Uniform Confidence Sets for Nonparametric Regression with Application to Cosmology

Christopher R. Genovese
Department of Statistics
Carnegie Mellon University
http://www.stat.cmu.edu/~genovese/

Larry Wasserman
Department of Statistics
Carnegie Mellon University

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

WMAP: The Cosmic Microwave Background

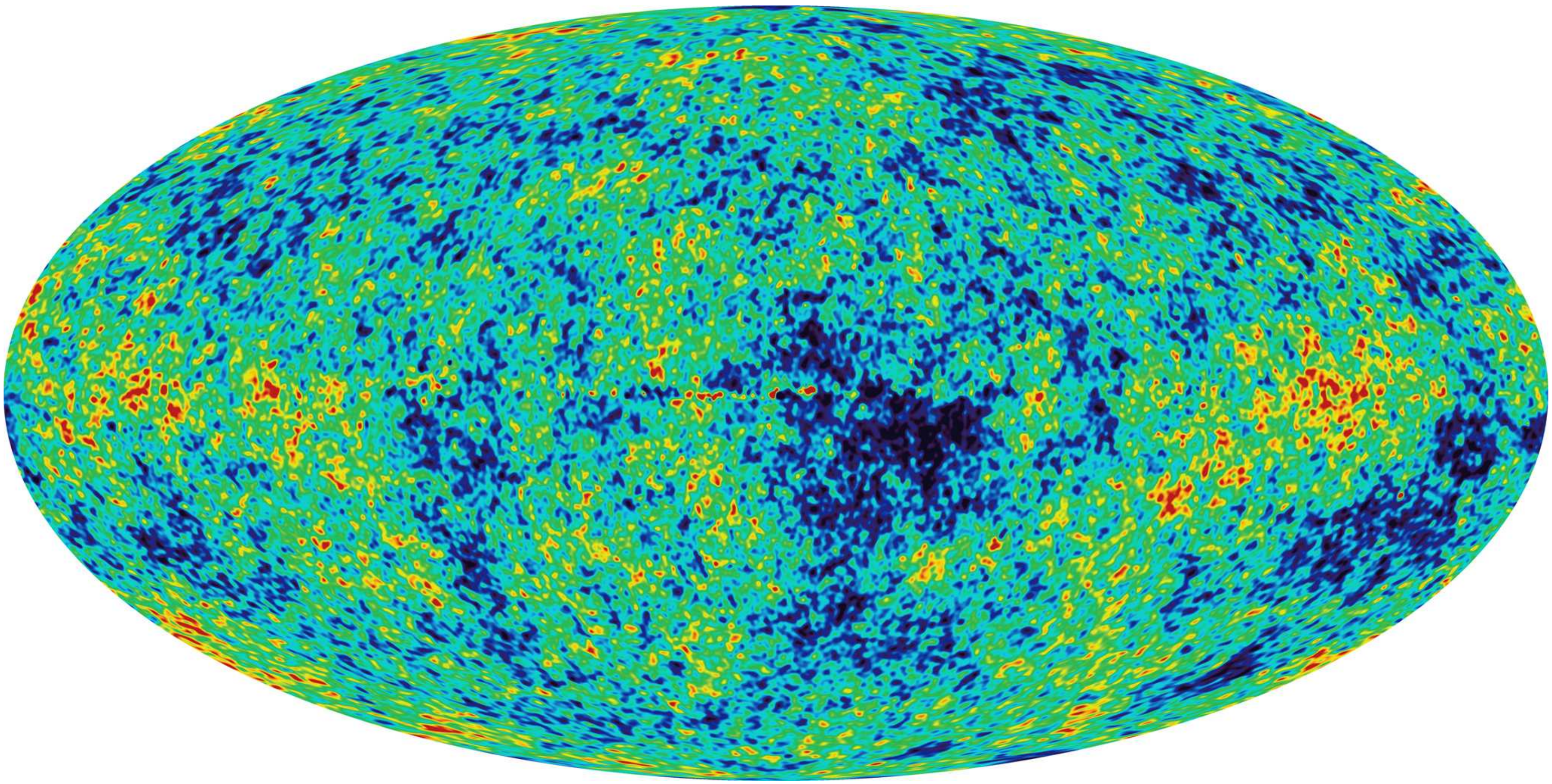


Image: NASA/WMAP Science Team

WMAP in the News

``Breakthrough of the Year, 2003'' — *Science*

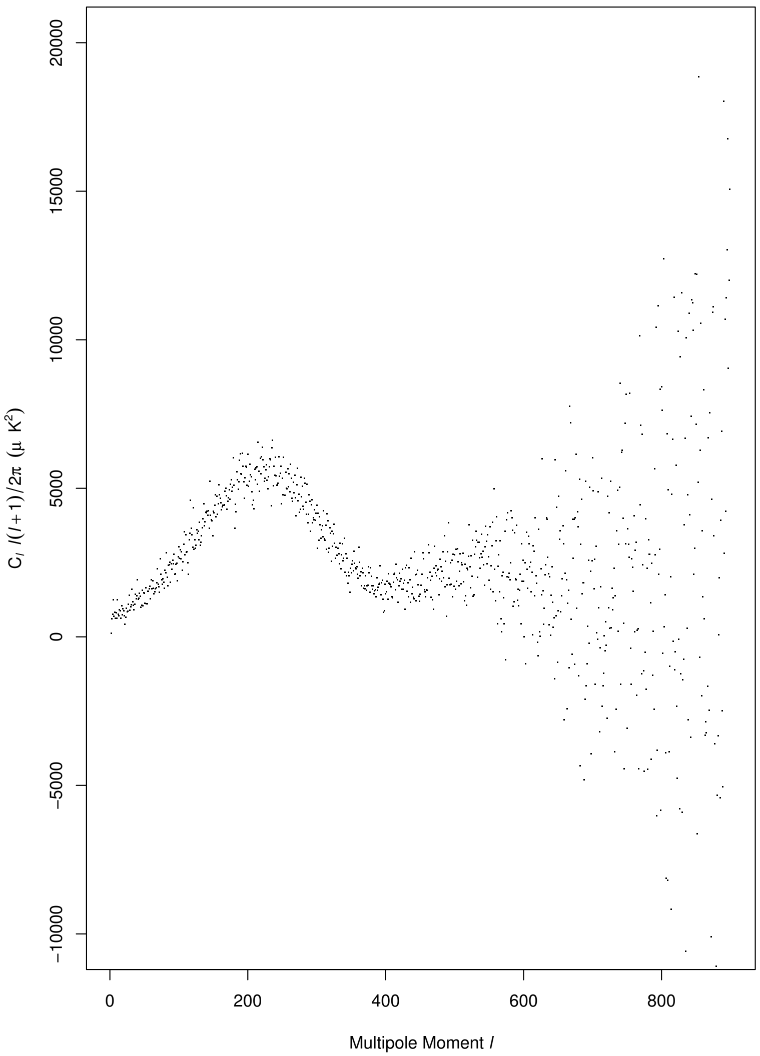
``Most precise, detailed map yet produced of universe just after its birth ... confirms Big Bang theory'' — *New York Times*

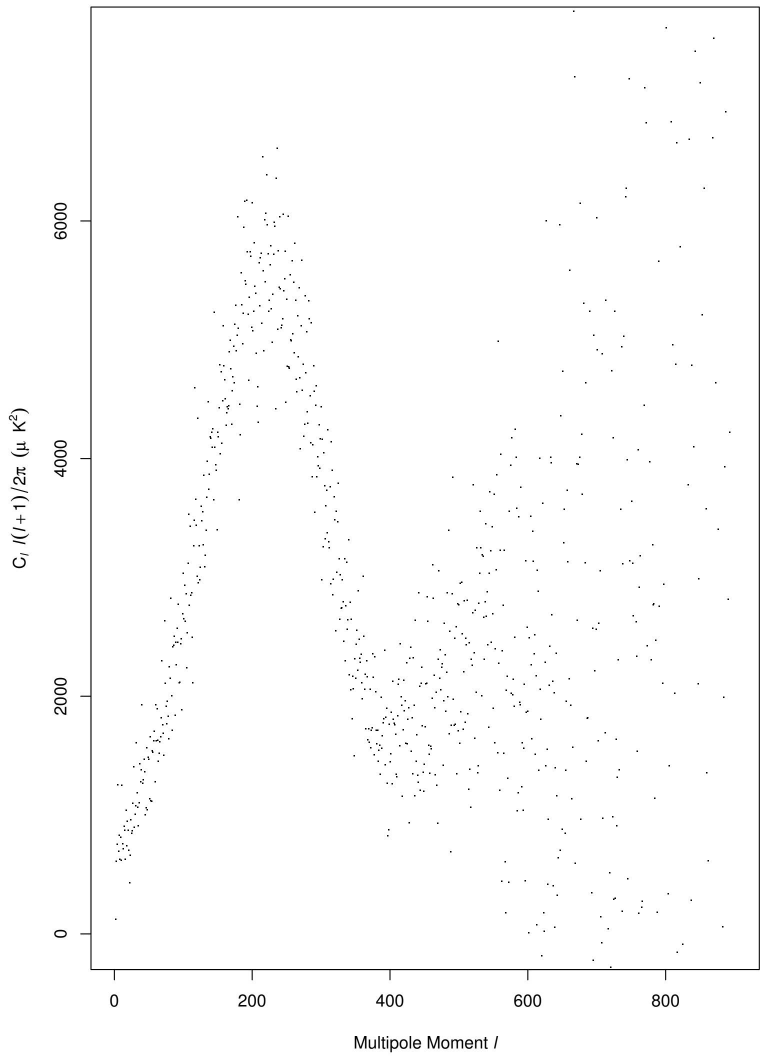
``As of today we know better than ever when the universe began, how it behaved in its earliest instants, how it has evolved since then, and everything it contains.'' — *Sky & Telescope*

``The WMAP data pinpointed -- with unprecedented accuracy -- the universe's age at 13.7 billion years; its flat shape; and its makeup of just 4 per cent "ordinary" matter, 23 per cent dark matter, and 73 per cent dark energy.'' — *New Scientist*

``I think every astronomer will remember where they were when they heard these results. ... I certainly will. This announcement represents a rite of passage for cosmology from speculation to precision science.'' — John Bahcall, Princeton astrophycist in *Washington Post*

It's Just Regression After All





Road Map

1. Constructing Confidence Sets for Unknown Functions
   - Simultaneity, Bias, and Relevance
   - Example: The Pivot-Ball Construction
   - Pivot-Ball Confidence Sets for Wavelet Bases
2. Making Inferences with Uniform Confidence Balls
   - Simultaneous Inferences about sets of functionals
   - Case Study: The Cosmic Microwave Background
   - Keeping Our Eyes on the Ball: Parametric Probes and Confidence Catalogs
3. Non-Adaptive Inference
   - If Confidence Bands Cannot Adapt ...
   - Confidence Catalogs
   - ϵ-Coverage
   - The Resolution-Uncertainty Tradeoff

Road Map

1. Constructing Confidence Sets for Unknown Functions
   - Simultaneity, Bias, and Relevance
   - Example: The Pivot-Ball Construction
   - Pivot-Ball Confidence Sets for Wavelet Bases
2. Making Inferences with Uniform Confidence Balls
   - Simultaneous Inferences about sets of functionals
   - Case Study: The Cosmic Microwave Background
   - Keeping Our Eyes on the Ball: Parametric Probes and Confidence Catalogs
3. Non-Adaptive Inference
   - If Confidence Bands Cannot Adapt . . .
   - Confidence Catalogs
   - ϵ-Coverage
   - The Resolution-Uncertainty Tradeoff

The Nonparametric Regression Problem

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

$$Y_i = f(x_i) + \epsilon_i$$

Assume E$\boldsymbol{\epsilon} = 0$ and Var $\boldsymbol{\epsilon} = \Sigma$.

Leading case: $x_i = i/n$ and $\Sigma = \sigma^2 I$ with σ^2 known.

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

Example: Sobolev ball

$$\mathcal{F} \equiv \mathcal{F}_p(C) = \left\{ f \in \mathcal{L}^2 \colon \int |f^{(p)}|^2 \leq C^2 \right\}.$$

Other examples: Besov space, Lipschitz class

Rate-Optimal Estimators

For a specified loss $L(\widehat{f}, f)$ (e.g., $\int(\widehat{f} - f)^2$ or $|\widehat{f}(x_0) - f(x_0)|^2$), want a procedure that achieves the minimax risk.

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 $\mathcal{F}_p$.

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

Adaptive Estimators

It's unsatisfying to depend too strongly on intangible assumptions such as whether $f \in \mathcal{F}_p$ or $f \in \mathcal{F}_q$.

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

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

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

without knowing p.

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

For example, certain wavelet shrinkage estimators are adaptive over restricted families of Besov spaces. (See Donoho and Johnstone 1998, Cai 1999, for example.)

Inference about the Unknown Function

But 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 $\mathsf{P}\Big\{\mathcal{C} \ni f\Big\} = 1-\alpha$.

Typically, $\mathcal{C}$ is the set of functions within a confidence band over all (or a finite set of) points in the domain.

Three challenges:

1. Bias
2. Simultaneity
3. Relevance

Bias

In nonparametric problems, using a rate-optimal tuning 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.

Simultaneity

Want inferences to hold simultaneously across $x_1, \dots, x_n$.

Often want to extend inferences to the whole object, which requires additional assumptions to constrain f between the x_is.

Bonferroni confidence bands $(y_i \pm \sigma \Phi^{-1}(\alpha/n))$ are unsatisfactory (grow with n, not smooth).

For confidence bands, one solution is the "volume of tubes" formula (Sun and Loader 1994).

If $\widehat{f}(x) = \sum_{i=1}^{n} \ell_i(x) Y_i$, then for a suitable class $\mathcal{F}$,

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

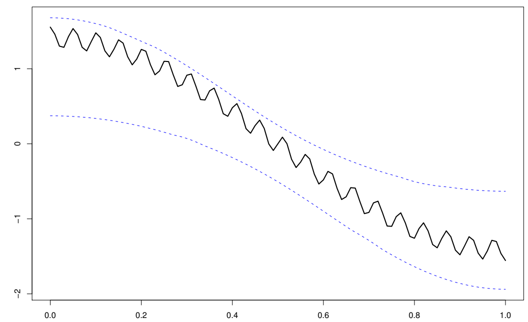
where c solves $\alpha = K_\ell \phi(c) + 2(1 - \Phi(c))$.

Must account for bias in general.

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: Uniform Coverage

For asymptotic confidence procedures, prefer uniform coverage:

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

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

Li (1989) showed that with no prior smoothness bound in $\mathcal{F}$, any $1-\alpha$ confidence sets for $f_n = (f(x_1), \ldots, f(x_n))$ of the form

$$\mathcal{C}_n = \left\{ f_n \in \mathbb{R}^n \colon\ n^{-1/2} \| f_n - \widehat{f}_n \| \leq s_n(y) \right\}$$

that are "asymptotically honest" in the sense of

$$\lim_{n \to \infty} \inf_{f \in \mathcal{F}} \mathsf{P}_f \left\{ \mathcal{C}_n \ni f \right\} \geq 1 - \alpha$$

must have $s_n \geq c n^{-1/4}$.

Remark: Adaptive Coverage

We would also like *adaptive* confidence procedures.

These would maintain coverage on $\mathcal{F}$ but could use the data to tailor the set's diameter to the unknown f, e.g., produce smaller balls for sufficiently smooth f.

Whether this is possible depends on the ability of random-sized confidence sets to outperform fixed-size sets.

Adaptive confidence sets exist in some spaces but not in others.

(See Low 1997, Juditsky 2002, Cai and Low 2004, Robbins and Van der Vaart 2004, Baraud 2004, and Genovese and Wasserman 2005).

More on this later.

Possible Approaches

X Estimate bias pointwise

Often increases variance more than it reduces bias.

X Undersmoothing

Requires additional calibration; typically non-uniform coverage.

$\surd$ Pivot-Ball Method (Beran and Dümbgen 1998)

Uniform asymptotic coverage for $\mathcal{L}^2$ confidence balls.
Supports functional search.

$\surd$ Subspace Pretesting (Baraud 2004)

Finite-sample coverage for ℓ^2 confidence balls.
Supports functional search.

$\surd$ Centering on Adaptive Estimator (Cai and Low 2004)

Adapts over Besov balls in certain ranges. Optimality results.

Other approaches include bounding global bias (Sun and Loader) and scale-space methods (Chaudhuri and Marron).

Pivot-Ball (Beran and Dümbgen 1998)

Let $\phi_1, \phi_2, \ldots$ be an orthonormal basis and write $f = \sum_j \theta_j \phi_j$.

Estimate θ_j by $\widehat{\theta}_j(\lambda)$, with tuning parameter vector λ.

0. Define loss $L_n(\lambda) = \sum_{j=1}^{n} (\widehat{\theta}_j(\lambda) - \theta_j)^2$.

   Let $S_n(\lambda)$ be an unbiased estimate of $\mathsf{E}L_n(\lambda)$.

   Choose $\widehat{\lambda}_n$ to minimize $S_n(\lambda)$.

1. Show that *pivot process* $B_n(\lambda) = \sqrt{n}(L_n(\lambda) - S_n(\lambda))$ converges weakly to Gaussian process with mean 0, cov. $K(s,t)$.

2. Find an estimator $\widehat{\tau}_n^2$ of $K(\widehat{\lambda}_n, \widehat{\lambda}_n)$ so that

$$\frac{B_n(\widehat{\lambda}_n)}{\widehat{\tau}_n} \rightsquigarrow \mathsf{N}(0,1).$$

Pivot-Ball (cont'd)

3. Conclude that $\mathcal{D}_n$ is an asymptotic $1-\alpha$ confidence set for $\boldsymbol{\theta}$:

$$\mathcal{D}_n = \left\{ \theta\colon\ \sum_{\ell=1}^{n} (\widehat{\theta}_n(\widehat{\lambda}_n) - \theta_\ell)^2 \leq \frac{\widehat{\tau}_n\, z_\alpha}{\sqrt{n}} + S_n(\widehat{\lambda}_n) \right\}.$$

is

4. Hence $\mathcal{C}_n = \left\{ f_n\colon\ \int (f_n - \widehat{f}_n)^2 \leq \dfrac{\widehat{\tau}_n z_\alpha}{\sqrt{n}} + S_n(\widehat{\lambda}_n) \right\}$ yields

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

for projection f_n onto first n coefficients.

5. With extra assumptions, can dilate $\mathcal{C}_n$ to cover f similarly.

Pivot-Ball (cont'd)

Beran and Dümbgen considered modulators

$$\widehat{\theta}(\lambda) = (\lambda_1\widetilde{\theta}_1, \ldots, \lambda_n\widetilde{\theta}_n),$$

where

$$1 \geq \lambda_1 \geq \cdots \geq \lambda_n \geq 0,$$

and

$$\widetilde{\theta}_j \approx \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_j(x_i).$$

Remark: Beran and Dümbgen (1998) stated the main result in ℓ^2, but in practice need Sobolev assumptions to (i) estimate σ and (ii) move from sequence to function space.

Extensions to the Pivot-Ball Method

- Wavelet bases (Genovese and Wasserman 2003)
  [next slide]

- Weighted-Loss/Nonconstant Variance (Genovese et al. 2004)
  [see CMB example]

- Density Estimation (Jang, Genovese, and Wasserman 2004)
  Uses p_n term basis expansions with, e.g., $p_n = o(n^{1/3})$.

Pivot-Ball Method: Wavelets

Write $f = \sum_k \alpha_{J_0,k}\varphi_j + \sum_{j=J_0}^{\infty}\sum_k \beta_{jk}\psi_{jk}$.

Wavelet coefficients characterize f within family of Besov spaces.

Donoho and Johnstone devised several shrinkage schemes that yield rate-optimal adaptive estimators over Besov spaces: Universal thresholding, global SureShrink, levelwise SureShrink.

Use pivot-ball construction to get confidence ball centered on wavelet-shrinkage estimator.

But wavelet basis functions unbounded, so pivot process is not asymptotically equicontinuous near zero.

Hence, must restrict thresholds to $[\varrho\widehat{\sigma}\rho_n, \widehat{\sigma}\rho_n]$, where $\rho_n = \sqrt{2\log n/n}$ and $1/\sqrt{2} < \varrho < 1$.

