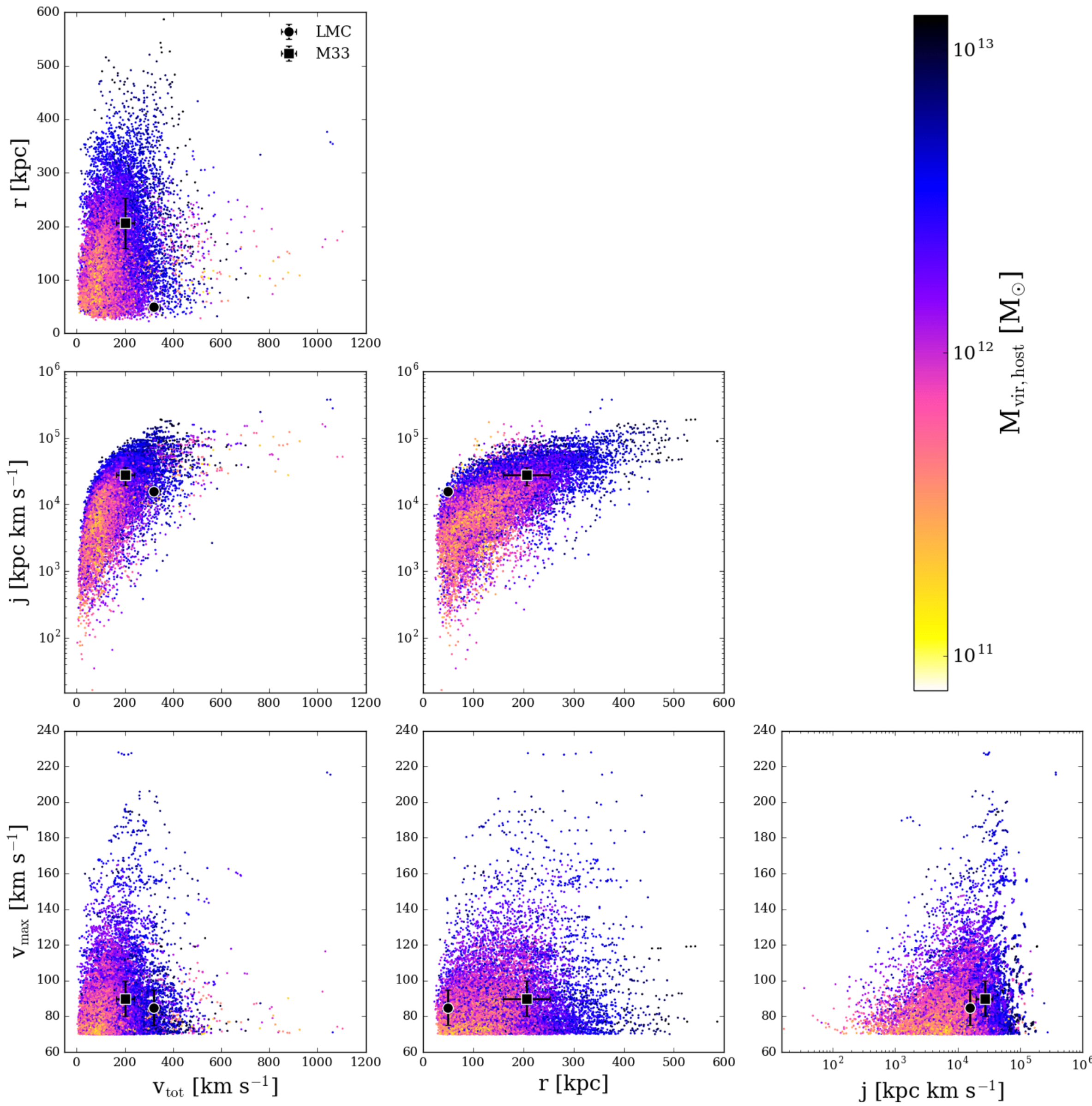
Astrostatistics: Friday 13 Feb 2019

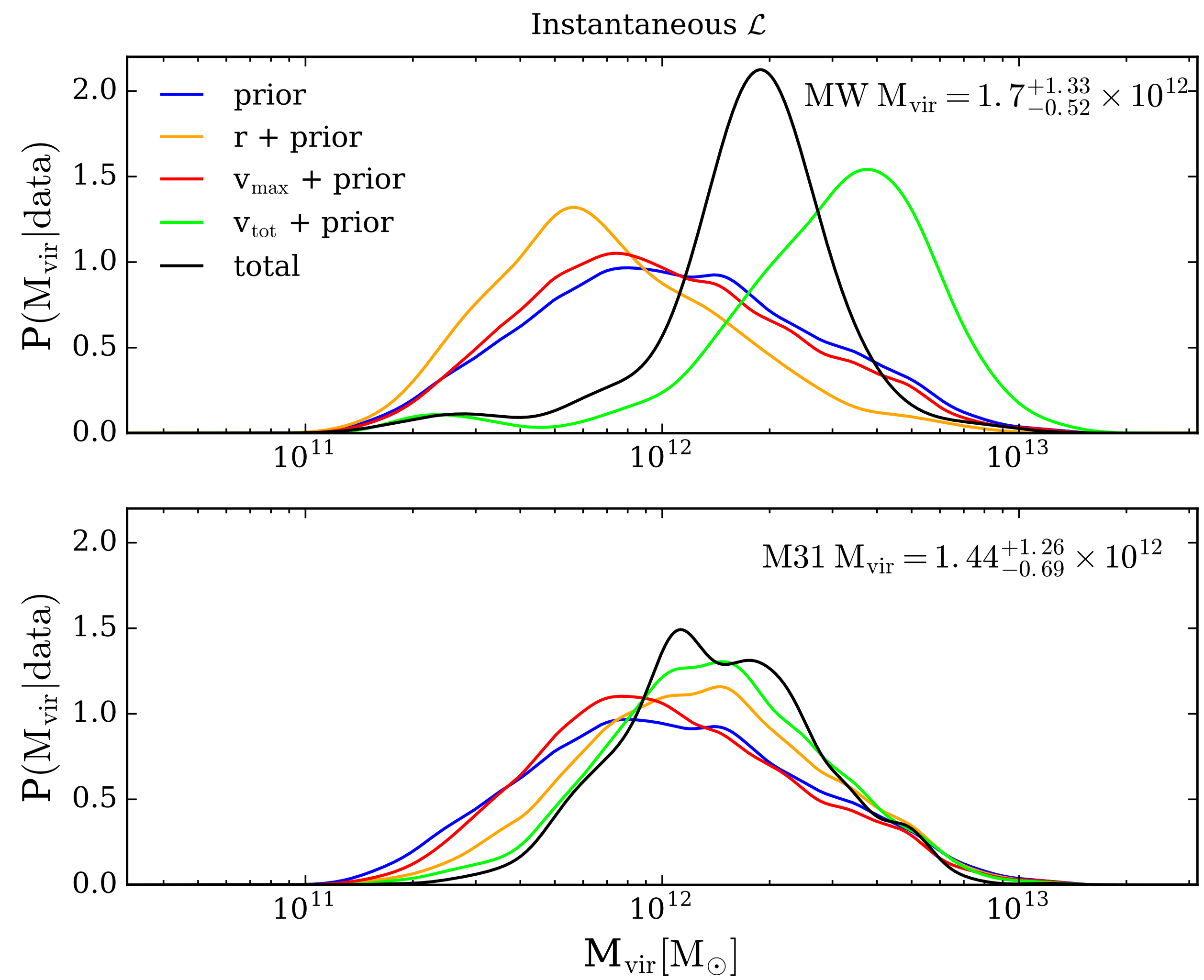
https://github.com/CambridgeAstroStat/PartIII-Astrostatistics-2019

- Example Sheet online, Ex Class Tue Feb 19, 1pm MR5
- Today: continue Bayesian computation / Monte Carlo Methods / MCMC
- MacKay: Ch 29-30; Bishop: Ch 11; Gelman
- Givens & Hoeting "Computational Statistics"
  (Free download through Cambridge Library iDiscover)
- Hogg & DFM, 2017 "Data analysis recipes: Using Markov Chain Monte Carlo." https://arxiv.org/abs/1710.06068

Review Astrostatistics Case Study:
Importance Weighing the Milky Way Galaxy
(Patel et al. 2017, 2018)

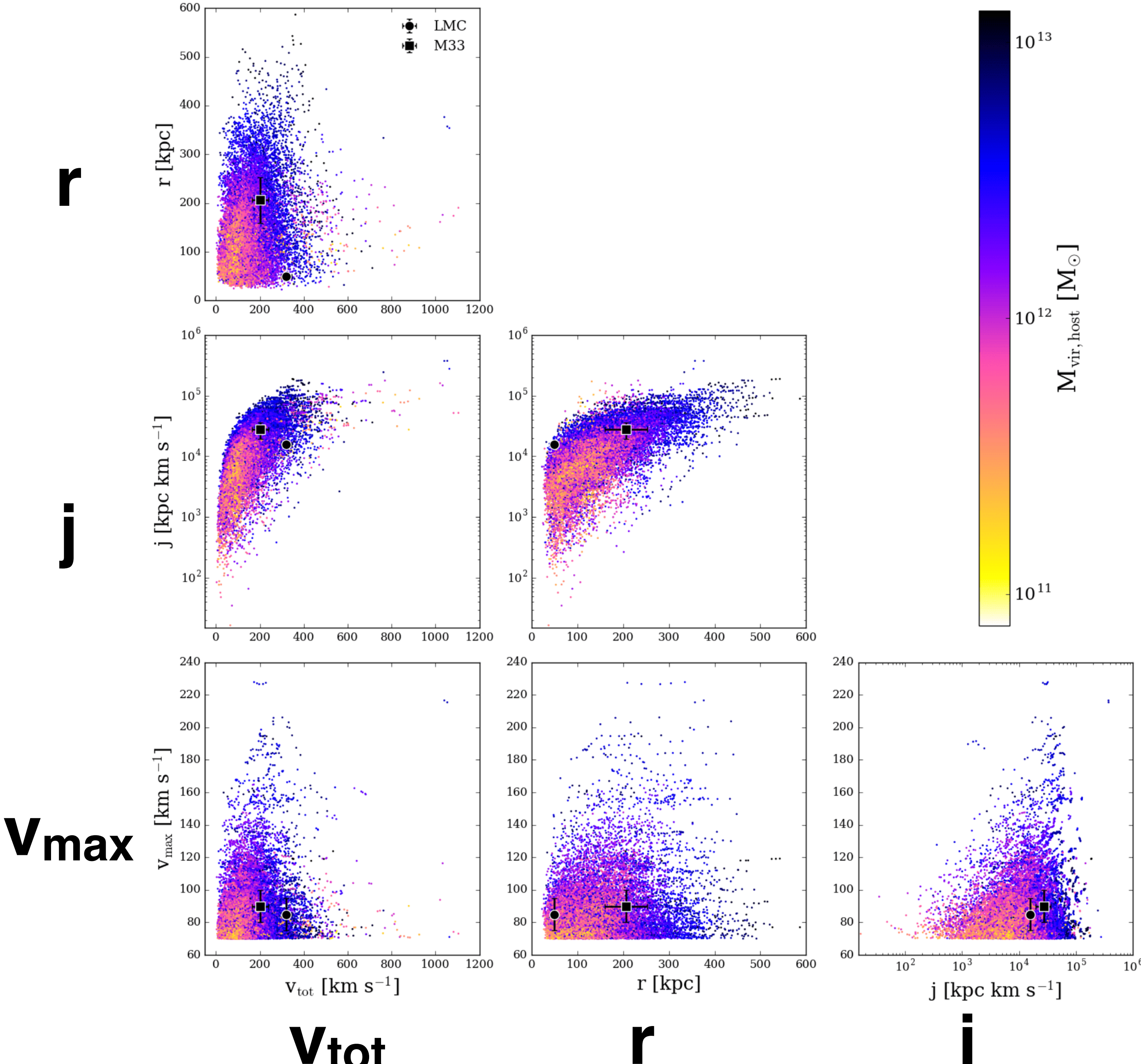


Simulation —> Prior



Data + Importance Sampling —> Posterior

Velocities ($v_{max,}$ v_{tot}), positions (r), momenta (j), of satellites are correlated with mass via galaxy formation physics in simulations (Prior)



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

M_{vir} = Mass of Galaxy

Parameters are:

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

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

Table 1. Observational data ($\boldsymbol{d}$) 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.

