Empirical Bayes Deconvolution Problem

Bradley Efron

Stanford University

Bayes Deconvolution Problem

- Unknown prior density $g(\theta)$ gives *unobserved* realizations

$$\Theta_1, \Theta_2, \dots, \Theta_N \overset{\text{iid}}{\sim} g(\theta)$$

- Each Θ_k gives *observed* $\quad X_k \sim p_{\Theta_k}(x) \qquad [p_\Theta(x) \text{ known}]$
- Marginal density

$$\boxed{f(x) = \int p_\theta(x) g(\theta)\, d\theta}$$

- Wish to estimate $g(\theta)$ from $X_1, X_2, \dots, X_N$

Three Familiar Cases

- Poisson $X_k \sim \text{Poi}(\Theta_k)$, $p_\theta(x) = e^{-\theta}\theta^x / x!$
- Normal $X_k \sim \mathcal{N}(\Theta_k, 1)$, $p_\theta(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\theta)^2}$
- Binomial $X_k \sim \text{Bi}(n_k, \Theta_k)$, $\underbrace{p_\theta(x) = \binom{n_k}{x} \theta^x (1-\theta)^{n_k - x}}_{\text{depends on } n_k}$

Robbins' Estimate (1956)

- Observe $X_k \sim \text{Poi}(\Theta_k)$, $k = 1, \ldots, N$
- $y_x = \#\{X_k = x\}$
- $\boxed{\hat{E}\{\Theta \mid X = x\} = (x+1)y_{x+1}/y_x}$

Claims x	0	1	2	3	4	5	6
Counts y	7840	1317	239	42	14	4	4
$\hat{E}\{\Theta \mid x\}$	0.17	0.36	0.53	1.3	1.4	6	1.2

- Don't need to estimate g!

More Ambitious Goal

- Estimate entire prior density $g(\theta)$
- Why?
- *Ensemble properties* of Θ's

$$\Pr\{\Theta > 2\}, \text{ etc.}$$

- Empirical Bayes: estimate full posterior distribution

$$\Pr\{\Theta > 2 | X = x\}, \text{ etc.}$$

- Asymptotics discouraging (Carroll and Hall, 1988)

g Modeling

- Idea   Model prior $g(\theta)$ as exponential family
- Discrete Θ-space: $\Theta \in \left\{\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(m)}\right\}$
- *Prior* $g_j = \Pr\left\{\Theta = \theta_{(j)}\right\}$
- $\boldsymbol{g} = (g_1, \ldots, g_m)$
- *Exponential family* $\boxed{\boldsymbol{g}_\alpha = e^{\boldsymbol{Q}\alpha} / a_\alpha}$

$$\left(g_{j\alpha} = e^{Q_j'\alpha} / a_\alpha,\ j = 1, \ldots, m\right)$$

- $\underset{m \times p}{\boldsymbol{Q}}$ structure matrix, α p-dimensional natural parameter

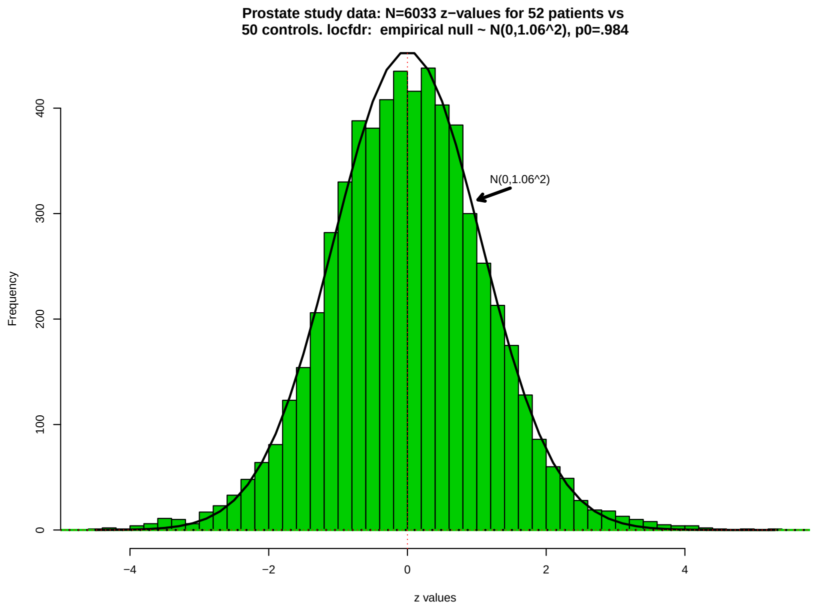
Maximum Likelihood Estimation of $\boldsymbol{g}_\alpha$

