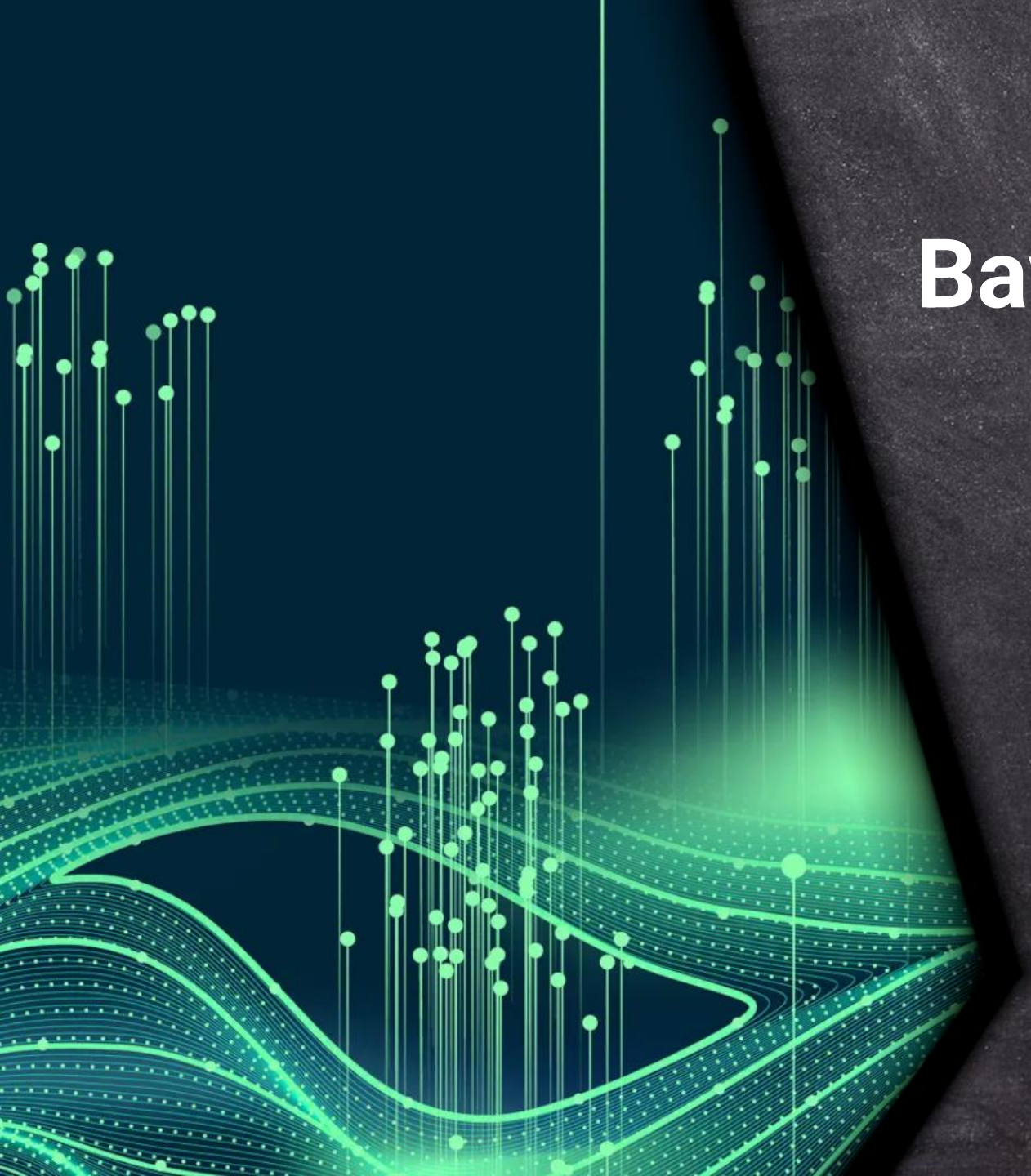


General review on Bayesian statistics and its applications

PhD Student
Stefano Menchiari

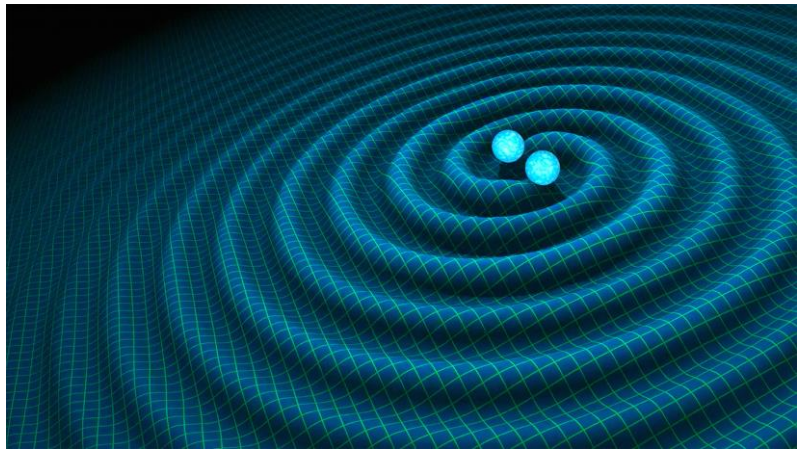
UNIVERSITÀ DI SIENA 1240

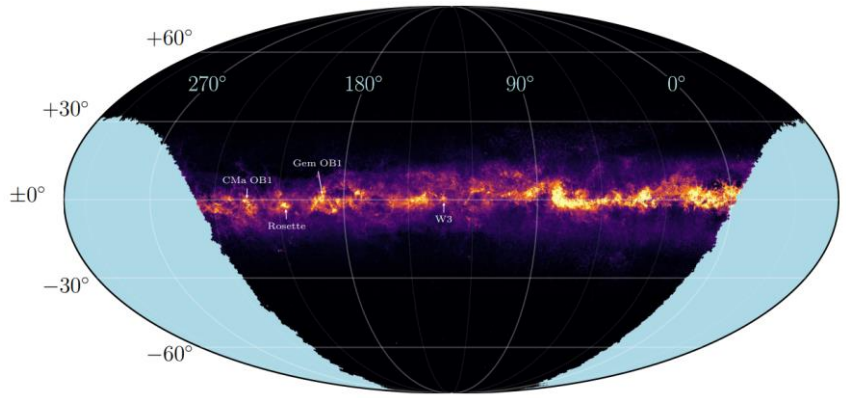
27/04/22

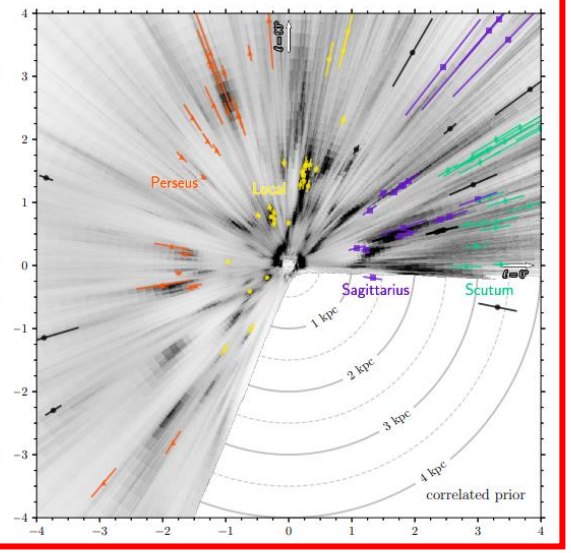
Outline

- The basic of Bayes Statistics
  - The Bayes theorem
  - Interpretation of Bayesian probability
- Understanding Bayes statistic with an example: Gravitational waves (GWs)
  - Prior distribution
  - Likelihood
  - Evidence
  - Posterior probability
- Additional example: Mapping the Milky Way
  - Test with mock stars
  - Dust maps
  - Cumulative distribution
- Final remarks
- Backup slides

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$





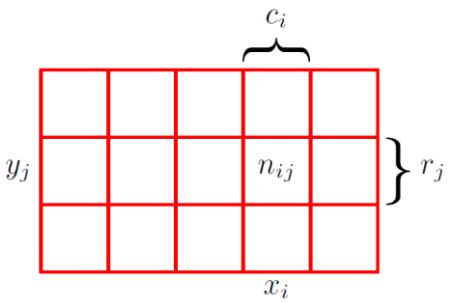


The Bayes theorem

The heart of Bayes statistics is the **Bayes theorem**.
The rules of Bayesian probability theory can be derived from just two basic rules (Cox, 1946)

Suppose to have two random variables **X** and **Y**, and consider a total of N trials in which we sample both X and Y.
Let the number of such trials in which $X=x_i$ and $Y=y_j$ be n_{ij}.
Also, let the number of trials in which $X=x_i$ (irrespective of the value that Y takes) be denoted by c_i

$$X = x_1, \ldots, x_i, \ldots, x_M$$
$$Y = y_1, \ldots, y_j, \ldots, y_L$$



$$P(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$ Joint probability

$$P(X = x_i) = \frac{c_i}{N} = \frac{\sum_j n_{ij}}{N}$$ Marginal probability

$$P(Y = y_j | X = x_i) = \frac{n_{ij}}{c_i}$$ Conditional probability

SUM RULE

$$P(X = x_i) = \sum_{j=1}^{L} P(X = x_i, Y = y_j)$$

PRODUCT RULE

$$P(X = x_i, Y = y_j) = P(Y = y_j | X = x_i) P(X = x_i)$$

Using the product rule, considering that P(X,Y)=P(Y,X), and applying in the last passage the sum rule