Data d =

	LMC μ	LMC σ	M33 μ	M33 σ
$v_{\rm max}^{\rm obs}$ (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.

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
x = latent (true) values
$\mathbf{M}_{\mathbf{vir}}$ = Mass of Galaxy

3.2.3 Importance sampling

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

(Ignore C)

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

Likelihood (observations)   Prior (samples from Simulation)

Importance Sampling

Parameters are: θ = (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:

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

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

The denominator of this equation is the normalization constant. If

Posterior Expectation of the Mass
(derive on board)

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

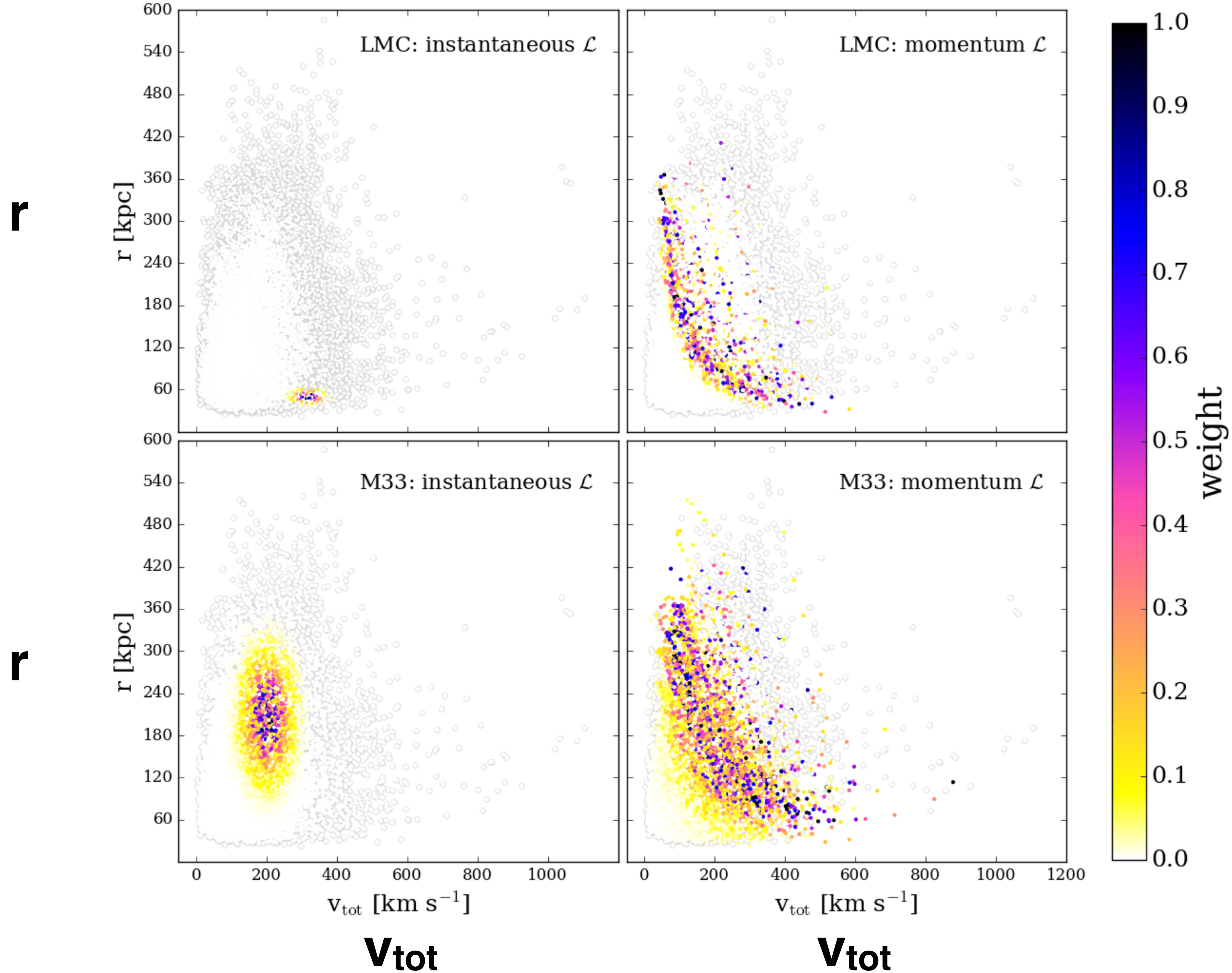
$$= \int f(M_{\rm vir})\, P(\boldsymbol{x}, M_{vir} | \boldsymbol{d}, \boldsymbol{C})\, {\rm d}\boldsymbol{x}\, {\rm d}M_{vir}$$

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

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

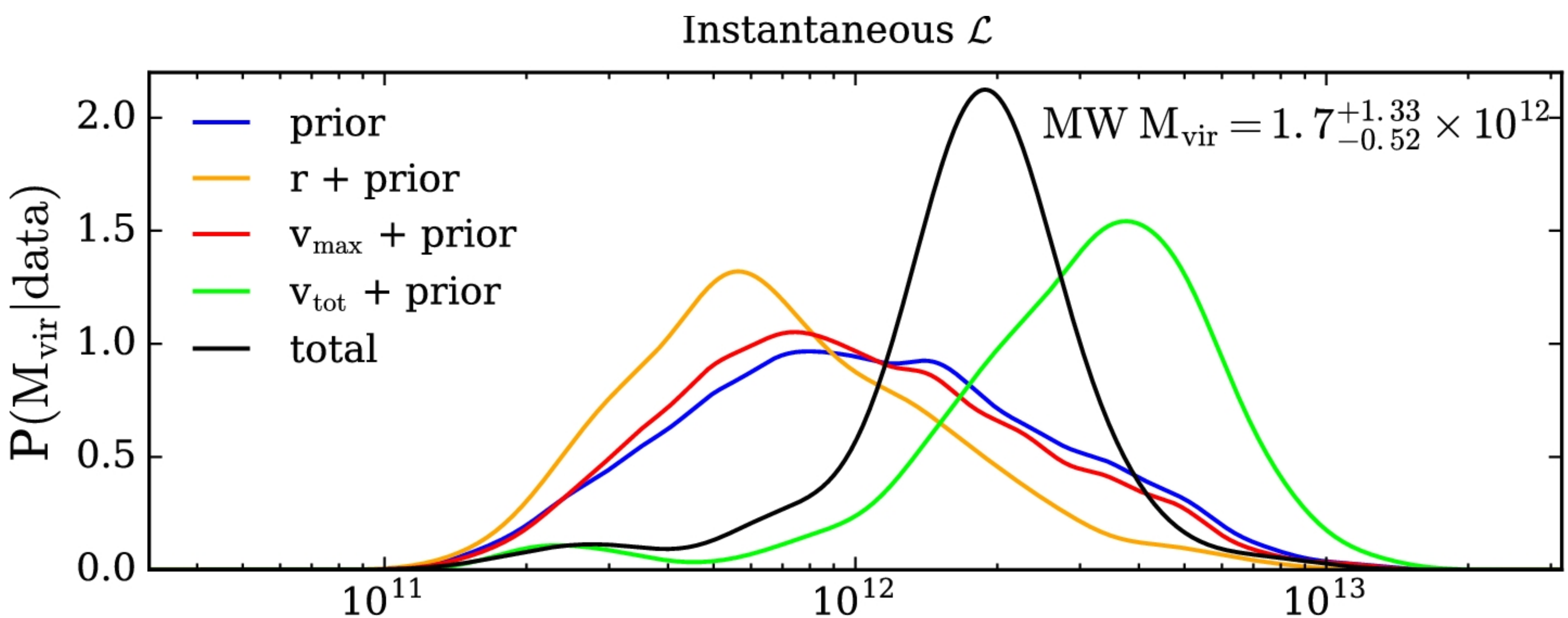
where $w_i = P(\boldsymbol{d} | \boldsymbol{x}_i) / \sum_{\rm j}^{\rm n} P(\boldsymbol{d} | \boldsymbol{x}_j)$ are importance weights.

Distribution of Importance Weights



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

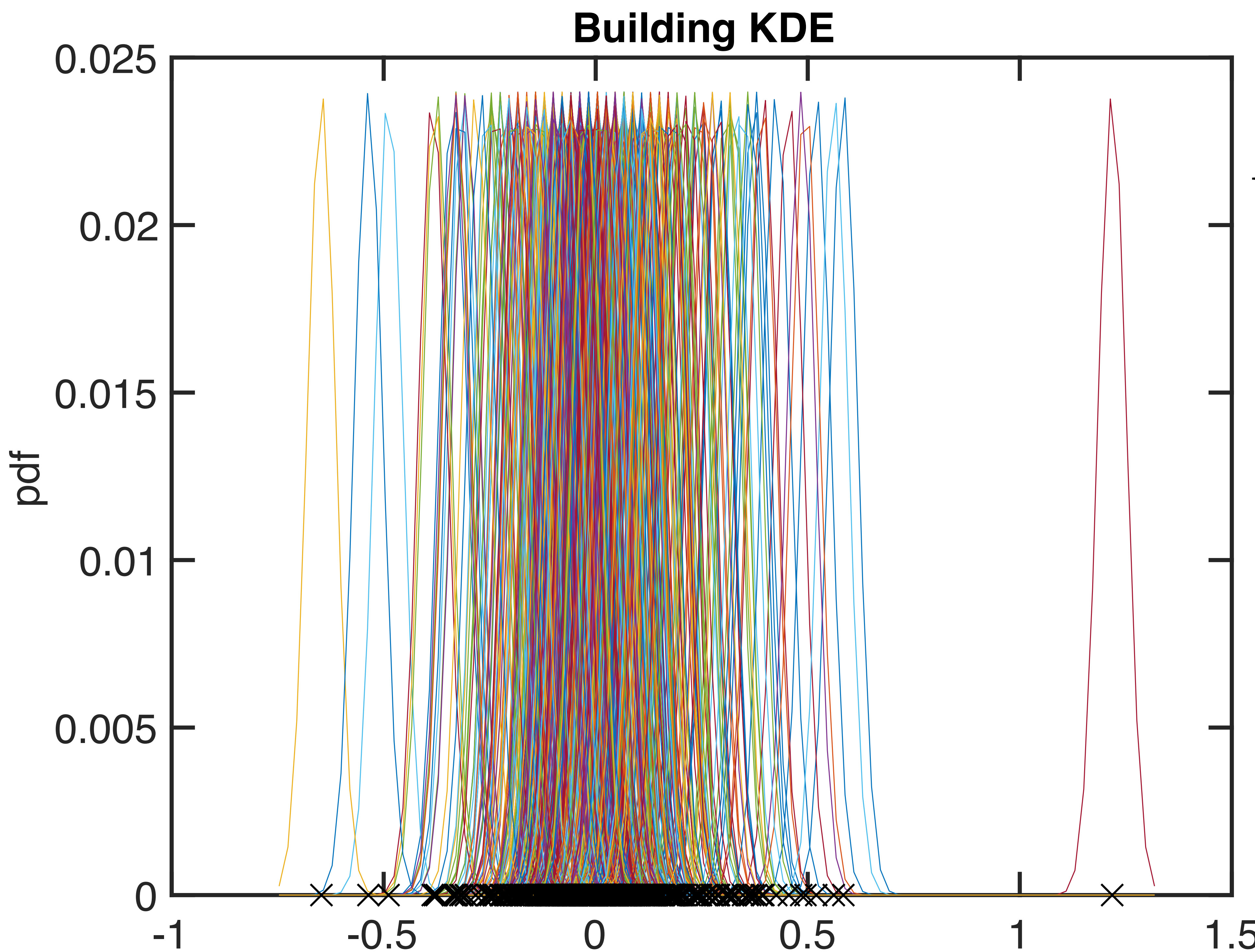
Posterior Density of Milky Way Galaxy Mass with KDE



