

Astrostatistics: 11 Mar 2020

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

- Example Class 3: Fri 13 Mar, 12pm, MR5
- MCMC tutorial: Speagle: “A Conceptual Introduction to Markov Chain Monte Carlo Methods” <https://arxiv.org/abs/1909.12313>
- Today: Last Lecture!
 - Hierarchical Bayesian Models / Shrinkage
 - Bayesian Model Comparison
 - MacKay, “Information Theory, Inference, and Learning Algorithms” <http://www.inference.org.uk/itila/book.html>
 - Bishop, “Pattern Recognition & Machine Learning” : <https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book>
 - Wall & Jenkins “Practical Statistics for Astronomers”
(free on iDiscover)

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

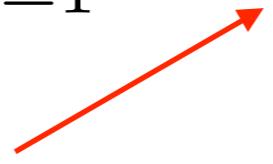
Joint Probability Density of ALL THE THINGS:

Data, Latent Variables, Hyperparameters

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

(Derive on board)

Measurement Likelihood



Population Dist'n / Prior

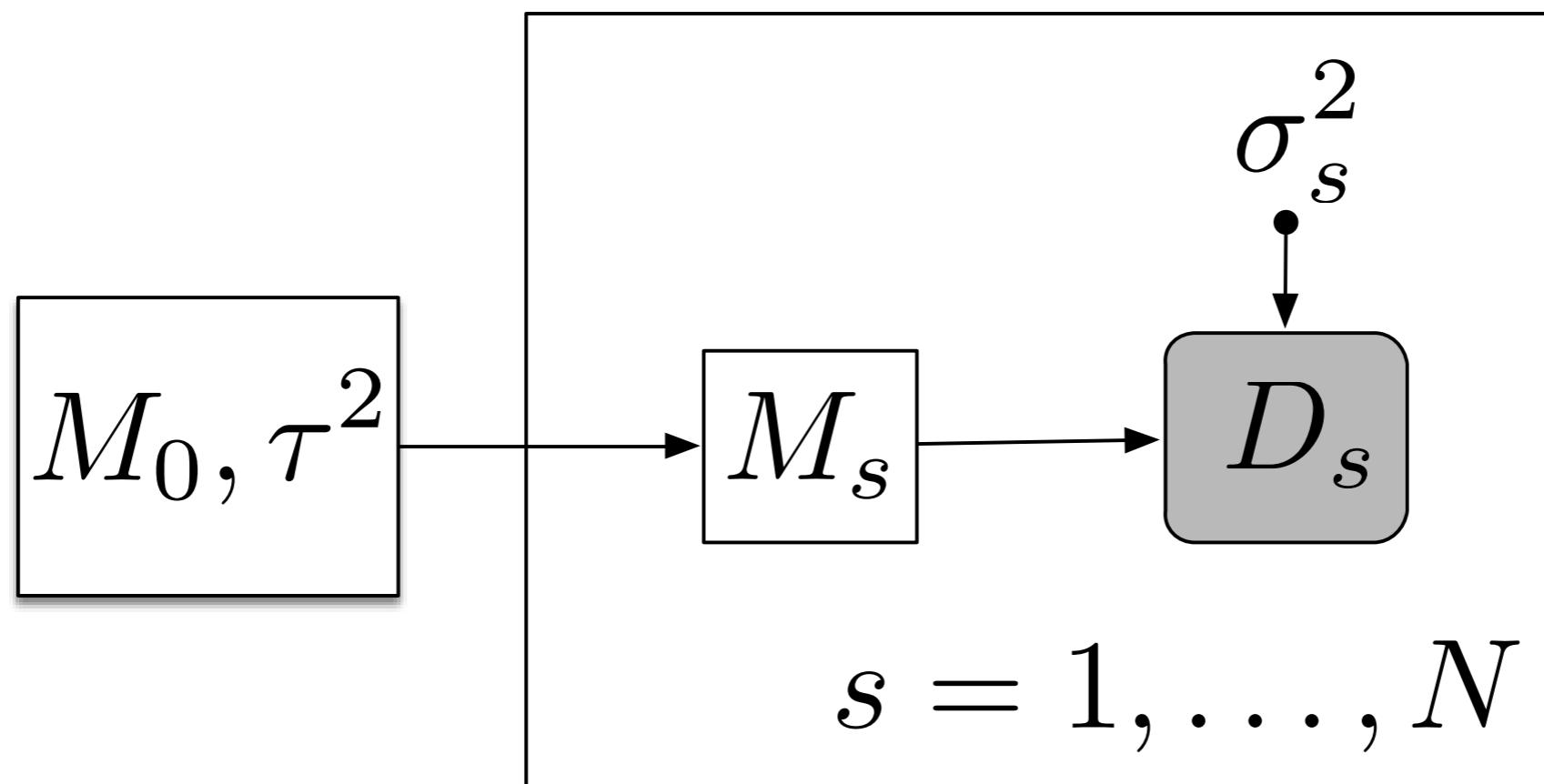


Hyperprior



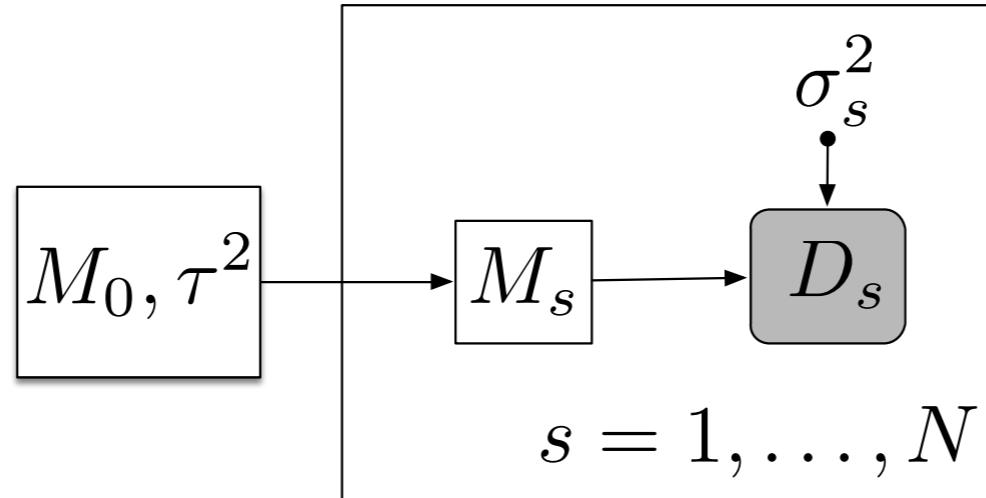
Sampling the Hierarchical Bayesian posterior

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

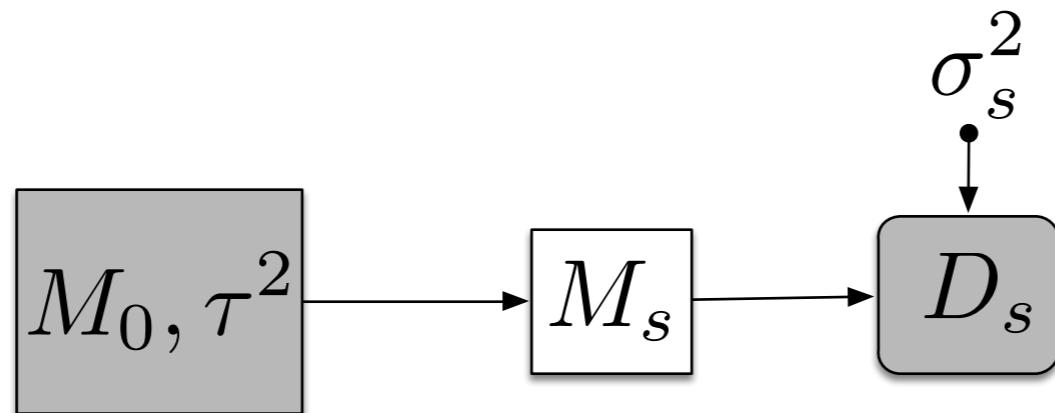


Utilise the Conditional Independence structure of PGM
to derive conditional posterior densities

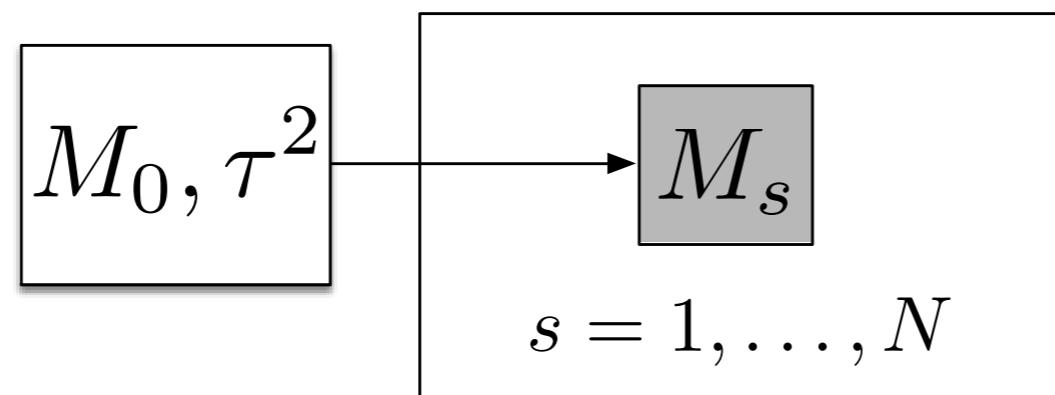
Gibbs sampling & Hierarchical Bayes



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



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

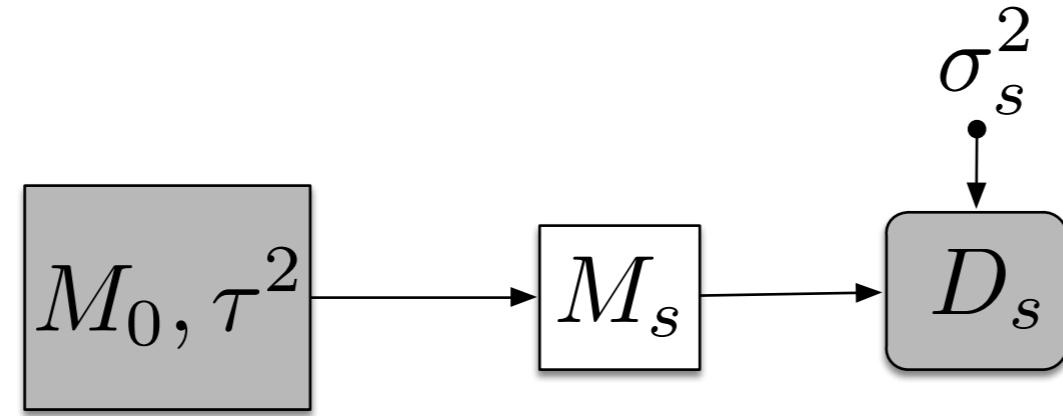


Gibbs sampling & Hierarchical Bayes

Utilises Conditional Independence structure of PGM to derive conditional posterior densities

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

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



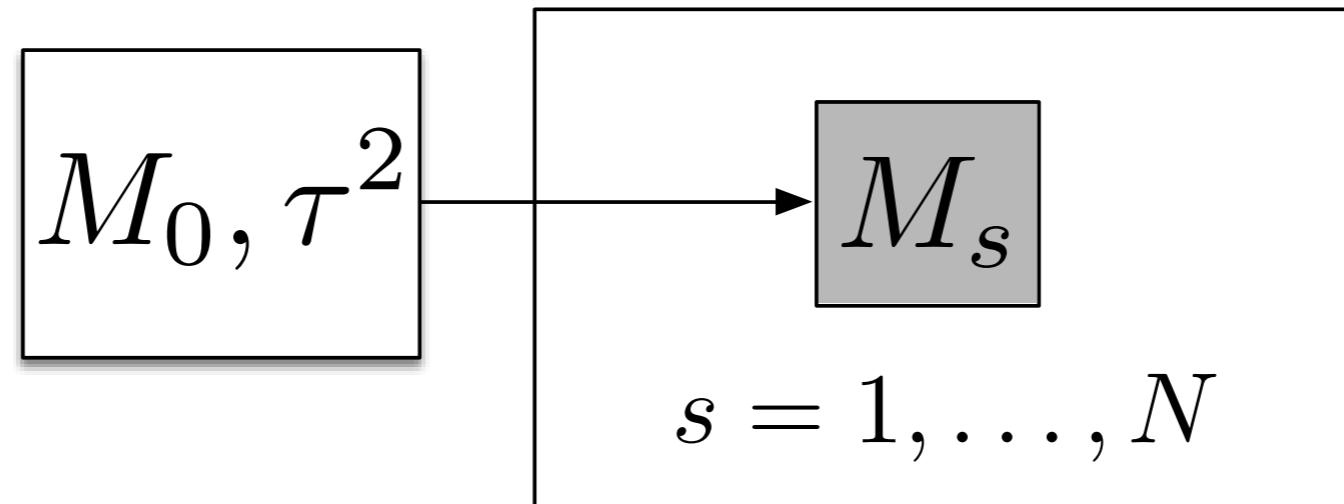
$$\begin{aligned} P(M_s | M_0, \tau^2, \{D_s\}) &\propto P(D_s | M_s) \times P(M_s | M_0, \tau^2) \\ &\propto N(D_s | M_s, \sigma_s^2) \times N(M_s | M_0, \tau^2) \end{aligned}$$

Gibbs sampling & Hierarchical Bayes

Utilises Conditional Independence structure of PGM to derive conditional posterior densities

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

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



Hyperprior: $P(M_0, \tau^2) = P(M_0)P(\tau^2) \propto 1$ $P(M_0) \propto 1$ $P(\tau^2) \propto 1$

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

(See Example Sheet 2, Problem 1)

(Gaussian)

(Inv- χ^2)

Gibbs sampling & Hierarchical Bayes

Utilises Conditional Independence structure of PGM to derive conditional posterior densities

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

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

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

Hyperprior: $P(M_0, \tau^2) \propto 1$ (Gaussian) (Inv- χ^2)

Draw from: $\tau^2 | \{M_s\} \sim \text{Inv-}\chi^2 \left(N - 3, \frac{(N - 1)}{(N - 3)} s^2 \right)$

