Some comments about the Highly-Adaptive LASSO with focus on survival data

Anders Munch
joint work with Thomas G., Helene, and Mark van der Laan

February 15, 2023

Outline

Setting and motivation

Multivariate càdlàg functions of bounded sectional variation norm

Some challenges with the exact definition of the estimator

Approximate minimization is sufficient

Motivation

Reasons for using HAL

- Theoretically important
- “Dimension-free” convergence rate
- Few assumptions imposed

Motivation

Reasons for using HAL

- Theoretically important
- "Dimension-free" convergence rate
- Few assumptions imposed

Better understanding of the estimator

- Multivariate càdlàg functions
- Sectional variation norm

Motivation

Reasons for using HAL

- Theoretically important
- "Dimension-free" convergence rate
- Few assumptions imposed

Better understanding of the estimator

- Multivariate càdlàg functions
- Sectional variation norm

Theoretical discussion

- In practice we will have to approximate the HAL estimator for computational reasons
- Still nice to know that the estimator we approximate has nice properties

Definition of the estimator

$\mathcal{D}_M([0,1]^d)$ the space of càdlàg functions $f\colon [0,1]^d \to \mathbb{R}$ with sectional variation norm bounded by M.

$\mathcal{O}$ the sample space

L a loss function, $L(f, O) \in \mathbb{R}_+$

The parameter of interest is the function minimizing the expected loss (risk)

$$f_0 = \operatorname*{argmin}_{f \in \mathcal{D}_M([0,1]^d)} P[L(f, \cdot)] = \operatorname*{argmin}_{f \in \mathcal{D}_M([0,1]^d)} \int_{\mathcal{O}} L(f, o) P(\mathrm{d}o).$$

We estimate f_0 with the function minimizing the empirical risk

$$\hat{f}_n = \operatorname*{argmin}_{f \in \mathcal{D}_M([0,1]^d)} \hat{\mathbb{P}}_n[L(f, \cdot)] = \operatorname*{argmin}_{f \in \mathcal{D}_M([0,1]^d)} \frac{1}{n} \sum_{i=1}^{n} L(f, O_i).$$

Multivariate càdlàg functions

Which direction is "left" when $d > 1$?

Multivariate càdlàg functions

Which direction is "left" when $d > 1$?

Definition (Neuhaus [1971])

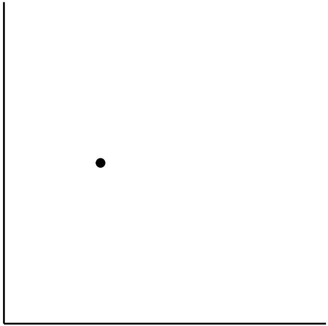
For a point $u \in [0,1]^d$ and a vertex $\mathbf{a} \in \{0,1\}^d$ look at quadrants $Q_{\mathbf{a}}(u)$ spanned by u and $\mathbf{a}$. The limit of $f(u_n)$ for $\{u_n\} \subset Q_{\mathbf{a}}(u)$, $u_n \to u$ should exist, and if $\mathbf{a} = (1, 1, \ldots, 1)$ then $\lim_{n\to\infty} f(u_n) = f(u)$.

Multivariate càdlàg functions

Which direction is "left" when $d > 1$?

Definition (Neuhaus [1971])

For a point $u \in [0,1]^d$ and a vertex $\mathbf{a} \in \{0,1\}^d$ look at quadrants $Q_{\mathbf{a}}(u)$ spanned by u and $\mathbf{a}$. The limit of $f(u_n)$ for $\{u_n\} \subset Q_{\mathbf{a}}(u)$, $u_n \to u$ should exist, and if $\mathbf{a} = (1,1,\dots,1)$ then $\lim_{n\to\infty} f(u_n) = f(u)$.

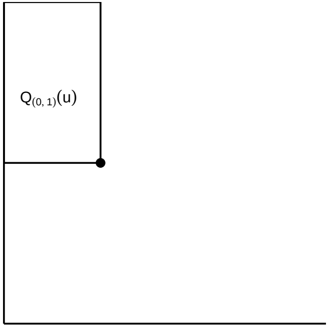


Multivariate càdlàg functions

Which direction is "left" when $d > 1$?

Definition (Neuhaus [1971])

For a point $u \in [0,1]^d$ and a vertex $\mathbf{a} \in \{0,1\}^d$ look at quadrants $Q_{\mathbf{a}}(u)$ spanned by u and $\mathbf{a}$. The limit of $f(u_n)$ for $\{u_n\} \subset Q_{\mathbf{a}}(u)$, $u_n \to u$ should exist, and if $\mathbf{a} = (1, 1, \ldots, 1)$ then $\lim_{n\to\infty} f(u_n) = f(u)$.

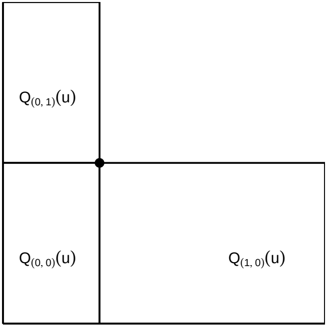


Multivariate càdlàg functions

Which direction is "left" when $d > 1$?

Definition (Neuhaus [1971])

For a point $u \in [0,1]^d$ and a vertex $\mathbf{a} \in \{0,1\}^d$ look at quadrants $Q_{\mathbf{a}}(u)$ spanned by u and $\mathbf{a}$. The limit of $f(u_n)$ for $\{u_n\} \subset Q_{\mathbf{a}}(u)$, $u_n \to u$ should exist, and if $\mathbf{a} = (1,1,\ldots,1)$ then $\lim_{n\to\infty} f(u_n) = f(u)$.

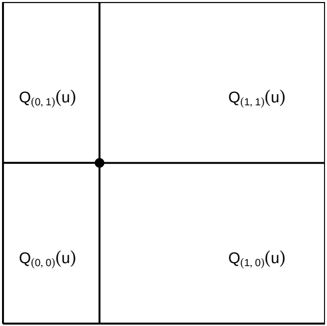


Multivariate càdlàg functions

Which direction is “left” when $d > 1$?

Definition (Neuhaus [1971])

For a point $u \in [0,1]^d$ and a vertex $\mathbf{a} \in \{0,1\}^d$ look at quadrants $Q_{\mathbf{a}}(u)$ spanned by u and $\mathbf{a}$. The limit of $f(u_n)$ for $\{u_n\} \subset Q_{\mathbf{a}}(u)$, $u_n \to u$ should exist, and if $\mathbf{a} = (1,1,\ldots,1)$ then $\lim_{n\to\infty} f(u_n) = f(u)$.



Càdlàg function with jumps

If we picture "continuous from the right with left-hand limits" in dimension $d = 1$ this does not seem very restrictive – for instance, "jumps" are allowed.

Càdlàg function with jumps

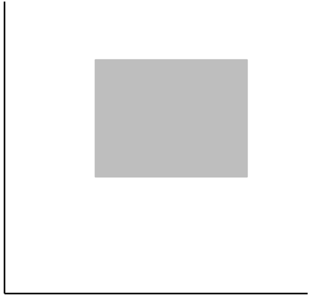
If we picture "continuous from the right with left-hand limits" in dimension $d = 1$ this does not seem very restrictive – for instance, "jumps" are allowed.

Don't trust $d = 1$!

Càdlàg function with jumps

If we picture "continuous from the right with left-hand limits" in dimension $d = 1$ this does not seem very restrictive – for instance, "jumps" are allowed.

