# Astrostatistics: Wed 19 Feb 2020

https://github.com/CambridgeAstroStat/PartIII-Astrostatistics-2020

- Today: continue Bayesian computation / Monte Carlo Methods:
  - MacKay: Ch 29-30; Bishop: Ch 11; Gelman BDA
  - Givens & Hoeting. "Computational Statistics" (Free through Cambridge Library iDiscover)
  - Importance Sampling
  - Case Study: Bayesian Estimates of the Mass of the Milky Way Galaxy
- Example Class 2, Thu Feb 27, 3:30pm MR13

# Monte Carlo Integration

Typically, we want to summarise the posterior and compute expectations of the form:

$$I = \mathbb{E}[f(\boldsymbol{\theta})|\boldsymbol{D}] = \int f(\boldsymbol{\theta})\, P(\boldsymbol{\theta}|\, \boldsymbol{D})\, d\boldsymbol{\theta}$$

Using m samples from the posterior:

$$\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}|\, \boldsymbol{D})$$

$$\hat{I} = \frac{1}{m}\sum_{i=1}^{m} f(\boldsymbol{\theta}_i) \longrightarrow \boldsymbol{I} \quad \text{(LLN for large m)}$$

Monte Carlo Error:

$$\mathrm{Var}[\hat{I}] = \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}[f(\boldsymbol{\theta})] = \frac{1}{m}\mathrm{Var}[f(\boldsymbol{\theta})] \approx \frac{1}{m}\widehat{\mathrm{Var}}[\{f(\boldsymbol{\theta}_i)\}]$$

# Importance Sampling

Objective: compute expectation wrt distribution P(θ)

$$I = \mathbb{E}[f(\theta)] = \int f(\theta) P(\theta)\, d\theta$$

Example: Posterior

$$P(\theta) = \frac{L(\theta)\pi(\theta)}{\int L(\theta)\pi(\theta)\, d\theta}$$

Likelihood

Prior

Can evaluate P(θ) but not sample from it.
Choose importance function Q(θ) you can evaluate and sample!

Importance Sampling estimate

$$\theta_1, \ldots, \theta_n \overset{iid}{\sim} Q(\theta)$$

$$\hat{I} = \sum_{i=1}^{n} f(\theta_i) w^*(\theta_i)$$

$$w_i^* \equiv w^*(\theta_i) \equiv P(\theta_i)/Q(\theta_i)$$

# Self-Normalised Importance Sampling

Objective: compute expectation:

$$I = \mathbb{E}[f(\theta)] = \int f(\theta) P(\theta)\, d\theta$$

but can only evaluate unnormalised $\tilde{P}(\theta)$

Example: unnormalised posterior $\tilde{P}(\theta) = L(\theta)\pi(\theta)$

Normalised posterior is $P(\theta) = \dfrac{\tilde{P}(\theta)}{Z_P}$

but cannot calculate: $Z_P = \int L(\theta)\pi(\theta)\, d\theta$

# Self-Normalised Importance Sampling

Objective: compute expectation:

$$I = \mathbb{E}[f(\theta)] = \int f(\theta) P(\theta)\, d\theta$$

$$I = \int f(\theta) P(\theta)\, d\theta = \int f(\theta) \frac{\tilde{P}(\theta)}{\int \tilde{P}(\theta)\, d\theta}\, d\theta = \frac{\int f(\theta) \frac{\tilde{P}(\theta)}{Q(\theta)} Q(\theta)\, d\theta}{\int \frac{\tilde{P}(\theta)}{Q(\theta)} Q(\theta)\, d\theta}$$

Draws from Impt Fcn Q(θ): $\theta_1, \ldots, \theta_n \overset{iid}{\sim} Q(\theta)$

Importance Sampling estimate:

$$\hat{I} = \frac{\sum_{i=1}^{n} f(\theta_i)\tilde{w}_i}{\sum_{i=1}^{n} \tilde{w}_i} = \sum_{i=1}^{n} f(\theta_i)\, w_i$$

Self-normalised Importance Weights

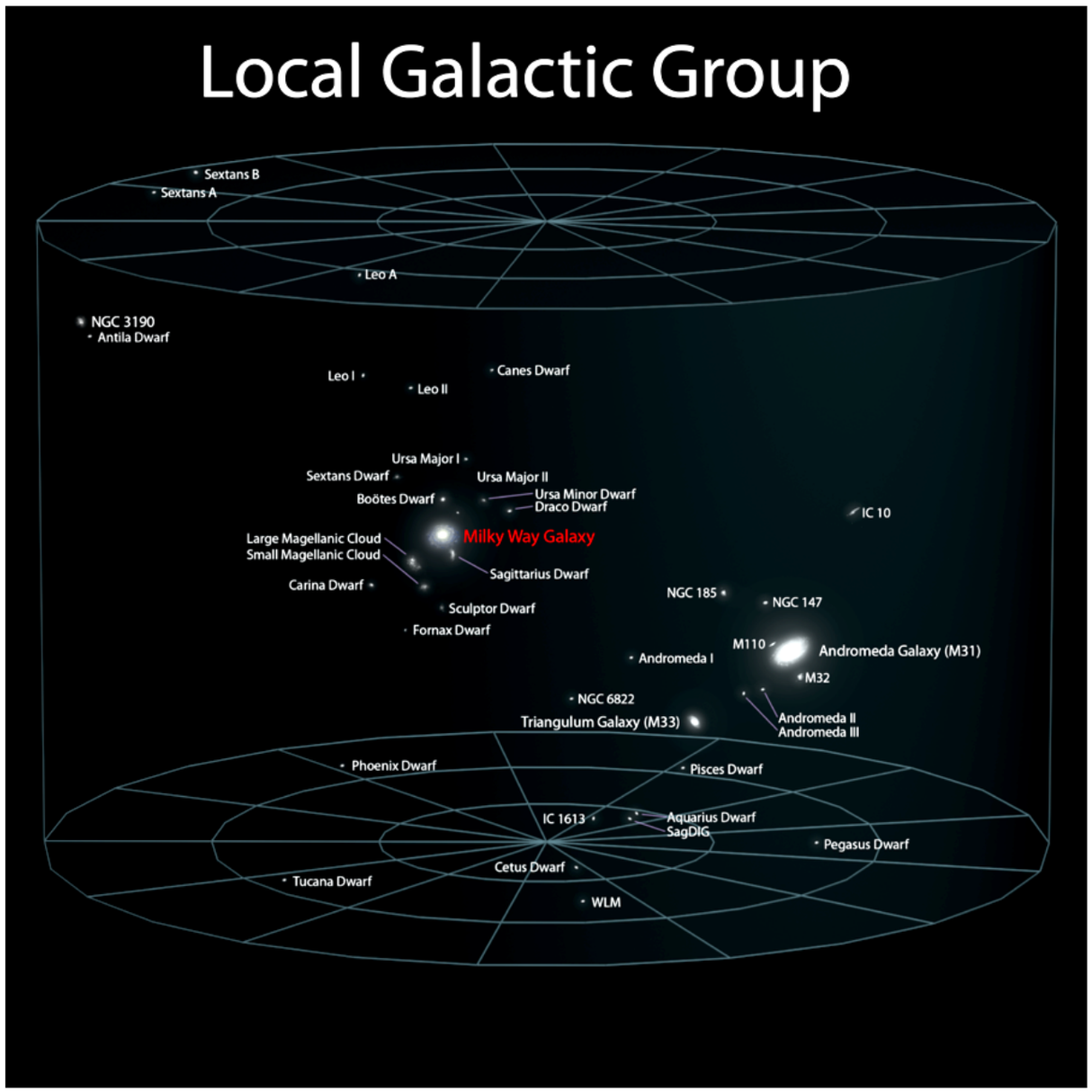
$$\tilde{w}_i \equiv \tilde{P}(\theta_i)/Q(\theta_i)$$

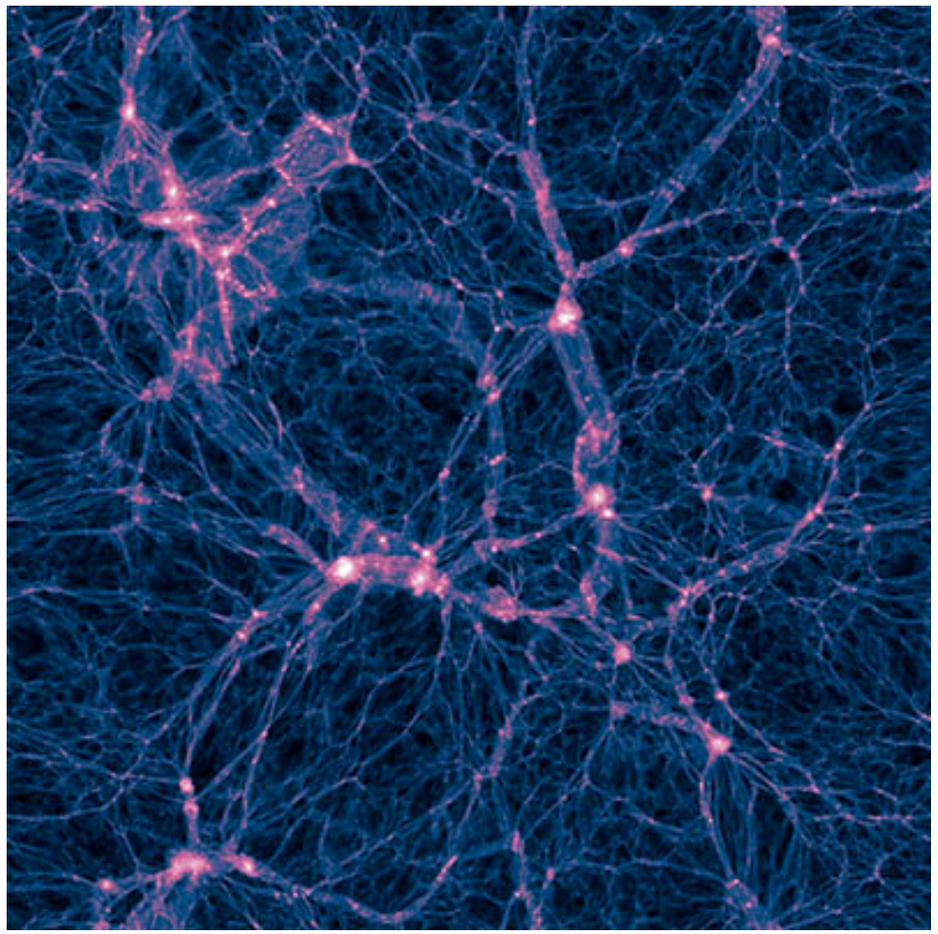
$$w_i \equiv \frac{\tilde{w}_i}{\sum_{j=1}^{n} \tilde{w}_j}$$

*Astrostatistics Case Study:*

# Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations

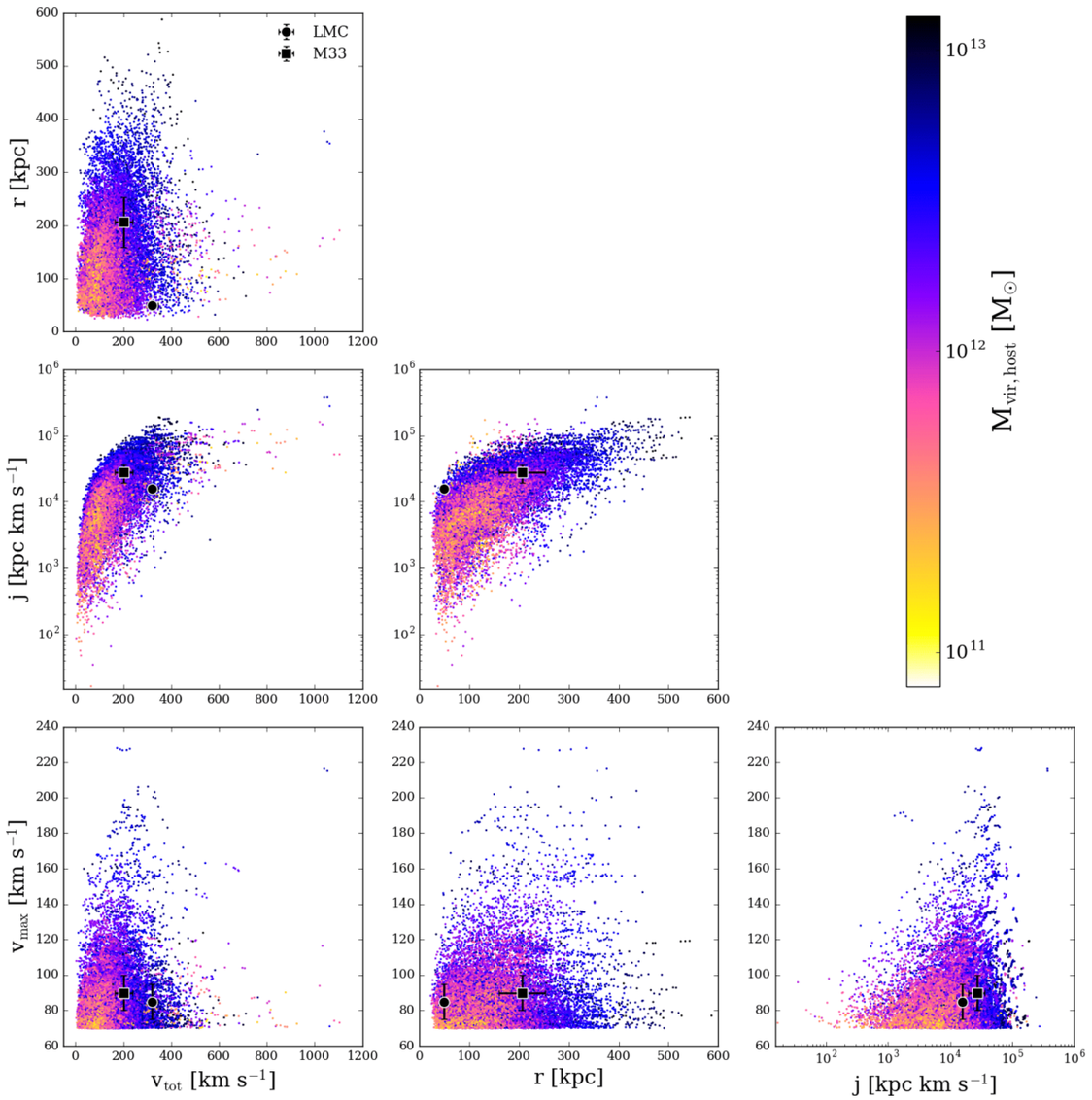
(Patel, Besla, & Mandel, 2017, 2018, arXiv:1703.05767, 1803.01878



