Unified Analyses of Populations of Sources

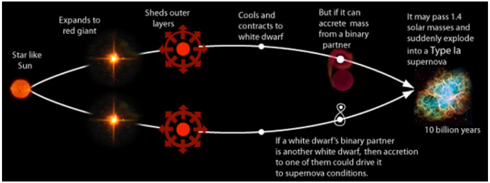
Advantages of “Shrinkage Estimates” in Astronomy

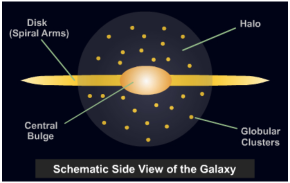
David A. van Dyk

Statistics Section, Imperial College London

University of Birmingham
April 2015

Populations of Sources



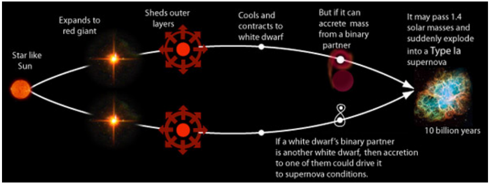


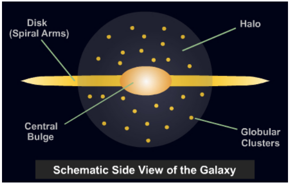
Estimating a property of each object in a population:

1. Intrinsic (absolute) magnitudes of Type Ia Super Novae. Or more simply: apparent magnitudes.
2. The ages of White Dwarfs in the galactic halo. Or more simply: ages of WDs in the galaxy.
3. Measured distance to Large Magellanic Cloud
   - ⋆ With different methods, each with their own systematics.

Imperial College London

Estimating Source Characteristics





Typical Strategy: Estimate the magnitude, distance, or age for each source in a separate data analysis.

Another Possibility: Preform unified analysis, modeling dist'n of magnitudes, distances, or ages among sources.

- Relative advantages depends on *quality of individual estimates* and *degree of homogeneity* in population.
- Discuss from Frequntist and Bayesian perspectives.

Imperial College London

Outline

1. All Roads Lead to Rome
   - Frequentist origins of shrinkage estimates
   - Bayesian hierarchical models
2. Example 1: Using SNIa to Fit Cosmological Models
   - Joint with Roberto Trotta, Xiyun Jiao, & Hikmatali Shariff
3. Example 2: Ages of White Dwarfs in the Galactic Halo
   - Joint work with Ted von Hippel & Shijing Si

Imperial College London

Outline

The Sample Mean

Suppose we wish to estimate a parameter, θ, from repeated measurement or a single source:

$$y_i \stackrel{\text{indep}}{\sim} \mathrm{N}(\theta, \sigma^2) \quad \text{for} \quad i = 1, \ldots, n$$

Eg: calibrating detector from n measures of known source.

An obvious estimator:

$$\hat{\theta}^{\text{naive}} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

What is not to like about the arithmetic average?

Imperial College London

Frequency Evaluation of an Estimator

- How far off is the estimator?

$$(\hat{\theta} - \theta)^2$$

- How far off do we expect it to be?

$$\text{MSE}(\hat{\theta}|\theta) = \text{E}\left[(\hat{\theta} - \theta)^2 \mid \theta\right] = \int \left(\hat{\theta}(y) - \theta\right)^2 f(y \mid \theta)dy$$

- This quantity is called the Mean Square Error of $\hat{\theta}$.
- An estimator is said to be inadmissible if there is an estimator that is uniformly better in terms of MSE:

$$\text{MSE}(\hat{\theta}|\theta) < \text{MSE}(\hat{\theta}^{\text{naive}}|\theta) \text{ for all } \theta.$$

Imperial College London

Inadmissibility of the Sample Mean

Suppose we wish to estimate more than one parameter:

$$y_{ij} \overset{\text{indep}}{\sim} \text{N}(\theta_j, \sigma^2) \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, G$$

The obvious estimator:

$$\hat{\theta}_j^{\text{naive}} = \frac{1}{n} \sum_{i=1}^{n} y_{ij} \quad \textit{is inadmissible if } G \geqslant 3.$$

The James-Stein Estimator dominates θ^{naive}:

$$\hat{\theta}_j^{\text{JS}} = \left(1 - \omega^{\text{JS}}\right) \hat{\theta}_j^{\text{naive}} + \omega^{\text{JS}} \nu \text{ for any } \nu$$

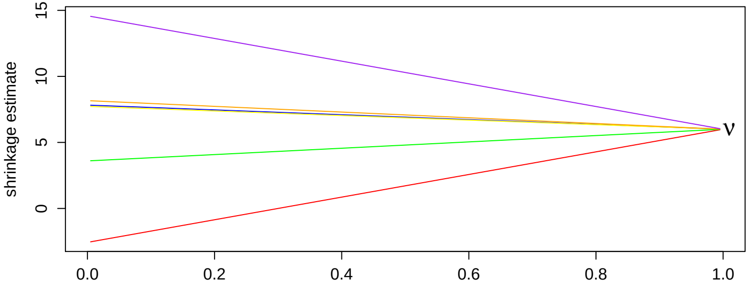
with $\omega^{\text{JS}} \approx \dfrac{\sigma^2/n}{\sigma^2/n + \tau_\nu^2}$ and $\tau_\nu^2 = \text{E}[(\theta_j - \nu)^2]$.

Specifically, $\omega^{\text{JS}} = (G-2)\sigma^2 \Big/ n \sum_{j=1}^{G} (\hat{\theta}_j^{\text{naive}} - \nu)^2$.

Shrinkage Estimators

James-Stein Estimator is a shrinkage estimator:

$$\hat{\theta}_j^{\text{JS}} = \left(1 - \omega^{\text{JS}}\right) \hat{\theta}_j^{\text{naive}} + \omega^{\text{JS}} \nu$$



θ_j highly variable /
data very precise

ω

data highly variable /
θ_j very similar (to ν)

To Whence To Shrink?

James-Stein Estimators

- Dominate the sample average for *any choice* of ν.
- Shrinkage is mild and $\hat{\theta}^{\text{JS}} \approx \hat{\theta}^{\text{naive}}$ for most ν.
- Can we choose ν to maximize shrinkage?

$$\hat{\theta}_j^{\text{JS}} = \left(1 - \omega^{\text{JS}}\right) \hat{\theta}_j^{\text{naive}} + \omega^{\text{JS}} \nu$$

with $\omega^{\text{JS}} \approx \dfrac{\sigma^2/n}{\sigma^2/n + \tau_\nu^2}$ and $\tau_\nu^2 = \mathrm{E}[(\theta_j - \nu)^2]$.

- Minimize τ_ν^2.

The optimal choice of ν is the average of the θ_j.

