
# Gaia CU9 Grand Challenge

Data Mining group, Spanish Virtual Observatory

30 de marzo de 2016

This presentation is based in a recent paper submitted to the Information Sciences journal: *Enabling Data Science in the Gaia Mission Archive: The Initial Mass Function and the Star Formation Rate* by D. Tapiador, A. Berihuete, L.M. Sarro, F. Julbe, E. Huedo.

# Table of contents

# What is the question?

# What is the question?

**General abstracts in papers related with GAIA:**

*The European Space Agency's Gaia mission will represent a tremendous discovery potential. It will create the largest and most precise three dimensional chart of our Galaxy, providing unprecedented position, parallax, proper motion, and radial velocity measurements for about one billion stars.*

BUT

can we offer the proper infrastructure and middleware upon which scientists will be able to do exploration and modeling with this huge data set?

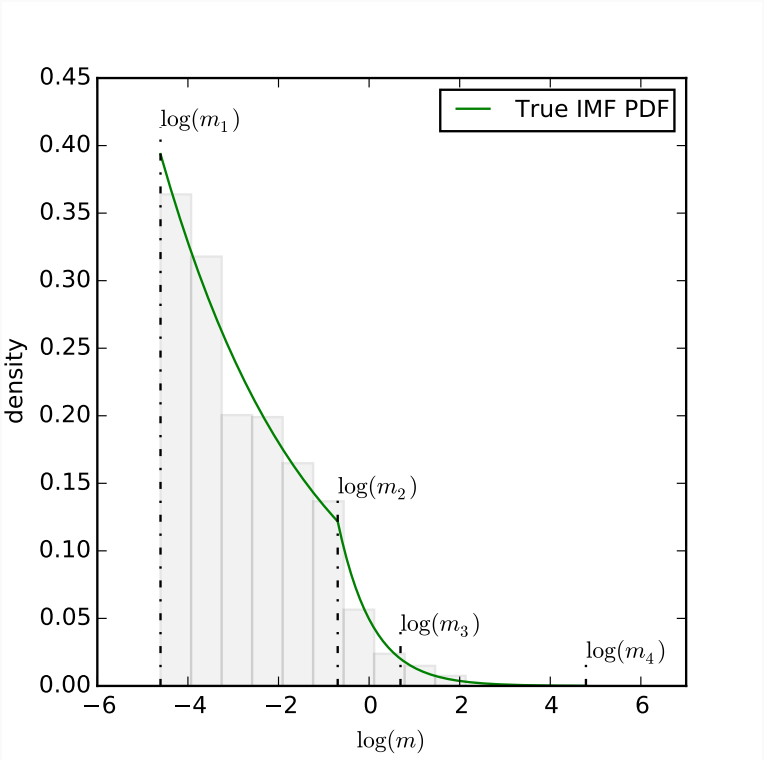
# What is the question?

In our paper:

*we present and contextualize these challenges, along with the decisions made and their justifications. Moreover, we exemplify these circumstances by building two probabilistic models using Hierarchical Bayesian Modelling*

# The probabilistic models

# The initial mass function




## Kroupa IMF

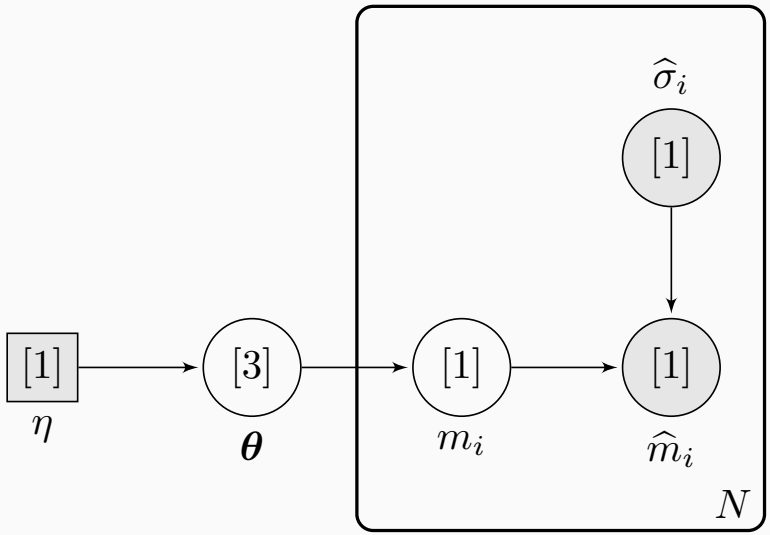
The Initial Mass Function (IMF) describes the distribution of initial true (**unknown**) masses for a population of stars. We establish the hypothesis that the IMF can be expressed as

$$\xi(m;\boldsymbol{\theta}) = c_j m^{-\theta_j},\ M_j < m \leq M_{j+1},$$

where, $j = 1, 2, 3$.

IMF PDF by setting $\boldsymbol{\theta} = (1.3, 2.3, 2.3)$

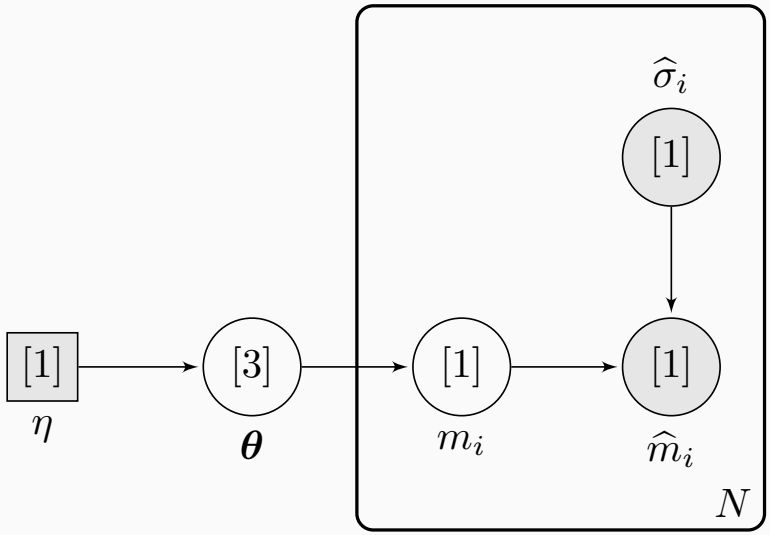
# The initial mass function




## Hierarchical Bayesian Model

The graphical model describes how the variables in the problem are related, stablishing a probabilistic framework in terms of a hierachical Bayesian model.

