Ideas for PDFs and the Banff Challenge

Chad M. Schafer

www.stat.cmu.edu/∼cschafer

Department of Statistics

Carnegie Mellon University

July 2010

The Core Collaborators

Philip B. Stark

Ann B. Lee

Peter E. Freeman

Susan M. Buchman

Joseph W. Richards

Work Supported by

NSF Grant #0707059, NASA AISR Grant, DOE Contract W-7405-Eng-48, ONR Grant N00014-08-1-073

The InCA Group: `www.incagroup.org`

Motivation

Cowan (2009), "Testing nature to the limit: the Large Hadron Collider," *Significance*, page 158:

"What the physicist would of course like to have is a test with maximal power with respect to a broad class of alternative hypotheses.

For a given signal model, for example, one would like to choose the acceptance and rejection regions based on the likelihood ratio

$$\frac{f_s(x)}{f_b(x)}.$$

Motivation

Cowan (2009), continued:

"In principle the signal and background theories should allow us to work out the required functions $f_s(x)$ and $f_b(x)$, but in practice the calculations are too difficult and we do not have explicit formulas for these.

What we have instead $\ldots$ are complicated Monte Carlo programs: that is, we can sample x to produce simulated signal and background events."

Motivation

Facing similar challenges in cosmology

How to estimate cosmological parameters when faced with complex model relating parameters to observable data?

Increasing use of simulation models

Motivation

Seek procedures (tests, confidence regions) that have "power against with maximal power with respect to a broad class of alternative hypotheses" that are physically feasible

The power tradeoff

Outline

$\Rightarrow$ Formalism:

Test Functions, Acceptance Probability Functions

$\Rightarrow$ Decision Theoretic Construction

$\Rightarrow$ From Theory to Practice

$\Rightarrow$ Related Problem in Cosmology

$\Rightarrow$ PDFs and the Banff Challenge

Formalism

Elements $\eta \in \Theta$ are parameter vectors that specify the distribution of the data:

In cosmology cases, $\eta = (H_0, \Omega_m, \Omega_\Lambda, \ldots)$

In case of estimating PDF, $\eta = (a_1, a_2, \ldots, a_{25})$

In the case of Banff Challenge, $\eta = ???$

Formalism

Test Function: $d(\eta, x)$ for $\eta \in \Theta$ and event data x

$$d(\eta, x) = \begin{cases} 1, & \text{if } \eta \text{ accepted when } x \text{ is observed} \\ 0, & \text{if } \eta \text{ rejected when } x \text{ is observed} \end{cases}$$

Of course,

$$d(\eta, x) = \begin{cases} 1, & \text{if } \eta \text{ included in confidence region} \\ 0, & \text{if } \eta \text{ excluded from confidence region} \end{cases}$$

Formalism

Acceptance Probability Function:

For $\theta, \eta \in \Theta$,

$$\begin{aligned}\gamma_d(\theta, \eta) &= \text{Probability test } d \text{ accepts } \eta \text{ when } \theta \text{ is truth} \\ &= \mathbf{P}_\theta(d(\eta, X) = 1)\end{aligned}$$

Formalism

Frequentists require choosing d such that

$$\gamma_d(\theta, \theta) \geq 1 - \alpha$$

for all $\theta \in \Theta$.

Bayesian credible regions satisfy

$$\int_\Theta \gamma_d(\theta, \theta)\, \pi(d\theta) = 1 - \alpha$$

for chosen prior π.

Decision Theoretic Construction

Neither of the above defines a unique choice for d.

Clearly, would prefer d that forces $\gamma_d(\theta, \theta)$ large while keeping $\gamma_d(\theta, \eta)$ small when $\theta \neq \eta$.

Propose decision theoretic considerations for choosing d.

Decision Theoretic Construction

Specify a nonnegative penalty function:

$$\phi(\theta, \eta) = \text{penalty for accepting } \eta \text{ when } \theta \text{ is truth}$$

Then define the loss function:

$$\mathbf{L}_d(\theta, x) = \int_\Theta \phi(\theta, \eta)\, d(\eta, x)\, d\eta$$

Note that $\mathbf{L}_d(\theta, x)$ is the accumulated penalties when d is used and data x is observed.

Decision Theoretic Construction

If choose $\phi(\theta, \eta) = 1$, then

$$\begin{aligned} \mathbf{L}_d(\theta, x) &= \int_\Theta d(\eta, x)\, d\eta \\ &= \text{Volume of confidence region,} \end{aligned}$$

a natural measure of precision.

Decision Theoretic Construction

If choose $\phi(\theta, \eta) = g(\eta)$, then

$$\begin{aligned} \mathbf{L}_d(\theta, x) &= \int_\Theta d(\eta, x)\, g(\eta)\, d\eta \\ &= \nu\text{-measure of confidence region,} \end{aligned}$$

where $g = d\nu / d\eta$.

Decision Theoretic Construction

For the PDF case, could choose

$$\phi(\theta, \eta) = \sum_{i=1}^{6} \| f_i(\theta) - f_i(\eta) \|,$$

where $f_i(\theta)$ is the i^{th} parton distribution function under parameters θ.

Decision Theoretic Construction

For the Banff challenge, could choose $\phi(\theta, \eta)$ as a function of whether or not the parameter vectors agree on their classification of background/signal:

If θ and η are both “background,” then make $\phi(\theta, \eta)$ small

If θ and η are both “signal,” then make $\phi(\theta, \eta)$ small

If θ is “signal” and η is “background,” then make $\phi(\theta, \eta)$ large

If θ is “background” and η is “signal,” then make $\phi(\theta, \eta)$ larger

Decision Theoretic Construction

Next define the risk function:

$$
\begin{aligned}
\mathbf{R}_d(\theta) &= \text{Expected loss when } \theta \text{ is truth} \\
&= \mathbf{E}_\theta[\mathbf{L}_d(\theta, X)] \\
&= \\
&= \\
&=
\end{aligned}
$$

Decision Theoretic Construction

Next define the risk function:

$$\begin{aligned}
\mathbf{R}_d(\theta) &= \text{Expected loss when } \theta \text{ is truth} \\
&= \mathbf{E}_\theta[\mathbf{L}_d(\theta, X)] \\
&= \int_{\mathcal{X}} \int_{\Theta} \phi(\theta, \eta)\, d(\eta, x)\, f_\theta(x)\, d\eta\, dx \\
&= \\
&=
\end{aligned}$$

Decision Theoretic Construction

Next define the risk function:

$$
\begin{aligned}
\mathbf{R}_d(\theta) &= \text{Expected loss when } \theta \text{ is truth} \\
&= \mathbf{E}_\theta[\mathbf{L}_d(\theta, X)] \\
&= \int_{\mathcal{X}} \int_{\Theta} \phi(\theta, \eta)\, d(\eta, x)\, f_\theta(x)\, d\eta\, dx \\
&= \int_{\Theta} \gamma_d(\theta, \eta)\, \phi(\theta, \eta)\, d\eta \\
&=
\end{aligned}
$$

Decision Theoretic Construction

Next define the risk function:

$$\begin{aligned}
\mathbf{R}_d(\theta) &= \text{Expected loss when } \theta \text{ is truth} \\
&= \mathbf{E}_\theta[\mathbf{L}_d(\theta, X)] \\
&= \int_{\mathcal{X}} \int_{\Theta} \phi(\theta, \eta)\, d(\eta, x)\, f_\theta(x)\, d\eta\, dx \\
&= \int_{\Theta} \gamma_d(\theta, \eta)\, \phi(\theta, \eta)\, d\eta \\
&= \text{Weighted average of acceptance} \\
& \text{probabilities}
\end{aligned}$$

Decision Theoretic Construction

Bayes risk for prior π:

$$\mathbf{B}_d(\pi) \equiv \int_\Theta \mathbf{R}_d(\theta)\, \pi(d\theta)$$

Neyman-Pearson Lemma: To minimize $\mathbf{B}_d(\pi)$,

$$d(\eta, x) = 1 \quad \text{if} \quad \frac{\int_\Theta f_\theta(x)\, \phi(\theta, \eta)\, \pi(d\theta)}{f_\eta(x)} \leq K_\eta$$

Denote this Bayes procedure d_π

Decision Theoretic Construction

Alternatively, one could seek d that is minimax, i.e. it minimizes

$$\max_{\theta \in \Theta} \mathbf{R}_d(\theta)$$

Either of these possibilities sets up a difficult computational problem.