Conjecture that SureShrink result holds for $\varrho > 0$ but it doesn't hold for $\varrho = 0$.

Pivot-Ball Method: Wavelets (cont'd)

The confidence ball $\mathcal{C}_n$ centered on the wavelet-shrinkage estimator has radius:

$$s_n^2 = \frac{\widehat{\sigma}^2 z_\alpha}{\sqrt{n/2}} + S_n(\widehat{\lambda}_n).$$

Confidence set radius $O(n^{-1/4})$ over Besov balls consistent with results of Li (1989) and Baraud (2004) for balls over $\mathbb{R}^n$.

This rate swamps subtle performance differences among estimators. What are the implications for inference here?

Remark: Cai and Low (2004) develop confidence balls that adapt over restricted ranges of Besov balls.

Road Map

1. Constructing Confidence Sets for Unknown Functions
   - Simultaneity, Bias, and Relevance
   - Example: The Pivot-Ball Construction
   - Pivot-Ball Confidence Sets for Wavelet Bases
2. Making Inferences with Uniform Confidence Balls
   - Simultaneous Inferences about sets of functionals
   - Case Study: The Cosmic Microwave Background
   - Keeping Our Eyes on the Ball: Parametric Probes and Confidence Catalogs
3. Non-Adaptive Inference
   - If Confidence Bands Cannot Adapt . . .
   - Confidence Catalogs
   - ϵ-Coverage
   - The Resolution-Uncertainty Tradeoff

Pivot-Ball Method: Inferences for Functionals

To make inferences for functionals of f, we can search $\mathcal{C}_n$:

$$\left(\inf_{f\in\mathcal{C}_n} T(f), \sup_{f\in\mathcal{C}_n} T(f)\right)$$

is a confidence set for $T(f)$.

If $\mathcal{T}$ is a set of functionals, then

$$\left\{\left(\inf_{f\in\mathcal{C}_n} T(f), \sup_{f\in\mathcal{C}_n} T(f)\right) : T \in \mathcal{T}\right\}$$

gives simultaneous intervals for all the functionals in $\mathcal{T}$.

This is useful for post-hoc exploration.

Pivot-Ball Method: Functionals (cont'd)

Fix a decreasing sequence $\Delta_n > 0$ and consider block-averages

$$\mathcal{T}_n = \left\{ T\colon\ T(f) = \frac{1}{b-a}\int_a^b f\,dx,\ 0 \le a < b \le 1,\ |b-a| \ge \Delta_n \right\}.$$

If $\mathcal{C}_n$ is the confidence ball, we can get simultaneous coverage on $\mathcal{T}_n$

$$\liminf_{n\to\infty}\ \inf_{f\in\mathcal{F}_{\eta,c}} \mathsf{P}\left\{ T(f) \in J_n(T) \text{ for all } T \in \mathcal{T}_n \right\} \ge 1-\alpha,$$

where $\mathcal{F}_{\eta,c}$ is large — a union of Besov spaces (see details).

The intervals are of the form

$$J_n(T) = \left(\inf_{f_n\in\mathcal{C}_n} T(f_n) - w_n,\ \sup_{f_n\in\mathcal{C}_n} T(f_n) + w_n \right),$$

where the $w_n \to 0$ at a rate inversely related to the rate at which $\Delta_n \to 0$.

Physics of the Early Universe

The Big Bang model posits an expanding universe that began hot and dense.
A concise history starting 13.7 billion years ago:

– Temperature $\approx$ 1 trillion K (about 1 second)

Density high enough to stop neutrinos

– Temperature $>$ 1 billion K (about 3 minutes)

Atoms cannot form. Space filled with a stew of photons, baryons (e.g., protons and neutrons), electrons, neutrinos, and other matter.

– Temperature 12000 K

Photons and baryons became coupled in a mathematically perfect fluid.
Dark matter begins to clump under gravity. Acoustic waves propagate.

– Temperature 3000 K (about 380,000 years). "Recombination"

Atoms form, photons are released.

– Temperature 2.7K (today). The Cosmic Microwave Background (CMB).

Photons released at recombination observed in microwave band.
Nearly uniform across the sky.

The Cosmic Microwave Background Today

The acoustic oscillations before recombination carried information about the geometry and composition of the early universe. (Can you hear the shape of the universe?)

We see this today as a pattern of hot and cold spots on the sky.

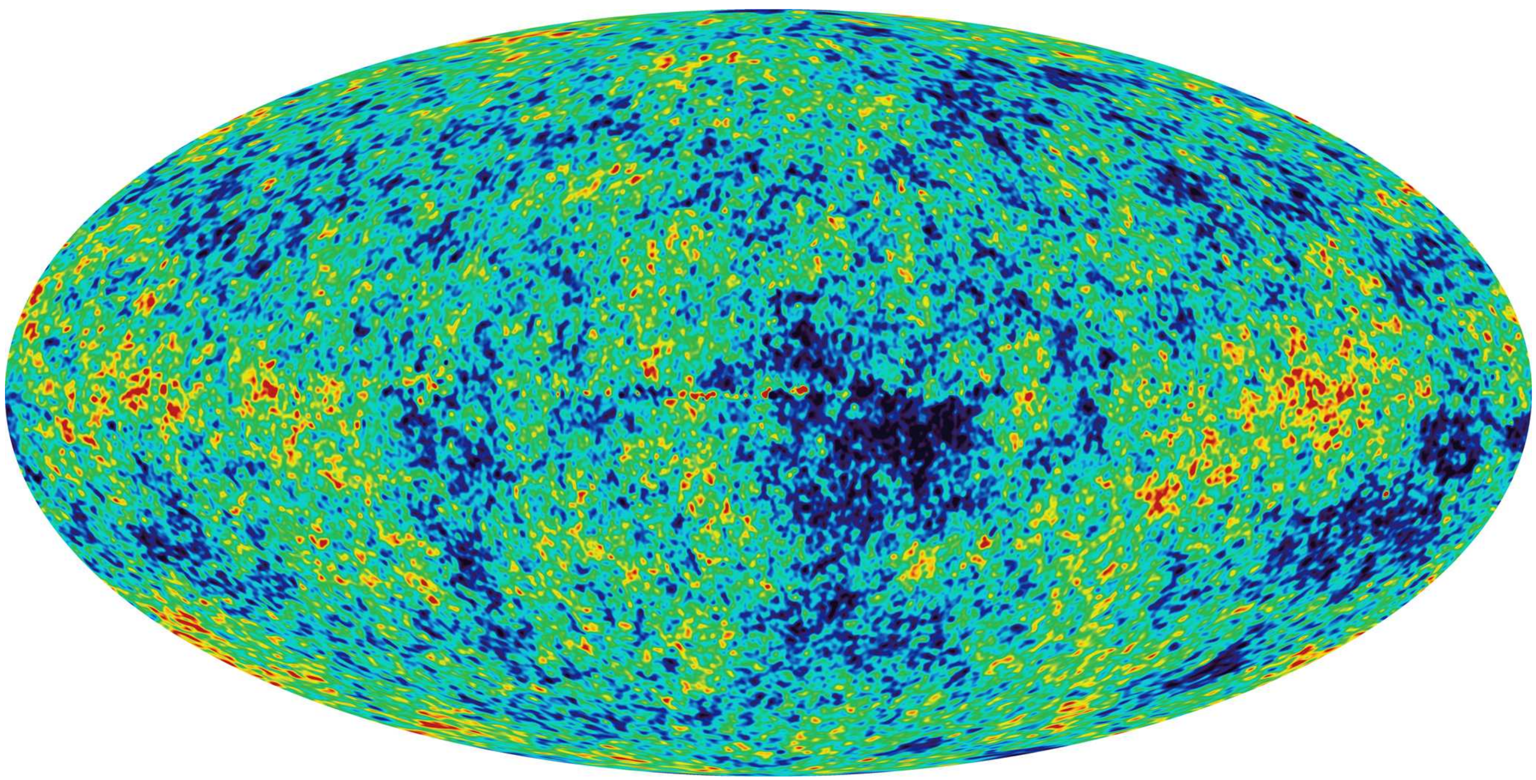


Image: NASA/WMAP Science Team

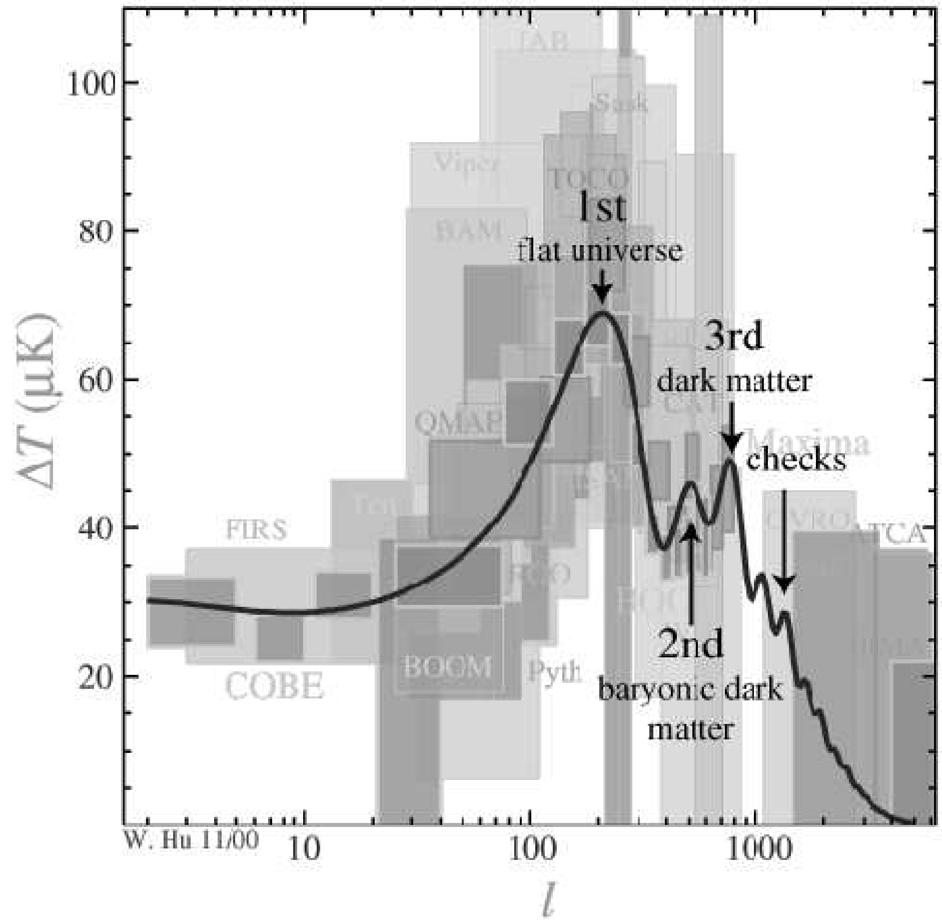
Physics of the Early Universe (cont'd)

Cosmologists decompose the sky map into spherical harmonics,

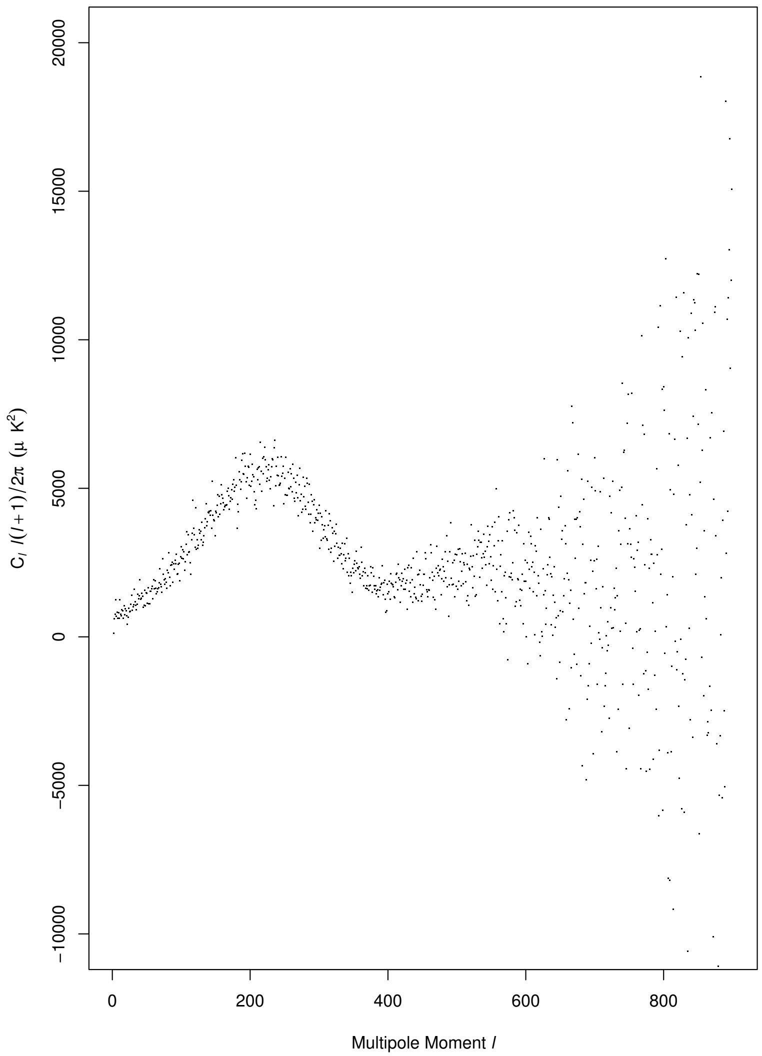
$$\frac{T - \overline{T}}{\overline{T}} = \sum_{\ell,m} a_{\ell,m} Y_{\ell,m},$$

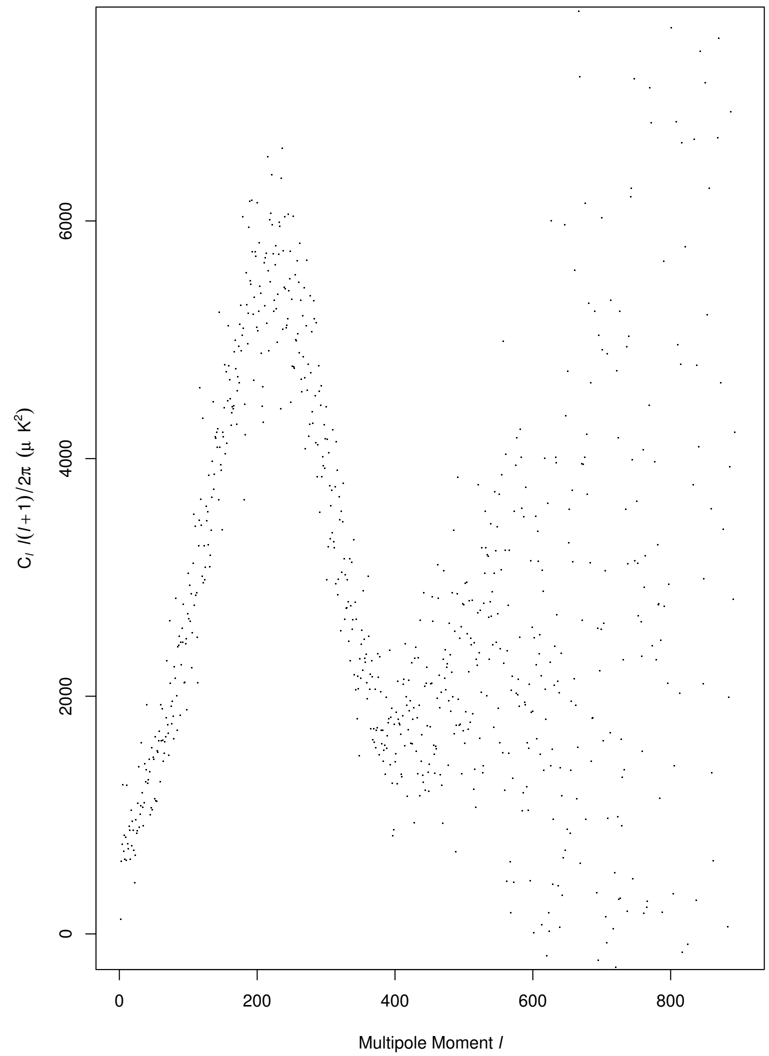
and compute the coefficient variance at each angular scale ℓ.

This is the raw estimated CMB "power spectrum" $f(\ell) = \widehat{C}_\ell$.