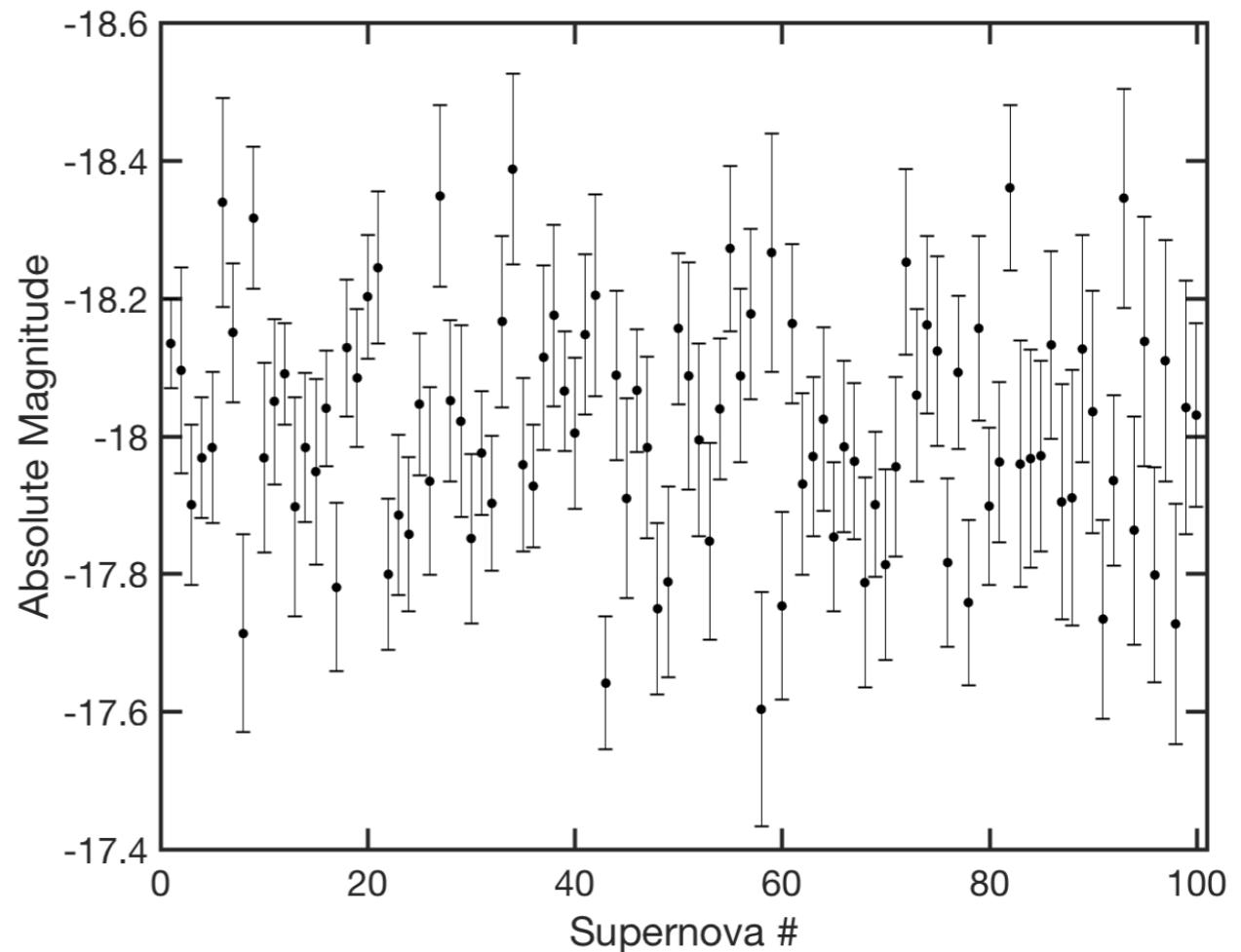
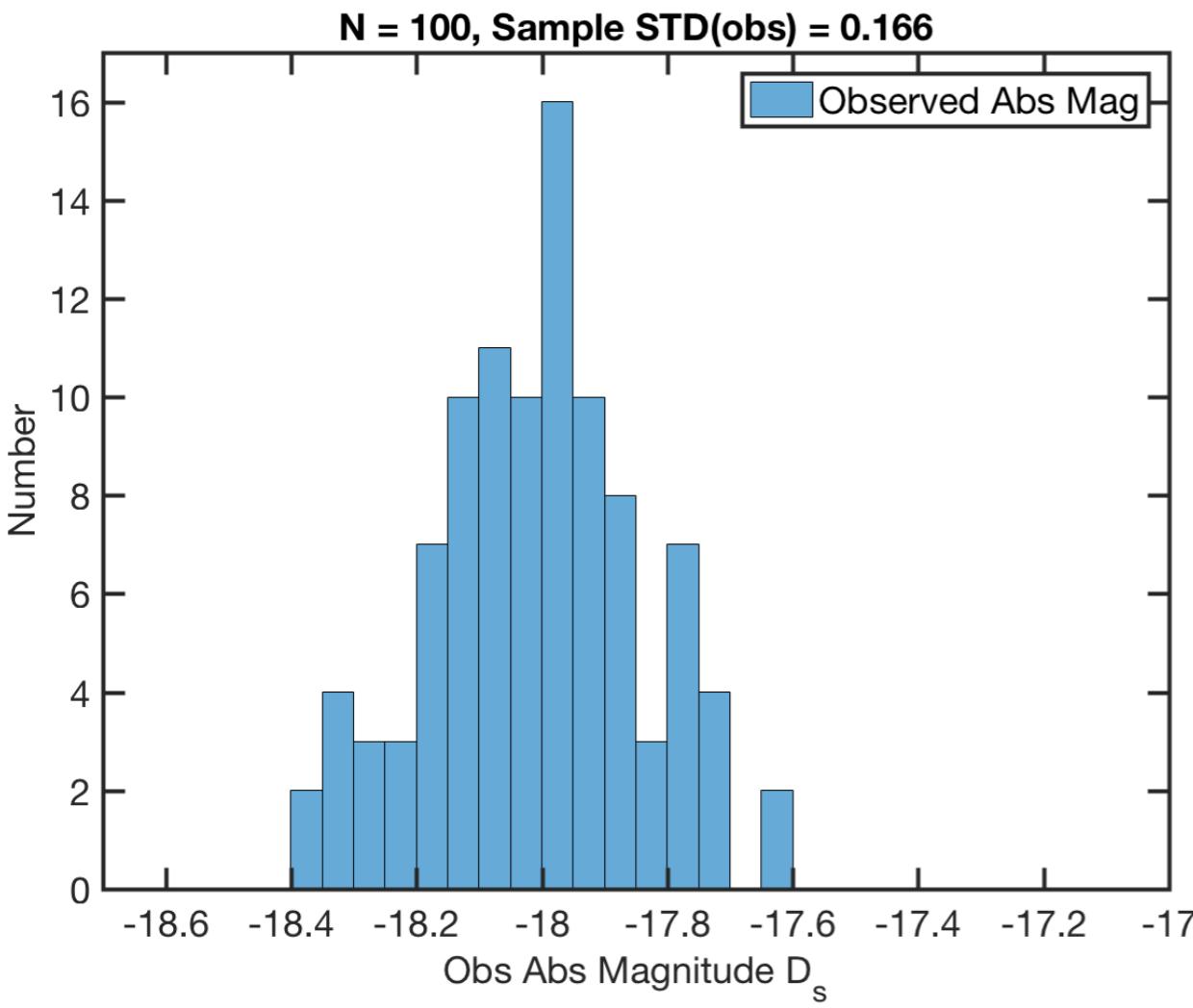
$$M_0 | \tau^2; \{M_s\} \sim N(\bar{M}, \tau^2/N)$$

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

Data Example

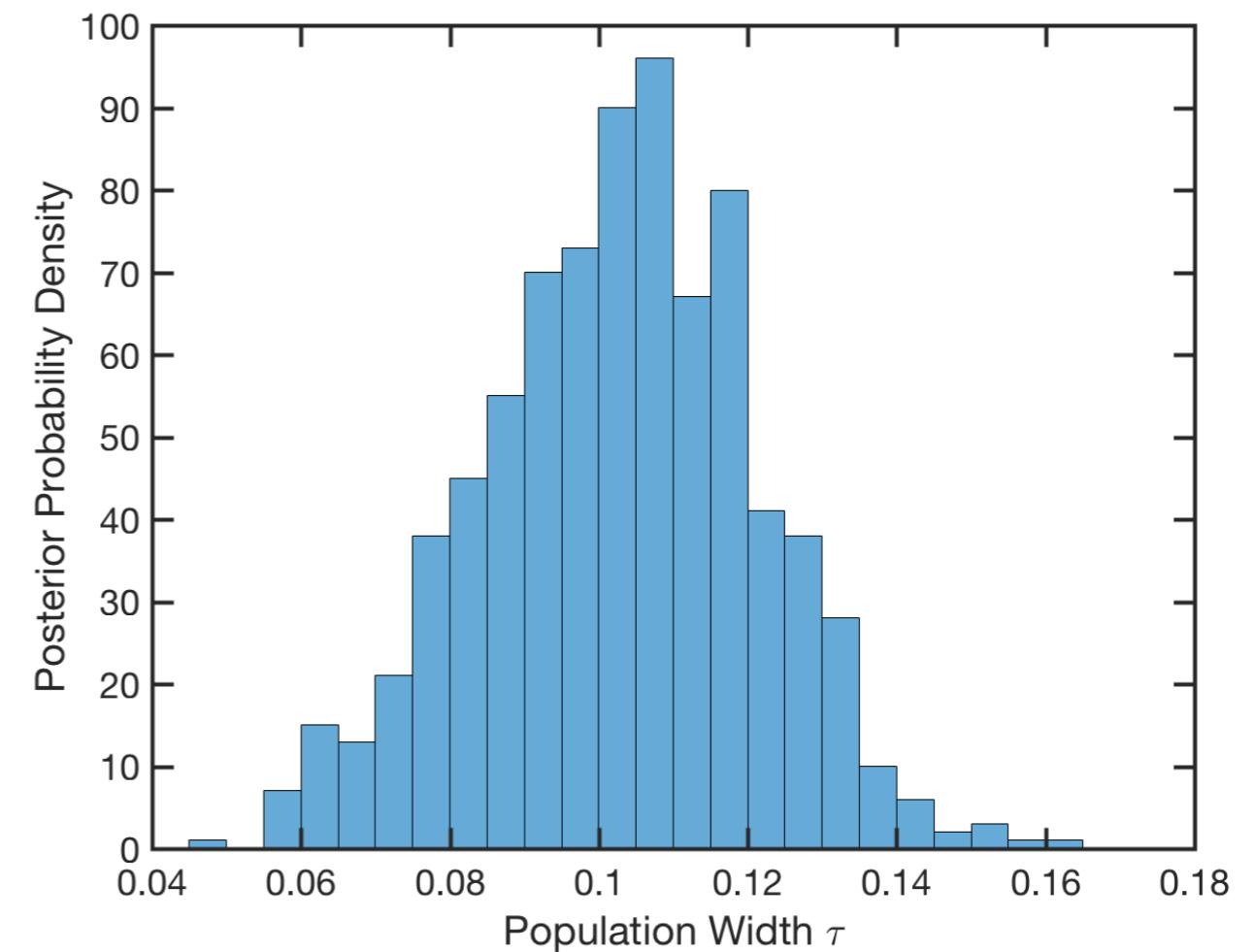
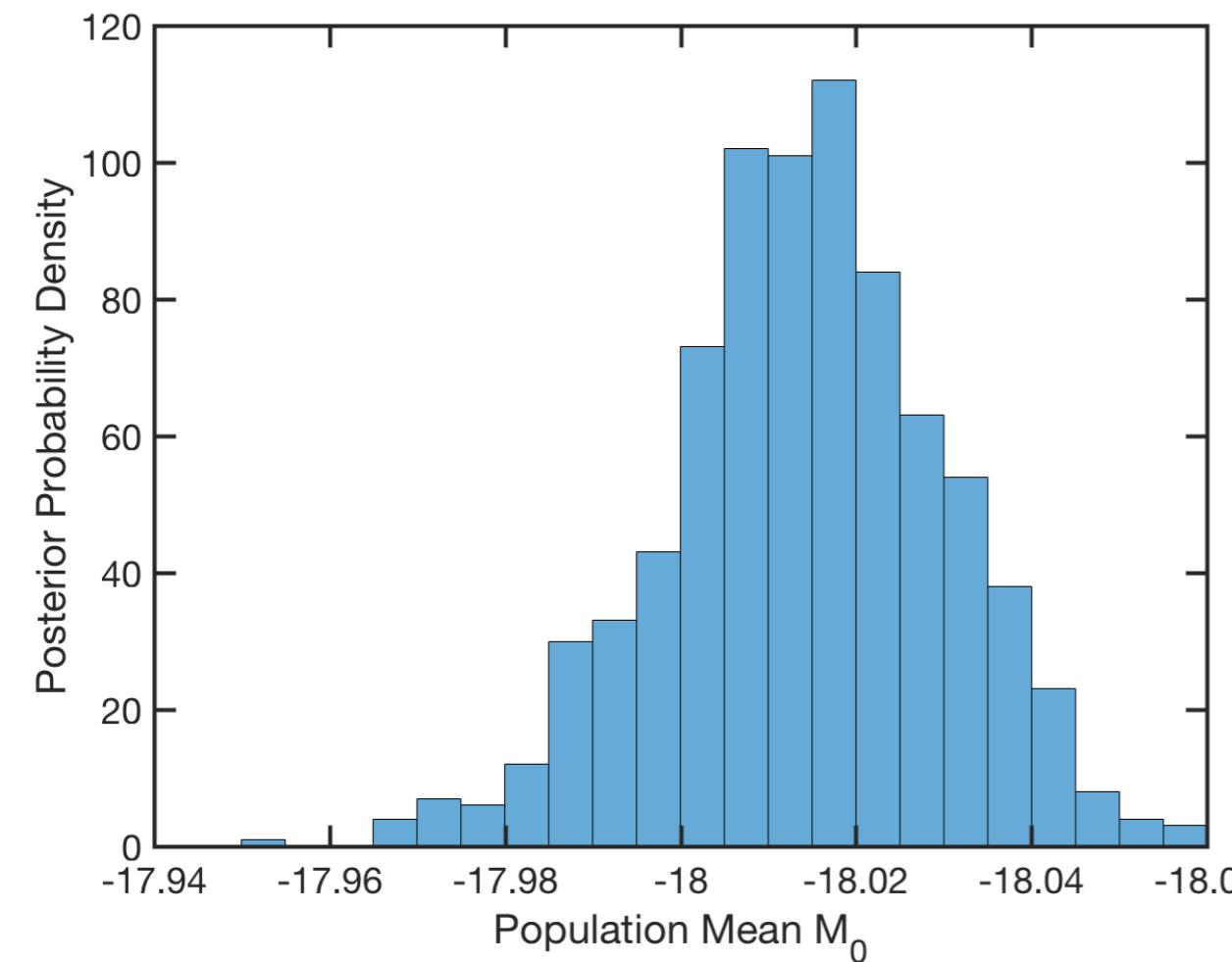
100 Supernovae with
app magnitude and
distance measurements

