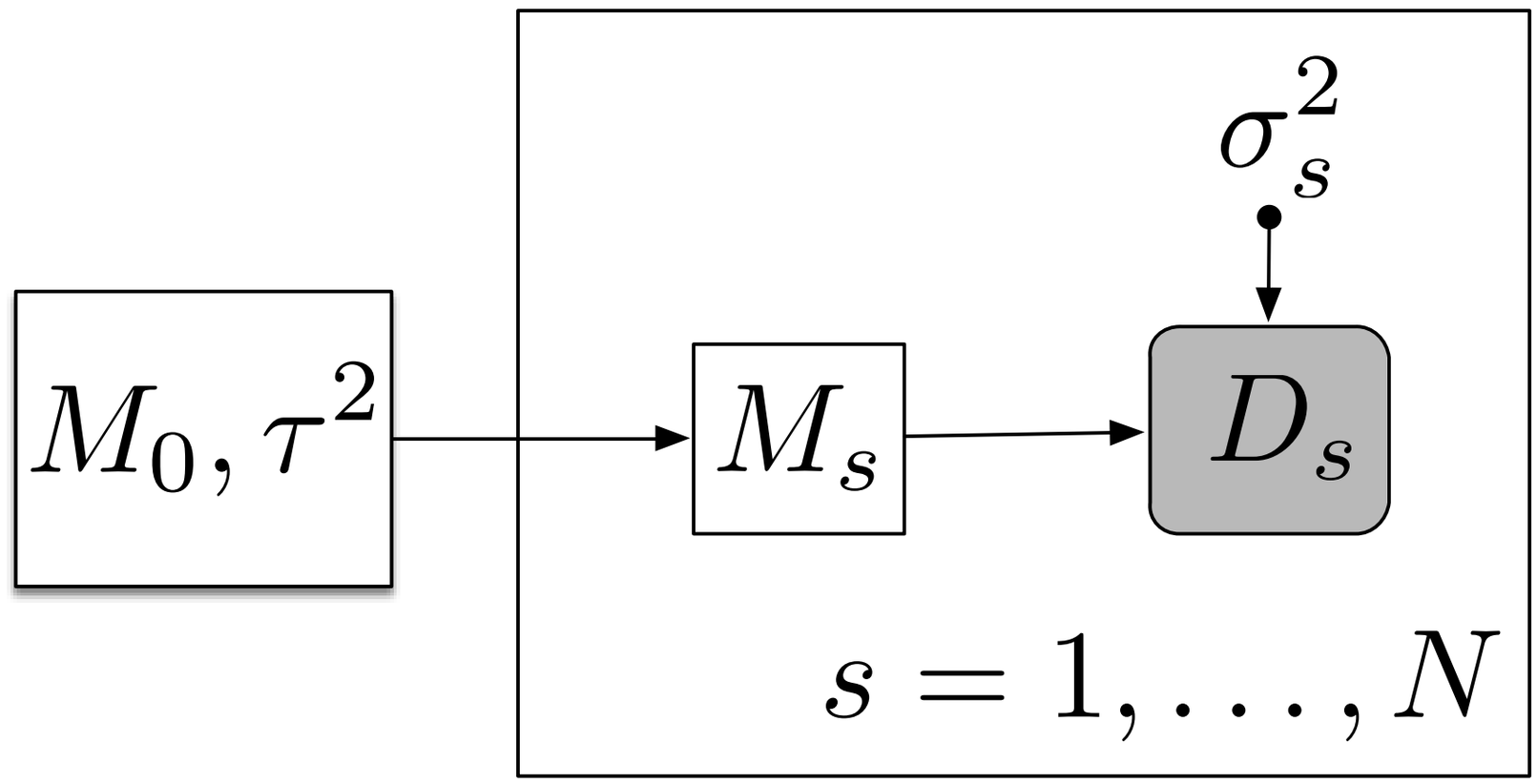
Astrostatistics: 14 Mar 2019

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

- Example Class, Fri 15 Mar 1pm: MR 12
- Today:
- Finish Hierarchical Bayes & Shrinkage Estimators
- Bayesian Model Comparison

Hierarchical Bayesian posterior

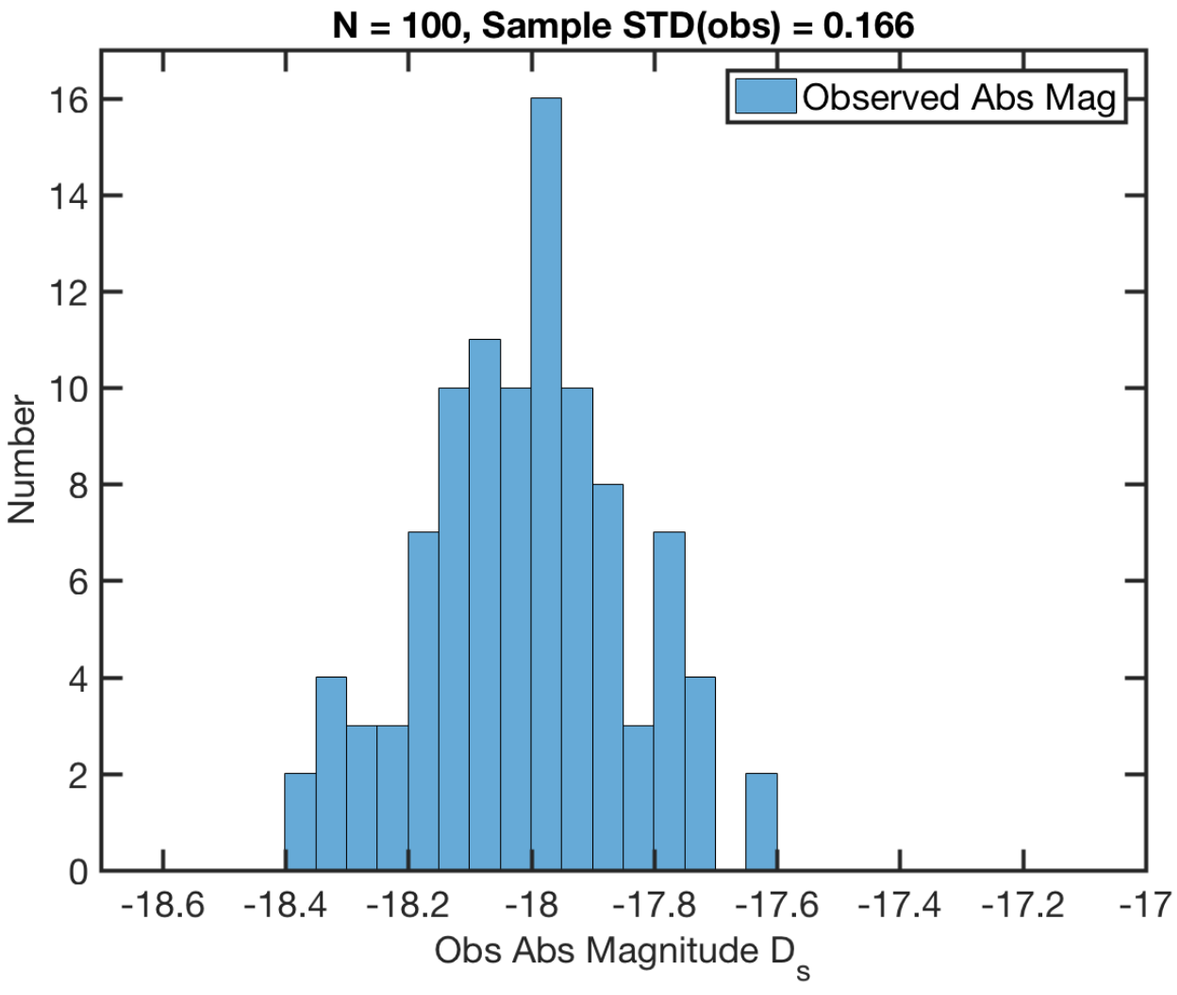
$$P(\{M_s\}, \boldsymbol{H}|\{D_s\}) \propto \Big[\prod_{s=1}^{N} P(D_s|M_s)P(M_s|M_0, \tau^2)\Big] \times P(\boldsymbol{H})$$

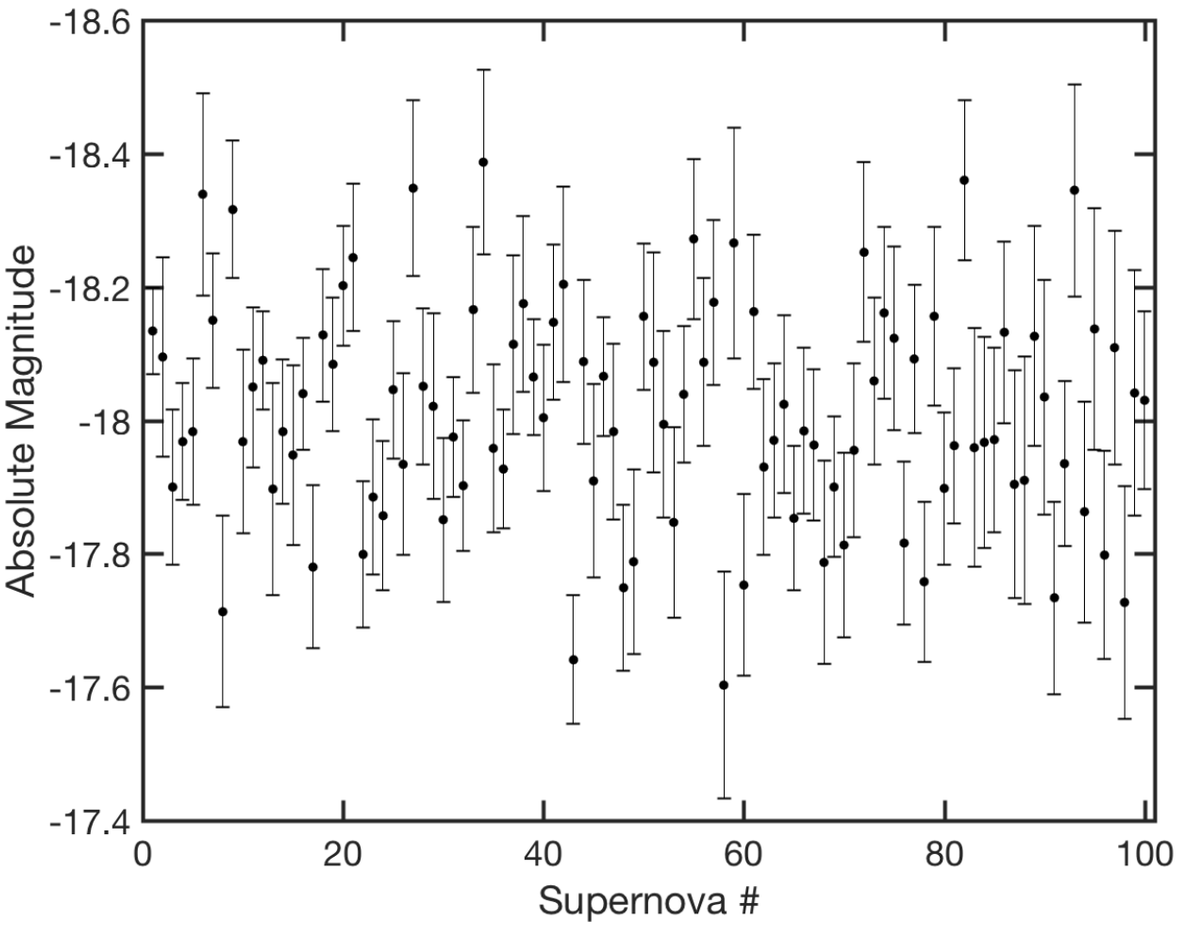


Utilise the Conditional Independence structure of PGM
to derive conditional posterior densities