Milky Way Mass M_{vir}

Kernel Density Estimation (KDE) (Smooth Histogram)

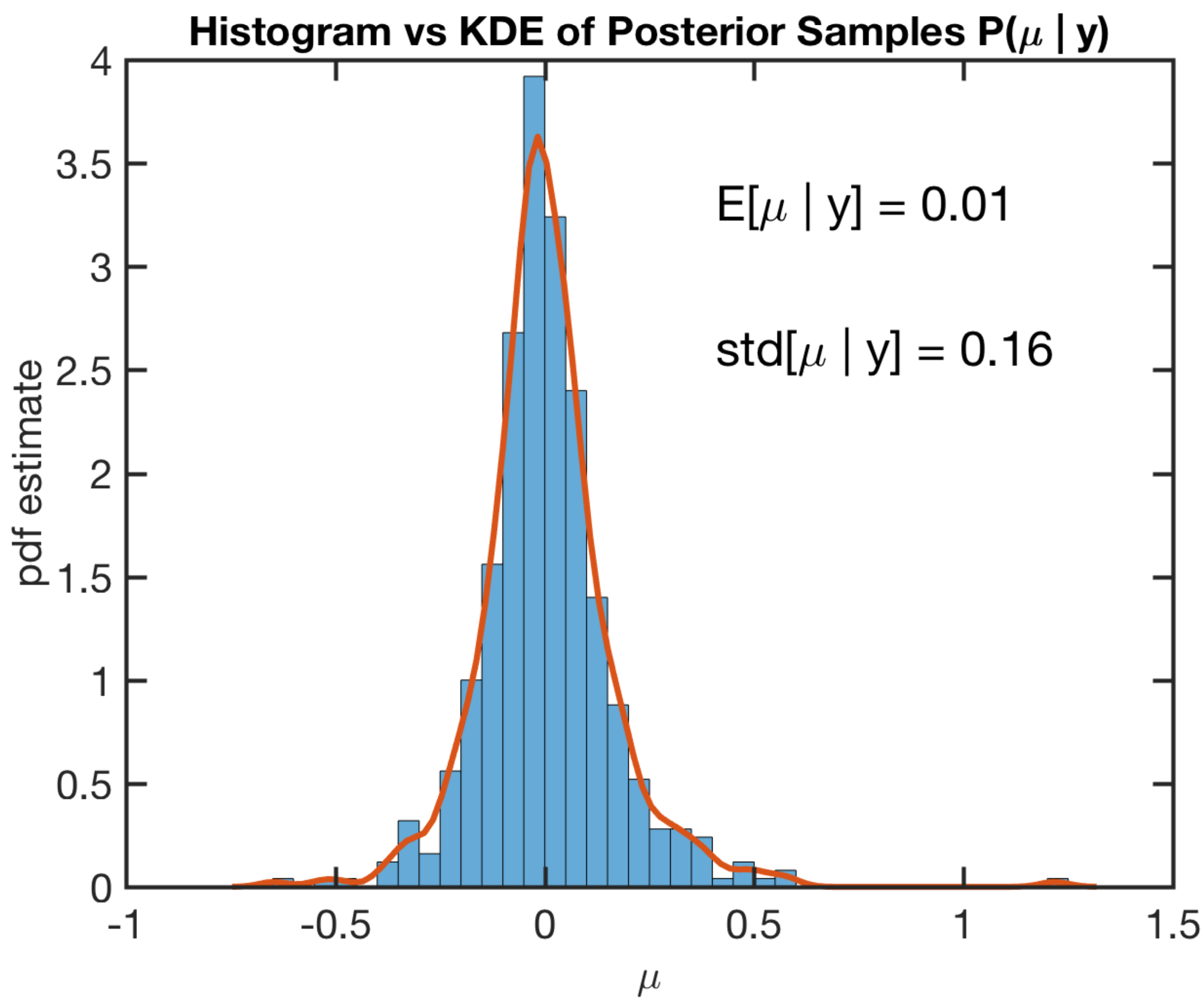
Each sample gets a Gaussian at the sample point with an "optimal" bandwidth bw (rule of thumb)



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

Kernel Density Estimation (KDE) (Smooth Histogram)

Then add them up and normalise pdf to 1



Weighted KDE

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

w_s = normalised importance weights

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

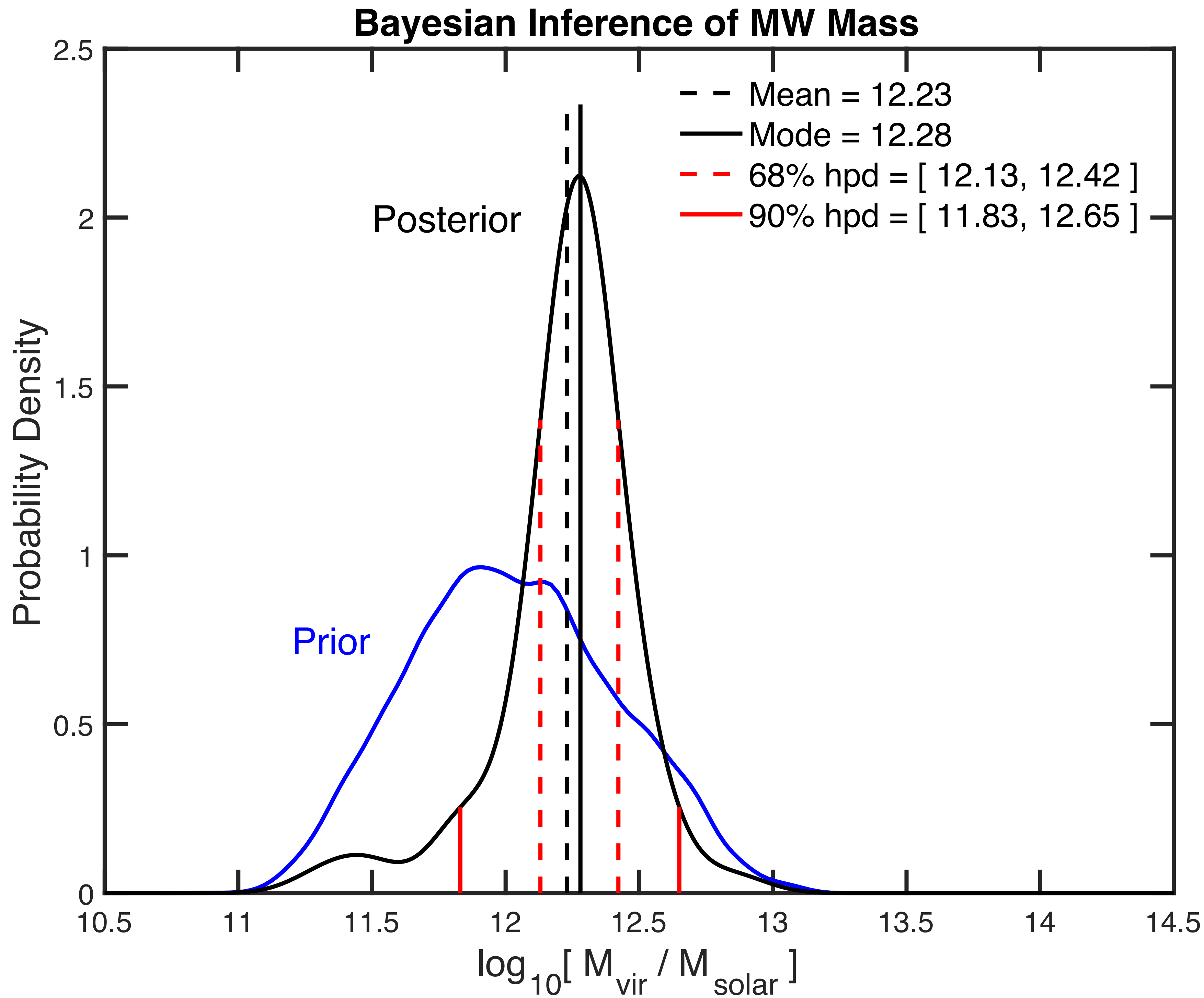
Estimate σ^2 from variance estimate

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

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

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

Highest Posterior Density Intervals



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

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: $\theta_i \sim P(\theta \,|D)$?

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

- Posterior simulation - Markov Chain Monte Carlo:

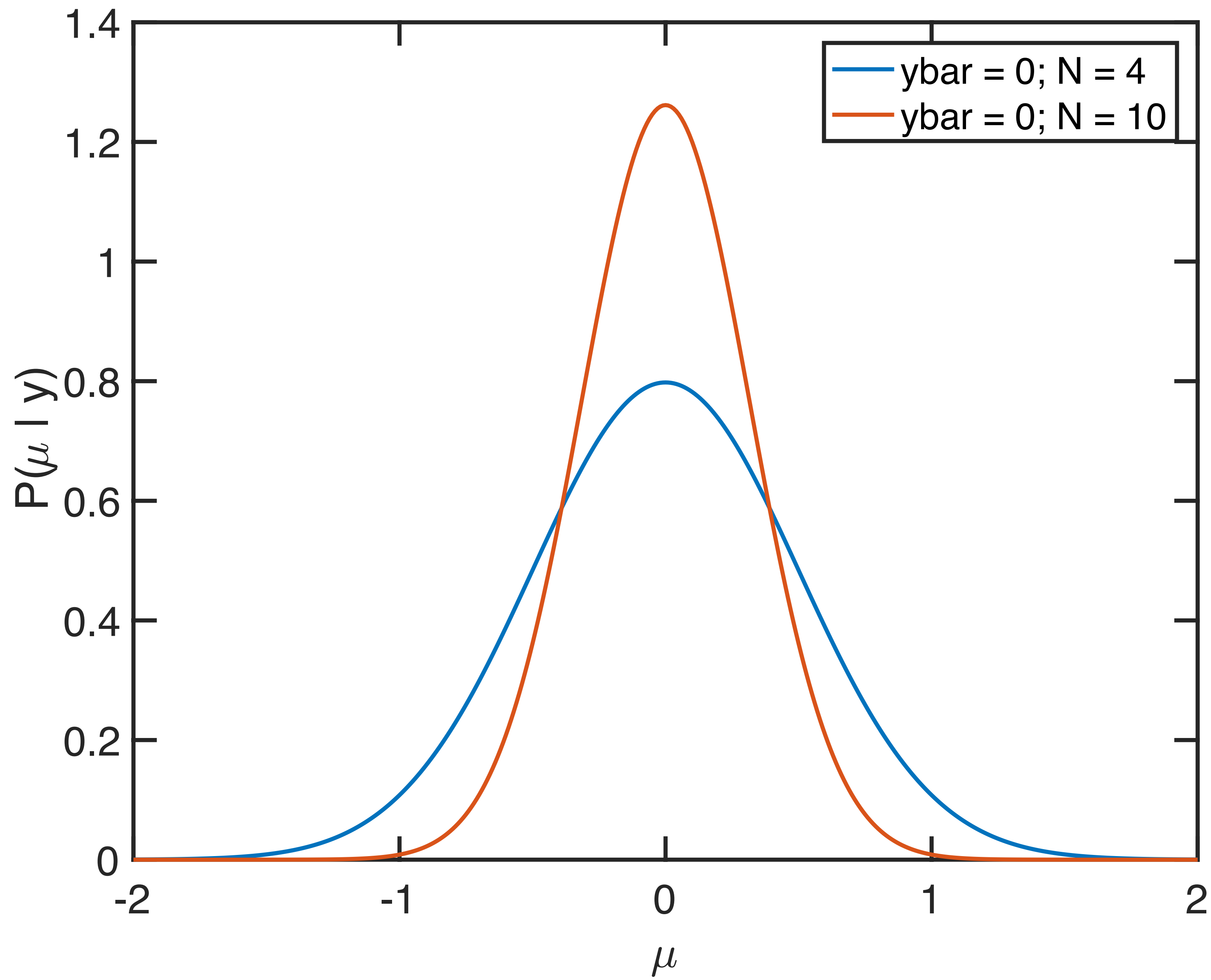
- Generate a correlated sequence (chain) of random variates (Monte Carlo) that (in a limit) are draws from the posterior distribution. The next value in the sequence only depends on the current values (Markov)

- 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 μ_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.

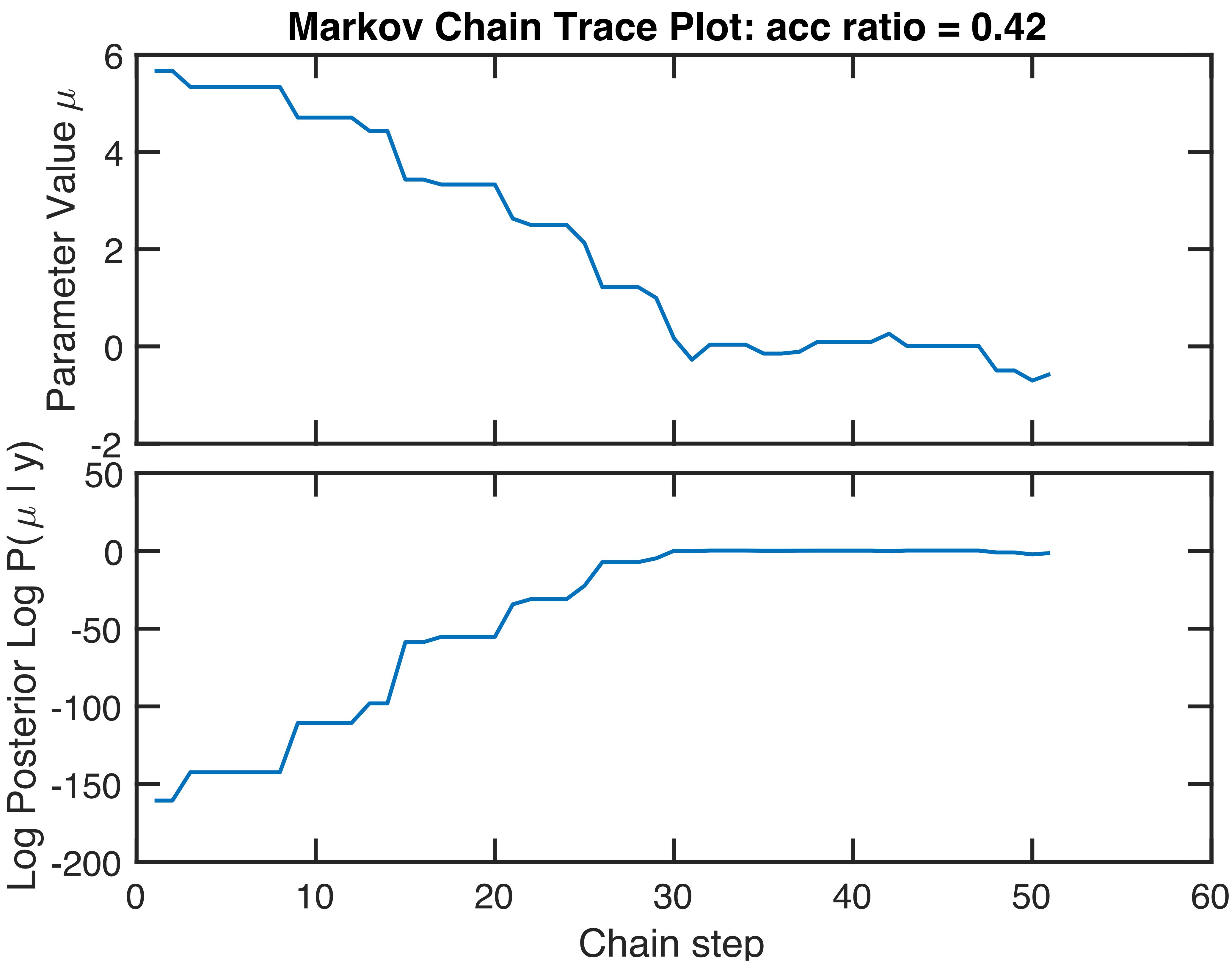
3. Evaluate ratio of posteriors at proposed vs current values. Metropolis Ratio $r = P(\mu_{prop} \mid \mathbf{y}) / P(\mu_i \mid \mathbf{y})$.

