Chapter 3. Efficiency in parametric models

March 3, 2022

Outline

Informal motivating discussion

Suppose that $X_1, \ldots, X_n$ are i.i.d. and let $\underline{X} = (X_1, \ldots, X_n)$. Let $S_\theta^{(n)}(\theta)$ and $\Lambda^{(n)}(\theta)$ be the score for θ and the tangent space at θ for a parametric model for the distribution of $\underline{X}$ indexed by θ.

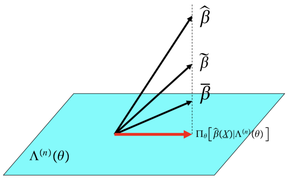
Suppose that $\widehat{\beta} \equiv \widehat{\beta}(\underline{X})$ is an unbiased estimator of an $\mathbb{R}^k$-valued parameter $\beta(\theta)$ in the parametric model.

From the discussion in the previous subsection we know that the variance of an estimator cannot exceed the variance of its projection into any space and in particular into the tangent space, that is,

$$
\begin{aligned}
&\mathrm{var}_\theta\{\widehat{\beta}(\underline{X})\} \\
&\geq \mathrm{var}_\theta\left\{\Pi_\theta\left[\widehat{\beta}(\underline{X}) \mid \Lambda^{(n)}(\theta)\right]\right\} \\
&= E\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\} \mathrm{var}_\theta\left\{S_\theta^{(n)}(\theta) S_\theta^{(n)}(\theta)^T\right\}^{-1} E\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\}^T \\
&= n^{-1} E\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\} I(\theta)^{-1} E\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\}^T .
\end{aligned}
$$

What makes unbiased estimators special is that as we shall now see, the projections of all unbiased estimators into the tangent space coincide. This, in turn, implies that the variance of the UNIQUE projection is a lower bound for the variance of ANY unbiased estimator of $\beta(\theta)$.

Graphically, if $\widehat{\beta}, \widetilde{\beta}$ and $\bar{\beta}$ are unbiased estimators of $\beta(\theta)$, then



We will see next that the variance of $\Pi_\theta\left[\widehat{\beta}(\underline{X}) \mid \Lambda^{(n)}(\theta)\right]$ is

$$n^{-1} C_{\mathcal{F}}(\theta).$$

The key fact that makes all unbiased estimators to have the same projection is that in regular parametric models indexed by θ,

$$\frac{\partial \beta(\theta)}{\partial \theta^T} = E_\theta\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\}. \tag{1}$$

This is so because if this identity is true, then

$$\begin{aligned}\Pi_\theta\left[\widehat{\beta}(\underline{X}) \mid \Lambda^{(n)}(\theta)\right] =& E\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\} \operatorname{var}_\theta\left\{S_\theta^{(n)}(\theta)\right\}^{-1} S_\theta^{(n)}(\theta) \\ =& \frac{\partial \beta(\theta)}{\partial \theta^T} \operatorname{var}_\theta\left\{S_\theta^{(n)}(\theta)\right\}^{-1} S_\theta^{(n)}(\theta),\end{aligned}$$

and the expression in the right hand side of the last equality does not depend on the estimator $\widehat{\beta}(\underline{X})$.

If (1) holds, then

$$\begin{aligned}&\operatorname{var}_\theta\left\{\Pi_\theta\left[\widehat{\beta}(\underline{X}) \mid \Lambda^{(n)}(\theta)\right]\right\} \\ =& E_\theta\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\} \operatorname{var}_\theta\left\{S_\theta^{(n)}(\theta)\right\}^{-1} E_\theta\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\}^T \\ =& E_\theta\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\}\left\{n \operatorname{var}_\theta\left[S_\theta(\theta)\right]\right\}^{-1} E_\theta\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\}^T \\ =& n^{-1} \frac{\partial \beta(\theta)}{\partial \theta^T} I(\theta)^{-1} \frac{\partial \beta(\theta)}{\partial \theta} \\ \equiv& n^{-1} C_{\mathcal{F}}(\theta).\end{aligned}$$

So,

$$\operatorname{var}_\theta\{\widehat{\beta}(\underline{X})\} \geq n^{-1} C_{\mathcal{F}}(\theta).$$

The matrix

$$\underbrace{C_{\mathcal{F}}(\theta)}_{k \times k} = \underbrace{\frac{\partial \beta(\theta)}{\partial \theta^T}}_{k \times p} \underbrace{I(\theta)^{-1}}_{p \times p} \underbrace{\frac{\partial \beta(\theta)}{\partial \theta}}_{p \times k},$$

is called the Cramer-Rao bound (per unit of sample) for $\beta(\theta)$ in model $\mathcal{F}$ at the law F_θ.

Informal derivation of (1)

Since for all θ,

$$\beta(\theta) = \int \widehat{\beta}(\underline{x}) f(\underline{x};\theta) dx,$$

and assuming exchange of differentiation and integration is possible, then

$$\begin{aligned}\frac{\partial\beta(\theta)}{\partial\theta^T} &= \int \widehat{\beta}(\underline{x}) \frac{\partial f(\underline{x};\theta)}{\partial\theta^T} d\underline{x} \\ &= \int \widehat{\beta}(\underline{x}) \left\{ \frac{\partial f(\underline{x};\theta)/\partial\theta^T}{f(\underline{x};\theta)} \right\} f(\underline{x};\theta) dx \\ &= \int \widehat{\beta}(\underline{x}) S_\theta^{(n)}(\theta)^T f(\underline{x};\theta) dx \\ &= E_\theta \left\{ \widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T \right\}.\end{aligned}$$

Informal derivation of (1)

The preceding derivation required that

(1) the derivative of $f(x;\theta)$ with respect to θ exists for all x and

(2) exchange of differentiation and integration be possible.

Indeed the result can be derived under much weaker conditions as stated in Lemma 3.3 below. These conditions include that of the mean squared differentiability of the square root of the density which we define next.

