The Bayes Deconvolution Problem

Bradley Efron

Stanford University

Bayes Deconvolution Problem

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

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

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

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

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

Two Familiar Cases

- Poisson $\quad X_k \sim \text{Poi}(\Theta_k), \quad p_\theta(x) = e^{-\theta}\theta^x/x!$

$$\left(\text{Robbins:}\quad \hat{E}\{\Theta|X=x\} = (x+1)\frac{\hat{f}(x+1)}{\hat{f}(x)};\quad \hat{f}(x) = \frac{\#\{X_k = x\}}{N}\right)$$

- Normal $\quad X_k \sim \mathcal{N}(\Theta_k, 1), \quad p_\theta(x) = \dfrac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}(x-\theta)^2}$

$$\left(\text{Tweedie, Stein, Brown:}\quad \hat{E}\{\Theta|X=x\} = x + \frac{d}{dx}\log\hat{f}(x)\right)$$

- Neither requires $\hat{g}$

More Ambitious Goal

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

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

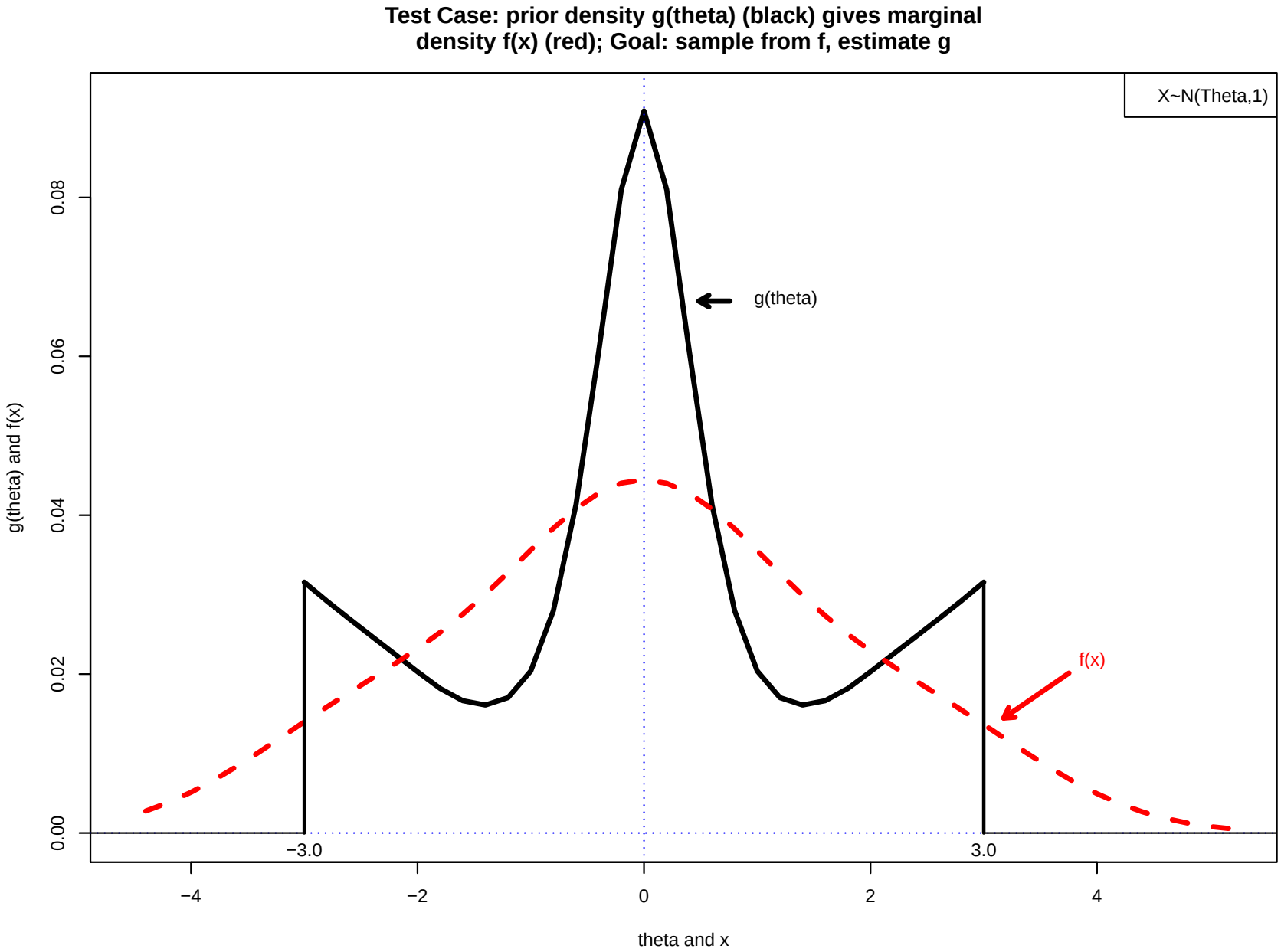
- *Empirical Bayes*: estimate full posterior distribution

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

- Asymptotics discouraging (Carroll and Hall, 1988)

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


Test Case: prior density g(theta) (black) gives marginal density f(x) (red); Goal: sample from f, estimate g
X~N(Theta,1)
g(theta)
f(x)
−3.0
3.0
g(theta) and f(x)
0.00
0.02
0.04
0.06
0.08
−4
−2
0
2
4
theta and x

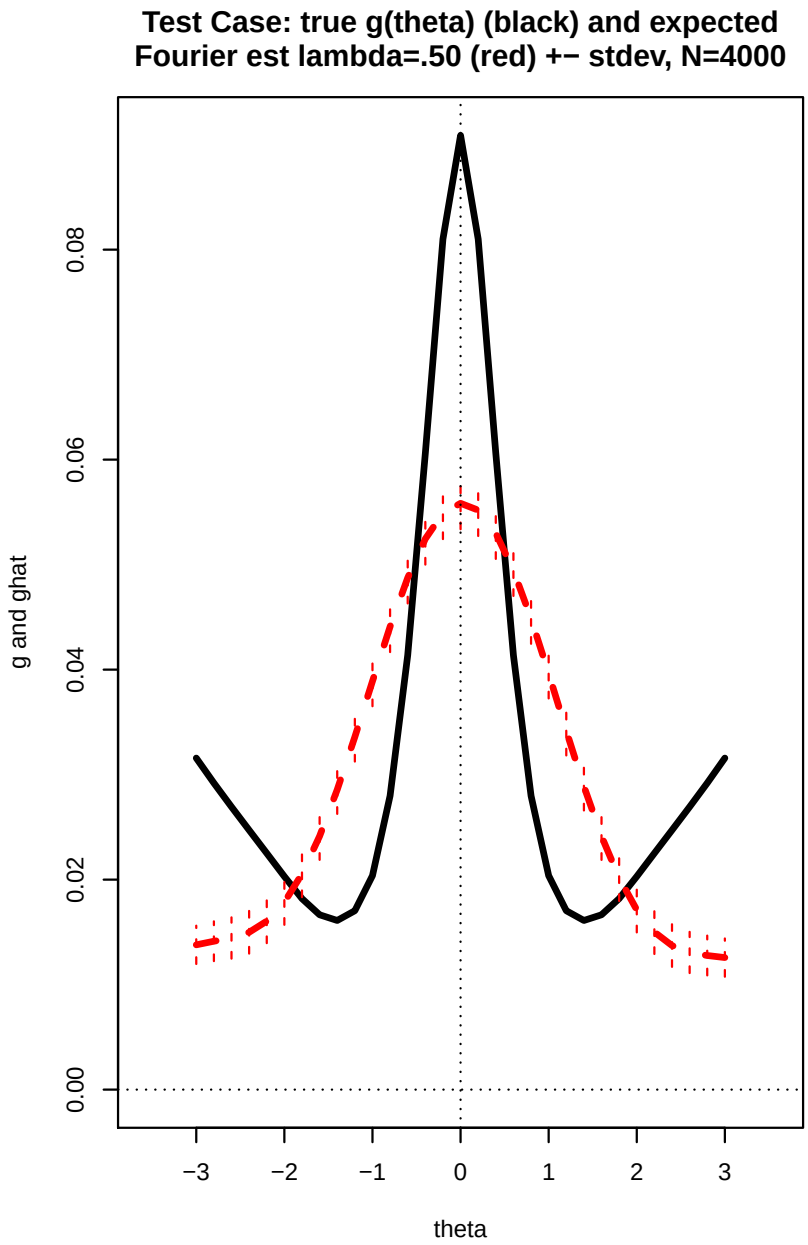
Fourier Method (Stefanski and Carroll, 1990)