Code/Data Demo

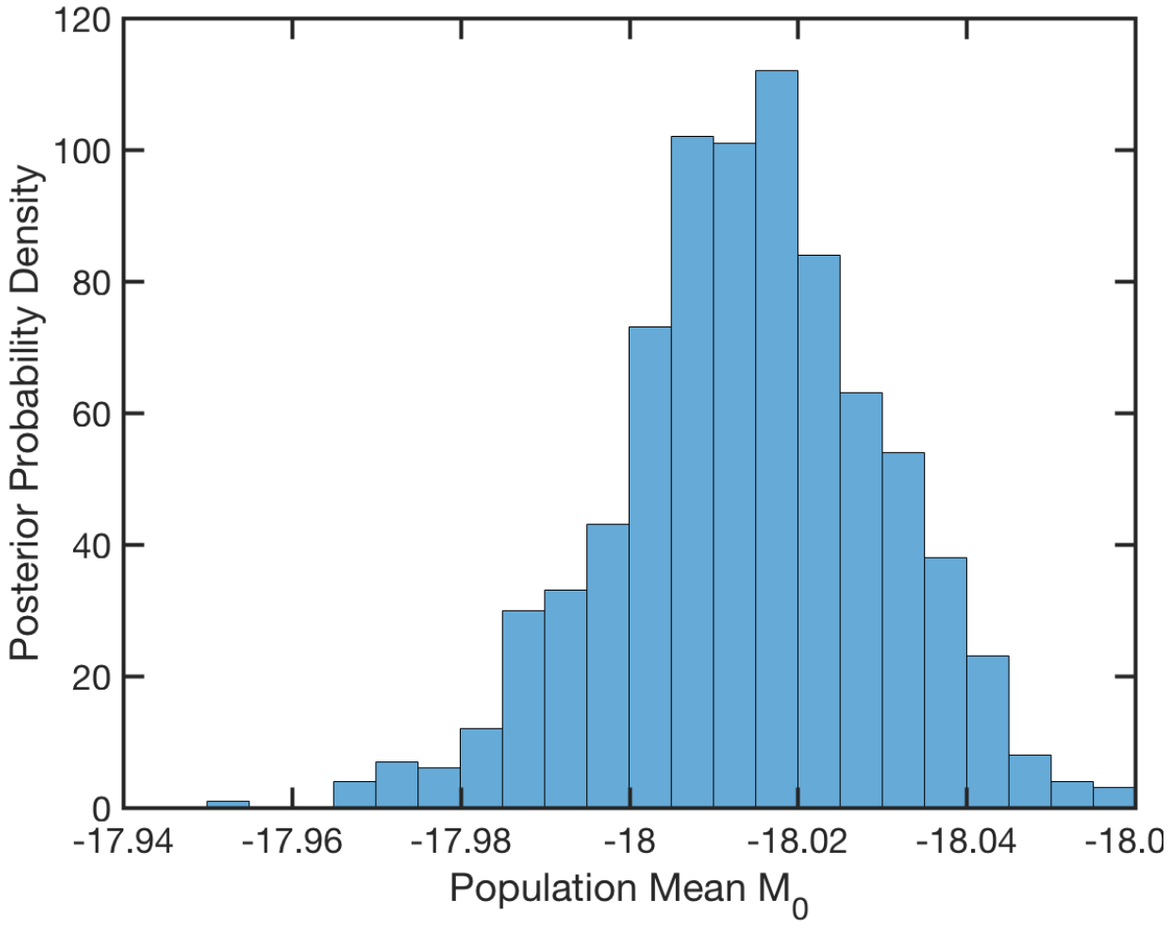
100 Supernovae with
app mag and
distance measurements

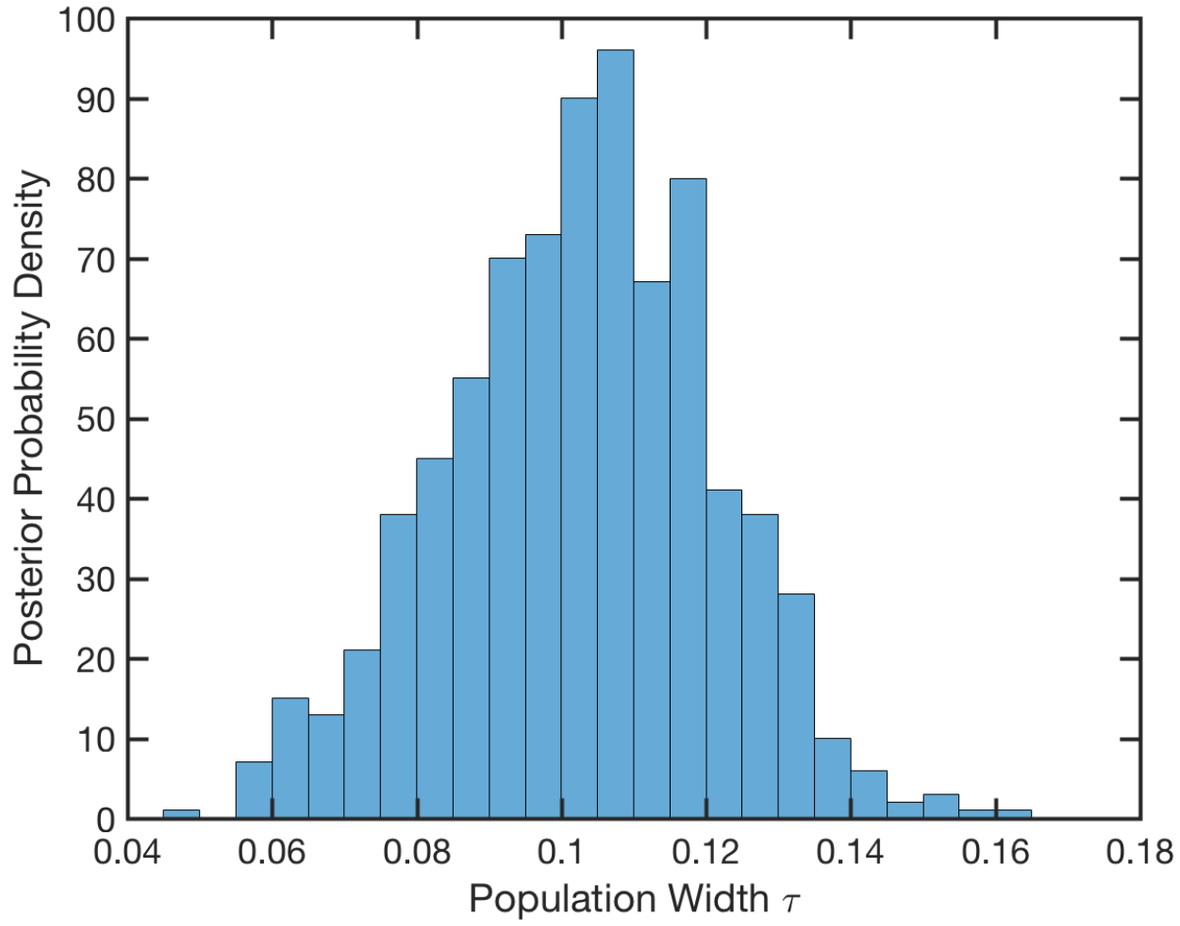




Hyperparameter Inference

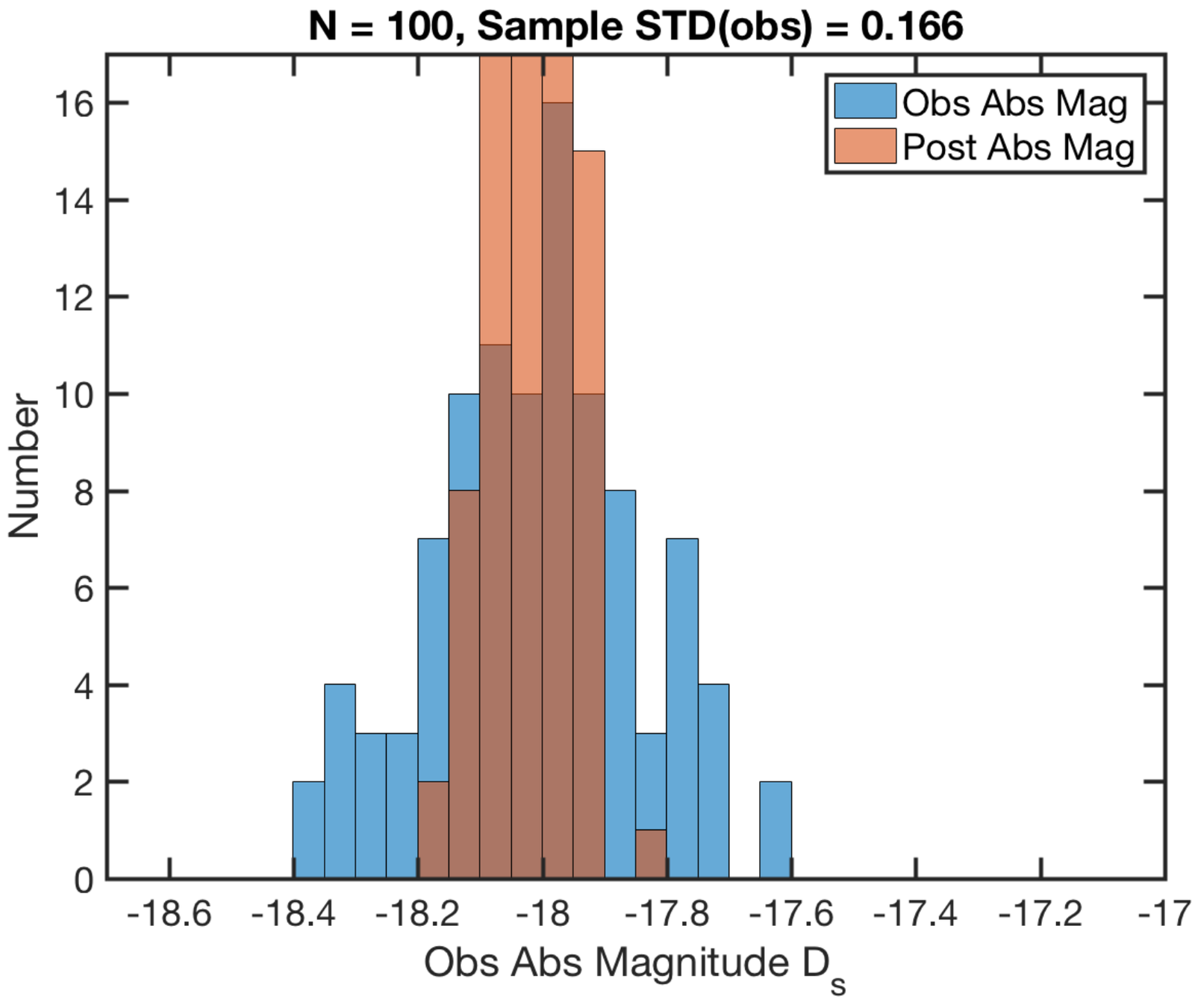
Marginal Posterior Histogram Estimates
from Gibbs Sampling in 102 dimensions