$$D_s = \hat{m}_s - \hat{\mu}_s$$



Hyperparameter Inference

Marginal Posterior Histogram Estimates
from Gibbs Sampling in 102 dimensions

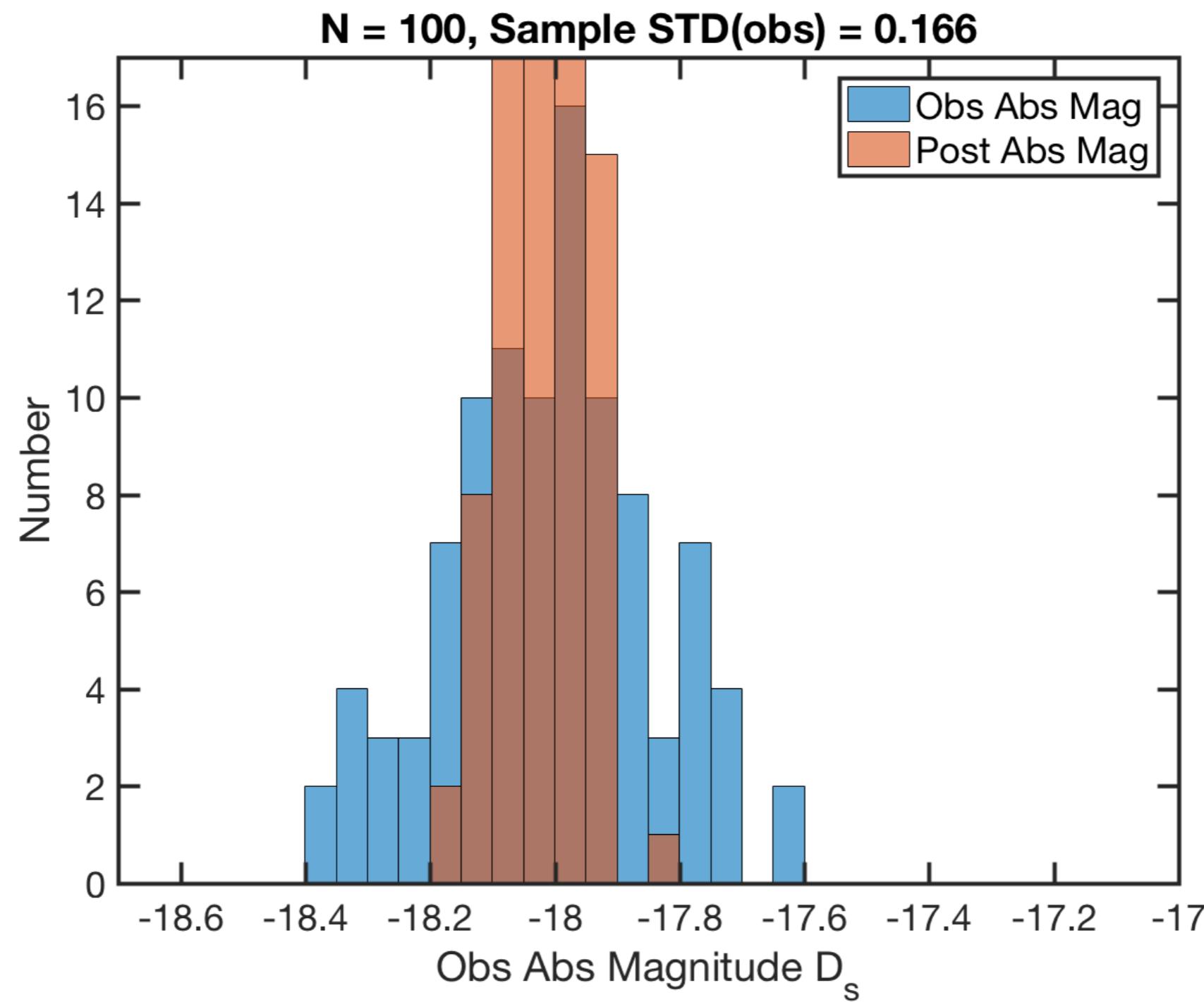


M_0

τ

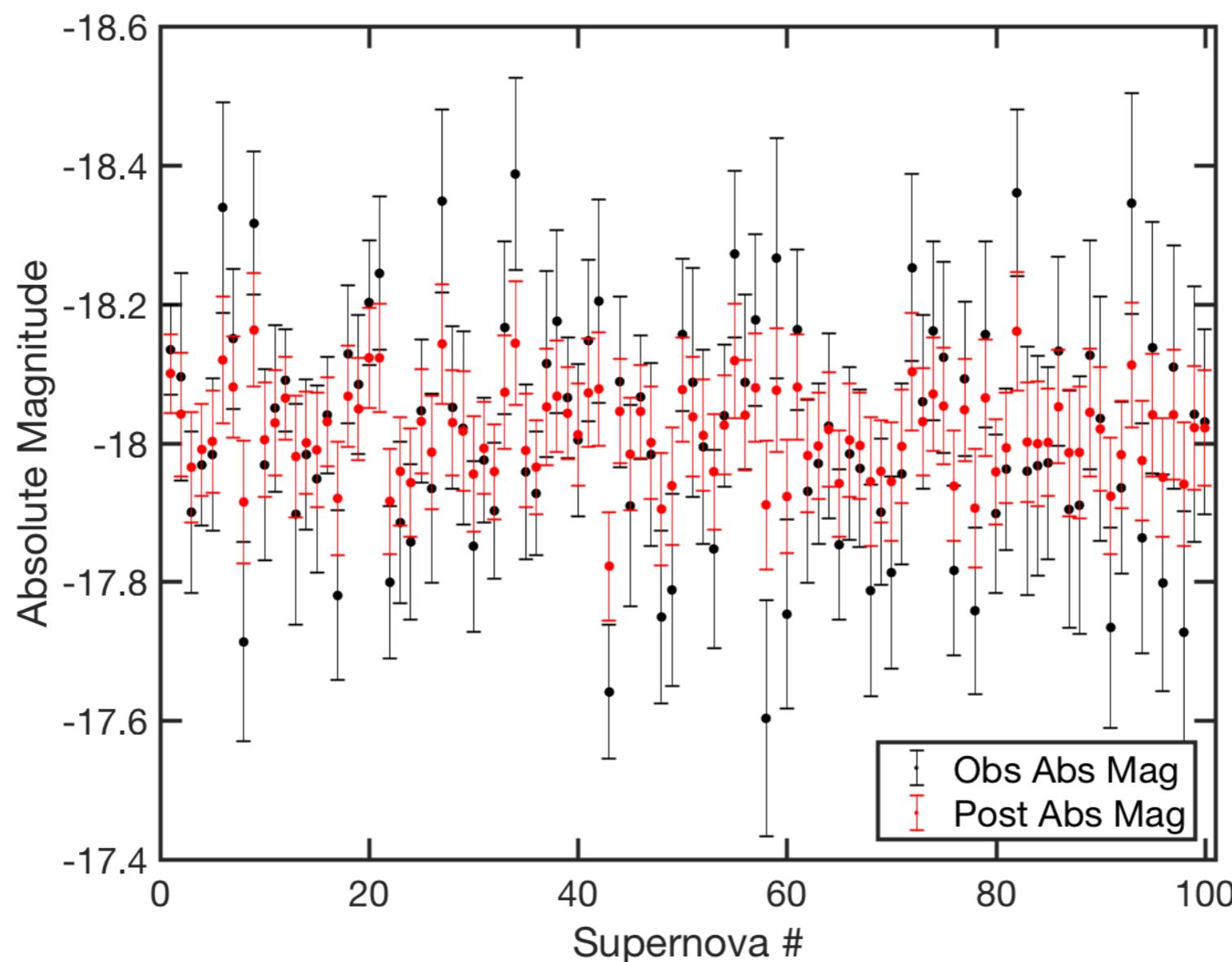
Latent Variable(s) Inference

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



Latent Variable(s) Inference

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

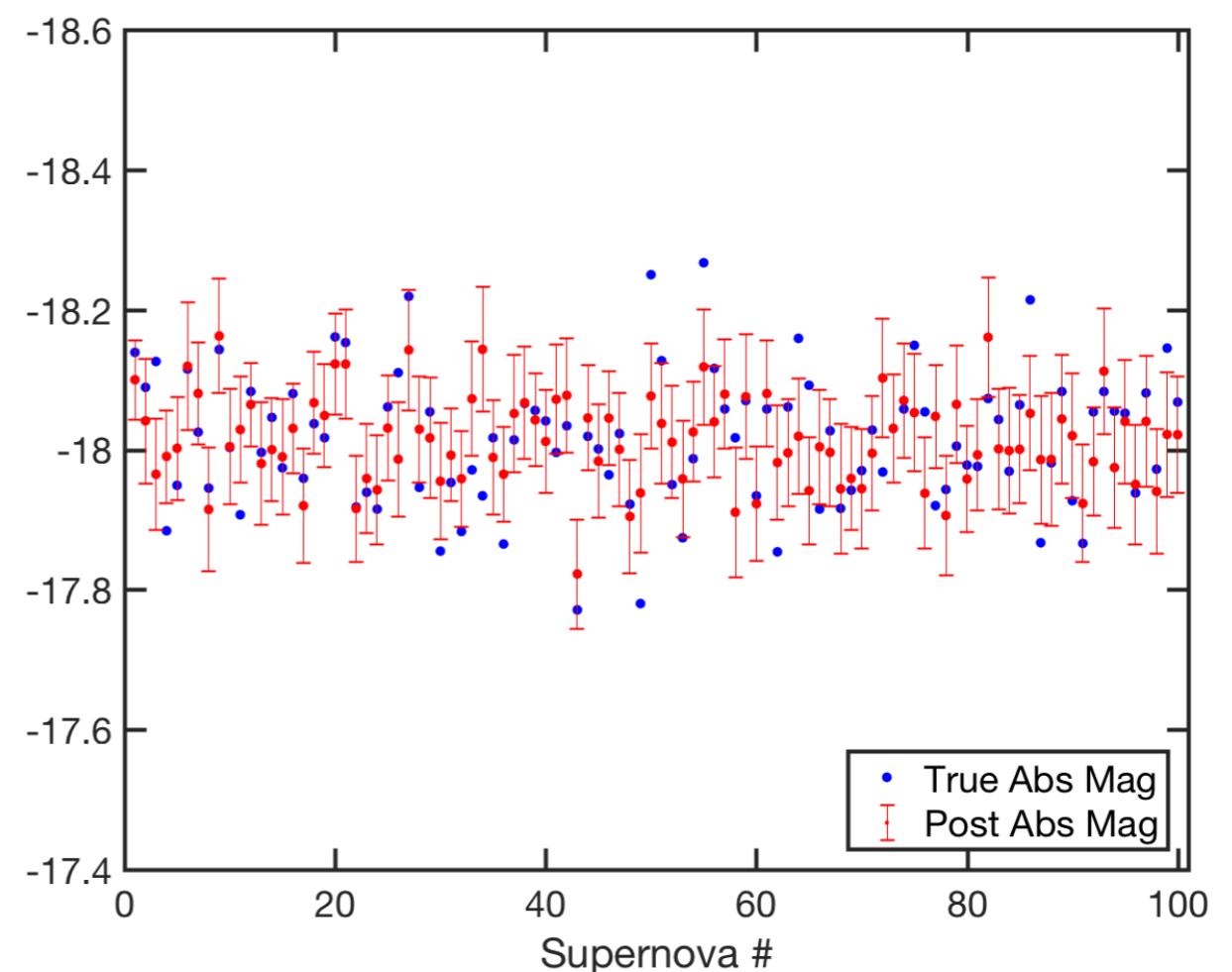
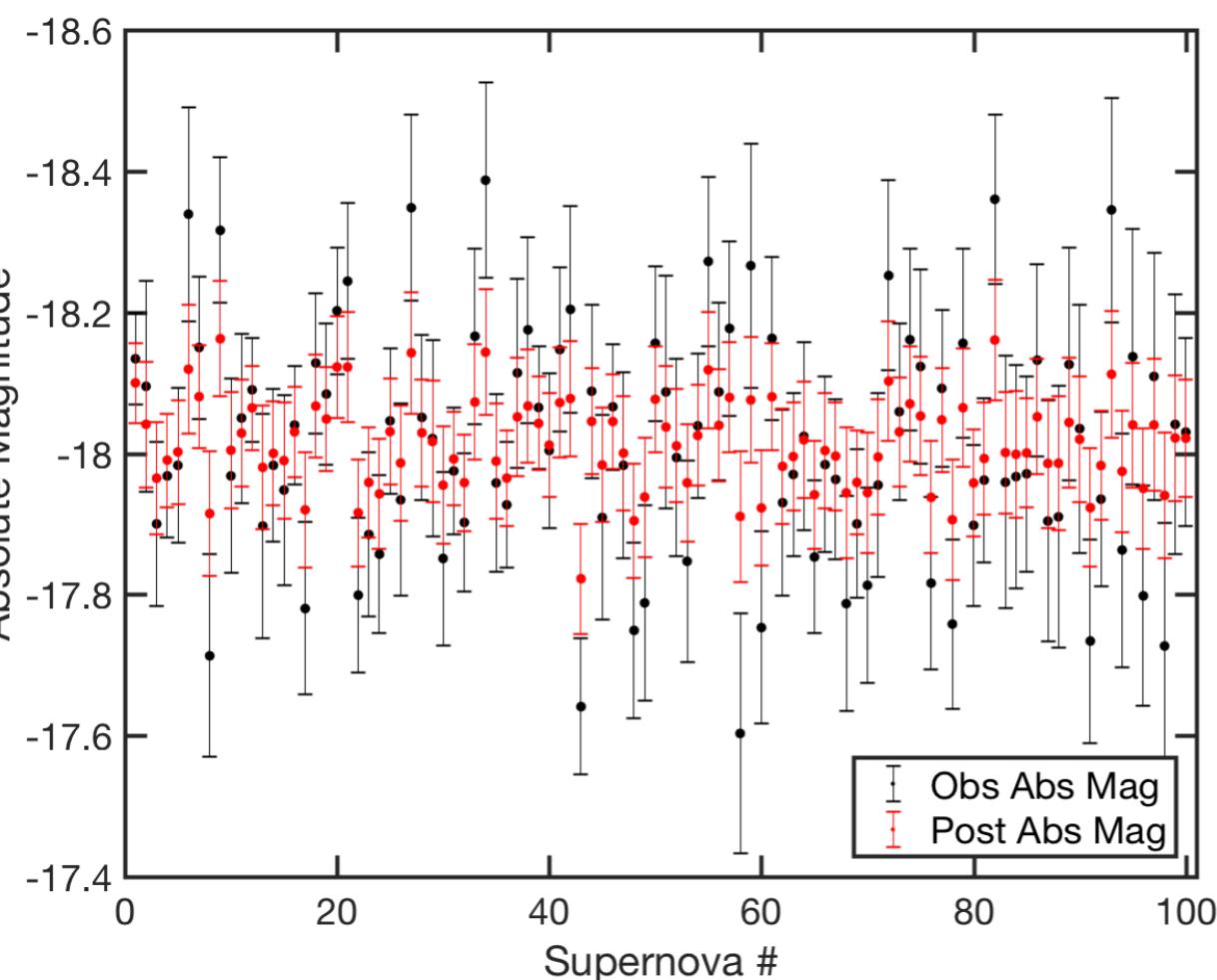


What's going on here?

Latent Variable(s) Inference

$$P(M_s | D)$$

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 $\{\text{MLE}_i\}$ 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

Shrinkage Estimators

- Bias estimator of individual to towards the population
- Leads to overall lower MSE than individual unbiased estimators
- Allows “sharing of information” between individuals to improve overall estimation

HB models: Partial Pooling, Shrinkage, and “Borrowing of Strength”

11.3 Multilevel models

255

Loredo & Hendry (2010)
arXiv:1911.12337

“Bayesian Multi-level modeling of cosmological populations”
in Bayesian Methods in Cosmology

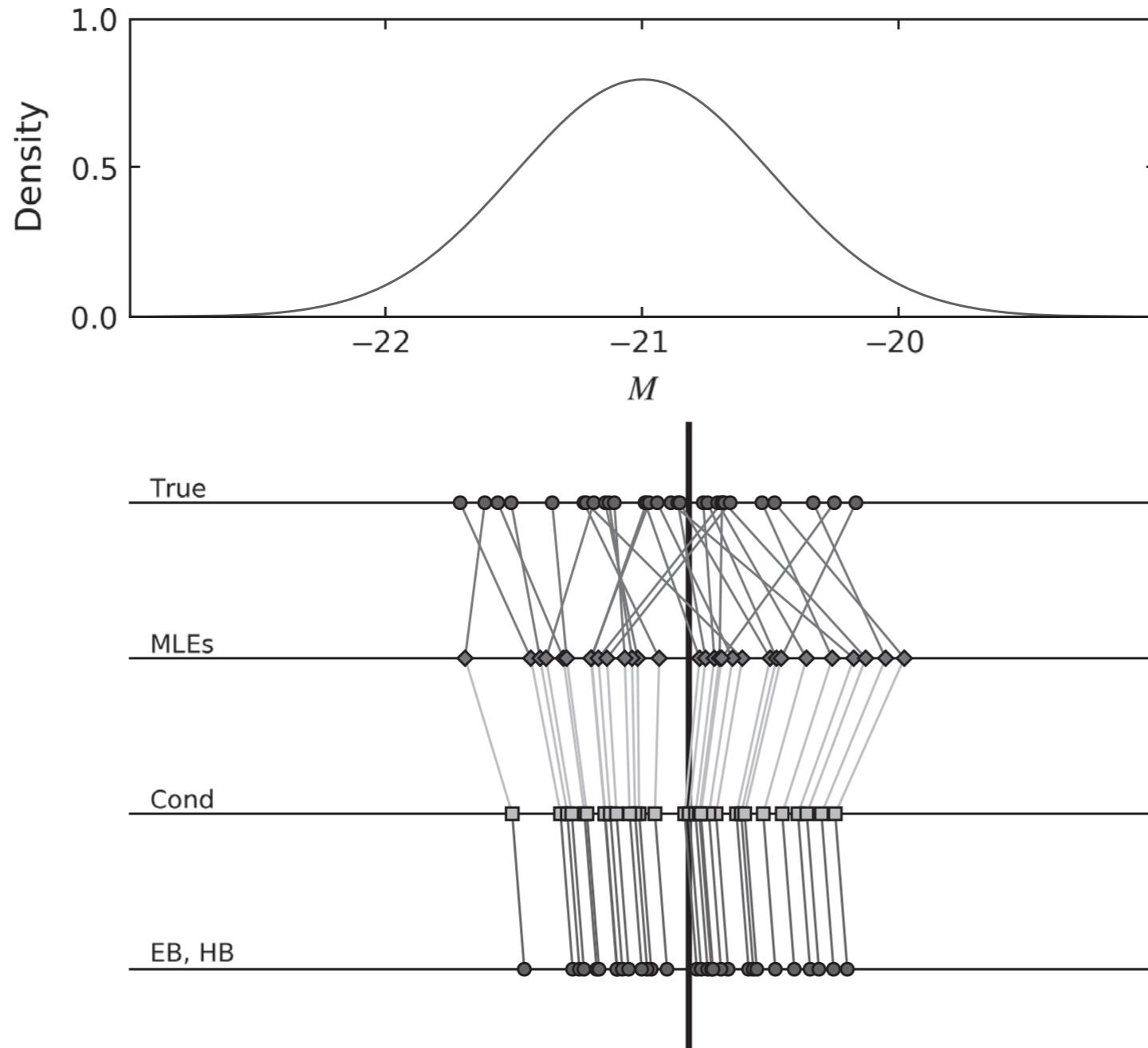
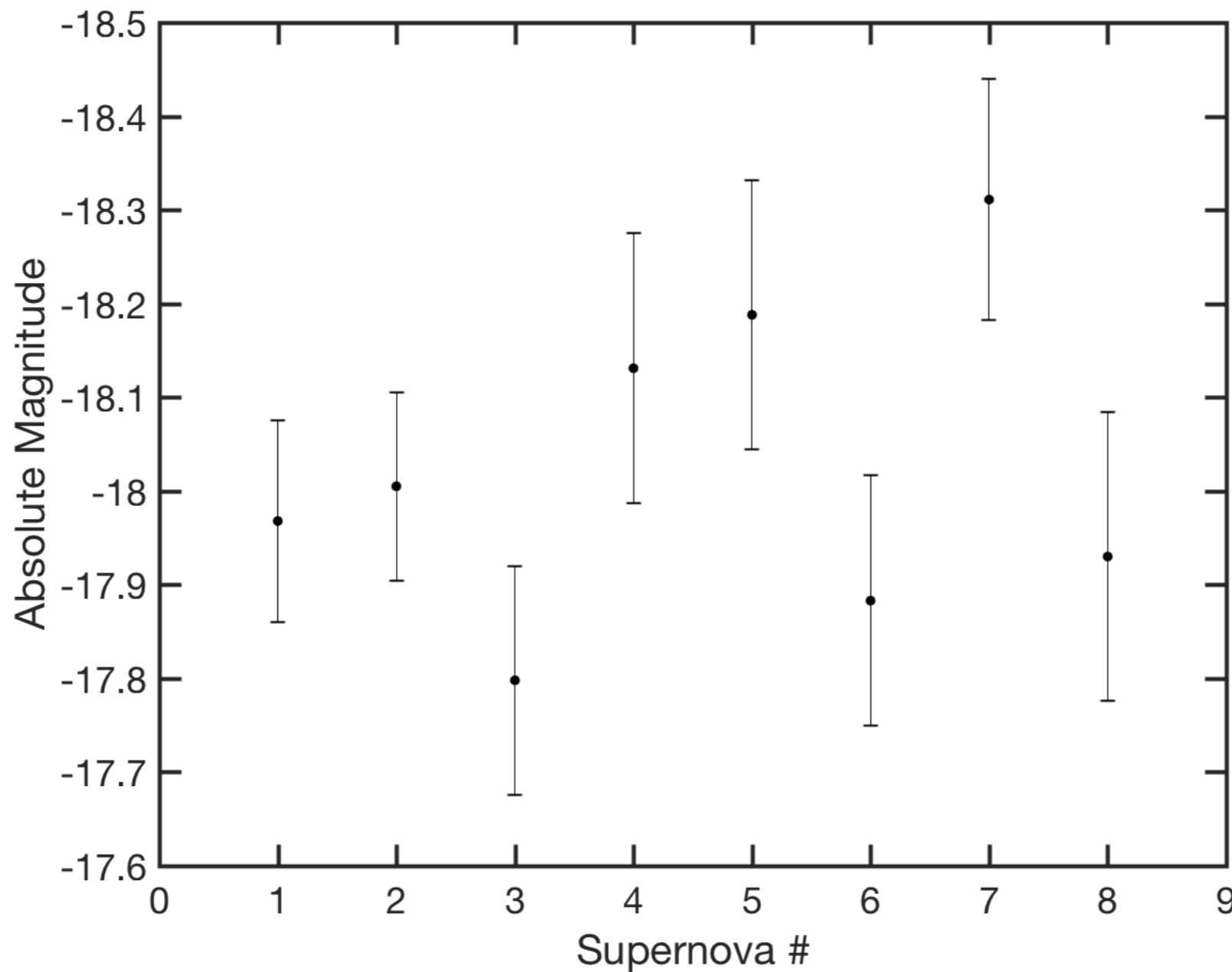


