Astrostatistics: 13 Mar 2019

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

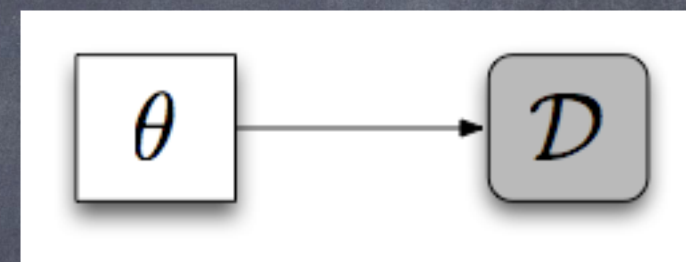
- Make-Up Lecture, Thu 14 Mar, 11am-12: MR13
- Example Class, Fri 15 Mar 1pm: MR 12
- Today:
- Hierarchical Bayes & Shrinkage Estimators

Aside on Product of Gaussian densities

Review:
Hierarchical Bayes and Probabilistic Graphical Models:
a visual way to understand complex statistical models

Simple Bayes:

$$\mathcal{D}|\,\theta \sim \mathrm{Model}(\theta)$$

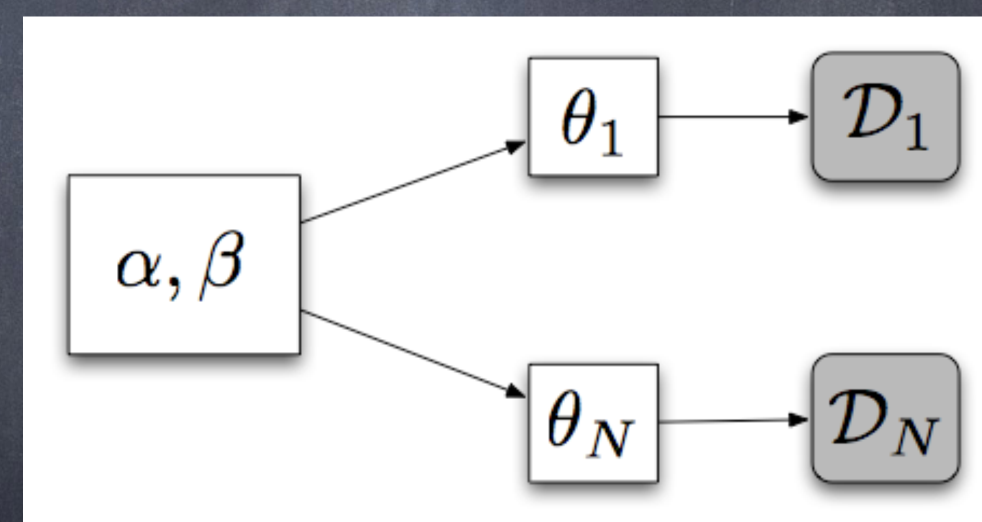


$$P(\theta|\,\mathcal{D}) \propto P(\mathcal{D}|\,\theta)P(\theta)$$

Hierarchical Bayes:

$$\mathcal{D}_i|\,\theta_i \sim \mathrm{Model}(\theta_i)$$

$$\theta_i|\,\alpha,\beta \sim \mathrm{PopModel}(\alpha,\beta)$$



$$P(\{\theta_i\},\alpha,\beta|\{\mathcal{D}_i\}) \propto \left[\prod_{i=1}^{N} P(\mathcal{D}_i|\,\theta_i)P(\theta_i|\alpha,\beta)\right]P(\alpha,\beta)$$

Build up complexity by layering conditional probabilities

Review: Hierarchical Bayesian "Normal-Normal" Model

Level 1: Population Distribution of Latent Variables (Absolute Mags)

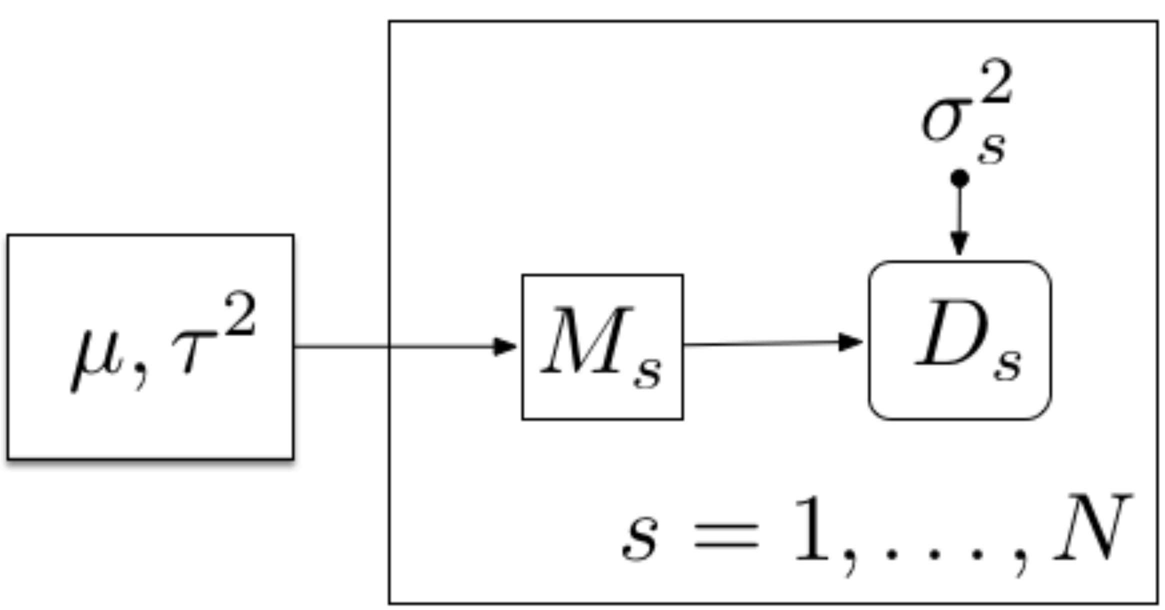
$$M_s \sim N(M_0, \tau^2) \quad s = 1 \dots N$$

Level 2 : Measurement Error Process

$$D_s | M_s \sim N(M_s, \sigma_s^2) \quad s = 1 \dots N$$

Joint Probability Density of Data, Latent Variables, **H**yperparameters

$$P(\{D_s\}, \{M_s\}, \boldsymbol{H} = M_0, \tau^2)$$



Joint factors into Conditional and Marginal densities based on Model Assumptions

Hierarchical vs Regular Bayes

- Could regard as just a general Bayesian inference problem in a very high dimensional parameter space, e.g.

$$\boldsymbol{\theta} = \{M_1, \ldots, M_N; M_0, \tau^2\} = \{\boldsymbol{M}; M_0, \tau^2\}$$

$$P(\boldsymbol{\theta}|\boldsymbol{D}) \propto P(\boldsymbol{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

$$P(\boldsymbol{\theta}|\boldsymbol{D}) \propto P(\boldsymbol{D}|\boldsymbol{M})P(\boldsymbol{M}|M_0, \tau^2)P(M_0, \tau^2)$$

$$P(\boldsymbol{\theta}|\boldsymbol{D}) \propto \left[\prod_{s=1}^{N} P(D_s|M_s)P(M_s|M_0, \tau^2)\right] P(M_0, \tau^2)$$

- However, special hierarchical structure is useful for modelling, estimation, and computation

- For large N, wouldn't want to do an N+2 dimensional Metropolis MCMC!

Hierarchical Bayesian "Normal-Normal" Model

Level 1: Population Distribution of Latent Variables (Absolute Mags)

$$M_s \sim N(M_0, \tau^2)$$

Population Dist'n / Prior

Level 2 : Measurement Error Process

$$D_s | M_s \sim N(M_s, \sigma_s^2)$$

Measurement Likelihood

Joint Probability Density of ALL THE THINGS:
Data, Latent Variables, Hyperparameters

$$P(\{D_s\}, \{M_s\}, \boldsymbol{H}) = \left[\prod_{s=1}^{N} P(D_s | M_s) P(M_s | M_0, \tau^2) \right] \times P(\boldsymbol{H})$$

(Derive on board)