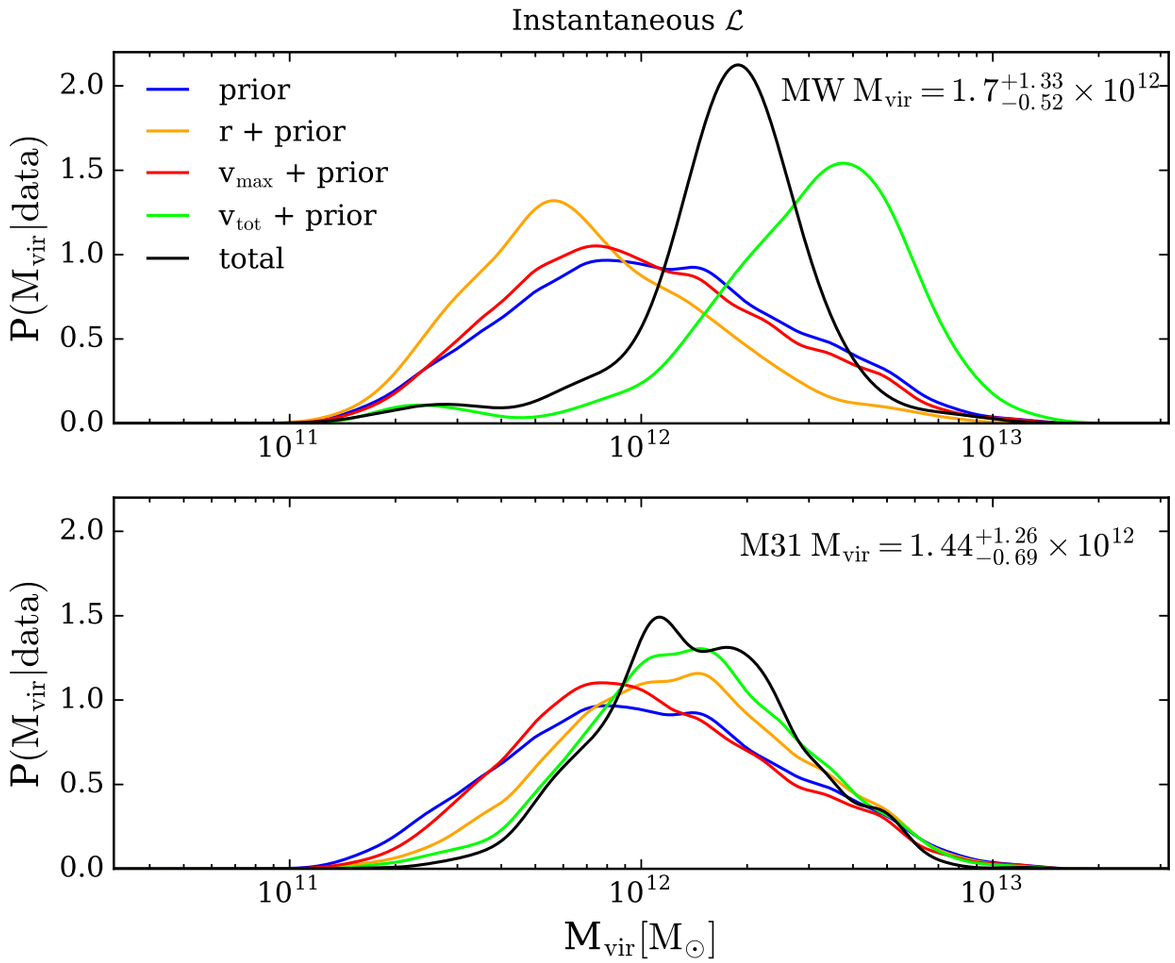



Illustris
Cosmological Simulation of Galaxy Formation

# *Astrostatistics Case Study:*
# Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations
# (Patel, Besla, & Mandel, 2017, 2018, arXiv:1703.05767, 1803.01878)