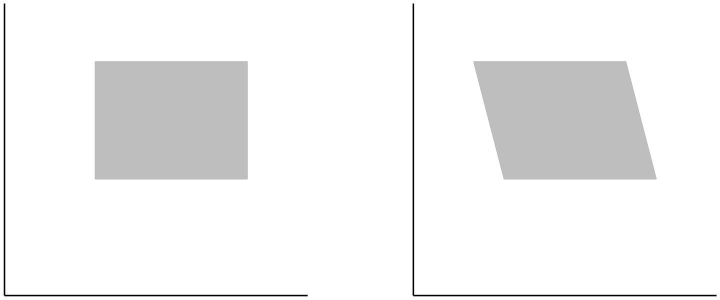
Don't trust $d = 1$!



Càdlàg function with jumps

If we picture "continuous from the right with left-hand limits" in dimension $d = 1$ this does not seem very restrictive – for instance, "jumps" are allowed.

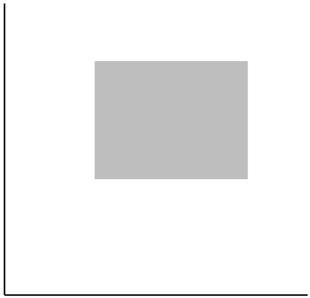
Don't trust $d = 1$!

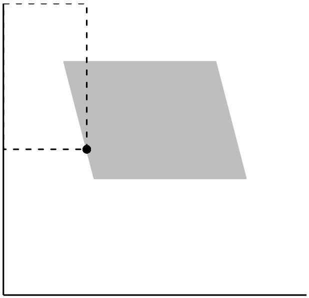


Càdlàg function with jumps

If we picture "continuous from the right with left-hand limits" in dimension $d = 1$ this does not seem very restrictive – for instance, "jumps" are allowed.

Don't trust $d = 1$!





Sectional variation norm

Sectional variation norm

In dimension $d = 1$, the variation norm of a function is

$$\|f\|_v = \sup_{\pi} \sum_{i=1}^{|\pi|} |f(t_i) - f(t_{i-1})|,$$

where the supremum is taken over all finite partitions $0 = t_0 < t_1 < \cdots < t_{|\pi|} = 1$.

Sectional variation norm

In dimension $d = 1$, the variation norm of a function is

$$\|f\|_v = \sup_\pi \sum_{i=1}^{|\pi|} |f(t_i) - f(t_{i-1})|,$$

where the supremum is taken over all finite partitions
$0 = t_0 < t_1 < \cdots < t_{|\pi|} = 1$.

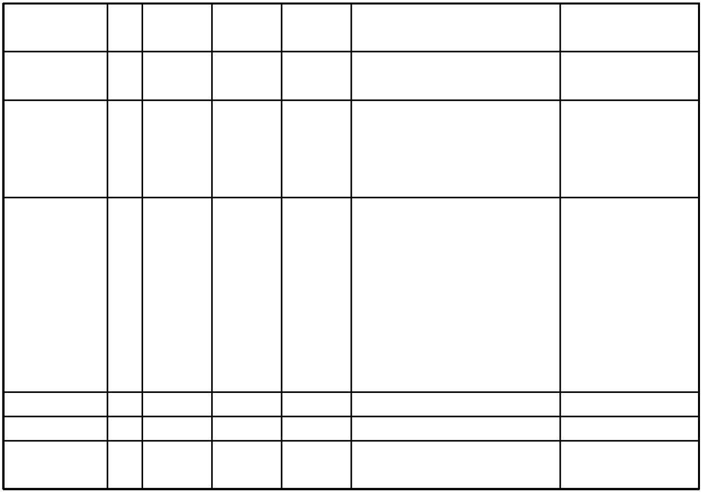
At first sight, a natural generalization seems to be the *Vitali variation*:

$$V^{(d)}(f) = \sup_\pi \sum_{A \in \pi} |\Delta(f; A)|,$$

where the supremum is taken over all "grid partitions" and $\Delta(f; A)$ is the *quasi-volume* that f assigns the rectangle A.

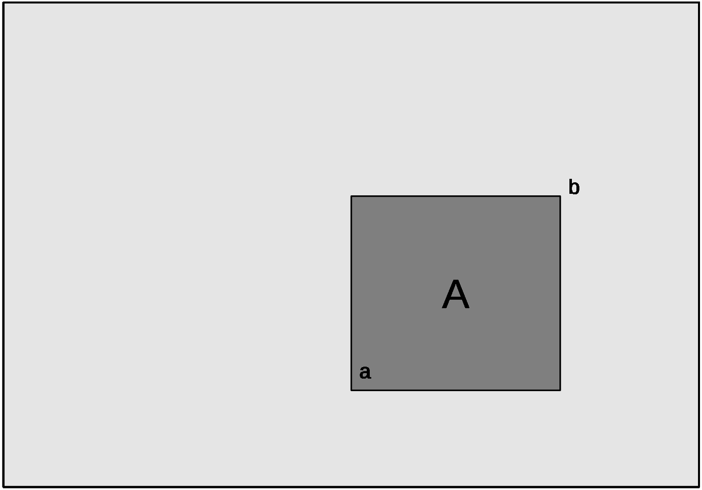
Vitali variation: $V^{(d)}(f) = \sup_{\pi} \sum_{A \in \pi} |\Delta(f; A)|$

Vitali variation: $V^{(d)}(f) = \sup_\pi \sum_{A \in \pi} |\Delta(f; A)|$



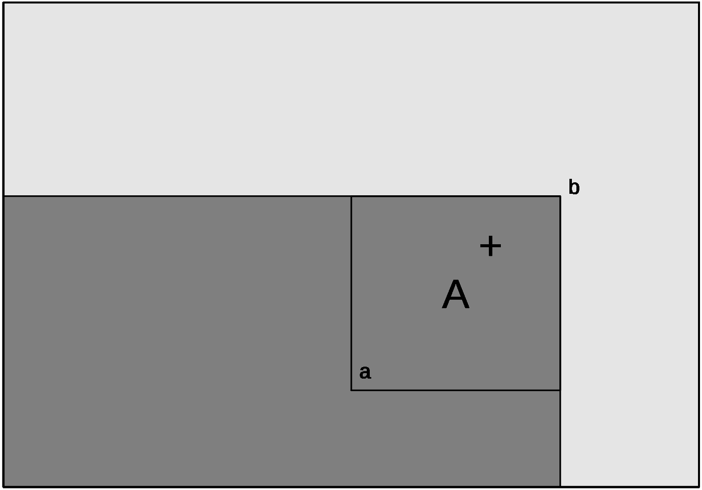
Vitali variation: $V^{(d)}(f) = \sup_\pi \sum_{A\in\pi} |\Delta(f;A)|$

When $d = 2$, $\quad \Delta(f;A) = f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2)$.