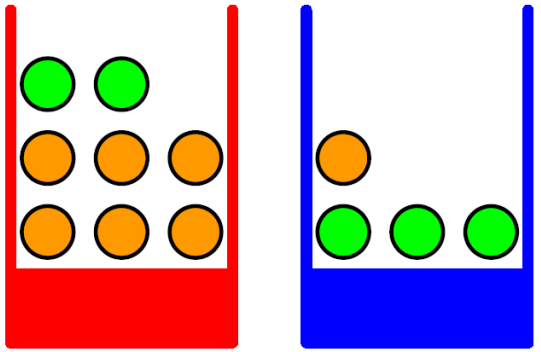
BAYES THEOREM

$$P(Y|X) = \frac{P(X|Y)\,P(Y)}{P(X)} = \frac{P(X|Y)\,P(Y)}{\sum_Y (X|Y)\,P(Y)}$$

Interpretation of Bayesian probability

Let us now consider the case where one has two boxes B (Y=B), one red (r) and one blue (b). Both boxes are filled with fruit (X=F), specifically, with oranges (o) and apples (a).

$$
\begin{aligned}
p(B=r) &= 4/10 \\
p(B=b) &= 6/10 \\
p(F=a|B=r) &= 1/4 \\
p(F=o|B=r) &= 3/4 \\
p(F=a|B=b) &= 3/4 \\
p(F=o|B=b) &= 1/4.
\end{aligned}
$$



Let's take into account the case where we are told that a piece of fruit has been selected and it is an orange, and we would like to know which box it came from.

We can answer the problem using Bayes theorem

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)}$$

Knowing that: $p(F=o) = \sum_B p(F=o|B)p(B)$

$$p(B=r|F=o) = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}$$

We can provide an important interpretation of Bayes' theorem as follows.

If we had been asked which box had been chosen before being told the identity of the selected item of fruit, then the most complete information we have available is p(B).
We call this the *prior probability* because it is the probability available before we observe the identity of the fruit.

Once we are told that the fruit is an orange, we can then use Bayes' theorem to compute the probability p(B|F), which we shall call the *posterior probability* because it is the probability obtained after we have observed F.

Nota bene: in this example, the prior probability of selecting the red box was 4/10, so that we were more likely to select the blue box than the red one. However, once we have observed that the piece of selected fruit is an orange, we find that the posterior probability of the red box is 2/3, so now it is more likely that the box we selected was, in fact, the red one.

Bayes' theorem is used to convert a prior probability into a posterior probability by incorporating the evidence provided by the observed data.

Bayesian framework through GWs

Let us now take a closer look to all the members appearing in the Bayes theorem

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\pi(\theta)}{\mathcal{Z}}$$

$\pi(\theta)$: **Prior distribution**

$\mathcal{L}(d|\theta)$: **Likelihood**

$\mathcal{Z}$: **Evidence**

$p(\theta|d)$: **Posterior distribution**

We will now examine one by one all the pieces of the Bayes theorem, and we will relay on the physics case of gravitational wave as a direct example for a better understanding

*From the investigation of a single signal to population studies (**see backup slides**) the analysis procedure is completely carried using the Bayes framework*

Gravitational waves in brief

The motion of two compact objects about to merge creates a distinct perturbation of the space-time metrics which propagates in space as a wave

$$g_{\mu\nu} = \eta_{\mu\nu} + h_{\mu\nu}$$

Metric tensor — Minkowsky metric — Perturbation

where: $|h_{\mu\nu}| \ll 1$

With some lengthy calculations, one can demonstrate that:

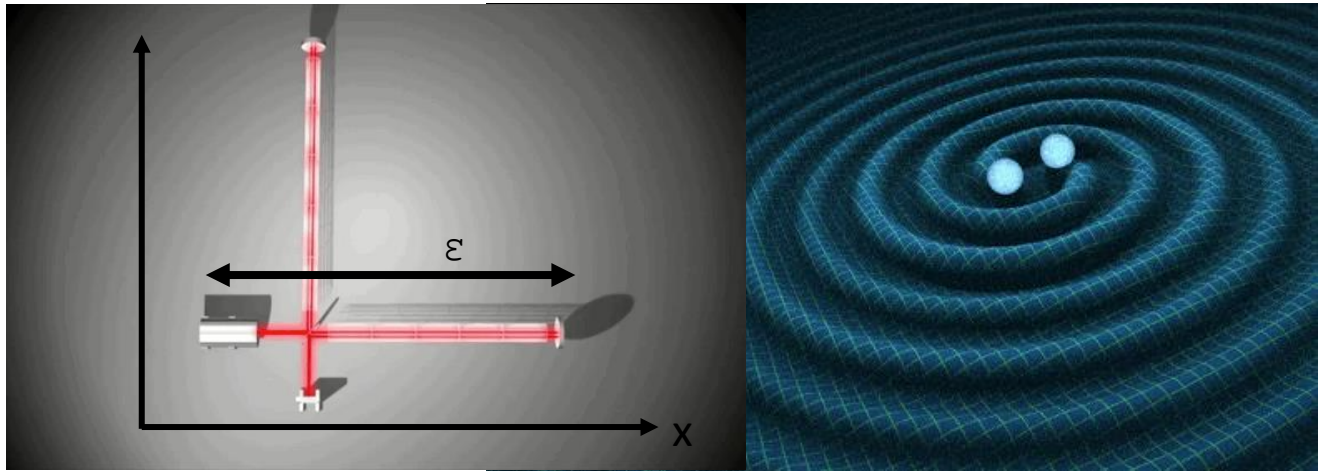
$$\Box\bar{h}_{\mu\nu} = \left(\partial^{\lambda}\partial_{\lambda}\right)\bar{h}_{\mu\nu} = \left(-\partial_t^2 + \nabla^2\right)\bar{h}_{\mu\nu} = 0 \Rightarrow \bar{h}_{\mu\nu} = A_{\mu\nu}e^{ik_{\lambda}x^{\lambda}}$$

where $\bar{h}$ is the trace reversed tensor.

If we consider two particles along the x axis, separated by a distance ε, the passage of the wave induces a change in the distance:

$$\Delta l = \int_0^n |g_{\mu\nu}dx^{\mu}dx^{\nu}|^{1/2} \simeq \varepsilon\left[1 + \frac{1}{2}\bar{h}_{xx}(x=0)\right]$$

(Credit Ligo)



Prior distribution

Priors express our present state of knowledge about the parameters of interest, which we wish to constrain by analyzing new data

Uninformative priors
They express our state of ignorance and have very little restricting power. Typically, their distributions are diffuse.

Informative priors
Characterized by very restricting distributions. They might come from the analysis of some previous data

There are some general recipes to keep in mind when constructing uninformative priors:

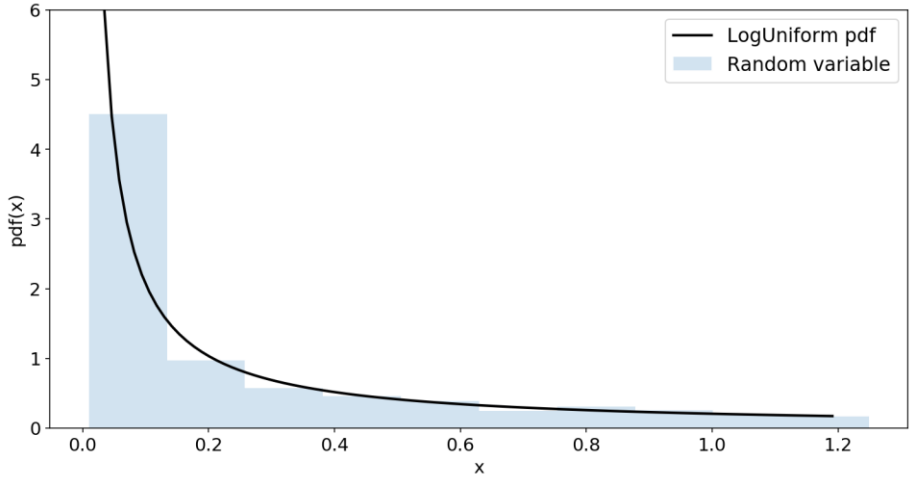
- *Laplace's principle of insufficient reason:* assigning equal probability to all possible values of the parameters
- *Invariance by transformation*: If the priors are uninformative, then, we should make the same Bayesian inference under a given transformation, which implies that the priors should also be invariant to the transformation
- Jeffreys rule: $p(\theta) \propto \det[\mathcal{I}(\theta)]^{1/2}$, where: $[\mathcal{I}(\theta)]_{ij} = \int \mathcal{L}(x|\theta) \frac{\partial^2}{\partial\theta_i \partial\theta_j} \ln\mathcal{L}(x|\theta)\, dx$
- Maximum entropy: given some constrains (i.e. $\sum p_i = 1$) on the prior, the prior should be chosen to be the distribution with the largest entropy S which follows these constraints

$$S = -\sum_i p_i \log(p_i) \Rightarrow$$

Example: without constrains the uniform distribution has the maximum S

Priors for Gravitational waves

Example: when considering the sky location of a BH merger, it is reasonable to choose an isotropic prior that weights each patch of sky as equally probable. One possible choice can be the uniform or log-uniform distributions



Warning about priors

The choice about is also leaded to the specific science case and must reflect the physics behind a given model.
Sometimes priors are chosen following only criteria of mathematical convenience.
This can induce sever biases

Likelihood

The likelihood function is something that we choose.

