

# Hierarchical Bayesian Modeling

Benjamin Johnson (Harvard/CfA)



My name is what? I use HBM for x  
we'll talk about HBM and some examples

## Introduction

- When is HBM appropriate? (with examples)
- Bayesian inference refresher/terminology/notation  
(individual/single-level)

## Hierarchical inference

- Individuals & Populations
- Conditional independence & graphs
- Inference of the prior (pooling)
- Improvements in individual posteriors (shrinkage)

Break & questions

## Hierarchical Examples

- Galaxy Mass Function
- Star Clusters

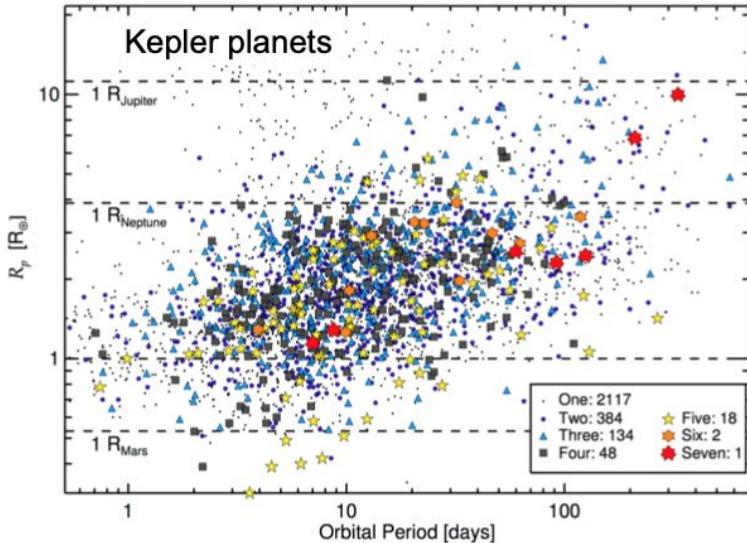
## Implementation Notes

- Gibbs sampling
- Pseudo-importance sampling
- Selection functions

# When is HBM appropriate?

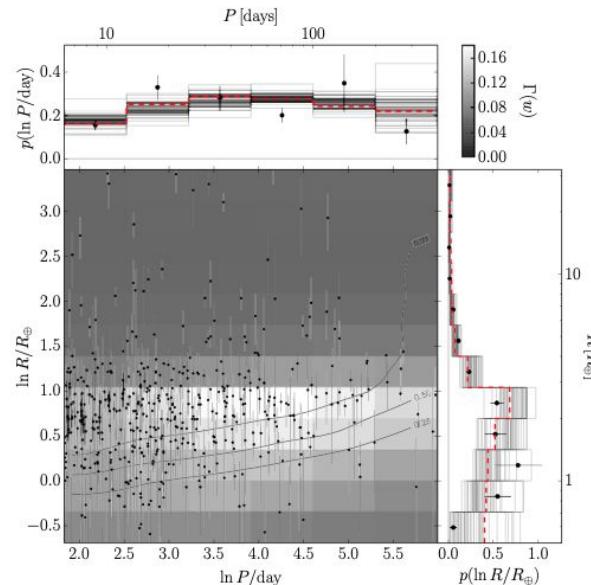
- You are interested in the ***population distribution*** of properties of individual objects, but the individual properties are uncertain.
- You want to combine many individual objects or observations to ***infer parameters of the population*** especially if the individuals carry different kinds of information.
- You want to use an observed population to ***improve estimates for individuals***

# When is HBM appropriate?



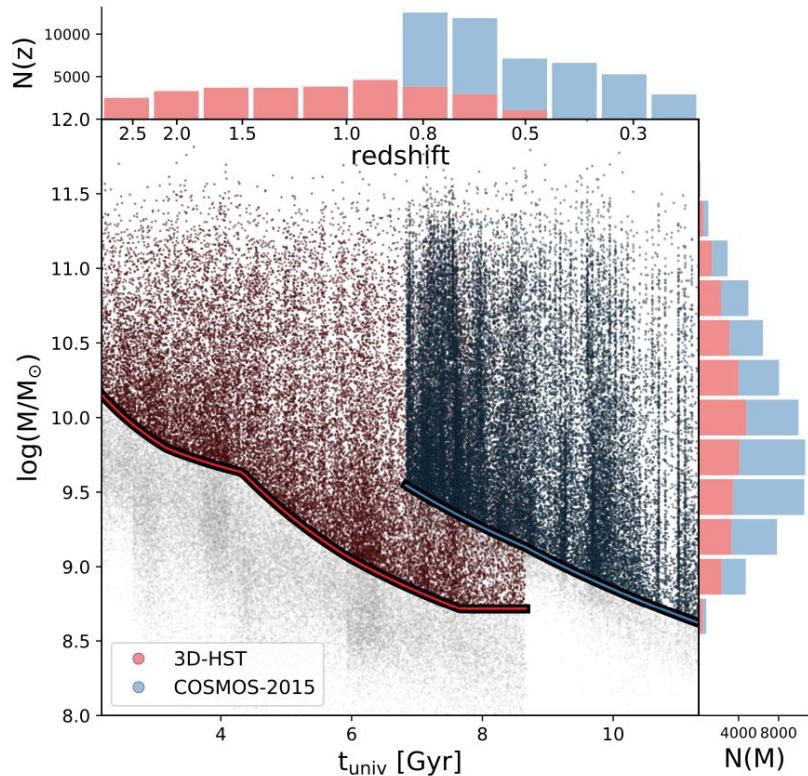
Lissauer, Dawson, & Tremaine, 2014

Given observations of many individual planets, what is the rate of earthlike planets around sunlike stars?



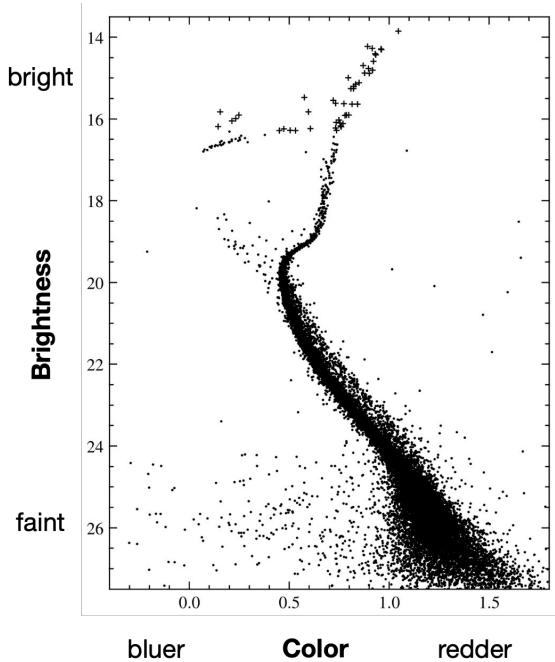
Foreman-Mackey et al. 2014

# When is HBM appropriate?



Given photometry and redshifts of many galaxies, what is the distribution of galaxy stellar masses, as a function of redshift?

# When is HBM appropriate?



Given photometry (and parallaxes) of individual member stars, what are the properties -- age, metallicity, distance -- of a star cluster?

# When is HBM appropriate?

Can also be used for “calibration”:

Inferring global ‘offset’ parameters from a population of many individuals while simultaneously applying the offset to individuals

# Bayesian Inference Refresher

Joint Probability

$$p(A, B) = \underbrace{p(A | B)}_{\text{conditional probability}} \times \underbrace{p(B)}_{\text{marginal probability}}$$

Bayes Theorem

$$p(A | B) p(B) = p(B | A) p(A)$$

Joint Probability for 3 variables

$$\begin{aligned} p(A, B, C) &= p(A, B | C) \times p(C) \\ &= p(A | B, C) \times p(B | C) \times p(C) \end{aligned}$$

# Bayesian inference refresher

## Conditional independence

A and B are independent conditional on C iff

$$p(A | B, C) = p(A | C) \quad \text{s.t.}$$

$$p(A, B, C) = p(A | C) \times p(B | C) \times p(C)$$

which also means

$$p(A, B | C) = p(A | C) \times p(B | C)$$

# Bayesian inference refresher

The standard Bayesian inference uses Bayes theorem as follows:

$$p(\theta | D) p(D) = p(D | \theta) p(\theta)$$

posterior      evidence      likelihood      prior

$\theta$   
parameters

$$D = \{d_j\}$$

dataset      individual data points  
or observations

The goal is to obtain an estimate for the **posterior** distribution.

# Bayesian inference refresher

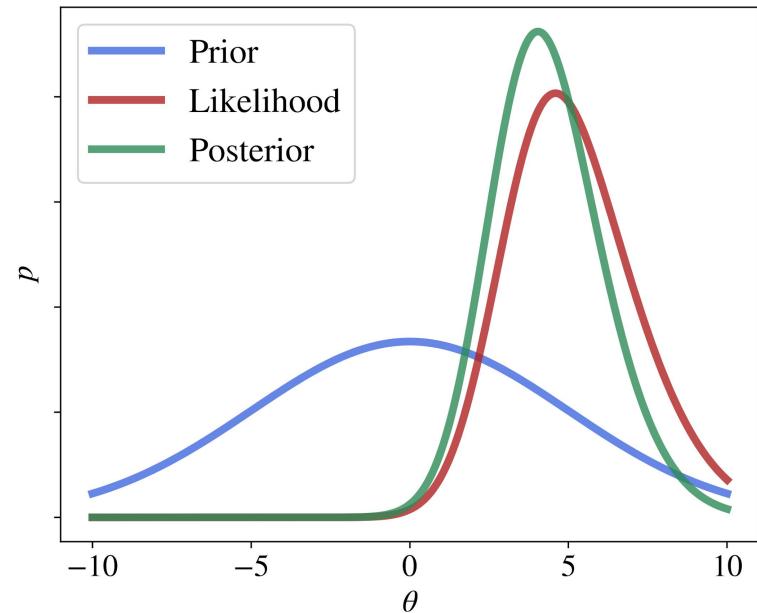
The standard Bayesian inference uses Bayes theorem as follows:

$$p(\theta | D) \propto p(\theta) p(D | \theta)$$



The prior may be physically motivated, or based on previous studies, or weakly informative (flat), but it is more or less ***arbitrarily constructed*** at this point.