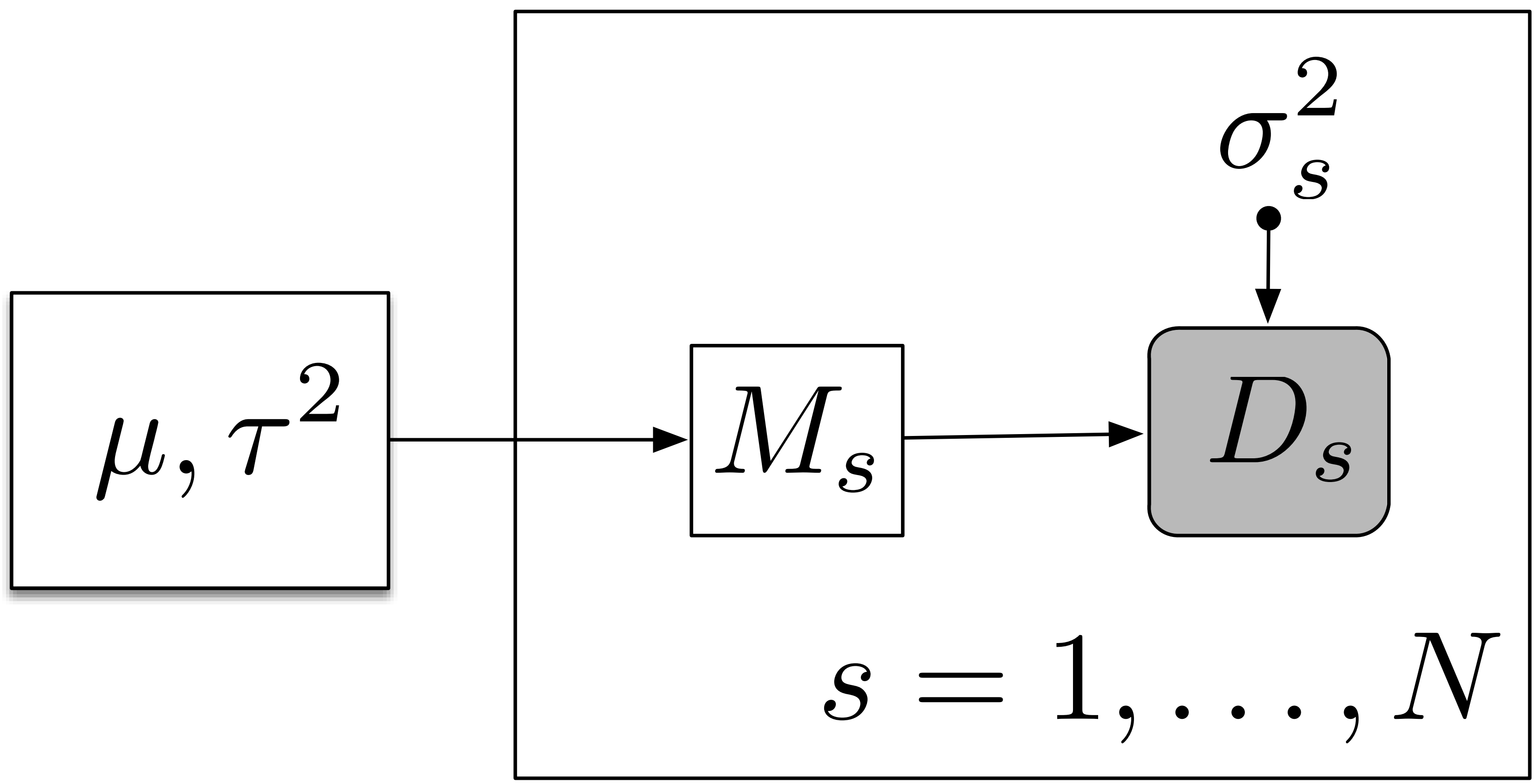
Measurement Likelihood

Population Dist'n / Prior

Hyperprior

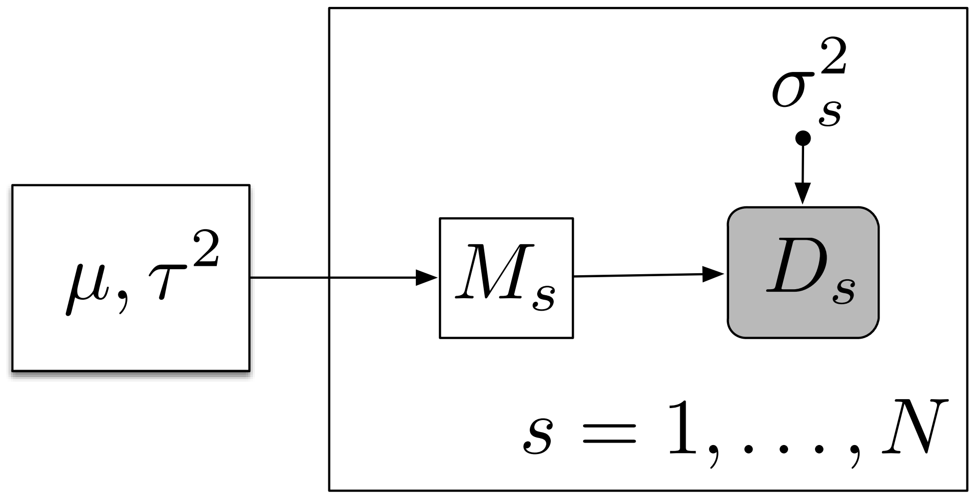
Sampling the Hierarchical Bayesian posterior

$$P(\{M_s\}, \boldsymbol{H} | \{D_s\}) \propto \Big[\prod_{s=1}^{N} P(D_s|M_s) P(M_s|M_0, \tau^2) \Big] \times P(\boldsymbol{H})$$

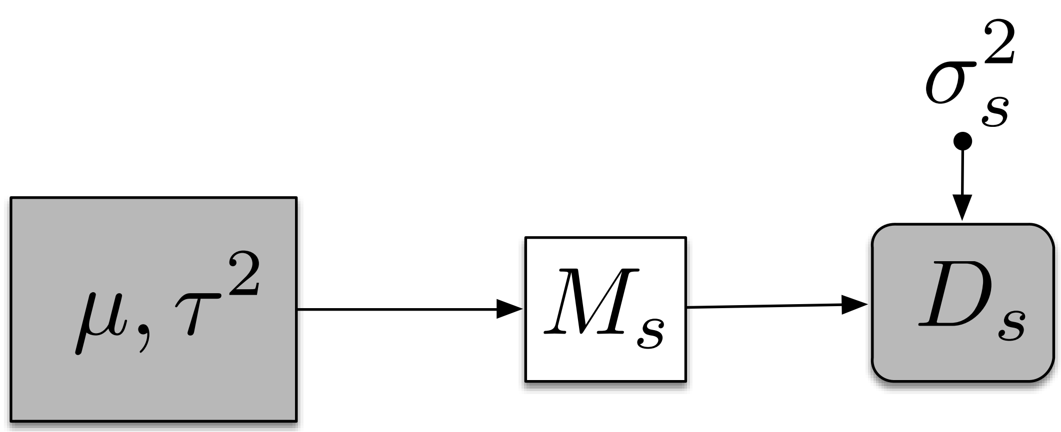


Utilises Conditional Independence structure of PGM to derive conditional posterior densities

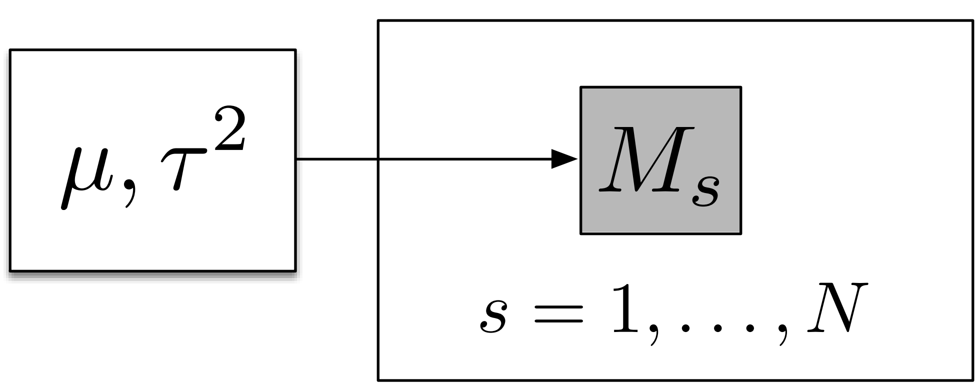
Gibbs sampling & Hierarchical Bayes



1. For s = 1 … N: Sample Latent Variables Conditional on Data and **Hyperparameters**



2. Sample Hyperparameters from Conditional on Data and Latent Variables:

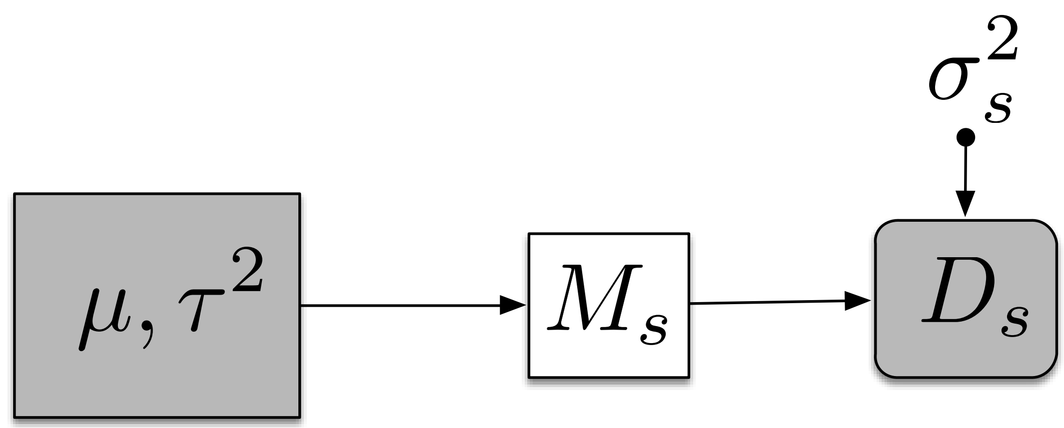


Gibbs sampling & Hierarchical Bayes

Utilises Conditional Independence structure of PGM to derive conditional posterior densities

$$P(\{M_s\}, \boldsymbol{H} | \{D_s\}) \propto \Big[\prod_{s=1}^{N} P(D_s | M_s) P(M_s | M_0, \tau^2) \Big] \times P(\boldsymbol{H})$$

