Local False Discovery Rate

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

April 16, 2020

Announcements

- HW 8: Due today at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- HW 9: Due April 23 at 5:00pm, email TA Scott Liang at ricestat533@gmail.com
- Lectures: Today, 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

Local False Discovery Rate (fdr)

Local fdr with Mixture Models

Fdr versus FWER Scaling

Outline

Local False Discovery Rate (fdr)

Local fdr with Mixture Models

Fdr versus FWER Scaling

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:

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

- The marginal distribution of z_i is

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

Local Fdr

- The local Fdr is

$$\text{fdr}(z) \equiv P(y = 0|z) = \frac{\pi_0 f_0(z)}{f(z)}$$

- It is "local" because reports false discovery rate at single point, rather than over region $\mathcal{Z}$.
- Uses of FDR, Fdr, fdr
  - FDR: Report set of p-values and associated FDR q
  - Fdr: Report tests in set $\mathcal{Z}$ and associated $\overline{\text{Fdr}}$
  - With fdr (local false discovery rate), report $\text{fdr}(z)$ for each hypothesis
    - More specifically report estimate $\widehat{\text{fdr}}(z)$

Estimation of fdr

$$\mathsf{fdr}(z) = \frac{\pi_0 f_0(z)}{f(z)}$$

- Need estimates of π_0 and $f(z)$
- Discussed estimation of π_0 in last lecture
- **Now:** Discuss estimation of $f(z)$
  - Sample $z_1, \ldots, z_n \sim f$
  - So this is a density estimation problem

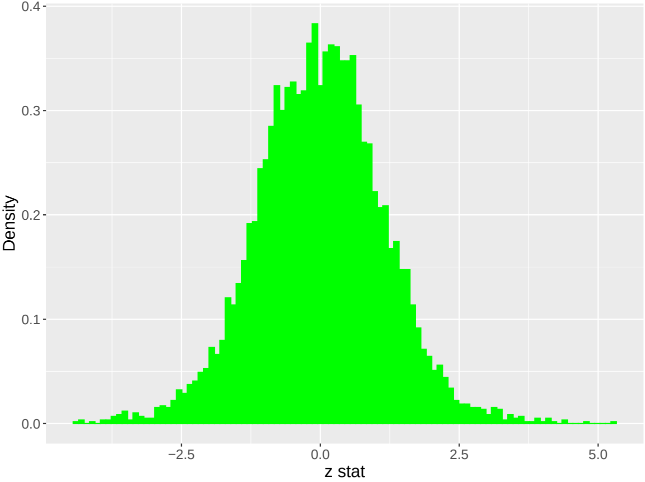
Kernel Density Estimation

- Kernel density estimate

$$\widehat{f}(z) = \frac{1}{hN} \sum_{i=1}^{N} K\left(\frac{z - z_i}{h}\right)$$

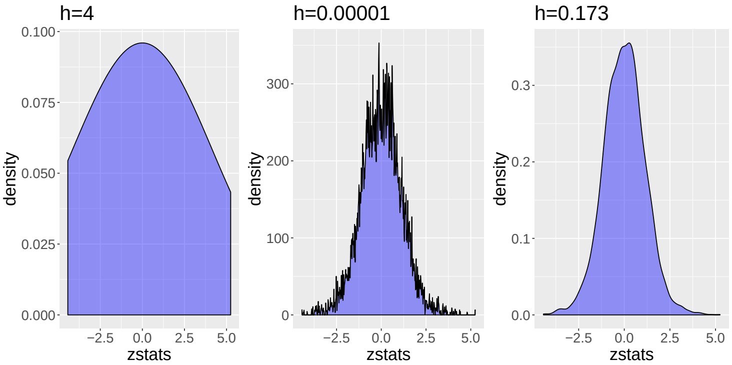
- K is the kernel function (often standard normal density)
- h is the bandwidth, controls how smooth density estimate is
- Usually: h estimated from the data to obtain appropriately smooth estimate
  - If h is very large $K\left(\frac{z-z_i}{h}\right) \approx 1/\sqrt{2\pi}$ for z in range of z_i. Then density will be constant over range of z_i
  - If h is very small $K\left(\frac{z-z_i}{h}\right) \approx 0$ at $z \neq z_i$. So density estimate will be point masses at z_i.

Prostate Data



Histogram of the prostate cancer z statistics.

