BIOST/STAT 578B
Modern inference in infinite-dimensional models

# Chapter 3: Overview of efficiency theory

Marco Carone
Department of Biostatistics
School of Public Health, University of Washington

Winter 2015

# Contents of this chapter

- Review of parametric efficiency theory
- Concept of tangent space
- Pathwise differentiability of statistical parameters and gradients
- Characterizing the set of influence functions
- Efficiency bounds and the efficient influence function (EIF)
- Impact and role of nuisance modeling in determining the EIF

# Review of parametric efficiency theory

In this chapter, we will give an exposition of **efficiency theory for estimating a finite-dimensional parameter in general models**. This will then guide our efforts to construct optimal estimators.

Suppose that $\mathcal{M} = \{P_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}$ is a regular parametric model and that all members of $\mathcal{M}$ are absolutely continuous relative to Lebesgue measure.

We observe $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_{\theta_0}$ with $\theta_0 \in \Theta$ and are interested in estimating the unknown scalar $\tau_0 := \tau(\theta_0)$.

Recall that the **Fisher information** for $\theta$ is defined as

$$\mathcal{I}(\theta) := P_\theta \left( \frac{\partial}{\partial \theta} \log p_\theta \right)^2 .$$

It is a measure of the curvature of the loglikelihood – the curvier, the more information there is about the parameter!

## Review of parametric efficiency theory

Hájek's convolution theorem states that

- if $\mathcal{M}$ is a sufficiently smooth model,
- if the information $\mathcal{I}(\theta_0) > 0$, and
- if $\tau_n$ is a regular estimator of $\tau_0$ with $n^{1/2}(\tau_n - \tau_0) \xrightarrow{d} Z$ ,

then $Z \overset{d}{=} Z_0 + \Delta_0$ for two independent variates $Z_0 \sim N(0, v_0)$ and $\Delta_0$, where

$$v_0(\mathcal{M}) = v_0 := \left( \frac{\partial}{\partial\theta}\tau(\theta)\bigg|_{\theta=\theta_0} \right)^2 \frac{1}{\mathcal{I}(\theta_0)} \ .$$

Based on the above, we have that

**the asymptotic variance of any regular estimator is no smaller than $v_0$.**

A regular estimator which achieves this bound asymptotically is said to be asymptotically efficient.

# Review of parametric efficiency theory

Suppose $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_0 \in \mathcal{M}$ and consider the parameter $\Psi : \mathcal{M} \to \mathbb{R}$. We wish to estimate $\psi_0 := \Psi(P_0)$ from the available data.

**If $\mathcal{M}$ is infinite-dimensional, what is the corresponding efficiency theory?**

A promising starting point:

**Estimation of $\psi_0$ in $\mathcal{M}$ should be no easier than in any (parametric) submodel through $P_0$.**

Let $\psi_n$ be a regular estimator of $\psi_0$ such that $n^{1/2}(\psi_n - \psi_0) \xrightarrow{d} Z$, and write $\sigma_0^2 := \mathrm{var}_{P_0}(Z) < +\infty$.

For any given $P \in \mathcal{M}$, denote by $\mathcal{S}_0(P)$ the set of all regular one-dimensional parametric submodels of $\mathcal{M}$ parametrized to go through $P$ at the origin.

## Review of parametric efficiency theory

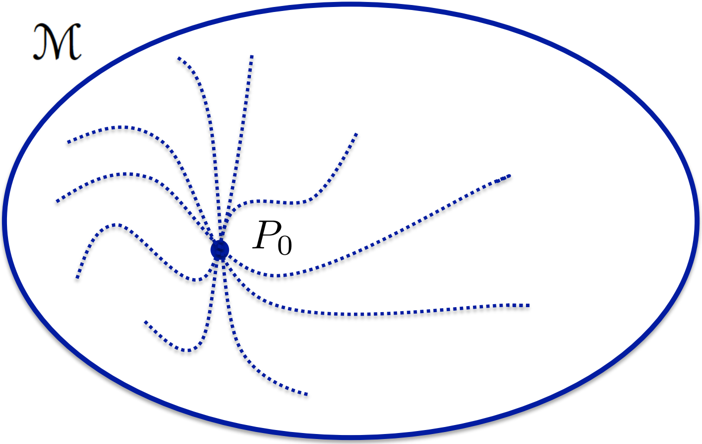
Suppose $\mathcal{H}$ is an index set for $\mathcal{S}_0(P_0)$. Then, for each $h \in \mathcal{H}$, we have that $\mathcal{M}_h = \{P_{\theta,h} : \theta \in \Theta\} \in \mathcal{S}_0(P_0)$, and furthermore, $\mathcal{M} = \cup_{h\in\mathcal{H}}\mathcal{M}_h$.

Since estimating $\psi_0$ over $\mathcal{M}$ is certainly no easier than over any possible $\mathcal{M}_h$, we can write that

$$\sigma_0^2 \;\geq\; \sup_{h\in\mathcal{H}} v_0(\mathcal{M}_h) \;\geq\; \sup_{h\in\mathcal{H}} \frac{\left(\frac{\partial}{\partial\theta}\Psi(P_{\theta,h})\big|_{\theta=0}\right)^2}{\mathcal{I}_{\mathcal{M}_h}(0)}\;,$$

where $\mathcal{I}_{\mathcal{M}_h}(0) := \; P_{\theta,h}\left(\frac{\partial}{\partial\theta}\log p_{\theta,h}\right)^2\Big|_{\theta=0}$ with $p_{\theta,h}$ denoting the density of $P_{\theta,h}$ is the Fisher information for estimating $\theta_0 = 0$ in the submodel $\mathcal{M}_h$.