It is a description of the measurement.

By writing down a likelihood, we implicitly introduce a noise model

The likelihood depends on all the model parameters.

Sometimes one could be interested in studying just few parameters.

In that case is useful to define the *marginalized* (integrated) likelihood over the parameters we are not interested in (called nuisance parameters)

Likelihood for Gravitational waves

For gravitational-wave astronomy, is typically assumed a Gaussian-noise likelihood:

$$\mathcal{L}(d|\theta) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{1}{2}\frac{|d-\mu(\theta)|^2}{\sigma^2}\right)$$

d: measured GW strain
θ: model parameters
σ: detector noise

$$\mu(\theta) = F_+(\mathrm{RA},\mathrm{DEC},\psi)h_+(\theta) + F_\times(\mathrm{RA},\mathrm{DEC},\psi)h_\times(\theta)$$

Marginalized likelihood for Gravitational waves

For the sake of simplicity, we can consider the case of a binary BH merging.

The strain signal can hence be described by a **θ=15** parameters model:

Intrinsic parameters (8):
Primary mass (m_1), secondary mass (m_2), primary spin vector (s_1), secondary spin vector (s_2)
Extrinsic parameters (7):
Inclination angle ( ), polarization angle (ψ), phase at coalescence (ϕ_c), right ascension (RA), declination (DEC), luminosity distance (D_L), time of coalescence (t)

- Distance marginalization
- Phase marginalization
- Time marginalization

Evidence

The evidence act as a normalization constant for the prior.

Calculating the evidence can be computationally challenging.

$$\mathcal{Z} \equiv \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta)$$

Model selection applied to GW

The most straightforward example is the comparison between a "signal model" (Z_N) and a "noise only model" (Z_S)

$$\mathcal{Z}_S \equiv \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta) \qquad \mathcal{Z}_N \equiv \mathcal{L}(d|0)$$

One additional example is to compare models based on different theories of general relativity

It is also possible to compare the same model but with different priors: BH merging with and without spin. The Bayes factor comparing these models would tell us if the data prefer spin.

$$\mathcal{Z}_{ns} = \int d\theta \mathcal{L}\,(d|\theta)\,\pi_{ns}\,(\theta) \qquad \mathcal{Z}_{\rm spin} = \int d\theta \mathcal{L}(d|\theta)\,\pi(\theta)$$

The evidence becomes crucial when comparing different models.

Assume to have two different models M_A and M_B, with a different number of parameters θ and ν. Model selection answers the question: which model is statistically preferred by the data and by how much?

It is possible to compare the two models by comparing their evidence:

$$\mathcal{Z}_A = \int d\theta \mathcal{L}(d|\theta, M_A)\,\pi(\theta)$$
$$\mathcal{Z}_B = \int d\nu \mathcal{L}(d|\nu, M_B)\,\pi(\nu)$$

$$\mathrm{BF}^A_B = \frac{\mathcal{Z}_A}{\mathcal{Z}_B} \qquad \log BF^A_B = \log \mathcal{Z}_A - \log \mathcal{Z}_B$$

Bayes Factor

If |logBF|>>1 then one model is preferred over the other. The sign of log BF tells us which model is preferred. A threshold of |logBF|=8 is often used as the level of "strong evidence" in favor of one hypothesis over another (Jeffreys, 1961)

Formally, the correct metric to compare two models is not the Bayes factor, but rather the odds

$$\mathcal{O}^A_B \equiv \frac{\mathcal{Z}_A}{\mathcal{Z}_B}\frac{\pi_A}{\pi_B}$$

Occam Factor

Bayesian evidence encodes two pieces of information: (1) how model fits data (likelihood), (2) the act of marginalization takes into account the size of the parameter space volume.

A model with a decent fit and a small prior volume often yields a greater evidence than a model with an excellent fit and a huge prior volume.

Bayes factor penalizes the more complicated model for being too complicated.

Posterior probability

The posterior distribution $p(\theta|d)$ is the probability density function for the continuous variable θ given the data d.

$$p(\theta|d) = \frac{\mathcal{L}(d|\theta)\,\pi(\theta)}{\mathcal{Z}}.$$

Calculating the posterior is a classical inverse problem, that are renown to be computationally challenging.

The curse of dimensionality, example: GWs

To calculate the posterior probability for the 15 parameters BH merger case we could naively think to create a grid with 10 bins in every dimension and evaluate the likelihood at each grid point. Even with this coarse resolution, our calculation It is computationally prohibitive to carry out, as it requires 10^{15} likelihood evaluations.
In double precision it means to calculate a ≈8000TB tensor

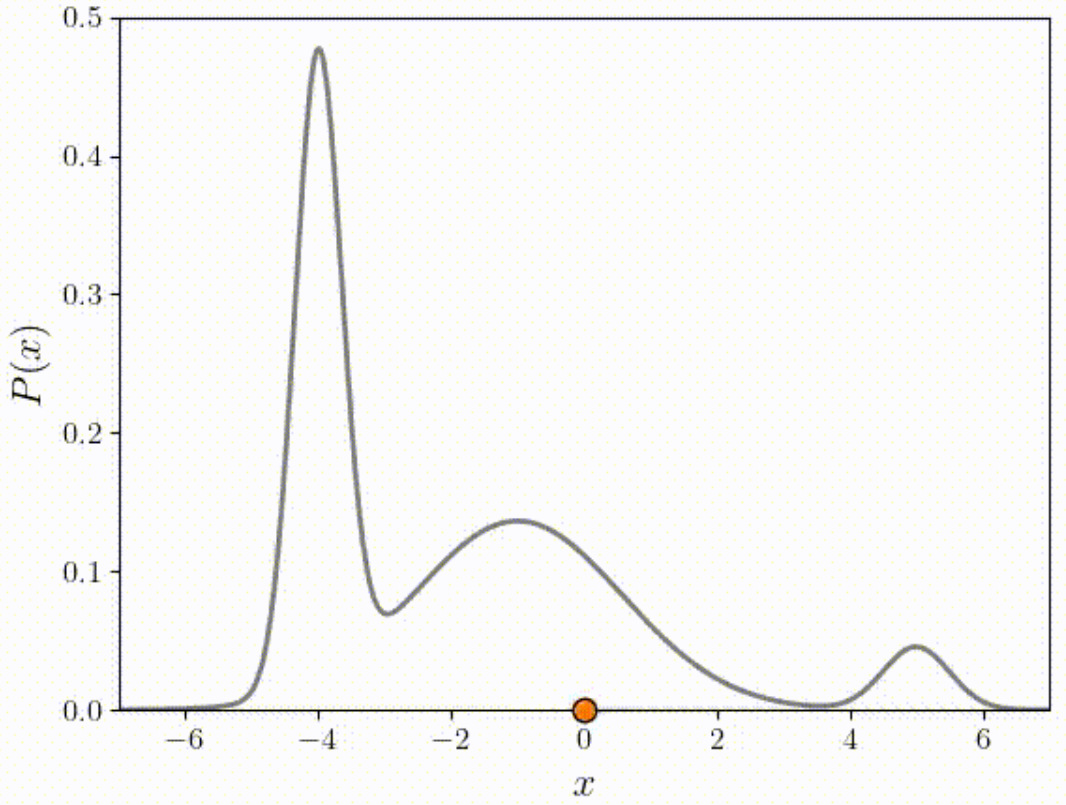
Sampling the posterior distribution

The posterior distribution can be obtained using stochastic samples. One of the most common approach is to use Markov Chain Monte Carlo (MCMC).

Example, _____ (MHA):

1. Initialize a random guess for $\theta=\theta_{=0}$
2. For t in n_steps:
   a) Generate $\theta'=\theta_{1}$+RandomNumber
   b) Calculate acceptance ratio R: $R = \min\left\{1; \frac{\mathcal{L}(d|\theta')\,\pi(\theta')}{\mathcal{L}(d|\theta_{t-1})\,\pi(\theta_{t-1})}\right\}$
   c) Set $\theta = \theta'$ if R<Uniform[0,1)

- Sometimes there could be a "burn in phase": sampling of the distribution begins after m_BurnIn_steps to avoid recording the initial less accurate samples
- One can think to sample a value every *k* steps ("lag")
- It is possible to build lots of variations, for example: Random walk Metropolis–Hastings (RWMH) where θ' does not depend on θ_{t-1}



Mapping the Milky Way

Using a Bayesian approach, it is possible to recreate a map of the distribution of dust in the Milky Way.

