Nonparametric Inference in Cosmology and Astrophysics: Biases and Variants

Christopher R. Genovese

Department of Statistics

Carnegie Mellon University

http://www.stat.cmu.edu/~genovese/

Collaborators: Larry Wasserman, Peter Freeman, Chad Schafer,

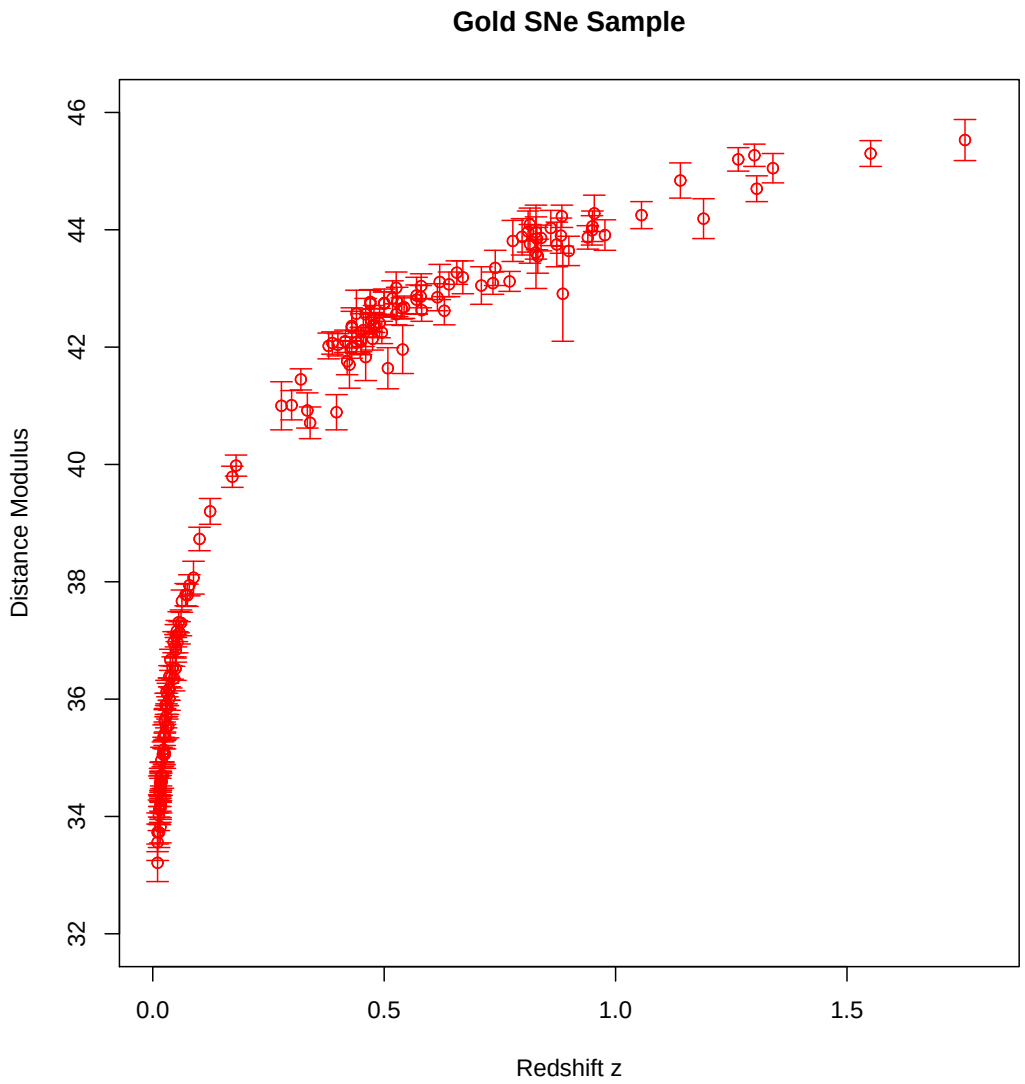
Alex Rojas, and the InCA group (www.incagroup.org)

Department of Statistics

Carnegie Mellon University

This work partially supported by NIH Grant 1 R01 NS047493-01 and NSF Grants ACI 0121671.

Example: Dark Energy



Example: Dark Energy (cont'd)

$$Y_i = r(z_i) + \epsilon_i, \qquad i = 1, \ldots, n.$$

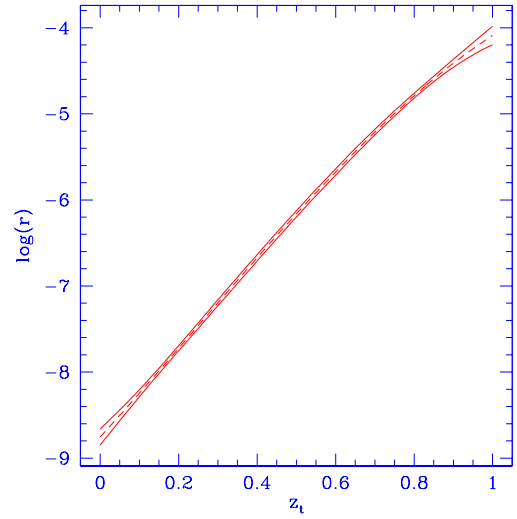
Y_i: distance measure derived from supernova luminosity

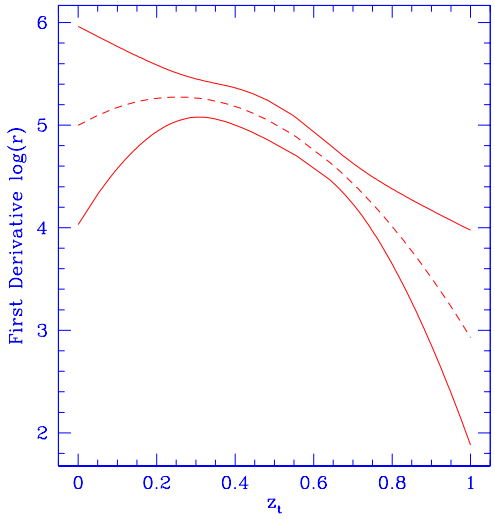
z_i: redshift

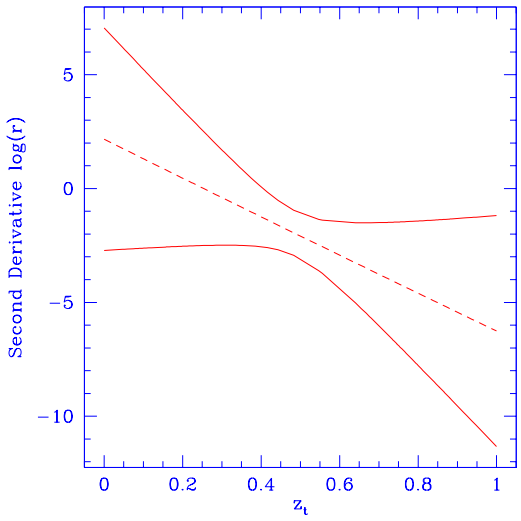
Want to make inferences about the Dark Energy Equation of State

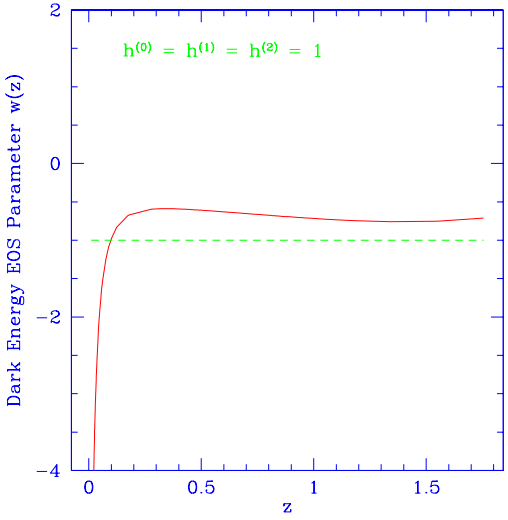
$$\begin{aligned} w(z) &= T(r, r', r'') \\ &= \frac{H_0^2 \Omega_M (1+z)^3 + \frac{2}{3} \frac{r''(z)}{(r'(z))^3}}{H_0^2 \Omega_M (1+z)^3 - \frac{1}{(r'(z))^2}} \end{aligned}$$

Example: Dark Energy (cont'd)