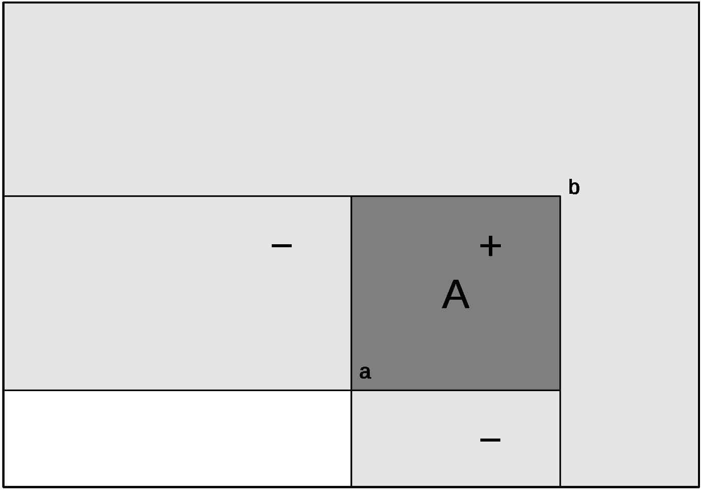
Vitali variation: $V^{(d)}(f) = \sup_\pi \sum_{A \in \pi} |\Delta(f; A)|$

When $d = 2$, $\quad \Delta(f; A) = f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2).$



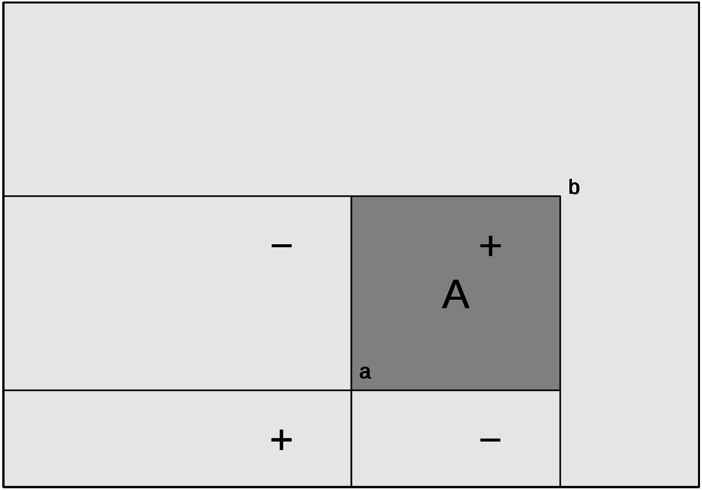
Vitali variation: $V^{(d)}(f) = \sup_{\pi} \sum_{A \in \pi} |\Delta(f; A)|$

When $d = 2$, $\quad \Delta(f; A) = f(b_1, b_2) - f(b_1, a_2) - f(a_1, b_2) + f(a_1, a_2).$



Vitali variation: $V^{(d)}(f) = \sup_\pi \sum_{A\in\pi} |\Delta(f;A)|$

When $d = 2$, $\quad \Delta(f;A) = f(b_1,b_2) - f(b_1,a_2) - f(a_1,b_2) + f(a_1,a_2)$.



Sectional variation norm (Hardy-Krause variation)

However, many interesting functions have Vitalli variation equal to 0, e.g., the function $f(x, y) = x$.

Sectional variation norm (Hardy-Krause variation)

However, many interesting functions have Vitalli variation equal to 0, e.g., the function $f(x, y) = x$.

We should also look at the function on the *lower dimensional faces*

$$U_s = \{(u_1, \ldots, u_d) \in [0, 1]^d : u_j = 0, j \notin s\},$$

for non-empty subsets $s \subset \{1, \ldots, d\}$.

Sectional variation norm (Hardy-Krause variation)

However, many interesting functions have Vitalli variation equal to 0, e.g., the function $f(x, y) = x$.

We should also look at the function on the *lower dimensional faces*

$$U_s = \{(u_1, \dots, u_d) \in [0,1]^d : u_j = 0, j \notin s\},$$

for non-empty subsets $s \subset \{1, \dots, d\}$.

We denote by $f_s \colon [0,1]^{|s|} \to \mathbb{R}$ the restriction of f to U_s and define the norm

$$\|f\|_v = \sum_s V^{(|s|)}(f_s),$$

where the sum is taken over all non-empty subsets $s \subset \{1, \dots, d\}$.

This is referred to as the *Hardy-Krause variation* by Fang et al. [2021] and the *sectional variation norm* by van der Laan [2017].

The sectional variation norm of smooth functions

In $d = 1$, if f is differentiable then

$$\|f\|_v = \int_0^1 |f'(x)| \, \mathrm{d}x.$$

The sectional variation norm of smooth functions

In $d = 1$, if f is differentiable then

$$\|f\|_v = \int_0^1 |f'(x)| \, \mathrm{d}x.$$

$\rightarrow$ Mild regularity condition.

The sectional variation norm of smooth functions

In $d = 1$, if f is differentiable then

$$\|f\|_v = \int_0^1 |f'(x)| \,\mathrm{d}x.$$

$\rightarrow$ Mild regularity condition.

Don't trust $d = 1$!

The sectional variation norm of smooth functions

In $d = 1$, if f is differentiable then

$$\|f\|_v = \int_0^1 |f'(x)| \, \mathrm{d}x.$$

$\rightarrow$ Mild regularity condition.

Don't trust $d = 1$!

In $d > 1$, if f is sufficiently smooth then

$$\|f\|_v = \sum_s \int_0^1 \cdots \int_0^1 \left| \frac{\partial^{|s|} f}{\partial x_1 \cdots \partial x_{|s|}} \right| \mathrm{d}x_1 \ldots x_{|s|}.$$

The sectional variation norm of smooth functions

In $d = 1$, if f is differentiable then

$$\|f\|_v = \int_0^1 |f'(x)| \, \mathrm{d}x.$$

$\rightarrow$ Mild regularity condition.

Don't trust $d = 1$!

In $d > 1$, if f is sufficiently smooth then

$$\|f\|_v = \sum_s \int_0^1 \cdots \int_0^1 \left| \frac{\partial^{|s|} f}{\partial x_1 \cdots \partial x_{|s|}} \right| \mathrm{d}x_1 \ldots x_{|s|}.$$

$\rightarrow$ Constraints on all mixed derivatives of order less than or equal to d.

The sectional variation norm of smooth functions

In $d = 1$, if f is differentiable then

$$\|f\|_v = \int_0^1 |f'(x)| \,\mathrm{d}x.$$

$\rightarrow$ Mild regularity condition.

Don't trust $d = 1$!

In $d > 1$, if f is sufficiently smooth then

$$\|f\|_v = \sum_s \int_0^1 \cdots \int_0^1 \left| \frac{\partial^{|s|} f}{\partial x_1 \cdots \partial x_{|s|}} \right| \mathrm{d}x_1 \ldots x_{|s|}.$$

$\rightarrow$ Constraints on all mixed derivatives of order less than or equal to d.

$\rightarrow$ The sum contains $\sum_{k=1}^n \binom{n}{k} = (2^d - 1)$ terms.

Implementation of the estimator

Implementation of the estimator

Gill et al. [1995] and van der Laan [2017] give the following representation of any $f \in \mathcal{D}_M([0,1]^d)$:

$$f(x) = \int_{[\mathbf{0},x]} \mathrm{d}f = f(0) + \sum_s \int_{(\mathbf{0}_s, x_s]} \mathrm{d}f_s.$$

Implementation of the estimator

Gill et al. [1995] and van der Laan [2017] give the following representation of any $f \in \mathcal{D}_M([0,1]^d)$:

$$f(x) = \int_{[\mathbf{0},x]} \mathrm{d}f = f(0) + \sum_s \int_{(\mathbf{0}_s, x_s]} \mathrm{d}f_s.$$

The norm $\|f\|_v$ is equal to the sum of the total variation of the measures on $(0(s), 1(s)]$ generated by the section f_s of f – hence the name.

Implementation of the estimator

Gill et al. [1995] and van der Laan [2017] give the following representation of any $f \in \mathcal{D}_M([0,1]^d)$:

$$f(x) = \int_{[0,x]} \mathrm{d}f = f(0) + \sum_s \int_{(0_s, x_s]} \mathrm{d}f_s.$$

The norm $\|f\|_v$ is equal to the sum of the total variation of the measures on $(0(s), 1(s)]$ generated by the section f_s of f – hence the name.