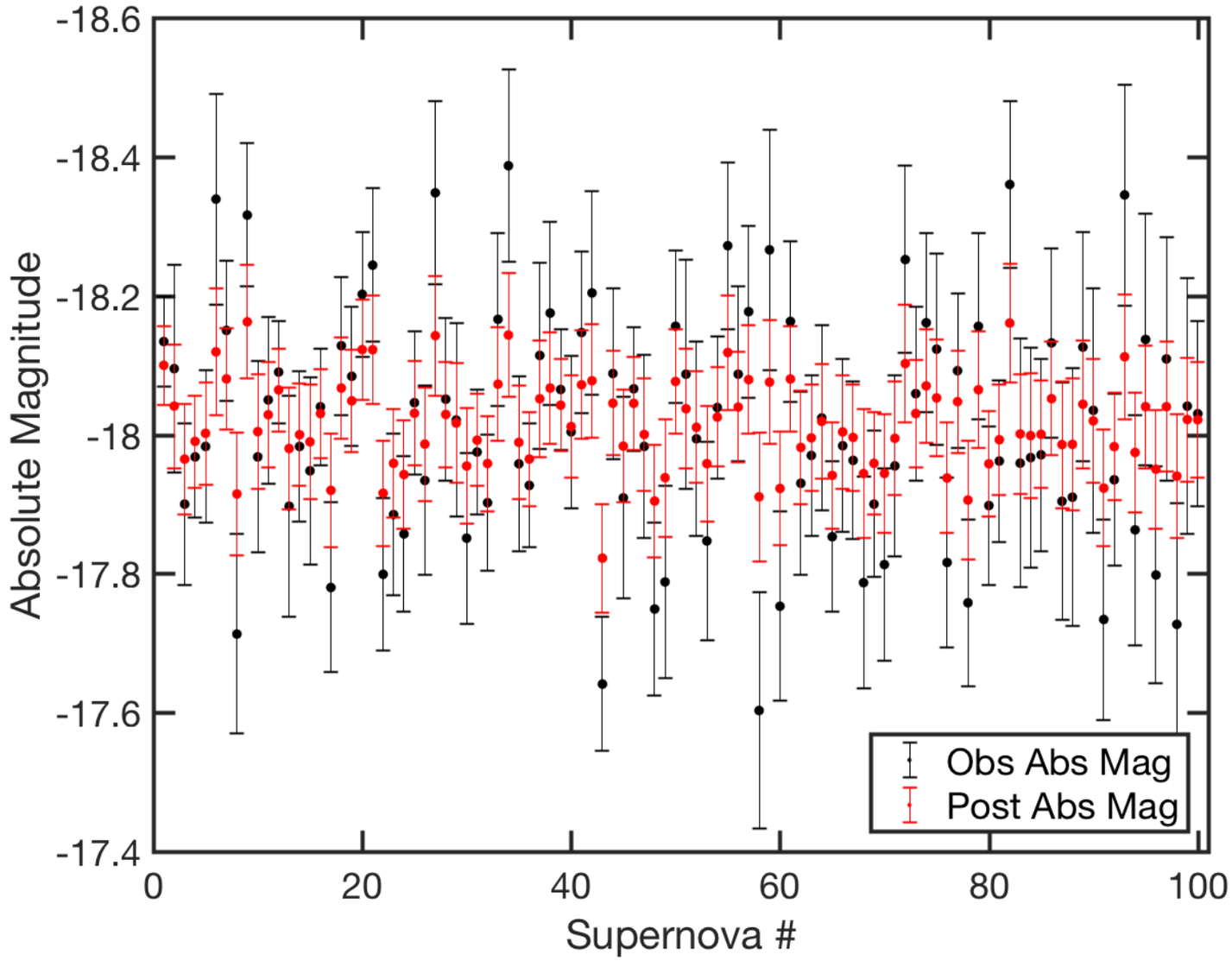
Latent Variable(s) Inference

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



Latent Variable(s) Inference

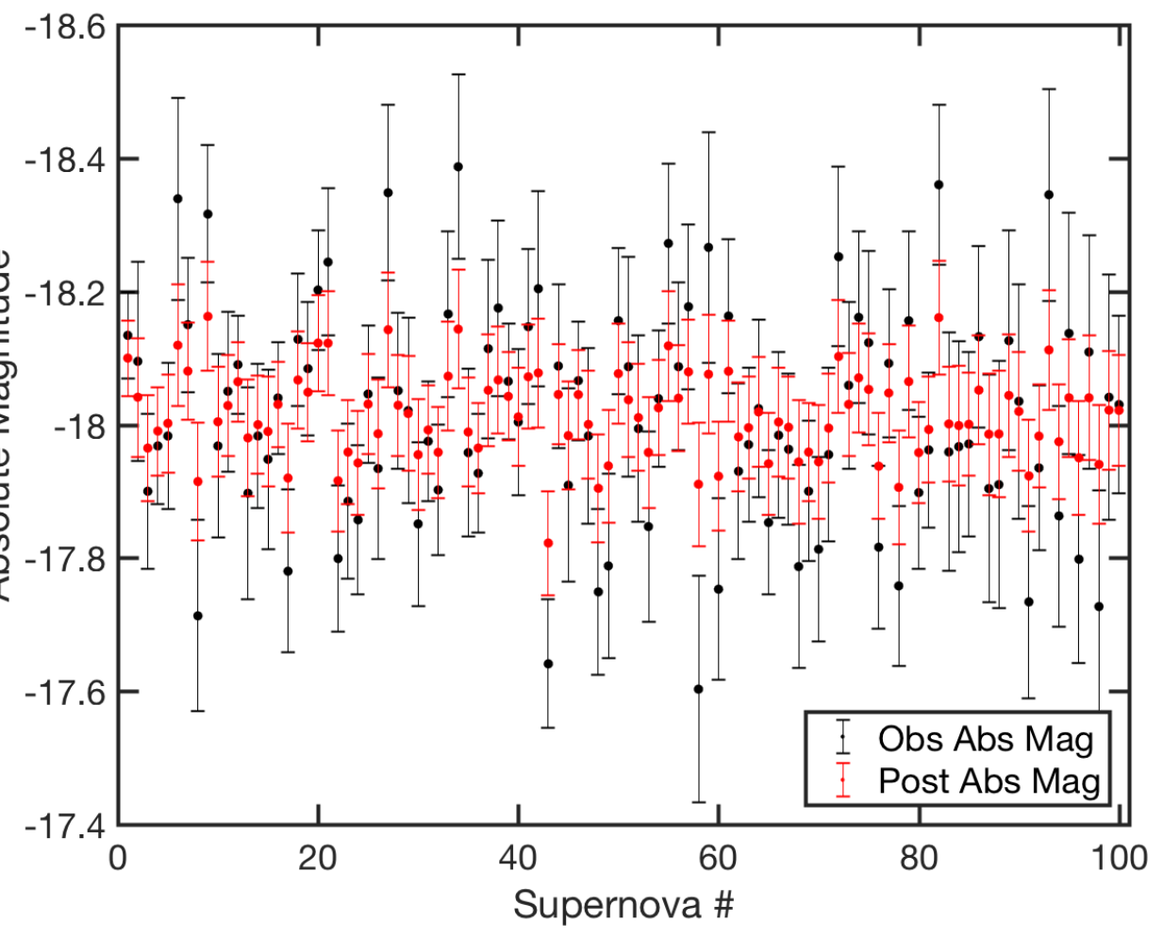
Histogram of $\mathbb{E}[M_s|\boldsymbol{D}]$ posterior estimates from MCMC

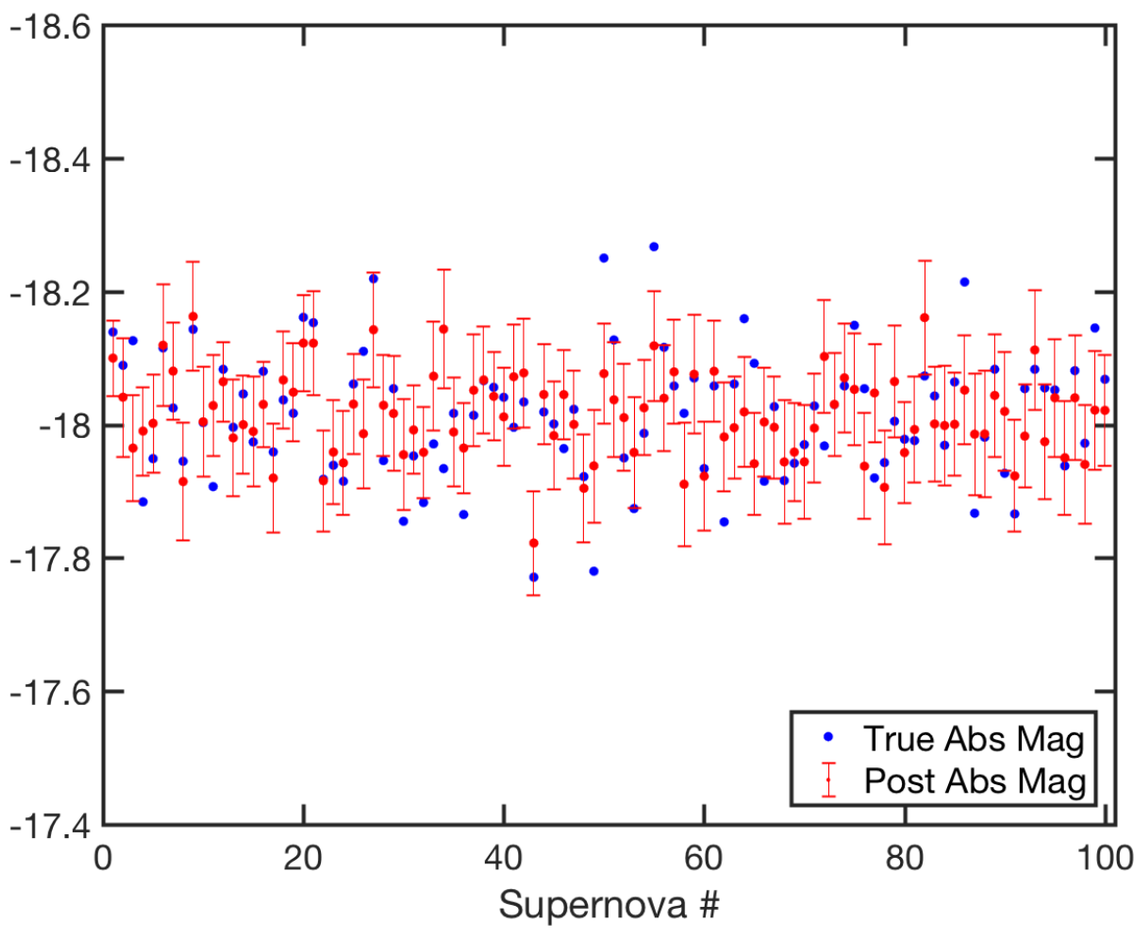


What's going on here?

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

Shrinkage Estimators

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

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

Loredo & Hendry (2010)

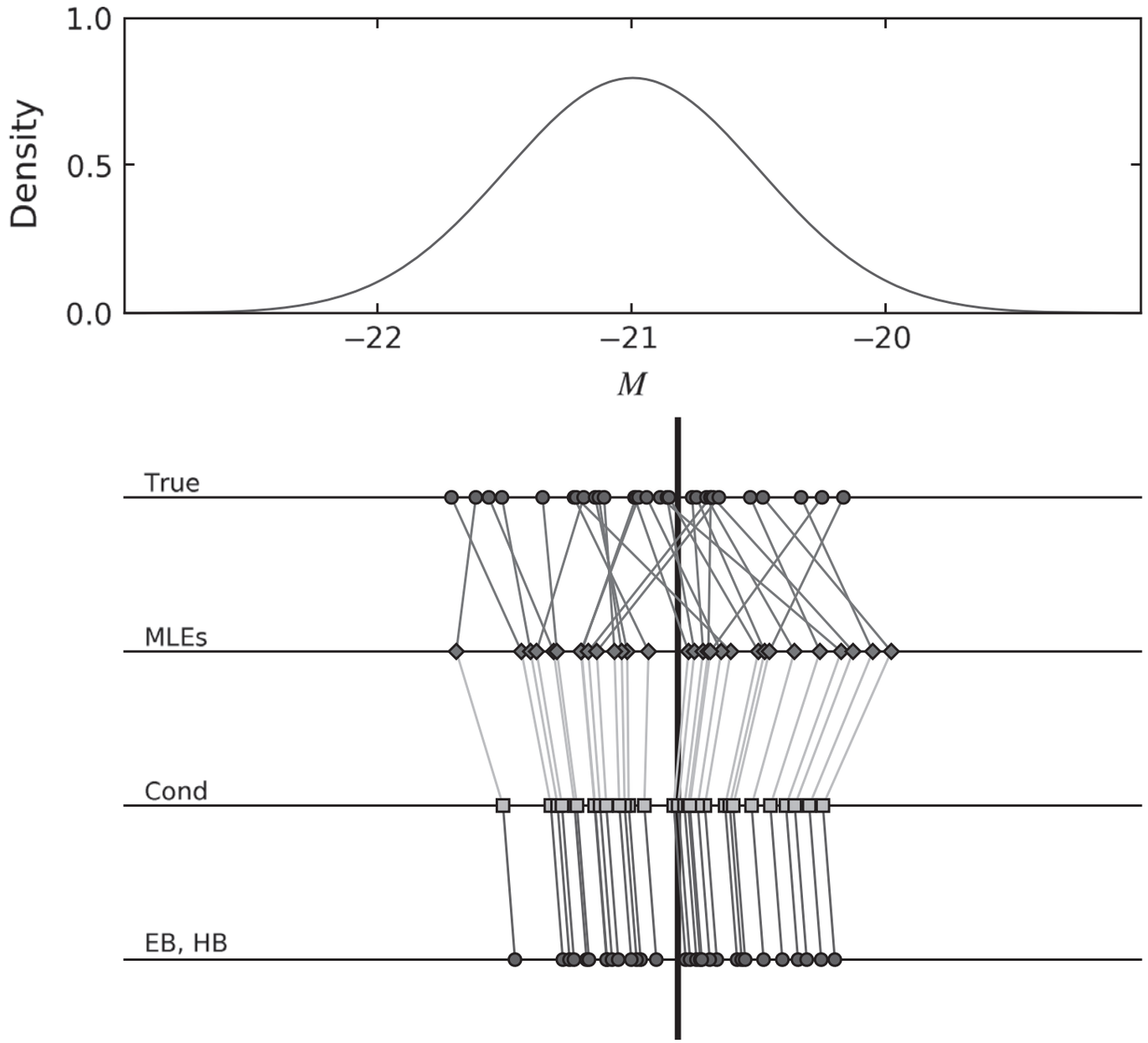
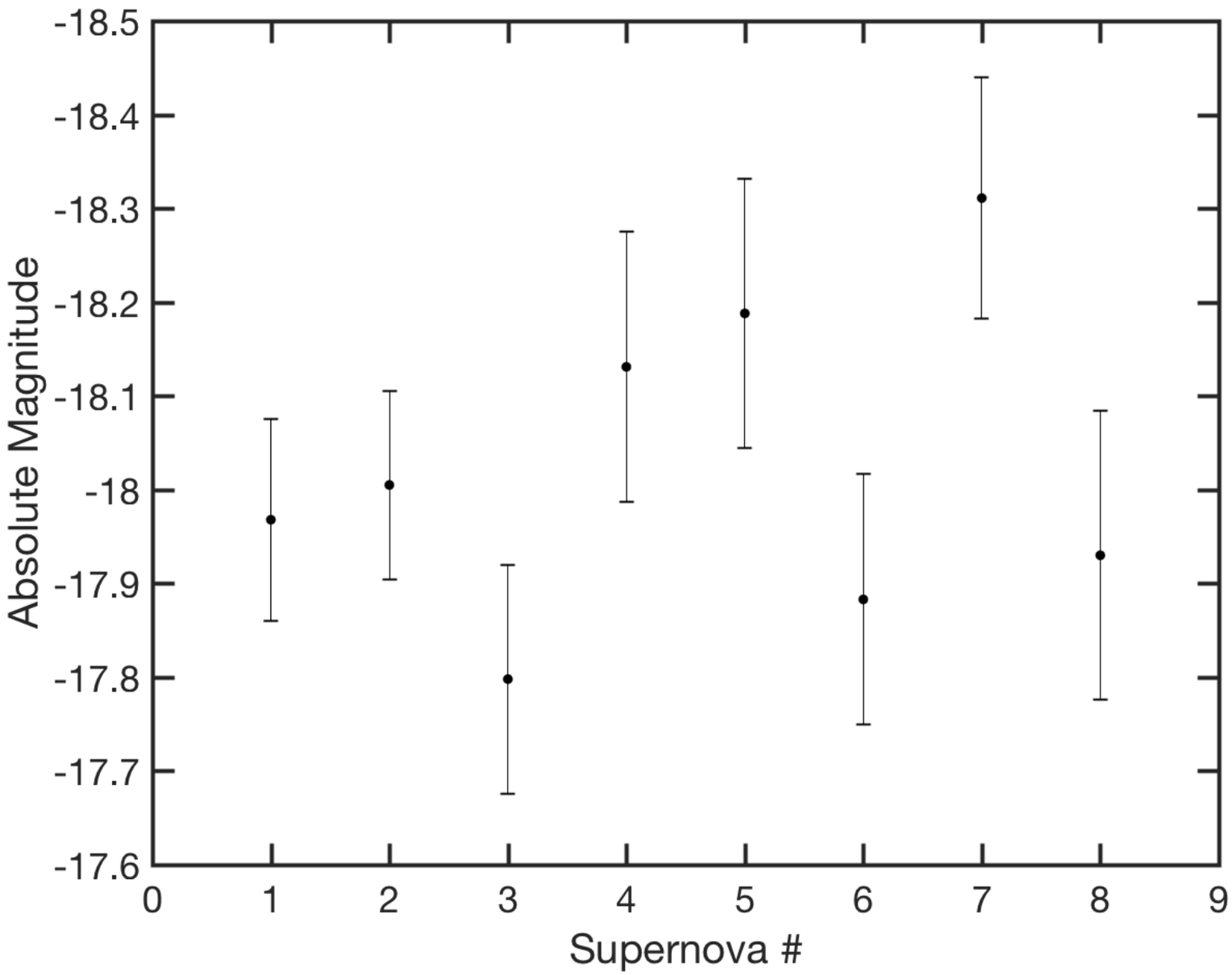


Fig. 11.2. Shrinkage in a simple normal–normal model. Top panel shows population distribution. 'True' axis shows M_i values of 30 samples. Remaining axes show estimates from measurements with $\sigma = 0.3$ normal error: MLEs, conditional (on the true mean), and empirical/hierarchical Bayes estimates.

