Optimal Inference with a Multidimensional Multiscale Statistic

Bodhisattva Sen[1]
Department of Statistics
Columbia University, New York
bodhi@stat.columbia.edu

Laboratoire de Probabilités, Statistique et Modélisation (LPSM)
Sorbonne Université, Jussieu
Paris, France

Joint work with Pratyay Datta (Columbia University)

22 June, 2018

[1]Supported by NSF grants DMS-1712822 and AST-1614743

Outline

- **Nonparametric regression**: Estimate $f : [0,1] \to \mathbb{R}$ from data

$$Y_i = f\left(\frac{i}{n}\right) + \epsilon_i, \qquad \epsilon_i \overset{i.i.d.}{\sim} N(0,1),\ i = 1, \ldots, n \tag{1}$$

Gaussian (continuous) white noise model

- Observe stochastic process $\{Z(t) : t \in [0,1]\}$ where

$$Z(t) = \sqrt{n} \int_0^t f(x)dx + W(t),^a \qquad t \in [0,1] \tag{2}$$

- $W(\cdot)$: *Brownian motion* on [0,1]; n: "*sample size*"

- **Goal**: Make (optimal) *inference* about $f \in L_1([0,1])$ (unknown)

[a] $\frac{1}{\sqrt{n}} \sum_{i=1}^{j} Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^{j} f(i/n) + \frac{1}{\sqrt{n}} \sum_{i=1}^{j} \epsilon_i$ " $\Leftrightarrow$ " $Z(t) = \sqrt{n} \int_0^t f(x)dx + W(t)$

- (1) and (2) are *asymptotically equivalent* (Brown and Low, 1996) under suitable conditions

- Any asymptotic solution to one of these problems will automatically yield a corresponding solution to the other

Testing the hypothesis $H_0 : f = 0$

- Consider testing

$$H_0 : f = 0 \quad \text{versus} \quad H_1 : f \neq 0 \in \mathbb{H}_{\beta,L}$$

- $\mathbb{H}_{\beta,L} := \left\{ g : [0,1] \to \mathbb{R} : |g(a) - g(b)| \leq L|a-b|^{\beta} \;\; \forall\, a, b \in [0,1] \right\}$ denotes the *Hölder class* of functions with $\beta \in (0,1]$ and $L > 0$

- For $k < \beta \leq k+1$, with an integer $k > 0$, $\mathbb{H}_{\beta,L}$ denotes the set of functions that are k times differentiable and whose k-th derivative belongs to $\mathbb{H}_{\beta-k,L}$

Testing for a constant signal in an interval

- Consider testing

$$H_0 : f = 0 \quad \text{versus} \quad H_1 : f = \mu\, \mathbb{I}_I$$

for *unknown* $\mu \in \mathbb{R}$ and $I \subseteq [0,1]$ *unknown interval*

Model: $Z(t) = \sqrt{n}\int_0^t f(x)dx + W(t), \quad t \in [0,1]$

The multiscale approach: Motivation

- The multiscale approach is based on the idea of *kernel averaging*
- $\psi : \mathbb{R} \to \mathbb{R}$ is a *kernel*[a]; define

$$\psi_{t,h}(x) := \psi\left(\frac{x-t}{h}\right), \qquad x \in [0,1], \quad h > 0$$

- *Kernel estimator*[b]: $\hat{f}_h(t) := \frac{1}{n^{1/2}h\langle 1,\psi\rangle}\int_0^1 \psi_{t,h}(x)dZ(x)$
- **Example**: If $\psi = \mathbb{I}_{[-1,1]}$, $\hat{f}_h(t) = \frac{1}{2\sqrt{n}h}[Z(t+h) - Z(t-h)]$
- **Discrete setting**: $Y_i = f(\frac{i}{n}) + \epsilon_i$, $i = 1, \ldots, n$, when $\psi = \mathbb{I}_{[-1,1]}$,

$$\hat{f}_h\left(\frac{i}{n}\right) = \frac{\sum_{s=i-\lfloor nh\rfloor}^{i+\lfloor nh\rfloor} Y_s}{2\lfloor nh\rfloor + 1}$$

---

[a] $\psi : \mathbb{R} \to \mathbb{R}$ is a measurable function s.t. (i) ψ is 0 outside $[-1,1]$; (ii) $\psi \in L_2(\mathbb{R})$, i.e., $\int_{\mathbb{R}} \psi^2(x)dx < \infty$; (iii) ψ is of finite total variation; and (iv) $\int_{\mathbb{R}} \psi(x)dx > 0$.

[b] For $g_1, g_2 \in L_2(\mathbb{R})$, $\langle g_1, g_2\rangle := \int_{\mathbb{R}} g_1(x)g_2(x)dx$; $\|g\|^2 := \int_{\mathbb{R}} g^2(x)dx$

- For *fixed* $t \in (0,1)$, we can test $f(t) = 0$ by rejecting for *extreme values* of the *normalized* statistic:

$$\hat{\Psi}(t,h) := \frac{\hat{f}_h(t)}{\sqrt{\mathrm{Var}(\hat{f}_h(t))}} = \frac{\int_0^1 \psi_{t,h}(x)dZ(x)}{\sqrt{h}\,\|\psi\|} \sim N(0,1) \ \ (\text{when } f \equiv 0)$$

- So, a naive approach to testing $H_0 : f \equiv 0$ could be to consider

$$\sup_{t \in [h, 1-h]} |\hat{\Psi}(t,h)|,$$

or

$$\sup_{h \in (0,1/2]} \sup_{t \in [h,1-h]} |\hat{\Psi}(t,h)|^{\dagger}$$

- **Discrete setting**: $Y_i = f(\frac{i}{n}) + \epsilon_i$, analogue of $\dagger$, when $\psi = \mathbb{I}_{[-1,1]}$, is

$$\max_{1 \le i < j \le n} \frac{\sum_{s=i}^{j} Y_s}{\sqrt{j-i+1}}$$

- The latter statistic *bypasses the choice* of the bandwidth h, and is a *multiscale statistic* (combines kernel estimators of various scales)

- Recall: $\hat{\Psi}(t,h) = \frac{\hat{f}_h(t)}{\sqrt{\mathrm{Var}(\hat{f}_h(t))}} = \frac{\int_0^1 \psi_{t,h}(x)dZ(x)}{\sqrt{h}\,\|\psi\|}$
- However, under H_0,

$$\sup_{h\in(0,1/2]} \sup_{t\in[h,1-h]} |\hat{\Psi}(t,h)| = +\infty \quad \text{a.s.}$$

- **Proof**: For $\psi = \mathbb{I}_{[-1,1]}$, take $h = 1/(2m)$, $m \geq 1$ integer. Then

$$\begin{aligned} \sup_{t\in[h,1-h]} |\hat{\Psi}(t,h)| &\geq \sup_{t\in\{\frac{1}{2m},\frac{3}{2m},\ldots,\frac{2m-1}{2m}\}} |\hat{\Psi}(t,h)| \\ &\stackrel{d}{=} \max\{|Z_1|,\ldots,|Z_m|\} = O_p(\sqrt{2\log m}) \end{aligned}$$

where $Z_i = \frac{1}{\sqrt{2h}}[W(\frac{2i}{2m}) - W(\frac{2i-2}{2m})] \overset{i.i.d.}{\sim} N(0,1)$

Normalized test statistic: $\hat{\Psi}(t,h) := \frac{1}{\sqrt{h}\,\|\psi\|}\int_0^1 \psi_{t,h}(x)dZ(x)$

Multiscale statistic [Dümbgen and Spokoiny (2001)]

- Multiscale statistic:

$$T(Z,\psi) := \sup_{h\in(0,1/2]}\ \sup_{t\in[h,1-h]} \frac{|\hat{\Psi}(t,h)| - \Gamma(2h)}{D(2h)}$$

where $\Gamma, D : (0,1] \to (0,\infty)$ are two functions defined as

$$\Gamma(r) := \sqrt{2\log(1/r)} \quad \text{and} \quad D(r) := (\log(e/r))^{-1/2}\log\log(e^e/r)$$

- $\Gamma(\cdot)$ — scale calibrating term that puts *different scales on equal footing*
- Prevents *smaller scales* from *dominating* the overall test statistic
- Resolves the *multiple testing* issue; improves *power*

Theorem [Dümbgen and Spokoiny (2001)]

Let $\psi : \mathbb{R} \to \mathbb{R}$ be a kernel. Then, when $f \equiv 0$, $T(Z,\psi) \equiv T(W,\psi) < \infty$ a.s.