Sneak peek: In HBM we can use the actual observed population to inform the prior



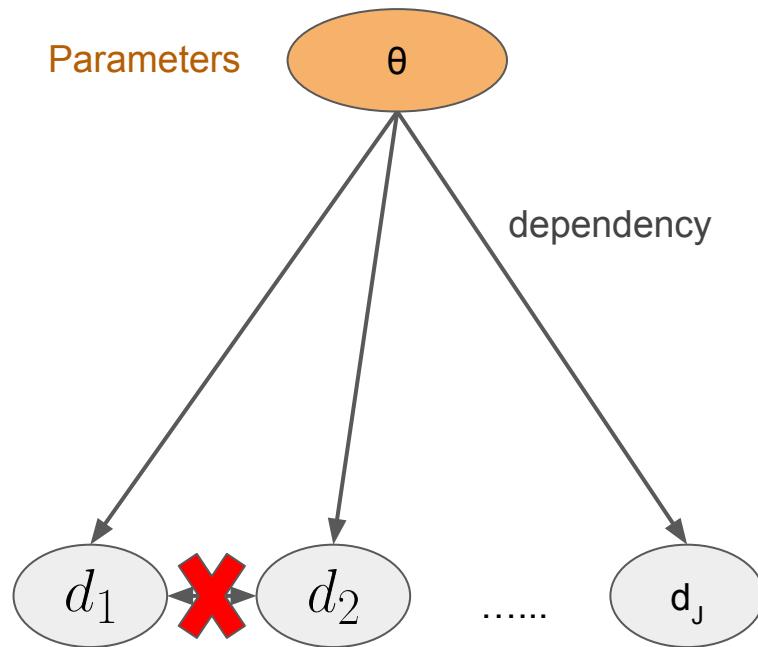
N.B. All normalized to integrate to 1

# Bayesian inference refresher

Iff the data points are **conditionally independent** of each other, we can rewrite this as something like

$$\begin{aligned} p(\theta | D) p(D) &= p(\theta) p(D | \theta) \\ &= p(\theta) p(d_1, d_2, \dots, d_J | \theta) \\ &= p(\theta) p(d_1 | \theta) p(d_2 | \theta) \dots p(d_J | \theta) \\ &= p(\theta) \prod_{j=1}^J p(d_j | \theta) \end{aligned}$$

conditional  
independence



N.B. If  $p(\theta)=1$  and the individual likelihoods are Gaussian, and you take the log of both sides, this quickly leads to the familiar chi-squared expression.

Data or observables

Now on to Hierarchical Inference

# HBM: individuals and populations

The previous few slides are relevant for inferring the properties of an *individual*. However, if you have many individuals, you may wish to:

- 1) Learn properties of the population **from** the individuals
- 2) Use the population to improve priors **for** the individuals



How do we do this self-consistently when we don't know the properties of the individuals before-hand?

Hierarchical Bayesian Modeling

# HBM: individuals and populations

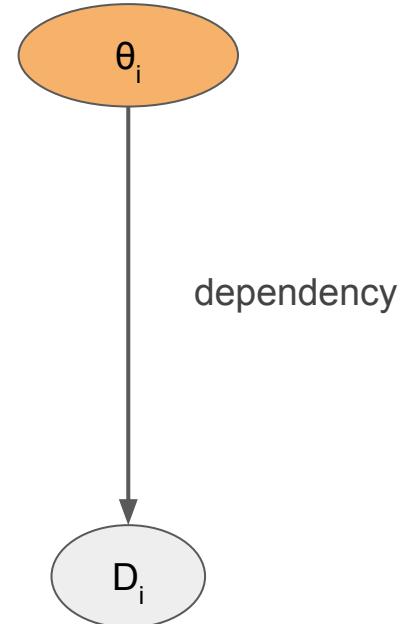
The equations on the previous few slides are relevant for inferring the properties of an individual.

If you have N individuals, you might write for one individual  $i$

$$p(\theta_i | D_i) p(D_i) = p(\theta_i) p(D_i | \theta_i)$$

Parameters  
of **one**  
individual

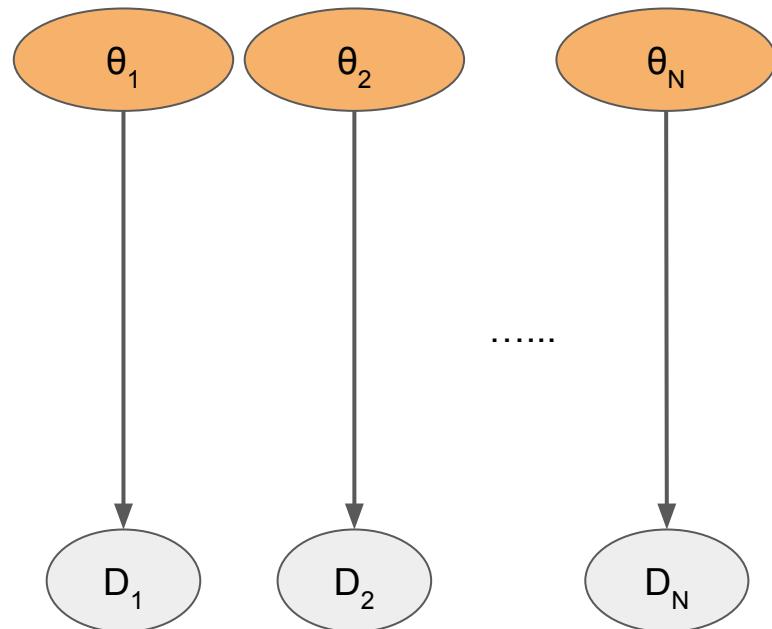
Data or  
observables  
for **one**  
individual



# HBM: individuals and populations

Because the individuals are independent, we can also write a simple expression for the probability of all the object parameters given all the data.

$$\begin{aligned} p(\{\theta_i\} | \{D_i\}) p(\{D_i\}) &= p(\{\theta_i\}) p(\{D_i\} | \{\theta_i\}) \\ &= p(\{\theta_i\}) p(D_1 | \theta_1) p(D_2 | \theta_2) \dots p(D_N | \theta_N) \\ &= p(\{\theta_i\}) \times \prod_{i=1}^N p(D_i | \theta_i) \\ &= p(\theta_1) p(\theta_2) \dots p(\theta_N) \times \prod_{i=1}^N p(D_i | \theta_i) \\ &= \prod_{i=1}^N p(\theta_i) p(D_i | \theta_i) \quad . \end{aligned}$$



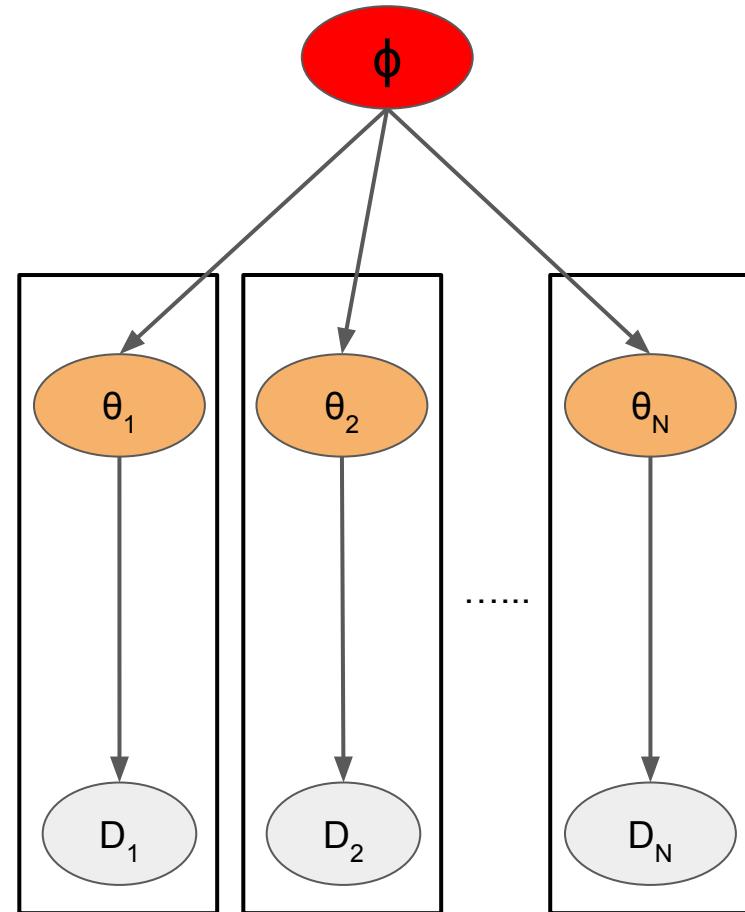
# HBM: individuals and populations

But, what if the distribution or probability of any  $\theta$  depends on properties of the population (e.g. the mass function)? This means that the prior probability on all the individuals  $p(\{\theta_i\})$  depends on some other, **population** parameters  $\phi$ . So now we have to write

$$\begin{aligned} p(\{\theta_i\}, \{D_i\}, \phi) &= p(\{D_i\}) p(\{\theta_i\}, \phi | \{D_i\}) \\ &= p(\phi) p(\{\theta_i\} | \phi) p(\{D_i\} | \{\theta_i\}, \phi) \end{aligned}$$

