Bayesian Statistical Methods for Astronomy

Part III: Model Building

David A. van Dyk

Statistics Section, Imperial College London

INAF - Osservatorio Astrofisico di Arcetri, September 2014

Outline

Recall Simple Multilevel Model

Example: Background contamination in a single bin detector

- Contaminated source counts: $y = y_S + y_B$
- Background counts: x
- Background exposure is 24 times source exposure.

A Poisson Multi-Level Model:

LEVEL 1: $y | y_B, \lambda_S \overset{\text{dist}}{\sim} \text{Poisson}(\lambda_S) + y_B$,

LEVEL 2: $y_B | \lambda_B \overset{\text{dist}}{\sim} \text{Pois}(\lambda_B)$ and $x | \lambda_B \overset{\text{dist}}{\sim} \text{Pois}(\lambda_B \cdot 24)$,

LEVEL 3: specify a prior distribution for λ_B, λ_S.

Each level of the model specifies a dist'n given unobserved quantities whose dist'ns are given in lower levels.

Multi-Level Models

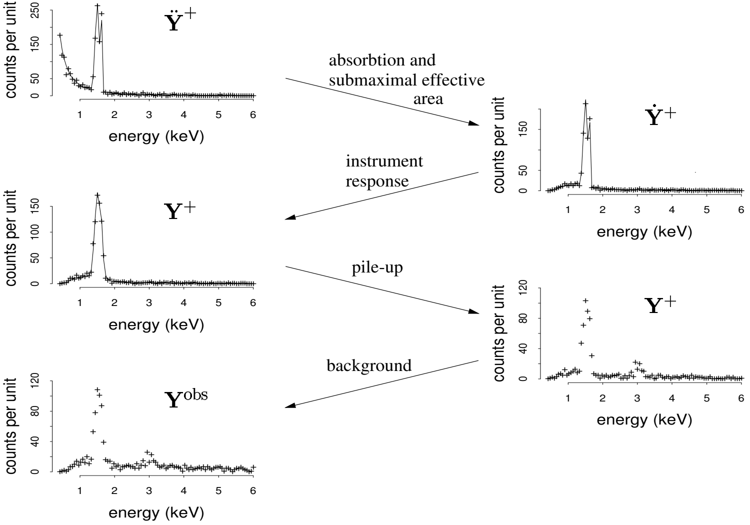
Definition

A ***multi-level model*** *is specified using a series of conditional distributions. The joint distribution can be recovered via the factorization theorem, e.g.,*

$$p_{XYZ}(x, y, z|\theta) = p_{X|YZ}(x|y, z, \theta_1)\ p_{Y|Z}(y|z, \theta_2)\ p_Z(z|\theta_3).$$

- This model specifics the joint distribution of X, Y, and Z, given the parameter $\theta = (\theta_1, \theta_2, \theta_3)$.
- The variables X, Y, and Z may consist of observed data, latent variables, missing data, etc.
- In this way we can combine models to derive an endless variety of *multi-level models*.

Example: High-Energy Spectral Modeling



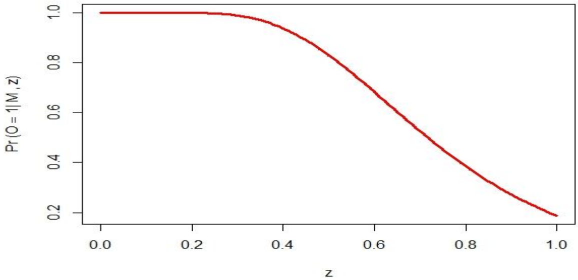
Outline

A Multilevel Model for Selection Effects

We wish to estimate a dist'n of absolute magnitudes, M_i,

- Suppose $M_i \sim \textsc{Norm}(\mu, \sigma^2)$, for $i = 1, \ldots, n$;
- But M_i is only observed if $M_i < F(z_i)$[1];
- Observe $N(< n)$ objects including z_i, $\theta = (\mu, \sigma^2)$ estimated.

(For $\mu = -19.3$ and $\sigma = 1$.)



[1] M_i observed if $< F(z_i) = 24 - \mu(z_i)$; $\mu(z_i)$ from Λ-CDM model ($\Omega_m = 0.3$, $\Omega_\kappa = 0$, $H_0 = 67.3$km/s/Mpc).

Model 1: Ignore Selection Effect

Likelihood: $M_i|\theta, z_i \sim \textsc{Norm}(\mu, \sigma^2)$, for $i = 1, \ldots, N$;

Prior: $\mu \sim \textsc{Norm}(\mu_0, \tau^2)$, and $\sigma^2 \sim \beta^2/\chi^2_\nu$;

Posterior: $\mu \mid (M_1, \ldots M_n, \sigma^2) \sim \textsc{Norm}(\cdot, \cdot)$ and

$\sigma^2 \mid (M_1, \ldots M_n, \mu) \sim \cdot/\chi^2$ (Details on next slide.)

Definition

If (some set of) conditional distributions of the prior and the posterior distributions are of the same family, the prior dist'n is called that likelihood's semi-congutate prior distribution.

Semi-conjugate priors are very amenable to the Gibbs sampler.

Gibbs Sampler for Model 1

Step 1: Update μ from its conditional posterior dist'n given σ^2:

$$\mu^{(t+1)} \sim \textsc{Norm}\left(\bar{\mu},\ s_\mu^2\right)$$

with

$$\bar{\mu} = \left(\frac{\sum_{i=1}^N M_i}{(\sigma^2)^{(t)}} + \frac{\mu_0}{\tau^2}\right) \Big/ \left(\frac{N}{(\sigma^2)^{(t)}} + \frac{1}{\tau^2}\right); \quad s_\mu^2 = \left(\frac{N}{(\sigma^2)^{(t)}} + \frac{1}{\tau^2}\right)^{-1}.$$

Step 2: Update σ^2 from its conditional posterior dist'n given μ:

$$(\sigma^2)^{(t+1)} \sim \left[\sum_{i=1}^N \left(M_i - \mu^{(t+1)}\right)^2 + \beta^2\right] \Big/ \chi^2_{N+\nu}.$$

In this case, resulting sample is nearly independent.

A Closer Look at Conditional Posterior: Step 1

Given σ^2:

Likelihood: $M_i|\theta, z_i \sim \textsc{Norm}(\mu, \sigma^2)$, for $i = 1, \ldots, N$;

Prior: $\mu \sim \textsc{Norm}(\mu_0, \tau^2)$

Posterior: $\mu \mid (M_1, \ldots M_n, \sigma^2) \sim \textsc{Norm}(\bar{\mu}, s_\mu^2)$ with

$$\bar{\mu} = \left(\frac{\sum_{i=1}^N M_i}{\sigma^2} + \frac{\mu_0}{\tau^2} \right) \Big/ \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2} \right); \quad s_\mu^2 = \left(\frac{N}{\sigma^2} + \frac{1}{\tau^2} \right)^{-1}.$$

- Posterior mean is a weighted average of sample mean $(\frac{1}{N}\sum_{i=1}^N M_i)$ and prior mean (μ_0), with weights $\frac{N}{\sigma^2}$ and $\frac{1}{\tau^2}$.
- Compare s_μ^2 with $\mathrm{Var}\left(\frac{1}{N}\sum_{i=1}^N M_i\right) = \frac{\sigma^2}{N}$.
- Reference prior sets $\mu_0 = 0$ and $\tau^2 = \infty$. (Improper and flat on μ.)