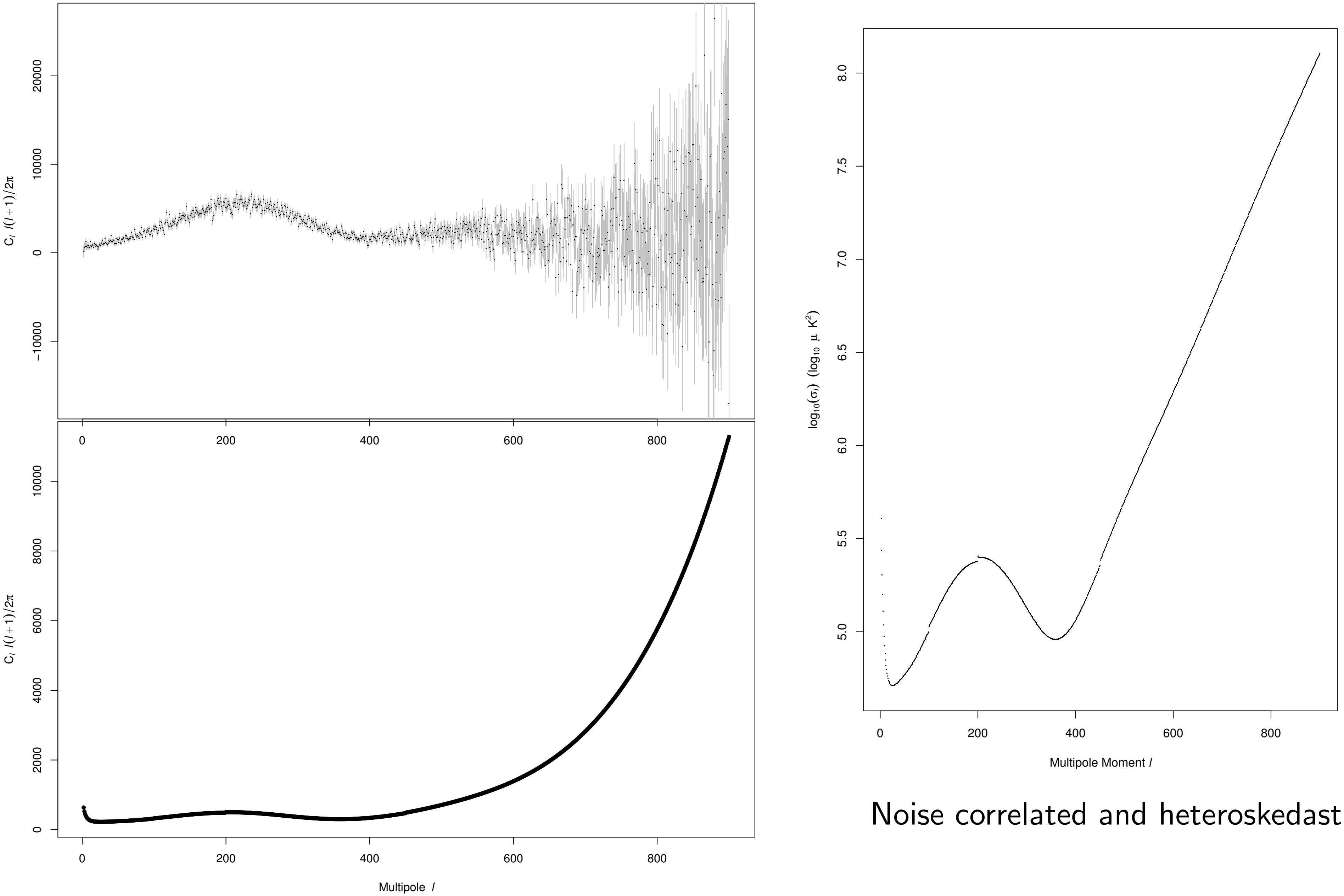


CMB Power Spectrum: WMAP Data



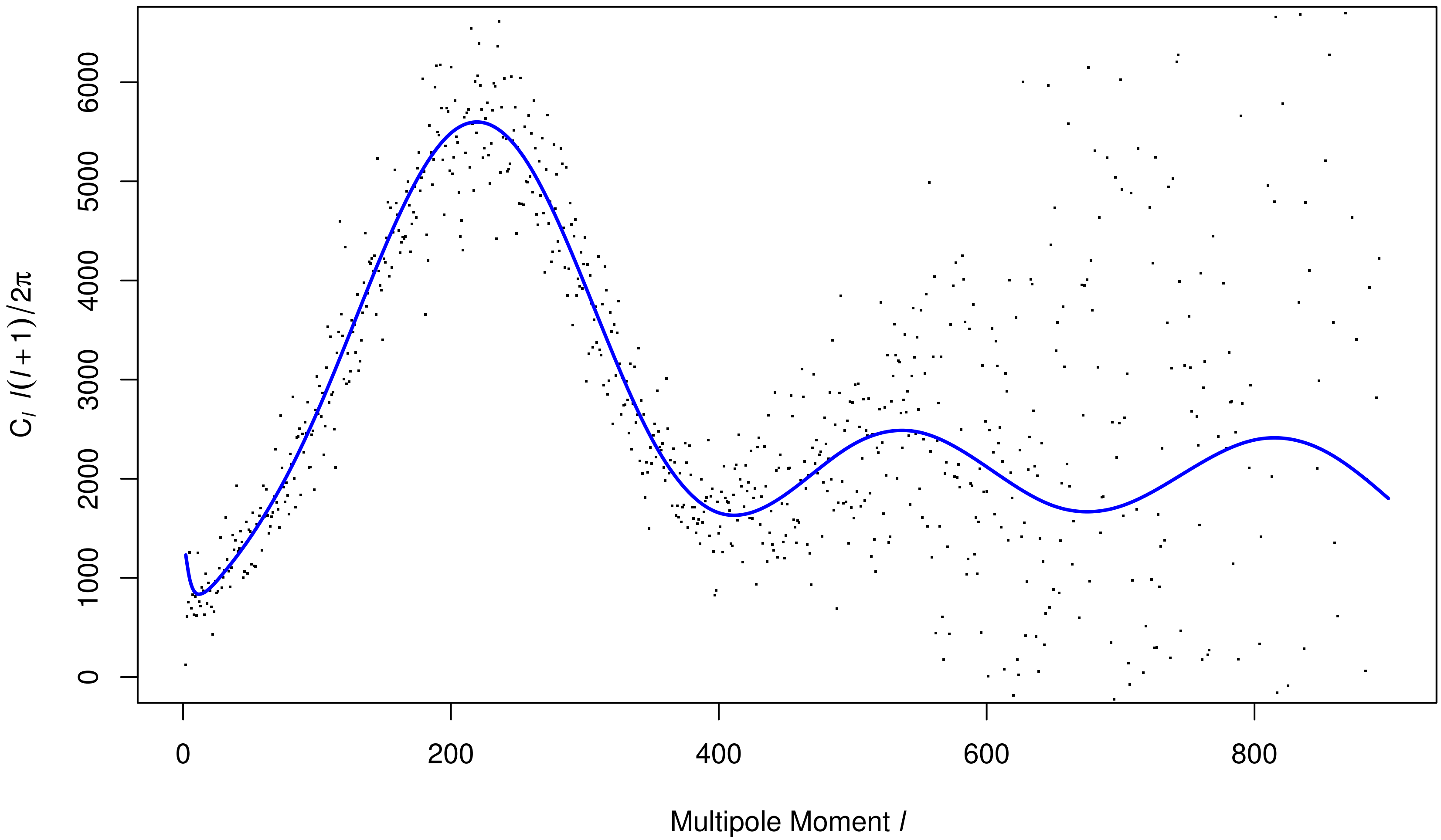


CMB Power Spectrum: WMAP Variances



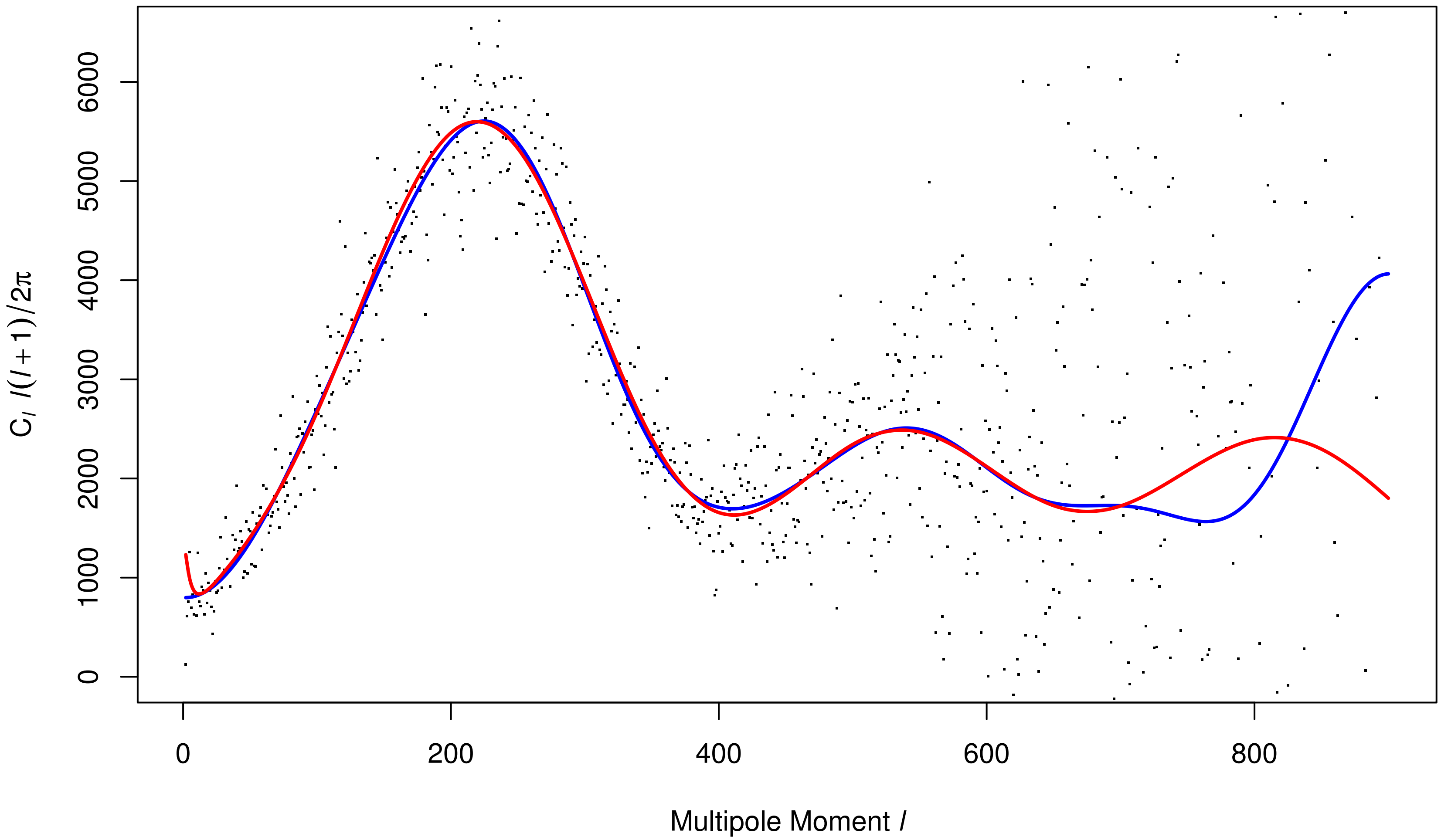
Noise correlated and heteroskedastic

Cosmological Models



- Low-dimensional model maps cosmological parameters to spectra.
- Ultimate goal: inferences about these cosmological parameters.
- Subsidiary goal: identify location, height, widths of peaks

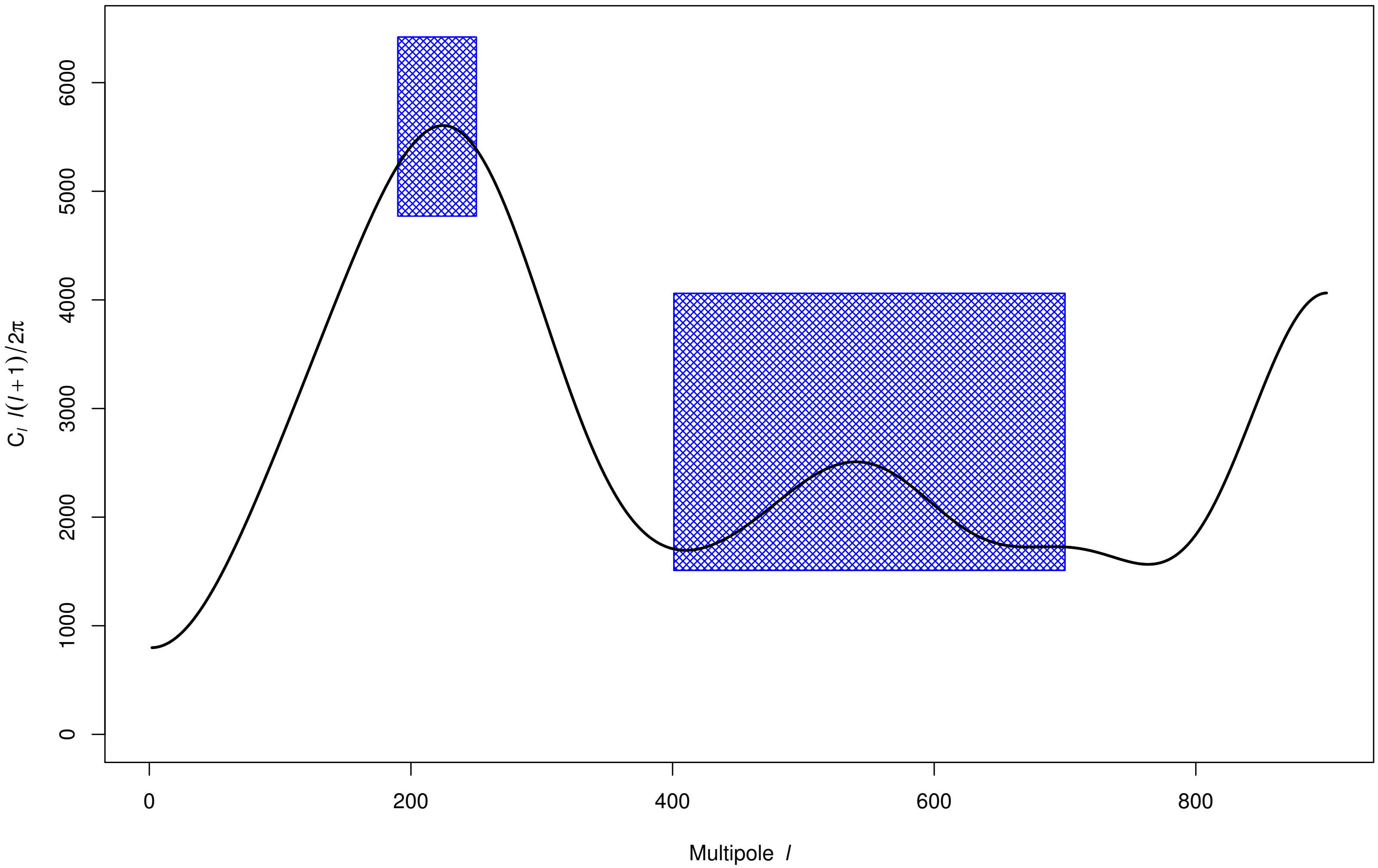
Confidence Ball Center vs Concordance Model



- Concordance model (red) is an MLE based on WMAP and four other data sets.
- Confidence ball based on weighted $\mathcal{L}^2$ loss and WMAP data only.
- Notable discrepancies between WMAP and theory (e.g., low ℓ turn-off).

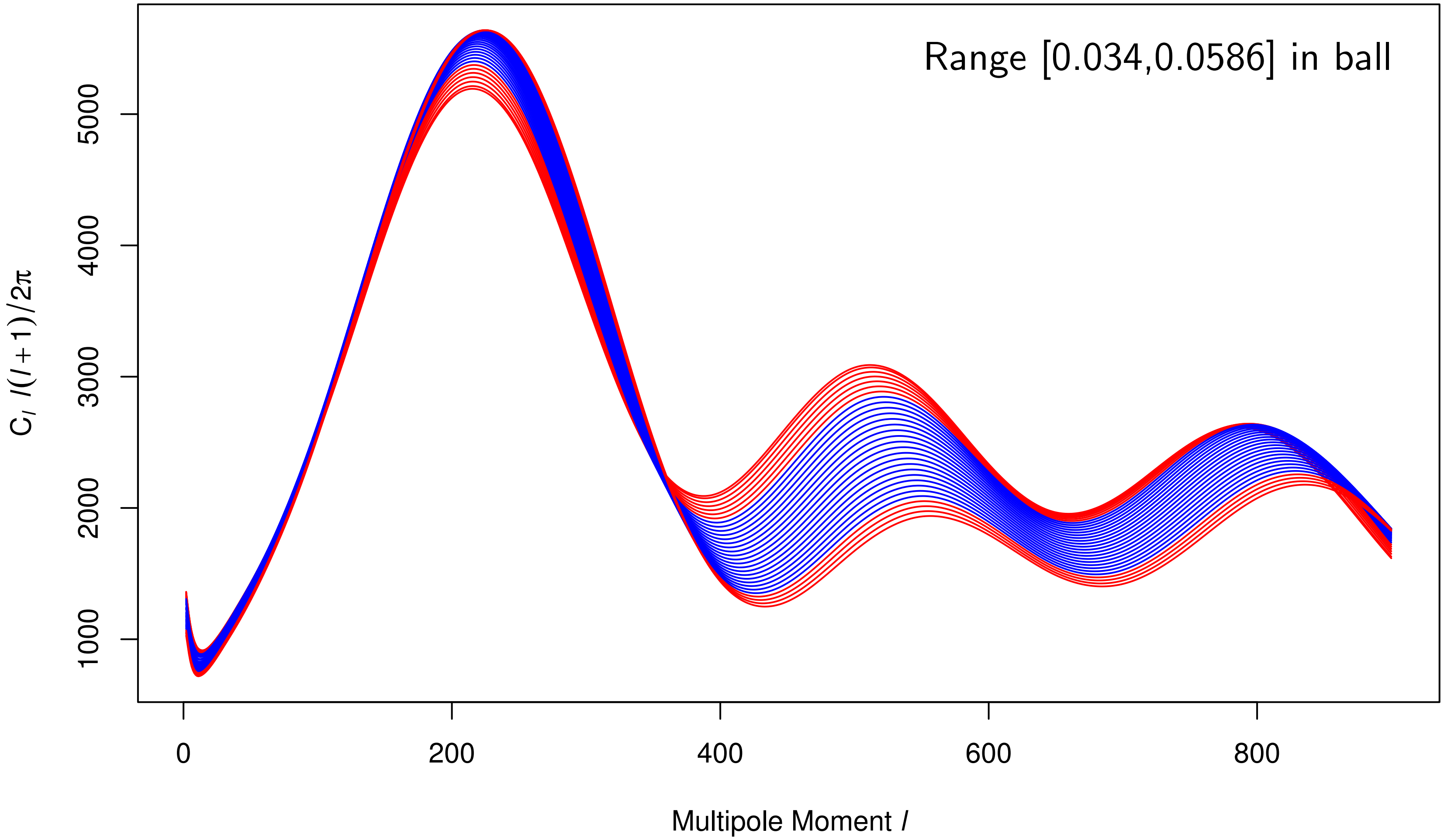
Eyes on the Ball I: Parametric Probes

Peak Heights, Peak Locations, Ratios of Peak Heights



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

Varied baryon fraction in CMBFAST keeping $\Omega_{\text{total}} = 1$

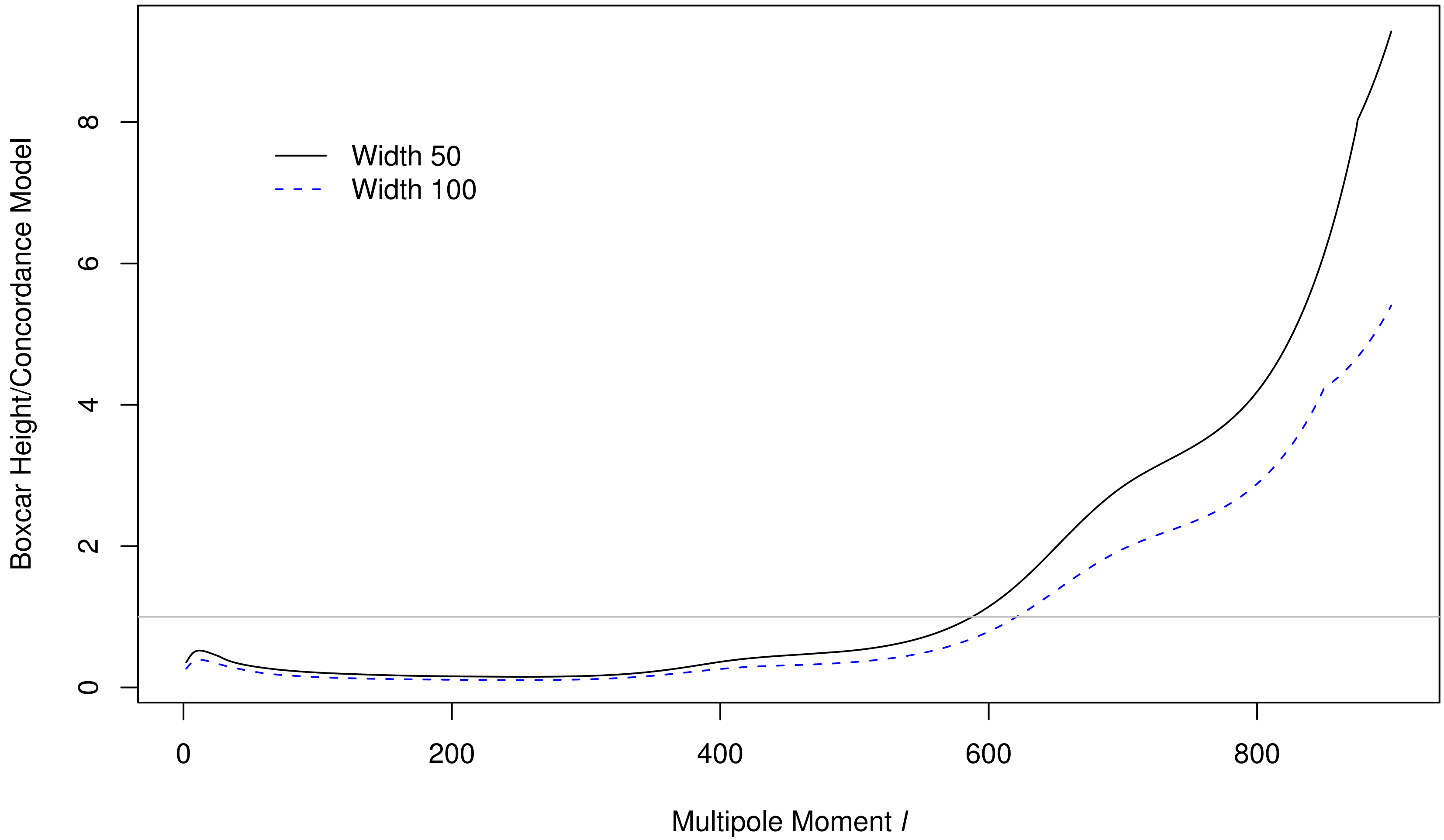


Extended search, over millions of spectra, in progress.