- $\alpha \longrightarrow \Theta \sim \boldsymbol{g}_\alpha \longrightarrow X \sim p_\Theta(x)$
- Marginal density $\boldsymbol{f}_\alpha(x)$
- $X \sim \mathcal{N}(\Theta, 1) \Rightarrow \boldsymbol{f}_\alpha = \boldsymbol{g}_\alpha * \mathcal{N}(0,1)$
- Observe $X_1, X_2, \ldots, X_N \sim \boldsymbol{f}_\alpha(x)$

  gives $\hat{\alpha}$, the marginal MLE, and $\boldsymbol{g}_{\hat{\alpha}}$

Prostate Study Example

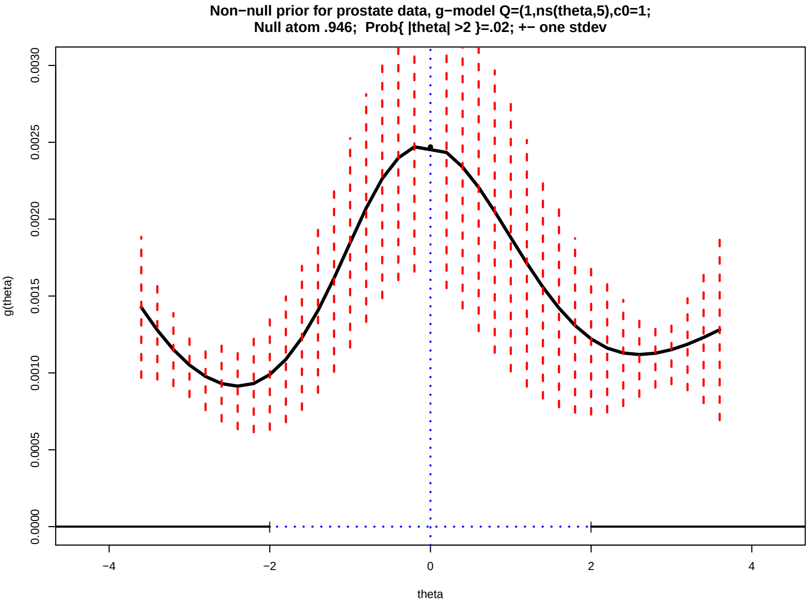
- 102 men, 52 patients and 50 controls
- $N = 6033$ genes
- Each gives "z value"
- $\boxed{z_i \sim \mathcal{N}(\theta_i, 1.06^2)}$ θ_i the "effect size"
- Wish to estimate $g(\theta)$, the effect size density


Prostate study data: N=6033 z−values for 52 patients vs
50 controls. locfdr: empirical null ~ N(0,1.06^2), p0=.984
N(0,1.06^2)
Frequency
0
100
200
300
400
−4
−2
0
2
4
z values

g-Modeling Estimate

- $\Theta \in \{-3.6, -3.4, \ldots, 3.6\}$
- $\boldsymbol{g}_\alpha = e^{Q\alpha} / a_\alpha$
- $Q = [\delta_0, ns(\theta, 5)]$ "spike and slab"
- Used "nlm" to find MLE $\hat{\alpha}$, giving
  - $\widehat{\Pr}\{\Theta = 0\} = 0.946$
  - $\widehat{\Pr}\{|\Theta| > 2\} = 0.02$

(`locfdr` atom 0.984 at $\Theta = 0$)


Non−null prior for prostate data, g−model Q=(1,ns(theta,5),c0=1;
Null atom .946; Prob{ |theta| >2 }=.02; +− one stdev
g(theta)
0.0000
0.0005
0.0010
0.0015
0.0020
0.0025
0.0030
−4
−2
0
2
4
theta

Discretizing the X Observations

- $X \in \left\{x_{(1)}, \ldots, x_{(n)}\right\}$
- $y_i = \#\left\{X_k = x_{(i)}\right\}$
- $\boldsymbol{y} = (y_1, \ldots, y_n)'$
- $f_i = \Pr\left\{X = x_{(i)}\right\}$
- $\boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f})$
- Let $p_{ij} = \Pr\left\{X = x_{(i)} \mid \Theta = \theta_{(j)}\right\}$ and $\underset{m \times n}{\boldsymbol{P}} = (p_{ij})$

$$\boxed{\underset{n}{\boldsymbol{f}} = \underset{m \times n}{\boldsymbol{P}} \; \underset{m}{\boldsymbol{g}}}$$

