Bayesian Statistical Methods for Astronomy

Part III: Model Building

David A. van Dyk

Statistics Section, Imperial College London

Center for Astrophysics | Harvard & Smithsonian
February 2025

Outline

1. Model Building
   - Multi-Level Models
   - Example: Selection Effects
   - Hierarchical Models and Shrinkage
2. Extended Modeling Examples
   - Hierarchical Model: Supernovae & Cosmology
   - Non-Representiative Data and StratLearn
   - Discussion

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

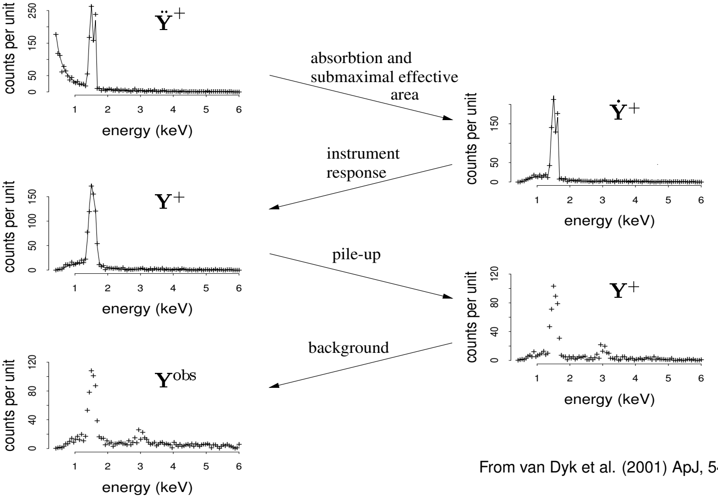
Definition

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



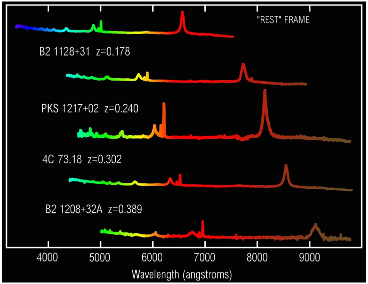
From van Dyk et al. (2001) ApJ, 548, 224-243

Outline

1. Model Building
   - Multi-Level Models
   - Example: Selection Effects
   - Hierarchical Models and Shrinkage
2. Extended Modeling Examples
   - Hierarchical Model: Supernovae & Cosmology
   - Non-Representiative Data and StratLearn
   - Discussion

The Expanding Universe

Redshift



http://www.noao.edu/image_gallery/html/im0566.html

For "nearby" objects,

$$z = \text{velocity}/c$$

$$\text{velocity} = H_0 \text{ distance.}$$

Hubble's Famous Diagram

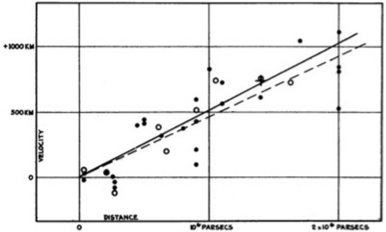


FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Radial velocities, corrected for solar motion, are plotted against distances estimated from involved stars and mean luminosities of nebulae in a cluster. The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.

Hubble (1929)

The Big Bang!

Distance Modulus in an Expanding Universe

Apparent magnitude - Absolute magnitude = Distance modulus:

$$m - M = \mu \quad \left[= 5 \log_{10}(\text{distance [Mpc]}) + 25 \right]$$

Relationship between μ and z

- For nearby objects,

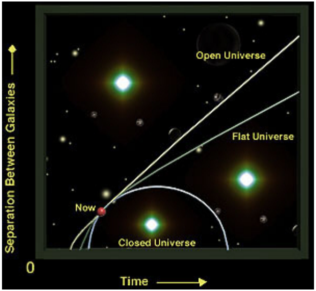
  $$\text{distance} = \mu \propto z.$$

  (Correcting for peculiar/local velocities.)

- For distant objects, involves expansion history of Universe:

  $$\mu = g(z, \Omega_\Lambda, \Omega_M, H_0)$$

  [function of density of dark energy and of total matter]



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

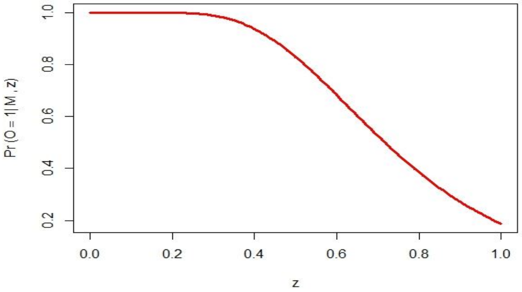
We observe M only if $m <$ (say) 24.
I.e., we observe M only if $M = m - \mu(z) < 24 - \mu(z) \equiv F(z)$.

A Multilevel Model for Selection Effects

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

- Suppose $M_i \sim \text{NORM}(\mu, \sigma^2)$, for $i = 1, \ldots, n$;
- But M_i is only observed if $M_i < F(z_i)$[1];   *[z is redshift, see next slide]*
- 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 ($\Omega_m = 0.3$, $\Omega_\kappa = 0$, $H_0 = 67.3$).

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

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

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

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

In this case, resulting sample is nearly independent.