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

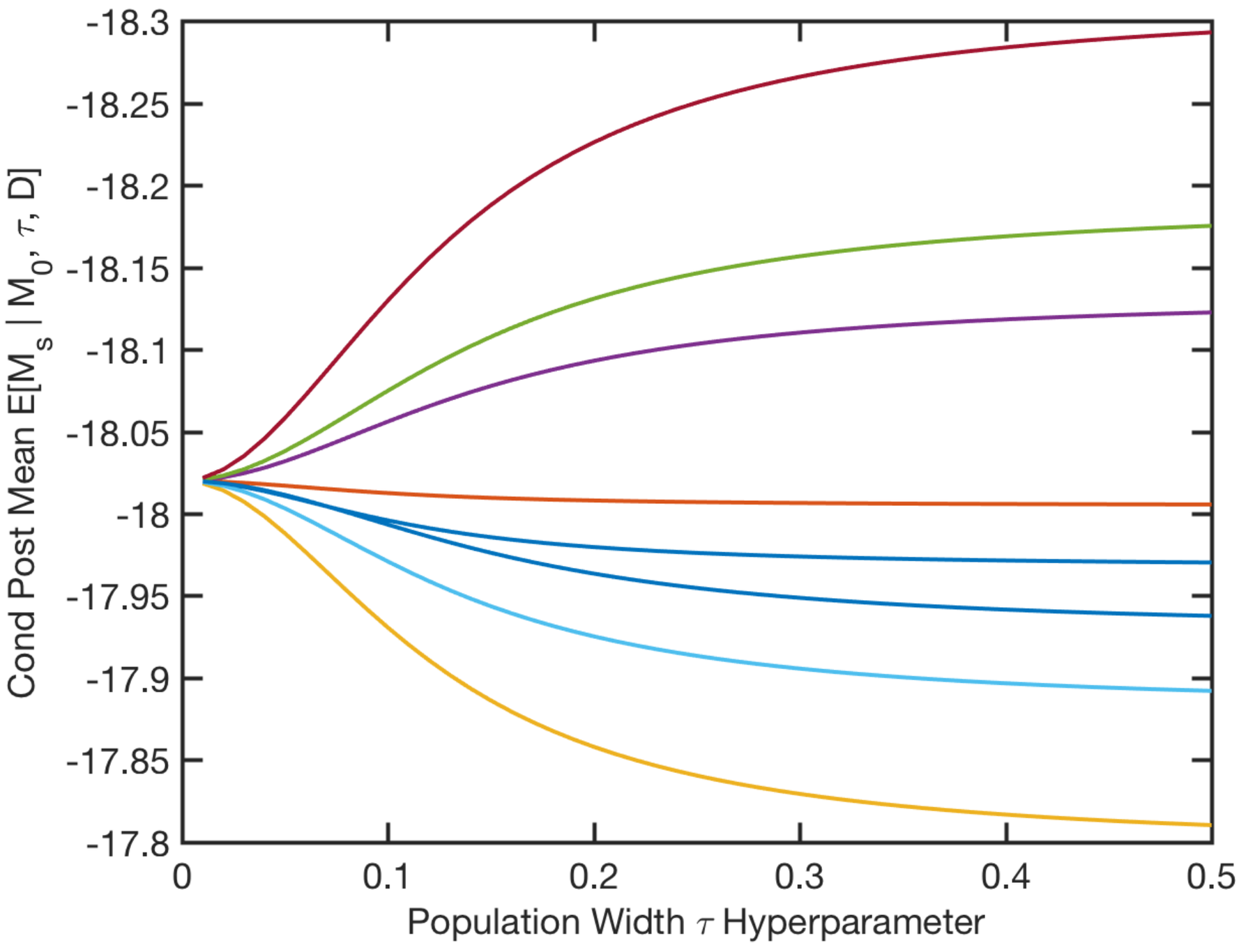
How does Hierarchical Bayes implement shrinkage?

First, let's shrink the dataset



Conditional Posterior of Individual

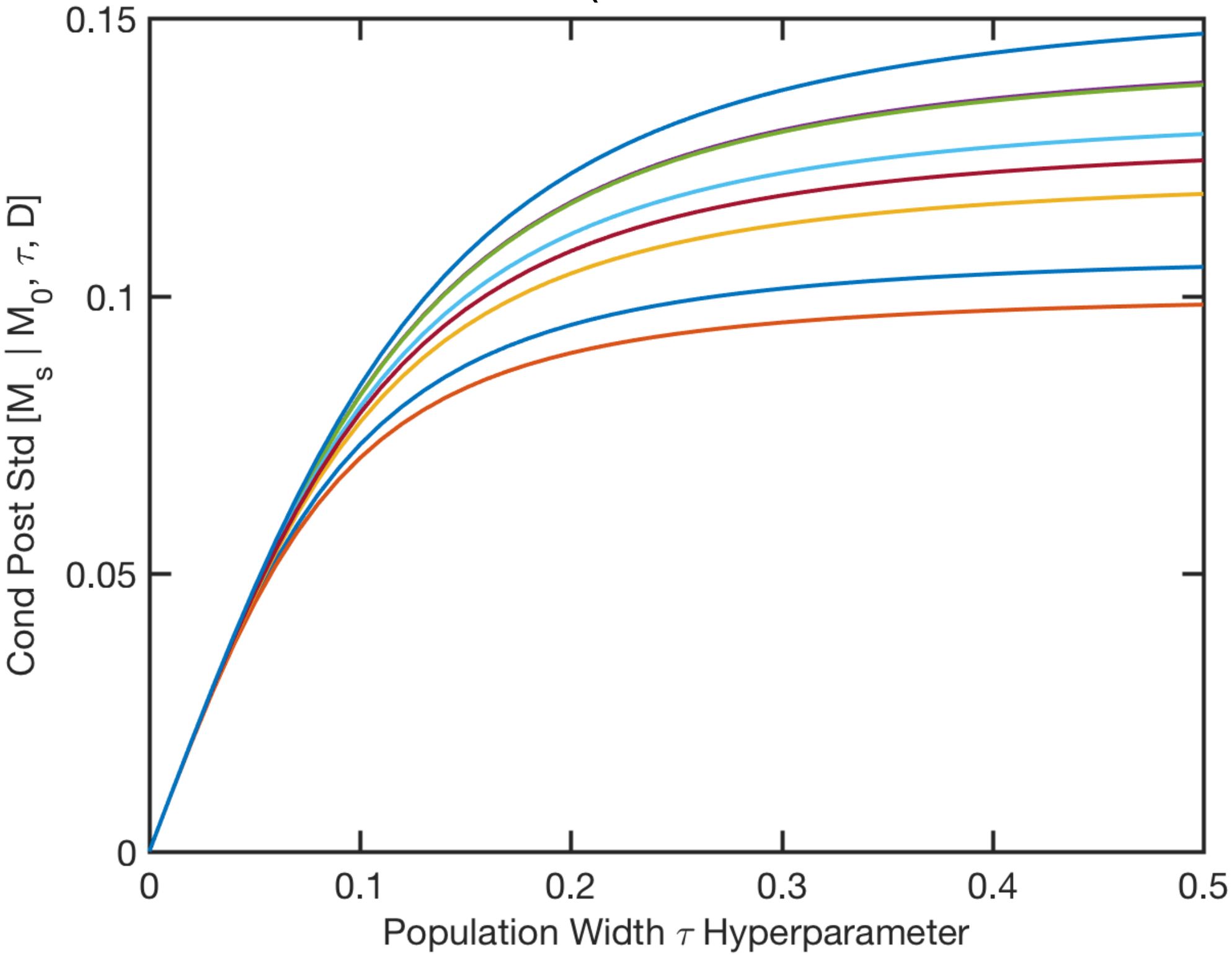
$P(M_s|\, M_0, \tau, \boldsymbol{D})$ (Posterior Mean)



τ controls the degree of shrinkage

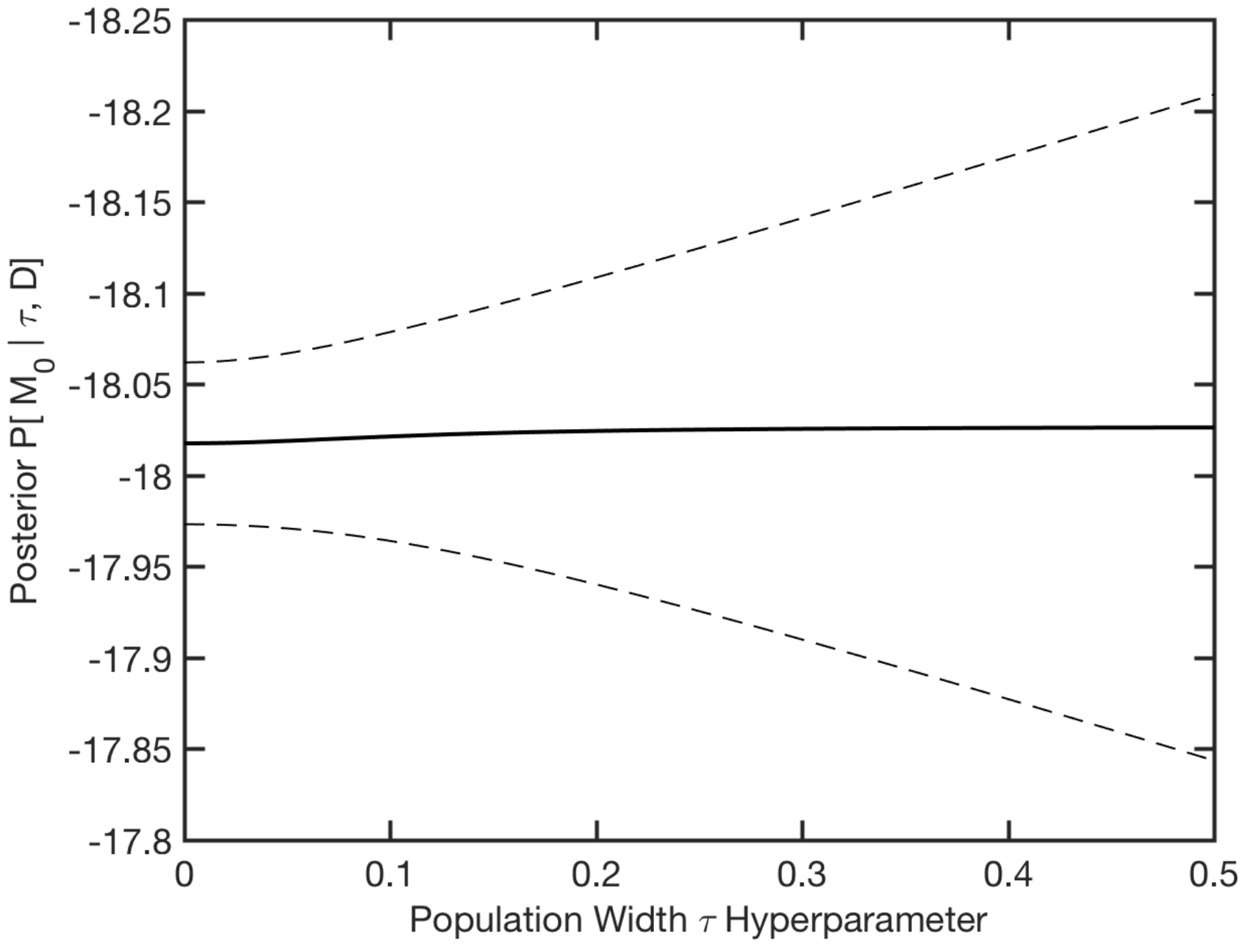
Conditional Posterior of Individual

$P(M_s|\, M_0, \tau, \boldsymbol{D})$ (Posterior Standard Deviation)

