Estimating the Proportion of True Nulls

James Long
jplong@mdanderson.org
Rice STAT 533 / GSBS 1283

April 14, 2020

Announcements

- HW 8: Due April 16 at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Lectures: Today, Thursday, April 21, April 23
- Take home exam (similar format to Exams 1 and 2)
- Lecture Format
  - Slides (plots / analyses in R)
  - .pdf and .R available on course website
- Lecture Structure
  - Microphones are muted when you enter the class.
  - But please ask questions, remember to unmute / mute
  - Let me know about audio issues (chat window or email if I am not responding)

Outline

Using Null Only Region

Parametric Mixture Model: Beta Uniform Mixture

Nonparametric Mixture Model

Data Comparison

Two Group Model

- Hypotheses $H_{01}, \ldots, H_{0N}$
- $\pi_0 =$ proportion of true nulls
- $\pi_1 = 1 - \pi_0 =$ proportion of true alternatives
- y_i is indicator H_{1i} is true
    - $y_i \sim Bernoulli(\pi_1)$
- z_i (or p_i) drawn from distribution:

$$\begin{aligned} &f_0(z) \text{ if } y_i = 0 \text{ (i.e. } H_{0i} \text{ is true)} \\ &f_1(z) \text{ if } y_i = 1 \text{ (i.e. } H_{1i} \text{ is true)} \end{aligned}$$

- The marginal distribution of z_i is

$$f(z) = \pi_0 f_0(z) + \pi_1 f_1(z)$$

π_0 Estimation

Thus far in course, always "estimate" π_0 with 1

- BH Algorithm: Specify q, then algorithm controls FDR at $q\pi_0 \leq q$
- Bayesian Fdr: Recall

$$\mathsf{Fdr}(\mathcal{Z}) = \frac{\pi_0 F_0(\mathcal{Z})}{F(\mathcal{Z})}$$

  estimated with

$$\overline{\mathsf{Fdr}}(\mathcal{Z}) = \frac{F_0(\mathcal{Z})}{\frac{1}{N}\sum_i 1_{z_i \in \mathcal{Z}}}$$

  so we are estimating an upper bound on $\mathsf{Fdr}(\mathcal{Z})$

Result: Replacing π_0 with 1 results in conservative procedures. Simple and good performance when $\pi_0 \approx 1$.

Reasons for estimating π_0

- Adaptive FDR Control:
  - Estimate π_0 with $\widehat{\pi}_0$
  - For FDR control at q, use BH with $q^* = q/\widehat{\pi}_0 > q$
  - $FDR \leq \pi_0 q^* = \pi_0 \frac{q}{\widehat{\pi}_0} \approx q$
  - Since $q^* > q$, cutoff is higher $\implies$ more rejections $\implies$ more power
- π_0 of inherent interest:
  - In gene expression problems comparing controls to cancer tissue, π_0 is the proportion of all genes that are differentially expressed in cancer.
  - Likely very different than the proportion of genes rejected by some FDR procedure. We only reject genes which were are fairly confident are differentially expression (e.g. control FDR at $q = 0.1$).
- Fdr estimates:
  - Can obtained consistent estimate of Fdr$(\mathcal{Z})$.

Outline

Using Null Only Region

Parametric Mixture Model: Beta Uniform Mixture

Nonparametric Mixture Model

Data Comparison

Method Overview

- Assumption: Region $\mathcal{A}_0$ such that

$$f_1(z) = 0 \text{ for } z \in \mathcal{A}_0$$

- Then

$$\begin{aligned} F(\mathcal{A}_0) &= \pi_0 F_0(\mathcal{A}_0) + \pi_1 F_1(\mathcal{A}_0) \\ &= \pi_0 F_0(\mathcal{A}_0) \end{aligned}$$