Definition 3.1. Let $f(x;\theta)$ be a density indexed by θ. Let $\Theta \subseteq \mathbb{R}^p$ be open. The map

$$\theta \in \Theta \to \sqrt{f(\cdot;\theta)}$$

is said to be mean squared differentiable at θ^* if there exists

$$s_\theta(x;\theta^*) = \left(s_{\theta_1}(x;\theta^*), \ldots, s_{\theta_p}(x;\theta^*)\right)^T$$

such that for all h in $\mathbb{R}^p$, when $h \to 0$,

$$\frac{1}{\|h\|^2} \int \left[\sqrt{f(x;\theta^*+h)} - \sqrt{f(x;\theta^*)} - \frac{1}{2} h^T s_\theta(x;\theta^*) \sqrt{f(x;\theta^*)} \right]^2 dx \to 0.$$

The vector $s_\theta(x;\theta^*)$ is called the score for θ at θ^*.

Typically, whenever

$$\frac{1}{\|h\|^2} \int \left[\sqrt{f(x;\theta^*+h)} - \sqrt{f(x;\theta^*)} - \frac{1}{2} h^T s_\theta(x;\theta^*) \sqrt{f(x;\theta^*)} \right]^2 dx \to 0 \quad h \to 0,$$

it holds that for μ-almost all x,

$$\begin{aligned}\frac{1}{2} s_\theta(x;\theta)\sqrt{f(x;\theta)} &= \partial\sqrt{f(x;\theta)}/\partial\theta \\ &= \frac{1}{2} \frac{\partial f(x;\theta)/\partial\theta}{\sqrt{f(x;\theta)}} \\ &= \frac{1}{2} \frac{\partial f(x;\theta)/\partial\theta}{f(x;\theta)} \sqrt{f(x;\theta)},\end{aligned}$$

so $s_\theta(x;\theta) = \partial \log f(x;\theta)/\partial\theta$.

In a Lemma below we will state that mean squared differentiability holds under first order differentiability of $\sqrt{f(x;\theta)}$ as a function of θ for each x and continuity of the information matrix.

However, indeed, mean squared differentiability can hold even without pointwise differentiability as the following example shows.

Example 3.1. Consider the double exponential density

$$f(x;\theta) = \frac{1}{2}\exp\{-|x-\theta|\}.$$

Then for each x, the map $\theta \to \sqrt{f(x;\theta)}$ is not differentiable at $\theta = x$, yet the map

$$\theta \to \sqrt{f(\cdot;\theta)}$$

is differentiable in mean square at θ^* with

$$s_\theta\left(x;\theta^*\right) = \mathrm{sg}\left(x-\theta^*\right).$$

For each x, the double exponential density

$$f(x;\theta) = \frac{1}{2}\exp\{-|x-\theta|\},$$

is infinitely differentiable with respect to θ at θ^* except for $x = \theta^*$.

One may think that differentiability with respect to θ of $f(x;\theta)$ for μ-almost all x, implies mean squared differentiability of $\theta \to \sqrt{f(\cdot;\theta)}$. This is however incorrect. A counterexample is the uniform density

$$f(x;\theta) = \theta^{-1}\mathrm{I}_{(0;\theta)}(x) \text{ with } \theta > 0.$$

For this model, the map $\theta \to \sqrt{f(\cdot;\theta)}$ is NOT mean squared differentiable at a given θ^*, even though infinitely differentiable with respect to θ at θ^* for all x, except for $x = \theta^*$.

Lemma 3.1. Suppose that at each fixed x the map

$$\theta \to \sqrt{f(\cdot;\theta)}$$

is continuously differentiable. Furthermore, suppose that the elements of the matrix

$$I(\theta) = \int \frac{\partial \log f(x;\theta)}{\partial\theta}\frac{\partial \log f(x;\theta)}{\partial\theta^T} f(x;\theta)dx$$

exist and is a continuous function of θ.
Then the map $\theta \to \sqrt{f(\cdot;\theta)}$ is mean squared differentiable and for each $\theta^* \in \Theta$,

$$s_\theta\left(x;\theta^*\right) = \left.\frac{\partial \log f(x;\theta)}{\partial\theta}\right|_{\theta=\theta^*}.$$

[1] This is Lemma 7.6 of van der Vaart, 2000.

Applying Lemma 3.1 it can be shown, for instance, that in exponential family models with densities

$$f(x;\theta) = c(\theta)h(x)\exp\left\{q(\theta)^T t(x)\right\}$$

satisfying that $q(\theta)$ is continuously differentiable, i.e. with continuous partial derivative, the map

$$\theta \to \sqrt{f(\cdot;\theta)}$$

is mean-squared differentiable at every θ such that $q(\theta)$ is in the interior of the natural parameter space (see Example 7.7 of van der Vaart, 2000).

The following Lemma states that the score has mean zero and finite variance.

Lemma 3.2. Suppose that Θ is an open subset of $\mathbb{R}^p$ and that the map

$$\theta \to \sqrt{f(\cdot;\theta)}$$

is mean squared differentiable at θ^*. Then,

- The score at θ has mean zero at θ^*, i.e.

$$\int s_\theta\left(x;\theta^*\right) f\left(x;\theta^*\right) dx = 0.$$

- All the entries of the matrix

$$I\left(\theta^*\right) = \int s_\theta\left(x;\theta^*\right) s_\theta\left(x;\theta^*\right)^T f\left(x;\theta^*\right) dx$$

exist. The matrix $I\left(\theta^*\right)$ is called the Information matrix for θ at θ^*.

[1] This is Lemma 7.2 of van der Vaart, 2000.

Back to the Cramer-Rao bound

Lemma 3.3. Let $X_1,\ldots,X_n$ i.i.d. and $T = t\left(X_1,\ldots,X_n\right)$ be a real valued measurable function. Let $\Theta \subset \mathbb{R}^p$ be open. Suppose that:

1) The map $\theta \in \Theta \to f(x;\theta)$ is continuous at θ^*(μ-a.e.),
2) The map $\theta \in \Theta \to \sqrt{f(x;\theta)}$ is mean squared differentiable at θ^*,
3) The map $\theta \in \Theta \to E_\theta\left(T^2\right)$ is continuous at θ^*.

