Debiasing and functional derivatives

Anders Munch

June 2, 2023

Outline

- Understand the efficient influence curve as the derivative of our target parameter
- Investigate the bias-variance trade-off with infinite-dimensional nuisance parameters
- See how the derivative interpretation can help us understand targeted/debiased estimation strategies

High-level perspective on a statistical problem

Example (the average treatment effect)

Given i.i.d. data $O_i = (X_i, A_i, Y_i) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}$, estimate

$$\mathbb{E}_P[f(1,W) - f(0,W)], \quad \text{with} \quad f(a,x) = \mathbb{E}_P[Y \mid A = a, X = x],$$

assuming $P(A = 1 \mid X = x) \in [\varepsilon, 1-\varepsilon]$ for all x for some fixed $\varepsilon > 0$.

High-level perspective on a statistical problem

Example (the average treatment effect)

Given i.i.d. data $O_i = (X_i, A_i, Y_i) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}$, estimate

$$\mathbb{E}_P[f(1,W) - f(0,W)], \quad \text{with} \quad f(a,x) = \mathbb{E}_P[Y \mid A = a, X = x],$$

assuming $P(A = 1 \mid X = x) \in [\varepsilon, 1-\varepsilon]$ for all x for some fixed $\varepsilon > 0$.

The targeted learning philosophy tells us that any statistical estimation problem is essentially characterized by

- Ψ the target parameter of interest
- $\mathcal{P}$ the model, i.e. ideally, the assumptions we are willing to make

High-level perspective on a statistical problem

Example (the average treatment effect)

Given i.i.d. data $O_i = (X_i, A_i, Y_i) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}$, estimate

$$\mathbb{E}_P\left[f(1,W) - f(0,W)\right], \quad \text{with} \quad f(a,x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right],$$

assuming $P(A = 1 \mid X = x) \in [\varepsilon, 1-\varepsilon]$ for all x for some fixed $\varepsilon > 0$.

The targeted learning philosophy tells us that any statistical estimation problem is essentially characterized by

$$\begin{aligned} \Psi &= \mathbb{E}_P\left[f(1,W) - f(0,W)\right] \\ \mathcal{P} &= \text{all distributions } P \text{ such that } P(A = 1 \mid X = x) \in [\varepsilon, 1-\varepsilon] \end{aligned}$$

High-level perspective on a statistical problem

Example (the average treatment effect)

Given i.i.d. data $O_i = (X_i, A_i, Y_i) \in \mathbb{R}^d \times \{0,1\} \times \mathbb{R}$, estimate

$$\mathbb{E}_P\left[f(1,W) - f(0,W)\right], \quad \text{with} \quad f(a,x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right],$$

assuming $P(A = 1 \mid X = x) \in [\varepsilon, 1-\varepsilon]$ for all x for some fixed $\varepsilon > 0$.

The targeted learning philosophy tells us that any statistical estimation problem is essentially characterized by

- Ψ the target parameter of interest
- $\mathcal{P}$ the model, i.e. ideally, the assumptions we are willing to make

What can we say about

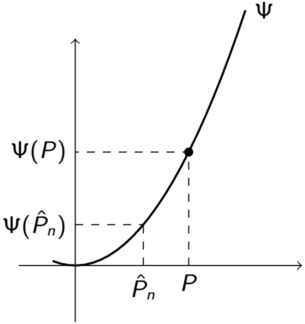
- the statistical problem $(\Psi, \mathcal{P})$?
- estimators of $\Psi(P)$ based on data generated by some $P \in \mathcal{P}$?

General approach – understand the derivative of Ψ

Understand the behavior of $\Psi(\hat{P}_n) - \Psi(P)$ through the derivative of Ψ.

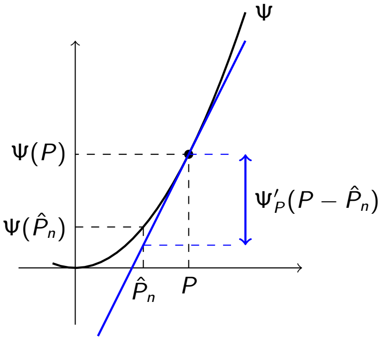
General approach – understand the derivative of Ψ

Understand the behavior of $\Psi(\hat{P}_n) - \Psi(P)$ through the derivative of Ψ.



General approach – understand the derivative of Ψ

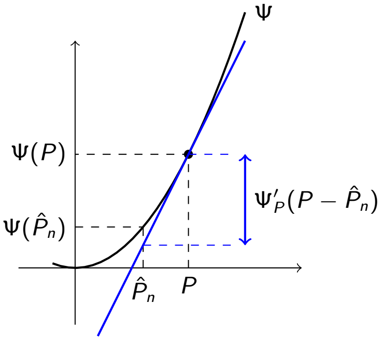
Understand the behavior of $\Psi(\hat{P}_n) - \Psi(P)$ through the derivative of Ψ.



- The derivative provides a local approximation of the map Ψ around P

General approach – understand the derivative of Ψ

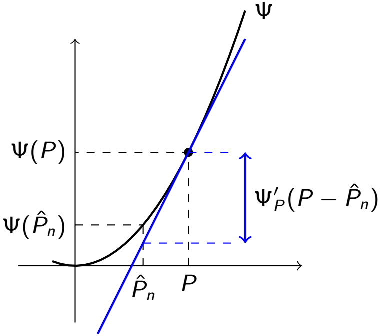
Understand the behavior of $\Psi(\hat{P}_n) - \Psi(P)$ through the derivative of Ψ.



- The derivative provides a local approximation of the map Ψ around P
- We can think of estimation of P as approaching P with our estimator $\hat{P}_n$

General approach – understand the derivative of Ψ

Understand the behavior of $\Psi(\hat{P}_n) - \Psi(P)$ through the derivative of Ψ.



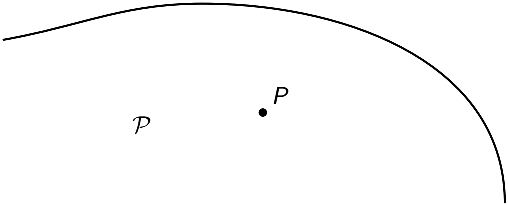
- The derivative provides a local approximation of the map Ψ around P
- We can think of estimation of P as approaching P with our estimator $\hat{P}_n$
- Thus, asymptotically, the derivative could give us a good idea about the behavior of $\Psi(\hat{P}_n)$

Geometric perspective

How can we talk about approaching P in $\mathcal{P}$?