## Review of parametric efficiency theory



# Review of parametric efficiency theory

Several questions naturally arise. . .

1. Do we really need to account for the whole index set $\mathcal{H}$?
   - Only the local behavior of $P_{\theta,h}$ around $\theta = 0$ seems to matter.
   - Could we index (equivalence classes of) submodels by their score at $\theta = 0$?
2. Do we have any grasp on the numerator and denominator?
   - This requires some "differentiability" of the path $\mathcal{M}_h$ and of $\Psi$ over paths.
3. Can the resulting maximization problem be performed explicitly?
   - Beyond this, is the resulting bound attainable?

# Concept of tangent space

How can we describe the **local behavior of the path $\mathcal{M}_h$ around $\theta = 0$?**

For simplicity, suppose all members of $\mathcal{M}$ are dominated by the same measure $\mu$. Let $\mathcal{M}_0 := \{p_\theta : \theta \in \Theta\} \in \mathcal{S}_0(P_0)$, and denote by $p_\theta$ the density of $P_\theta$ relative to $\mu$.

If $p_\theta$ is smooth enough in $\theta$ around $\theta = 0$, we might expect that

$$\frac{p_\theta(o)}{p_0(o)} = 1 + \theta g(o) + \theta r_\theta(o) \qquad (\star)$$

with $g(o) := \frac{\partial}{\partial\theta} p_\theta(o)\big|_{\theta=0} / p_0(o)$ and $r_\theta \to 0$ in an appropriate sense.

# Concept of tangent space

$$\frac{p_\theta(o)}{p_0(o)} = 1 + \theta g(o) + \theta r_\theta(o)$$

A few observations:

- $g$ is simply the score of $\theta$ at $\theta = 0$ in $\mathcal{M}_0$;
- $g$ completely determines the (first-order) local behavior of $p_\theta$ around $\theta = 0$, and is the 'direction' from which $P_\theta$ approaches $P_0$ as $\theta \to 0$;
- formally, we refer to *differentiability in quadratic mean*, i.e., there exists some 'score' $g$ such that

$$\int \left( \frac{\sqrt{p_\theta} - \sqrt{p_0}}{\theta} - \frac{1}{2} g \sqrt{p_0} \right)^2 d\mu \to 0 \ .$$

# Concept of tangent space

In order to understand local deviations from $p_0$ in $\mathcal{M}$, it becomes clear that we need to enumerate all possible $g$. This leads us to the following concept.

For $P \in \mathcal{M}$, denote by $L_2^0(P)$ the collection of all real-valued functions $f$ defined on the support of $P$ and such that $Pf = 0$ and $Pf^2 < +\infty$.
If we endow $L_2^0(P)$ with the so-called *covariance inner product*

$$(f_1, f_2) \mapsto \langle f_1, f_2 \rangle_P := P(f_1 f_2) \ ,$$

it is easy to verify that $L_2^0(P)$ is a Hilbert space.

The **tangent set** of $\mathcal{M}$ at $P$ is the set of elements $g \in L_2^0(P)$ arising in $(\star)$ for some submodel in $\mathcal{S}_0(P)$. The closure of its linear span is called the **tangent space** of $\mathcal{M}$ at $P$ and will be denoted by $T_{\mathcal{M}}(P) \subseteq L_2^0(P)$.

## Concept of tangent space

**An important case: a nonparametric model**

Suppose $\mathcal{M}_*$ consists of all $d$-variate probability distributions dominated by $\mu$. Then, it follows that $T_{\mathcal{M}_*}(P) = L_2^0(P)$.

To see this, take any $h \in L_2^0(P)$ and define pointwise the density function

$$p_\theta(o) := c(\theta)^{-1}\,\text{expit}[2\theta h(o)]p(o)$$

relative to $\mu$, where we have set $c(\theta) := \int \text{expit}[2\theta h(o)]dP(o)$.

If $P_\theta$ is the distribution corresponding to $p_\theta$, then $\mathcal{M}_h := \{P_\theta : \theta \in \mathbb{R}\}$ is an element of $\mathcal{S}_0(P)$ with score for $\theta$ at $\theta = 0$ equal to $h$. So, $L_2^0(P) \subseteq T_{\mathcal{M}_*}(P)$.

If $T_{\mathcal{M}}(P) = L_2^0(P)$ at each $P \in \mathcal{M}$, we say that $\mathcal{M}$ is a **nonparametric model**, even if $\mathcal{M} \subsetneq \mathcal{M}_*$. If $T_{\mathcal{M}}(P)$ is finite-dimensional at each $P \in \mathcal{M}$, we say that $\mathcal{M}$ is a **parametric model**. Otherwise, $\mathcal{M}$ is a **semiparametric model**.