Suggests estimating f by estimating the measures $\mathrm{d}f_s$ with weighted empirical measures in the sections s:

$$f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s}\psi_{i,s}(x), \quad \text{with} \quad \psi_{i,s}(x) = \mathbb{1}\{X_i(s) \leq x(s)\},$$

with $\|\beta\|_1 = \sum |\beta_{s,i}| \leq M$.

Implementation of the estimator

Gill et al. [1995] and van der Laan [2017] give the following representation of any $f \in \mathcal{D}_M([0,1]^d)$:

$$f(x) = \int_{[0,x]} \mathrm{d}f = f(0) + \sum_s \int_{(0_s, x_s]} \mathrm{d}f_s.$$

The norm $\|f\|_v$ is equal to the sum of the total variation of the measures on $(0(s), 1(s)]$ generated by the section f_s of f – hence the name.

Suggests estimating f by estimating the measures $\mathrm{d}f_s$ with weighted empirical measures in the sections s:

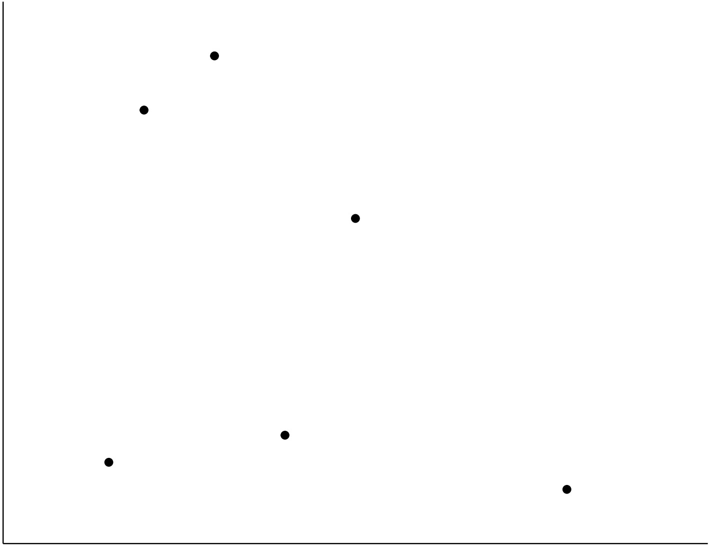
$$f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s} \psi_{i,s}(x), \quad \text{with} \quad \psi_{i,s}(x) = \mathbb{1}\{X_i(s) \leq x(s)\},$$

with $\|\beta\|_1 = \sum |\beta_{s,i}| \leq M$.

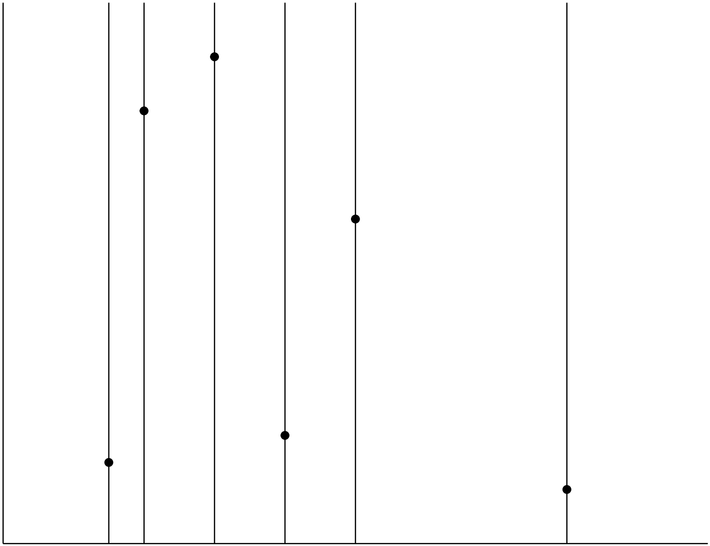
This can be phrased as the LASSO problem

$$\operatorname*{argmin}_\beta \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{such that} \quad \|\beta\|_1 \leq M.$$