Simulation —> Prior

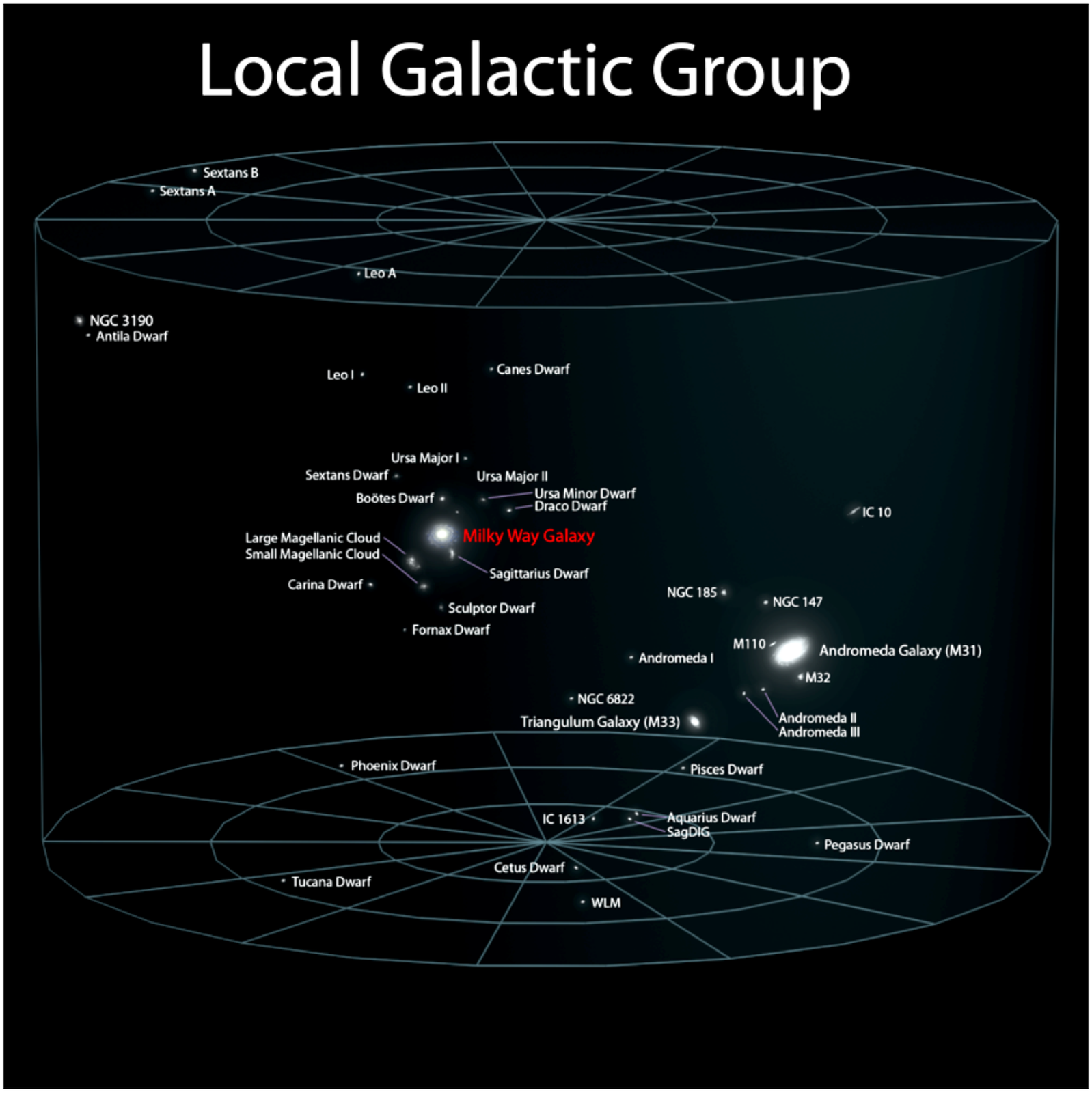



- Bayesian Inference
- Importance Sampling
- Kernel Density Estimation

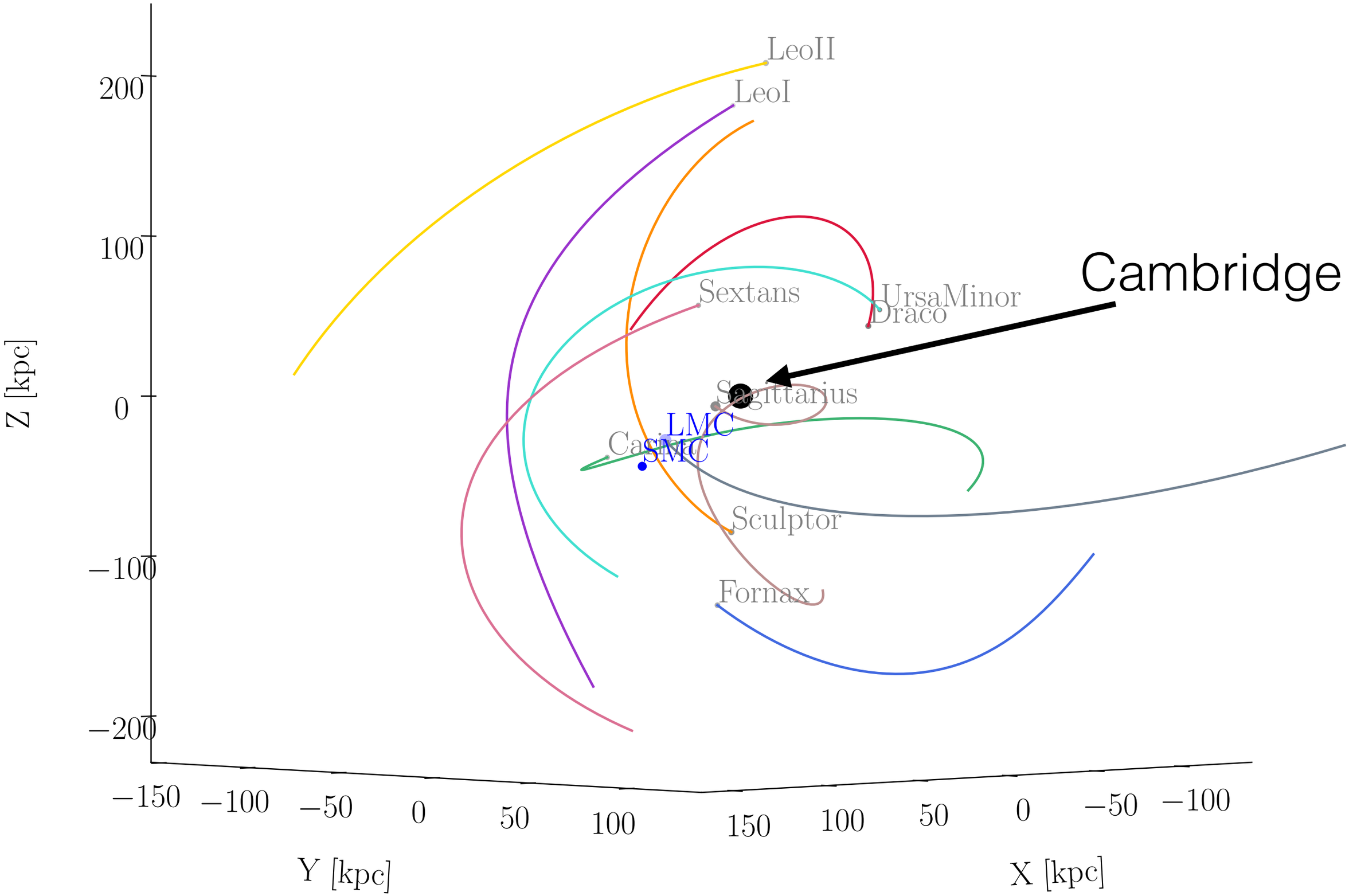
Illustris Cosmological Simulation
Movie

http://www.illustris-project.org/movies/illustris_movie_cube_sub_frame.mp4

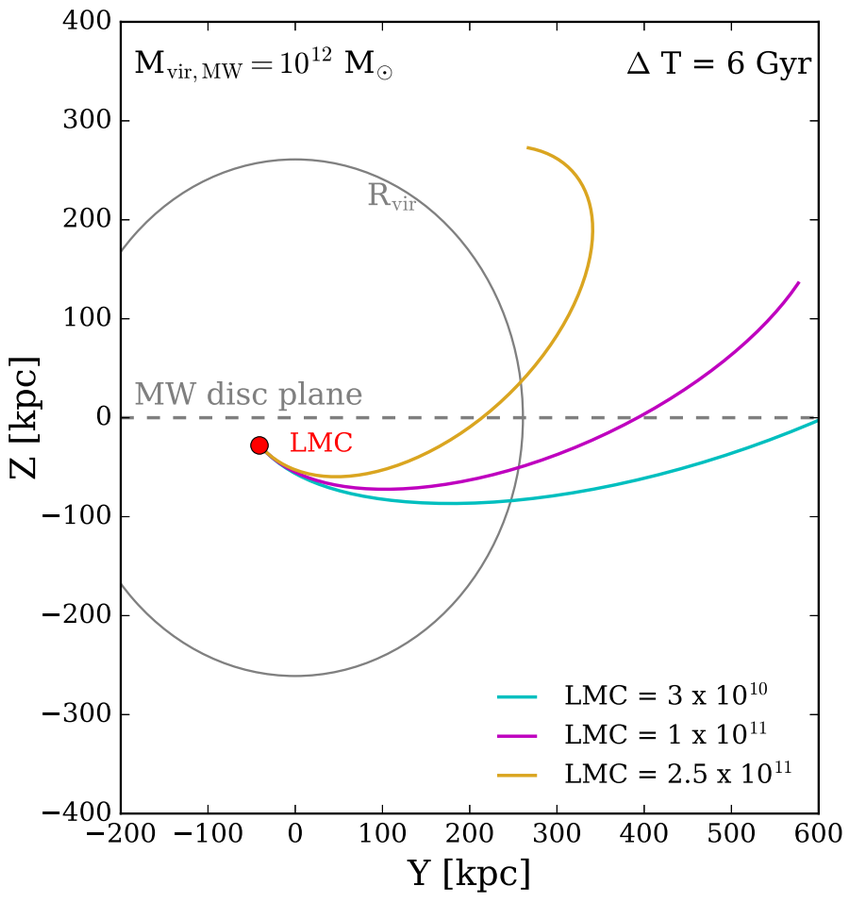
# Milky Way has satellite galaxies

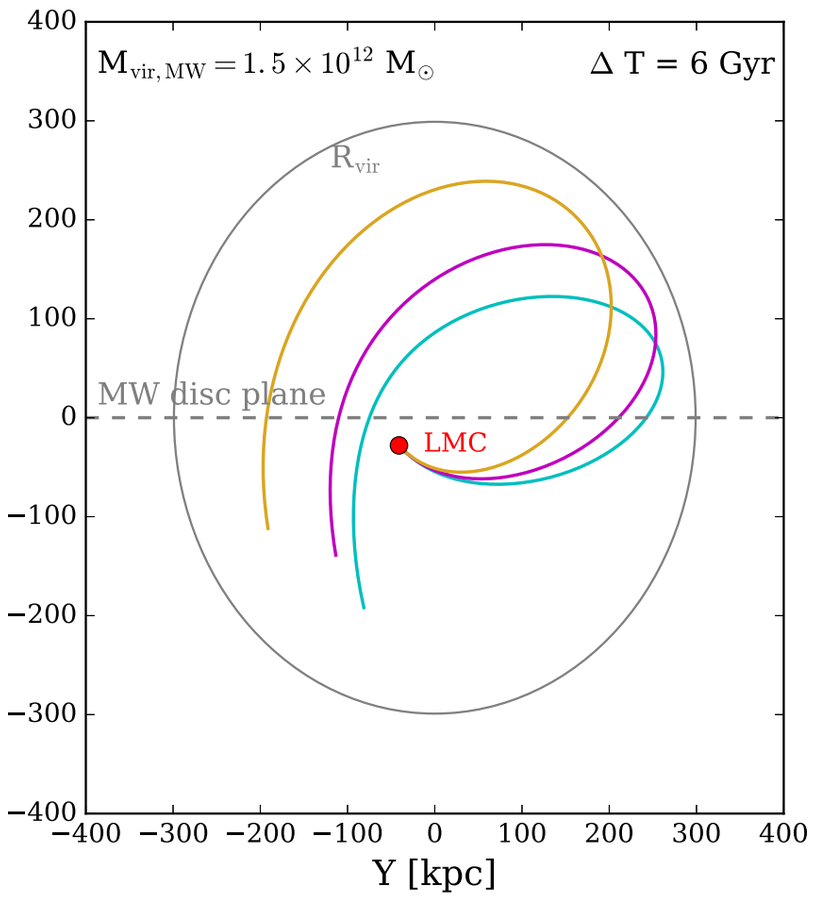


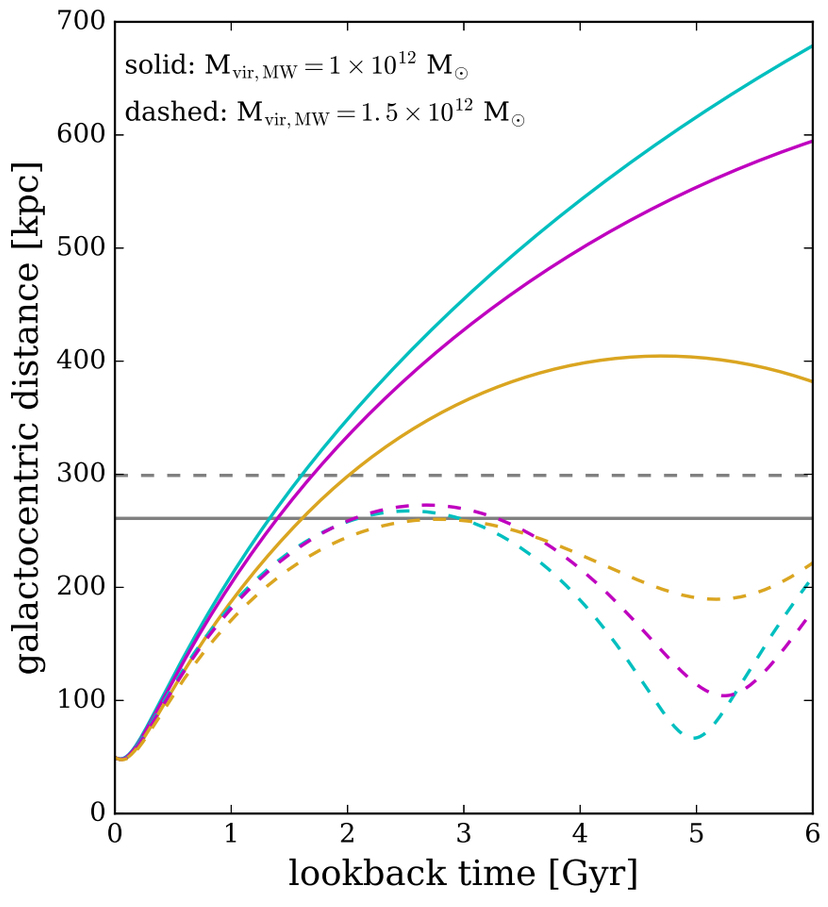
# Satellite Galaxies are moving around (us)



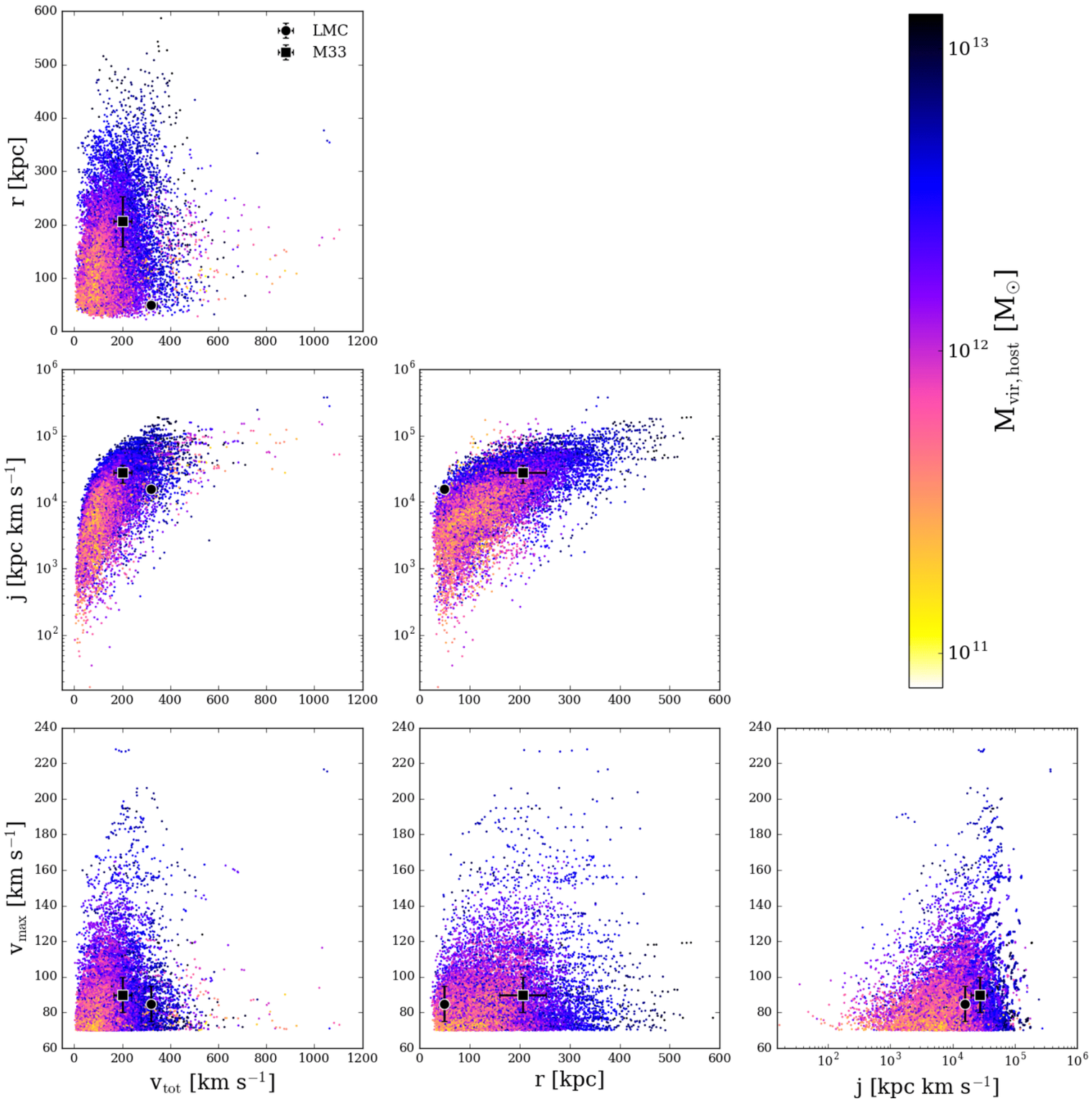
# Their trajectories depend on the Milky Way Mass



# Velocities (v), positions (r), momenta (j), of satellites are correlated with central Galaxy Mass via galaxy formation physics in simulations (Prior)




x = latent (true) values of v, r, j

$M_{vir}$ = Mass of Galaxy

Parameters are:

$\theta = (x, M_{vir})$

Prior density cannot be evaluated, only sampled!

# We can measure the (v, r, j) of MW's biggest satellite, Large Magellanic Cloud (LMC)

**Table 1.** Observational data (***d***) for the LMC and M33 used to build likelihoods in the Bayesian inference scheme include the maximum circular velocity, current separation from the host galaxy and total velocity relative to the host galaxy.

| | LMC $\mu$ | LMC $\sigma$ | M33 $\mu$ | M33 $\sigma$ |
|---|---|---|---|---|
| $v_{\rm max}^{\rm obs}$ (km s$^{-1}$) | 85[a] | 10 | 90[b] | 10 |
| $r^{\rm obs}$ (kpc) | 50 | 5 | 203 | 47 |
| $v_{\rm tot}^{\rm obs}$ (km s$^{-1}$) | 321 | 24 | 202 | 38 |
| $j^{\rm obs}$ (kpc km s$^{-1}$) | 15 688 | 1788 | 27 656 | 8219 |

*Notes.* [a]The maximal circular velocity of the LMC's halo rotation curve is adopted from Besla et al. (2012).
[b]M33's halo rotation curve maximum is duplicated from van der Marel et al. (2012b).
M33's position, velocity and their errors are adopted from Paper I (table 1), and references within.

**Data d =**

**Measurement Likelihood**

$$\mathcal{L}(\boldsymbol{x}|\,\boldsymbol{d}) = N(v_{\rm max}^{\rm obs}|\,v_{\rm max},\sigma_v^2)\times N(r^{\rm obs}|r,\sigma_r^2)\times N(v^{\rm obs}|v_{\rm tot},\sigma_{\rm v}^2), \quad (8)$$

where

$$N(y|\,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \quad (9)$$

# How do we combine these measurements (likelihood) with the joint prior on P(v, r, j, M) from the Simulations?

d = measurements of satellite properties

x = latent (true) values of satellite properties

$M_{vir}$ = Mass of Galaxy

*3.2.3 Importance sampling*

Now that the prior and likelihood have been defined, we return to Bayes' theorem:

(Ignore C)

$$P(\boldsymbol{x}, M_{\rm vir}|\,\boldsymbol{d}, C) \propto P(\boldsymbol{d}|\,\boldsymbol{x}) \times P(\boldsymbol{x}, M_{\rm vir}|\,\boldsymbol{C}), \quad (11)$$

Likelihood (observations)

Prior (samples from Simulation)

# Importance Sampling

Parameters are: $\theta = (x, M_{vir})$

measured data are: d

Expectations of functions of the physical parameters under the posterior PDF are approximated as sums over the $n$ samples as follows:

$$\int f(\boldsymbol{\theta}) P(\boldsymbol{x}, M_{\rm vir}|\, \boldsymbol{d}, \boldsymbol{C})\, {\rm d}\boldsymbol{\theta} = \frac{\int f(\boldsymbol{\theta}) P(\boldsymbol{d}|\, \boldsymbol{x}) P(\boldsymbol{x}, \underline{M}_{\rm vir}|\boldsymbol{C})\, {\rm d}\boldsymbol{\theta}}{\int P(\boldsymbol{d}|\, \boldsymbol{x}) P(\boldsymbol{x}, M_{\rm vir}|\, \boldsymbol{C})\, {\rm d}\boldsymbol{\theta}}$$

$$\approx \frac{\sum_j^n f(\boldsymbol{\theta}_j) P(\boldsymbol{d}|\, \boldsymbol{x}_j)}{\sum_j^n P(\boldsymbol{d}|\, \boldsymbol{x}_j)}. \qquad (12)$$

The denominator of this equation is the normalization constant. If

Sum over Samples from Prior P(x, $M_{vir}$)

# Importance Sampling: Computing Posterior Mean of Galaxy Mass

$$\int f(\boldsymbol{M}_{\rm vir}) P(\boldsymbol{M}_{\rm vir} | \, \boldsymbol{d}, \boldsymbol{C}) \, {\rm d}\boldsymbol{M}_{\rm vir}$$

$$= \int f(\boldsymbol{M}_{\rm vir}) \, P(\boldsymbol{x}, \boldsymbol{M}_{vir} | \, \boldsymbol{d}, \boldsymbol{C}) \, {\rm d}\boldsymbol{x} \, {\rm d}\boldsymbol{M}_{vir}$$

$$\approx \frac{\sum_j^n f(\boldsymbol{M}^j_{\rm vir}) \, P(\boldsymbol{d} | \, \boldsymbol{x}_j)}{\sum_j^n P(\boldsymbol{d} | \, \boldsymbol{x}_j)}$$