# Concept of tangent space

Suppose that $O := (X, Y, Z) \sim P_{X,Y,Z} \in \mathcal{M}$ and that

$$\mathcal{M} = \mathcal{M}_X \otimes \mathcal{M}_{Y|X} \otimes \mathcal{M}_{Z|Y,X} \ ,$$

with $\mathcal{M}_X$, $\mathcal{M}_{Y|X}$ and $\mathcal{M}_{Z|Y,X}$ models for $P_X$, $P_{Y|X}$ and $P_{Z|Y,X}$, respectively, so that orthogonal components of $P_{X,Y,Z}$ are modeled orthogonally.

The total tangent space can be written as the direct sum

$$T_{\mathcal{M}}(P) = T_{\mathcal{M}_X}(P) \oplus T_{\mathcal{M}_{Y|X}}(P) \oplus T_{\mathcal{M}_{Z|Y,X}}(P)$$

of partial tangent spaces. This decomposition is very useful in many contexts.

For any $v \in T_{\mathcal{M}}(P)$, we have that

$$\Pi_{T_{\mathcal{M}}(P)} v = \Pi_{T_{\mathcal{M}_X}(P)} v + \Pi_{T_{\mathcal{M}_{Y|X}}(P)} v + \Pi_{T_{\mathcal{M}_{Z|Y,X}}(P)} v \ ,$$

where $\Pi_{\mathcal{R}}$ denotes projection onto $\mathcal{R}$.

# Concept of tangent space

If $\mathcal{M}_X$, $\mathcal{M}_{Y|X}$ and $\mathcal{M}_{Z|Y,X}$ are nonparametric, respectively, then

$$\begin{aligned}
T_{\mathcal{M}_X}(P) &= \{x \mapsto s(x) : P_X s = 0\} \\
T_{\mathcal{M}_{Y|X}}(P) &= \{(y,x) \mapsto s(y,x) : P_{Y|X} s = 0\} \\
T_{\mathcal{M}_{Z|Y,X}}(P) &= \{(z,y,x) \mapsto s(z,y,x) : P_{Z|Y,X} s = 0\} \ .
\end{aligned}$$

Furthermore, it is easy to verify that, if $(z,y,x) \mapsto s(z,y,x) \in L_2^0(P)$, then

$$\begin{aligned}
\Pi_{T_{\mathcal{M}_X}(P)} s &= x \mapsto E_P\left[s(Z,Y,X)|X=x\right] \\
\Pi_{T_{\mathcal{M}_{Y|X}}(P)} s &= (y,x) \mapsto E_P\left[s(Z,Y,X)|X=x, Y=y\right] - E_P\left[s(Z,Y,X)|X=x\right] \\
\Pi_{T_{\mathcal{M}_{Z|Y,X}}(P)} s &= (z,y,x) \mapsto s(z,y,x) - E_P\left[s(Z,Y,X)|X=x, Y=y\right]
\end{aligned}$$

## Concept of tangent space

As an example, suppose $O := (X, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the class of all bivariate distributions on $\mathbb{R}^2$ under which $X$ **and** $Y$ **are independent**.

What is the corresponding tangent space $T_{\mathcal{M}}(P)$?

Approach #1: **(conditional decomposition)**

We know from the previous slides that we may write

$$T_{\mathcal{M}}(P) = T_{\mathcal{M}_X}(P) \oplus T_{\mathcal{M}_{Y|X}}(P) \ .$$

Since the model $\mathcal{M}_X$ is nonparametric, we have that $T_{\mathcal{M}_X}(P) = L_2^0(P_X)$.

Because of independence, the model $\mathcal{M}_{Y|X}$ is simply the unrestricted model $\mathcal{M}_Y$ for the marginal of $Y$ – this is simply given by $L_2^0(P_Y)$.

It follows then that $T_{\mathcal{M}}(P) = L_2^0(P_X) + L_2^0(P_Y)$.

# Concept of tangent space

Approach #2: **(direct fluctuation approach)**

Suppose $p$ is the density of $P$ and take $\{p_\theta : \theta \in \Theta\}$ as a one-dimensional parametric submodel of $\mathcal{M}$ such that $p_{\theta=0} = p$. It must then be that for

$$p_\theta(x, y) = p_{X,\theta}(x) p_{Y,\theta}(y)$$

for every $(x, y)$ and $\theta$ for some marginal densities $p_{X,\theta}$ and $p_{Y,\theta}$ satisfying that $p_{X,\theta=0} = p_X$ and $p_{Y,\theta=0} = p_Y$.

It follows then

$$\frac{\partial}{\partial\theta} \log p_\theta(x, y)\bigg|_{\theta=0} = \frac{\partial}{\partial\theta} \log p_{X,\theta}(x)\bigg|_{\theta=0} + \frac{\partial}{\partial\theta} \log p_{Y,\theta}(y)\bigg|_{\theta=0},$$

which suggests that $T_{\mathcal{M}}(P) \subseteq T_{\mathcal{M}_X}(P) + T_{\mathcal{M}_Y}(P)$. Given scores $s_X \in L_2^0(P_X)$ and $s_Y \in L_2^0(P_Y)$, we see that $p_\theta(x, y) = [1 + \theta s_X(x)][1 + \theta s_Y(y)]p(x, y)$ has score $s_X(x) + s_Y(y)$, and so, $T_{\mathcal{M}}(P) \supseteq T_{\mathcal{M}_X}(P) + T_{\mathcal{M}_Y}(P)$.