Then, the partial derivatives of the map $\theta \in \Theta \to E_\theta(T)$ exist at θ^* and satisfy

$$\left.\frac{\partial E_\theta(T)}{\partial \theta}\right|_{\theta=\theta^*} = E_{\theta^*}\left\{T S_\theta^{(n)}\left(\theta^*\right)\right\},$$

where $S_\theta^{(n)}\left(\theta^*\right) = \sum_{i=1}^n s_\theta\left(X_i;\theta^*\right)$.

Applying the previous result to $T = \widehat{\beta}(\underline{X})$, we have that if $\widehat{\beta}(\underline{X})$ is unbiased, $\theta \to E_\theta\left(\widehat{\beta}^2\right)$ is continous at θ^* and $f(x;\theta)$ satisfies conditions 1 and 2 of Lemma 3.3, then at $\theta = \theta^*$ it holds that

$$\frac{\partial \beta(\theta)}{\partial \theta^T} = E_\theta\left\{\widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^T\right\},$$

because since $\widehat{\beta}(\underline{X})$ is unbiased,

$$\beta(\theta) = E_\theta\{\widehat{\beta}(\underline{X})\}.$$

Definition 3.2. A parametric model is regular iff there exists a parameterization indexed by $\theta \in \Theta$ such that
1) Θ is an open subset of $\mathbb{R}^p$,
2) For each θ^* in Θ, the map $\theta \to f(x;\theta)$ is continuous at θ^*(μ-a.e.),
3) The map $\theta \to \sqrt{f(\cdot;\theta)}$ is mean squared differentiable at all $\theta \in \Theta$,
4) The Information Matrix $I(\theta)$ is non-singular for all $\theta \in \Theta$.

- We know that the variance of unbiased estimators of $\beta(\theta)$ is no smaller than the Cramer-Rao bound.
- Then we would expect that the variance of the limiting distribution of estimators $\widehat{\beta}_n$ such that $\widehat{\beta}_n - \beta(\theta)$ converges (at an appropriate rate) to mean zero random distribution would also be no less than the Cramer-Rao lower bound.
- This is "almost true" but not quite true!
- One can construct "superefficient" estimators of $\beta(\theta)$ that have variance equal to the Cramer-Rao bound at most values of θ in the model but at some values of θ have indeed variance smaller than the Cramer-Rao bound.
- These estimators have excellent performance (as measured by mean squared error) at a θ at the expense of very poor performance in a neighborhood of θ.

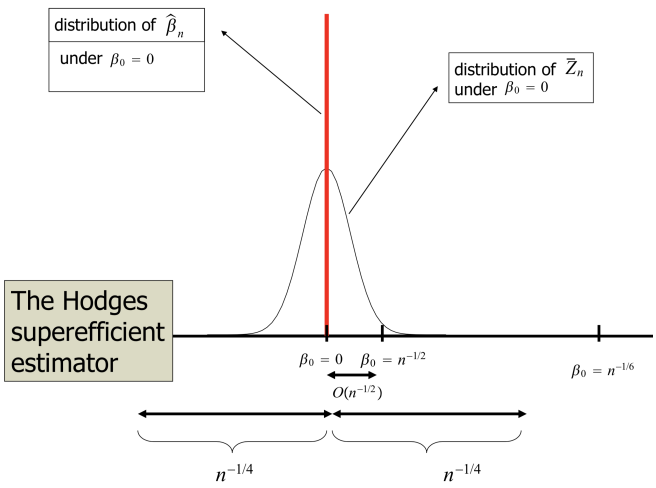
Example of a super-efficient estimator (Hodges)

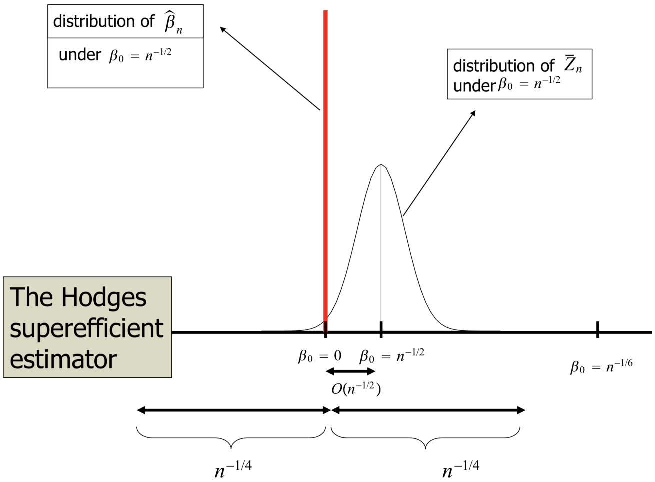
$Z_1, \dots, Z_n \overset{\text{iid}}{\sim} N(\beta, 1)$,

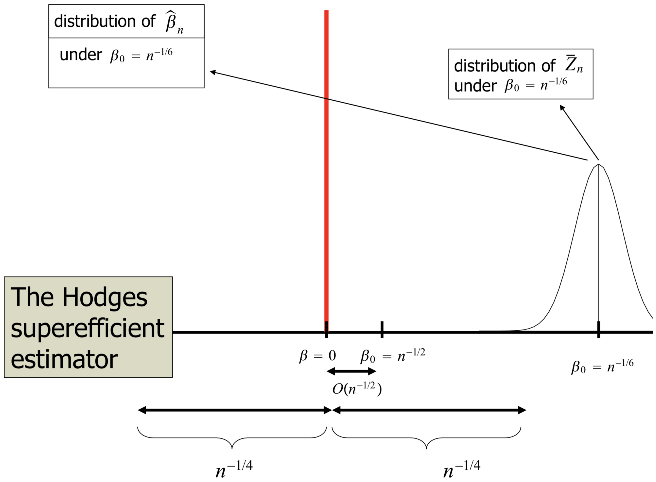
$$\widehat{\beta}_n = \begin{cases} \bar{Z}_n & \text{if } |\bar{Z}_n| > n^{-1/4} \\ 0 & \text{if } |\bar{Z}_n| \leq n^{-1/4} \end{cases}.$$

