Lecture 4: Bias-variance trade-off with infinite-dimensional nuisance parameters

Anders Munch

September 27, 2021

Outline

A statistical estimation problem

We call a collection of probability measures $\mathcal{P}$ together with a functional $\Psi\colon \mathcal{P} \to \mathbb{R}$ *a statistical estimation problem*.

A statistical estimation problem

We call a collection of probability measures $\mathcal{P}$ together with a functional $\Psi \colon \mathcal{P} \to \mathbb{R}$ *a statistical estimation problem*.

Example (Average treatment effect)

We are given n iid. samples of $O \sim \mathrm{P}$, with $\mathrm{P} \in \mathcal{P}$ where $O = (X, A, Y)$, with $X \in \mathbb{R}^d$, $A \in \{0, 1\}$, and $Y \in \{0, 1\}$. We want to estimate the average treatment effect

$$\mathbb{E}_{\mathrm{P}} \left[f_P(1, X) - f_P(0, X) \right],$$

with $f_P(a, x) := \mathbb{E}_{\mathrm{P}} \left[Y \mid A = a, X = x \right]$. The target parameter is

$$\Psi(\mathrm{P}) = \mathbb{E}_{\mathrm{P}} \left[f_{\mathrm{P}}(1, X) - f_{\mathrm{P}}(0, X) \right].$$

A statistical estimation problem

We call a collection of probability measures $\mathcal{P}$ together with a functional $\Psi \colon \mathcal{P} \to \mathbb{R}$ *a statistical estimation problem*.

Example (Average treatment effect)

We are given n iid. samples of $O \sim \mathrm{P}$, with $\mathrm{P} \in \mathcal{P}$ where $O = (X, A, Y)$, with $X \in \mathbb{R}^d$, $A \in \{0, 1\}$, and $Y \in \{0, 1\}$. We want to estimate the average treatment effect

$$\mathbb{E}_{\mathrm{P}}\left[f_P(1, X) - f_P(0, X)\right],$$

with $f_P(a, x) := \mathbb{E}_{\mathrm{P}}\left[Y \mid A = a, X = x\right]$. The target parameter is

$$\Psi(\mathrm{P}) = \mathbb{E}_{\mathrm{P}}\left[f_{\mathrm{P}}(1, X) - f_{\mathrm{P}}(0, X)\right].$$

Target and nuisance parameters

Target and nuisance parameters

Target parameter is low-dimensional, scientifically meaningful.

Target and nuisance parameters

Target parameter is low-dimensional, scientifically meaningful.

Nuisance parameters are needed to express the target parameter.

Target and nuisance parameters

Target parameter is low-dimensional, scientifically meaningful.

Nuisance parameters are needed to express the target parameter.

Example (Average treatment effect)

We can express the ATE as

$$\Psi(P) = \tilde{\Psi}(f, \mu) := \int_{\mathbb{R}^p} [f(1, x) - f(0, x)] \, \mathrm{d}\mu(x),$$

where μ is the marginal distribution of X. The nuisance parameters are f and μ.

Target and nuisance parameters

Target parameter is low-dimensional, scientifically meaningful.

Nuisance parameters are needed to express the target parameter.

Example (Average treatment effect)

We can express the ATE as

$$\Psi(P) = \tilde{\Psi}(f, \mu) := \int_{\mathbb{R}^p} [f(1, x) - f(0, x)] \, \mathrm{d}\mu(x),$$

where μ is the marginal distribution of X. The nuisance parameters are f and μ. This immediately suggests the target estimator

$$\hat{\psi}_n^{\text{g-formula}} = \Psi(\hat{f}_n, \hat{\mu}_X).$$

For instance, if we use $\hat{\mu}_X = \hat{\mathbb{P}}_n$ we have

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

Hence we just have to select a nuisance estimator $\hat{f}_n$.

Nuisance parameter estimation and plug-in strategy

In many interesting cases the nuisance parameter is a conditional mean

Nuisance parameter estimation and plug-in strategy

In many interesting cases the nuisance parameter is a conditional mean

$\rightarrow$ Estimation of the nuisance parameter reduces to the standard statistical / machine learning task of fitting a prediction model.

Nuisance parameter estimation and plug-in strategy

In many interesting cases the nuisance parameter is a conditional mean

- $\rightarrow$ Estimation of the nuisance parameter reduces to the standard statistical / machine learning task of fitting a prediction model.
- $\rightarrow$ This is a very well studied problem, and a zoo of estimators exists for these kinds of problems (most machine learning problems are of this type)

Nuisance parameter estimation and plug-in strategy

In many interesting cases the nuisance parameter is a conditional mean

- $\rightarrow$ Estimation of the nuisance parameter reduces to the standard statistical / machine learning task of fitting a prediction model.
- $\rightarrow$ This is a very well studied problem, and a zoo of estimators exists for these kinds of problems (most machine learning problems are of this type)
- $\rightarrow$ Provides estimators of the target parameter by plugging in a nuisance estimator

Nuisance parameter estimation and plug-in strategy

In many interesting cases the nuisance parameter is a conditional mean

- $\rightarrow$ Estimation of the nuisance parameter reduces to the standard statistical / machine learning task of fitting a prediction model.
- $\rightarrow$ This is a very well studied problem, and a zoo of estimators exists for these kinds of problems (most machine learning problems are of this type)
- $\rightarrow$ Provides estimators of the target parameter by plugging in a nuisance estimator

Such a plug-in strategy can perform poorly, even when the nuisance estimator is optimized for the nuisance problem.

Nuisance parameter estimation and plug-in strategy

In many interesting cases the nuisance parameter is a conditional mean

→ Estimation of the nuisance parameter reduces to the standard statistical / machine learning task of fitting a prediction model.

→ This is a very well studied problem, and a zoo of estimators exists for these kinds of problems (most machine learning problems are of this type)

→ Provides estimators of the target parameter by plugging in a nuisance estimator

Such a plug-in strategy can perform poorly, even when the nuisance estimator is optimized for the nuisance problem.

We will demonstrate this with the following a toy example.

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F_P(x) = P(X \leq x)$ for unknown $P \in \mathcal{P}$.

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F_P(x) = P(X \leq x)$ for unknown $P \in \mathcal{P}$. Our target parameter is then $\Psi(P) = F_P(x)$ which we can express as

$$\Psi(P) = \tilde{\Psi}(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z, \quad \text{for} \quad P = f \cdot \lambda,$$

because of our assumption about $\mathcal{P}$.

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F_P(x) = P(X \leq x)$ for unknown $P \in \mathcal{P}$. Our target parameter is then $\Psi(P) = F_P(x)$ which we can express as

$$\Psi(P) = \tilde{\Psi}(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z, \quad \text{for} \quad P = f \cdot \lambda,$$

because of our assumption about $\mathcal{P}$. If we wanted to use "machine learning" for this problem, we could use a kernel estimator

$$\hat{f}_n(x) = \hat{\mathbb{P}}_n[k_h(X, x)] = \frac{1}{n}\sum_{i=1}^{n} k_h(X_i, x),$$

to estimate the density f, where $h \in \mathbb{R}_+$ is the bandwidth (a tuning parameter).