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

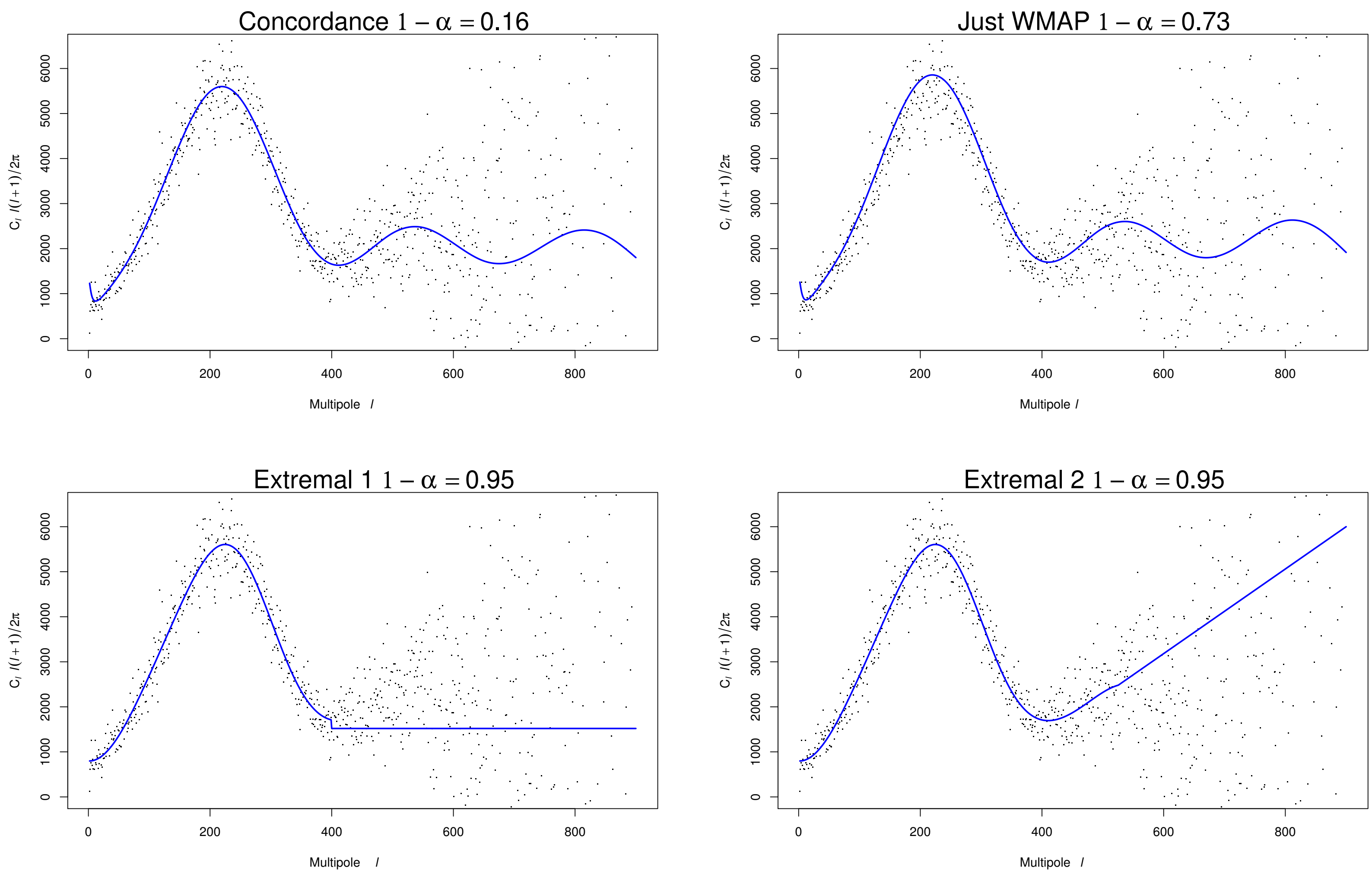
How much can spectrum change locally within the confidence ball?

For each ℓ, add to Concordance model a multiple of a boxcar of given width centered at ℓ. Increase height until it leaves the 95% ball.



Eyes on the Ball II: Model Checking

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



Eyes on the Ball III: Confidence Catalogs

- Our confidence set construction does not impose constraints based on prior knowledge.

  Instead: form ball first and impose constraints at will.

- Raises the possibility of viewing inferences *as a function* of prior assumptions.

  The confidence ball creates a mapping from prior assumptions to inferences; we call this a confidence catalog.

- Example: Constraints on peak curvature over range defined by reasonable parametric models.

Road Map

1. Constructing Confidence Sets for Unknown Functions
   – Simultaneity, Bias, and Relevance
   – Example: The Pivot-Ball Construction
   – Pivot-Ball Confidence Sets for Wavelet Bases

2. Making Inferences with Uniform Confidence Balls
   – Simultaneous Inferences about sets of functionals
   – Case Study: The Cosmic Microwave Background
   – Keeping Our Eyes on the Ball: Parametric Probes and Confidence Catalogs

3. Non-Adaptive Inference
   – If Confidence Bands Cannot Adapt . . .
   – Confidence Catalogs
   – ϵ-Coverage
   – The Resolution-Uncertainty Tradeoff

Adaptive and Non-Adaptive Confidence Sets

Let $f = (f_1, \ldots, f_n)$ with $f_i = f(x_i)$ and consider a confidence set of the form

$$\mathcal{C} = \left\{ f \colon \frac{1}{n} \sum_{i=1}^{n} (f_i - \widehat{f}_i)^2 \leq s_n \right\} \quad \text{with} \quad \inf_{f \in \mathbb{R}^n} \mathsf{P}_f\{\mathcal{C} \ni f\} \geq 1 - \alpha.$$

Then, results in Baraud (2004) and Cai and Low (2004) show that

$$\inf_{f \in \mathbb{R}^n} \mathsf{E}_f(s_n) \geq C_1 n^{-\frac{1}{4}} \quad \text{and} \quad \sup_{f \in \mathbb{R}^n} \mathsf{E}_f(s_n) \geq C_2,$$

for $C_1, C_2 > 0$. Moreover, the $n^{-1/4}$ rate can be achieved.

This shows that a random-diameter $\mathcal{L}^2$ confidence balls can improve on a fixed-diameter balls, which necessarily has rate $O(1)$.

Confidence Bands Cannot Adapt

For $\mathcal{L}^p$ confidence balls $2 \leq p < \infty$, the corresponding rate is $n^{-1/2p}$, showing less adaptivity as p increases.

For confidence bands (or $\mathcal{L}^\infty$ balls), there is no rate adaptivity.

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 typical 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.

If Confidence Bands Cannot Adapt . . .

Adaptation for confidence balls is possible, though limited.

Because confidence bands are not adaptive, the width of the bands will be determined by the worst-case assumption contemplated.

This is both arbitrary and unsatisfying:

- The results are extremely sensitive to the assumptions made.
- There is rarely enough information to constrain this worst-case assumption too strictly.
- No benefit is provided by the advanced adaptive estimators because the width of the bands swamps the differences among good and not-so-good estimators.

Moreover, the "unknown σ" problem can have a substantial impact on the best rate.

What strategies can we use for non-adaptive inference?

Confidence Catalogs

View the product of inference not as a single confidence set but as a mapping from assumptions to confidence sets.

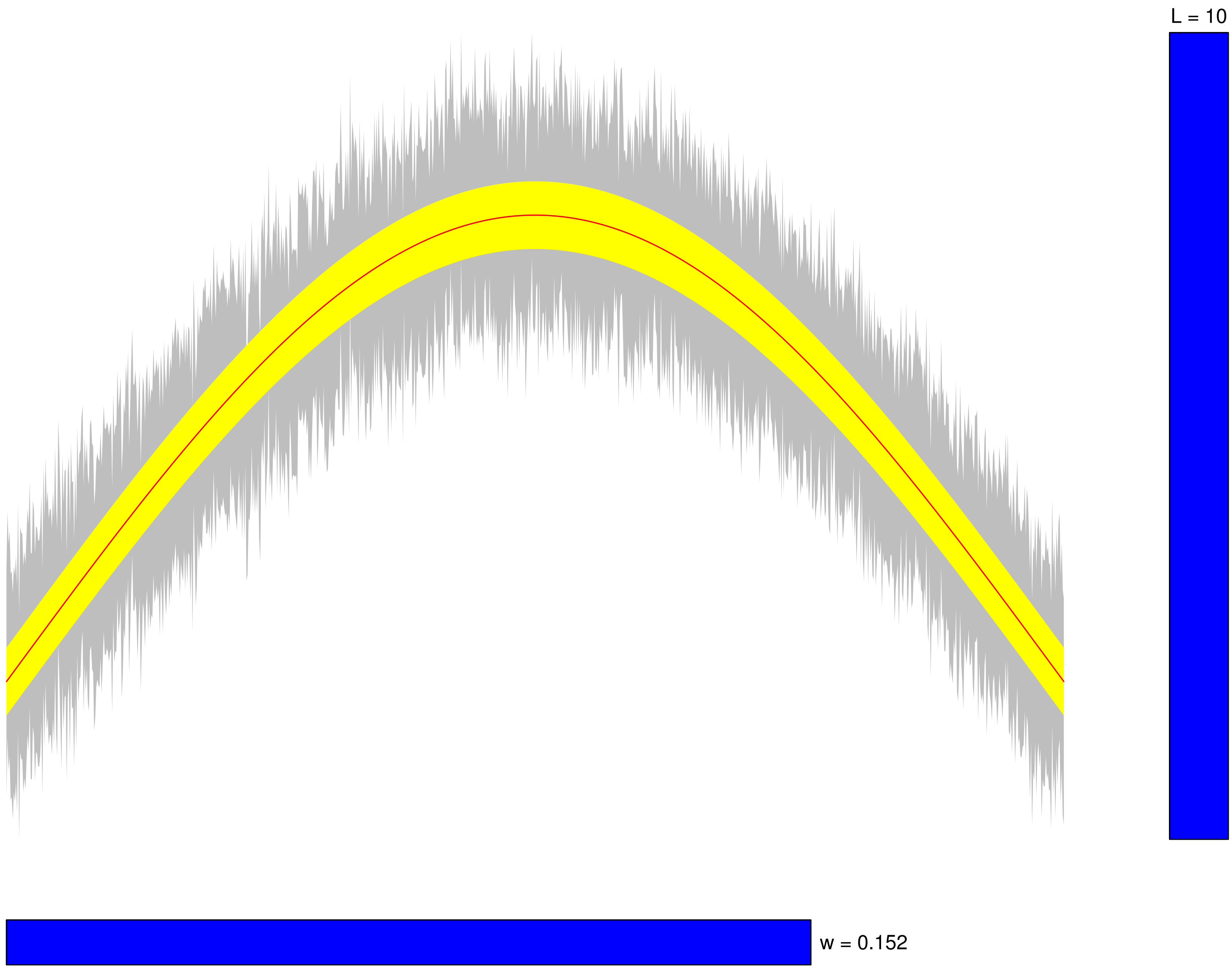
Let $\mathcal{A}$ be a finite-dimensional space parametrizing the assumptions.

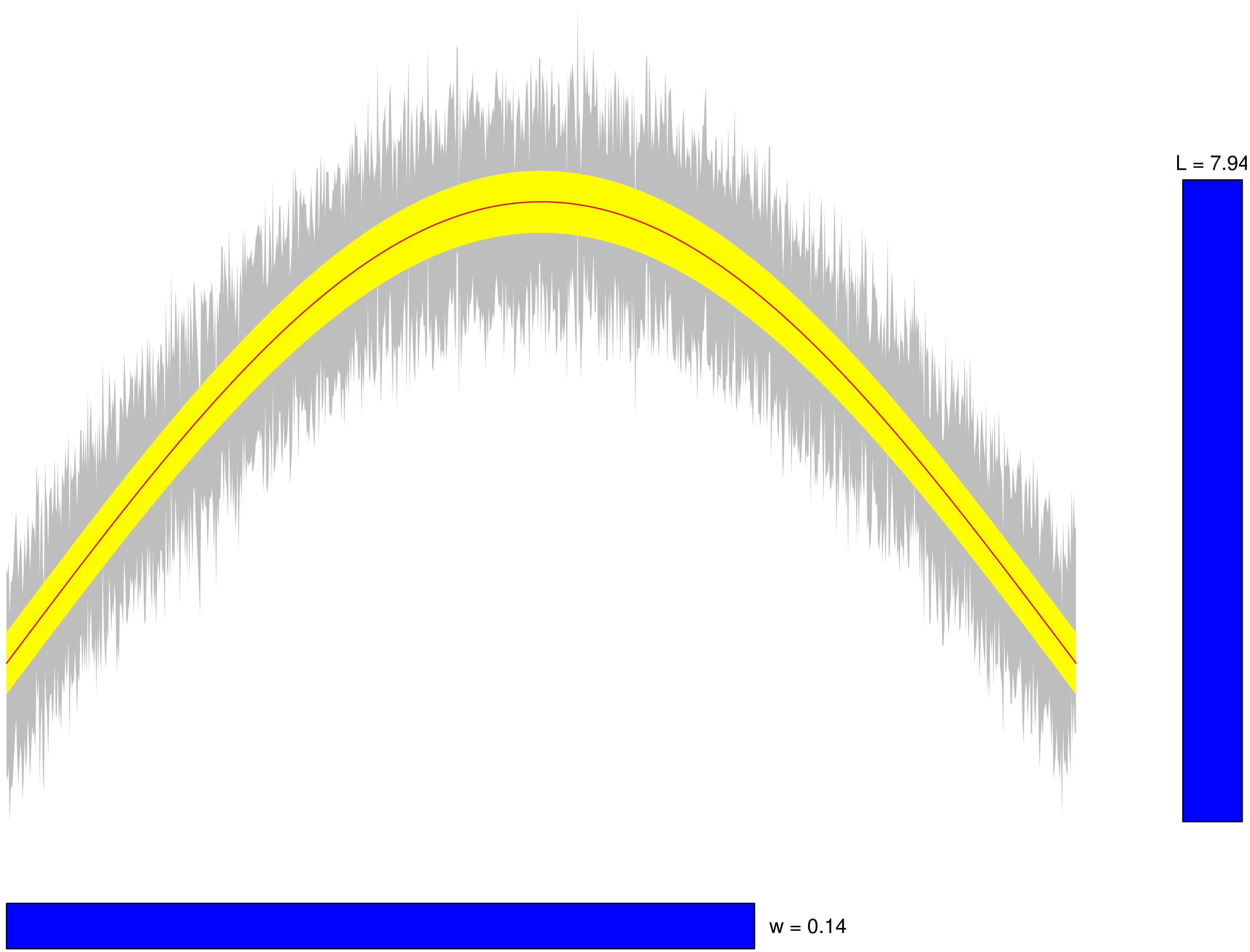
Define a *confidence catalog* C_α to be a mapping from $\mathcal{A}$ to a collection of uniform $1-\alpha$ confidence sets for f.

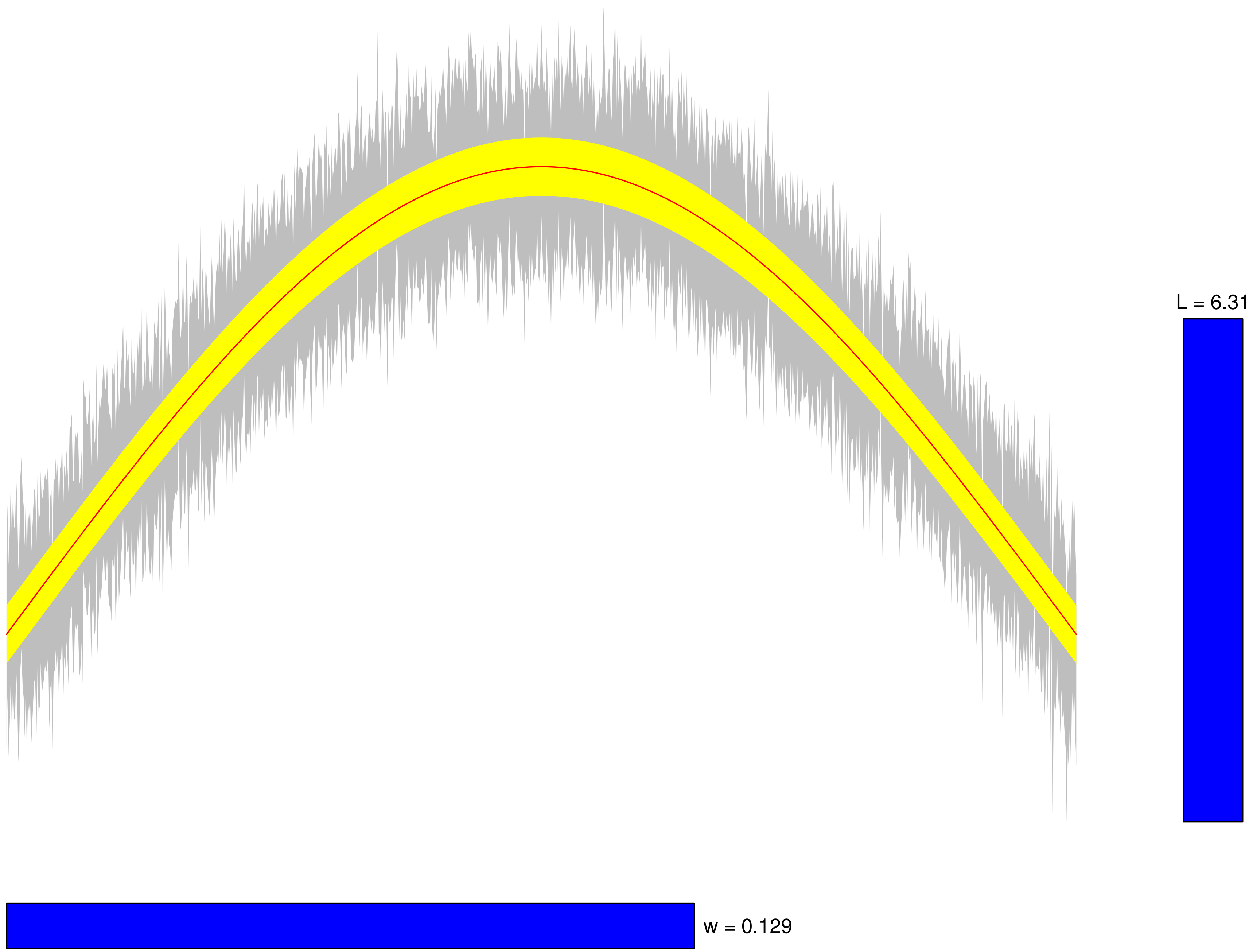
Examples:

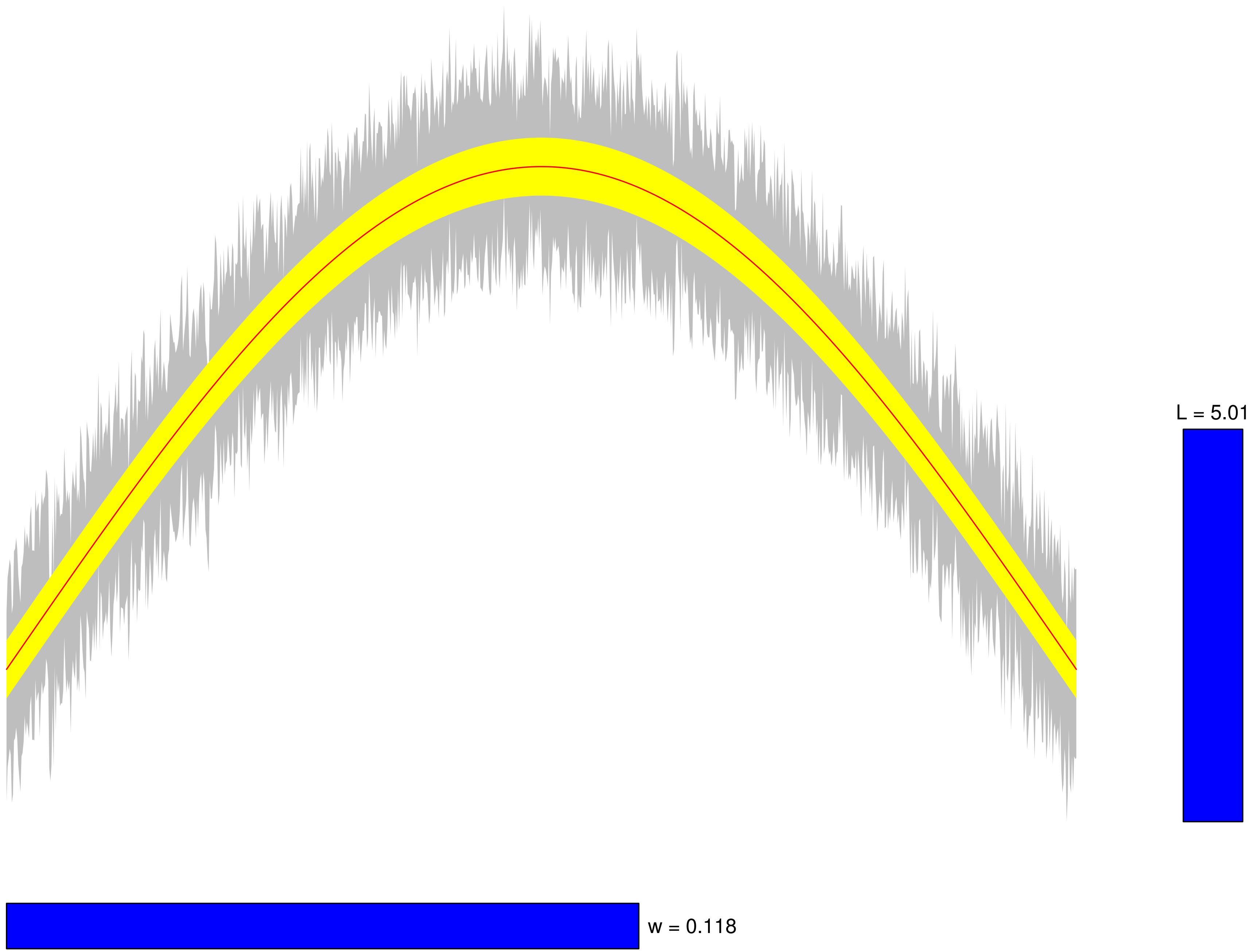
- Lipschitz Catalog. Assume $f \in \mathcal{F}_L$ for some $L > 0$ where $\mathcal{F}_L = \{f : |f(x) - f(y)| \leq L|x-y|, \forall x, y\}$. Let $\mathcal{A} = (0, \infty)$ index L.
- Local Extremes Catalog. Assume that f has two continuous derivatives and k local extremes. Then, $\mathcal{A} = \{0, 1, 2, \ldots\}$.