Multivariate Gaussian (continuous) white noise model

- Observe stochastic process $\{Z(t) : t \in [0,1]^d\}$ ($d \geq 1$) where

$$Z(t_1, \ldots, t_d) = \sqrt{n} \int_0^{t_1} \ldots \int_0^{t_d} f(x_1, \ldots, x_d) dx_d \ldots dx_1 + W(t_1, \ldots, t_d)$$

- $W(\cdot)$: *Brownian sheet*[a] on $[0,1]^d$ (unobserved)

- **Goal**: Make (optimal) *inference* about f (unknown)

---

[a] A d-dimensional *Brownian sheet* is a *mean-zero Gaussian process* $\{W(t) : t \in [0,1]^d\}$ with covariance kernel

$$\mathrm{Cov}(W(t_1, \ldots, t_d), W(s_1, \ldots, s_d)) = \Pi_{i=1}^d \min(t_i, s_i), \tag{3}$$

for $(t_1, \ldots, t_d), (s_1, \ldots, s_d) \in [0,1]^d$. The Brownian sheet is the d-dimensional counterpart of the *standard Brownian motion*; see Khoshnevisan (2002).

Kernel estimator

- $\psi : \mathbb{R}^d \to \mathbb{R}$ is a *kernel*[2]
- *Smoothing bandwidths*: $h = (h_1, \ldots, h_d) \in (0, 1/2]^d$
- Define: $\psi_{t,h}(x) := \psi\left(\frac{x_1 - t_1}{h_1}, \ldots, \frac{x_d - t_d}{h_d}\right), \quad x = (x_1, \ldots, x_d) \in [0,1]^d$
- *Kernel estimator*[3]:

$$\hat{f}_h(t) := \frac{1}{\sqrt{n}(\Pi_{i=1}^d h_i)\langle 1, \psi\rangle} \int_{[0,1]^d} \psi_{t,h}(x) dZ(x)$$

Discrete setting: $d = 2$

- $Y_{i,j} = f(\frac{i}{m}, \frac{j}{m}) + \epsilon_{i,j}, \quad i, j = 1, \ldots, m; \quad m^2 = n$
- If $\psi = \mathbb{I}_{[-1,1]^2}, \quad \hat{f}_h(\frac{i_1+j_1}{2m}, \frac{i_2+j_2}{2m}) := \frac{\sum_{k=i_1}^{j_1} \sum_{l=i_2}^{j_2} Y_{k,l}}{(j_1 - i_1 + 1)(j_2 - i_2 + 1)}, \quad h_k = \frac{j_k - i_k}{2m}, k = 1, 2$

---

[2] $\psi : \mathbb{R} \to \mathbb{R}$ is a meas. fnc. s.t. (i) ψ is 0 outside $[-1,1]^d$; (ii) $\psi \in L_2(\mathbb{R}^d)$, i.e., $\int_{\mathbb{R}^d} \psi^2(x)dx < \infty$; (iii) ψ is of finite total HK-variation; and (iv) $\int_{\mathbb{R}^d} \psi(x)dx > 0$.

[3] For $g_1, g_2 \in L_2(\mathbb{R}^d)$, $\langle g_1, g_2\rangle := \int_{\mathbb{R}^d} g_1(x)g_2(x)dx$; $\|g\|^2 := \int_{\mathbb{R}^d} g^2(x)dx$

- *Normalized* test statistic (for testing $H_0 : f(t) = 0$):

$$\hat{\Psi}(t,h) := \frac{\hat{f}_h(t)}{\sqrt{\mathrm{Var}(\hat{f}_h(t))}} = \frac{1}{(\Pi_{i=1}^d h_i)^{1/2}\|\psi\|} \int_{[0,1]^d} \psi_{t,h}(x) dZ(x)$$

- For any $h := (h_1, \ldots, h_d) \in (0, 1/2]^d$, we define

$$A_h := \{t \in \mathbb{R}^d : h_i \leq t_i \leq 1 - h_i, \quad \text{for } i = 1, \ldots, d\}$$

- So, a naive approach to testing $H_0 : f \equiv 0$ could be to consider

$$\sup_{t \in A_h} |\hat{\Psi}(t,h)|, \qquad \text{or} \qquad \sup_{h \in (0,1/2]^d} \sup_{t \in A_h} |\hat{\Psi}(t,h)|$$

- The latter statistic *bypasses the choice* of the bandwidth h, and is a *multiscale statistic* (combines kernel estimators of various scales)

- However (under $H_0 : f \equiv 0$),

$$\sup_{h>0} \sup_{t \in A_h} |\hat{\Psi}(t,h)| = +\infty \quad \text{a.s.}$$

Normalized test statistic: $\hat{\Psi}(t,h) = \frac{1}{(\Pi_{i=1}^d h_i)^{1/2}\|\psi\|} \int_{[0,1]^d} \psi_{t,h}(x) dZ(x)$

Multiscale statistic [Datta and S. (2018a)]

- *Multiscale statistic*[a]: When $f \equiv 0$

$$T(Z,\psi) := \sup_{h\in(0,1/2]^d} \sup_{t\in A_h} \frac{|\hat{\Psi}(t,h)| - \Gamma(2^d \Pi_{i=1}^d h_i)}{D(2^d \Pi_{i=1}^d h_i)} < \infty \quad \text{a.s.}$$

where $\Gamma, D : (0,1] \to (0,\infty)$ are two functions defined as

$$\Gamma(r) := \sqrt{2\log(1/r)} \quad \text{and} \quad D(r) := (\log(e/r))^{-1/2} \log\log(e^e/r)$$

- $\Gamma(\cdot)$ — scale calibrating term that puts *different scales on equal footing*
- Prevents smaller scales from dominating the overall test statistic
- Resolves the *multiple testing* issue; improves *power*

[a] see Datta and S. (2018a); https://arxiv.org/abs/1806.02194

Multiscale statistic: $T(Z,\psi) = \sup_{h\in(0,1/2]^d} \sup_{t\in A_h} \frac{|\hat{\Psi}(t,h)| - \Gamma(2^d \Pi_{i=1}^d h_i)}{D(2^d \Pi_{i=1}^d h_i)}$

Computational Issues ($d = 2$)

- Discrete setting: $Y_{i,j} = f(\frac{i}{m}, \frac{j}{m}) + \epsilon_{i,j}, \quad i,j = 1,\ldots,m; \quad m^2 = n$
- $T_m(Z, \mathbb{I}_{[-1,1]^2}) := \max\limits_{\substack{1\le i_1<j_1\le m \\ 1\le i_2<j_2\le m}} \dfrac{\frac{\sum_{k=i_1}^{j_1}\sum_{l=i_2}^{j_2} Y_{k,l}}{\sqrt{(j_1-i_1+1)(j_2-i_2+1)}} - \Gamma\left(\frac{(j_1-i_1+1)(j_2-i_2+1)}{m^2}\right)}{D\left(\frac{(j_1-i_1+1)(j_2-i_2+1)}{m^2}\right)}$
- Computing $T_m(Z, \mathbb{I}_{[-1,1]^2})$ can be *expensive*; $O(n^2) \equiv O(m^4)$
- For a *suitable* collection of *rectangles* $\mathcal{R}$ we can compute

$$T_{\mathcal{R}}(Z, \mathbb{I}_{[-1,1]^2}) = \max_{R\in\mathcal{R}} \frac{\frac{Ave(Y_{k,l}:k,l\in R)}{\sqrt{m^2|R|}} - \Gamma(|R|)}{D(|R|)}; \qquad |R| = \text{ area of } R$$