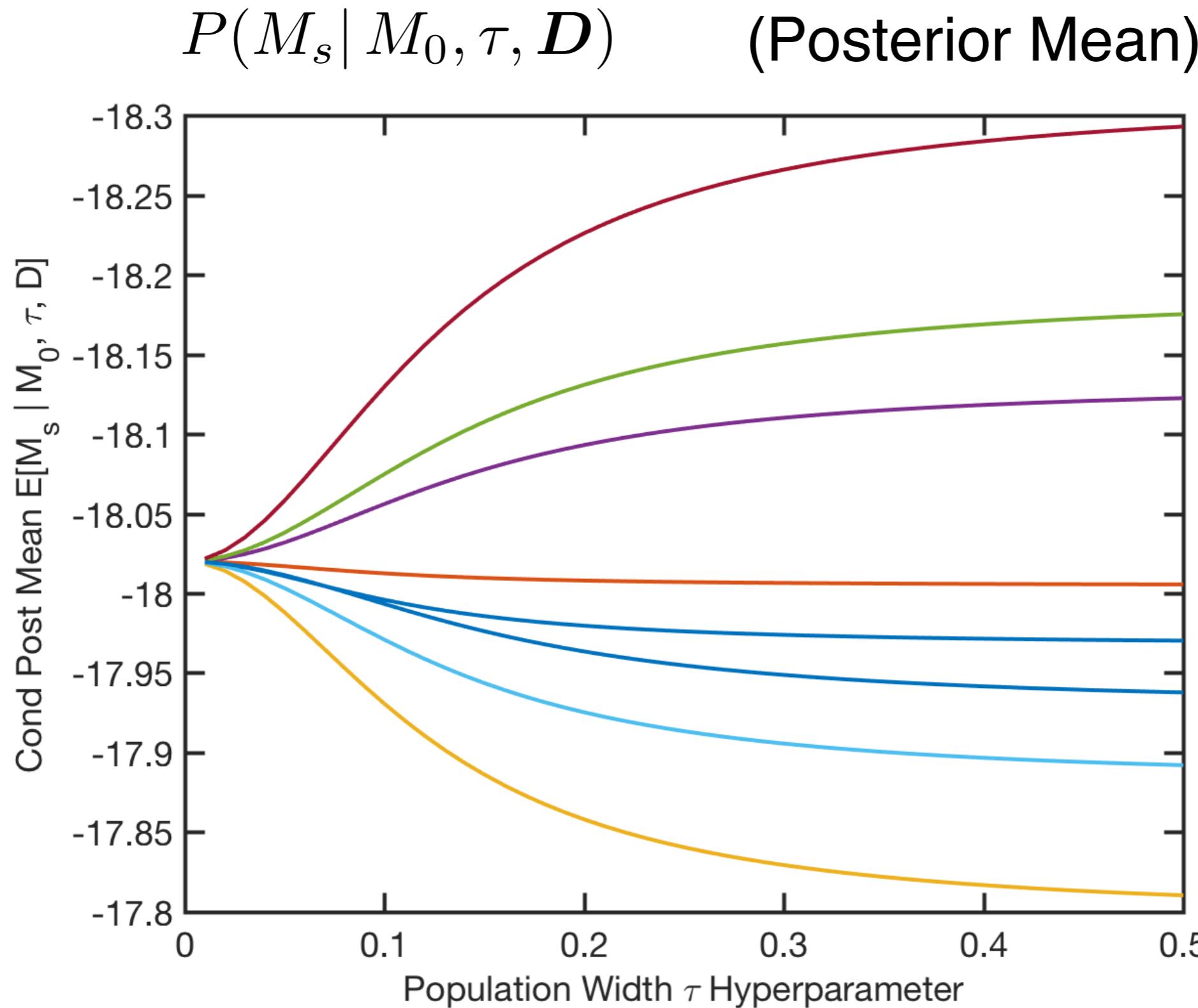
Fig. 11.2. Shrinkage in a simple normal–normal model. Top panel shows population distribution. ‘True’ axis shows M_i values of 30 samples. Remaining axes show estimates from measurements with $\sigma = 0.3$ normal error: MLEs, conditional (on the true mean), and empirical/hierarchical Bayes estimates.

How does Hierarchical Bayes implement shrinkage?

First, let's shrink the dataset to 8 supernovae

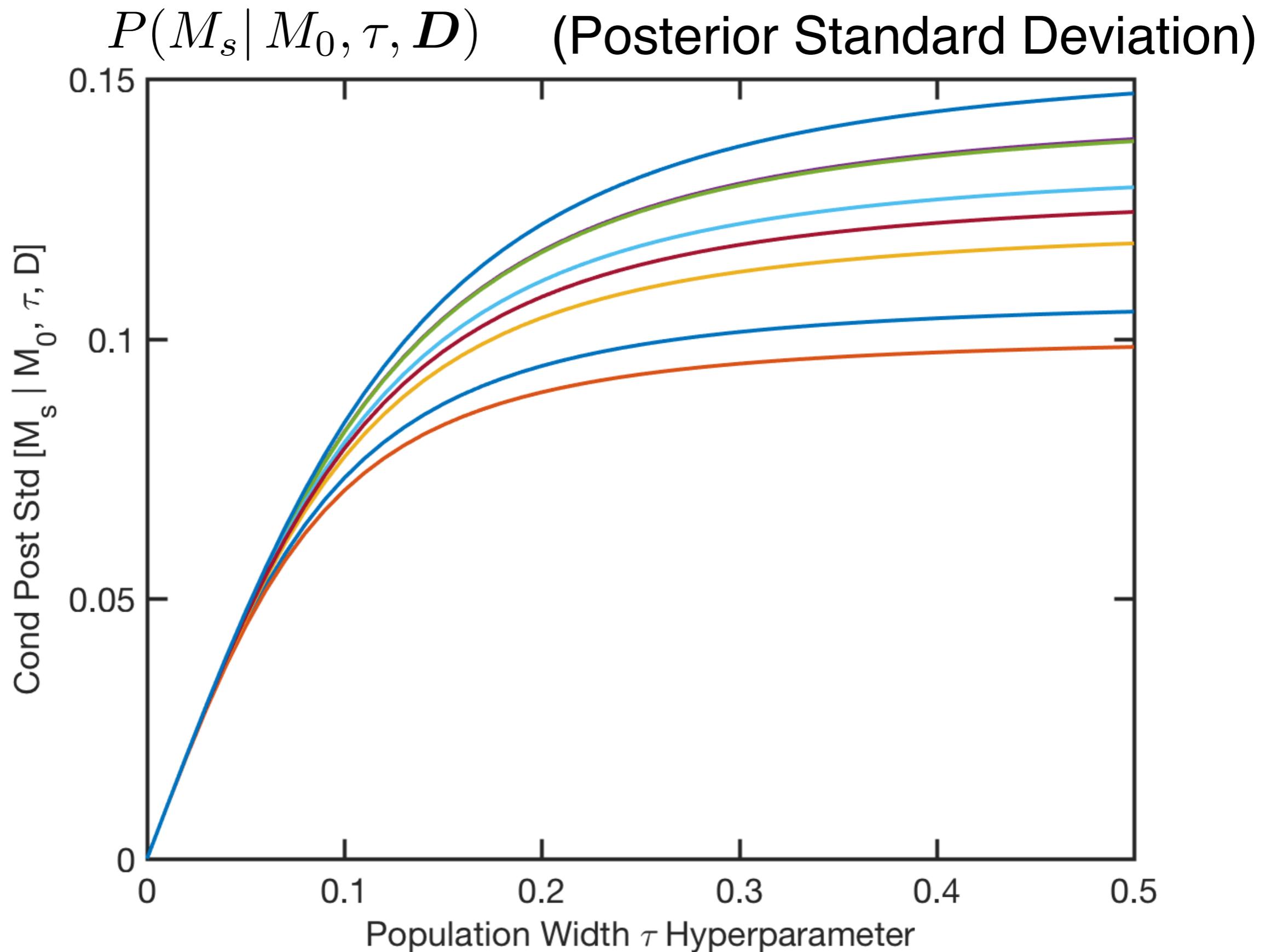


If you know M_0, τ : conditional posterior of individual M_s



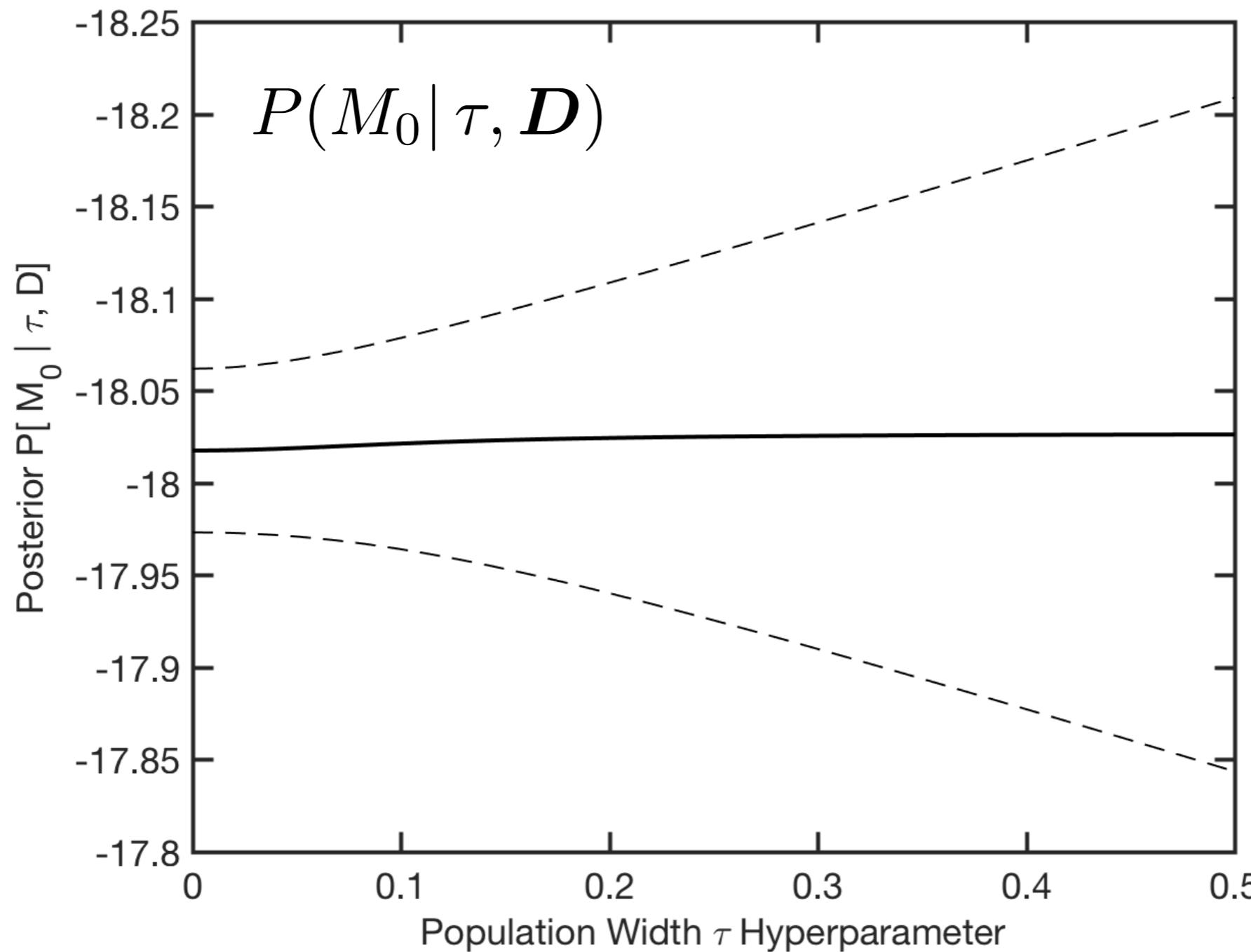
τ controls degree of shrinkage toward M_0

Conditional posterior of individual M_s

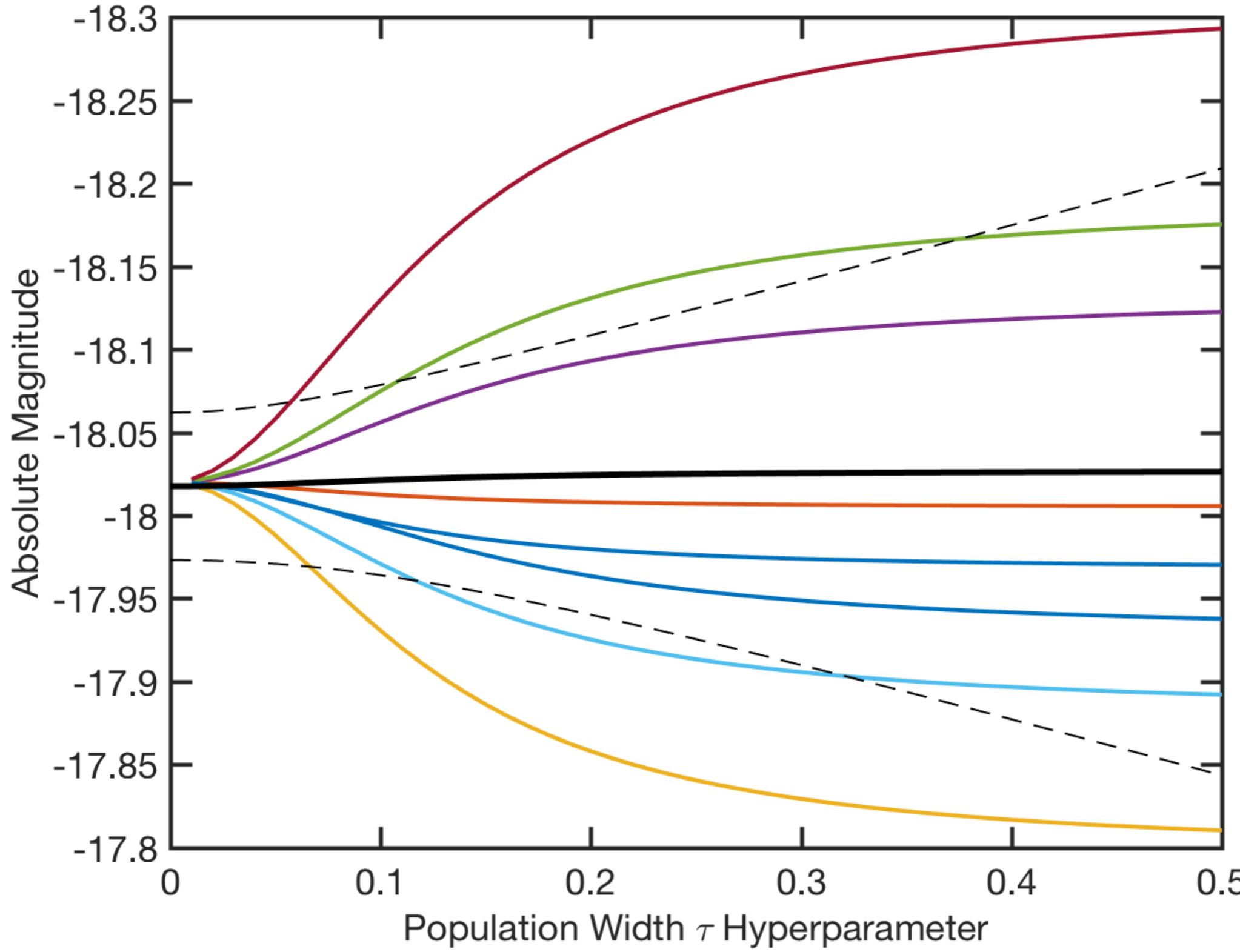


Marginalise out individuals M_s : Posterior of Population hyperparameters:

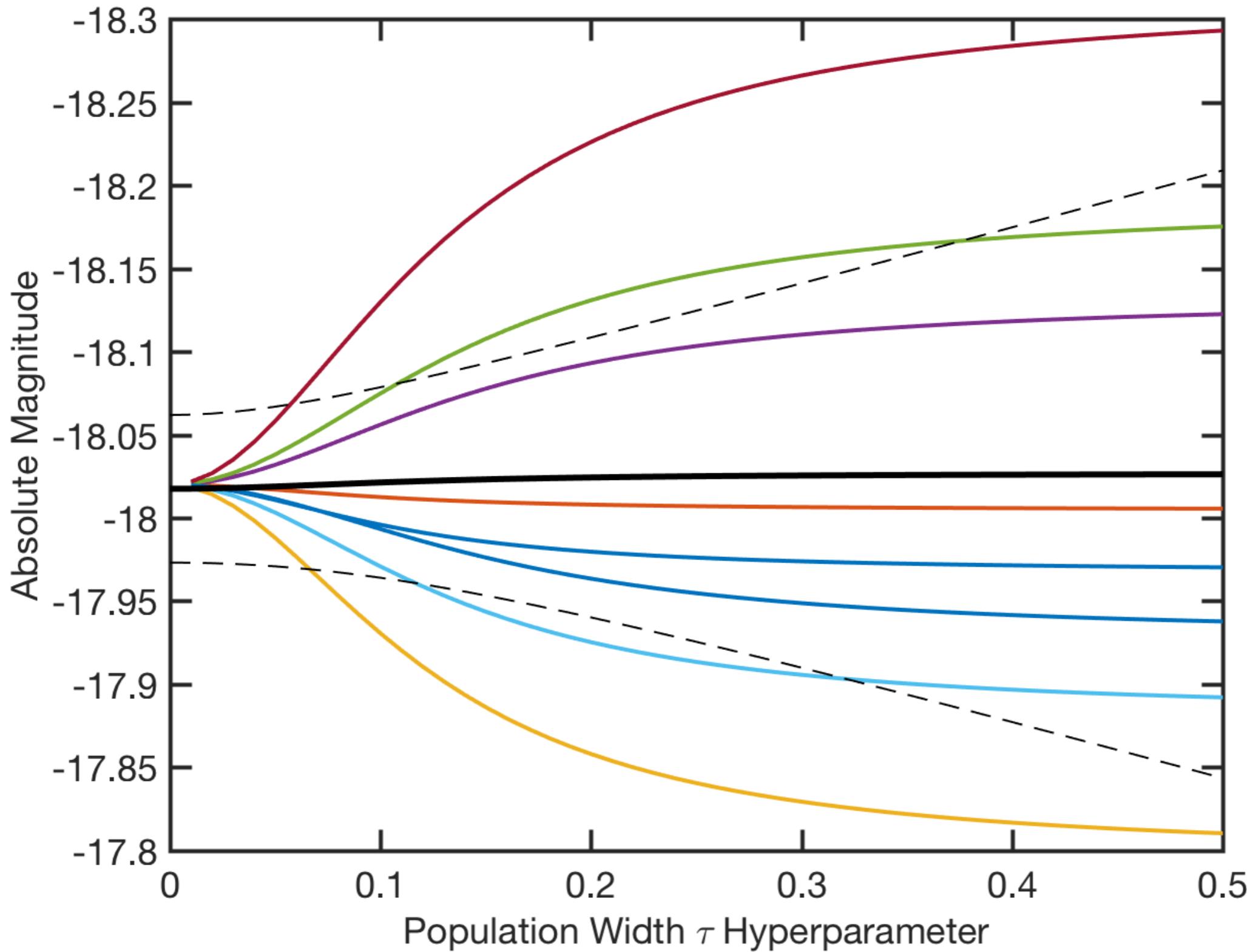
$$P(M_0, \tau | D) = P(M_0 | \tau, D) \times P(\tau | D)$$



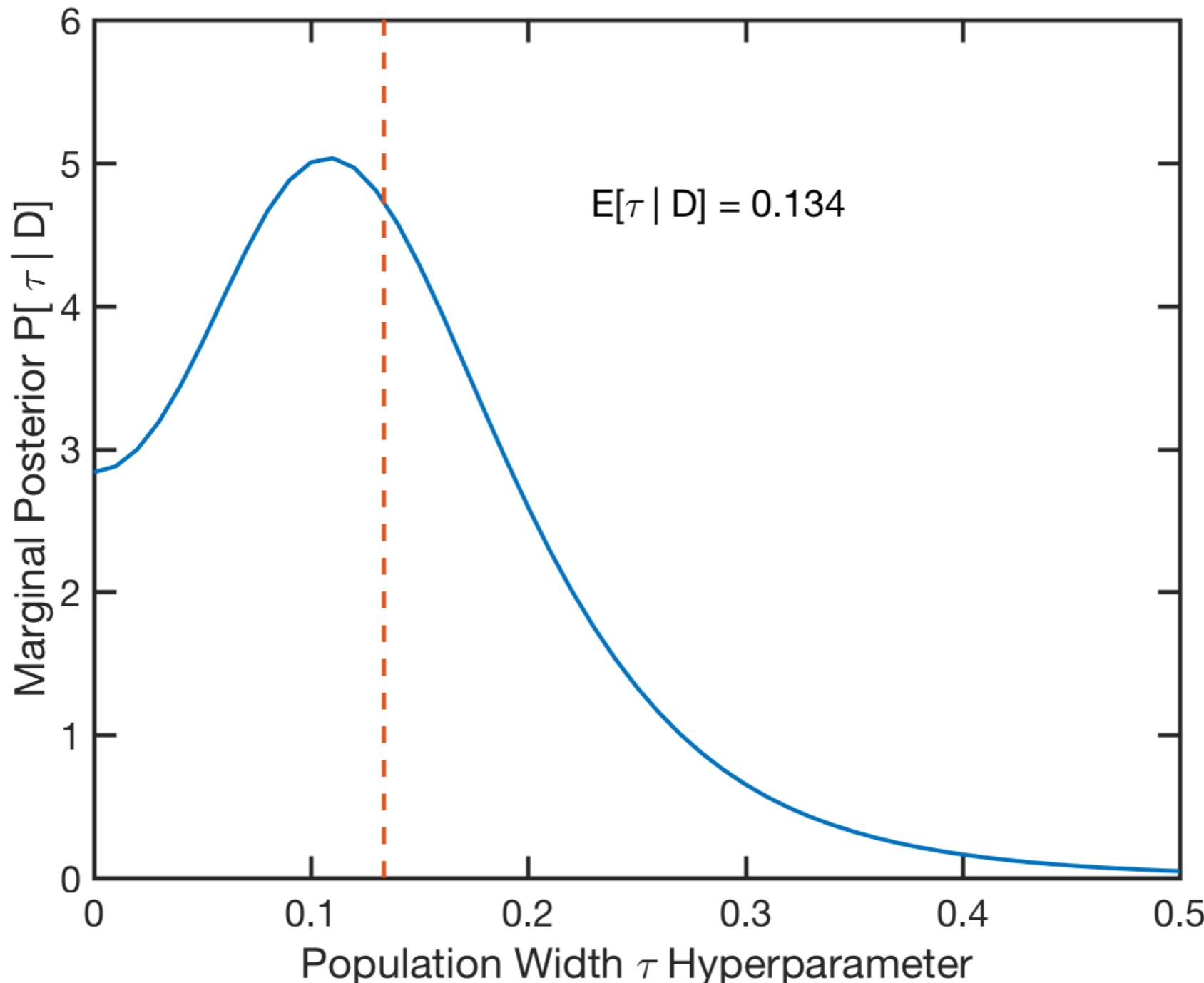
(Conditional) Pop Mean $P(M_0 | \tau, D)$
vs. Individual Means $P(M_s | M_0, \tau, D)$



How much shrinkage to apply?

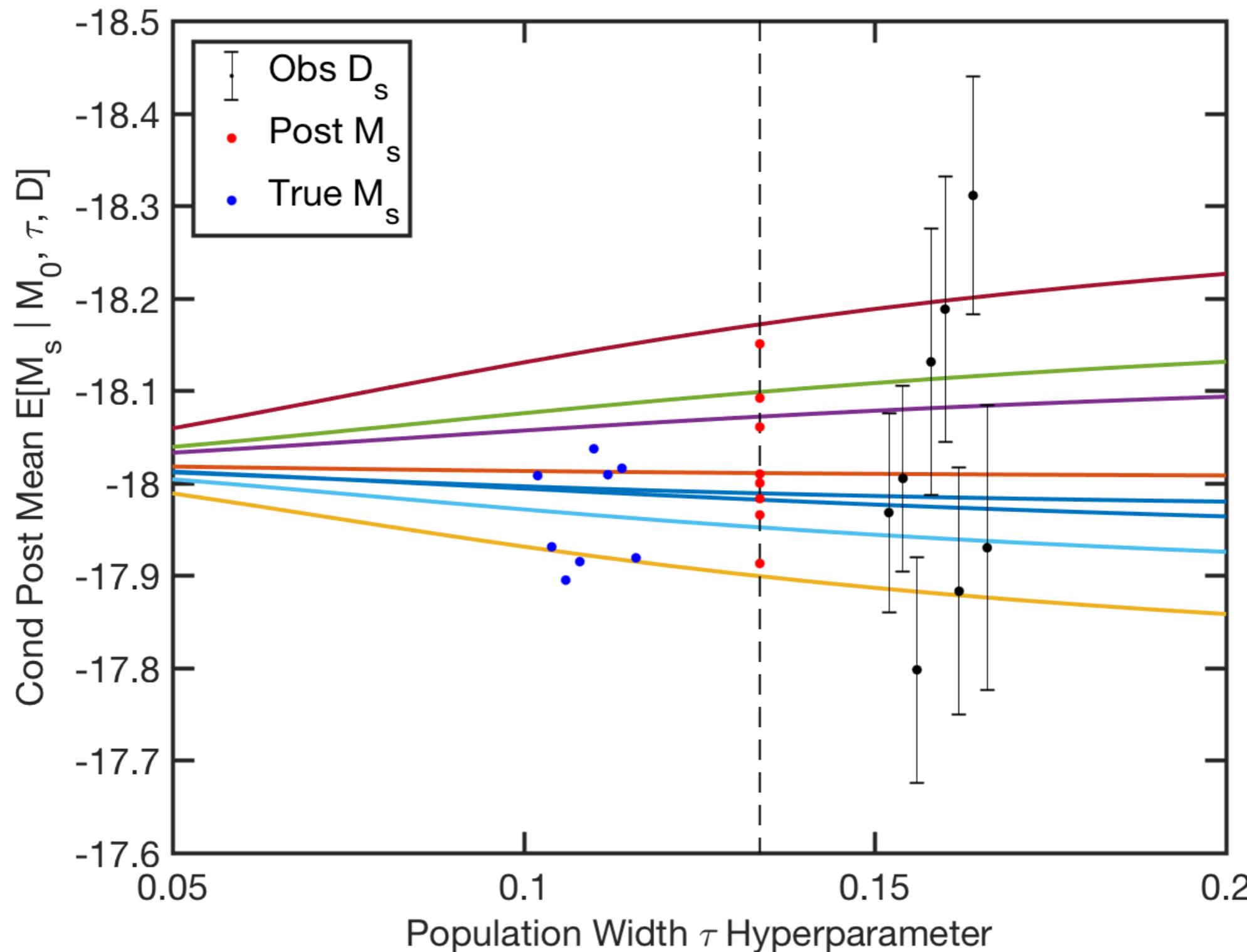


Hierarchical Bayes gives you the posterior density $P(\tau | D)$

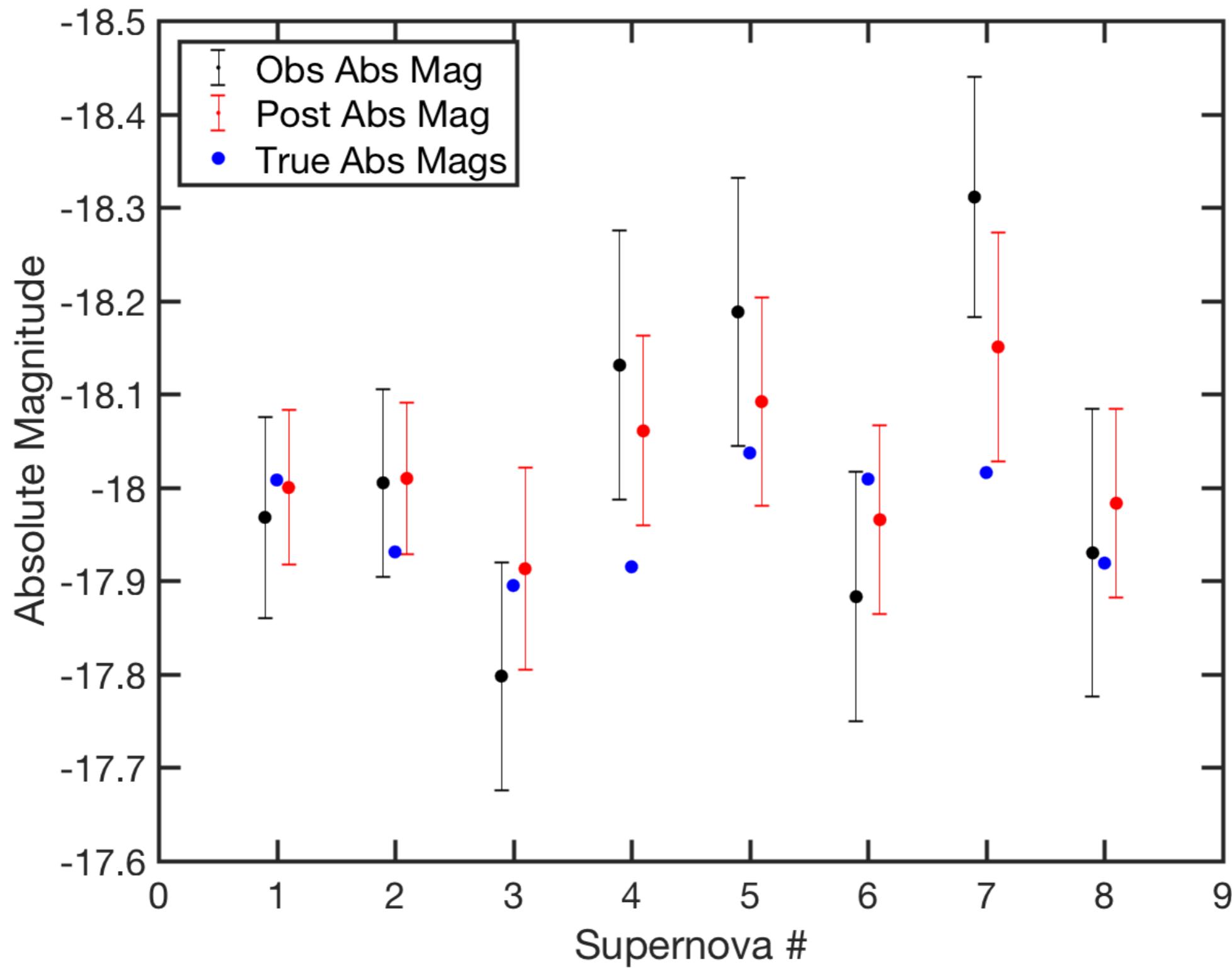


τ controls the degree of shrinkage

How does Hierarchical Bayes implement shrinkage?