A Closer Look at Conditional Posterior: Step 2

Given μ:

Likelihood: $M_i|\theta, z_i \sim \textsc{Norm}(\mu, \sigma^2)$, for $i = 1, \ldots, N$;

Prior: $\sigma^2 \sim \beta^2/\chi^2_\nu$;

Posterior:

$$(\sigma^2)^{(t+1)} \mid (M_1, \ldots M_n, \mu) \sim \left[\sum_{i=1}^{N} \left(M_i - \mu^{(t+1)}\right)^2 + \beta^2\right] \Big/ \chi^2_{N+\nu}.$$

- The prior has the affect of adding ν additional data points with variance β^2.
- Reference prior sets $\nu = \beta^2 = 0$. (Improper and flat on $\log(\sigma^2)$.)

Model 2: Account for Selection Effect

Likelihood: The distribution of the observed magnitudes:

$$p(M_i|O_i = 1, \theta, z_i) = \frac{\Pr(O_i = 1|M_i, z_i, \theta) p(M_i|\theta, z_i)}{\int \Pr(O_i = 1|M_i, z_i, \theta) p(M_i|\theta, z_i) \mathrm{d}M_i};$$

Here

- $M_i|\theta, z_i \sim \textsc{Norm}(\mu, \sigma^2)$ and
- $\Pr(O_i = 1|M_i, z_i, \theta)) = \Pr(M_i < F(z_i)|\theta)$

So $M_i|(O_i = 1, \theta, z_i) \sim \textsc{TrunNorm}[\mu, \sigma^2; F(z_i)]$.

Prior: $\mu \sim \textsc{Norm}(\mu_0, \tau^2)$, $\sigma^2 \sim \beta^2/\chi^2_\nu$;

Posterior: Prior is not conjugate, posterior is not standard.

MH within Gibbs for Model 2

Neither step of the Gibbs Sampler is a standard dist'n:

Step 1: Update μ from its conditional dist'n given σ^2

Use Random-Walk Metropolis with a $\textsc{Norm}(\mu^{(t)}, s_1^2)$ proposal distribution.

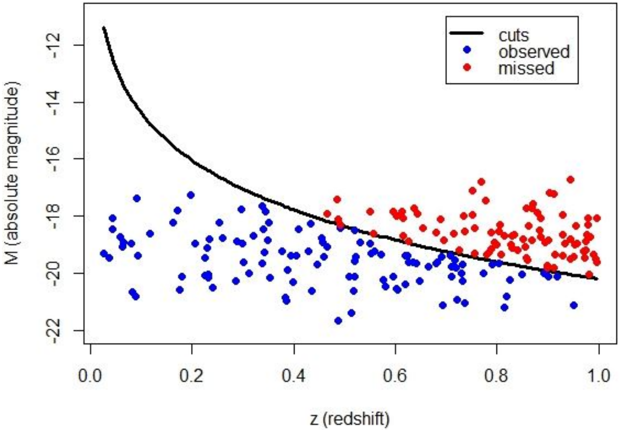
Step 2: Update σ^2 from its conditional dist'n given μ

Use Random-Walk Metropolis Hastings with a $\textsc{LogNorm}\left[\log\left(\sigma^{2\ (t)}\right), s_2^2\right]$ proposal distribution.

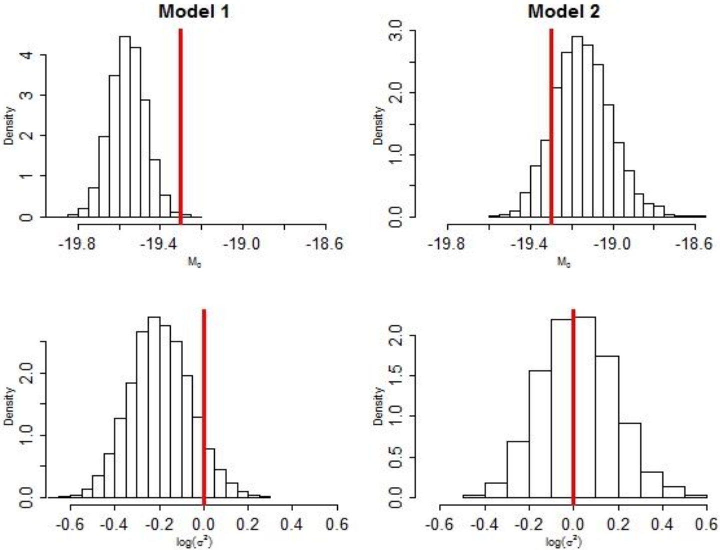
Adjust s_1^2 and s_2^2 to obtain an acceptance rate of around 40%.

Simulation Study I

- Sample $M_i \sim \text{NORM}(\mu = -19.3, \sigma = 1)$ for $i = 1, \ldots, 200$.
- Sample z_i from $p(z) \propto (1 + z)^2$, yielding $N = 112$.

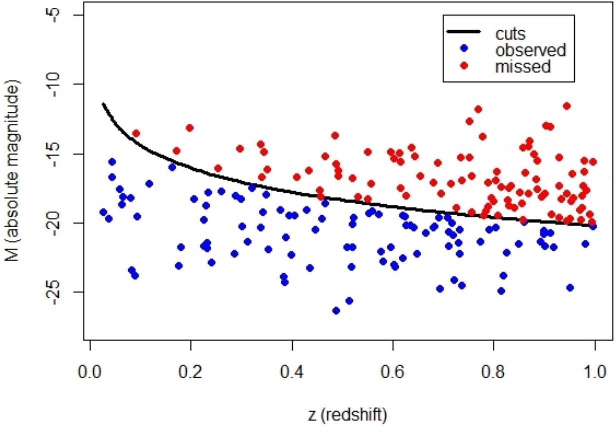


Simulation I ($\mu_0 = -19.3$, $\sigma_m = 20$, $\nu = 0.02$, $\beta^2 = 0.02$)

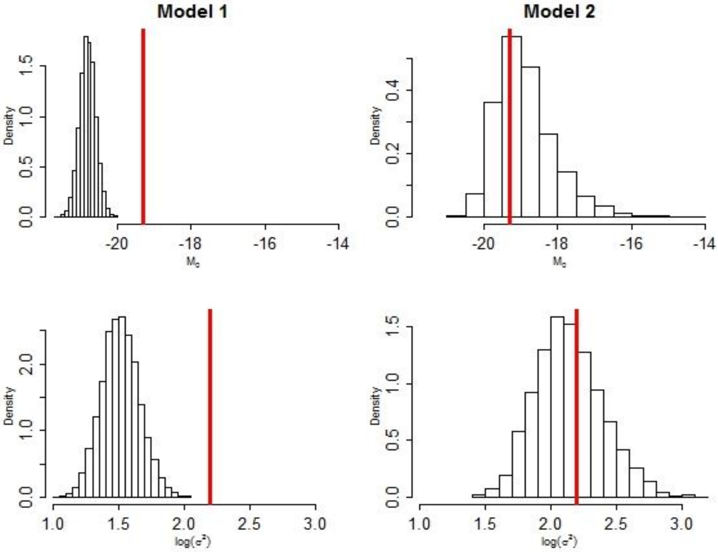


Simulation Study II

- Sample $M_i \sim \textsc{Norm}(\mu = -19.3, \sigma = 3)$ for $i = 1, \ldots, 200$.
- Sample z_i from $p(z) \propto (1 + z)^2$, yielding $N = 101$.