Toy example: Integrated kernel density

$\mathcal{P}$ consist all probability measures with continuous Lebesgue-density (this is an infinite-dimensional space). We want to estimate $F_P(x) = P(X \leq x)$ for unknown $P \in \mathcal{P}$. Our target parameter is then $\Psi(P) = F_P(x)$ which we can express as

$$\Psi(P) = \tilde{\Psi}(f) := \int_{-\infty}^{x} f(z)\,\mathrm{d}z, \quad \text{for} \quad P = f \cdot \lambda,$$

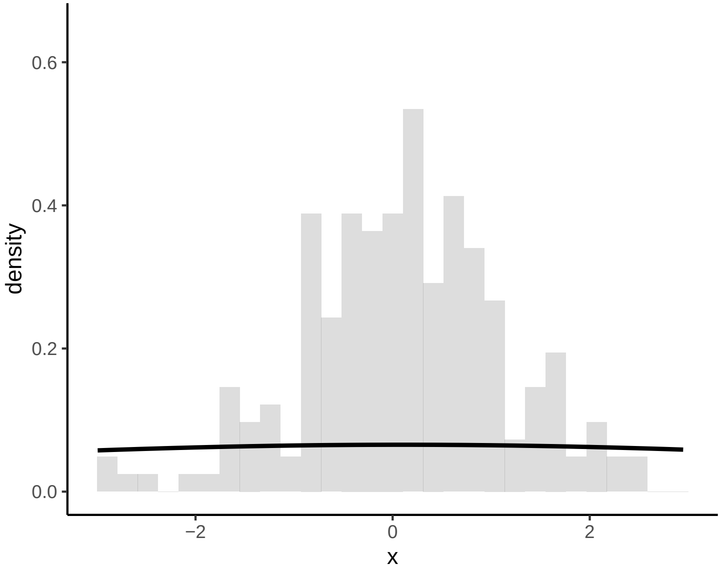
because of our assumption about $\mathcal{P}$. If we wanted to use "machine learning" for this problem, we could use a kernel estimator

$$\hat{f}_n(x) = \hat{\mathbb{P}}_n[k_h(X, x)] = \frac{1}{n}\sum_{i=1}^{n} k_h(X_i, x),$$

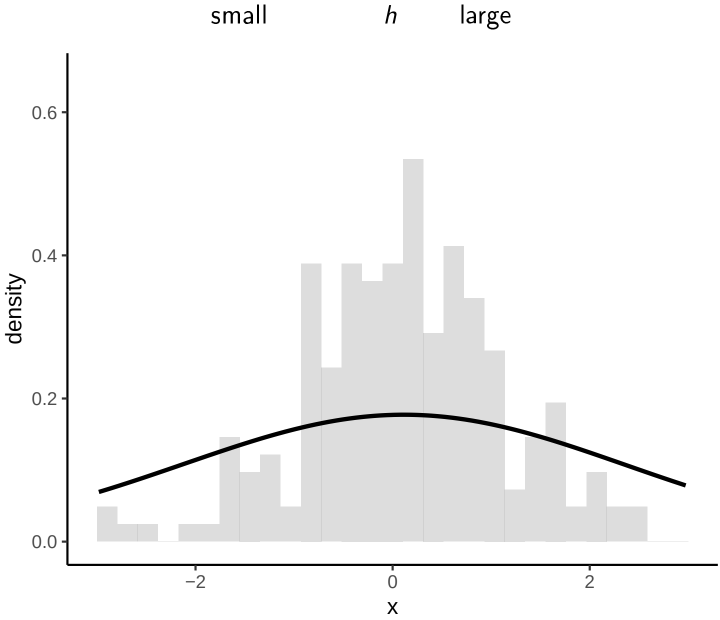
to estimate the density f, where $h \in \mathbb{R}_+$ is the bandwidth (a tuning parameter). We could then obtain the target estimator $\hat{\psi}_n = \tilde{\Psi}(\hat{f}_n)$.