A Closer Look at Conditional Posterior: Step 1

Given σ^2:

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

Prior: $\mu \sim \text{NORM}(\mu_0, \tau^2)$

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

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

- Posterior mean is a weighted average of sample mean ($\frac{1}{N}\sum_{i=1}^N M_i$) and prior mean ($\mu_0$), with weights $\frac{N}{\sigma^2}$ and $\frac{1}{\tau^2}$.
- Compare s_μ^2 with $\text{Var}\left(\frac{1}{N}\sum_{i=1}^N M_i\right) = \frac{\sigma^2}{N}$.
- 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
- $\Pr(O_i = 1|M_i, z_i, \theta)) = \text{Indicator}\{M_i < F(z_i)\}$

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

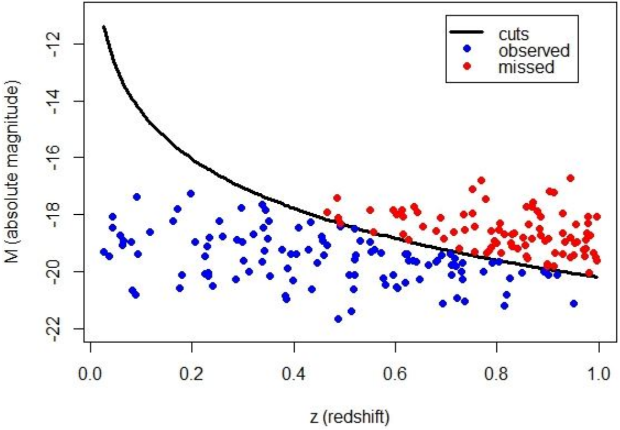
Use Random-Walk Metropolis Hastings with a $\textsc{LogNorm}\left[\log\left(\sigma^{2\ (t)}\right), s_2^2\right]$ proposal distribution.[2]

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

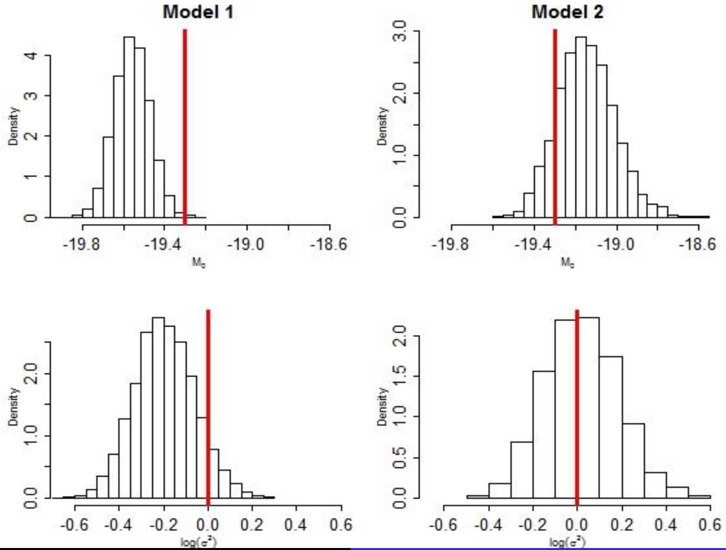
[2] If $X \sim \textsc{LogNorm}(\mu, \sigma^2)$ then $\log(X) \sim \textsc{Norm}(\mu, \sigma^2)$.

Simulation Study I

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

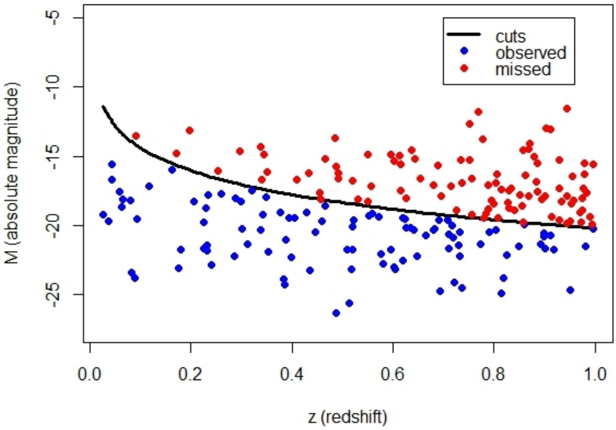


Simulation I ($\mu_0 = -19.3$, $\sigma_m = 20$, $\nu = 0.02$, $\beta^2 = 0.02$)

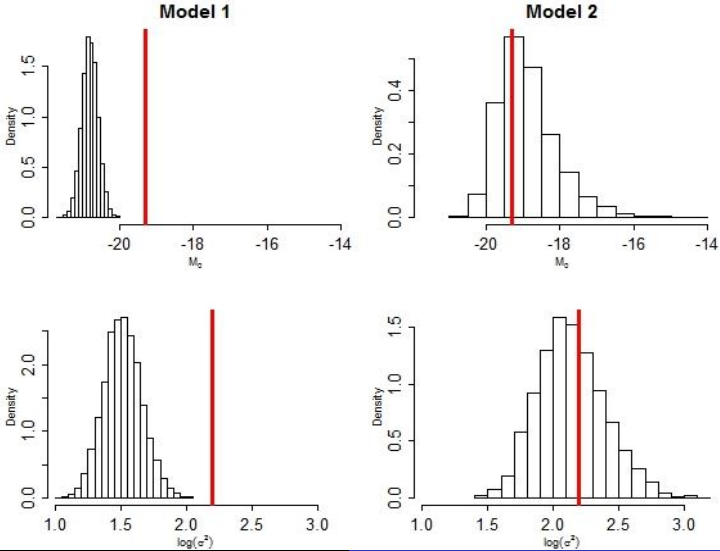


Simulation Study II

- 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:

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

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:

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

Mean Square Error: An Illustration

EXAMPLE: Suppose $H \sim \textsc{Binomial}(n = 3, \pi)$.

> **Recall:**
> If $H|n, \pi \stackrel{\text{dist}}{\sim} \textsc{Binomial}(n, \pi)$ and $\pi \stackrel{\text{dist}}{\sim} \textsc{Beta}(\alpha, \beta)$
>
> then $\pi|H, n \stackrel{\text{dist}}{\sim} \textsc{Beta}(h + \alpha, n - h + \beta)$.