1. For s = 1 … N: Sample Latent Variables Conditional on Data and **Hyperparameters**



$$P(M_s | \, \mu, \tau^2, D_S) \propto P(D_s | \, M_s) \times P(M_s | M_0, \tau^2)$$

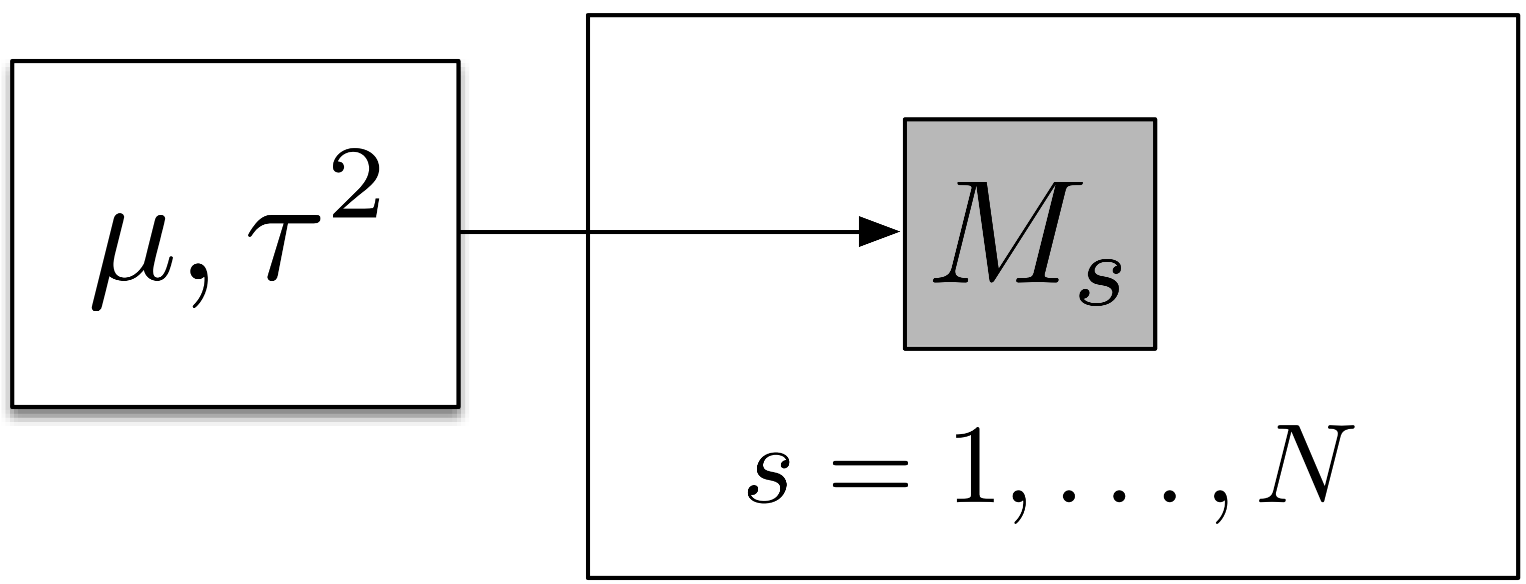
$$\propto N(D_s | \, M_s, \sigma^2) \times N(M_s | M_0, \tau^2)$$

Gibbs sampling & Hierarchical Bayes

Utilises Conditional Independence structure of PGM to derive conditional posterior densities

$$P(\{M_s\}, \boldsymbol{H} | \{D_s\}) \propto \left[\prod_{s=1}^{N} P(D_s|M_s) P(M_s|M_0, \tau^2) \right] \times P(\boldsymbol{H})$$

2. Sample Hyperparameters from Conditional on Data and Latent Variables:



Hyperprior: $P(\mu, \tau^2) \propto (\tau^2)^{-1/2}$

$$P(M_0, \tau^2 | \{M_s\}; \{D_s\}) = P(M_0, \tau^2 | \{M_s\}) = P(M_0 | \tau^2, \{M_s\}) P(\tau^2 | \{M_s\})$$

(See Example Sheet 1, Problem 4) (Gaussian) (Inv-χ^2)

Gibbs sampling & Hierarchical Bayes

Utilises Conditional Independence structure of PGM to derive conditional posterior densities

$$P(\{M_s\}, \boldsymbol{H} | \{D_s\}) \propto \left[\prod_{s=1}^{N} P(D_s | M_s) P(M_s | M_0, \tau^2) \right] \times P(\boldsymbol{H})$$

2. Sample Hyperparameters from Conditional on Data and Latent Variables:

$$P(M_0, \tau^2 | \{M_s\}; \{D_s\}) = P(M_0, \tau^2 | \{M_s\}) = P(M_0 | \tau^2, \{M_s\}) P(\tau^2 | \{M_s\})$$

(Gaussian) (Inv-χ^2)

Hyperprior: $P(\mu, \tau^2) \propto (\tau^2)^{-1/2}$

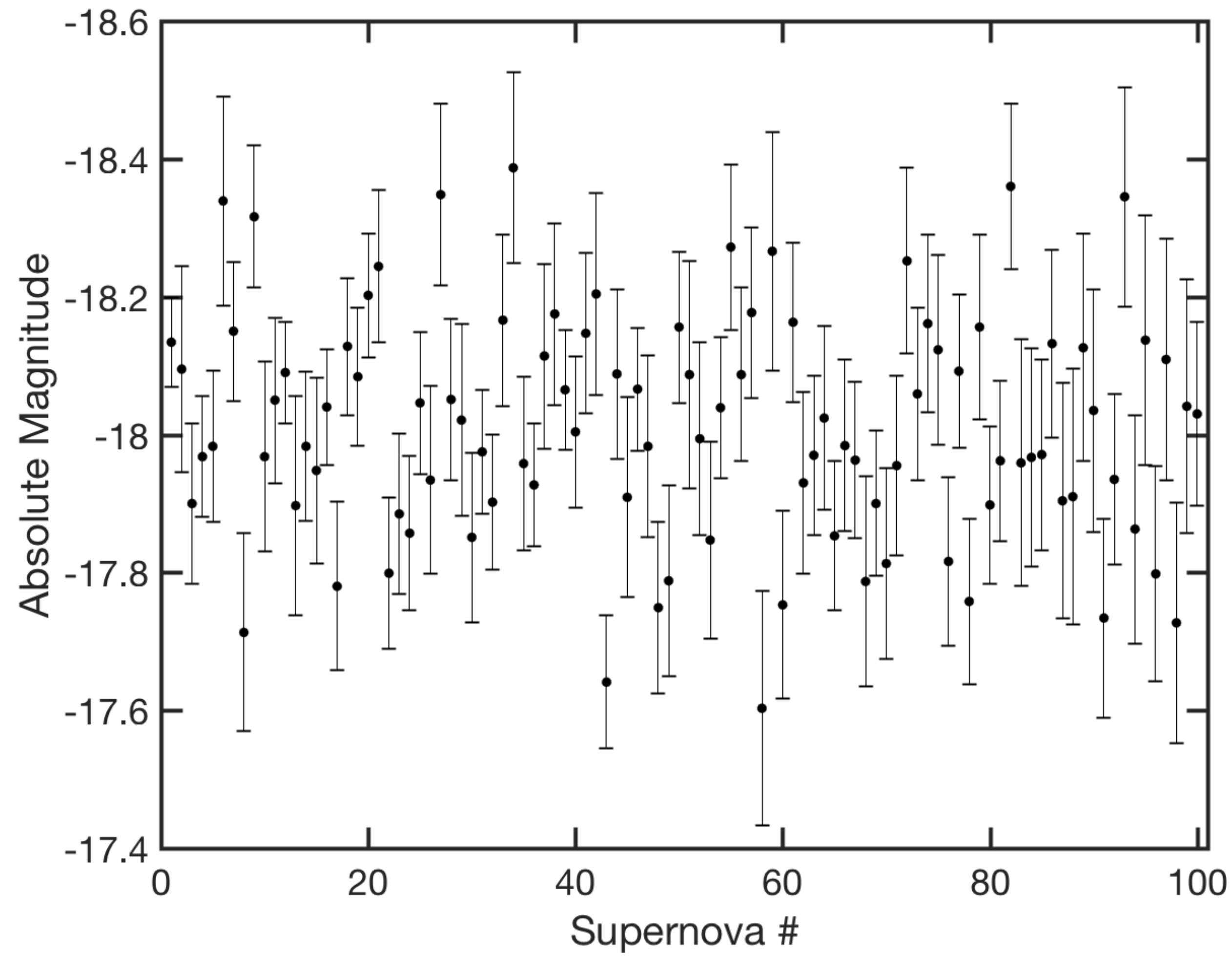
Conditional Posteriors: $\tau^2 | \{M_s\} \sim \text{Inv-}\chi^2 \left(N - 2, \frac{(N-1)}{(N-2)} s^2 \right)$