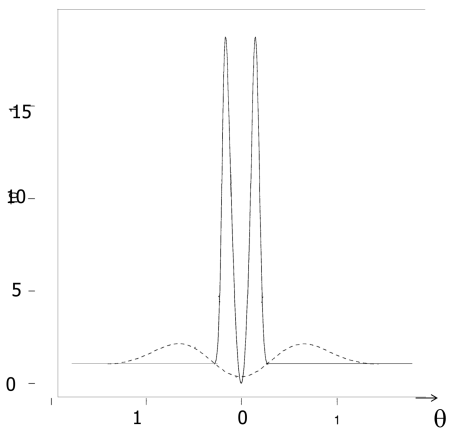
Then

$$\sqrt{n}\left(\widehat{\beta}_n - \beta\right) \underset{n\to\infty}{\overset{D(F_\beta)}{\to}} \begin{cases} N(0,1) & \text{if } \beta \neq 0 \\ 0 & \text{if } \beta = 0 \end{cases}.$$









MSE of Hodges Estimator

--- : sample Size 10

— : sample Size 100

Source: van Der Vaart, Asymptotic Statistics, Fig 8.1

The preceeding example shows that to formulate a theory of asymptotic efficiency it is not enough to study the convergence of estimators at each fixed θ.

We must analyze simultaneously, the distribution of the estimator under different values of the parameter, as the sample size increases.

Yet, the example also shows that somehow the different parameter values that matter, are indeed those in "shrinking neighborhoods"of a parameter value, that shrink with n at an appropriate rate.

This concept can be made rigorous through the notion of local asymptotic normality.

Assume $\theta \in \mathbb{R}$. If $f(x;\theta)$ is three times differentiable wrt θ, then for some δ_x,

$$\log \frac{f(x;\theta+h)}{f(x;\theta)} = hs_\theta(x;\theta) + \frac{1}{2}h^2 \frac{d^2}{d\theta^2}\log f(x;\theta) + \frac{1}{6}h^3 \frac{d^3}{d\theta^3}\log f(x;\theta+\delta_x).$$

So,

$$\log \prod_{i=1}^n \frac{f(X_i;\theta+h/\sqrt{n})}{f(X_i;\theta)} = \frac{h}{\sqrt{n}}\sum_{i=1}^n s_\theta(X_i;\theta) + \frac{h^2}{2n}\sum_{i=1}^n \frac{d^2}{d\theta^2}\log f(X_i;\theta) + \mathrm{Re}_n,$$

where for some $|\delta_{X_i}| < |h|/\sqrt{n}$,

$$\mathrm{Re}_n = \frac{1}{\sqrt{n}}\frac{1}{6}h^3 \frac{1}{n}\sum_{i=1}^n \frac{d^3}{d\theta^3}\log f(X_i;\theta+\delta_{X_i}).$$

So, if $|\log f(x;\theta+\eta)| < g(x)$ for all η in a neighborhood of θ for some $g(x)$ such that $E_F[g(X)] < \infty$,

$$\log \prod_{i=1}^n \frac{f(X_i;\theta+h/\sqrt{n})}{f(X_i;\theta)} = hI(\theta)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n I(\theta)^{-1}s_\theta(X_i;\theta)\right] - \frac{h^2}{2}I(\theta) + o_{p_\theta}(1).$$

By the CLT, the expression in squared brackets converges to $Z \sim N\left(0, I(\theta)^{-1}\right)$.

The identity shows that the likelihood ratio process (process in that it is indexed by θ) converges to

$$\log\left[\frac{\exp\left\{-\frac{(Z-h)^2}{2I(\theta)^{-1}}\right\}}{\exp\left\{-\frac{Z^2}{2I(\theta)^{-1}}\right\}}\right] = hI(\theta)Z - h^2\frac{1}{2}I(\theta).$$

The last display is the likelihood ratio based on just one observation Z in the model $\mathcal{F}_{\text{normal}} \equiv \left\{N\left(h, I(\theta)^{-1}\right) : h \in \mathbb{R}\right\}$.

The preceding derivation was carried out making too strong assumptions (e.g. existence of third derivative of $\log f(x;\theta)$ and θ scalar).

In fact, the likelihood ratio expansion holds in all regular parametric models.

Proposition 3.1. Suppose that $\mathcal{F} \equiv \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ is a regular parametric model, then for $h \in \mathbb{R}^p$,

$$\begin{aligned}\log \prod_{i=1}^n \left\{\frac{f(X_i;\theta+h/\sqrt{n})}{f(X_i;\theta)}\right\} =& h^T I(\theta)\left[\frac{1}{\sqrt{n}}\sum_{i=1}^n I(\theta)^{-1}s_\theta(X_i;\theta)\right] \\ & -\frac{1}{2}h^T I(\theta)h + o_{p_\theta}(1).\end{aligned}$$

[1]This is part of Theorem 7.2 of van der Vaart, 2000.

The preceding approximation suggests that, with large n, it should be as hard to estimate h from n iid observations drawn from a law in the "local model"

$$\mathcal{F}_{\text{local}} \equiv \{f(x;\theta+h/\sqrt{n}) : h \in \mathbb{R}\}$$

as it should be to estimate h from one observation drew from the Normal model

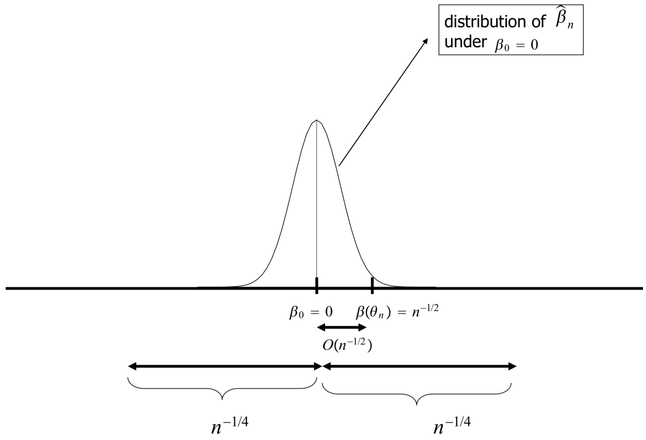
$$\mathcal{F}_{\text{normal}} \equiv \left\{N\left(h, I(\theta)^{-1}\right) : h \in \mathbb{R}\right\}.$$

Remark: note that in the local model $\mathcal{F}_{\text{local}}$, θ is fixed and regarded as known and h is the unknown parameter.