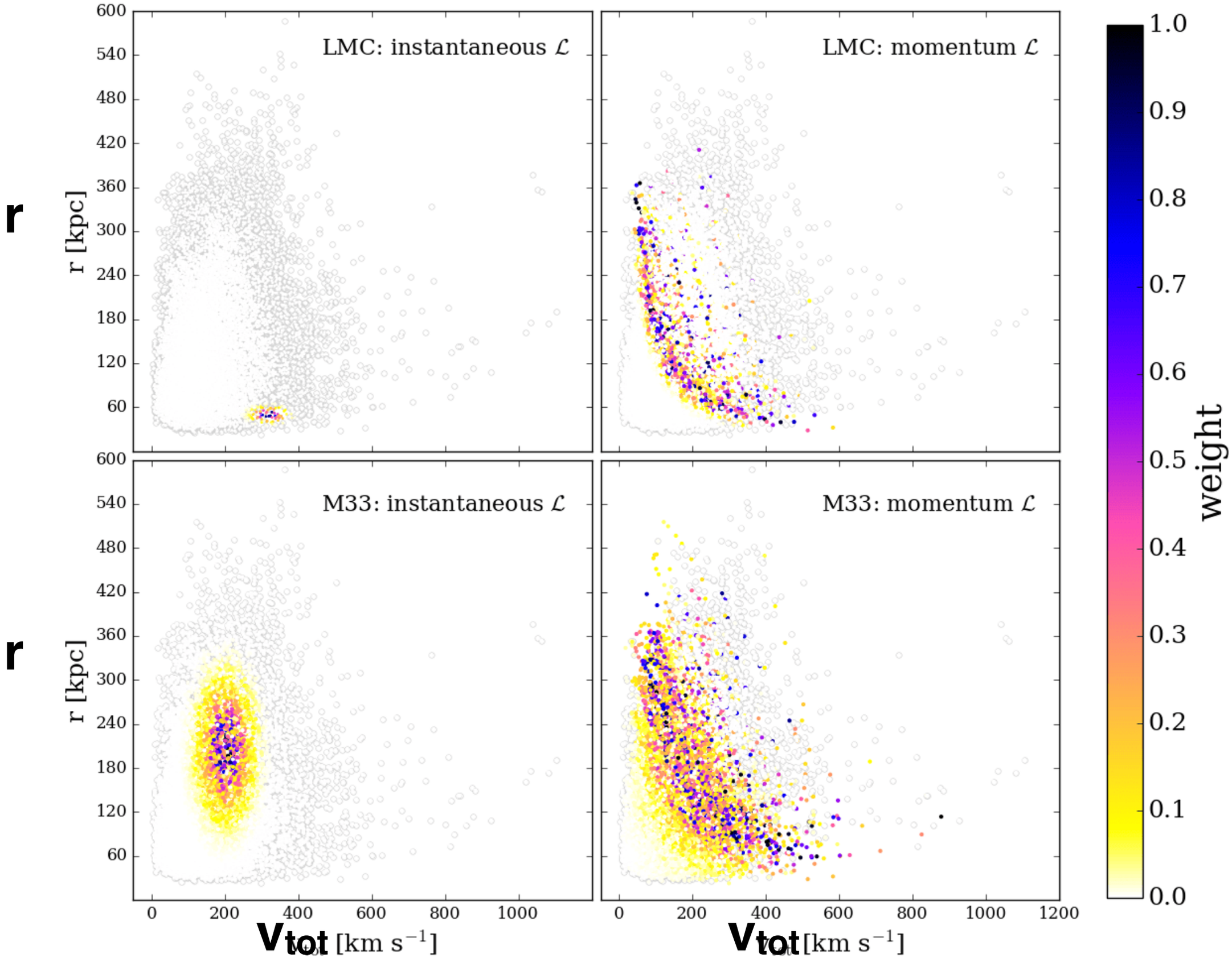
$$= \sum_j^n f(\boldsymbol{M}^j_{\rm vir}) \, w_j,$$

Impt Weight is proportional to likelihood of each sample

(Self-normalised)

where $w_i = P(\boldsymbol{d} | \, \boldsymbol{x}_i) / \sum_{\rm j}^{\rm n} P(\boldsymbol{d} | \, \boldsymbol{x}_j)$ are importance weights.

# Distribution of Importance Weights

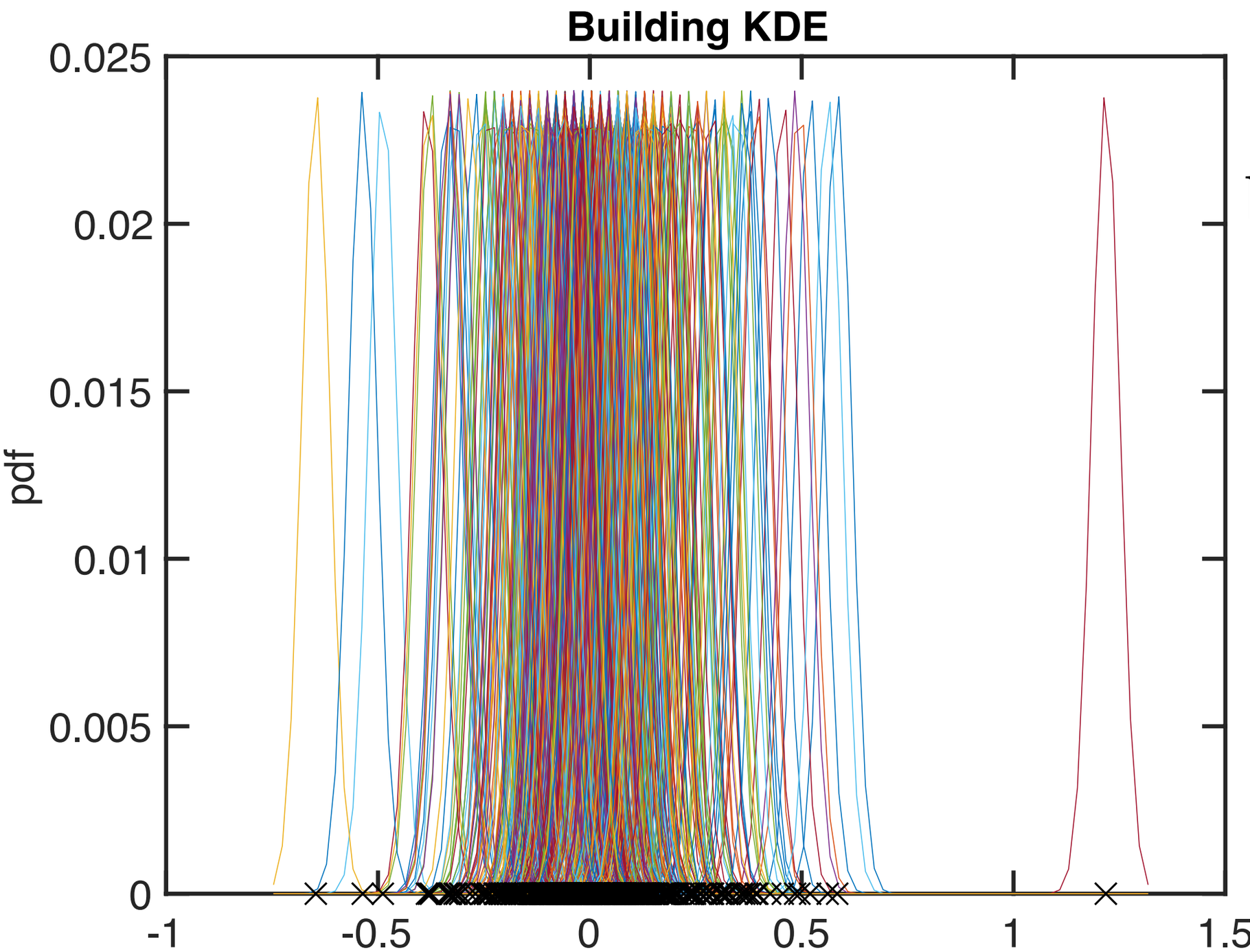



**r**

**r**

# Kernel Density Estimation (KDE) (Smooth Histogram)

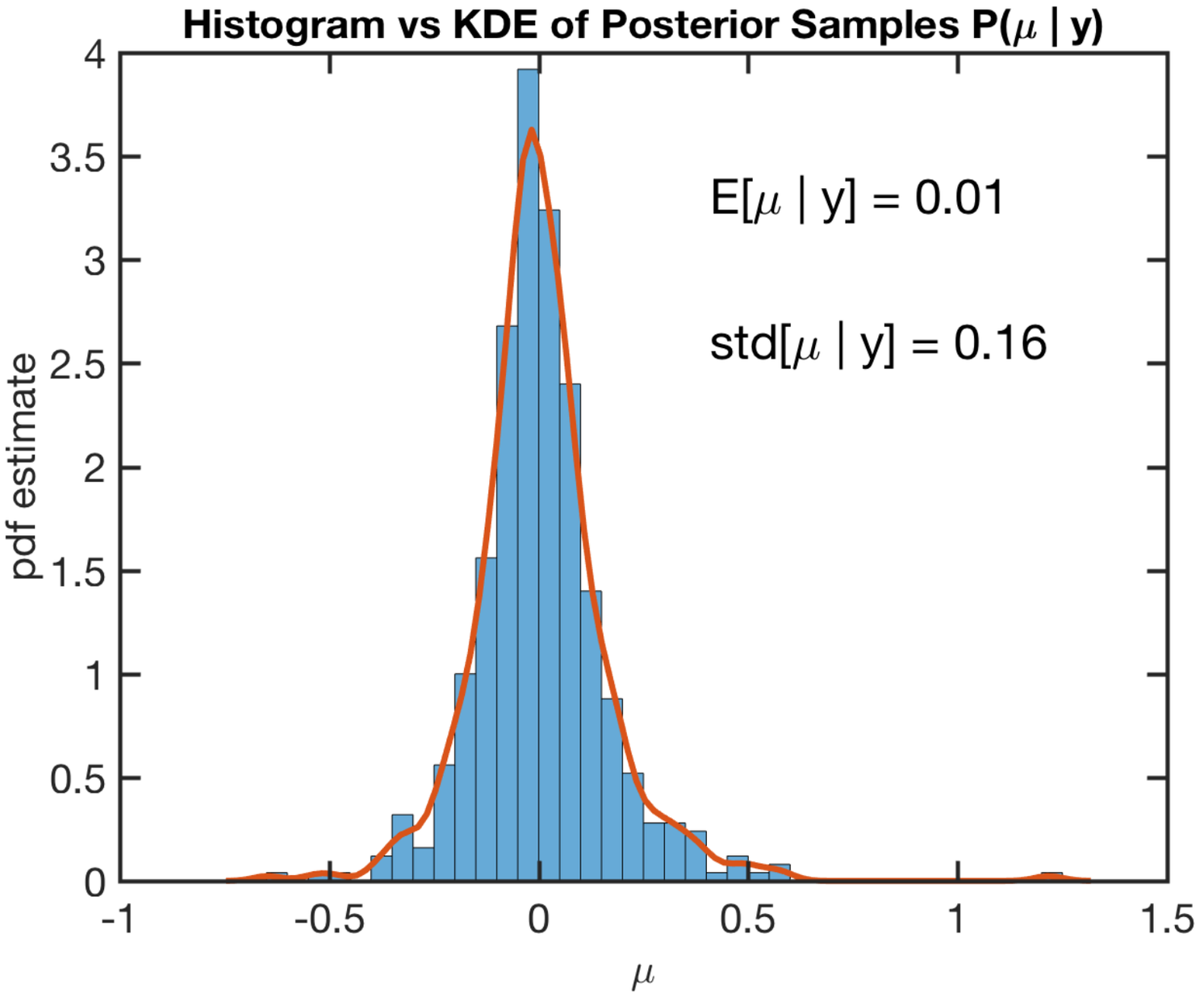
Each sample gets a Gaussian at the sample point
with an “optimal” bandwidth bw (rule of thumb)




$$\mathrm{bw} = \left(\frac{4\sigma^5}{3n}\right)^{1/5}$$

# Kernel Density Estimation (KDE) (Smooth Histogram)

Then add them up and normalise pdf to 1



# Weighted KDE

$$\mathrm{wkde}(\theta) = \sum_{s=1}^{m} w_s \times N(\theta \,|\, \theta_s, \mathrm{bw}^2)$$

$w_s$ = normalised importance weights

bandwidth: Silverman's Rule of Thumb: $\mathrm{bw} = \left(\frac{4\sigma^5}{3n}\right)^{1/5}$