The method is based on the reddening induced by dust on the starlight

The reddening profile (E, defined as color excess in some pair of passbands, i.e. E[B-V]) at a given distance (μ) depends on a set of parameters ($\boldsymbol{\alpha}$, for example the dust density). Knowing the photometry $\mathbf{m_i}$ of a certain star, we can write:

$$p(\boldsymbol{\alpha}\,|\{\boldsymbol{m}\}) = \frac{p(\{\boldsymbol{m}\}|\,\boldsymbol{\alpha})p(\boldsymbol{\alpha})}{p(\{\boldsymbol{m}\})} \Rightarrow p(\boldsymbol{\alpha}\,|\{\boldsymbol{m}\}) \propto p(\boldsymbol{\alpha}) \prod_i p(\boldsymbol{m}_i|\,\boldsymbol{\alpha})$$

We wish to determine the reddening profile E, so we can rewrite the posterior as:

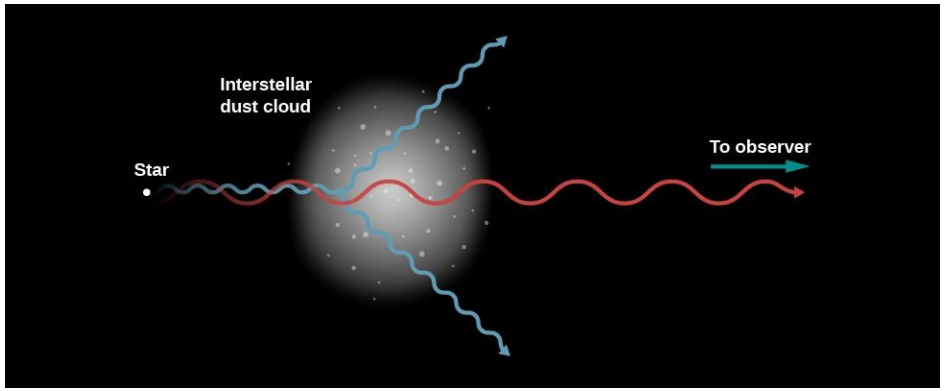
$$p(\boldsymbol{\alpha}\,|\{\boldsymbol{m}\}) \propto p(\boldsymbol{\alpha}) \prod_i \int d\mu_i \; p(\mu_i, E(\mu_i;\boldsymbol{\alpha})|\,\boldsymbol{m}_i)$$

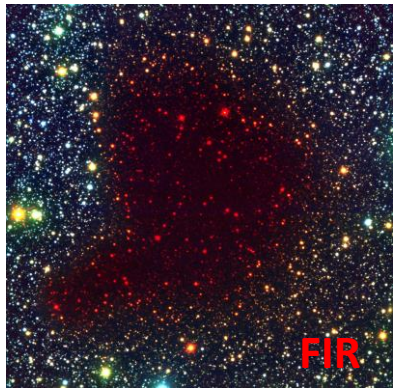
$$p(\mu_i, E_i|\boldsymbol{m}_i) \equiv \frac{1}{Z_i} \int d\boldsymbol{\Theta}_i \; p(\boldsymbol{m}_i|\mu_i, \boldsymbol{\Theta}_i, E_i) p(\mu_i, \boldsymbol{\Theta}_i)$$

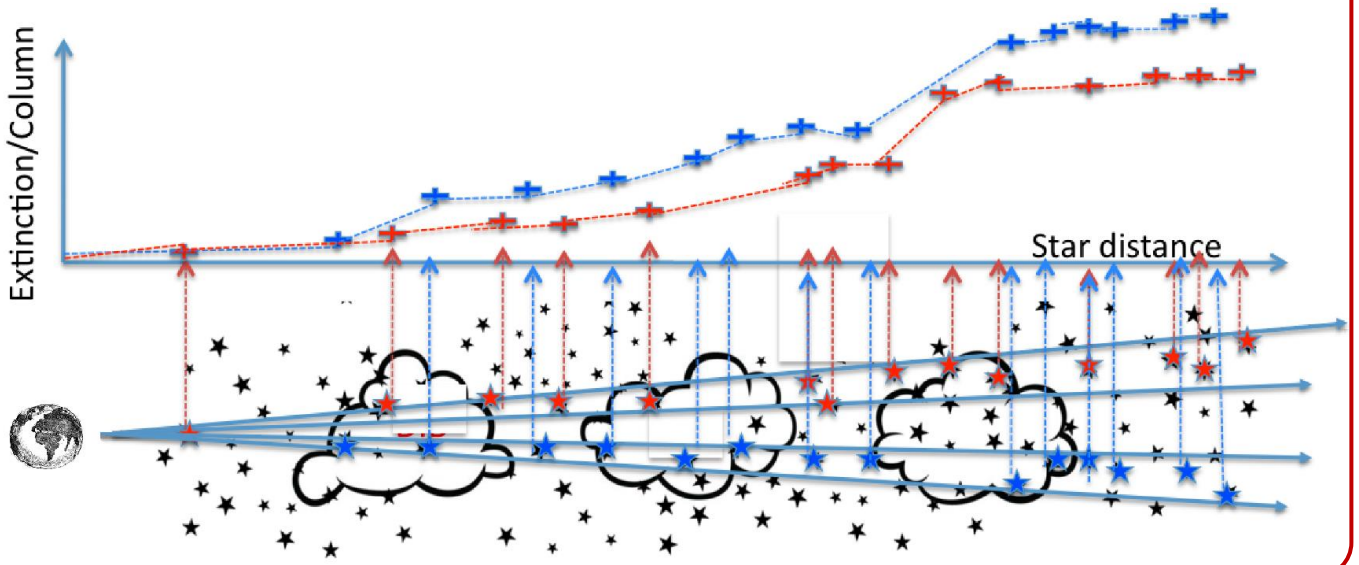
where $p(\mu_i, E_i \mid m_i)$ is a marginalized likelihood over the stellar types Θ. Finally, the reddening profile E is parametrized as a picewise linear linear function in μ ($\alpha_i = \Delta E^{(i)}$)

Dust reddening

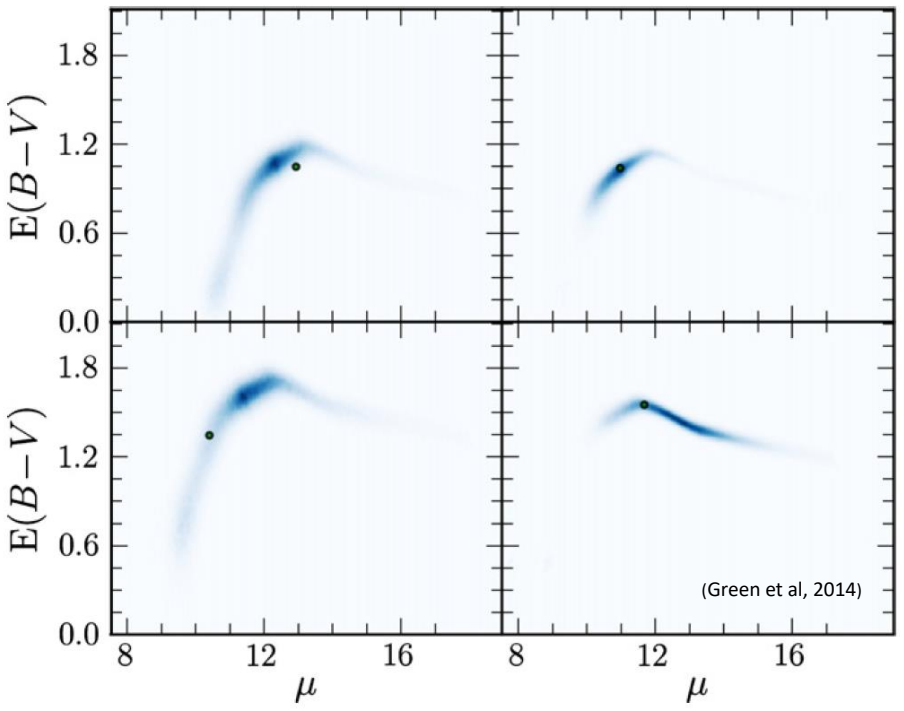
- Dust grains efficiently scatters short wavelength photons ("blue light").
- Dust clouds reduces the magnitude of stars in high-frequency bands ("reddening")
- The amount of reddening is proportional to the amount of dust along the line of sight