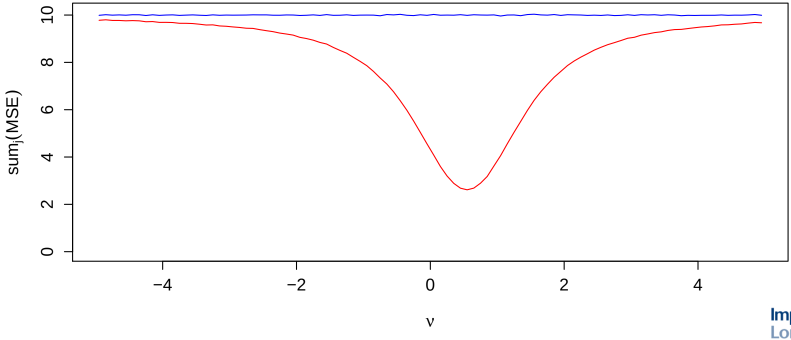
Imperial College London

Illustration

Suppose:

- $y_j \sim \mathrm{N}(\theta_j, 1)$ for $j = 1, \ldots, 10$
- θ_j are evenly distributed on [0,1]

$\mathrm{MSE}(\hat{\theta}^{\text{naive}})$ versus $\mathrm{MSE}(\hat{\theta}^{\text{JS}})$:

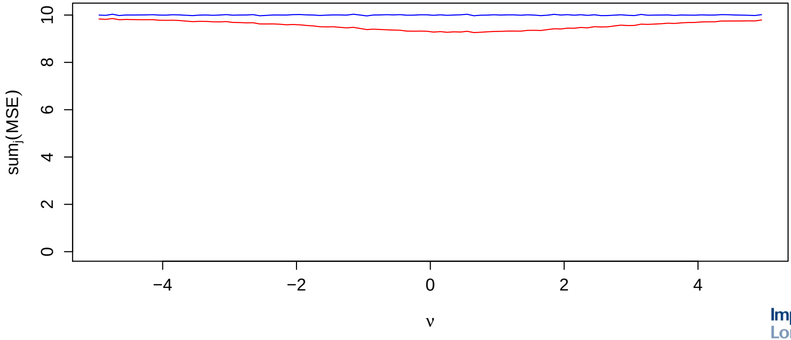


Illustration

Suppose:

- $y_j \sim \mathrm{N}(\theta_j, 1)$ for $j = 1, \ldots, 10$
- θ_j are evenly distributed on **[-4,5]**

$\mathrm{MSE}(\hat{\theta}^{\mathrm{naive}})$ versus $\mathrm{MSE}(\hat{\theta}^{\mathrm{JS}})$:



Imperial College London

Intuition

1. It you are estimating more than two parameters, it is always better to use shrinkage estimators.
2. Optimally shrink toward average of the parameters.
3. Most gain when the naive (non-shrinkage) estimators
   - are noisy (σ^2 is large)
   - are similar (τ^2 is small)
4. Bayesian versus Frequentist:
   - From frequentist point of view this is somewhat problematic.
   - From a Bayesian point of view this is an opportunity!
5. James-Stein is a milestone in statistical thinking.
   - Results viewed as paradoxical and counterintuitive.
   - James and Stein are geniuses.

Imperial College London

Outline

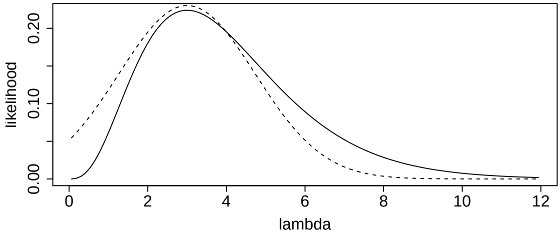
1. All Roads Lead to Rome
   - Frequentist origins of shrinkage estimates
   - Bayesian hierarchical models
2. Example 1: Using SNIa to Fit Cosmological Models
   - Joint with Roberto Trotta, Xiyun Jiao, & Hikmatali Shariff
3. Example 2: Ages of White Dwarfs in the Galactic Halo
   - Joint work with Ted von Hippel & Shijing Si

Bayesian Statistical Analyses: Likelihood

Likelihood Functions: The distribution of the data given the model parameters. E.g., $Y \sim \text{Poisson}(\lambda_S)$:

$$\text{likelihood}(\lambda_S) = e^{-\lambda_S} \lambda_S^Y / Y!$$

Maximum Likelihood Estimation: Suppose $Y = 3$



The likelihood and its normal approximation.

Can estimate λ_S and its error bars.

Imperial College London

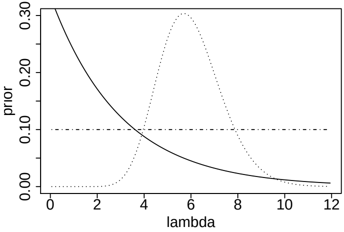
Bayesian Analyses: Prior and Posterior Dist'ns

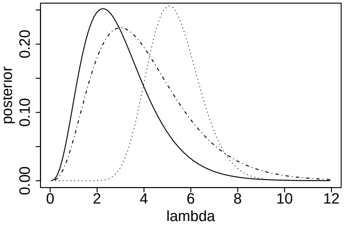
Prior Distribution: Knowledge obtained *prior* to current data.

Bayes Theorem and Posterior Distribution:

$$\text{posterior}(\lambda \mid Y) \propto \text{likelihood}(\lambda; Y) \times \text{prior}(\lambda)$$

Combine past and current information:





Bayesian analyses rely on probability theory

Bayesian Perspective

Back to shrinkage....

Frequentists tend to avoid quantities like:

1. $\mathrm{E}(\theta_j)$ and $\mathrm{Var}(\theta_j)$
2. $\mathrm{E}\left[(\theta_j - \nu)^2\right]$

From a Bayesian point of view it is quite natural to consider

1. the *prior* distribution of a parameter or
2. the *common distribution of a group of parameters*.

Models that are formulated in terms of the latter are
Hierarchical Models.

Imperial College London

A Simple Bayesian Hierarchical Model

Suppose

$$y_{ij}|\theta_j \stackrel{\text{indep}}{\sim} \mathrm{N}(\theta_j, \sigma^2) \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, G$$

with

$$\theta_j \stackrel{\text{indep}}{\sim} \mathrm{N}(\mu, \tau^2).$$

Let $\phi = (\sigma^2, \tau^2, \mu)$

$$\mathrm{E}(\theta_j \mid Y, \phi) = (1 - \omega^{\mathrm{HB}})\hat{\theta}^{\mathrm{naive}} + \omega^{\mathrm{HB}}\mu \text{ with } \omega^{\mathrm{HB}} = \frac{\sigma^2/n}{\sigma^2/n + \tau^2}.$$

The Bayesian perspective

- automatically picks the best ν,
- provides model-based estimates of ϕ,
- requires priors be specified for σ^2, τ^2, and μ.

Imperial College London

Color Correction Parameter for SNIa Lightcurves

SNIa light curves vary systematically across color bands.

- Measure how peaked the color distribution is.
- Details in the next section!!
- A hierarchical model:

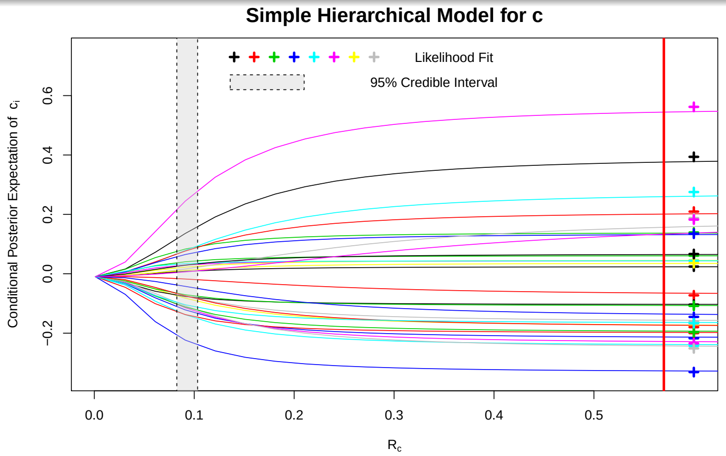
$$\hat{c}_j | c_j \overset{\text{indep}}{\sim} \text{N}(c_j, \sigma_j^2) \ \text{ for } \ j = 1, \ldots, 288$$

with

$$c_j \overset{\text{indep}}{\sim} \text{N}(c_0, R_c^2) \ \text{ and } \ p(c_0, R_c) \propto 1.$$