- Walther (2010): $O(n\log^2 n)$ rectangles — is almost *linear* in n; for $d = 1$ $\exists$ many clever algorithms, Frick et al. (2015) ...
- We will *not* address the computational issues in this talk

Theorem [Datta and S. (2018a)]

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a kernel. Then, when $f \equiv 0$,

$$T(Z, \psi) := \sup_{h \in (0,1/2]^d} \sup_{t \in A_h} \frac{|\hat{\Psi}(t,h)| - \Gamma(2^d \Pi_{i=1}^d h_i)}{D(2^d \Pi_{i=1}^d h_i)} < \infty \quad \text{a.s.}$$

Some comments

- Seems to us that the above result was not known before
- Let
$$\Gamma_V(r) := (2V \log(1/r))^{1/2}.$$
Our multiscale statistic the *penalization term* corresponds to $V = 1$
- König et al. (2018) use $V = 2d - 1 + \epsilon$, for any $\epsilon > 0$
- Yields *optimal* tests when $d > 1$
- *Cannot* have $V < 1$! Our choice of the *penalization term* is optimal

Optimality against Hölder class of functions $\mathbb{H}_{\beta,L}$

$\beta \in (0,1]$

$$\mathbb{H}_{\beta,L} := \left\{ g : [0,1]^d \to \mathbb{R} : |f(a) - f(b)| \leq L\|a-b\|^{\beta} \;\; \forall\, a,b \in [0,1]^d \right\}$$

$\beta > 1$

- For $k = (k_1, k_2, \ldots, k_d) \in \mathbb{N}^d$ set $\|k\|_1 := \sum_{i=1}^{d} k_i$.
- $\mathbb{H}_{\beta,L}$ is the set of all functions $f : [0,1]^d \to \mathbb{R}$ having *all partial derivatives of order* $\lfloor \beta \rfloor$ on $[0,1]^d$ such that

$$\sum_{0 \leq \|k\|_1 \leq \lfloor \beta \rfloor} \sup_{x \in [0,1]^d} \left| \frac{\partial^{\|k\|_1} f(x)}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} \right| \leq L$$

and

$$\sum_{\|k\|_1 = \lfloor \beta \rfloor} \left| \frac{\partial^{\|k\|_1} f(a)}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} - \frac{\partial^{\|k\|_1} f(b)}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}} \right| \leq L\|a-b\|^{\beta - \lfloor \beta \rfloor} \quad \forall\, a,b \in [0,1]^d.$$

Optimality against Hölderian alternatives

Test: $H_0 : f = 0$ versus $H_1 : f \neq 0 \in \mathbb{H}_{\beta,L}$

Theorem [Datta and S. (2018a)]

- Consider $T(Z, \psi)$ with $\psi \equiv \psi_\beta$ ($\psi_\beta(x) = \mathbf{1}\{\|x\| \leq 1\}(1 - \|x\|^\beta)$, $\beta \leq 1$)
- Let $\rho_n = \left(\frac{\log n}{n}\right)^{\beta/(2\beta+d)}$ and $c_* := \left(\frac{2dL^{d/\beta}}{(2\beta+d)\|\psi_\beta\|^2}\right)^{\beta/(2\beta+d)}$.
- (*Impossibility*) For arbitrary tests ϕ_n with $\mathbb{E}_0[\phi_n(Z)] \leq \alpha$,[a]

$$\limsup_{n\to\infty} \inf_{g\in\mathbb{H}_{\beta,L}:\|g\|_\infty \geq (1-\epsilon_n)c_*\rho_n} \mathbb{E}_g[\phi_n(Z)] \leq \alpha$$

- (*Optimality*[b]) Let $J_n := [(c_*\rho_n/L)^{1/\beta}, 1 - (c_*\rho_n/L)^{1/\beta}]^d$. Then,

$$\lim_{n\to\infty} \inf_{g\in\mathbb{H}_{\beta,L}:\|g\|_{J_n,\infty} \geq (1+\epsilon_n)c_*\rho_n} \mathbb{P}_g[T(Z, \psi_\beta) \geq \kappa_\alpha] = 1$$

where κ_α is the $(1-\alpha)$'th quantile of $T(W, \psi_\beta)$

[a] Here $\epsilon_n \to 0$ and $\epsilon_n\sqrt{\log n} \to \infty$ as $n \to \infty$

[b] Here β needs to be known but the method is adaptive to L

Proposition (When $\beta \in (0,1]$ is unknown) [Datta and S. (2018a)]

Let $\psi_1(x) = \mathbf{1}\{\|x\| \le 1\}(1 - \|x\|)$. Let $T_1 \equiv T(Y, \psi_1)$ be the multiscale statistic with kernel ψ_1. Define

$$\rho_n := \left(\frac{\log n}{n}\right)^{\frac{\beta}{2\beta+d}} \quad \text{and} \quad M > \left(\frac{2dL^{d/\beta}\|\psi_1\|^2}{(2\beta+d)\langle\psi_1, \psi_\beta\rangle^2}\right)^{\frac{\beta}{2\beta+d}}.$$

Let $J_n := [(M\rho_n/L)^{1/\beta}, 1 - (M\rho_n/L)^{1/\beta}]^d$. Then we have

$$\lim_{n\to\infty} \inf_{g \in \mathbb{H}_{\beta,L} : \|g\|_{J_n,\infty} \ge M\rho_n} \mathbb{P}_g(T > \kappa_\alpha) = 1$$

where κ_α is the $(1-\alpha)$ quantile of $T(W, \psi_1)$.

- Thus our test is *rate optimal* over arbitrary Hölder classes
- Adaptivity with respect to *both* β and L, is still an open problem
- Can extend the argument to *anisotropic* Hölder classes with different smoothness along each coordinate axis

Optimality against constant signals

Test: $H_0 : f = 0$ versus $H_1 : f_n = \mu_n \mathbb{I}_{B_n}$ (μ_n, B_n *unknown*)
where B_n is an *hyperrectangle* with sides *parallel* to the coordinate axes

Theorem [Datta and S. (2018)] (Signal on large scales)

- Let $T \equiv T(Z, \psi_0)$ where $\psi_0 = \mathbb{I}_{[-1,1]^d}$. Consider the case

$$\liminf_{n\to\infty} |B_n| > 0^a.$$

- (Impossibility) Let ϕ_n be any test of level $\alpha \in (0,1)$. Then, for any $f_n = \mu_n \mathbb{I}_{B_n}$ s.t. $\limsup_{n\to\infty} |\mu_n|\sqrt{n|B_n|} < \infty$,

$$\limsup_{n\to\infty} \mathbb{E}_{f_n}[\phi_n(Z)] < 1$$

- (Optimality) For the proposed multiscale test based on T, we have

$$\lim_{n\to\infty} \inf_{f_n : \lim \sqrt{n|B_n|}|\mu_n| = \infty} \mathbb{P}_{f_n}(T > \kappa_\alpha) = 1.$$

[a] $|B_n|$ denotes the Lebesgue measure of the set B_n

Test: $H_0 : f = 0$ versus $H_1 : f_n = \mu_n \mathbb{I}_{B_n}$ (μ_n, B_n unknown)

Theorem [Datta and S. (2018)] (Signal on small scales)