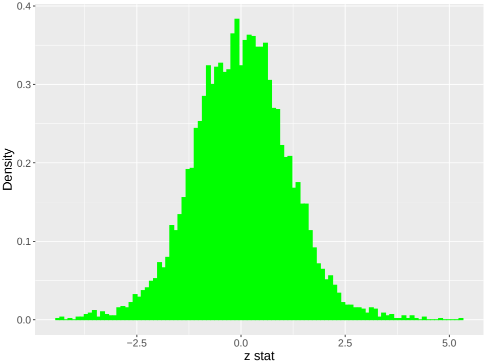
- Use plug–in estimator

$$\widehat{\pi}_0 = \frac{N^{-1} \sum_i 1_{z \in \mathcal{A}_0}}{F_0(\mathcal{A}_0)}$$

- If Assumption true, $\widehat{\pi}_0$ unbiased, asymptotically normal for π_0.
- If Assumption false, $\widehat{\pi}_0$ biased high:

$$\widehat{\pi}_0 \to \frac{\pi_0 F_0(\mathcal{A}_0) + \pi_1 F_1(\mathcal{A}_0)}{F_0(\mathcal{A}_0)} = \pi_0 + \pi_1 \frac{F_1(\mathcal{A}_0)}{F_0(\mathcal{A}_0)}$$

Selecting Region $\mathcal{A}_0$

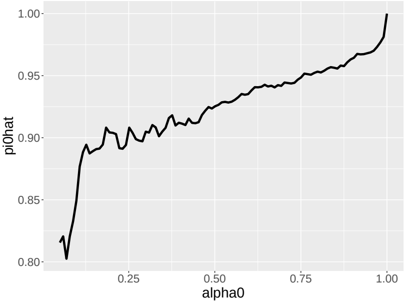


- For $z \sim N(0,1)$ under H_0, z near 0 mostly null because non-nulls should have large absolute test statistics
- Suggests

$$\mathcal{A}_0(\alpha_0) = \left[\Phi^{-1}(0.5 - \alpha_0/2), \Phi^{-1}(0.05 + \alpha_0/2)\right]$$

for some α_0.

Selecting Region $\mathcal{A}_0$



- At each $\mathcal{A}_0$ (alternatively α_0), these are **estimated upper bounds** on π_0.
  - Upper bound

$$\pi_0(\mathcal{A}_0) \equiv \frac{F(\mathcal{A}_0)}{F_0(\mathcal{A}_0)} \leq \pi_0$$

  - Estimated Upper Bound

$$\widehat{\pi}_0(\mathcal{A}_0) \leq \frac{N^{-1}\sum_i 1_{z_i \in \mathcal{A}_0}}{F_0(\mathcal{A}_0)} \; ? \; \pi_0$$

- For small α_0, more uncertainty in estimate.

Uncertainty in Estimate

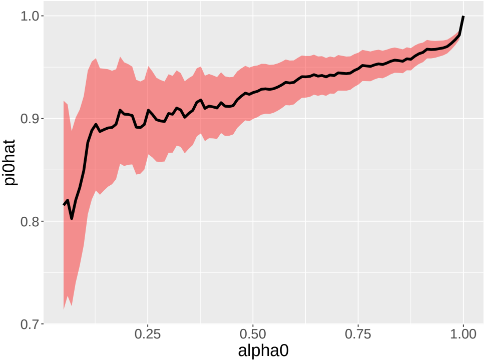
- Estimator asymptotically normal with

$$s.d.(\widehat{\pi}_0(\mathcal{A}_0)) = \frac{\sqrt{F(\mathcal{A}_0)(1 - F(\mathcal{A}_0))}}{\sqrt{N}F_0(\mathcal{A}_0)}$$

- 95% Confidence Interval

$$\widehat{\pi}_0 \pm 2 \times \frac{\sqrt{N^{-2} \sum 1_{z_i \in \mathcal{A}_0}(N - \sum 1_{z_i \in \mathcal{A}_0})}}{\sqrt{N}F_0(\mathcal{A}_0)}$$

Uncertainty in Estimate



- Efron chooses $\alpha_0 = 0.5$, $\mathcal{A}_0 = [-0.67, 0.67]$, $\widehat{\pi}_0 =$ 0.925
- At each α_0 we have an **estimate** of an **upper bound** on π_0.
  - If these were not estimates (black curve actual upper bounds), just take smallest value (lowest upper bound is best)
  - But a lot of uncertainty, especially for $\alpha_0 < 0.2$, so these upper bounds are a bit dangerous to use.