Consider four estimates of π:

i) $\hat{\pi}_1 = H/n$, the maximum likelihood estimator of π;

ii) $\hat{\pi}_2 = \mathrm{E}(\pi|H)$, where π has prior distribution $\pi \sim \mathrm{Beta}(1, 1)$

iii) $\hat{\pi}_3 = \mathrm{E}(\pi|H)$, where π has prior distribution $\pi \sim \mathrm{Beta}(1, 4)$

iv) $\hat{\pi}_4 = \mathrm{E}(\pi|H)$, where π has prior distribution $\pi \sim \mathrm{Beta}(4, 1)$

Frequency Properties of Estimators and Intervals

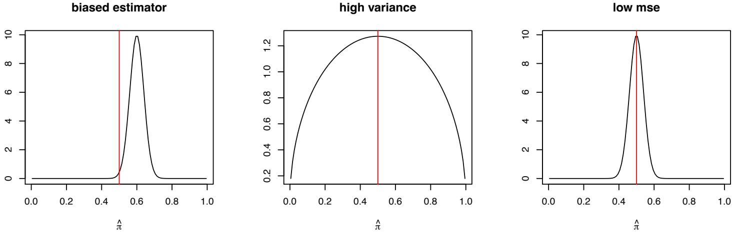
Remember: If the data is a random sample of all possible data, the estimator $\hat{\pi}_i$ is also random. It has a distribution, mean, and variance.

We can evaluate the $\hat{\pi}_i$ as an estimator of π in terms of its

bias: $\mathrm{E}(\hat{\pi}_i \mid \pi) - \pi$ (Is bias bad??)

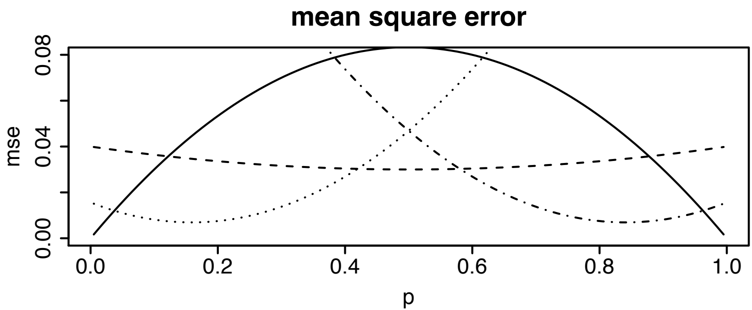
variance: $\mathrm{E}\left[\left(\hat{\pi}_i - \mathrm{E}(\hat{\pi}_i \mid \pi)\right)^2 \mid \pi\right]$

mean square error: $\mathrm{E}\left[(\hat{\pi}_i - \pi)^2 \mid \pi\right] = \text{bias}^2 + \text{variance}$



ıth

MSE of Four Estimators of Binomial Probability



Solid: MLE **Dashed:** BETA(1,1) **Dotted:** BETA(1,4) **Mixed:** BETA(4,1)

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

Inadmissibility of the Sample Mean

Suppose we wish to estimate more than one parameter:

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

The obvious estimator:

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

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

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

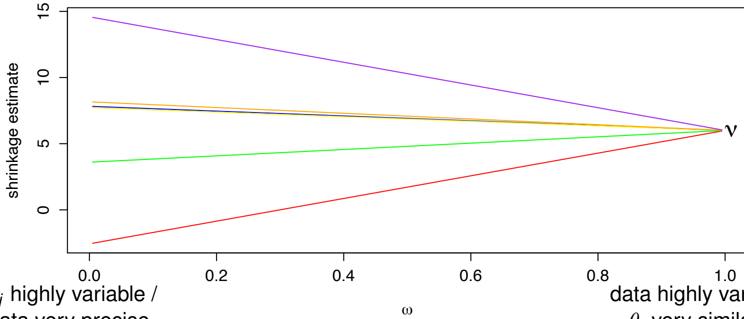
with $\omega^{\text{JS}} \approx \dfrac{\sigma^2/n}{\sigma^2/n + \tau_\nu^2}$ and $\tau_\nu^2 = \mathrm{E}[(\theta_i - \nu)^2]$.

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:

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



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

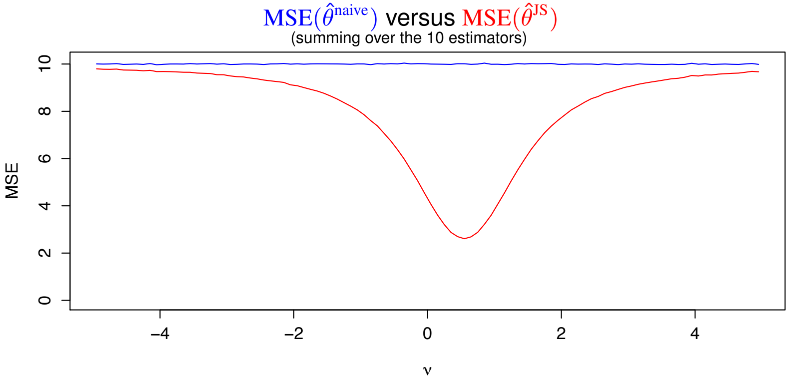
- Minimize τ^2.

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

Illustration

Suppose:

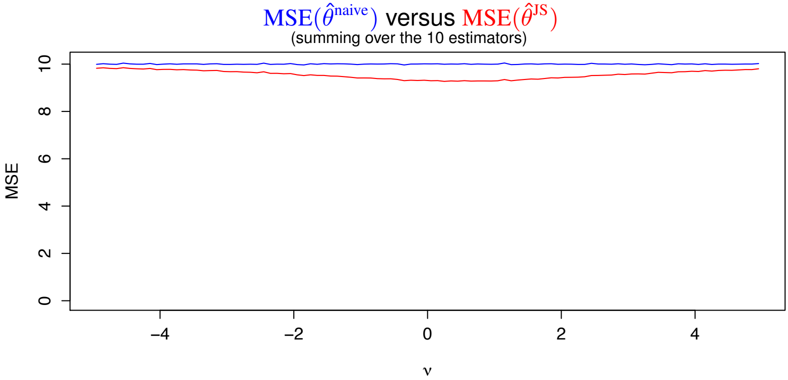
- $y_j \sim \textsc{Norm}(\theta_j, 1)$ for $j = 1, \ldots, 10$
- θ_j are evenly distributed on [0,1]



Illustration

Suppose:

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



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

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

with

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

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.