- MLE: $\alpha \longrightarrow \boldsymbol{g}_\alpha = e^{Q\alpha}/a_\alpha \longrightarrow \boldsymbol{f}_\alpha = \boldsymbol{P}\boldsymbol{g}_\alpha \rightarrow \boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}_\alpha)$

Fisher Information Calculations

- Define $W_{ij} = g_{\alpha j}\left(\frac{p_{ij}}{f_{\alpha i}} - 1\right)$
- $\underset{n\times m}{\boldsymbol{W}_{\alpha}} = (W_{ij})$
- *Expected Fisher information* at $\alpha = \hat{\alpha}$:

$$\boxed{\mathcal{I}_{\hat{\alpha}} = Q'\left\{\boldsymbol{W}'_{\hat{\alpha}}\,\mathrm{diag}(N\boldsymbol{f}_{\hat{\alpha}})\boldsymbol{W}_{\hat{\alpha}}\right\}Q}$$

Regularization and Accuracy for g Models

- $\hat{\alpha} = \arg\max_{\alpha}\{\ell_{\alpha} - s_{\alpha}\} \qquad (s_{\alpha} = c_0 \|\alpha\|)$

$$\hat{\alpha} - \alpha \,\dot{\sim}\, \Big[\underbrace{-\big(\hat{I} + \ddot{s}_{\hat{\alpha}}\big)^{-1} \dot{s}_{\hat{\alpha}},}_{\widehat{\text{Bias}}} \quad \underbrace{\big(\hat{I} + \ddot{s}_{\hat{\alpha}}\big)^{-1} \hat{I} \big(\hat{I} + \ddot{s}_{\hat{\alpha}}\big)^{-1}}_{\widehat{\text{Cov}}} \Big]$$

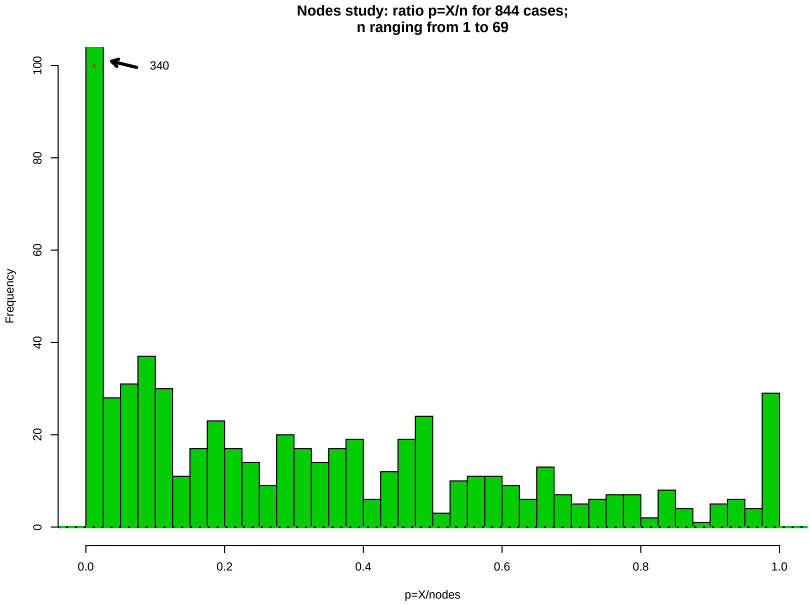
- Letting $\hat{R} = [\operatorname{diag}(\hat{g}) - \hat{g}\hat{g}']\, Q$,