Simulation II ($\mu_0 = -19.3$, $\sigma_m = 20$, $\nu = 0.02$, $\beta^2 = 0.02$)



Outline

Frequentists Origins of Hierarchical Models

Suppose we wish to estimate a parameter, θ, from repeated measurements:

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

E.g.: calibrating a 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?

Frequency Evaluation of an Estimator

- How far off is the estimator?

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

- How far off do we expect it to be?

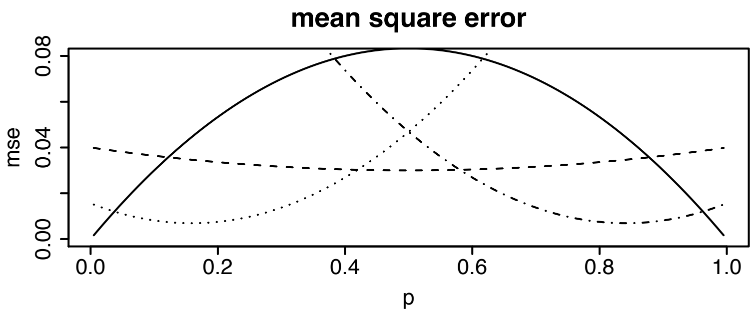
$$\mathrm{MSE}(\hat{\theta}|\theta) = \mathrm{E}\left[(\hat{\theta} - \theta)^2 \mid \theta\right] = \int \left(\hat{\theta}(y) - \theta\right)^2 f_Y(y|\theta)dy$$

- 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:

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

MSE of Four Estimators of Binomial Probability

Recall Simulation Study:



- The MSE (of all four estimators) depends on true p.
- In this case: no evidence of inadmissiblity.

Inadmissibility of the Sample Mean

Suppose we wish to estimate more than one parameter:

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

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}:

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

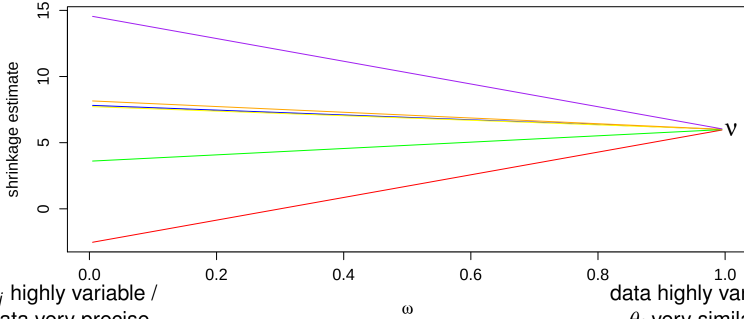
with $\omega^{\text{JS}} \approx \dfrac{\sigma^2/n}{\sigma^2/n + \tau_\nu^2}$ and $\tau_\nu^2 = \text{E}[(\theta_i - \nu)^2]$.

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

data highly variable / θ_j very similar (to ν)

To Where Should We 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 $\tau_\nu^2 = \mathrm{E}[(\theta_i - \nu)^2]$.

- Minimize τ^2.

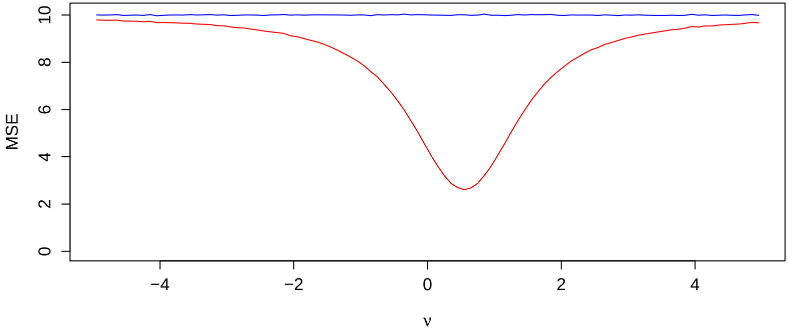
The optimal choice of ν is the average of the θ_j.

Illustration

Suppose:

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

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

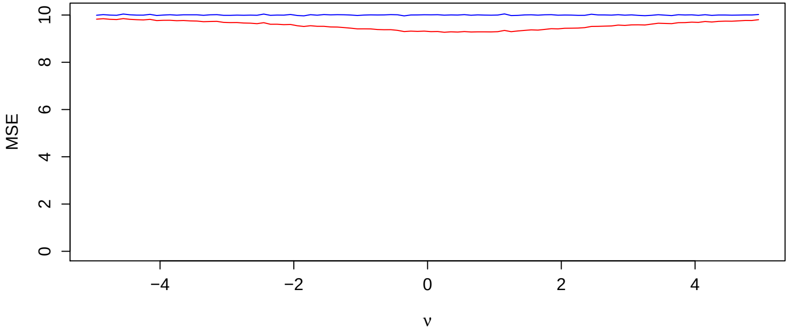


Illustration

Suppose:

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

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



Intuition

1. If 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.

Bayesian Perspective

Frequentist 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 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.

A Simple Bayesian Hierarchical Model

Suppose

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

with

$$\theta_j \overset{\text{indep}}{\sim} \text{NORM}(\mu, \tau^2).$$

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

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

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.

Color Correction Parameter for SNIa Lightcurves

SNIa light curves vary systematically across color bands.