From Theory to Practice

Instead, only limit $\mathbf{R}_d$ over densities of the form

$$f(x) = \sum_{i=1}^{p} \rho_i f_i(x)$$

where $f_1, f_2, \ldots, f_p$ are user-specified basis densities.

The nonnegative mixing coefficients ρ_i satisfy

$$\sum_{i=1}^{p} \rho_i = 1.$$

From Theory to Practice

Schafer and Stark (2009):

Monte Carlo algorithm for approximating $d(\eta, x)$ that minimizes the maximum value of $\mathbf{R}_d(\theta)$ under the assumption that

$$f(x) = \sum_{i=1}^{p} \rho_i f_i(x).$$

Considers the case $\phi(\theta, \eta) = 1$, but theory extends

From Theory to Practice

Minimax Expected Size (MES) procedure

Pratt (1961):

$$
\begin{aligned}
\mathbf{R}_d(\theta) &= \int_\Theta \gamma_d(\theta, \eta)\, d\eta \\
&= \int_\Theta \mathrm{P}_\theta(\eta \text{ in confidence set})\, d\eta \\
&= \text{Expected volume of confidence set}
\end{aligned}
$$

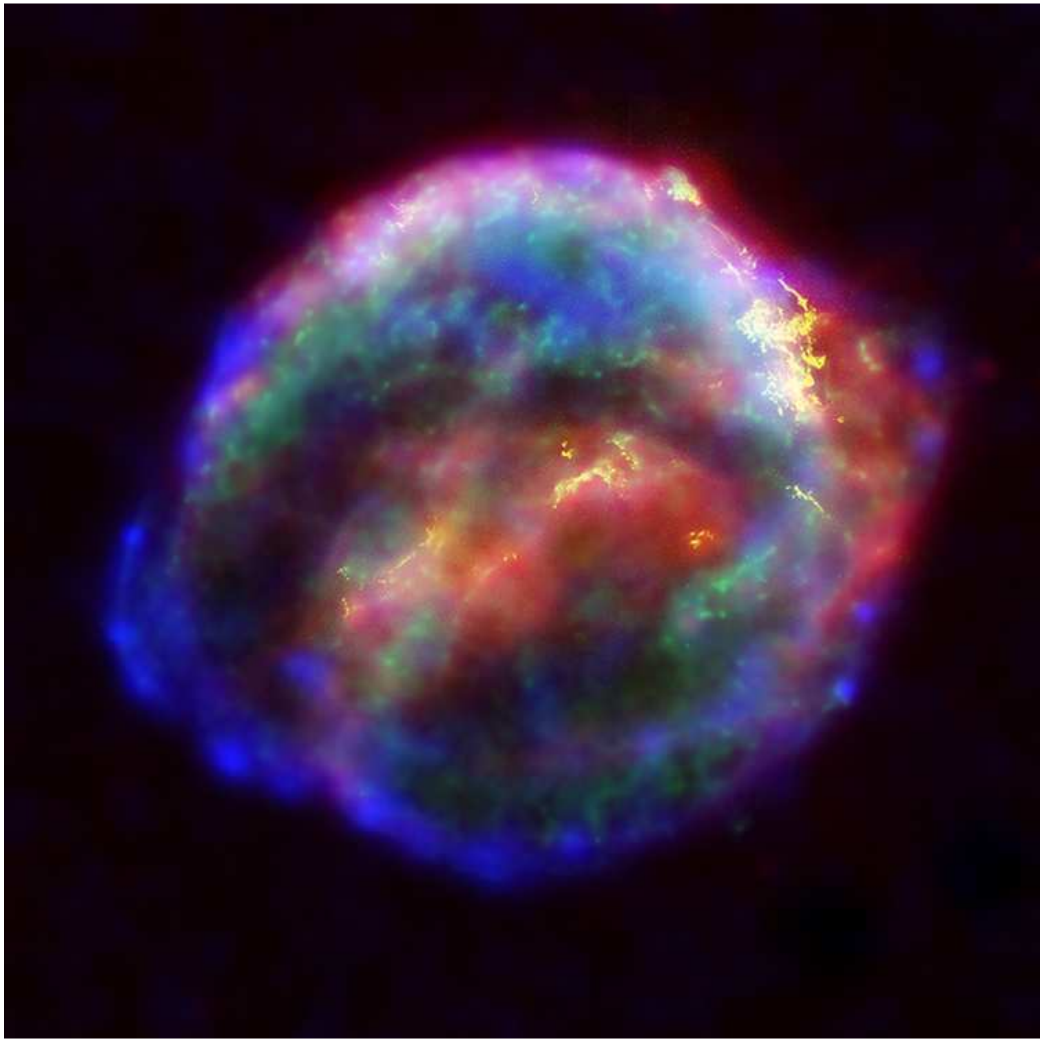
Type Ia Supernovae Analysis

Type Ia Supernovae are exploding stars, and standard candles

Observe redshift (z) and apparent magnitude (m)

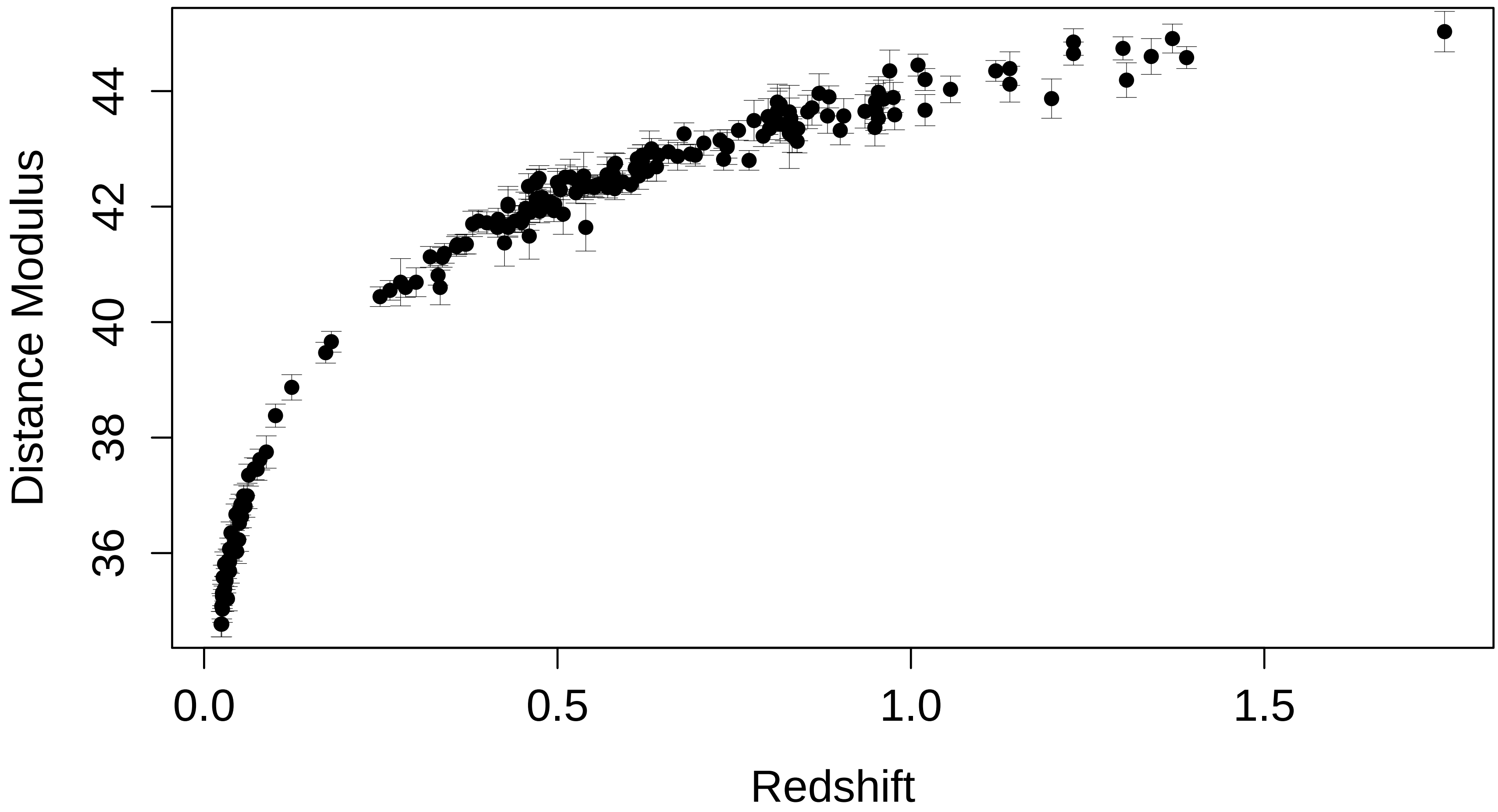
Theory predicts relationship between redshift and distance modulus as a function of cosmological parameters