- Let $T \equiv T(Z, \psi_0)$ where $\psi_0 = \mathbb{I}_{[-1,1]^d}$. Consider the case

$$\lim_{n\to\infty} |B_n| = 0.$$

- (Impossibility) Let

$$\mathcal{G}_n^- := \left\{ f_n = \mu_n \mathbb{I}_{B_n} : |\mu_n|\sqrt{n|B_n|} = (1-\epsilon_n)\sqrt{2\log(1/|B_n|)} \right\}.$$

Let ϕ_n be any test of level $\alpha \in (0,1)$, then[a]

$$\limsup_{n\to\infty} \inf_{f_n \in \mathcal{G}_n^-} \mathbb{E}_{f_n}[\phi_n(Z)] \le \alpha$$

- (Optimality) Let

$$\mathcal{G}_n^+ := \left\{ f_n = \mu_n \mathbb{I}_{B_n} : |\mu_n|\sqrt{n|B_n|} \ge (1+\epsilon_n)\sqrt{2\log(1/|B_n|)} \right\},$$

then for our multiscale test we have

$$\lim_{n\to\infty} \inf_{f_n \in \mathcal{G}_n^+} \mathbb{P}_{f_n}(T > \kappa_\alpha) = 1$$

[a] Here, we assume that $\epsilon_n \to 0$ and $\epsilon_n\sqrt{2\log(1/|B_n|)} \to \infty$ as $n \to \infty$

Summary

- Proposed a *multidimensional multiscale statistic*
- Showed that the statistic is *finite a.s.* (Under H_0)
- Illustrated its *optimality* in testing $H_0 : f \equiv 0$ against *Hölder classes of functions*
- Optimal for detecting *constant signals* on a rectangle, both at *large* and *small* scales
- The *scan* statistic, *average* likelihood ratio statistic are *NOT* optimal

Multivariate shape-restricted regression ($d \geq 1$)

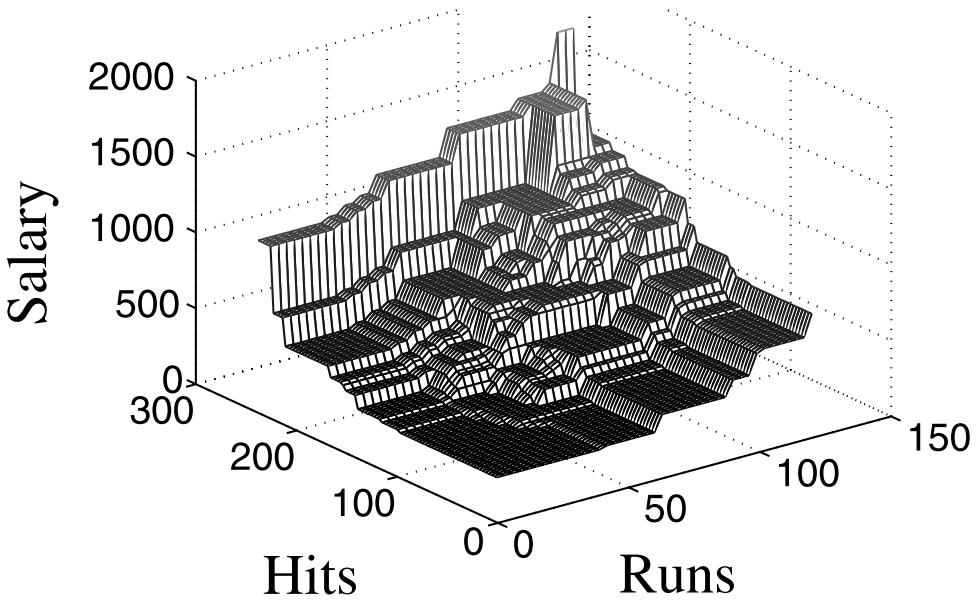
- **Model**: $Y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n \quad (x_1, \ldots, x_n \in [0,1]^d)$
- **Unknown**: $f : [0,1]^d \to \mathbb{R}$, $f \in \mathcal{F}$ is *shape-constrained*

Multivariate isotonic function [Chatterjee et al. (2018), Han et al. (2017+)]

- $f : [0,1]^d \to \mathbb{R}$ is *isotonic* (coordinate-wise nondecreasing) if

$$f(u_1, \ldots, u_d) \leq f(v_1, \ldots, v_d) \quad \text{when} \quad u_i \leq v_i \text{ for } i = 1, \ldots, d$$

- $\mathcal{F}_1 := \{f : [0,1]^d \to \mathbb{R} : f \text{ is isotonic}\}$

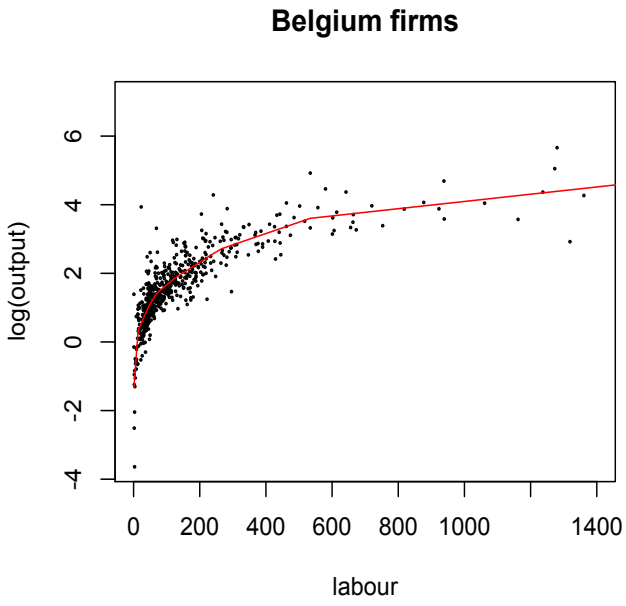


Multivariate convex function [Kuosmanen (2008), Seijo and S. (2011)]

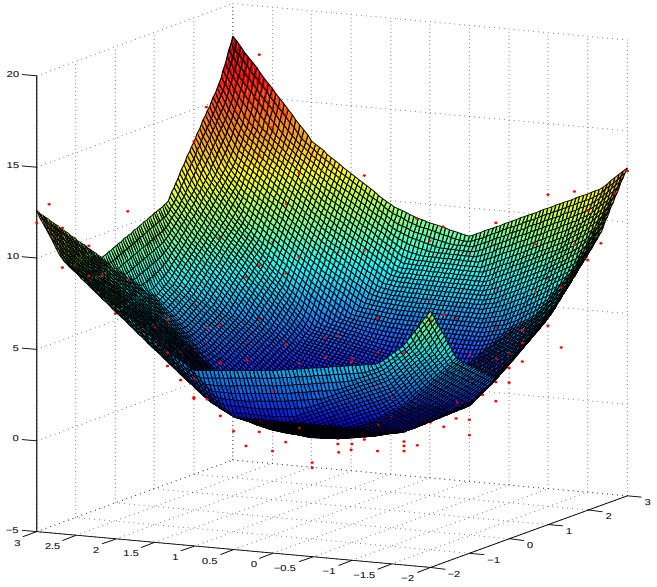
- $f : [0,1]^d \to \mathbb{R}$ is *convex* iff

$$f(\lambda u + (1-\lambda)v) \leq \lambda f(u) + (1-\lambda) f(v), \quad \forall \lambda \in [0,1],\ \forall u, v \in [0,1]^d$$

- $\mathcal{F}_2 := \{f : [0,1]^d \to \mathbb{R} : f \text{ is convex}\}$



Concave regression ($d = 1$)



Convex regression ($d = 2$)

Goals

- Construct *confidence band* for $f \in \mathcal{F}$ ($\mathcal{F}_1$ or $\mathcal{F}_2$) with *guaranteed coverage*, i.e., construct $(\hat{\ell}, \hat{u})$: $[0,1]^d \to \mathbb{R} \times \mathbb{R}$ s.t.

