Bayes and Empirical Bayes Information (Learning from the experience of others)

Bradley Efron

Stanford University

A Jar of Biased Coins

- Coin A 45 flips, 18 heads: $\hat{p}_A = 18/45$ $(= .400)$
- Other coins $\hat{p}_B,\ \hat{p}_C,\ \hat{p}_D, \ldots$
- Coins are different and all flips independent
- Question Can $\hat{p}_B,\ \hat{p}_C,\ \hat{p}_D, \ldots$ help you estimate p_A?
- Answer Yes!
  (*Empirical Bayes* circa 1950: Robbins, Stein, Good, Turing)

Bayes Theorem

- *Prior density* $g(\mu)$ for unknown parameter μ
- *Sampling density* $f_\mu(x)$ for observed data x

Bayes Rule $\boxed{g(\mu|x) = g(\mu)f_\mu(x)/f(x)}$

where $f(x) = \int g(\mu)f_\mu(x)\, d\mu$, the *marginal density* of x

- Empirical Bayes Try to learn $g(\mu)$ from "other" data

The 18 Baseball Players

	Player	hits/AB	$\hat{p}$ ("MLE")	$\tilde{p}$ ("James–Stein")	$\boldsymbol{p}$ ("Truth")
1.	Clemente	18/45	.400	.290	.346
2.	Robinson	17/45	.378	.286	.298
3.	Howard	16/45	.356	.281	.276
⋮	⋮	⋮	⋮	⋮	⋮
17.	Munson	8/45	.178	.244	.316
18.	Alvis	7/45	.156	.239	.200
	Grand Average		*.265*	*.265*	*.265*

Total Prediction Error

- *MLE* $\sum_{1}^{18} (p_i - \hat{p}_i)^2 = 0.075$
- James–Stein $\sum_{1}^{18} (p_i - \tilde{p}_i)^2 = 0.021$
- Ratio 3.5
- No fluke!

The James–Stein Rule (1961)

- *Observe* $x_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_0^2)$ for $i = 1, 2, \ldots, N \geq 4$,

  i.e., independent normals, with $f_{\mu_i}(x_i) = \left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right) \exp\left\{-\frac{1}{2}\left(\frac{x_i - \mu_i}{\sigma_0}\right)^2\right\}$

- James–Stein $\bar{x} = \dfrac{\sum x_i}{N}$ $\hat{V} = \dfrac{\sum (x_i - \bar{x})^2}{N - 3}$

- *MLE* $\hat{\mu}_i = x_i$ • *JS* $\tilde{\mu}_i = \bar{x} + \left[1 - \sigma_0^2 / \hat{V}\right](x_i - \bar{x})$

Theorem *JS always beats MLE in expected squared error:*

$$E\left\{\sum (\mu_i - \tilde{\mu}_i)^2\right\} < E\left\{\sum (\mu_i - \hat{\mu}_i)^2\right\}$$

Bayesian Justification

- If *a priori* $\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(M, A)$ and $x_i|\mu_i \overset{\text{ind}}{\sim} \mathcal{N}(\mu_i, \sigma_0^2)$:
  then $x_i \sim \mathcal{N}(M, V)$ with $V = A + \sigma_0^2$ and

$$E\{\mu_i|x_i\} = M + \left[1 - \sigma_0^2/V\right](x_i - M)$$

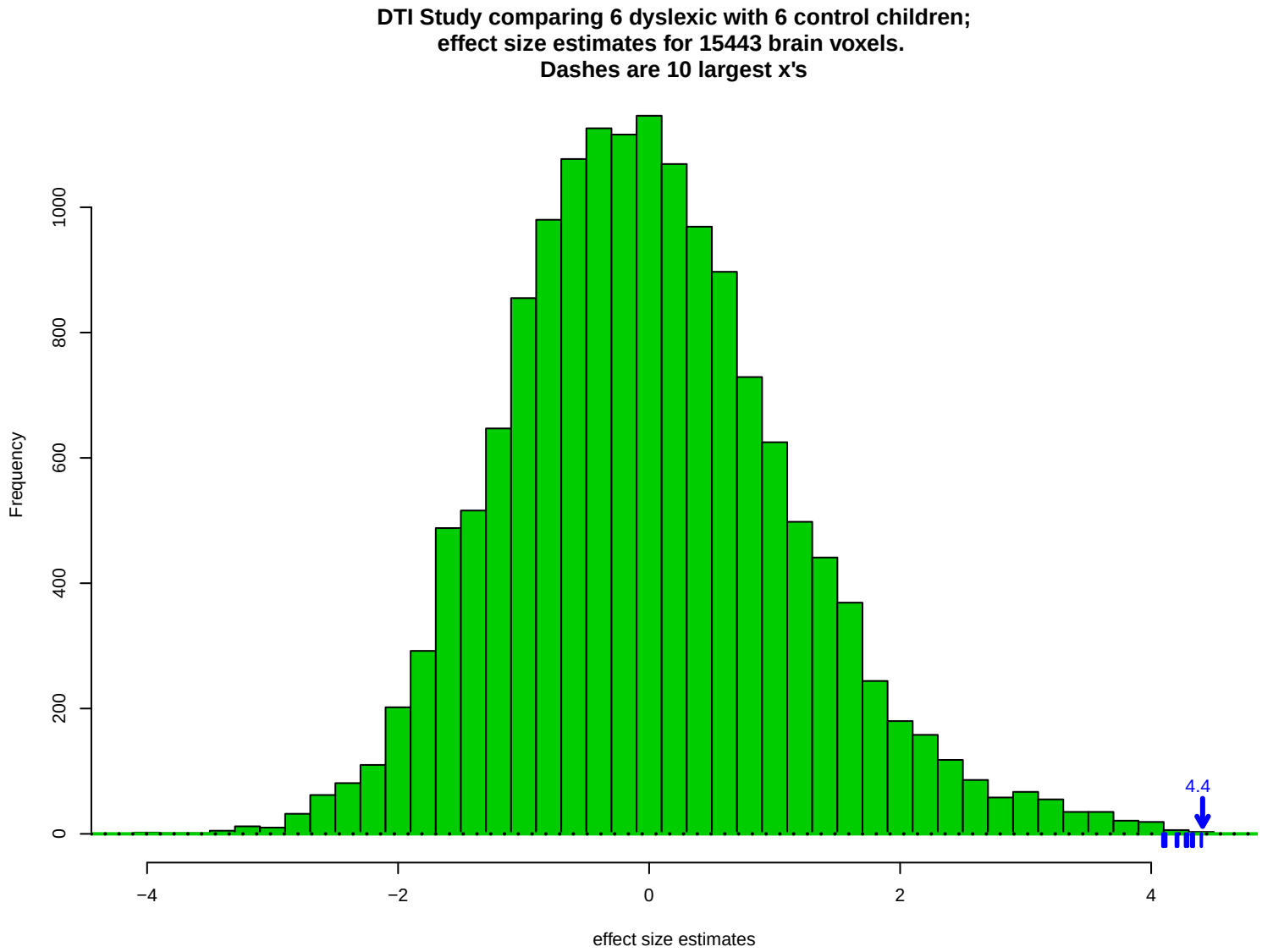
- James–Stein: $\bar{x}$ unbiased for M, and $1/\hat{V}$ unbiased for $1/V$
- "Empirical Bayes"