## Concept of tangent space

Suppose that $O \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ is the **parametric model** $\{P_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^p$ is open and convex.

What is the corresponding tangent space $T_{\mathcal{M}}(P_\theta)$?

For each smooth submodel of $\mathcal{M}$ through $P_\theta$, there is some $u \in \mathbb{R}^p$ such that $\mathcal{M}_{\theta,u} := \{P_{\theta,\epsilon} := P_{\theta+\epsilon u} : \epsilon\} \subseteq \mathcal{M}$ locally approximates $P_\theta$. Setting $\nu_\theta(\epsilon) := \theta + \epsilon u$, the score for $\epsilon$ at $\epsilon = 0$ is then

$$\begin{aligned} s_{\theta,u}(o) &:= \left.\frac{\partial}{\partial \epsilon} \log p_{\theta+\epsilon u}(o)\right|_{\epsilon=0} = \left.\frac{\partial}{\partial \nu_\theta(\epsilon)} \log p_{\nu_\theta(\epsilon)}(o) \cdot \frac{\partial}{\partial \epsilon}\nu_\theta(\epsilon)\right|_{\epsilon=0} \\ &= u^\top \frac{\partial}{\partial \theta} \log p_\theta(o) \ . \end{aligned}$$

The tangent space $T_{\mathcal{M}}(P)$ is simply given by $\{s_{\theta,u} : u \in \mathbb{R}^p\}$ – this is nothing but the linear span of the components of the usual score function.

# Pathwise differentiability of statistical parameters and gradients

As we see from the numerator of the generalized Crámer-Rao bound, the development of a general efficiency theory requires that the statistical parameter $\Psi : \mathcal{M} \to \mathbb{R}$ be differentiable in some appropriate fashion.

A notion of differentiability valid over an arbitrary model space is needed.

- Common types require a locally convex model space.
- In semiparametric and parametric models, models are often not so.

How can we define differentiability over a possibly complex model space?

Over a parametric path, usual differentiability of real functions suffices. Can we extend this to a general model by defining derivatives over all parametric paths?

## Pathwise differentiability of statistical parameters and gradients

A parameter $\Psi : \mathcal{M} \to \mathbb{R}$ is **pathwise differentiable** at $P \in \mathcal{M}$ if there exists a continuous linear map $\dot{\Psi}_P : L_2^0(P) \to \mathbb{R}$ such that, for every $h \in T_{\mathcal{M}}(P)$,

$$\frac{\partial}{\partial \theta} \Psi(P_\theta) \bigg|_{\theta=0} = \dot{\Psi}_P(h)$$

for each regular one-dimensional parametric submodel $\{P_\theta : \theta \in \Theta\}$ through $P$ at $\theta = 0$ and with score for $\theta$ at $\theta = 0$ equal to $h$.

Any element $D(P) \in L_2^0(P)$ such the pathwise derivative can be represented as

$$\dot{\Psi}_P(h) = \langle D(P), h \rangle_P = P\left[D(P)h\right]$$

for each $h \in T_{\mathcal{M}}(P)$ is called a **gradient** of $\Psi$ at $P$ relative to $T_{\mathcal{M}}(P)$.

**A few observations on pathwise differentiability:**

- The pathwise derivative depends on the chosen path only through its associated score at $\theta = 0$.
- The Riez Representation Theorem guarantees the existence of a gradient.
- There is a direct parallel here between the pathwise derivative over general model spaces and the directional derivative in multivariate calculus.
  - If $f : \mathbb{R}^p \to \mathbb{R}$, $\vec{u}$ is a unit vector in $\mathbb{R}^p$ and $x$ is a point in $\mathbb{R}^p$, the directional derivative of $f$ at $x$ in the direction of $\vec{u}$ is given by

    $$D_{\vec{u}} f(x) = \vec{\nabla} f(x) \cdot \vec{u} \ ,$$

    an inner product between the gradient of $f$ at $x$ and a directional vector.
  - The function and location are disentangled from the direction of motion.
  - This parallel explains the use of the term *gradient* to describe $D(P)$.

# Pathwise differentiability of statistical parameters and gradients

**Unless $\mathcal{M}$ is nonparametric, there are many gradients.**

Denote by $\mathcal{G}_{\mathcal{M}}(P) \subset L_2^0(P)$ the set of gradients of $\Psi$ at $P$ relative to model $\mathcal{M}$.

If $D_0(P)$ is any given gradient, then

$$\mathcal{G}_{\mathcal{M}}(P) = \left\{ D(P) = D_0(P) + q(P) : q(P) \in T_{\mathcal{M}}(P)^{\perp} \right\},$$

where $T_{\mathcal{M}}(P)^{\perp}$ is the orthogonal complement of $T_{\mathcal{M}}(P)$ in $L_2^0(P)$.

There is only one gradient, say $D^*(P)$, in $T_{\mathcal{M}}(P)$ – it is referred to as the **canonical gradient**. It is found by projecting any gradient $D(P)$ into $T_{\mathcal{M}}(P)$:

$$D^*(P) = \Pi_{T_{\mathcal{M}}(P)} D(P) \text{ for each } D(P) \in \mathcal{G}_{\mathcal{M}}(P).$$

**Statistical inference in infinite-dimensional models relies heavily on knowledge of gradients**, and the **canonical gradient is critical for efficiency**.