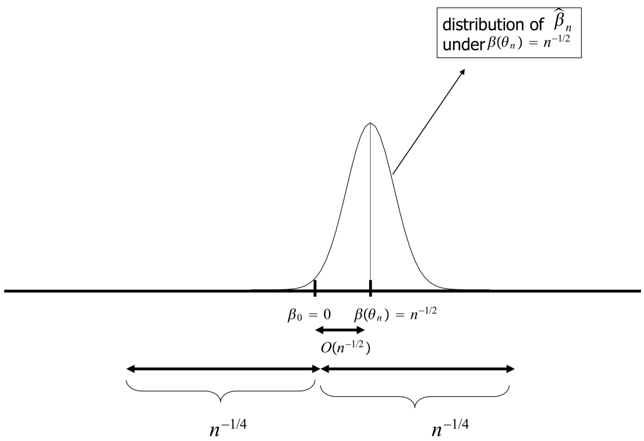
Definition 3.3. Given a collection of densities $\mathcal{F} \equiv \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ and n iid observations $X_1^{(n)}, \ldots, X_n^{(n)}$, an estimator sequence $\widehat{\beta}_n$ based on $X_1^{(n)}, \ldots, X_n^{(n)}$ is said to be a regular estimator of an $\mathbb{R}^k$ valued parameter $\beta(\theta)$ at the law F_{θ^*} (or simply, at θ^*), iff there exists a law G_θ such that for all $h \in \mathbb{R}^k$,

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(\theta + h/\sqrt{n})\right\} \overset{\theta+h/\sqrt{n}}{\to} G_\theta.$$

Regular estimators



Regular estimators



Warning: the class of regular estimators excludes the Hodges estimator (which is a good thing!) but it also excludes shrinkage estimators (which, is not necessarily a good thing!).

More on this later....

The following result, due to Hajek and known as the Convolution Theorem, should not come as a surprise...

Hajek's convolution theorem in parametric models

Proposition 3.2. Let $\mathcal{F} \equiv \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a regular parametric model and let $\beta(\theta)$ an $\mathbb{R}^k$-valued parameter that a differentiable function of θ.
If $\widehat{\beta}_n$ is a regular estimator of $\beta(\theta)$ at F_θ then

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(\theta)\right\} \overset{D(F_\theta)}{\to} U^* + U,$$

where

$$U^* \sim N_k\left(0, C_{\mathcal{F}}(\theta)\right),$$

U is independent of U^* and $C_{\mathcal{F}}(\theta)$ is the CR bound for $\beta(\theta)$ in model $\mathcal{F}$.

[1]This is Theorem 8.8 of van der Vaart, 2000.

Efficient estimators in regular parametric models

Definition 3.4. In a regular parametric model $\mathcal{F}$, an estimator $\widehat{\beta}_n$ of a differentiable parameter $\beta(\theta)$ is locally asymptotically efficient at F_θ if it is regular at F_θ and

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(\theta)\right\} \overset{D(F_\theta)}{\to} N_k\left(0, C_{\mathcal{F}}(\theta)\right).$$

The estimator is globally asymptotically efficient if it is locally asymptotically efficient at all θ.

Can we always find globally asymptotically efficient estimators?

We will see later in the class that if the map $\theta \to f(x;\theta)$ satisfies one additional "smoothness"condition, then the answer is yes, and a globally efficient estimator is, of course, the maximum likelihood estimator.

The convolution theorem establishes a precise way in which $N_k\left(0, C_{\mathcal{F}}(\theta)\right)$ can be interpreted as an asymptotic bound for the limiting distribution of estimators.

Yet, is regularity of an estimator a sufficiently compelling criterion that we should "always"prefer, with large samples, asymptotically efficient estimators, like the MLE, to other estimators?

Another look at the normal experiment will raise a warning about restricting attention solely to regular estimators and in particular, about declaring the MLE asymptotically optimal!

Let us then consider the normal experiment and to simplify the discussion, let us assume that the inferential problem is to estimate the means of p independent normals with variance 1.

That is, we are given one single draw Z from $N_p\left(h, Id_{p\times p}\right)$ and the inferential problem is to estimate the $p \times 1$ vector h.

This problem is probably one of the most studied problems in Statistics!

If $p \geq 3$ it is well known that the Stein's shrinkage estimator:

$$\hat{h} = \left\{1 - \left[(p-2)/\|Z\|^2\right]\right\} Z$$

where $\|Z\|^2 = \sum_i Z_i^2$ satisfies for all h,

$$E_h\left[\|\hat{h} - h\|^2\right] < E_h\left[\|Z - h\|^2\right] = \mathrm{var}_h(Z) = p.$$

So, even though it is the case that Z is minimax for the joint quadratic loss $\ell(x) = \|x\|^2$, i.e. for any estimator $\widetilde{h}$

$$\text{sup}_h E_h \left[||\widetilde{h} - h||^2 \right] \geq E_h \left[\|Z - h\|^2 \right],$$

it is also the case that for Stein's estimator $\hat{h}$,

$$E_h \left[\|Z - h\|^2 \right] > E_h \left[\|\hat{h} - h\|^2 \right].$$

Of course, this can only happen if $\text{sup}_h E_h \left[||\hat{h} - h||^2 \right]$ equals the risk of Z, so Stein's estimator is also minimax. The important take home message is that Stein's estimator has risk everywhere strictly smaller than that of Z.

Remark: The estimator Z is said to be inadmissible for the joint quadratic loss, because there exists at least one estimator whose risk is less than or equal that of Z for all h, and for at least one h, it is strictly smaller. By the way, it can be shown that Stein's estimator is also inadmissible! See Lehmann and Casella, Theory of Point Estimation, chapter 5.

We will now show an example (taken from van der Vaart, 2000, sec. 8.8) of an experiment with an estimator that is asymptotically efficient (and hence regular) at all θ except at one θ^*, and such that in the local experiment at θ^*, it converges to Stein's estimator in the normal experiment.

Such estimator, which we will show now, is NOT regular at θ^*, because Stein's estimator is NOT equivariant in law at such θ^*, but in large samples under values of $\theta^* + h/\sqrt{n}$ has strictly smaller risk (for joint quadratic loss) than the MLE.