Prostate KDE Estimate of f

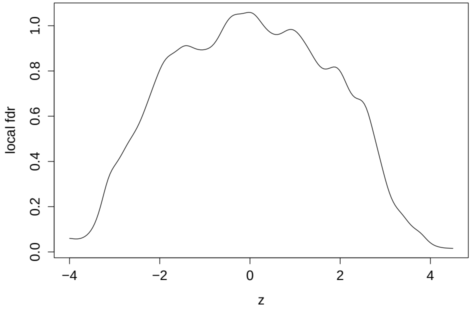


Left) Bandwidth too large Center) Bandwidth too small Right) Reasonable bandwidth.

Prostate Local fdr with KDE

Using $h = 0.173$ compute:

$$\widehat{fdr}(z) = \frac{\widehat{\pi} f_0(z)}{\widehat{f}(z)}$$



- Probably too wiggly.
- Could try increasing bandwidth
- Or use a different density estimation method.

Flexible MLE Density Estimation

- $f(z) = e^{\sum_{j=0}^{J} \beta_j z^j}$, J controls flexibility of model
- $f(z) > 0$
- β_0 chosen to normalize density

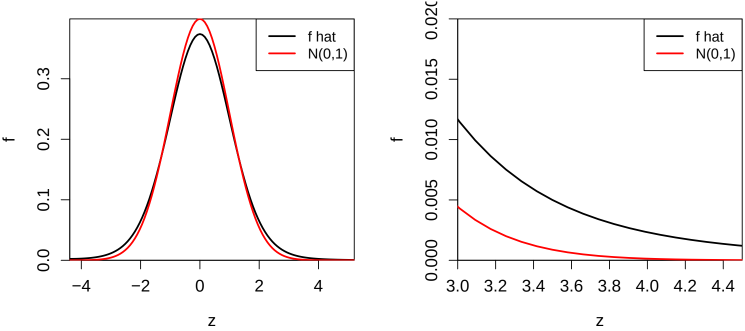
$$\beta_0 = -\log \int_{-\infty}^{\infty} e^{\sum_{j=1}^{J} \beta_j z^j} dz$$

- Estimate $\beta_1, \ldots, \beta_J$ via MLE

$$\widehat{\beta} = \underset{\beta}{\operatorname{argmax}} \prod_{i=1}^{N} e^{\sum_{j=0}^{J} \beta_j z_i^j}$$

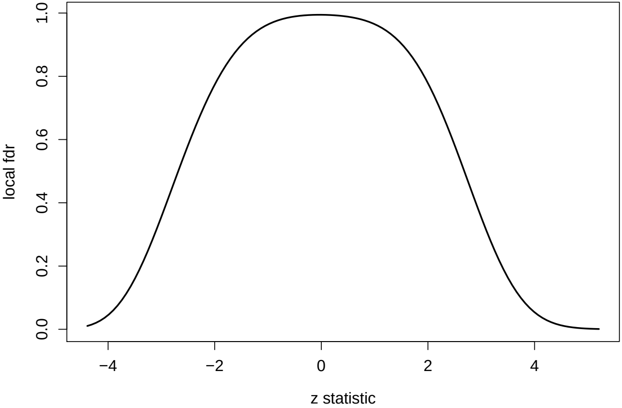
  - Efron approximates MLE using Poisson regression
  - Partition space of test statistics into equal width bins $\mathcal{Z} = \cup_{k=1}^{K} \mathcal{Z}_k$
  - $x_k =$ center of bin $\mathcal{Z}_k$
  - $y_k = \sum 1_{z_i \in \mathcal{Z}_k}$
  - $y_k \sim_{iid}$ Poisson(ν_k) where $\log(\nu_k) = \sum_{j=0}^{J} \beta_j x_k^j$

Prostate $\widehat{f}$ with Poisson Regression $J = 7$



$$\widehat{\mathsf{fdr}}(z) = \frac{\widehat{\pi}_0 \phi(z)}{\widehat{f}(z)}$$

Prostate Local fdr with Poisson Regression



Local fdr fairly symmetric about 0.