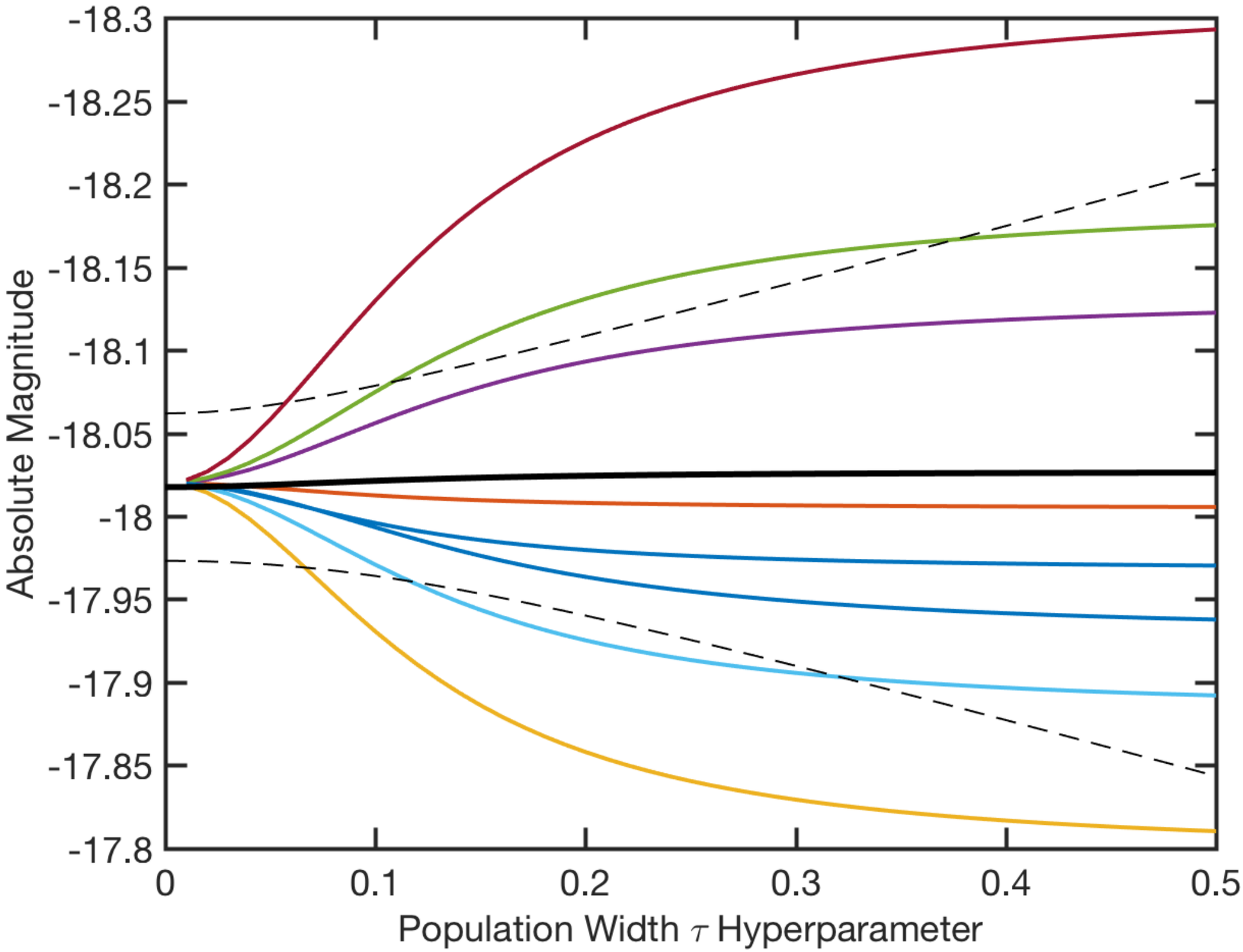


Conditional Posterior of Population

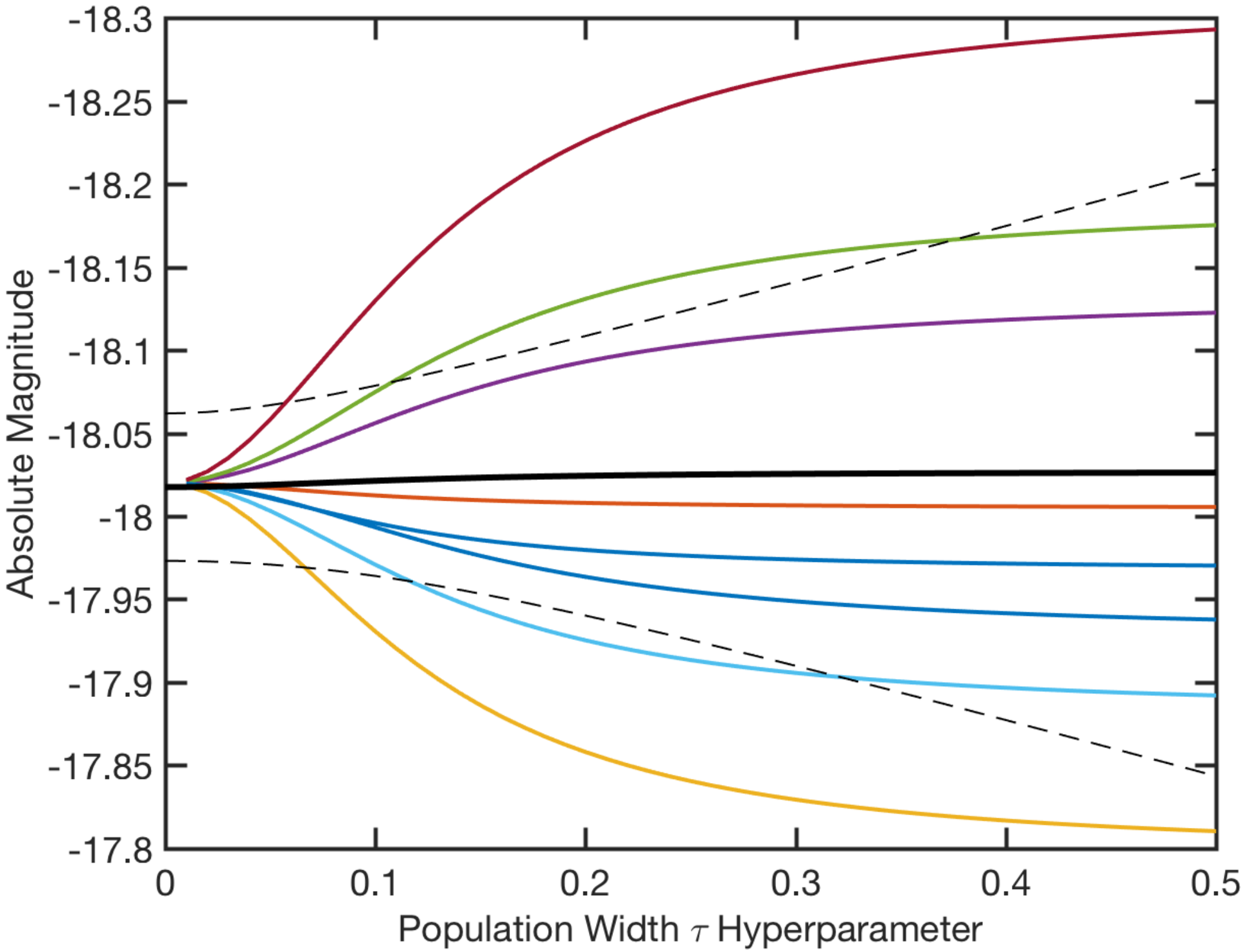
$$P(M_0|\,\tau, \boldsymbol{D})$$



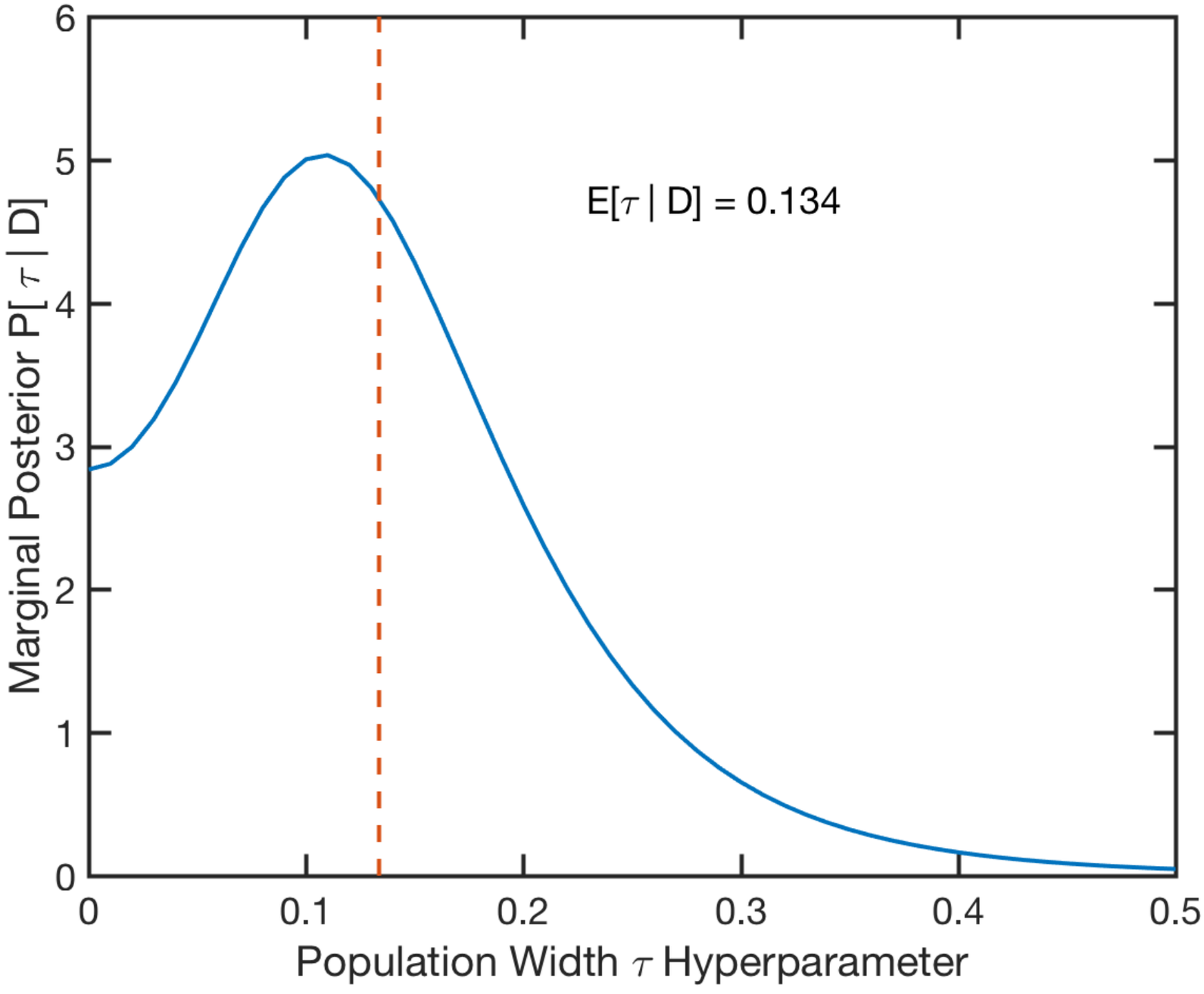
(Conditional) Pop Mean vs. Individual Means



How much shrinkage to apply?