A Diffusion Tensor Imaging Study

- 6 dyslexic children and 6 normal controls
- Each measured at $N = 15443$ brain locations (*voxels*)
- x_i compares dyslexics with controls at voxel i:

$$x_i \sim \mathcal{N}\left(\mu_i, \sigma_0^2\right)$$

- μ_i the "effect size"
- $\sigma_0 = 1.06$
- x_i's range from -4.0 to 4.4


DTI Study comparing 6 dyslexic with 6 control children;
effect size estimates for 15443 brain voxels.
Dashes are 10 largest x's
Frequency
0
200
400
600
800
1000
4.4
−4
−2
0
2
4
effect size estimates

The Winner's Curse
("Selection bias," "Regression to the mean")

- *10 largest x_i's* $4.40, 4.32, 4.30, \ldots, 4.09$
- Question What are the 10 corresponding effect sizes μ_i?
- Curse μ_i's usually (much) smaller than x_i's
- Bayes estimates are immune to Winner's Curse
- Empirical Bayes?

Tweedie's Formula (1956)

- *Bayes model* $\mu \sim g(\cdot)$ *a priori* and $x|\mu \sim \mathcal{N}\left(\mu, \sigma_0^2\right)$
- *Marginal density* $f(x) = \int_{-\infty}^{\infty} f_0(x-\mu) g(\mu)\, d\mu$

  with $f_0(x)$ the $\mathcal{N}(0, \sigma_0^2)$ density
- Tweedie's Formula

$$\boxed{E\{\mu|x\} = x + \sigma_0^2 l'(x)}$$

  - $l'(x) = \dfrac{d}{dx} \log f(x)$

$\uparrow$ Est. effect $\quad \uparrow$ MLE $\quad \nwarrow$ Bayes correction

- *Advantage* works directly with f, not g

Empirical Bayes Implementation

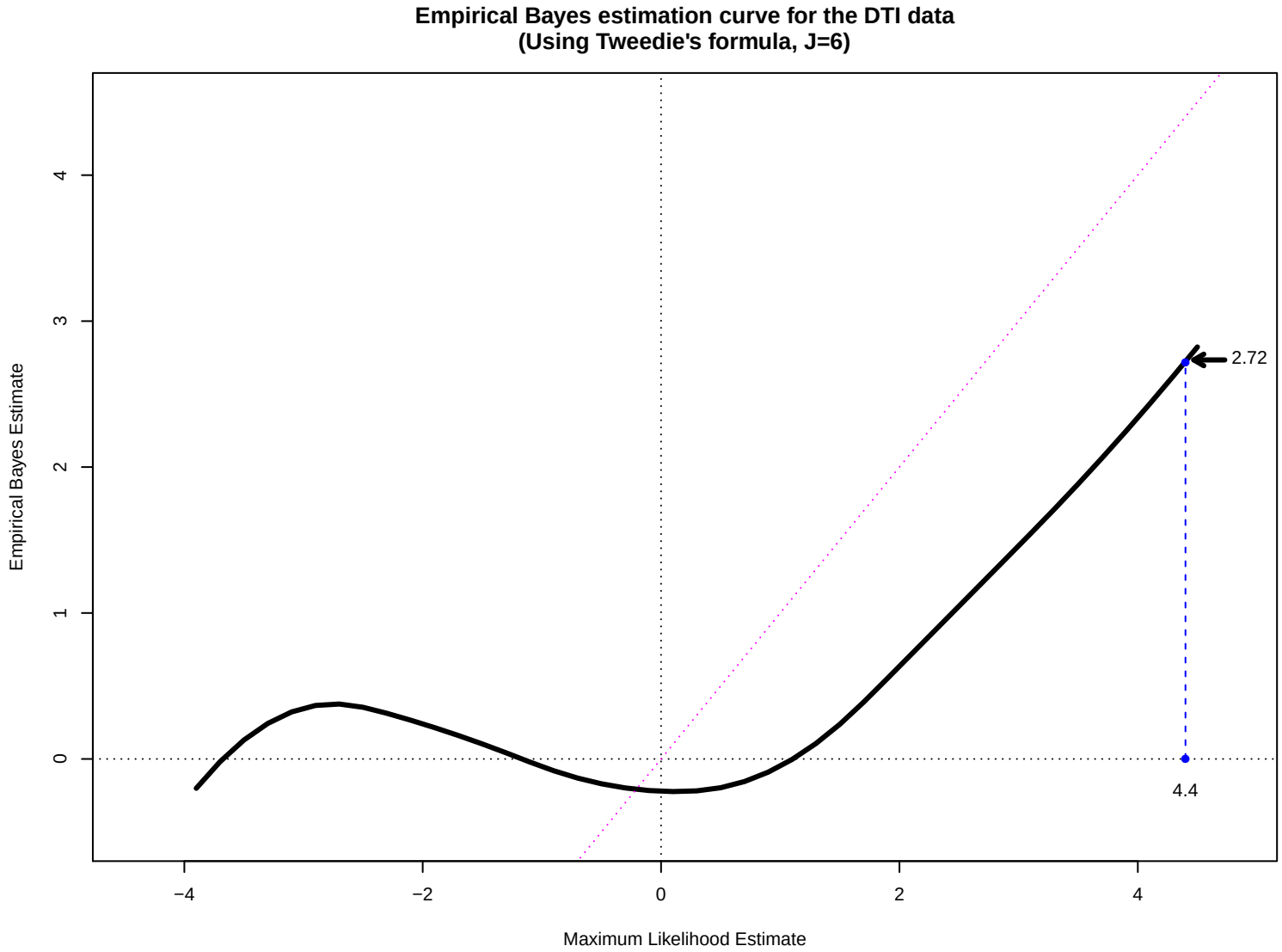
- We don't know prior $g(\mu)$ or mixture $f(x)$ but can estimate $\hat{f}(x)$ and $\hat{l} = \log \hat{f}(x)$ from histogram of all N x's:

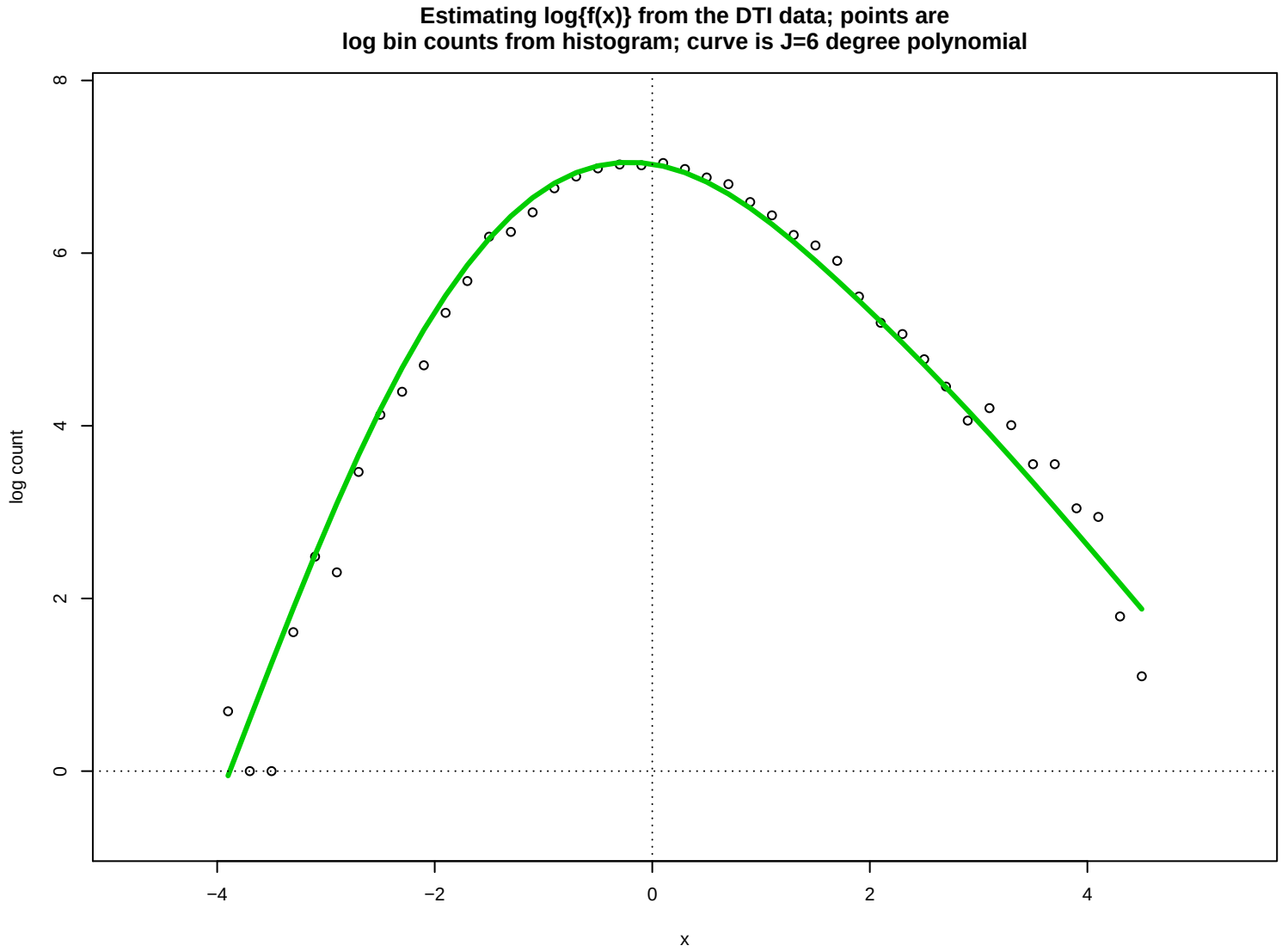
$$\hat{E}\{\mu_i | x_i\} = x_i + \sigma_0^2 \hat{l}'(x_i)$$

- *Parametric model* $\log f(x) = \sum_{j=0}^{J} \beta_j x^j$ [$J = 2$: normal]

- *MLE* $\hat{l}(x) = \sum_{j=0}^{J} \hat{\beta}_j x^j$ so $\hat{l}'(x) = \sum_{j=1}^{J} j\hat{\beta}_j x^{j-1}$

  [$J = 2$: James–Stein]


Empirical Bayes estimation curve for the DTI data
(Using Tweedie's formula, J=6)
Empirical Bayes Estimate
Maximum Likelihood Estimate
2.72
4.4
−4
−2
0
2
4
0
1
2
3
4


Estimating log{f(x)} from the DTI data; points are
log bin counts from histogram; curve is J=6 degree polynomial
log count
x
0
2
4
6
8
−4
−2
0
2
4

Does "Tweedie" Lift the Winner's Curse?