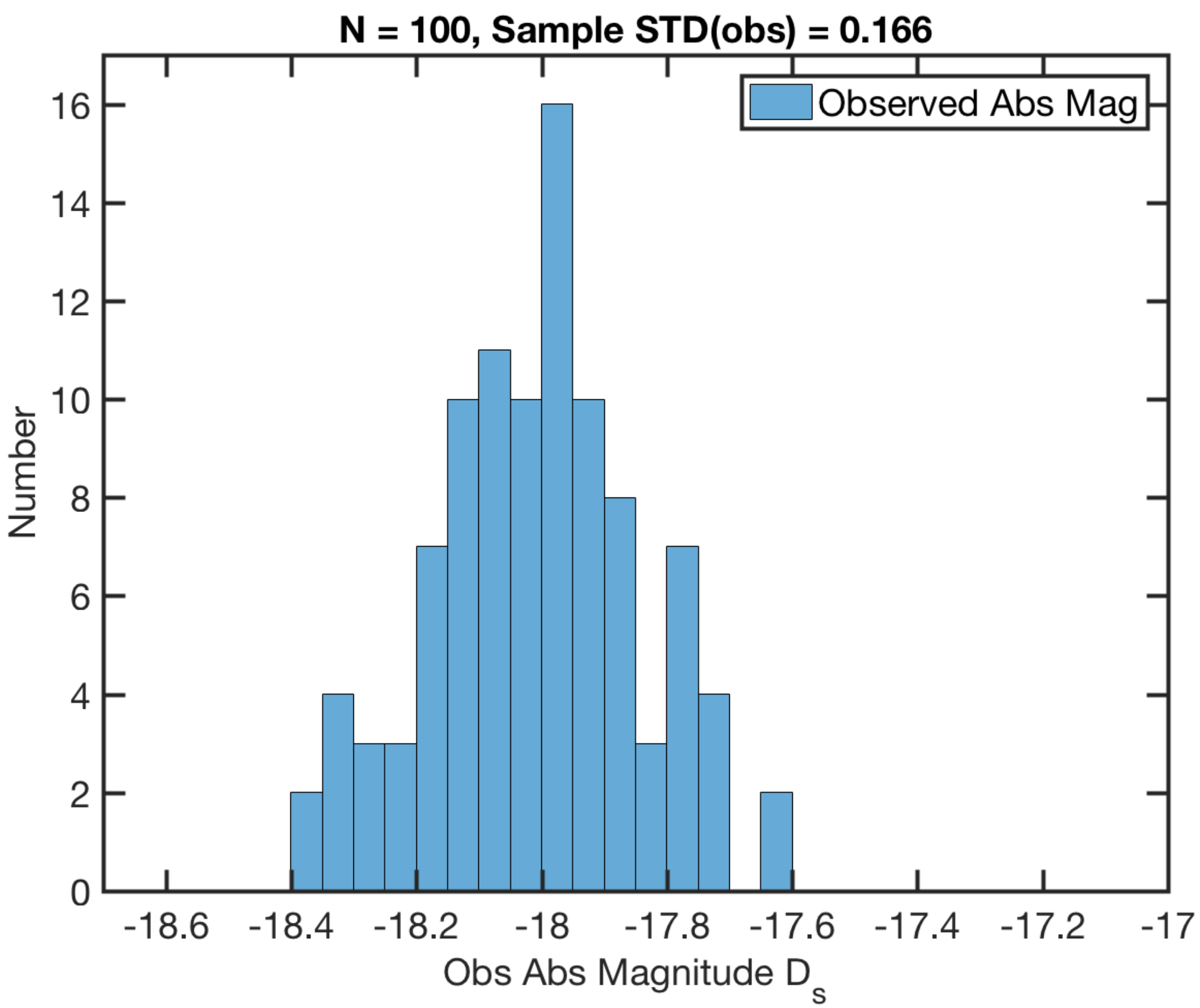
$$M_0 | \tau^2; \{M_s\} \sim N(\bar{M}, \tau^2 / N)$$

$$\bar{M} \equiv \frac{1}{N} \sum_{s=1}^{N} M_s \qquad s^2 \equiv \frac{1}{N-1} \sum_{s=1}^{N} (M_s - \bar{M})^2$$