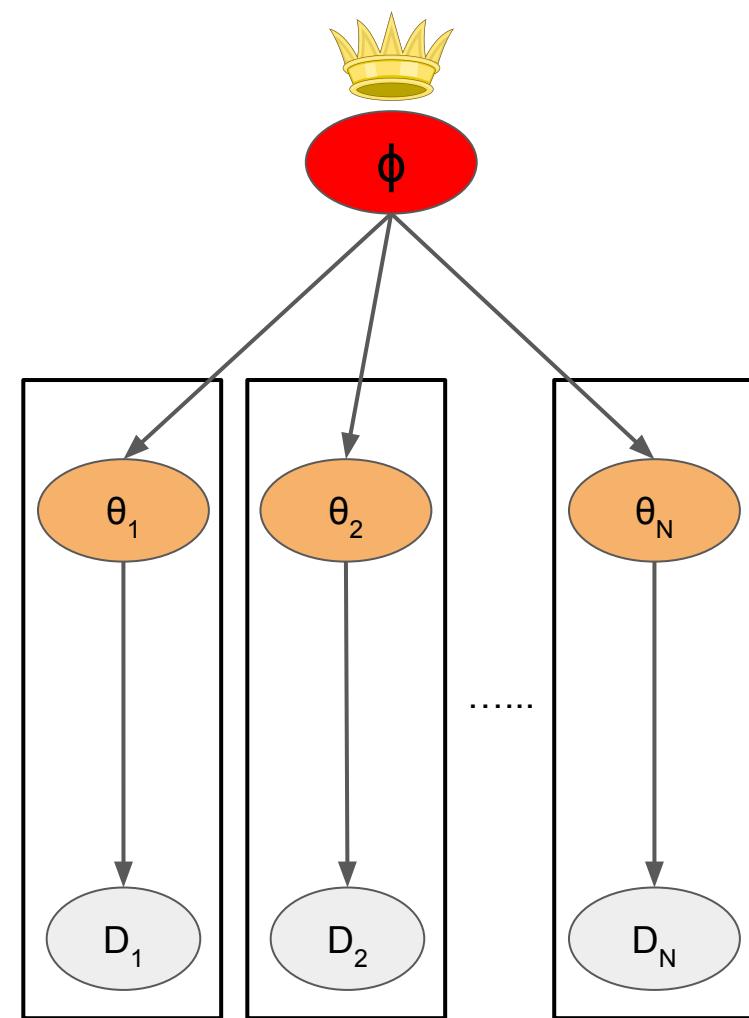
And iff the  $\theta_i$  are conditionally independent, we can write this as

$$\begin{aligned} p(\{D_i\}) p(\{\theta_i\}, \phi | \{D_i\}) &= p(\phi) p(\{\theta_i\} | \phi) p(\{D_i\} | \{\theta_i\}, \phi) \\ &= p(\phi) \prod_{i=1}^N p(\theta_i | \phi) p(D_i | \theta_i, \phi) \end{aligned}$$



# HBM: individuals and populations

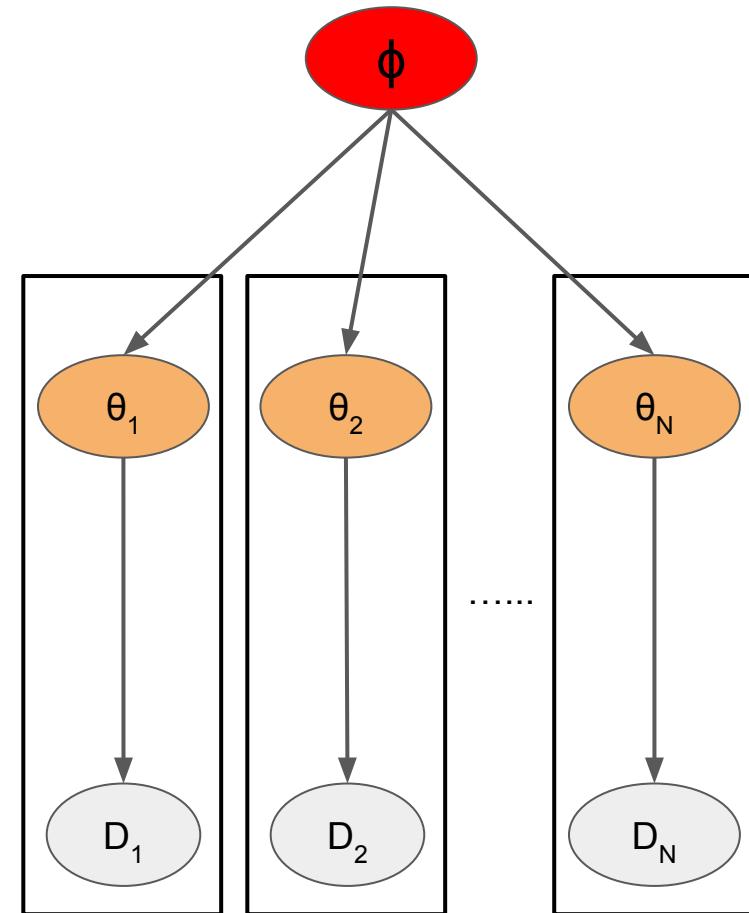
Now we have a ***hierarchy*** of parameters



# HBM: individuals and populations

$$p(\{\theta_i\}, \phi | \{D_i\}) = p(\phi) \prod_{i=1}^N p(\theta_i | \phi) p(D_i | \theta_i)$$

$$p(\Theta, \phi | \mathbf{D}) \propto p(\phi) p(\Theta | \phi) p(\mathbf{D} | \Theta, \phi)$$

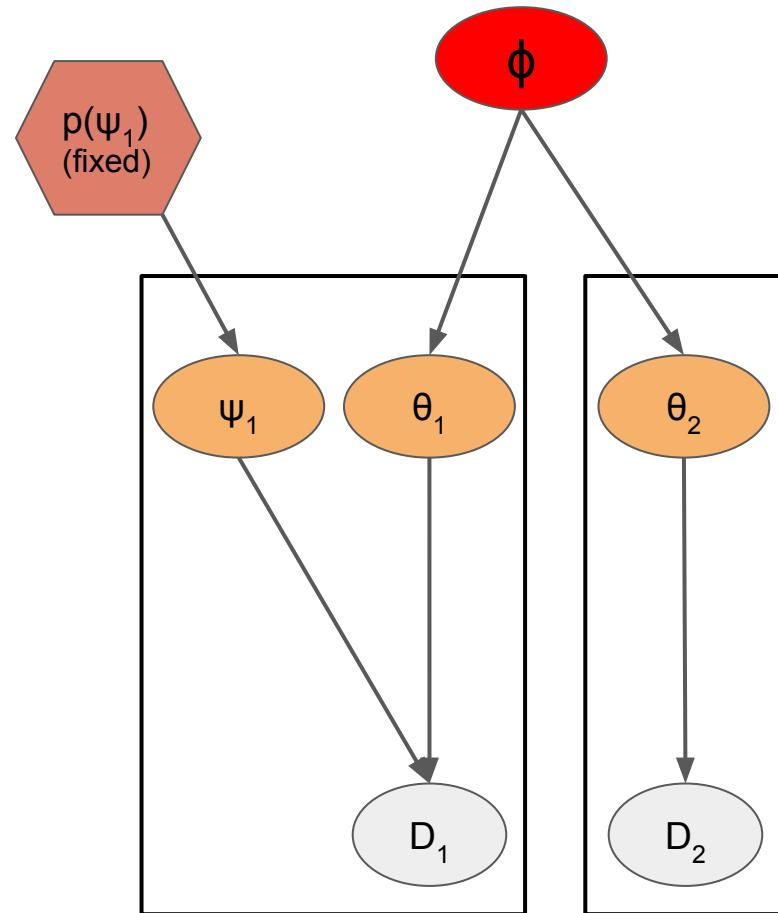


# HBM: individuals and populations

Note that each individual may have additional parameters ( $\psi$ ) that are *not* informed by the population.

These are effectively *marginalized* over when inferring  $\phi$

$$p(D_i | \theta_i) = \int d\psi_i p(D_i | \theta_i, \psi_i) p(\psi_i)$$



## HBM: individuals and populations

Since we have many individuals, we should be able to infer the value of  $\phi$  (marginalizing over all the individual parameters)

$$p(\phi | \mathbf{D}) \propto p(\phi) \int d\Theta p(\Theta | \phi) p(\mathbf{D} | \Theta, \phi)$$

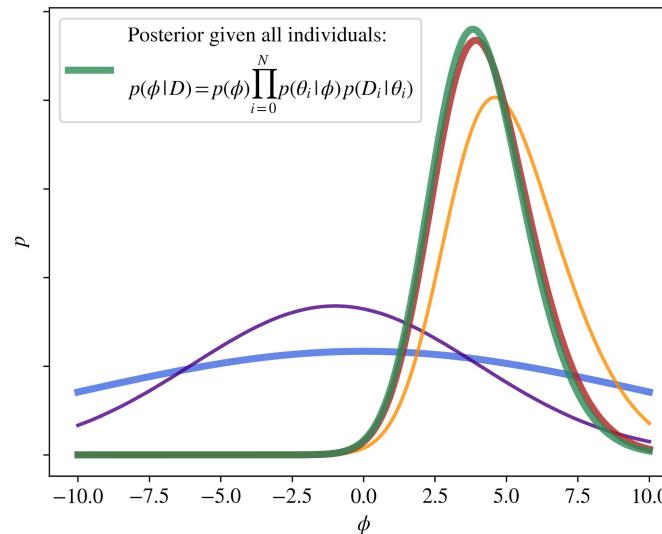
This is called ***pooling*** (since we are pooling estimates from individuals together)