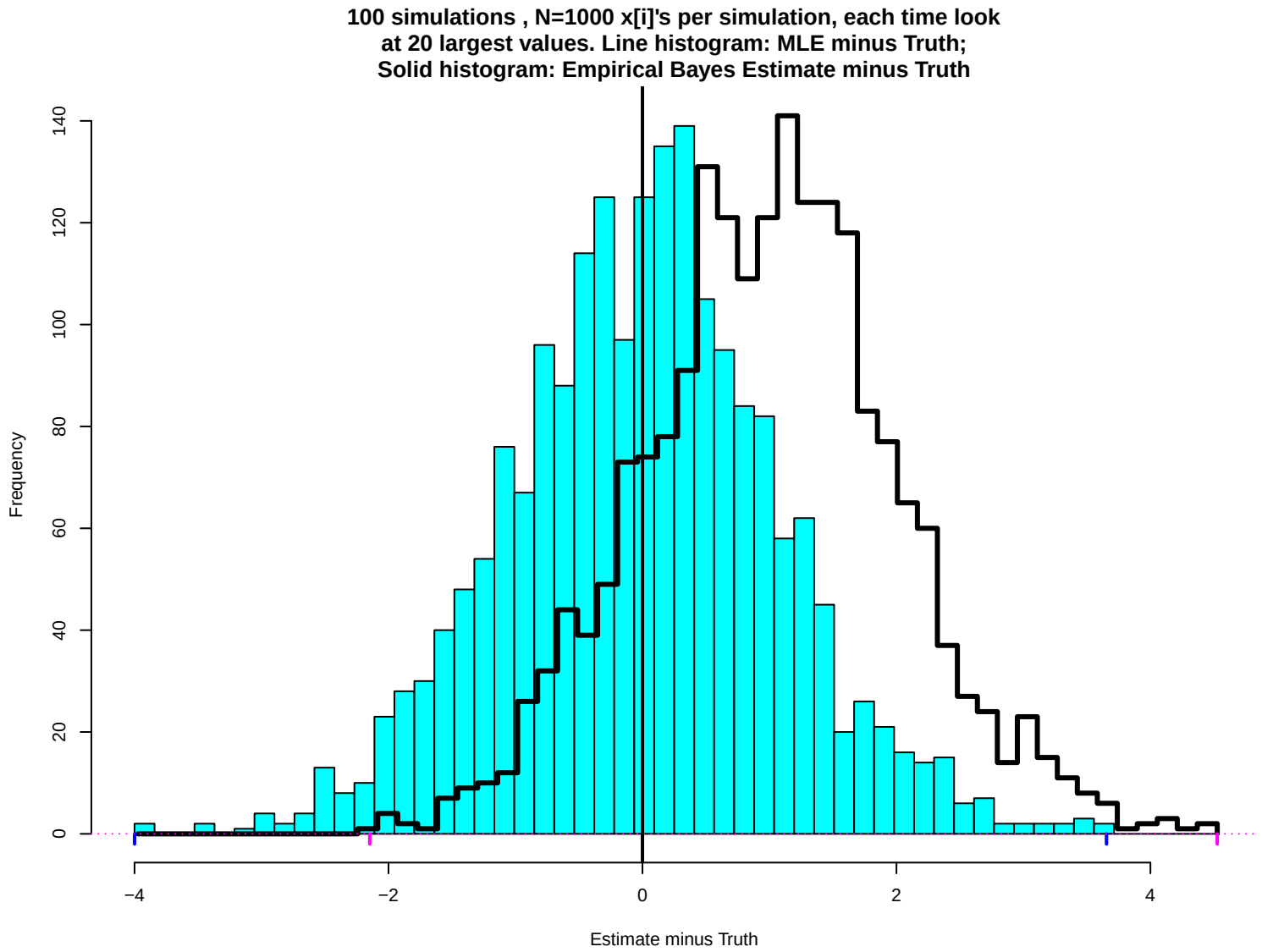
- *Simulations* prior $g(\mu) = e^{-\mu}$ $[\mu > 0]$ and $x|\mu \sim \mathcal{N}(\mu, 1)$:

$$\{(\mu_i, x_i),\ i = 1, 2, \ldots, N = 1000\}$$

- Compute Tweedie estimates $\{\tilde{\mu}_i\}$
- Repeat 100 times, each time comparing

$$x_i - \mu_i \quad \text{with} \quad \tilde{\mu}_i - \mu_i$$

for 20 largest x_i values


100 simulations , N=1000 x[i]'s per simulation, each time look
at 20 largest values. Line histogram: MLE minus Truth;
Solid histogram: Empirical Bayes Estimate minus Truth
Frequency
0
20
40
60
80
100
120
140
−4
−2
0
2
4
Estimate minus Truth

Empirical Bayes Regret

- For a given observed value x_0:
  true Bayes estimate $\mu^{\dagger} = x_0 + l'(x_0)$
  must be more accurate than EB estimate $\tilde{\mu} = x_0 + \hat{l}'_{\boldsymbol{x}}(x_0)$
- *Regret*(x_0) $E\left\{(\mu - \tilde{\mu})^2 - \left(\mu - \mu^{\dagger}\right)^2 \middle| x_0\right\}$

$$= E\left[\hat{l}'_{\boldsymbol{x}}(x_0) - l'(x_0)\right]^2$$

- *Asymptotically* $\mathrm{Reg}(x_0) \doteq \dfrac{c(x_0)}{N}$ $\left[\hat{l}'_{\boldsymbol{x}}\right.$ method determines $\left.c(x_0)\right]$