Posterior vs. individual MLEs (data) vs. truth



Another Hierarchical Bayesian model: Linear Regression with measurement errors

8, and Kelly et al. 2007, The Astrophysical Journal, 665, 1506). Consider the probabilistic generative model described in class:

$$\xi_i \sim N(\mu, \tau^2) \quad (1)$$

$$\eta_i | \xi_i \sim N(\alpha + \beta \xi_i, \sigma^2) \quad (2)$$

$$x_i | \xi_i \sim N(\xi_i, \sigma_{x,i}^2) \quad (3)$$

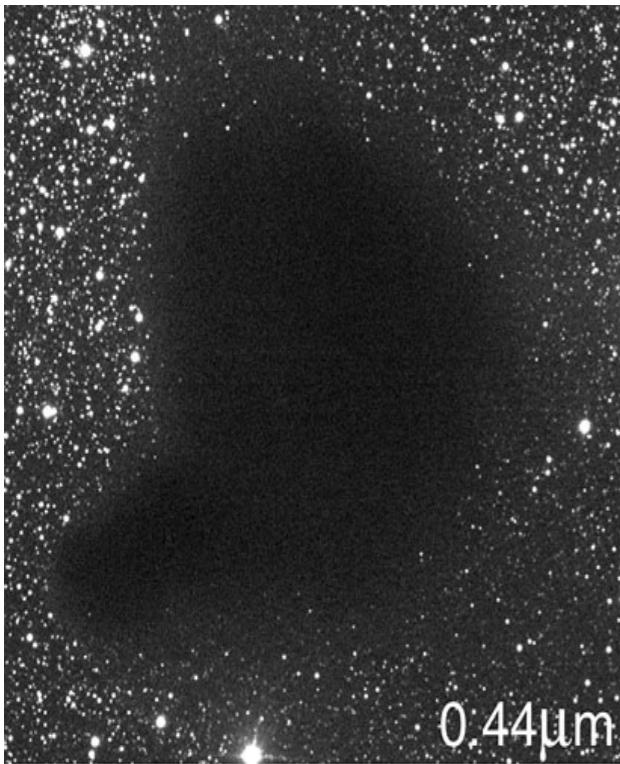
$$y_i | \eta_i \sim N(\eta_i, \sigma_{y,i}^2) \quad (4)$$

The astronomer measures values $\mathcal{D} = \{x_i, y_i\}$ with known measurement error variances $\{\sigma_{x,i}^2, \sigma_{y,i}^2\}$, for $i = 1, \dots, N$ quasars.

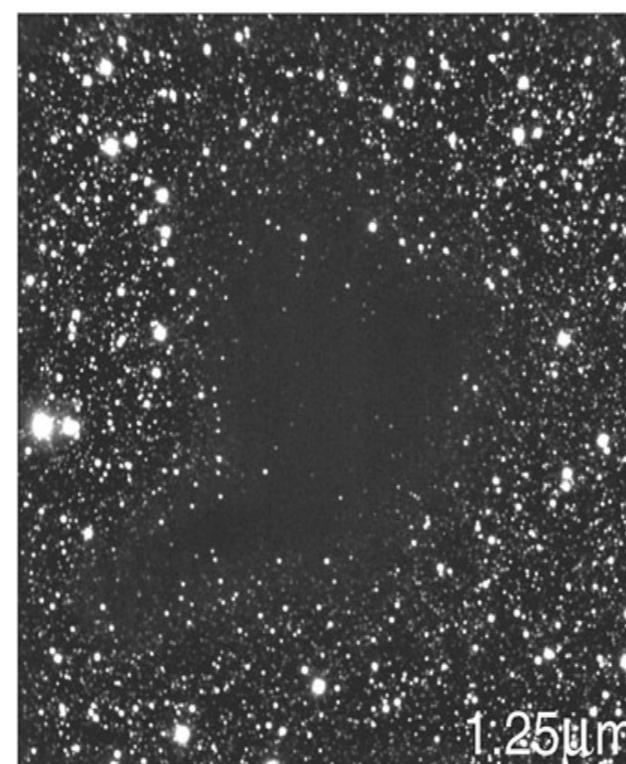
(Draw PGM)

Example: supernova colours
Interstellar Dust is a real physical effect

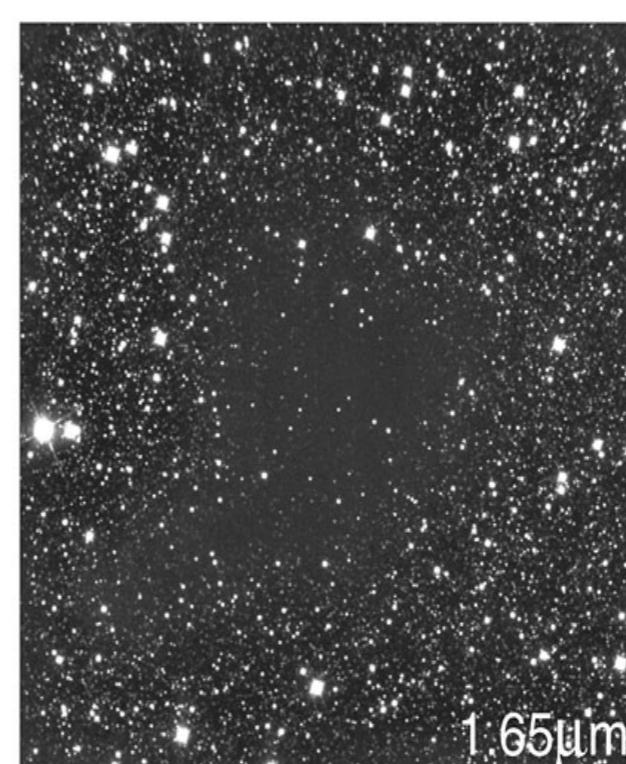
Optical light



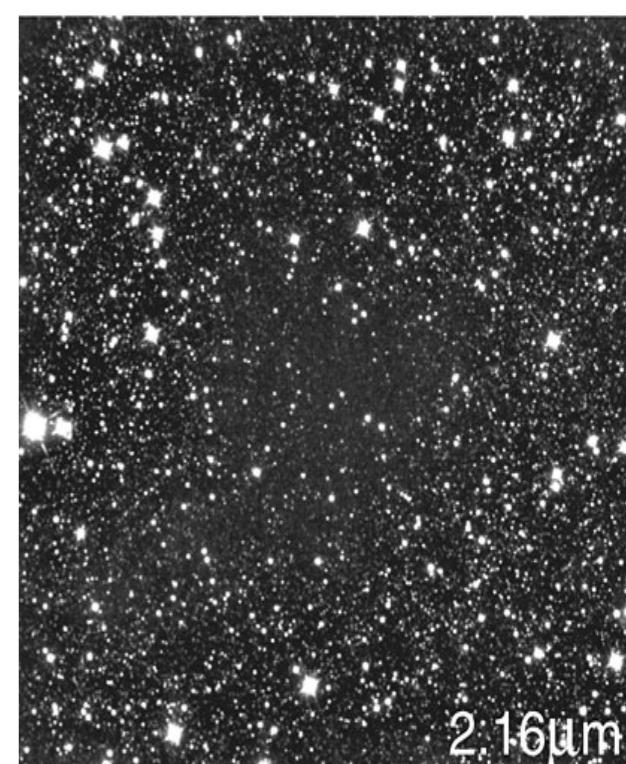
B-band



J-band



H-band

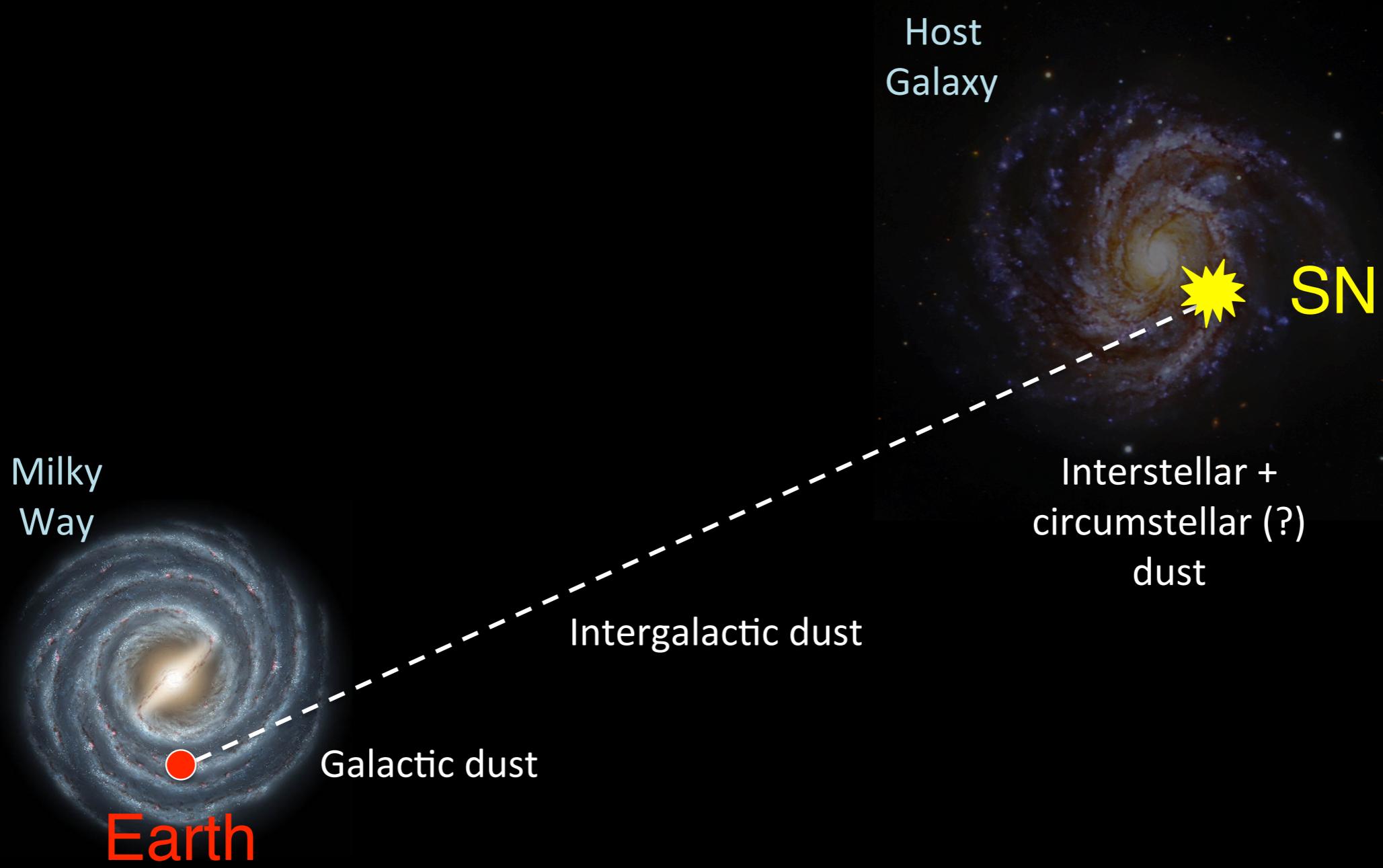


K-band

Near Infrared Wavelengths

Seeing through interstellar dust

From Dust to Dust



What about the host galaxy dust?

Dust Absorption vs. Wavelength of Light

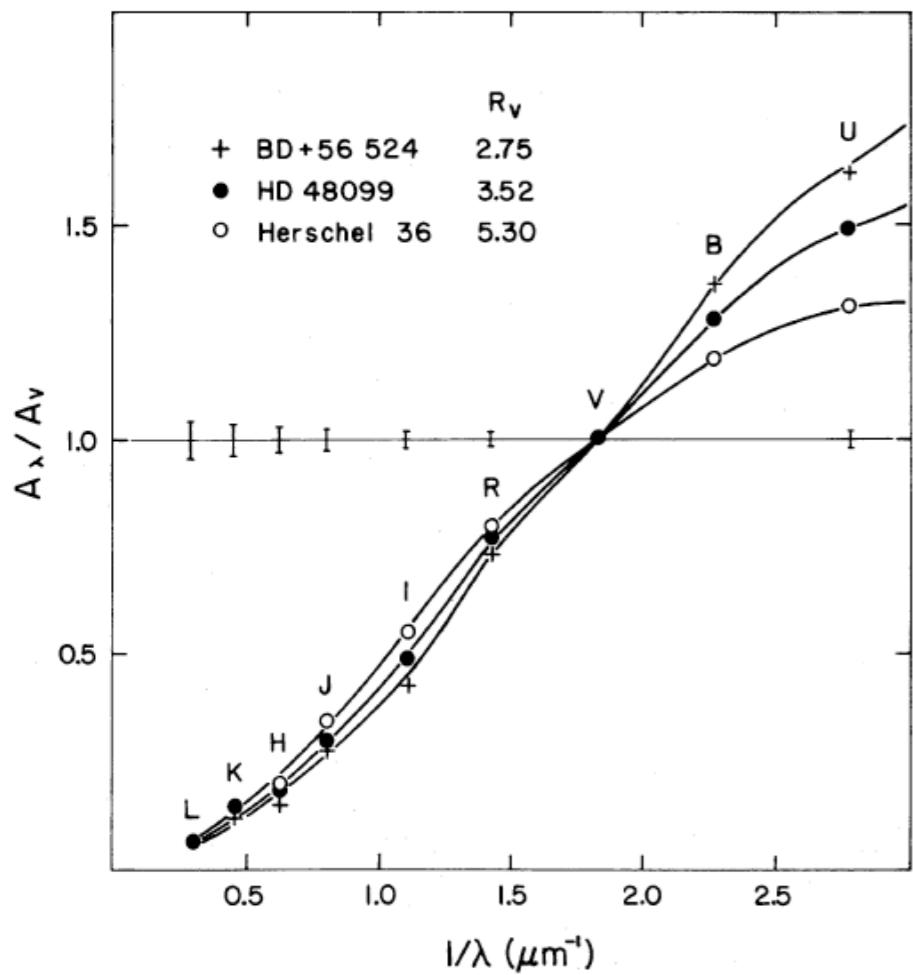


FIG. 3.—Comparison between the mean optical/NIR R_V -dependent extinction law from eqs. (2) and (3) and three lines of sight with largely separated R_V values. The wavelength position of the various broad-band filters from which the data were obtained are labeled (see Table 3). The “error” bars represent the computed standard deviation of the data about the best fit of $A(\lambda)/A(V)$ vs. R_V^{-1} with $a(x) + b(x)/R_V$ where $x \equiv \lambda^{-1}$. The effect of varying R_V on the shape of the extinction curves is quite apparent, particularly at the shorter wavelengths.