# HBM: individuals and populations

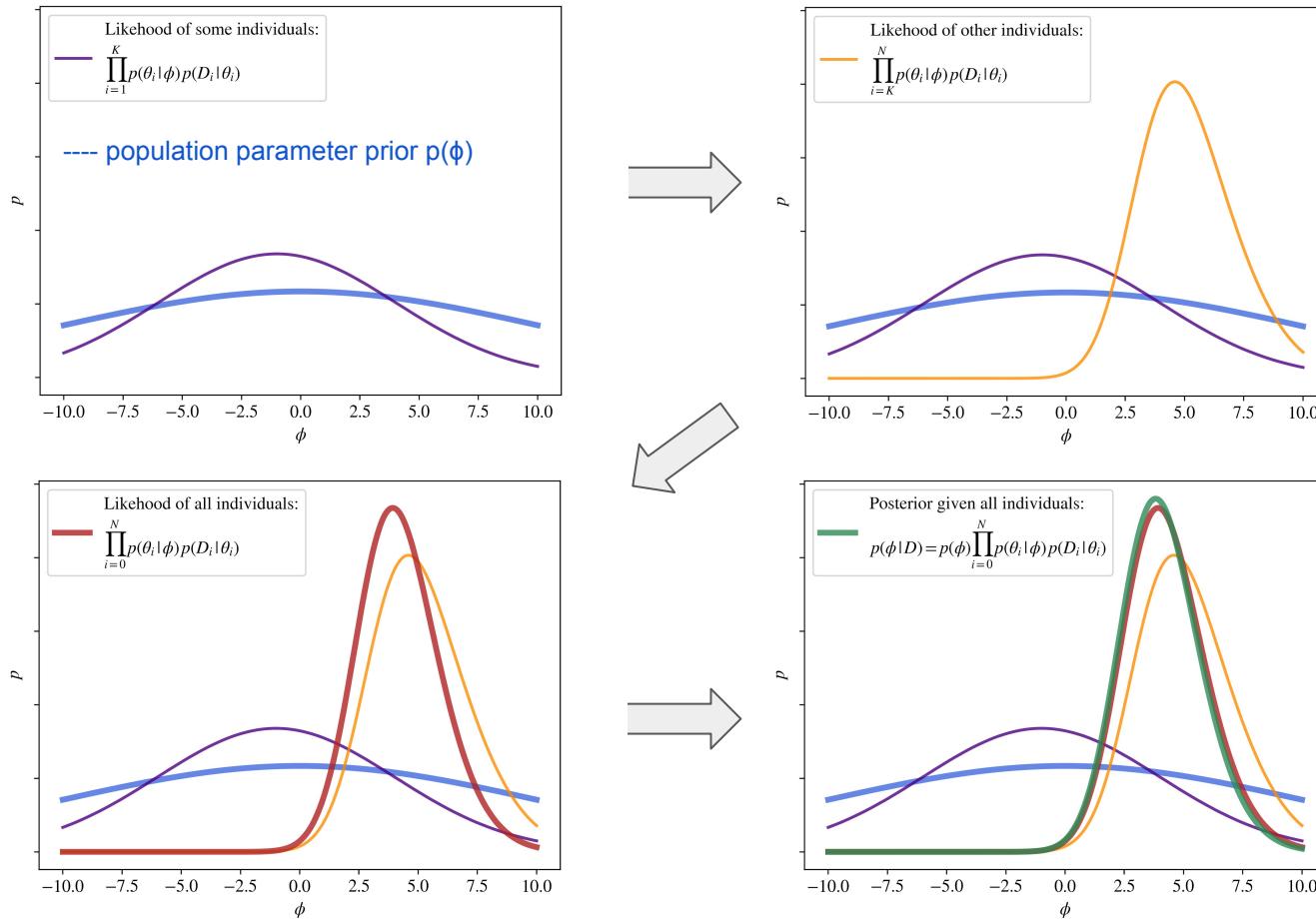
Since we have many individuals, we should be able to infer the value of  $\phi$  (marginalizing over all the individual parameters)

This is called ***pooling*** (since we are pooling estimates from individuals together)

The constraints on  $\phi$  will come from the most *informative* individuals, but are applied to *all* individuals



# HBM: individuals and populations *pooling*



# HBM: individuals and populations

Since we can infer and constrain  $\phi$  using the pooled estimate, the priors for each  $\theta_i$  are better known (often narrower) and we get better inference of  $\theta_i$  than if we treated the objects separately.

This effect of pooling is called **shrinkage** since the individual estimates will ‘shrink’ towards values predicted by the rest of the population.

This is a key advantage of HBM!

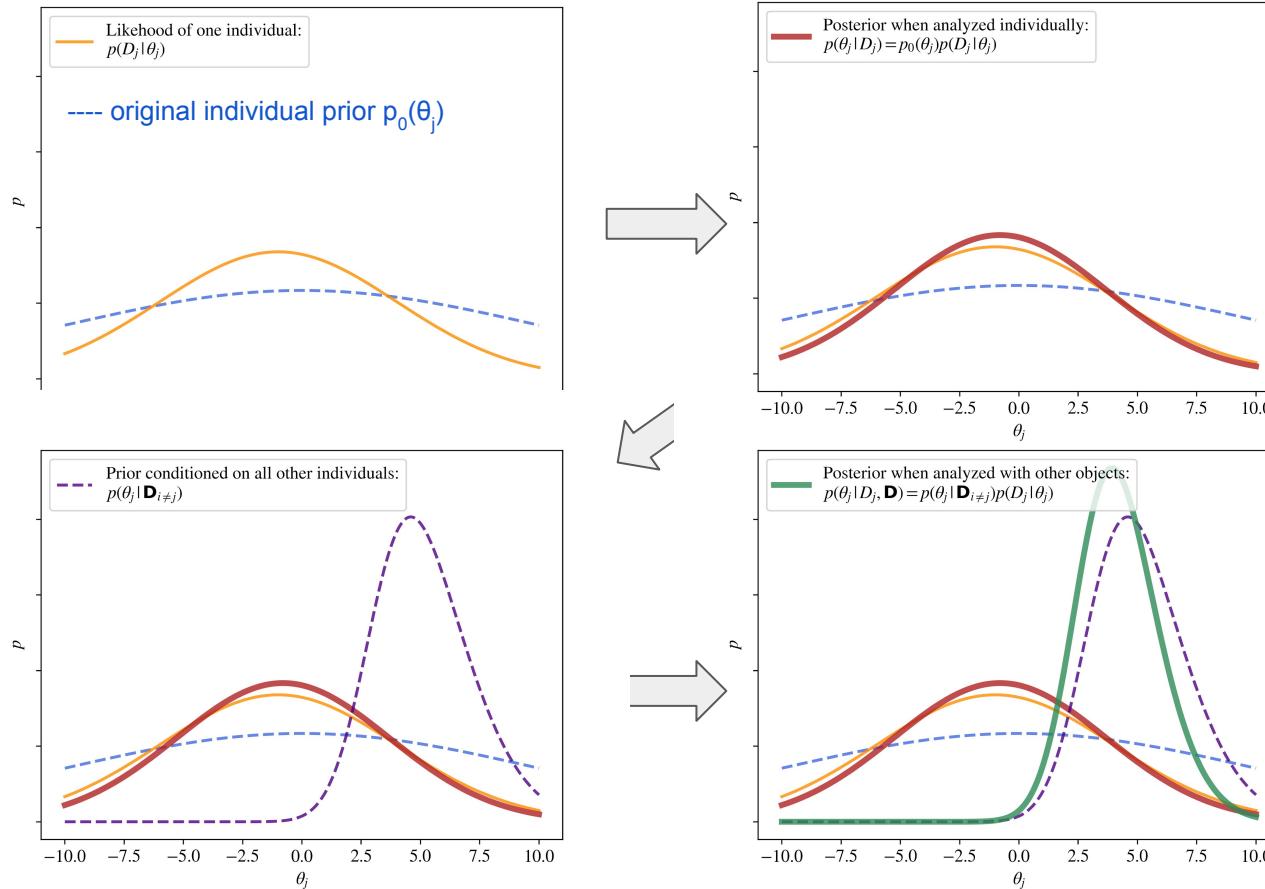
Expressed mathematically as marginalization over all *other* individual level parameters; looks complex but often easy to do in practice with MCMC samples

$$p(\theta_j | D_j, \mathbf{D}_{i \neq j}) \propto \int d\phi p(D_j | \theta_j) p(\theta_j | \phi) p(\phi | \mathbf{D}_{i \neq j})$$

The equation is annotated with curly braces under the integral terms. The first brace groups the entire integral term  $\int d\phi$  and the term  $p(\theta_j | \phi)$ , labeled "posterior". The second brace groups the term  $p(D_j | \theta_j)$ , labeled "likelihood of jth data". The third brace groups the term  $p(\phi | \mathbf{D}_{i \neq j})$ , labeled "posterior for  $\phi$  given all other data".

# HBM: individuals and populations

*shrinkage*





# Break & Questions



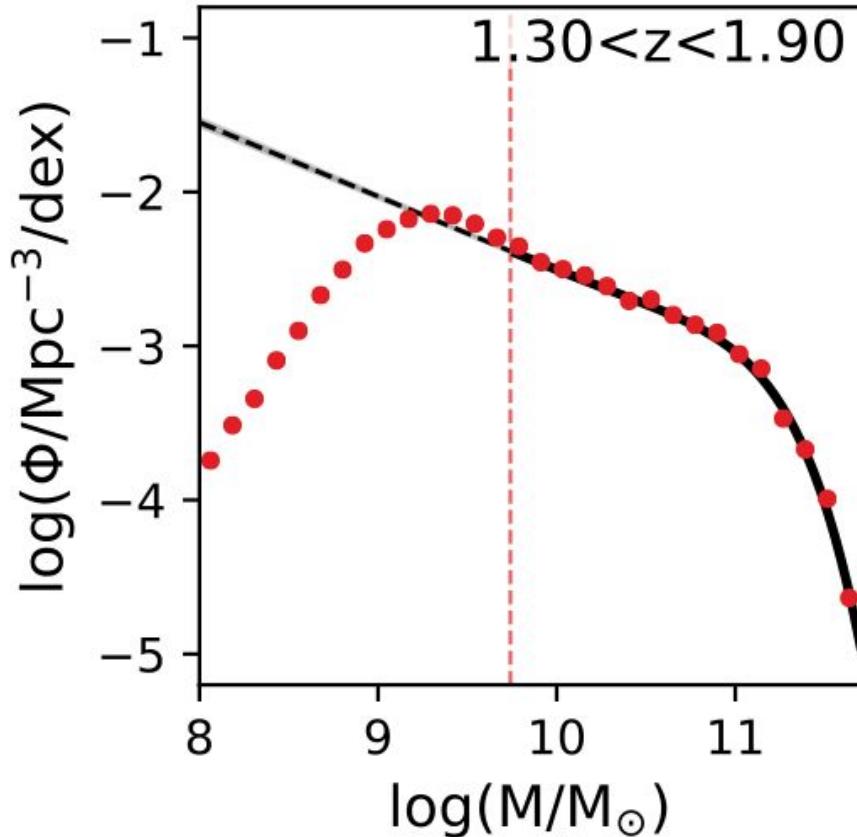
# HBM: Examples

- 1) Galaxy Stellar Mass Function (Leja, Speagle, BDJ et al. 2020)
- 2) Star cluster properties

## HBM: Galaxy stellar mass function

A very common use for HBM is to determine the prior probability of individual properties from the distribution of observed examples.