Type Ia Supernovae Analysis



Credit: NASA, ESA, R. Sankrit and W. Blair (Johns Hopkins University)

Type Ia Supernovae Analysis



From Riess, et al. (2007), 182 Type Ia Supernovae

Type Ia Supernovae Analysis

Simple, flat cosmology, two parameter model:

$$\mu(z \mid \theta) = 5\log_{10}\left(\frac{c(1+z)}{H_0}\int_0^z \frac{du}{\sqrt{\Omega_m(1+u)^3 + (1-\Omega_m)}}\right) + 25$$

Observed pairs (z_i, Y_i) are realizations of

$$Y_i = \mu(z_i \mid \theta) + \sigma_i \epsilon_i,$$

where the ϵ_i are i.i.d. standard normal.

Type Ia Supernovae Analysis

To establish link with previous notation:

$\theta = (H_0, \Omega_m)$, the two cosmological parameters.

Θ is the range of the cosmological parameters considered physically possible. We assume $60 \leq H_0 \leq 90$ and $500 \leq \Omega_m H_0^2 \leq 2500$

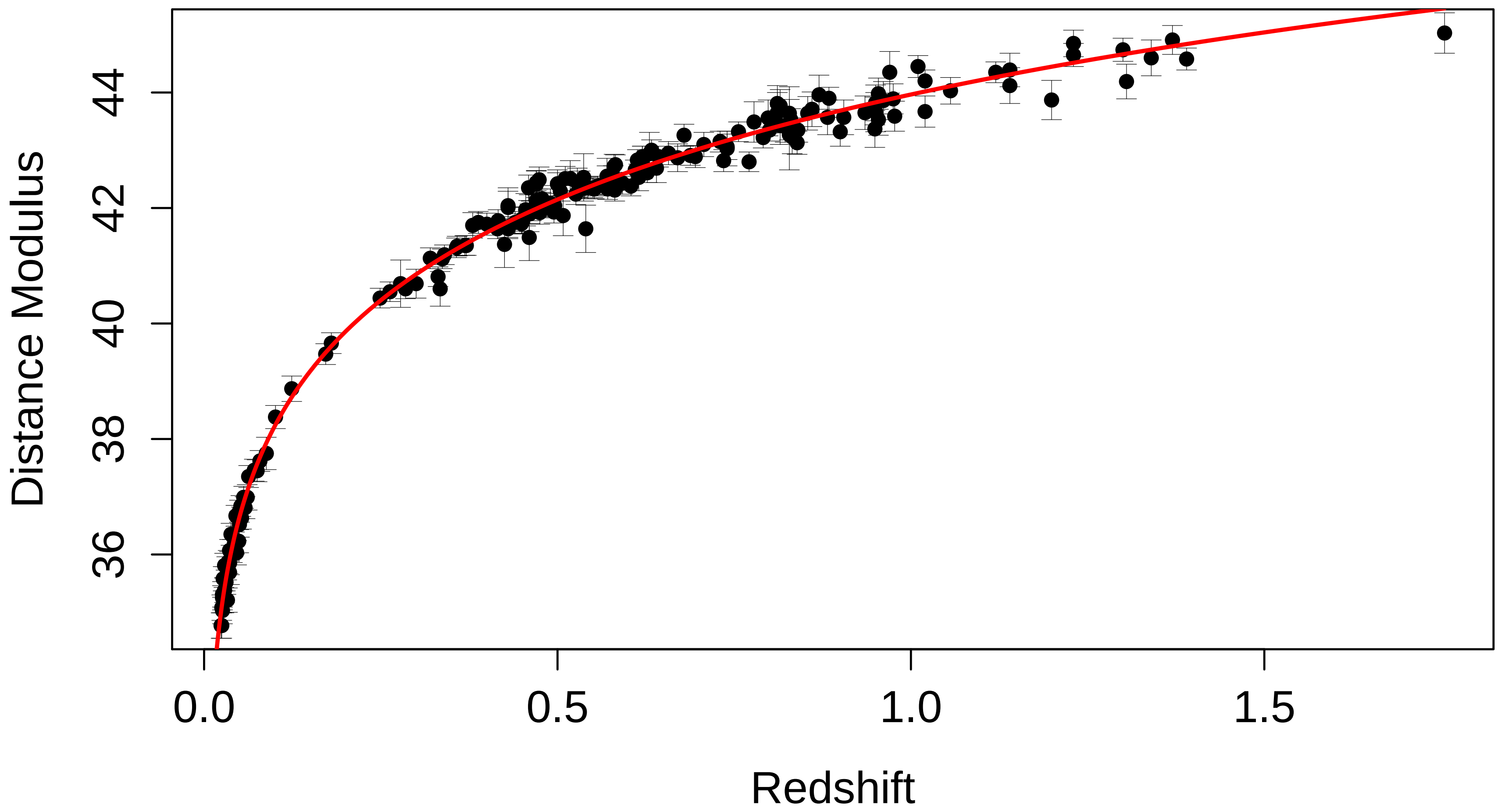
Type Ia Supernovae Analysis

x is the collection of all 182 pairs (z_i, Y_i)

$f_\theta(x)$ is the multivariate normal distribution with mean and covariance given by the "complex" model

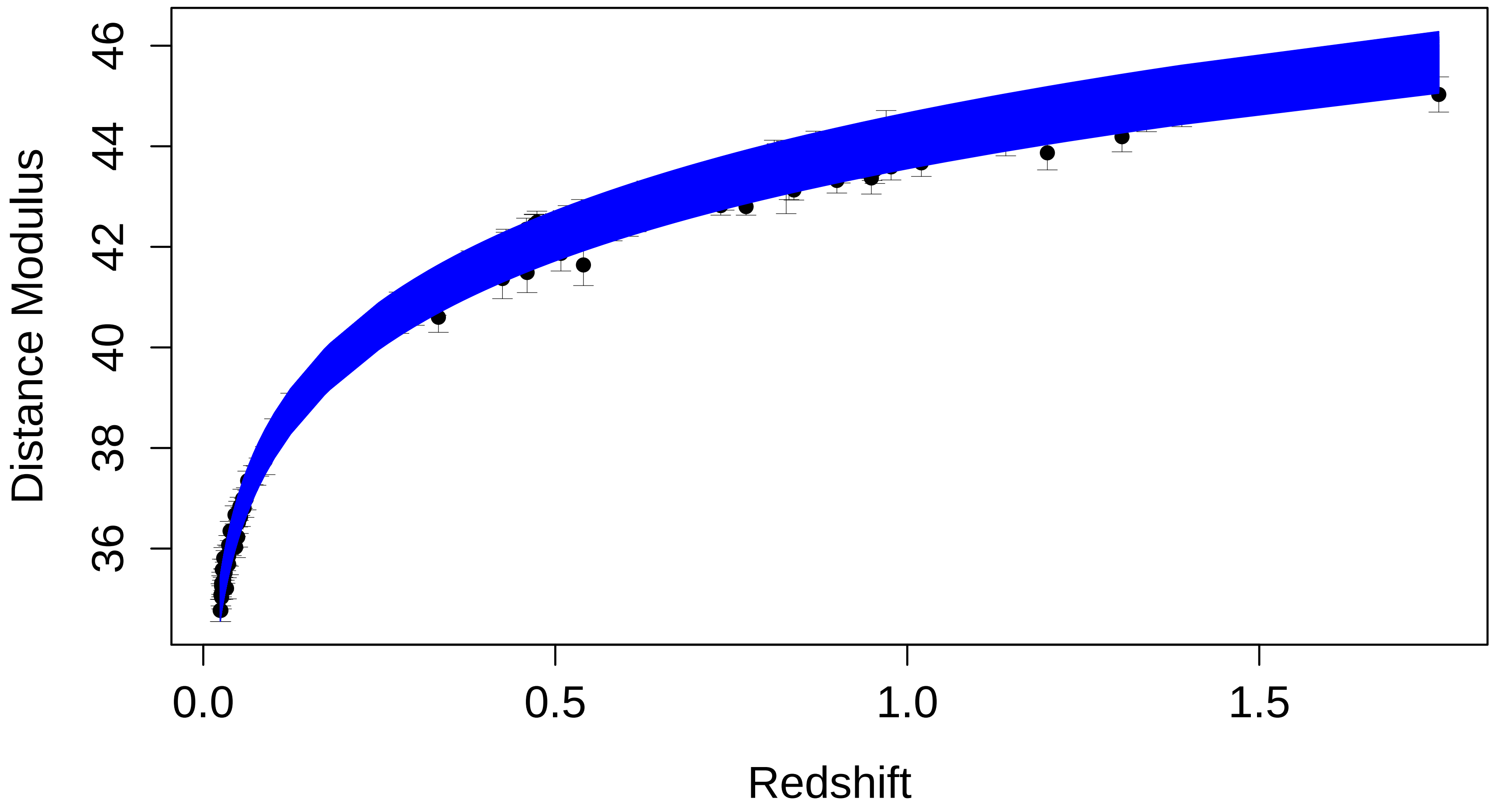
The objective is to construct a 95% confidence region for (H_0, Ω_m) that minimizes $\gamma_d(\theta, \eta)$ to the extent possible