- Color Correction: Measure the peakedness of color dist'n.
- Details in the next section!!
- A hierarchical model:

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

with

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

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

Fitting the Hierarchical Model with Gibbs Sampler

$$\boxed{\begin{array}{l} \hat{c}_j | c_j \stackrel{\text{indep}}{\sim} \text{NORM}(c_j, \sigma_j^2) \ \text{ for } \ j = 1, \ldots, G \\ c_j \stackrel{\text{indep}}{\sim} \text{NORM}(c_0, R_c^2) \ \text{ and } \ p(c_0, R_c) \propto 1. \end{array}}$$

To Derive the Gibbs Sampler Note:

1. Given (c_0, R_C^2), a standard Gaussian model for each j:

$$\hat{c}_j | c_j \stackrel{\text{indep}}{\sim} \text{NORM}(c_j, \sigma_j^2) \ \text{ with } \ c_j \stackrel{\text{indep}}{\sim} \text{NORM}(c_0, R_c^2).$$

2. Given $c_1, \ldots, c_G$, another standard Gaussian model:

$$c_j \stackrel{\text{indep}}{\sim} \text{NORM}(c_0, R_c^2) \ \text{ with } \ p(c_0, R_c) \propto 1.$$

Fitting the Hierarchical Model with Gibbs Sampler

The Gibbs Sampler:

Step 1: Sample $c_1, \ldots c_G$ from their joint posterior given (c_0, R_C^2):

$$c_j^{(t)} \,\Big|\, (\hat{c}_j, c_0^{(t-1)}, (R_C^2)^{(t-1)}) \sim \textsc{Norm}\left(\mu_j,\ s_j^2\right)$$

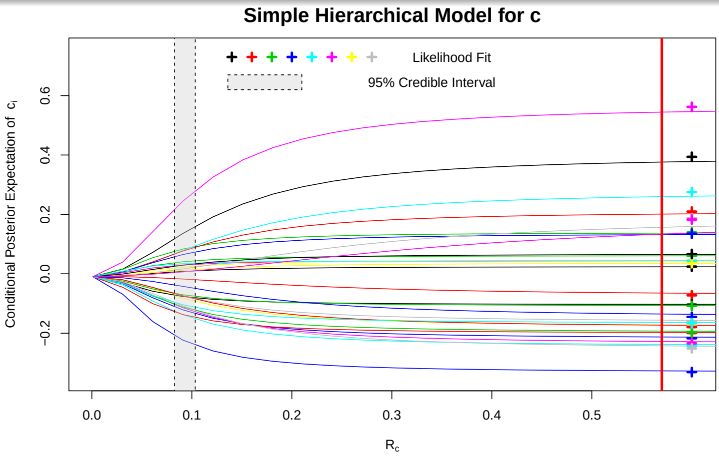
$$\mu_j = \left(\frac{\hat{c}_j}{\sigma_j^2} + \frac{c_0^{(t-1)}}{(R_C^2)^{(t-1)}}\right) \Big/ \left(\frac{1}{\sigma_j^2} + \frac{1}{(R_C^2)^{(t-1)}}\right); \quad s_j^2 = \left(\frac{1}{\sigma^2} + \frac{1}{(R_C^2)^{(t-1)}}\right)^{-1}.$$

Step 2: Sample (c_0, R_C^2) from their joint posterior given $c_1, \ldots c_G$:

$$(R_C^2)^{(t)} \big| (c_1^{(t)}, \ldots, c_G^{(t)}) \sim \frac{\sum_{j=1}^G (c_j^{(t)} - \bar{c})^2}{\chi^2_{G-2}} \quad \text{with} \quad \bar{c} = \frac{1}{G}\sum_{j=1}^G c_j^{(t)}$$

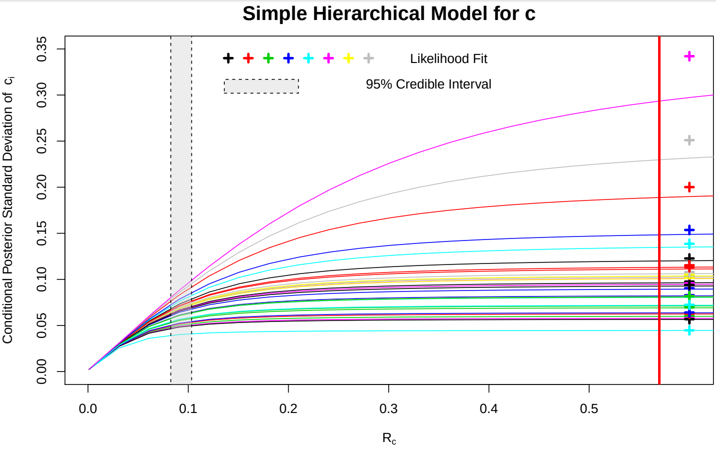
$$c_0^{(t)} \big| (c_1^{(t)}, \ldots, c_G^{(t)}), (R_C^2)^{(t)} \sim \textsc{Norm}\left(\bar{c}, (R_C^2)^{(t)}/G\right)$$

Shrinkage of the Fitted Color Correction



Pooling may dramatically change fits.

Standard Deviation of the Fitted Color Correction



Borrowing strength for more precise estimates.

The Bayesian Perspective

Advantages of Bayesian Perspective:

- The advantage of James-Stein estimation is automatic.
  James and Stein had to find their 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.

The Choice of Prior Distribution

Suppose

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

with

$$\theta_j \overset{\text{indep}}{\sim} \textsc{Norm}(\mu, \tau^2).$$

- Reference prior for normal variance: $p(\sigma^2) \propto 1/\sigma^2$, flat on $\log(\sigma^2)$
- Using this prior for the level-two variance,

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

leads to an improper posterior distribution:

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

Outline

Type Ia Supernovae as Standardizable Candles

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

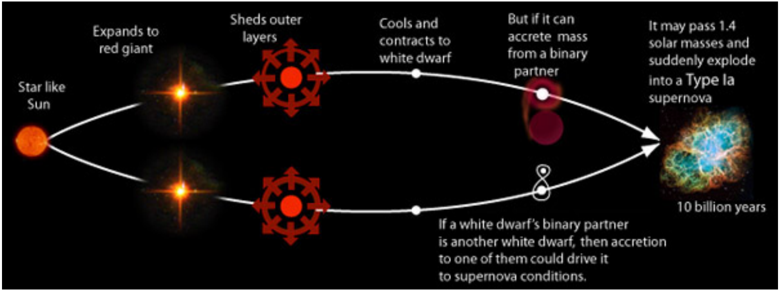


Image Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

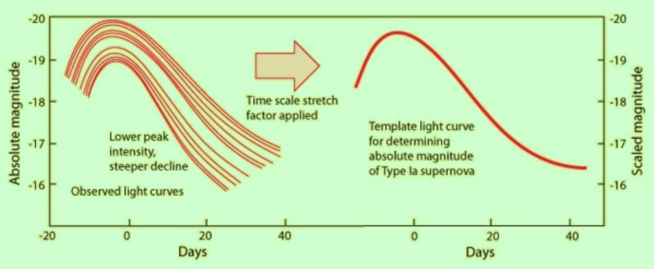
Due to their common "flashpoint", SN1a have similar absolute magnitudes:

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

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)



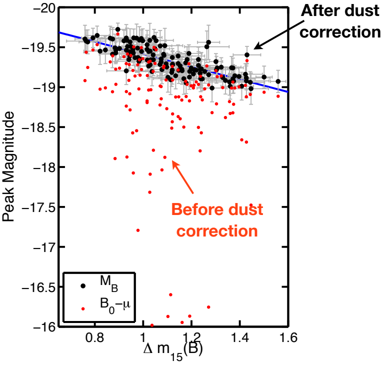
Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

Phillips Corrections

- Recall:
  $M_j \sim \textsc{Norm}(M_0, \sigma^2_{\text{int}})$.
- Regression Model:
  $$M_j = -\alpha x_j + \beta c_j + M_j^\epsilon,$$
  with $M_j^\epsilon \sim \textsc{Norm}(M_0, \sigma^2_\epsilon)$.
- $\sigma^2_\epsilon \leqslant \sigma^2_{\text{int}}$
- Including x_i and c_i reduces variance and increases precision of estimates.