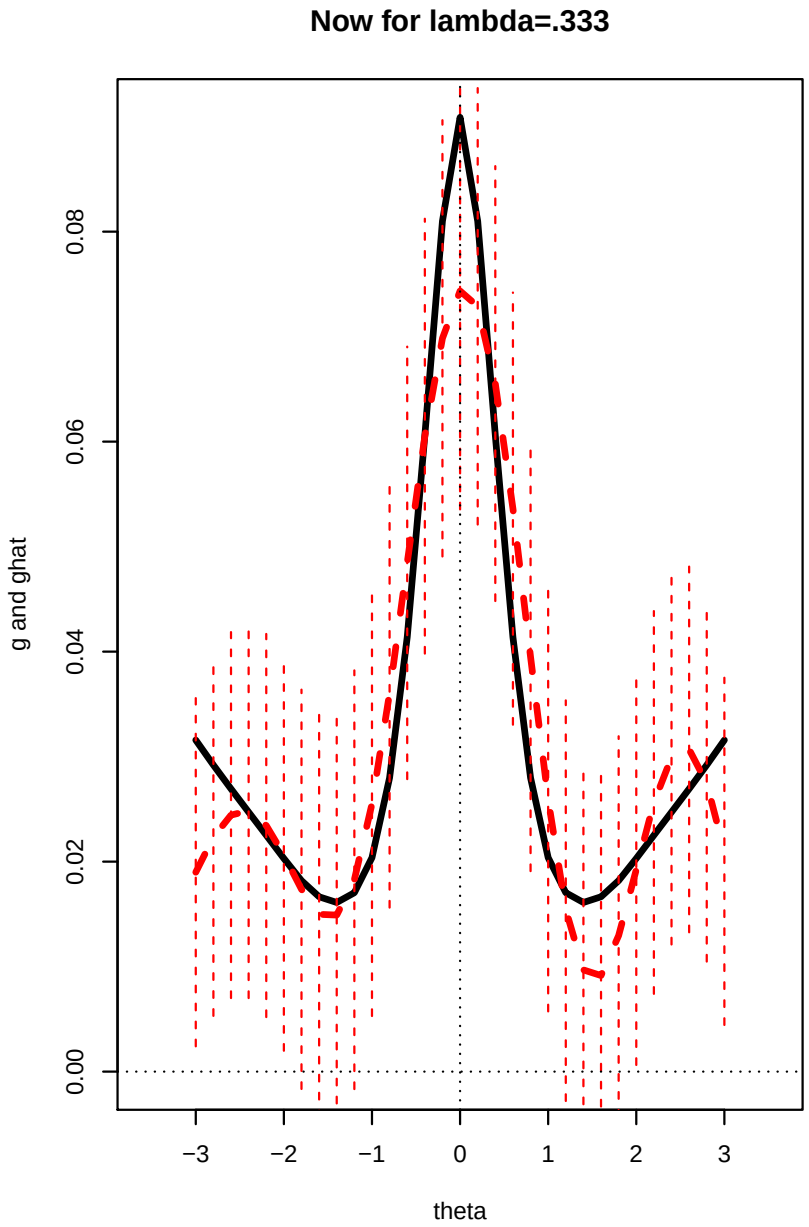
- $X \sim \mathcal{N}(\Theta, 1)$: $\mathcal{F}(f) = \mathcal{F}(g)e^{-t^2/2}$ ($\mathcal{F}$ = Fourier transform)
- Smoothing

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

- *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

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


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

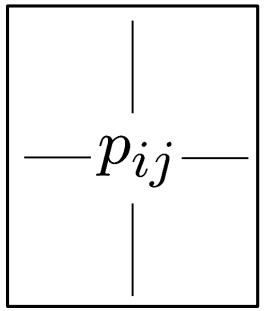

Now for lambda=.333
g and ghat
0.00
0.02
0.04
0.06
0.08
−3
−2
−1
0
1
2
3
theta

Discretization

- Θ in $(\theta_1, \theta_2, \ldots, \theta_m)$ and X in $(x_1, x_2, \ldots, x_n)$
- $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)$ and $\boldsymbol{f} = (f_1, f_2, \ldots, f_n)$
- $\underset{m\times n}{\boldsymbol{P}} = (p_{ij})$ with $p_{ij} = \Pr\{X = x_i | \Theta = \theta_j\}$
- Then $\boxed{\boldsymbol{f} = \boldsymbol{P}\boldsymbol{g}}$
- *Test case*

  $\boldsymbol{\theta} = (-3, -2.8, \ldots, 3)$ and $\boldsymbol{x} = (-4.4, -4.2, \ldots, 5.2)$

  $m = 31,\ n = 49$ and $p_{ij} = \varphi(x_i - \theta_j) \cdot 0.20$



Inversion and Regularization

- $\boldsymbol{f} = \boldsymbol{P}\boldsymbol{g}$ with svd $\boldsymbol{P} = \boldsymbol{U}\boldsymbol{d}\boldsymbol{V}'$
- $\boldsymbol{g} = \boldsymbol{P}^{-}\boldsymbol{f}$ with $\boldsymbol{P}^{-} = \boldsymbol{V}\frac{1}{\boldsymbol{d}}\boldsymbol{U}'$
- Regularization

$$\boxed{\boldsymbol{g}_r = \boldsymbol{P}_r^{-}\boldsymbol{f}} \qquad \text{with } \boldsymbol{P}_r^{-} = \boldsymbol{V}_r\frac{1}{\boldsymbol{d}_r}\boldsymbol{U}_r'$$

(first r eigens; bias/variance trade-off)

- *Stef–Carroll*:

$$\boxed{\boldsymbol{g} = \boldsymbol{k}_\lambda\boldsymbol{f}} \qquad \text{with } \boldsymbol{k}_\lambda(i,j) = k_\lambda(x_i - \theta_j)$$

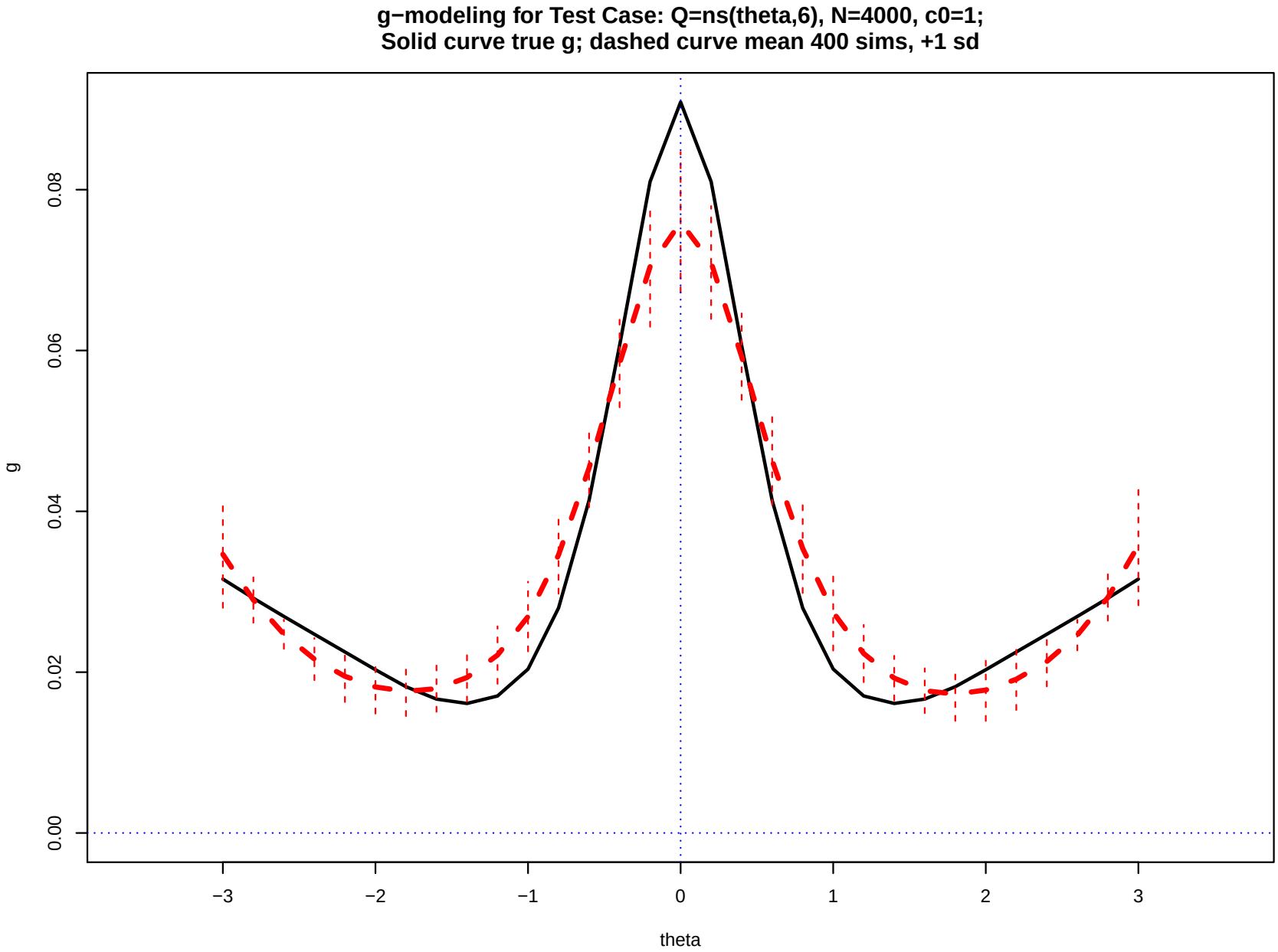
($\lambda = 1/2$ like $r = 4$; $\lambda = 1/3$ like $r = 7$)