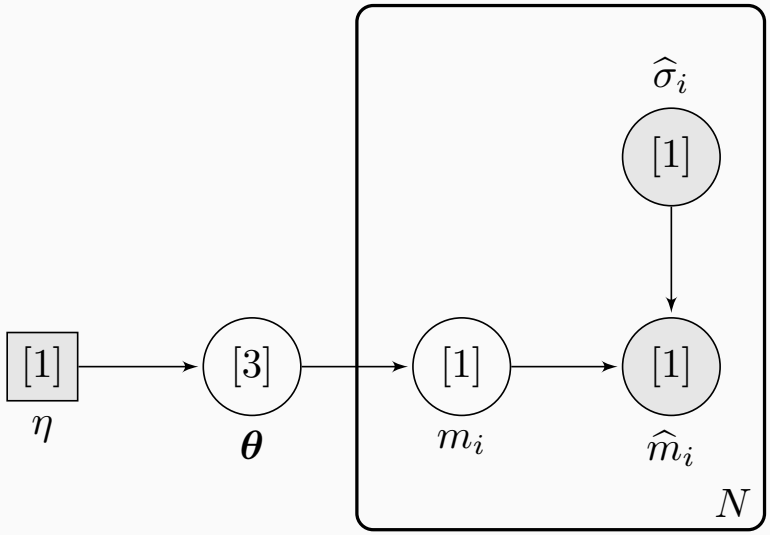
# The initial mass function




## So what?

In our HBM, the vector of model parameters $\boldsymbol{\theta}$ is itself treated as a random variable, the distribution of which we aim to infer. By inferring the probability distribution of $\boldsymbol{\theta}$ given the data $\mathcal{D}$, we infer the IMF.

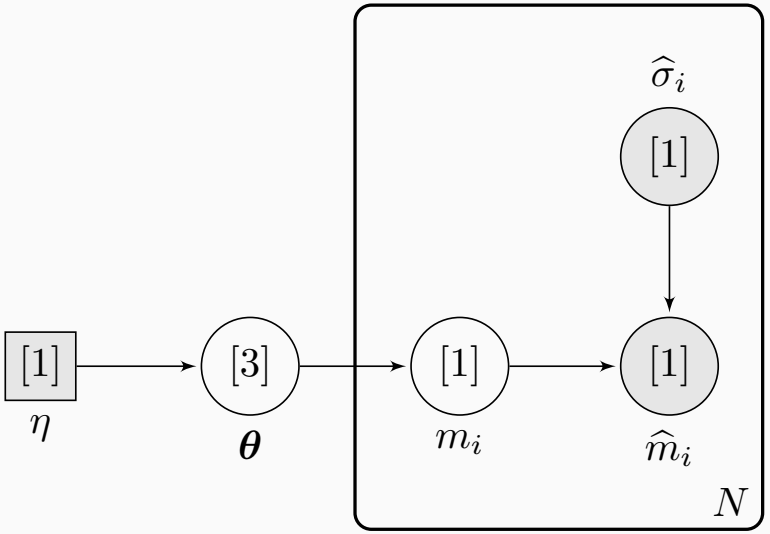
# The initial mass function




## So what?

The main goal of astronomers is to move in the inverse way: from observations to model parameters. In our example, from the masses and their uncertainties, $\mathcal{D} = \{\hat{m}_i, \hat{\sigma}_i\}_{i=1}^{N}$, to the parameter $\boldsymbol{\theta}$.

# The initial mass function




**How can we do that?**

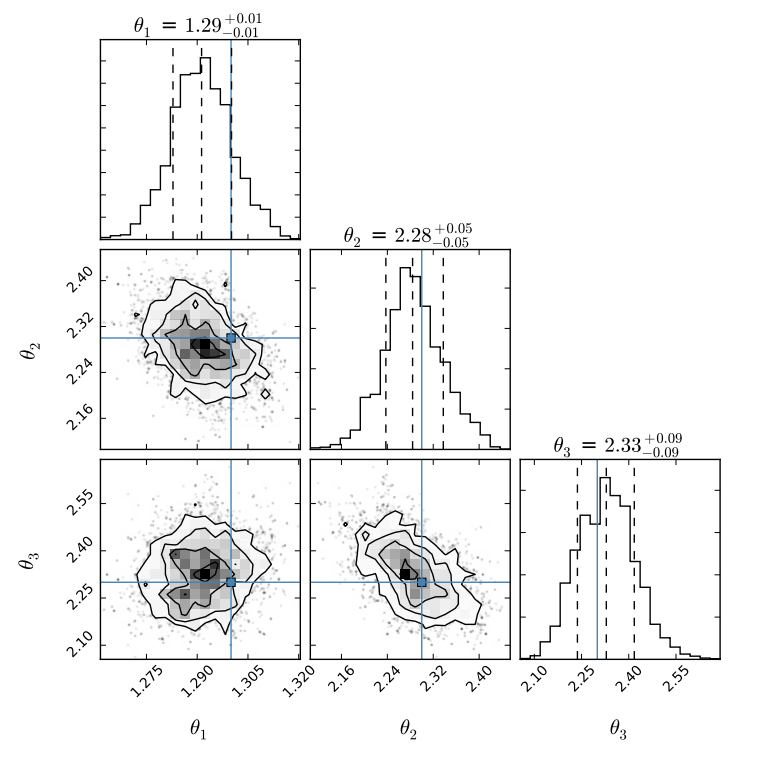
First, doing Bayes,

$$p(\boldsymbol{\theta}|\mathcal{D},\eta) \propto \prod_{i=1}^{N} p(\hat{m}_i|\hat{\sigma}_i,\boldsymbol{\theta})p(\boldsymbol{\theta}|\eta),$$

but also marginalising the likelihood:

$$\begin{aligned} p(\hat{m}_i|\hat{\sigma}_i,\boldsymbol{\theta}) &= \int p(\hat{m}_i,m|\hat{\sigma}_i,\boldsymbol{\theta})\mathrm{d}m \\ &= \int p(\hat{m}_i|m,\hat{\sigma}_i,\boldsymbol{\theta})p(m|\hat{\sigma}_i,\boldsymbol{\theta})\mathrm{d}m \\ &= \int p(\hat{m}_i|m,\hat{\sigma}_i)p(m|\boldsymbol{\theta})\mathrm{d}m. \end{aligned}$$