This example then illustrates that not all "irregular"estimators are bad... Some can even "beat"the MLE even in large samples.

Example 3.2. Consider the p-variate normal model with unknown mean and known var equal $Id_{p\times p}$, $\mathcal{F} = \{N(\theta, Id_{p\times p}) : \theta \in \mathbb{R}^p\}$. Let $X_1, \dots X_n$ be iid from a law in $\mathcal{F}$. Consider the estimator of θ

$$\widehat{\theta}_n = \left(1 - \frac{p-2}{n\left\|\bar{X}_n\right\|^2}\right)\bar{X}_n.$$

If $\theta \neq 0$, then by the LLN,

$$\frac{(p-2)}{\left\|\bar{X}_n\right\|^2}\bar{X}_n \overset{P_{\theta+h/\sqrt{n}}}{\rightarrow} \frac{(p-2)}{\|\theta\|^2}\theta \text{ therefore } \frac{1}{\sqrt{n}}\frac{(p-2)}{\left\|\bar{X}_n\right\|^2}\bar{X}_n \overset{P_{\theta+h/\sqrt{n}}}{\rightarrow} 0,$$

so by Slutsky's theorem,

$$\sqrt{n}\left\{\widehat{\theta}_n - (\theta + h/\sqrt{n})\right\} = \sqrt{n}\left\{\bar{X}_n - (\theta + h/\sqrt{n})\right\} - \frac{1}{\sqrt{n}}\frac{(p-2)}{\left\|\bar{X}_n\right\|^2}\bar{X}_n$$

$\overset{\theta+h/\sqrt{n}}{\rightarrow} N(0, Id_{p\times p})$, thus showing that the estimator is regular and efficient at any $\theta \neq 0$.

When $\theta = 0$,

$$\begin{aligned}
\frac{1}{\sqrt{n}}\frac{(p-2)}{\left\|\bar{X}_n\right\|^2}\bar{X}_n &= (p-2)\frac{\sqrt{n}\bar{X}_n}{\left\|\sqrt{n}\bar{X}_n\right\|^2} \\
&= (p-2)\frac{\sqrt{n}\left(\bar{X}_n - h/\sqrt{n}\right) + h}{\left\|\sqrt{n}\left(\bar{X}_n - h/\sqrt{n}\right) + h\right\|^2} \\
&\overset{\theta+h/\sqrt{n}}{\rightarrow} (p-2)\frac{Z}{\|Z\|^2} \text{ with } Z \sim N(h, 1),
\end{aligned}$$

and

$$\sqrt{n}\left(\bar{X}_n - h/\sqrt{n}\right) \overset{\theta+h/\sqrt{n}}{\rightarrow} Z - h \text{ with } Z \sim N(h, 1),$$

so

$$\sqrt{n}\left\{\widehat{\theta}_n - (\theta + h/\sqrt{n})\right\} \overset{\theta+h/\sqrt{n}}{\rightarrow} \left[Z - (p-2)\frac{Z}{\|Z\|^2}\right] - h.$$

We conclude that when $\theta = 0$,

$$\sqrt{n}\left\{\widehat{\theta}_n - (\theta + h/\sqrt{n})\right\} \overset{\theta+h/\sqrt{n}}{\to} \left[Z - (p-2)\frac{Z}{\|Z\|^2}\right] - h,$$

for $Z \sim N(h, Id_{p\times p})$. Because the limiting r.v. has distribution that changes with h, we conclude that $\widehat{\theta}_n$ is not regular.

Yet, the limiting r.v. is Stein's estimator minus its mean, and as we have seen this difference has uniformly smaller risk (for joint squared error loss) than that of the best equivariant estimator Z, i.e. for all h,

$$E_h\left[\left\|\left[Z - (p-2)\frac{Z}{\|Z\|^2}\right] - h\right\|^2\right] < E_h\left[\|Z - h\|^2\right].$$

Then, since $\sqrt{n}\left\{\bar{X}_n - h/\sqrt{n}\right\} \sim N(0, Id_{p\times p})$, we have

$$E_h\left[\|Z - h\|^2\right] = E_0\left[\|Z\|^2\right] = E_{h/\sqrt{n}}\left[\left\|\sqrt{n}\left\{\bar{X}_n - h/\sqrt{n}\right\}\right\|^2\right],$$

and we conclude that $\hat{\theta}_n$ has asymp. strictly smaller risk than the MLE $\bar{X}_n$ in local neighborhoods of $\theta = 0$.

The preceeding example illustrates why settling on a criterion for what constitutes asymptotic optimality is so difficult and controversial.

It is important that you keep this in mind when we define later asymptotic efficiency in semiparametric models, as the definition we will give will require regularity of the estimator in a sense to be defined later.

One may interpret the lesson from the example as telling that one should not regard the MLE as asymptotically the "best"estimator. However, we will now see an additional result that tempers this conclussion.

The result implies that only in a set of Lebesgue measure 0 can the MLE have worst asymptotic performance than an irregular estimator.

Proposition 3.3. Suppose that $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ is a regular parametric model and $\beta(\theta)$ is an $\mathbb{R}^k$-valued parameter that is a differentiable function of θ at all $\theta \in \Theta$. Suppose that $\widehat{\beta}_n$ satisfies for each $\theta \in \Theta$,

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(\theta)\right\} \overset{\theta}{\to} G_\theta,$$

for some law G_θ. Then for for all $\theta \in \Theta$ except at most for θ in a set of Lebesgue measure $0, G_\theta$ is the distribution of $U_\theta^* + U_\theta$ where

$$U_\theta^* \sim N\left(0, C_{\mathcal{F}}(\theta)\right),$$

and

U_θ^* and U_θ are independent.

Definition 3.5. A sequence of estimators $\widehat{\beta}_n \equiv \widehat{\beta}_n(X_1, \ldots, X_n)$ of a parameter

$$\beta(\cdot) : \mathcal{F} \to \mathbb{R}^k$$

is asymptotically linear at F iff there exists a random vector $\varphi_F(X)$ (possibly depending on F) with mean zero and finite variance under F such that

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \varphi_F(X_i) + o_{p,F}(1),$$

where $o_{p,F}(1)$ denotes a sequence converging to 0 in probability under F. The function $\varphi_F(\cdot)$ is referred to as the influence function of the estimator $\widehat{\beta}_n$ at F.