4. If μ_{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 $\boldsymbol{\mu_{i+1} = \mu_i}$ & include in chain]**.

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

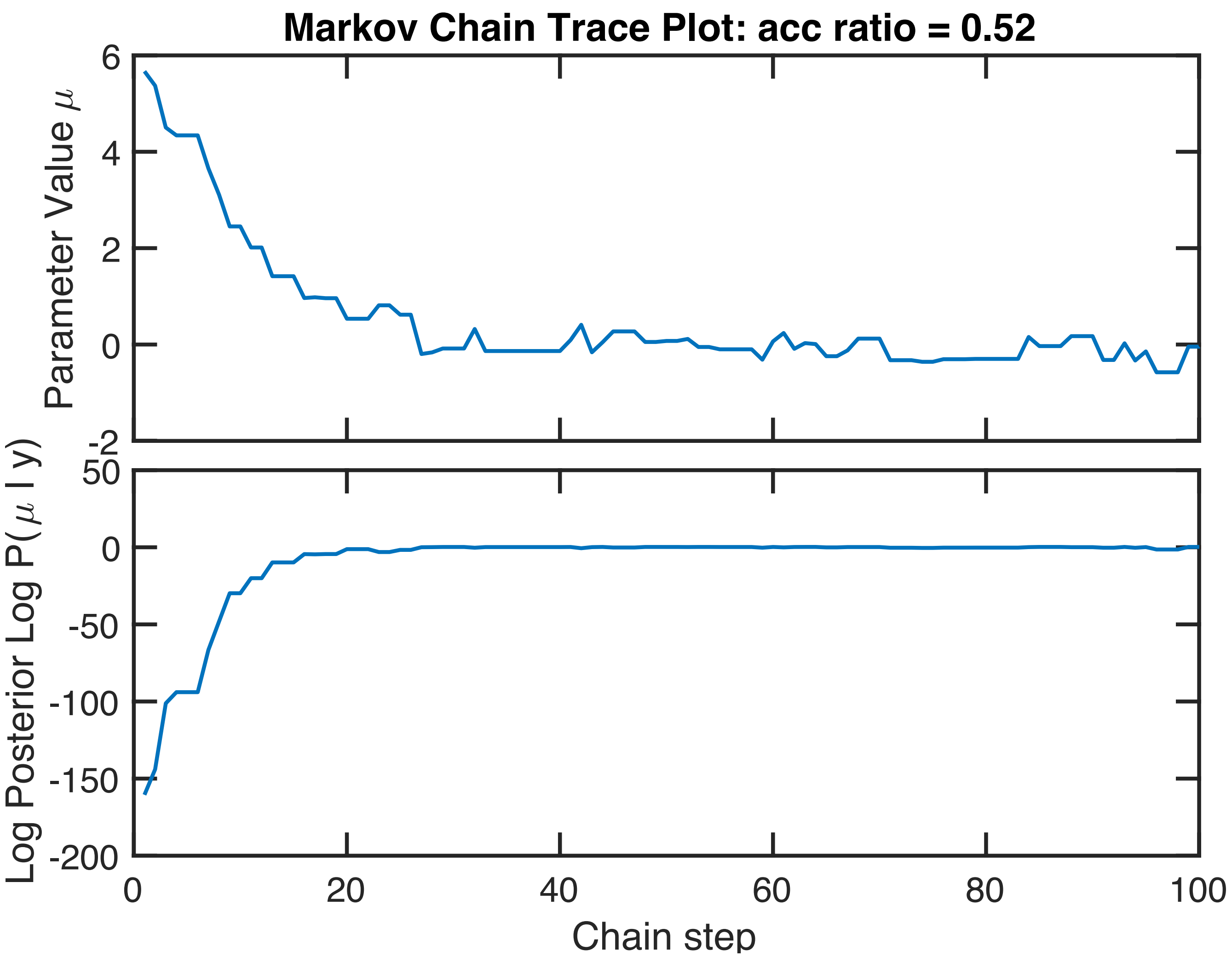
Code demo: metropolis1.m

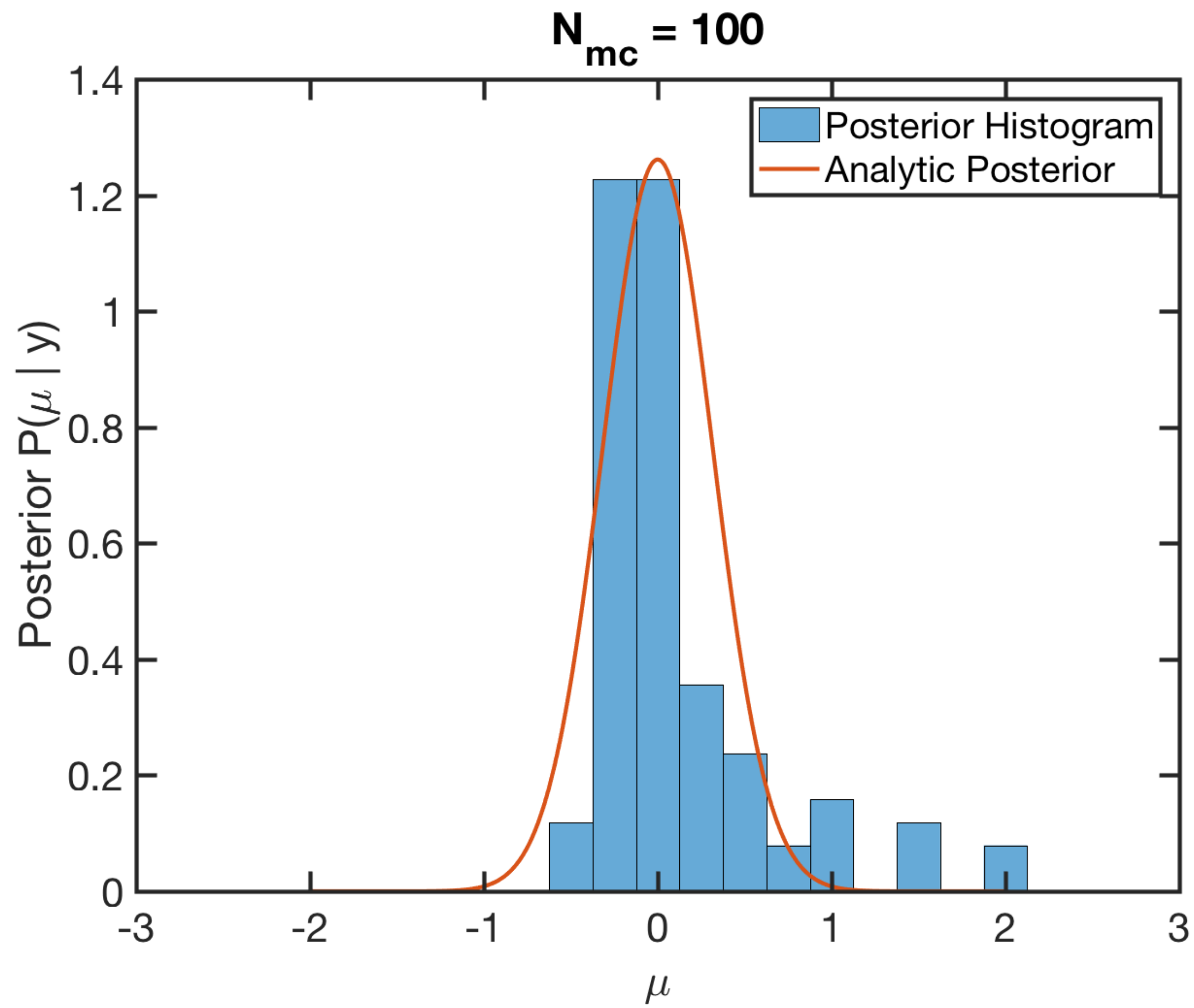
First 50 iterations



Code demo: metropolis1.m

First 100 iterations



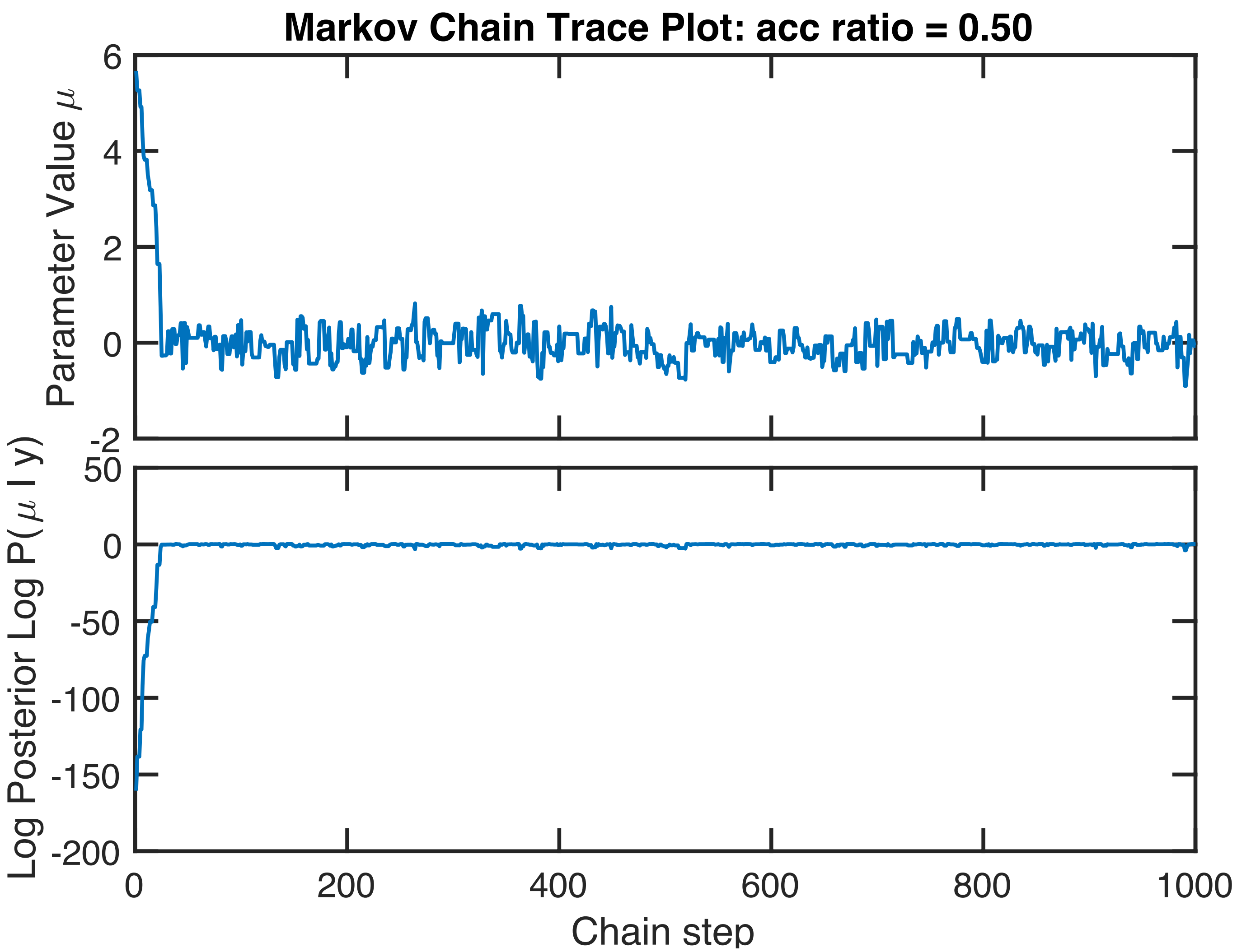


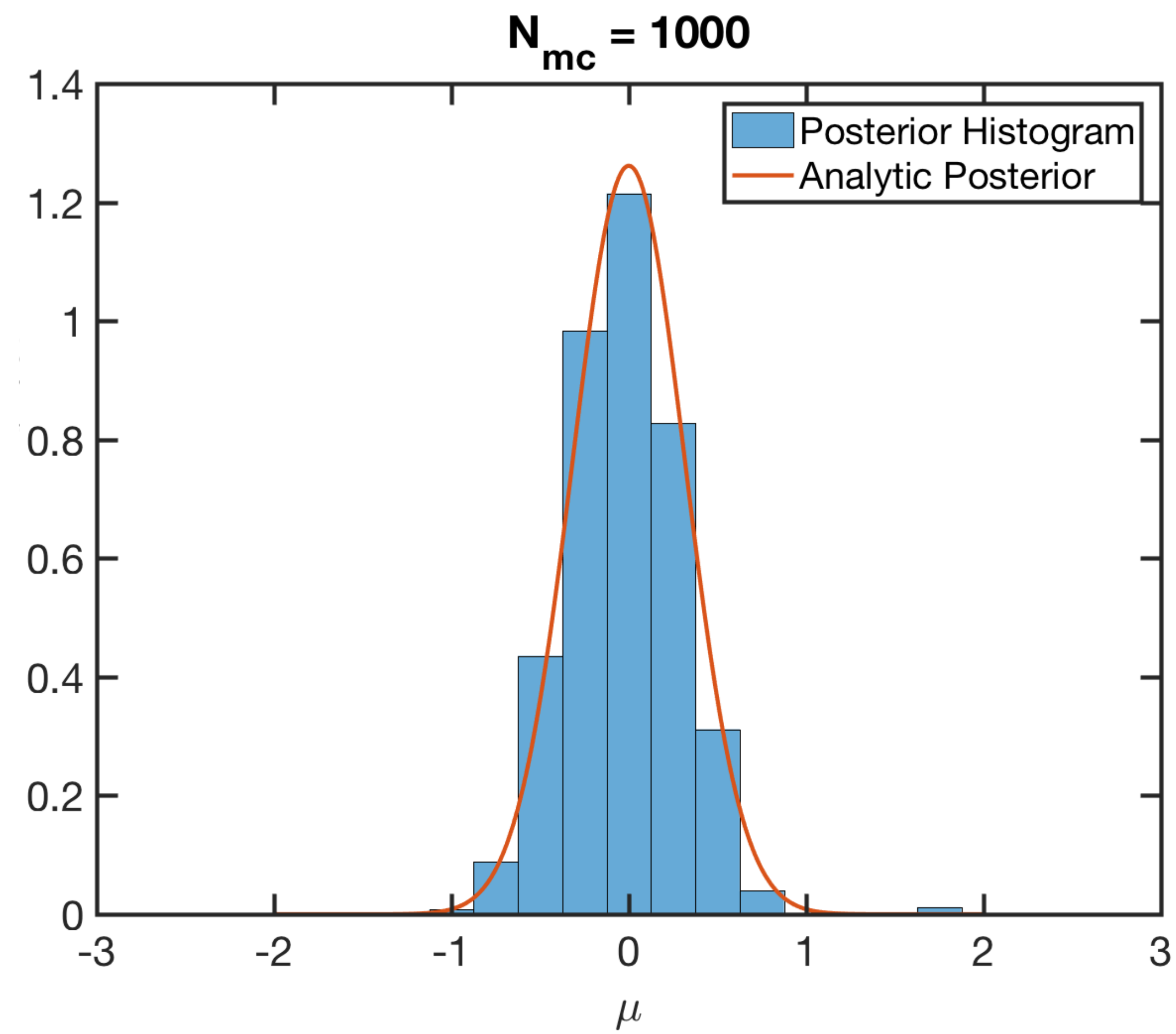
Trace Plot

Posterior

Code demo: metropolis1.m

First 1000 iterations



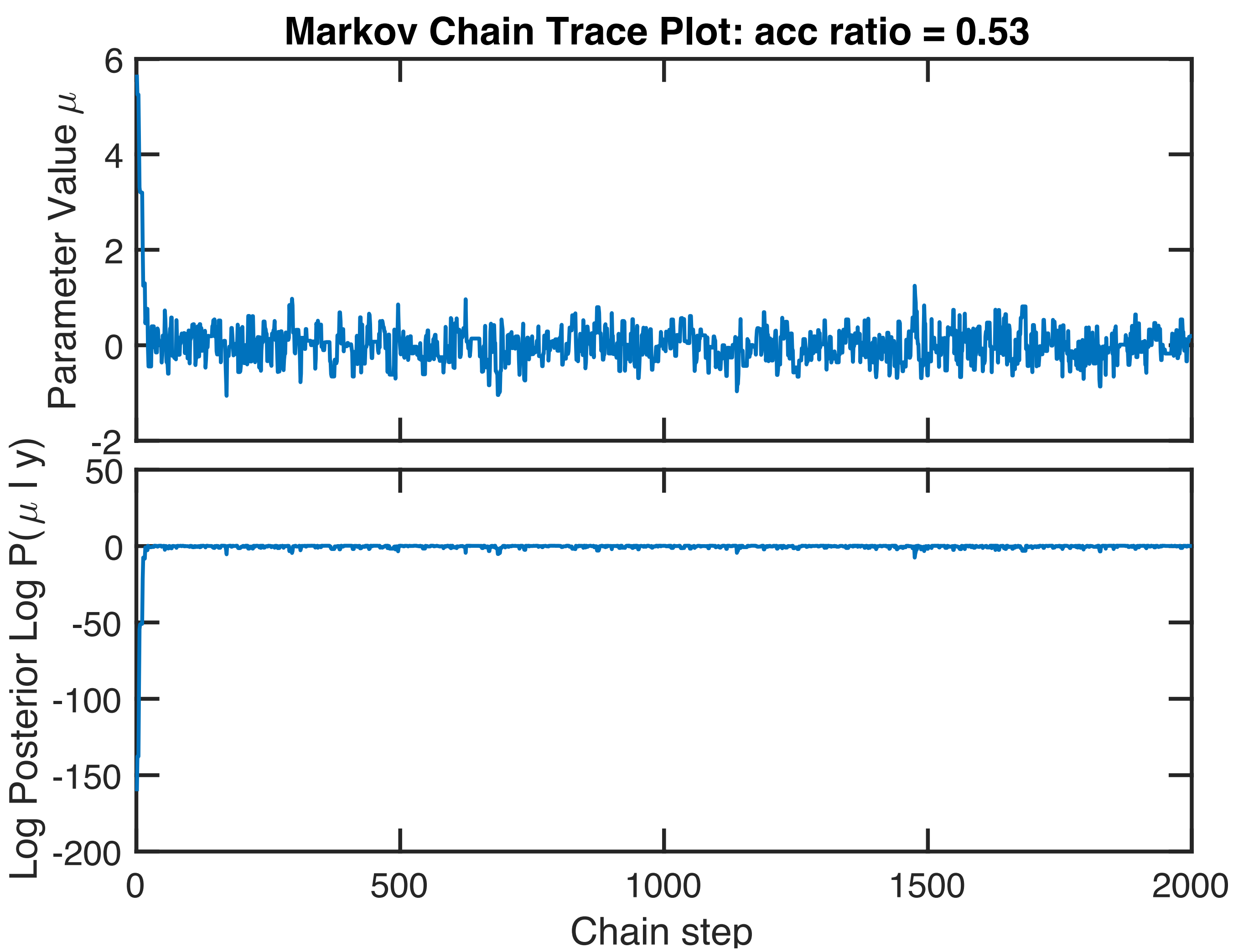


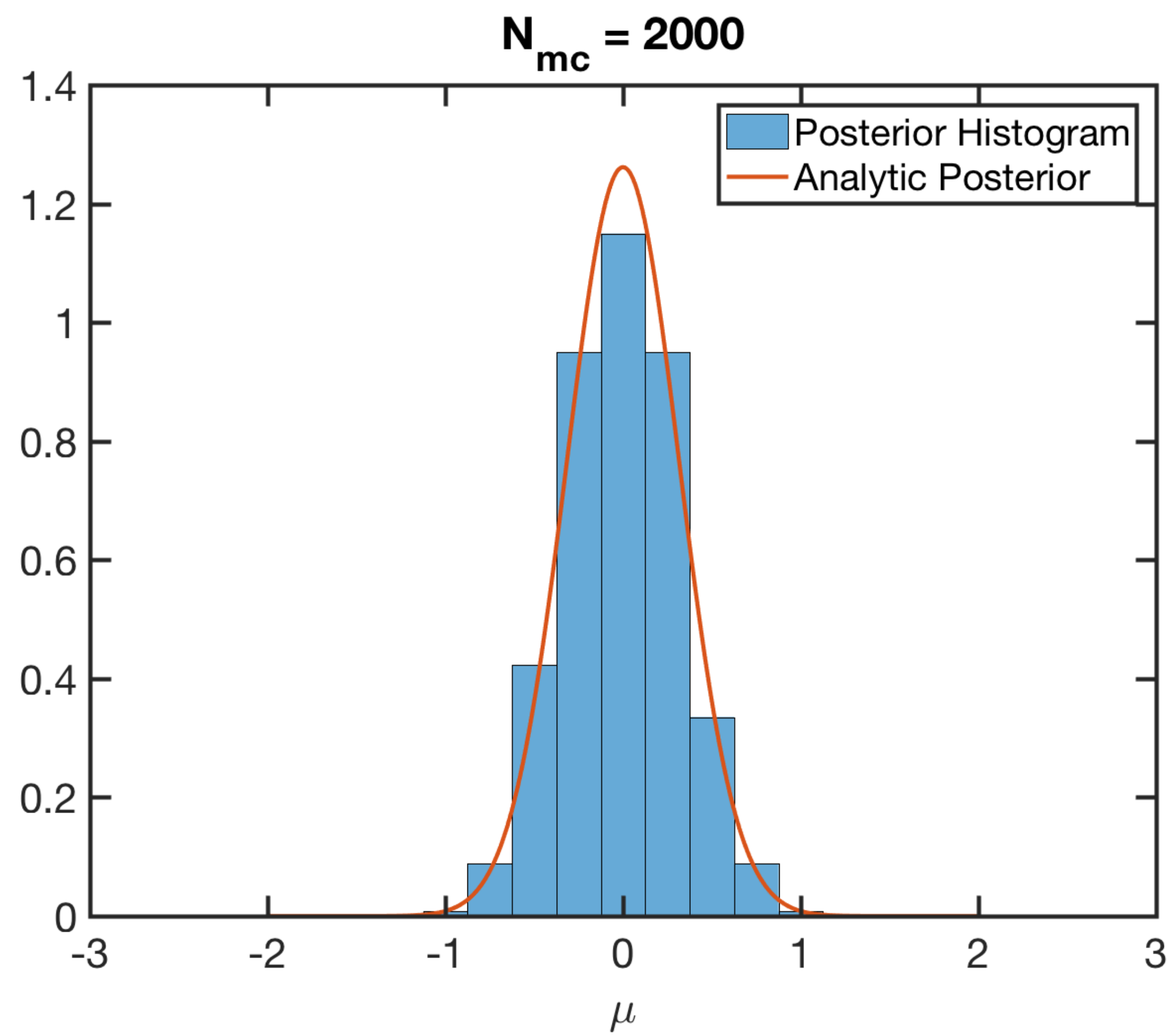
Trace Plot

Posterior

Code demo: metropolis1.m

2000 iterations



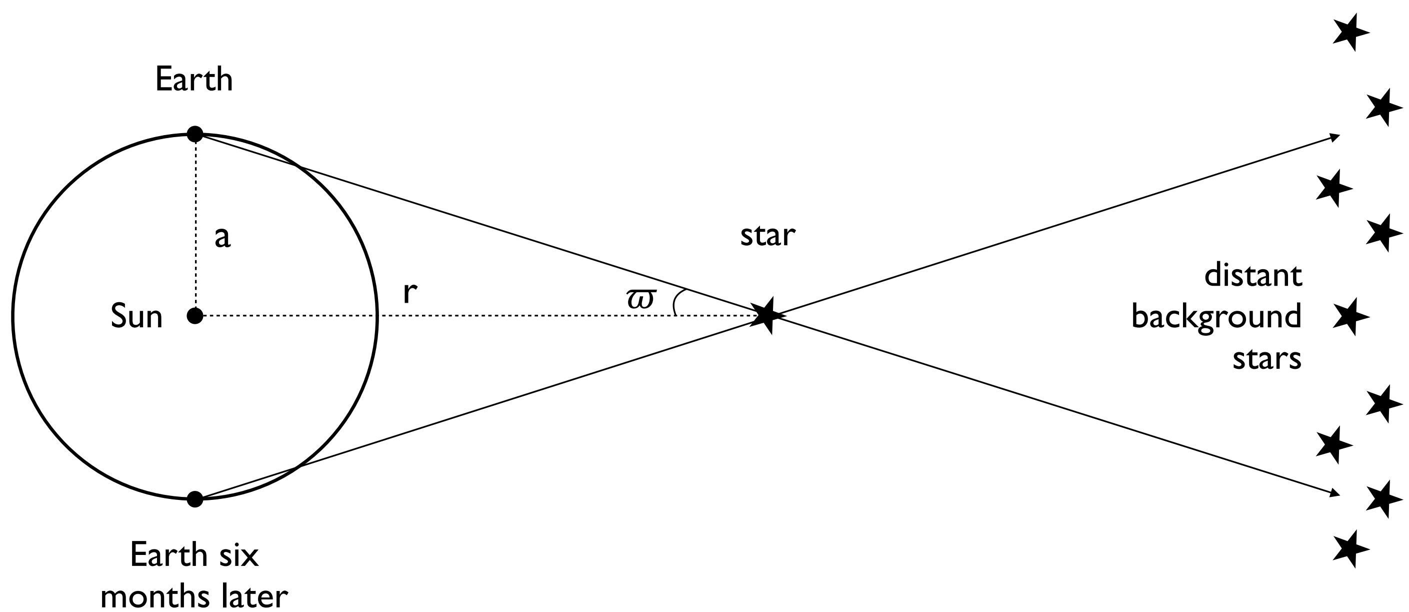


Trace Plot

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

Parallax Example
Likelihood:

$$P(\varpi\,|\,r) \;=\; \frac{1}{\sigma_{\varpi}\sqrt{2\pi}}\exp\left[-\frac{1}{2\sigma_{\varpi}^{2}}\left(\varpi-\frac{1}{r}\right)^{2}\right] \quad \text{where} \quad \sigma_{\varpi}>0,$$



Parallax Angle

$$\frac{\omega}{\text{arcsec}} = \frac{\text{parsec}}{r}$$

Distance

The parallax ϖ of a star is the apparent angular displacement of that star (relative to distant background stars) due to the orbit of the Earth about the Sun. More precisely, the parallax is the angle subtended by the Earth's orbital radius a as seen from the star. As parallaxes are extremely small angles ($\varpi \ll 1$), $\varpi = a/r$ to a very good approximation. When ϖ is 1 arcsecond, r is defined as the *parsec*, which is about 3.1×10^{13} km. In this sketch the size of the Earth's orbit has been greatly exaggerated compared to the distance to the star, and the distance to the background stars in reality is orders of magnitude larger again.