# The initial mass function




**How can we do that?**

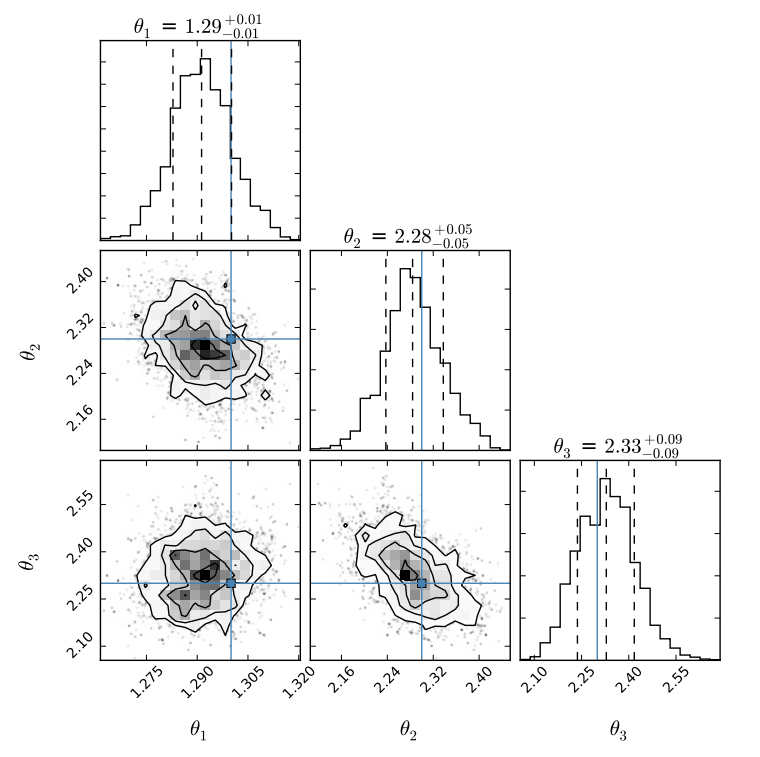
We draw independent and identical distribution samples from

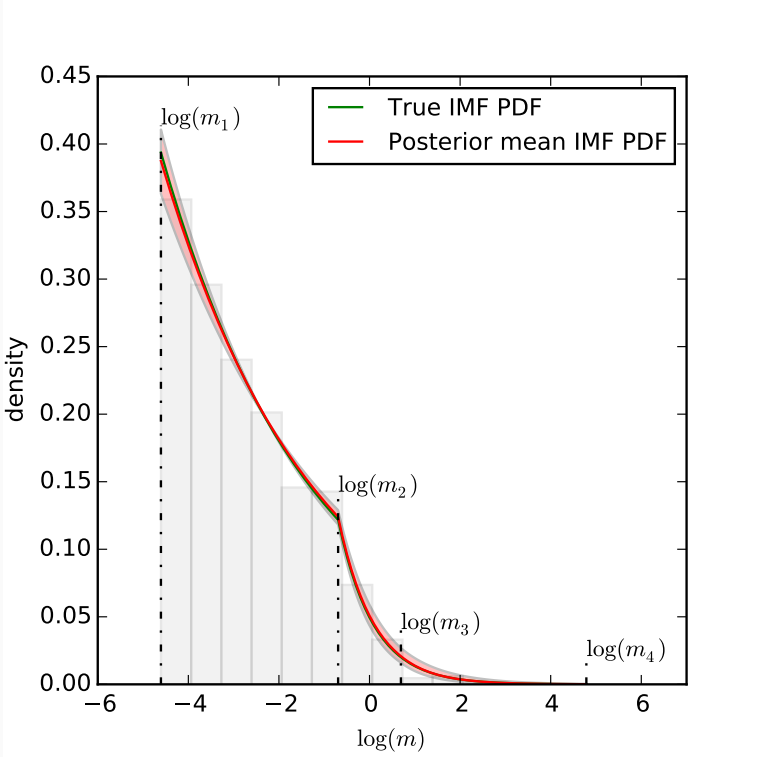
$$p(\boldsymbol{\theta}|\mathcal{D},\eta) \propto \prod_{i=1}^{N} p(\hat{m}_i|\hat{\sigma}_i,\boldsymbol{\theta})p(\boldsymbol{\theta}|\eta),$$

by using emcee algorithm (`http://dan.iel.fm/emcee/current/`)

The true value was $\boldsymbol{\theta} = (1.3, 2.3, 2.3)$, and the estimated value $\hat{\boldsymbol{\theta}} = (1.29^{+0.01}_{-0.01}, 2.28^{+0.05}_{-0.05}, 2.33^{+0.09}_{-0.09})$ by calculating the quantiles 0.16, 0.50 and 0.84 of the posterior samples.

# The initial mass function




The true value was $\boldsymbol{\theta} = (1.3, 2.3, 2.3)$, and the estimated value $\hat{\boldsymbol{\theta}} = (1.29^{+0.01}_{-0.01}, 2.28^{+0.05}_{-0.05}, 2.33^{+0.09}_{-0.09})$ obtained by calculating the quantiles 0.16, 0.50 and 0.84 of the posterior samples.