Mandel et al (2011)

Low-z calibration sample



Light curve stretch

Brighter SNIa are slower decliners over time.

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,

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

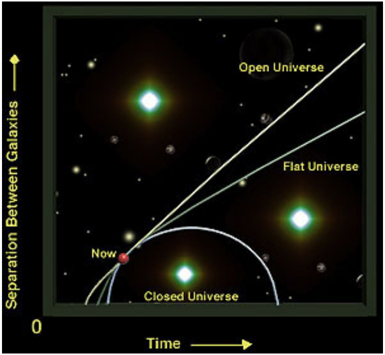
(Correcting for peculiar/local velocities.)

- For distant objects, involves expansion history of Universe:

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

.

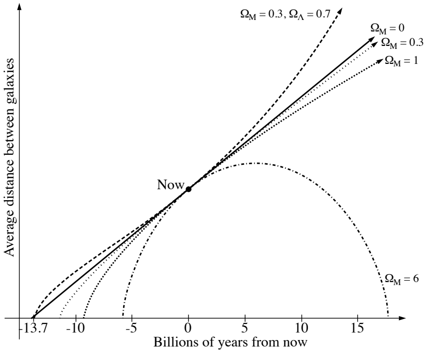
- We use peak B band magnitudes.



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

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.

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

Level 2:

1. $c_j \sim \textsc{Norm}(c_0, R_c^2)$
2. $x_j \sim \textsc{Norm}(x_0, R_x^2)$
3. The conditional dist'n of m_{Bj} given c_j and x_j is specified via

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

with $\mu_j = g(z_j, \Omega_\Lambda, \Omega_M, H_0)$ and $M_j^\epsilon \sim \textsc{Norm}(M_0, \sigma_\epsilon^2)$.

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

Regression With Measurement Errors

The above model encompasses measurement error model:

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

$$\begin{pmatrix} \hat{c}_j \\ \hat{x}_j \\ \hat{m}_{Bj} \end{pmatrix} \sim \textsc{Norm}\left\{ \begin{pmatrix} c_j \\ x_j \\ m_{Bj} \end{pmatrix}, \; \hat{C}_j \right\}.$$

Level 2: *[Omitting hierarchical and cosmological components]*
The conditional dist'n of m_{Bj} given c_j and x_j is specified via

$$m_{Bj} = M_0 - \alpha x_j + \beta c_j + M_j^{\epsilon} \;\; \text{with} \;\; M_j^{\epsilon} \sim \textsc{Norm}(0, \sigma_{\epsilon}^2).$$

Level 3: Priors on M_0, α, β, σ_{ϵ}^2, and (hierarchical? on) c_j and x_j.

We can simply model the complexity and fit the resulting model using MCMC.

Other Model Features

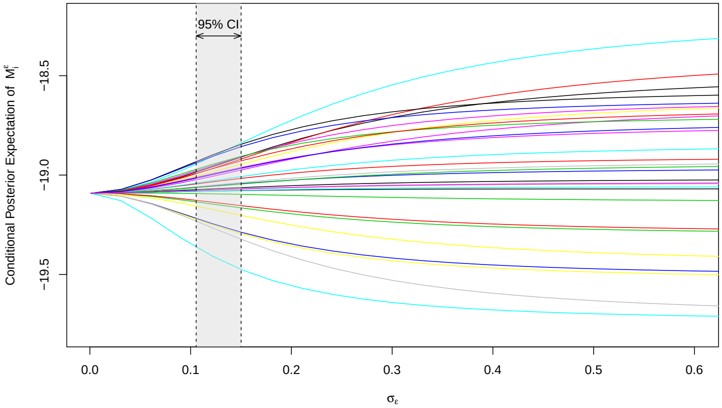
Results are based on an SDSS (2009) sample of 288 SNIa.

In our full analysis, we also

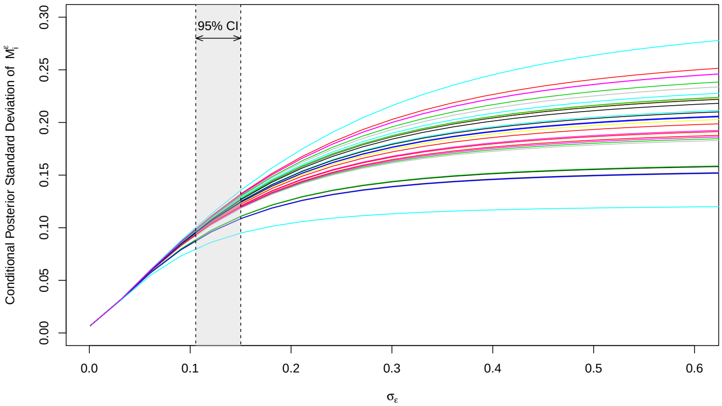
1. account for systematic errors that have the effect of correlating observation across supernovae,
2. allow the mean and variance of M_i^ϵ to differ for galaxies with stellar masses above or below 10^{10} solar masses,
3. include a model component that adjusts for selection effects, and
4. use a larger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