Introducing physical constraints into the prior

$$P(r) = \begin{cases} \dfrac{1}{2L^3}\, r^2 e^{-r/L} & \text{if } r > 0 \\ 0 & \text{otherwise} \end{cases}$$

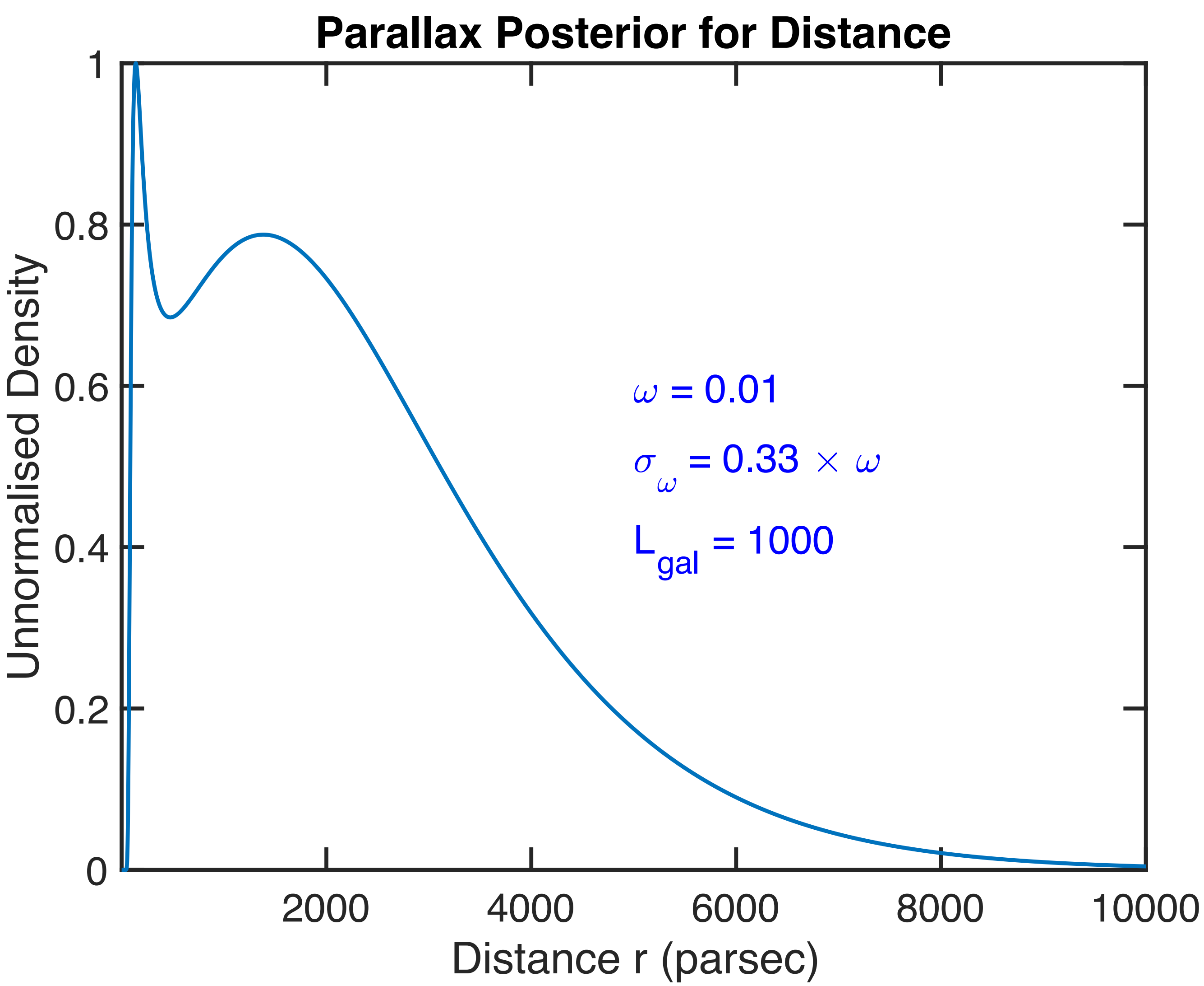
Exponential decrease in density of stars with Galactic length scale L

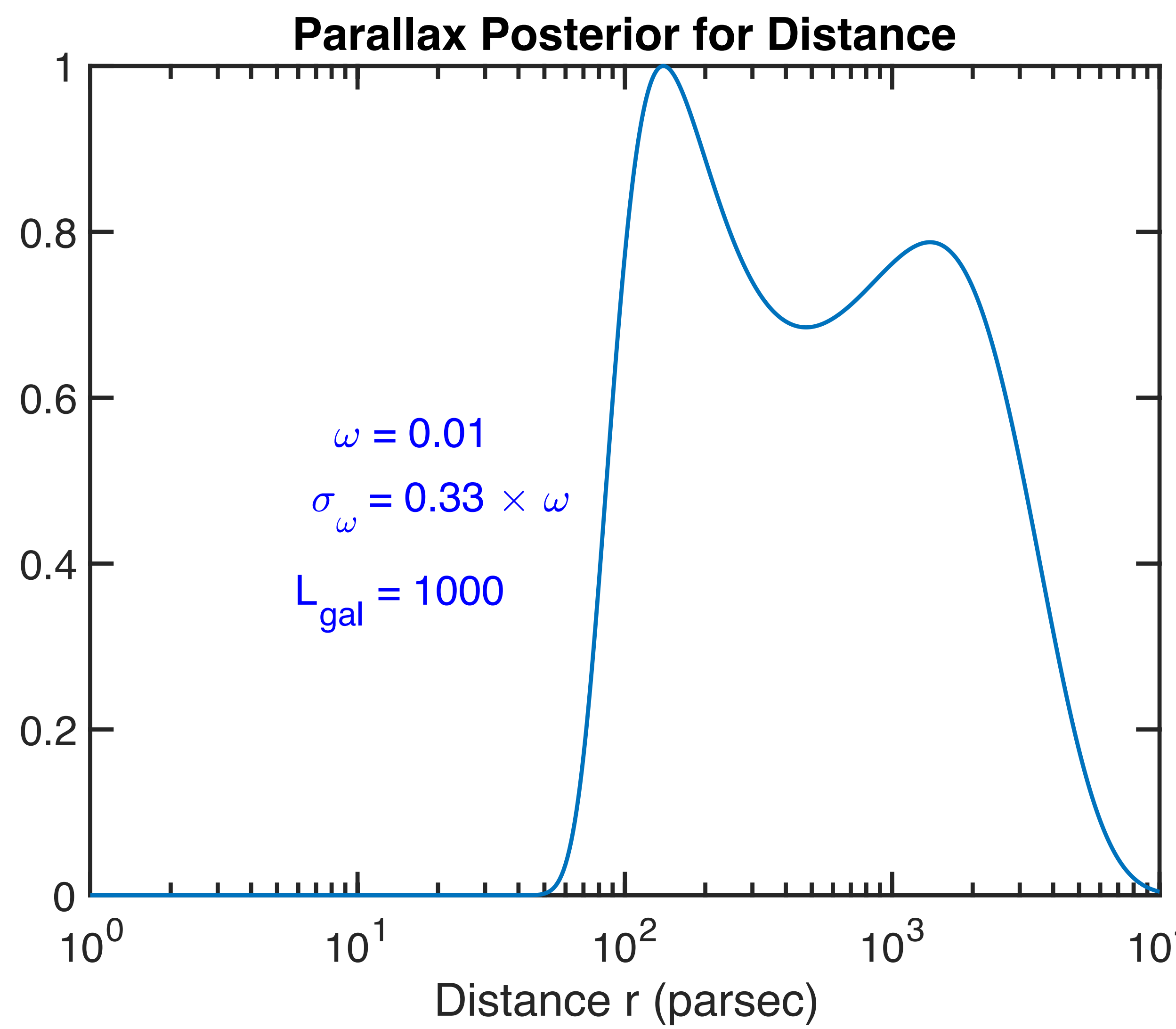
$$P(r|\,\omega) \propto P(\omega|r) \times P(r)$$

Unnormalised Posterior:

$$P^*_{r^2 e^{-r}}(r|\varpi, \sigma_\varpi) = \begin{cases} \frac{r^2 e^{-r/L}}{\sigma_\varpi} \exp\left[-\frac{1}{2\sigma_\varpi^2}\left(\varpi - \frac{1}{r}\right)^2\right] & \text{if } r > 0 \\ 0 & \text{otherwise} \end{cases} .$$