Visualize kernel estimator

small h large

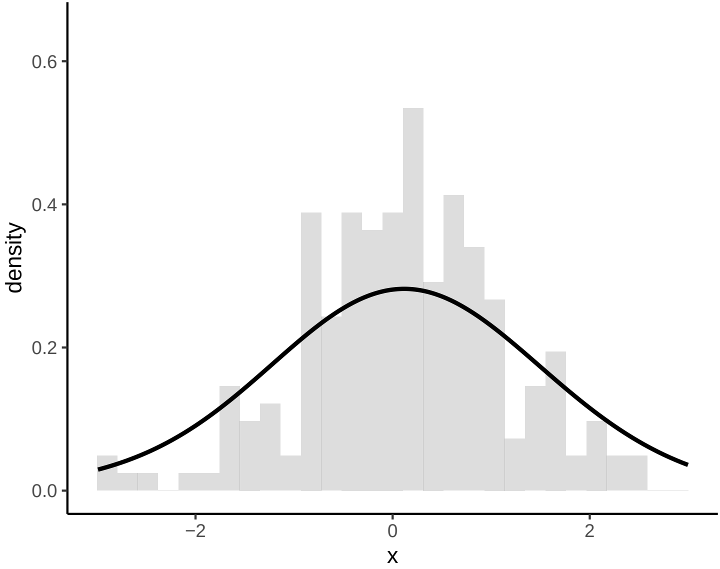


Visualize kernel estimator



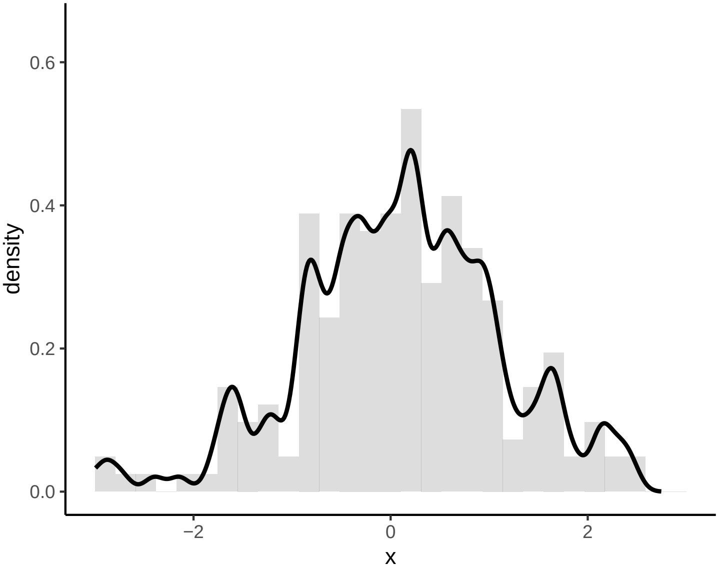
Visualize kernel estimator

small h large



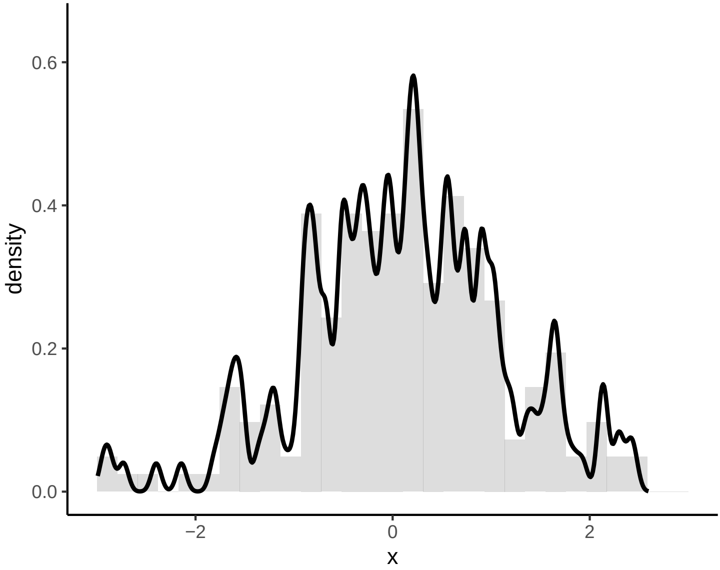
Visualize kernel estimator

small h large



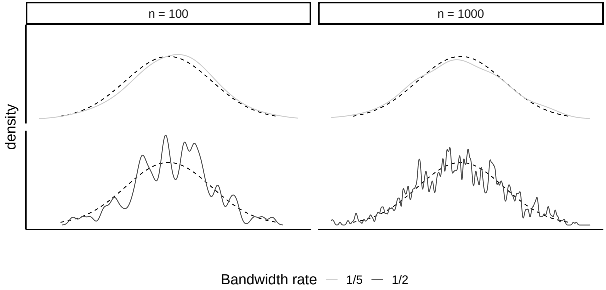
Visualize kernel estimator

small h large

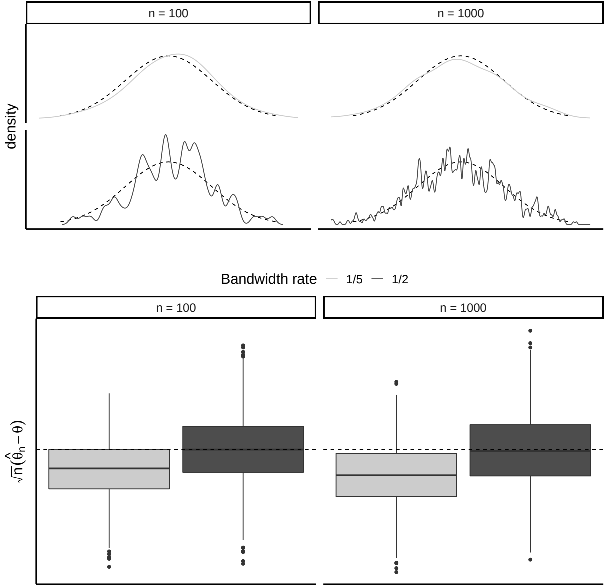


How does this work in practice?

How does this work in practice?



How does this work in practice?



What happened?

The bias-variance trade-off for the nuisance parameter f is

$$\mathrm{MSE}(\hat{f}_n) = C_1 h^4 + C_2(nh)^{-1} + \mathcal{O}(h^2) + \mathcal{O}(n^{-1}),$$

where $n \to \infty$ and $h \to 0$. This implies that the optimal value for the bandwidth h is $h \asymp n^{-1/5}$ [van der Vaart, 2000, chp. 24].

What happened?

The bias-variance trade-off for the nuisance parameter f is

$$\mathrm{MSE}(\hat{f}_n) = C_1 h^4 + C_2 (nh)^{-1} + \mathcal{O}(h^2) + \mathcal{O}(n^{-1}),$$

where $n \to \infty$ and $h \to 0$. This implies that the optimal value for the bandwidth h is $h \asymp n^{-1/5}$ [van der Vaart, 2000, chp. 24].

For the target parameter $\int_{-\infty}^{x} f(z)\,\mathrm{d}\lambda(z)$ the bias-variance "trade-off" becomes

$$\mathrm{MSE}(\hat{\theta}_n) = \mathcal{O}(n^{-1}) + \mathcal{O}(hn^{-1}) + \mathcal{O}(h^4).$$

The optimal value of h is now found by picking h as small as possible.

What happened?

The bias-variance trade-off for the nuisance parameter f is

$$\mathrm{MSE}(\hat{f}_n) = C_1 h^4 + C_2(nh)^{-1} + \mathcal{O}(h^2) + \mathcal{O}(n^{-1}),$$

where $n \to \infty$ and $h \to 0$. This implies that the optimal value for the bandwidth h is $h \asymp n^{-1/5}$ [van der Vaart, 2000, chp. 24].

For the target parameter $\int_{-\infty}^{x} f(z)\,\mathrm{d}\lambda(z)$ the bias-variance "trade-off" becomes

$$\mathrm{MSE}(\hat{\theta}_n) = \mathcal{O}(n^{-1}) + \mathcal{O}(hn^{-1}) + \mathcal{O}(h^4).$$

The optimal value of h is now found by picking h as small as possible.

Using $h = 0$ can be interpreted as just using the empirical distribution function $\hat{F}_n$, i.e.,

$$\int_{-\infty}^{x} \hat{f}_n(z)\,\mathrm{d}z = \hat{\mathbb{P}}_n\left[\int_{-\infty}^{x} k_h(X_i, z)\,\mathrm{d}z\right] \longrightarrow \hat{\mathbb{P}}_n[\mathbb{1}(X_i \leq x)] =: \hat{F}_n(x).$$

The bias-variance tradeoff revisited

Consider a statistical estimation problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \tilde{\Psi}(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$.

The bias-variance tradeoff revisited

Consider a statistical estimation problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \tilde{\Psi}(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(\tilde{\Psi}(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] \pm \mathrm{P}[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right\},
\end{aligned}$$

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

The bias-variance tradeoff revisited

Consider a statistical estimation problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \tilde{\Psi}(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(\tilde{\Psi}(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] \pm \mathrm{P}[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right\},
\end{aligned}$$

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)]$ determines the (main) variance –

The bias-variance tradeoff revisited

Consider a statistical estimation problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \tilde{\Psi}(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(\tilde{\Psi}(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] \pm \mathrm{P}[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right\},
\end{aligned}$$

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)]$ determines the (main) variance – often $\mathbb{G}_n[\varphi(O, \hat{\nu}_n)] = \mathbb{G}_n[\varphi(O, \nu)] + o_P(1)$

The bias-variance tradeoff revisited

Consider a statistical estimation problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \tilde{\Psi}(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(\tilde{\Psi}(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] \pm \mathrm{P}[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right\},
\end{aligned}$$

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)]$ determines the (main) variance – often
$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)] = \mathbb{G}_n[\varphi(O, \nu)] + o_P(1)$

$\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)$ is bias

The bias-variance tradeoff revisited

Consider a statistical estimation problem $(\mathcal{P}, \Psi)$ for which we can write $\Psi(\mathrm{P}) = \tilde{\Psi}(\mathrm{P}, \nu) = \mathrm{P}[\varphi(O, \nu)]$. We have

$$\begin{aligned}
\sqrt{n}\left(\hat{\theta}_n - \theta\right) &= \sqrt{n}\left(\tilde{\Psi}(\hat{\mathbb{P}}_n, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \sqrt{n}\left(\hat{\mathbb{P}}_n[\varphi(O, \hat{\nu}_n)] \pm \mathrm{P}[\varphi(O, \hat{\nu}_n)] - \mathrm{P}[\varphi(O, \nu)]\right) \\
&= \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)\right\},
\end{aligned}$$

with $\mathbb{G}_n := \sqrt{n}(\hat{\mathbb{P}}_n - \mathrm{P})$ the empirical process.