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

Shrinkage Estimates in Hierarchical Model



Shrinkage Errors in Hierarchical Model



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

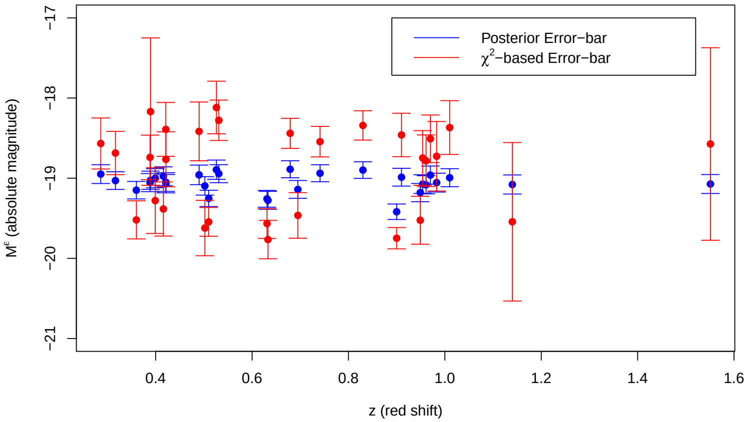
we have

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

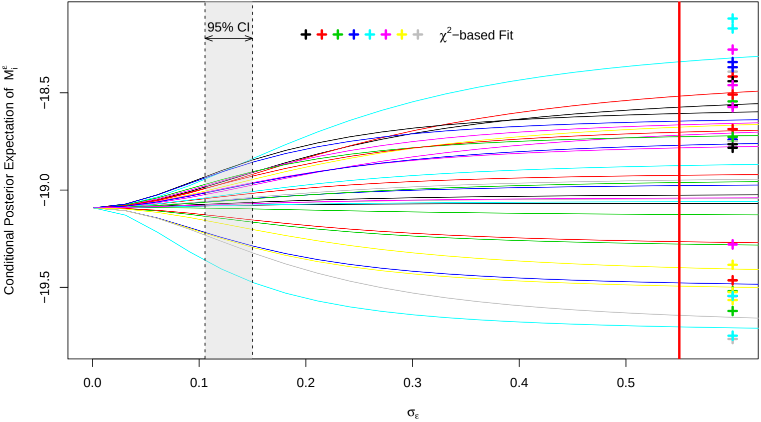
with error

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

Comparing the Estimates

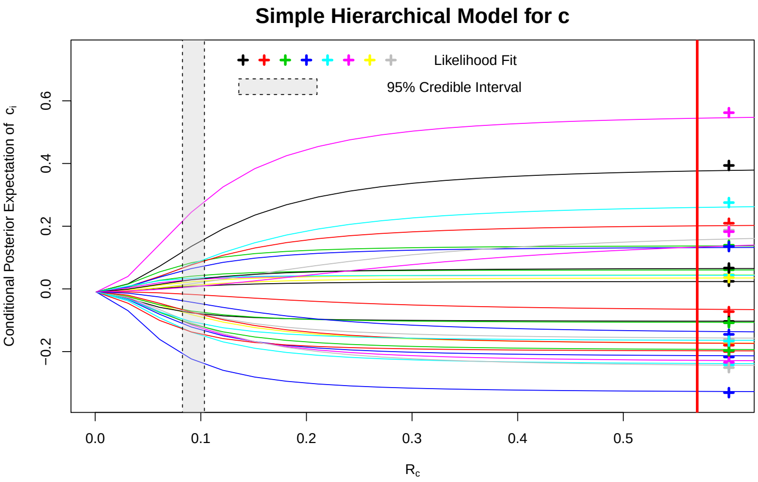


Comparing the Estimates

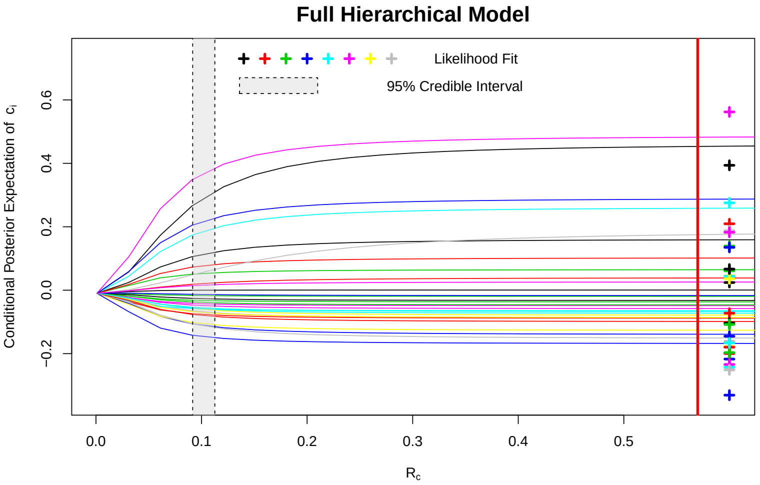


Offset estimates even without shrinkage.

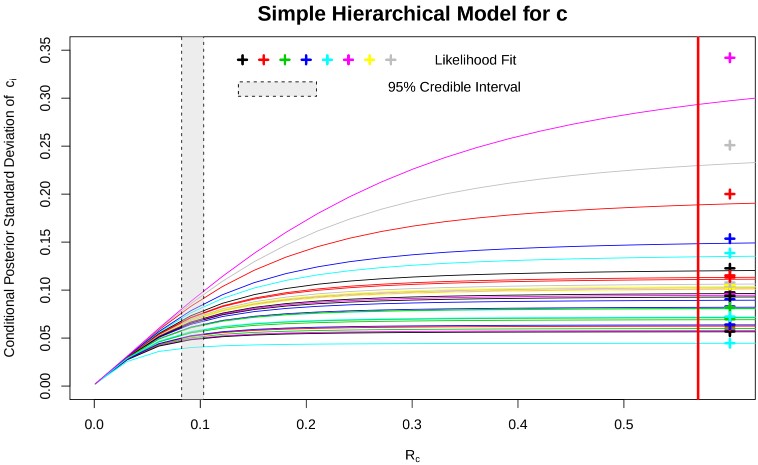
Fitting a simple hierarchical model for c_i



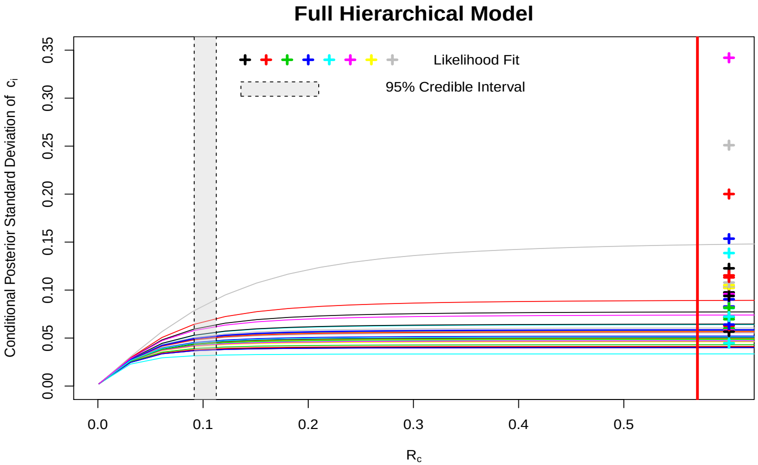
Additional shrinkage due to regression



Errors under simple hierarchical model for c_i



Reduced errors due to regression



Model Checking

We model:

$$m_{Bi} = g(z_i, \Omega_\Lambda, \Omega_M, H_0) - \alpha x_i + \beta c_i + M_i^\epsilon$$

How good of a fit is the cosmological model, $g(z_i, \Omega_\Lambda, \Omega_M, H_0)$?

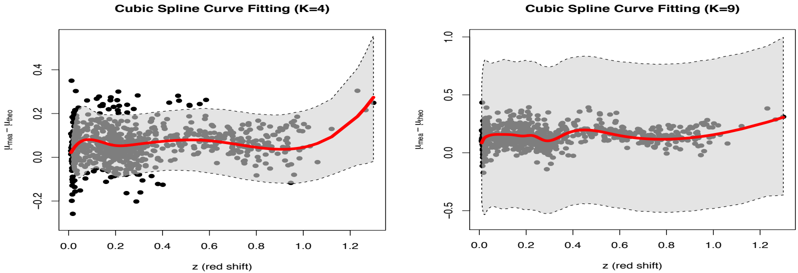
We can check the model by adding a cubic spline term:

$$m_{Bi} = g(z_i, \Omega_\Lambda, \Omega_M, H_0) + h(z_i) + M_i^\epsilon - \alpha x_i + \beta c_i + M_i^\epsilon$$

where, $h(z_i)$ is cubic spline term with K knots.