Type Ia Supernovae Analysis



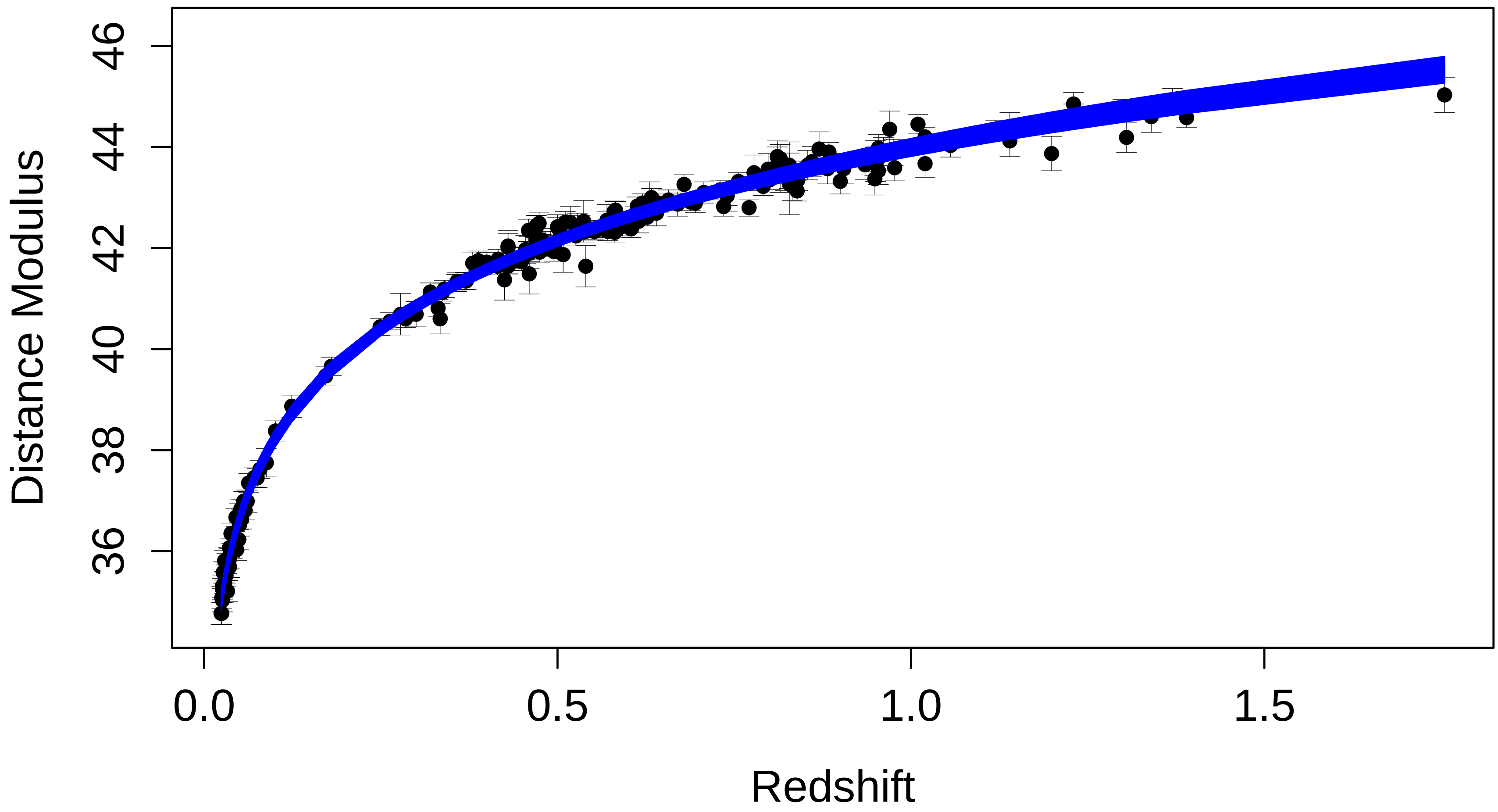
Curve is case where $H_0 = 72.76$ and $\Omega_m = 0.341$ (the MLE)

Type Ia Supernovae Analysis



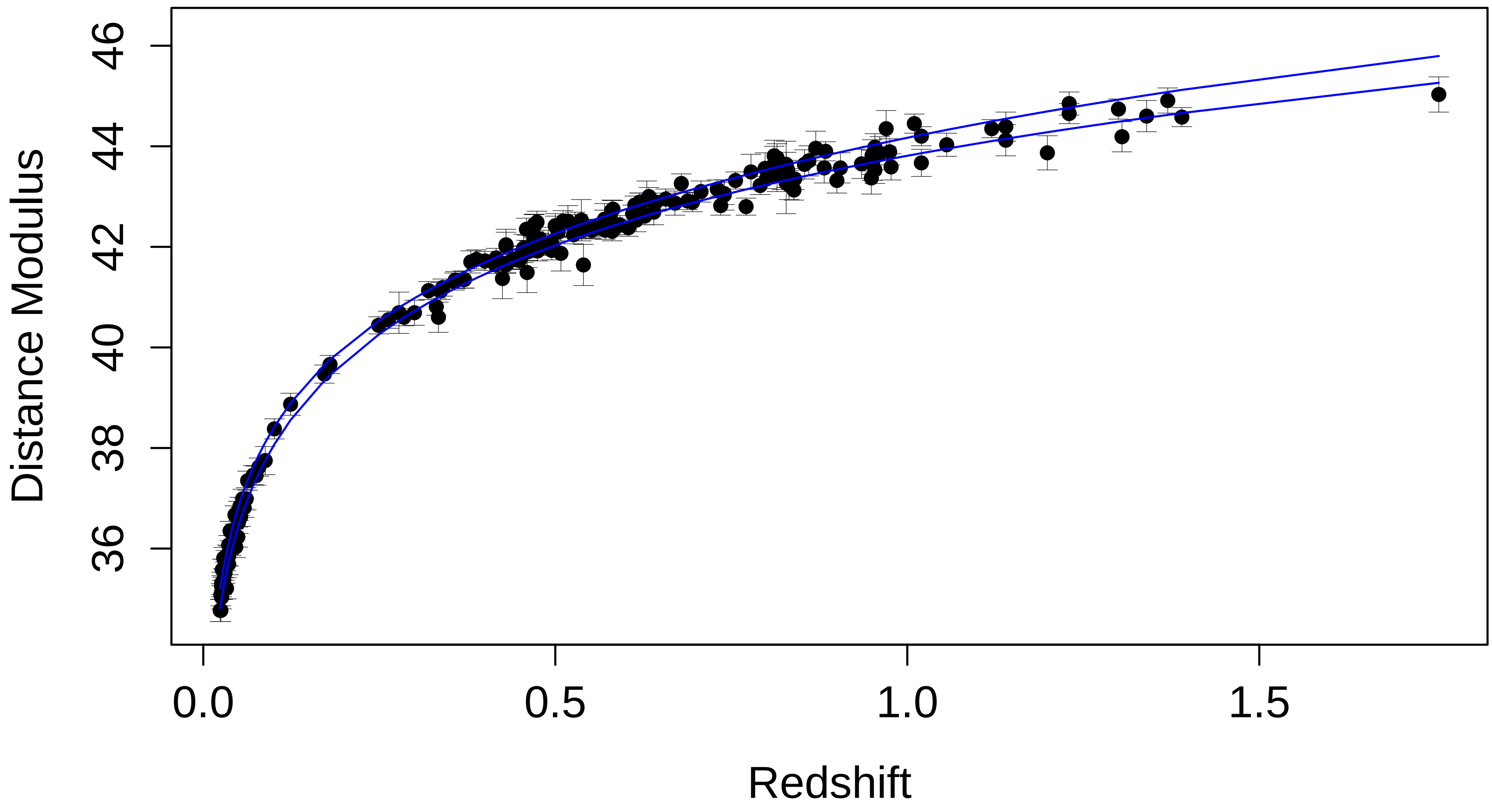
The collection of tested theories: $d(\eta, x)$ is for each η depicted