# The star formation rate

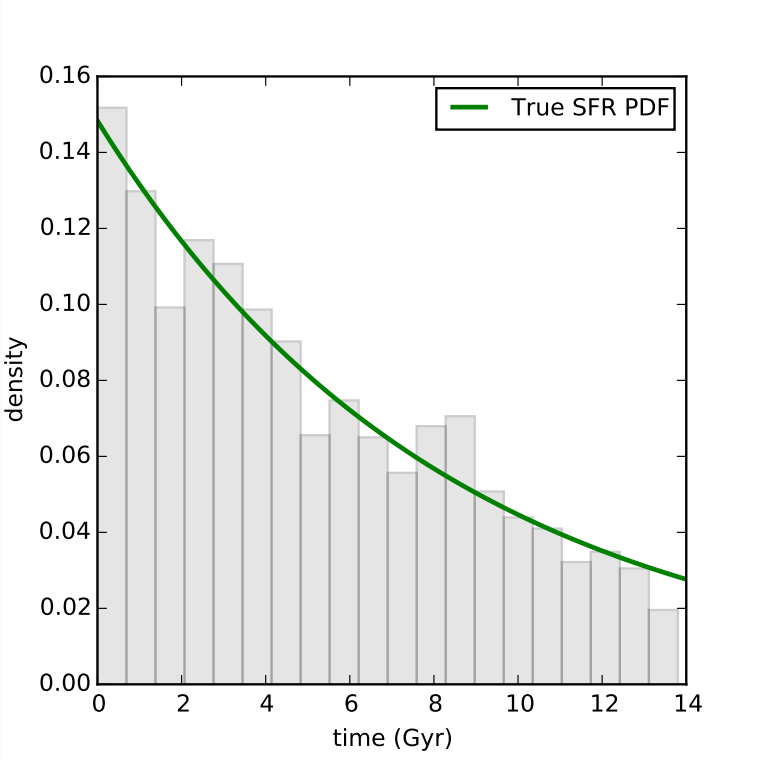
The Star Formation Rate (SFR) is defined as $\zeta(t)$, the total mass (expressed in units of the solar mass) transformed into stars at time $t$. We would like to be able to infer any arbitrary shape of the SFR function, except maybe discontinuous ones. If we assume continuity and smoothness of the SFR, we can define it as:

$$\zeta(t) = \sum_{k=1}^{K} w_k \phi_k(t) = \boldsymbol{\phi}^T(t) \cdot \boldsymbol{w},$$

we can choose

$$\boldsymbol{\phi}(t) = (e^{-0.5(t-t_{01})^2}, \quad e^{-0.5(t-t_{02})^2}, \ldots, e^{-0.5(t-t_{0K})^2})^T$$

# The star formation rate




## SFR PDF

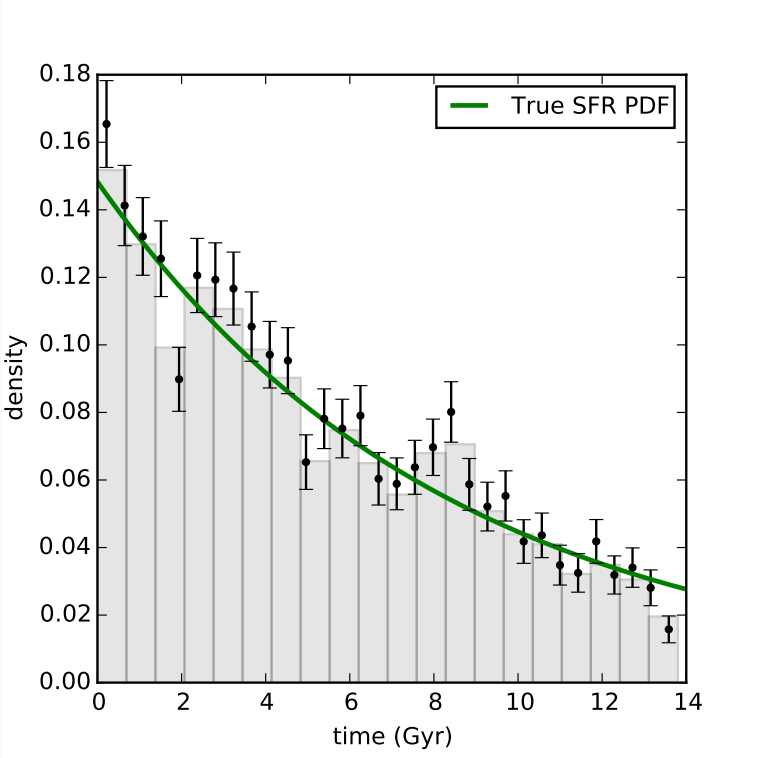
Instead of $\zeta(t)$ which has units of solar masses per unit time, we will work with $\zeta'(t)$ defined as the fraction of stars (of whatever mass) created per unit time:

$$\zeta'(t) = \frac{\sum_k w_k \phi_k(t)}{\sum_k w_k \int_a^b \phi_k(t)\,\mathrm{d}t}.$$

Thus, the SFR defines a non-stationary Poisson process with intensity function $\zeta'(t)$, $t \geq 0$. Figure shows

$$\zeta'(t) = \frac{0.12}{(1 - \exp(-13.8 \cdot 0.12))} \exp(-0.12t)$$

# The star formation rate




## SFR PDF

Given the data set available from Gaia (the posterior samples of mass and age $p_i(m, (13.8 - t)|\mathcal{D})$ produced by the FLAMES module) we could estimate the function $\zeta'$ at a set of times $\boldsymbol{t}_1 = \{t_{11}, t_{12}, \ldots, t_{1Q}\}$ as

$$\hat{\zeta}'(t_{1j}) = \frac{1}{N} \int_{m_{min}}^{m_{max}} \sum_{i=1}^{N} p_i(m, (13.8 - t_{1j})|\mathcal{D}) \cdot \mathrm{d}m.$$

Function $\zeta'(t)$ is a non-stationary Poisson process. Furthermore $\zeta'(t)$ and $\hat{\zeta}'(t)$ will be related by

$$\hat{\zeta}'(t) \sim \mathcal{P}(\zeta'(t)) \sim \mathcal{N}(\zeta'(t), \sigma(t)),$$