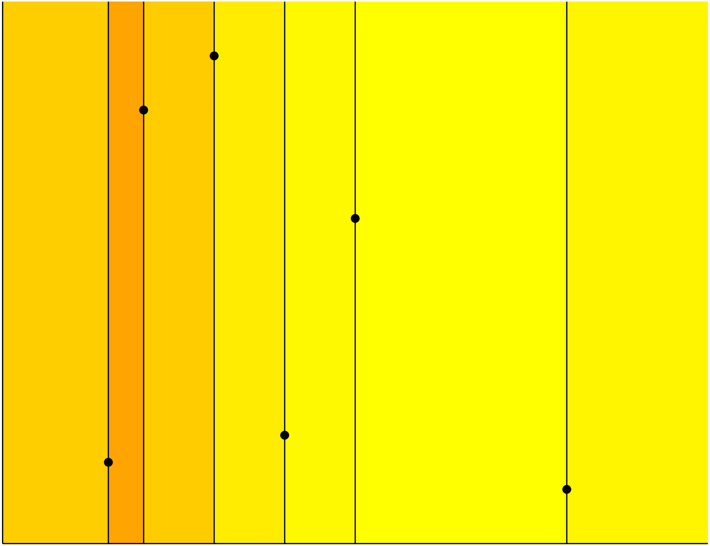
Basis functions for the HAL estimator



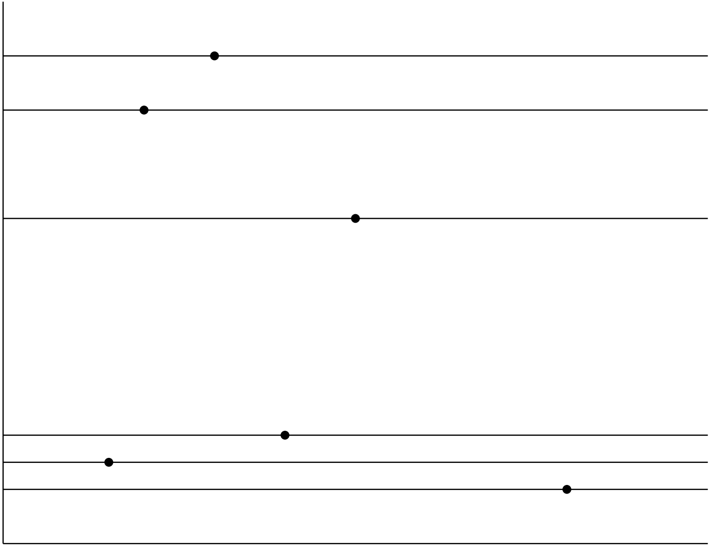
Basis functions for the HAL estimator



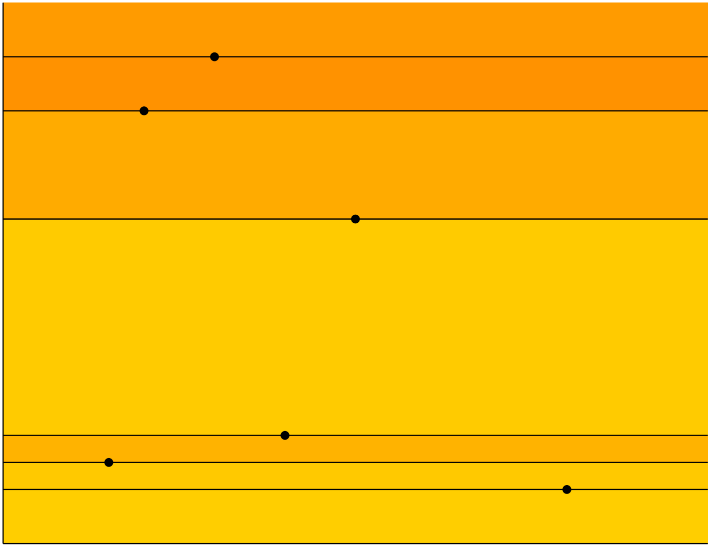
Basis functions for the HAL estimator



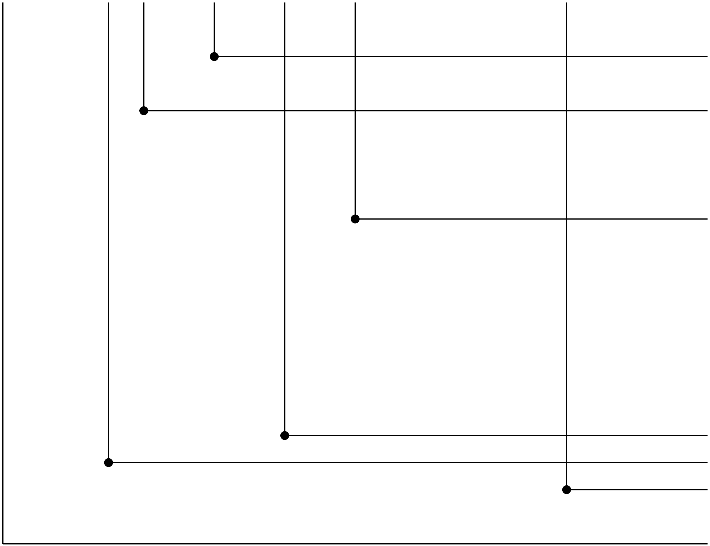
Basis functions for the HAL estimator



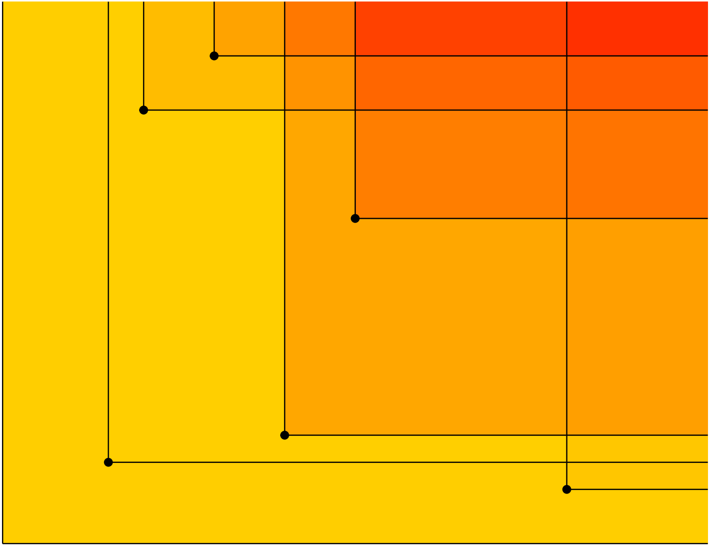
Basis functions for the HAL estimator



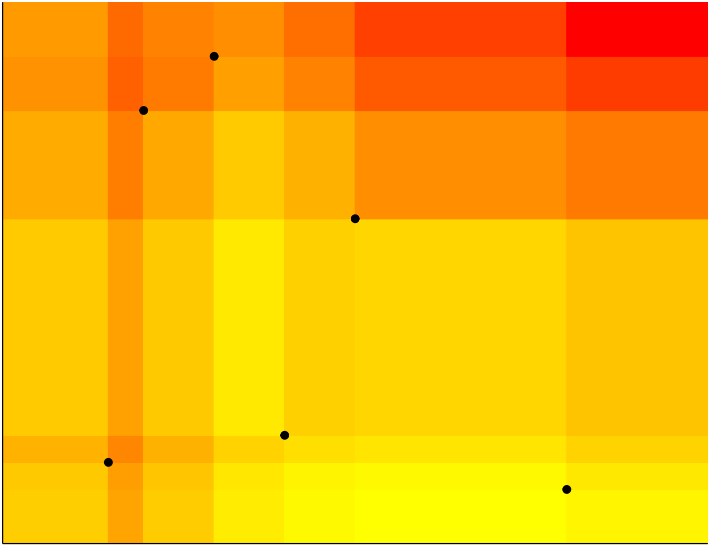
Basis functions for the HAL estimator



Basis functions for the HAL estimator



Basis functions for the HAL estimator



The solution to the minimization problem

$$\hat{\beta}_n = \underset{\beta:\|\beta\|_1 \leq M}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{with} \quad f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s}\psi_{i,s}(x).$$

The solution to the minimization problem

$$\hat{\beta}_n = \underset{\beta:\|\beta\|_1 \leq M}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{with} \quad f_\beta = \beta_0 + \sum_s \sum_{i=1}^{n} \beta_{i,s}\psi_{i,s}(x).$$

It seems to be common wisdom that

$$f_{\hat{\beta}_n} = \hat{f}_n = \underset{f \in \mathcal{D}_M([0,1]^d)}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f, \cdot)].$$

The solution to the minimization problem

$$\hat{\beta}_n = \underset{\beta:\|\beta\|_1 \leq M}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{with} \quad f_\beta = \beta_0 + \sum_s \sum_{i=1}^{n} \beta_{i,s}\psi_{i,s}(x).$$

It seems to be common wisdom that

$$f_{\hat{\beta}_n} = \hat{f}_n = \underset{f \in \mathcal{D}_M([0,1]^d)}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f, \cdot)].$$

✓ when $d = 1$ and, for instance, $L(f, (X, Y)) = \{f(X) - Y\}^2$

The solution to the minimization problem

$$\hat{\beta}_n = \underset{\beta:\|\beta\|_1 \le M}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{with} \quad f_\beta = \beta_0 + \sum_s \sum_{i=1}^{n} \beta_{i,s}\psi_{i,s}(x).$$

It seems to be common wisdom that

$$f_{\hat{\beta}_n} = \hat{f}_n = \underset{f \in \mathcal{D}_M([0,1]^d)}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f, \cdot)].$$

✓ when $d = 1$ and, for instance, $L(f, (X, Y)) = \{f(X) - Y\}^2$

Given f, construct $\bar{f}$ as the piece-wise constant function such that $\bar{f}(X_i) = f(X_i)$ and $\bar{f}(0) = f(0)$.

The solution to the minimization problem

$$\hat{\beta}_n = \underset{\beta:\|\beta\|_1 \le M}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{with} \quad f_\beta = \beta_0 + \sum_s \sum_{i=1}^{n} \beta_{i,s}\psi_{i,s}(x).$$

It seems to be common wisdom that

$$f_{\hat{\beta}_n} = \hat{f}_n = \underset{f \in \mathcal{D}_M([0,1]^d)}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f, \cdot)].$$

✓ when $d = 1$ and, for instance, $L(f, (X, Y)) = \{f(X) - Y\}^2$

Given f, construct $\bar{f}$ as the piece-wise constant function such that $\bar{f}(X_i) = f(X_i)$ and $\bar{f}(0) = f(0)$. Then

$$L(f, (X_i, Y_i)) = L(\bar{f}, (X_i, Y_i)), \quad \text{for all} \quad i = 1, \ldots, n,$$

The solution to the minimization problem

$$\hat{\beta}_n = \underset{\beta:\|\beta\|_1 \le M}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f_\beta, \cdot)], \quad \text{with} \quad f_\beta = \beta_0 + \sum_s \sum_{i=1}^{n} \beta_{i,s}\psi_{i,s}(x).$$

It seems to be common wisdom that

$$f_{\hat{\beta}_n} = \hat{f}_n = \underset{f \in \mathcal{D}_M([0,1]^d)}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f, \cdot)].$$