Type Ia Supernovae Analysis



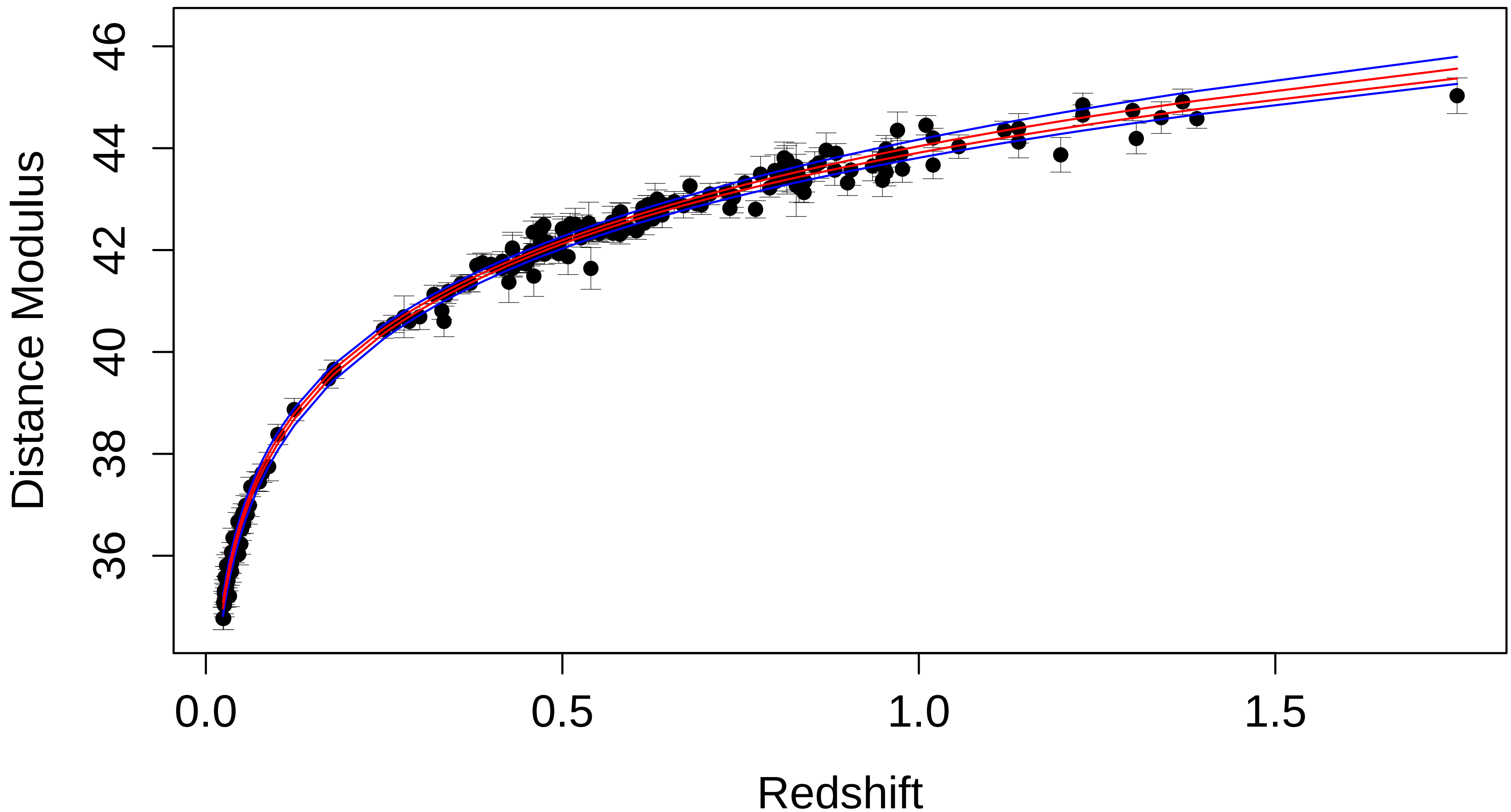
Those accepted by a chi-squared test

Type Ia Supernovae Analysis



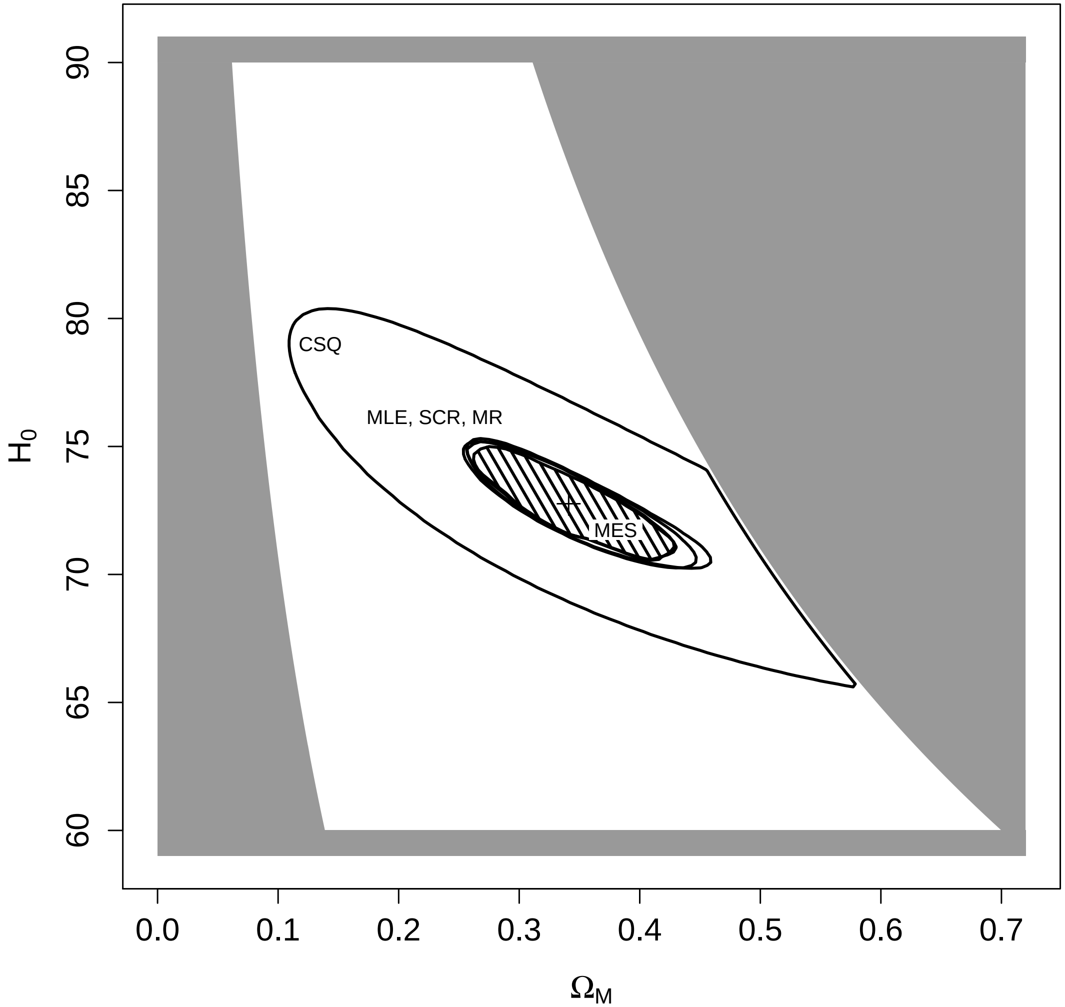
The range of those accepted by the chi-squared test