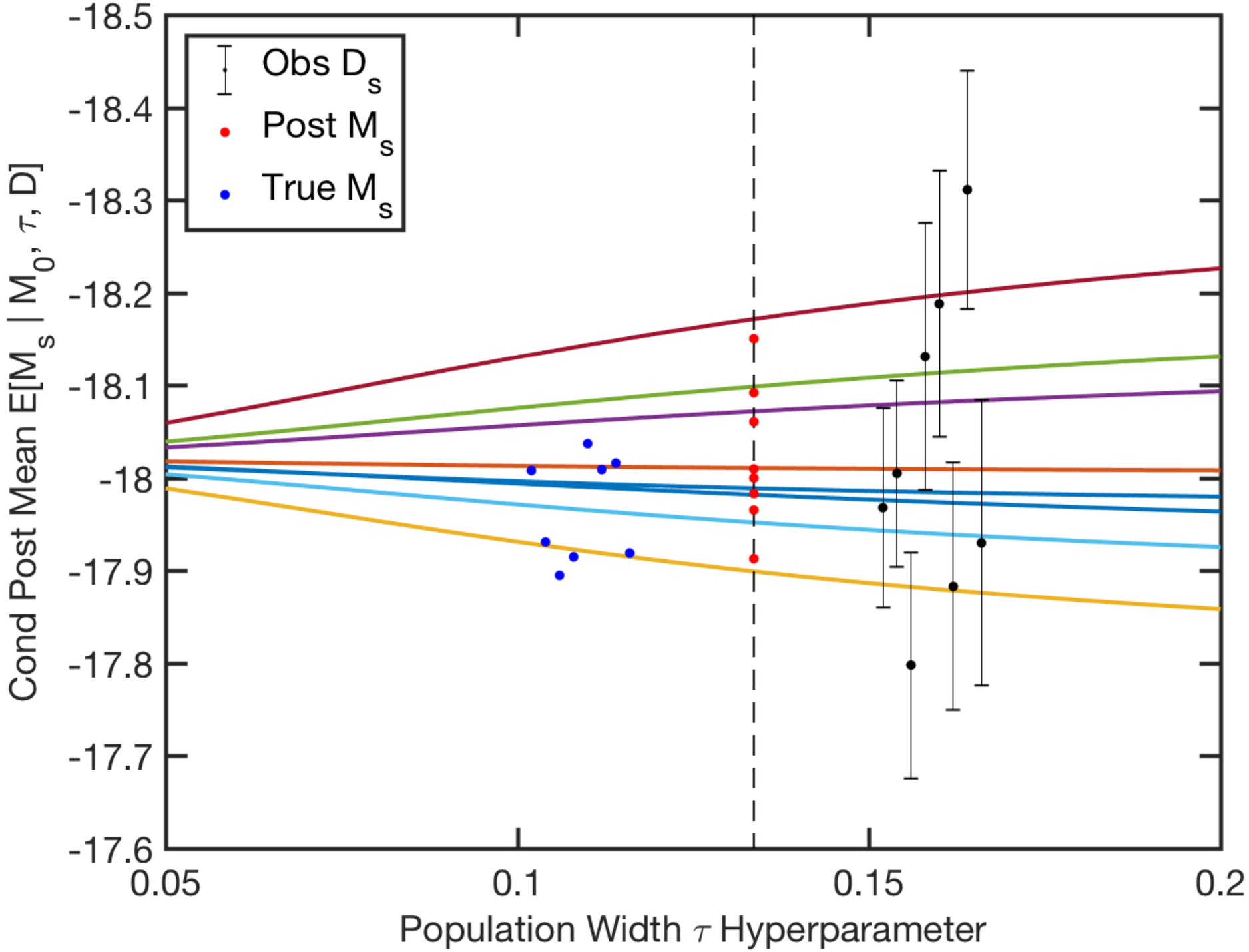


Hierarchical Bayes gives you the posterior density in τ

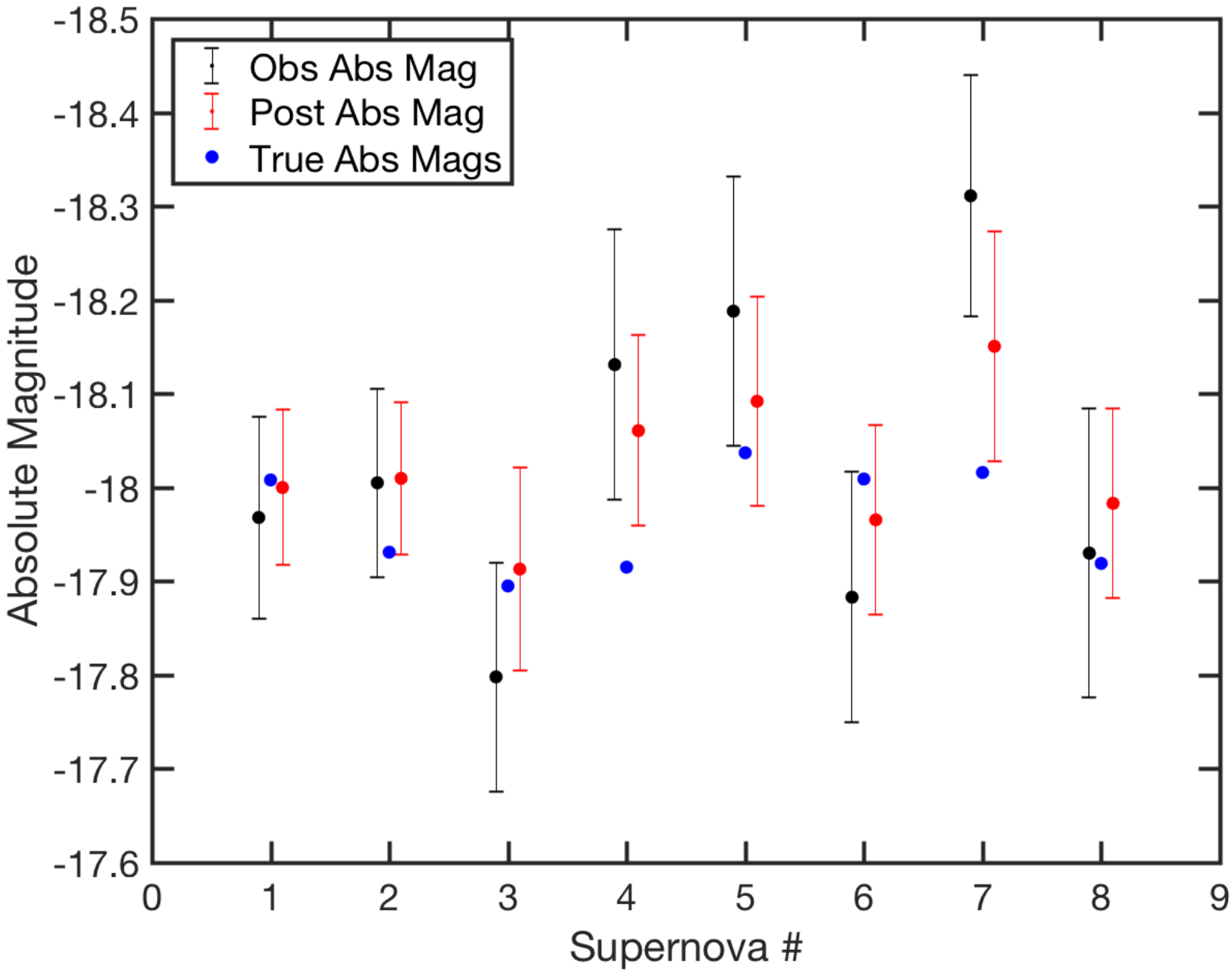


τ controls the degree of shrinkage

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:

$$\xi_i \sim N(\mu, \tau^2) \tag{1}$$

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

$$x_i|\xi_i \sim N(\xi_i, \sigma^2_{x,i}) \tag{3}$$

$$y_i|\eta_i \sim N(\eta_i, \sigma^2_{y,i}) \tag{4}$$

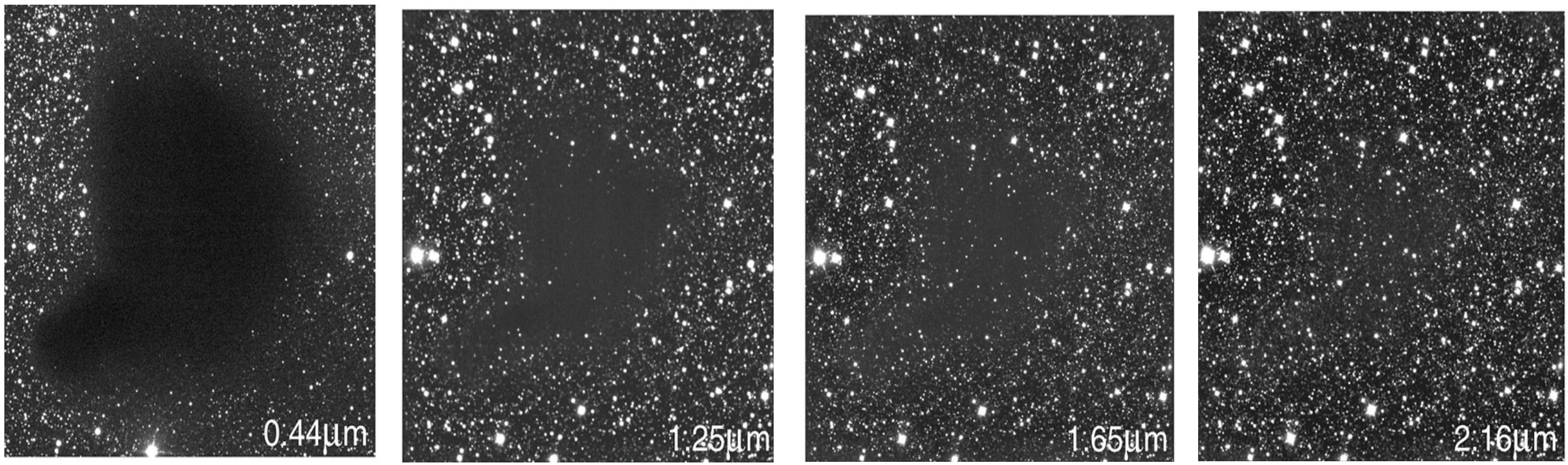
The astronomer measures values $\mathcal{D} = \{x_i, y_i\}$ with known measurement error variances $\{\sigma^2_{x,i}, \sigma^2_{y,i}\}$, for $i = 1, \ldots, N$ quasars.

(Draw PGM)

Example: supernova colours
Interstellar Dust is a real physical effect

Optical light

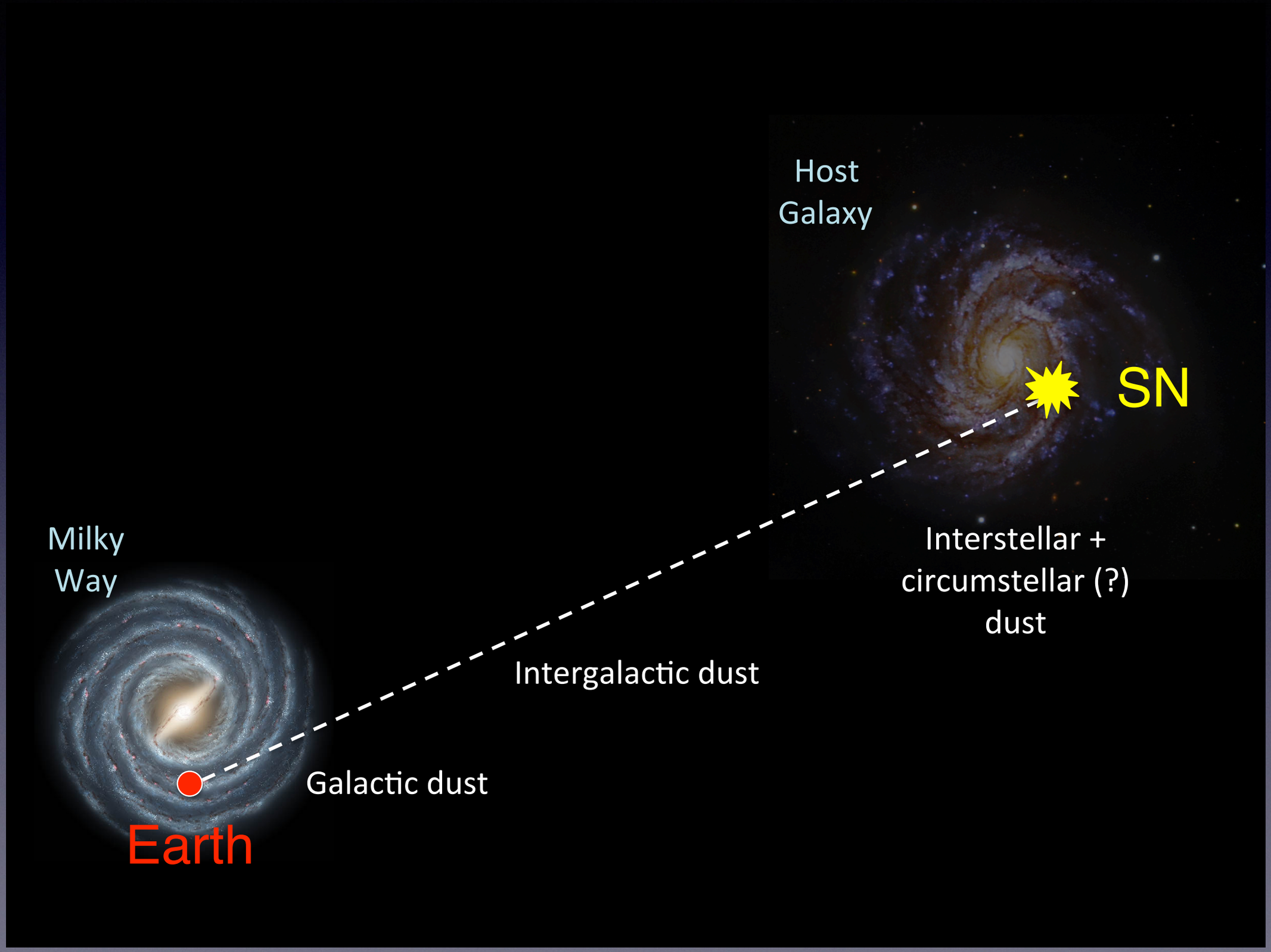
Near Infrared Wavelengths



B-band J-band H-band K-band

Seeing through interstellar dust

From Dust to Dust



What about the host galaxy dust?