Outline

Using Null Only Region

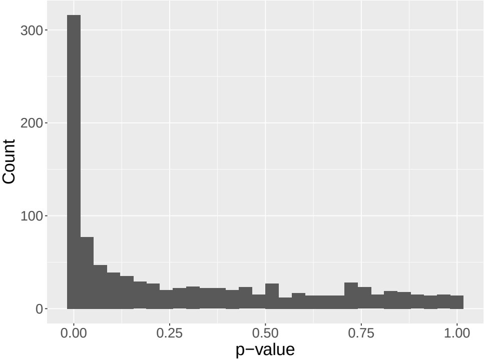
Parametric Mixture Model: Beta Uniform Mixture

Nonparametric Mixture Model

Data Comparison

Kidney Cancer Example

- For each gene, associate expression level with survival time in Cox model
- Obtain ~ 1000 p-values
- **Goal:** Estimate π_0

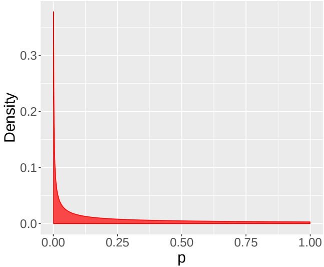


Mixture Model

- p_i are drawn from

$$f(p) = \pi_0 \underbrace{f_0(p)}_{\text{Unif}[0,1]} + (1-\pi_0) \underbrace{f_1(p)}_{\text{unknown}}$$

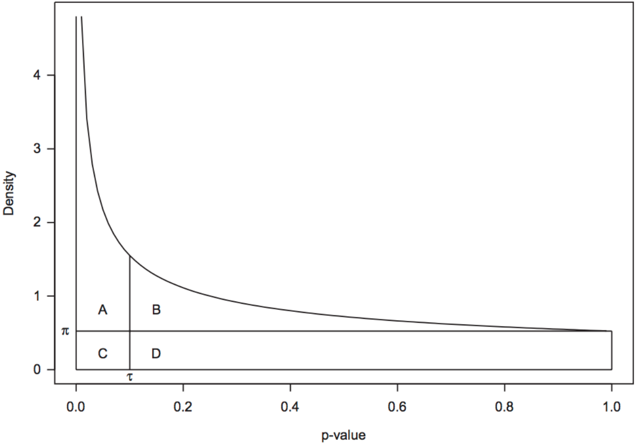
- Choose some parametric model for f_1
- $Beta(\alpha, 1)$ may be reasonable choice



Beta(0.3,1)

Proposed in (Pounds, Stan, and Stephan W Morris. 2003) Bioinformatics.

BUM Model



Parameter Estimation in BUM Model

- Two parameters π_0 and α
- Estimate with maximum likelihood
- Mixture models typically do not have closed form solutions for MLE
- Use quasi-newton (e.g. BFGS) or EM Algorithm
- Obtain uncertainties on π_0 from Fisher information matrix

Outline

Using Null Only Region

Parametric Mixture Model: Beta Uniform Mixture

Nonparametric Mixture Model

Data Comparison

Mixture Model

- z_i (or p_i) are drawn from

$$F(z) = \pi_0 \underbrace{F_0(z)}_{\text{known}} + (1 - \pi_0) \underbrace{F_1(z)}_{\text{unknown}}$$

- Parametric model (such as beta) imposes restrictions on shape of F_1
- Non–parametric estimation of F_1 offers increased flexibility
- Semi–parametric problem: parametric estimation of π_0 and non-parametric estimation of F_1

Identifiability

(π_0, F_1) are not jointly identifiable given sample $z_1, \ldots, z_N \sim F$

- Suppose (π_0', F_1') the true value of the parameters

$$F(z) = \pi_0' F_0(z) + (1 - \pi_0') F_1'(z)$$

- Setting $(\pi_0 = 0, F_1 = F)$ will generate same data. Interpretation: There are no true nulls and the observed test statistic distribution is entirely generated by true alternatives.
- More generally let $\pi_0^* < \pi_0'$ and define

$$F_1^*(z) = \frac{F(z) - \pi_0^* F_0(z)}{1 - \pi_0^*}$$

Then F_1^* is a valid cdf and

$$\pi_0' F_0(z) + (1 - \pi_0') F_1'(z) =_d \pi_0^* F_0(z) + (1 - \pi_0^*) F_1^*(z)$$