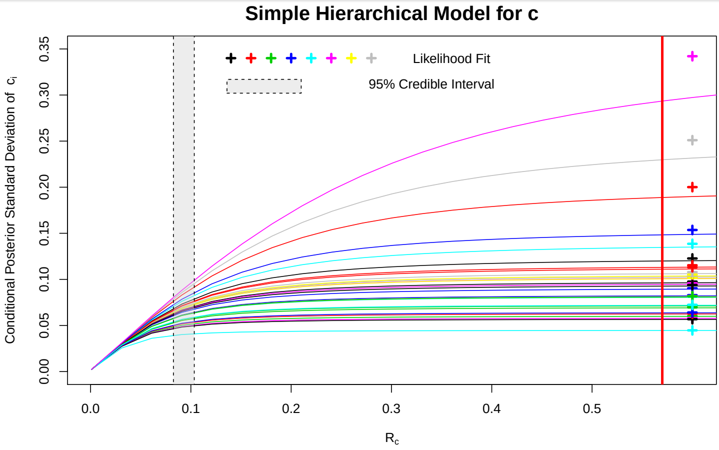
- The measurement variances, σ_j^2 are assumed known.
- We could estimate each c_j via $\hat{c}_j \pm \sigma_j$, or...

Shrinkage of the Fitted Treatment Effects



Pooling may dramatically change fits.

Standard Deviation of the Fitted Treatment Effects



Borrowing strength for more precise estimates.

Imperial College London

The Bayesian Perspective

Advantages of Bayesian Perspective:

- The advantage of James-Stein estimation is automatic.
  James-Stein had to find the estimator!
- Bayesians have a method to generate estimators.
  Even frequentists like this!
- General principle is easily tailored to any problem.
- Specification of level two model *may* not be critical.
  James-Stein derived same estimator using only moments.

Cautions:

- Results can depend on prior distributions for parameters that reside deep within the model, and far from the data.

Imperial College London

The Choice of Prior Distribution

Suppose

$$y_{ij}|\theta_j \stackrel{\text{indep}}{\sim} \mathrm{N}(\theta_j, \sigma^2) \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, G$$

with

$$\theta_j \stackrel{\text{indep}}{\sim} \mathrm{N}(\mu, \tau^2).$$

- Std non-informative prior for normal variance: $p(\sigma^2) \propto 1/\sigma^2$.
- Using this prior for the level-two variance,

$$p(\tau^2) \propto 1/\tau^2$$

leads to an improper posterior distribution:

$$p(\tau^2|y) \propto p(\tau^2) \sqrt{\frac{\mathrm{Var}(\mu|y,\tau)}{(\sigma^2+\tau^2)^G}} \exp\left\{ \sum_{j=1}^{G} - \frac{(\bar{y}_{\cdot j} - \mathrm{E}(\mu|y,\tau^2))^2}{2(\sigma^2+\tau^2)} \right\}$$

Imperial College London

Outline

1. All Roads Lead to Rome
   - Frequentist origins of shrinkage estimates
   - Bayesian hierarchical models
2. Example 1: Using SNIa to Fit Cosmological Models
   - Joint with Roberto Trotta, Xiyun Jiao, & Hikmatali Shariff
3. Example 2: Ages of White Dwarfs in the Galactic Halo
   - Joint work with Ted von Hippel & Shijing Si

Imperial College London

Outline

Imperial College London

Type Ia Supernovae as Standardizable Candles

If mass surpasses "Chandrasekhar threshold" of $1.44 M_\odot$...

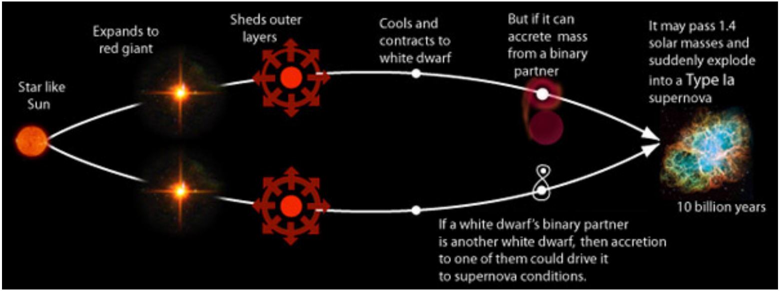


Image Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

Due to their common "flashpoint", SN1a have similar absolute magnitudes:

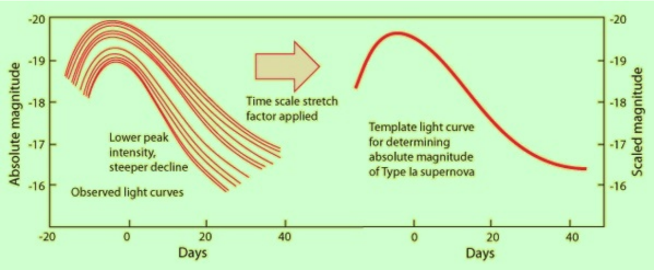
$$M_j \sim \mathrm{N}(M_0, \sigma^2_{\text{int}}).$$

Imperial College London

Predicting Absolute Magnitude

SN1a absolute magnitudes are correlated with characteristics of the explosion / light curve:

- x_j: rescale light curve to match mean template
- c_j: describes how flux depends on color (spectrum)



Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

Imperial College London

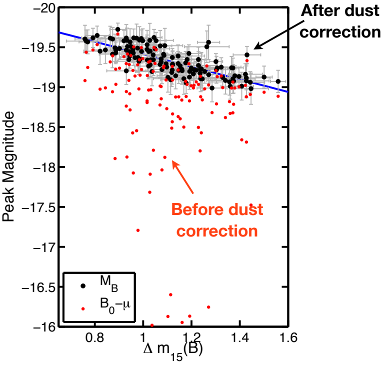
Phillips Corrections

- Recall: $M_j \sim \mathrm{N}(M_0, \sigma^2_{\rm int})$.
- Regression Model:

$$M_j = -\alpha x_j + \beta c_j + M_j^{\epsilon},$$

  with $M_j^{\epsilon} \sim \mathrm{N}(M_0, \sigma^2_{\epsilon})$.
- $\sigma^2_{\epsilon} \leqslant \sigma^2_{\rm int}$
- Including x_j and c_j reduces variance and increases precision of estimates.

Low-z calibration sample



Light curve stretch

Mandel et al (2011)

Brighter SNIa are slower decliners over time.

Imperial College London

Distance Modulus in an Expanding Universe

Apparent mag depends on absolute mag & distance modulus:

$$m_{Bj} \;=\; \mu_j + M_j \;=\; \mu_j + M_j^{\epsilon} - \alpha x_j + \beta c_j$$

Relationship between μ_i and z_i

- For nearby objects,

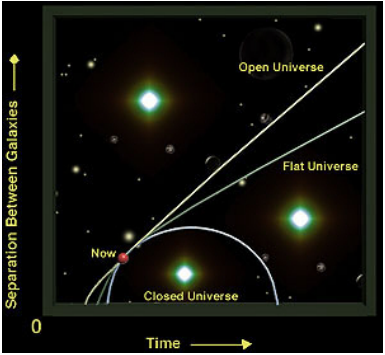
$$z_j = \text{velocity}/c$$
$$\text{velocity} = H_0 \text{ distance.}$$

(Correcting for peculiar/local velocities.)

- For distant objects, involves expansion history of Universe:

$$\mu_j = g(z_j, \Omega_\Lambda, \Omega_M, H_0)$$
$$= 5\log_{10}(\text{distance[Mpc]}) + 25$$

- We use peak B band magnitudes.