✓ when $d = 1$ and, for instance, $L(f, (X, Y)) = \{f(X) - Y\}^2$

Given f, construct $\bar{f}$ as the piece-wise constant function such that $\bar{f}(X_i) = f(X_i)$ and $\bar{f}(0) = f(0)$. Then

$$L(f, (X_i, Y_i)) = L(\bar{f}, (X_i, Y_i)), \quad \text{for all} \quad i = 1, \dots, n,$$

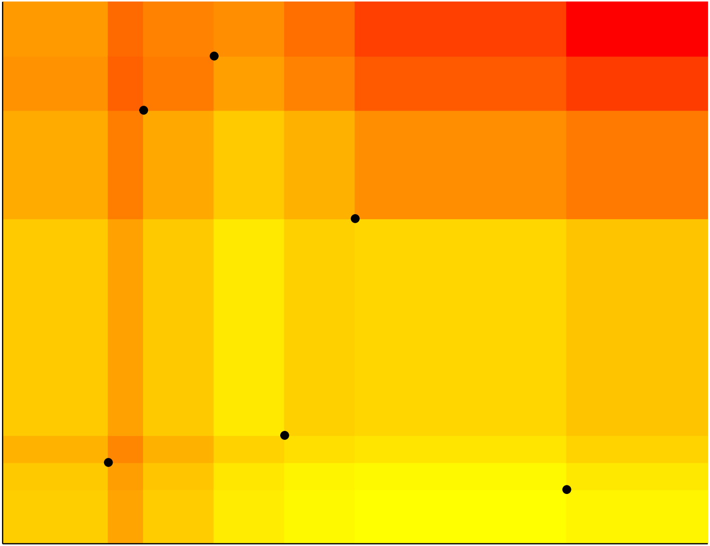
and, with $0 = X_{(0)} \le X_{(1)} \cdots \le X_{(n)}$,

$$\begin{aligned}\|\bar{f}\|_v = \sum_{i=1}^{n} |\bar{f}(X_{(i)}) - \bar{f}(X_{(i-1)})| &= \sum_{i=1}^{n} |f(X_{(i)}) - f(X_{(i-1)})| \\ &\le \sup_\pi \sum_{i=1}^{|\pi|} |f(t_i) - f(t_{i-1})| = \|f\|_v.\end{aligned}$$

Not clear same trick works when $d > 1$

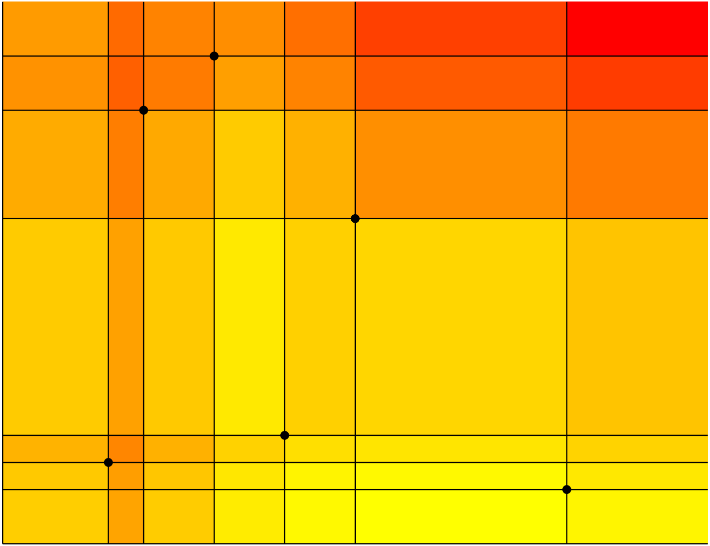
Not clear same trick works when $d > 1$ – don't trust $d = 1$!

Not clear same trick works when $d > 1$ – don't trust $d = 1$!



Not clear same trick works when $d > 1$ – don't trust $d = 1$!

We have $6 \times 3 + 1 = 17$ basis functions but $(6+1)^2 = 49$ rectangles.



Not clear same trick works when $d > 1$ – don't trust $d = 1$!

Fang et al. [2021] formally show that when L is the squared error loss, the minimizer

$$\hat{f}_n = \underset{f \in \mathcal{D}_M([0,1]^d)}{\operatorname{argmin}} \hat{\mathbb{P}}_n[L(f, \cdot)]$$

can be taken to be the solution to a LASSO problem using indicator functions as basis functions.

However, they need up to $\asymp n^d$ basis functions whereas the HAL estimator is made up of only $n \times (2^d - 1) + 1 \asymp n$ basis functions.

The estimator $\hat{f}_n$ in the survival setting

The estimator $\hat{f}_n$ in the survival setting

Consider estimation of the hazard for the survival time T.

$$\begin{aligned} \text{Data} \quad & O = (\tilde{T}, \Delta),\ \tilde{T} = T \wedge C,\ \Delta \in \{0,1\} \\ \text{Hazard} \quad & h = e^f,\ f \in \mathcal{D}_M([0,1]) \\ \text{Loss} \quad & L(f, O) = \int_0^{\tilde{T}} e^{f(s)} - \Delta f(\tilde{T}) \end{aligned}$$

The estimator $\hat{f}_n$ in the survival setting

Consider estimation of the hazard for the survival time T.

$$\begin{aligned} \text{Data} \quad & O = (\tilde{T}, \Delta),\ \tilde{T} = T \wedge C,\ \Delta \in \{0,1\} \\ \text{Hazard} \quad & h = e^f,\ f \in \mathcal{D}_M([0,1]) \\ \text{Loss} \quad & L(f, O) = \int_0^{\tilde{T}} e^{f(s)} - \Delta f(\tilde{T}) \end{aligned}$$

If there is an $i \in \{1, \ldots, n-1\}$ such that $f(\tilde{T}_{(i)}) > f(\tilde{T}_{(i+1)})$ then f is not the minimizer of $\hat{\mathbb{P}}_n[L(f, \cdot)]$ over $\mathcal{D}_M([0,1])$.

The estimator $\hat{f}_n$ in the survival setting

Consider estimation of the hazard for the survival time T.

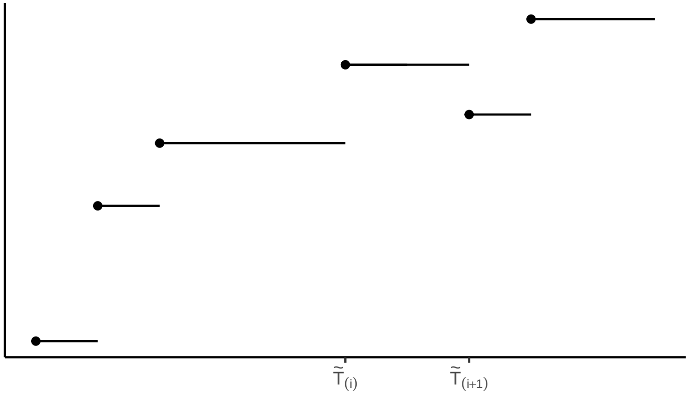
$$\begin{aligned} \text{Data} \quad & O = (\tilde{T}, \Delta),\ \tilde{T} = T \wedge C,\ \Delta \in \{0, 1\} \\ \text{Hazard} \quad & h = e^f,\ f \in \mathcal{D}_M([0, 1]) \\ \text{Loss} \quad & L(f, O) = \int_0^{\tilde{T}} e^{f(s)} - \Delta f(\tilde{T}) \end{aligned}$$

If there is an $i \in \{1, \ldots, n-1\}$ such that $f(\tilde{T}_{(i)}) > f(\tilde{T}_{(i+1)})$ then f is not the minimizer of $\hat{\mathbb{P}}_n[L(f, \cdot)]$ over $\mathcal{D}_M([0, 1])$.