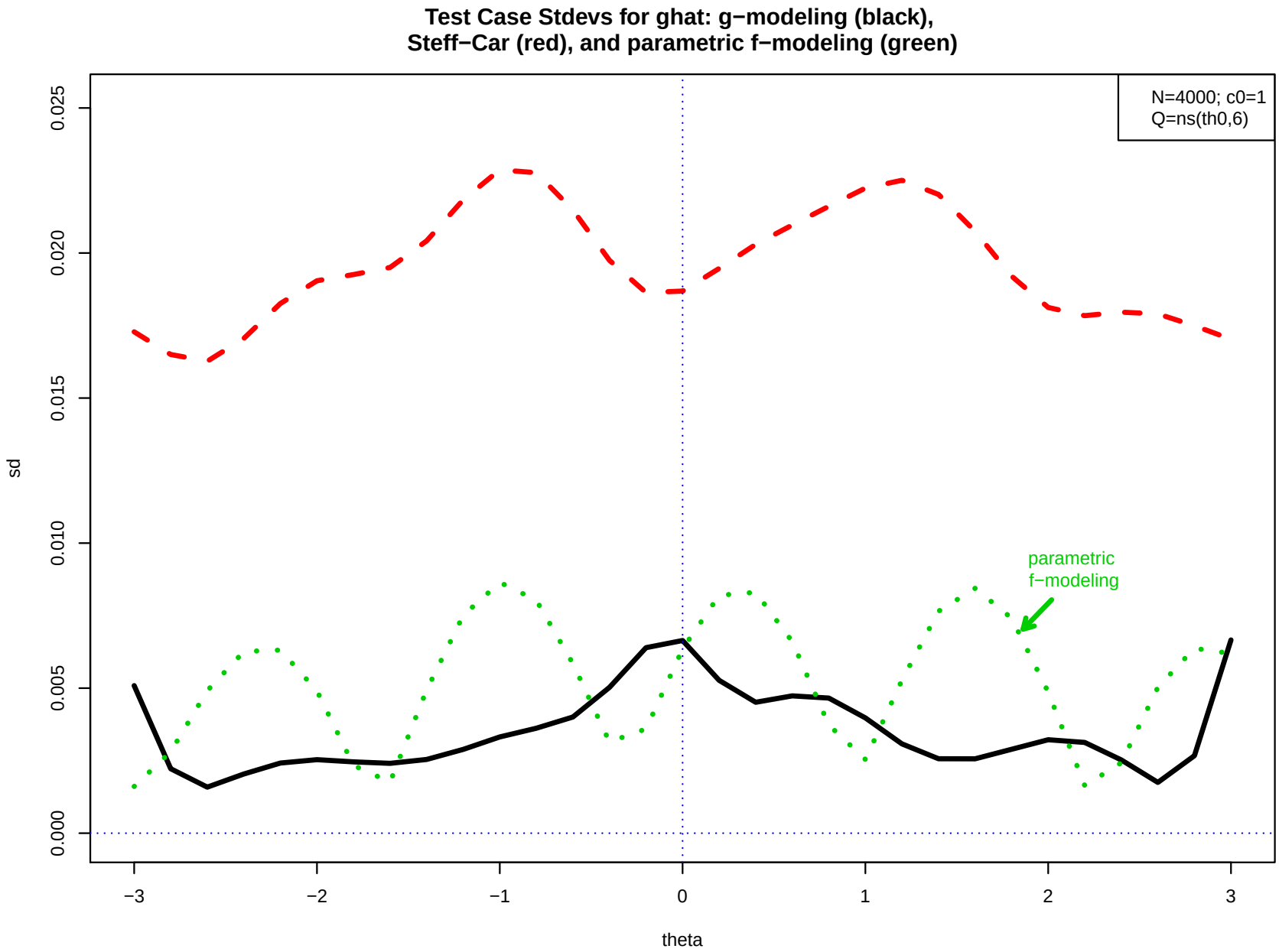
g Modeling

- Idea Model $\boldsymbol{g} = (g_1, g_2, \ldots, g_m)$ as exponential family on $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$:

$$\boldsymbol{g}_\alpha = e^{\boldsymbol{Q}\alpha}/a_\alpha \qquad \left[\underset{m\times p}{\boldsymbol{Q}} \text{ structure matrix, } \alpha \text{ natural parameter}\right]$$

- $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ vector of counts: $y_i = \#\{X_k = x_i\}$
- MLE $\alpha \to \boldsymbol{g}_\alpha \to \boldsymbol{f}_\alpha = \boldsymbol{P}\boldsymbol{g}_\alpha \to \boldsymbol{y} \sim \text{Mult}_n(N, \boldsymbol{f}_\alpha) \xrightarrow{\text{mle}} \hat{\alpha}$
- *Test case* $\boldsymbol{Q} = ns(\theta, 6)$


g−modeling for Test Case: Q=ns(theta,6), N=4000, c0=1;
Solid curve true g; dashed curve mean 400 sims, +1 sd
g
0.00
0.02
0.04
0.06
0.08
−3
−2
−1
0
1
2
3
theta


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

Fisher Information Calculations

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

$$\underset{p\times p}{\hat I} = -\ddot\ell_{\hat\alpha} = Q'\left\{\boldsymbol{W}'_{\hat\alpha}\,\mathrm{diag}(\boldsymbol{y})\boldsymbol{W}_{\hat\alpha} - \mathrm{diag}(\boldsymbol{W}_{\hat\alpha}\boldsymbol{y})\right\}Q$$

- *Expected Fisher information* at $\alpha = \hat\alpha$:

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

Regularization and Accuracy for g Models

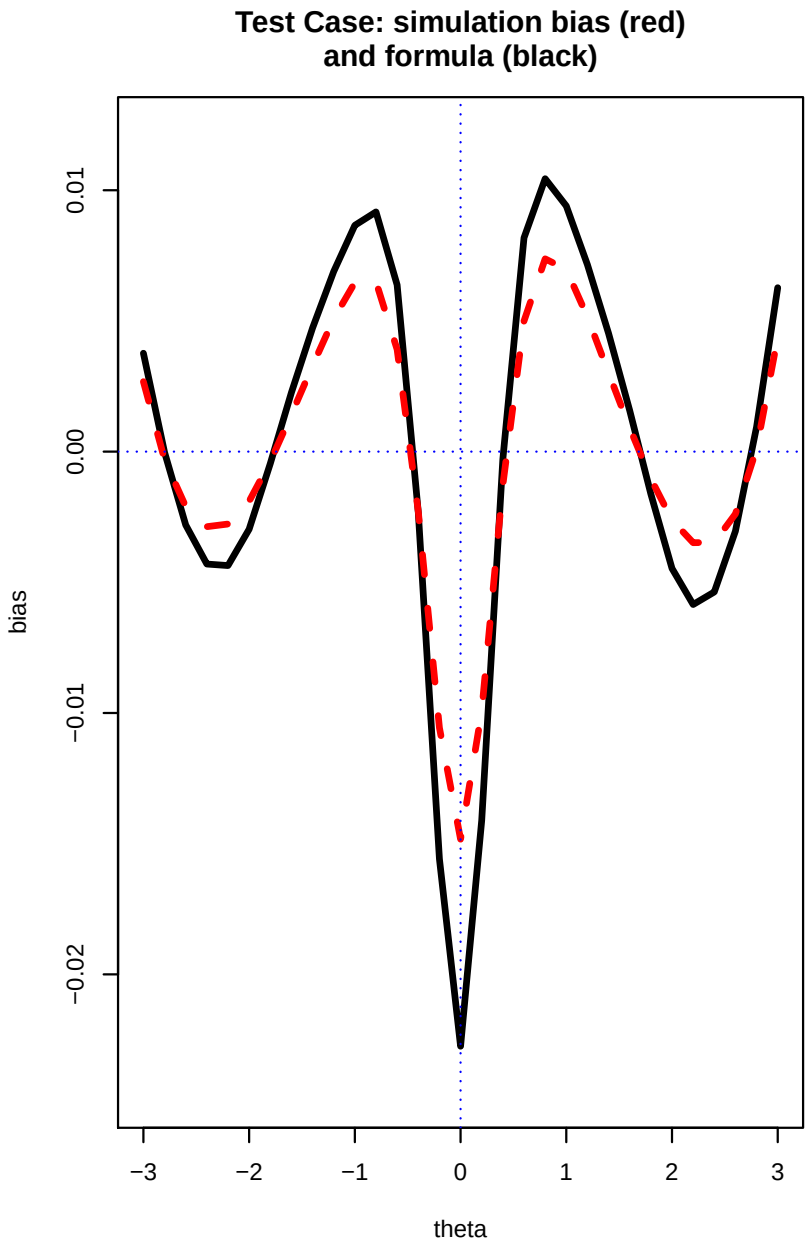
- $\hat{\alpha} = \arg\max_\alpha\{\ell_\alpha - s_\alpha\} \qquad (s_\alpha = c_0\|\alpha\|)$