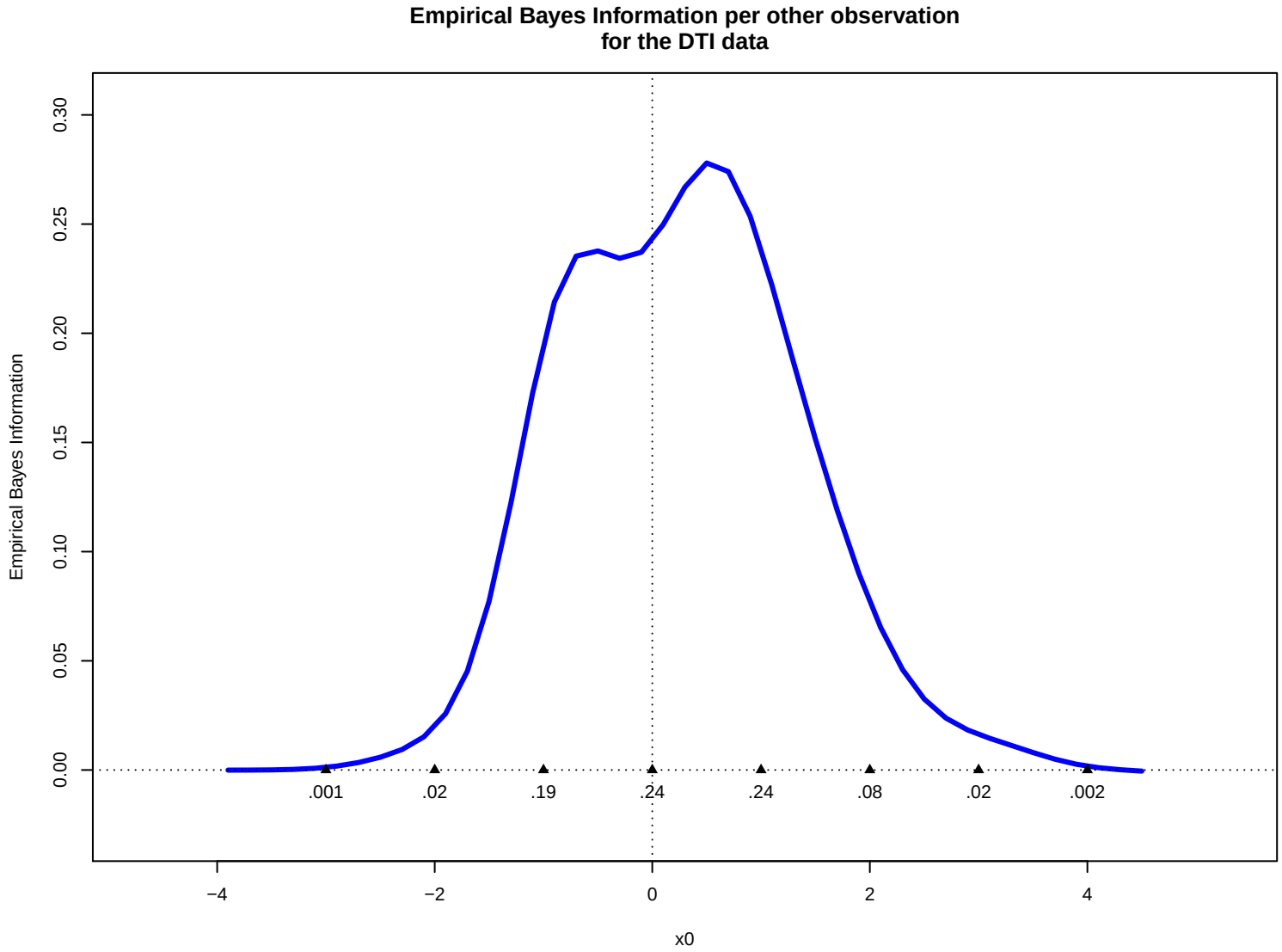
Empirical Bayes Information

- $N = 1 : \hat{\mu}_i = x_i$ (MLE)
- $N = \infty : \mu_i^{\dagger} = x_i - l'(x_i)$ (Bayes)
- $N : \tilde{\mu}_i = x_i - \hat{l}'_{\boldsymbol{x}}(x_i)$ (EB)