$\mathbb{G}_n[\varphi(O, \hat{\nu}_n)]$ determines the (main) variance – often $\mathbb{G}_n[\varphi(O, \hat{\nu}_n)] = \mathbb{G}_n[\varphi(O, \nu)] + o_P(1)$

$\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu)$ is bias $\rightarrow$ This is what ruined the naive plug-in strategy!

Finite- versus infinite-dimensional nuisance parameters

This phenomena only occurs when there are infinite-dimensional nuisance parameters (e.g., smooth functions, shape-constrained densities, etc.) – or more precisely, when we do not have parametric ($n^{-1/2}$) rate of convergence for the nuisance parameter estimator.

Finite- versus infinite-dimensional nuisance parameters

This phenomena only occurs when there are infinite-dimensional nuisance parameters (e.g., smooth functions, shape-constrained densities, etc.) – or more precisely, when we do not have parametric ($n^{-1/2}$) rate of convergence for the nuisance parameter estimator.

For finite-dimensional nuisance parameters ν (and under suitable regularity conditions), if

$$\sqrt{n}(\hat{\nu}_n - \nu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

then, by the delta method, also

$$\sqrt{n}(\Psi(\hat{\nu}_n) - \Psi(\nu)) \rightsquigarrow \mathcal{N}(0, \dot{\Psi}\sigma^2).$$

Finite- versus infinite-dimensional nuisance parameters

This phenomena only occurs when there are infinite-dimensional nuisance parameters (e.g., smooth functions, shape-constrained densities, etc.) – or more precisely, when we do not have parametric ($n^{-1/2}$) rate of convergence for the nuisance parameter estimator.

For finite-dimensional nuisance parameters ν (and under suitable regularity conditions), if

$$\sqrt{n}(\hat{\nu}_n - \nu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

then, by the delta method, also

$$\sqrt{n}(\Psi(\hat{\nu}_n) - \Psi(\nu)) \rightsquigarrow \mathcal{N}(0, \dot{\Psi}\sigma^2).$$

Even more, if $\hat{\nu}_n$ is efficient $\Psi(\hat{\nu}_n)$ will also be efficient [van der Vaart, 2000, Bickel et al., 2003].

Finite- versus infinite-dimensional nuisance parameters

This phenomena only occurs when there are infinite-dimensional nuisance parameters (e.g., smooth functions, shape-constrained densities, etc.) – or more precisely, when we do not have parametric ($n^{-1/2}$) rate of convergence for the nuisance parameter estimator.

For finite-dimensional nuisance parameters ν (and under suitable regularity conditions), if

$$\sqrt{n}(\hat{\nu}_n - \nu) \rightsquigarrow \mathcal{N}(0, \sigma^2)$$

then, by the delta method, also

$$\sqrt{n}(\Psi(\hat{\nu}_n) - \Psi(\nu)) \rightsquigarrow \mathcal{N}(0, \dot{\Psi}\sigma^2).$$

Even more, if $\hat{\nu}_n$ is efficient $\Psi(\hat{\nu}_n)$ will also be efficient [van der Vaart, 2000, Bickel et al., 2003].

When ν is infinite-dimensional this no longer holds $\rightarrow$ we have to do something different.

Functional derivatives (brief digression)

Functional derivatives and von Mises expansions are useful for analyzing and handling the issues we have encountered [Serfling, 1980].

Defining a functional derivative

What is a derivative?

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_P$ of the map Ψ at $P \in \mathcal{P}$, i.e.,

$$\left\| \Psi(P + \varepsilon_n g_n) - \Psi(P) - \dot{\Psi}_P(\varepsilon_n g_n) \right\| = \mathcal{O}(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_P$ of the map Ψ at $P \in \mathcal{P}$, i.e.,

$$\left\| \Psi(P + \varepsilon_n g_n) - \Psi(P) - \dot{\Psi}_P(\varepsilon_n g_n) \right\| = \mathcal{O}(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (and operators) Ψ.

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_P$ of the map Ψ at $P \in \mathcal{P}$, i.e.,

$$\left\| \Psi(P + \varepsilon_n g_n) - \Psi(P) - \dot{\Psi}_P(\varepsilon_n g_n) \right\| = o(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (and operators) Ψ.

- For which g_n should this hold? Along "lines", "paths", or "uniformly" (g_n fixed, converging, or bounded)?

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_P$ of the map Ψ at $P \in \mathcal{P}$, i.e.,

$$\left\| \Psi(P + \varepsilon_n g_n) - \Psi(P) - \dot{\Psi}_P(\varepsilon_n g_n) \right\| = o(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (and operators) Ψ.

- For which g_n should this hold? Along "lines", "paths", or "uniformly" (g_n fixed, converging, or bounded)?
- Which norm on $\mathcal{P}$ should we use?

Defining a functional derivative

What is a derivative?

A linear approximation $\dot{\Psi}_P$ of the map Ψ at $P \in \mathcal{P}$, i.e.,

$$\left\| \Psi(P + \varepsilon_n g_n) - \Psi(P) - \dot{\Psi}_P(\varepsilon_n g_n) \right\| = o(\varepsilon_n),$$

when $\varepsilon_n \to 0$.

This expression also makes sense for functionals (and operators) Ψ.

- For which g_n should this hold? Along "lines", "paths", or "uniformly" (g_n fixed, converging, or bounded)?
- Which norm on $\mathcal{P}$ should we use?
- Is there a natural space $\mathcal{M}$ in which to embed $\mathcal{P}$?

Gâteaux and Hadamard differentiability

The weakest kind of differentiability is *Gâteaux differentiability*. When $\Psi\colon \mathcal{P} \to \mathbb{R}$ the Gâteaux derivative $\dot{\Psi}_P$ is the directional derivative

$$\dot{\Psi}_P(g) = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon g). \tag{1}$$

Gâteaux and Hadamard differentiability

The weakest kind of differentiability is *Gâteaux differentiability*. When $\Psi\colon \mathcal{P} \to \mathbb{R}$ the Gâteaux derivative $\dot{\Psi}_P$ is the directional derivative

$$\dot{\Psi}_P(g) = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon g). \tag{1}$$

A stronger type is *Hadamard differentiability* which demands that the linear approximation is valid along any converging sequence $g_n \to g$.