Model Checking

Fitted cubic spline, $h(z)$, and its errors:



Can use similar methods to compare with competing cosmological models.

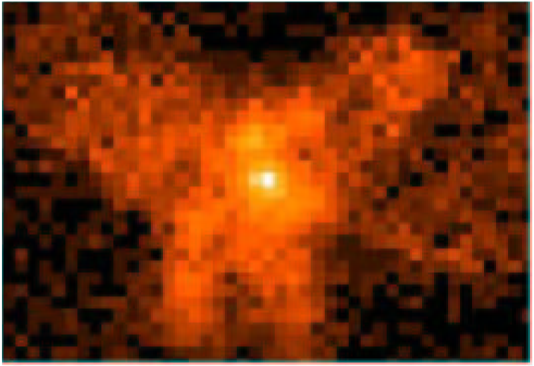
Discussion

- Estimation of groups of parameters describing populations of sources 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”.

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

Outline

X-ray Image Analysis

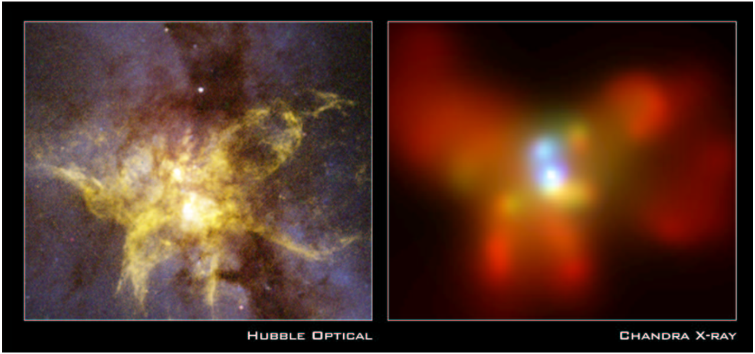


NGC 6240

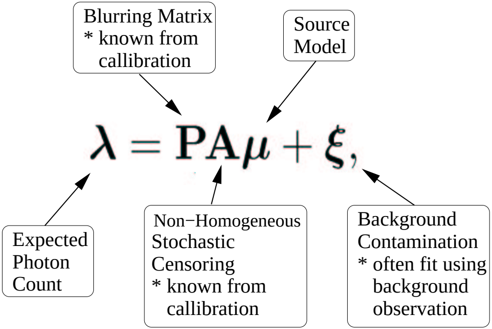
- Photon counts in each of a large number of image pixels.
- We use Poisson models for the photon counts.
- Blurring, detector sensitivity, background contamination.

X-ray Image Analysis

Optical and (smoothed) X-ray Images of NGC 6240:

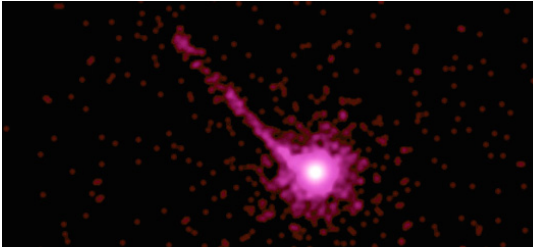


Bayesian Deconvolution



- Pixel counts: $Y_i \stackrel{\text{indep}}{\sim} \textsc{Poisson}(\lambda_i)$, for $i = 1, \ldots, n$.
- P is the point spread function.
- A describes detector sensitivity.
- ξ is an $n \times 1$ vector of expected background counts.
- μ is the image of the astronomical source.

A Model for the Source Image



A useful model for the source image, μ must allow for

1. Known or presumed structures such as point sources for concentrated X-ray emitters.
2. Irregular and unpredictable structure in extended emission.

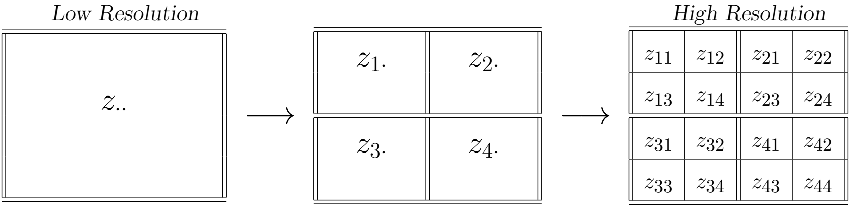
We may want to conduct a statistical tests for evidence of an extended source.

A Smoothing Prior for the Extended Source

Imagine counts not subject to blurring, detector sensitivity, or background:

$$Z_i \overset{\text{indep}}{\sim} \textsc{Poisson}(\mu_i), \quad \textit{for} \ \ i = 1, \ldots, n.$$

The Nowak-Kolaczyk Multiscale Model:



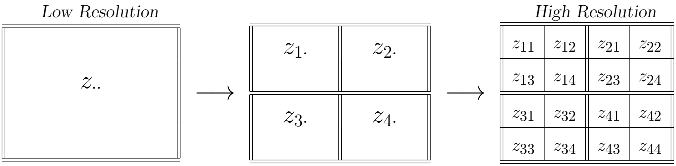
$z_{\cdot\cdot} \sim \text{Poisson}(\mu)$
$\mu \sim \text{Gamma}\{(\alpha_0, \beta_1)\}$

$\boldsymbol{z}_{,\cdot} | z_{\cdot\cdot} \sim \text{Multinomial}(\boldsymbol{p}_1)$
$\boldsymbol{p}_1 \sim \text{Dirich.}\{(\alpha_1, \alpha_1, \alpha_1, \alpha_1)\}$

$\boldsymbol{z}_{i,} | z_{i\cdot} \sim \text{Multinomial}(\boldsymbol{p}_{2i})$
$\boldsymbol{p}_{2i} \sim \text{Dirich.}\{(\alpha_2, \alpha_2, \alpha_2, \alpha_2)\}$

The Dirichlet Prior Distribution

The Nowak-Kolaczyk Multiscale Model:



$z_{\cdot\cdot} \sim \text{Poisson}(\mu)$
$\mu \sim \text{Gamma}\{(\alpha_0, \beta_1)\}$

$\boldsymbol{z}_{,\cdot}|z_{\cdot\cdot} \sim \text{Multinomial}(\boldsymbol{p}_1)$
$\boldsymbol{p}_1 \sim \text{Dirich.}\{(\alpha_1, \alpha_1, \alpha_1, \alpha_1)\}$

$\boldsymbol{z}_{i,}|z_{i\cdot} \sim \text{Multinomial}(\boldsymbol{p}_{2i})$
$\boldsymbol{p}_{2i} \sim \text{Dirich.}\{(\alpha_2, \alpha_2, \alpha_2, \alpha_2)\}$

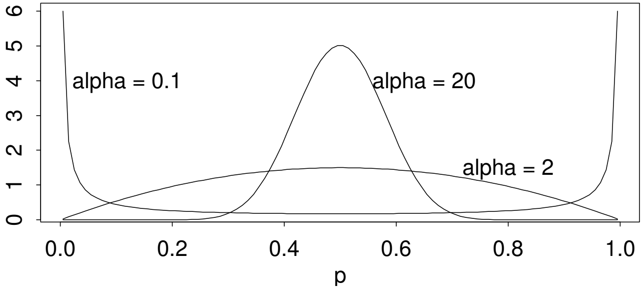
- The Dirichlet is a generalization of the beta distribution.
- It is the conjugate prior for a multinomial probability vector.
- The Dirchlet priors on $\boldsymbol{p}$ in the Nowak-Kolaczyk model have expected value $(0.25, 0.25, 0.25, 0.25)$.
- This choice of prior favors a smooth reconstructed image.