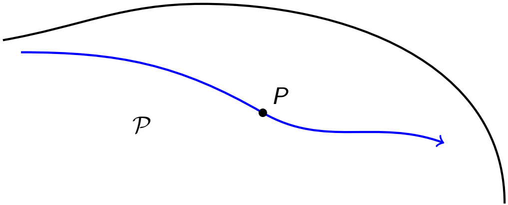
Geometric perspective

How can we talk about approaching P in $\mathcal{P}$?
$\rightarrow$ submodels $\{P_t\} \subset \mathcal{P}$ such that $P_{t=0} = P$



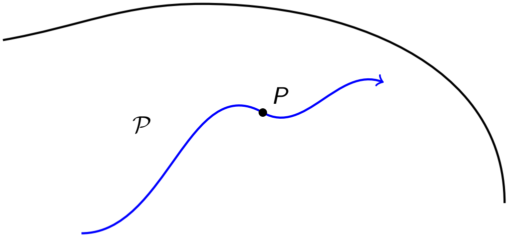
Geometric perspective

How can we talk about approaching P in $\mathcal{P}$?
$\rightarrow$ submodels $\{P_t\} \subset \mathcal{P}$ such that $P_{t=0} = P$



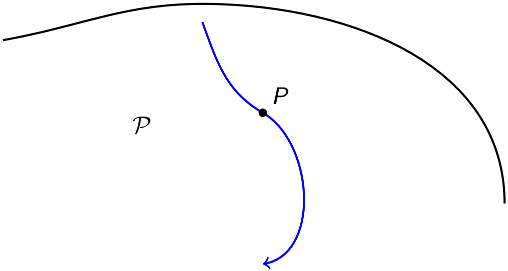
Geometric perspective

How can we talk about approaching P in $\mathcal{P}$?
$\rightarrow$ submodels $\{P_t\} \subset \mathcal{P}$ such that $P_{t=0} = P$



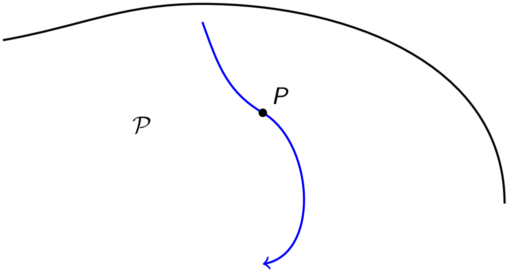
Geometric perspective

How can we talk about approaching P in $\mathcal{P}$?
$\rightarrow$ submodels $\{P_t\} \subset \mathcal{P}$ such that $P_{t=0} = P$



Geometric perspective

How can we talk about approaching P in $\mathcal{P}$?
$\rightarrow$ submodels $\{P_t\} \subset \mathcal{P}$ such that $P_{t=0} = P$



Using finite-dimensional submodels we know how to talk about the *likelihood* and the *score function* of such models. The tangent space is the collection of all score functions,

$$\dot{\mathcal{P}}_P = \overline{\mathrm{span}}\{\dot{\ell}_0\}, \quad \text{where} \quad \dot{\ell}_0 = \frac{\partial}{\partial t}\Big|_{t=0} \log(p_t), \quad P_t = p_t \cdot \mu.$$

Gradients

One can show that $\dot{\mathcal{P}}_P \subset \mathcal{L}_0^2(P) = \{f \in \mathcal{L}^2(P) : P[f] = 0\}$, where $\mathcal{L}^2(P)$ is the Hilbert space of P-square integrable functions with inner product $\langle f, g \rangle_P = P[fg] = \mathbb{E}_P\,[f(O)g(O)]$.

Gradients

One can show that $\dot{\mathcal{P}}_P \subset \mathcal{L}_0^2(P) = \{f \in \mathcal{L}^2(P) : P[f] = 0\}$, where $\mathcal{L}^2(P)$ is the Hilbert space of P-square integrable functions with inner product $\langle f, g \rangle_P = P[fg] = \mathbb{E}_P[f(O)g(O)]$.

A *gradient* is a function $g \in \mathcal{L}_0^2(P)$ such that

$$\frac{\partial}{\partial t}\bigg|_{t=0} \Psi(P_t) = \langle g, \dot{\ell}_0 \rangle_P, \quad \text{for all submodels } \{P_t\} \text{ with score } \dot{\ell}_0.$$

There can be many gradients but there is a unique gradient φ_P such that $\varphi_P \in \dot{\mathcal{P}}_P$. This is called the *canonical gradient*.

Gradients

One can show that $\dot{\mathcal{P}}_P \subset \mathcal{L}_0^2(P) = \{f \in \mathcal{L}^2(P) : P[f] = 0\}$, where $\mathcal{L}^2(P)$ is the Hilbert space of P-square integrable functions with inner product $\langle f, g\rangle_P = P[fg] = \mathbb{E}_P\left[f(O)g(O)\right]$.

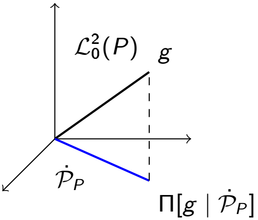
A *gradient* is a function $g \in \mathcal{L}_0^2(P)$ such that

$$\frac{\partial}{\partial t}\bigg|_{t=0} \Psi(P_t) = \langle g, \dot{\ell}_0\rangle_P, \quad \text{for all submodels } \{P_t\} \text{ with score } \dot{\ell}_0.$$

There can be many gradients but there is a unique gradient φ_P such that $\varphi_P \in \dot{\mathcal{P}}_P$. This is called the *canonical gradient*.