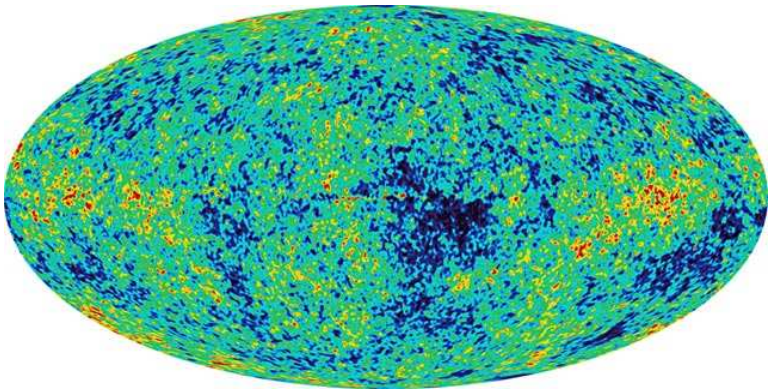


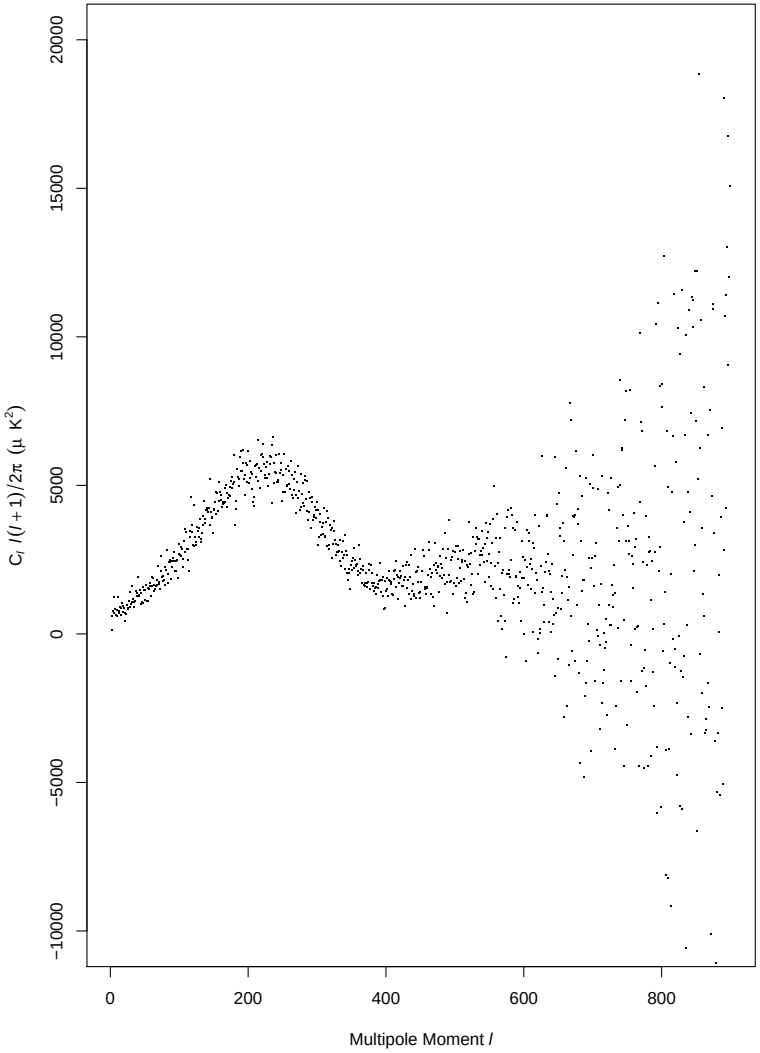


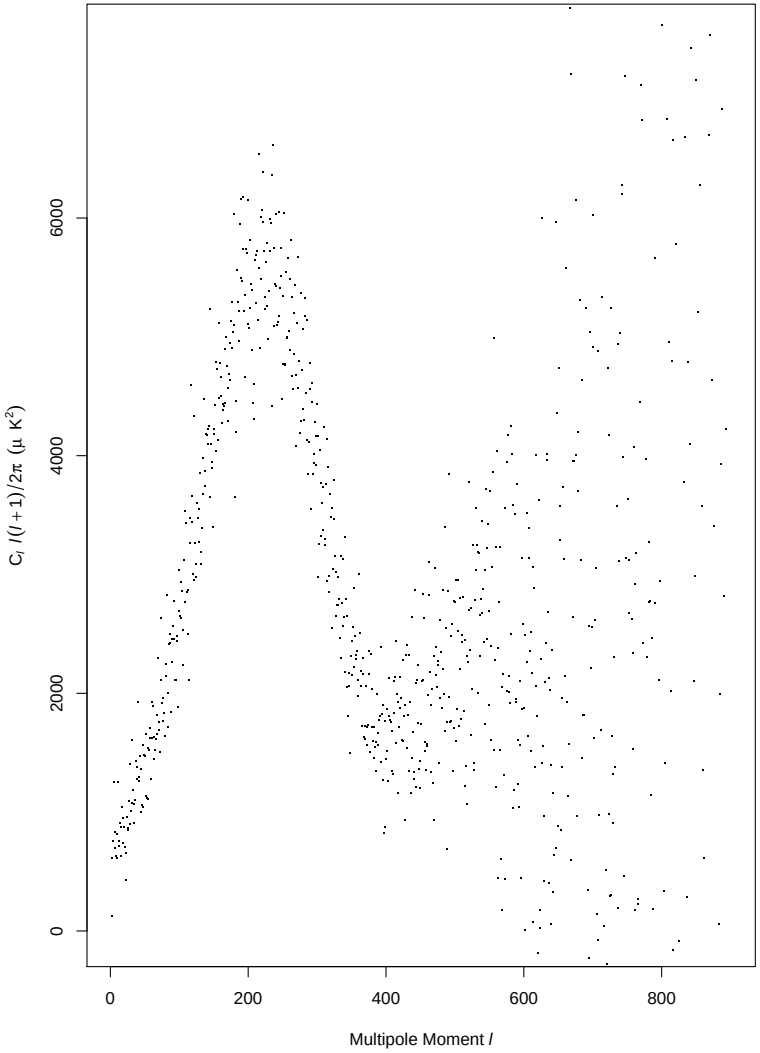




Example: CMB




20000
15000
10000
5000
0
−5000
−10000
$C_l\ /(l+1)/2\pi\ (\mu\ K^2)$
0
200
400
600
800
Multipole Moment l


6000
4000
2000
0
$C_l\ /(l+1)/2\pi\ (\mu\ K^2)$
0
200
400
600
800
Multipole Moment l

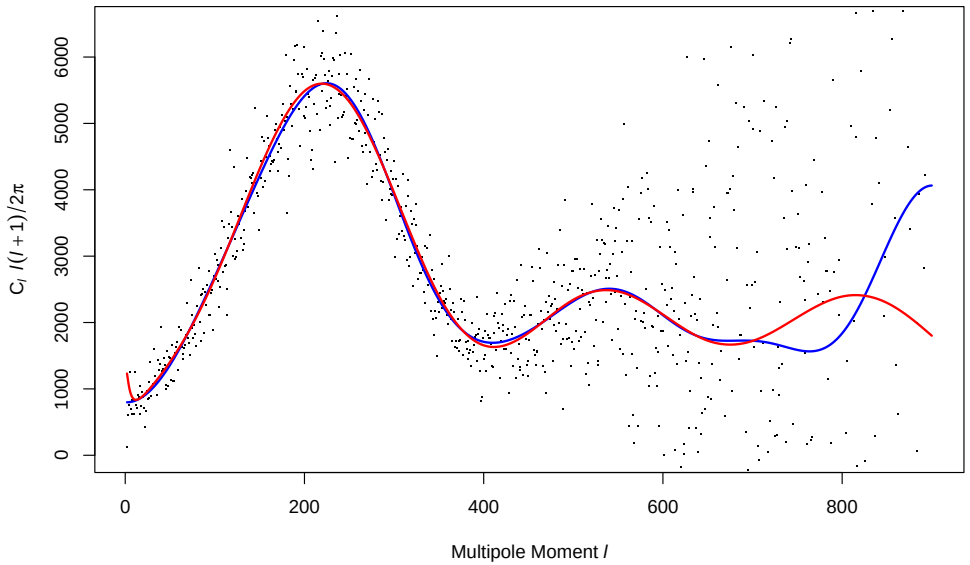
Example: CMB Power Spectrum (cont'd)

$$Y_i = f(x_i) + \epsilon_i, \qquad i = 1, \ldots, n$$

Y_i: estimated power spectrum

x_i: multipole index

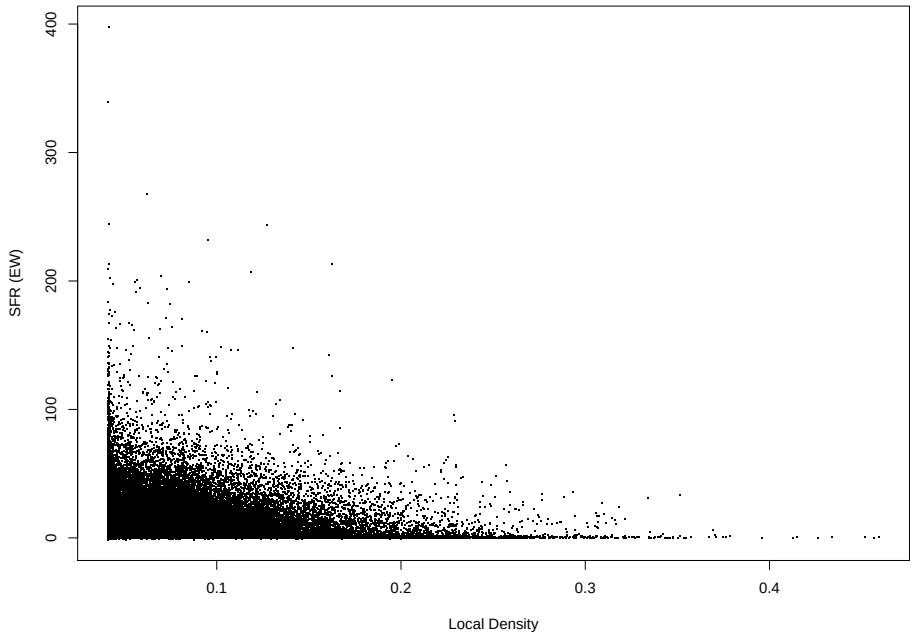
Want to make inferences about features of f such as location and relative heights of peaks.

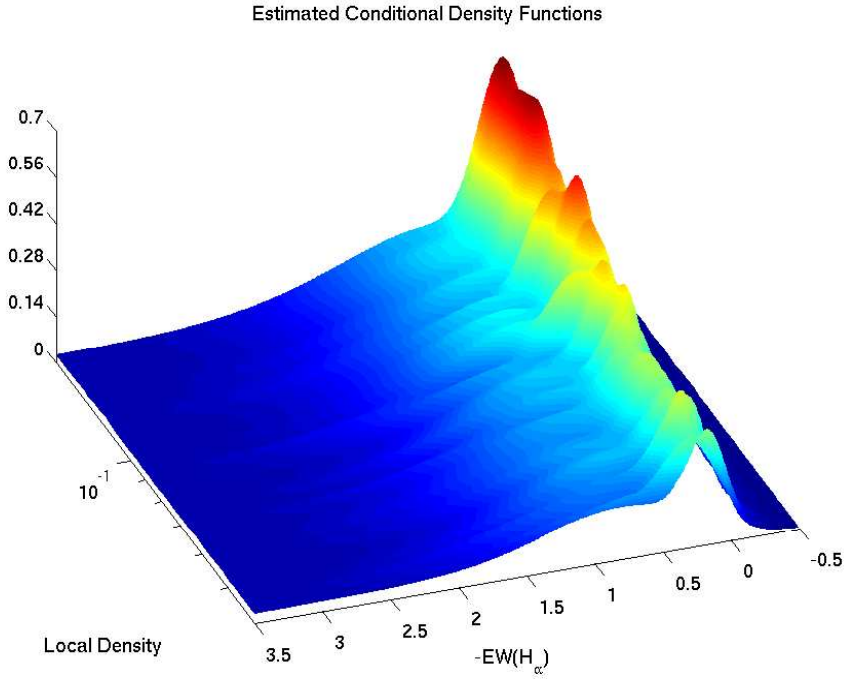


Example: Galaxy Star Formation Rate

A galaxy's evolution is affected by its local environment.

$$\begin{aligned} \text{SFR} \qquad & S_i = f(D_i) + \epsilon_i \\ \text{Density} \qquad & D_i = \widehat{\rho}(x_i) \approx \rho(x_i) + \delta_i. \end{aligned}$$





Why Nonparametric?