Test with mock stars

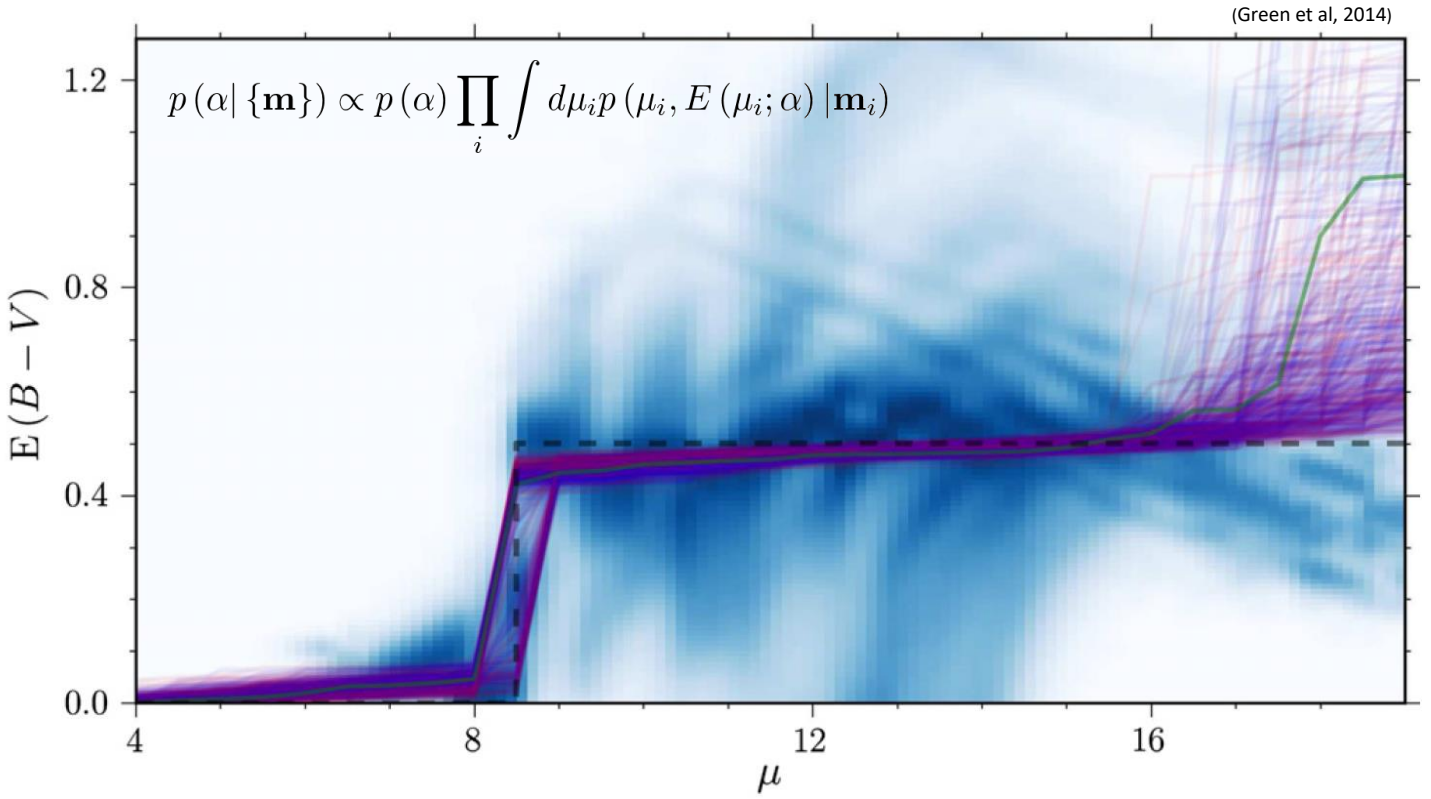


Distance and reddening estimates for four simulated stars. The posterior in distance and reddening is shown as a heat map, the "true" distances and reddenings for the stars are shown as green dots.

$$p(\mu_i, E_i|\boldsymbol{m}_i) \equiv \frac{1}{Z_i} \int d\boldsymbol{\Theta}_i \; p(\boldsymbol{m}_i|\mu_i, \boldsymbol{\Theta}_i, E_i) p(\mu_i, \boldsymbol{\Theta}_i)$$

- Inferred posterior densities of 150 simulated stars stacked on top of one another. This shows the information fed into the second stage of the analysis, where the reddening as a function of distance is recovered from the individual stellar probability densities.
- The curves show possible reddening profiles, conditioned on the mock photometry. The green curve traces the most probable reddening profile. For this example, a single cloud of depth $E(B-V) = 0.5$ at distance modulus $\mu = 8.5$ (dashed black line) has been used.
- A priori, having no reddening away from the cloud is unlikely, and this induces a slight gradual increase in inferred reddening beyond the cloud.

(Green et al, 2014)