The canonical gradient can be found as the projection of any gradient onto the tangent space,

$$\varphi_P = \Pi[g \mid \dot{\mathcal{P}}_P].$$



Derivative and the chain rule

If $t \mapsto P_t \in \mathbb{R}^k$ and $\Psi \colon \mathbb{R}^k \to \mathbb{R}$, the chain rule tells us that

$$(\Psi \circ P)'(t) = (\Psi'(P_t))^T P_t' = \left\langle \Psi'(P_t), P_t' \right\rangle .$$

Derivative and the chain rule

If $t \mapsto P_t \in \mathbb{R}^k$ and $\Psi\colon \mathbb{R}^k \to \mathbb{R}$, the chain rule tells us that

$$(\Psi \circ P)'(t) = (\Psi'(P_t))^T P_t' = \left\langle \Psi'(P_t), P_t' \right\rangle .$$

Using that

$$\frac{\partial}{\partial t}\Big|_{t=0} \log(p_t) p = \frac{\partial}{\partial t}\Big|_{t=0} p_t,$$

the (canonical) gradient fulfills

$$\frac{\partial}{\partial t}\Big|_{t=0} \Psi(P_t) = \langle \varphi_P, \dot{\ell}_0 \rangle_P = \left\langle \varphi_P, \frac{\partial}{\partial t}\Big|_{t=0} \log(p_t) \right\rangle_P = \left\langle \varphi_P, \frac{\partial}{\partial t}\Big|_{t=0} p_t \right\rangle_\mu .$$

Derivative and the chain rule

If $t \mapsto P_t \in \mathbb{R}^k$ and $\Psi \colon \mathbb{R}^k \to \mathbb{R}$, the chain rule tells us that

$$(\Psi \circ P)'(t) = (\Psi'(P_t))^T P_t' = \left\langle \Psi'(P_t), P_t' \right\rangle .$$

Using that

$$\frac{\partial}{\partial t}\Big|_{t=0} \log(p_t) p = \frac{\partial}{\partial t}\Big|_{t=0} p_t,$$

the (canonical) gradient fulfills

$$\frac{\partial}{\partial t}\Big|_{t=0} \Psi(P_t) = \langle \varphi_P, \dot{\ell}_0 \rangle_P = \left\langle \varphi_P, \frac{\partial}{\partial t}\Big|_{t=0} \log(p_t) \right\rangle_P = \left\langle \varphi_P, \frac{\partial}{\partial t}\Big|_{t=0} p_t \right\rangle_\mu .$$

Define the canonical gradient as a function such that the chain rule holds

$$\Updownarrow$$

Define a suitable type of derivative and then prove that the chain rule holds

Submodels and information bounds

For a one-dimensional submodel $\{P_t\}$ the Cramér-Rao bound states

$$\mathrm{Var}[\hat{\Psi}_n] \geq \frac{\left(\frac{\partial}{\partial t}\Big|_{t=0} \Psi(P_t)\right)^2}{P[\dot{\ell}_0^2]} =: V(\{P_t\}, \Psi),$$

for any (suitably regular) estimator $\hat{\Psi}_n$. The *information bound* for the finite-dimensional model $\{P_t\}$ is $\mathcal{I}(\{P_t\}, \Psi) = V(\{P_t\}, \Psi)^{-1}$.

The information bound for the full model $\mathcal{P}$ is

$$\mathcal{I}(\mathcal{P}, \Psi) = \inf_{\{P_t\}} \mathcal{I}(\{P_t\}, \Psi).$$

Submodels and information bounds

For a one-dimensional submodel $\{P_t\}$ the Cramér-Rao bound states

$$\mathrm{Var}[\hat{\Psi}_n] \geq \frac{\left(\frac{\partial}{\partial t}\big|_{t=0}\Psi(P_t)\right)^2}{P[\dot{\ell}_0^2]} =: V(\{P_t\}, \Psi),$$

for any (suitably regular) estimator $\hat{\Psi}_n$. The *information bound* for the finite-dimensional model $\{P_t\}$ is $\mathcal{I}(\{P_t\}, \Psi) = V(\{P_t\}, \Psi)^{-1}$.

The information bound for the full model $\mathcal{P}$ is

$$\mathcal{I}(\mathcal{P}, \Psi) = \inf_{\{P_t\}} \mathcal{I}(\{P_t\}, \Psi).$$

$\mathcal{I}(\mathcal{P}, \Psi)^{-1} = P[\varphi_P^2]$, where φ_P is the canonical gradient.

Influence functions and RAL estimators

An estimator $\hat{\Psi}_n$ of the parameter Ψ under the model $\mathcal{P}$, is called *asymptotically linear* with *influence function* $\mathrm{IF}(\cdot, P) \in \mathcal{L}^2(P)$, if $P[\mathrm{IF}(\cdot, P)] = 0$ for all $P \in \mathcal{P}$, and

$$\hat{\Psi}_n - \Psi = \mathbb{P}_n[\mathrm{IF}(\cdot, P)] + o_P(n^{-1/2}).$$

In particular, the asymptotic variance of $\hat{\Psi}_n$ is $P[\mathrm{IF}(\cdot, P)^2]$.

Influence functions and RAL estimators

An estimator $\hat{\Psi}_n$ of the parameter Ψ under the model $\mathcal{P}$, is called *asymptotically linear* with *influence function* $\mathrm{IF}(\cdot, P) \in \mathcal{L}^2(P)$, if $P[\mathrm{IF}(\cdot, P)] = 0$ for all $P \in \mathcal{P}$, and

$$\hat{\Psi}_n - \Psi = \mathbb{P}_n[\mathrm{IF}(\cdot, P)] + o_P(n^{-1/2}).$$

In particular, the asymptotic variance of $\hat{\Psi}_n$ is $P[\mathrm{IF}(\cdot, P)^2]$.

The influence function IF of any regular asymptotically linear (RAL) estimator is a gradient. Hence $\varphi_P = \Pi[\mathrm{IF} \mid \dot{\mathcal{P}}_P]$.

Influence functions and RAL estimators

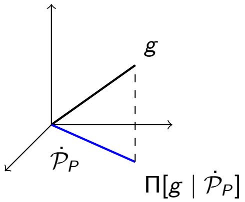
An estimator $\hat{\Psi}_n$ of the parameter Ψ under the model $\mathcal{P}$, is called *asymptotically linear* with *influence function* $\mathrm{IF}(\cdot, P) \in \mathcal{L}^2(P)$, if $P[\mathrm{IF}(\cdot, P)] = 0$ for all $P \in \mathcal{P}$, and

$$\hat{\Psi}_n - \Psi = \mathbb{P}_n[\mathrm{IF}(\cdot, P)] + o_P(n^{-1/2}).$$

In particular, the asymptotic variance of $\hat{\Psi}_n$ is $P[\mathrm{IF}(\cdot, P)^2]$.

The influence function IF of any regular asymptotically linear (RAL) estimator is a gradient. Hence $\varphi_P = \Pi[\mathrm{IF} \mid \dot{\mathcal{P}}_P]$.

This implies that a RAL estimator with φ_P as influence function will be *efficient* – it has lowest possible asymptotic variance among all RAL estimators.