Type Ia Supernovae Analysis



The range of those accepted by MES

Type Ia Supernovae Analysis



Schafer and Stark (2009)

Type Ia Supernovae Analysis

Practical Concern:

How to choose the the basis densities f_i?

In this case, use a set of densities f_θ for p values of θ

Ideally, the distributions are evenly "spread out"

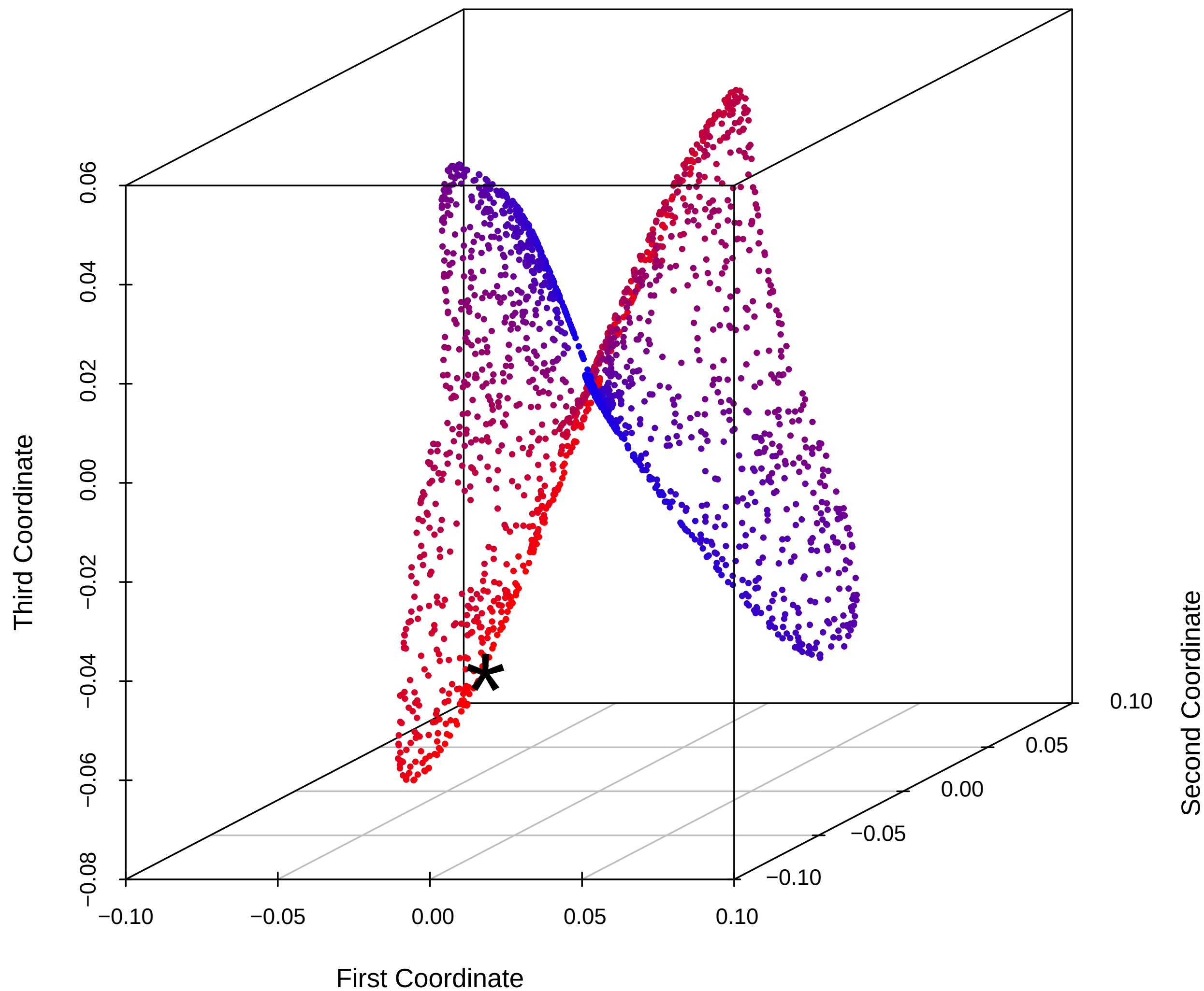
Type Ia Supernovae Analysis

The Hellinger Distance between distributions:

$$\mathcal{H}(f,g) = \sqrt{\frac{1}{2}\int\left(\sqrt{f(x)} - \sqrt{g(x)}\right)^2 dx}$$

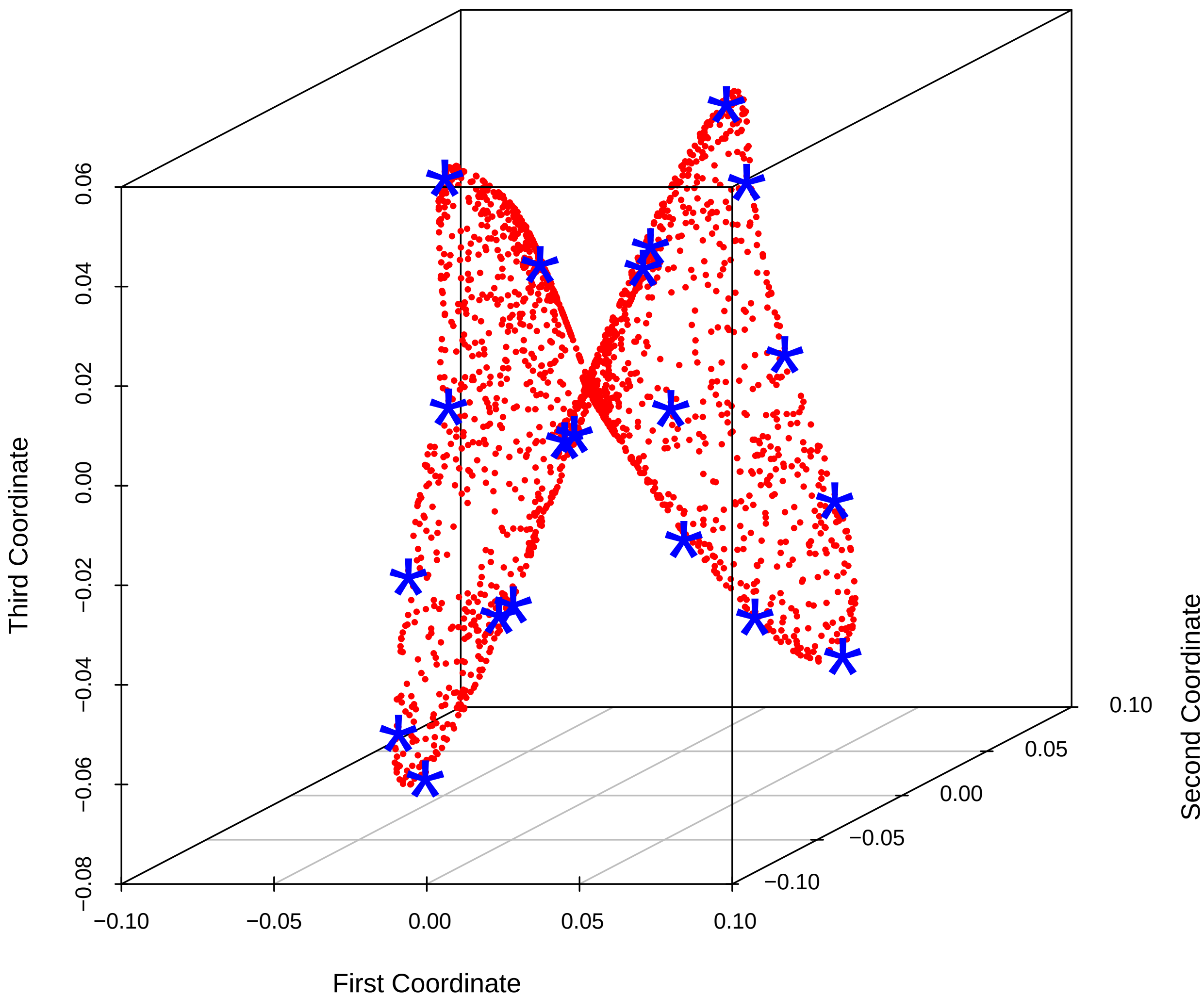
Note that $0 \leq \mathcal{H}(f,g) \leq 1$