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:

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

with

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

- 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

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

To Derive the Gibbs Sampler Note:

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

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

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

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

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):

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

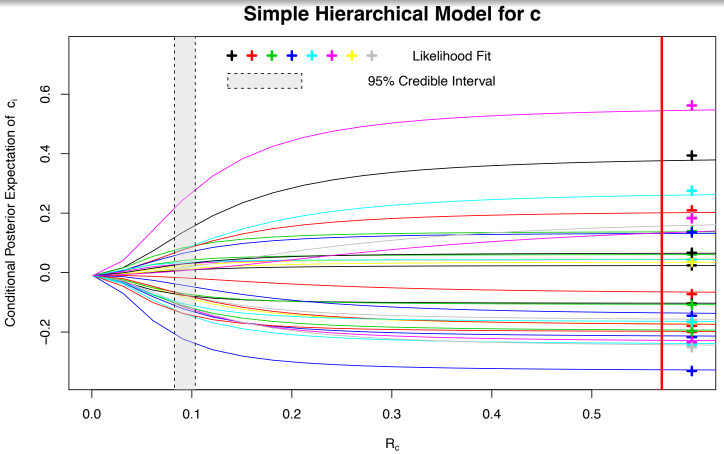
$$\mu_j = \left(\frac{\hat{c}_j}{\sigma_j^2} + \frac{c_0^{(t-1)}}{(R_C^2)^{(t-1)}}\right) \Big/ \left(\frac{1}{\sigma_j^2} + \frac{1}{(R_C^2)^{(t-1)}}\right); \quad s_j^2 = \left(\frac{1}{\sigma_j^2} + \frac{1}{(R_C^2)^{(t-1)}}\right)^{-1}.$$

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

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

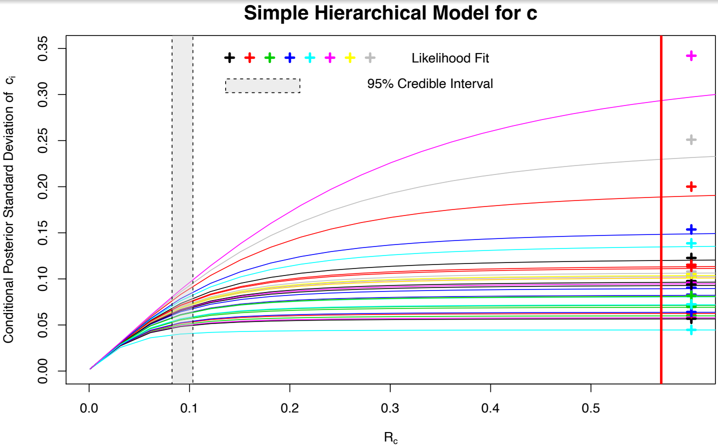
$$c_0^{(t)} \big| (c_1^{(t)}, \ldots, c_G^{(t)}), (R_C^2)^{(t)} \sim \textrm{NORM}\left(\bar{c}, (R_C^2)^{(t)}/G\right)$$

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,

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

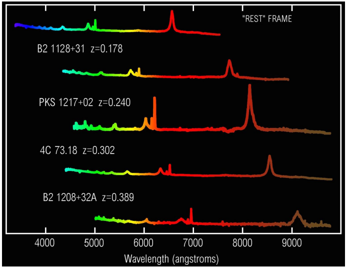
leads to an improper posterior distribution:

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

Outline

The Expanding Universe

Redshift



http://www.noao.edu/image_gallery/html/im0566.html

For "nearby" objects,

$$z = \text{redshift} \propto \text{velocity}$$
$$= H_0 \text{ distance.}$$

Hubble's Famous Diagram

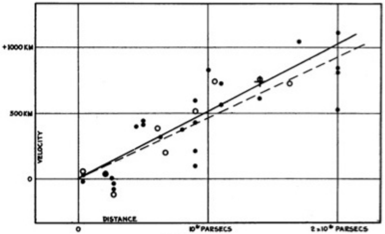


FIGURE 1

Velocity-Distance Relation among Extra-Galactic Nebulae.

Radial velocities, corrected for solar motion, are plotted against distances estimated from involved stars and mean luminosities of nebulae in a cluster. The black discs and full line represent the solution for solar motion using the nebulae individually; the circles and broken line represent the solution combining the nebulae into groups; the cross represents the mean velocity corresponding to the mean distance of 22 nebulae whose distances could not be estimated individually.

Hubble (1929)

The Big Bang!

Distance Modulus in an Expanding Universe

Apparent magnitude - Absolute magnitude = Distance modulus:

$$m - M = \mu \quad \left[= 5\log_{10}(\text{distance [Mpc]}) + 25 \right]$$

Relationship between μ and z

- For nearby objects,

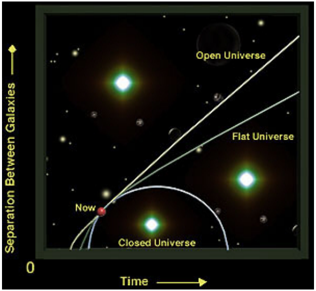
  $$\text{distance} = \mu \propto z.$$

  (Correcting for peculiar/local velocities.)

- For distant objects, involves expansion history of Universe:

  $$\mu = g(z, \Omega_\Lambda, \Omega_M, H_0)$$

  [function of density of dark energy and of total matter]



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

If we observe both m and M we can infer μ and the cosmological parameters.