Gâteaux and Hadamard differentiability

The weakest kind of differentiability is *Gâteaux differentiability*. When $\Psi\colon \mathcal{P} \to \mathbb{R}$ the Gâteaux derivative $\dot{\Psi}_P$ is the directional derivative

$$\dot{\Psi}_P(g) = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon g). \tag{1}$$

A stronger type is *Hadamard differentiability* which demands that the linear approximation is valid along any converging sequence $g_n \to g$. A Hadamard differentiable map is also called *compactly differentiable*, because this requirement is equivalent to a uniformly valid linear approximation over compact subsets [Reeds, 1976].

Gâteaux and Hadamard differentiability

The weakest kind of differentiability is *Gâteaux differentiability*. When $\Psi\colon \mathcal{P} \to \mathbb{R}$ the Gâteaux derivative $\dot{\Psi}_P$ is the directional derivative

$$\dot{\Psi}_P(g) = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon g). \tag{1}$$

A stronger type is *Hadamard differentiability* which demands that the linear approximation is valid along any converging sequence $g_n \to g$. A Hadamard differentiable map is also called *compactly differentiable*, because this requirement is equivalent to a uniformly valid linear approximation over compact subsets [Reeds, 1976]. This makes Hadamard differentiability well-suited for statistical problems [Gill et al., 1989, van der Vaart, 2000].

Gâteaux and Hadamard differentiability

The weakest kind of differentiability is *Gâteaux differentiability*. When $\Psi\colon \mathcal{P} \to \mathbb{R}$ the Gâteaux derivative $\dot{\Psi}_P$ is the directional derivative

$$\dot{\Psi}_P(g) = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon g). \tag{1}$$

A stronger type is *Hadamard differentiability* which demands that the linear approximation is valid along any converging sequence $g_n \to g$. A Hadamard differentiable map is also called *compactly differentiable*, because this requirement is equivalent to a uniformly valid linear approximation over compact subsets [Reeds, 1976]. This makes Hadamard differentiability well-suited for statistical problems [Gill et al., 1989, van der Vaart, 2000].

If Ψ is Hadamard differentiability, Ψ is also Gâteaux differentiability, and in that case the Hadamard and Gâteaux derivative are identical.

The efficient influence function

(bag to the statistical estimation problem)

Hadamard differentiability and the canonical gradient

The "gradient" of $\Psi \colon \mathcal{P} \to \mathbb{R}$ is called the *canonical gradient* or *efficient influence function* of a statistical estimation problem $(\mathcal{P}, \Psi)$, and it is a fundamental object for semi-parametric efficiency theory – we will see why in a moment. More formally we define (following Bickel et al. [1993]):

Hadamard differentiability and the canonical gradient

The "gradient" of $\Psi\colon \mathcal{P} \to \mathbb{R}$ is called the *canonical gradient* or *efficient influence function* of a statistical estimation problem $(\mathcal{P}, \Psi)$, and it is a fundamental object for semi-parametric efficiency theory – we will see why in a moment. More formally we define (following Bickel et al. [1993]):

Definition (Tangent space)

The *tangent space* $\dot{\mathcal{P}}_P$ for the model $\mathcal{P}$ at $P \in \mathcal{P}$ is the (closed linear span of the) collection of (Hadamard) derivatives $\dot{P}_\varepsilon$ for all one-dimensional parametric submodel $P_\varepsilon \subset \mathcal{P}$ with $P_0 = P$.

Hadamard differentiability and the canonical gradient

The "gradient" of $\Psi \colon \mathcal{P} \to \mathbb{R}$ is called the *canonical gradient* or *efficient influence function* of a statistical estimation problem $(\mathcal{P}, \Psi)$, and it is a fundamental object for semi-parametric efficiency theory – we will see why in a moment. More formally we define (following Bickel et al. [1993]):

Definition (Tangent space)

The *tangent space* $\dot{\mathcal{P}}_P$ for the model $\mathcal{P}$ at $P \in \mathcal{P}$ is the (closed linear span of the) collection of (Hadamard) derivatives $\dot{P}_\varepsilon$ for all one-dimensional parametric submodel $P_\varepsilon \subset \mathcal{P}$ with $P_0 = P$.

Definition (Canonical gradient)

If $\Psi \colon \mathcal{P} \to \mathbb{R}$ is Hadamard differentiable at P tangential to the tangent space $\dot{\mathcal{P}}_P$,[1] we refer to the Hadamard derivative $\dot{\Psi}_P$ as the *canonical gradient of the statistical estimation problem*.

Hadamard differentiability and the canonical gradient

The "gradient" of $\Psi\colon \mathcal{P} \to \mathbb{R}$ is called the *canonical gradient* or *efficient influence function* of a statistical estimation problem $(\mathcal{P}, \Psi)$, and it is a fundamental object for semi-parametric efficiency theory – we will see why in a moment. More formally we define (following Bickel et al. [1993]):

Definition (Tangent space)

The *tangent space* $\dot{\mathcal{P}}_P$ for the model $\mathcal{P}$ at $P \in \mathcal{P}$ is the (closed linear span of the) collection of (Hadamard) derivatives $\dot{P}_\varepsilon$ for all one-dimensional parametric submodel $P_\varepsilon \subset \mathcal{P}$ with $P_0 = P$.

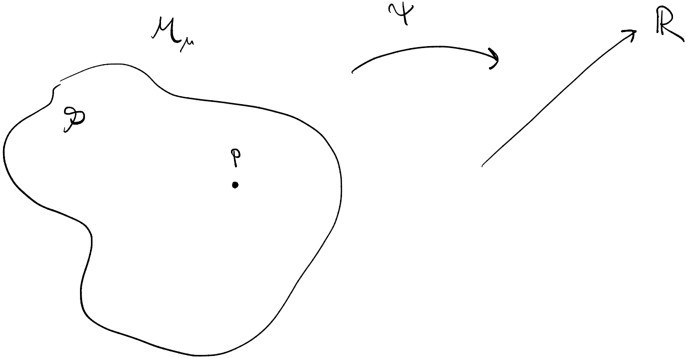
Definition (Canonical gradient)

If $\Psi\colon \mathcal{P} \to \mathbb{R}$ is Hadamard differentiable at P tangential to the tangent space $\dot{\mathcal{P}}_P$,[1] we refer to the Hadamard derivative $\dot{\Psi}_P$ as the *canonical gradient of the statistical estimation problem*.

[1] This just means that the Ψ and $\dot{\Psi}_P$ need only be defined on the subsets $\mathcal{P} \subset \mathcal{M}$ and $\dot{\mathcal{P}}_P \subset \mathcal{M}$, respectively.

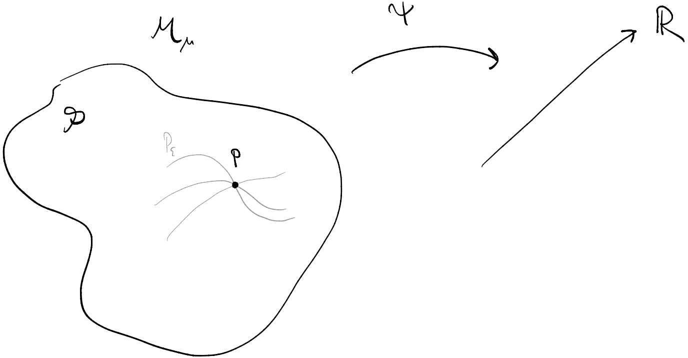
Visualization

Think of the gradient of a function defined on manifold (e.g., a surface).