Type Ia Supernovae Data



“Theories” are spaced by their similarity

Type Ia Supernovae Analysis



Ideally, the f_i would be representative of all the possible truths

Banff Challenge

Could marginalize over the nuisance parameters $\beta_{\text{signal}}, \beta_{\text{background}}, \epsilon_{\text{signal}}, \epsilon_{\text{background}}, \mathcal{L}$

I interpreted x to be the individual event data.

Define ξ as the parameter

$$\xi = \begin{cases} 1, & \text{if from signal} \\ 0, & \text{if from background} \end{cases}$$

Banff Challenge

Assume that

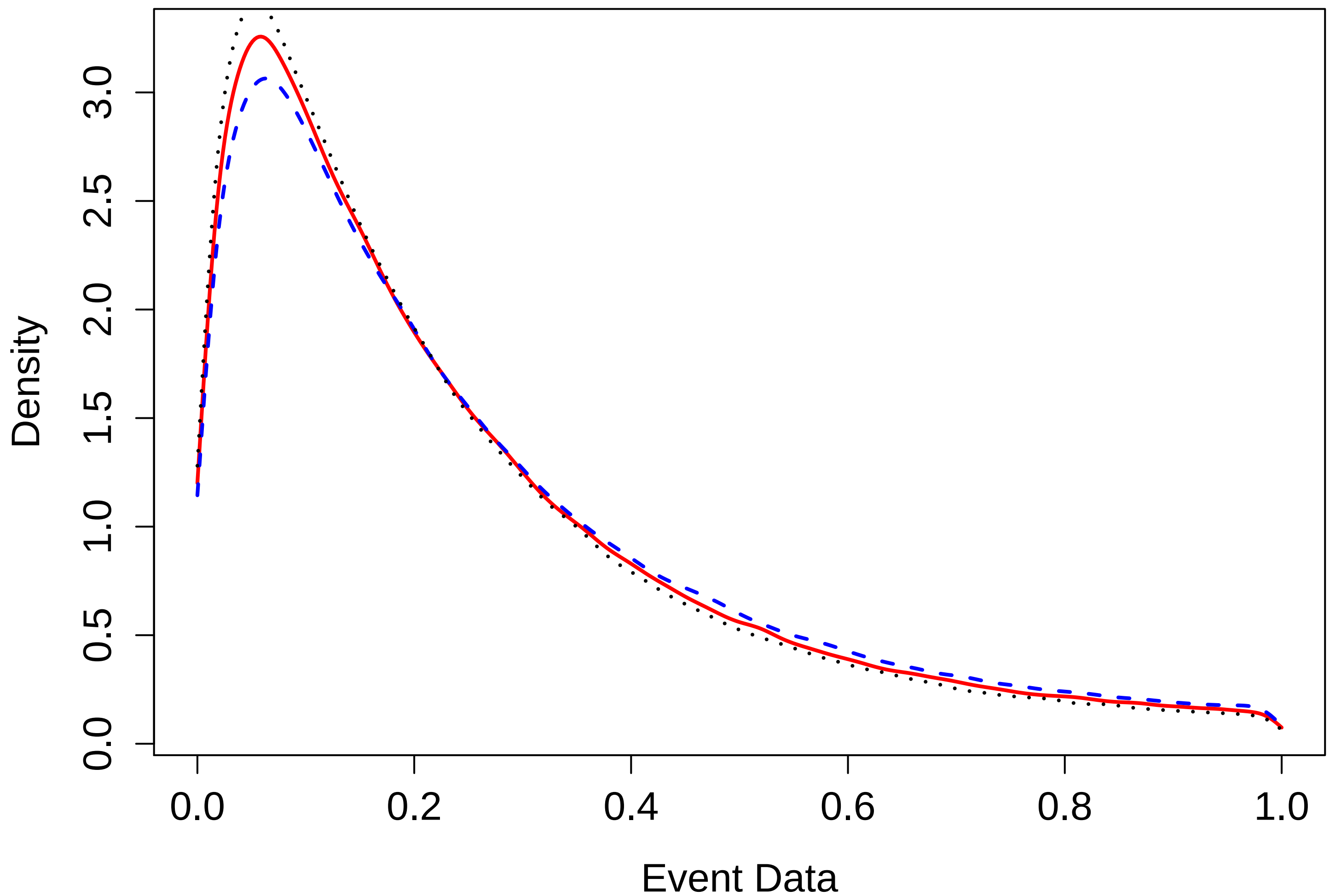
$$f_b(x) = \sum_{i=1}^{3} \alpha_i f_{\text{background},i}(x)$$

and

$$f_s(x) = \sum_{i=1}^{3} \tau_i f_{\text{signal},i}(x)$$

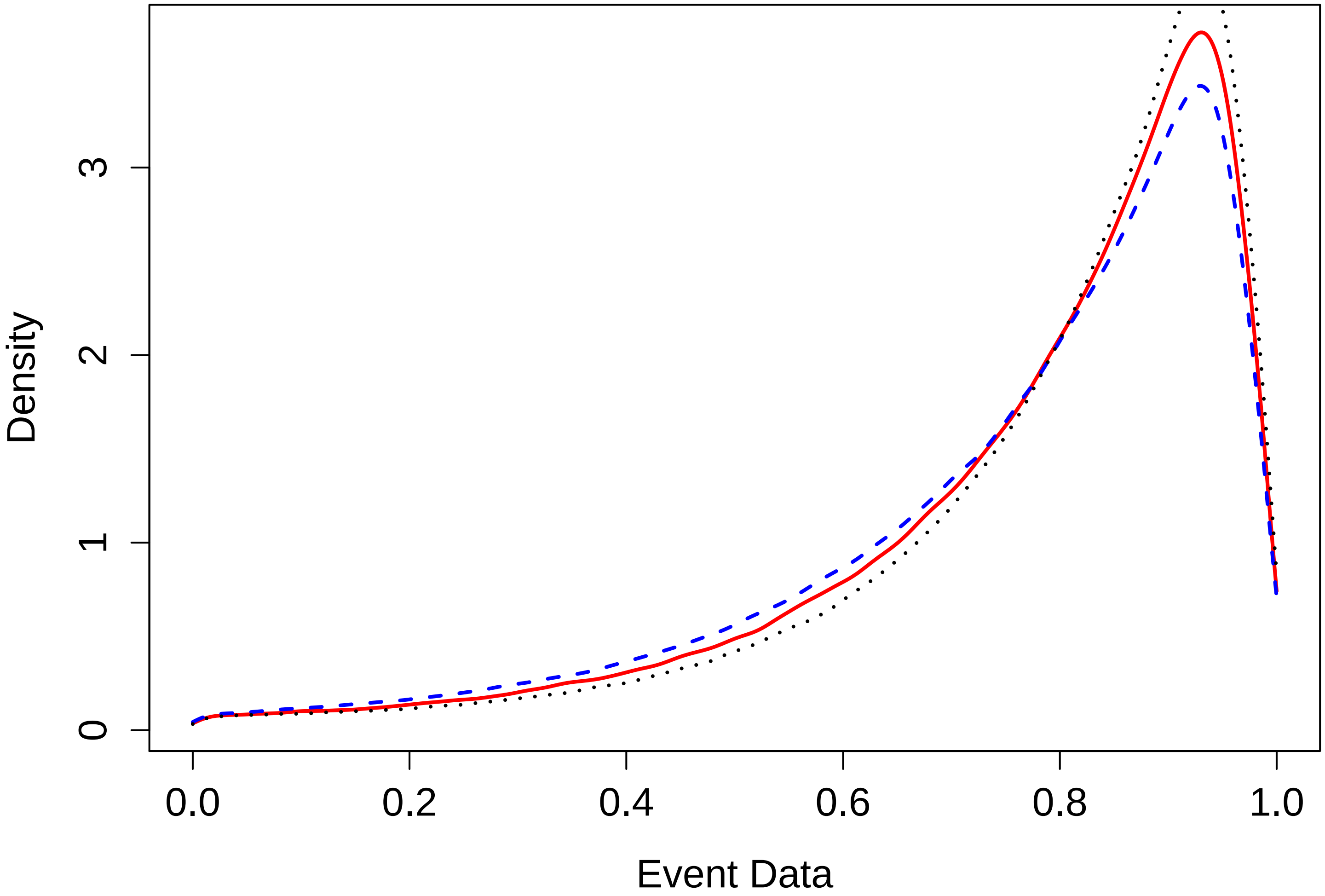
as a way of compensating for uncertainty in these densities.