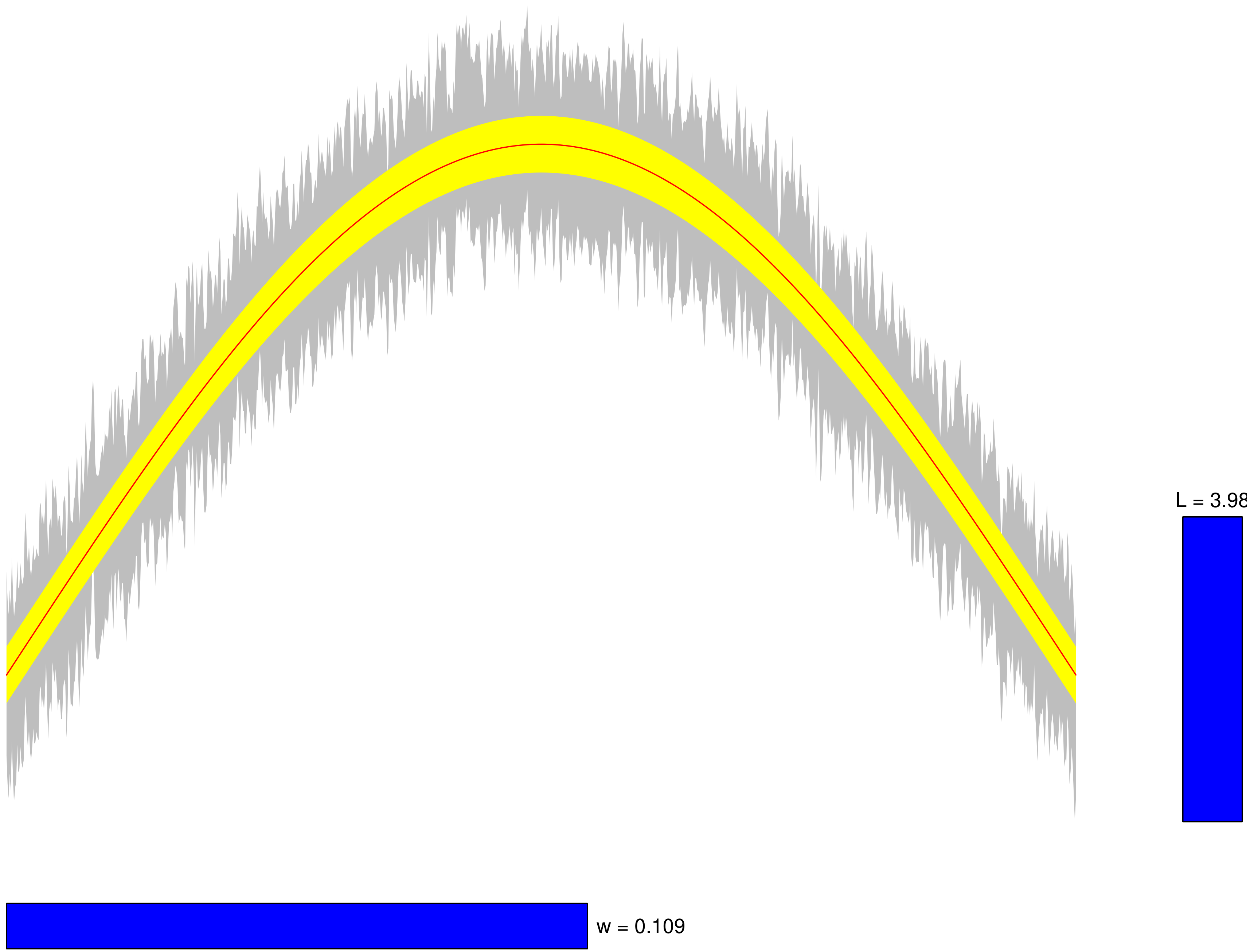
Typical question: What do I have to assume to detect a particular feature? C_α^{-1}(feature significant)