The galaxy stellar mass function (GSMF) is a key constraint on theories of galaxy evolution, describing the ***distribution*** of galaxy stellar masses.



## HBM: Galaxy stellar mass function

The galaxy stellar mass function (GSMF) is a key constraint on theories of galaxy evolution, describing the ***distribution*** of galaxy stellar masses.

It is often described by a 3-parameter function, the Schechter function:

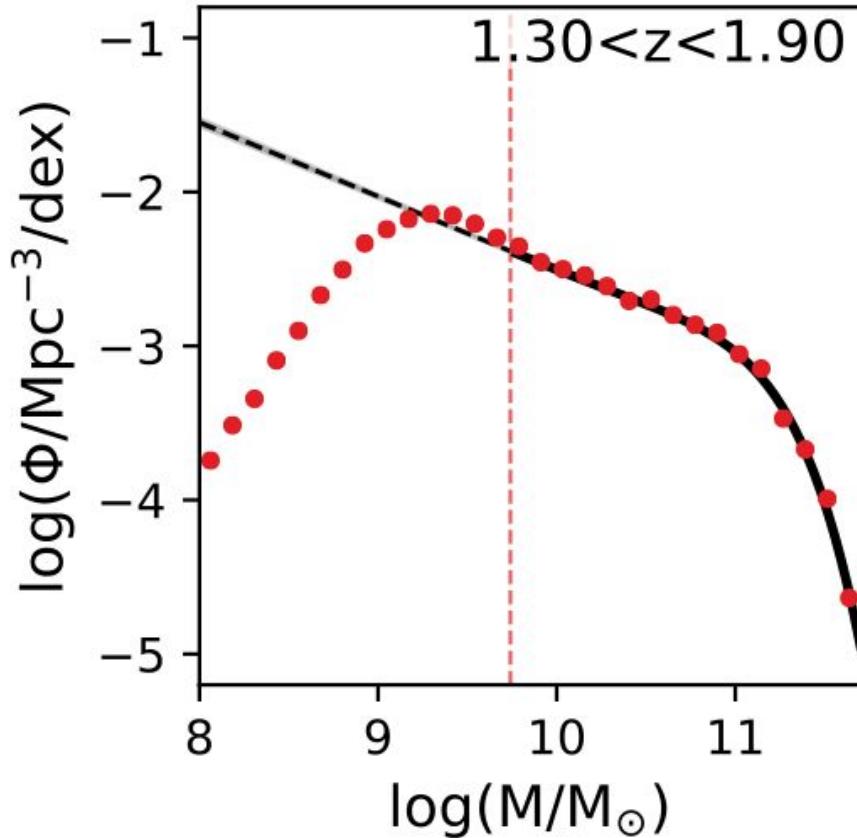
$$\Phi(M) = \rho_* (M/M_*)^\alpha e^{-M}$$

$M_*$  : knee

$\alpha$  : faint end slope

$\rho_*$  : normalization

The parameters of this function may evolve with redshift as galaxies grow and merge  
(see Leja et al. 2020 for details)  
but for simplicity we consider a single redshift.



# HBM: Galaxy stellar mass function

## Observables

Photometry of each galaxy, grism redshifts:

$$D_i = \{\text{magnitudes}\}, \text{redshift}$$



## Individual

For each galaxy:

$$\theta_i = (\text{Mass}, \text{redshift}, \text{metallicity}, A_V, \text{SFH}, \text{etc..})$$

## Population

For all the galaxies:

$$\phi = (\alpha, M_*)$$

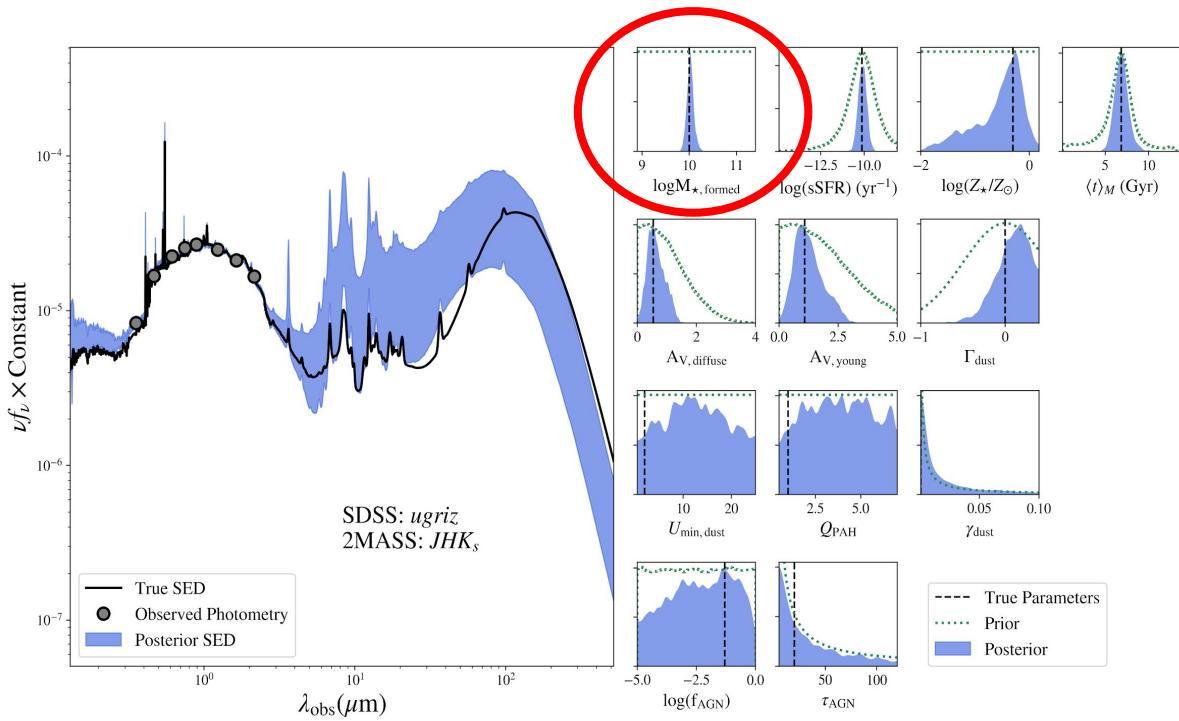
We neglect  $\rho_*$  for the moment

# HBM: Galaxy stellar mass function

We can use stellar population synthesis modeling to infer the properties of individual galaxies from observed magnitudes.

In this case we have a very weakly informative prior.

$$p(\log M_i) = \text{constant}$$



Johnson et al 2020

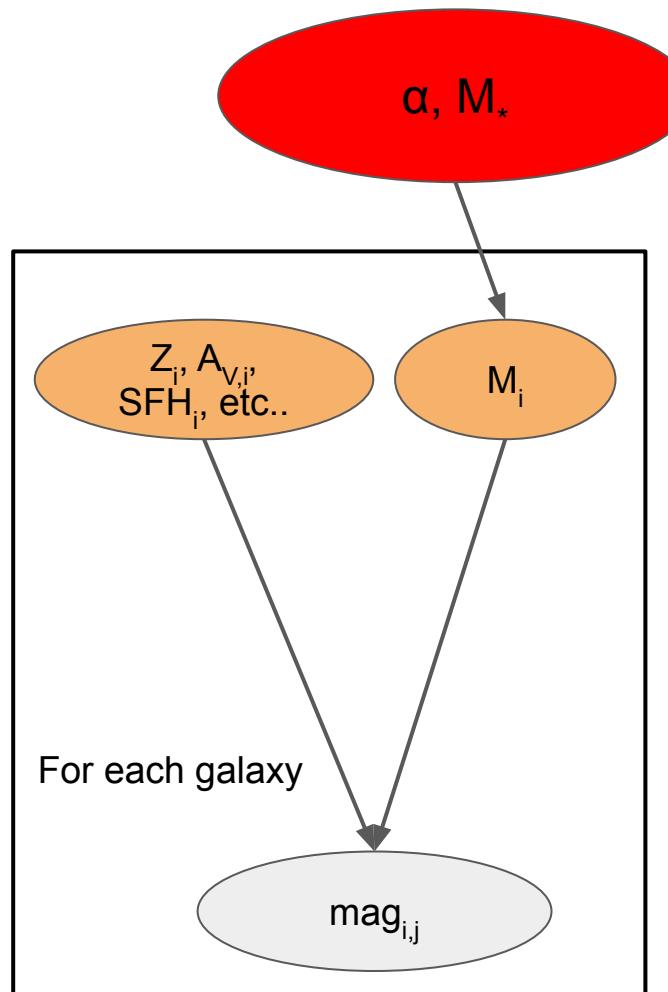
## HBM: Galaxy stellar mass function

But we can combine all the galaxies to obtain a constraint on the GSMF parameters, which forms a more *realistic* prior.

$$\begin{aligned} p(M_i | \phi) &= p(M_i | \alpha, M_*) \\ &= \frac{(M/M_*)^\alpha e^{-M/M_*}}{\Gamma(\alpha + 1, M_c/M_*)} \end{aligned}$$

This is just the Schechter function normalized to 1 by the term in the denominator.  $M_c$  is the completeness limit.