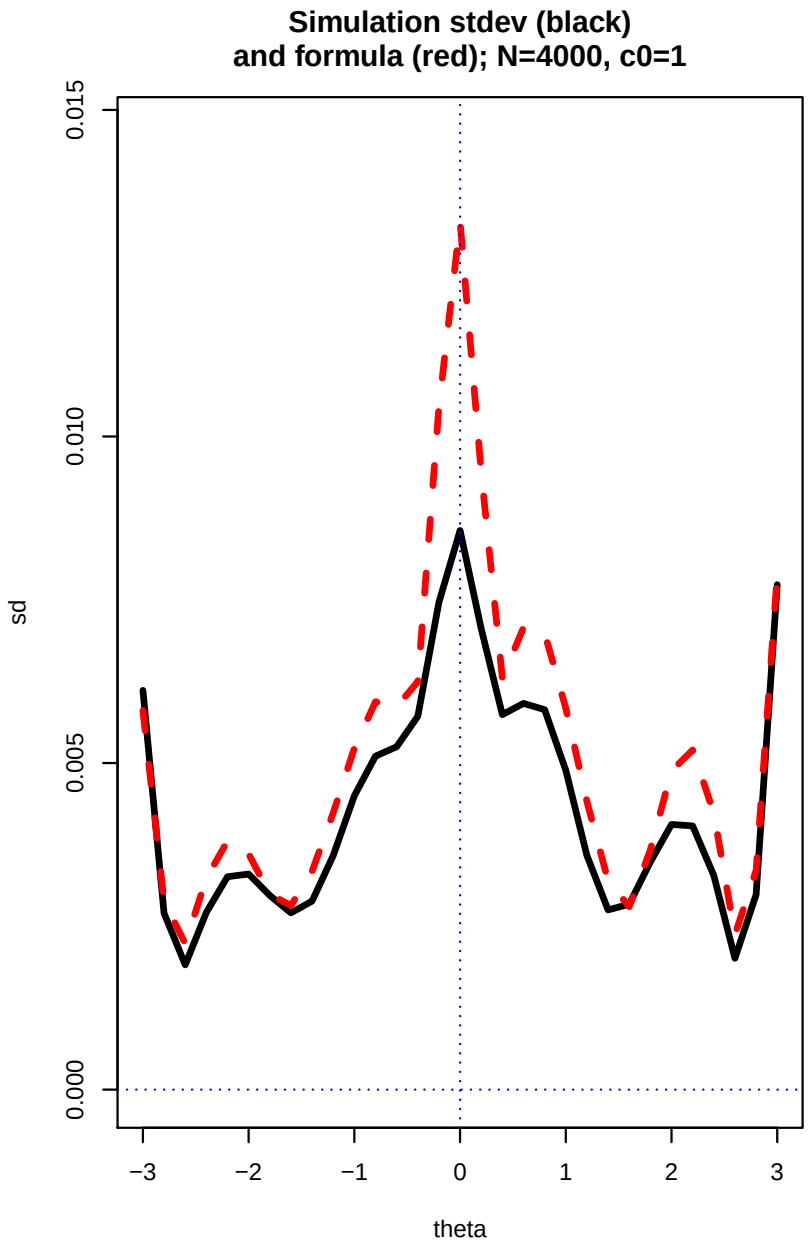
$$\hat{\alpha} - \alpha \doteq \Big[\underbrace{-\left(\hat{I} + \ddot{s}_{\hat{\alpha}}\right)^{-1}\dot{s}_{\hat{\alpha}}}_{\widehat{\text{Bias}}},\quad \underbrace{\left(\hat{I} + \ddot{s}_{\hat{\alpha}}\right)^{-1}\hat{I}\left(\hat{I} + \ddot{s}_{\hat{\alpha}}\right)^{-1}}_{\widehat{\text{Cov}}}\Big]$$

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

$$g_{\hat{\alpha}} - g \doteq \left(\hat{R}\cdot\widehat{\text{Bias}},\ \hat{R}\cdot\widehat{\text{Cov}}\cdot\hat{R}'\right)$$

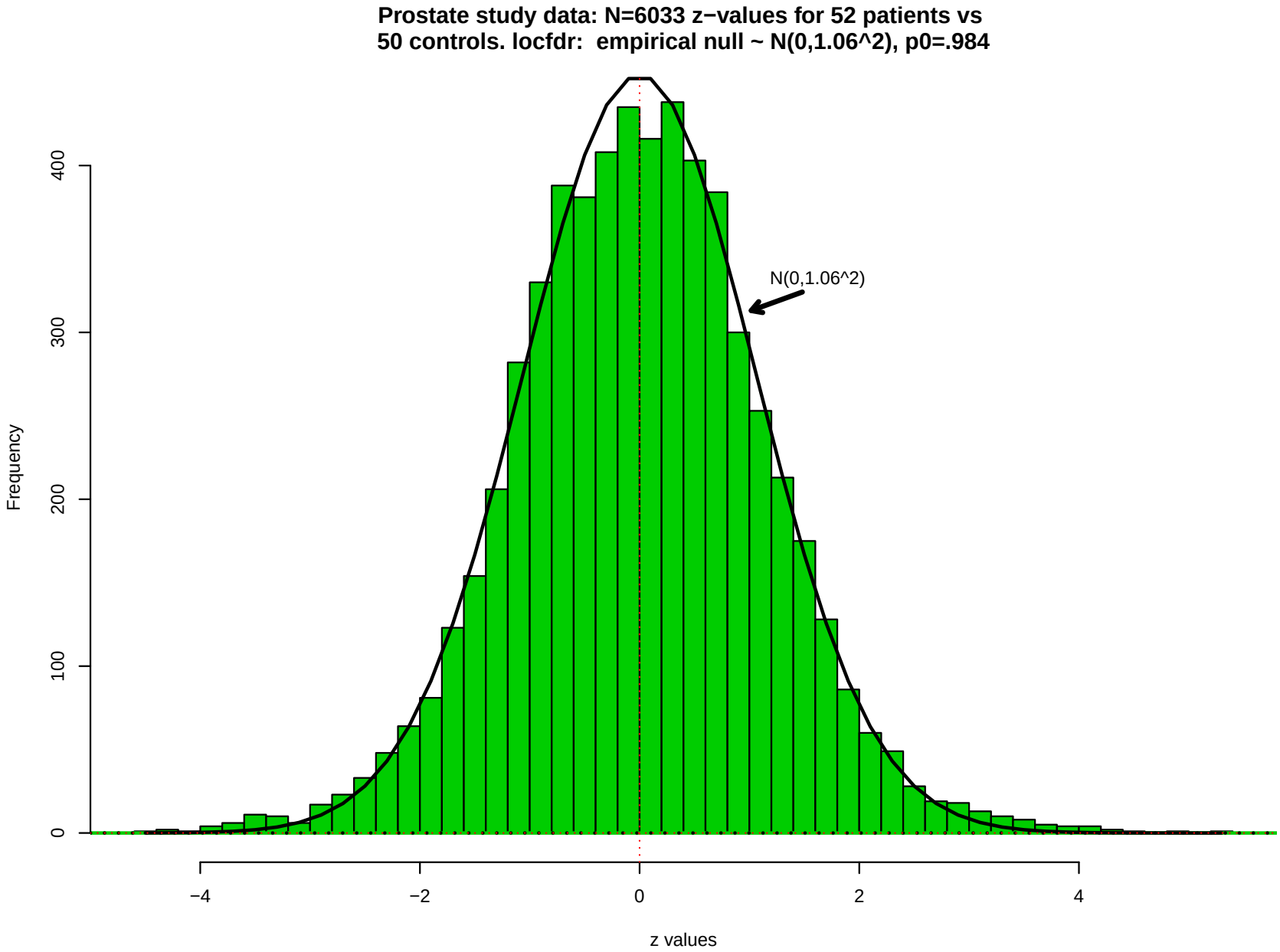
$$\left[c_0 = 1 \text{ made } \text{tr}(\ddot{s}_{\hat{\alpha}})/\,\text{tr}(\hat{I}) = 0.03\right]$$


Test Case: simulation bias (red)
and formula (black)
bias
0.01
0.00
−0.01
−0.02
−3
−2
−1
0
1
2
3
theta


Simulation stdev (black)
and formula (red); N=4000, c0=1
sd
0.015
0.010
0.005
0.000
−3
−2
−1
0
1
2
3
theta