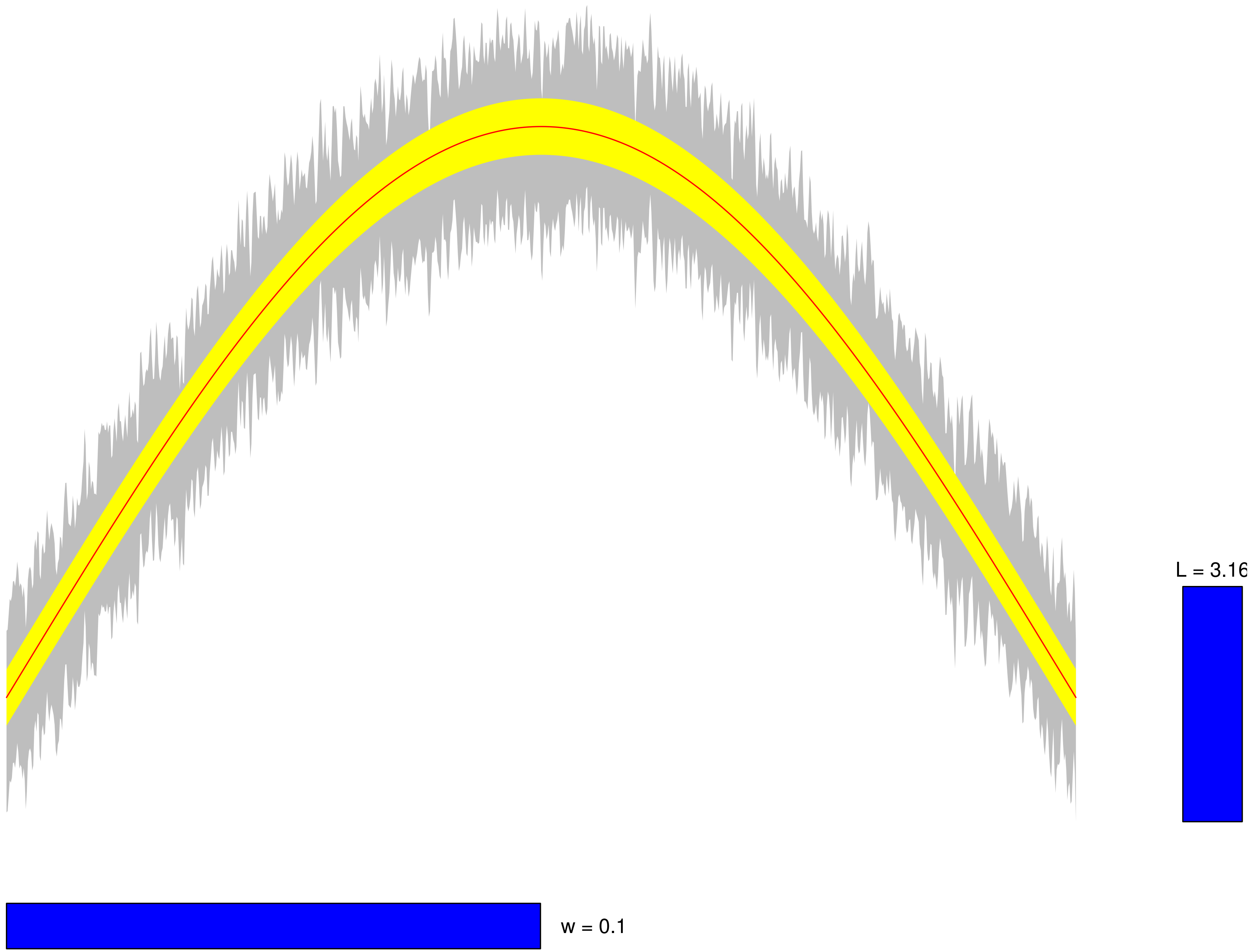

L = 10
w = 0.152

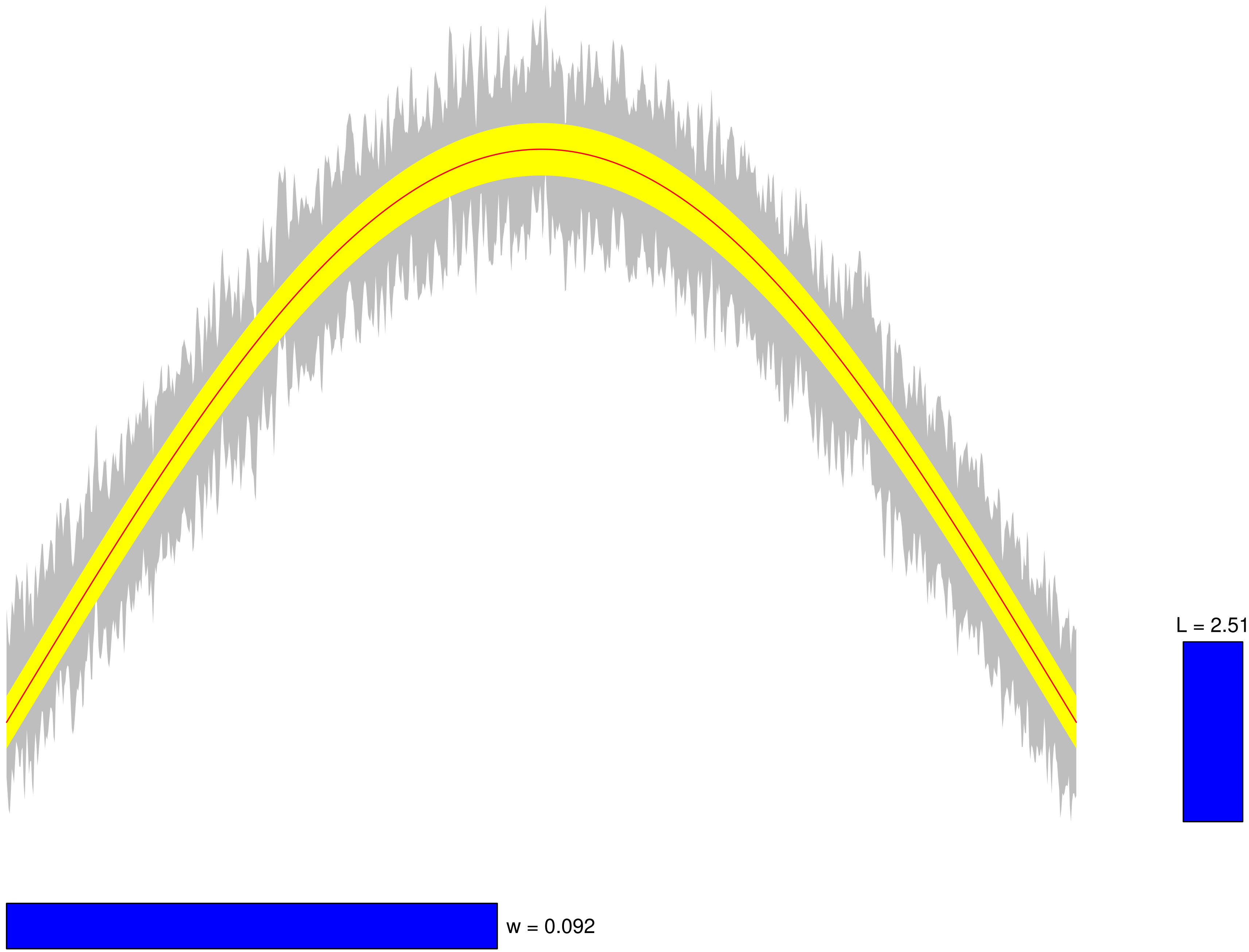

L = 7.94
w = 0.14

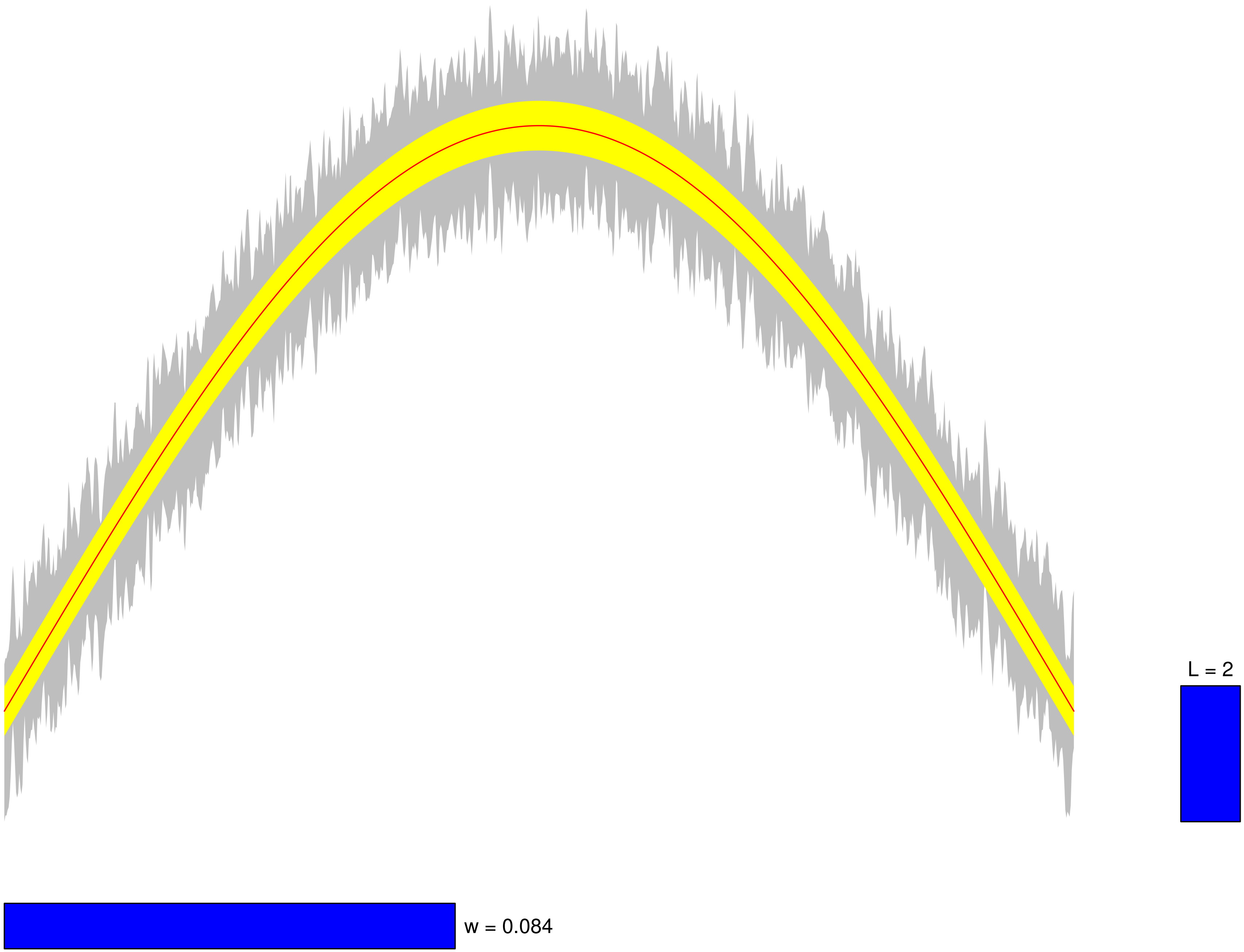

L = 6.31
w = 0.129

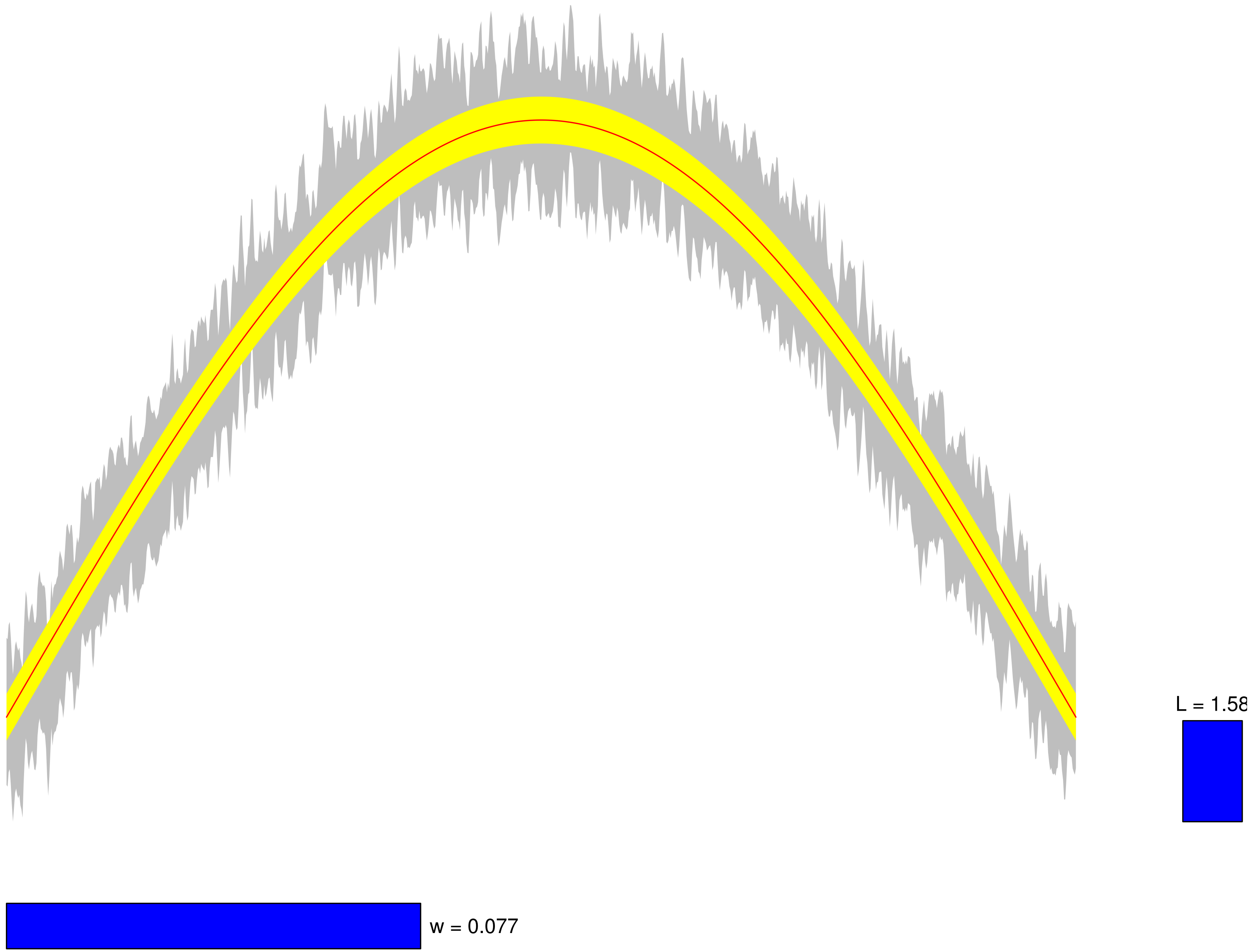

L = 5.01
w = 0.118

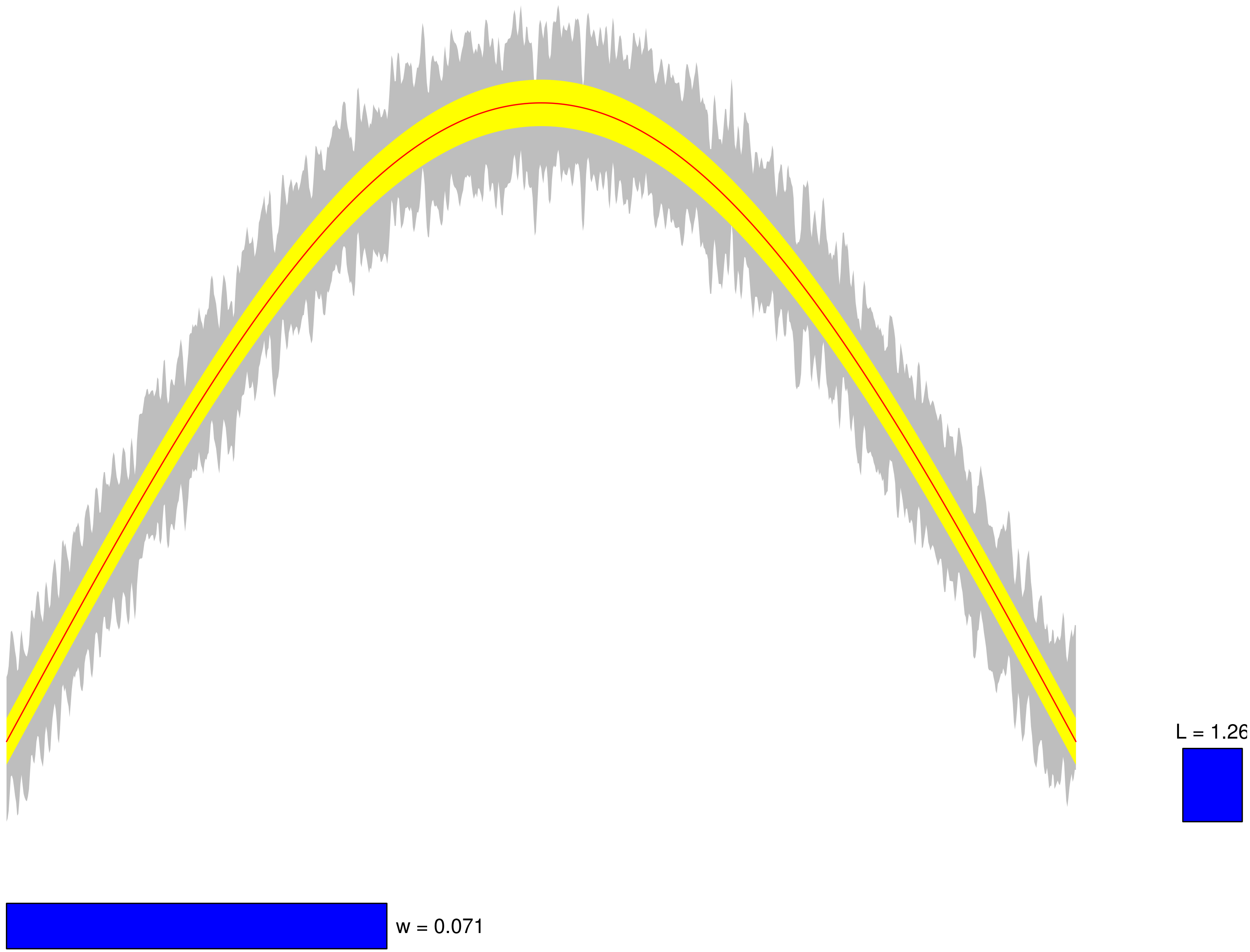

L = 3.98
w = 0.109

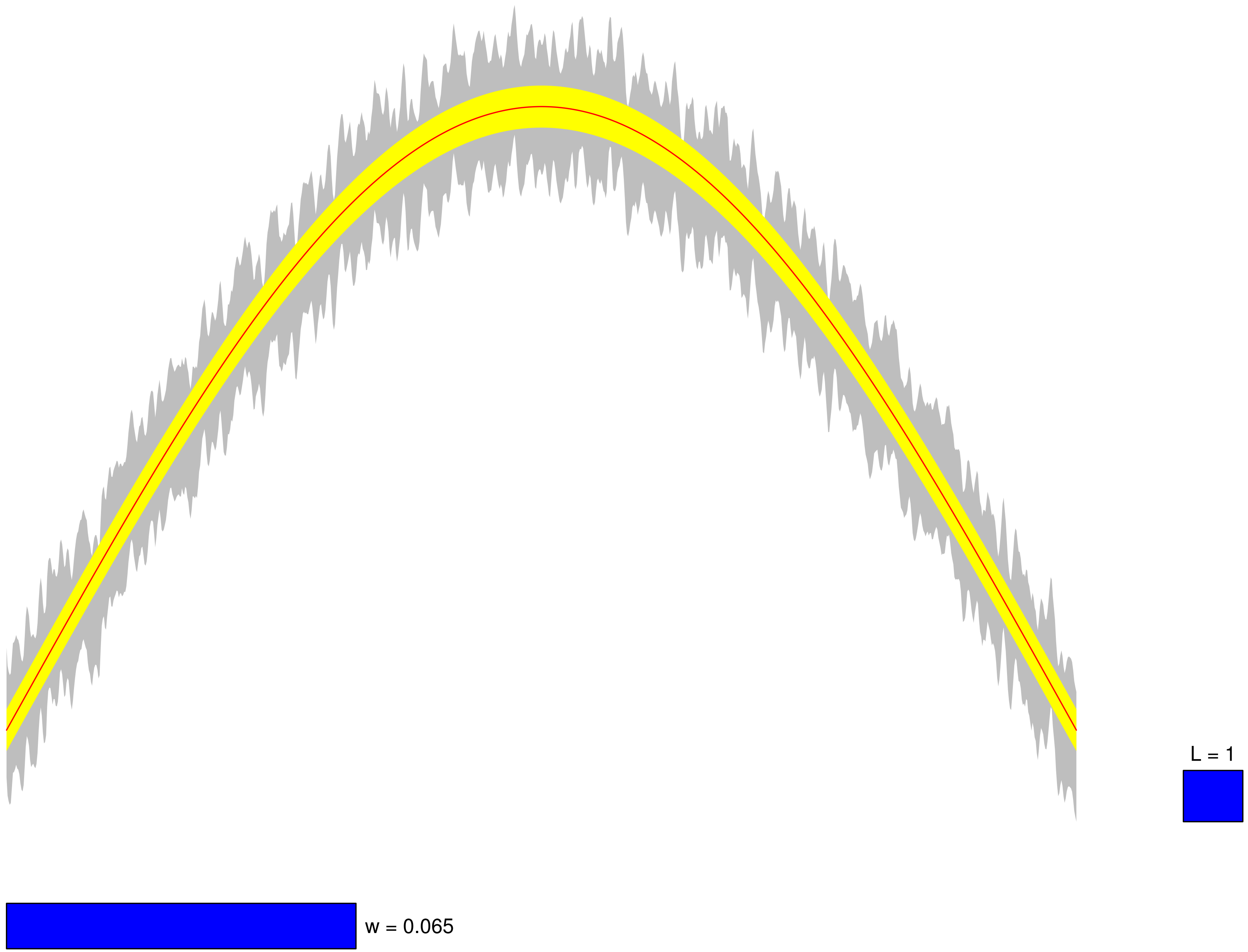

L = 3.16
w = 0.1

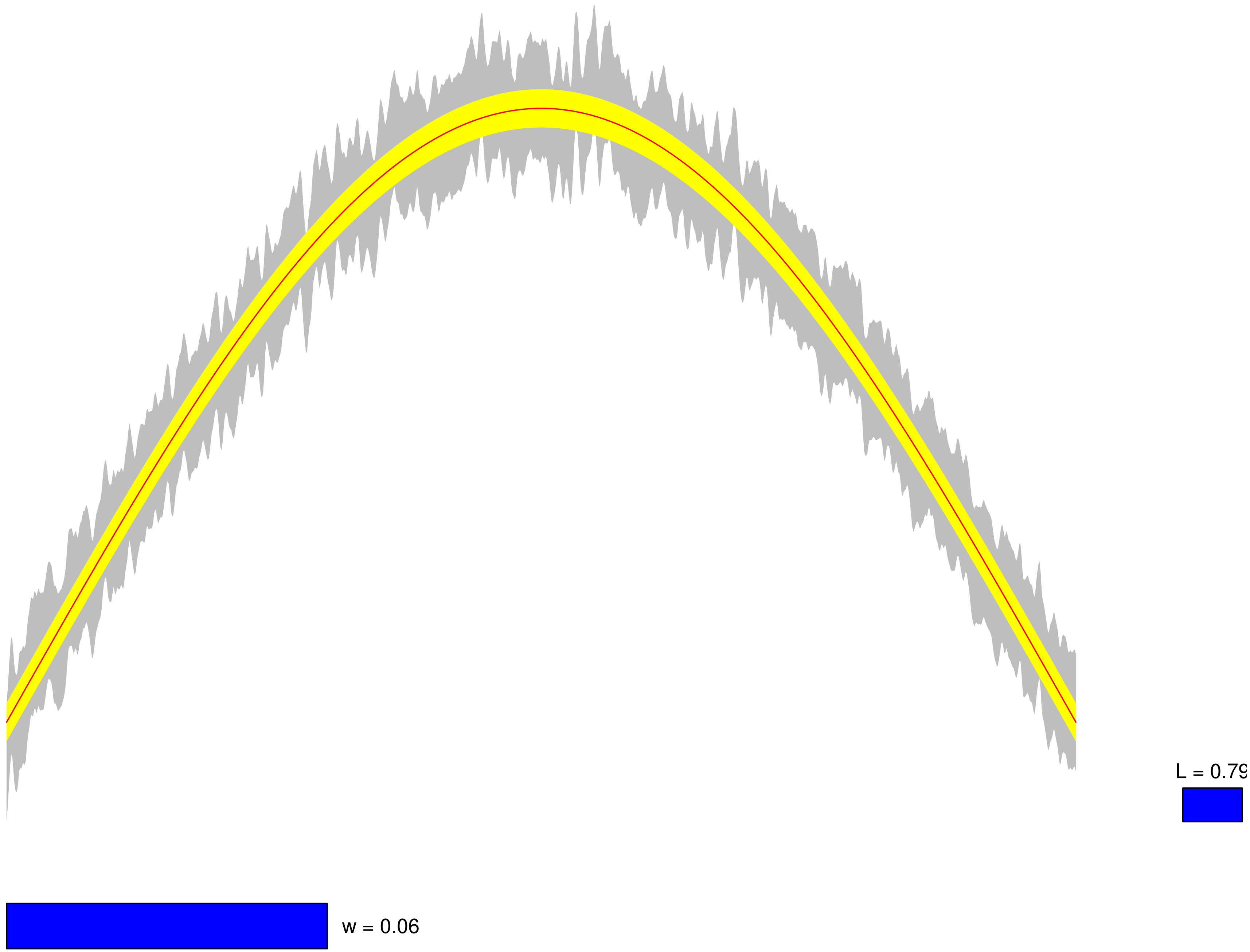

L = 2.51
w = 0.092

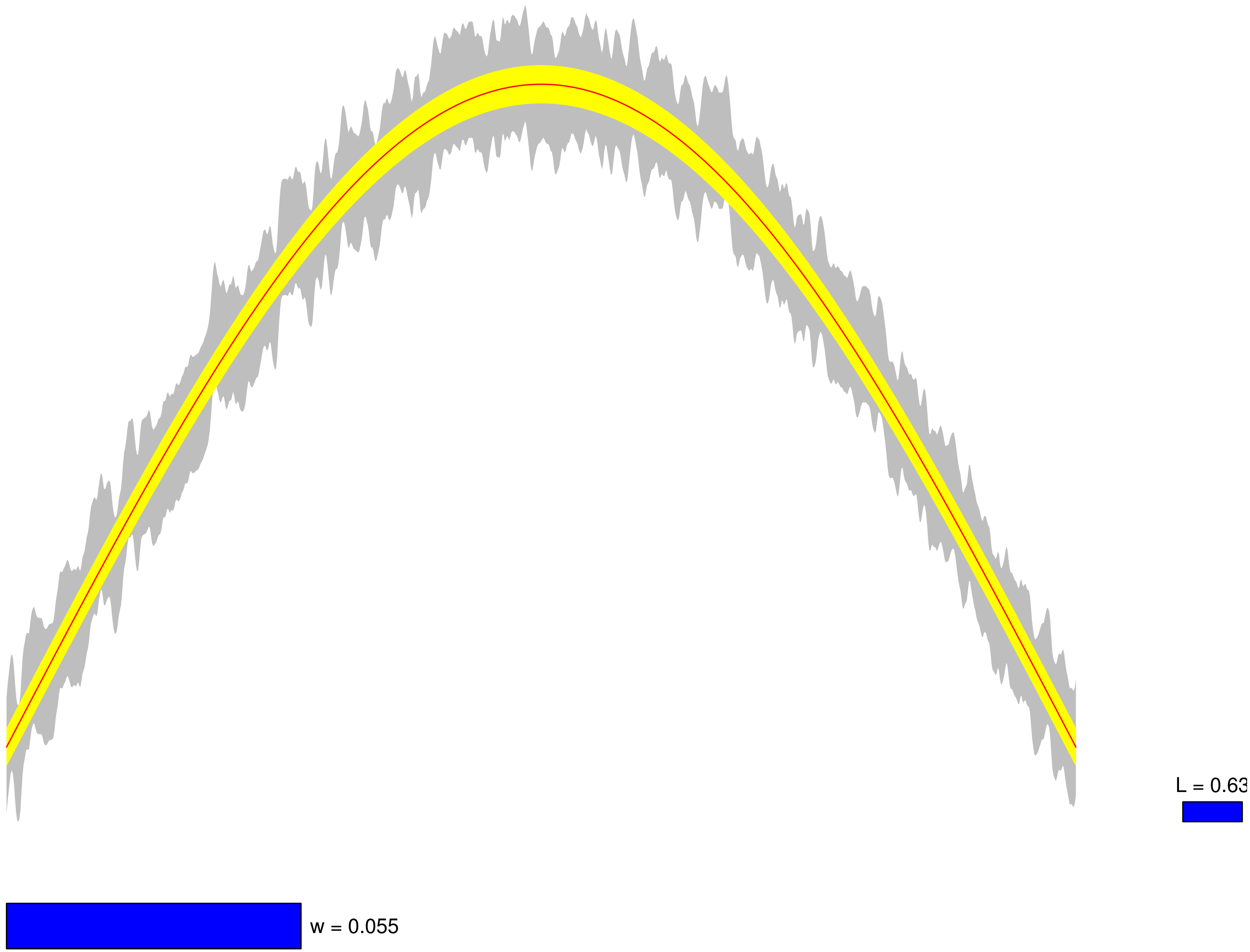

L = 2
w = 0.084

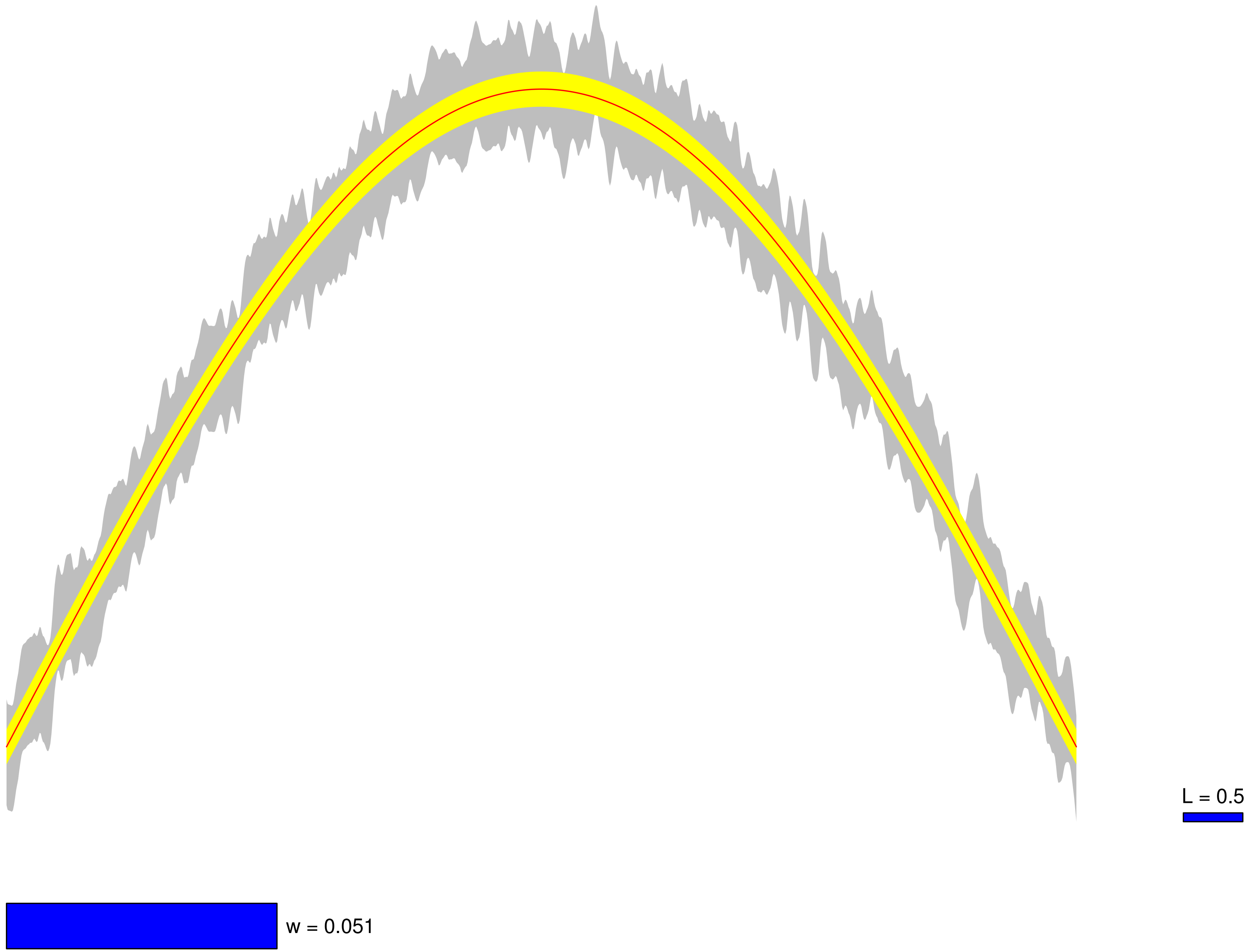

L = 1.58
w = 0.077


L = 1.26
w = 0.071

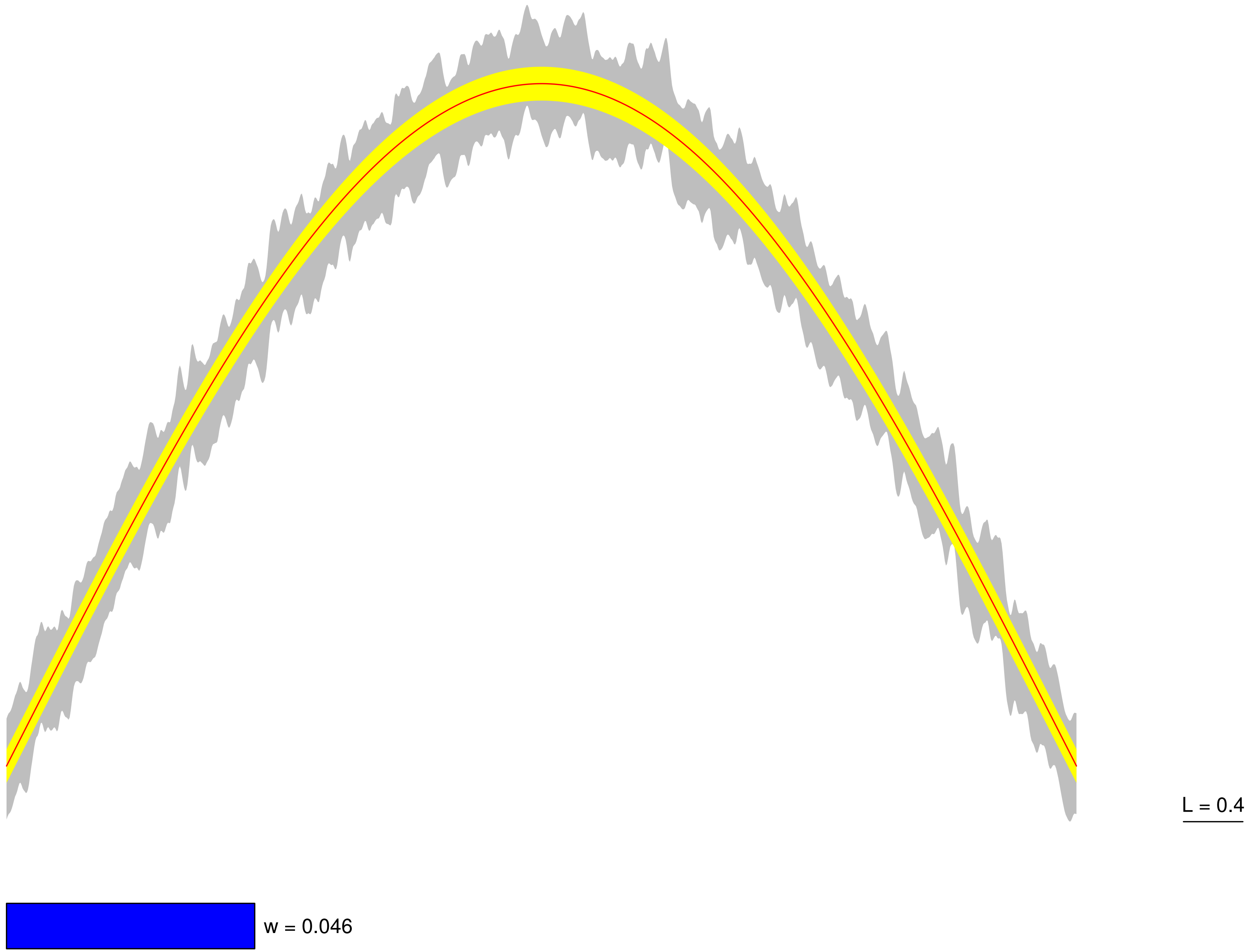

L = 1
w = 0.065


L = 0.79
w = 0.06


L = 0.63
w = 0.055

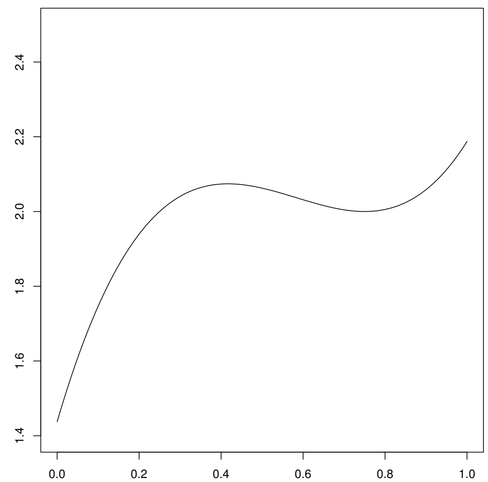


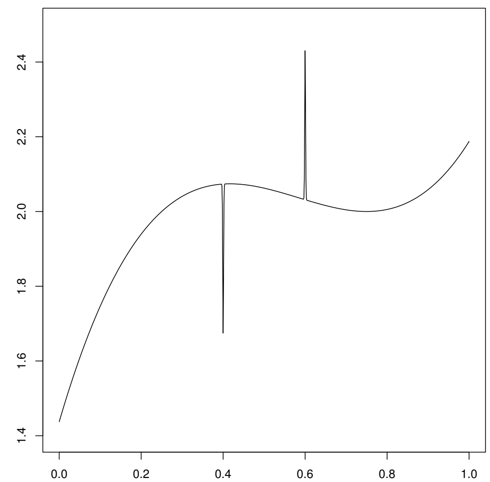
L = 0.5
w = 0.051


L = 0.4
w = 0.046

ϵ-Coverage

But perhaps the standard notion of coverage is too stringent. Missing only a few sharp features might be considered a success.





Consider weakening "coverage" to include only the features for which we have power to test.

Define a partial order $g \preceq f$ that measures "complexity", as derived for instance from a confidence catalog.

ϵ-Coverage (cont'd)

Examples of partial orders:

1. Linear Sieve

   Let $\phi_1, \phi_2, \ldots$ denote an orthonormal basis on $[0, 1]$.

   If $f = \sum_{j=1}^{n} \theta_j \phi_j$ and $g = \sum_{j=1}^{n} \beta_j \phi_j$, write $g \preceq f$ if for all $1 \leq j \leq n$, $\sum_{k=1}^{j} \beta_k^2 \leq \sum_{k=1}^{j} \theta_k^2$.

2. k-Local Extremes.

   $g \preceq f$ if g has fewer (inclusive) local extremes than f.

3. Wavelet Sieve

Given an inaccuracy measure $d(f, g)$, define ϵ-coverage by

$$\text{coverage}_\epsilon(\ell, u) = \inf_{f \in \mathbb{R}^n} \sup_{\substack{g \preceq f \\ d(f,g) \leq \epsilon}} \mathsf{P}_f \left\{ \ell \leq g \leq u \right\}$$

We want a procedure that gives $\text{coverage}_\epsilon(\ell, u) \geq 1 - \alpha$.

ϵ-Coverage: Linear Sieve Example

Define

- $d(f, g) = \|f - g\|_2$,
- $T_j = n \sum_{k=j+1}^{n} \widehat{\theta}_j^2 / \sigma^2$ for each j,
- $\widehat{J} = \min\{j : \ T_j \leq \chi^2_{\frac{\alpha}{2n}, n-j}\}$,
- (ℓ_j, u_j), which are $1 - \frac{\alpha}{2n}$ confidence bands for model j, and
- $B = (\ell, u) \equiv (\ell_{\widehat{J}}, u_{\widehat{J}})$.

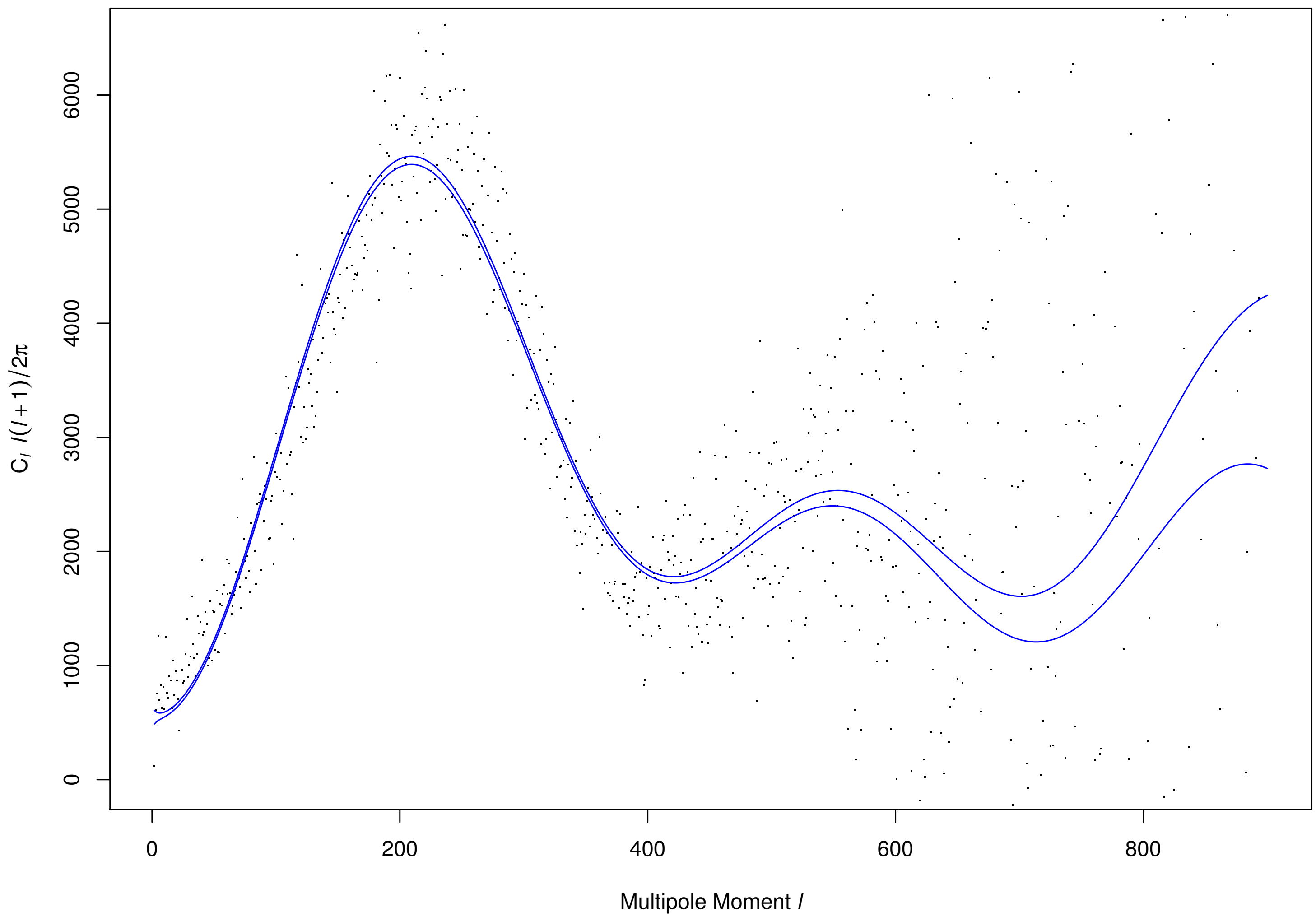
Theorem. There are constants C_{α, n_0} such that if $n \geq n_0$ and

$$\epsilon \geq C_{\alpha, n_0} \, \sigma \, n^{-1/4} \sqrt{2 \log n},$$

then $B = (\ell, u)$ is a $1 - \alpha$ ϵ-confidence band for f.

In progress: similar results for the k-Local Extremes ordering and several flavors of Wavelet sieve.

Example: CMB Power Spectrum



Resolution Limited Inference

There is a resolution-uncertainty relation that is connected to, though distinct from, the usual bias various trade-off.

Consider making inferences about f at varying levels of resolution.