Type Ia Supernovae

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

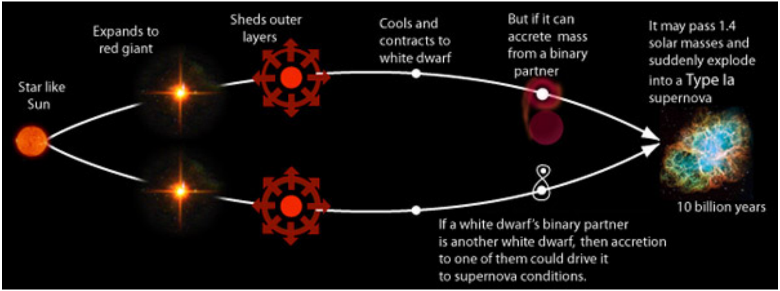


Image Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

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

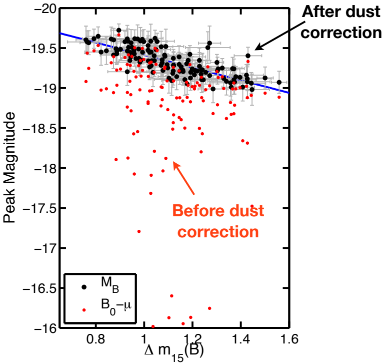
$$M_j \sim \textsc{Norm}(M_0, \sigma^2_{\text{int}}).$$

Non-linear Regression: $m_{Bj} = g(z_j, \Omega_\Lambda, \Omega_M, H_0) + M_j$

Phillips Corrections

- Recall:
  $M_j \sim \textsc{Norm}(M_0, \sigma^2_{\text{int}})$.
- Regression:
  $$M_j = -\alpha x_j + \beta c_j + M_j^{\epsilon},$$
  - $M_j^{\epsilon} \sim \textsc{Norm}(M_0, \sigma^2_{\epsilon})$.
  - x_j is a LC stretch
  - c_j is color correction.
- $\sigma^2_{\epsilon} \leqslant \sigma^2_{\text{int}}$
- Reduce variance, increase precision of estimates.

Low-z calibration sample



Mandel et al (2011)

Light curve stretch

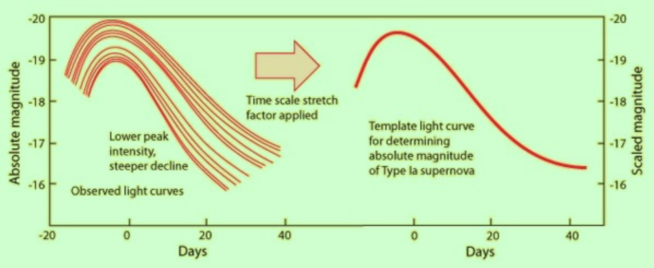
Brighter SNIa are slower decliners over time.

Non-linear Regression: $m_{Bj} = g(z_j, \Omega_\Lambda, \Omega_M, H_0) + \alpha x_j + \beta c_j + M_j^{\epsilon}$

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

A Hierarchical Model.

Level 1:[3] 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

[3] Shariff et al (2016). BAHAMAS: SNIa Reveal Inconsistencies with Standard Cosmology. ApJ 827, 1.

Other Model Features

Results are based on an SDSS (2009) sample of 288 SNIa.[4]

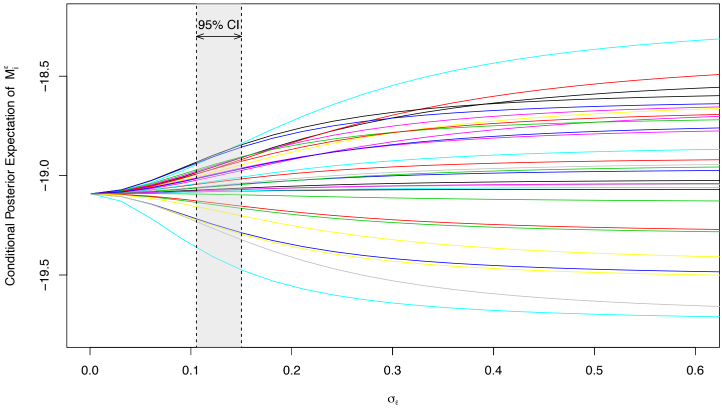
In our full analysis, we also

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, and
3. use a larger JLA sample[5] of 740 SNIa observed with SDSS, HST, and SNLS.

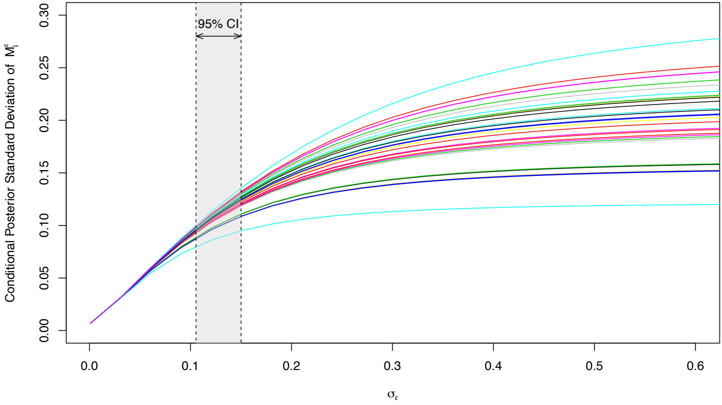
[4] Shariff et al (2016). BAHAMAS: New SNIa Analysis Reveals Inconsistencies with Standard Cosmology. ApJ 827, 1.

[5] 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

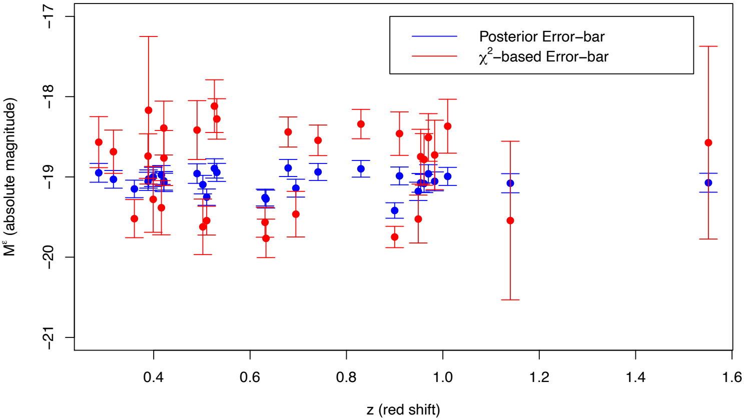
we have

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

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



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, $g(z_i, \Omega_\Lambda, \Omega_M, H_0)$?