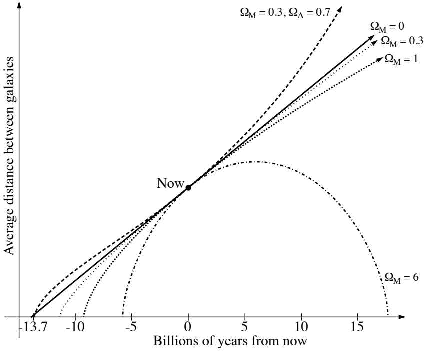


http://skyserver.sdss.org/dr1/en/astro/universe/universe.asp

Imperial College London

Accelerating Expansion of the Universe

- 2011 Physics Nobel Prize: discovery that expansion rate is increasing.
- Dark Energy is the principle theorized explanation of accelerated expansion.
- Ω_Λ: density of dark energy (describes acceleration).
- Ω_M: total matter.



A Hierarchical Model

Level 1: c_j, x_j, and m_{Bj} are observed with error.

$$\begin{pmatrix} \hat{c}_j \\ \hat{x}_j \\ \hat{m}_{Bj} \end{pmatrix} \sim \mathrm{N} \left\{ \begin{pmatrix} c_j \\ x_i \\ m_{Bj} \end{pmatrix}, \hat{C}_j \right\}$$

with $m_{Bj} = \mu_j + M_j^\epsilon - \alpha x_j + \beta c_j$ and $\mu_j = g(z_j, \Omega_\Lambda, \Omega_M, H_0)$

Level 2:

1. $c_j \sim \mathrm{N}(c_0, R_c^2)$
2. $x_j \sim \mathrm{N}(x_0, R_x^2)$
3. $M_j^\epsilon \sim \mathrm{N}(M_0, \sigma_\epsilon^2)$

Level 3: Priors on α, β, Ω_Λ, Ω_M, H_0, c_0, R_c^2, x_0, R_x^2 M_0, σ_ϵ^2

Other Model Features

Results are based on an SDSS (2009) sample of 288 SNIa.

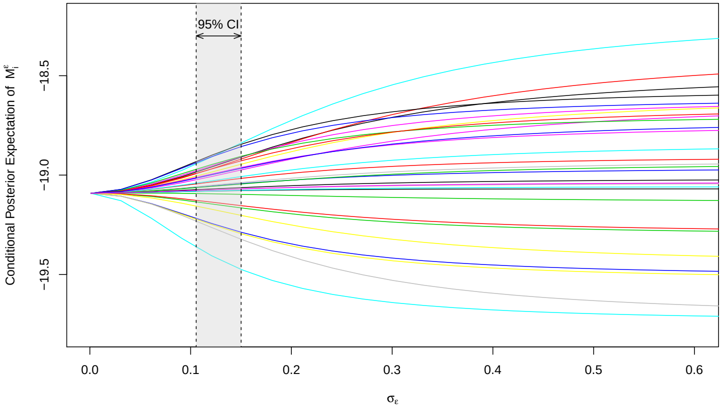
In our full analysis, we also

1. account for systematic errors that have the effect of correlating observation across supernovae,
2. allow the mean and variance of M_i^ϵ to differ for galaxies with stellar masses above or below 10^{10} solar masses,
3. include a model component that adjusts for selection effects, and
4. use a larger JLA sample[1] of 740 SNIa observed with SDSS, HST, and SNLS.

Imperial College London

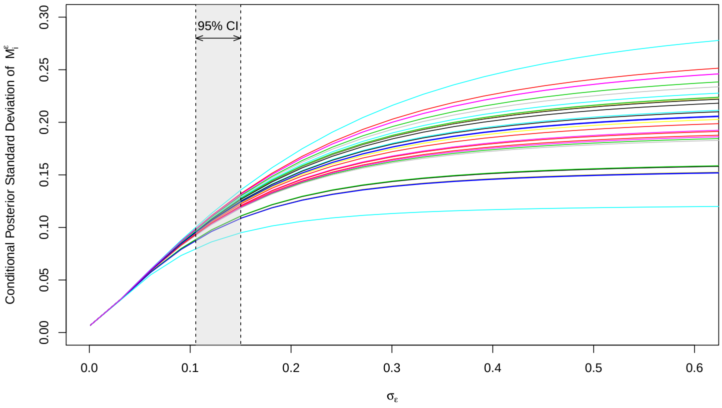
[1] Betoule, et al., 2014, arXiv:1401.4064v1

Shrinkage Estimates in Hierarchical Model



Imperial College London

Shrinkage Errors in Hierarchical Model



Imperial College London

Fitting Absolute Magnitudes Without Shrinkage

Under the model, absolute magnitudes are given by

$$M_j^{\epsilon} = m_{Bj} - \mu_j + \alpha x_j - \beta c_j \text{ with } \mu_i = g(z_j, \Omega_\Lambda, \Omega_M, H_0)$$

Setting

1. $\alpha, \beta, \Omega_\Lambda$, and Ω_M to their minimum χ^2 estimates,
2. $H_0 = 72 km/s/Mpc$, and
3. m_{Bj}, x_j, and c_j to their observed values

we have

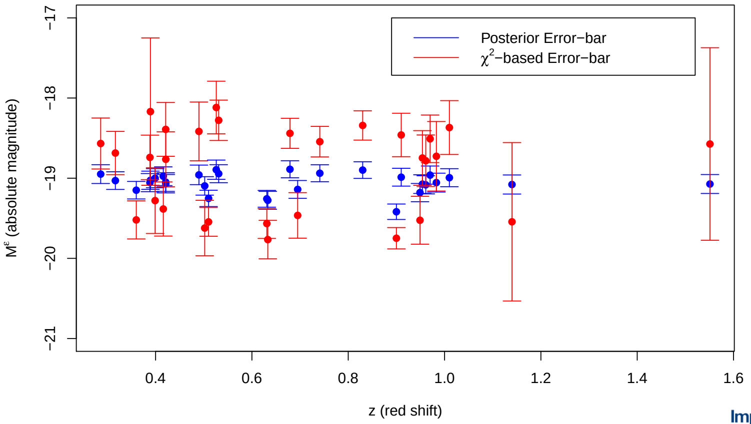
$$\hat{M}_j^{\epsilon} = \hat{m}_{Bi} - g(\hat{z}_j, \hat{\Omega}_\Lambda, \hat{\Omega}_M, \hat{H}_0) + \hat{\alpha}\hat{x}_j - \hat{\beta}\hat{c}_j$$

with error

$$\approx \sqrt{\mathrm{Var}(\hat{m}_{Bj}) + \hat{\alpha}^2 \mathrm{Var}(\hat{x}_j) + \hat{\beta}^2 \mathrm{Var}(\hat{c}_j)}$$