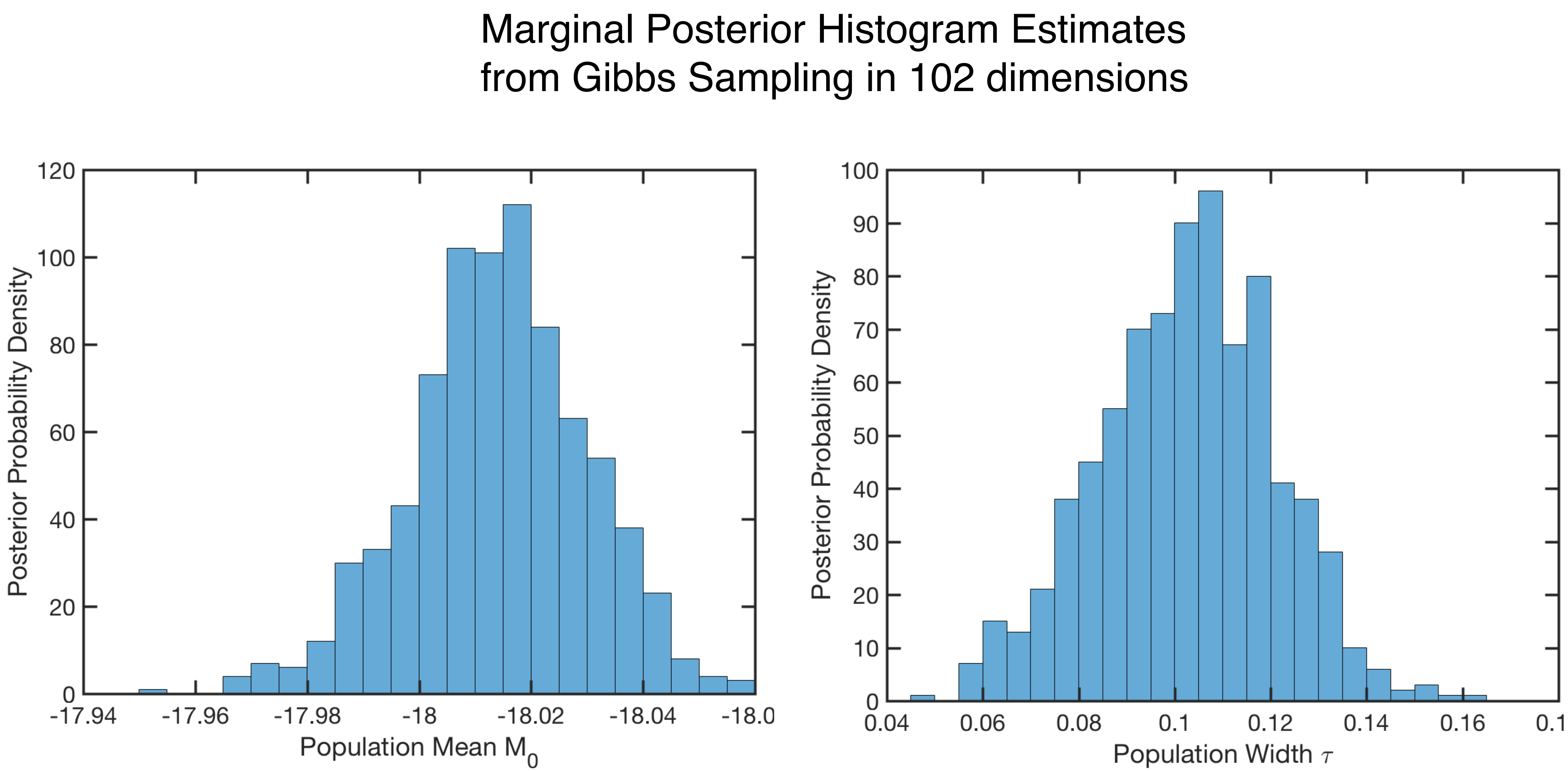
Code/Data Demo

100 Supernovae with
app mag and
distance measurements



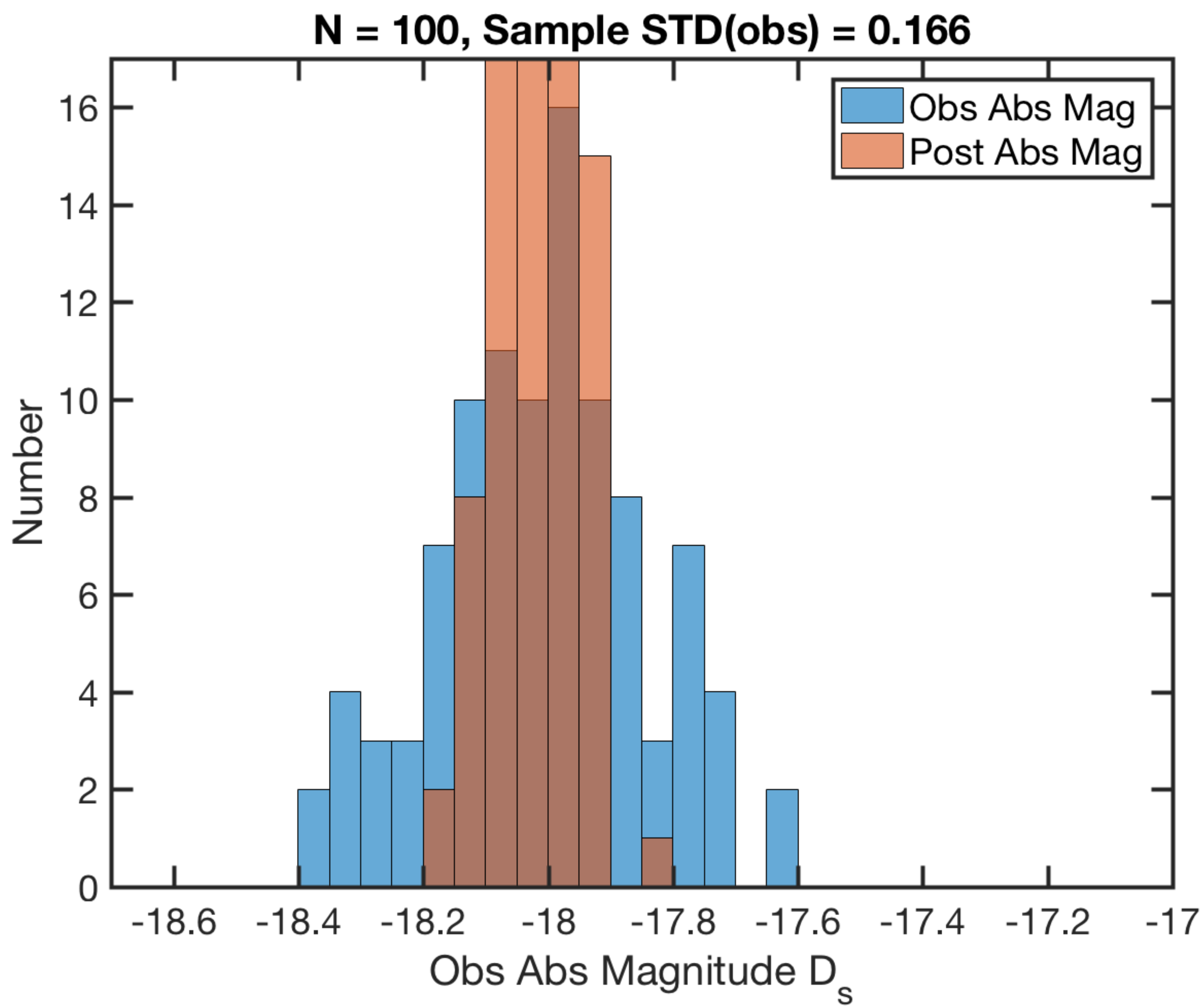


Hyperparameter Inference



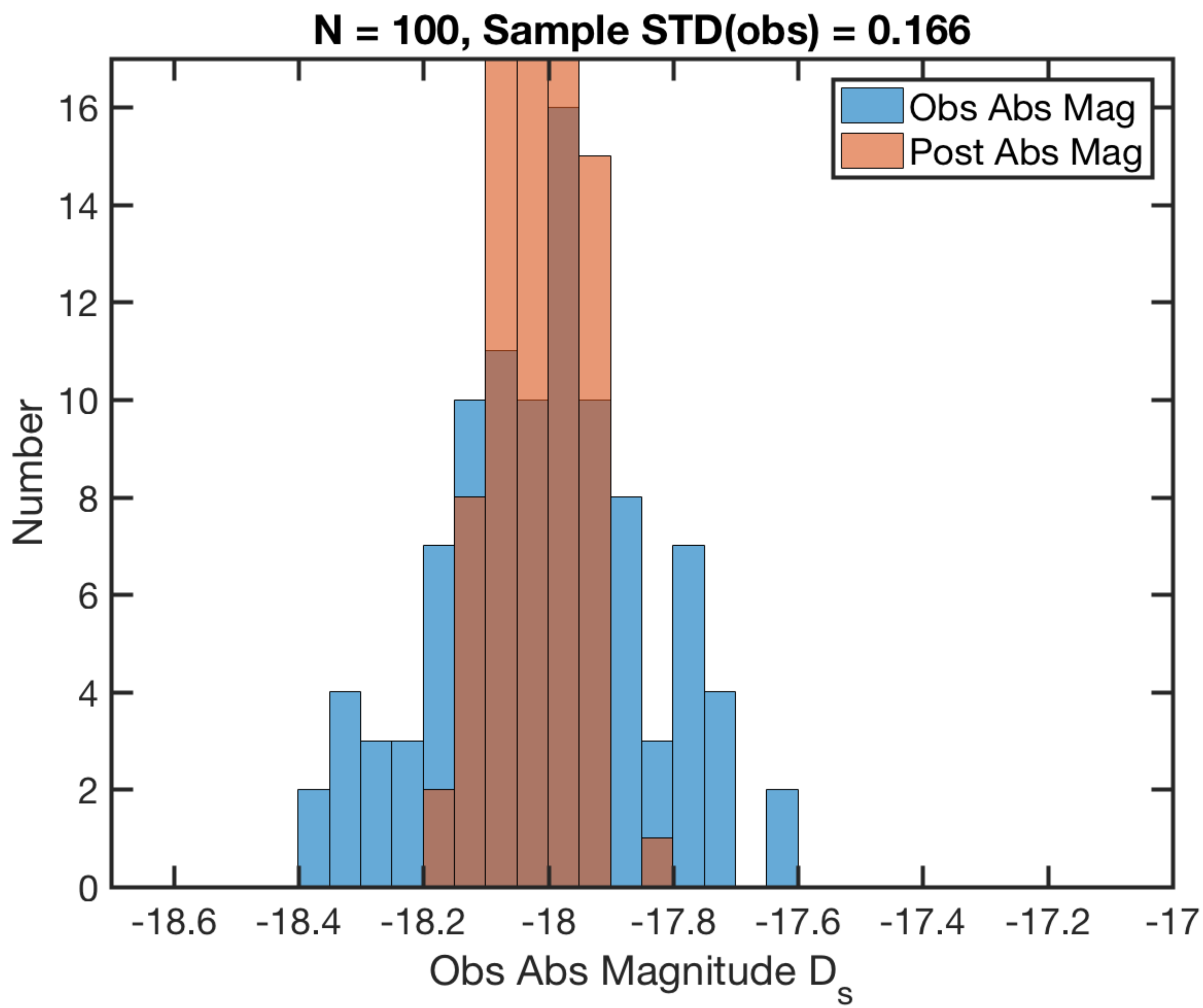
Latent Variable Inference

Histogram of $\mathbb{E}[M_s|\boldsymbol{D}]$ Estimates from MCMC



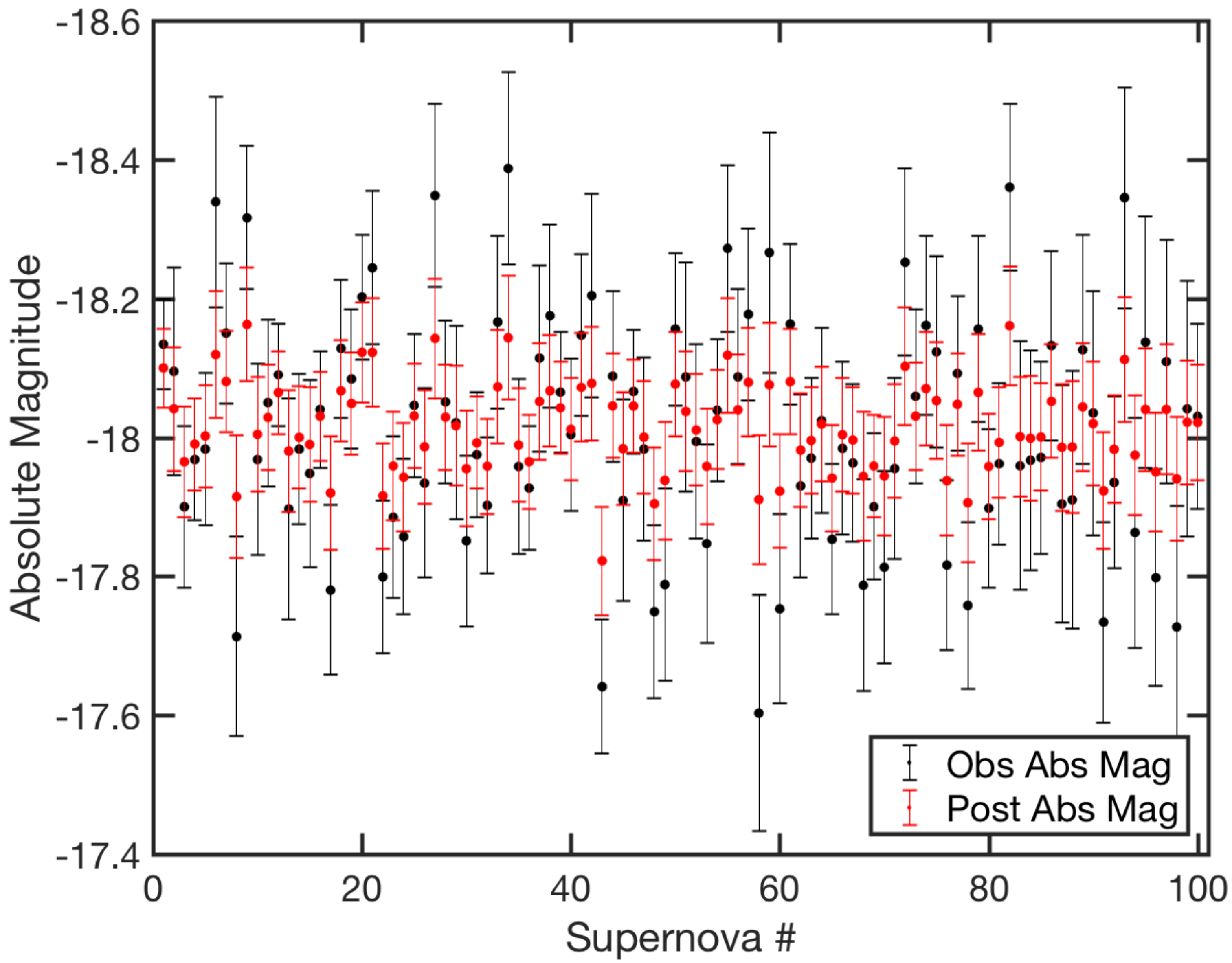
Latent Variable(s) Inference

Histogram of $\mathbb{E}[M_s|\boldsymbol{D}]$ posterior estimates from MCMC



Latent Variable(s) Inference

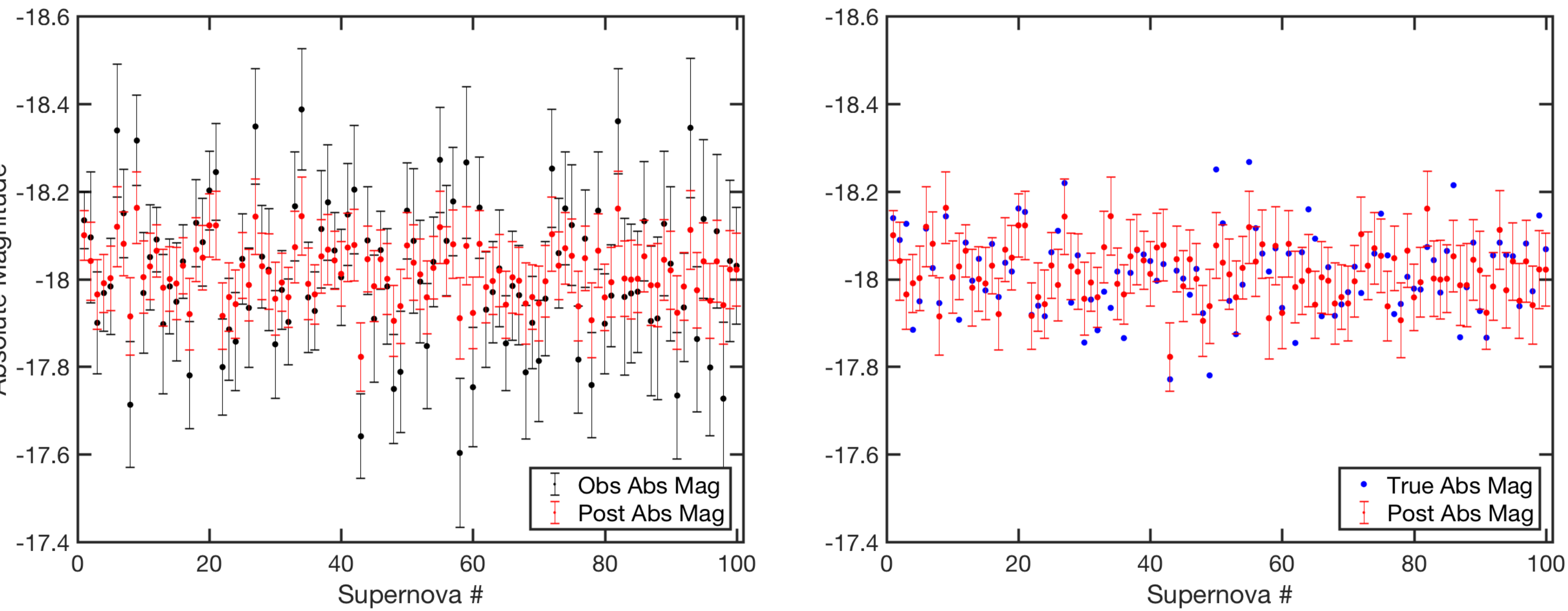
Histogram of $\mathbb{E}[M_s|\boldsymbol{D}]$ posterior estimates from MCMC



What's going on here?

Latent Variable(s) Inference

Comparing the Data, Posterior Latent Mags vs. true Mags



What's going on here?

HB models: Partial Pooling, Shrinkage, and "Borrowing of Strength"

- Common Sense Procedure:
  - Analyze each individual object's data D_i separately and get each individual MLE_i estimate (with error)
  - "Plug-in" all $\{MLE_i\}$ to estimate population hyperparameters

- Problem: Each individual θ_i estimate may be unbiased but collectively give a biased estimate of population (e.g. variance).
- Solution: Use HB to model and infer individuals & population simultaneously and get better estimates of both

Aside on Shrinkage (derivation on board)

HB models: Partial Pooling, Shrinkage, and "Borrowing of Strength"

Loredo & Hendry (2010)