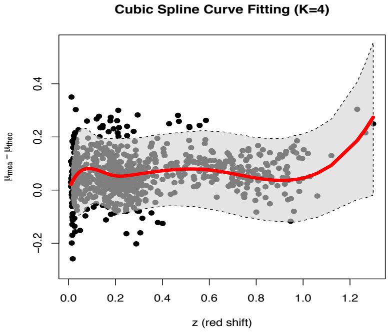
We can check the model by adding a cubic spline term:

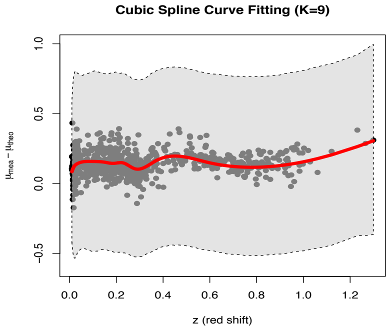
$$m_{Bi} = g(z_i, \Omega_\Lambda, \Omega_M, H_0) + h(z_i) - \alpha x_i + \beta c_i + M_i^\epsilon$$

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.

Classification of Sources

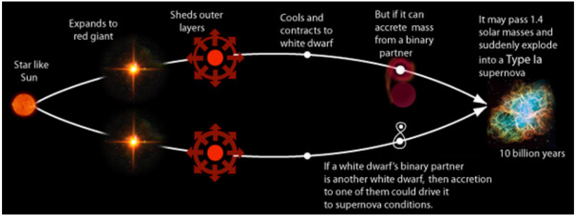


Image Credit: http://hyperphysics.phy-astr.gsu.edu/hbase/astro/snovcn.html

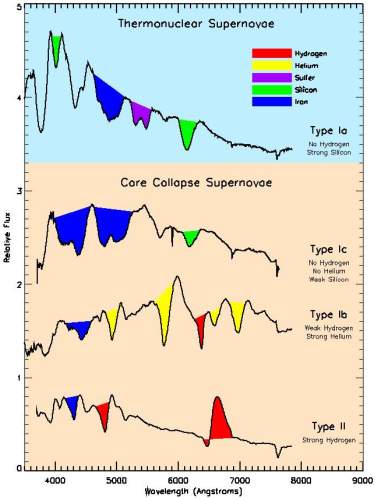
Due to common “flashpoint”, SN1a have similar absolute magnitudes:

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

Non-linear Regression: $m_{Bj} = g(z_j, \Omega_\Lambda, \Omega_M, H_0) + M_j$

It is critical that we are able to identify a sample of Type 1a Supernovae.

Identifying Type Ia SN is Critical



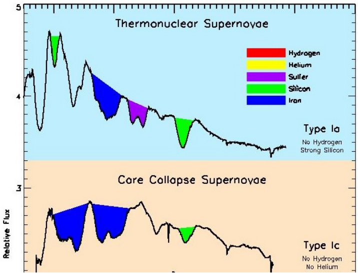
http://supernova.lbl.gov/~dnkasen/tutorial/

Spectral Classification

- **Type Ia**
  - Reignition of nuclear fusion in WD.
  - No Hydrogen, strong Silicon
- **Others**
  - Gravitational collapse in massive stellar core.

Spectroscopic and Photometric Data

Spectroscopic Redshift

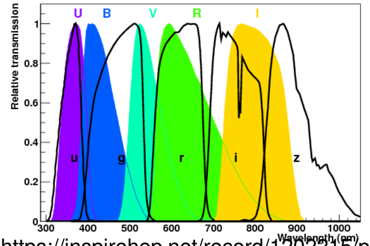


http://www.noao.edu/image_gallery/html/im0566.html

Can we Train a Classifier

- Train on Spectroscopic
- Target = Photometric

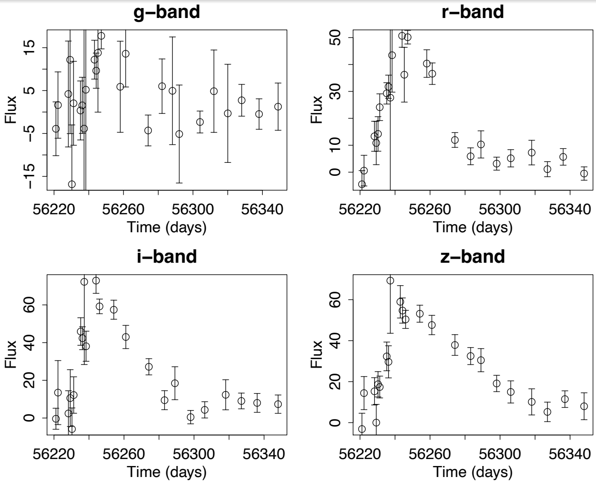
Photometric Redshift



https://inspirehep.net/record/1202215/plots

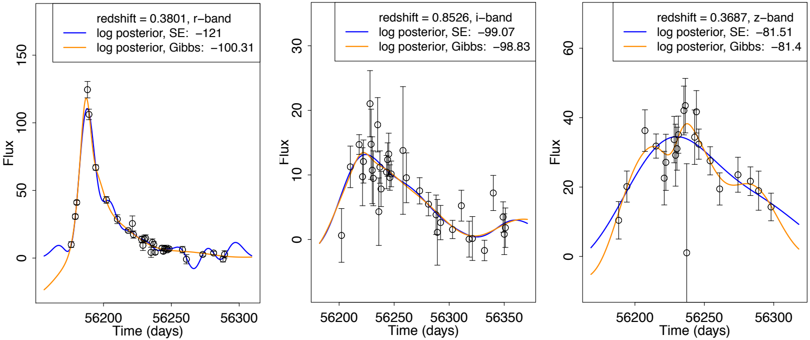
- Integrated average in each passband.
- More readily available, but far less informative.