Estimate $\sigma^2$ from variance estimate

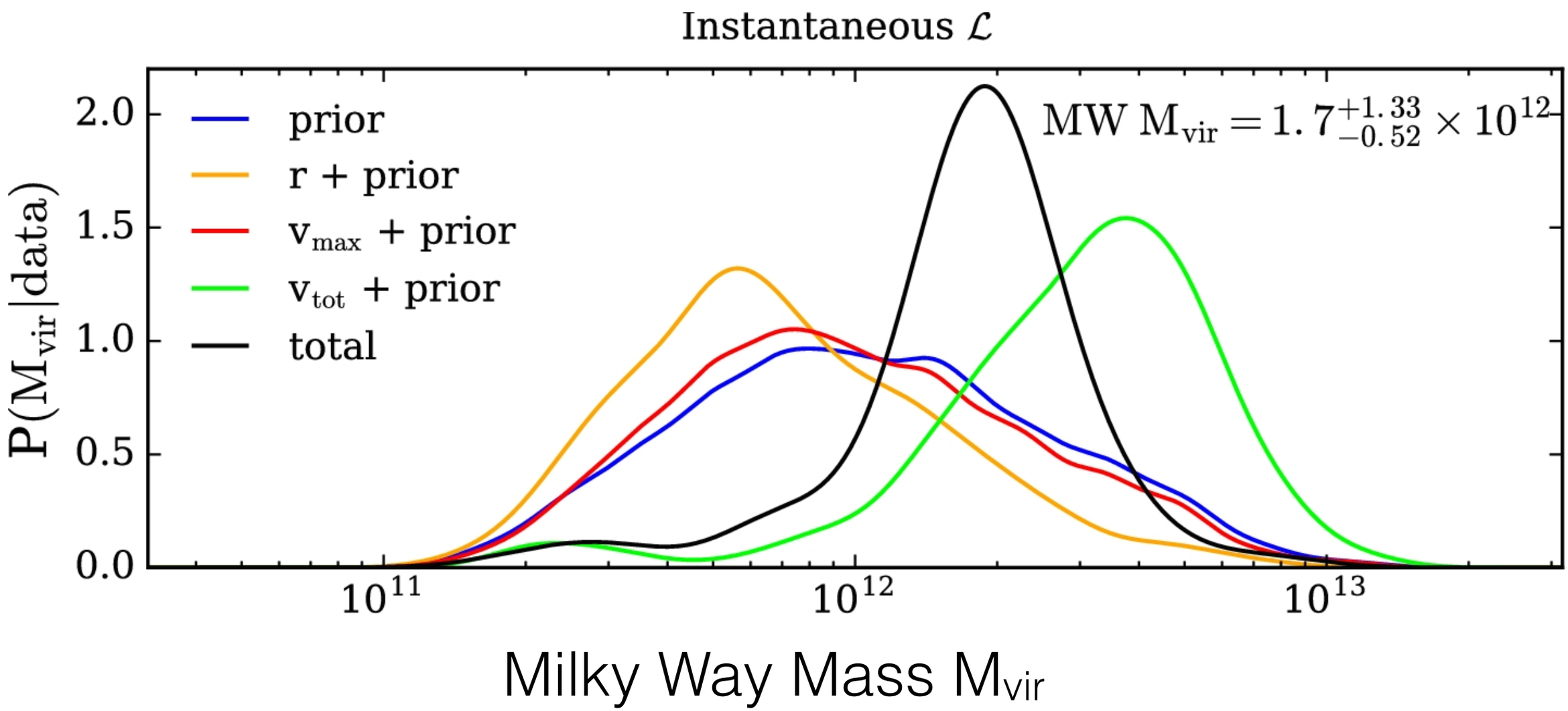
Use effective sample size (ESS) for **n**

if equal weights: $w_i = 1/m$, reduces to

$$\mathrm{kde}(\theta) = \sum_{s=1}^{m} \frac{1}{m} \times N(\theta \,|\, \theta_s, \mathrm{bw}^2)$$

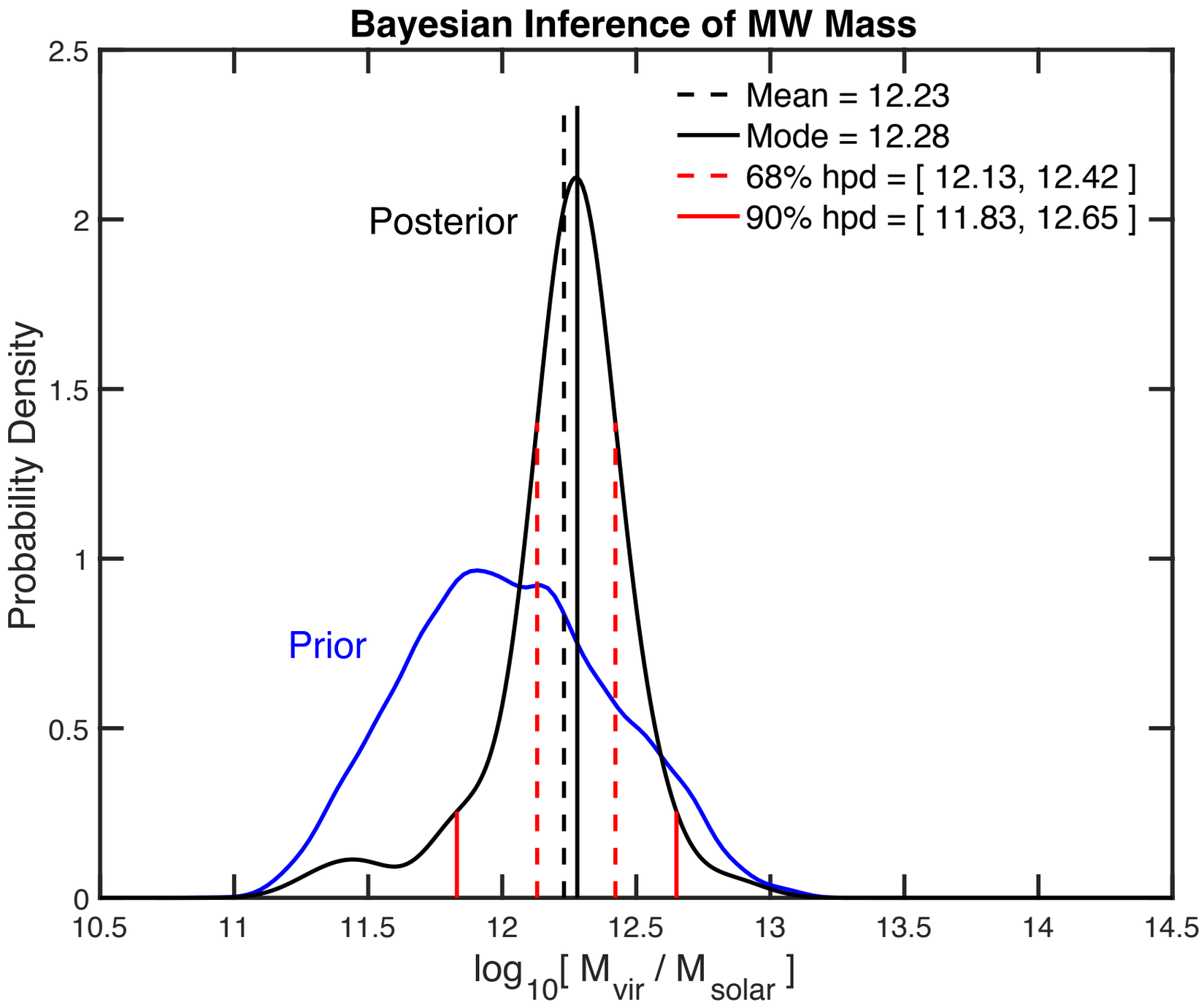
# Bayesian estimates of the Milky Way and Andromeda masses using high-precision astrometry and cosmological simulations (Patel, Belsa & Mandel 2017)

## Posterior of Milky Way Galaxy Mass with weighted KDE



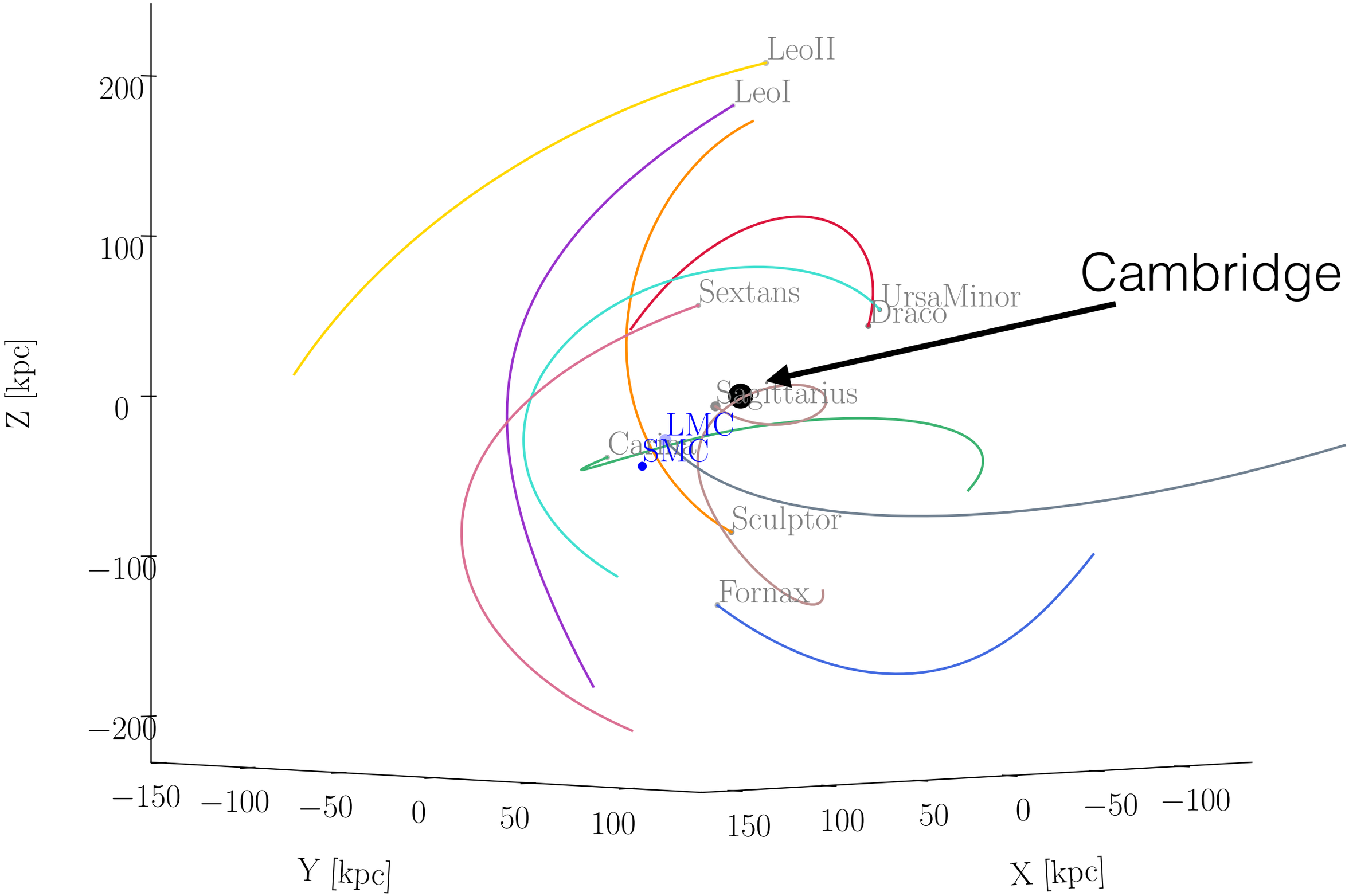

Using only the biggest satellite (Large Magellanic Cloud)

# Highest Posterior Density Intervals

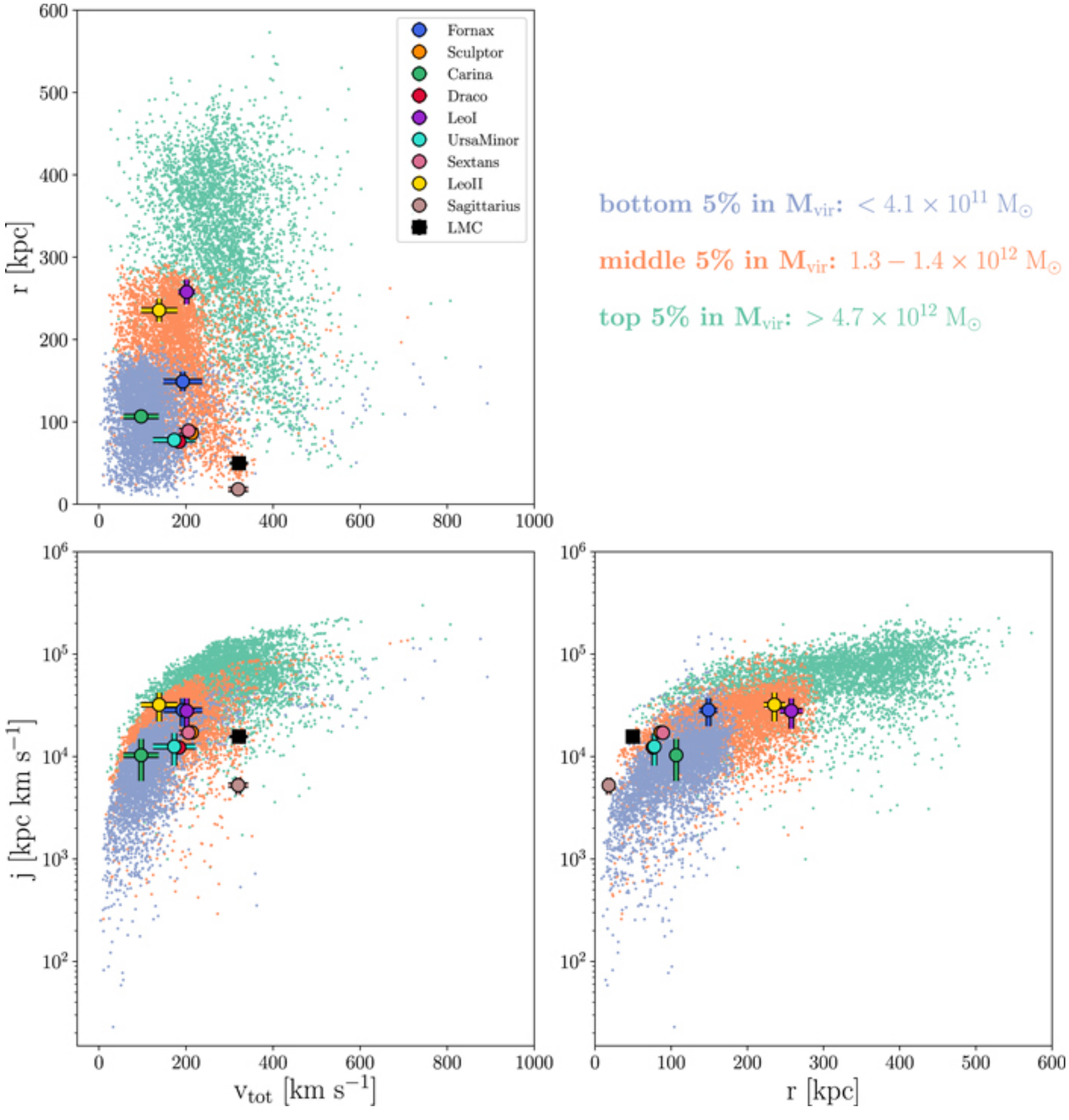



x% HPD = Highest Posterior Density x% credible region
= interval(s) with highest density containing x% of posterior