# The star formation rate

Let $\zeta'_{t_1}$ be a finite restriction of $\zeta'$ to values at reference times $\boldsymbol{t}_1 = [t_{11}, \ldots, t_{1Q}]$. Then,

$$p(\hat{\zeta}'_{t_1}|\zeta'_{t_1}) = p(\hat{\zeta}'_{t_1}|\boldsymbol{w}, \boldsymbol{\phi}_{t_1}) = \mathcal{N}(\hat{\zeta}'_{t_1}; \boldsymbol{\phi}_{t_1}^T \boldsymbol{w}, \Sigma_{\boldsymbol{\epsilon}}),$$

where $\Sigma_{\boldsymbol{\epsilon}}$ is a $Q \times Q$ diagonal matrix containing $\sigma^2(t_{1i})$ in the diagonal, and $\boldsymbol{\phi}_{t_1}$ is the $K \times Q$ matrix

$$\boldsymbol{\phi}_{t_1} = \begin{pmatrix} 1 & \phi_1(t_{12}) & \ldots & \phi_1(t_{1Q}) \\ \phi_2(t_{11}) & 1 & \ldots & \phi_2(t_{1Q}) \\ \ldots & \ldots & \ldots & \ldots \\ \phi_K(t_{11}) & \phi_K(t_{12}) & \ldots & 1 \end{pmatrix}.$$

# The star formation rate

Let us assume a $K$-dimensional multivariate normal distribution for the prior probability distribution of $\boldsymbol{w}$,

$$p(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}; \boldsymbol{\mu}_w, \Sigma_w),$$

and consider the vector $\zeta'_{\boldsymbol{t}_2}$ for some finite sequence of times $\boldsymbol{t}_2 = \{t_{21}, t_{22}, \ldots, t_{2M}\}$ that includes and extends the times $\boldsymbol{t}_1$ (that is, $\boldsymbol{t}_1 \subset \boldsymbol{t}_2$). It is possible to reorder $\zeta'_{\boldsymbol{t}_2}$ as $(\zeta'_{\boldsymbol{t}}, \zeta'_{\boldsymbol{t}_1})^T$, where $\boldsymbol{t}$ contains the points in $\boldsymbol{t}_2$ excluding $\boldsymbol{t}_1$. Furthermore, $(\zeta'_{\boldsymbol{t}}, \zeta'_{\boldsymbol{t}_1})^T = \zeta'_{\boldsymbol{t}_2} = \phi^T_{\boldsymbol{t}_2} \boldsymbol{w}$ is also distributed as a multivariate normal because it is a linear combination of $\boldsymbol{w}$.

# The star formation rate

In this situation, we can apply the property of closure under conditioning thus obtaining

$$p(\zeta_t'|\hat{\zeta}_{t_1}', \boldsymbol{\phi}_{t_1}) = \mathcal{N}(\zeta_t'; \boldsymbol{\mu}^{\text{post}}, \Sigma^{\text{post}}),$$

where

$$\begin{aligned}
\boldsymbol{\mu}^{\text{post}} &= \boldsymbol{\phi}_t^T \boldsymbol{\mu}_w + \boldsymbol{\phi}_t^T \Sigma_w \boldsymbol{\phi}_{t_1} (\boldsymbol{\phi}_{t_1}^T \Sigma_w \boldsymbol{\phi}_{t_1} + \Sigma_\varepsilon)^{-1} (\hat{\zeta}_{t_1}' - \boldsymbol{\phi}_{t_1}^T \mu_w), \\
\Sigma^{\text{post}} &= \boldsymbol{\phi}_t^T \Sigma_w \boldsymbol{\phi}_t - \boldsymbol{\phi}_t^T \Sigma_w \boldsymbol{\phi}_{t_1} (\boldsymbol{\phi}_{t_1}^T \Sigma_w \boldsymbol{\phi}_{t_1} + \Sigma_\varepsilon)^{-1} \boldsymbol{\phi}_{t_1}^T \Sigma_w \boldsymbol{\phi}_t.
\end{aligned}$$

At this point we see that the size of the matrices increases with the number of features $K$, which enters the calculations via $\boldsymbol{w}$.

# The star formation rate

In order to avoid the problem of increasing the number of fatures $K$, we have worked with Gaussian Process:

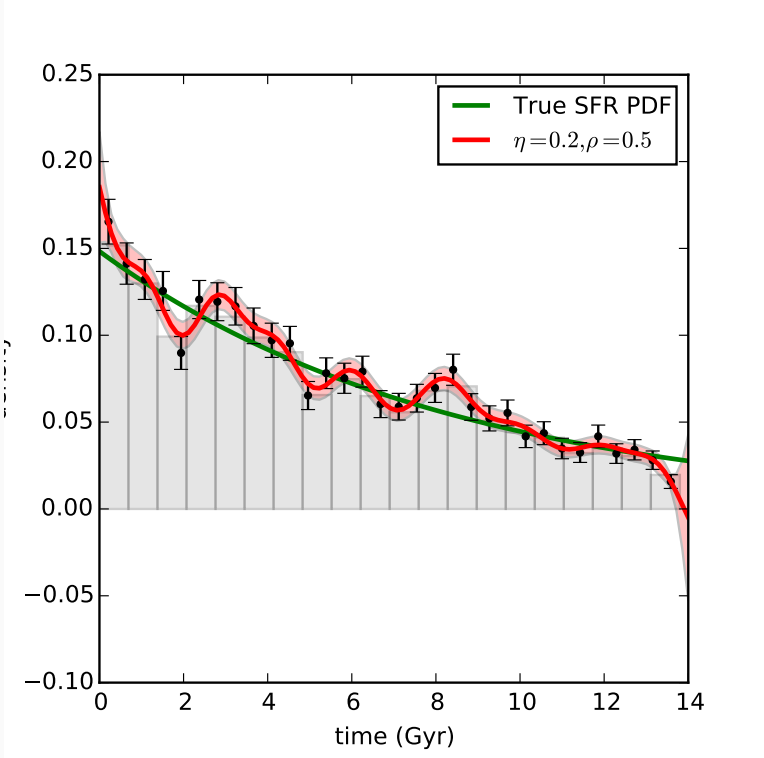
**Definition**

Let $\nu : \mathbb{X} \to \mathbb{R}$ be any function, called the mean function, and let $\kappa : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ be a Mercel kernel. A Gaussian Process $p(f) = \mathcal{GP}(f; \nu, \kappa)$ is a probability distribution over the function $f : \mathbb{X} \to \mathbb{R}$, such that every finite restriction to function values $f_X := [f_{x_1}, \ldots, f_{x_K}] = [f(x_1), f(x_2), \ldots, f(x_K)]$ is Gaussian distributed $p(f_X) = \mathcal{N}(f_X; \nu_X, \kappa_{XX})$