At low resolution, discover few features because of large degree of smoothing.

At high resolution, discover few features because of large uncertainties.

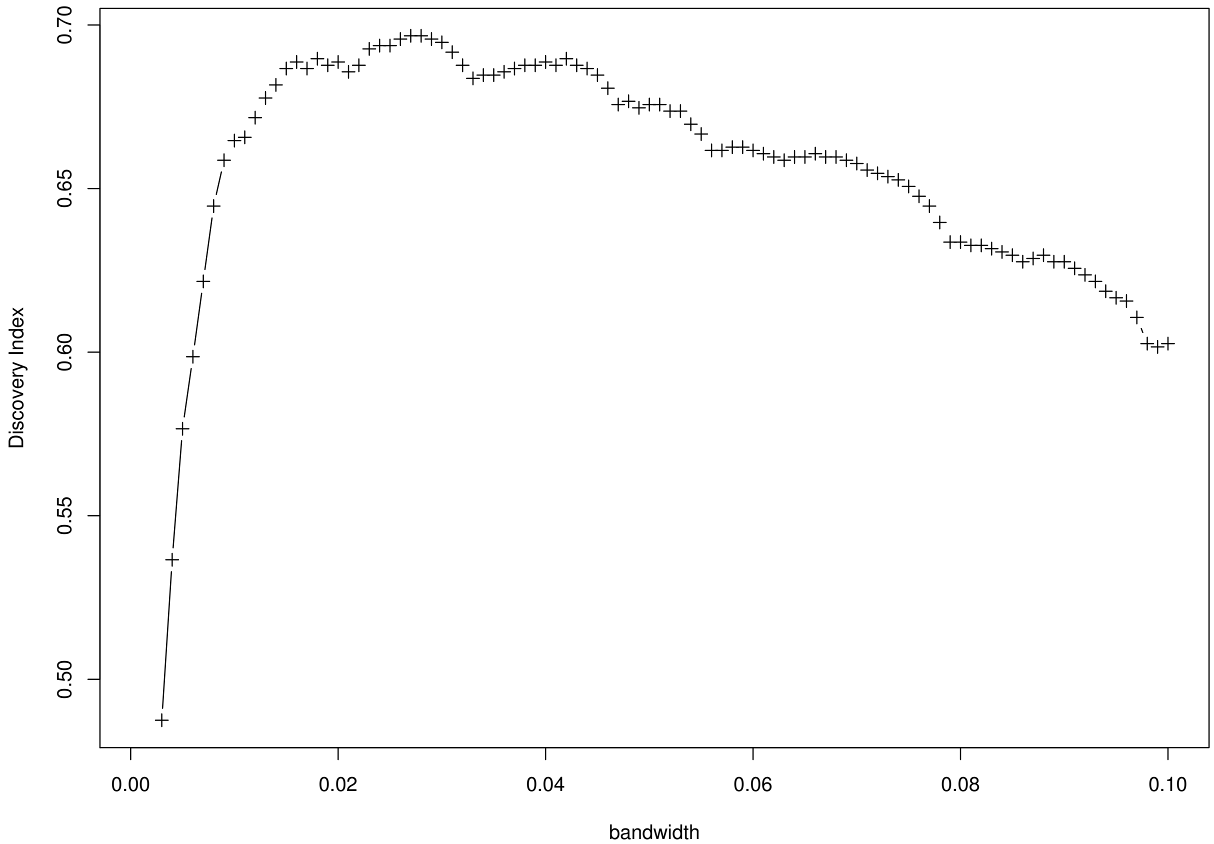
For example, given confidence bands (ℓ, u) (and mean zero data for simplicity), define the Discovery Index

$$D = \int 1\Big\{(\ell(x), u(x)) \not\ni 0\Big\}\, dx.$$

If h is the bandwidth for a local linear estimator, then we'd expect D to be small for both small and large h. Find h to maximize D.

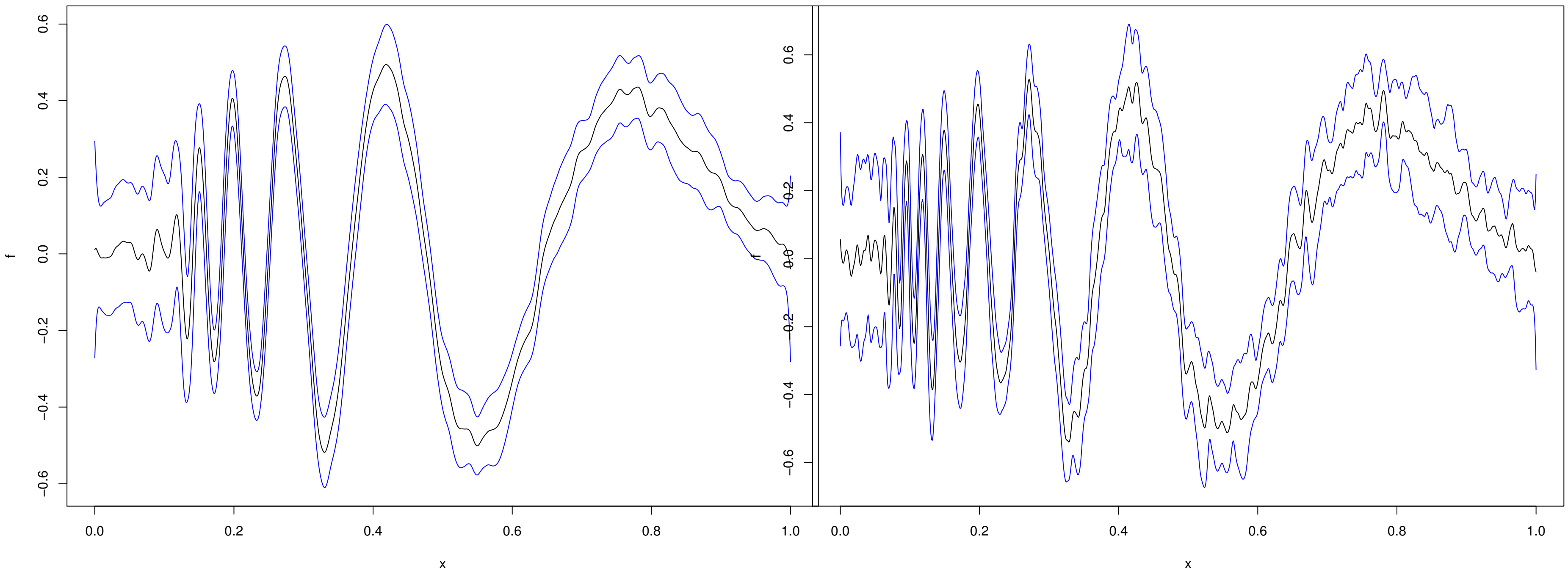
Resolution Limited Inference (cont'd)

Example: Doppler



$\widehat{h}_{\mathrm{DI}} \approx 0.27$, $\widehat{h}_{\mathrm{CV}} \approx 0.12$.

Resolution Limited Inference: (cont'd)



$\widehat{h}_{\mathrm{DI}}$ $\widehat{h}_{\mathrm{CV}}$

Road Map

1. Constructing Confidence Sets for Unknown Functions
   - Simultaneity, Bias, and Relevance
   - Example: The Pivot-Ball Construction
   - Pivot-Ball Confidence Sets for Wavelet Bases
2. Making Inferences with Uniform Confidence Balls
   - Simultaneous Inferences about sets of functionals
   - Case Study: The Cosmic Microwave Background
   - Keeping Our Eyes on the Ball: Parametric Probes and Confidence Catalogs
3. Non-Adaptive Inference
   - If Confidence Bands Cannot Adapt . . .
   - Confidence Catalogs
   - ϵ-Coverage
   - The Resolution-Uncertainty Tradeoff

Take-Home Points

- Inference as Search
  - Build confidence sets with uniform coverage and post-hoc protection.
  - Constrain and search post-hoc
  - Can yield effective and practically useful inferences.
- Adaptation
  - It is possible to construct adaptive confidence balls, though adaptation is limited (and delicate).
  - Confidence bands, in most cases, cannot adapt.
  - Without adaptation, inferences driven by worst-case assumptions.
- Strategies for Non-Adaptive Inference
  - Consider the impact of the assumptions explicitly. (Need this to be efficiently realized.)
  - Change the notion of coverage.
  - Tune inferences to appropriate resolution level.

Appendix I: Tube Formula

Special case: $f(x) = \langle \ell(x), \theta \rangle$. Then, for $T(x) = \ell(x)/\|\ell(x)\|$,

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

By conditioning on $\|\epsilon\|$ this reduces to finding the volume of a tube on the sphere S^{n-1} around the image of T.

Approximation found by Hotelling (1939), generalized by Weyl (1939), well studied since then.

Must account for bias in general.

Appendix II

Pivot-Ball Simulations

Pivot-Ball Method: Simple Simulation

Test functions:

$$f_0(x) = 0$$

$$f_1(x) = 2(6.75)^3 x^6 (1-x)^3$$

$$f_2(x) = \begin{cases} 1.5 & \text{if } 0 \le x < 0.3 \\ 0.5 & \text{if } 0.3 \le x < 0.6 \\ 2.0 & \text{if } 0.6 \le x < 0.8 \\ 0.0 & \text{otherwise.} \end{cases}$$

Let $\alpha = .05$, $n = 1024$, $\sigma = 1$, and use 5000 iterations.