1. When we don't have a well-justified parametric (finite-dimensional) model for the object of interest.

2. When we have a well-justified parametric model but have enough data to go after even more detail.

3. When we can do as well (or better) more simply.

4. As a way of assessing sensitivity to model assumptions.

Goal: make sharp inferences about unknown functions with a minimum of assumptions.

This involves estimation procedures, but we also need an accurate assessment of uncertainty.

Road Map

1. Smoothing
2. The Six Biases
3. Confidence Sets

Road Map

1. Smoothing
2. The Six Biases
3. Confidence Sets

The Nonparametric Regression Problem

Observe data (X_i, Y_i) for $i = 1, \ldots, n$ where

$$Y_i = f(X_i) + \epsilon_i,$$

where $\mathsf{E}(\epsilon_i) = 0$ and the X_is can be fixed (x_i) or random.

Leading cases: 1. $x_i = i/n$ and $\mathsf{Cov}(\boldsymbol{\epsilon}) \equiv \Sigma = \sigma^2 I$.

2. X_i IID g and $\mathsf{Cov}(\boldsymbol{\epsilon}) \equiv \Sigma = \sigma^2 I$.

Key Assumption: $f \in \mathcal{F}$ for some infinite dimensional space $\mathcal{F}$.

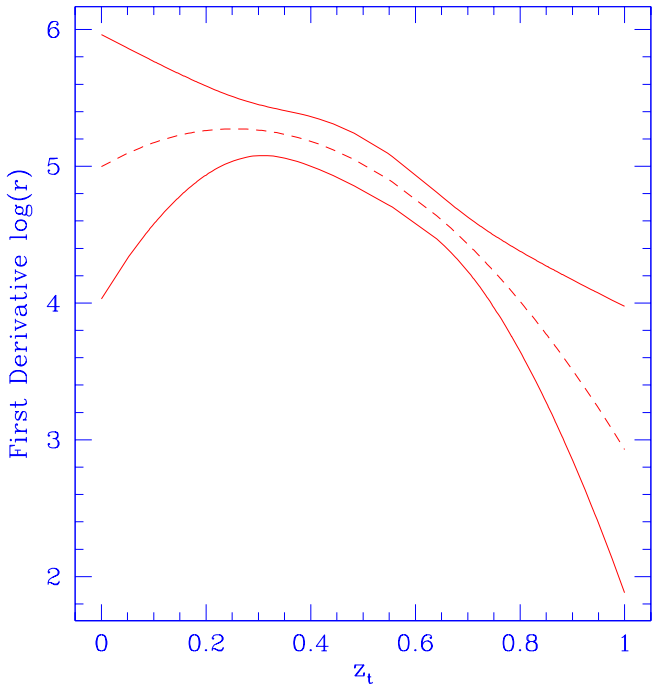
Examples

1. Sobolev: $\mathcal{F} \equiv \mathcal{W}_p(C) = \{f\colon \int |f|^2 < \infty \text{ and } \int |f^{(p)}|^2 \leq C^2\}$
2. Lipschitz: $\mathcal{F} \equiv \mathcal{H}(A) = \{f\colon |f(x) - f(y)| \leq A|x - y|, \text{ for all } x, y\}$

Goal: Make inferences about f or about specific features of f.

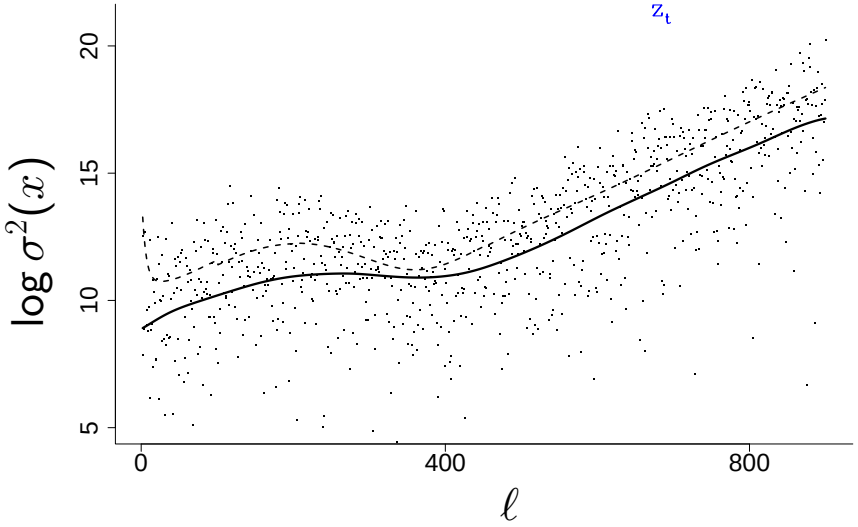
Variants of the Problem

- Inference for Derivatives of f
- Estimating Variance functions
- Regression in High dimensions
- Inferences about specific functionals of f



Related Problems:

- Density Estimation
- Spectral Density Estimation



Rate-Optimal Estimators

Choose a performance measure, or risk function, e.g.,
$R(\widehat{f}, f) = \mathsf{E} \int (\widehat{f} - f)^2$ or $R(\widehat{f}, f) = \mathsf{E}|\widehat{f}(x_0) - f(x_0)|^2)$

Want $\widehat{f}$ that minimizes worst-case risk over $\mathcal{F}$ (minimax).

But typically must settle for achieving the optimal minimax rate of convergence r_n:

$$\inf_{\widehat{f}_n} \sup_{f \in \mathcal{F}} R(\widehat{f}_n, f) \asymp r_n$$

In infinite-dimensional problems, $r_n \sqrt{n} \to \infty$.

For example, $r_n = n^{-\frac{2p}{2p+1}}$ on $\mathcal{W}_p$.

Rate-optimal estimators exist for a wide variety of spaces and risk functions.

Smoothing Methods

- (Quasi-) Linear Methods
  - Kernels and Local Polynomial Regression
  - Roughness Penalty Regularization

    ($\widehat{f}_n = \arg\min_{\xi} \sum_{i=1}^{n} (Y_i - \xi(X_i))^2 + \lambda Q(f)$, e.g., smoothing splines)
  - Basis Decomposition
- Nonlinear Methods
  - Wavelet Shrinkage
  - Variable Bandwidth Kernels

    (Attempt to adapt spatially by changing smoothing over domain; appealing but hard.)
- Others
  - Scale-Space Methods (e.g., SiZer)

    (Consider all levels of smoothing simultaneously.)

(Quasi-) Linear Smoothers

An estimator $\widehat{f}_n$ is a *linear smoother* if there exists functions $s(x) = (s_1(x), \ldots, s_n(x))$ such that

$$\widehat{f}_n(x) = \sum_{j=1}^{n} s_j(x) Y_j.$$

Writing $\widehat{\boldsymbol{f}}_{\boldsymbol{n}} = (\widehat{f}_n(x_1), \ldots, \widehat{f}_n(x_n))$ and $S_{ij} = s_j(x_i)$, we have

$$\widehat{\boldsymbol{f}}_{\boldsymbol{n}} = S\boldsymbol{Y}.$$

For nonparametric smoothers, the function $s(x) \equiv s_h(x)$ depends on a free *smoothing parameter* h that governs complexity of $\widehat{f}$.

The smoothing parameter must be selected, usually from the data. A quasi-linear smoother is linear conditional on h.

(Quasi-) Linear Smoothers: Kernels

The Nadaraya-Watson Kernel estimator, for suitable function K and bandwidth h, is

$$\widehat{f}_n(x) = \frac{\sum_{i=1}^n K\left(\frac{X_i - x}{h}\right) Y_i}{\sum_{j=1}^n K\left(\frac{X_j - x}{h}\right)} = \sum_{i=1}^n s_i(x)\, Y_i.$$

This weighted average is defined by a *local* optimization problem: for each x, find $a_0 \equiv a_0(x)$ to minimize

$$\sum_{i=1} K\left(\frac{X_i - x}{h}\right) (Y_i - a_0)^2.$$

This is fitting a local constant.

(Quasi-) Linear Smoothers: Local Polynomials

Consider higher order approximation for u near x:

$$f(u) \approx a_0 + a_1(u-x) + \cdots + a_p \frac{(u-x)^p}{p!} \equiv p_x(u;a).$$

Let $\widehat{a} \equiv \widehat{a}(x)$ minimize

$$\sum_{i=1} K\left(\frac{X_i - x}{h}\right) (Y_i - p_x(X_i;a))^2.$$

Then,

$$\widehat{f}_n(x) = p_x(x;\widehat{a}) = \widehat{a}_0(x).$$

The $p = 0$ case gives back the kernel estimator.

Remarks on Local Polynomials

- Although $\widehat{f}_n(x) = \widehat{a}_0(x)$, this is not simply fitting a local constant.
- Performance is insensitive to the choice of kernel K but highly sensitive to the choice of smoothing parameter h.
- Local polynomial estimators for $p > 0$ automatically correct some of the biases inherent in the kernel estimator.
- Work well for estimating derivatives. For estimating the vth derivative, prefer $p - v$ odd.

Nonlinear Smoothers: Wavelet Shrinkage

Wavelets provide a sparse representation for a wide class of functions.

$$f = \sum_k \alpha_{J_0,k}\varphi_j + \sum_{j=J_0}^{\infty}\sum_k \beta_{jk}\psi_{jk}$$
$$= \text{Coarse Part} + \text{Successively Refined Details}$$

Schematic:

1. Fast Discrete Wavelet Transform (DWT): $\begin{pmatrix}\widetilde{\alpha}\\ \widetilde{\beta}\end{pmatrix} = W\mathbf{Y}$.

2. Nonlinear shrinkage of detail coefficients: $\widehat{\beta} = \eta_\lambda(\widetilde{\beta})$. For example,

$$\widehat{\beta} = \text{sgn}(\widetilde{\beta})\left(|\widetilde{\beta}| - \lambda\right)_+.$$

3. Invert Transform: $\widehat{\boldsymbol{f}} = W^{-1}(\widetilde{\alpha}, \widehat{\beta})^T$.

Nonlinear Smoothers: Wavelet Shrinkage (cont'd)

Result: Adaptive Estimators

The same procedure is nearly optimal (asymptotic minimax) across a range of assumed spaces.