- Empirical Bayes information $\boxed{\mathcal{I}(x_0) = \dfrac{1}{c(x_0)}}$

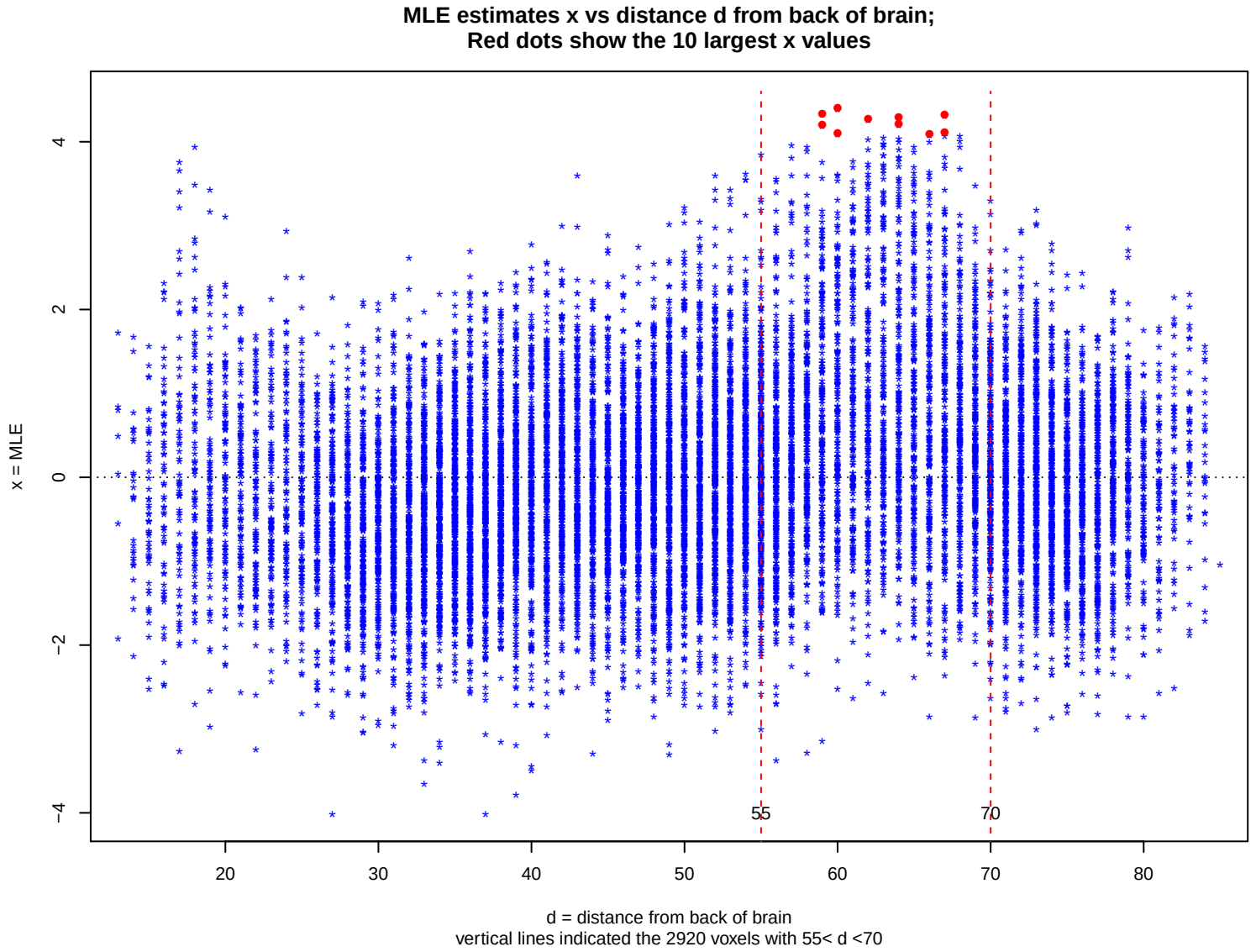
  (amount of info for estimating $E\{\mu|x_0\}$ per other observation)

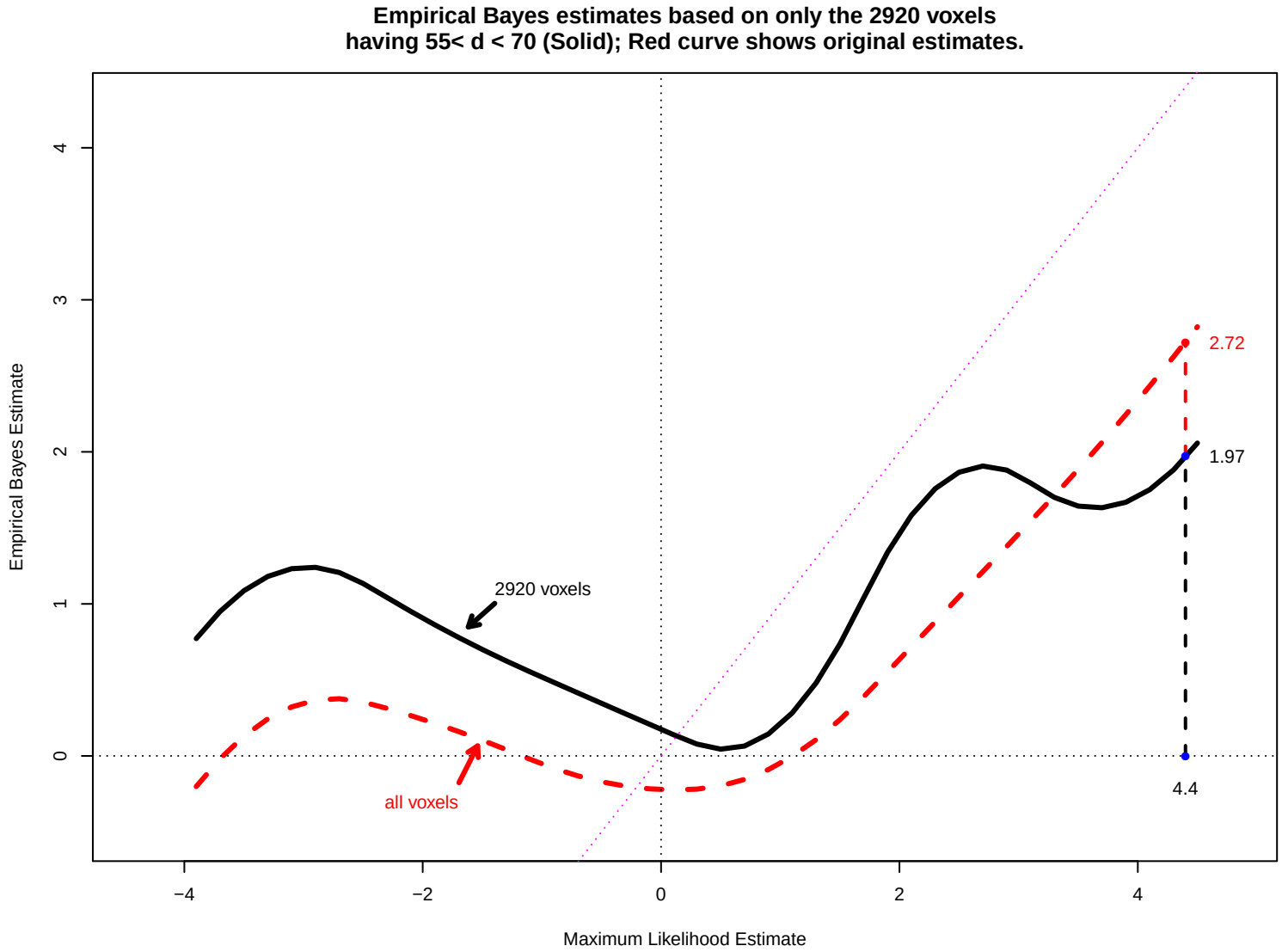
- *Bayes regret* $\mathrm{Reg}(x_0) \doteq \dfrac{1}{N\mathcal{I}(x_0)}$