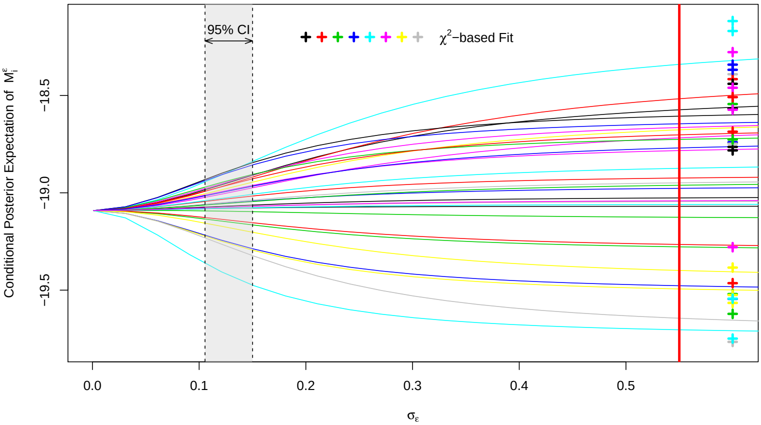
Imperial College London

Comparing the Estimates



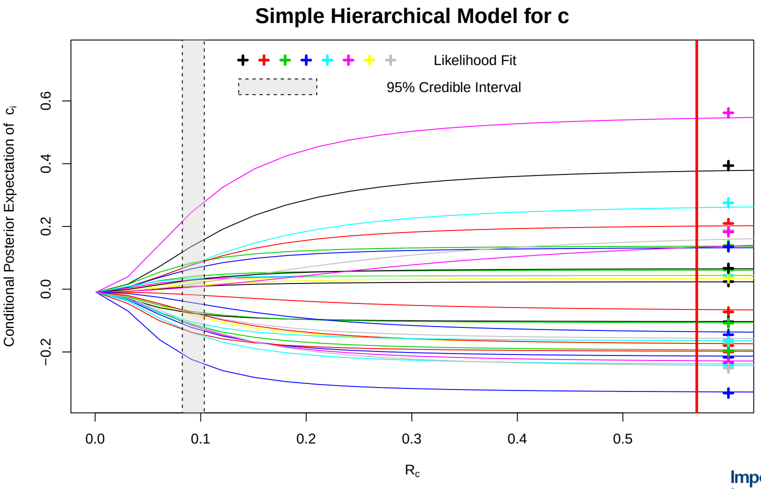
Imperial College London

Comparing the Estimates



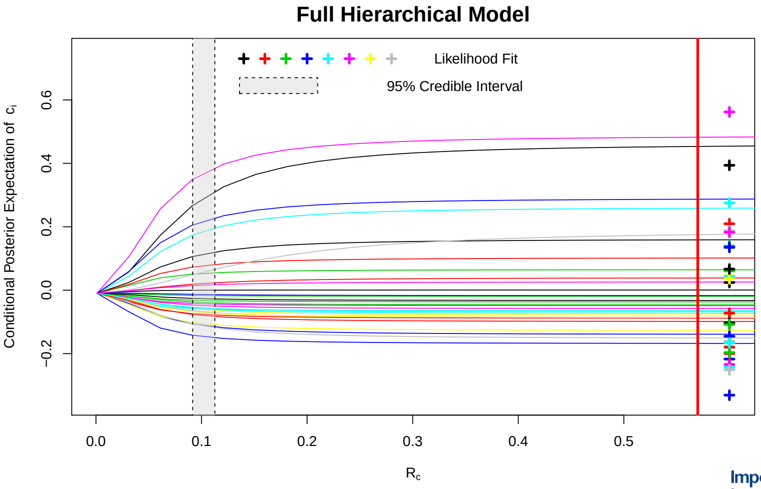
Offset estimates even without shrinkage.

Fitting a simple hierarchical model for c_i



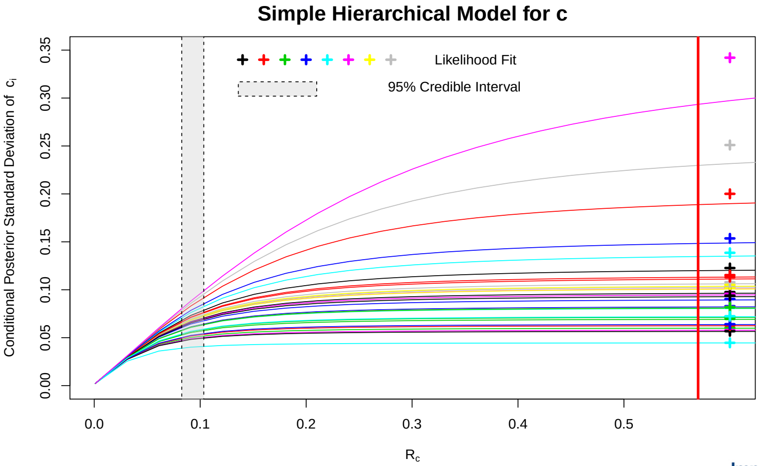
Imperial College London

Additional shrinkage due to regression



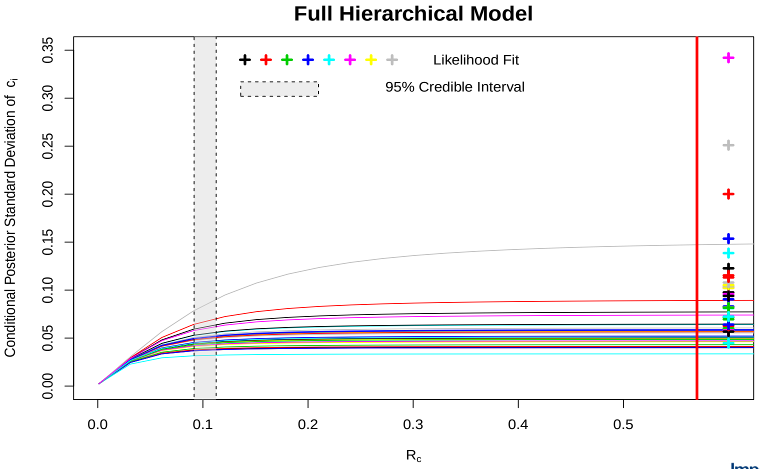
Imperial College London

Errors under simple hierarchical model for c_i

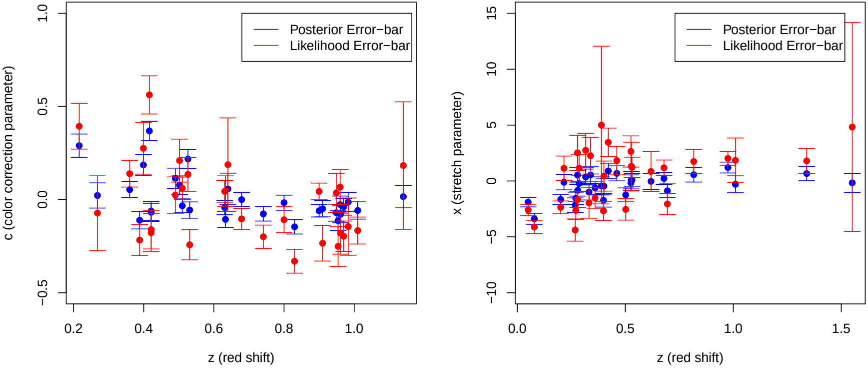


Imperial College London

Reduced errors due to regression



Comparing the Estimates of c_i and x_i



Imperial College London

Outline

1. All Roads Lead to Rome
   - Frequentist origins of shrinkage estimates
   - Bayesian hierarchical models
2. Example 1: Using SNIa to Fit Cosmological Models
   - Joint with Roberto Trotta, Xiyun Jiao, & Hikmatali Shariff
3. Example 2: Ages of White Dwarfs in the Galactic Halo
   - Joint work with Ted von Hippel & Shijing Si

Imperial College London

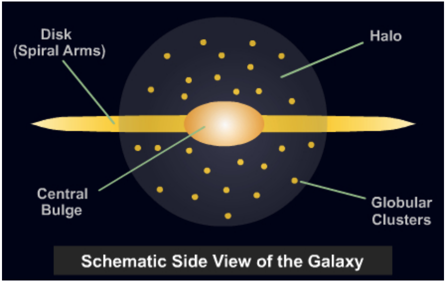
Outline

1. All Roads Lead to Rome
   - Frequentist origins of shrinkage estimates
   - Bayesian hierarchical models
2. Example 1: Using SNIa to Fit Cosmological Models
   - Joint with Roberto Trotta, Xiyun Jiao, & Hikmatali Shariff
3. Example 2: Ages of White Dwarfs in the Galactic Halo
   - Joint work with Ted von Hippel & Shijing Si

Imperial College London

Visitors from the Galactic Halo

- Age of galactic *halo* or *disk* can be estimated with their older stars.
- Halo stars pass through the galactic disk as they orbit the central bulge.