$\Longrightarrow$ The empirical risk minimizer $\hat{f}_n$ is in general either not well-defined or a very bad estimator.

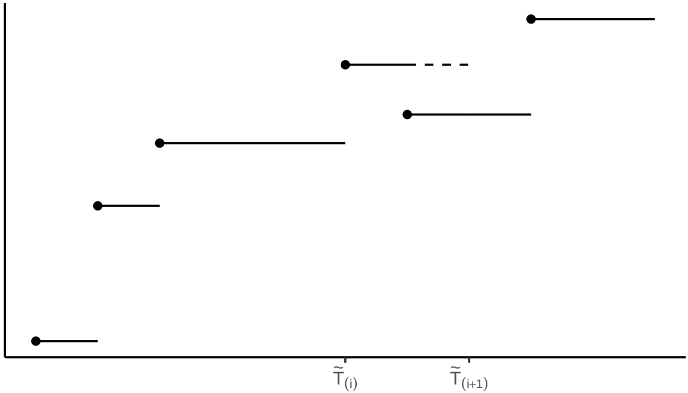
Proof by picture

$$L(f, O_i) = \int_0^{\tilde{T}_i} e^{f(s)} - \Delta_i f(\tilde{T}_i), \qquad \hat{\mathbb{P}}_n[L(f, \cdot)] = \frac{1}{n}\sum_{i=1}^{n} L(f, O_i)$$



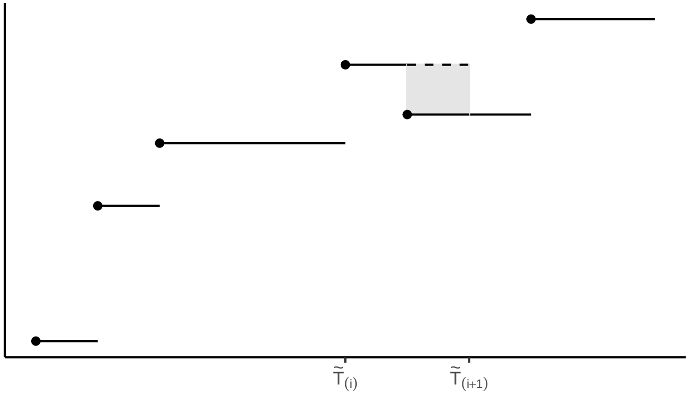
Proof by picture

$$L(f, O_i) = \int_0^{\tilde{T}_i} e^{f(s)} - \Delta_i f(\tilde{T}_i), \qquad \hat{\mathbb{P}}_n[L(f,\cdot)] = \frac{1}{n}\sum_{i=1}^{n} L(f, O_i)$$



Proof by picture

$$L(f, O_i) = \int_0^{\tilde{T}_i} e^{f(s)} - \Delta_i f(\tilde{T}_i), \qquad \hat{\mathbb{P}}_n[L(f, \cdot)] = \frac{1}{n}\sum_{i=1}^{n} L(f, O_i)$$



Need results for $f_{\hat{\beta}_n}$ instead of $\hat{f}_n$

Need results for $f_{\hat{\beta}_n}$ instead of $\hat{f}_n$

(Note that $f_{\hat{\beta}_n}$ is well-defined in the survival setting.)

Deriving convergence rates using empirical processes theory

Deriving convergence rates using empirical processes theory

The convergence rate for an empirical loss minimizer over a function space $\mathcal{F}$ can be read off from the *modulus of continuity* of the empirical process $\mathbb{G}_n = \sqrt{n}(\hat{\mathbb{P}}_n - P)$ over the space

$$\mathcal{L} = \{L(f, \cdot) - L(f_0, \cdot) : f \in \mathcal{F}\},$$

which is defined as

$$\varphi_n(\delta) = \mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{L}(\delta)}\right], \quad \text{where} \quad \|\mathbb{G}_n\|_{\mathcal{L}(\delta)} = \sup_{h \in \mathcal{L}(\delta)} |\mathbb{G}_n[h]|,$$

and $\mathcal{L}(\delta) = \{h \in \mathcal{L} : \|h\| \leq \delta\}$.

Deriving convergence rates using empirical processes theory

The convergence rate for an empirical loss minimizer over a function space $\mathcal{F}$ can be read off from the *modulus of continuity* of the empirical process $\mathbb{G}_n = \sqrt{n}(\hat{\mathbb{P}}_n - P)$ over the space

$$\mathcal{L} = \{L(f, \cdot) - L(f_0, \cdot) : f \in \mathcal{F}\},$$

which is defined as

$$\varphi_n(\delta) = \mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{L}(\delta)}\right], \quad \text{where} \quad \|\mathbb{G}_n\|_{\mathcal{L}(\delta)} = \sup_{h \in \mathcal{L}(\delta)} |\mathbb{G}_n[h]|,$$

and $\mathcal{L}(\delta) = \{h \in \mathcal{L} : \|h\| \leq \delta\}$.

The modulus φ_n can be controlled by the covering or bracketing entropy for $\mathcal{F}$. When $\mathcal{F} = \mathcal{D}_M([0,1]^d)$ this leads to the convergence rate

$$\|\hat{f}_n - f_0\| = \mathcal{O}_P(r_n), \quad \text{for} \quad r_n = n^{-1/3} \log(n)^{2(d-1)/3}.$$

Deriving convergence rates using empirical processes theory

The convergence rate for an empirical loss minimizer over a function space $\mathcal{F}$ can be read off from the *modulus of continuity* of the empirical process $\mathbb{G}_n = \sqrt{n}(\hat{\mathbb{P}}_n - P)$ over the space

$$\mathcal{L} = \{L(f, \cdot) - L(f_0, \cdot) : f \in \mathcal{F}\},$$

which is defined as

$$\varphi_n(\delta) = \mathbb{E}\left[\|\mathbb{G}_n\|_{\mathcal{L}(\delta)}\right], \quad \text{where} \quad \|\mathbb{G}_n\|_{\mathcal{L}(\delta)} = \sup_{h \in \mathcal{L}(\delta)} |\mathbb{G}_n[h]|,$$

and $\mathcal{L}(\delta) = \{h \in \mathcal{L} : \|h\| \leq \delta\}$.

The modulus φ_n can be controlled by the covering or bracketing entropy for $\mathcal{F}$. When $\mathcal{F} = \mathcal{D}_M([0,1]^d)$ this leads to the convergence rate

$$\|\hat{f}_n - f_0\| = \mathcal{O}_P(r_n), \quad \text{for} \quad r_n = n^{-1/3} \log(n)^{2(d-1)/3}.$$

Exact minimization is not needed – we just need the estimator $f_n^\star$ to fulfill

$$\hat{\mathbb{P}}_n[L(f_n^\star, \cdot)] \leq \hat{\mathbb{P}}_n[L(f_0, \cdot)] + \mathcal{O}_P(r_n^2).$$

This holds for $f_{\hat{\beta}_n}$: $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(f_0, \cdot)] + \mathcal{O}_P(r_n^2)$

This holds for $f_{\hat{\beta}_n}$: $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(f_0, \cdot)] + \mathcal{O}_P(r_n^2)$

Write

$$f_0(x) = f_0(0) + \sum_s \int_{(0_s, x_s]} \mathrm{d}f_{0,s} = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \, \mathrm{d}P_s,$$

and define

$$\tilde{f}_n = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \, \mathrm{d}\mathbb{P}_{s,n}.$$

This holds for $f_{\hat{\beta}_n}$: $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(f_0, \cdot)] + \mathcal{O}_P(r_n^2)$

Write

$$f_0(x) = f_0(0) + \sum_s \int_{(0_s, x_s]} \mathrm{d}f_{0,s} = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \, \mathrm{d}P_s,$$

and define

$$\tilde{f}_n = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \, \mathrm{d}\mathbb{P}_{s,n}.$$

The function $\tilde{f}_n$ is on the form $f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s}\psi_{i,s}(x)$, and by the law of large numbers $\|\tilde{f}_n\|_v \xrightarrow{P} \|f_0\|_v$.