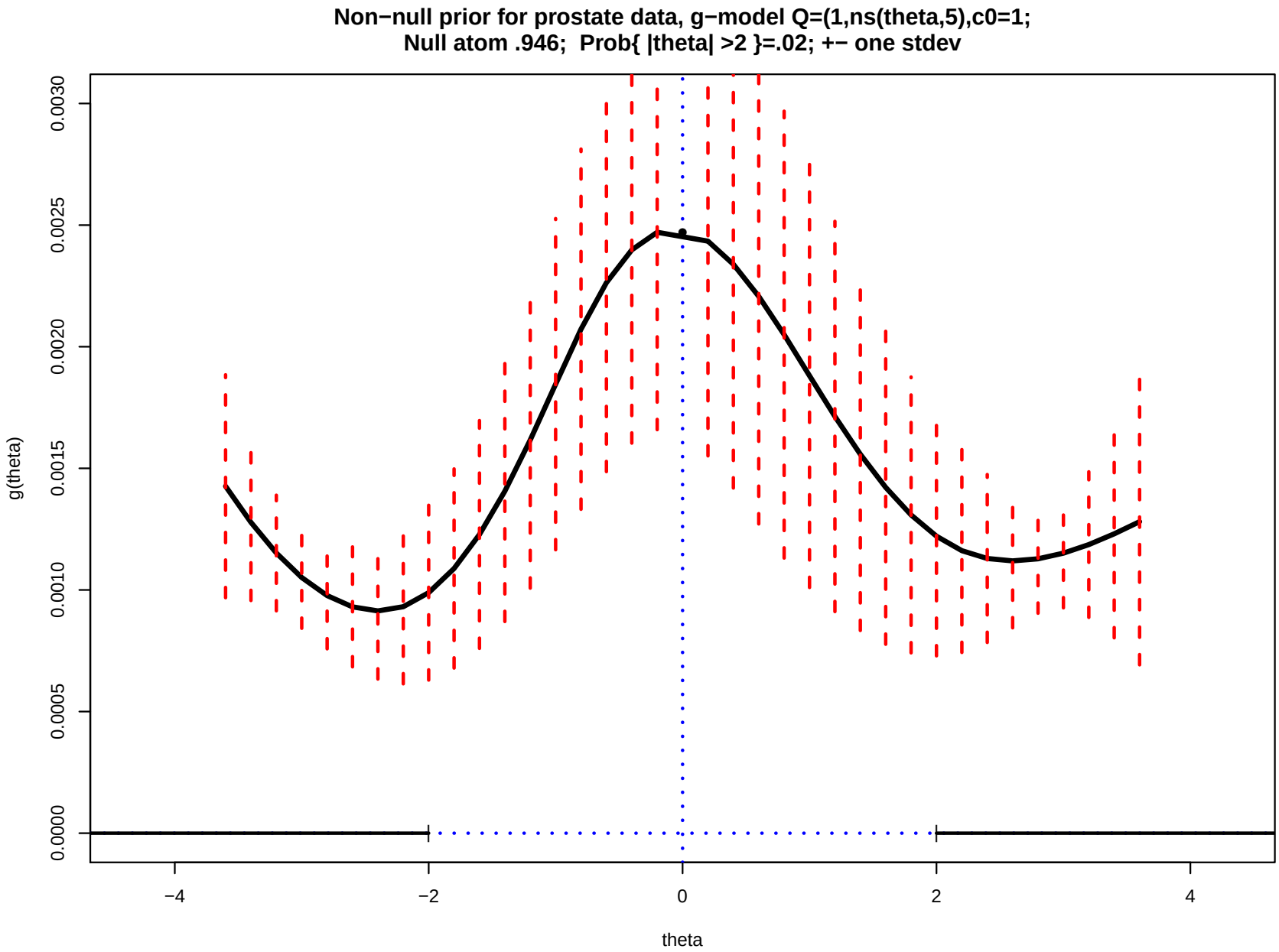
Parametric f Modeling

- *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 $\boldsymbol{f}$
- Slide 13 $\hat{\boldsymbol{f}} = \texttt{glm}(\boldsymbol{y} \sim \textit{ns}(\boldsymbol{x}, 6), \texttt{Poisson})\texttt{\$est}$
- Need $X_k = \Theta_k + \epsilon_k$, with ϵ_k iid noise
- Need $g(\theta)$ smooth


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

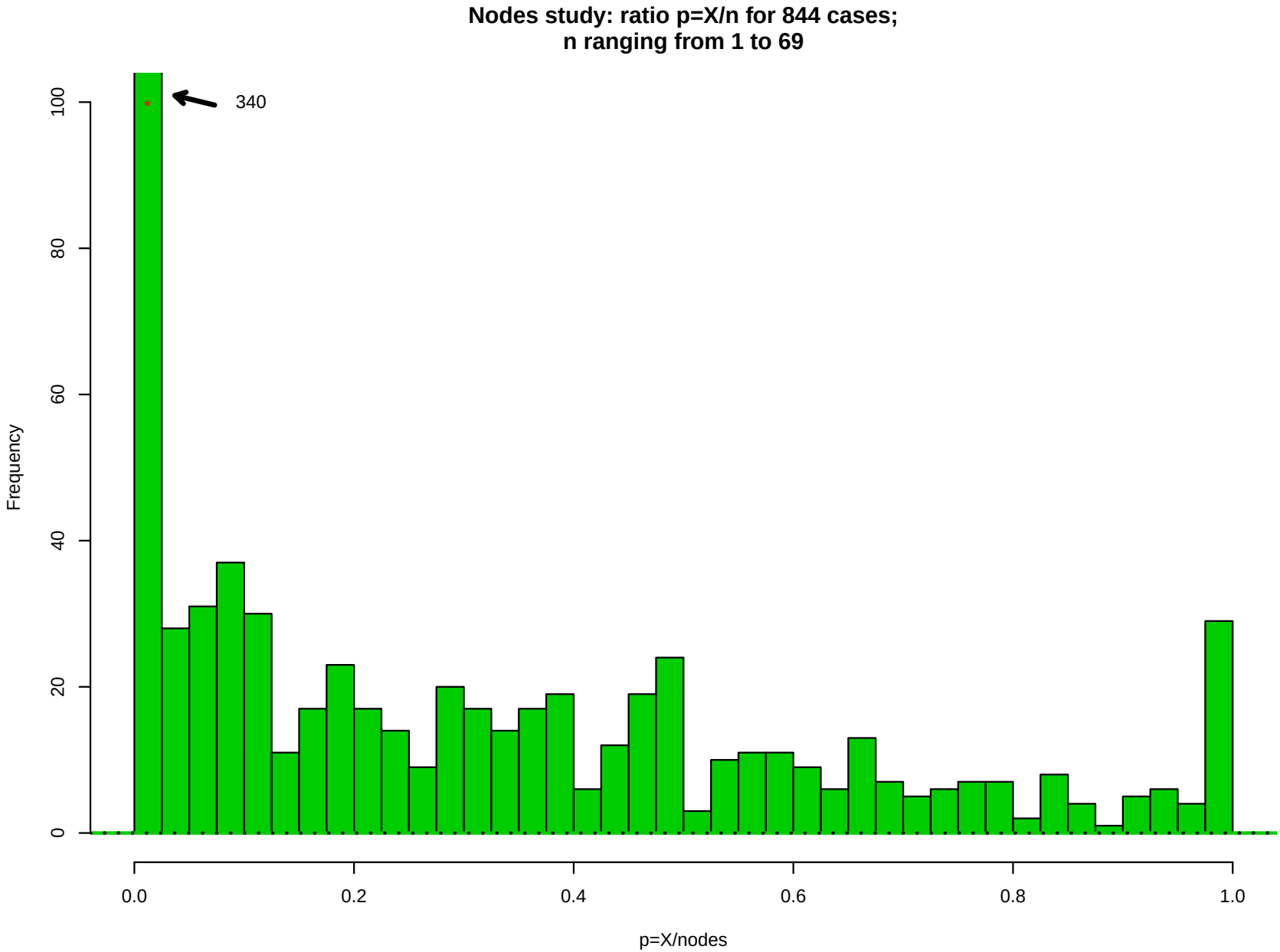
Estimating $g(\theta)$ for the Prostate Study