Summary so far

- The geometric perspective is useful because it allows us to talk about how difficult a statistical problem is through the information bound.
- The differential perspective is useful because it provides us with a completely description of the asymptotic behavior of any (RAL) estimator.
- It even suggests a strategy for constructing efficient estimators.

Summary so far

- The geometric perspective is useful because it allows us to talk about how difficult a statistical problem is through the information bound.
- The differential perspective is useful because it provides us with a completely description of the asymptotic behavior of any (RAL) estimator.
- It even suggests a strategy for constructing efficient estimators.

$\rightarrow$ We move on to talk a bit more about estimation, in particular ...

Estimating low-dimensional target parameters using estimators of infinite-dimensional nuisance parameters

The naïve plug-in strategy

Often we can write

$$\Psi(P) = \tilde{\Psi}(Q(P)),$$

for some nuisance parameter Q.

Natural strategy: Estimate Q with $\hat{Q}_n$ and use $\hat{\Psi}_n = \tilde{\Psi}(\hat{Q}_n)$.

The naïve plug-in strategy

Often we can write

$$\Psi(P) = \tilde{\Psi}(Q(P)),$$

for some nuisance parameter Q.

Natural strategy: Estimate Q with $\hat{Q}_n$ and use $\hat{\Psi}_n = \tilde{\Psi}(\hat{Q}_n)$.

Example (Average treatment effect)

We can write

$$\Psi(P) = \tilde{\Psi}(f(P), \mu(P)), \quad \text{with} \quad \tilde{\Psi}(f, \mu) = \int f(1, x) - f(0, x) \, \mathrm{d}\mu(x),$$

so here $Q = (f, \mu)$. This suggests using

$$\hat{\Psi}_n = \tilde{\Psi}(\hat{f}_n, \mathbb{P}_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_n(1, X_i) - \hat{f}_n(0, X_i).$$

The naïve plug-in strategy

Often we can write

$$\Psi(P) = \tilde{\Psi}(Q(P)),$$

for some nuisance parameter Q.

Natural strategy: Estimate Q with $\hat{Q}_n$ and use $\hat{\Psi}_n = \tilde{\Psi}(\hat{Q}_n)$.

Example (Average treatment effect)

We can write

$$\Psi(P) = \tilde{\Psi}(f(P), \mu(P)), \quad \text{with} \quad \tilde{\Psi}(f, \mu) = \int f(1, x) - f(0, x) \, \mathrm{d}\mu(x),$$

so here $Q = (f, \mu)$. This suggests using

$$\hat{\Psi}_n = \tilde{\Psi}(\hat{f}_n, \mathbb{P}_n) = \frac{1}{n} \sum_{i=1}^{n} \hat{f}_n(1, X_i) - \hat{f}_n(0, X_i).$$

When $Q \in \mathbb{R}^k$ then, under regularity conditions, we have

- if $\hat{Q}_n$ is asymptotically linear so is $\tilde{\Psi}(\hat{Q}_n)$
- if $\hat{Q}_n$ is efficient so is $\tilde{\Psi}(\hat{Q}_n)$

Infinite-dimensional nuisance parameter

In general, if we use flexible data-adaptive methods to estimate the nuisance parameter, the "naïve" plug-in strategy does not work well.

Infinite-dimensional nuisance parameter

In general, if we use flexible data-adaptive methods to estimate the nuisance parameter, the "naïve" plug-in strategy does not work well.

Example (Kernel density plug-in)

Consider the problem of estimating

$$\Psi(P) = P(X \leq x), \quad \text{for some fixed} \quad x \in \mathbb{R}.$$

We assume that P has a Lebesgue density and write

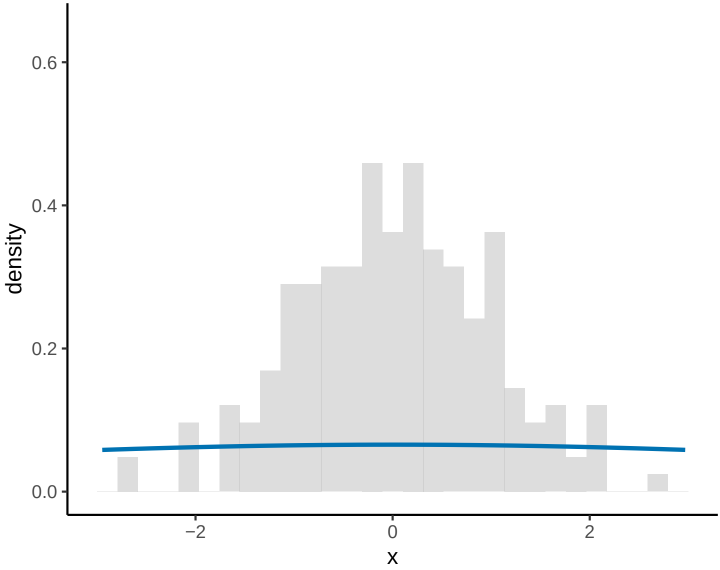
$$\Psi(P) = \tilde{\Psi}(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z.$$

We decide to estimate Ψ by first estimating the density f with a kernel-based density estimator $\hat{f}_h$. We then obtain an estimator of Ψ as

$$\hat{\Psi}_n = \tilde{\Psi}(\hat{f}_h) = \int_{-\infty}^{x} \hat{f}_h(z)\,\mathrm{d}z.$$