This holds for $f_{\hat{\beta}_n}$: $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(f_0, \cdot)] + \mathcal{O}_P(r_n^2)$

Write

$$f_0(x) = f_0(0) + \sum_s \int_{(0_s, x_s]} \mathrm{d}f_{0,s} = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \,\mathrm{d}P_s,$$

and define

$$\tilde{f}_n = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \,\mathrm{d}\mathbb{P}_{s,n}.$$

The function $\tilde{f}_n$ is on the form $f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s}\psi_{i,s}(x)$, and by the law of large numbers $\|\tilde{f}_n\|_v \xrightarrow{P} \|f_0\|_v$.

Hence if $\|f_0\|_v < M$ then $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(\tilde{f}_n, \cdot)]$ with prob. $\to 1$, so

$$\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] - \hat{\mathbb{P}}_n[L(f_0, \cdot)] \leq \hat{\mathbb{P}}_n[L(\tilde{f}_n, \cdot)] - \hat{\mathbb{P}}_n[L(f_0, \cdot)] \quad \text{with prob. } \to 1.$$

This holds for $f_{\hat{\beta}_n}$: $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(f_0, \cdot)] + \mathcal{O}_P(r_n^2)$

Write

$$f_0(x) = f_0(0) + \sum_s \int_{(0_s, x_s]} \mathrm{d}f_{0,s} = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \, \mathrm{d}P_s,$$

and define

$$\tilde{f}_n = f_0(0) + \sum_s \int_{(0_s, x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s} \, \mathrm{d}\mathbb{P}_{s,n}.$$

The function $\tilde{f}_n$ is on the form $f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s}\psi_{i,s}(x)$, and by the law of large numbers $\|\tilde{f}_n\|_v \xrightarrow{P} \|f_0\|_v$.
Hence if $\|f_0\|_v < M$ then $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] \leq \hat{\mathbb{P}}_n[L(\tilde{f}_n, \cdot)]$ with prob. $\to 1$, so

$$\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n}, \cdot)] - \hat{\mathbb{P}}_n[L(f_0, \cdot)] \leq \hat{\mathbb{P}}_n[L(\tilde{f}_n, \cdot)] - \hat{\mathbb{P}}_n[L(f_0, \cdot)] \quad \text{with prob. } \to 1.$$

$$\|\tilde{f}_n - f_0\|_\infty = n^{-1/2} \sum_s \|\mathbb{G}_{s,n}\|_{\mathcal{D}} = \mathcal{O}_p(n^{-1/2}).$$

This holds for $f_{\hat{\beta}_n}$: $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n},\cdot)] \leq \hat{\mathbb{P}}_n[L(f_0,\cdot)] + \mathcal{O}_P(r_n^2)$

Write

$$f_0(x) = f_0(0) + \sum_s \int_{(0_s,x_s]} \mathrm{d}f_{0,s} = f_0(0) + \sum_s \int_{(0_s,x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s}\,\mathrm{d}P_s,$$

and define

$$\tilde{f}_n = f_0(0) + \sum_s \int_{(0_s,x_s]} \frac{\mathrm{d}f_{0,s}}{\mathrm{d}P_s}\,\mathrm{d}\mathbb{P}_{s,n}.$$

The function $\tilde{f}_n$ is on the form $f_\beta = \beta_0 + \sum_s \sum_{i=1}^n \beta_{i,s}\psi_{i,s}(x)$, and by the law of large numbers $\|\tilde{f}_n\|_v \xrightarrow{P} \|f_0\|_v$.

Hence if $\|f_0\|_v < M$ then $\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n},\cdot)] \leq \hat{\mathbb{P}}_n[L(\tilde{f}_n,\cdot)]$ with prob. $\to 1$, so

$$\hat{\mathbb{P}}_n[L(f_{\hat{\beta}_n},\cdot)] - \hat{\mathbb{P}}_n[L(f_0,\cdot)] \leq \hat{\mathbb{P}}_n[L(\tilde{f}_n,\cdot)] - \hat{\mathbb{P}}_n[L(f_0,\cdot)] \quad \text{with prob. } \to 1.$$

$$\|\tilde{f}_n - f_0\|_\infty = n^{-1/2} \sum_s \|\mathbb{G}_{s,n}\|_{\mathcal{D}} = \mathcal{O}_p(n^{-1/2}).$$

Combine this with a bound on $\varphi_n(\delta)$ for $\mathcal{D}_M([0,1]^d)$ to obtain

$$\hat{\mathbb{P}}_n[L(\tilde{f}_n,\cdot)] - \hat{\mathbb{P}}_n[L(f_0,\cdot)] = \mathcal{O}_P(r_n^2).$$

Conclusion and discussion

- We obtain the wanted convergence rate when using $f_{\hat{\beta}_n}$ as our estimator instead of $\hat{f}_n$.

Conclusion and discussion

- We obtain the wanted convergence rate when using $f_{\hat{\beta}_n}$ as our estimator instead of $\hat{f}_n$.
- This is an *asymptotic* result – no focus on finite sample bounds.

Conclusion and discussion

- We obtain the wanted convergence rate when using $f_{\hat{\beta}_n}$ as our estimator instead of $\hat{f}_n$.
- This is an *asymptotic* result – no focus on finite sample bounds.
- Would $\hat{f}_n$ (when it is defined) perform better than $f_{\hat{\beta}_n}$? – or is the reduction in the number of basis functions actually attractive in finite samples?

Conclusion and discussion

- We obtain the wanted convergence rate when using $f_{\hat{\beta}_n}$ as our estimator instead of $\hat{f}_n$.
- This is an *asymptotic* result – no focus on finite sample bounds.
- Would $\hat{f}_n$ (when it is defined) perform better than $f_{\hat{\beta}_n}$? – or is the reduction in the number of basis functions actually attractive in finite samples?
- Can we reduce the number of basis functions further? – would be computationally attractive.

Conclusion and discussion

- We obtain the wanted convergence rate when using $f_{\hat{\beta}_n}$ as our estimator instead of $\hat{f}_n$.
- This is an *asymptotic* result – no focus on finite sample bounds.
- Would $\hat{f}_n$ (when it is defined) perform better than $f_{\hat{\beta}_n}$? – or is the reduction in the number of basis functions actually attractive in finite samples?
- Can we reduce the number of basis functions further? – would be computationally attractive.
- What kind of constraints are put on functions in $\mathcal{D}_M([0,1]^d)$ that are continuous but not much smoother?

References

B. Fang, A. Guntuboyina, and B. Sen. Multivariate extensions of isotonic regression and total variation denoising via entire monotonicity and hardy–krause variation. *The Annals of Statistics*, 49(2):769–792, 2021.

R. D. Gill, M. J. Laan, and J. A. Wellner. Inefficient estimators of the bivariate survival function for three models. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 545–597, 1995.

G. Neuhaus. On weak convergence of stochastic processes with multidimensional time parameter. *The Annals of Mathematical Statistics*, 42(4):1285–1295, 1971.

M. van der Laan. A generally efficient targeted minimum loss based estimator based on the highly adaptive lasso. *The international journal of biostatistics*, 13(2), 2017.