$$\mathbb{P}_f \left(\hat{\ell}(t) \leq f(t) \leq \hat{u}(t) \quad \text{for all } t \in [0,1]^d \right) \geq 1 - \alpha, \quad \forall\, n,\ \forall\, f \in \mathcal{F}$$

- The *width*

$$\sup_{t \in A_n} [\hat{u}(t) - \hat{\ell}(t)], \qquad A_n \subset [0,1]^d$$

  of the constructed confidence band should be *as "small" as possible*

Summary of results

- Use the proposed *multiscale statistic* to construct confidence band[a]
- Our confidence band *adapts* to the *unknown smoothness* and *intrinsic dimension* of the underlying function class
- Exhibits *parametric* ($n^{-1/2}$) rates for certain "*simple*" f
- In some cases confidence band has (almost) *minimum possible width*

[a] Our approach generalizes Dümbgen (2003) *beyond* $d = 1$

Multivariate Gaussian (continuous) white noise model

- Observe stochastic process $\{Z(t) : t \in [0,1]^d\}$ ($d \geq 1$) where

$$Z(t_1, \ldots, t_d) = \sqrt{n} \int_0^{t_1} \ldots \int_0^{t_d} f(x_1, \ldots, x_d) dx_d \ldots dx_1 + W(t_1, \ldots, t_d)$$

- $W(\cdot)$: *Brownian sheet*[4] on $[0,1]^d$ (unobserved)
- $f(\cdot)$: *Unknown* function of interest

[4] A d-dimensional *Brownian sheet* is a *mean-zero Gaussian process* $\{W(t) : t \in [0,1]^d\}$ with covariance kernel

$$\mathrm{Cov}(W(t_1, \ldots, t_d), W(s_1, \ldots, s_d)) = \Pi_{i=1}^d \min(t_i, s_i), \tag{4}$$

for $(t_1, \ldots, t_d), (s_1, \ldots, s_d) \in [0,1]^d$. The Brownian sheet is the d-dimensional counterpart of the *standard Brownian motion*; see Khoshnevisan (2002).

- **Model**: $dZ(t) = \sqrt{n} f(t)dt + dW(t), \qquad t \in [0,1]^d, \quad f$ unknown

- $\psi : \mathbb{R}^d \to \mathbb{R}$ is a *kernel*; $\qquad \psi_{t,h}(x) := \psi\left(\frac{x_1 - t_1}{h_1}, \ldots, \frac{x_d - t_d}{h_d}\right)$

- *Kernel estimator* of f: $\qquad \hat{f}_h(t) = \frac{\int_{[0,1]^d} \psi_{t,h}(x) dZ(x)}{\sqrt{n}(\Pi_{i=1}^d h_i)\langle 1, \psi \rangle}$

- Elementary calculations show that

$$\begin{aligned} \mathbb{E}[\hat{f}_h(t)] &= \frac{\int_{[0,1]^d} \psi_{t,h}(x) f(x) dx}{(\Pi_{i=1}^d h_i)\langle 1, \psi \rangle}, \\ \mathrm{Var}(\hat{f}_h(t)) &= \frac{\|\psi\|^2}{n(\Pi_{i=1}^d h_i)\langle 1, \psi \rangle^2} \end{aligned}$$

- $\frac{\hat{f}_h(t) - \mathbb{E}(\hat{f}_h(t))}{\sqrt{\mathrm{Var}(\hat{f}_h(t))}} = \frac{\frac{\int_{[0,1]^d} \psi_{t,h}(x)(dZ(x) - \sqrt{n} f(x) dx)}{\sqrt{n}(\Pi_{i=1}^d h_i)\langle 1, \psi \rangle}}{\frac{\|\psi\|}{\sqrt{n(\Pi_{i=1}^d h_i)}\langle 1, \psi \rangle}} = \frac{\int_{[0,1]^d} \psi_{t,h}(x) dW(x)}{\sqrt{\Pi_{i=1}^d h_i}\|\psi\|} \sim N(0,1), \ \forall t, h$

- **Question**: How do we control the *random fluctuations* for this kernel estimator *uniformly in t and h?*

- **Recall**: $\frac{\hat{f}_h(t)-\mathbb{E}(\hat{f}_h(t))}{\sqrt{\mathrm{Var}(\hat{f}_h(t))}} = \frac{\int_{[0,1]^d} \psi_{t,h}(x)dW(x)}{(\Pi_{i=1}^d h_i)^{1/2}\|\psi\|}$

- For $h := (h_1, \ldots, h_d) \in (0, 1/2]^d$, $A_h = \{t \in \mathbb{R}^d : h_i \leq t_i \leq 1 - h_i\}$

Theorem [Datta and S. (2018a)] — Multiscale statistic

Let $\psi : \mathbb{R}^d \to \mathbb{R}$ be a kernel. Then,

$$T(\psi) := \sup_{h>0} \sup_{t \in A_h} \left(\frac{\int_{[0,1]^d} \psi_{t,h}(x)dW(x)}{(\Pi_{i=1}^d h_i)^{1/2}\|\psi\|} - \Gamma(2^d \Pi_{i=1}^d h_i) \right) < \infty \quad \text{a.s.}$$

where $\Gamma : (0, 1] \to (0, \infty)$ is defined as

$$\Gamma(r) := \sqrt{2\log(1/r)}$$

- $\Gamma(\cdot)$ — scale calibrating term that puts *different scales on equal footing*

- Prevents *smaller scales* from dominating the overall test statistic

Question: What about the *bias* of these kernel estimators?

What does it mean for f to be shape-restricted?

- Assume $f \in \mathcal{F}$ (e.g., $\mathcal{F}$ could be $\mathcal{F}_1$ or $\mathcal{F}_2$)
- We say that $\mathcal{F}$ is *shape-restricted* if we can find *kernels* $\psi^{(l)}$ and $\psi^{(u)}$ such that the corresponding kernel estimators $\hat{f}_h^l(t)$ and $\hat{f}_h^u(t)$ satisfy

$$\mathbb{E}[\hat{f}_h^l(t)] \leq f(t) \leq \mathbb{E}[\hat{f}_h^u(t)] \quad \forall\, h, \forall\, t \in A_h, \text{and } \forall\, f \in \mathcal{F}$$

- For example, when $f \in \mathcal{F}_1$ one such trivial choice could be $\psi^{(l)} = \mathbb{I}_{[-1,0]^d}$ and $\psi^{(u)} = \mathbb{I}_{[0,1]^d}$
- Due to the known shape-constraint on $\mathcal{F}$ we can *bound* the *bias* of the kernel estimators

Question: How to construct the confidence band?

- $\mathcal{F}$ is *shape-restricted*, i.e., $\exists$ kernels $\psi^{(l)}$ and $\psi^{(u)}$ such that

$$\mathbb{E}(\hat{f}_h^l(t)) \leq f(t) \leq \mathbb{E}(\hat{f}_h^u(t)) \quad \text{for all } h > 0,\ t \in A_h \text{ and } f \in \mathcal{F}$$

- $T(\psi) = \sup_{h>0} \sup_{t \in A_h} \left(\frac{\hat{f}_h(t) - \mathbb{E}(\hat{f}_h(t))}{\sqrt{\mathrm{Var}(\hat{f}_h(t))}} - \Gamma(2^d \Pi_{i=1}^d h_i) \right) < \infty$ a.s.