Photometric Lightcurve Data



- Supernova photometric classification challenge (Kessler, 2010).
- Irregular observation times: interpolate for comparison

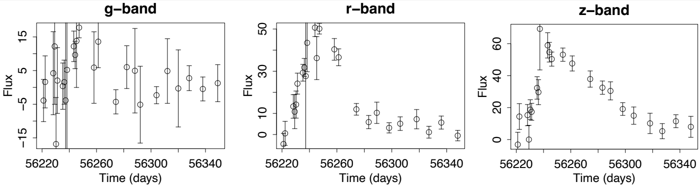
Gaussian Process Interpoloation



- Squared Exponential kernel: $K_{\mathrm{se}}(t,s) = \tau^2 \exp\left(-\frac{1}{2}\frac{(t-s)^2}{l^2}\right)$.
- Gibbs kernel: allows the length scale l to vary over time.
- Gibbs kernel appears to overfit, we use SE.

Photometric Classification of Supernovae

Data:



E.g., Supernova photometric classification challenges, such as Kessler (2010).

Classifier: Interpolate with Gaussian Process → Identify features w/ Diffusion Maps → Classify using Random Forest [6]

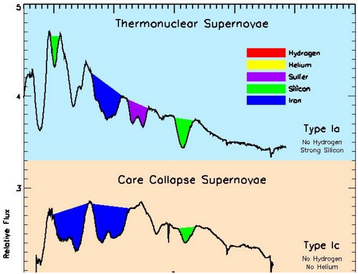
- Gaussian process fit of LCs (four color bands, g, r, i, z)
- Diffusion map, plus redshift and a measure of brightness, to extract **102 covariates**
- Random forest: cross validation to select hyperparameter

[6] Revsbech, Trotta, and van Dyk (2018). STACCATO: A Novel Solution to Supernova Photometric Classification with Biased Training Samples, **473**, 3969-3986.

Outline

Spectroscopic and Photometric Data

Spectroscopic Redshift

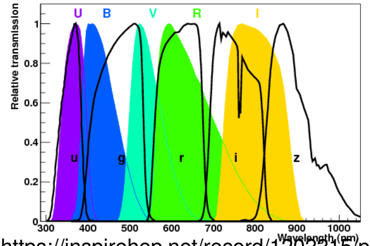


http://www.noao.edu/image_gallery/html/im0566.html

Can we Train a Classifier

- Train on Spectroscopic
- Target = Photometric

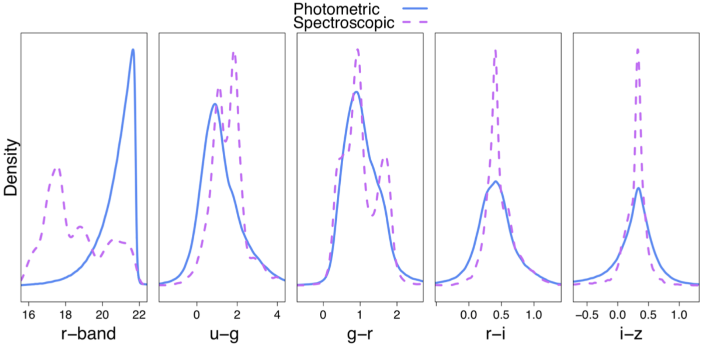
Photometric Redshift



https://inspirehep.net/record/1202215/plots

- Integrated average in each passband.
- More readily available, but far less informative.

Spectroscopic Training Set Not Representative



A General Challenge

- **Aim:** use training set (x, y) to predict target set (y from x).
- Spectroscopic data more available for bright/near objects.
- These object differ systematically from population.

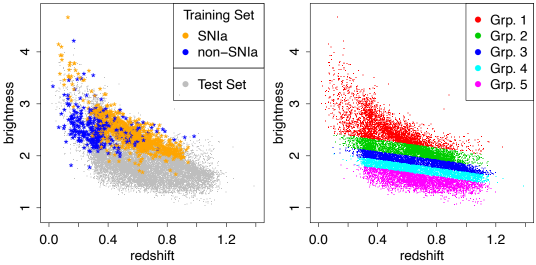
[Image Credit: Izbicki, Lee, Freeman, 2017, AoAS]

Covariate Shift

We Assume Covariate Shift:

$$p_{\text{training}}(y \mid x) = p_{\text{target}}(y \mid x) \quad \textbf{but} \quad p_{\text{training}}(x) \neq p_{\text{target}}(x)$$

Supernovae classification:



Learning methods must be adapted to account for non-representative training data.

Does a new drug improve health outcomes?

Causal Inference:

- Split subjects: treatment ($Z = 1$) and control ($Z = 0$) group
- What if treatment group differs systematically from control group, e.g., in terms of x.

$$p_{\text{treatment}}(x) \stackrel{?}{=} p_{\text{control}}(x)$$

- Randomiziation is the gold standard, not always possible.

Propensity Scores:

- Rosenbaum and Rubin (1983) define propensity scores:

$$e(x) = \Pr(Z = 1 \mid x).$$

- Demonstrate that $e(x)$ is a *balancing score*:

$$p_{\text{treatment}}(x \mid e(x)) = p_{\text{control}}(x \mid e(x)).$$

... easy to diagnose in practice!