- Kilic et al. (ApJ, 2010) identified three nearby old halo white dwarfs in the SDSS; we have a sample of five.

We would like to model the white dwarf colors to estimate their age and the age of galactic halo.

Imperial College London

Fitting Dist'n of Stellar Ages in Galactic Halo

We observe seven photometric magnitudes for each WD:

$$(X_{1j}, \ldots X_{7j}) \sim \mathrm{MVN}\Big(G(\theta_j), V\Big)$$

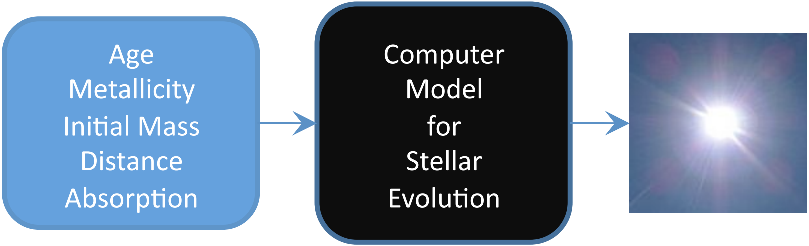
where $\theta_j = (\log_{10}(\text{age}_j),\ \text{distance}_j,\ \text{mass}_j)$ and

$$\log_{10}(\text{age}_j) \sim \mathrm{N}(\mu, \tau^2).$$

- If the WD are a representative sample, μ and τ^2 are the population mean and variance for galactic halo.
- Even if sample is not representative, hierarchical model produces estimators with better statistical properties.

Imperial College London

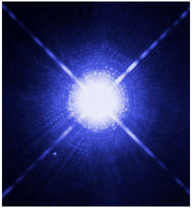
Computer Model for Main Sequence (& RG) Evolution



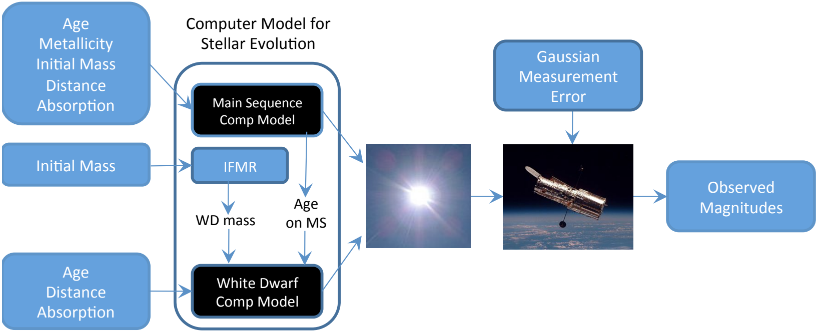
- Computer model predicts how the emergent and apparent spectra evolve as a function of input parameters.
- We observe photometric magnitudes, the apparent luminosity in each of several wide wavelength bands.

Imperial College London

White Dwarfs Physics



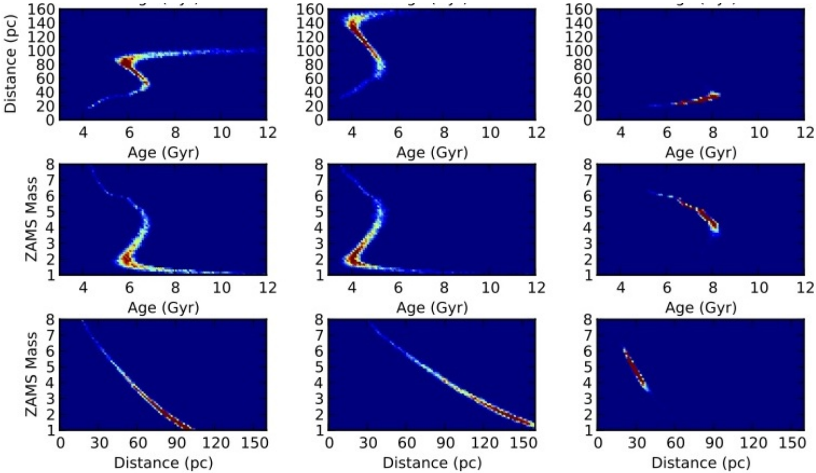
- White dwarf spectra are not predicted from MS/RG models
- Different physical processes require different models:
  1. Computer Model for White Dwarf Cooling
  2. Computer Model for White Dwarf Atmosphere
  3. Initial Final Mass Relationship (IFMR)



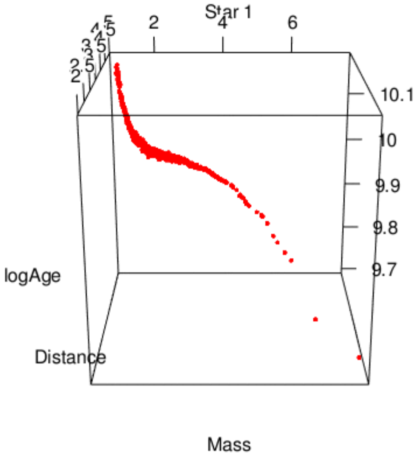
A parametric model for the IFMR forms a bridge between the computer models.

Imperial College London

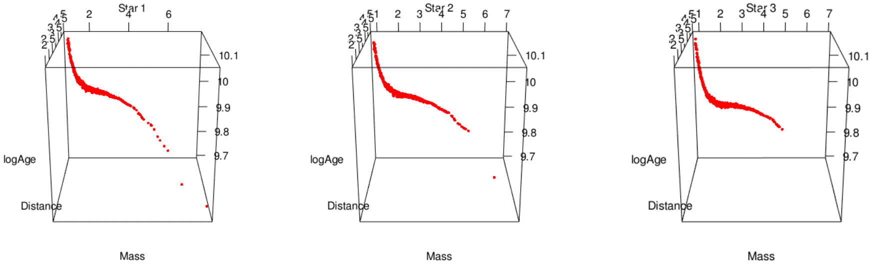
Complex Posterior Distributions



Complex Posterior Distributions



Fitting Each WD Individually (Kilic's sample)



Posterior distributions exhibit similar structure and similar fitted parameter values.

Imperial College London

Fitting the Population Distribution of Halo WDs

Model:

$$(X_{1j}, \dots X_{7j}) \sim \text{MVN}\Big(G(\theta_j), V\Big)$$

where $\theta_j = (\log_{10}(\text{age}_j),\ \text{distance}_j,\ \text{mass}_j)$ and

$$\log_{10}(\text{age}_j) \sim \text{N}(\mu, \tau^2).$$

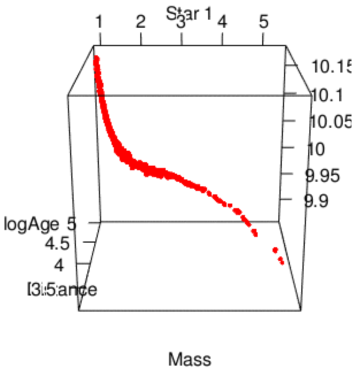
Maximum a posterior estimates:

- $\hat{\mu} = 10.065$ (11.6 gigayears)
- $\widehat{\log_{10}\tau} = \log_{10}(0.053)$
- 95% range: $(9.1, 14.8)$ gigayears