- Therefore, for every $h \in (0, 1/2]^d$,

$$\begin{aligned} f(t) &= \{f(t) - \mathbb{E}(\hat{f}_h^l(t))\} + \{\mathbb{E}(\hat{f}_h^l(t)) - \hat{f}_h^l(t)\} + \hat{f}_h^l(t) \\ &\geq \hat{f}_h^l(t) - \frac{\|\psi^{(l)}\| \left(T(\psi^{(l)}) + \Gamma(2^d \Pi_{i=1}^d h_i) \right)}{\langle 1, \psi^{(l)} \rangle (n \Pi_{i=1}^d h_i)^{1/2}} \end{aligned}$$

- Similarly, we have, for every $h \in (0, 1/2]^d$,

$$f(t) \leq \hat{f}_h^u(t) + \frac{\|\psi^{(u)}\| \left(T(-\psi^{(u)}) + \Gamma(2^d \Pi_{i=1}^d h_i) \right)}{\langle 1, \psi^{(u)} \rangle (n \Pi_{i=1}^d h_i)^{1/2}}$$

- Let κ_α is the $(1 - \alpha)$ quantile of the combined statistic

$$T^\star = \max(T(\psi^{(l)}), T(-\psi^{(u)}))$$

- Define our confidence band as follows: for $t \in [0,1]^d$, let

$$\hat{\ell}(t) := \sup_{h>0:t\in A_h} \left\{ \hat{f}_h^l(t) - \frac{\|\psi^{(l)}\| \left(\kappa_\alpha + \Gamma(2^d \Pi_{i=1}^d h_i)\right)}{\langle 1, \psi^{(l)} \rangle (n \Pi_{i=1}^d h_i)^{1/2}} \right\}$$

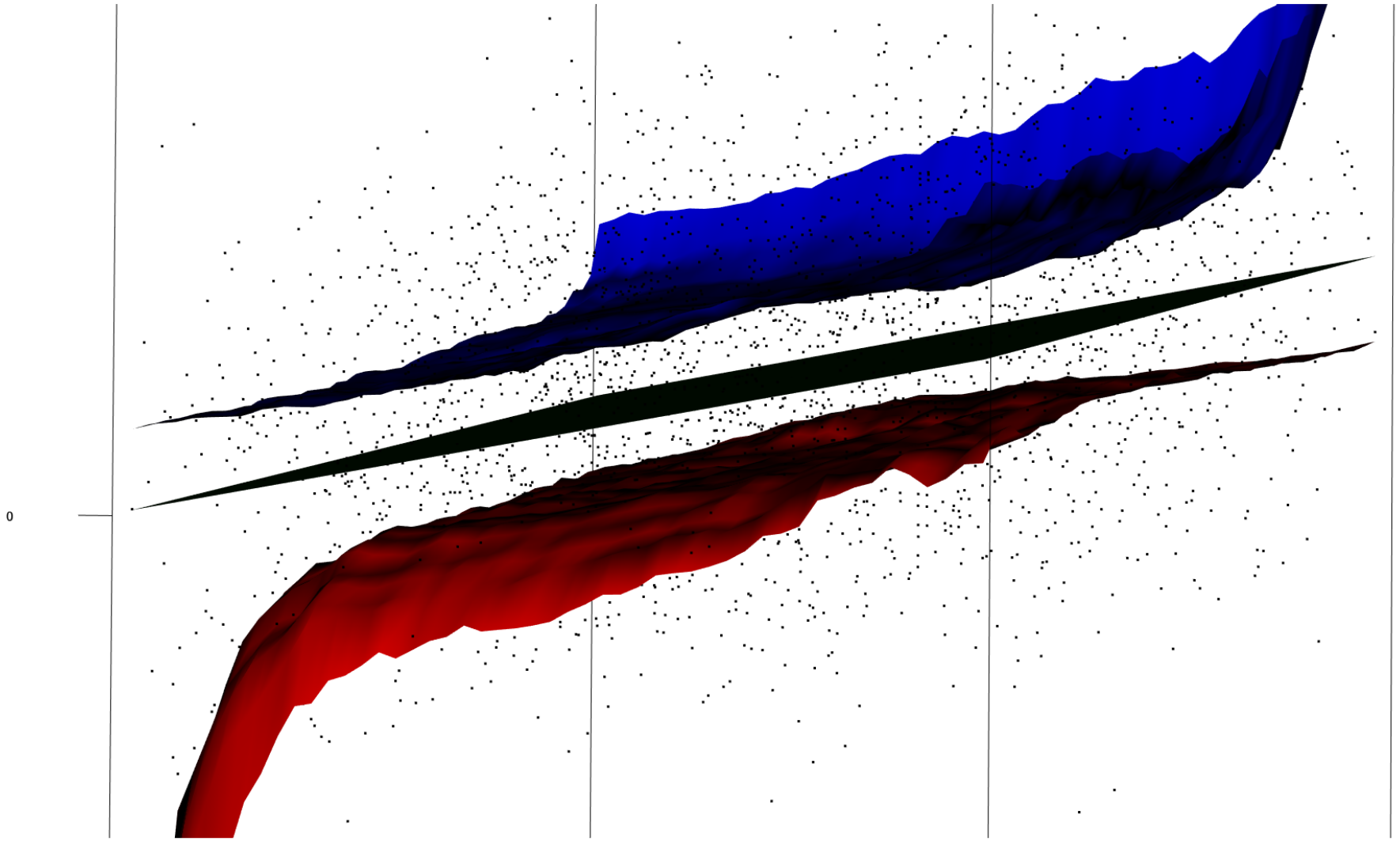
and

$$\hat{u}(t) := \inf_{h>0:t\in A_h} \left\{ \hat{f}_h^u(t) + \frac{\|\psi^{(u)}\| \left(\kappa_\alpha + \Gamma(2^d \Pi_{i=1}^d h_i)\right)}{\langle 1, \psi^{(u)} \rangle (n \Pi_{i=1}^d h_i)^{1/2}} \right\}$$

Theorem

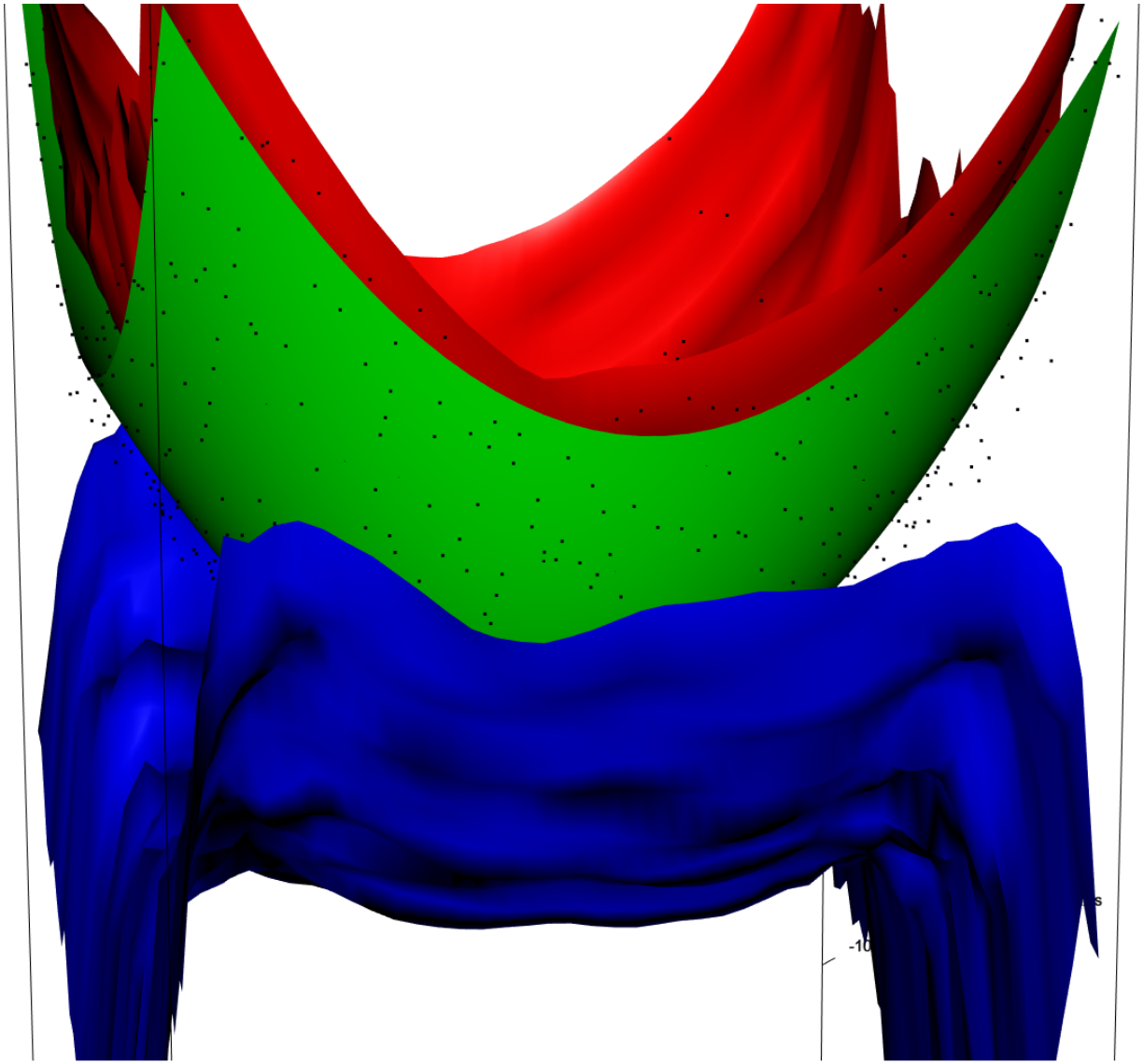
Let $\mathcal{F}$ be shape-restricted. Let $(\hat{\ell}, \hat{u}) : [0,1]^d \to \mathbb{R} \times \mathbb{R}$ be defined as above. Then

