frac{\alpha}{\alpha+n-1}$ draw X_n from H,
- with probability $\frac{1}{\alpha+n-1}$ select on of previous values.

This particular system has many interesting properties. Most importantly it has some nice properties with respect to the parameter α. If α has a significant value we are getting more unique samples (first step), while for low values of α we have more repetitions (hence the parameter is called concentration).

Samples from DP

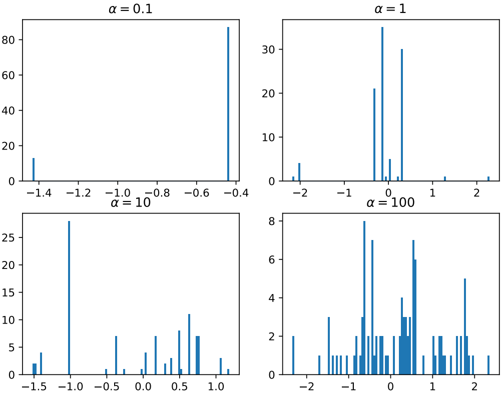
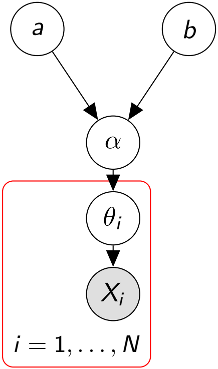


Figure: Histogram of samples from the DP. For smallest α we have only two unique samples!

DP GMM



$$
\begin{aligned}
\alpha &\sim \mathrm{Gamma}(a, b) \\
\theta_i &\sim DP(\alpha, H) \\
x_i &\sim F(\theta_i)
\end{aligned}
$$

Because values in the Dirichlet process repeat we will have mixture model with number of clusters equall to unique θ. This number **is random in nature**. In general we expect, that number of unique clusters is roughly $\alpha \log\left(1 + \frac{N}{\alpha}\right)$.

Stick-Breaking

There is no easy way to use the previous method in case of the inference. Hence in order to solve the problem other representation is more useful. Let's assume, that we are creating an infinite amount of samples. **Unique samples** will be denoted as θ_i and amount of samples equal to θ_i is n_i. It turns out that weights

$$\pi_i = \frac{n_i}{\sum_0^\infty n_i}$$

are well defined and their distribution is known analytically.

Stick-Breaking

$$\beta_n \sim \mathrm{Beta}(1, \alpha)$$

$$\pi_n = \prod_{i=1}^{n-1} (1 - \beta_i) \beta_n$$

This is the so-called Stick-Breaking process. It turns out that this particular way of DP realization is much easier to use in practice as we can directly compute likelihood. This model is well-defined with an infinite number of clusters hence it is often called **Infinite Mixture Model**.

How does it work?

- initialize β_i up to some value K,
- compute weights up to K,
- use simple GMM prescription for the rest of the components.

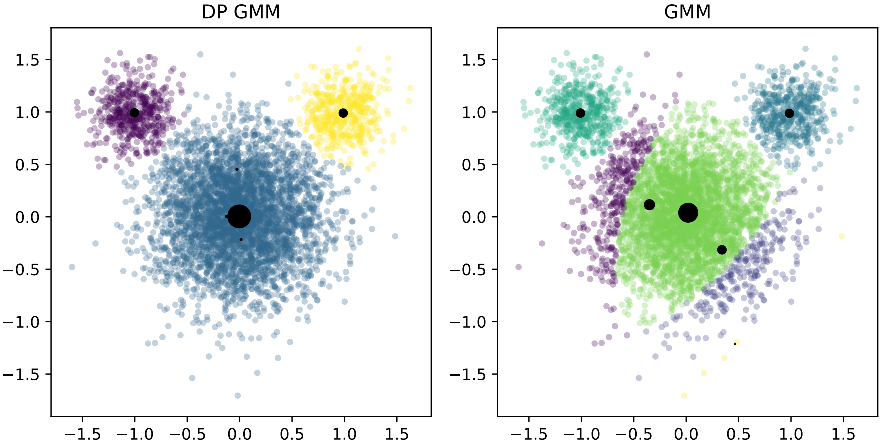
Models are very similar, what is the main difference? Instead of using normal weights we are replacing them with those generated using Stick-Breaking method. Do we get any benefits form such a complicated procedure?

Testing

Lets try simple setup:

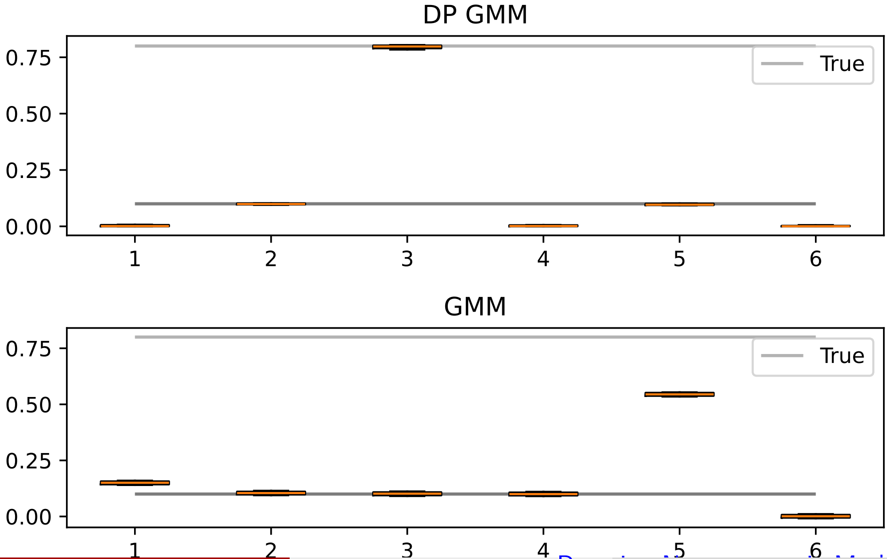
- Mickey Mouse dataset (3 gaussians),
- DP GMM - Hamiltonian Monte Carlo sampling,
- GMM - EM algorithm used to infer MAP values.
- Each model is used with 6 components, which one will give better description of the dataset?

Predictions



Colour - which cluster given point belongs to, black dot - center of cluster (sizes corresponds to associated weights).

Predicted weights for the components



We can see, that DPGMM is much more regularized and actually finds the configuration used to generate data contrary to simple GMM. There are multiple setups where it is impossible to run the model many times. Therefore, computing BIC for different K values in terms of the GMM may be impossible. This makes this particular model very popular in the case of huge datasets.

Mixture Models

The true potential of the Dirichlet Process is unveiled in so-called population models. There are various examples, especially in the field of bioinformatics. In general, the potential is still there and there are interesting ways to explore it. Idea: modelling microlensing populations. Let's:

- parametrize potential populations of objects (for example power-law mass distribution with cutoff values),
- assign probability weights to populations using the Stick-Breaking method,
- compute observable properties of the combined population (for example distribution of Einstein timescales),
- compare t_E distribution with observed one and infer number and properties of unique populations in the model.

Easy population statistics with a "learned" number of components!

Other use cases & conclusions

- Dirichlet Process can be used in the efficient manner for the clustering and other mixture models where number of components is unknown.
- Other applications include regression and density estimation.
- In principle various extensions to Dirichlet process exist and can be used in various setups for the inference tasks.

Thank you for your attention!

References to the topics discussed during the meeting on the next page

Bonus

- Most popular application - Topic modelling using LDA (50000+ citations).
- Astronomical motivation - Nonparametric Black Hole mass inference.
- Implementation using Pyro PPL.