Dust Absorption vs. Wavelength of Light

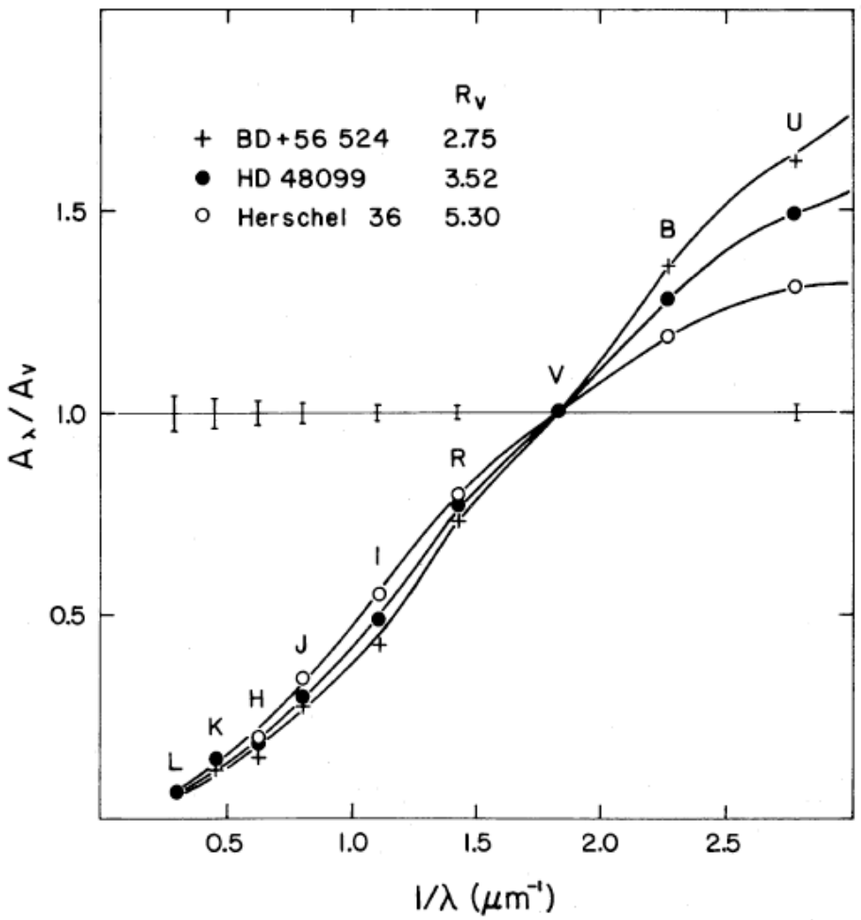


FIG. 3.—Comparison between the mean optical/NIR R_V-dependent extinction law from eqs. (2) and (3) and three lines of sight with largely separated R_V values. The wavelength position of the various broad-band filters from which the data were obtained are labeled (see Table 3). The "error" bars represent 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 ~ Amount of Dust Absorption (dimming) in B-band (450nm)
  - $R_V = A_V/E(B\text{-}V)$ ~ Wavelength Dependence of Dust Absorption

Another example: Supernova colours

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

Intrinsic Colour:

$$C_s = M^B_{\mathrm{int},s} - M^V_{\mathrm{int},s}$$

(Reddening)

$$E_s \equiv A^B_s - A^V_s$$

Apparent Colour:

$$O_s = m^B_s - m^V_s = C_s + A^B_s - A^V_s = C_s + E_s$$

$$C_s \sim N(\mu_C, \sigma^2_C) \qquad E_s \sim \mathrm{Exponen}(\tau)$$

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

Supernova colours

Intrinsic Colour Population

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

Dust Reddening Population

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

Individual Measurement Error

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

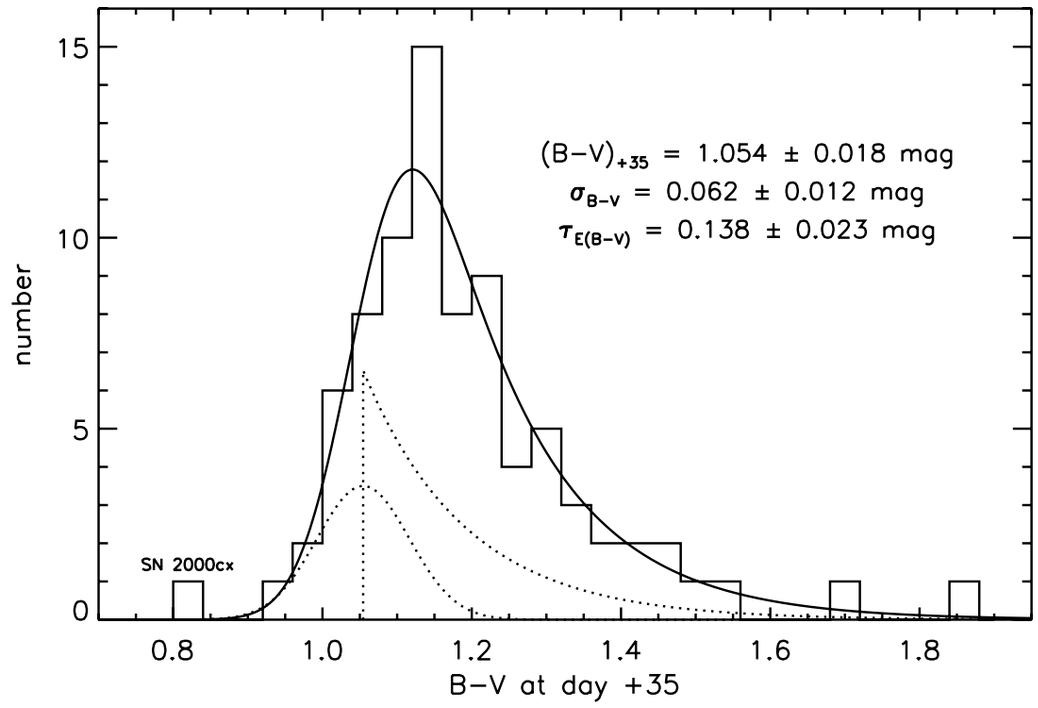
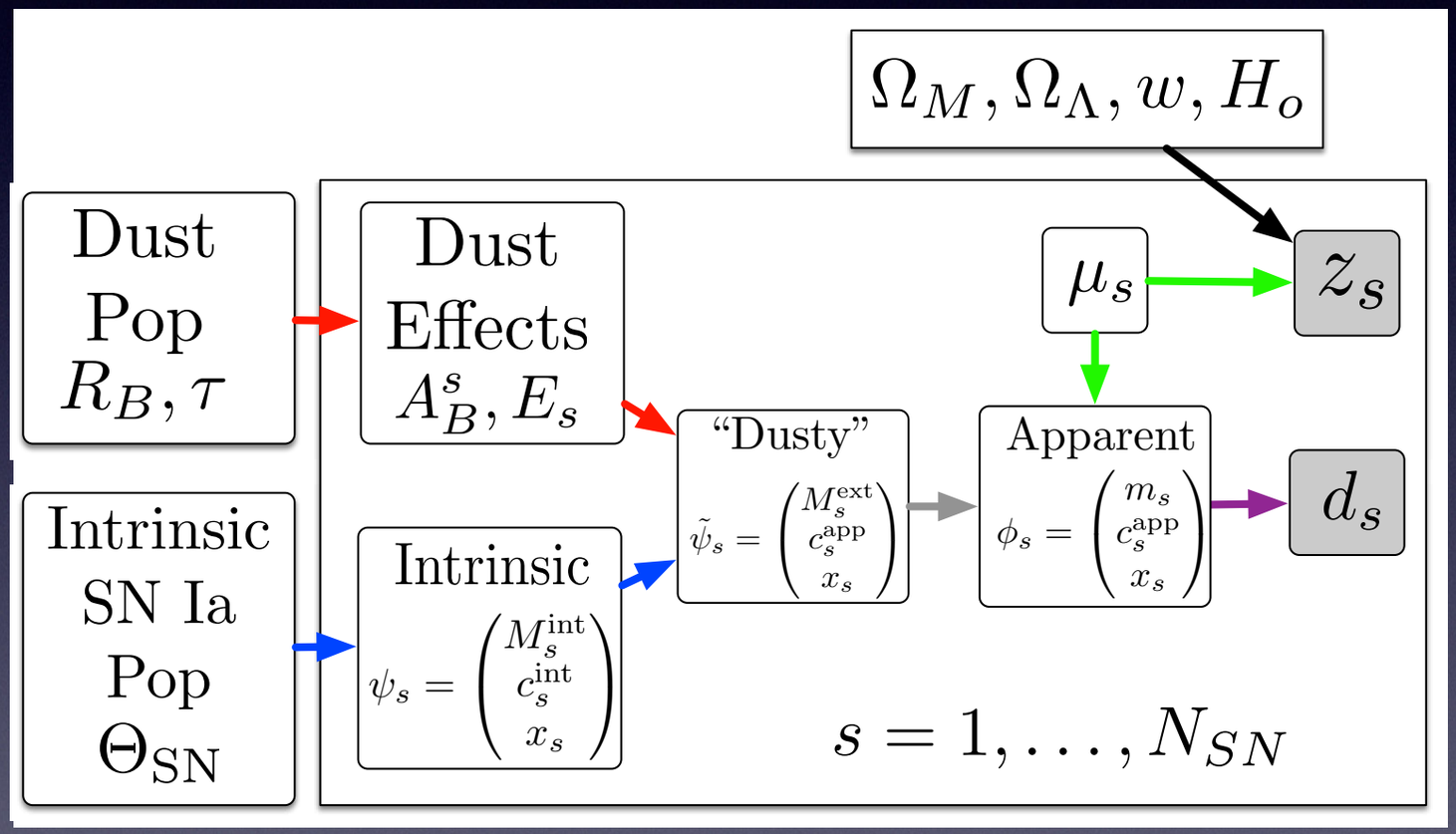


FIG. 6.—Histogram of 82 SNe Ia with well-measured late-time $B - V$ color evolution. The data were corrected for Galactic extinction and the K-correction and referenced to +35 days after B maximum, adopting a late-time color evolution slope of -0.0118 mag day^{-1}. The maximum likelihood fit model is 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



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 ΛCDM 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 $\Omega_{\rm m}h^2$ (physical matter density), $\Omega_{\rm b}h^2$ (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
$\Omega_{\rm m}h^2$	0.128 ± 0.008	0.126 ± 0.005	0.132 ± 0.004
$\Omega_{\rm b}h^2$	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]
$m_{\rm dm}$	(warm) dark matter mass
$w+1$	dark energy equation of state
$\mathrm{d}w/\mathrm{d}z$	redshift dependence of w
	[or more complex parameterization of dark energy evolution]
$c_{\rm S}^2 - 1$	effects of dark energy sound speed
$1/r_{\rm top}$	topological identification scale
	[or more complex parameterization of non-trivial topology]
$\mathrm{d}\alpha/\mathrm{d}z$	redshift dependence of the fine structure constant
$\mathrm{d}G/\mathrm{d}z$	redshift dependence of the gravitational constant
$\mathrm{d}n/\mathrm{d}\ln k$	running of the scalar spectral index
$k_{\rm cut}$	large-scale cut-off in the spectrum
$A_{\rm feature}$	amplitude of spectral feature (peak, dip or step) ...
$k_{\rm feature}$	... and its scale
	[or adiabatic power spectrum amplitude parameterized in N bins]
$f_{\rm NL}$	quadratic contribution to primordial non-Gaussianity
	[or more complex parameterization of non-Gaussianity]
r	tensor-to-scalar ratio
$r + 8n_{\rm T}$	violation of the inflationary consistency equation
$\mathrm{d}n_{\rm T}/\mathrm{d}\ln k$	running of the tensor spectral index
$\mathcal{P}_S$	CDM isocurvature perturbation S ...
n_S	... and its spectral index ...
$\mathcal{P}_{S\mathcal{R}}$	... and its correlation with adiabatic perturbations ...
$n_{S\mathcal{R}} - n_S$	... and the spectral index of that correlation
	[or more complicated multi-component isocurvature perturbation]
$G\mu$	cosmic string component of perturbations

Bayesian Model Comparison

Parameter Estimation: Posterior on parameters

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

Model Selection: Posterior on Models:

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

(Needs proper prior!)