$$\mathbb{P}_f \left(\hat{\ell}(t) \leq f(t) \leq \hat{u}(t) \quad \text{for all } t \in [0,1]^d \right) \geq 1 - \alpha, \quad \forall\, n,\ \forall\, f \in \mathcal{F}$$



Confidence band for the function $f(x_1, x_2) = x_1 + x_2$ assuming $f \in \mathcal{F}_1$ and sample size $n = 50^2$

Confidence band for the function $f(x_1, x_2) = \mathbb{I}(x_1 \geq 0.5)$ assuming $f \in \mathcal{F}_1$ and sample size $n = 50^2$



Confidence band for the function $f(x_1, x_2) = 40((x_1 - 0.5)^2 + (x_2 - 0.5)^2)$ assuming $f \in \mathcal{F}_2$ and sample size $n = 40^2$

$f(x_1, x_2)$	Category	Coverage probability
0	isotonic	0.95
$x_1 + x_2$	isotonic	0.97
$20(x_2 + x_2)$	isotonic	1.00
$\mathbb{I}(x \geq 0.5)$	isotonic	0.97
0	convex	0.95
$x_1 + x_2$	convex	0.96
$10(x_1 + x_2)$	convex	0.95
$(x_1 - 0.5)^2 + (x_2 - 0.5)^2$	convex	0.98

Table: Estimated coverage probabilities of nominal 95% confidence bands for different functions

Goal: Construct *adaptive* (and *optimal*) confidence band for appropriate subclasses of $\mathcal{F}_1$ and $\mathcal{F}_2$

Rate of convergence of a confidence band

- We say that the confidence band $\{(\ell(t), u(t)) : t \in [0,1]^d\}$, with coverage probability $1-\alpha$, has *rate of convergence* γ_n on the set A_n for a class of functions $\mathcal{G}$ if

$$\inf_{f \in \mathcal{G}} \mathbb{P}_f \left(\sup_{t \in A_n} (u(t) - \ell(t)) \leq \Delta \gamma_n \right) \geq 1 - \alpha, \qquad \text{for all } n$$

- Δ is a constant *not depending on n* (but may depend on α and $\mathcal{G}$)

Note: Without any *qualitative* assumptions on f constructing confidence bands (with *guaranteed coverage* for all $f \in \mathcal{G}$) is an *ill-posed* problem, even when $d = 1$; see e.g., Donoho (1988) ...

Result: Our confidence band *optimally adapts* to *unknown smoothness*

Theorem (Datta and S. (2018b))

Suppose that $f \in \mathcal{G}$ *where* $\mathcal{G}$ *is either* $\mathcal{F}_1 \cap \mathbb{H}_{\beta,L}$ *(*$0 < \beta \leq 1$ *and* $L > 0$ *unknowns) or* $\mathcal{F}_2 \cap \mathbb{H}_{\beta,L}$ *(*$1 < \beta \leq 2$ *and* $L > 0$ *unknowns). Then,*

$$\inf_{f \in \mathcal{G}} \mathbb{P}_f \left(\sup_{t \in A_{(\epsilon_n, \ldots, \epsilon_n)}} (\hat{u}(t) - \hat{\ell}(t)) \leq \Delta \left[\frac{\log(en)}{n} \right]^{\frac{\beta}{2\beta+d}} \right) \geq 1 - \alpha \quad \forall\, n$$

for some constant Δ *depending only on* $L, \beta, \psi^{(l)}, \psi^{(u)}, \alpha$ *and* $\epsilon_n = (\log(en)/n)^{1/(2\beta+d)}$ *and* $A_{(\epsilon_n, \ldots, \epsilon_n)} := \{t \in \mathbb{R}^d : \epsilon_n \leq t_i \leq 1 - \epsilon_n\}$

- This shows that the *rate of convergence* of our confidence band is $(\log n/n)^{\beta/(2\beta+d)}$ for the class $\mathcal{F}_i \cap \mathbb{H}_{\beta,L}$ $(i = 1, 2)$ on $A_{(\epsilon_n, \ldots, \epsilon_n)}$
- This rate is indeed *optimal*!
- Construction of $(\hat{\ell}, \hat{u})$ *does NOT* need knowledge of β and L

Result: Our confidence band *automatically adapts* to the unknown *intrinsic dimension* (# of relevant variables) and *smoothness* of f

- Define the following function classes ($1 \leq k \leq d$ is an integer):

$$\mathcal{G}_k \quad := \quad \Big\{ f \in L_1([0,1]^d) : f(x) = g(x_j : j \in \mathcal{P}),\ \forall\, x = (x_1, \ldots, x_d),$$
$$\mathcal{P} \subseteq \{1, \ldots, d\} \text{ with } |\mathcal{P}| = k \leq d \Big\}$$

Function class	Rate of convergence	Corresponding set[5]
$\mathcal{G}_k \cap \mathcal{F}_i \cap \mathbb{H}_{\beta,L}$	$(\log(en)/n)^{\beta/(2\beta+k)}$	A_ε : any $\varepsilon > 0$

[5] $A_{(\epsilon,\ldots,\epsilon)} := \{(t_1, \ldots, t_d) \in \mathbb{R}^d : \epsilon \leq t_i \leq 1 - \epsilon\}$

Result: *Parametric* rate of convergence when f is *locally constant/affine*

- Suppose that $f \in \mathcal{F}_1$ and that the set

$$I_4(f,\varepsilon) := \left\{ t \in [0,1]^d : B_\infty(t,\varepsilon) \subset [0,1]^d, f \text{ is } \textit{constant} \text{ on } B_\infty(t,\varepsilon) \right\} \neq \emptyset$$

where $\varepsilon = (\epsilon_1, \ldots, \epsilon_d) > 0$ and $B_\infty(t,\varepsilon) = \Pi_{i=1}^d (t_i - \epsilon_i, t_i + \epsilon_i)$

- $I_4(f,\varepsilon)$ denotes the subset of $[0,1]^d$ such that f is *constant* around an *ε-neighborhood* of each of its elements