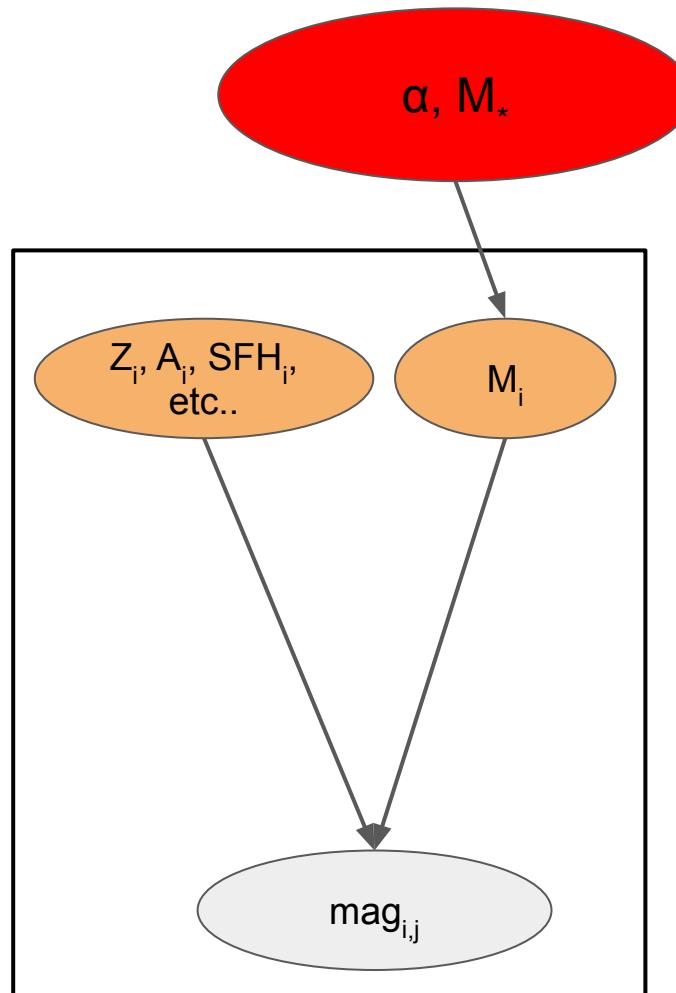
NOTE: If you are interested in the normalization parameter ( $\rho_*$ ) then you need another term in the probability related to the expected total number of objects: see Leja+20 or Foreman-Mackey+14 for details



## HBM: Galaxy stellar mass function

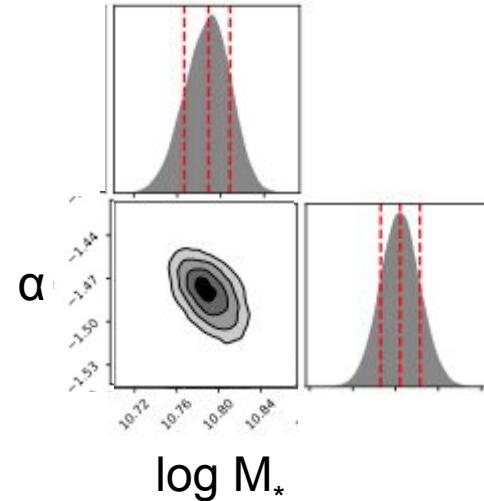
But we can combine all the galaxies to obtain a constraint on the GSMF parameters, which forms a more accurate prior.

$$p(\{M_i\}, \alpha, M_* | \mathbf{D}) \propto p(\alpha, M_*) \prod_{i=1}^N \frac{(M_i/M_*)^\alpha e^{-M_i/M_*}}{\Gamma(\alpha + 1, M_c/M_*)} p(D_i | M_i)$$



## HBM: Galaxy stellar mass function

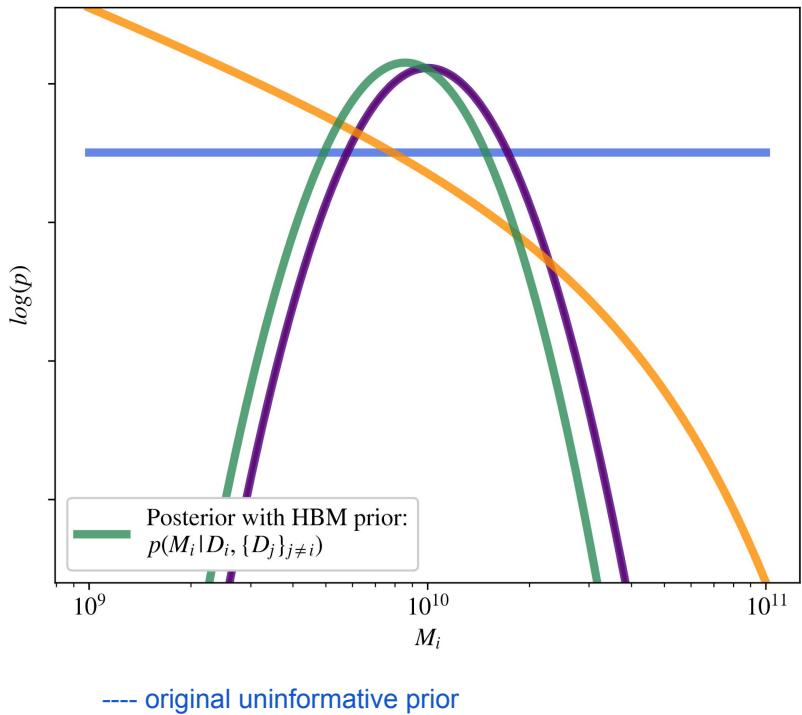
**But** we can combine all the galaxies to obtain a constraint on the GSMF parameters, which forms a more accurate prior.



## HBM: Galaxy stellar mass function

This changes (improves) the prior and posterior for the mass of individual galaxies - **shrinkage**

Though in this case the posterior does not get much narrower, it based on a more realistic prior.



# HBM: Galaxy stellar mass function

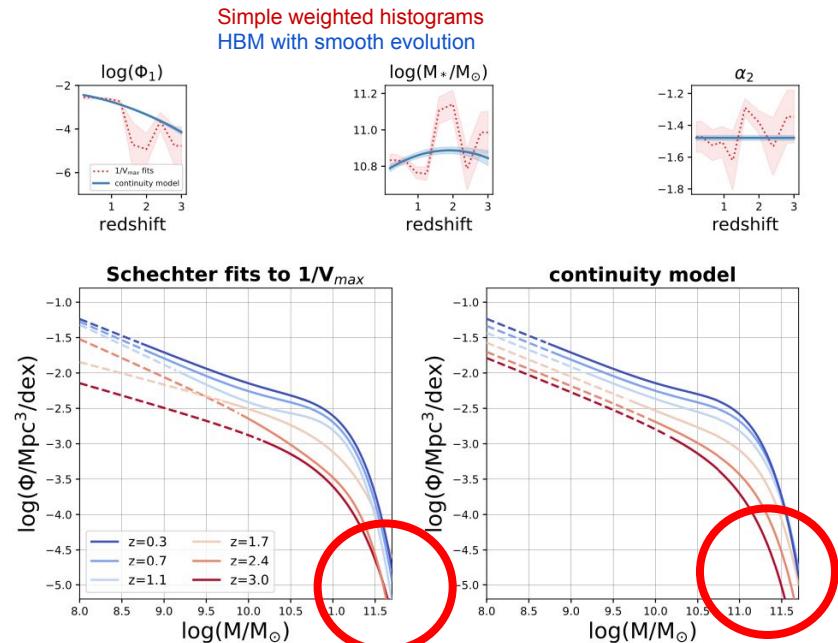
Other benefits:

- 1) When considering many redshifts, you can impose a smoothness constraint on the parameter evolution. E.g.

$$M_*(z) = M_{*,0} + M_{*,1} \times z$$

Then the GSMF at any redshift incorporates information from galaxies at all redshifts, and avoids binning effects.

- 2) Accounts for measurement uncertainties and uncertainties on individual galaxy masses



Leja, Speagle, BDJ et al., 2020

## HBM: Galaxy stellar mass function

Implementation Note:

This now involves sampling for  $2 + N_{\text{galaxies}} \times (N_\psi + 1)$  parameters!!!

In practice we performed inference to obtain samples of individual posteriors based on the initial uninformative prior  $p_0(M_i)$ , and then **re-weighted** each sample  $k$  to get the HBM weighted likelihood.

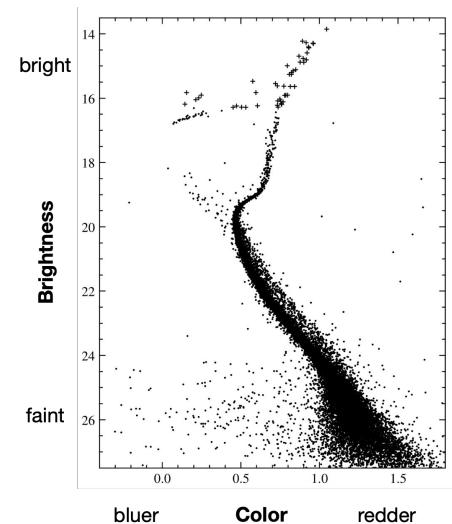
This requires having ‘enough’ samples near the bulk of the new likelihood (and knowing the original prior distribution)

$$w_{i,k} = \frac{p(M_{i,k} | \alpha, M_*)}{p_0(M_{i,k})}$$

## HBM: Star clusters

Star clusters are gravitationally bound groups of stars, **co-located** and assumed to be **coeval** and born with the **same metallicity** and elemental abundances.

However, each star has a different initial mass -- distributed according to the **IMF** -- and is therefore in a different stage of stellar evolution at the present time, and has different physical parameters (luminosity, effective temperature, radius) and therefore different values of the observables.