Banff Challenge



Challenge One, the three background distributions

Banff Challenge



Challenge One, the three signal distributions

Banff Challenge

Now have six effective parameters

When $\xi = 0$, $\sum_{i=1}^{3} \alpha_i = 1$ and $\tau_i = 0$

When $\xi = 1$, $\sum_{i=1}^{3} \tau_i = 1$ and $\alpha_i = 0$

Basis densities can be the $f_{\text{background},i}(x)$ and $f_{\text{signal},i}(x)$

Banff Challenge

The penalty function only penalizes accpeting cases when $\xi = 0$ when, in fact, $\xi = 1$, and vice-versa.

Banff Challenge

Can test both the case where...

... $\xi = 0$, and get the p-value p_0

... $\xi = 1$, and get the p-value p_1

PDFs

Could work well, if willing to assume

$$\sum_{j=1}^{35}\sum_{i=1}^{N_j}\left(\frac{\text{data}_i - \text{theory}_i}{\text{error}_i}\right)^2$$

is the log-likelihood of a normal.

Handles the complexity of the relationship between parameters and PDFs well.

Conclusion

Formalism for considering frequentist confidence procedures

How to work with complex models?

Practical issues

Ideas for the PDFs and Banff Challenge

Approximating the LFA

Schafer and Stark (2009):
The least favorable alternative is approximated via Monte Carlo simulations

Sample from parameter space Θ, sample from data space under each of these theories

Set up a “matrix game” in which statistician chooses d, and nature chooses π

Approximating the LFA

Goal: Estimate $\mathrm{B}(\pi, d_\pi)$ for fixed π

Estimate Type II Error probabilities: $P_\theta[d_\pi(\eta, X) = 1]$

If $X \sim f_\eta$, then

$$\mathrm{E}\left[\left(\frac{f_\theta(X)}{f_\eta(X)}\right) d_\pi(\eta, X)\right] = P_\theta[d_\pi(\eta, X) = 1]$$

Approximating the LFA

Goal: Estimate $\mathrm{B}(\pi, d_\pi)$ for fixed π

Estimate Type II Error probabilities: $P_\theta[d_\pi(\eta, X) = 1]$

If $X \sim f_\eta$, then

$$\mathrm{E}\left[\left(\frac{f_\theta(X)}{f_\eta(X)}\right) d_\pi(\eta, X)\right] = P_\theta[d_\pi(\eta, X) = 1]$$

and

$$\mathrm{E}\left[\left(\int_\Theta \frac{f_\theta(X)}{f_\eta(X)}\,\pi(d\theta)\right) d_\pi(\eta, X)\right] = \int_\Theta P_\theta[d_\pi(\eta, X) = 1]\,\pi(d\theta)$$

Approximating the LFA

If $X \sim f_\eta$, then

$$\mathrm{E}\left[\left(\int_\Theta \frac{f_\theta(X)}{f_\eta(X)}\,\pi(d\theta)\right) d_\pi(\eta, X)\right] = \int_\Theta P_\theta[d_\pi(\eta, X) = 1]\,\pi(d\theta)$$

but

$$\int_\Theta \frac{f_\theta(X)}{f_\eta(X)}\,\pi(d\theta)$$

is distributed as desired test statistic under the null

Use Monte Carlo to estimate $d_\pi(\eta, \cdot)$

Approximating the LFA

If $X \sim f_\eta$, then

$$\mathrm{E}\left[\left(\int_\Theta \frac{f_\theta(X)}{f_\eta(X)}\,\pi(d\theta)\right) d_\pi(\eta, X)\right] = \int_\Theta P_\theta[d_\pi(\eta, X) = 1]\,\pi(d\theta)$$

but

$$\int_\Theta \left[\int_\Theta P_\theta[d_\pi(\eta, X) = 1]\,\pi(d\theta)\right] \nu(d\eta) = \mathrm{B}(\pi, d_\pi)$$

Another level of MC: randomly choose η to estimate outer integral – This defines $\widehat{\mathrm{B}}(\pi)$

Approximating the LFA

$$\widehat{\mathrm{B}}(\pi) = \sum_k \mathbf{d_k}'\mathbf{A}(\eta_k)\pi$$

"Nature" chooses π and "Statistician" chooses $\mathbf{d_k}$

The (i, j) entry of $\mathbf{A}(\eta_k)$ is

$$\frac{f_{\theta_j}(x_i)}{f_{\eta_k}(x_i)}$$

Approximating the LFA

$$\begin{array}{cc}
 & \begin{pmatrix} \pi_1 & \pi_2 & \cdots & \pi_p \end{pmatrix} \\
\begin{pmatrix} d(\eta, X_1) \\ d(\eta, X_2) \\ \vdots \\ d(\eta, X_n) \end{pmatrix} &
\begin{pmatrix}
\mathrm{LR}[\theta_1/\eta, X_1] & \mathrm{LR}[\theta_2/\eta, X_1] & \cdots & \mathrm{LR}[\theta_p/\eta, X_1] \\
\mathrm{LR}[\theta_1/\eta, X_2] & \mathrm{LR}[\theta_2/\eta, X_2] & \cdots & \mathrm{LR}[\theta_p/\eta, X_2] \\
\vdots & \vdots & \ddots & \vdots \\
\mathrm{LR}[\theta_1/\eta, X_n] & \mathrm{LR}[\theta_2/\eta, X_n] & \cdots & \mathrm{LR}[\theta_p/\eta, X_n]
\end{pmatrix}
\end{array}$$

$$\mathrm{LR}[\theta/\eta, x] \equiv \frac{f_\theta(x)\phi(\theta, \eta)}{f_\eta(x)}$$

Matrix Games

Matrix game characterized by payoff matrix $\mathbf{A}$

Player one chooses row i, player two column j

Player one pays player two $\mathbf{A}(i, j)$

Example: Matching Pennies

$$\mathbf{A} \equiv \begin{array}{cc} & \begin{array}{cc} \mathrm{H} & \mathrm{T} \end{array} \\ \left(\begin{array}{cc} 1 & -1 \\ -1 & 1 \end{array} \right) & \begin{array}{c} \mathrm{H} \\ \mathrm{T} \end{array} \end{array}$$

Optimal strategy is mixed: randomly choose heads or tails.

Matrix Games

Takes the form

$$\widehat{\mathrm{B}}(\pi) = \sum_k \mathbf{d_k}'\mathbf{A}(\eta_k)\pi$$

"Nature" chooses π and "Statistician" chooses $\mathbf{d_k}$

Brown-Robinson algorithm handles statistician's complicated strategy space.