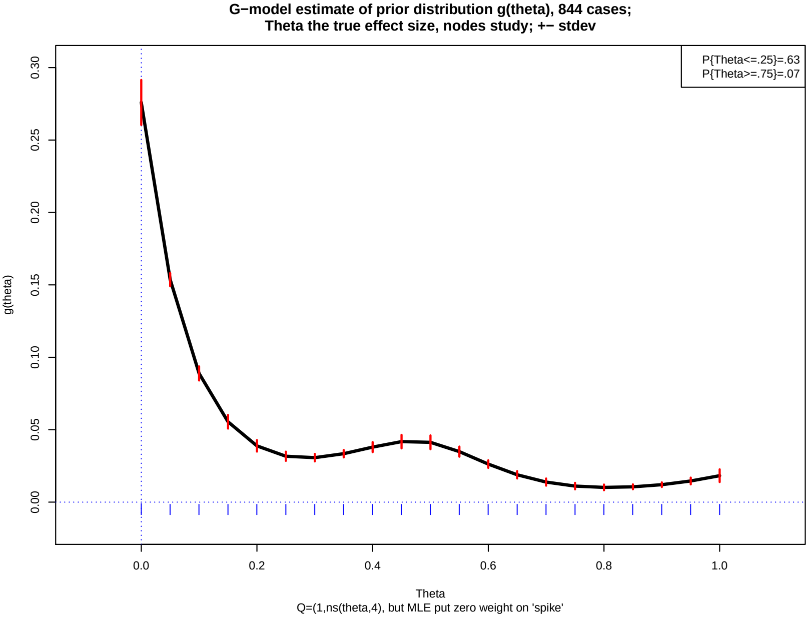
$$g_{\hat{\alpha}} - g \,\dot{\sim}\, \Big(\hat{R} \cdot \widehat{\text{Bias}},\ \hat{R} \cdot \widehat{\text{Cov}} \cdot \hat{R}'\Big)$$

$\Big[c_0 = 1 \text{ made } \operatorname{tr}(\ddot{s}_{\hat{\alpha}}) / \operatorname{tr}(\hat{I}) = 0.03 \Big]$

A Binomial Example

- 844 cancer patients: n_k lymph nodes removed; X_k found positive
- Binomial model $\quad X_k \sim \text{Bi}(n_k, \Theta_k) \qquad [\theta \in (0, .05, .10, \ldots, 1)]$
- *g modeling* $\qquad$ prior $\boldsymbol{g} = e^{\boldsymbol{Q}\alpha}/a_\alpha \quad$ with $Q = [\delta_0, ns(\theta, 4)]$
- $P_{844\times 21} : P_{kj} = \binom{n_k}{x_k} \theta_j^{x_k} (1-\theta_j)^{n_k - x_k}; \quad \boldsymbol{f}_\alpha = \boldsymbol{P}\boldsymbol{g}_\alpha$
- MLE $\qquad \ell_\alpha = \sum_{k=1}^{844} \log(f_{\alpha k}) \qquad [\text{sd's from } -\ddot{\ell}_{\hat{\alpha}}]$
- Fan (1991): binomial easier than normal


Nodes study: ratio p=X/n for 844 cases;
n ranging from 1 to 69
340
Frequency
0
20
40
60
80
100
0.0
0.2
0.4
0.6
0.8
1.0
p=X/nodes

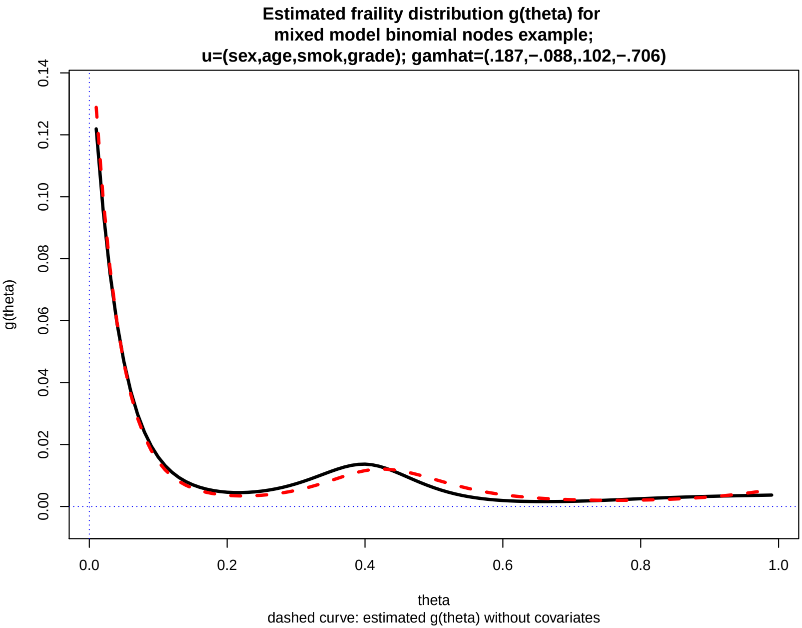

G−model estimate of prior distribution g(theta), 844 cases;
Theta the true effect size, nodes study; +− stdev
P{Theta<=.25}=.63
P{Theta>=.75}=.07
g(theta)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0
Theta
Q=(1,ns(theta,4), but MLE put zero weight on 'spike'