# There are many dwarf satellite galaxies



# Prior Distribution of dynamical properties with central galaxy mass



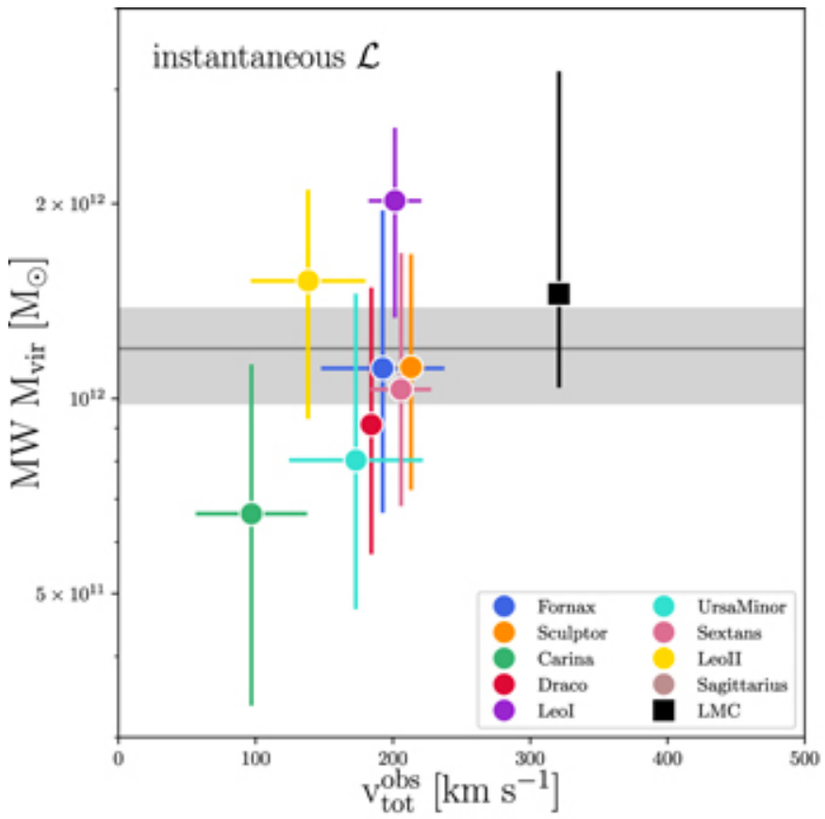

**bottom 5% in $M_{vir}$:** $< 4.1 \times 10^{11}\ M_\odot$

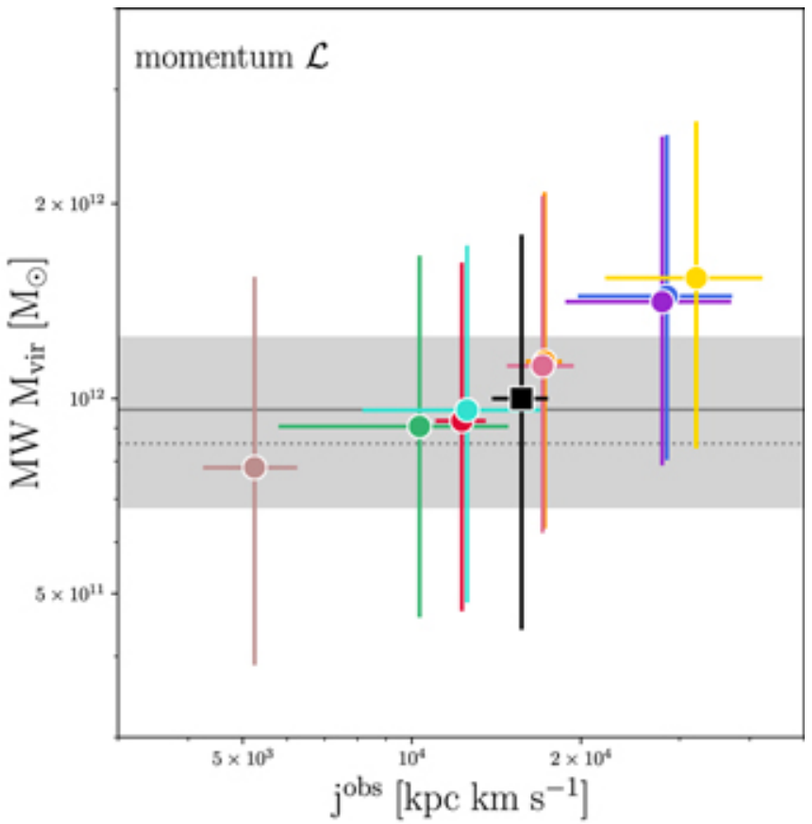
**middle 5% in $M_{vir}$:** $1.3 - 1.4 \times 10^{12}\ M_\odot$

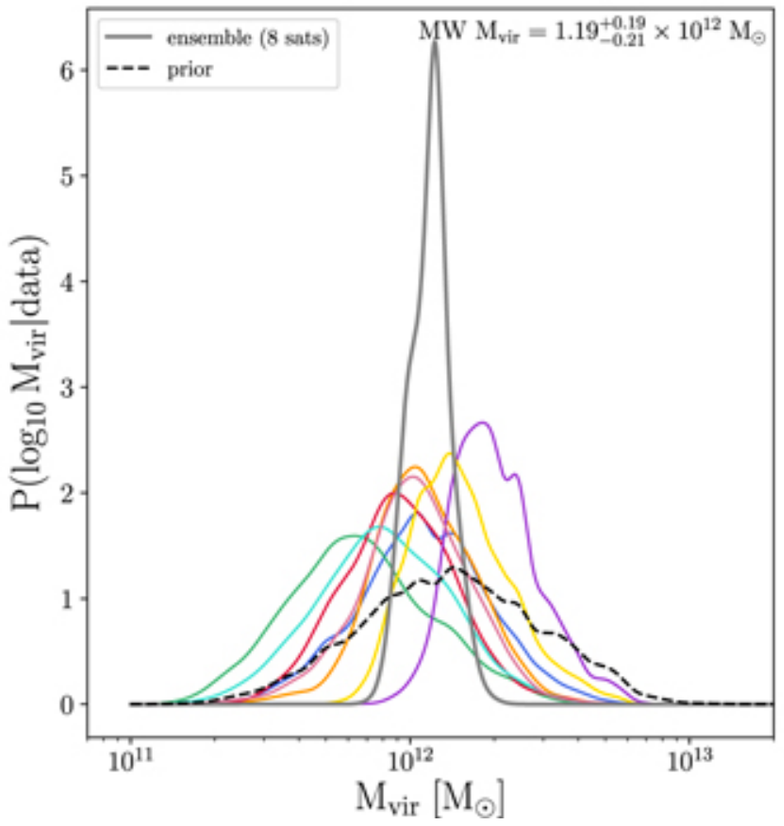
**top 5% in $M_{vir}$:** $> 4.7 \times 10^{12}\ M_\odot$

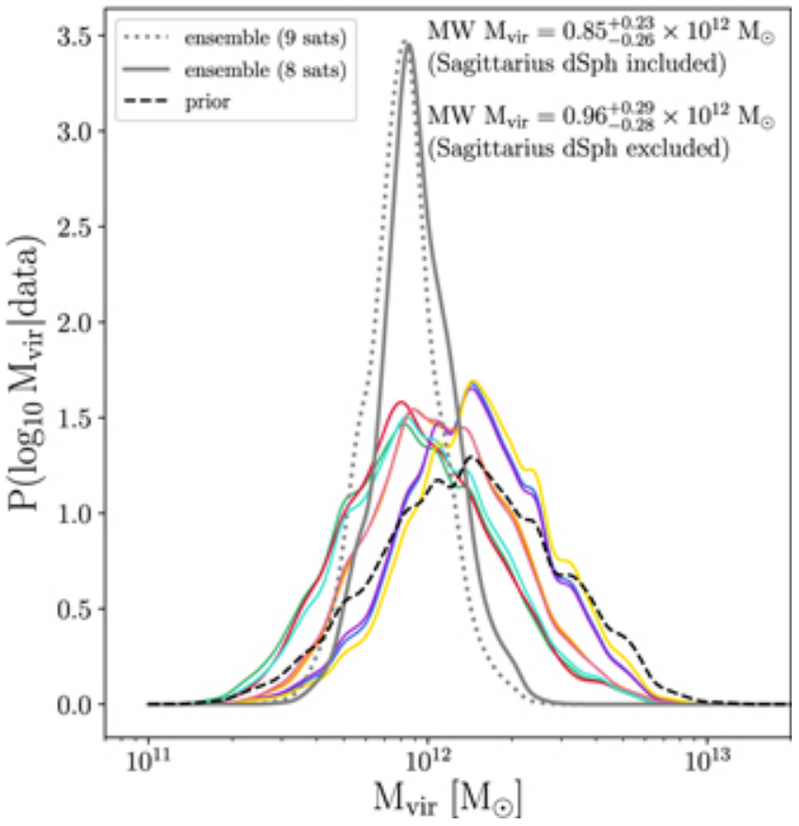
# Combined Posterior of Milky Way Mass

Patel, Belsa, Mandel & Sohn 2018



# Markov Chain Monte Carlo (MCMC)

# Evaluating the Posterior P($\theta$ |D)

- Simple likelihoods/conjugate priors admit analytic solutions to the posterior

- Simple models may allow direct draws: $\theta_i \sim P(\theta | D)$
  i.e. “Direct simulation”

- Small numbers of parameters (p): Evaluate posterior on a p-dimensional grid.  Wouldn’t recommend for p > 3).

- Realistic models with many parameters p:
  Markov Chain Monte Carlo is workhorse

# Monte Carlo Integration

Typically, we want to compute expectations of the form:

$$\mathbb{E}[f(\boldsymbol{\theta}|\, D)] = \int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} \approx \frac{1}{K} \sum_{i=1}^{K} f(\boldsymbol{\theta}_i)$$

Using m samples from the posterior:

$\theta_i \sim P(\theta \mid D)$

# What if you can't directly sample the posterior: $\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta} | D)$?

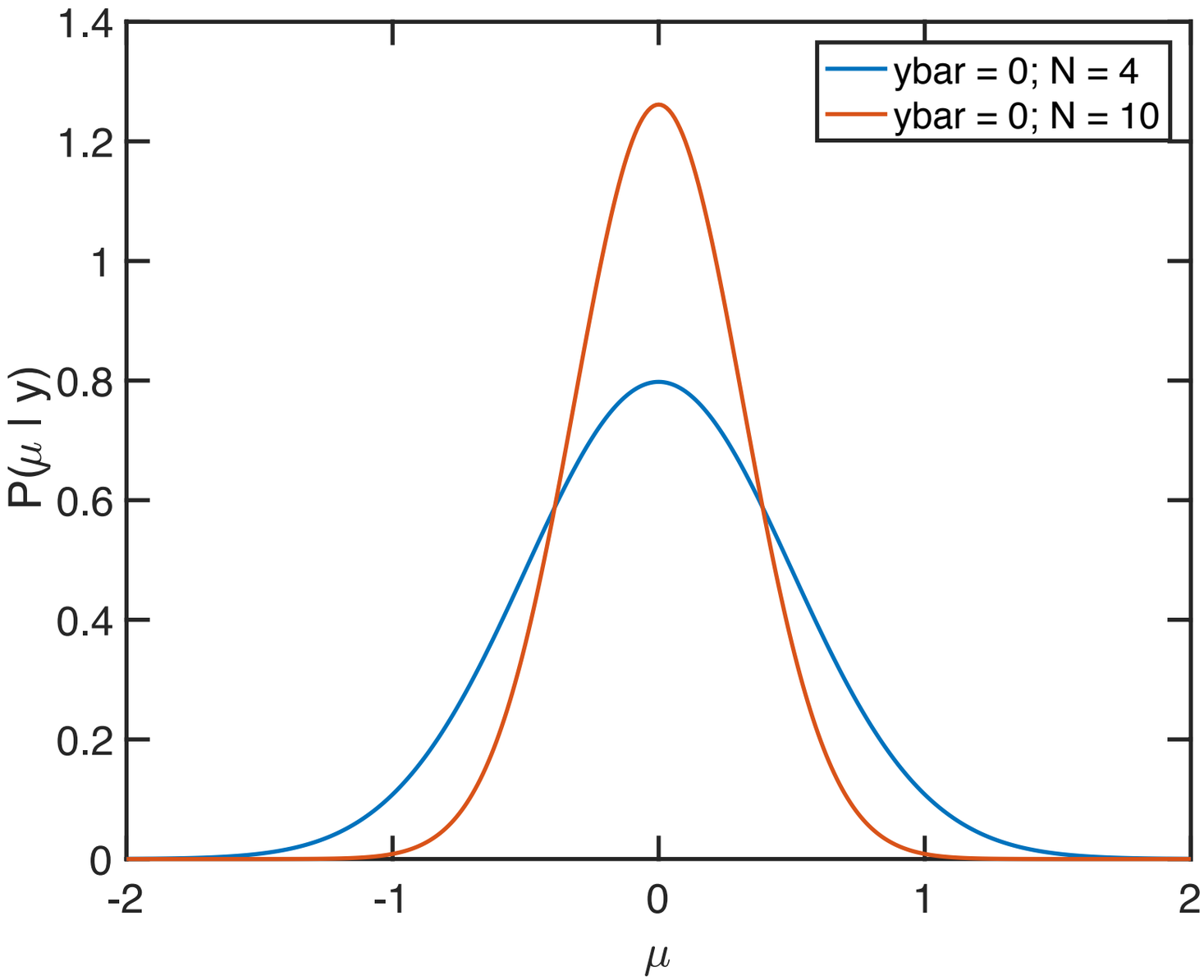
$$\mathbb{E}[f(\boldsymbol{\theta})|\, D] = \int f(\boldsymbol{\theta}) P(\boldsymbol{\theta}|\, D)\, d\boldsymbol{\theta} \approx \frac{1}{m} \sum_{i=1}^{m} f(\boldsymbol{\theta}_i)$$

- Posterior simulation - Markov Chain Monte Carlo:
  - Generate a correlated sequence (chain) of random variates (Monte Carlo) that (in a limit) are draws from the posterior distribution. The next value in the sequence only depends on the current values (Markov)
  - Algorithm cleverly constructed to ensure distribution of chain values —> posterior dist'n = stationary dist'n in the long-run (explain how next week).