Propensity for Selection to Training Set

Setup:

- We wish to predict y from x in target set.
- Use prediction function, $f(x)$, estimated in training set.
- In this context we define the propensity score:

$$e(x) = \Pr(\text{training set} \mid x), \text{ with } 0 < e(x) < 1.$$

Result:

Because $e(x)$ is a balancing score, under covariate shift,

$$p_{\text{target}}(x, y \mid e(x)) = p_{\text{train}}(x, y \mid e(x)).$$

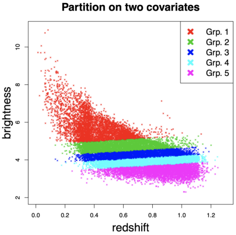
I.e, given $e(x)$ the joint test and target distributions are equal. It follows, that for any loss function $\ell(f(x), y)$,

$$\mathrm{E}_{\text{target}}[\ell(f(x), y) \mid e(x)] = \mathrm{E}_{\text{train}}[\ell(f(x), y) \mid e(x)].$$

StratLearn: Improved Learning under Covariate Shift

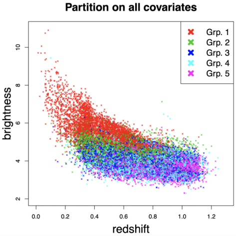
Propensity scores

- **Estimate:**
  $\hat{e}(x) = \Pr(\text{target set} \mid \text{covariates})$
- **Check:** $p_{\text{train}}(x \mid \hat{e}(x)) = p_{\text{target}}(x \mid \hat{e}(x))$
- Given $e(x)$, expected loss of predictor, $f(x)$, is same in target & training sets.



StratLearn

- Stratify training & target sets on $\hat{e}(x)$.
- Classify data separately in each strata.



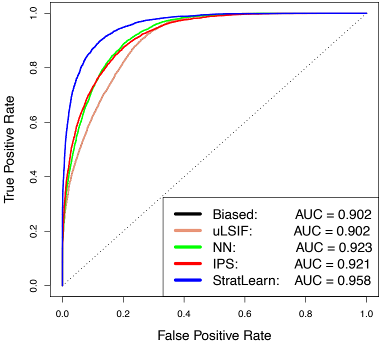
Reduce covariate shift and thus expected classification/prediction error.

Reference: Autenrieth, van Dyk, Trotta, and Stenning (2023). Stratified Learning: A General-Purpose Statistical Method for Improved Learning under Covariate Shift, SADM, 1-16.

Results for Supernova Classification

ROC for StratLearn and several existing weighting methods.

- "Biased" ignores Covariate Shift.
- With an unbiased training set AUC = 0.965.



Weighting Methods for Cov Shift

$$\mathrm{E}_{\text{target}}[\ell(f(x), y)] = \mathrm{E}_{\text{train}}\left[\frac{p_{\text{target}}(x)}{p_{\text{train}}(x)}\ell(f(x), y)\right]$$

- KLIEP (Sugiyama et al., 2008)
- uLSIF (Kanamori et al.. 2009);
- NN: Nearest-Neighbor (Kremer et al.. 2015);
- IPS: probabilistic classification (Kanamori et al.. 2009);

Unfortunately, large weights are highly variable and cause unreliable target predictions.

Example: Photo-z Conditional Density Estimation

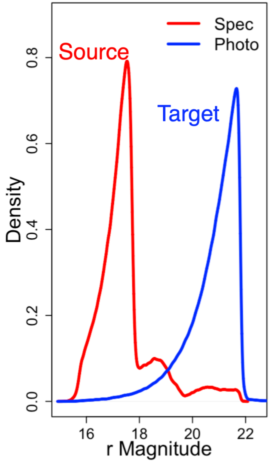
Objective:
Conditional density estimation $f(z|x)$ of redshift given photometric magnitudes.

Significant covariate shift in magnitudes.

Data (following Izbicki et al., 2017):

- 468k galaxies (Sheldon et al. 2012), spectroscopic redshift, 5 photometric magnitudes.
- Create non-representative training set.
- Add $k \in \{10, 50\}$ i.i.d. Gaussian covariates.

What is the effect of high-dimensional irrelevant covariates?



Example: Photo-z conditional density estimation

***Generalized* risk** (Izbicki, et al., 2017):

$$\hat{R}(\hat{f}) = \frac{1}{n_{\text{target}}} \sum_{i=1}^{n_{\text{target}}} \int \hat{f}^2(z|x_{\text{target}}^{(i)})dz - \frac{2}{n_{\text{train}}} \sum_{i=1}^{n_{\text{train}}} \hat{f}(z_{\text{train}}^{(i)}|x_{\text{train}}^{(i)})\hat{w}(x_{\text{train}}^{(i)}),$$

Conditional density estimation models:

- hist-NN, ker-NN, Series
- Comb (combination model):

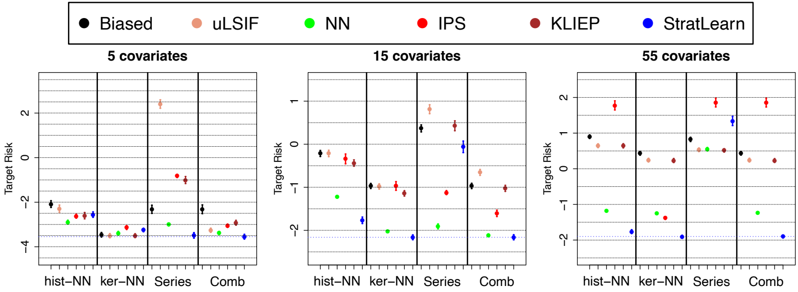
$$\hat{f}^{\alpha}(z|x) = \sum_{k=1}^{p} \alpha_k \hat{f}_k(z|x),$$

[where $\alpha_i \geqslant 0$ and $\sum_{k=1}^{p} \alpha_k = 1$.]

StratLearn:

- Minimize risk separately in each stratum (with $w(x) \equiv 1$).
- Optimize α separately for each strata (with $w(x) \equiv 1$).

Photo-z: Stress Test:



Target risk of photometric redshift estimates, using different sets of predictors.

StratLearn is especially advantageous with high dimensional covariates.

Outline

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)