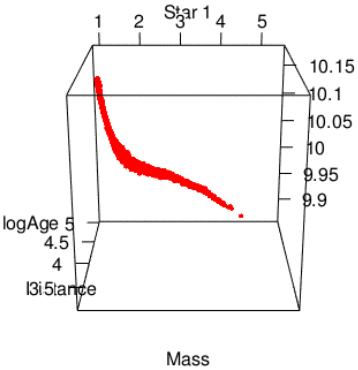
Imperial College London

Suppose $\mathrm{sd}(\log_{10}(\mathrm{age})) = 0.009$

Individual Fit



Hierarchical Fit

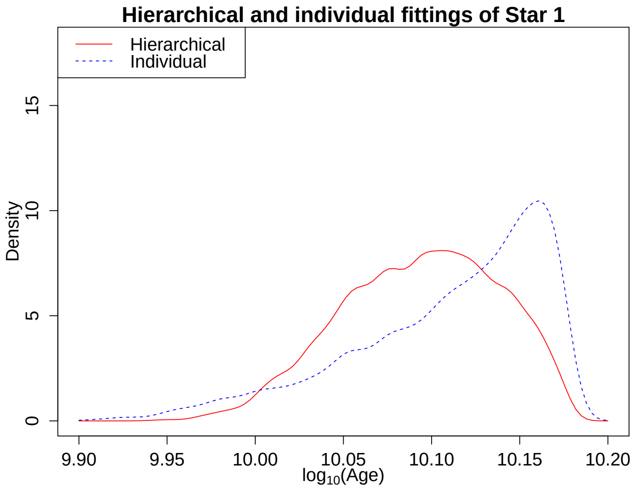


Effect of Shrinkage for one halo WD.

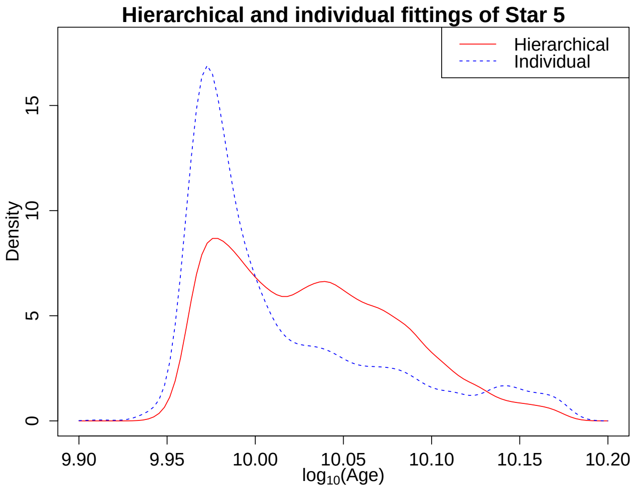
Here we exaggerate the shrinkage by using $\tau = 0.009 < \hat{\tau}$.

Imperial College London

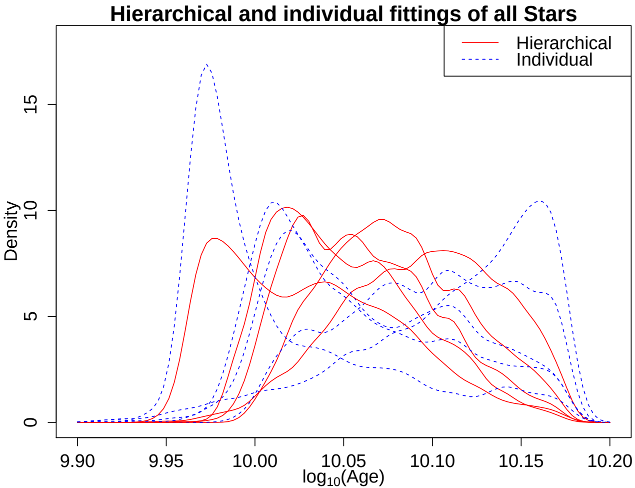
Shrinkage in the Posterior of Age (with fitted τ)



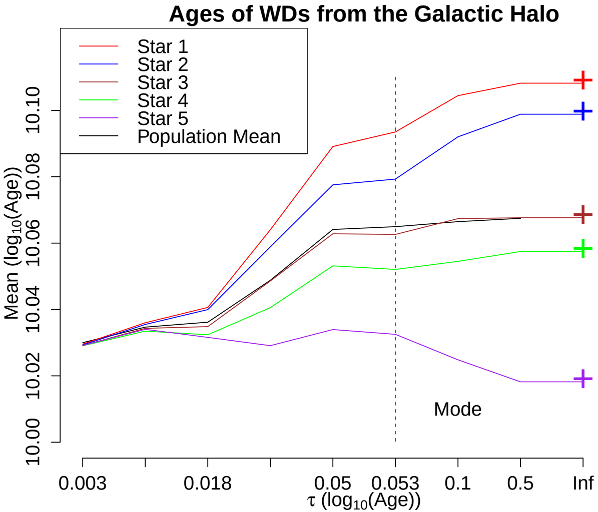
Shrinkage in the Posterior of Age (with fitted τ)



Shrinkage in the Posterior of Age (with fitted τ)

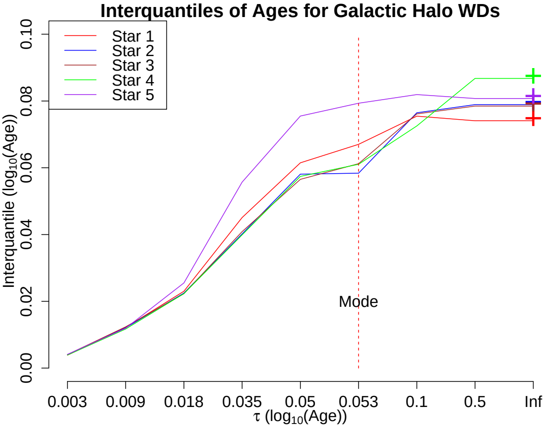


Sensitivity of Results to $\mathrm{Var}(\log_{10}(\mathrm{age}))$



Imperial College London

Sensitivity of Results to $\mathrm{Var}(\log_{10}(\mathrm{age}))$



Gaia will provide photometric magnitudes for hundreds of galactic halo WDs.

Imperial College London

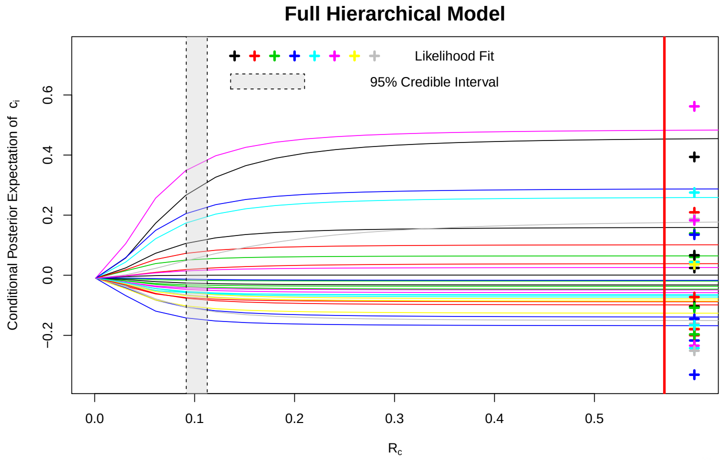
Discussion

- Estimation of groups of parameters describing populations of sources is not uncommon in astronomy.
- These parameters may or may not be of primary interest.
- Modeling the distribution of object-specific parameters can dramatically reduce both error bars and MSE ...
- ... especially with noisy observations of similar objects.
- Shrinkage estimators are able to "borrow strength".
- May be little cost of freeing object-specific parameters (e.g., metallicity or distance of stars in a cluster).