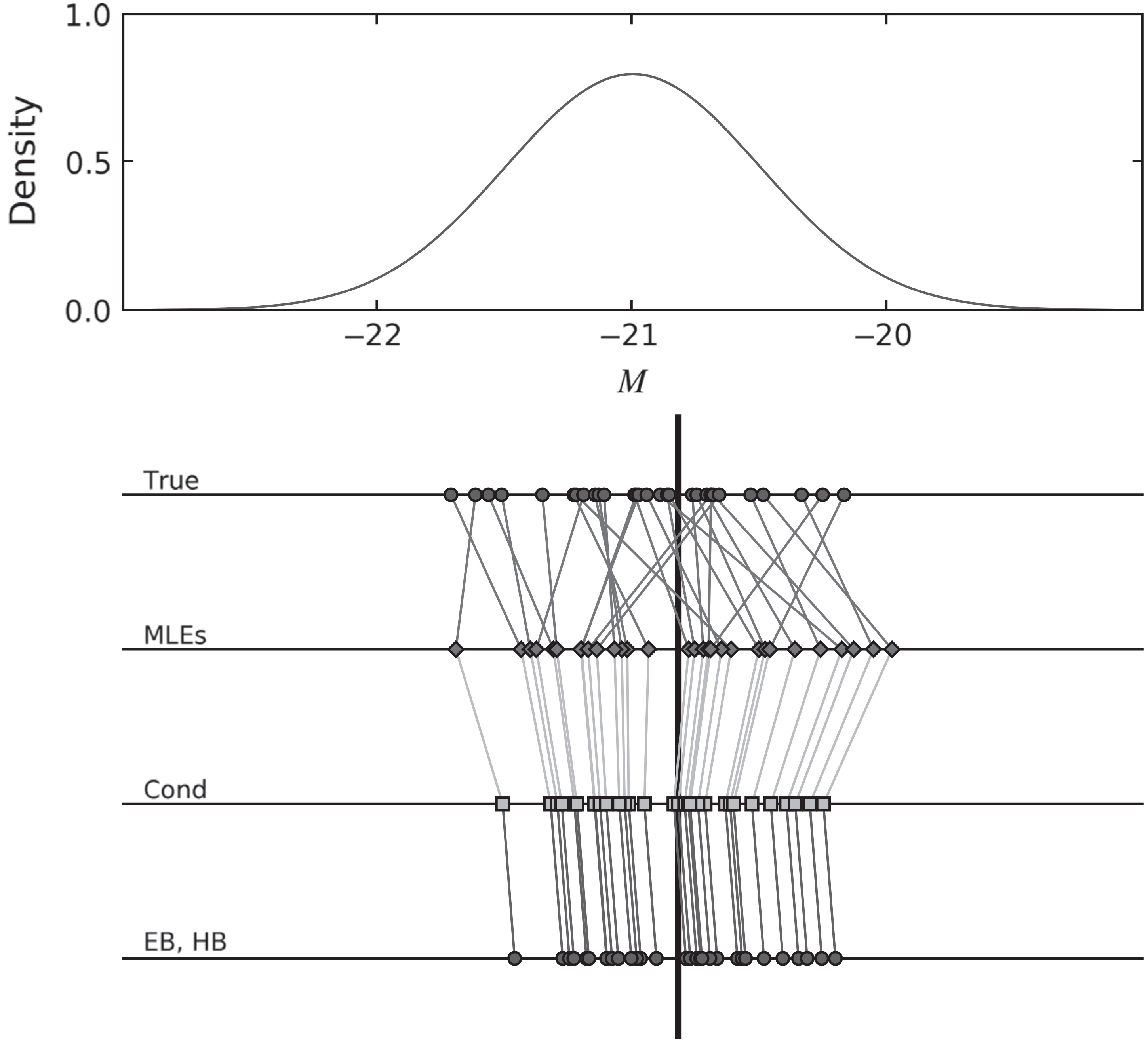
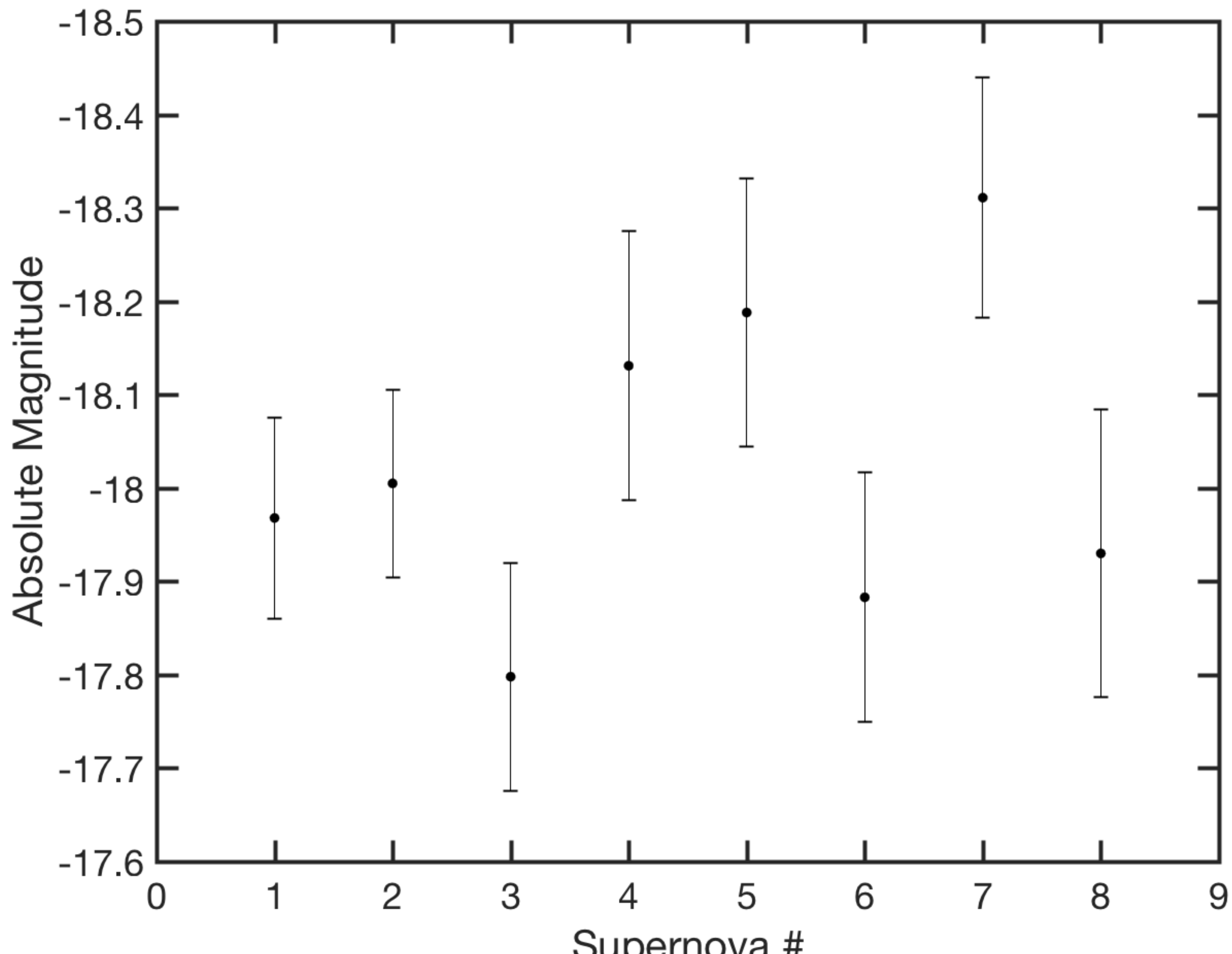


Fig. 11.2. Shrinkage in a simple normal–normal model. Top panel shows population distribution. 'True' axis shows M_i values of 30 samples. Remaining axes show estimates from measurements with $\sigma = 0.3$ normal error: MLEs, conditional (on the true mean), and empirical/hierarchical Bayes estimates.

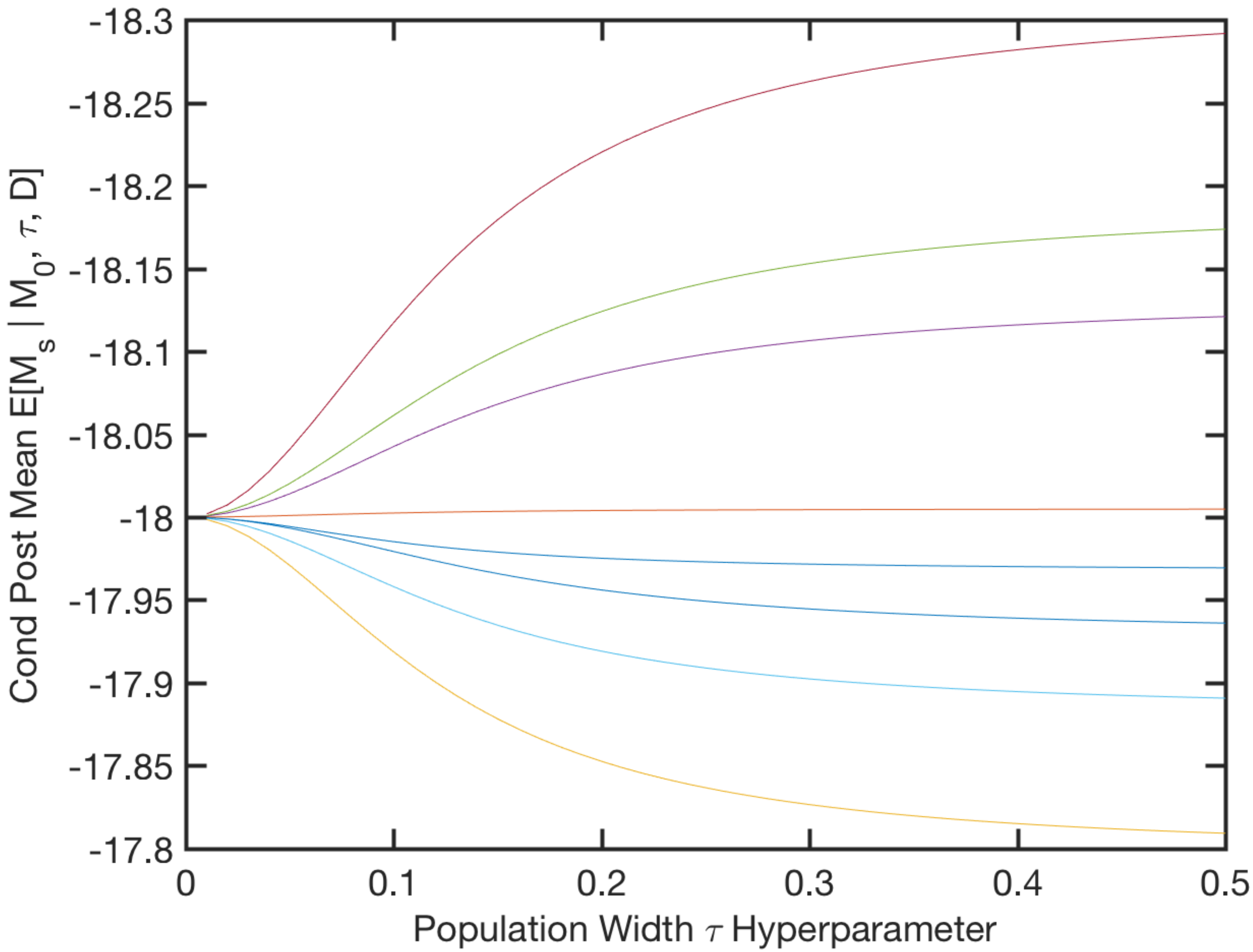
"Bayesian Multi-level modeling of cosmological populations" in Bayesian Methods in Cosmology

How does Hierarchical Bayes implement shrinkage?