**For a given parameter and model, how can we identify a gradient?**

We can, for example, use the definition of pathwise differentiability directly.

1. Take a smooth one-dimensional parametric submodel $\{P_\theta : \theta \in \Theta\} \subseteq \mathcal{M}$ with $P_{\theta=0} = P$ and score $h \in T_{\mathcal{M}}(P)$ at $\theta = 0$.
2. Compute $\frac{\partial}{\partial\theta}\Psi(P_\theta)\big|_{\theta=0}$ and express it as $P\left[D_\diamond(P)h\right]$, with $D_\diamond(P)$ not depending on the particular submodel chosen.
3. Recenter $D_\diamond(P)$ by $PD_\diamond(P)$, that is, take $D(P) := D_\diamond(P) - PD_\diamond(P)$.

**Relationship between gradients in nested models**

An easy but important fact is that

$$\mathcal{G}_{\mathcal{M}_2}(P) \subseteq \mathcal{G}_{\mathcal{M}_1}(P) \text{ whenever } \mathcal{M}_1 \subseteq \mathcal{M}_2.$$

In practice, this implies that we can always relax $\mathcal{M}$ to a nonparametric model in step 1 above, provided $\Psi$ is properly defined or can be extended there.

This allows us to use very simple submodels, including

1. $p_\theta(o) := [1 + \theta h(o)]\, p(o)$;
2. $p_\theta(o) := \exp[\theta h(o)]\, p(o)/c(\theta)$, where $c(\theta) := \int \exp[\theta h(o)]\, dP(o)$;
3. $p_\theta(o) := \operatorname{expit}[2\theta h(o)]\, p(o)/c(\theta)$, where $c(\theta) := \int \operatorname{expit}[2\theta h(o)]\, dP(o)$.

Submodels 1 and 2 generally require that $h$ be bounded, whereas submodel 3 does not. For the sake of computing a gradient, submodel 1 generally suffices.

**Example 1: a general moment**

Suppose that $\Psi(P) := Pf_0$ for a fixed and known function $f_0$.

With $p_\theta(o) := [1 + \theta h(o)]\, p(o)$, we find $\Psi(P_\theta) = \Psi(P) + \theta P\,(f_0 h)$ and so,

$$\frac{\partial}{\partial\theta}\Psi(P_\theta)\bigg|_{\theta=0} = P\,(f_0 h) = P\,[(f_0 - Pf_0)\,h]\,.$$

Thus, $D(P)(o) := f_0(o) - \Psi(P)$ is a gradient of $\Psi$ at $P$. Furthermore, it is the canonical gradient if the model for $P$ is nonparametric.

**Example 2: the average density value**

Suppose that $\Psi(P) := \int p^2(u)du = Pp$, where $p$ is the Lebesgue density of $P$.

With $p_\theta(o) := [1 + \theta h(o)]\, p(o)$, we find $\Psi(P_\theta) = \Psi(P) + 2\theta P(ph) + \theta^2 P(ph^2)$ and so,

$$\frac{\partial}{\partial\theta}\Psi(P_\theta)\bigg|_{\theta=0} = P\,(2ph) = P\,[2\,(p - Pp)\,h]\,.$$

Thus, $D(P)(o) := 2\,[p(o) - \Psi(P)]$ is a gradient of $\Psi$ at $P$. Again, this is the canonical gradient of $\Psi$ in a nonparametric model for $P$.

# Characterizing the set of influence functions

To understand the relevance of computing gradients of a statistical parameter, we first define the notion of regularity.

An estimator $\psi_n$ of $\psi_0 := \Psi(P_0)$ is **locally regular** at $P_0$ if for any $g \in T_{\mathcal{M}}(P_0)$, there is a path $\{P_\theta\}$ through $P_0$ at $\theta = 0$ and with score $g$ at $\theta = 0$ such that, under sampling from $P_{n^{-1/2}}$ and $P_0$, respectively,

$$n^{1/2}\left(\psi_n - \psi_{0n}\right) \quad \text{and} \quad n^{1/2}\left(\psi_n - \psi_0\right)$$

have the same limit distribution, where we denote $\psi_{0n} := \Psi(P_{n^{-1/2}})$.

If this holds uniformly over $\mathcal{M}$, then $\psi_n$ is said to be **regular** over $\mathcal{M}$.

This guarantees that small perturbations in the data-generating distribution do not affect the limiting distribution of the estimator.

## Characterizing the set of influence functions

Key result #1: **influence functions are gradients**

Suppose that $\psi_n$ is an asymptotically linear estimator of $\psi_0 := \Psi(P_0)$ with influence function $\phi_{P_0}$. Then, the following statements are equivalent:

1. $\Psi$ is pathwise differentiable and $\phi_P$ is a gradient of $\Psi$ at $P$;
2. the estimator $\psi_n$ is regular.