Current State of the Art: Johnstone and Silverman (2005)

– Put mixture prior on each β

– Posterior median yields threshold

– Works with translation invariant transform

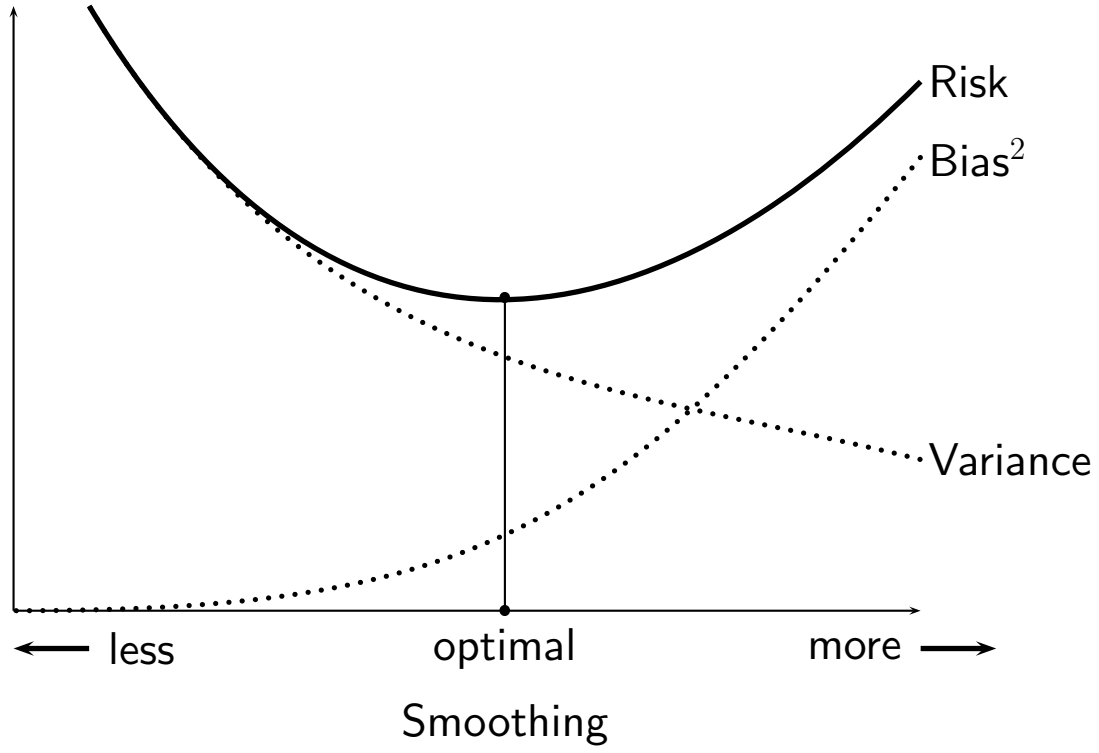
– Handles boundaries, estimates derivatives, excellent performance.

The Bias-Variance Tradeoff

All of these methods have a smoothing parameter that determines the complexity of the fit.

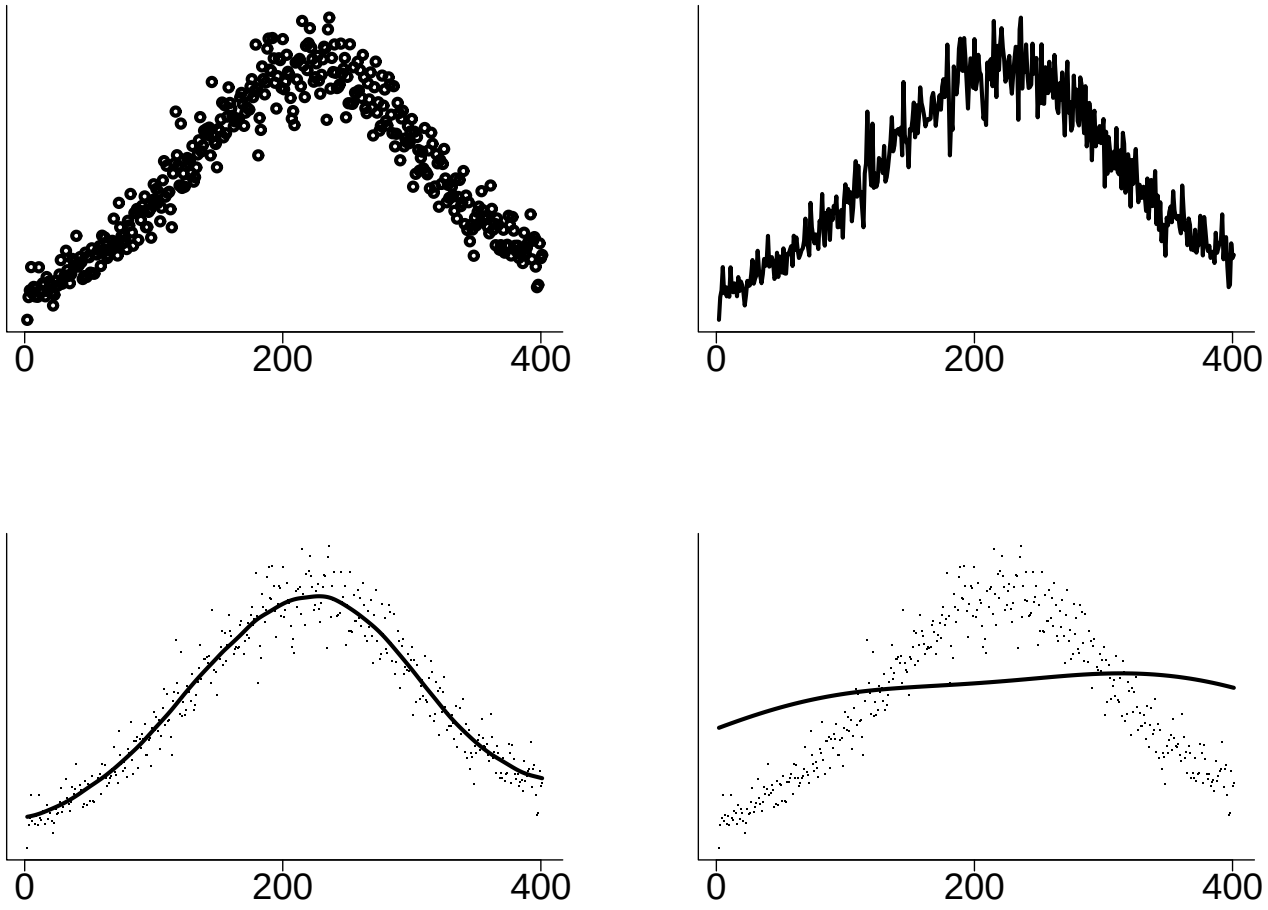
Key goal: Choose the correct level of smoothing.

$$R(\widehat{f}, f) = \mathsf{E} \int (f(x) - \widehat{f}(x))^2 \, dx = \int \mathsf{bias}^2(x) \, dx + \int \mathsf{variance}(x) \, dx$$



The Bias-Variance Tradeoff (cont'd)

Example: CMB Spectrum



Choosing the Smoothing Parameter

1. Plug-in

   Asymptotically optimal bandwidths are often of the form $h_* = C(f)r_n$ for $r_n \to 0$.

   Then, $\widehat{h}_{\mathrm{PI}} = C(\widetilde{f})r(n)$ for a pilot estimate $\widetilde{f}$.

   Do not perform well in general (Loader 1999).

2. Cross-Validation

   Divide $\{1, \ldots, n\}$ into m disjoint subsets $S_1, \ldots, S_m$.

   Let $\widehat{f}^{[-\ell]}$ be the estimate obtained with *the same procedure* but omitting the subset of data (X_i, Y_i) with $i \in S_\ell$.

Choosing the Smoothing Parameter (cont'd)

Then $\widehat{\mathsf{PE}}(h) = \frac{1}{m}\sum_{\ell=1}^{m}\frac{1}{\#(S_\ell)}\sum_{i\in S_\ell}(Y_i - \widehat{f}^{[-\ell]}(X_i))^2$.

When $m = n$ and $S_\ell = \{\ell\}$, this becomes

$$\widehat{\mathsf{PE}}(h) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \widehat{f}(X_i)}{1 - S_{ii}(h)}\right)^2,$$

where $S(h)$ is the smoothing matrix.

Choose $\widehat{h}$ to minimize $\widehat{\mathsf{PE}}(h)$. (There are several variants.)

Then, $\mathsf{PE}(\widehat{h}) \leq c_1 \inf_h \mathsf{PE}(h) + o(1)$.

Choosing the Smoothing Parameter (cont'd)

3. Risk Estimation

   Would like h to minimize the risk $R(\widehat{f}_h, f) = \int (\widehat{f}_h - f)^2$.

   Find $\widehat{R}(h)$ such that $\mathsf{E}\widehat{R}(h) = R(\widehat{f}_h, f)$.

   Choose $\widehat{h}$ to minimize $\widehat{R}(h)$.

   In many cases, can show that this approximates true minimum.

   Example: Basis coefficients $\widehat{f}_h = \sum_{k=0}^{h} \beta_k \phi_k$, $h = 0, \ldots n$.

$$R(\widehat{f}_h, f) = h\frac{\sigma^2}{n} + \sum_{k=h+1}^{n} \beta_k^2 \longleftrightarrow \widehat{R}(h) = h\frac{\sigma^2}{n} + \sum_{k=h+1}^{n} (\beta_k^2 - \frac{\sigma^2}{n}).$$

Whatever the method, $\dfrac{\widehat{h}_n - h_n}{h_n} \to 0$ slowly.

Estimating the tuning parameter is an intrinsically hard problem.

Road Map

1. Smoothing
2. The Six Biases
3. Confidence Sets

The Six Biases

1. Model Bias
2. Design Bias
3. Boundary Bias
4. Derivative Bias
5. Measurement-Error Bias
6. Coverage Bias

Most of these can be dealt with well-understood statistical techniques.

Model Bias

Without assumptions, there can be no conclusions.
– John Tukey

The performance of nonparametric procedures can depend on the space $\mathcal{F}$ assumed to contain f.

Choosing $\mathcal{F}$ too restrictively incurs a bias from the un-modeled component of f; choosing $\mathcal{F}$ too expansively can lead the procedure's to be driven by implausible cases.

And how can we justify the abstract choice of a particular space or ball size etc. in terms of concrete data and prior information?

Model Bias (cont'd): Adaptive Estimators

It's unsatisfying to depend too strongly on intangible assumptions such as whether $f \in \mathcal{W}_p(C)$ or $f \in \mathcal{H}(A)$.

Instead, we want procedures to *adapt* to the unknown smoothness.