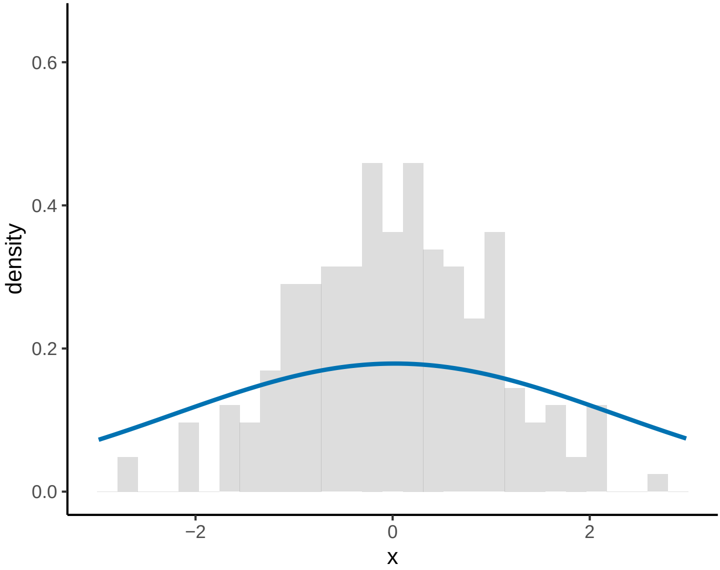
Kernel estimator and bandwidth

small h large



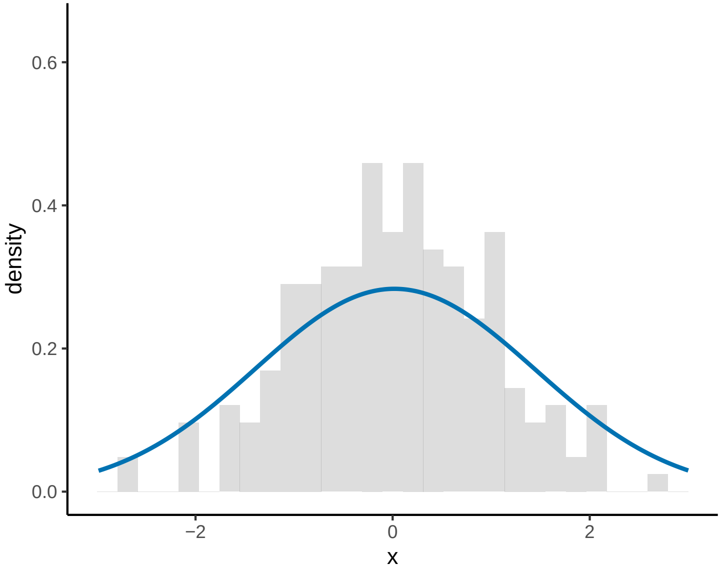
Kernel estimator and bandwidth

small h large



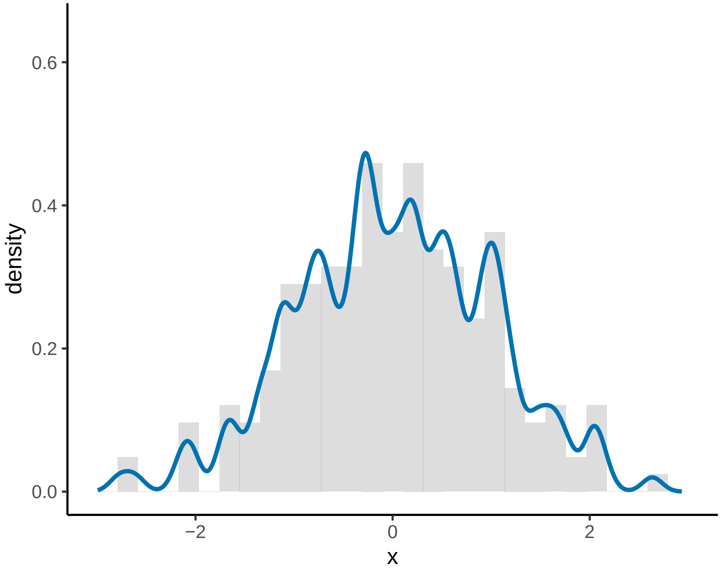
Kernel estimator and bandwidth

small h large



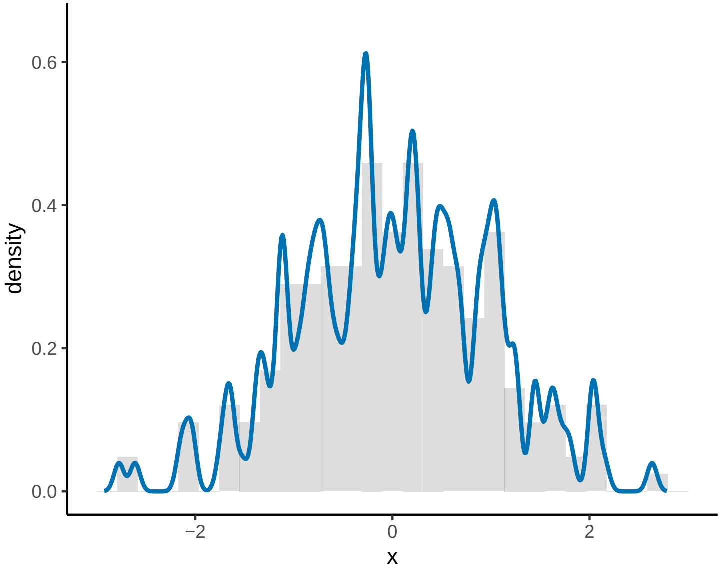
Kernel estimator and bandwidth

small h large

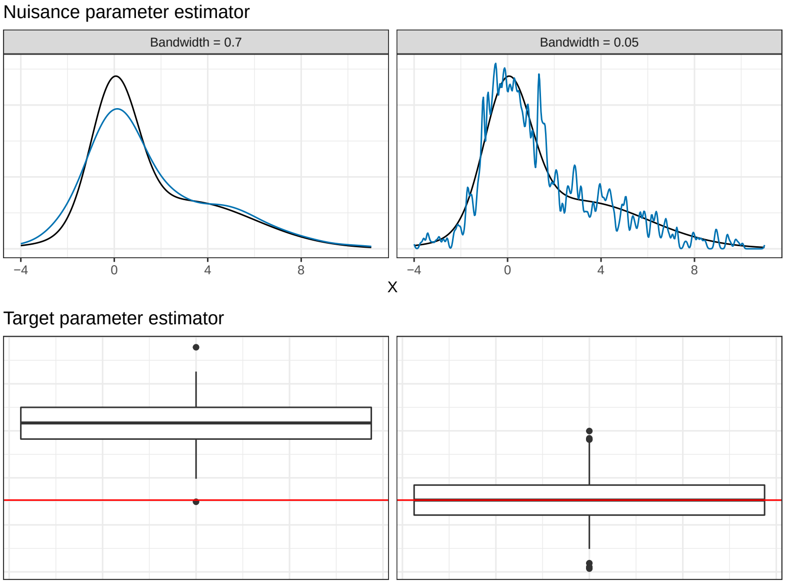


Kernel estimator and bandwidth

small h large

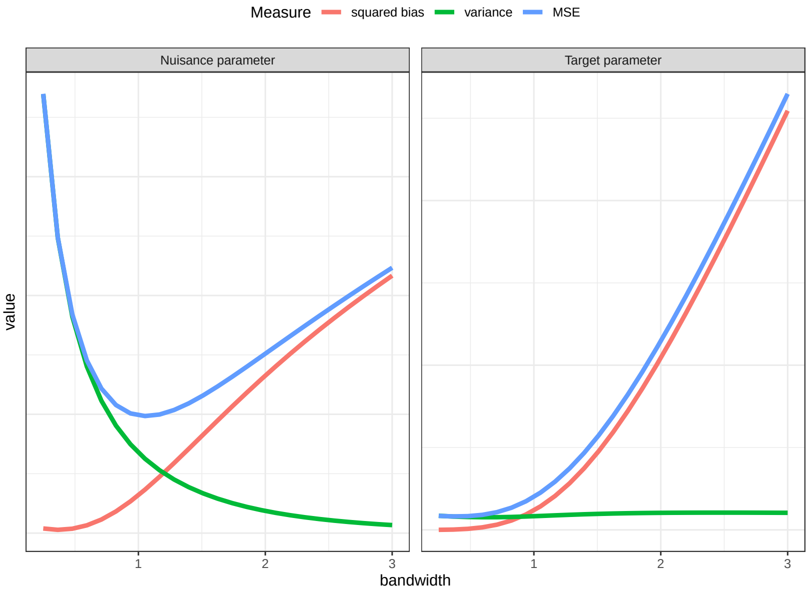


How does this work?



Exercise

Conclusion from the exercise – bias-variance trade-off



Decomposition

For a general problem, we can write

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) + \mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] + o_P(1),$$

where $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P)$ is the empirical process and φ the canonical gradient, when $\varphi_{\hat{P}_n} \xrightarrow{P} \varphi_P$. Informally, imagine that $\mathbb{G}_n$ and $\hat{P}_n$ are independent; then

$$\mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] \sim \mathcal{N}\left(0, P\left[\left(\varphi_{\hat{P}_n} - \varphi_P\right)^2\right]\right) \rightsquigarrow 0.$$

Decomposition

For a general problem, we can write

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) + \mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] + o_P(1),$$

where $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P)$ is the empirical process and φ the canonical gradient, when $\varphi_{\hat{P}_n} \xrightarrow{P} \varphi_P$. Informally, imagine that $\mathbb{G}_n$ and $\hat{P}_n$ are independent; then

$$\mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] \sim \mathcal{N}\left(0, P\left[\left(\varphi_{\hat{P}_n} - \varphi_P\right)^2\right]\right) \rightsquigarrow 0.$$

Thus

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = \mathbb{G}_n[\varphi_P] + o_P(1) - n^{1/2}\mathbb{P}_n[\varphi_{\hat{P}_n}] \\ + n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}]\right)$$

Decomposition

For a general problem, we can write

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) + \mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] + o_P(1),$$

where $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P)$ is the empirical process and φ the canonical gradient, when $\varphi_{\hat{P}_n} \xrightarrow{P} \varphi_P$. Informally, imagine that $\mathbb{G}_n$ and $\hat{P}_n$ are independent; then

$$\mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] \sim \mathcal{N}\left(0, P\left[\left(\varphi_{\hat{P}_n} - \varphi_P\right)^2\right]\right) \rightsquigarrow 0.$$

Thus