# Simple Gaussian mean μ (where we know the answer)

$$y_i \sim N(\mu, \sigma^2 = 1), i = 1 \ldots N$$

$$P(\mu) \propto 1 \qquad P(\mu|\,\boldsymbol{y}) = N(\mu|\,\bar{y}, \sigma^2/N)$$



# Simplest MCMC: Metropolis Algorithm

1. Choose a random starting point $\mu_0$

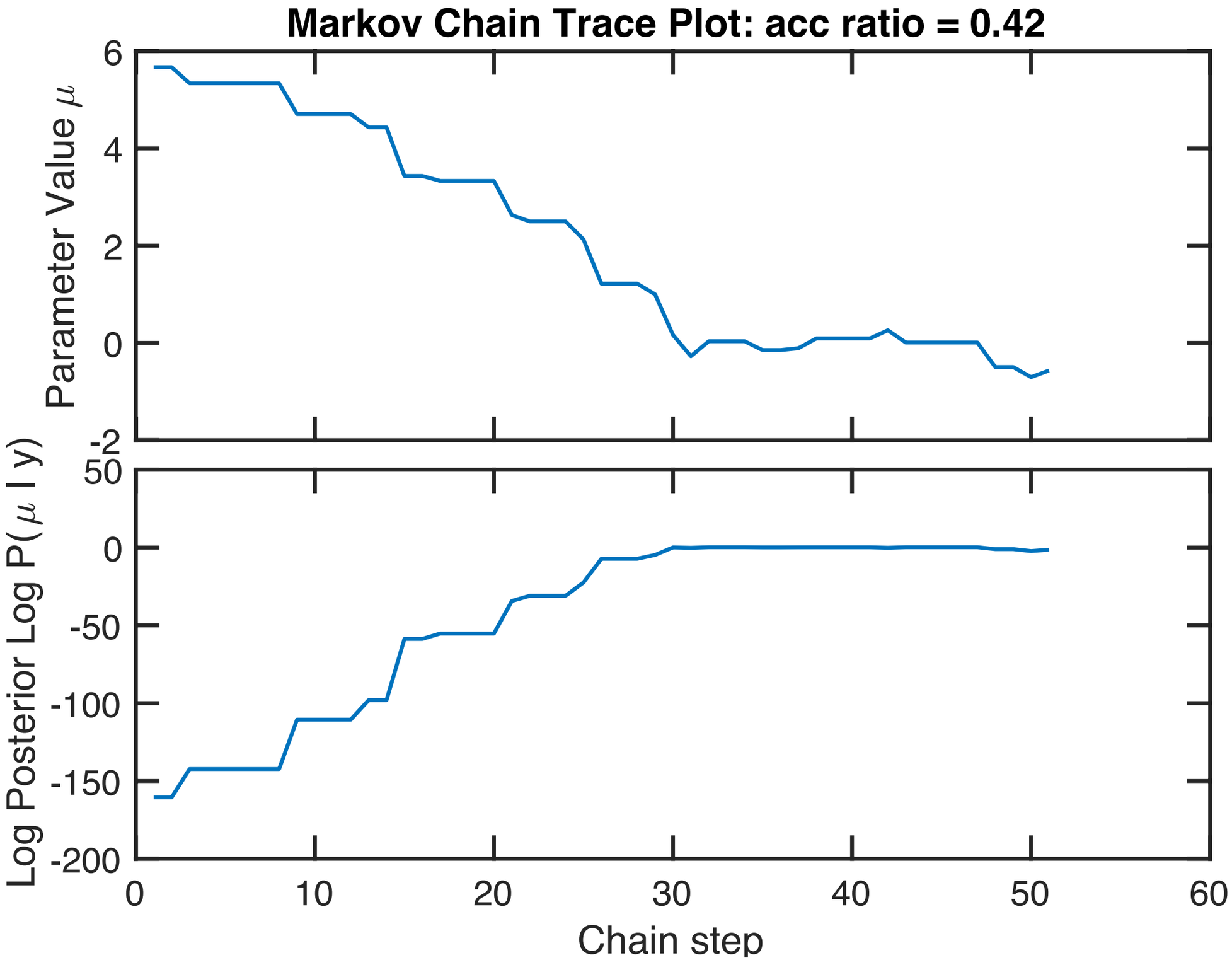
2. At step i = 1…N, propose a new parameter value $\mu_{prop} \sim N(\mu_i, \boldsymbol{\tau}^2)$. The proposal scale $\boldsymbol{\tau}$ is chosen cleverly.

3. Evaluate ratio of posteriors at proposed vs current values. Metropolis Ratio $r = P(\mu_{prop} \mid \mathbf{y}) / P(\mu_i \mid \mathbf{y})$.

4. If $\mu_{prop}$ is a better solution (higher posterior), $r > 1$, accept the new value $\mu_{i+1} = \mu_{prop}$. Else accept with probability r (i.e. accept with probability min(r,1). **[If not accept, stay at same value $\mu_{i+1} = \mu_i$ & include in chain]**.

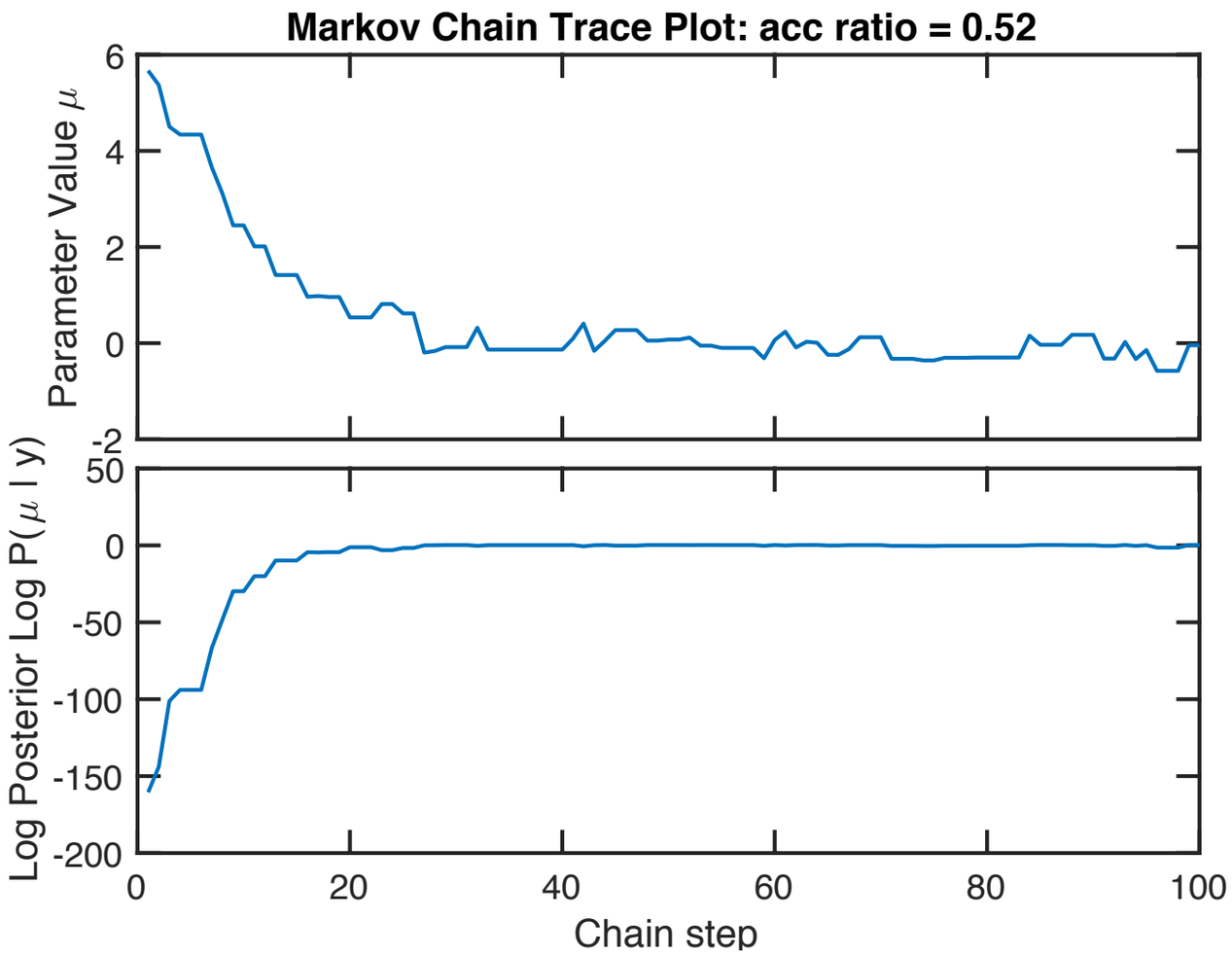
5. Repeat steps 2-4 until reach some measure of convergence and gather enough samples to compute your inference