For example, $\widehat{f}_n$ is a *(rate) adaptive procedure* over the $\mathcal{W}_p$ spaces if when $f \in \mathcal{W}_p$

$$\widehat{f}_n \to f \text{ at rate } n^{-2p/2p+1}$$

without knowing p.

Rate adaptive estimators exist over a variety of function families and over a range of norms (or semi-norms).

Adaptive confidence sets? (later)

Design Bias

Some procedures affected by clustering or positioning of X_is.

Example: Kernel and local linear estimators have

$$\text{variance}(x) = \frac{1}{nh_n} \frac{\sigma^2(x) \int K^2(u)du}{g(x)} + o_P\left(\frac{1}{nh_n}\right).$$

But, the kernel estimator $p = 0$ has bias

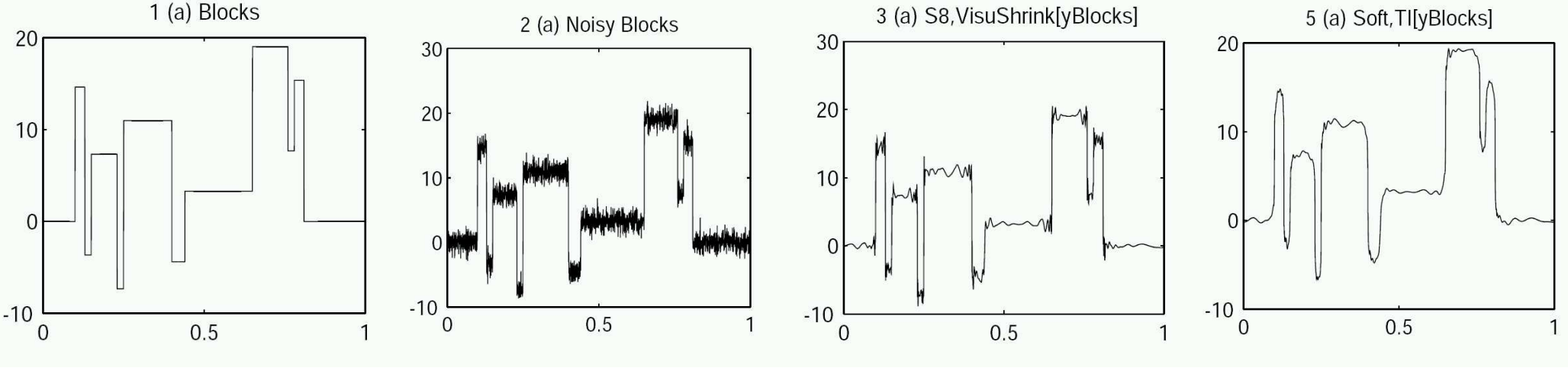
$$h_n^2 \left(\frac{1}{2} f''(x) + \underbrace{\frac{f'(x)g'(x)}{g(x)}}_{\text{design bias}} \right) \int u^2 K(u)du + o_P(h^2)$$

whereas the local linear estimator $p = 1$ has asymptotic bias

$$h_n^2 \left(\frac{1}{2} f''(x) \right) \int u^2 K(u)du + o_P(h^2).$$

Design Bias (cont'd)

Another Example: Wavelets and choice of origin



from Coifman and Donoho (1995)

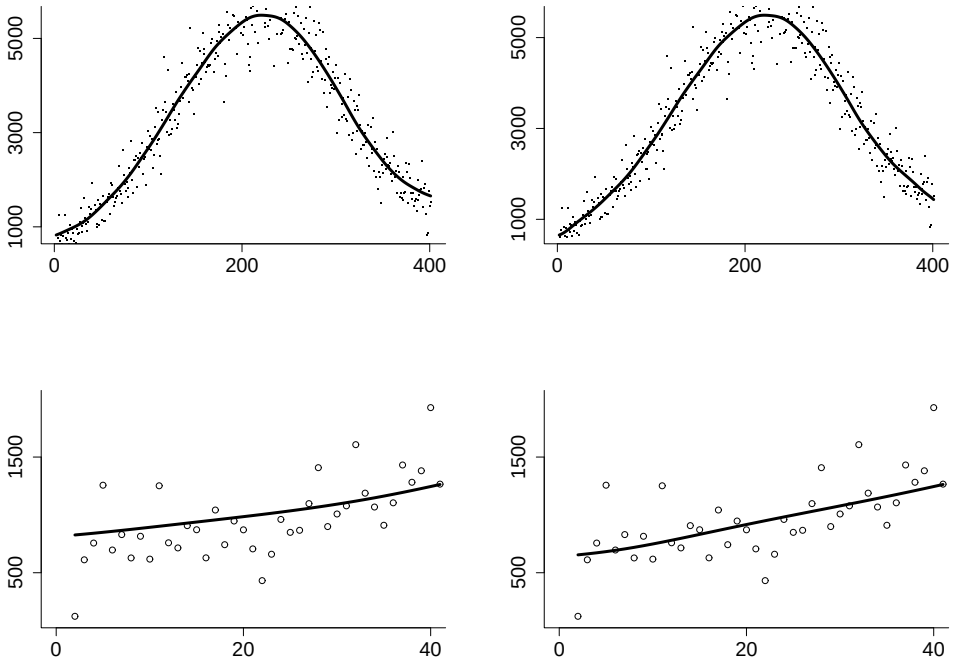
One solution: Translation Invariant wavelet transform

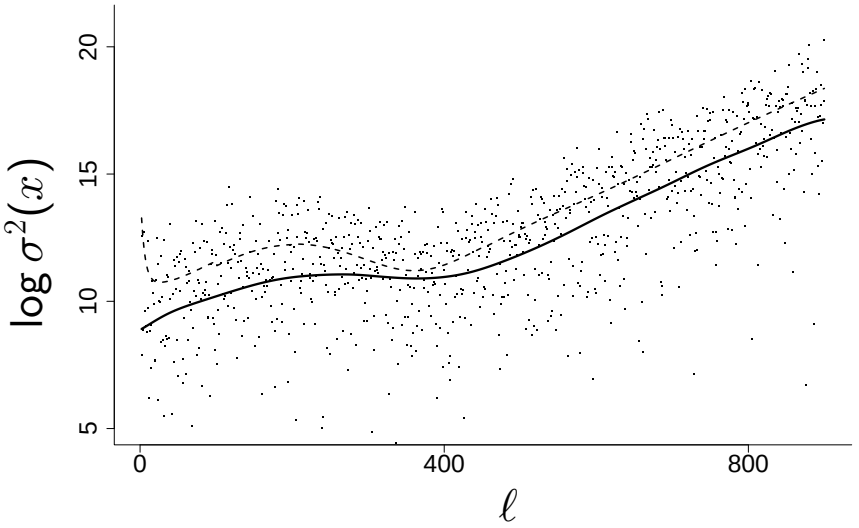
$$\widehat{f} = \frac{1}{n}\sum_{\Delta=1}^{n} \text{Shift}_{\Delta}^{-1} \circ \text{DWT}^{-1} \circ \text{Threshold} \circ \text{DWT} \circ \text{Shift}_{\Delta},$$

where Shift_{Δ} is a circular shift of an n vector by Δ components.

Boundary Bias

Many methods exhibit nontrivial biases near data boundaries.





This is potentially important in problems such as the CMB and Dark Energy because boundary behavior is of scientific interest.

Boundary biases get worse in high-dimensions because a greater proportion of points lie near boundaries.

Boundary Bias (cont'd)

Example: Kernel Smoothing

The kernel "hits" fewer points near the boundary.

Local polynomial for $p \geq 1$ is an easy fix, called "automatic boundary carpentry".

Example: Periodic Wavelets

Standard orthonormal wavelets on an interval are periodic, but can lead to significant bias when the function is not periodic.

Cohen, Daubechies, Jawerth and Vial (1993) construction: a preconditioning step that creates boundary wavelets.

Derivative Bias

When estimating derivatives f', f'', ..., $f^{(d)}$, note that $(\widehat{f}_n)^{(d)}$ is NOT a good estimator of $f^{(d)}$.

Illustration: Let $\phi_0(x) = 1$ and $\phi_k(x) = \sqrt{2}\cos(\pi k x)$, for $k \geq 1$.

If $f = \sum_{k \geq 0} \beta_k \phi_k$ on $[0, 1]$, then $f'' = \sum_{k > 0} -\pi^2 k^2 \beta_k \phi_k$.

Let $\widehat{\beta}_k = \frac{1}{n} \sum_{i=0}^{n} Y_i \phi_k(x_i) \approx N(\beta_k, \sigma^2/n)$.

Then, $\widehat{\beta}_k^{(d)} \equiv -\pi^2 k^2 \widehat{\beta}_k \approx N(-\pi^2 k^2 \beta_k, \pi^4 k^4 \frac{\sigma^2}{n})$. Ouch.

Inference for derivatives requires different levels of smoothing.

Derivative Bias (cont'd)

With local polynomial estimation, this takes a manageable form.

Recall $p_x(u; a) = a_0 + a_1(u - x) + \cdots + a_p \dfrac{(u-x)^p}{p!}$

and $\widehat{f}_n(x) = \widehat{a}_0(x)$.

Then, $\widehat{f}_n^{(d)}(x) = \widehat{a}_d(x)$. Note $\widehat{a}_0^{(d)} \neq \widehat{a}_d$.

Choose $p - d$ odd (e.g., $p = d + 1$) to avoid boundary and design bias.

The (asymptotically) optimal bandwidth for d derivatives is then

$$h_d(k) = C(k, p)\, h_0$$

where $C(k, p)$ is a known constant.

What about confidence sets? There it's not so easy.

Example: Dark Energy.

Measurement-Error Bias

Errors on the covariates can lead to counter-intuitive effects.

Simplified Example: Galaxy Star Formation Rates versus Local Density

$$\begin{array}{ll} \text{SFR} & S_i = f(D_i) + \epsilon_i \\ \text{Density} & D_i = \widehat{\rho}(x_i) \approx \rho(x_i) + \delta_i. \end{array}$$

Common result is *attenuation bias*, but the opposite is possible (Carroll, Ruppert, and Stefanski 1995).

Bias depends on the nature and extent of the errors.

Measurement-Error Bias

Observe:

$$Y_i = f(X_i) + \epsilon_i$$
$$\widetilde{X}_i = X_i + U_i$$

If we use the $(\widetilde{X}_i, Y_i)$ to estimate f, the estimator will be inconsistent

$$\int (\widehat{f}_n(x) - f)^2 \not\to 0.$$

Not enough to simply increase the error bars on Y.