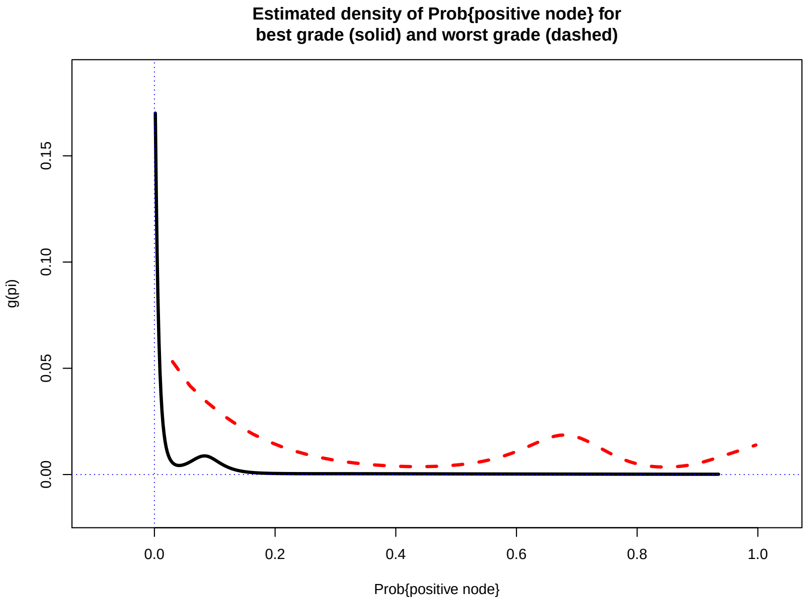
Binomial Nodes Example with Covariates

- Now $X_k \sim \text{Bi}(n_k, \pi_k)$ with

$$\boxed{\pi_k = \lambda^{-1}\left\{\lambda(\Theta_k) + \boldsymbol{U}_k'\gamma\right\}}$$

- $\lambda =$ logistic transform
- $\boldsymbol{U} =$ covariate vector
- $\Theta_k =$ "frailty" for patient k
- Generalized Mixed Model


Estimated fraility distribution g(theta) for
mixed model binomial nodes example;
u=(sex,age,smok,grade); gamhat=(.187,−.088,.102,−.706)
g(theta)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.0
0.2
0.4
0.6
0.8
1.0
theta
dashed curve: estimated g(theta) without covariates


Estimated density of Prob{positive node} for
best grade (solid) and worst grade (dashed)
g(pi)
0.00
0.05
0.10
0.15
0.0
0.2
0.4
0.6
0.8
1.0
Prob{positive node}

Fourier Method (Stefanski and Carroll, 1990)

- $X \sim \mathcal{N}(\Theta, 1)$: $\mathcal{F}(f) = \mathcal{F}(g) e^{-t^2/2}$ ($\mathcal{F}$ = Fourier transform)
- Smoothing

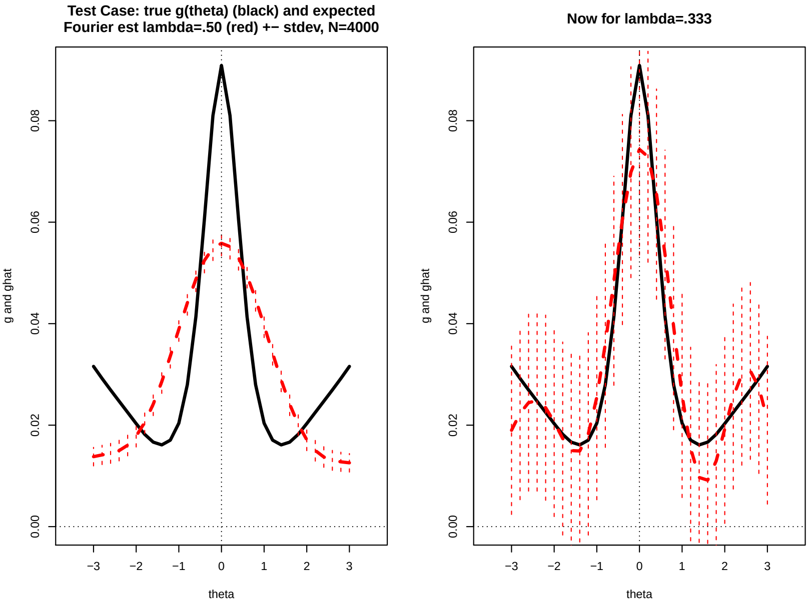
$$\hat{f}(x) = \frac{1}{N} \sum_{k=1}^{N} \sin\left(\frac{X_k - x}{\lambda}\right) \Big/ (X_k - x)$$