- Absorption of light (dimming) depends on λ , causing reddening
- Interstellar lines of sight to SN in different galaxies can pass through different random amounts of dust
- Key Parameters of Interstellar Dust (different for each SN)
 - $A_B \sim$ Amount of Dust Absorption (dimming) in B-band (450nm)
 - $R_V = A_V/E(B-V) \sim$ Wavelength Dependence of Dust Absorption

Another example: Supernova colours

$$m_s = M_{\text{int},s} + \mu_s + A_s \quad (\text{A} = \text{Extinction} = \text{Dimming})$$

Intrinsic Colour: (Reddening)

$$C_s = M_{\text{int},s}^B - M_{\text{int},s}^V \quad E_s \equiv A_s^B - A_s^V$$

Apparent Colour:

$$O_s = m_s^B - m_s^V = C_s + A_s^B - A_s^V = C_s + E_s$$

$$C_s \sim N(\mu_C, \sigma_C^2) \quad E_s \sim \text{Exponen}(\tau)$$

$$\hat{O}_s = N(C_s + E_s, \sigma_s^2)$$

Supernova colours

Intrinsic Colour Population

$$C_s \sim N(\mu_C, \sigma_C^2)$$

Dust Reddening Population

$$E_s \sim \text{Exponen}(\tau)$$

Individual Measurement Error

$$\hat{O}_s = N(C_s + E_s, \sigma_s^2)$$

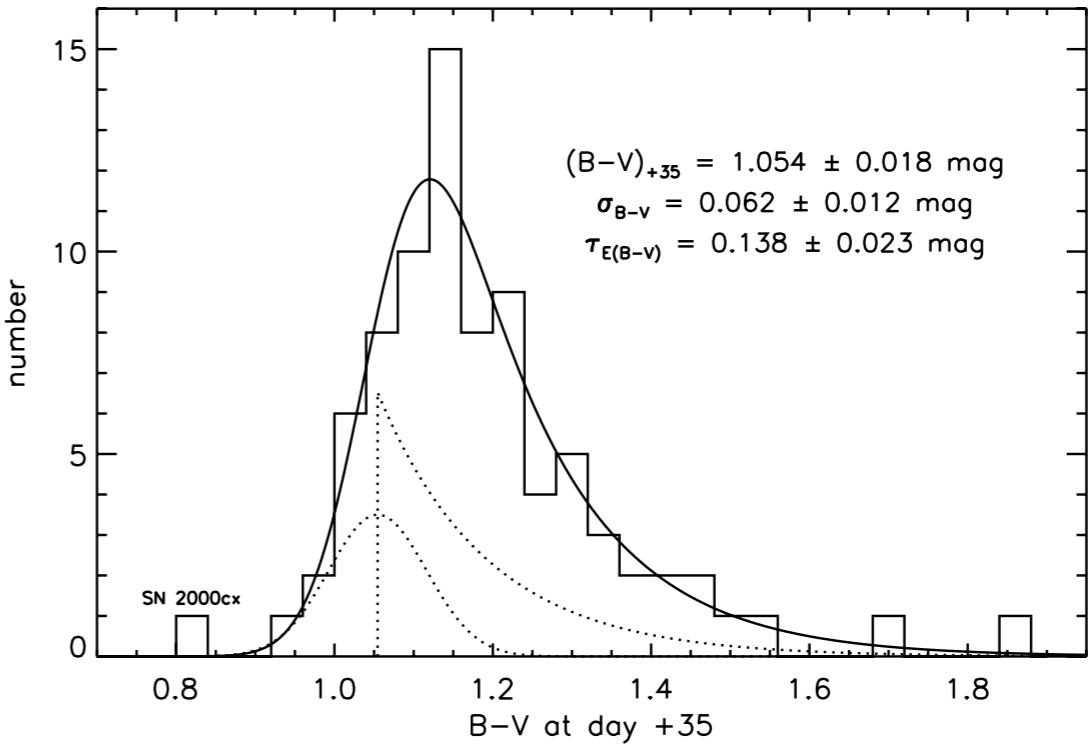
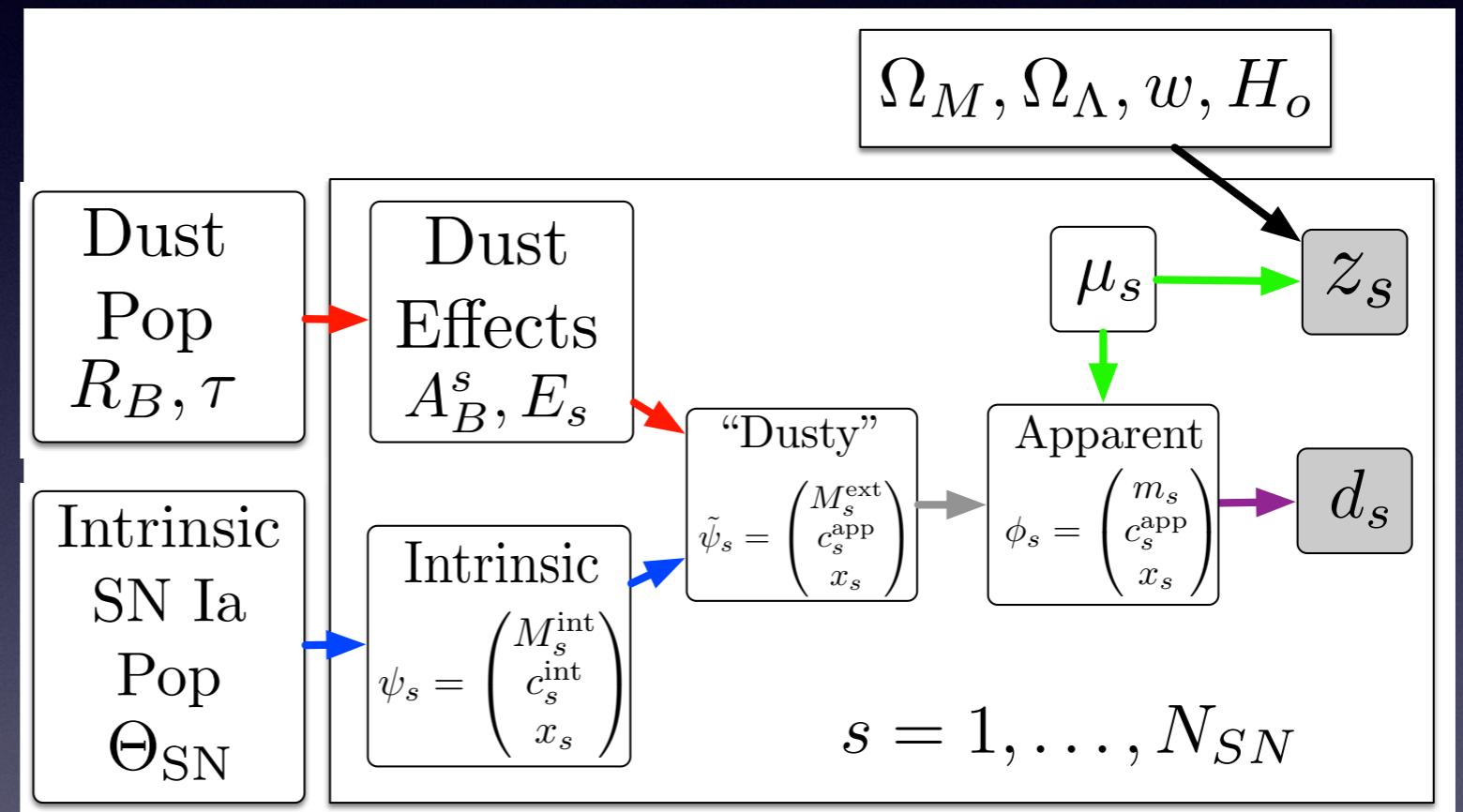


FIG. 6.—Histogram of 82 SNe Ia with well-measured late-time $B - V$ color evolution. The data were corrected for Galactic extinction and the K -correction and referenced to +35 days after B maximum, adopting a late-time color evolution slope of $-0.0118 \text{ mag day}^{-1}$. The maximum likelihood fit model is shown as the solid line; it is the convolution of the dotted lines shown (at an arbitrary scale for clarity). SN 2000cx, a clear outlier, was not included in the fit.

Example Application: The Type Ia Supernova Color-Magnitude Relation and Host Galaxy Dust: A Simple Hierarchical Bayesian Model

SN 2014J KAIT/LOSS color image



arXiv:1609.04470
(Mandel et al. 2017)

Model Comparison & Selection

- Number of Spectral Lines in a (noisy) spectrum?
- Clustering/mixture models - how many clumps?
- Time Series - Curve fitting
 - Is there a trend?
 - Complexity/order/degree of best model
 - which GP kernel best explains the data?
- Cosmology - standard (8-parameter) cosmological model vs more exotic (more parameters) models?

Cosmological Model Selection

Table 4.1. *Parameter constraints of the Standard Cosmological Model, reproduced from Spergel et al. (2007, WMAP collaboration) with some additional rounding. The values quoted are mean values and 68 per cent confidence intervals. All columns assume the Λ CDM cosmology with a power-law initial spectrum, no tensors, spatial flatness, and a cosmological constant as dark energy. Three different data combinations are shown to highlight the extent to which this choice matters. The parameters are $\Omega_m h^2$ (physical matter density), $\Omega_b h^2$ (physical baryon density), h (Hubble parameter), n (density perturbation spectral index), τ (optical depth to last-scattering surface), and σ_8 (density perturbation amplitude).*

	WMAP alone	WMAP + 2dF	WMAP + all
$\Omega_m h^2$	0.128 ± 0.008	0.126 ± 0.005	0.132 ± 0.004
$\Omega_b h^2$	0.0223 ± 0.0007	0.0222 ± 0.0007	0.0219 ± 0.0007
h	0.73 ± 0.03	0.73 ± 0.02	$0.704^{+0.015}_{-0.016}$
n	0.958 ± 0.016	0.948 ± 0.015	0.947 ± 0.015
τ	0.089 ± 0.030	0.083 ± 0.028	$0.073^{+0.027}_{-0.028}$
σ_8	0.76 ± 0.05	0.74 ± 0.04	0.78 ± 0.03

Standard LCDM Cosmological Model

Cosmological Model Selection: which extra parameters are warranted by data?

Table 4.2. *Candidate parameters: those that might be relevant for cosmological observations, but for which there is presently no convincing evidence requiring them. They are listed so as to take the value zero in the base cosmological model. Those above the line are parameters of the background homogeneous cosmology, and those below describe the perturbations. Of the latter set, the first five refer to adiabatic perturbations, the next three to tensor perturbations, and the remainder to isocurvature perturbations. This table is taken from Liddle (2004).*

Ω_k	spatial curvature
$N_\nu - 3.04$	effective number of neutrino species (CMBFAST definition)
m_{ν_i}	neutrino mass for species ‘ i ’ [or more complex neutrino properties]
m_{dm}	(warm) dark matter mass
$w + 1$	dark energy equation of state
dw/dz	redshift dependence of w [or more complex parameterization of dark energy evolution]
$c_S^2 - 1$	effects of dark energy sound speed
$1/r_{\text{top}}$	topological identification scale [or more complex parameterization of non-trivial topology]
$d\alpha/dz$	redshift dependence of the fine structure constant
dG/dz	redshift dependence of the gravitational constant
$dn/d \ln k$	running of the scalar spectral index
k_{cut}	large-scale cut-off in the spectrum
A_{feature}	amplitude of spectral feature (peak, dip or step) ...
k_{feature}	... and its scale
f_{NL}	[or adiabatic power spectrum amplitude parameterized in N bins] quadratic contribution to primordial non-Gaussianity [or more complex parameterization of non-Gaussianity]
r	tensor-to-scalar ratio
$r + 8n_T$	violation of the inflationary consistency equation
$dn_T/d \ln k$	running of the tensor spectral index
\mathcal{P}_S	CDM isocurvature perturbation S ...
n_S	... and its spectral index ...
\mathcal{P}_{SR}	... and its correlation with adiabatic perturbations ...
$n_{SR} - n_S$... and the spectral index of that correlation
$G\mu$	[or more complicated multi-component isocurvature perturbation] cosmic string component of perturbations

Bayesian Model Comparison

Parameter Estimation: Posterior on parameters

$$P(\boldsymbol{\theta}_1 | \mathbf{D}, M_1) = \frac{P(\mathbf{D} | \boldsymbol{\theta}_1, M_1) P(\boldsymbol{\theta}_1 | M_1)}{P(\mathbf{D} | M_1)}$$

Model Selection: Posterior on Models:

$$P(\mathbf{D} | M_1) = \int P(\mathbf{D} | \boldsymbol{\theta}_1, M_1) P(\boldsymbol{\theta}_1 | M_1) d\boldsymbol{\theta}_1$$

(Needs proper prior!)

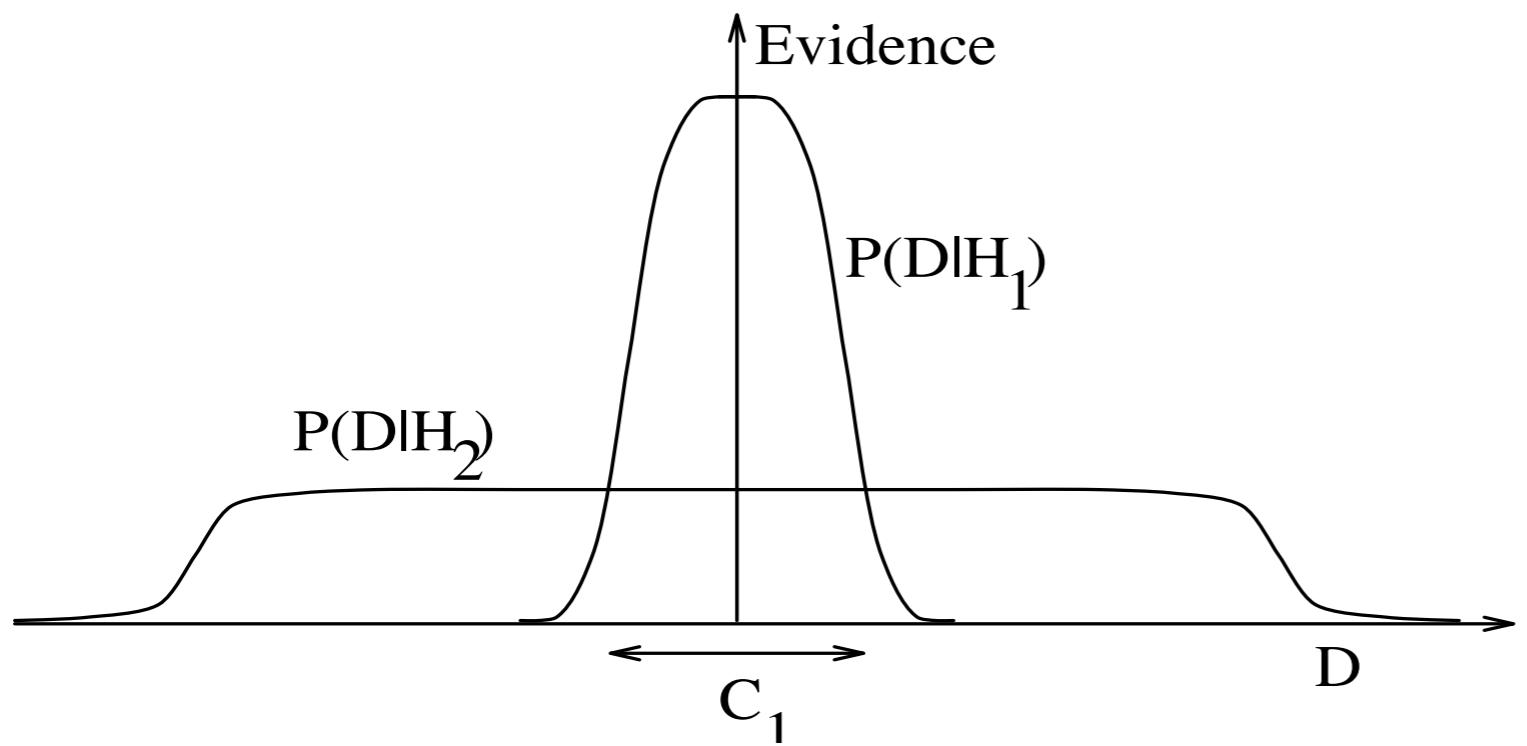
Posterior Odds Ratio

$$\frac{P(M_1 | \mathbf{D})}{P(M_2 | \mathbf{D})} = \frac{P(\mathbf{D} | M_1)}{P(\mathbf{D} | M_2)} \times \frac{P(M_1)}{P(M_2)}$$

$$\text{BayesFactor}_{12} = \frac{P(\mathbf{D} | M_1)}{P(\mathbf{D} | M_2)}$$

Bayesian Model Comparison (switch to Bishop's slides)

Occam's Razor



(MacKay, Ch 28)

Figure 28.3. Why Bayesian inference embodies Occam's razor. This figure gives the basic intuition for why complex models can turn out to be less probable. The horizontal axis represents the space of possible data sets D . Bayes' theorem rewards models in proportion to how much they *predicted* the data that occurred. These predictions are quantified by a normalized probability distribution on D . This probability of the data given model \mathcal{H}_i , $P(D | \mathcal{H}_i)$, is called the evidence for \mathcal{H}_i . A simple model \mathcal{H}_1 makes only a limited range of predictions, shown by $P(D | \mathcal{H}_1)$; a more powerful model \mathcal{H}_2 , that has, for example, more free parameters than \mathcal{H}_1 , is able to predict a greater variety of data sets. This means, however, that \mathcal{H}_2 does not predict the data sets in region \mathcal{C}_1 as strongly as \mathcal{H}_1 . Suppose that equal prior probabilities have been assigned to the two models. Then, if the data set falls in region \mathcal{C}_1 , the *less powerful* model \mathcal{H}_1 will be the *more probable* model.