Empirical Bayes Information per other observation
for the DTI data
Empirical Bayes Information
0.00
0.05
0.10
0.15
0.20
0.25
0.30
.001
.02
.19
.24
.24
.08
.02
.002
−4
−2
0
2
4
x0

Learning from the Experience of Others

- *Which "others" ?*
  - All the other baseball players?
  - All the other voxels?
- *Next* DTI effect size MLEs x_i plotted versus d_i, distance from back of brain


MLE estimates x vs distance d from back of brain;
Red dots show the 10 largest x values
x = MLE
4
2
0
−2
−4
55
70
20
30
40
50
60
70
80
d = distance from back of brain
vertical lines indicated the 2920 voxels with 55< d <70


Empirical Bayes estimates based on only the 2920 voxels
having 55< d < 70 (Solid); Red curve shows original estimates.
Empirical Bayes Estimate
Maximum Likelihood Estimate
2.72
1.97
4.4
2920 voxels
all voxels

Relevance Functions

- Let $\rho_0(i)$ be *relevance* of "other" case i to target case i_0
  $\Big[$e.g., $\rho_0(i) = e^{-|d_i - 65|/10}$ for target case having $d = 65\Big]$
- Extended Tweedie Formula

$$\tilde{\mu}_{i_0} = x_{i_0} + \hat{l}'(x_{i_0}) + \frac{d}{dx} \log \hat{R}_0(x_{i_0})$$

where $\hat{R}_0(x) = \hat{E}\{\rho_0(i)|x\}$.

False Discovery Rates (1995)

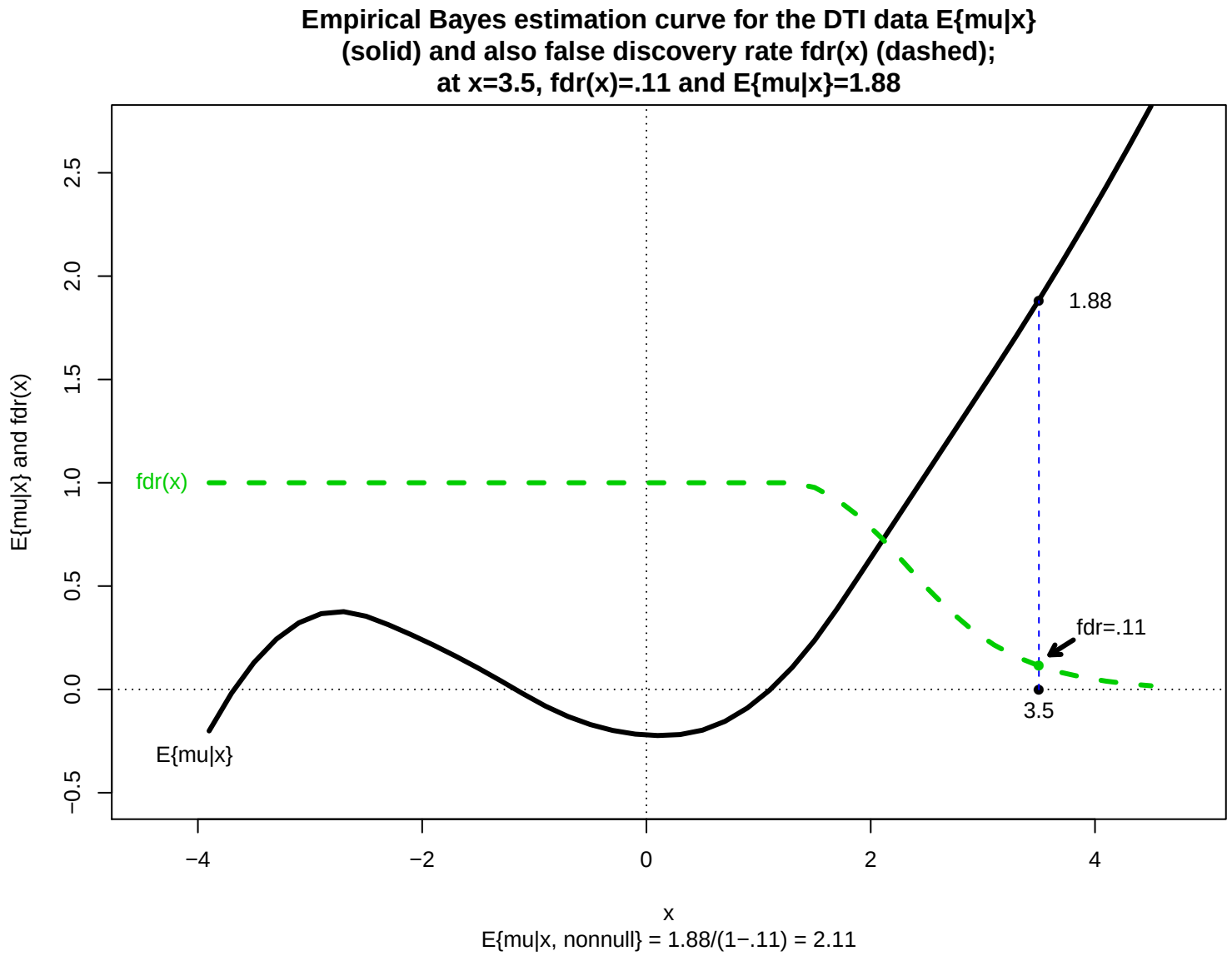
- *Tweedie* $\mu \sim g(\cdot)$ observe $x|\mu \sim \mathcal{N}(\mu, \sigma_0^2)$
- Testing $\pi_0 = \Pr\{\mu = 0\}$ ["null"; often π_0 near 1]
- Local false discovery rate $= \Pr\{\text{null} \,|\, x\}$