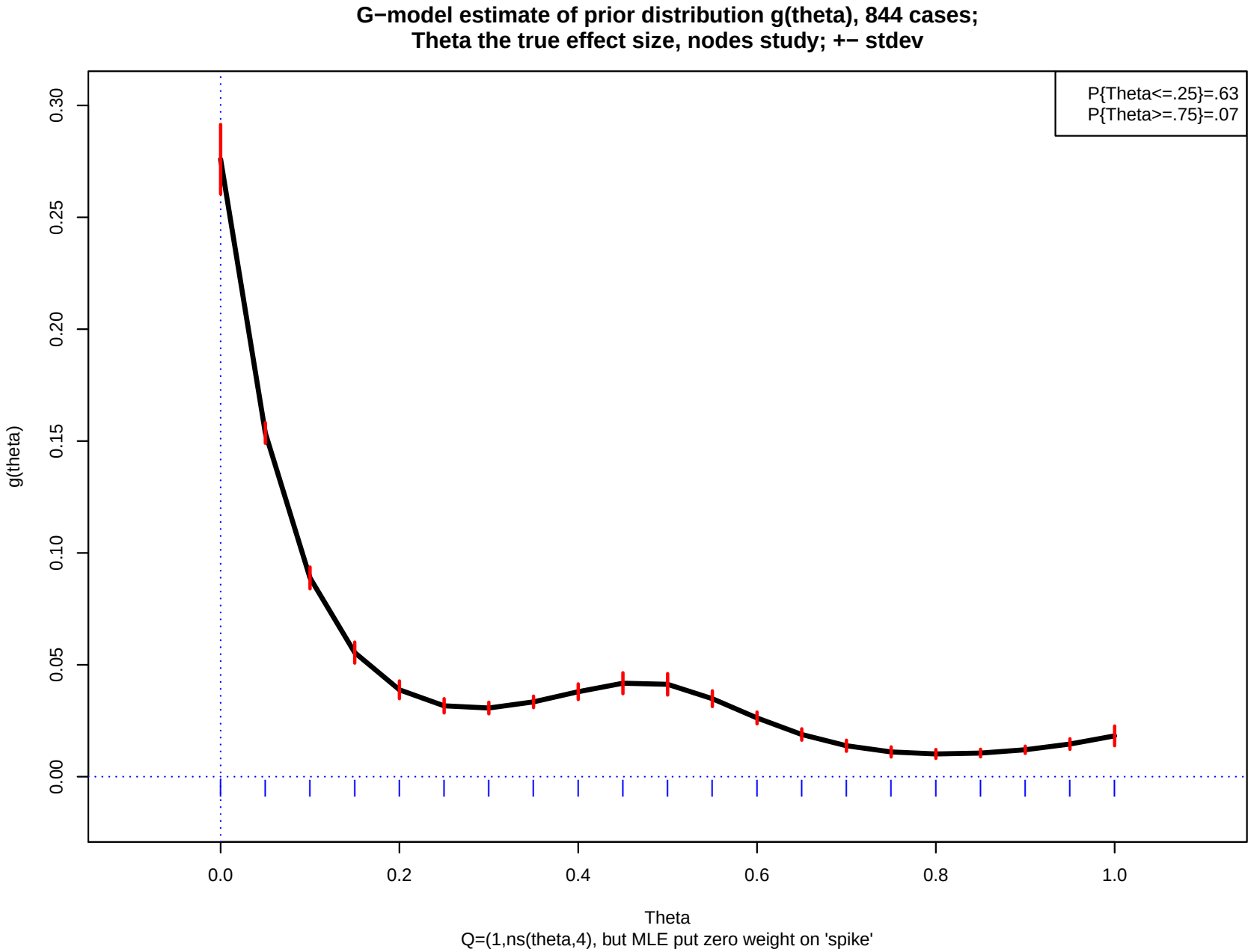
- z-value for kth gene $\quad z_k \sim \mathcal{N}(\Theta_k, 1.06^2) \quad (\Theta_k =$ "effect size")
- g model $\quad Q = (\mathbf{1}, ns(\theta, 5)) \quad$ ("spike and slab")
- Tried $c_0 = 0.5,\ 1,\ 2$
- *Accuracy formula* gave $\widehat{\text{sd}}$, and $\widehat{\text{Bias}} \doteq 0$


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

A Binomial Example

- 844 cancer patients: n_k lymph nodes removed; X_k found positive
- *Binomial model* $X_k \sim \text{binom}(n_k, \Theta_k)$ $[\boldsymbol{\theta} = (0, 0.05, 0.10, \ldots, 1)]$
- g modeling Prior $\boldsymbol{g} = e^{\boldsymbol{Q}\alpha}/a_\alpha$ with $Q = (1, ns(\boldsymbol{\theta}, 4))$
- $P_{844\times 21}$: $P_{kj} = \binom{n_k}{x_k}\theta_j^{x_k}(1-\theta_j)^{n_k - x_k}$; $\boldsymbol{f}_\alpha = \boldsymbol{P}\boldsymbol{g}_\alpha$
- MLE $\ell_\alpha = \sum_{k=1}^{844} \log(f_{\alpha k})$ [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

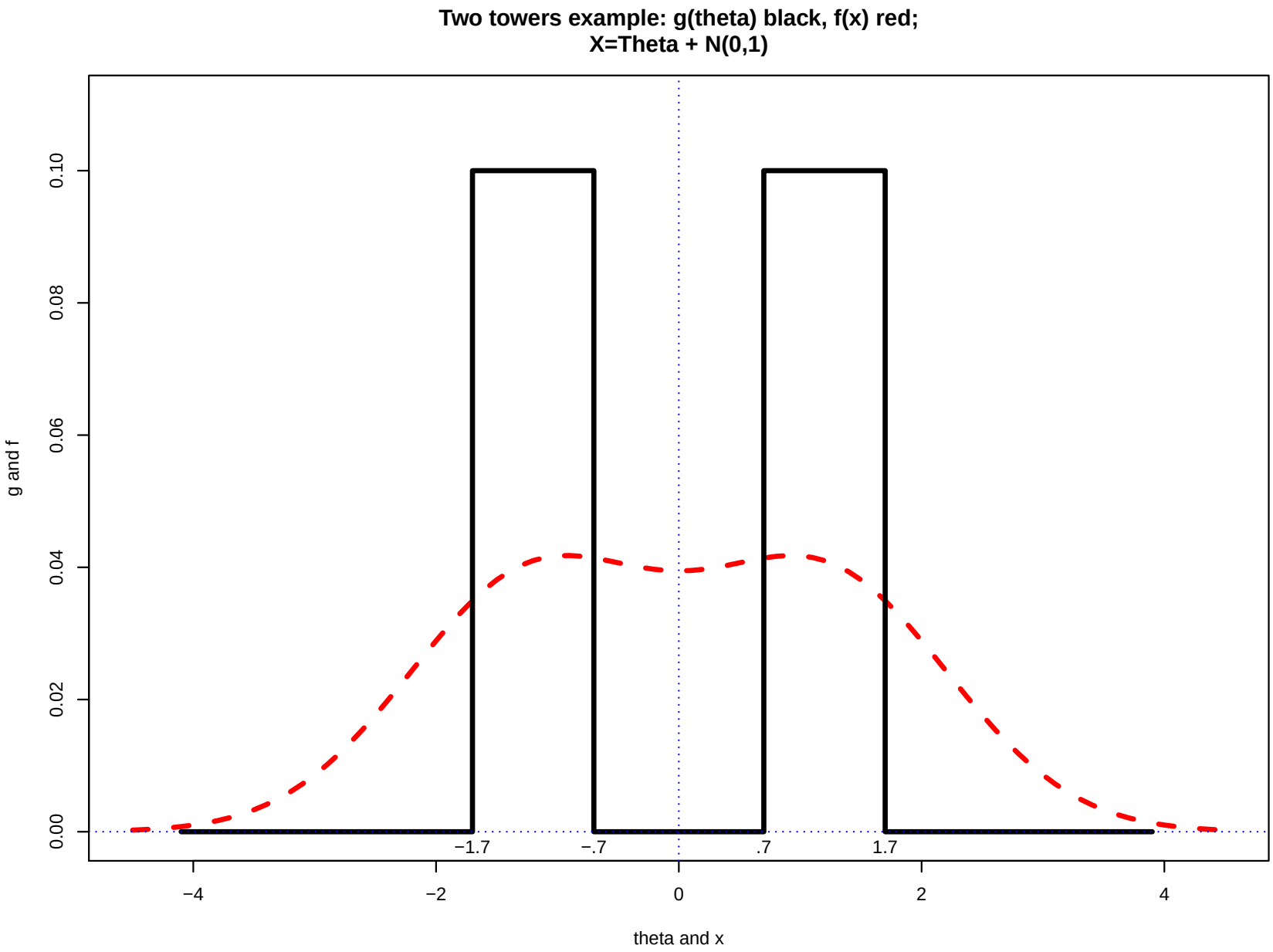

G−model estimate of prior distribution g(theta), 844 cases;
Theta the true effect size, nodes study; +− stdev
P{Theta<=.25}=.63
P{Theta>=.75}=.07
g(theta)
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0
0.2
0.4
0.6
0.8
1.0
Theta
Q=(1,ns(theta,4), but MLE put zero weight on 'spike'