Example

$$\begin{aligned} \nu_X &= 0 \\ \kappa_{X,X}(i,j) &= \kappa(x_j, x_j) = \phi^T(x_i)\phi(x_j). \end{aligned}$$

# The star formation rate



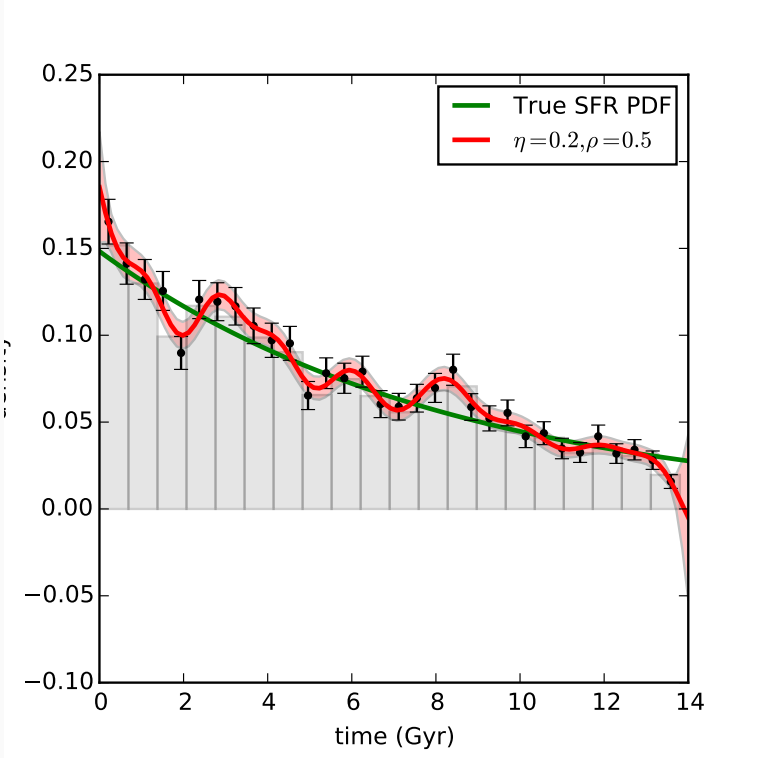

## How can we do that?

We tackle the problem of inferring the SFR PDF as a GP, by substituting every term which includes the feature function in $\boldsymbol{\mu}^{\text{post}}$ and $\Sigma^{\text{post}}$ in the posterior distribution, with their mean and kernel functions counterparts, namely

$$\begin{aligned}\boldsymbol{\mu}^{\text{post}} &= \nu_t - \kappa_{tt_1}\left(\kappa_{t_1t_1} + \Sigma_{\boldsymbol{\varepsilon}}\right)^{-1}\left(\hat{\zeta}'_{t_1} - \nu_{t_1}\right)\\ \Sigma^{\text{post}} &= \kappa_{tt} - \kappa_{tt_1}\left(\kappa_{t_1t_1} + \Sigma_{\boldsymbol{\varepsilon}}\right)^{-1}\kappa_{t_1t}\end{aligned}$$

where $\kappa_{tt}(i,j) = \eta^2 \exp(-\rho^2(t_i - t_j)^2)$ is a Mercel kernel and $\nu_t = \mathbf{0}$ is the mean function.

# The star formation rate



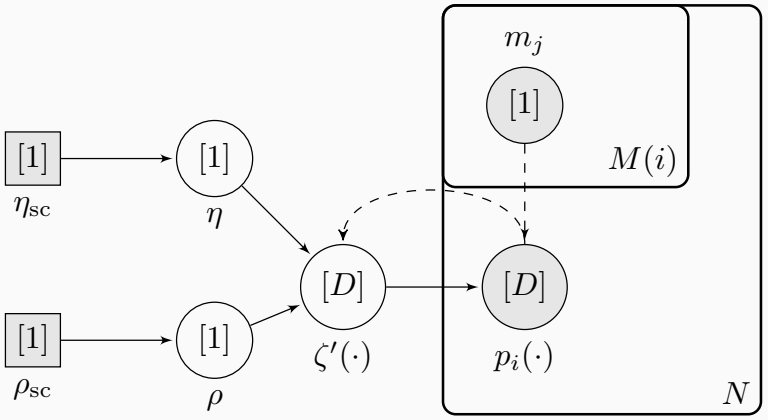

## How can we do that?

We tackle the problem of inferring the SFR PDF as a GP, by substituting every term which includes the feature function in $\boldsymbol{\mu}^{\text{post}}$ and $\Sigma^{\text{post}}$ in the posterior distribution, with their mean and kernel functions counterparts, namely

$$\begin{aligned}\boldsymbol{\mu}^{\text{post}} &= \nu_t - \kappa_{tt_1}\left(\kappa_{t_1t_1} + \Sigma_{\boldsymbol{\varepsilon}}\right)^{-1}\left(\hat{\zeta}'_{t_1} - \nu_{t_1}\right)\\ \Sigma^{\text{post}} &= \kappa_{tt} - \kappa_{tt_1}\left(\kappa_{t_1t_1} + \Sigma_{\boldsymbol{\varepsilon}}\right)^{-1}\kappa_{t_1t}\end{aligned}$$

where $\kappa_{tt}(i,j) = \eta^2 \exp(-\rho^2(t_i - t_j)^2)$ is a Mercel kernel and $\nu_t = \mathbf{0}$ is the mean function.