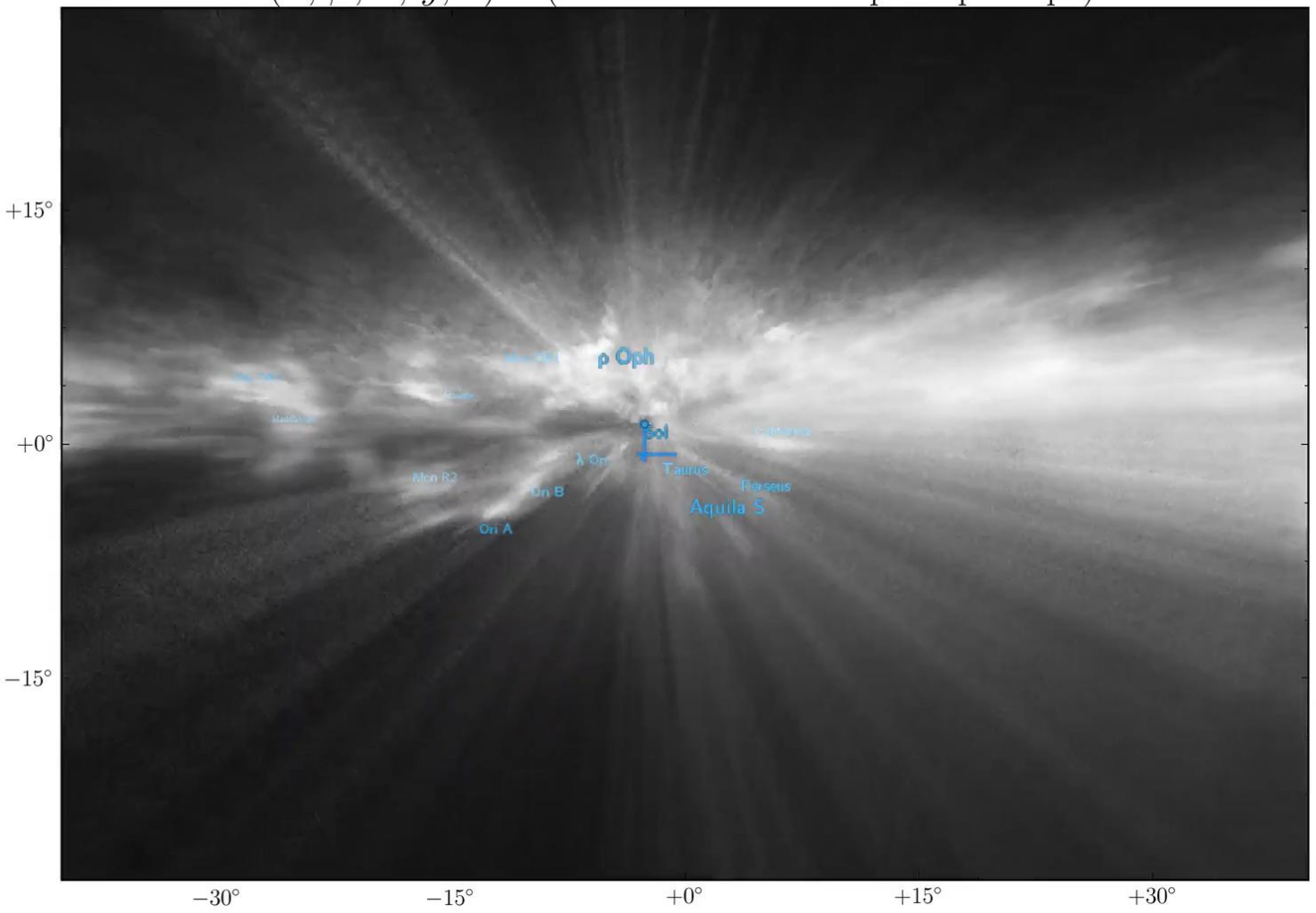


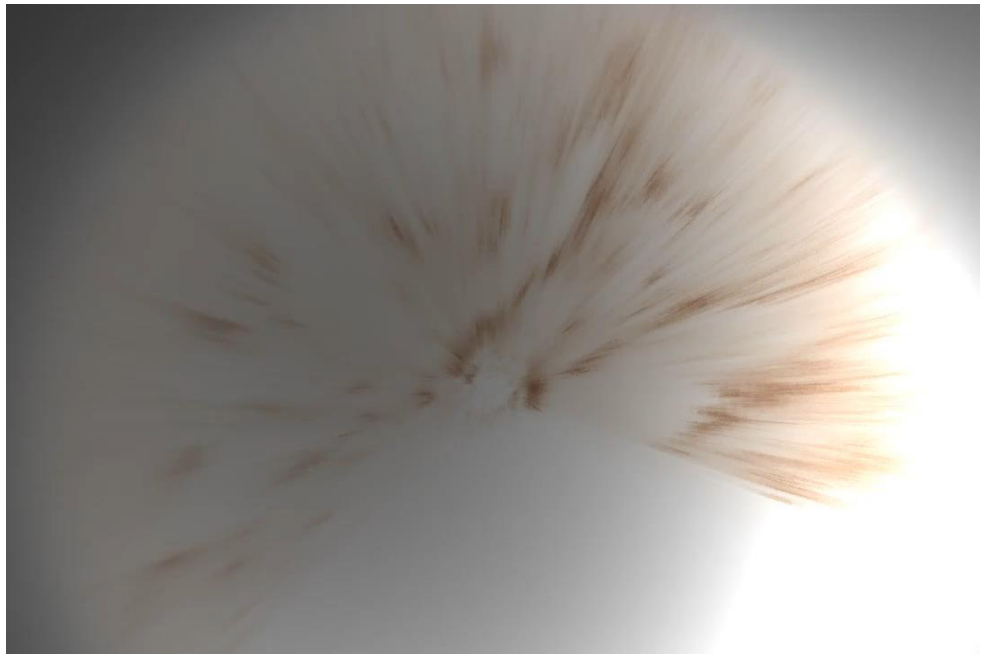
Rendering of dust distribution

Two video showing the 3D distribution of dust in the Milky Way

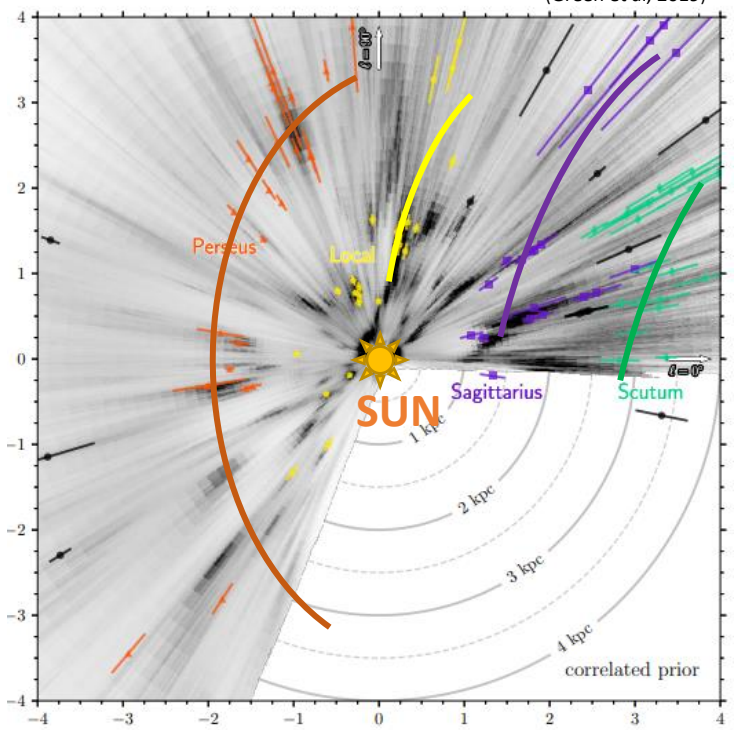
- Bobbing through galactic plane
- Milky Way tour

$(\alpha,\ \beta,\ x,\ y,\ z) = (97.1^\circ\ \ -177.0^\circ\ \ 700\,\mathrm{pc}\ 69\,\mathrm{pc}\ 70\,\mathrm{pc})$





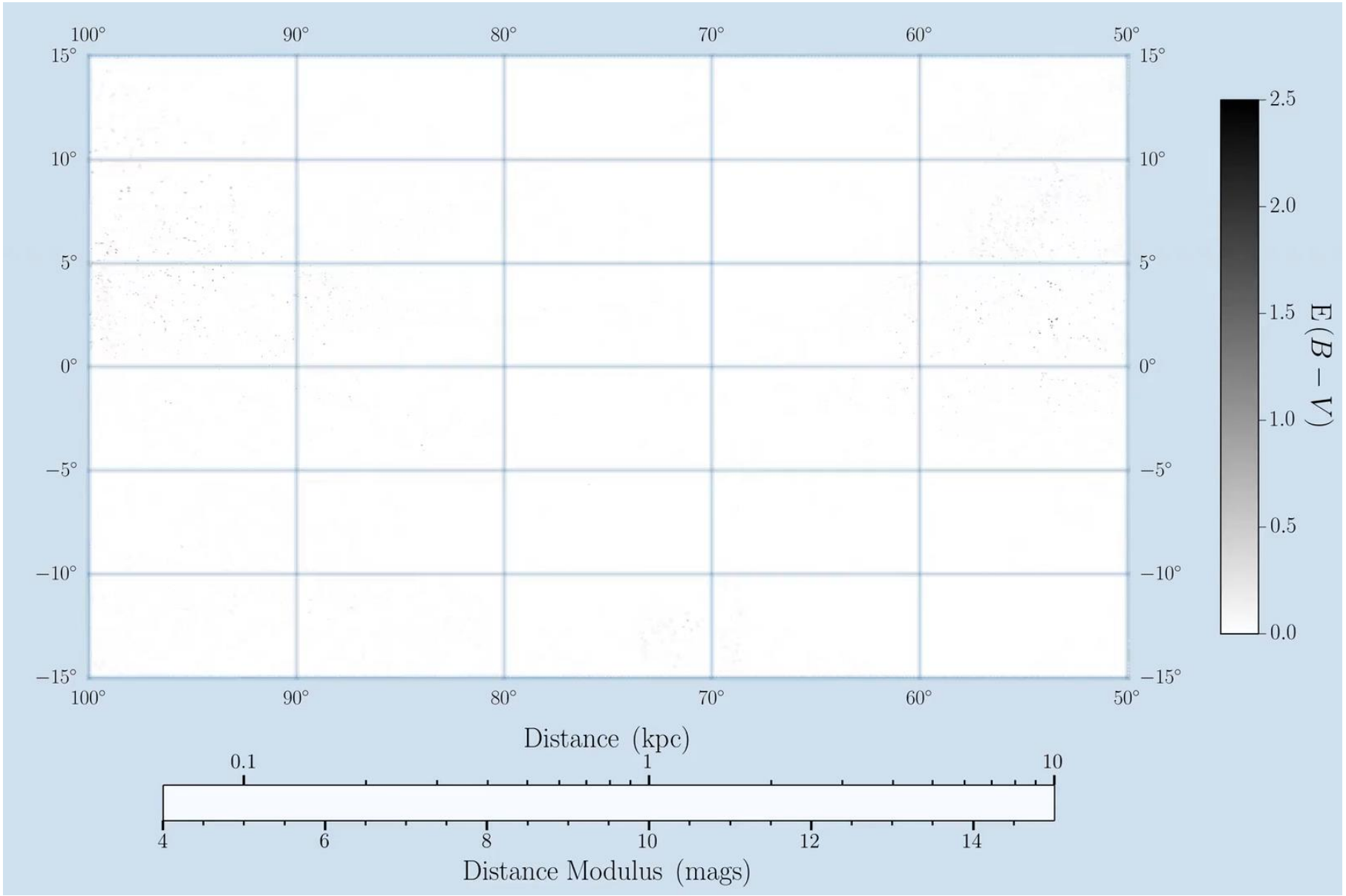
(Green et al, 2019)



⇒ Towards Galactic Center

Cumulative reddening distribution

(Green et al, 2019)



Final Remarks

- Bayesian statistical methods start with existing 'prior' beliefs, and update these using data to give 'posterior' beliefs.
- The core of Bayes statistics is Bayes theorem, composed by 4 main ingredients:
  - *Prior distribution*: express the present state of knowledge
  - *Likelihood*: description of the measurement
  - *Evidence*: used to compare different models
  - *Posterior distribution*: probability for a given model parameters knowing the data
- Bayes statistics is nowadays widely used in almost all branch of science. Two examples applied to astrophysics are:
  - Gravitational waves: from the analysis of single events to population studies
  - Studies of the Milky Way structure

References:

- C. M. Bishop, "*Pattern recognition and machine learning*", 2006
- S. Sharma, "*Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy*", 2017
- Thrane and Talbot, "*An introduction to Bayesian inference in gravitational-wave astronomy: parameter estimation, model selection, and hierarchical models*", 2020
- Green et. al., "*Measuring distances and reddenings for a billion stars: toward a 3d dust map from pan-starrs 1*", 2014
- Green et. al., "*A 3D Dust Map Based on Gaia, Pan-STARRS 1 and 2MASS*", 2019

Backup Slides

Hierarchical Models

Hierarchical Bayesian inference is a formalism, which allows us to go beyond **individual events** in order to study **population properties**