"Estimation of a two-component mixture model with applications to multiple testing." Patra and Sen. JRSSB 2016

Identifiability

- Instead of estimating π_0, estimate:

$$\pi_0' = \max_{\pi_0 \in [0,1]} \{\pi_0 : \frac{F(z) - \pi_0 F_0(z)}{1 - \pi_0} \text{ is a valid c.d.f. } \}$$

- π_0' is the largest component of F_0 which can be removed from F while still producing a valid c.d.f.

$$F_1'(z) = \frac{F(z) - \pi_0' F_0(z)}{1 - \pi_0'}$$

Outline of Estimation Strategy

- $\widehat{F}$ is empirical c.d.f. of $z_1, \ldots, z_n$
- Define

$$\widehat{F}_{1,\pi_0}(z) = \frac{\widehat{F}(z) - \pi_0 F_0(z)}{1 - \pi_0}$$

  Note: $\widehat{F}_{1,\pi_0}(z)$ may not be c.d.f.

- Find closest c.d.f. to $\widehat{F}_{1,\pi_0}(z)$ via **isotonic regression**

$$\check{F}_{1,\pi_0}(z) = \underset{\text{c.d.f. } W}{\operatorname{argmin}} \int_z (W(z) - \widehat{F}_{1,\pi_0}(z))^2 d\widehat{F}_{1,\pi_0}(z)$$

- Measure distance:

$$\gamma(\pi_0) = d(\check{F}_{1,\pi_0}(z), \widehat{F}_{1,\pi_0}(z)) = \int_z (\check{F}_{1,\pi_0}(z) - \widehat{F}_{1,\pi_0}(z))^2 \check{F}_{1,\pi_0}(z)$$

- Select largest π_0 such that $\gamma(\pi_0)$ is small
  - Suggested strategy $\widehat{\pi}_0 = \underset{\pi_0}{\operatorname{argmax}} \, \gamma''(\pi_0)$

"Estimation of a two-component mixture model with applications to multiple testing." Patra and Sen. JRSSB 2016

Outline

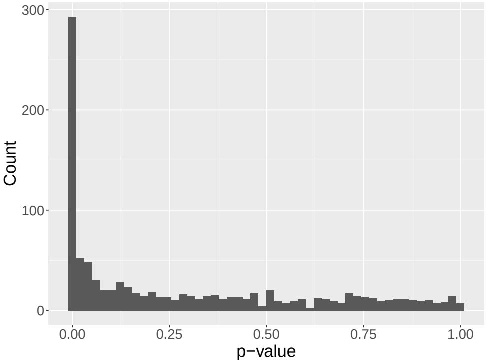
Using Null Only Region

Parametric Mixture Model: Beta Uniform Mixture

Nonparametric Mixture Model

Data Comparison

Data

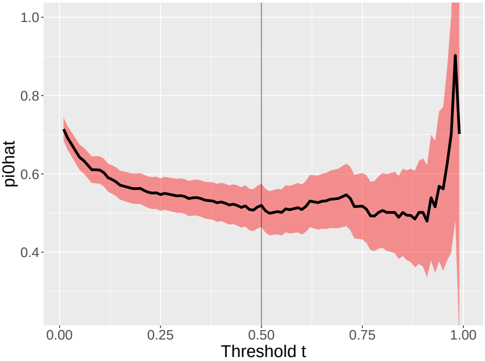


1. Null Only Region
2. BUM
3. Nonparametric Mixture Model of Patra–Sen

Null Only Method

Define $\mathcal{A}_0 = (t, 1]$. Then:

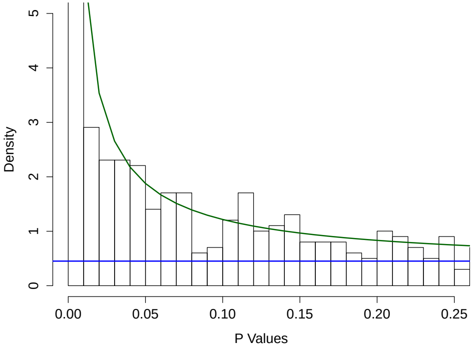
$$\widehat{\pi}_0(\mathcal{A}_0) = \frac{N^{-1}\sum_i 1_{p \in \mathcal{A}_0}}{F_0(\mathcal{A}_0)} = \frac{N^{-1}\sum_i 1_{p_i > t}}{1 - t}$$



Choose $t = 0.5$. $\widehat{\pi}_0 = 0.52$

Bayesian Uniform Mixture

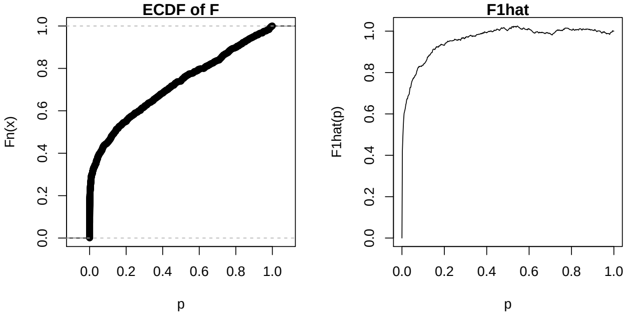
```
> library(ClassComparison)
> out <- Bum(ps)
> par(mar=c(5,5,1,1))
> hist(out,xlim=c(0,0.25),ylim=c(0,5),
+       cex.lab=1.3,cex.axis=1.3)
```



$$\widehat{\pi}_0 = 0.452$$

Nonparametric Mixture via Isotonic Regression

$$F(p) = \pi_0 F_0(p) + (1 - \pi_0) F_1(p)$$

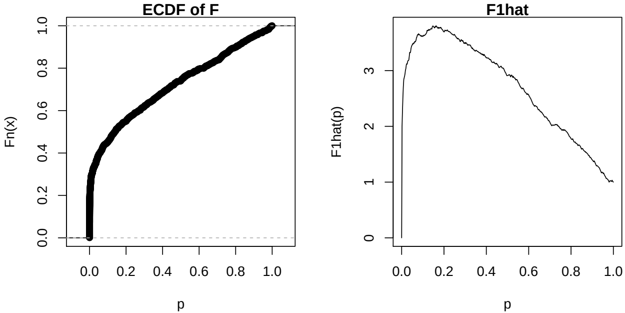


- Left plot: Empirical cdf of p-values
- $\widehat{\pi}_0 = 0.522$
- Right plot:

$$\widehat{F}_1(p) = \frac{\widehat{F}(p) - \widehat{\pi}_0 F_0(p)}{1 - \widehat{\pi}_0}$$

Nearly (up to sampling error) non-decreasing. Looks like a cdf.

Consider $\widehat{\pi}_0 = 0.9$



- Left plot: Empirical cdf of p-values
- $\widehat{\pi}_0 = 0.9$
- Right plot:

$$\widehat{F}_1(p) = \frac{\widehat{F}(p) - \widehat{\pi}_0 F_0(p)}{1 - \widehat{\pi}_0}$$

Does not look at all like cdf! $\widehat{\pi}_1$ Estimate too large.

Review / Summary / Further Directions

- Nonparametric model of Patra / Sen can produce confidence intervals for π_0
- Many other π_0 estimation methods:
  - "Adaptive linear step-up procedures that control the false discovery rate" Biometrika. Benjamini et al 2006
  - "Estimating the proportion of true null hypotheses, with application to DNA microarray data. JRSSB. Langaas et al 2005
  - "A direct approach to false discovery rates." JRSSB. Storey 2002
  - . . . .
- Efron is somewhat skeptical of putting a lot of effort into π_0 estimation: "The exact choice of $\widehat{\pi}_0$ is not crucial. A much more crucial and difficult issue is the appropriate choice of the null density f_0."
  - "It is inappropriate to be concerned about mice when there are tigers abroad." - George Box
- **Thursday:** Local Fdr, Sections 5.1 and 5.2 in Efron