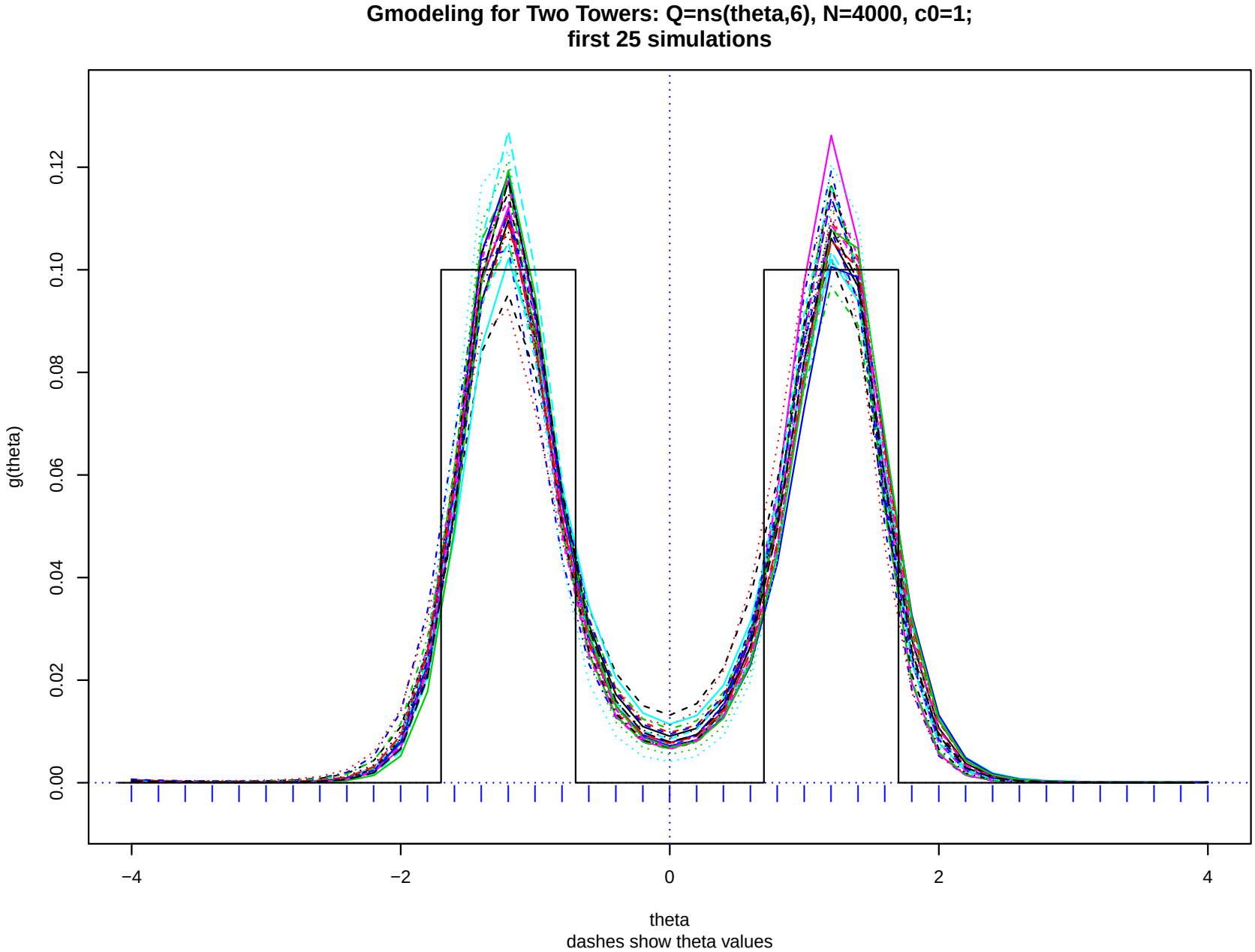
A Difficult Example

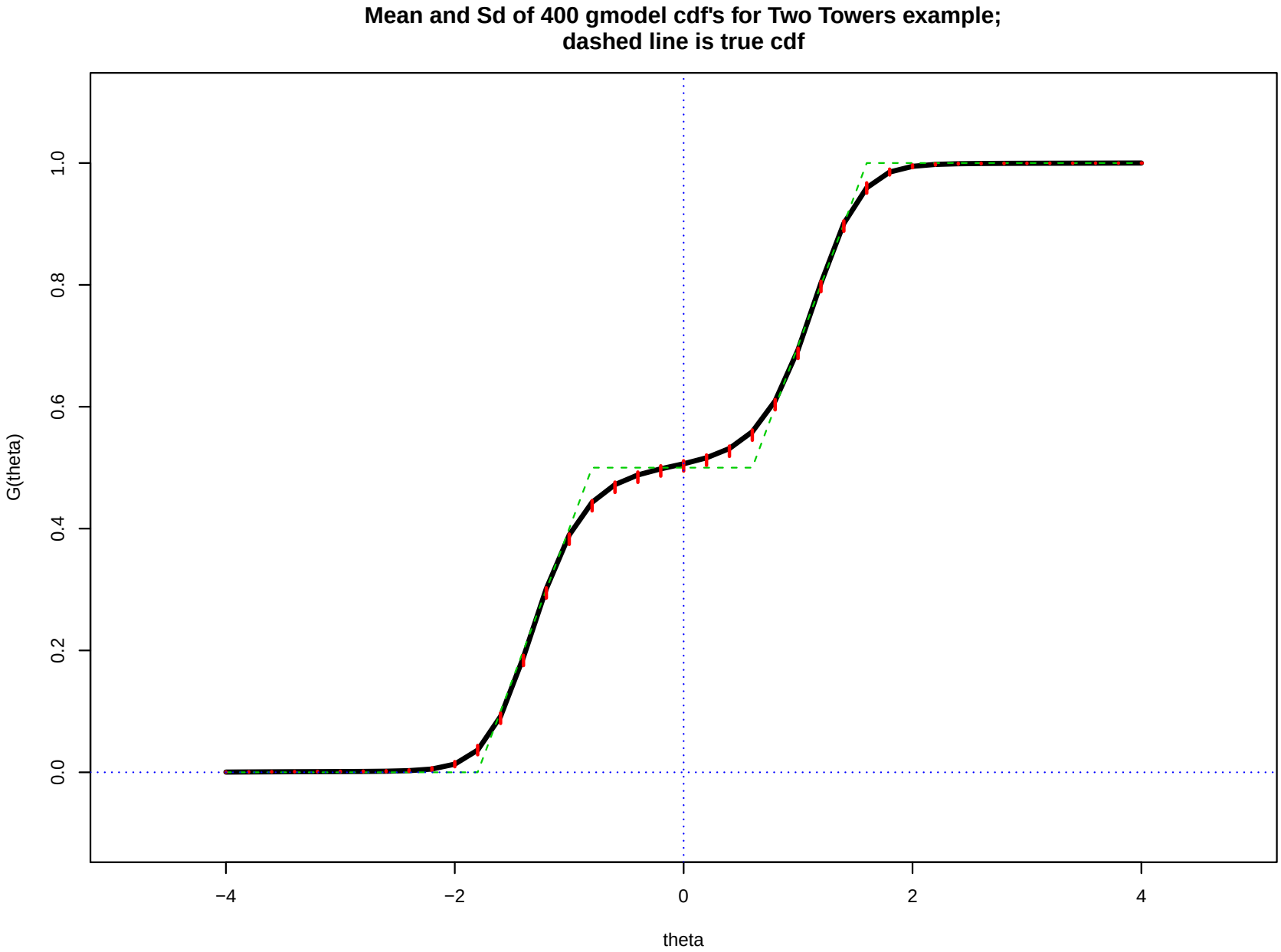
- *"Two Towers"*

$$g(\theta) = \begin{cases} 1 & \text{on } [-1.7, -0.7] \text{ and } [0.7, 1.7] \\ 0 & \text{otherwise} \end{cases}$$

- $X \sim \mathcal{N}(\Theta, 1)$
- g modeling $\quad \boldsymbol{\theta} = (-4, -3.8, -3.6, \ldots, 3.8, 4)$
- $Q = ns(\boldsymbol{\theta}, 6), \quad c_0 = 1$


Two towers example: g(theta) black, f(x) red;
X=Theta + N(0,1)
g and f
0.00
0.02
0.04
0.06
0.08
0.10
−1.7
−.7
.7
1.7
−4
−2
0
2
4
theta and x


Gmodeling for Two Towers: Q=ns(theta,6), N=4000, c0=1;
first 25 simulations
g(theta)
0.00
0.02
0.04
0.06
0.08
0.10
0.12
−4
−2
0
2
4
theta
dashes show theta values