# The star formation rate

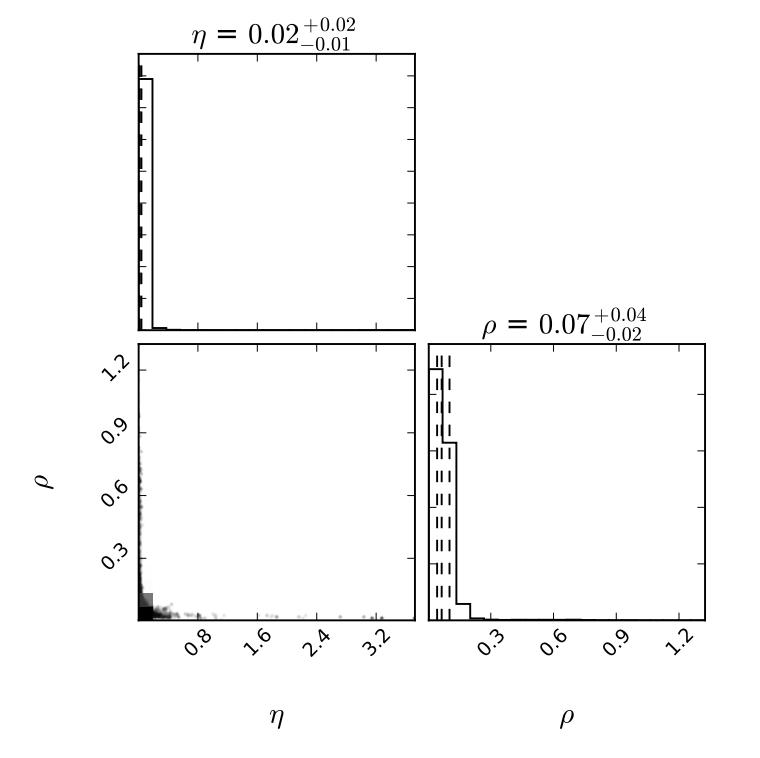


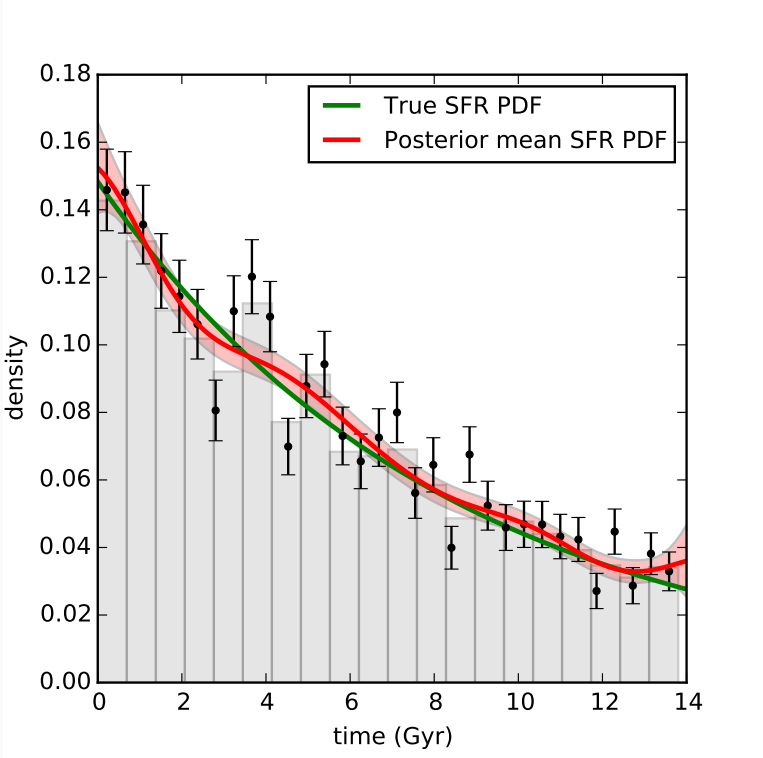

## Hierarchical model

The posterior distribution of the HBM parameters in the simple case of a single star observed would be

$$
\begin{aligned}
p(\eta, \rho | p_1(t), \zeta'_t, \hat{\zeta}'_{t_1}, \Sigma_\varepsilon) \quad \propto \quad & p(p_1(t) | \zeta'_t, \hat{\zeta}'_{t_1}, \Sigma_\varepsilon) \cdot \\
& p(\zeta'_t | \hat{\zeta}'_{t_1}, \Sigma_\varepsilon, \eta, \rho) \cdot \\
& p(\eta) \cdot p(\rho),
\end{aligned}
$$

# The star formation rate




The true SFR PDF (green line) was

$$\zeta'(t) = \frac{0.12}{(1 - \exp(-13.8 \cdot 0.12))} \exp(-0.12t)$$

# Results

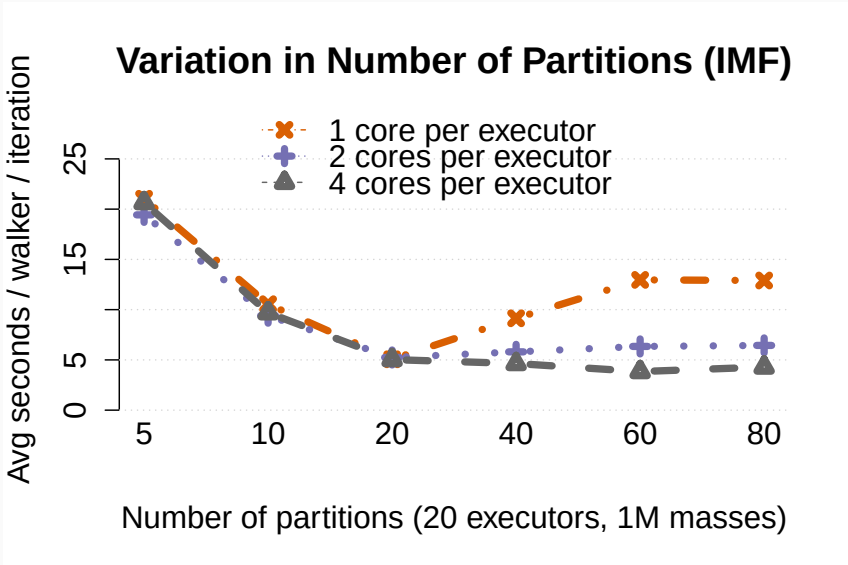
# Results

Tools:

We used emcee algorithm in Spark platform coding in python.