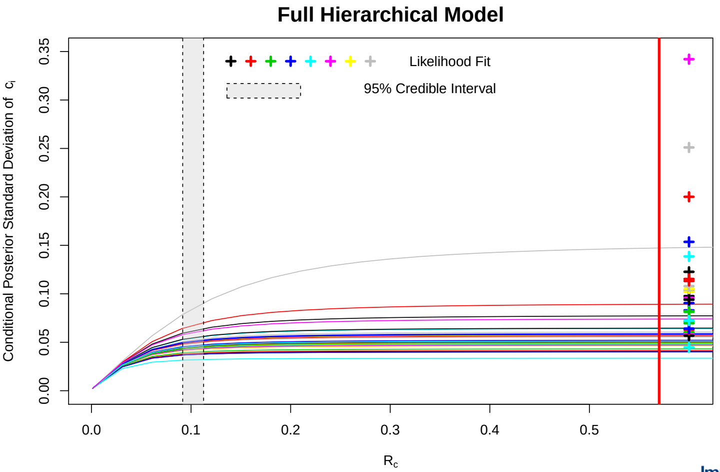
Don't throw away half of your toolkit!!
(Bayesian and Frequency methods)

Imperial College London

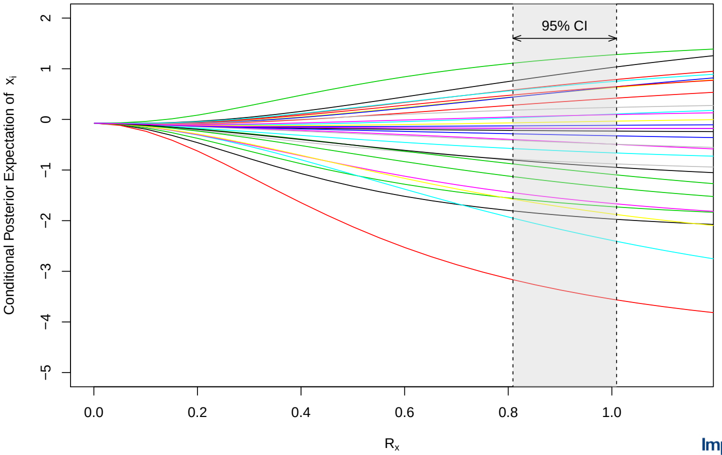
Shrinkage Estimates of c_i in Hierarchical Model



Shrinkage Errors of c_i in Hierarchical Model

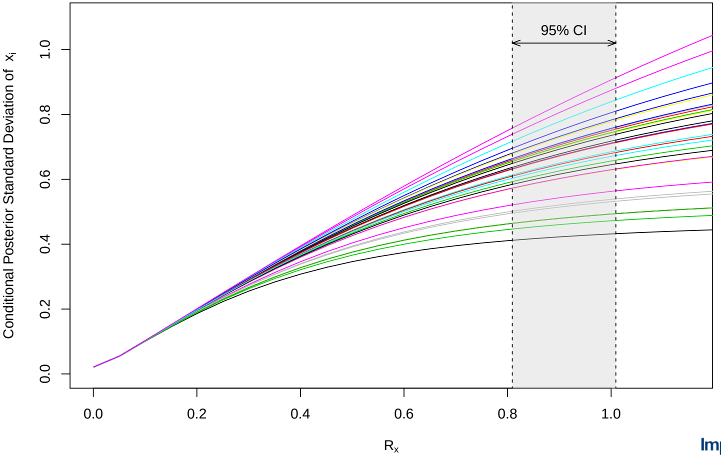


Shrinkage Estimatesof x_i in Hierarchical Model



Imperial College London

Shrinkage Errors of x_i in Hierarchical Model



Imperial College London

A Non-Astronomical Example

The Educational Testing Service studied the effects of coaching programs on SAT-V scores in eight US high schools:[2]

$$y_j|\theta_j \overset{\text{indep}}{\sim} \mathrm{N}(\theta_j, \sigma_j^2) \text{ for } j = 1, \ldots, 8$$

with

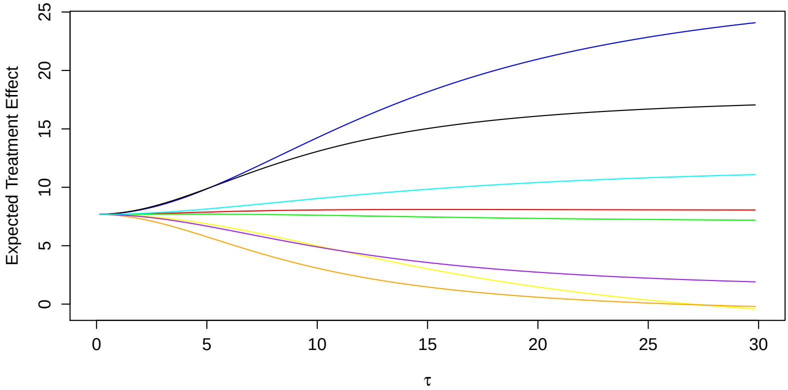
$$\theta_j \overset{\text{indep}}{\sim} \mathrm{N}(\mu, \tau^2) \text{ and } p(\mu, \tau) \propto 1.$$

The y_j are estimated treatment effects

- based on preliminary analyses
- adjust for PSAT (V and M) scores
- standard errors on estimated treatment effects are regarded as known

[2] From Gelman et al. (2013), Bayesian Data Analysis, 3rd Edition, §5.5.

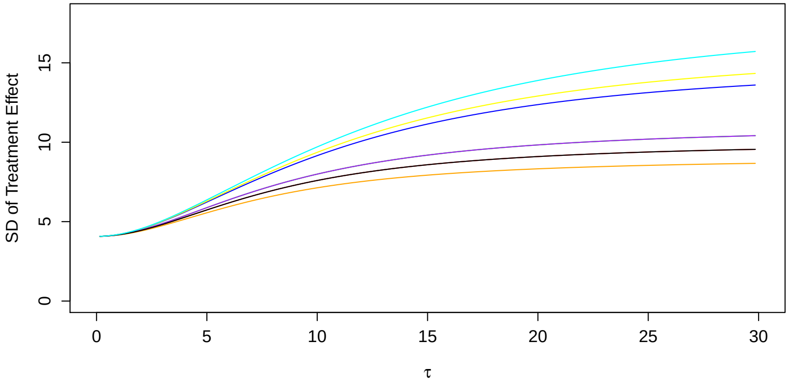
Shrinkage of the Fitted Treatment Effects



Pooling may dramatically change fitted effects.

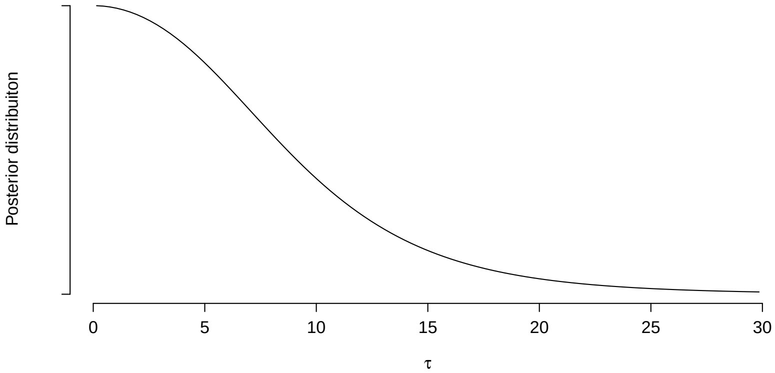
Imperial College London

Standard Deviation of the Fitted Treatment Effects



Pooling results in more precise estimates.

Fitting the Standard Deviation of the Treatment Effects



Fitted τ determines the degree of pooling.