Interpreting the Smoothing Parameters

The Multiscale prior is specified in terms of the Dirichlet smoothing hyperparameters: $(\alpha_1, \alpha_2, \ldots, \alpha_K)$.

- Different values at each level of resolution.
- Larger α_k encourage more smoothing ("prior counts").
- We put a hierarchical prior on these smoothing parameters.

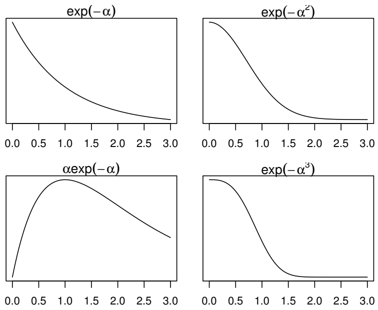
Using binary splits and the beta distribution for illustration:



Fitting the Smoothing Parameters

We use a common prior on the smoothing parameters.

- Too much mass near zero leads to numerical instability. (Priors that put all mass in one quadrant.)
- Too much mass far from zero results in too much smoothing.
- A compromise: $\alpha_k \sim \exp(-\delta\alpha^3/3)$
- Exact shape of the prior matters less than its general features.



Summary of the Hierarchical / Multilevel Model

Level 1: Blurring, varying Sensitivity, and Background:

$$\boldsymbol{\lambda} = \mathbf{PA}\boldsymbol{\mu} + \boldsymbol{\xi}$$

Level 2: The image, $\boldsymbol{\mu}$, combines known features and a multiscale model for unknown features.

Level 3: The flexible multiscale model parameterized via a nested set of 2 by 2 tables.

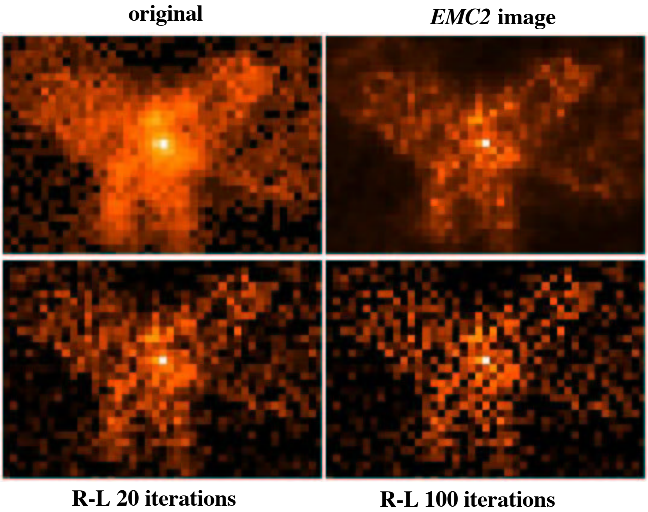
Level 4: The smoothing prior shrinks the probailites in the tables toward $(0.25, 0.25, 0.25.0.25)$.

$$\boldsymbol{p}_k \sim \text{Dirich.}\{(\alpha_k, \alpha_k, \alpha_k, \alpha_k)\}$$

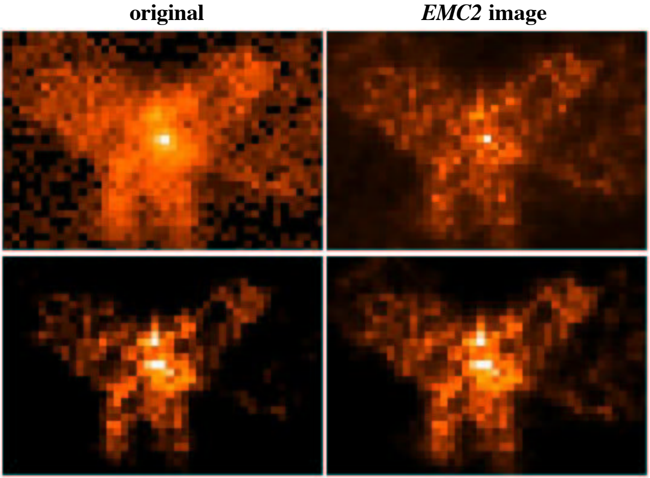
The degree of smoothing is governed by the α_k.

Level 5: Fit the smoothing parameters hierarchically, tuning their prior for good performance: $\alpha_k \sim \exp(-\delta\alpha^3/3)$.

Results



Results

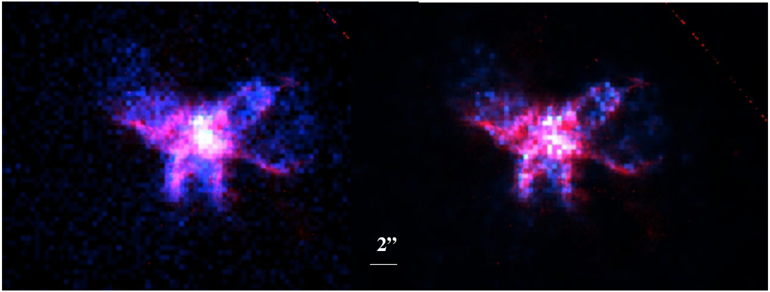


EMC2 significance map: ***3 sigma*** *EMC2* significance map: ***1 sigma***

Results

[3]

Chandra (blue) and HST H-alpha (red)



original ***EMC2***

[3] Esch, D. N., Connors, A., Karovska, M., and van Dyk, D. A. (2004). An Image Reconstruction Technique with Error Estimates. *The Astrophysical Journal*, **610**, 1213–1227.

Thanks...

Stellar Evolution:

- Nathan Stein
- David Stenning
- Shijing Si
- Elizabeth Jeffery
- William H. Jefferys
- Ted von Hippel

SNIa Cosmology:

- Xiyun Jiao
- Hikmatali Shariff
- Roberto Trotta

X-ray Image Analysis:

- David Esch (original work)
- Nathan Stein (recent work)
- Alanna Connors
- Vinay Kashyap
- Aneta Siegminowska

And

The CHASC International AstroStatistics Collaboration