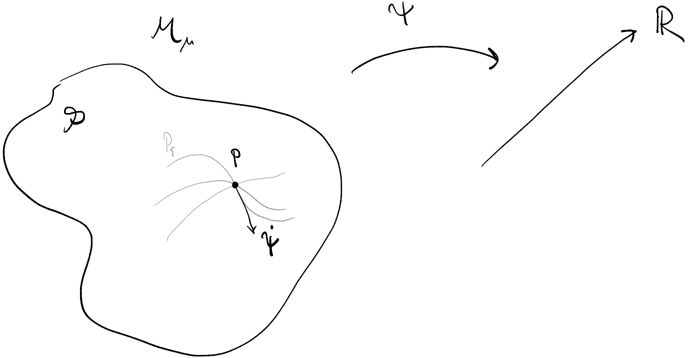
Visualization

Think of the gradient of a function defined on manifold (e.g., a surface).



Visualization

Think of the gradient of a function defined on manifold (e.g., a surface).



Alternative characterization and connection to estimation

Alternative characterization and connection to estimation

The tangent space can be represented as

$$\dot{\mathcal{P}}_P = \overline{\mathrm{span}}\{\dot{\ell}_0\} \subset \mathcal{L}_0^2(P), \quad \text{where} \quad \dot{\ell}_0 = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(p_\varepsilon),$$

and $\mathcal{L}_0^2(P)$ is the space of all functions $g \in \mathcal{L}^2(P)$ and $P[g] = 0$.

Alternative characterization and connection to estimation

The tangent space can be represented as

$$\dot{\mathcal{P}}_P = \overline{\mathrm{span}}\{\dot{\ell}_0\} \subset \mathcal{L}_0^2(P), \quad \text{where} \quad \dot{\ell}_0 = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(p_\varepsilon),$$

and $\mathcal{L}_0^2(P)$ is the space of all functions $g \in \mathcal{L}^2(P)$ and $P[g] = 0$.

By Riesz' representation theorem, and the (Hadamard) chain rule, this implies that the canonical gradient (efficient influence function) $\dot{\Psi}_P$ can be identified with an element $\varphi_P \in \dot{\mathcal{P}}_P \subset \mathcal{L}_0^2(P)$ such that

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P_\varepsilon) = \dot{\Psi}_P\left(\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} P_\varepsilon\right) = \dot{\Psi}_P(\dot{\ell}_0) = \langle \varphi_P, \dot{\ell}_0 \rangle_{\mathrm{P}}, \tag{2}$$

holds for any differentiable submodel P_ε with score function $\dot{\ell}_0$, where $= \langle \varphi_P, \dot{\ell}_0 \rangle_{\mathrm{P}} = \int \varphi_P(o)\dot{\ell}_0(o)\, \mathrm{d}P(o)$.

Alternative characterization and connection to estimation

The tangent space can be represented as

$$\dot{\mathcal{P}}_P = \overline{\mathrm{span}}\{\dot{\ell}_0\} \subset \mathcal{L}_0^2(P), \quad \text{where} \quad \dot{\ell}_0 = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(p_\varepsilon),$$

and $\mathcal{L}_0^2(P)$ is the space of all functions $g \in \mathcal{L}^2(P)$ and $P[g] = 0$.

By Riesz' representation theorem, and the (Hadamard) chain rule, this implies that the canonical gradient (efficient influence function) $\dot{\Psi}_P$ can be identified with an element $\varphi_P \in \dot{\mathcal{P}}_P \subset \mathcal{L}_0^2(P)$ such that

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P_\varepsilon) = \dot{\Psi}_P\left(\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} P_\varepsilon\right) = \dot{\Psi}_P(\dot{\ell}_0) = \langle \varphi_P, \dot{\ell}_0 \rangle_{\mathrm{P}}, \tag{2}$$

holds for any differentiable submodel P_ε with score function $\dot{\ell}_0$, where $= \langle \varphi_P, \dot{\ell}_0 \rangle_{\mathrm{P}} = \int \varphi_P(o)\dot{\ell}_0(o)\,\mathrm{d}P(o)$.

Note the function φ_P is determined *both* by the map Ψ *and* the model $\mathcal{P}$: Many functions $\tilde{\varphi}_P \in \mathcal{L}_0^2(P)$ might fulfill (2), but only one will be contained in the tangent space.

Alternative characterization and connection to estimation

The tangent space can be represented as

$$\dot{\mathcal{P}}_P = \overline{\mathrm{span}}\{\dot{\ell}_0\} \subset \mathcal{L}_0^2(P), \quad \text{where} \quad \dot{\ell}_0 = \frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \log(p_\varepsilon),$$

and $\mathcal{L}_0^2(P)$ is the space of all functions $g \in \mathcal{L}^2(P)$ and $P[g] = 0$.

By Riesz' representation theorem, and the (Hadamard) chain rule, this implies that the canonical gradient (efficient influence function) $\dot{\Psi}_P$ can be identified with an element $\varphi_P \in \dot{\mathcal{P}}_P \subset \mathcal{L}_0^2(P)$ such that

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P_\varepsilon) = \dot{\Psi}_P\left(\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} P_\varepsilon\right) = \dot{\Psi}_P(\dot{\ell}_0) = \langle \varphi_P, \dot{\ell}_0 \rangle_{\mathrm{P}}, \tag{2}$$

holds for any differentiable submodel P_ε with score function $\dot{\ell}_0$, where $= \langle \varphi_P, \dot{\ell}_0 \rangle_{\mathrm{P}} = \int \varphi_P(o)\dot{\ell}_0(o)\,\mathrm{d}P(o)$.

Note the function φ_P is determined *both* by the map Ψ *and* the model $\mathcal{P}$: Many functions $\tilde{\varphi}_P \in \mathcal{L}_0^2(P)$ might fulfill (2), but only one will be contained in the tangent space.

In the important special case of a *fully non-parametric model* $\mathcal{P}$, $\dot{\mathcal{P}}_P = \mathcal{L}_0^2(P)$ and thus (2) alone uniquely identifies the function φ_P.

Efficient estimation and the canonical gradient

Efficient estimation and the canonical gradient

Definition (RAL estimators)

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

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

Efficient estimation and the canonical gradient

Definition (RAL estimators)

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

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

Theorem (Efficient influence function)

The RAL estimator with lowest possible asymptotic variance has the canonical gradient as its influence function.

Efficient estimation and the canonical gradient

Definition (RAL estimators)

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

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

Theorem (Efficient influence function)

The RAL estimator with lowest possible asymptotic variance has the canonical gradient as its influence function.