## HBM: Star clusters

This naturally lends itself to a hierarchical model where the individuals are the stars

with **observables**

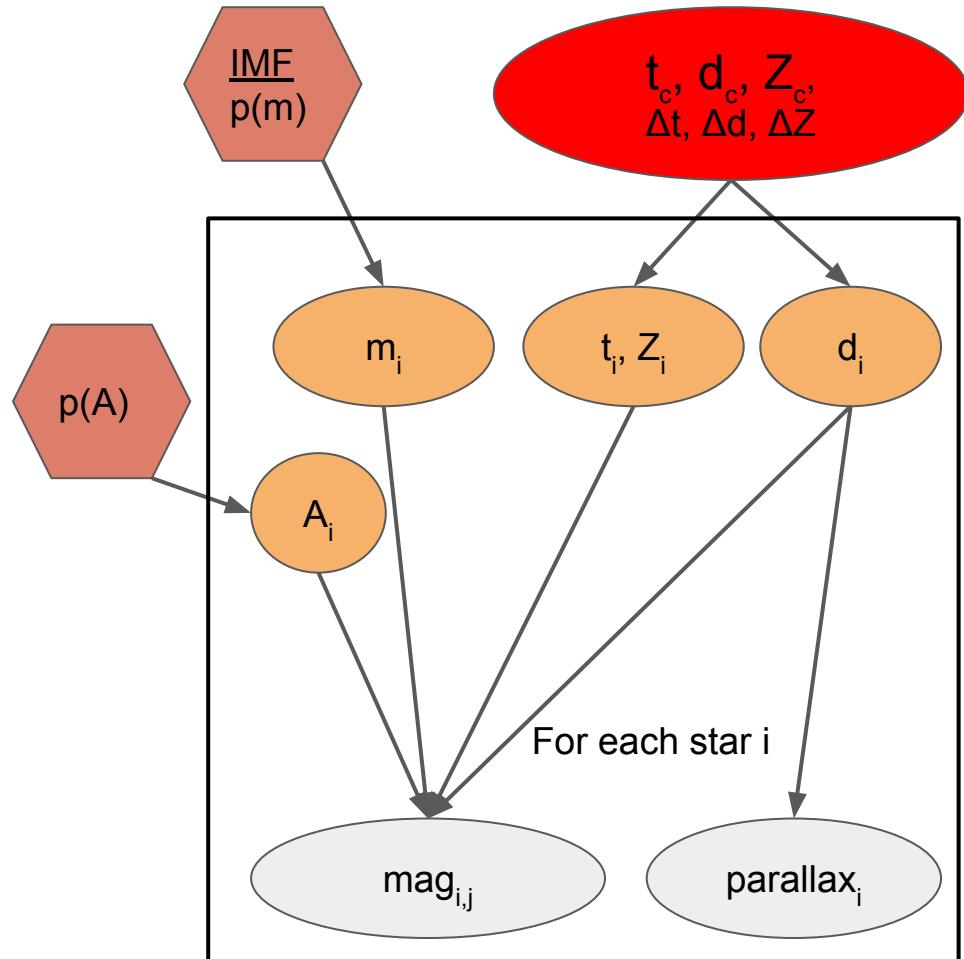
$D_i$  = magnitude, colors [, parallax]

and fundamental **individual parameters**

$\theta_i$  = (distance, age, metallicity, mass,  $A_V$ )

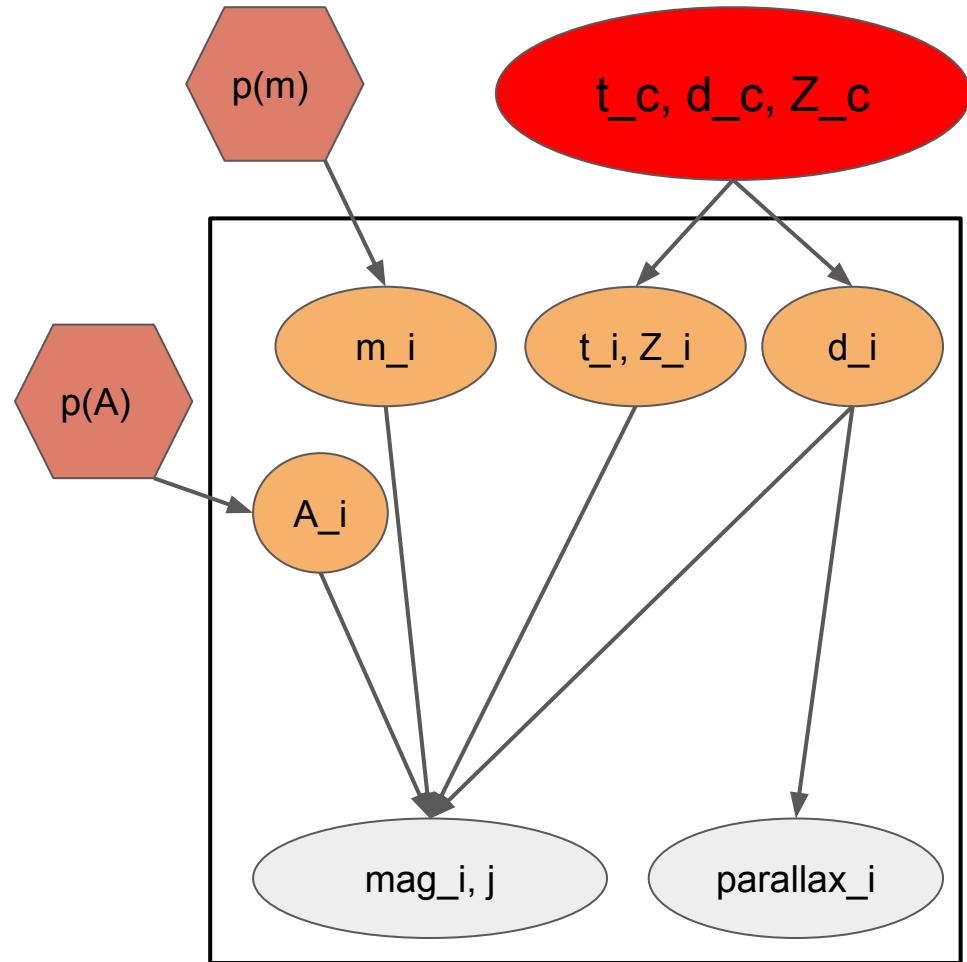
The **population parameters** are the cluster parameters:

$\phi$  = (cluster age, cluster distance, cluster metallicity, widths of each distribution)



## HBM: Star clusters

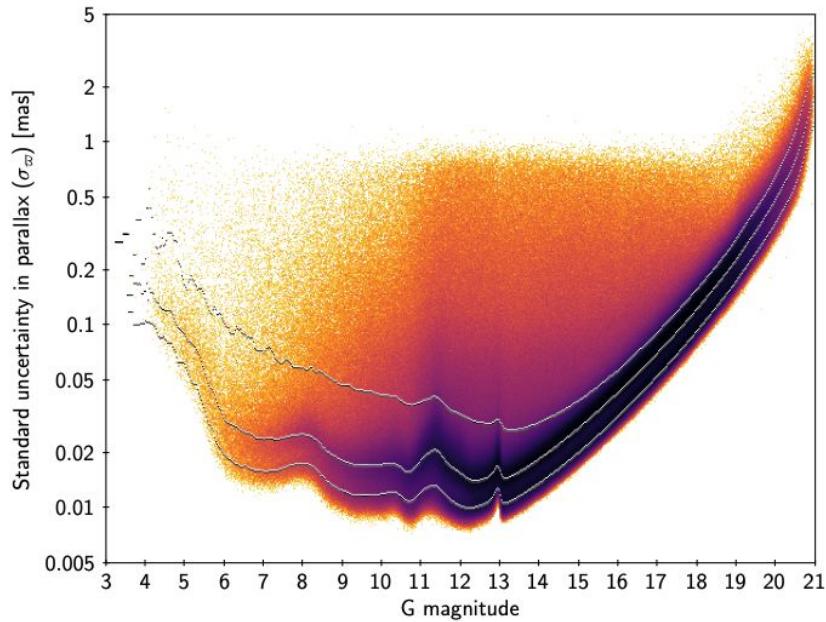
Equation here



## HBM: Star clusters

Pooling information:

The brightest stars will have the best parallax and be most informative about ***distance***



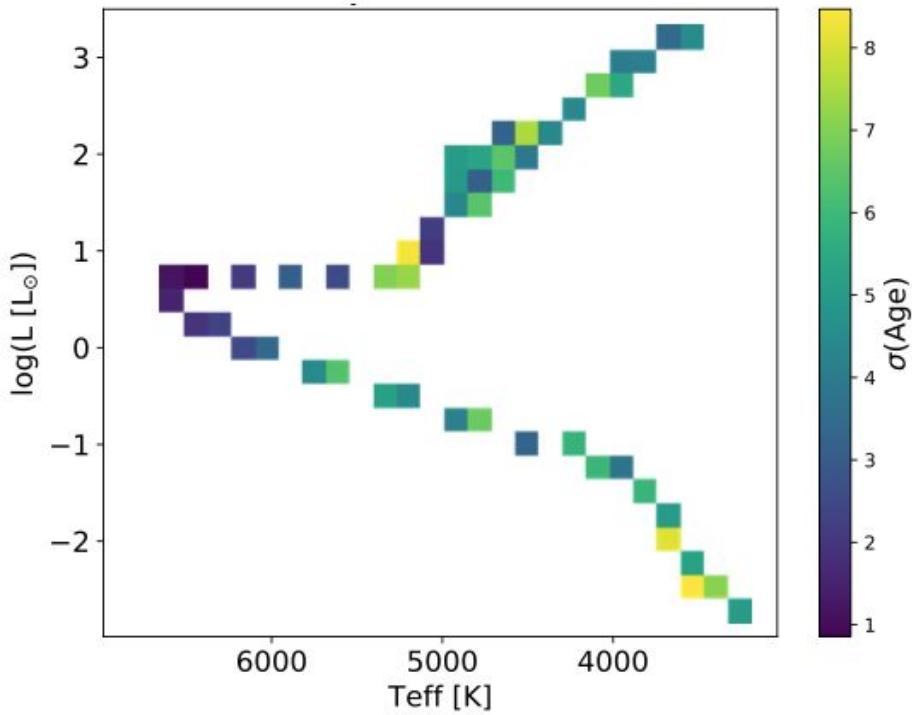
Lindgren et al. 2020

## HBM: Star clusters

Pooling information:

The brightest stars will have the best parallax and be most informative about ***distance***

The bluest MS stars are very informative about ***age*** (given a distance)