The population properties of some set of events is described by the shape of the prior

$$\underbrace{p\left(\Lambda|d\right)}_{\text{Hyper-posterior}} = \frac{\mathcal{L}\left(d|\Lambda\right)\pi\left(\Lambda\right)}{\mathcal{Z}_\Lambda}$$

$$\mathcal{L}(d|\Lambda) = \int d\theta\, \mathcal{L}(d|\theta)\, \pi(\theta|\Lambda)$$

$$\underbrace{\mathcal{Z}_\Lambda}_{\text{Hyper-evidence}} \equiv \int d\Lambda\, \mathcal{L}(d|\Lambda)\, \underbrace{\pi(\Lambda)}_{\text{Hyper-prior}}$$

If we consider the analysis of a set of events **d**

$$p_{\rm tot}(\Lambda|\vec{d}) = \frac{\mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\,\pi(\Lambda)}{\int d\Lambda\, \mathcal{L}_{\rm tot}(\vec{d}|\Lambda)\,\pi(\Lambda)}$$

$$\mathcal{L}_{\rm tot}(\vec{d}|\Lambda) = \prod_i^N \int d\theta_i\, \mathcal{L}(d_i|\theta_i)\,\pi(\theta_i|\Lambda)$$

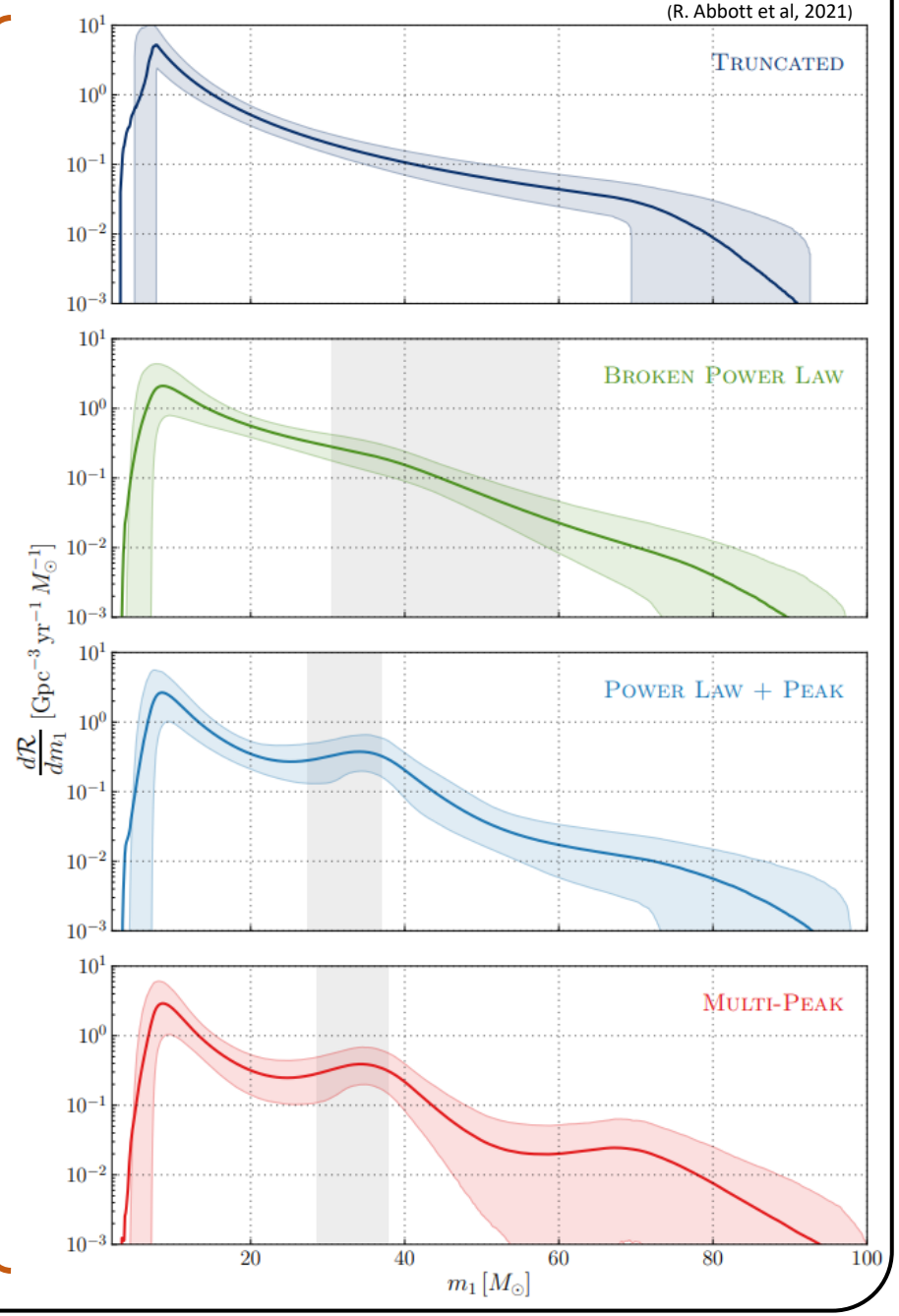
Similarly to standard Bayesian analysis, one can make model comparison by comparing the evidence for different hyper-models

Calculating hyper-evidence can be computationally intensive!

One possible way is to break the integrals into individual integral for each event and subsequently recombine them ("recycling")

Herarchical Models for GWs

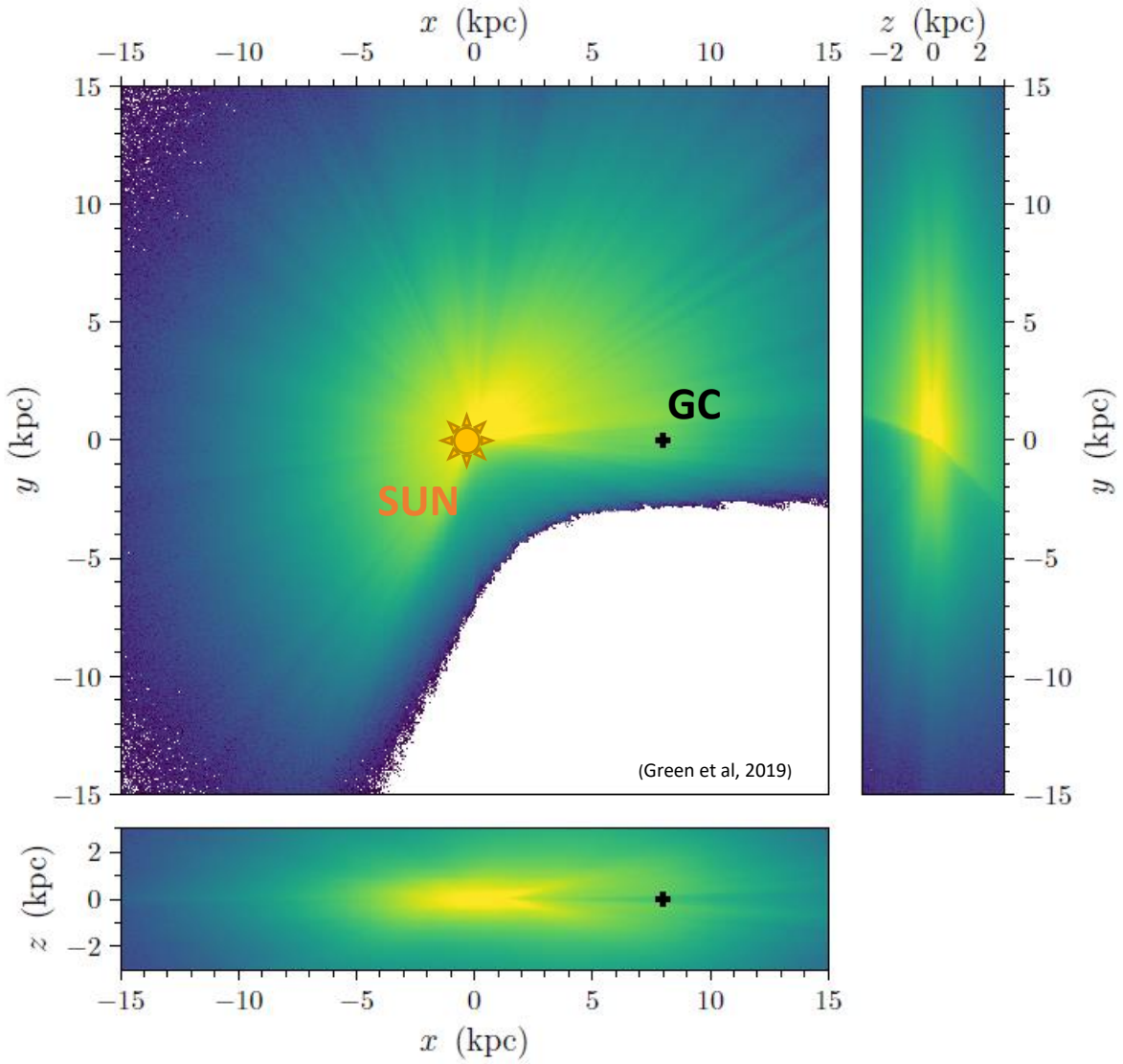
(R. Abbott et al, 2021)