Key result #2: **gradients are influence functions** (Klaassen, 1987)

Under certain regularity conditions and for a given gradient $\phi_P$, the following statements are equivalent:

1. an asymptotically linear estimator of $\psi_0$ with influence function $\phi_{P_0}$ exists;
2. it is possible to estimate $\phi_{P_0}$ consistently (in an appropriate sense).

**Why is this relevant information?**

- If we wish to construct regular asymptotically linear (RAL) estimators, then this suggests that
  - we must restrict ourselves to pathwise differentiable parameters;
  - studying the pathwise derivative of our parameter is critical.
- A gradient can be found by computing the influence curve of an estimator known to be RAL.
  - For this, it is sometimes useful to consider the discrete setting as a guide.
  - Example: suppose $O_i := (W_i, A_i, Y_i)$, $O_1, O_2, \ldots, O_n \overset{iid}{\sim} P_0$ and consider

$$\psi_n := \frac{1}{n} \sum_{k=1}^{n} \frac{\frac{1}{n} \sum_{i=1}^{n} Y_i A_i I(W_i = W_k)}{\frac{1}{n} \sum_{i=1}^{n} A_i I(W_i = W_k)}$$

    as an estimator of $\psi_0 := E_{P_0} E_{P_0}(Y|A = 1, W)$ when $W$ has finite support.

## Efficiency bounds and the efficient influence function (EIF)

This link between influence functions and gradients is critical to **establishing efficiency bounds** in arbitrary models.

Say $\psi_n$ is a RAL estimator of $\psi_0$. Then, $n^{1/2}(\psi_n - \psi_0)$ has asymptotic variance $P_0 D(P_0)^2$ for some gradient $D(P_0) \in L_2^0(P_0)$.

- We can represent any $D$ as $D^* + H$ for some $H \in T_{\mathcal{M}}^{\perp}(P)$ – in fact, we can take $H(P) := \Pi_{T_{\mathcal{M}}^{\perp}(P)} D(P)$. This allows us to write that

$$\begin{aligned} P_0 D(P_0)^2 &= P_0 D^*(P_0)^2 + 2P_0 D^*(P_0) H(P_0) + P_0 H(P_0)^2 \\ &= P_0 D^*(P_0)^2 + P_0 H(P_0)^2 \\ &\geq P_0 D^*(P_0)^2 \ . \end{aligned}$$

- This lower bound is exactly achieved whenever $D(P_0) = D^*(P_0)$.

# Efficiency bounds and the efficient influence function (EIF)

**Characterization of an efficient estimator:**

A regular asymptotically linear estimator $\psi_n$ of $\psi_0$ is efficient

if and only if

$$\psi_n = \psi_0 + \frac{1}{n}\sum_{i=1}^{n} D^*(P_0)(O_i) + o_P(n^{-1/2}).$$

For this, the canonical gradient is often referred to as the **efficient influence function**: it is the influence function of any efficient RAL estimator of $\psi_0$.

For such an estimator $\psi_n$, the asymptotic variance of $n^{1/2}(\psi_n - \psi_0)$ is exactly equal to $\sigma_0^2 := P_0 D^*(P_0)^2$.

How does this relate to the generalized Crámer-Rao bound?

# Efficiency bounds and the efficient influence function (EIF)

If $\{P_{\theta,g} : \theta \in \Theta\}$ is a one-dimensional parametric submodel through $P_0$ at $\theta = 0$ and with score $g$ for $\theta$ at $\theta = 0$, the Crámer-Rao lower bound is

$$\frac{\left(\frac{\partial}{\partial\theta}\Psi(P_{\theta,g})\big|_{\theta=0}\right)^2}{\mathcal{I}_{\mathcal{M}_g}(0)} = \frac{(P_0 D^*(P_0) g)^2}{P_0 g^2} \leq \frac{P_0 D^*(P_0)^2 P_0 g^2}{P_0 g^2} = P_0 D^*(P_0)^2.$$

The generalized Crámer-Rao bounder should then satisfy that

$$\sup_{g \in T_{\mathcal{M}}(P)} \frac{\left(\frac{\partial}{\partial\theta}\Psi(P_{\theta,g})\big|_{\theta=0}\right)^2}{\mathcal{I}_{\mathcal{M}_g}(0)} \leq P_0 D^*(P_0)^2 \ .$$

Since this upper bound side is achieved by a submodel with $g = D^*(P)$, it defines the efficiency bound.

Any such submodel is said to be a **least-favorable parametric submodel**.

# Efficiency bounds and the efficient influence function (EIF)

As an example, suppose that $O := (X, Y) \sim P_0 \in \mathcal{M}$, where $\mathcal{M}$ consists of all bivariate distributions on $\mathbb{R}^2$ under which $X$ **and** $Y$ **are independent**. We wish to estimate a moment $P_0 f_0$ for fixed $f_0$ using $n$ independent draws from $P_0$.

We found before that the tangent space here is $T_{\mathcal{M}}(P) = L_2^0(P_X) + L_2^0(P_Y)$. Given $s \in L_2^0(P)$, we can verify that the projection of $s$ onto $T_{\mathcal{M}}(P)$ is simply given pointwise by $\Pi_{T_{\mathcal{M}}(P)} s(x, y) = E_P[s(x, Y)] + E_P[s(X, y)]$.