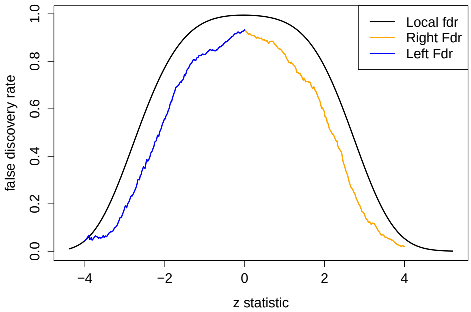
Local fdr versus Fdr

- Let $\mathcal{Z}_R = (z, \infty)$ and

$$\overline{\mathsf{Fdr}}_R(\mathcal{Z}_R) = \frac{\widehat{\pi}_0(1-\Phi(z))}{N^{-1}\sum_{i=1}^N 1_{z_i>z}}$$

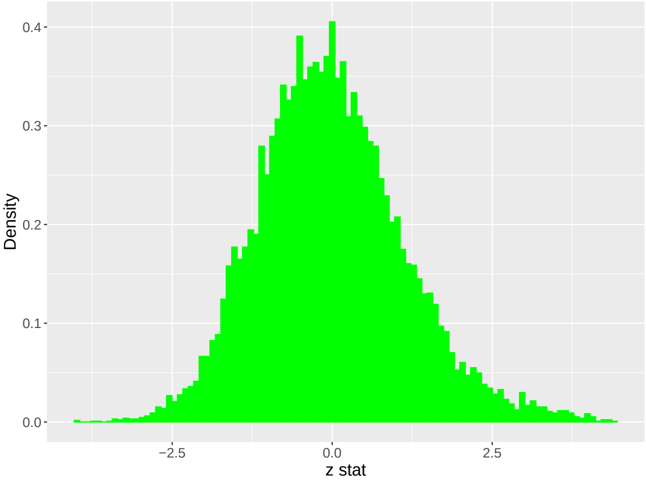
- Let $\mathcal{Z}_L = (-\infty, z)$ and

$$\overline{\mathsf{Fdr}}_L(\mathcal{Z}_R) = \frac{\widehat{\pi}_0\Phi(z)}{N^{-1}\sum_{i=1}^N 1_{z_i<z}}$$



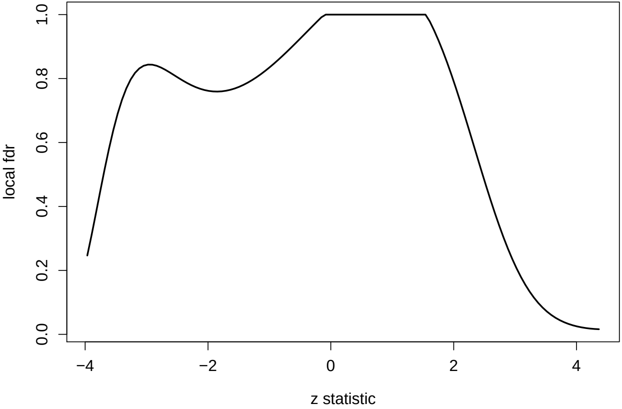
Note that at given z, Fdr always less than local fdr.

DTI Local fdr with Poisson Regression



DTI z–statistics contain substantial asymmetry. More signal on the right.

DTI Local fdr with Poisson Regression

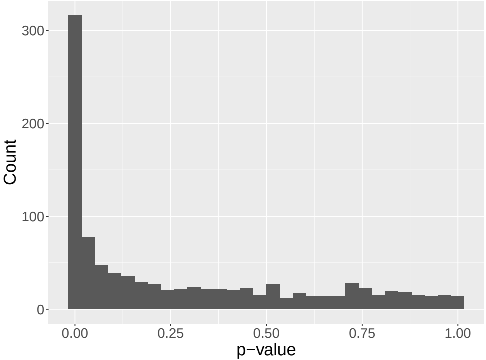


Almost no signal on the left.

Outline

Kidney Cancer p–values

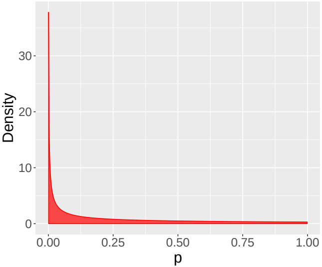
- For each gene, associate expression level with survival time in Cox model
- Obtain ~ 1000 p-values
- **Goal:** Estimate local fdr at each p–value



Mixture Model

p_i are drawn from

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



Beta(0.3,1)