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = \mathbb{G}_n[\varphi_P] + o_P(1) - n^{1/2}\mathbb{P}_n[\varphi_{\hat{P}_n}] + n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}]\right)$$

One-step $\hat{\Psi}_n^* = \Psi(\hat{P}_n) + \mathbb{P}_n[\varphi_{\hat{P}_n}]$

Decomposition

For a general problem, we can write

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) + \mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] + o_P(1),$$

where $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P)$ is the empirical process and φ the canonical gradient, when $\varphi_{\hat{P}_n} \xrightarrow{P} \varphi_P$. Informally, imagine that $\mathbb{G}_n$ and $\hat{P}_n$ are independent; then

$$\mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] \sim \mathcal{N}\left(0, P\left[\left(\varphi_{\hat{P}_n} - \varphi_P\right)^2\right]\right) \rightsquigarrow 0.$$

Thus

$$\begin{aligned} n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) &= \mathbb{G}_n[\varphi_P] + o_P(1) - n^{1/2}\mathbb{P}_n[\varphi_{\hat{P}_n}] \\ &\quad + n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}]\right) \end{aligned}$$

One-step $\hat{\Psi}_n^* = \Psi(\hat{P}_n) + \mathbb{P}_n[\varphi_{\hat{P}_n}]$

TMLE $\Psi(\hat{P}_n^*)$ with $\hat{P}_n^*$ such that $n^{1/2}\mathbb{P}_n[\varphi_{\hat{P}_n^*}] = o_P(1)$

Decomposition

For a general problem, we can write

$$n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) = n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) + \mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] + o_P(1),$$

where $\mathbb{G}_n := n^{1/2}(\mathbb{P}_n - P)$ is the empirical process and φ the canonical gradient, when $\varphi_{\hat{P}_n} \xrightarrow{P} \varphi_P$. Informally, imagine that $\mathbb{G}_n$ and $\hat{P}_n$ are independent; then

$$\mathbb{G}_n[\varphi_P - \varphi_{\hat{P}_n}] \sim \mathcal{N}\left(0, P\left[\left(\varphi_{\hat{P}_n} - \varphi_P\right)^2\right]\right) \rightsquigarrow 0.$$

Thus

$$\begin{aligned} n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P)\right) &= \mathbb{G}_n[\varphi_P] + o_P(1) - n^{1/2}\mathbb{P}_n[\varphi_{\hat{P}_n}] \\ &\quad + n^{1/2}\left(\Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}]\right) \end{aligned}$$

One-step $\hat{\Psi}_n^* = \Psi(\hat{P}_n) + \mathbb{P}_n[\varphi_{\hat{P}_n}]$

TMLE $\Psi(\hat{P}_n^*)$ with $\hat{P}_n^*$ such that $n^{1/2}\mathbb{P}_n[\varphi_{\hat{P}_n^*}] = o_P(1)$

The remainder term vanishes if only $\hat{P}_n = P + o_P(n^{-1/4})$!

Functional Taylor expansion

For $f \to \mathbb{R} \to \mathbb{R}$ differentiable at $a \in \mathbb{R}$ we have

$$f(x) = f(a) + f'(a)(x - a) + \mathrm{Rem}(a, x)$$

with $\mathrm{Rem}(a, x) = o(|x - a|)$ – when f is smooth enough the remainder is of second order, i.e., $\mathrm{Rem}(a, x) = o((x - a)^2)$.