By Slutzky's theorem if $\widehat{\beta}_n$ is an asymptotically linear estimator with influence function $\varphi_F(X)$ then

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} \overset{D(F_\theta)}{\to} N_k\left(0, \mathrm{var}\left\{\varphi_F(X)\right\}\right).$$

Remark 1: an asymptotically linear estimator has a unique influence function (you will show this as hwk).

Remark 2: the property of asymptotic linearity is NOT associated with a model, it is a property at each, single law F, regardless if F is assumed to belong to a model $\mathcal{F}$.

Asymptotically linear estimators are important because many estimators used in practice are asymptotically linear. For instance, Z-estimators typically are, as we will argue next.

Furthermore, importantly, asymptotically efficient estimators are indeed asymptotically linear as the next result from van der Vaart, 2000, states.

Thus, by restricting attention to the class of regular asymptotically linear estimators, and searching for the one whose influence function has the smallest variance we do not risk loosing efficiency.

Lemma 3.4. Suppose that $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ is a regular parametric model and $\beta(\theta)$ is an $\mathbb{R}^k$-valued parameter that is a differentiable function of θ. Then, if $\widehat{\beta}_n$ is asymptotically efficient, it must be asymptotically linear.

Remark: I mentioned earlier that under very mild reg. conditions (that we will see later) the MLE of $\beta(\theta)$ is asymptotically efficient. Then, by Lemma 3.4, it must be asymptotically linear. In a few slides we will show that its influence function at F_θ is equal to

$$\varphi_{\text{eff},F_\theta}(X) = -\left[\partial\beta(\theta)/\partial\theta^T\right] I(\theta)^{-1} S_\theta(X;\theta).$$

[1]This is a corollary of Lemma 8.14 of van der Vaart, 2000.

Estimating equations

Suppose that $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ is a parametric or semiparametric model and let $\beta(.) : \mathcal{F} \to \mathbb{R}^k$.

Suppose that $u(X;\beta)$ is a known $\mathbb{R}^k$-valued vector function of X and β such that for each $\theta \in \Theta$ there exists a unique $\beta(\theta) \equiv \beta(F_\theta)$ verifying

$$E_\theta[u(X, \beta(\theta))] = 0.$$

Suppose that $\widehat{\beta}_n$ based on $X_1, \ldots, X_n$ iid, is any β solving approximately the equation

$$\sum_i u(X_i, \beta) = 0. \qquad (2)$$

Note: many estimators used in practice solve (2) exactly. We will nevertheless allow that the equation be solved approximately (giving precise conditions to the degree of approximation) because for some important estimators such as the sample median, $u(x;\cdot)$ is discontinuous and an exact solution of (2) does not exist. Estimators solving eq like (2) are called Z-estimators in van der Vaart, 2000.

Even though we will study in detail later in the course the asymptotic theory for Z-estimators, at this point it is important that I highlight the main steps of the approach used to analyze Z-estimators as these will help understand the "regularity conditions"that suffice for Z-estimators to be asymptotically linear.

Define for any β,

$$e(\beta) \equiv \int u(x;\beta) f(x) dx = E_F[u(X;\beta)],$$
$$\widehat{e}(\beta) \equiv \mathbb{P}_n[u(X;\beta)] = n^{-1}\sum_{i=1}^{n} u(X_i;\beta).$$

Suppose that $\widehat{\beta}_n$ is an approximate solution of the estimating equation in that

$$\sqrt{n}\widehat{e}\left(\widehat{\beta}_n\right) = o_p(1).$$

Now, recalling that, by definition, $e(\beta(F)) = 0$, write

$$\begin{aligned} o_p(1) =& \sqrt{n}\widehat{e}\left(\widehat{\beta}_n\right) \\ =& \underbrace{\sqrt{n}\left\{\widehat{e}\left(\widehat{\beta}_n\right) - e\left(\widehat{\beta}_n\right)\right\} - \sqrt{n}\{\widehat{e}(\beta(F)) - e(\beta(F))\}}_{A_n} \\ &+ \underbrace{\sqrt{n}\{\widehat{e}(\beta(F)) - e(\beta(F))\}}_{B_n} + \underbrace{\sqrt{n}\left\{e\left(\widehat{\beta}_n\right) - e(\beta(F))\right\}}_{C_n}. \end{aligned}$$

- (a). If $E_F\left[u(X;\beta(F))^2\right] < \infty$, then by the CLT, $B_n \to N(0, V_F)$, so $B_n = O_p(1)$.
- (b). If $e(\beta)$ is differentiable at $\beta = \beta(F)$, then

  $$C_n = [\dot{e}(\beta(F))]\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} + \sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| o(1),$$

  where $\dot{e}(\beta) \equiv \partial e(\beta)/\partial\beta^T$.
- (c). Suppose that we can show that $A_n = o_p(1)$.

Under (a), (b) and (c), we have

$$[\dot{e}(\beta(F))]\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} = -\left\{B_n + \sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| o(1) + o_p(1)\right\},$$

from where

$$\left\|[\dot{e}(\beta(F))]\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\}\right\| = O_p(1) + \sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| o(1).$$

If, in addition, $\dot{e}(\beta(F))$ is invertible, then

$$\begin{aligned} \sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| &\le \left\|[\dot{e}(\beta(F))]^{-1}\right\| \left\|[\dot{e}(\beta(F))]\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\}\right\| \\ &= O_p(1) + \sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| o(1). \end{aligned}$$

Re-arranging terms in the last inequality we obtain

$$\sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| \le O_p(1)/\{1 - o(1)\},$$

which implies that

$$\sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| o(1) \le o(1)O_p(1)/\{1 - o(1)\} = o_p(1).$$

Thus, under the conditions of the preceeding slide, the identity

$$[\dot{e}(\beta(F))]\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} = -\left\{B_n + \sqrt{n}\left\|\widehat{\beta}_n - \beta(F)\right\| o(1) + o_p(1)\right\}$$

implies that

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} = -[\dot{e}(\beta(F))]^{-1}B_n + o_p(1),$$

or equivalently

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} = -\left[\partial E_F[u(X;\beta)]/\,\partial\beta^T\big|_{\beta=\beta(F)}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u\left(X_i;\beta(F)\right) + o_p(1).$$

We have thus shown the following proposition.