Posterior Odds Ratio

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

$$\mathrm{BayesFactor}_{12} = \frac{P(\boldsymbol{D}|\, M_1)}{P(\boldsymbol{D}|\, M_2)}$$

Bayesian Model Comparison (switch to Bishop's slides)

Occam's Razor

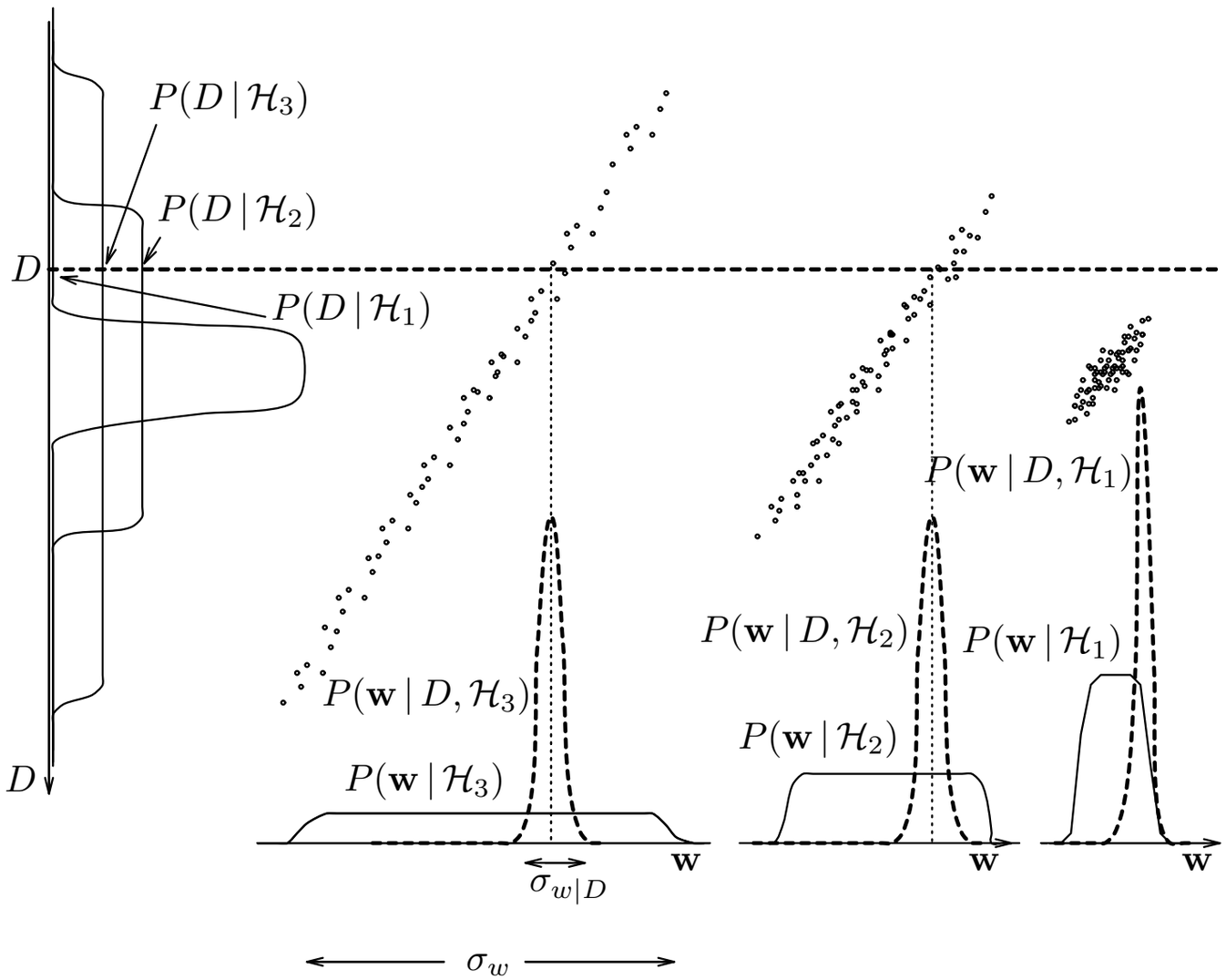


Figure 28.6. A hypothesis space consisting of three exclusive models, each having one parameter $\mathbf{w}$, and a one-dimensional data set D. The 'data set' is a single measured value which differs from the parameter $\mathbf{w}$ by a small amount of additive noise. Typical samples from the joint distribution $P(D, w, \mathcal{H})$ 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 $\mathbf{w}$ 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

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

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

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

$$P^*(\boldsymbol{\theta}) = P(\boldsymbol{D}|\boldsymbol{\theta}) P(\boldsymbol{\theta})$$

Find MAP estimate: $\boldsymbol{\theta}_0 = \text{argmax}_{\boldsymbol{\theta}} \ln P^*(\boldsymbol{\theta})$

Taylor Expansion

$$\ln P^*(\boldsymbol{\theta}) \approx \ln P^*(\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{A} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \ldots$$

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

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

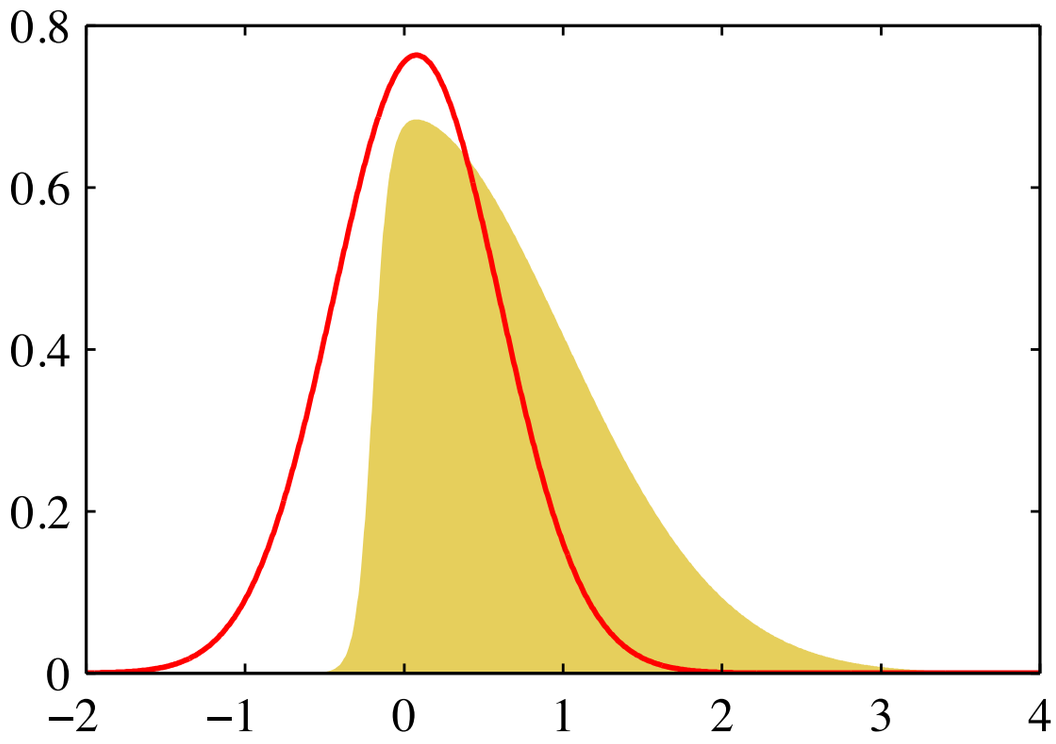
The Laplace Approximation

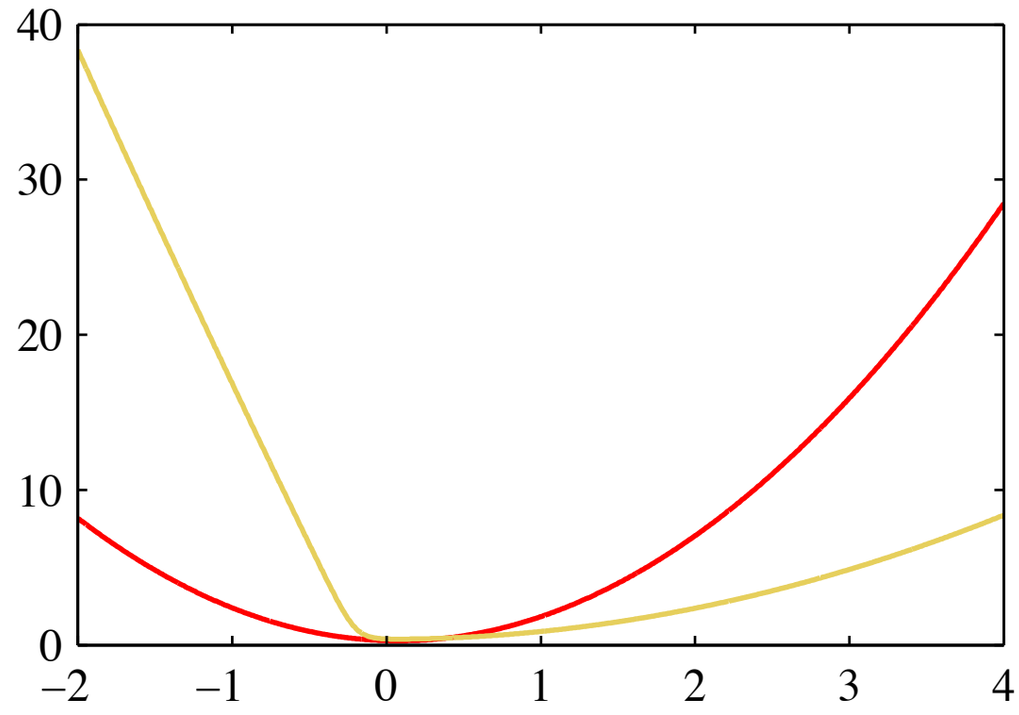
Taylor Expansion

$$\ln P^*(\boldsymbol{\theta}) \approx \ln P^*(\boldsymbol{\theta}_0) - \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^T \boldsymbol{A}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \ldots$$

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

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





The Laplace Approximation

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

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

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

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

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

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

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

The Laplace Approximation

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

Evidence Best-Fit Likelihood Occam Factor

(Ratio of Posterior to Prior Width)

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: $\{y_i, \lambda_i\}$

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

Example: using Laplace Approximation

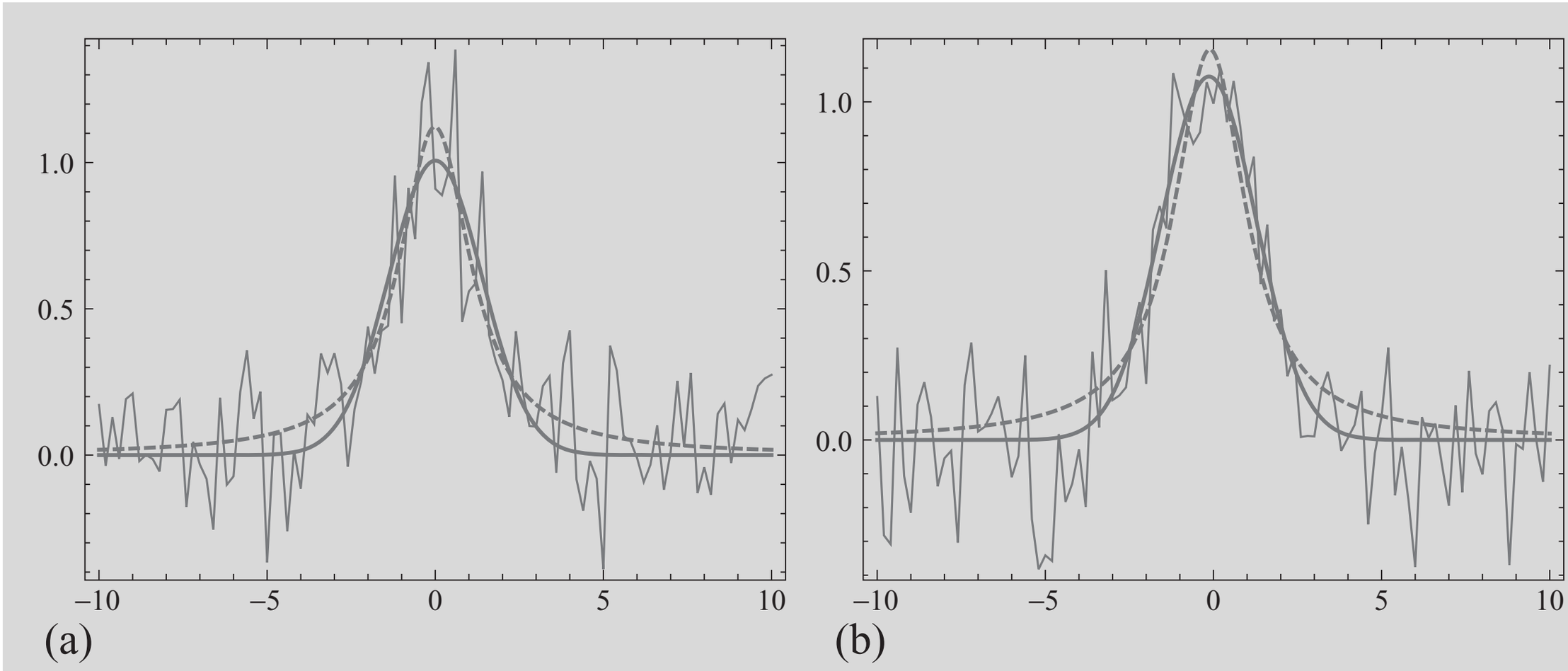


Figure 7.1 Spectral-line simulations. These are at a signal-to-noise ratio, in the peak of the line, of 5. (a) The Lorentzian fit is favoured at odds of 30:1, whereas (b) the Gaussian is favoured at odds of 100:1.

Wall & Jenkins 2012

Ways to calculate Evidence

- Analytic (really simple problems)
- Monte Carlo: probably slow if likelihood is peaked

$$\boldsymbol{\theta}_i \sim P(\boldsymbol{\theta}_i)$$

$$P(\boldsymbol{D}) \approx \sum_{i=1}^{M} P(D|\boldsymbol{\theta}_i)$$

- Harmonic Mean Estimator (bad!)
- Laplace Approximation
- Thermodynamic Integration
- Nested Sampling