Functional Taylor expansion

For $f \to \mathbb{R} \to \mathbb{R}$ differentiable at $a \in \mathbb{R}$ we have

$$f(x) = f(a) + f'(a)(x - a) + \mathrm{Rem}(a, x)$$

with $\mathrm{Rem}(a, x) = o(|x - a|)$ – when f is smooth enough the remainder is of second order, i.e., $\mathrm{Rem}(a, x) = o((x - a)^2)$.

Similarly, we can write

$$\begin{aligned} \Psi(P) &= \Psi(\hat{P}_n) + \langle \varphi_{\hat{P}_n}, p - \hat{p}_n \rangle_\mu + \mathrm{Rem}(\hat{P}_n, P) \\ &= \Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] - \hat{P}_n[\varphi_{\hat{P}_n}] + \mathrm{Rem}(\hat{P}_n, P) \\ &= \Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] + \mathrm{Rem}(\hat{P}_n, P) \end{aligned}$$

$$\implies \quad |\mathrm{Rem}(\hat{P}_n, P)| = \left| \Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}] \right|$$

Functional Taylor expansion

For $f \to \mathbb{R} \to \mathbb{R}$ differentiable at $a \in \mathbb{R}$ we have

$$f(x) = f(a) + f'(a)(x - a) + \mathrm{Rem}(a, x)$$

with $\mathrm{Rem}(a, x) = o(|x - a|)$ – when f is smooth enough the remainder is of second order, i.e., $\mathrm{Rem}(a, x) = o((x - a)^2)$.

Similarly, we can write

$$\begin{aligned} \Psi(P) &= \Psi(\hat{P}_n) + \langle \varphi_{\hat{P}_n}, p - \hat{p}_n \rangle_\mu + \mathrm{Rem}(\hat{P}_n, P) \\ &= \Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] - \hat{P}_n[\varphi_{\hat{P}_n}] + \mathrm{Rem}(\hat{P}_n, P) \\ &= \Psi(\hat{P}_n) + P[\varphi_{\hat{P}_n}] + \mathrm{Rem}(\hat{P}_n, P) \\ \implies \quad |\mathrm{Rem}(\hat{P}_n, P)| &= \left| \Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}] \right| \end{aligned}$$

So if Ψ is smooth enough we would expect

$$\begin{aligned} \mathrm{Rem}(\hat{P}_n, P) &= o_P(\|\hat{P}_n - P\|^2) \\ \implies n^{1/2} \left(\Psi(\hat{P}_n) - \Psi(P) + P[\varphi_{\hat{P}_n}] \right) &= o_P(1) \quad \text{if} \quad \hat{P}_n = P + o_P(n^{-1/4}). \end{aligned}$$

Summary so far

- Targeted or "debiased" learning essentially works by calculating and correcting the first order asymptotic bias due to estimation of an (infinite-dimensional) nuisance parameter.
- This is done using a functional Taylor expansion.
- The canonical gradient is an important tool in this regard, because it is the derivative of the map $\Psi \colon \mathcal{P} \to \mathbb{R}$.
- The targeting/debiasing step is important when working with flexible, data-adaptive estimators of infinite-dimensional nuisance parameters (as we saw in the exercise)

The tangent space and the canonical gradient

The canonical gradient depends on the tangents space $\dot{\mathcal{P}}_P$ ($\varphi_P = \Pi[g \mid \dot{\mathcal{P}}_P]$).

Changing assumptions $\implies$ changes information bound.

The tangent space and the canonical gradient

The canonical gradient depends on the tangents space $\dot{\mathcal{P}}_P$ ($\varphi_P = \Pi[g \mid \dot{\mathcal{P}}_P]$).

Changing assumptions $\implies$ changes information bound.

- A non-parametric model will have $\dot{\mathcal{P}}_P = \mathcal{L}_0^2(P)$
- A parametric model will have a finite-dimensional $\dot{\mathcal{P}}_P$
- Semi-parametric models can have infinite-dimensional $\dot{\mathcal{P}}_P \subsetneq \mathcal{L}_0^2(P)$

The tangent space and the canonical gradient

The canonical gradient depends on the tangents space $\dot{\mathcal{P}}_P$ ($\varphi_P = \Pi[g \mid \dot{\mathcal{P}}_P]$).

Changing assumptions $\implies$ changes information bound.

- A non-parametric model will have $\dot{\mathcal{P}}_P = \mathcal{L}_0^2(P)$
- A parametric model will have a finite-dimensional $\dot{\mathcal{P}}_P$
- Semi-parametric models can have infinite-dimensional $\dot{\mathcal{P}}_P \subsetneq \mathcal{L}_0^2(P)$

The non-parametric case is important because it implies that that all RAL estimators are efficient and asymptotically equivalent.

The tangent space and the canonical gradient

The canonical gradient depends on the tangents space $\dot{\mathcal{P}}_P$ ($\varphi_P = \Pi[g \mid \dot{\mathcal{P}}_P]$).

Changing assumptions $\implies$ changes information bound.

- A non-parametric model will have $\dot{\mathcal{P}}_P = \mathcal{L}_0^2(P)$
- A parametric model will have a finite-dimensional $\dot{\mathcal{P}}_P$
- Semi-parametric models can have infinite-dimensional $\dot{\mathcal{P}}_P \subsetneq \mathcal{L}_0^2(P)$

The non-parametric case is important because it implies that that all RAL estimators are efficient and asymptotically equivalent.

Changing assumptions does not *always* imply changing the information bound.

- Smoothness- or shape-constraints often does not change $\dot{\mathcal{P}}_P$
- Independence assumptions can change $\dot{\mathcal{P}}_P$ without changing φ_P

This type of information have no effect asymptotically – but it might still be relevant to incorporate to improve finite sample performance.

Strategy for finding (candidate for) the canonical gradient

We can find a candidate for the canonical gradient in a nonparametric model by calculating the directional derivative of Ψ in the direction of the Dirac measure δ_{O_i}.[1]

[1] See also Hines et al. [2022] and Ichimura and Newey [2015].