The extra bias $\sigma_U^2 \left(\frac{g'(x)}{g(x)} f'(x) + \frac{f''(x)}{2} \right) \not\to 0.$

Measurement-Error Bias (cont'd)

Toy Illustration of Attenuation Bias:

$Y_i = \beta X_i + \epsilon_i$, but we observe (W_i, Y_i) where $W_i = X_i + U_i$.

Then, least squares estimate consistent for $\widetilde{\beta} = \beta \dfrac{\sigma_X^2}{\sigma_X^2 + \sigma_U^2} < \beta$.

Also $\text{Var}(Y \mid W) = \sigma_\epsilon^2 + \beta^2 \dfrac{\sigma_X^2 \sigma_U^2}{\sigma_X^2 + \sigma_U^2}$.

This is not just an issue of variance – *both* variance and bias are affected.

Measurement-Error Bias (cont'd)

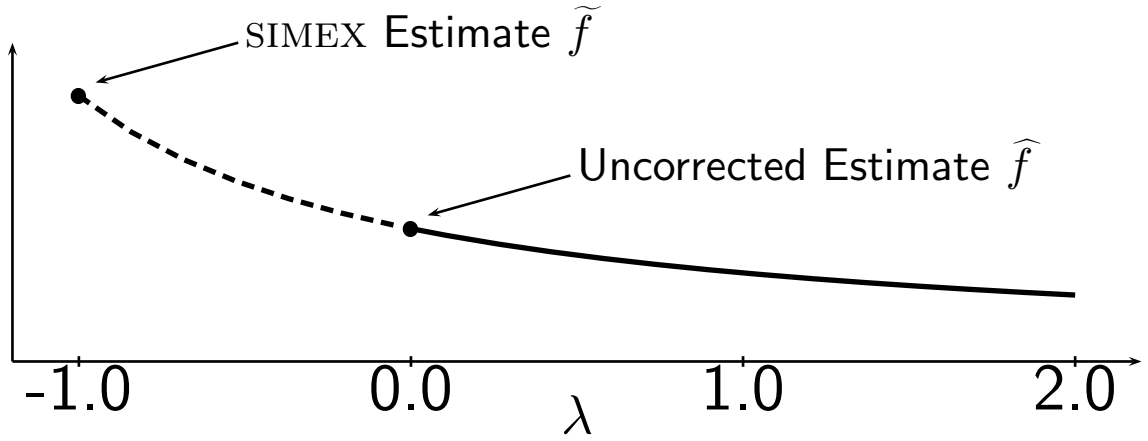
One solution: SIMEX (Cook and Stefanski 1994)

Key idea: determine effect of error empirically via simulation.

Toy Illustration Revisited: Given data sets with measurement error variances by factors of $(1 + \lambda_m)$ for $0 = \lambda_1 < \lambda_1 < \cdots < \lambda_m$

Get a regression for $r(\lambda) = \frac{\beta^2\sigma_X^2}{\sigma_X^2+(1+\lambda)\sigma_U^2}$. Want to estimate $r(-1)$.

In general, new data sets made from old by adding noise.



Measurement-Error Bias (cont'd)

Another Solution: Special kernels (Fan and Truong 1990)

$$\widehat{f}_n(x) = \frac{\sum_{i=1}^n K_n\left(\frac{x-\widetilde{X}_i}{h_n}\right) Y_i}{\sum_{i=1}^n K_n\left(\frac{x-\widetilde{X}_i}{h_n}\right)}$$

where

$$K_n(x) = \frac{1}{2\pi}\int e^{-itx}\frac{\phi_K(t)}{\phi_U(t/h_n)}dt,$$

where ϕ_K is the Fourier transform of a kernel K and ϕ_U is the characteristic function of U.

This is a standard kernel estimator except for the unusual kernel K_n, chosen to reduce bias.

How to do confidence bands/spheres for this case? We are currently working on that.

Coverage Bias

Using a rate-optimal smoothing parameter gives

$$\text{bias}^2 \approx \text{var}.$$

Loosely, if $\widetilde{f} = \mathsf{E}\widehat{f}$ and $s = \sqrt{\mathsf{Var}\,\widehat{f}}$, then

$$\frac{\widehat{f} - f}{s} = \frac{\widehat{f} - \widetilde{f}}{s} + \frac{\widetilde{f} - f}{s} \approx \mathsf{N}(0,1) + \frac{\text{bias}}{\sqrt{\text{var}}}.$$

So, "$\widehat{f} \pm 2s$" undercovers.

Two common solutions in the literature:

– Bias Correction: Shift confidence set by estimated bias.

– Undersmoothing: Smooth so that var dominates bias^2.

Road Map

1. Smoothing
2. The Six Biases
3. Confidence Sets

Confidence Sets

In practice, we usually need more than $\widehat{f}$.

We want to make *inferences* about features of f:
shape, magnitude, peaks, inclusion, derivatives.

One approach: construct a $1-\alpha$ confidence set for f,
a random set $\mathcal{C}$ such that $\mathsf{P}\Big\{\mathcal{C} \ni f\Big\} = 1-\alpha$.

Usually $\mathcal{C}$ is a ball or a band around some $\widehat{f}$.

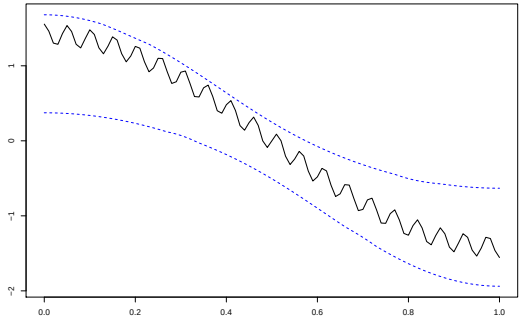
Three challenges:

1. Bias
2. Simultaneity
3. Relevance

Relevance

In small samples, confidence balls and bands need not constrain all features of interest.

For example, number of peaks:



Alternative: confidence intervals for *specific functionals* of f

Two practical problems:

1. Many relevant functionals (e.g., peak locations) hard to work with.
2. One often ends up choosing functionals post-hoc.

Better to obtain construct a confidence set for the whole object with post-hoc protection for inferences about many functionals.

Remark: What We Want

1. For asymptotic confidence procedures, prefer uniform coverage:

$$\sup_{f \in \mathcal{F}} \left| \mathsf{P}\Big\{ \mathcal{C}_n \ni f \Big\} - (1 - \alpha) \right| \to 0.$$

This ensures that the coverage error depends only on n, not on f.

2. We would also like adaptive confidence procedures.

That is, maintain coverage on $\mathcal{F}$ but can use the data to tailor the set's diameter to the unknown f.

Confidence Bands Cannot Adapt

Confidence bands are of the form $\mathcal{C} = \{f\colon\ L \leq f \leq U\}$ for some random functions L and U.

Unfortunately, confidence bands whose width is determined from the data cannot do better than fixed-width bands.

Let $\mathcal{D}$ denote fixed-diameter confidence band. Then,

$$\liminf_{n\to\infty} \frac{\inf_{\mathcal{C}} \inf_{f\in\mathcal{F}} \mathsf{E}_f(s_n(\mathcal{C}))}{\inf_{\mathcal{D}} \inf_{f\in\mathcal{F}} \mathsf{E}_f(s_n(\mathcal{D}))} > 0.$$

This continues to hold (Low 1997, Genovese and Wasserman 2005) even when smoothness constraints are imposed.

Bottom line: For commonly used smoothers, neither the width nor the tuning parameter of the optimal confidence bands depends on the data.

Building Confidence Bands: Volume of Tubes

If $\widehat{f}(x) = \sum_{i=1}^n \ell_i(x) Y_i$, for weights $\ell(x) = (\ell_1(x), \ldots, \ell_n(x))$, then

$$\inf_{f \in \mathcal{F}} \mathsf{P}\Big\{ \widehat{f}(x) - c\widehat{\sigma}\, \|\ell(x)\| \leq f(x) \leq \widehat{f}(x) + c\widehat{\sigma}\, \|\ell(x)\|, \forall x \Big\} = 1 - \alpha,$$

for suitable class $\mathcal{F}$ (Sun and Loader 1994). The constant c solves the equation $\alpha = K_\ell \phi(c) + 2(1 - \Phi(c))$.

Special case: $f(x) = \sum_{i=1}^n \ell_i(x)\theta_i$.

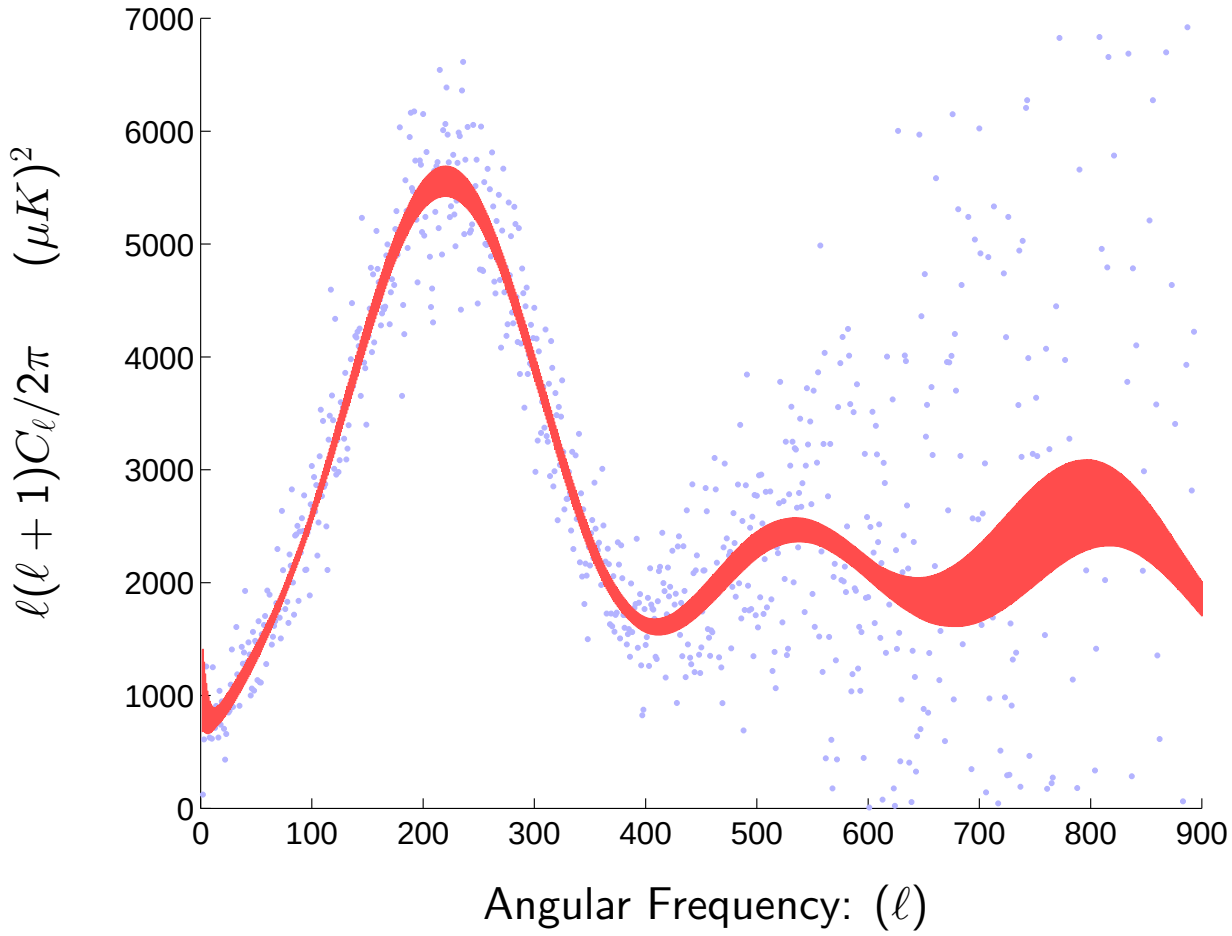
Then, $|\widehat{f}(x) - f(x)| = \left|\sum_{i=1}^n \ell_i(x)\epsilon_i\right| = |\langle \ell(x), \epsilon \rangle|$, so