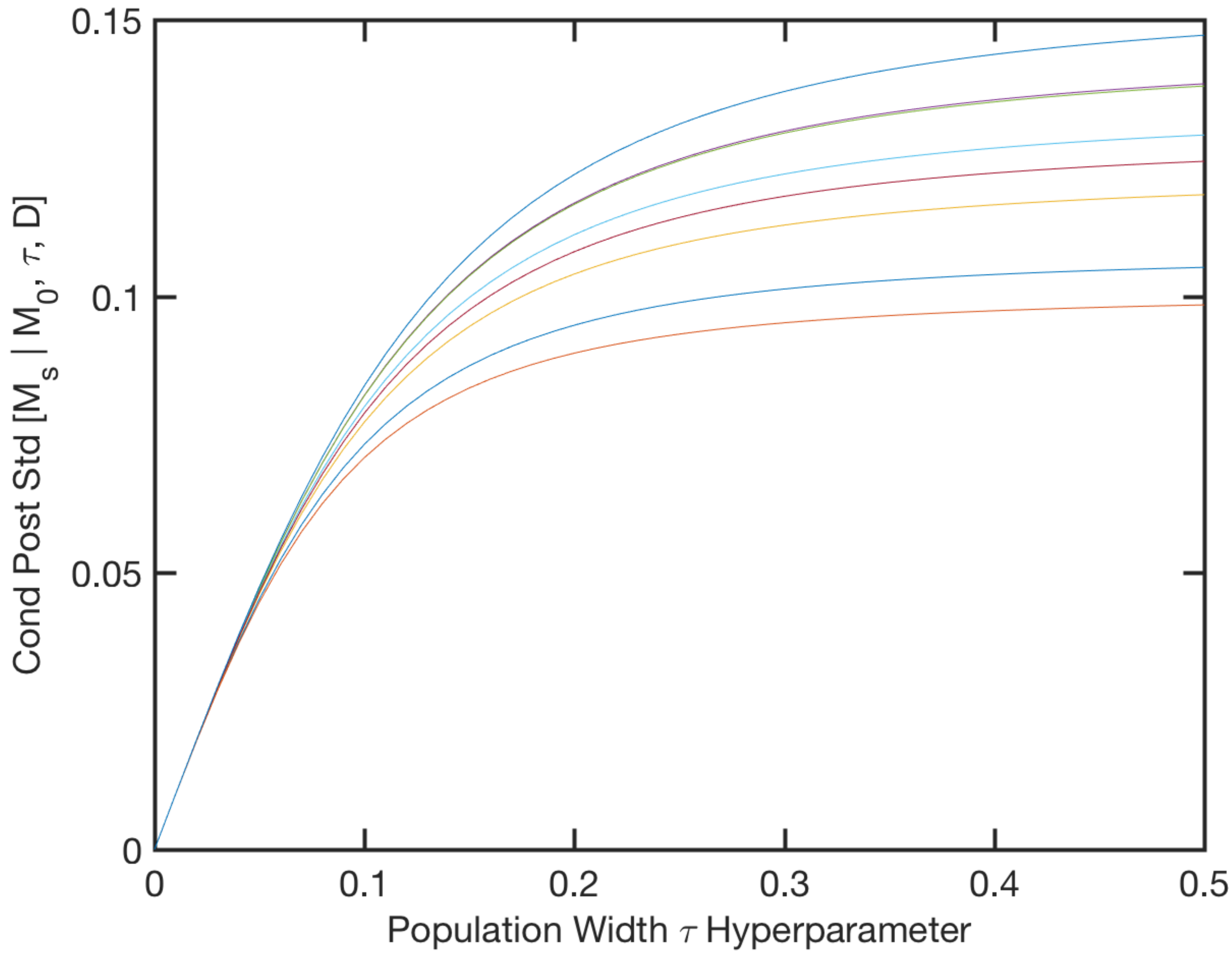
First, let's shrink the dataset



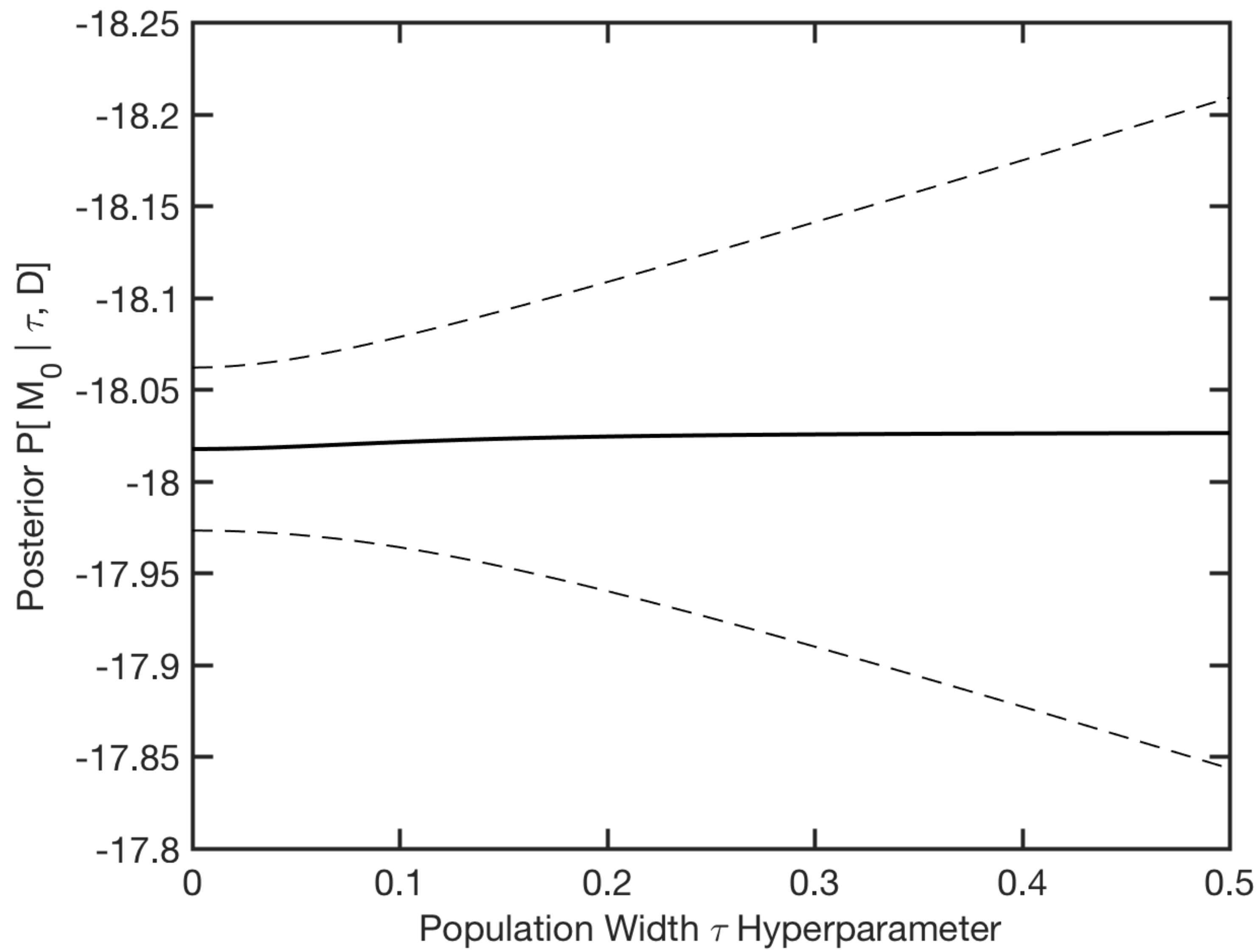
How does Hierarchical Bayes implement shrinkage?



How does Hierarchical Bayes implement shrinkage?



How does Hierarchical Bayes implement shrinkage?



How does Hierarchical Bayes implement shrinkage?