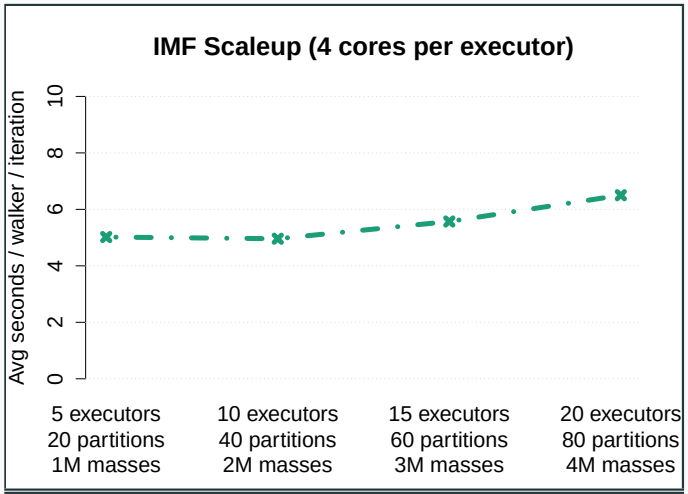


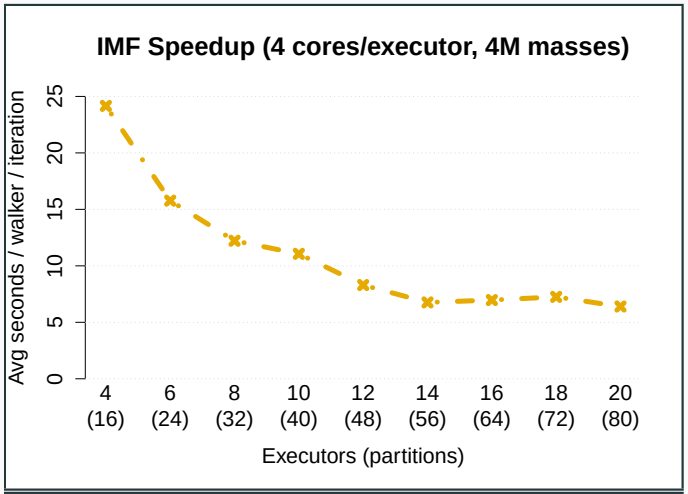
# Results

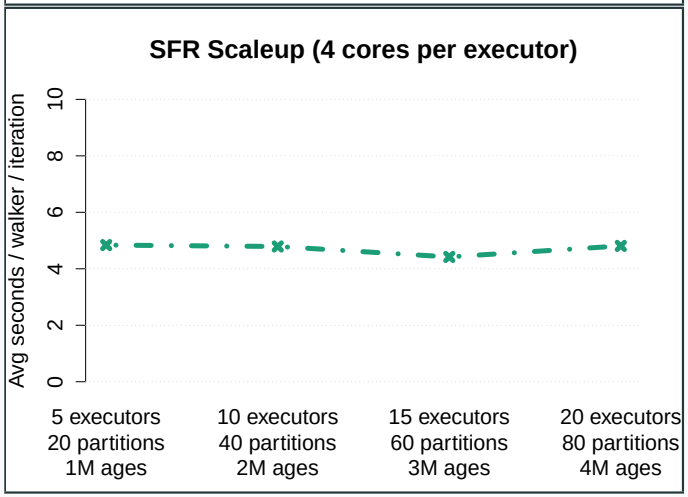


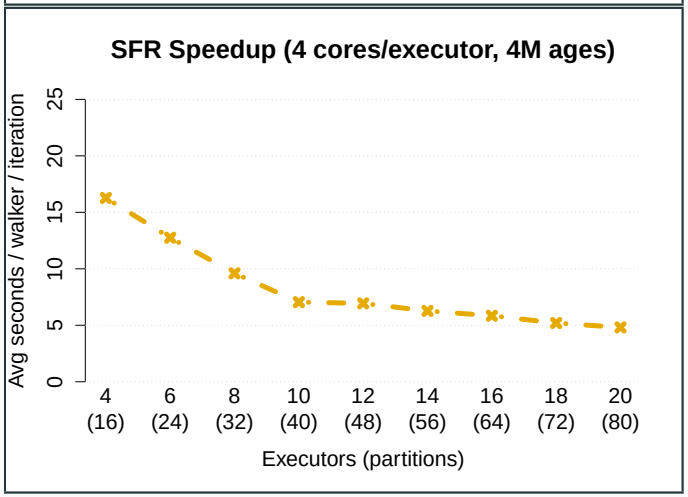

- To confirm the tasks (processing the likelihood) were homogeneous in execution time, we set a given number of executors.
- In order to distribute the load among all executors/cores available, we repartitioned the RDD containing the masses.

# Results


IMF Scaleup (4 cores per executor)
Avg seconds / walker / iteration
0
2
4
6
8
10
5 executors
20 partitions
1M masses
10 executors
40 partitions
2M masses
15 executors
60 partitions
3M masses
20 executors
80 partitions
4M masses


IMF Speedup (4 cores/executor, 4M masses)
Avg seconds / walker / iteration
0
5
10
15
20
25
4
(16)
6
(24)
8
(32)
10
(40)
12
(48)
14
(56)
16
(64)
18
(72)
20
(80)
Executors (partitions)


SFR Scaleup (4 cores per executor)
Avg seconds / walker / iteration
0
2
4
6
8
10
5 executors
20 partitions
1M ages
10 executors
40 partitions
2M ages
15 executors
60 partitions
3M ages
20 executors
80 partitions
4M ages


SFR Speedup (4 cores/executor, 4M ages)
Avg seconds / walker / iteration
0
5
10
15
20
25
4
(16)
6
(24)
8
(32)
10
(40)
12
(48)
14
(56)
16
(64)
18
(72)
20
(80)
Executors (partitions)

Questions?