Occam's Razor

350

28 — Model Comparison and Occam's Razor

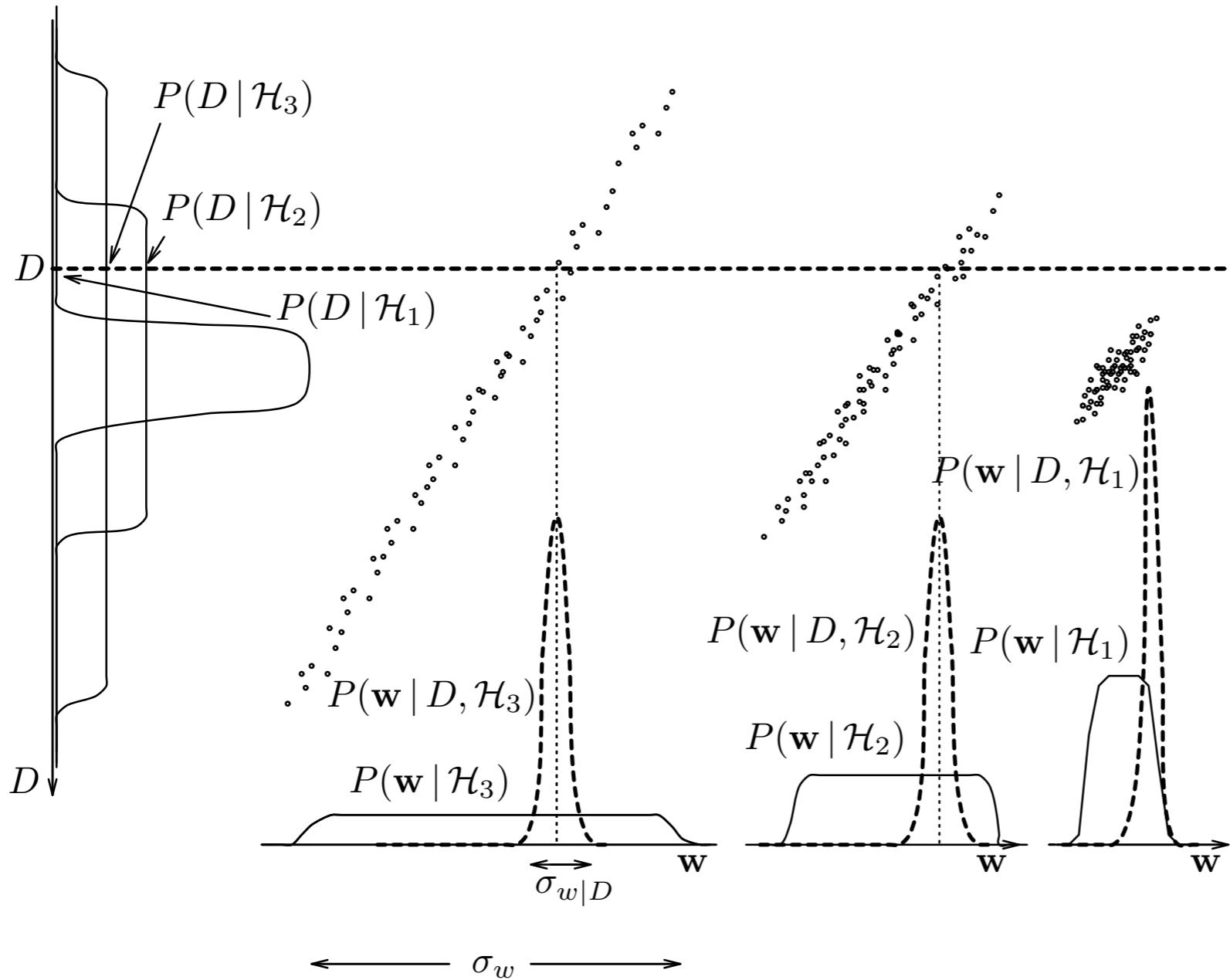


Figure 28.6. A hypothesis space consisting of three exclusive models, each having one parameter \mathbf{w} , and a one-dimensional data set D . The ‘data set’ is a single measured value which differs from the parameter \mathbf{w} by a small amount of additive noise. Typical samples from the joint distribution $P(D, w, \mathcal{H})$ are shown by dots. (N.B., these are not data points.) The observed ‘data set’ is a single particular value for D shown by the dashed horizontal line. The dashed curves below show the posterior probability of w for each model given this data set (cf. figure 28.3). The evidence for the different models is obtained by marginalizing onto the D axis at the left-hand side (cf. figure 28.5).

(MacKay, Ch 28)

Interpreting the Bayes Factor / Evidence Ratio: The Jeffreys Scale

$$\Delta \ln E = \ln \text{BF}$$

Model selection and multi-model inference

	Jeffreys	This volume
$\Delta \ln E < 1$	Not worth more than a bare mention.	
$1 < \Delta \ln E < 2.5$	Significant.	Weak.
$2.5 < \Delta \ln E < 5$	Strong to very strong.	Significant.
$5 < \Delta \ln E$	Decisive.	Strong.

The Laplace Approximation

Evidence: $P(D) = \int P^*(\theta|D) d\theta$

Unnormalised Posterior: $P^*(\theta|D) = P(D|\theta)P(\theta)$

Find MAP estimate: $\theta_0 = \operatorname{argmax}_{\theta} \ln P^*(\theta|D)$

Taylor Expansion:

$$\ln P^*(\theta|D) \approx \ln P^*(\theta_0|D) - \frac{1}{2}(\theta - \theta_0)^T A(\theta - \theta_0) + \dots$$

Hessian at mode: $A_{ij} = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln P^*(\theta|D) \Big|_{\theta=\theta_0}$

$$P^*(\theta|D) \approx P^*(\theta_0|D) \times \exp \left(-\frac{1}{2}(\theta - \theta_0)^T A(\theta - \theta_0) \right)$$

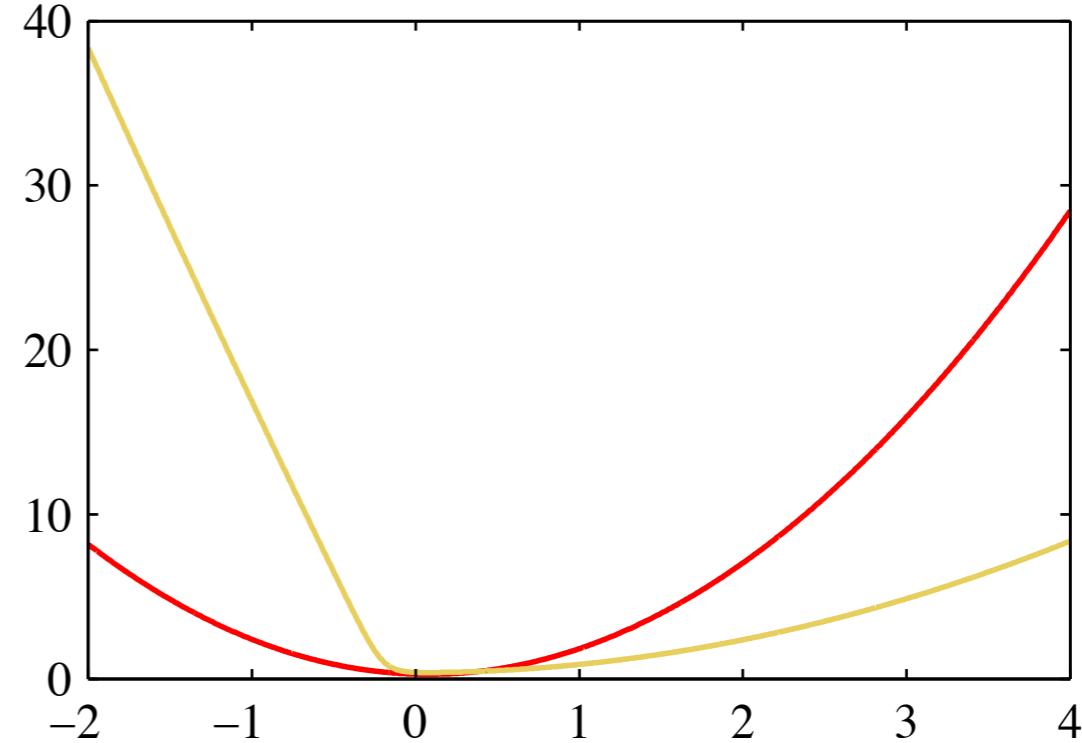
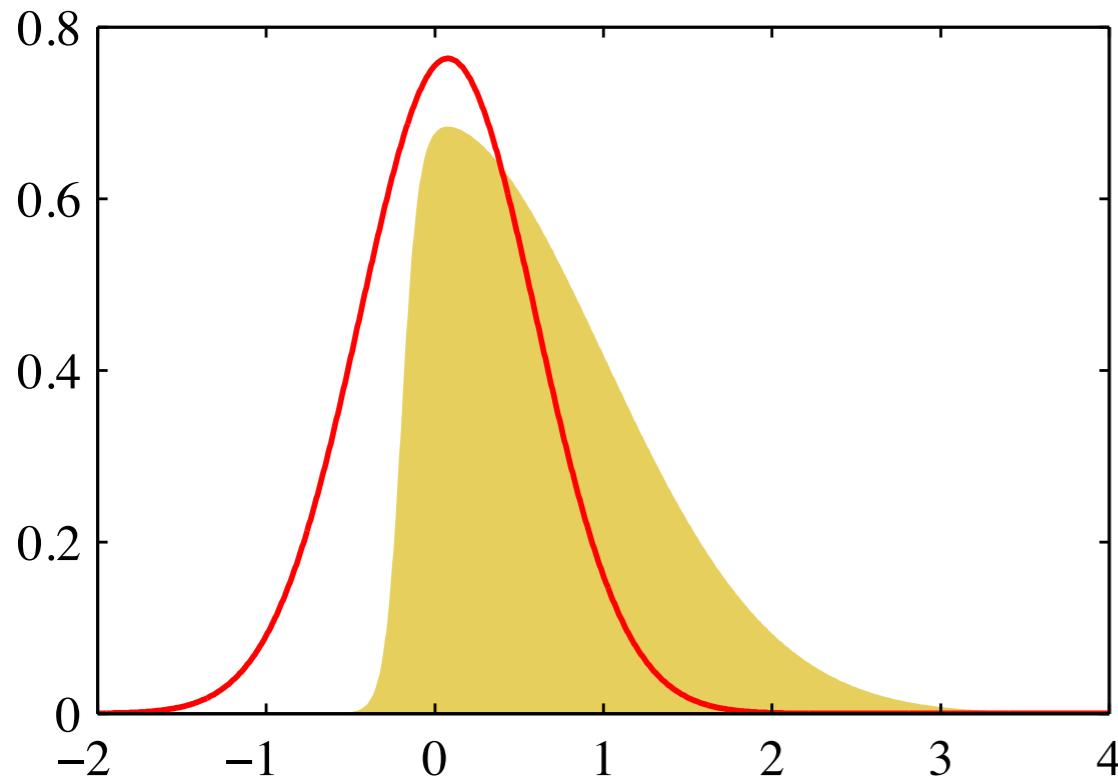
The Laplace Approximation

Taylor Expansion

$$\ln P^*(\boldsymbol{\theta}|\mathcal{D}) \approx \ln P^*(\boldsymbol{\theta}_0|\mathcal{D}) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \dots$$

$$A_{ij} = -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \ln P^*(\boldsymbol{\theta}|\mathcal{D}) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$$

$$P^*(\boldsymbol{\theta}|\mathcal{D}) \approx P^*(\boldsymbol{\theta}_0|\mathcal{D}) \times \exp \left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \right)$$



The Laplace Approximation

$$P(\mathbf{D}) = \int P^*(\boldsymbol{\theta}|\mathbf{D}) d\boldsymbol{\theta} \quad P^*(\boldsymbol{\theta}|\mathbf{D}) = P(\mathbf{D}|\boldsymbol{\theta})P(\boldsymbol{\theta})$$

(Suppose prior is wide enough to well-contain the peak)

$$P^*(\boldsymbol{\theta}|\mathbf{D}) \approx P^*(\boldsymbol{\theta}_0|\mathbf{D}) \times \exp\left(-\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \mathbf{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right)$$

$$P(\mathbf{D}) \approx P^*(\boldsymbol{\theta}_0|\mathbf{D}) \times |2\pi \mathbf{A}^{-1}|^{1/2} \int N(\boldsymbol{\theta}|\boldsymbol{\theta}_0, \mathbf{A}^{-1}) d\boldsymbol{\theta}$$

$$P(\mathbf{D}) \approx P^*(\boldsymbol{\theta}_0|\mathbf{D}) \times \det(\mathbf{A}/2\pi)^{-1/2}$$

$$P(\mathbf{D}) \approx P(\mathbf{D}|\boldsymbol{\theta}_0) \times P(\boldsymbol{\theta}_0) \times \det(\mathbf{A}/2\pi)^{-1/2}$$

$$\mathbf{A} = -\nabla\nabla \ln P(\boldsymbol{\theta}|\mathbf{D})$$

The Laplace Approximation

$$P(D) \approx P(D|\theta_0) \times P(\theta_0) \times \det(A/2\pi)^{-1/2}$$

Evidence	Best-Fit Likelihood	Occam Factor (Ratio of Posterior to Prior Width)
----------	---------------------	---

$$\Sigma_{\text{post}} = A^{-1}$$

Suppose:

$$P(\theta) = N(\theta | \theta_{\text{prior}}, \Sigma_{\text{prior}})$$

$$\text{Occam Factor} \propto |2\pi\Sigma_{\text{prior}}|^{-1/2} \times |2\pi\Sigma_{\text{post}}|^{1/2}$$

$$\propto \frac{|\Sigma_{\text{post}}|^{1/2}}{|\Sigma_{\text{prior}}|^{1/2}}$$

Example: Spectral Line profile

Line Model:

Gaussian: $H_1 : f(\lambda) = \alpha_1 + \alpha_2 \exp\left(-\frac{(\lambda - \alpha_4)^2}{2\alpha_3^2}\right)$

Lorentzian: $H_2 : f(\lambda) = \beta_1 + \beta_2 \left(1 + \frac{(\lambda - \beta_4)^2}{\beta_3^2}\right)^{-1}$

Measurement Likelihood for data: $\{y_i, \lambda_i\}$

$$P(D|\theta) = \prod_{i=1}^N N(y_i | f(\lambda_i; \theta), \sigma_i^2)$$

Example: using Laplace Approximation

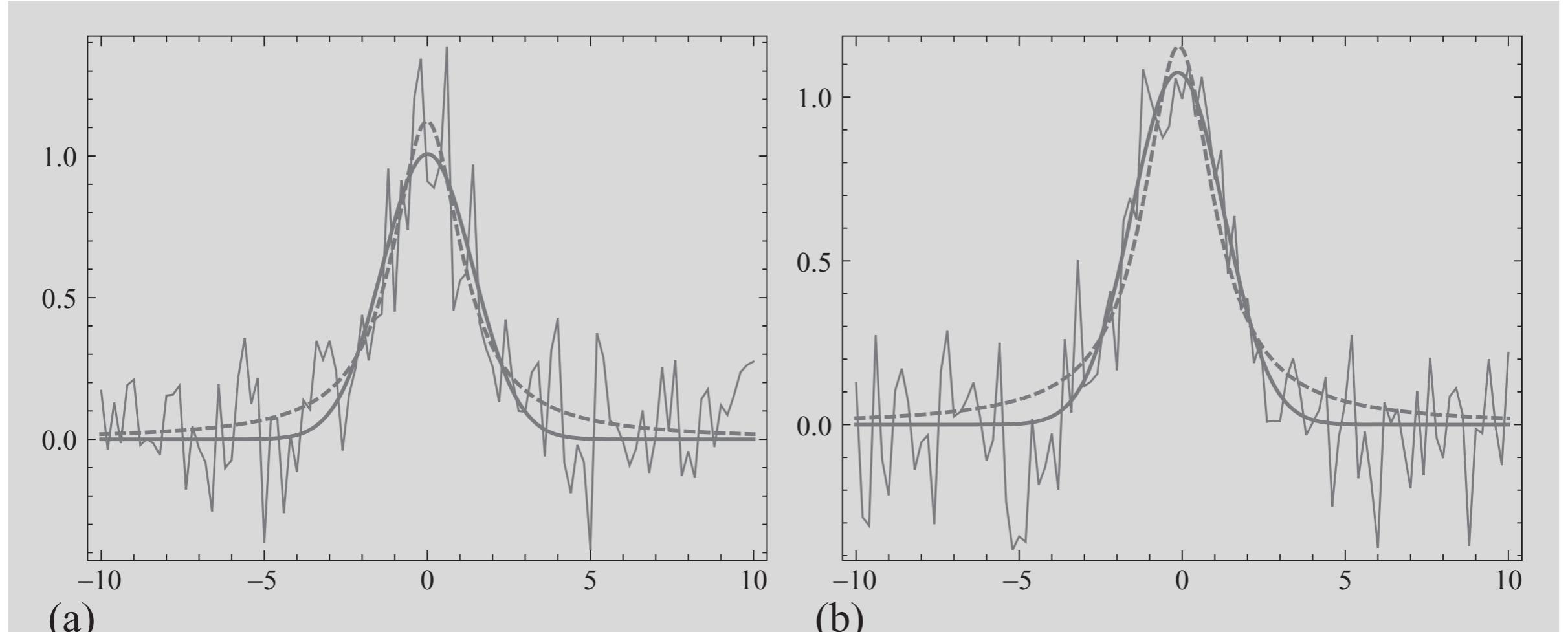


Figure 7.1 Spectral-line simulations. These are at a signal-to-noise ratio, in the peak of the line, of 5. (a) The Lorentzian fit is favoured at odds of 30:1, whereas (b) the Gaussian is favoured at odds of 100:1.

Ways to calculate Evidence

- Analytic (really simple problems)
- Monte Carlo: probably slow if likelihood is peaked
$$\theta_i \sim P(\theta_i)$$
$$P(D) \approx \sum_{i=1}^M P(D|\theta_i)$$
- Harmonic Mean Estimator (bad!)
- Laplace Approximation
- Thermodynamic Integration
- Nested Sampling

Example: Nested Sampling

- <http://chi-feng.github.io/mcmc-demo/app.html?algorithm=RadFriends-NS&target=banana>