Information bound

The *information bound* for estimating Ψ in the model $\mathcal{P}$ is

$$\mathcal{I}(\mathcal{P}, \Psi) := \inf_{P_\varepsilon} \{\mathcal{I}(P_\varepsilon, \Psi)\}, \quad \text{with} \quad \mathcal{I}(P_\varepsilon, \Psi) := \frac{P[\dot{\ell}_0^2]}{(\partial_0 \Psi(P_\varepsilon))^2}.$$

It holds that $\mathcal{I}(\mathcal{P}, \Psi)^{-1} = P[\varphi^2]$.

Debiasing and the canonical gradient

Recall the statistical estimation problem $\Psi(P) = \tilde{\Psi}(P, \nu) = P[\varphi(O, \nu)]$ and the decomposition

$$\sqrt{n}(\hat{\theta}_n - \theta) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(P, \hat{\nu}_n) - \tilde{\Psi}(P, \nu)\right\}.$$

Debiasing and the canonical gradient

Recall the statistical estimation problem $\Psi(P) = \tilde{\Psi}(P, \nu) = P[\varphi(O, \nu)]$ and the decomposition

$$\sqrt{n}(\hat{\theta}_n - \theta) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(P, \hat{\nu}_n) - \tilde{\Psi}(P, \nu)\right\}.$$

If φ is the canonical gradient then the Gâteaux derivative of $\nu \mapsto \tilde{\Psi}(P, \nu)$ is zero.[2]

Debiasing and the canonical gradient

Recall the statistical estimation problem $\Psi(P) = \tilde{\Psi}(P, \nu) = P[\varphi(O, \nu)]$ and the decomposition

$$\sqrt{n}(\hat{\theta}_n - \theta) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(P, \hat{\nu}_n) - \tilde{\Psi}(P, \nu)\right\}.$$

If φ is the canonical gradient then the Gâteaux derivative of $\nu \mapsto \tilde{\Psi}(P, \nu)$ is zero.[2] Hence a von Mises expansion suggests that

$$\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu) = 0 + \mathcal{O}_P(\|\hat{\nu}_n - \nu\|^2).$$

Debiasing and the canonical gradient

Recall the statistical estimation problem $\Psi(P) = \tilde{\Psi}(P, \nu) = P[\varphi(O, \nu)]$ and the decomposition

$$\sqrt{n}(\hat{\theta}_n - \theta) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(P, \hat{\nu}_n) - \tilde{\Psi}(P, \nu)\right\}.$$

If φ is the canonical gradient then the Gâteaux derivative of $\nu \mapsto \tilde{\Psi}(P, \nu)$ is zero.[2] Hence a von Mises expansion suggests that

$$\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu) = 0 + \mathcal{O}_P(\|\hat{\nu}_n - \nu\|^2).$$

In particular, we can obtain $n^{-1/2}$ convergence of $\hat{\theta}_n$ even though $\hat{\nu}_n$ only converges at rate $n^{-1/4}$.

Debiasing and the canonical gradient

Recall the statistical estimation problem $\Psi(P) = \tilde{\Psi}(P, \nu) = P[\varphi(O, \nu)]$ and the decomposition

$$\sqrt{n}(\hat{\theta}_n - \theta) = \mathbb{G}_n[\varphi(O, \hat{\nu}_n)] + \sqrt{n}\left\{\tilde{\Psi}(P, \hat{\nu}_n) - \tilde{\Psi}(P, \nu)\right\}.$$

If φ is the canonical gradient then the Gâteaux derivative of $\nu \mapsto \tilde{\Psi}(P, \nu)$ is zero.[2] Hence a von Mises expansion suggests that

$$\tilde{\Psi}(\mathrm{P}, \hat{\nu}_n) - \tilde{\Psi}(\mathrm{P}, \nu) = 0 + \mathcal{O}_P(\|\hat{\nu}_n - \nu\|^2).$$

In particular, we can obtain $n^{-1/2}$ convergence of $\hat{\theta}_n$ even though $\hat{\nu}_n$ only converges at rate $n^{-1/4}$.

[2] This property is referred to as *Neyman orthogonality*, and is a central component of "debiased machine learning"; see Chernozhukov et al. [2018] for details.

Summary

- We have seen that the "naive" plug-in strategy can lead to estimators of the target parameter with poor performance.

Summary

- We have seen that the "naive" plug-in strategy can lead to estimators of the target parameter with poor performance.
- Instead we should aim at constructing estimators that have the canonical gradient (efficient influence function) as their influence function.

Summary

- We have seen that the "naive" plug-in strategy can lead to estimators of the target parameter with poor performance.
- Instead we should aim at constructing estimators that have the canonical gradient (efficient influence function) as their influence function.
- This provides estimators with a bias-variance trade-off that is optimized for the *target parameter* instead of the nuisance parameter, because such estimators are *efficient* and have *vanishing first order bias*, asymptotically.

Exercises

How do we find the efficient influence function for a statistical estimation problem?

Introduction: Finding the canonical gradient for a given Ψ

We can (informally and heuristically[3]) find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O, where δ_O is the Dirac measure in the point O:

[3] To make the arguments more precise, see Ichimura and Newey [2015].

Introduction: Finding the canonical gradient for a given Ψ

We can (informally and heuristically[3]) find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O, where δ_O is the Dirac measure in the point O:

For any $h \in \mathbb{R}_+$, define the sub-model $P^h_\varepsilon := P + \varepsilon K^h_O$, where $K^h_O \to \delta_O$, $h \to 0$.

[3] To make the arguments more precise, see Ichimura and Newey [2015].

Introduction: Finding the canonical gradient for a given Ψ

We can (informally and heuristically[3]) find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O, where δ_O is the Dirac measure in the point O:

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

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

[3] To make the arguments more precise, see Ichimura and Newey [2015].

Introduction: Finding the canonical gradient for a given Ψ

We can (informally and heuristically[3]) find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O, where δ_O is the Dirac measure in the point O:

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

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

Thus, if Ψ is Hadamard differentiable, we can use (2) to write

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

[3] To make the arguments more precise, see Ichimura and Newey [2015].

Introduction: Finding the canonical gradient for a given Ψ

We can (informally and heuristically[3]) find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O, where δ_O is the Dirac measure in the point O:

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

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

Thus, if Ψ is Hadamard differentiable, we can use (2) to write

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

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

$$\frac{\partial}{\partial\varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon\delta_O) = \varphi_P(O).^4$$

[3] To make the arguments more precise, see Ichimura and Newey [2015].

Introduction: Finding the canonical gradient for a given Ψ

We can (informally and heuristically[3]) find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O, where δ_O is the Dirac measure in the point O:

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

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

Thus, if Ψ is Hadamard differentiable, we can use (2) to write

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

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

$$\frac{\partial}{\partial \varepsilon}\Big|_{\varepsilon=0} \Psi(P + \varepsilon \delta_O) = \varphi_P(O).^4$$

[3] To make the arguments more precise, see Ichimura and Newey [2015].