For comparison, χ^2 radius is 1.074.

Pivot-Ball Method: Simple Simulation (cont'd)

σ known:

Method	Function	Coverage	Average Radius
SureShrink (levelwise)	f_0	0.944	0.268
	f_1	0.940	0.289
	f_2	0.927	0.395
Modulator (cosine)	f_0	0.931	0.253
	f_1	0.930	0.259
	f_2	0.905	0.318

σ unknown:

Method	Function	Coverage
SureShrink (levelwise)	f_0	0.954
	f_1	0.953
	f_2	0.929
Modulator (cosine)	f_0	0.999
	f_1	0.999
	f_2	0.997

Appendix III

Details on simultaneous coverage of block average functionals with Pivot-Ball confidence set.

Pivot-Ball Method: Functionals, Details

For $\eta, c > 0$, define

$$\mathcal{F}_{\eta,c} = \bigcup_{p,q \geq 1} \bigcup_{\gamma \geq 1/2+\eta} \mathcal{B}_{p,q}^{\varsigma(\gamma)}(c),$$

with $\varsigma(\gamma) = \gamma + (1/p - 1/2)_+$. The parameter η is an increment of smoothness required only in the non-sparse case ($p \geq 2$).

Let

$$\kappa = \sup\left\{\#\{\psi_{jk}(x) \neq 0 : 0 \leq k < 2^j\} \colon 0 \leq x \leq 1, j \geq J_0\right\}$$

be the maximal number of $\psi_{j.}$ that "hit" a given point.

Pivot-Ball Method: Functionals, Details (cont'd)

Theorem. Let ψ, ϕ be compactly supported wavelets with $\kappa < \infty$ and $\|\psi\|_1 < \infty$.

If for some $0 \leq \zeta \leq 1$,

$$\Delta_n^{-1} = o(n^\zeta / (\log n)^{\lfloor \zeta \rfloor})$$

then for any sequence $w_n \geq 0$ satisfying

$$w_n \to 0 \qquad \text{and} \qquad \liminf_{n\to\infty} w_n n^{1-\zeta} (\log n)^{\lfloor \zeta \rfloor} > 0,$$

we have

$$\liminf_{n\to\infty} \inf_{f\in\mathcal{F}_{\eta,c}} \mathrm{P}\Big\{ T(f) \in J_n(T) \text{ for all } T \in \mathcal{T}_n \Big\} \geq 1 - \alpha.$$

The conditions are satisfied by most standard wavelet functions.

Appendix IV

Subspace Pretesting: Baraud 2004 and extensions.

Subspace Pretesting (Baraud 2004)

Write $\boldsymbol{Y} = \boldsymbol{f} + \sigma\boldsymbol{\epsilon}$ where $\boldsymbol{f} = (f(x_1), \ldots, f(x_n))$.

Baraud procedure constructs finite-sample confidence ball for $\boldsymbol{f}$.

Ideal Construction: Control $\|\boldsymbol{f} - \widehat{\boldsymbol{f}}\|$ uniformly over $\boldsymbol{f}$ for good $\widehat{\boldsymbol{f}}$.

But this doesn't work.

Let $S \subset \mathbb{R}^n$ be a subspace of dimension $< n$ and define π_S to be orthogonal projection onto S.

If $\widehat{\boldsymbol{f}} \equiv \pi_S \boldsymbol{Y}$, then $\|\boldsymbol{f} - \widehat{\boldsymbol{f}}\|^2 = \|(I - \pi_S)\boldsymbol{f}\|^2 + \sigma^2 \|\pi_S \boldsymbol{\epsilon}\|^2$.

We usually cannot bound $\|(I - \pi_S)\boldsymbol{f}\|^2$ a priori.

Instead: Use pretest to control $\|(I - \pi_S)\boldsymbol{f}\|^2$. Specifically, use $(I - \pi_S)\boldsymbol{Y}$ to test $\boldsymbol{f} \in S$ versus $\boldsymbol{f} \notin S$.

When don't reject $\boldsymbol{f} \in S$, then $\|(I - \pi_S)\boldsymbol{f}\|^2$ is small with high-probability.

Subspace Pretesting (cont'd)

Let $\mathcal{S}$ be a collection of subspaces S such that $\mathbb{R}^n \in \mathcal{S}$.

Example: $f(x) = \sum_j \theta_j \phi_j(x)$ and S_j corresponds to j-term partial sums.

For $S \in \mathcal{S}$, let $\widehat{\boldsymbol{f}}_S = \pi_S \boldsymbol{Y}$ and choose α_S such that $\sum_S \alpha_S \le \alpha$.

Choose tests and radii ρ_S for $S \in \mathcal{S}$ so that

$$\mathsf{P}\Big\{\mathsf{Ball}(\widehat{\boldsymbol{f}}_S, \rho_S) \not\ni \boldsymbol{f} \text{ and Don't reject } H_0 : \boldsymbol{f} \in S\Big\} \le \alpha_S.$$

If $\widehat{S} = \mathrm{argmin}_{S \in \mathcal{S}}\, \rho_S$, then

$$\mathsf{P}\Big\{\mathsf{Ball}(\widehat{\boldsymbol{f}}_{\widehat{S}}, \rho_{\widehat{S}}) \not\ni \boldsymbol{f}\Big\} \le \sum \mathsf{P}\Big\{\boldsymbol{f} \notin \mathsf{Ball}(\widehat{\boldsymbol{f}}_S, \rho_S) \text{ and Don't reject S}\Big\}$$
$$\le \sum \alpha_S \le \alpha.$$

Can get a smaller set by taking intersection of balls.

Extensions to Subspace Pretesting: Bands

Let $\phi_1, \phi_2, \ldots$ be bounded, ortho. basis on $[0, 1]$, e.g., cosine basis.

Assume $f = \sum_{j=1}^{J} \theta_j \phi_j$ for specified $J \equiv J_n$. Often take $J_n = n$.

Consider subspaces S_d of partial sums, and let $\widehat{f}_d = \sum_{j=1}^{d} \widehat{\theta}_j \phi_j$, where $\widehat{\theta}_j \approx$ ind $N(\theta_j, \sigma^2/n)$.

Define $a_d = \sqrt{\sum_{j=1}^{d} \phi_j^2}$ and $b_d = \sqrt{\sum_{j=d+1}^{J} \phi_j^2}$.

We consider confidence bands

$$\mathcal{B} = \cap_d \mathcal{B}_d \quad \text{where} \quad \mathcal{B}_d = \left\{ f\colon\, |f(x) - \widehat{f}_d(x)| \leq \frac{\sigma}{\sqrt{n}} \delta_d a_d(x) \right\}.$$

Following Baraud, choose tests and define δ_ds so that

$$\mathsf{P}\left\{ \mathcal{B}_d \not\ni f \text{ and } S_d \text{ not rejected} \right\} \leq \alpha_d,$$

then $\mathcal{B}$ is a $1 - \alpha$ confidence set that accounts for the bias.

Subspace Pretesting: Bands (cont'd)

Define the following:

1. Normalized basis functions $T_{dj} = \phi_j / a_d$ and $\widetilde{T}_{dj} = \phi_j / b_d$.
2. Bias function $z_{\theta,d} = \frac{b_d}{a_d} \sum_{j=d+1}^{J} \theta_j \widetilde{T}_{dj}$.
3. Maximum bias
$$W_d = \left\| \frac{\widehat{f}_d - f}{a_d} \right\|_\infty = \sup_x \left| \frac{\sigma}{\sqrt{n}} \sum_{j=1}^{d} Z_j T_{dj}(x) + z_{\theta,d}(x) \right|$$
4. Test statistic
$$U_d = \left\| \frac{\widehat{f}_d - \widehat{f}_J}{a_d} \right\|_\infty = \sup_x \left| \frac{\sigma}{\sqrt{n}} \frac{b_d(x)}{a_d(x)} \sum_{j=1}^{d} Z_j \widetilde{T}_{dj}(x) + z_{\theta,d}(x) \right|.$$
5. Distributions $\overline{G}_{\theta,d}(w) = \mathsf{P}\Big\{ W_d > w \Big\}$ and $H_{\theta,d}(u) = \mathsf{P}\Big\{ U_d \leq u \Big\}$.
6. Critical value for test $c_d = H_{0,d}^{-1}(1 - \gamma_d)$ where $0 < \gamma_d < 1 - \alpha_d$.

Subspace Pretesting: Bands (cont'd)

Set

$$\delta_d = \sup_\theta \frac{\sqrt{n}}{\sigma} \overline{G}_{\theta,d}^{-1}\left(\frac{\alpha_d}{H_{\theta,d}(c_d)}\right).$$

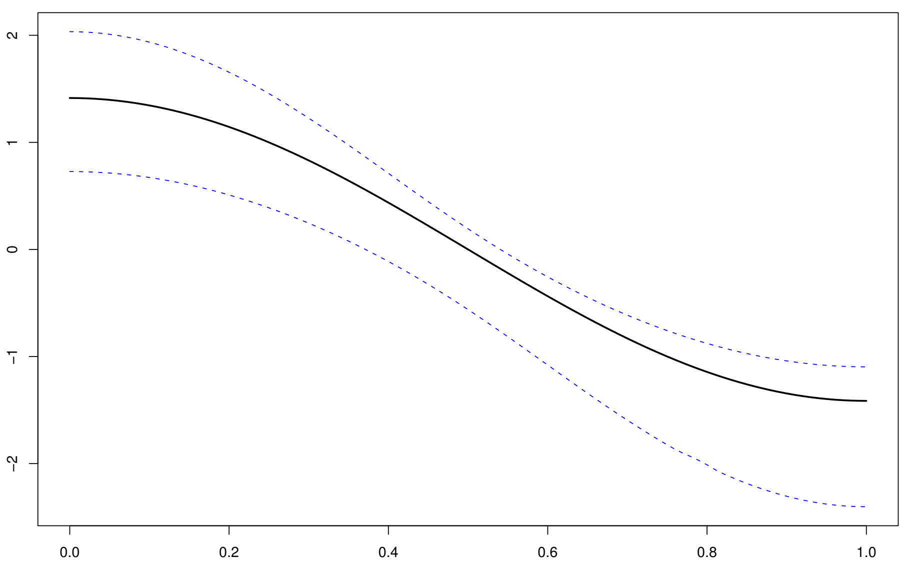
We have

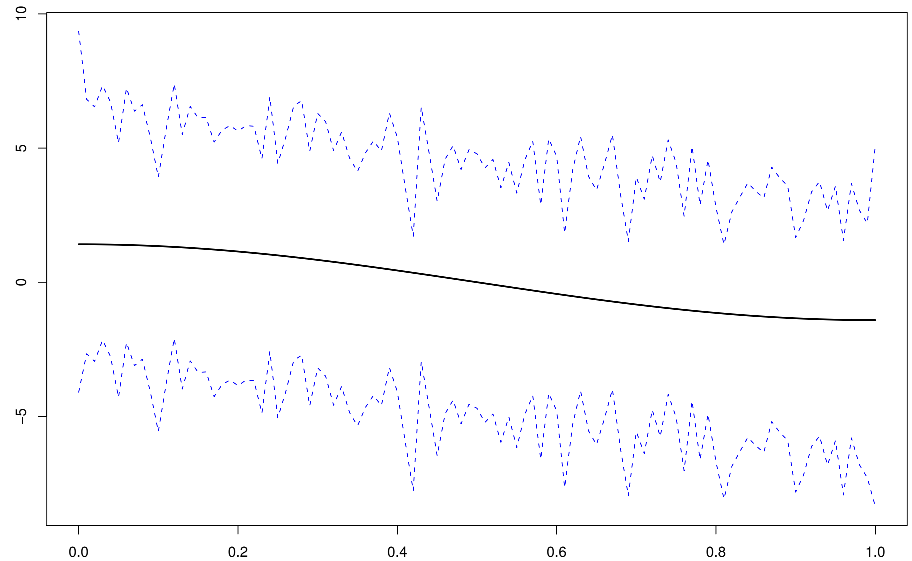
$$\begin{aligned}
\mathsf{P}\Big\{\mathcal{B}_d \not\ni f, U_d \le c_d\Big\} &= \mathsf{P}\Big\{\mathcal{B}_d \not\ni f\Big\}\mathsf{P}\Big\{U_d \le c_d\Big\} \\
&= \mathsf{P}\left\{\left\|\frac{\widehat{f}_d - f}{a_d}\right\|_\infty > \frac{\sigma}{\sqrt{n}}\delta_d\right\} H_{\theta,d}(c_d) \\
&= \overline{G}_{\theta,d}\Big(\frac{\sigma}{\sqrt{n}}\delta_d\Big) H_{\theta,d}(c_d) \\
&\le \alpha_d.
\end{aligned}$$

By approximation, can reduce the δ_d feasible computation.

Not rate adaptive but produces good, smooth bands.

Pre-testing Band Examples



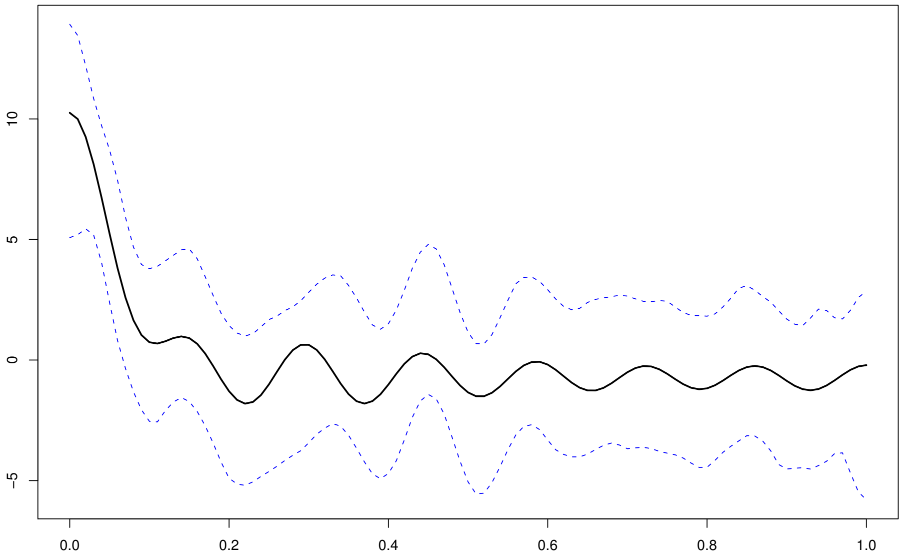


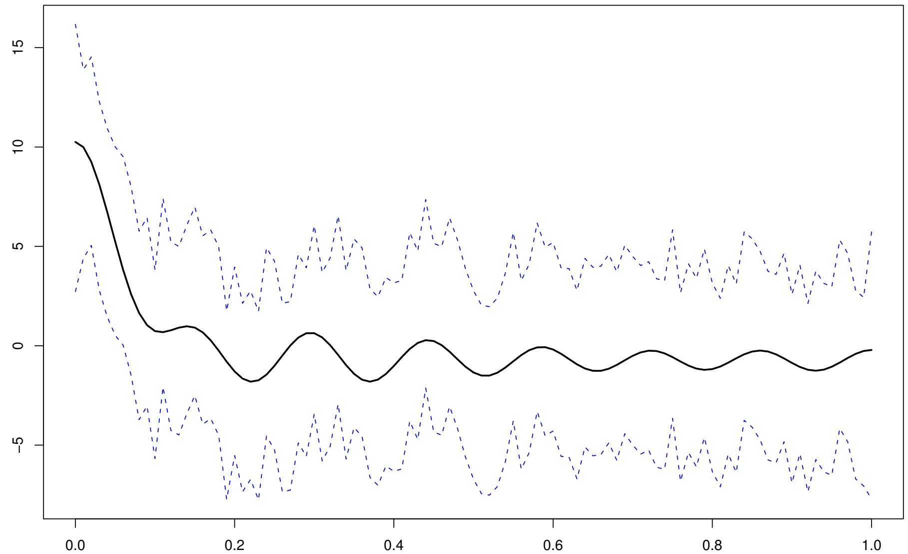
Approximate Pre-testing Band

Full basis

($n = 100$, $\alpha = 0.05$)

Pre-testing Band Examples (cont'd)



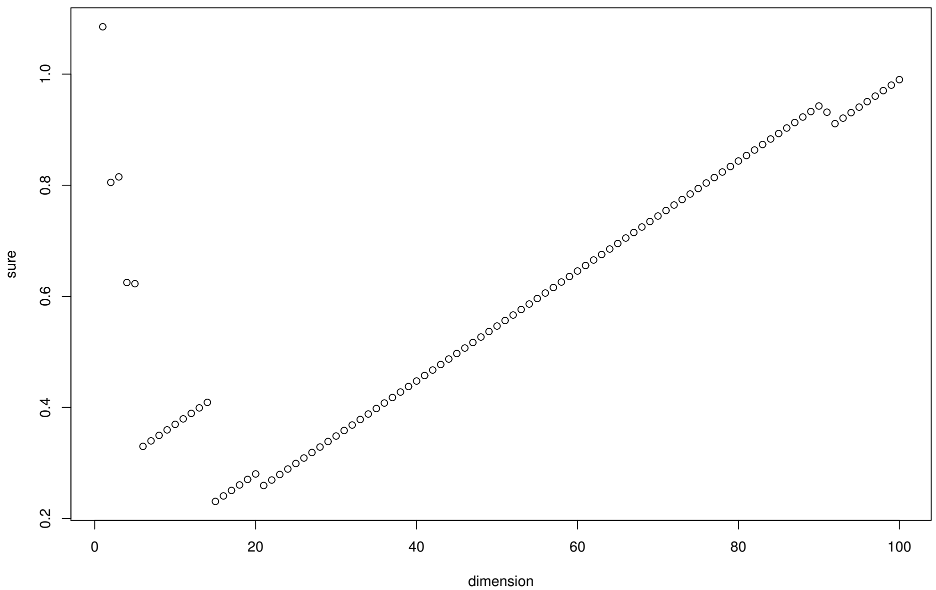


Approximate Pre-testing Band

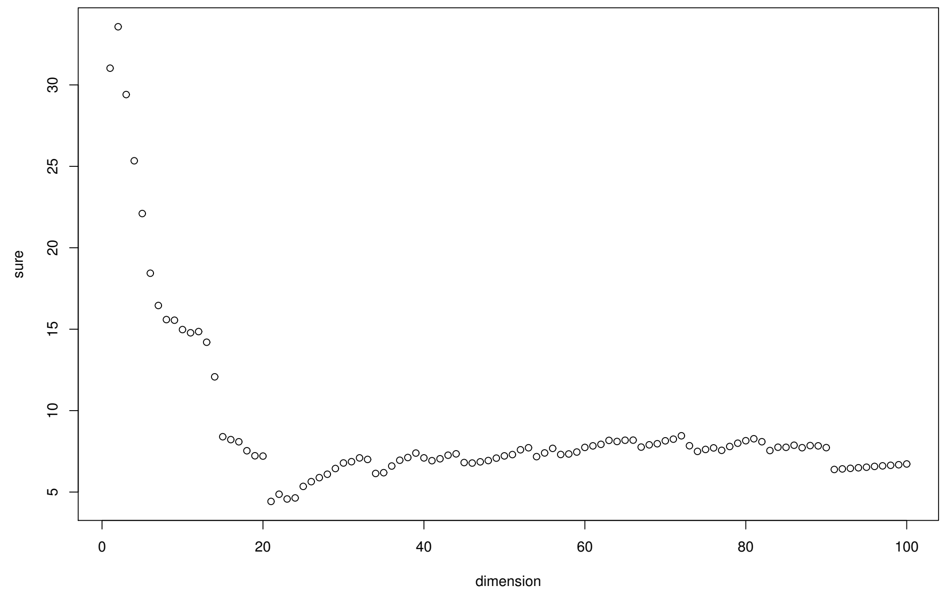
Full basis

($n = 100$, $\alpha = 0.05$)

Pre-testing Band Examples (cont'd)



SURE



Maximum Width