Strategy for finding (candidate for) the canonical gradient

We can find a candidate for the canonical gradient in a nonparametric model by calculating the directional derivative of Ψ in the direction of the Dirac measure δ_{O_i}.[1]

For any $h \in \mathbb{R}_+$, define the sub-model $P_\varepsilon^h := P + \varepsilon K_{O_i}^h$, where $K_{O_i}^h \to \delta_{O_i}$, $h \to 0$. The score function of this model is

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(\mathrm{d}P + \varepsilon\, \mathrm{d}K_{O_i}^h) = \frac{\mathrm{d}K_{O_i}^h}{\mathrm{d}P}.$$

[1] See also Hines et al. [2022] and Ichimura and Newey [2015].

Strategy for finding (candidate for) the canonical gradient

We can find a candidate for the canonical gradient in a nonparametric model by calculating the directional derivative of Ψ in the direction of the Dirac measure δ_{O_i}.[1]

For any $h \in \mathbb{R}_+$, define the sub-model $P_\varepsilon^h := P + \varepsilon K_{O_i}^h$, where $K_{O_i}^h \to \delta_{O_i}$, $h \to 0$. The score function of this model is

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(\mathrm{d}P + \varepsilon\, \mathrm{d}K_{O_i}^h) = \frac{\mathrm{d}K_{O_i}^h}{\mathrm{d}P}.$$

Using the property of a gradient we have

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P_\varepsilon^h) = \left\langle \varphi_P, \frac{\mathrm{d}K_{O_i}^h}{\mathrm{d}P} \right\rangle_P = \int \varphi_P(o) \frac{\mathrm{d}K_{O_i}^h(o)}{\mathrm{d}P(o)}\, \mathrm{d}P(o) = \int \varphi_P(o)\, \mathrm{d}K_{O_i}^h(o).$$

[1] See also Hines et al. [2022] and Ichimura and Newey [2015].

Strategy for finding (candidate for) the canonical gradient

We can find a candidate for the canonical gradient in a nonparametric model by calculating the directional derivative of Ψ in the direction of the Dirac measure δ_{O_i}.[1]

For any $h \in \mathbb{R}_+$, define the sub-model $P_\varepsilon^h := P + \varepsilon K_{O_i}^h$, where $K_{O_i}^h \to \delta_{O_i}$, $h \to 0$. The score function of this model is

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(\mathrm{d}P + \varepsilon\, \mathrm{d}K_{O_i}^h) = \frac{\mathrm{d}K_{O_i}^h}{\mathrm{d}P}.$$

Using the property of a gradient we have

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P_\varepsilon^h) = \left\langle \varphi_P, \frac{\mathrm{d}K_{O_i}^h}{\mathrm{d}P} \right\rangle_P = \int \varphi_P(o) \frac{\mathrm{d}K_{O_i}^h(o)}{\mathrm{d}P(o)}\, \mathrm{d}P(o) = \int \varphi_P(o)\, \mathrm{d}K_{O_i}^h(o).$$

Letting $h \to 0$, we get a candidate for the efficient influence curve:

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon \delta_{O_i}) = \lim_{h \to 0} \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P_\varepsilon^h) = \lim_{h \to 0} \int \varphi_P(o)\, \mathrm{d}K_{O_i}^h(o) = \varphi_P(O_i).$$

[1] See also Hines et al. [2022] and Ichimura and Newey [2015].

Example – estimating the mean

Consider the parameter

$$\Psi(P) = \mathbb{E}_P[O] = \int o \, \mathrm{d}P(o).$$

Example – estimating the mean

Consider the parameter

$$\Psi(P) = \mathbb{E}_P[O] = \int o \,\mathrm{d}P(o).$$

Let O_i be a fixed point and consider

$$\begin{aligned}
\Psi(P + \varepsilon\delta_{O_i}) &= \int o \,\mathrm{d}[P + \varepsilon\delta_{O_i}](o) \\
&= \int o \,\mathrm{d}P(o) + \varepsilon \int o\delta_{O_i}(o) \\
&= \int o \,\mathrm{d}P(o) + \varepsilon O_i.
\end{aligned}$$

Example – estimating the mean

Consider the parameter

$$\Psi(P) = \mathbb{E}_P\left[O\right] = \int o \, \mathrm{d}P(o).$$

Let O_i be a fixed point and consider

$$\begin{aligned}
\Psi(P + \varepsilon\delta_{O_i}) &= \int o \, \mathrm{d}[P + \varepsilon\delta_{O_i}](o) \\
&= \int o \, \mathrm{d}P(o) + \varepsilon \int o\delta_{O_i}(o) \\
&= \int o \, \mathrm{d}P(o) + \varepsilon O_i .
\end{aligned}$$

Taking the derivative with respect to ε we get

$$\frac{\partial}{\partial\varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon\delta_{O_i}) = \frac{\partial}{\partial\varepsilon}\Big|_{\varepsilon=0} \left\{ \int o \, \mathrm{d}P(o) + \varepsilon O_i \right\} = 0 + 1 \cdot O_i = O_i$$

Example – estimating the mean

Consider the parameter

$$\Psi(P) = \mathbb{E}_P\left[O\right] = \int o \,\mathrm{d}P(o).$$

Let O_i be a fixed point and consider

$$\begin{aligned}\Psi(P + \varepsilon\delta_{O_i}) &= \int o \,\mathrm{d}[P + \varepsilon\delta_{O_i}](o) \\ &= \int o \,\mathrm{d}P(o) + \varepsilon \int o\delta_{O_i}(o) \\ &= \int o \,\mathrm{d}P(o) + \varepsilon O_i.\end{aligned}$$

Taking the derivative with respect to ε we get

$$\frac{\partial}{\partial\varepsilon}\Big|_{\varepsilon=0}\Psi(P + \varepsilon\delta_{O_i}) = \frac{\partial}{\partial\varepsilon}\Big|_{\varepsilon=0}\left\{\int o \,\mathrm{d}P(o) + \varepsilon O_i\right\} = 0 + 1 \cdot O_i = O_i$$

Thus $\tilde{\varphi}_P(o) = o$ is a candidate – to make it integrate to 0 we can use

$$\varphi_P(o) = \tilde{\varphi}_P(o) - \int \tilde{\varphi}_P \,\mathrm{d}P(o) = o - \Psi(P).$$

Exercise

References

O. Hines, O. Dukes, K. Diaz-Ordaz, and S. Vansteelandt. Demystifying statistical learning based on efficient influence functions. *The American Statistician*, 76(3): 292–304, 2022.

H. Ichimura and W. K. Newey. The influence function of semiparametric estimators. *arXiv preprint arXiv:1508.01378*, 2015.