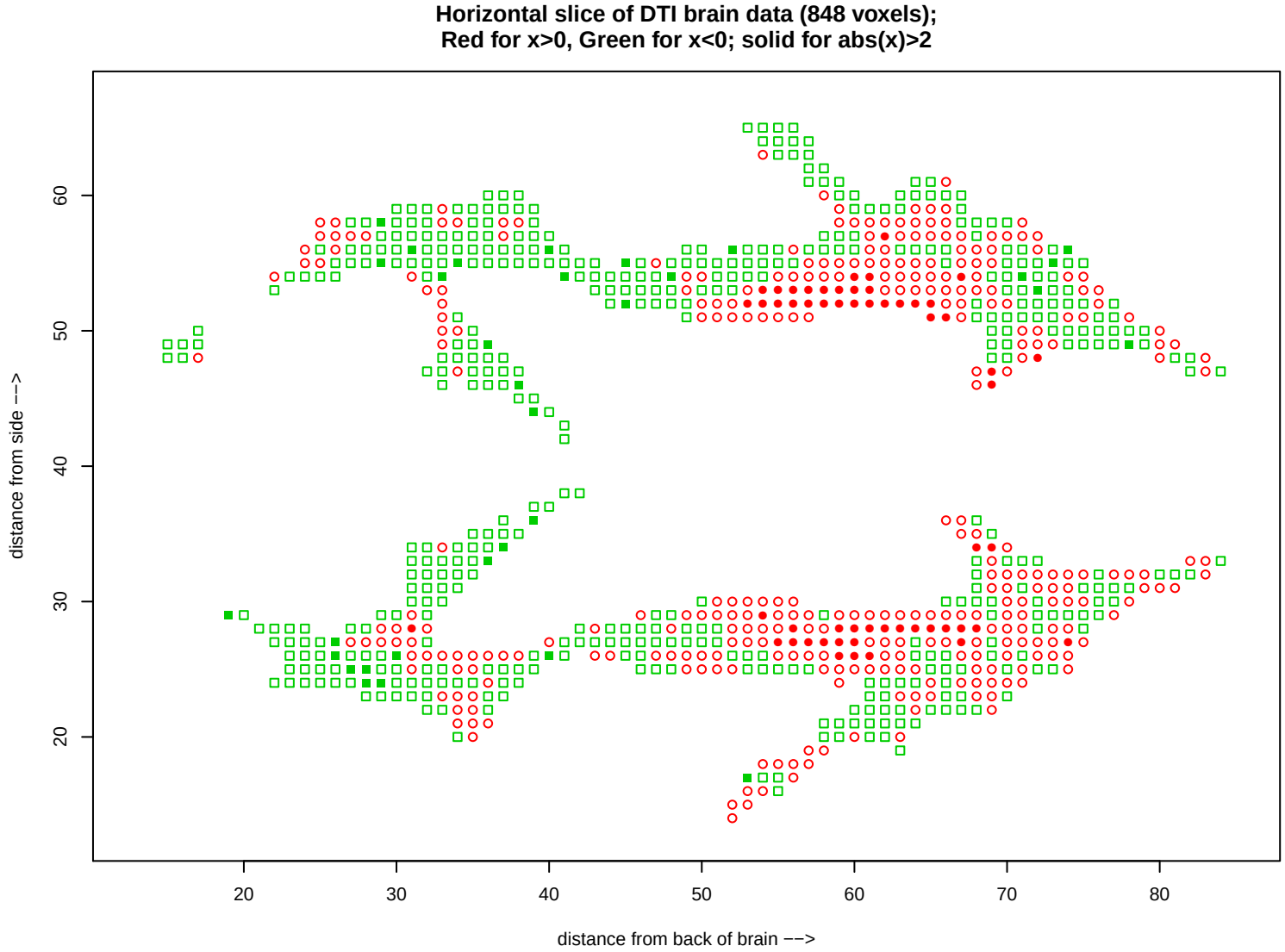
$$\boxed{\text{fdr}(x) = \pi_0 f_0(x)/f(x)}$$

- $f_0(x) = \dfrac{e^{-\frac{1}{2}\left(\frac{x}{\sigma_0}\right)^2}}{\sqrt{2\pi\sigma_0^2}}$ and $f(x) =$ mixture density

Tweedie and the fdr

- $-\log \text{fdr}(x) = \dfrac{x^2}{2\sigma_0^2} + l(x) + \text{constant} \quad [l(x) = \log f(x)]$
- $\boxed{\dfrac{d}{dx} - \log \text{fdr}(x) = \dfrac{1}{\sigma_0^2}\left[x + \sigma_0^2 l'(x)\right] = \dfrac{E\{\mu|x\}}{\sigma_0^2}}$
- locfdr (language R) estimates $\hat{l}(x)$ as before $\Rightarrow \widehat{\text{fdr}}(x)$
- fdr requires $\hat{l}$; Tweedie requires $\hat{l}'$


Empirical Bayes estimation curve for the DTI data E{mu|x}
(solid) and also false discovery rate fdr(x) (dashed);
at x=3.5, fdr(x)=.11 and E{mu|x}=1.88
E{mu|x} and fdr(x)
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
fdr(x)
1.88
fdr=.11
3.5
E{mu|x}
−4
−2
0
2
4
x
E{mu|x, nonnull} = 1.88/(1−.11) = 2.11


Horizontal slice of DTI brain data (848 voxels);
Red for x>0, Green for x<0; solid for abs(x)>2
distance from side −−>
20
30
40
50
60
distance from back of brain −−>
20
30
40
50
60
70
80

References

http://stat.stanford.edu/~brad/papers/

- "Tweedie's formula and selection bias"
- "Microarrays, empirical Bayes, and the two-groups model"
- "The future of indirect evidence"
- "Empirical Bayes estimates for large-scale prediction problems"