$$\alpha = \mathsf{P}\left\{ \sup_x \left| \frac{\widehat{f}(x) - f(x)}{\|\ell(x)\|} \right| > c\sigma \right\} = \mathsf{P}\left\{ \sup_x \left| \left\langle \frac{\ell(x)}{\|\ell(x)\|}, \frac{\epsilon}{\|\epsilon\|} \right\rangle \right| > \frac{c\sigma}{\|\epsilon\|} \right\}.$$

Reduces to finding the volume of a tube on the sphere S^{n-1}.

Building Confidence Bands: Parametric

Approximately minimum expected size parametric confidence bands (Schafer 2004)



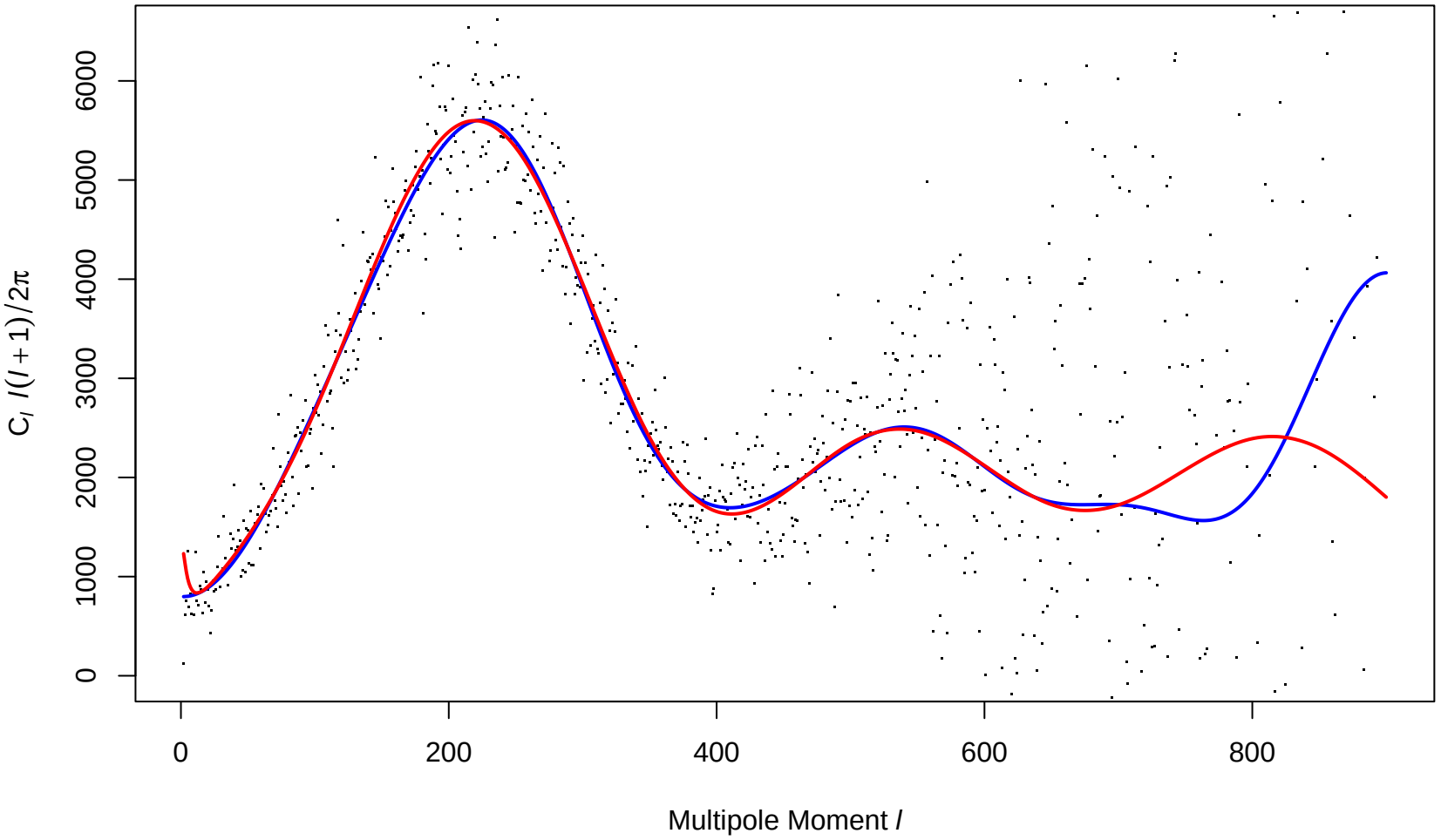
Confidence Balls

1. Expand $f = \sum_k \beta_k \phi_k$ in orthonormal basis (e.g., cosine basis).
2. Shrink naive estimators by $\widehat{\beta}_k = \lambda_k \widetilde{\beta}_k$, $1 \geq \lambda_1 \geq \cdots \geq \lambda_n$.
3. Choose λ by minizing estimated risk $\widehat{R}(\lambda)$.
4. $$\mathcal{C}_n = \left\{ \beta : \sum_k (\widehat{\beta}_k - \beta_k)^2 \leq \frac{k_\alpha}{\sqrt{n}} + \widehat{R}(\widehat{\lambda}) \right\}.$$

Confidence balls *can* adapt to unknown smoothness.

Can impose constraints post-hoc and get valid inferences.

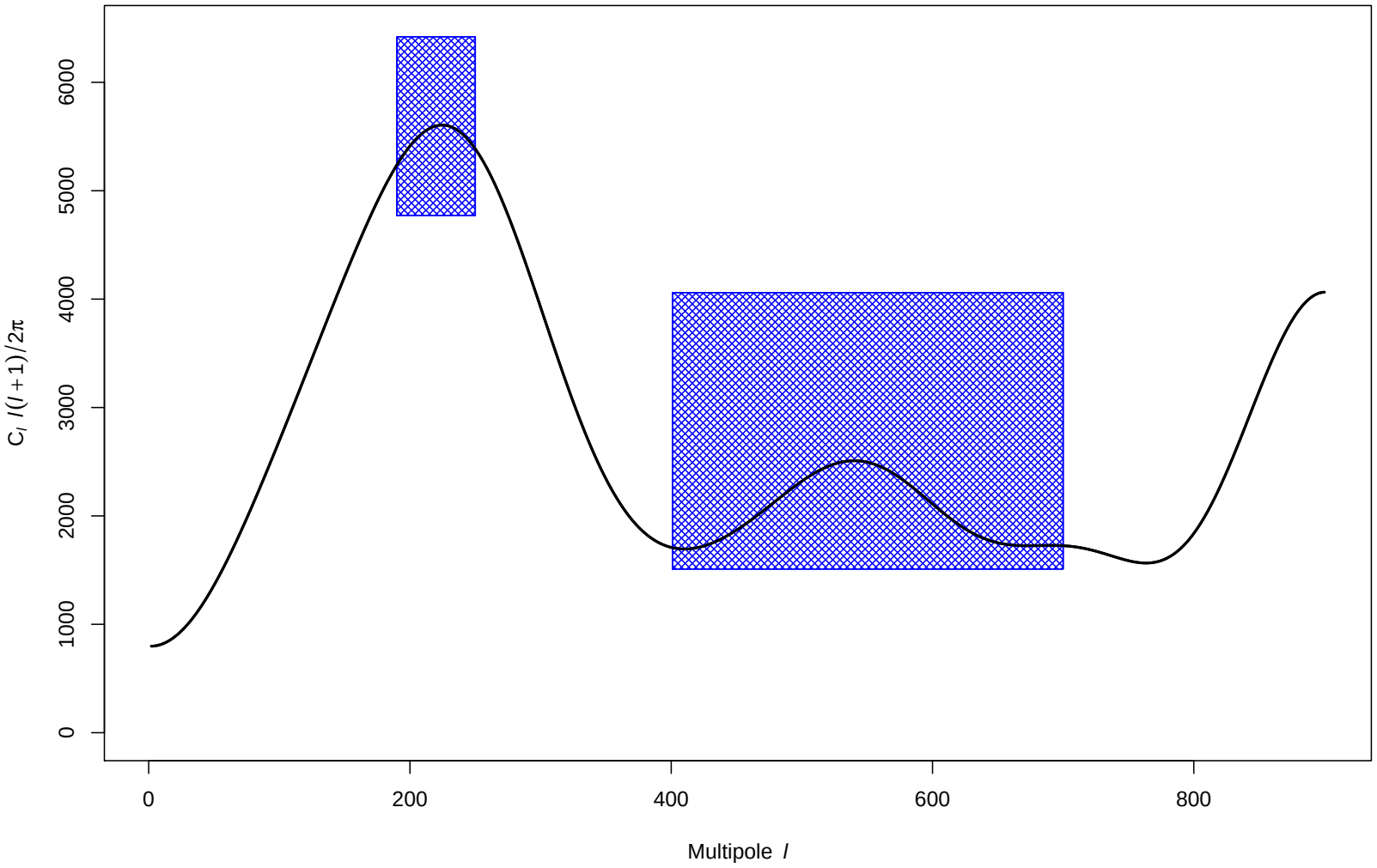
Confidence Ball Center vs Concordance Model



- Concordance model is an MLE based on WMAP and four other data sets.
- Confidence ball center based on WMAP data only.