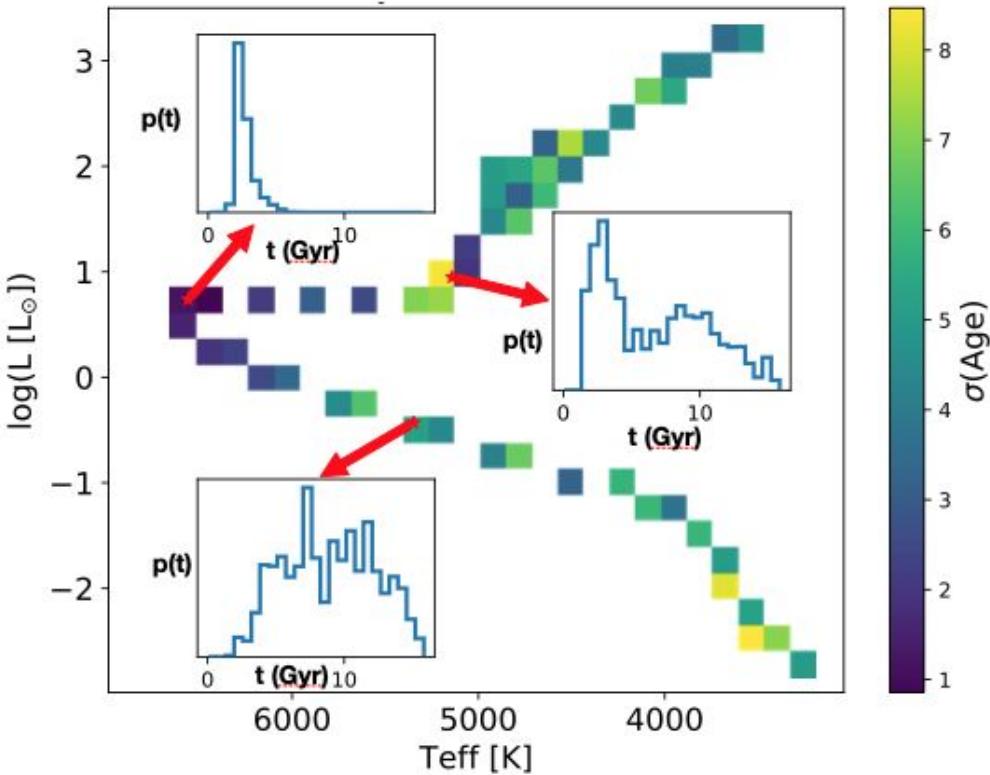
see e.g Cargile, Conroy, BDJ et al. 2020

## HBM: Star clusters

Pooling information:

The brightest stars will have the best parallax and be most informative about **distance**

The bluest MS stars are very informative about **age** (given a distance)



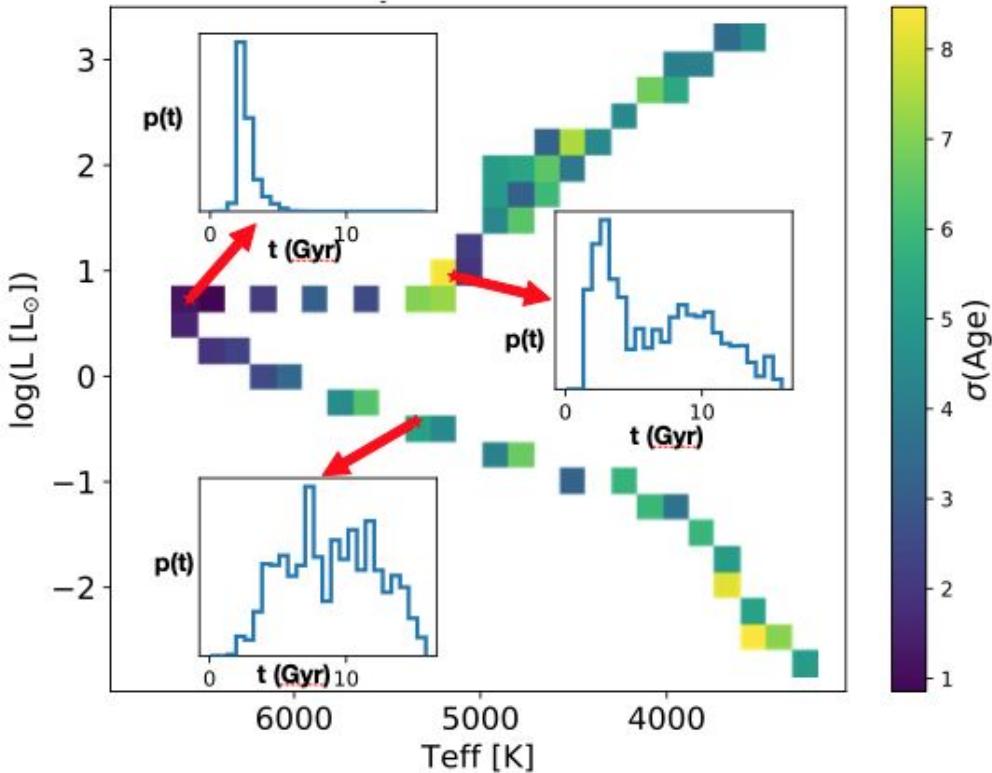
## HBM: Star clusters

Pooling information:

The brightest stars will have the best parallax and be most informative about **distance**

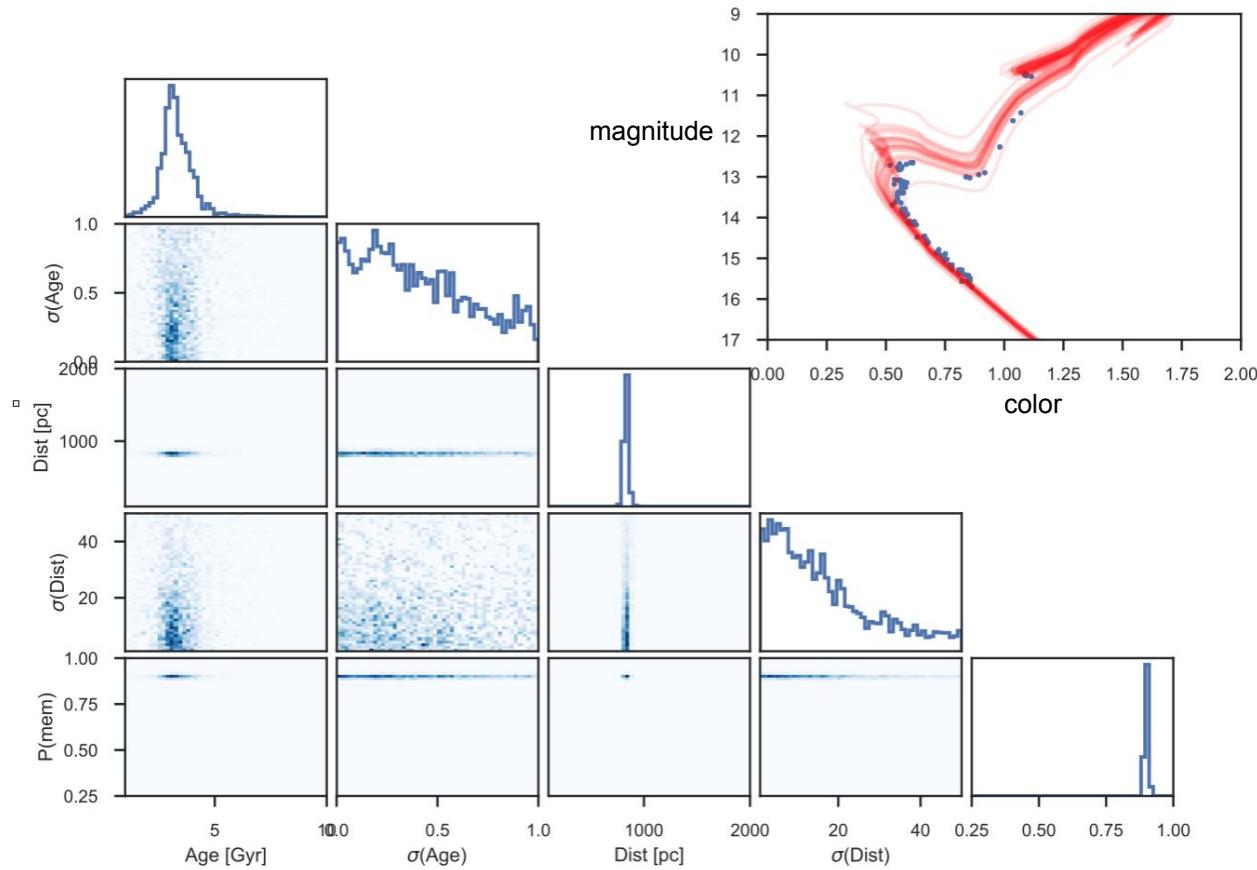
The bluest MS stars are very informative about **age** (given a distance)

All stars are informative about **metallicity** (given age and distance)



# HBM: Star clusters

Results (posterior density) for M67  
using Gaia DR1 data



## HBM: Star clusters

PLOT HERE showing shrinkage of mass or  
FeH estimates



## HBM: Practical advice

Sampling can be challenging, since there are many parameters (population parameters plus all individual parameters.)

Clever strategies often required.

Pseudo-importance sampling (e.g. Hogg et al. 2010, Foreman-Mackey et al. 2014)

- re-weight existing samples based on new priors, requires samples to exist where necessary and knowledge of priors

Gibbs sampling (e.g. Mandel 2011, Gelman et al.)

- MCMC sample parameters of individuals conditional on population parameters, then sample population parameters conditional on individuals, repeat ad-infinitum
- Exploits conditional independence of the graphs

Hamiltonian Monte Carlo (R. Neal 2012 <https://arxiv.org/abs/1206.1901> , M. Betancourt 2017)

- use gradients to enable sampling in large numbers of parameters (individual & population) simultaneously

## HBM: Practical advice

- Carefully consider dependencies in your model.
- Beware selection effects; these can be included in the model but may be complex.
- Usually good to try on a mock population first!

General Resources:

Bayesian Data Analysis (textbook)  
Gelman, et al.  
(advanced & detailed)

Krushke & Vanpaemel (2015)  
<https://jkkweb.sitehost.iu.edu/articles/KruschkeVanpaemel2015.pdf>

Loredo 2013  
<https://arxiv.org/abs/1208.3036>

Loredo & Hendry 2019 (excellent!)  
<https://arxiv.org/abs/1911.12337>

Some notable examples in Astronomy Literature:

von Hippel, T. et al. 2006 (star clusters)  
<https://arxiv.org/abs/astro-ph/0603493>

van Dyk et al 2009 (star clusters)  
<https://arxiv.org/abs/0905.2547>

Andreon & Hurn 2010  
(galaxy cluster scaling relations)  
<https://arxiv.org/abs/1001.4639>

Foreman-Mackey et al. 2014 (exoplanets)  
<https://arxiv.org/abs/1406.3020>

Wolfgang et al 2016 (exoplanet scaling)  
<https://arxiv.org/abs/1504.07557>

Mandel et al 2011 (supernovae)  
<https://arxiv.org/abs/1011.5910>