[4] One should then verify that the found the candidate φ_P fulfills (2) for *all* parametric sub-model. In addition, if we impose restrictions on $\mathcal{P}$ such that $\dot{\mathcal{P}}_P$ is a proper subset of $\mathcal{L}_0^2(P)$ we also need to check that $\varphi_P \in \dot{\mathcal{P}}_P$.

Exercise 1 – efficient influence function of the toy example

Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_X for $\Psi(P) = F_P(x)$ where

$$F_P(x) = P(X \leq x) = \int_{-\infty}^{x} \mathrm{d}P(z)$$

is the cumulative distribution function at x.

Exercise 1 – efficient influence function of the toy example

Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_X for $\Psi(P) = F_P(x)$ where

$$F_P(x) = P(X \leq x) = \int_{-\infty}^{x} \mathrm{d}P(z)$$

is the cumulative distribution function at x.

Strategy

1. Write the target parameter $\Psi(P)$ explicitly as a functional of P.

Exercise 1 – efficient influence function of the toy example

Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_X for $\Psi(P) = F_P(x)$ where

$$F_P(x) = P(X \leq x) = \int_{-\infty}^{x} \mathrm{d}P(z)$$

is the cumulative distribution function at x.

Strategy

1. Write the target parameter $\Psi(P)$ explicitly as a functional of P.
2. Calculate the ordinary derivative of

$$\varepsilon \longmapsto \Psi(P + \varepsilon\delta_X).$$

and evaluate at $\varepsilon = 0$.

Exercise 2 – the treatment-specific mean response

Let $O = (Y, A, X) \sim P$. Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O with

$$\Psi(P) = \mathbb{E}_P\left[f_P(1, X)\right] = \int f_P(1, x)\,\mathrm{d}\mu_P(x),$$

with $f_P(a, x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right]$ and μ_P the marginal distribution of X.

Exercise 2 – the treatment-specific mean response

Let $O = (Y, A, X) \sim P$. Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O with

$$\Psi(P) = \mathbb{E}_P\left[f_P(1, X)\right] = \int f_P(1, x)\,\mathrm{d}\mu_P(x),$$

with $f_P(a, x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right]$ and μ_P the marginal distribution of X.

Strategy

1. Write the target parameter $\Psi(P)$ explicitly as a functional of P.

Exercise 2 – the treatment-specific mean response

Let $O = (Y, A, X) \sim P$. Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O with

$$\Psi(P) = \mathbb{E}_P\left[f_P(1, X)\right] = \int f_P(1, x)\,\mathrm{d}\mu_P(x),$$

with $f_P(a, x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right]$ and μ_P the marginal distribution of X.

Strategy

1. Write the target parameter $\Psi(P)$ explicitly as a functional of P.
   1.1 Write f_P and μ_P explicitly as functions of P.

Exercise 2 – the treatment-specific mean response

Let $O = (Y, A, X) \sim P$. Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O with

$$\Psi(P) = \mathbb{E}_P\left[f_P(1, X)\right] = \int f_P(1, x)\,\mathrm{d}\mu_P(x),$$

with $f_P(a, x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right]$ and μ_P the marginal distribution of X.

Strategy

1. Write the target parameter $\Psi(P)$ explicitly as a functional of P.
   - 1.1 Write f_P and μ_P explicitly as functions of P.
   - 1.2 Insert the these into the definition of Ψ above.

Exercise 2 – the treatment-specific mean response

Let $O = (Y, A, X) \sim P$. Find a candidate for the efficient influence function by calculating the Gâteaux derivative of Ψ at δ_O with

$$\Psi(P) = \mathbb{E}_P\left[f_P(1, X)\right] = \int f_P(1, x)\,\mathrm{d}\mu_P(x),$$

with $f_P(a, x) = \mathbb{E}_P\left[Y \mid A = a, X = x\right]$ and μ_P the marginal distribution of X.

Strategy

1. Write the target parameter $\Psi(P)$ explicitly as a functional of P.
   - 1.1 Write f_P and μ_P explicitly as functions of P.
   - 1.2 Insert the these into the definition of Ψ above.
2. Calculate the ordinary derivative of

$$\varepsilon \longmapsto \Psi(P + \varepsilon\delta_O).$$

   and evaluate at $\varepsilon = 0$.

Exercise 2 – first step

Express $\Psi(P) = \int_{\mathcal{X}} f(1, x)\mu_P(\mathrm{d}x)$ explicitly as a functional of P:

Exercise 2 – first step

Express $\Psi(P) = \int_{\mathcal{X}} f(1, x)\mu_P(\mathrm{d}x)$ explicitly as a functional of P: We can write

$$\mu_P(x) = \sum_{a=0}^{1} \int_{\mathcal{Y}} P(\mathrm{d}y, a, x),$$

Exercise 2 – first step

Express $\Psi(P) = \int_{\mathcal{X}} f(1,x)\mu_P(\mathrm{d}x)$ explicitly as a functional of P: We can write

$$\mu_P(x) = \sum_{a=0}^{1} \int_{\mathcal{Y}} P(\mathrm{d}y, a, x),$$

and

$$f(1,x) = \int_{\mathcal{Y}} y \frac{P(\mathrm{d}y, 1, x)}{\int_{\mathcal{Y}} P(\mathrm{d}y, 1, x)},$$

Exercise 2 – first step

Express $\Psi(P) = \int_{\mathcal{X}} f(1,x)\mu_P(\mathrm{d}x)$ explicitly as a functional of P: We can write

$$\mu_P(x) = \sum_{a=0}^{1} \int_{\mathcal{Y}} P(\mathrm{d}y, a, x),$$

and

$$f(1,x) = \int_{\mathcal{Y}} y \frac{P(\mathrm{d}y, 1, x)}{\int_{\mathcal{Y}} P(\mathrm{d}y, 1, x)},$$

so

$$\Psi(P_\varepsilon) = \int_{\mathcal{X}} \left\{ \frac{\int_{\mathcal{Y}} y \, P_\varepsilon(\mathrm{d}y, 1, x)}{\int_{\mathcal{Y}} P_\varepsilon(\mathrm{d}y, 1, x)} \left(\sum_{a=0}^{1} \int_{\mathcal{Y}} P_\varepsilon(\mathrm{d}y, a, \mathrm{d}x) \right) \right\}.$$

References

P. J. Bickel, C. A. Klaassen, P. J. Bickel, Y. Ritov, J. Klaassen, J. A. Wellner, and Y. Ritov. *Efficient and adaptive estimation for semiparametric models*, volume 4. Johns Hopkins University Press Baltimore, 1993.

P. J. Bickel, Y. Ritov, et al. Nonparametric estimators which can be "plugged-in". *The Annals of Statistics*, 31(4):1033–1053, 2003.

V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

R. D. Gill, J. A. Wellner, and J. Præstgaard. Non-and semi-parametric maximum likelihood estimators and the von mises method (part 1)[with discussion and reply]. *Scandinavian Journal of Statistics*, pages 97–128, 1989.

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

J. Reeds. On the definition of von mises functionals. *Ph. D. dissertation, Harvard Univ.*, 1976.

R. J. Serfling. *Approximation theorems of mathematical statistics*, volume 162. John Wiley & Sons, 1980.

A. W. van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.