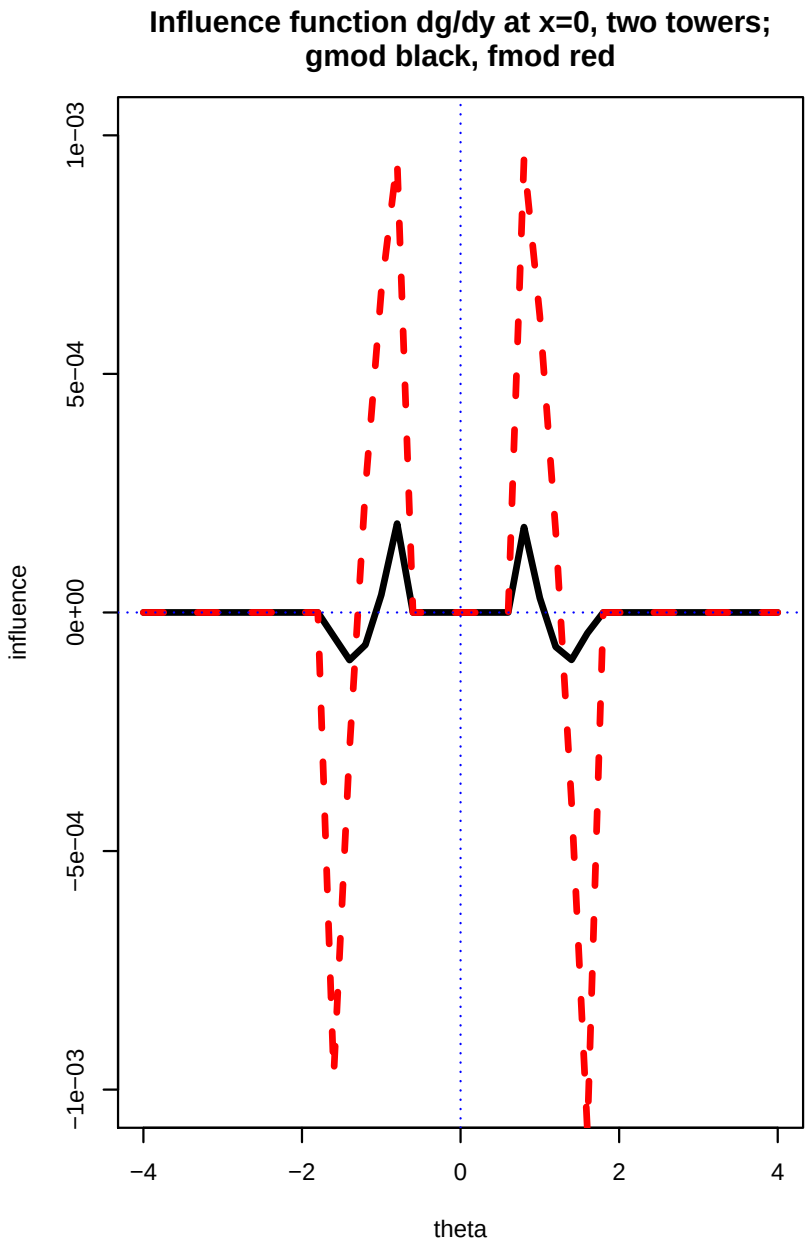

Mean and Sd of 400 gmodel cdf's for Two Towers example;
dashed line is true cdf
G(theta)
0.0
0.2
0.4
0.6
0.8
1.0
−4
−2
0
2
4
theta

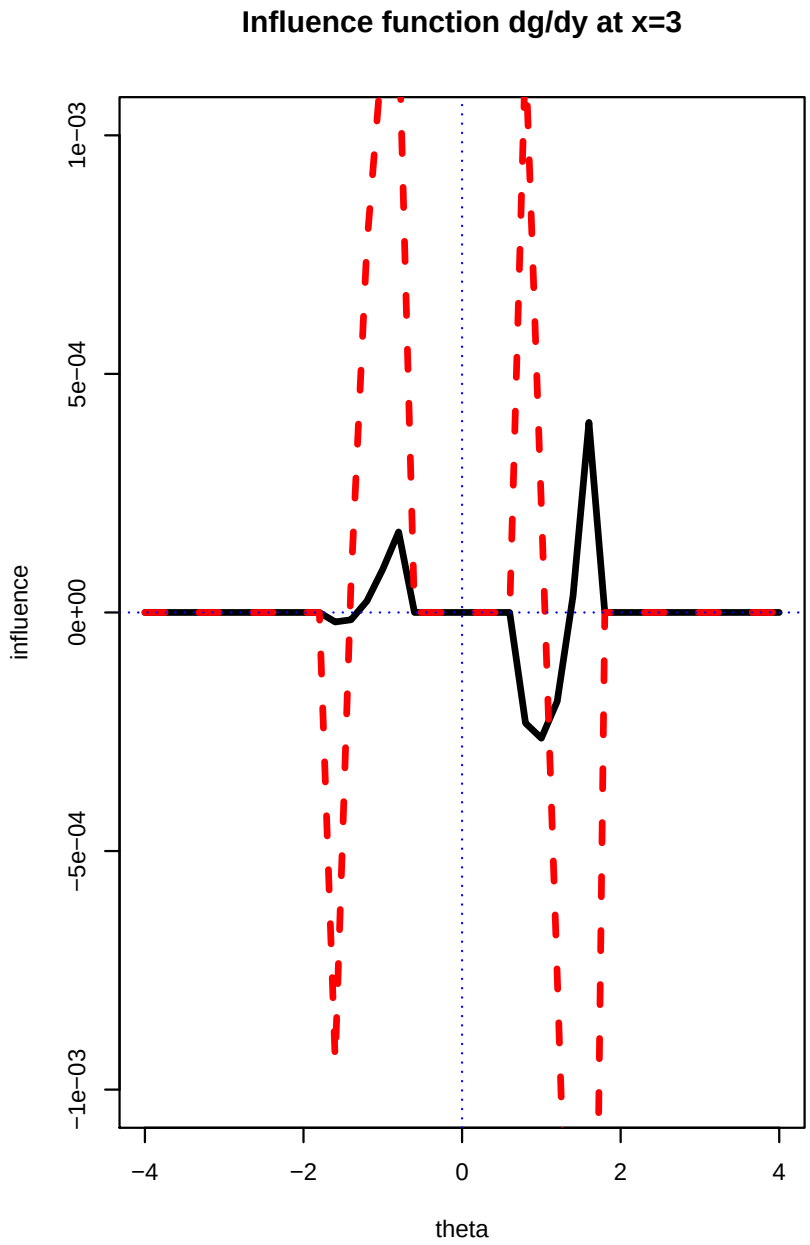
Influence Functions

- Derivative matrix of $\hat{\boldsymbol{g}}$ with respect to $\boldsymbol{y}$
- *g* model $\quad \dfrac{d\hat{\boldsymbol{g}}}{d\boldsymbol{y}} = \left[\operatorname{diag}\left(\boldsymbol{f}_{\hat{\alpha}}\right) - \boldsymbol{f}_{\hat{\alpha}}\boldsymbol{f}'_{\hat{\alpha}}\right] Q \left(\hat{l} + \ddot{s}_{\hat{\alpha}}\right)^{-1} Q' \boldsymbol{W}'_{\hat{\alpha}}$
- *f* model $\quad \tilde{\boldsymbol{g}} = \boldsymbol{k}_\lambda \hat{\boldsymbol{f}}, \qquad \hat{\boldsymbol{g}} = \tilde{\boldsymbol{g}}/\tilde{g}_+,$

  and $\hat{\boldsymbol{f}}$ from `glm(`$\boldsymbol{y} \sim X$`,Poisson)`

$$\frac{d\hat{\boldsymbol{g}}}{d\boldsymbol{y}} = \frac{1}{N}\operatorname{diag}(\hat{\boldsymbol{g}})\left[\operatorname{diag}\left(\frac{1}{\hat{\boldsymbol{g}}}\right) - \frac{1}{\tilde{g}_+}\right]\operatorname{diag}\left(\hat{\boldsymbol{f}}\right)\left[X\left(X'\operatorname{diag}\left(\hat{\boldsymbol{f}}\right)X\right)^{-1}X'\right]$$


Influence function dg/dy at x=0, two towers;
gmod black, fmod red
influence
1e−03
5e−04
0e+00
−5e−04
−1e−03
−4
−2
0
2
4
theta


Influence function dg/dy at x=3
influence
1e−03
5e−04
0e+00
−5e−04
−1e−03
−4
−2
0
2
4
theta

Empirical Bayes Information

- $\text{CoeffVar}\{\hat{g}(\theta)\} \doteq 1/N i_\theta$ $\quad$ ($i_\theta =$ "information in single X")
- *Test case* (g model)

θ	-1	0	1	2	3
i_θ	.0038	.0117	.0030	.0034	.0040
$N_{0.1}$	26,300	8500	32,800	23,100	25,000

($N_{0.1} = \#X_k$ needed for cv$= 0.1$)

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. (2011). Tweedie's formula and selection bias. *J. Amer. Statist. Assoc.* 106: 1602–1614, doi: 10.1198/jasa.2011.tm11181.

Efron, B. (2014). Two modeling strategies for empirical Bayes estimation. *Statist. Sci.* To appear.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. *Ann. Statist.* 19: 1257–1272, doi: 10.1214/aos/1176348248.

Stefanski, L. and Carroll, R. J. (1990). Deconvoluting kernel density estimators. *Statistics* 21: 169–184, doi: 10.1080/02331889008802238.