Using as initial gradient (relative to $\mathcal{M}$) the nonparametric EIF of $\Psi$, known to be $D(P) := f_0 - P_0 f_0$, we obtain the EIF of $\Psi$ relative to $\mathcal{M}$ as

$$\Pi_{T_{\mathcal{M}}(P)} D(P)(x, y) = E_P[f_0(x, Y)] + E_P[f_0(X, y)] \ .$$

This allows us to check, for example, that $F_n^*(x_0, y_0) = F_{X,n}(x_0) F_{Y,n}(y_0)$ is an asymptotically efficient estimator of $F_0(x_0, y_0) := P_0 I_{(-\infty, x_0] \times (-\infty, y_0]}$, where $F_{X,n}$ and the $F_{Y,n}$ are the empirical marginal CDFs based on the $X$ and $Y$ samples, respectively, whereas the empirical bivariate CDF at $(x_0, y_0)$ is not!!!

# Impact and role of nuisance modeling in determining the EIF

**Not all restrictions on $\mathcal{M}$ have an impact on the efficiency bound.**

Say $P = Qg$ with $\Psi : \mathcal{M} \to \mathbb{R}$ depending on $P$ through $Q$ alone, and that $P \in \mathcal{M}$ iff $Q \in \mathcal{M}_Q$ and $f \in \mathcal{M}_g$, i.e., $Q$ and $g$ are variationally independent.

A few important observations follow:

- the total tangent space can be expressed as $T_{\mathcal{M}}(P) = T_{\mathcal{M}_Q}(P) \oplus T_{\mathcal{M}_g}(P)$;
- shrinking $T_{\mathcal{M}_g}(P)$ generally enlarges $T_{\mathcal{M}}^{\perp}(P)$;
- $\mathcal{G}_{\mathcal{M}}(P)$ and $T_{\mathcal{M}_g}(P)$ are orthogonal to each other;
- since the EIF is strictly contained in $T_{\mathcal{M}_Q}(P)$, it is not affected by any shrinking of $T_{\mathcal{M}_g}(P)$.

Conclusion:

**Even though restrictions on $\mathcal{M}_g$ generally yield more gradients, in no way do they impact the EIF.**

Thus, to find EIF, we may as well do as if $g$ were completely known!

## Impact and role of nuisance modeling in determining the EIF

Example: **estimating the mean counterfactual**

Writing $P_O = P_{Y|A,W} P_{A|W} P_W$, we note that $\Psi$ does not depend on $P_{A|W}$ – the latter is an **orthogonal nuisance parameter**.

Restrictions on the model for $P_{A|W}$ do not change the EIF. If $\mathcal{M}_*$ is the fully nonparametric model and $\mathcal{M}_0$ is the same model with additional knowledge that $g$ is completely known (i.e., $g = g_0$), the EIF in $\mathcal{M}_*$ and $\mathcal{M}_0$ are the same.

In $\mathcal{M}_0$, the estimator $\frac{1}{n}\sum_{i=1}^{n} \frac{Y_i A_i}{g_0(W_i)}$ can be used – it has influence function

$$D(P)(o) := \frac{ya}{g_0(w)} - \Psi(P) \; .$$

This is a gradient in $\mathcal{M}_0$ (but not in $\mathcal{M}$). Upon projecting it onto the tangent space

$$T_{\mathcal{M}_0}(P) = T_{\mathcal{M}_{Y|A,W}}(P) + T_{\mathcal{M}_W}(P) \; ,$$

we recover the EIF $D^*(P)(o) := \frac{a}{g(w)}\left[y - \bar{Q}(w)\right] + \bar{Q}(w) - \Psi(P)$.

# Impact and role of nuisance modeling in determining the EIF

Suppose that $P = Qg$ with $\Psi : \mathcal{M} \to \mathbb{R}$ depending on $P$ through $Q$ alone, and that $P \in \mathcal{M}$ iff $Q \in \mathcal{M}_Q$ and $g \in \mathcal{M}_g$.

For given $g \in \mathcal{M}_g$, define the model $\mathcal{M}(g) = \{P = Qg : Q \in \mathcal{M}_Q\}$.

(Theorem 2.3 of van der Laan & Robins, 2003) Provided that

1. $\psi_n(g_0)$ is an asymptotically linear estimator of $\psi_0 := \Psi(P_0)$ in $\mathcal{M}(g_0)$, say with influence function $\mathrm{IC}_{P_0}$;
2. $\psi_n(g_n) - \psi_0 = \psi_n(g_0) - \psi_0 + \chi(g_n) - \chi(g_0) + o_P(n^{-1/2})$ for some functional $\chi$;
3. $\chi(g_n)$ is an efficient estimator of $\chi(g_0)$ in the model $\mathcal{M}$,

the estimator $\psi_n(g_n)$ of $\psi_0$ is asymptotically linear with influence function

$$\mathrm{IC}^*_{P_0} := \mathrm{IC}_{P_0} - \Pi_{T_{\mathcal{M}_g}(P_0)} \mathrm{IC}_{P_0} \; .$$

# Impact and role of nuisance modeling in determining the EIF

Why is this an **improvement in efficiency**?