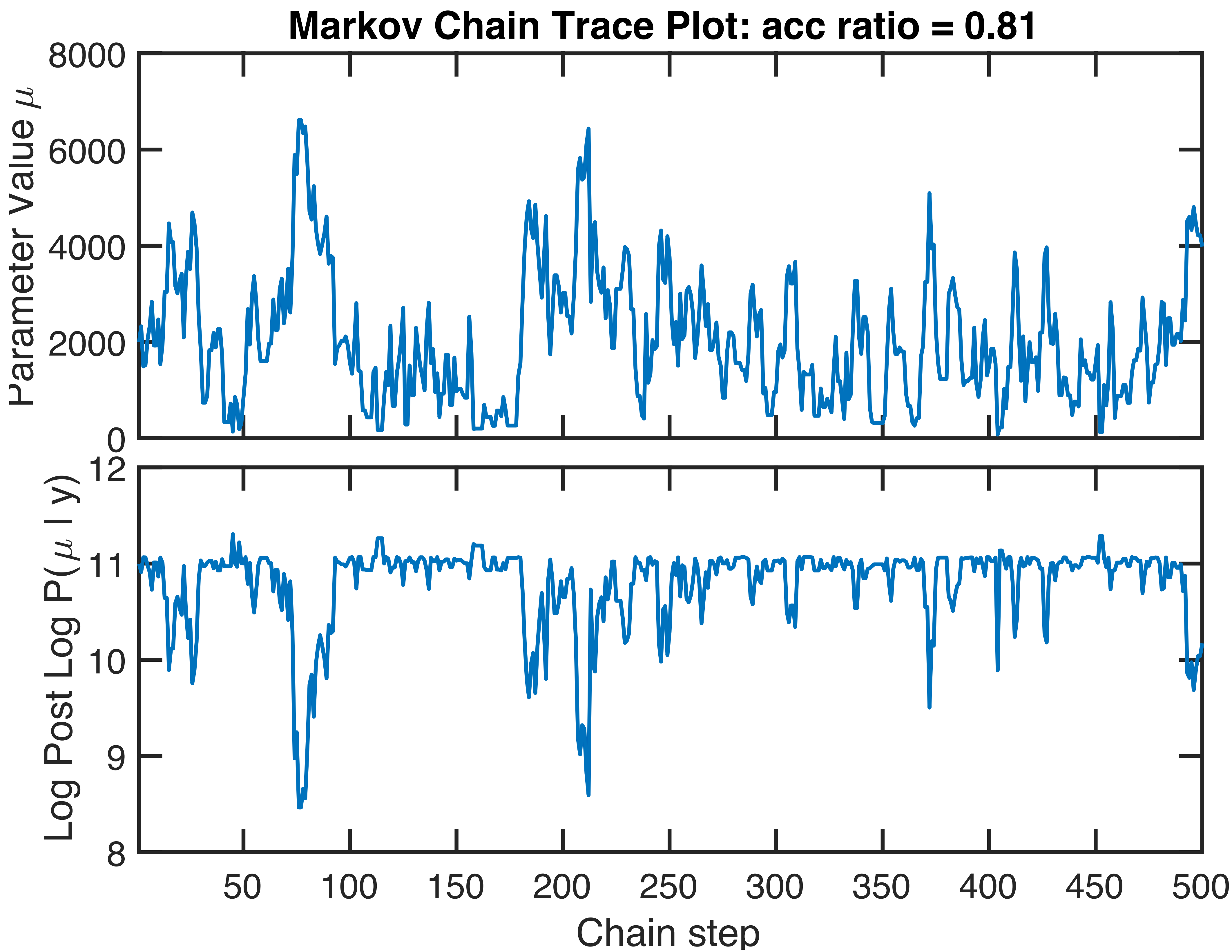
Parallax Example



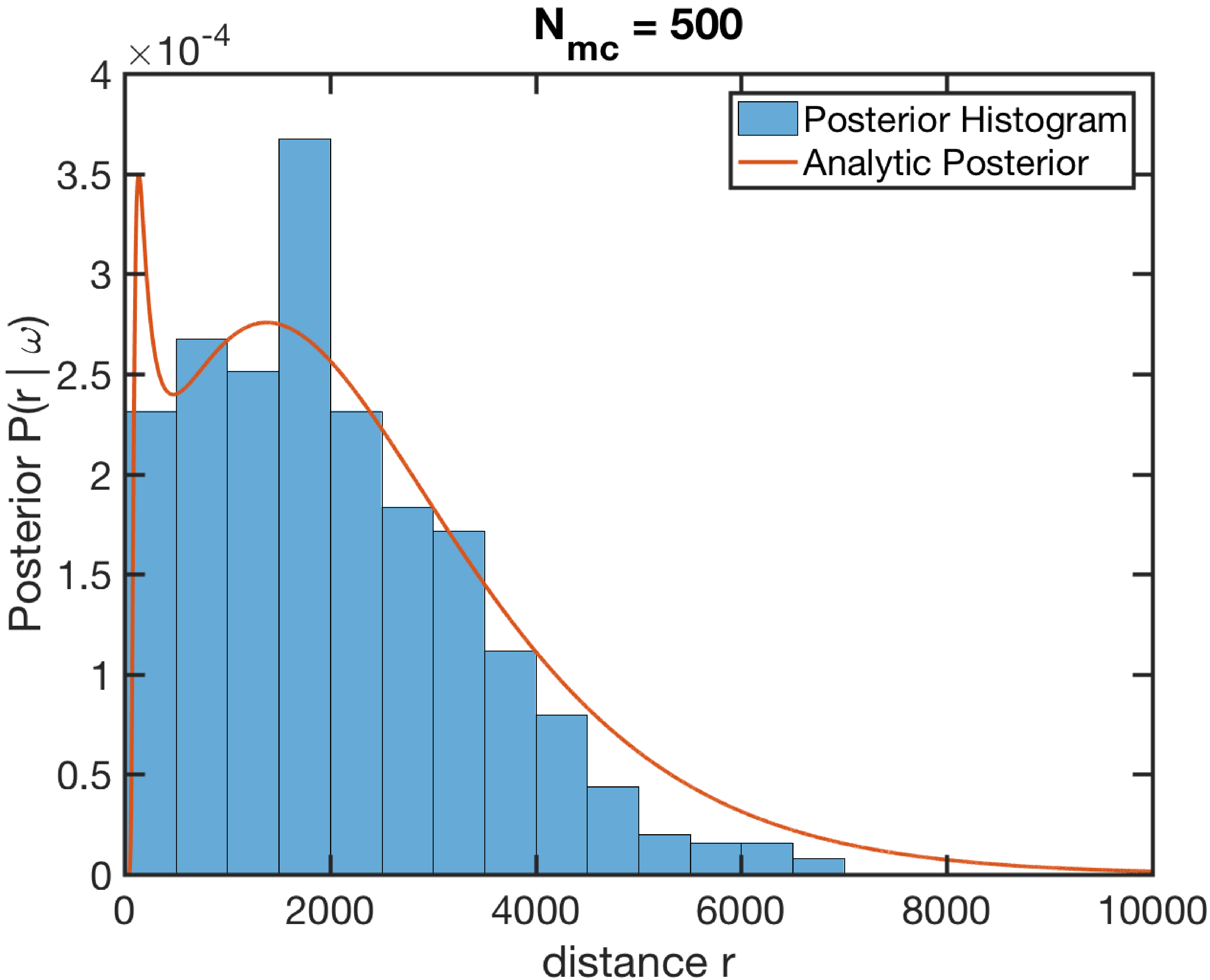


Log Scale

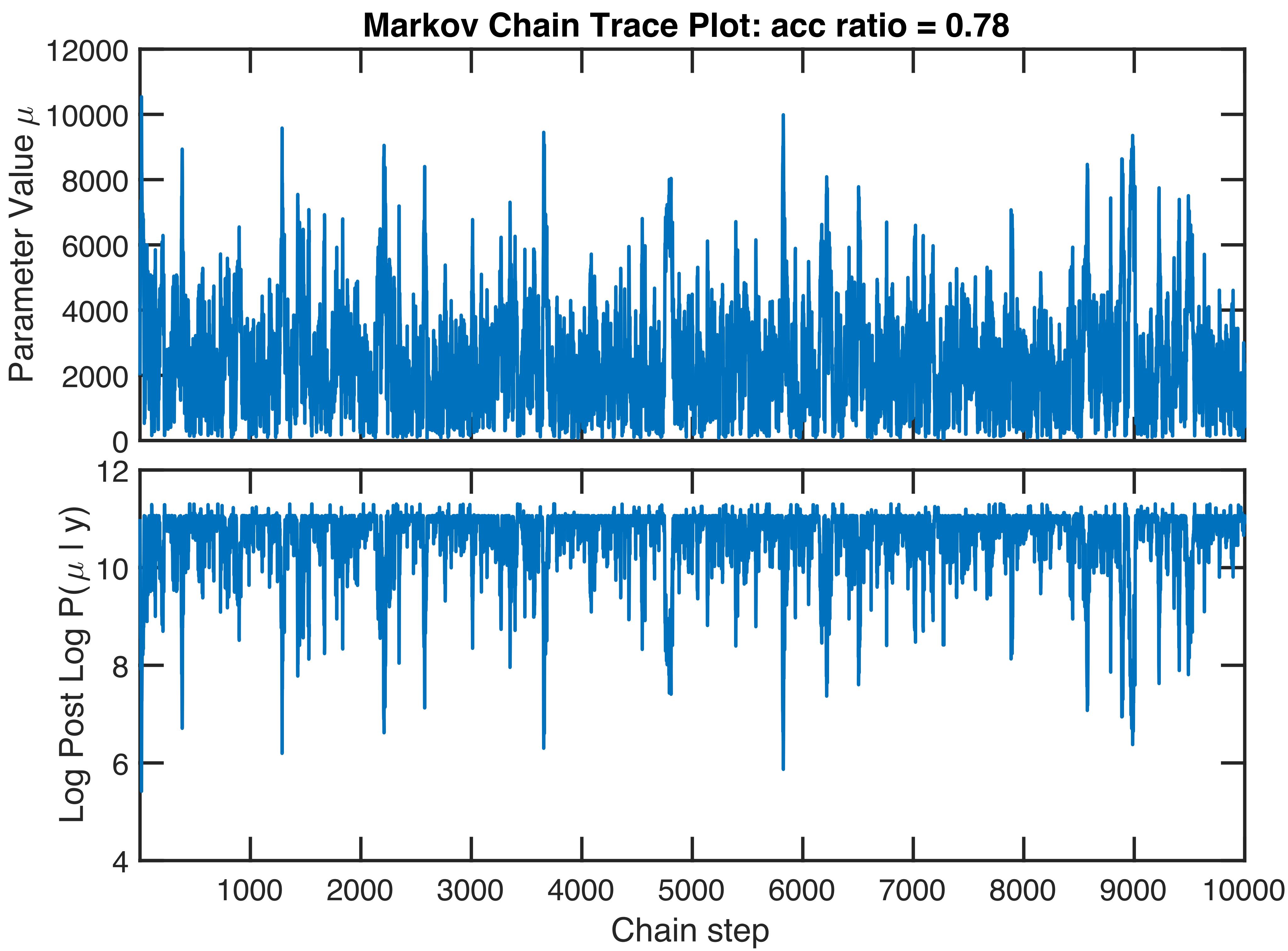
Parallax Example



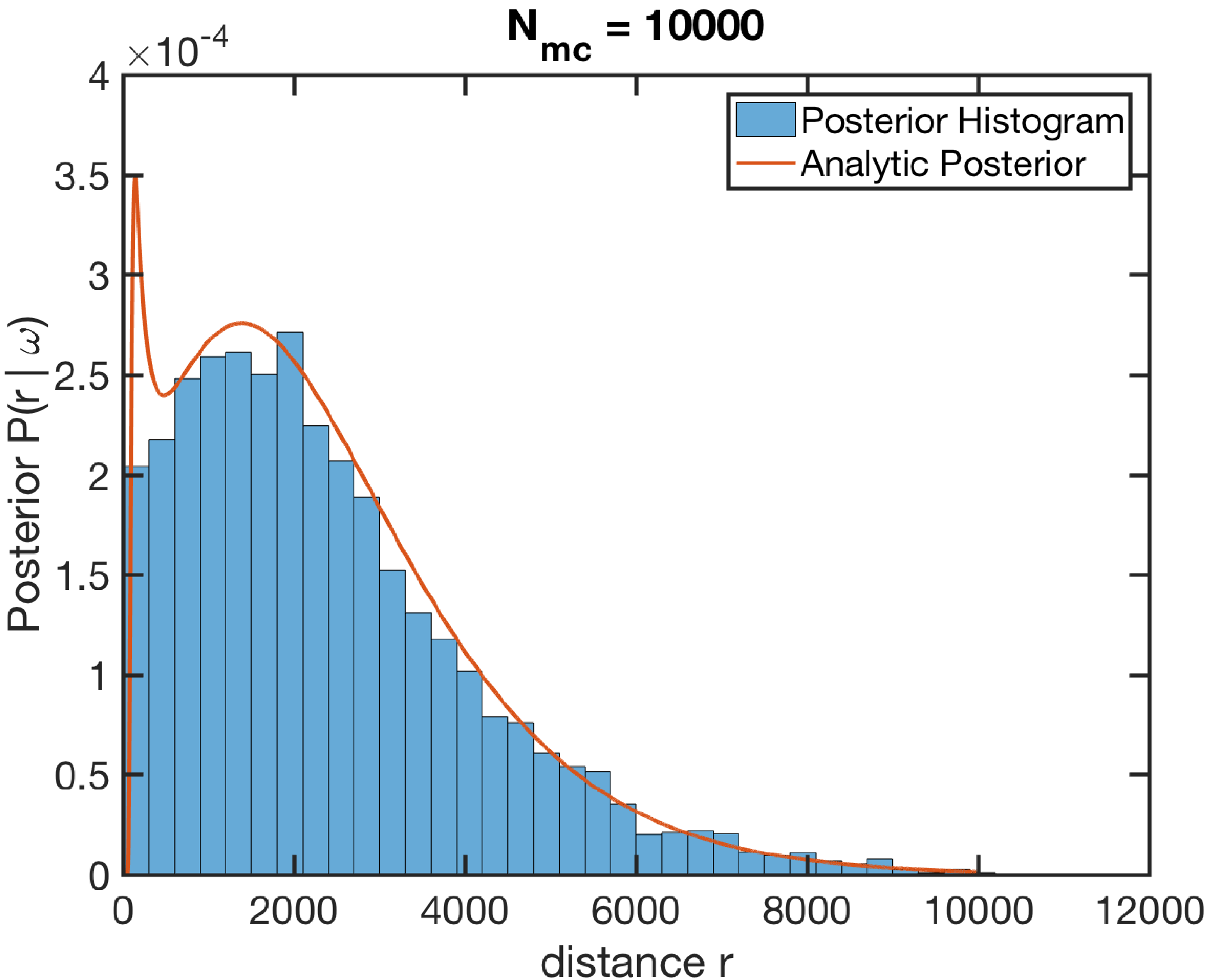
Parallax Example



Parallax Example



Parallax Example



Mapping the Posterior P(θ |D)

- Markov Chain Monte Carlo (MCMC)
- Last time: 1D Metropolis algorithm
- Now:
    - Drawing Multivariate Gaussian random variables
    - N-D Metropolis Algorithm
    - Rules of thumb for proposal scale
    - assessing convergence (G-R Ratio)
    - Metropolis-Hastings algorithm
    - Gibbs sampling