# Code demo: metropolis1.m
# First 50 iterations

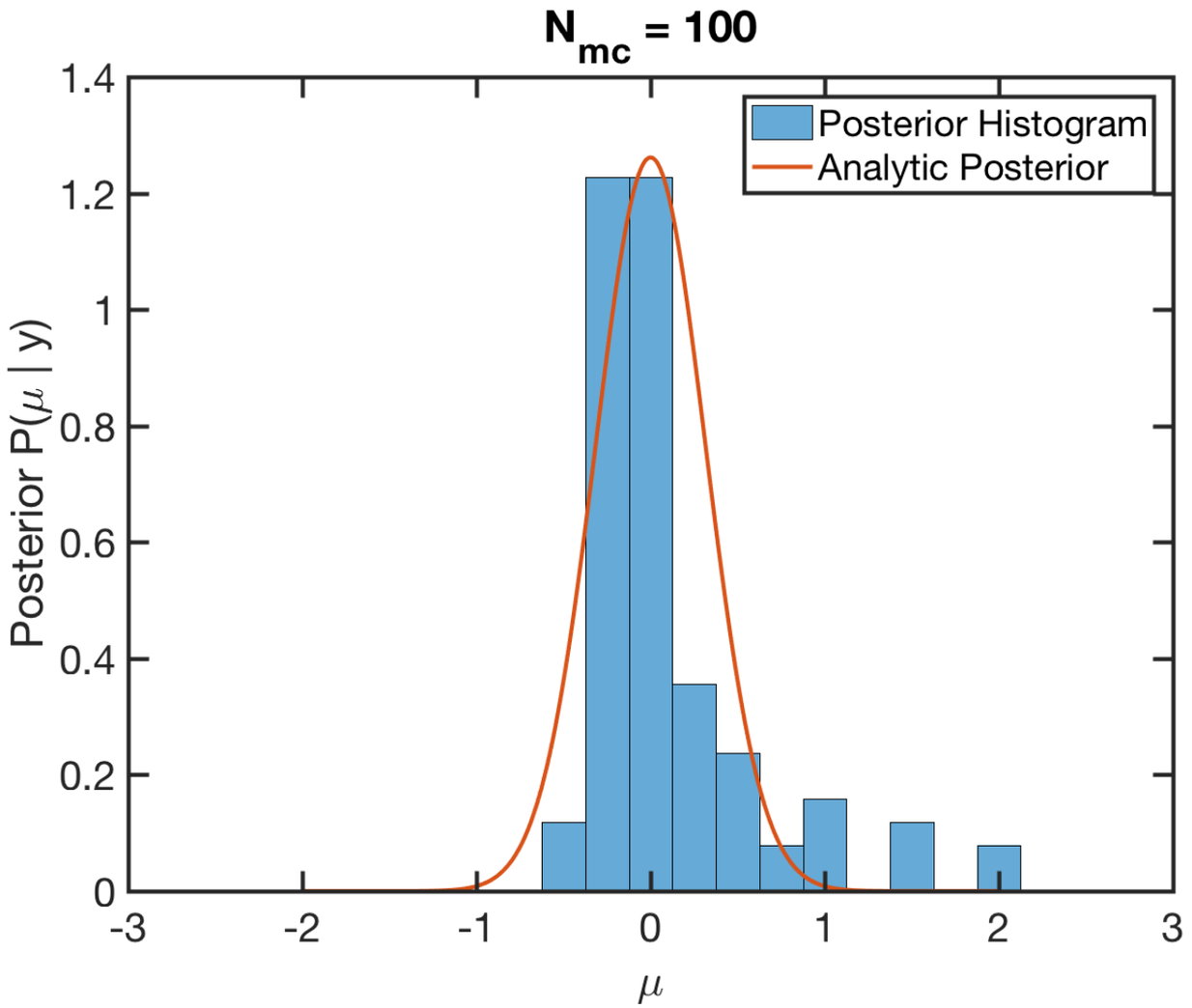


# Code demo: metropolis1.m

## First 100 iterations



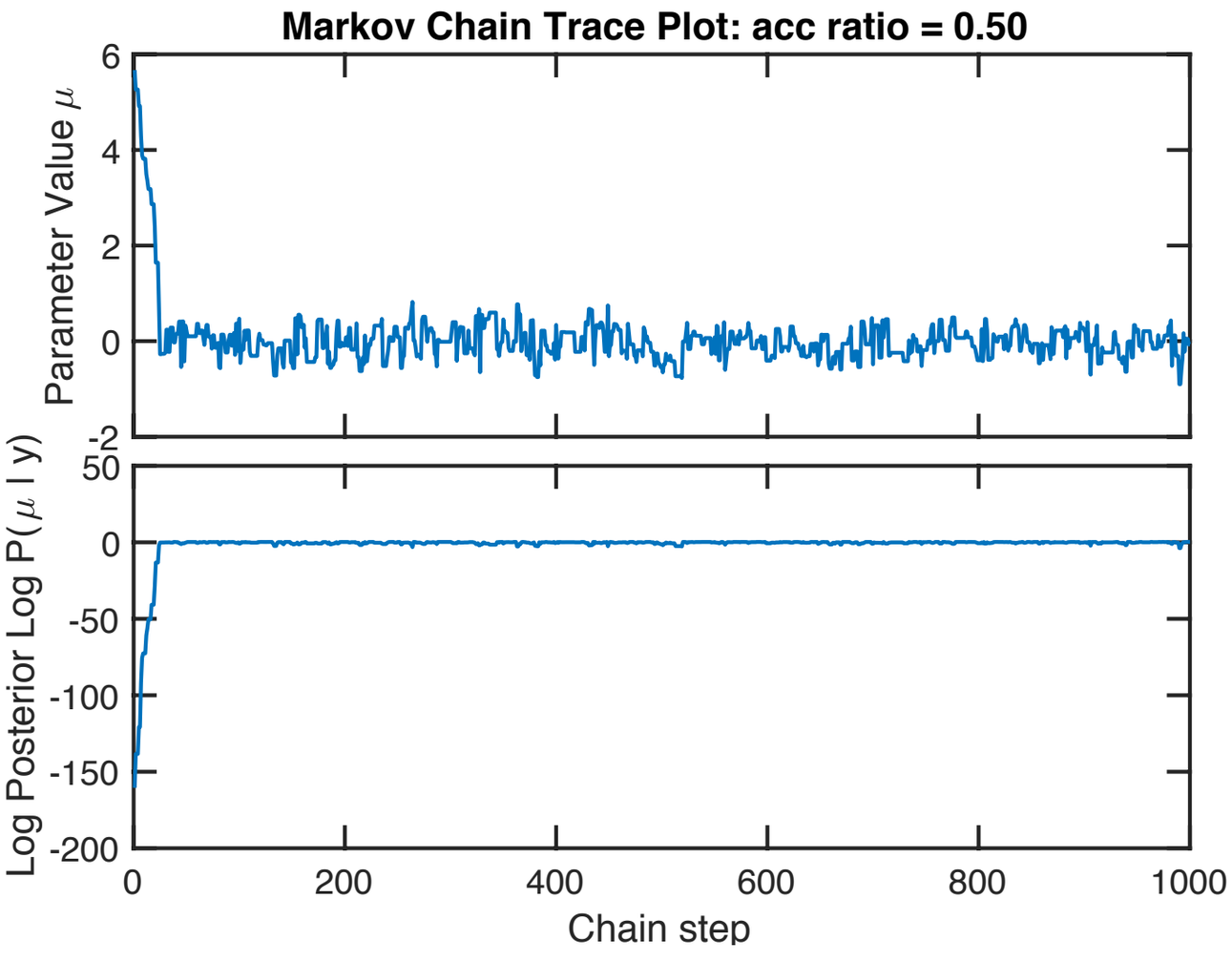

Trace Plot



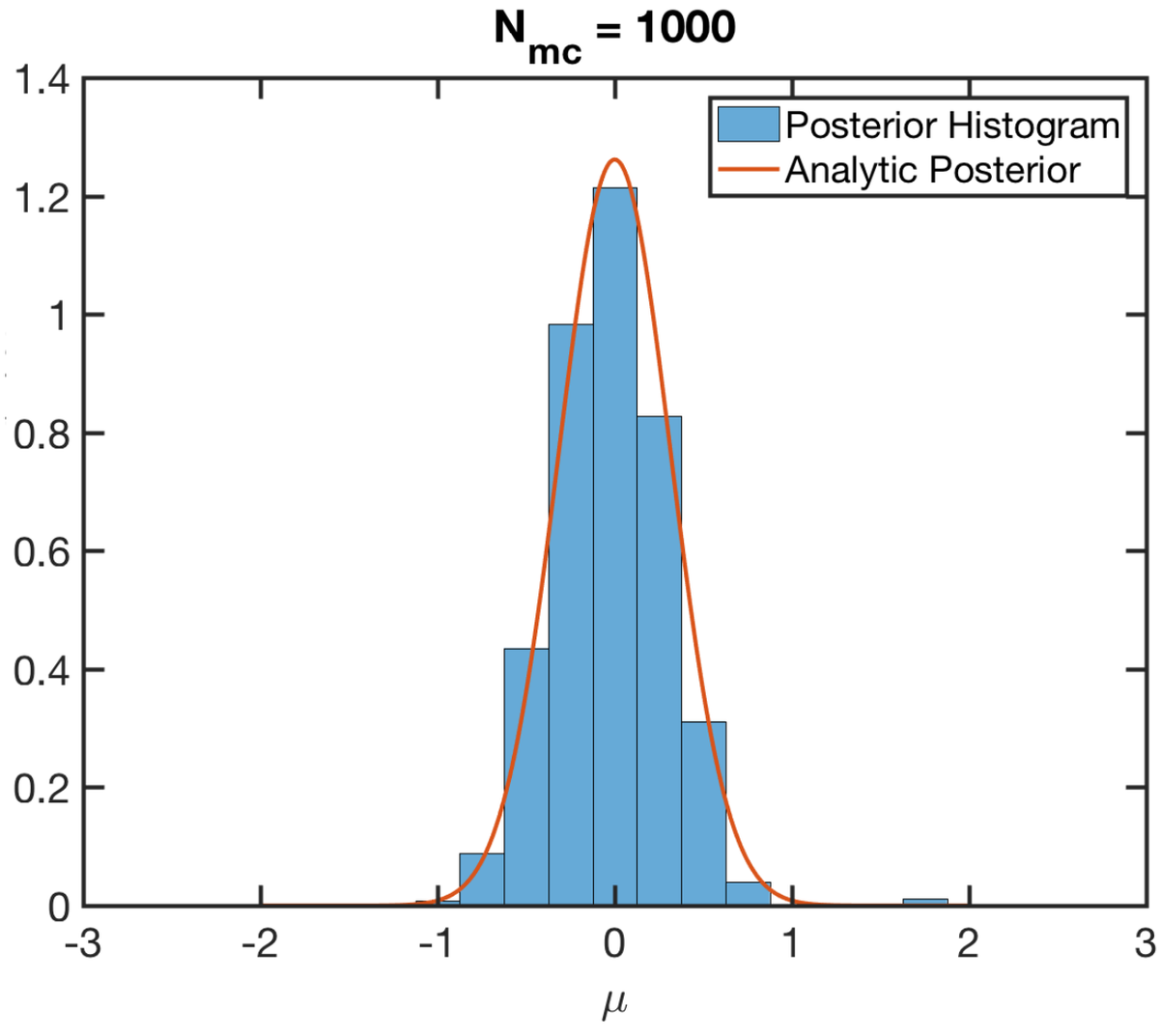

Posterior

# Code demo: metropolis1.m

## First 1000 iterations



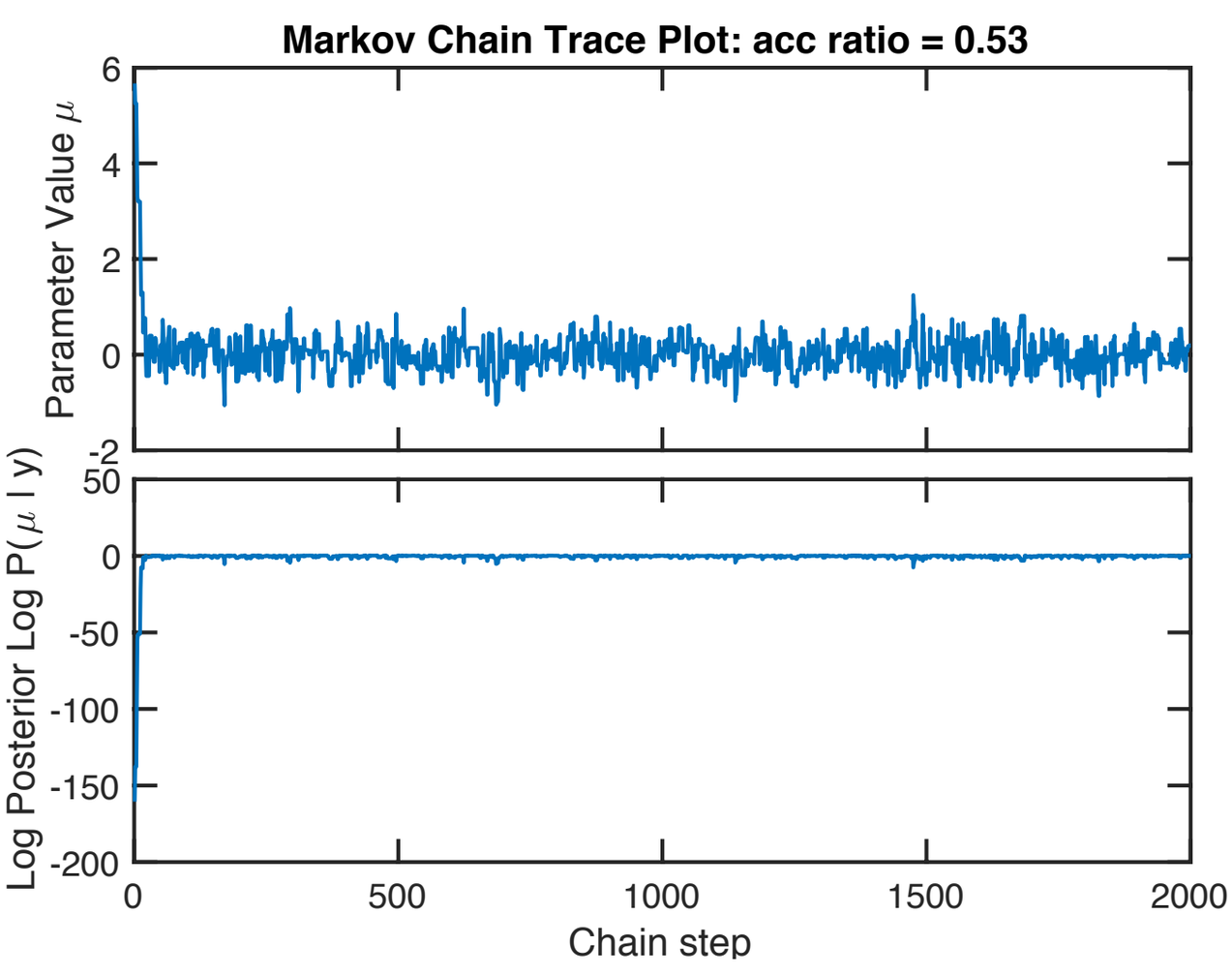

Trace Plot



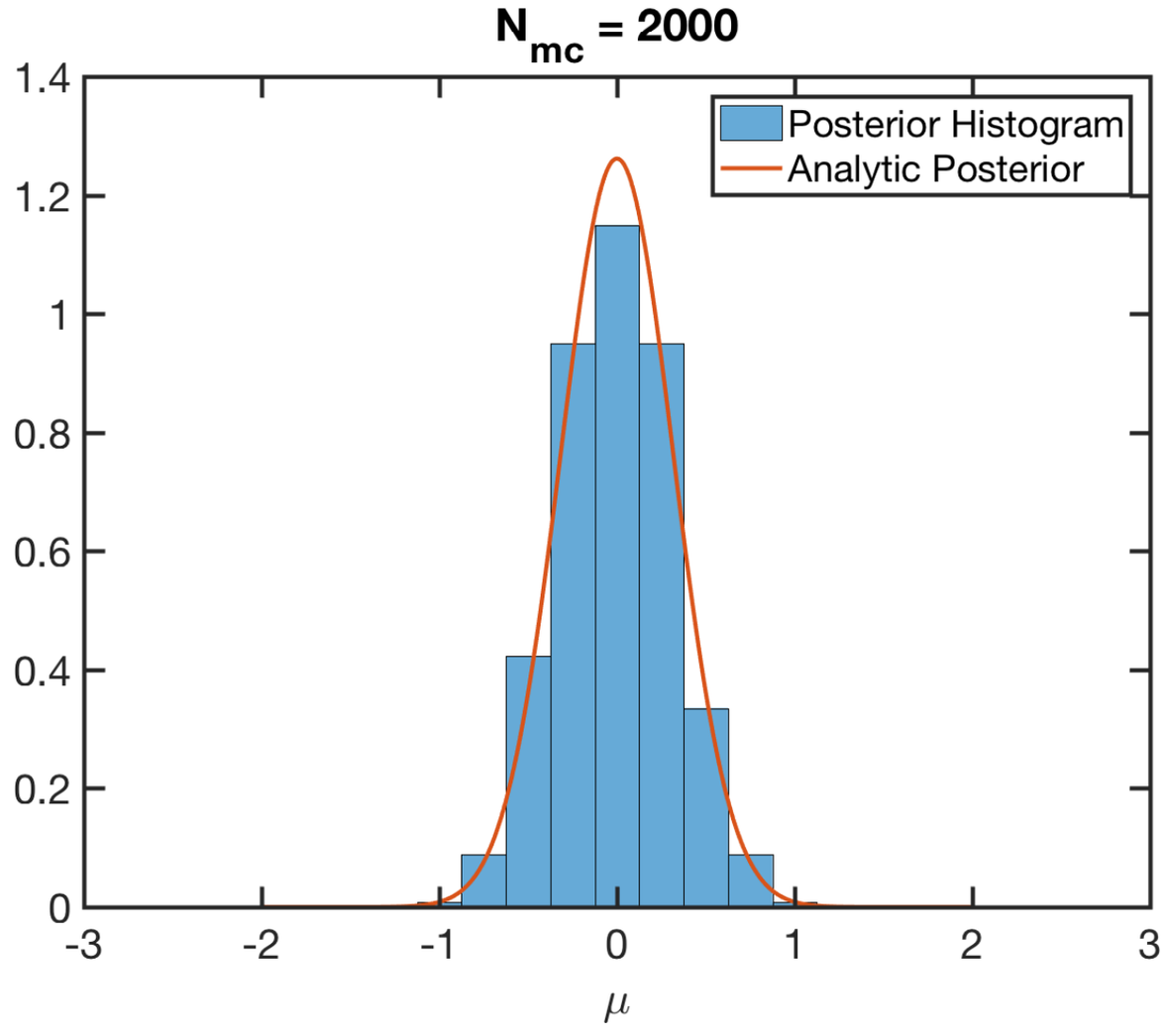

Posterior

# Code demo: metropolis1.m

# 2000 iterations




Trace Plot




Posterior histogram of
500 samples
after cutting 50% burn-in
& thinning by 2