$$\pi(m_1|\alpha) \propto m_1^\alpha$$

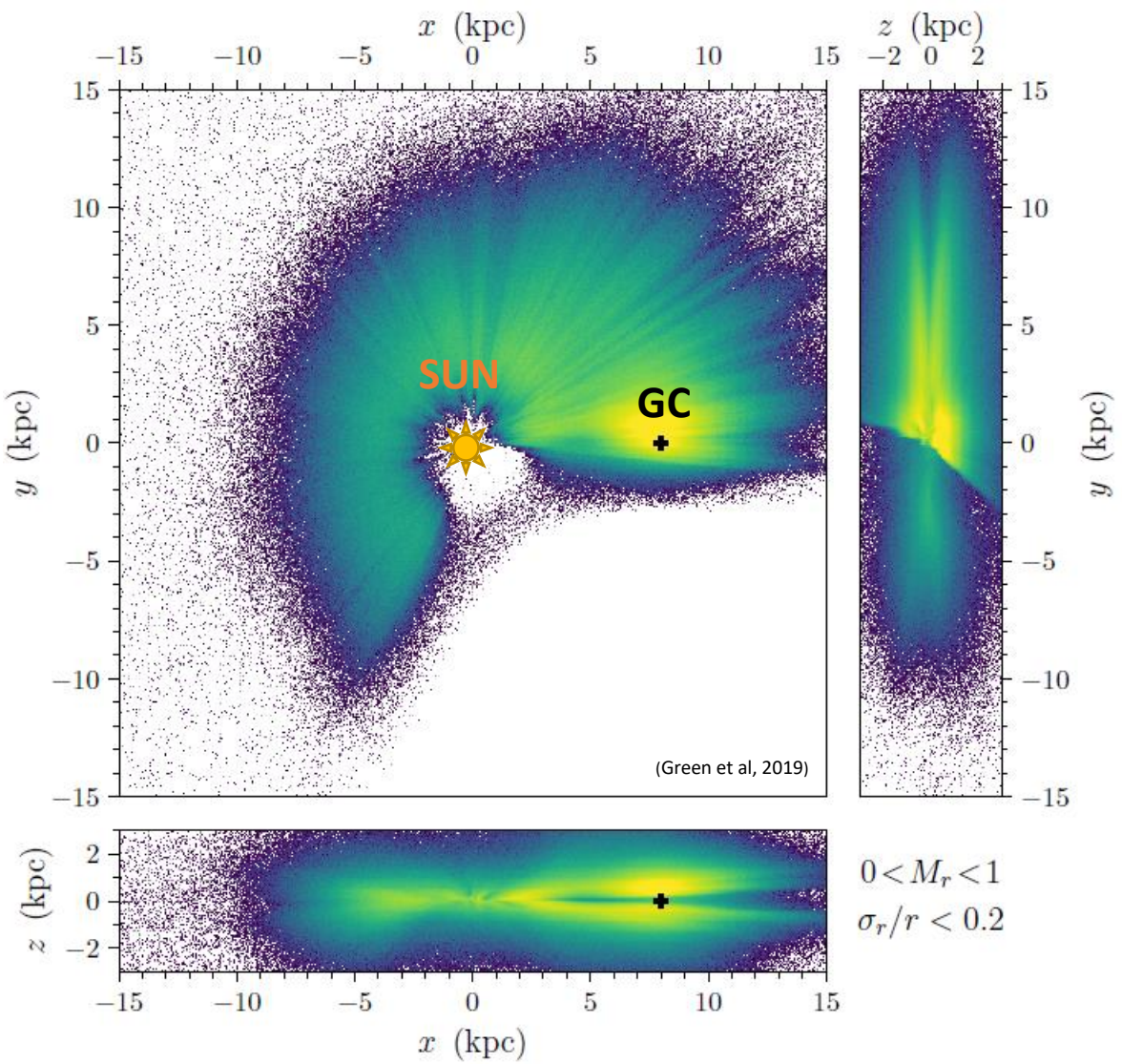
It is possible to consider diverse mass distribution. Different models are build considering various physical effects: PLP (Gaussian peak introduced to account pile-up from pulsational pair instability SN), MP (account for hierarchical merging), …

Stars distribution



Distribution of stars

Distribution of stars of absolute magnitude 0 < Mr < 1 with well-determined distances

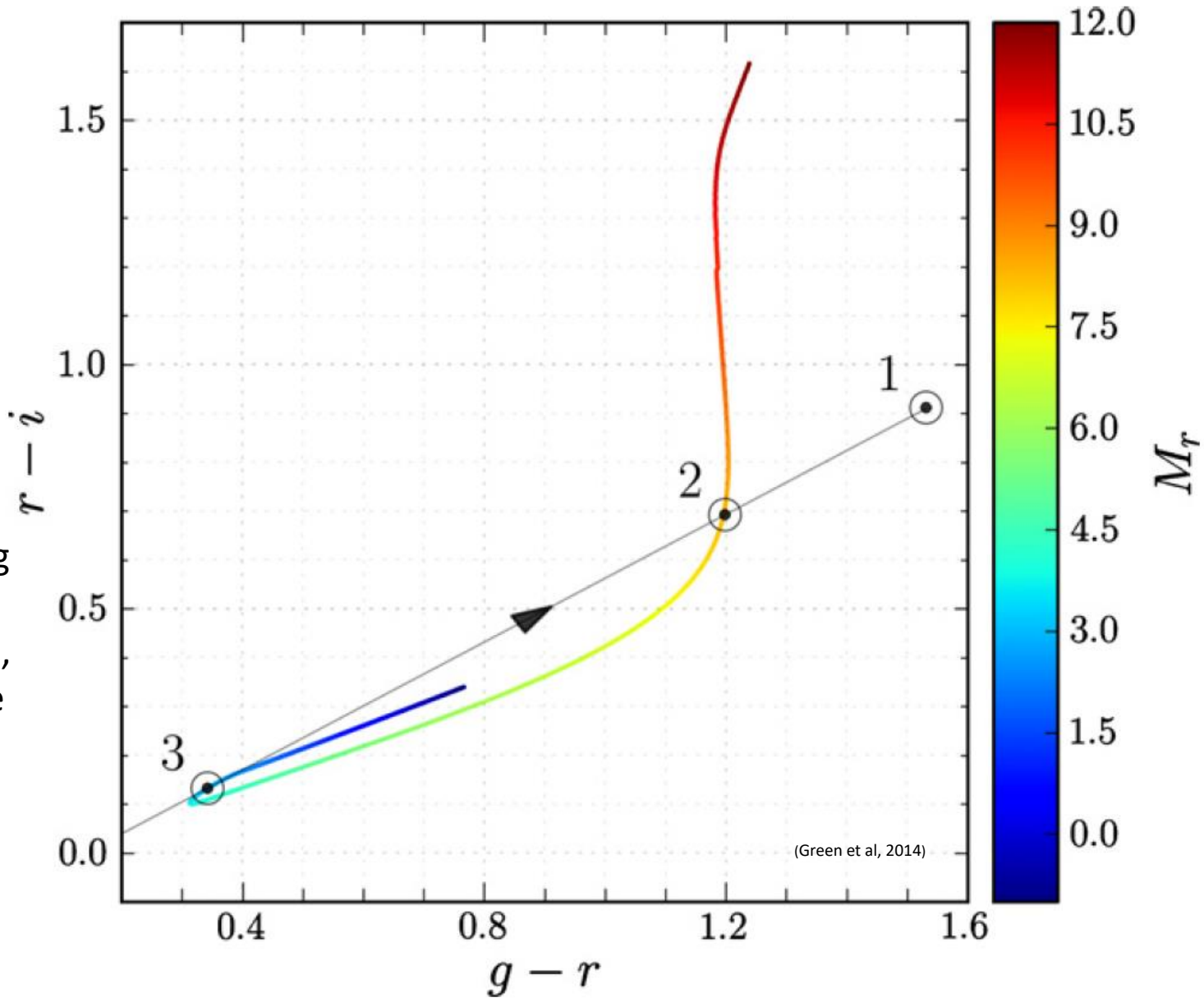


Photometric parallaxes

- A star is observed at location (1) in color–color space.
- Its dereddened colors may lie along any point on the gray line, parallel to the reddening vector.
- The intersections of this line with the model stellar locus, labeled (2) and (3), represent the most likely intrinsic stellar types.
- The posterior density for the star will thus have two modes—one at larger distance and lesser reddening (2) and one at smaller distance and greater reddening (3). For simplicity, we assume Solar metallicity in this example.

This is how one would make a distance and reddening determination by eye. Our more rigorous Bayesian method takes into account photometric uncertainties, as well as priors on stellar type and Galactic structure

Filter	Wavelength (Angstroms)
Ultraviolet (u)	3543
Green (g)	4770
Red (r)	6231
Near Infrared (i)	7625
Infrared (z)	9134



The Solar system neighborhood

(Green et al, 2019)