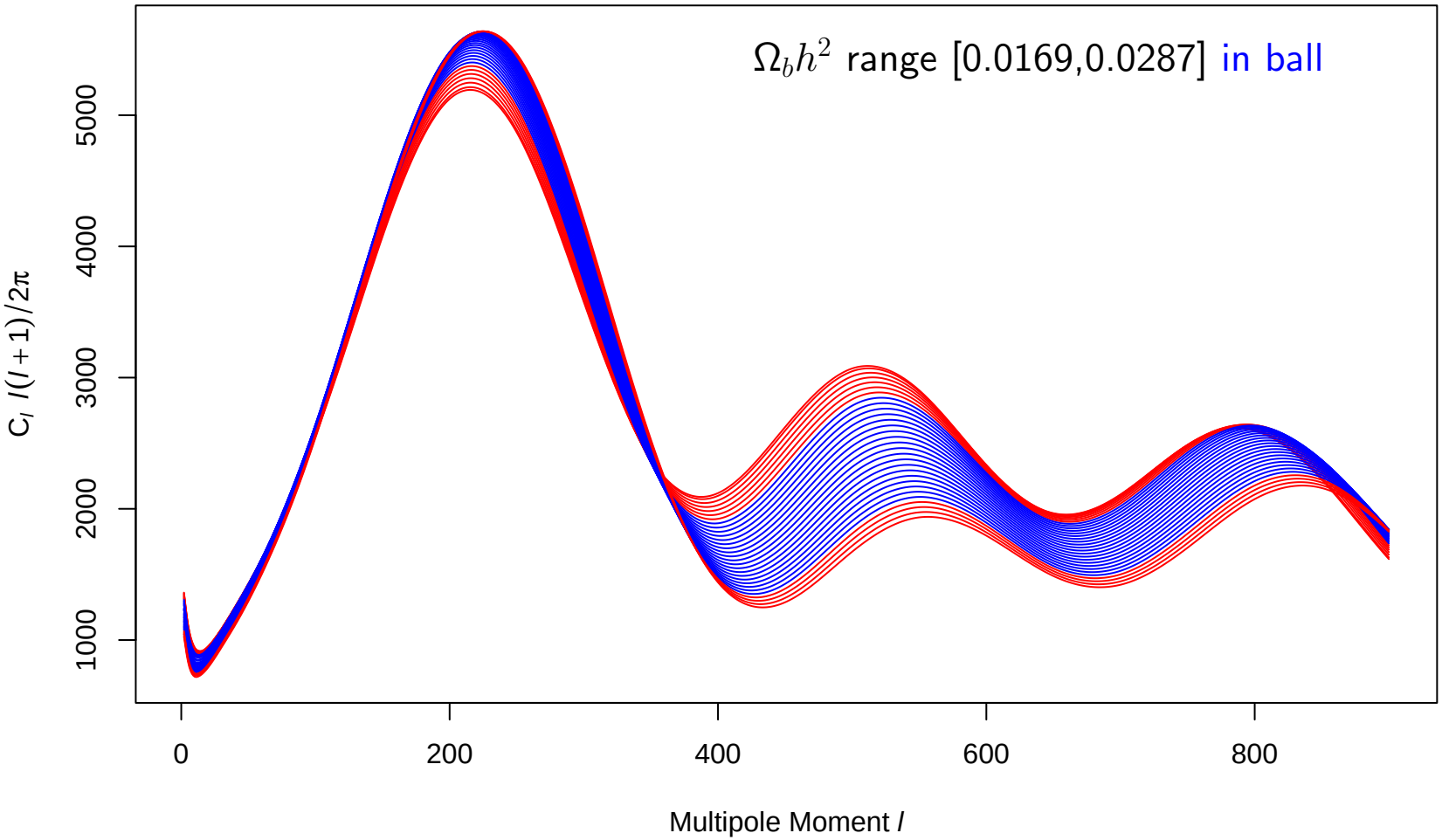
Eyes on the Ball I: Parametric Probes

Simultaneous 95% CIs on Peak Heights, Locations, and Height Ratios



Eyes on the Ball I: Parametric Probes (cont'd)

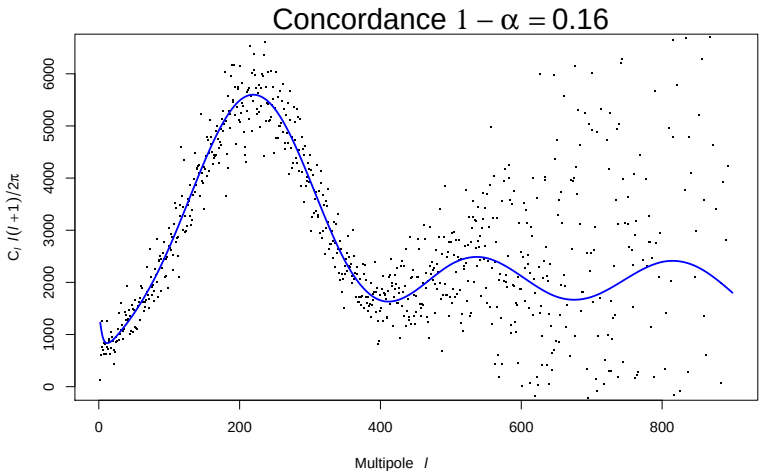
Varied baryon fraction ($\Omega_b h^2$) keeping $\Omega_{\rm total} \equiv 1$

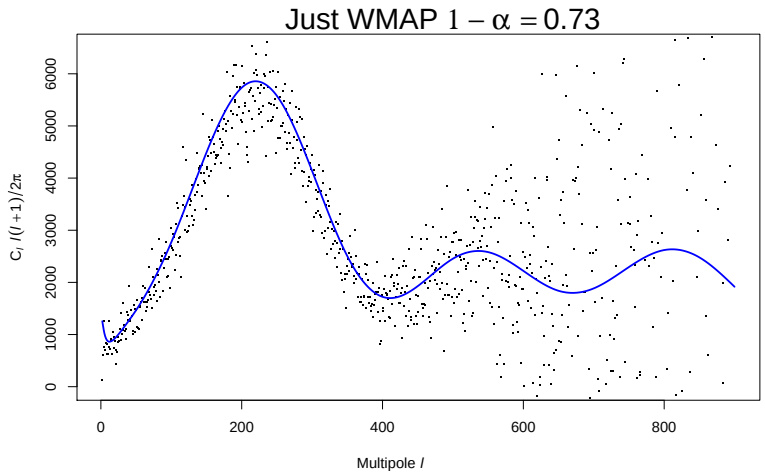


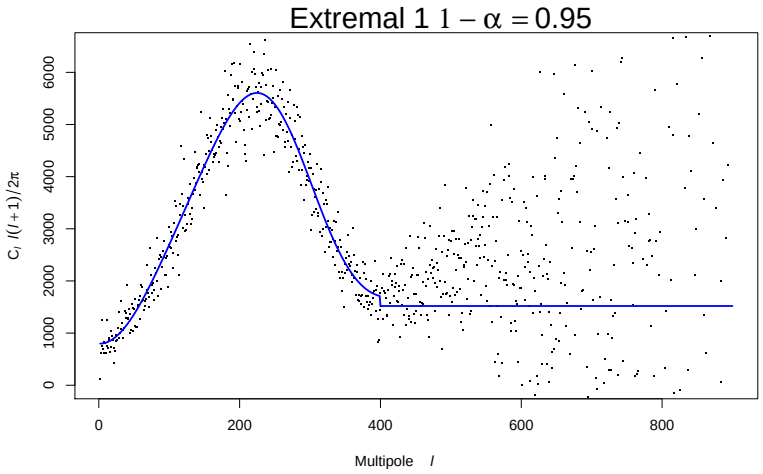
Extended search over millions of spectra (Bryan et al. 2006).

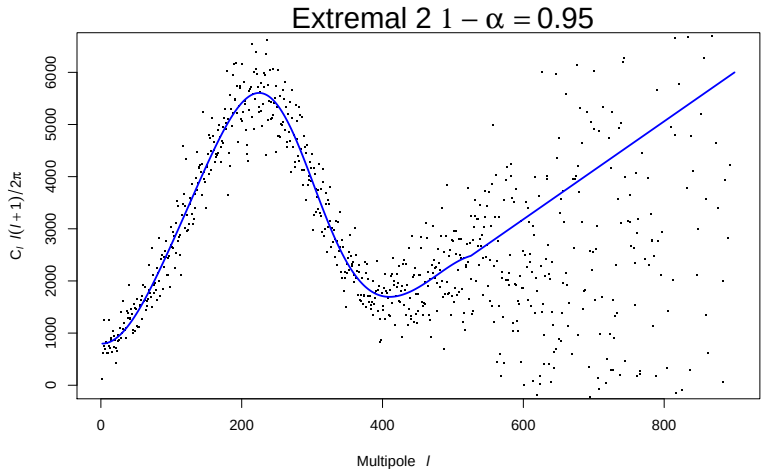
Eyes on the Ball II: Model Checking

Inclusion in the confidence ball provides simultaneous goodness-of-fit tests for parametric (or other) models.









Comment: Coverage and Posterior Probability

- Frequentist Confidence Set $\mathcal{C}$

$$\min_f \mathsf{P}\Big\{\mathcal{C} \ni f\Big\} \geq 1-\alpha. \tag{1}$$

- Bayesian Posterior Region $\mathcal{B}$

$$\mathsf{P}\Big\{f \in \mathcal{B} \mid \mathsf{Data}\Big\} \geq 1-\alpha. \tag{2}$$

- In nonparametric problems, can have (2) hold and yet have

$$\min_f \mathsf{P}\Big\{\mathcal{B} \ni f\Big\} \approx 0. \tag{3}$$

Road Map

1. Smoothing
2. The Six Biases
3. Confidence Sets

Take-Home Points

- The crucial decision in nonparametric estimation is choosing the correct amount of smoothing.
- We must avoid the six biases, but fortunately methods exist to deal with most of them.
- Estimates alone are not enough, also need an assesment of uncertainty. Various types of confidence sets can be constructed, but their dependence on the assumptions can be delicate.