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

π_0 Estimate

- Recall Beta$(\alpha, 1)$ density is:

$$f(p|\alpha) = \frac{p^{\alpha-1}}{B(\alpha, 1)}$$

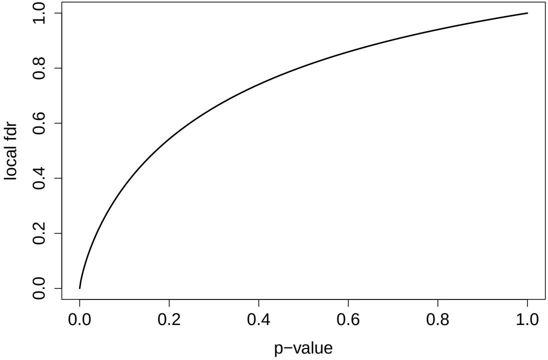
- Since $\alpha < 1$ for modeling p–value distributions, $f(p|\alpha)$ decreasing in p
- $f(1|\alpha) = \alpha$
- So there is an additional α uniform component which can be removed from Beta$(\alpha, 1)$
- So can define:

$$\pi_0 = L + (1 - L)\alpha$$

- Assumption: p-value density under H_a is 0 at $p = 1$

BUM Model Local fdr

- $\widehat{L}$ and $\widehat{\alpha}$ are MLEs of L and α
- $\widehat{\pi}_0 = \widehat{L} + (1 - \widehat{L})\widehat{\alpha}$
- $$\widehat{\mathsf{fdr}}(p) = \frac{\widehat{\pi}_0 f_0(p)}{\widehat{f}(p)} = \frac{\widehat{\pi}_0}{\widehat{L} + (1 - \widehat{L})\frac{p^{\widehat{\alpha}-1}}{B(\widehat{\alpha},1)}}$$



Outline

Fdr/fdr Asymptotics in N

Question: Assuming two group model, as N increases, how do inferences for hypothesis H_{0i} with z-statistic z change?

- Local fdr:

$$\widehat{\text{fdr}}(z) = \frac{\widehat{\pi}_0 f_0(z)}{\widehat{f}(z)}$$

  As N increases, variance of estimates $\widehat{\pi}_0$ and $\widehat{f}$ decrease. But should not dramatically change $\widehat{\text{fdr}}(z)$ (supposing original N reasonably large).

- Fdr: Bayesian False Discovery Rate of region $\mathcal{Z}$ is

$$\overline{\text{Fdr}}(\mathcal{Z}) = \frac{\widehat{\pi}_0 F_0(\mathcal{Z})}{\widehat{F}(\mathcal{Z})}$$

  If $z \in \mathcal{Z}$ will continue (as N increases) to reject H_{0i} and $\overline{\text{Fdr}}(\mathcal{Z})$ will converge to $\text{Fdr}(\mathcal{Z})$

Message: Assuming two group model, do not pay a penalty for larger N for Fdr, local fdr, and FDR. In fact, larger N helpful because estimators have smaller variance.

Bonferroni Asymptotics in N

- $p = 1 - F_0(z)$ (right sided p-value)
- Bonferroni rejects if
$$p < \frac{\alpha}{N}$$
- So increasing N may change rejection decision.
- Rejection threshold converging to 0 rather than any fixed quantity.
- Similar story for Holm / Hochberg (exercise 5.6 in textbook)
- This is general problem with FWER control procedures.

Two Group Model Violations

Question: If Fdr/fdr/FDR do not pay price for larger N (in fact estimators have smaller variance), why not just throw all possible hypotheses together from all sorts of experiments?

- Result: Violation of Assumptions of 2-group model
- Example:
  - $N_1 = 1000$ gene panel of genes thought to be associated with cancer
    - Two group model parameters: π_{01}, f_{11}
  - About $N_2 = 20000$ genes in second panel, not known to be associated with cancer
    - Two group model parameters: π_{02}, f_{12}
  - Very likely $\pi_{02} > \pi_{01}$ (more true nulls in second panel) and f_{12} more concentrated near 0 than f_{11} (smaller effect sizes in second panel)
  - So merging these two data sets will result in larger local fdr at given z and higher Fdr for set $\mathcal{Z}$ than analyzing only the first set

Summary / Preview

- Local fdr is the probability the null is true given the test statistic (or p–value).
- In practice, can combine FDR with local fdr
  - Report all hypotheses with FDR < 0.1
  - For these hypotheses, report $\widehat{\text{fdr}}$
- Thus far we have assumed null distribution f_0 is known
  - When testing 1000s of hypotheses, can estimate f_0 from distribution of test statistics
  - Chapter 6 in Efron, cover on Tuesday