- Stef–Carroll: $\hat{g}(\theta) = \mathcal{F}^{-1}\left\{\mathcal{F}\left(\hat{f}\right) e^{t^2/2}\right\}$
- Kernal form $\hat{g}(\theta) = \frac{1}{N} \sum_{k=1}^{N} k_\lambda(X_k - \theta)$ where

$$\boxed{k_\lambda(x) = \frac{1}{\pi} \int_0^{1/\lambda} e^{t^2/2} \cos(tx)\, dt}$$

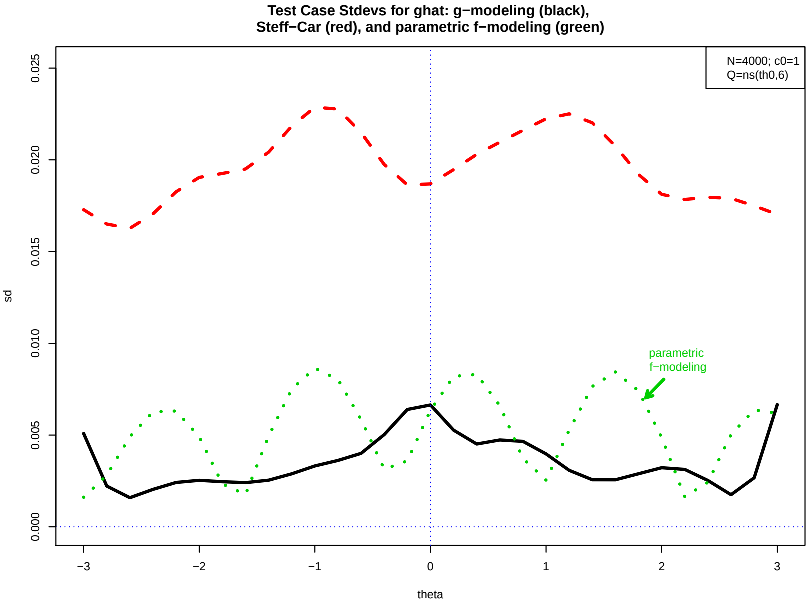
A Test Case

- Normal model $X_k \sim \mathcal{N}(\Theta_k, 1)$ with $\Theta \in [-3, 3]$
- $g(\theta)$ an equal mixture of $\mathcal{N}(0, 0.5^2)$ and a symmetric density proportional to $|\theta|$
- Gives triangular-shaped marginal $f(x)$
- *Goal* Sample from $f(\cdot)$, estimate g
- $N = 4000$


Test Case: true g(theta) (black) and expected
Fourier est lambda=.50 (red) +− stdev, N=4000
Now for lambda=.333
g and ghat
0.00
0.02
0.04
0.06
0.08
−3
−2
−1
0
1
2
3
theta

Parametric f Modeling

- Stef–Carroll: $\hat{\boldsymbol{g}} = \boldsymbol{k}_\lambda \bar{f}$ where $\bar{f}$ empirical density $\boldsymbol{y}/N$
- Instead take $\hat{\boldsymbol{g}} = \boldsymbol{k}_\lambda \hat{\boldsymbol{f}}$, $\hat{\boldsymbol{f}}$ parametric estimate of $\boldsymbol{f}$
- Next Slide $\hat{\boldsymbol{f}} = \texttt{glm(}\boldsymbol{y} \sim \texttt{ns(}\boldsymbol{x}\texttt{,6),Poisson)\$est}$
- Need $X_k = \Theta_k + \epsilon_k$, with ϵ_k iid noise


Test Case Stdevs for ghat: g−modeling (black),
Steff−Car (red), and parametric f−modeling (green)
N=4000; c0=1
Q=ns(th0,6)
parametric
f−modeling
sd
0.000
0.005
0.010
0.015
0.020
0.025
−3
−2
−1
0
1
2
3
theta

Summary

- Nonparametric f modeling: deprecated
- Parametric f modeling: OK for smooth g, additive noise (preferred for Robbins/Tweedie situations)
- g modeling: flexible and reasonably efficient for a wide variety of situations

References

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. *J. Amer. Statist. Assoc.* 83: 1184–1186.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* 29: 285–301.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* 19: 1257–1272.

Gholami, S., Janson, L., Worhunsky, D., et al. (2015). Number of lymph nodes removed and survival after gastric cancer resection: An analysis from the US Gastric Cancer Collaborative. *J. Amer. Coll. Surg.* 221: 291–299.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* 21: 169–184.