We can decompose the entire space as the direct sum

$$\begin{aligned} L_2^0(P) &= T_{\mathcal{M}}^{\perp}(P) \oplus T_{\mathcal{M}}(P) \\ &= T_{\mathcal{M}}^{\perp}(P) \oplus T_{\mathcal{M}_Q}(P) \oplus T_{\mathcal{M}_g}(P) . \end{aligned}$$

In general, $\mathrm{IC}_{P_0}$ has a component in each of these three summands, each given by an appropriate projection. Specifically, setting

$$v_0 := \Pi_{T_{\mathcal{M}}^{\perp}(P_0)} \mathrm{IC}_{P_0}, \quad v_1 := \Pi_{T_{\mathcal{M}_Q(P_0)}} \mathrm{IC}_{P_0} \ \text{ and } \ v_2 := \Pi_{T_{\mathcal{M}_g(P_0)}} \mathrm{IC}_{P_0} ,$$

we have that $\mathrm{IC}_{P_0} = v_0 + v_1 + v_2$ while $\mathrm{IC}_{P_0}^* = v_0 + v_1$. Since all summands are orthogonal, we have that

$$P_0 \mathrm{IC}_{P_0}^2 = P_0 \mathrm{IC}_{P_0}^{*2} + P_0 v_2^2 \geq P_0 \mathrm{IC}_{P_0}^{*2} .$$

Of course, there is still not optimal if $v_0 \not\equiv 0$!

# Impact and role of nuisance modeling in determining the EIF

Recall the example at the end of Chapter 2: estimating a mean counterfactual using an **IPTW estimator with known or estimated propensity score**.

The influence function $\mathrm{IC}^*_{P_0}$ for the IPTW estimator using an estimator $g_n := g_{\theta_n}$ of the true propensity $g_0 := g_{\theta_0}$ based on the parametric model $\{g_\theta : \theta \in \Theta\}$ is given by

$$\mathrm{IC}^*_{P_0} = \mathrm{IC}_{P_0} + \gamma_0 \phi_{\theta_0} \ ,$$

where $\mathrm{IC}_{P_0}$ is the influence function of the IPTW estimator using the known $g_0$, $\gamma_0 = -\int \bar{Q}(w,1) \, \frac{\partial}{\partial\theta} \log g_\theta(w)\big|_{\theta=\theta_0} dQ_{W,0}(w)$ and $\phi_{\theta_0}$ is the influence function of $\theta_n$.

We derived this result in Chapter 2. We can show that

$$\gamma_0 \phi_{\theta_0} = -\Pi_{T_{\mathcal{M}_g}(P_0)} \mathrm{IC}_{P_0} \ ,$$

thereby directly establishing the previous theorem in this context.

## Impact and role of nuisance modeling in determining the EIF

Indeed, we can verify each of the following facts:

- the score for the conditional distribution of $A$ given $W$ is given
$$s_{A|W,\theta}(a,w) := \frac{\frac{\partial}{\partial\theta} g_\theta(w)}{g_\theta(w)[1-g_\theta(w)]}\left[a - g_\theta(w)\right];$$
- $\langle \mathrm{IC}_{P_0}, s_{A|W,\theta_0}\rangle_{P_0} = P_0\left[\mathrm{IC}_{P_0} s_{A|W,\theta_0}\right] = -\gamma_0;$
- $\langle s_{A|W,\theta_0}, s_{A|W,\theta_0}\rangle_{P_0} = P_0 s^2_{A|W,\theta_0} = -P_0\left(\frac{\partial}{\partial\theta} s_{A|W,\theta}\big|_{\theta=\theta_0}\right);$
- the tangent space $T_{\mathcal{M}_g}(P_0)$ of the model for the conditional distribution of $A$ given $W$ is simply given by $\{\alpha s_{A|W,\theta_0} : \alpha \in \mathbb{R}\}$;
- if $\theta_n$ is asymptotically efficient, then $\phi_{\theta_0} = s_{A|W,\theta_0}/\langle s_{A|W,\theta_0}, s_{A|W,\theta_0}\rangle_{P_0}$;
- the projection of $\mathrm{IC}_{P_0}$ onto $T_{\mathcal{M}_g}(P_0)$ is given by
$$\frac{\langle \mathrm{IC}_{P_0}, s_{A|W,\theta_0}\rangle_{P_0}}{\langle s_{A|W,\theta_0}, s_{A|W,\theta_0}\rangle_{P_0}} s_{A|W,\theta_0} = -\gamma_0 \phi_{\theta_0}\ .$$

# Relevant references


- Klaassen, CAJ (1987). Consistent estimation of the influence function of locally asymptotically linear estimators. Annals of Statistics.
- Pfanzagl, J (1990). Estimation in semiparametric models. Springer. (*Chapter 1*)
- Newey, W (1991). The asymptotic variance of semiparametric estimators. Working paper, Department of Economics, Massachusetts Institute of Technology.
- Bickel, PJ, Klaassen, CAJ, Ritov Y & Wellner, JA (1993). Efficient and adaptive estimation for semiparametric models. Springer. (*Chapter 3*)
- van der Laan, MJ & Robins, JM (2003). Unified methods for censored longitudinal data and causality. Springer. (*Chapters 1-2*)
- van der Laan, MJ & Rose, S (2013). Targeted learning: causal inference for observational and experimental data. Springer. (*Appendices 4, 5 and 7*)