- Our confidence band has *parametric* rate of convergence on $I_4(f,\varepsilon)$, i.e.,

$$\mathbb{P}_f \left(\sup_{t \in I_4(f,\varepsilon)} (\hat{u}(t) - \hat{\ell}(t)) \leq \Delta_\varepsilon n^{-1/2} \right) \geq 1 - \alpha, \quad \text{for all } n$$

Similar result holds when $f \in \mathcal{F}_2$ and $I_4(f,\varepsilon)$ is replaced by

$$I_5(f,\varepsilon) := \left\{ t \in [0,1]^d : B_\infty(t,\varepsilon) \subset [0,1]^d, f \text{ is } \textit{affine} \text{ on } B_\infty(t,\varepsilon) \right\} \neq \emptyset$$

Result: Our confidence band has (almost) *minimal width (locally)*

- Suppose that $f \in \mathcal{F}_1$ *continuously differentiable* in a *neighborhood* U of $x_0 \in (0,1)^d$ s.t.

$$\tilde{L}[x_0] := \left[\Pi_{i=1}^d \frac{\partial}{\partial x_i} f(x_1, \ldots, x_d) \Big|_{x=x_0} \right]^{1/d} > 0.$$

- For *any* confidence band $\{(\ell(t), u(t)) : t \in [0,1]^d\}$ for f with *guaranteed coverage* $(1-\alpha)$ for the class $\mathcal{F}_1$, and for any $\epsilon \in (0,1)$,

$$\liminf_{n\to\infty} \mathbb{P}_f \left(\sup_{x \in U} [f(x) - \ell(x)] \geq (1-\epsilon) \Delta^{(l)} \tilde{L}^{\frac{d}{2+d}}[x_0] \left(\frac{\log(en)}{n} \right)^{\frac{1}{2+d}} \right) \geq 1 - \alpha$$

- For our constructed confidence band $(\hat{\ell}, \hat{u})$, for any $\epsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}_f \left((f - \hat{\ell})(x_0) \leq (1+\epsilon) \Delta^{(l)} \tilde{L}^{\frac{d}{2+d}}[x_0] \left(\frac{\log(en)}{n} \right)^{\frac{1}{2+d}} \right) = 1$$

- *Similar* results also exist when considering the difference $\hat{u} - f$

Summary

- Proposed a *multiscale statistic* in multidimension
- Constructed *confidence bands* for *shape-restricted regression functions* that have *guaranteed coverage*
- Our confidence band *adapts* to the *unknown smoothness* and *intrinsic dimension* of the (shape-restricted) function class
- In some cases our band has (almost) the *minimum possible width*

Thank you very much!

Questions?

References

- Datta, P., and **Sen, B.** (2018a). Optimal Inference with a Multidimensional Multiscale Statistic. (see https://arxiv.org/abs/1806.02194)
- Datta, P., and **Sen, B.** (2018b). Adaptive Confidence Bands in Multivariate Nonparametric Shape-restricted Regression. (working paper)

Theorem [Datta and S. (2018)]

Let X be a *stochastic process* on a *pseudometric space* $(\mathscr{F}, \rho)$ with *continuous* sample paths. Suppose that the following 3 conditions hold:

(a) There is a function $\sigma : \mathscr{F} \to (0, 1]$ and a constant $K \geq 1$ such that

$$\mathbb{P}\big(X(a) > \sigma(a)\eta\big) \leq K \exp(-\eta^2/2) \qquad \text{for all } \eta > 0, \text{ for all } a \in \mathscr{F}.$$

Moreover, $\sigma^2(b) \leq \sigma^2(a) + \rho^2(a, b)$ for all $a, b \in \mathscr{F}$.

(b) For some constants $L, M \geq 1$,

$$\mathbb{P}\big(|X(a) - X(b)| > \rho(a, b)\eta\big) \leq L \exp(-\eta^2/M) \quad \forall\, \eta > 0,\ \forall\, a, b \in \mathscr{F}.$$

(c) For some constants $A, B, V, p > 0$,

$$N((\delta u)^{1/2}, \{a \in \mathscr{F} : \sigma^2(a) \leq \delta\}) \leq A \frac{(\log(e/\delta))^p}{u^B \delta^V} \quad \forall\, u, \delta \in (0, 1].$$

Then the random variable

$$S(X) := \sup_{a \in \mathscr{F}} \frac{X^2(a)/\sigma^2(a) - 2V \log(1/\sigma^2(a))}{\log\log(e^e/\sigma^2(a))} < \infty \ \text{ a.s.}$$

To show: $T(W,\psi) := \sup_{h\in(0,1/2]^d} \sup_{t\in A_h} \frac{|\hat{\Psi}(t,h)| - \Gamma(2^d \Pi_{i=1}^d h_i)}{D(2^d \Pi_{i=1}^d h_i)} < \infty$ a.s.

Last result shows $S(X) := \sup_{a\in\mathscr{F}} \frac{X^2(a)/\sigma^2(a) - 2V\log(1/\sigma^2(a))}{\log\log(e^e/\sigma^2(a))} < \infty$ a.s.

Application of the previous theorem to our problem

- $\mathscr{F} := \{(t,h) \in \mathbb{R}^d \times (0,1/2]^d : h_i \le t_i \le 1 - h_i,\ \forall\, i = 1,2,\ldots,d\}$
- *Stochastic process*: $X(t,h) := \sqrt{2^d \Pi_{i=1}^d h_i}\ \hat{\Psi}(t,h) \quad \forall\,(t,h) \in \mathscr{F}$
- *Pseudometric*:

  $\rho^2((t,h),(t',h')) := |B_\infty(t,h) \triangle B_\infty(t',h')|$,[a] $\quad \forall\,(t,h),(t',h') \in \mathscr{F}$,

  where $B_\infty(t,h) := \Pi_{i=1}^d (t_i - h_i, t_i + h_i)$
- $A \triangle B := (A \cap B^c) \cup (A^c \cap B)$ is the *symmetric difference* of A and B
- $\sigma^2(t,h) := |B_\infty(t,h)| = 2^d \Pi_{i=1}^d h_i, \qquad$ for $(t,h) \in \mathscr{F}$

[a] $|B|$ denotes the Lebesgue measure of the set B

The following result shows that for the above defined pseudometric space $(\mathscr{F}, \rho)$ the *packing number* condition (c) holds

Lemma (Datta and S. (2018))

Let $\mathscr{F}, \rho(\cdot,\cdot)$ and $\sigma(\cdot)$ be as described above. Then, for all $u, \delta \in (0,1]$,

$$N\left((u\delta)^{1/2}, \{(t,h) \in \mathscr{F} : \sigma^2(t,h) \leq \delta\}\right) \leq K u^{-2d} \delta^{-1} (\log(e/\delta))^{d-1}$$

for some constant K depending only on d.

Multiscale statistic: A generalization of Lévy's modulus of continuity ($d = 1$)

- Multiscale statistic: For $\Gamma(r) := \sqrt{2\log(1/r)}$,

$$T(Z,\psi) := \sup_{h\in(0,1/2]} \sup_{t\in[h,1-h]} \frac{|\hat{\Psi}(t,h)| - \Gamma(2h)}{D(2h)}$$

- Consider $\psi = \mathbb{I}_{[-1,1]}$ (and $f \equiv 0$), then

$$\hat{\Psi}(t,h) = (2h)^{-1/2}[W(t+h) - W(t-h)]$$

- *Lévy's modulus of continuity*:

$$\lim_{h\to 0} \sup_{t\in[h,1-h]} \frac{|\hat{\Psi}(t,h)|}{\Gamma(2h)} \equiv \lim_{h\to 0} \sup_{t\in[h,1-h]} \frac{|W(t+h) - W(t-h)|}{\sqrt{(2h)2\log(1/(2h))}} = 1$$