Proposition 3.4. if $\widehat{\beta}_n$ satisfies $n^{-1/2}\sum_{i=1}^{n} u\left(X_i;\widehat{\beta}_n\right) = o_p(1)$ and
i) $\left[\partial E_F[u(X;\beta)]/\,\partial\beta^T\big|_{\beta=\beta(F)}\right]$ exists and is invertible,
ii) $E_F\left[u(X;\beta(F))^2\right] < \infty$,
iii) regularity conditions that we will study later in the course hold so that $A_n = o_p(1)$,
then the estimator $\widehat{\beta}_n$ satisfies

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} = -\left[\partial E_F[u(X;\beta)]/\,\partial\beta^T\big|_{\beta=\beta(F)}\right]^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u\left(X_i;\beta(F)\right) + o_p(1).$$

Therefore, $\widehat{\beta}_n$ is asymptotically linear at F with influence function given by

$$\varphi_F(X) = -\left[\partial E_F[u(X;\beta)]/\,\partial\beta^T\big|_{\beta=\beta(F)}\right]^{-1} u(X;\beta(F)).$$

Example 3.3 (Median). Let $\mathcal{F}$ be the model for the law of a r.v. X which only assumes that the law is absolutely continuous with respect to the Lebesgue measure and the density $f(x)$ is continous.
Let $\beta(F)$ be the median of F, i.e. solution to $F(\beta) = 1/2$. Define the sample median $\widehat{\beta}_n$ as $X_{((n+1)/2)}$ if n is odd and as $\left(X_{(n/2)} + X_{(n/2+1)}\right)/2$ if n is even where $X_{(j)}$ is the j^{th} order statistics. It is easy to check that with such definition, $\widehat{\beta}_n$ solves

$$\sum_{i=1}^{n} \mathrm{sg}\left(\beta - X_i\right) = 1 \text{ if } n \text{ is odd,}$$

$$\sum_{i=1}^{n} \mathrm{sg}\left(\beta - X_i\right) = 0 \text{ if } n \text{ is even.}$$

So, $\widehat{\beta}_n$ solves $n^{-1/2}\sum_{i=1}^{n}\left[u\left(X_i,\beta\right)\right] = o_p(1)$ with $u(X,\beta) = \mathrm{sg}(\beta - X) \equiv 2\mathrm{I}(X \le \beta) - 1$.

Then,

$$\begin{aligned} E_F[u(X,\beta)] &= E_F[2\mathrm{I}(X \le \beta) - 1] \\ &= 2F(\beta) - 1, \end{aligned}$$

so,

$$E_F[u(X,\beta)]/\,\partial\beta|_{\beta=\beta(F)} = 2f(\beta(F)).$$

Later in the course we will see that the term A_n in the expansion of the estimating function in slide 150 is $O_p(1)$ when

$$u(X,\beta) = 2\mathrm{I}(X \le \beta) - 1.$$

So, invoking proposition 3.4 we conclude that the sample median is asymptotically linear and its influence function at F is given by

$$\varphi_F(X) = -\frac{\mathrm{I}(X \le \beta(F)) - 1/2}{f(\beta(F))}.$$

The next Lemma, provides a characterization of the set of influence functions of asymptotically linear estimators which are also regular (throughout referred to as regular asymptotically linear estimators and abbreviated as RAL).

The Lemma is a consequence of an important result in asymptotic statistics, known as Le-Cam's third Lemma. Le-Cam's third lemma gives us a very useful "automatic"tool for finding the limit distributions of estimators under local alternatives. I will state Le Cam's lemma after.

Lemma 3.5. Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta\}$ be a regular parametric model. Suppose that $\widehat{\beta}_n$ is an asymptotically linear estimator at $F = F_{\theta^*}$ of a parameter $\beta(\theta)$ which is differentiable at θ^*. Let $\varphi_F(X)$ be the influence function of $\widehat{\beta}_n$ Then, $\widehat{\beta}_n$ is regular at θ^* if and only if

$$\left.\frac{\partial \beta(\theta)}{\partial \theta^T}\right|_{\theta=\theta^*} = E_{\theta^*}\left\{\varphi_{F_{\theta^*}}(X) S_\theta\left(\theta^*\right)\right\},$$

where $S_\theta\left(\theta^*\right) = s_\theta(X;\theta^*)$ is the score for θ at θ^*.

Remark: Note that the if part of this Lemma is like an "asymptotic version" of Lemma 3.3.

Note: this lemma is essentially Theorem 2.2 of Newey, 1990 , except that Newey requires that the map $\theta \to E_\theta\left[\varphi_{F_{\theta^*}}(X)\varphi_{F_{\theta^*}}(X)^T\right]$ is continuous at θ^* and according to the proof I give below does not appear to be needed for the validity of the Lemma.

As an application of Lemma 3.5 let us study conditions that ensure the regularity of estimators $\widehat{\beta}_n$ that are solutions of unbiased estimating equations $\sum_i u\left(X_i;\beta\right) = 0$. We have seen that, under the conditions of Proposition 3.4, $\widehat{\beta}_n$ is asymptotically linear with infl. fcn at θ^*,

$$\varphi_{F_{\theta^*}}(X) = -\left\{\left.\frac{\partial}{\partial \beta^T} E_\theta\{u(X;\beta)\}\right|_{\beta=\beta(\theta^*)}\right\}^{-1} u\left(X;\beta\left(\theta^*\right)\right).$$

Thus, if θ is in $\mathbb{R}^p$, from Lemma 3.5, we have that the requirement for regularity of $\widehat{\beta}_n$ at θ^* is that

$$\begin{aligned} &\left.\frac{\partial}{\partial \theta^T}\beta(\theta)\right|_{\theta=\theta^*} \\ &= -\left\{\left.\frac{\partial}{\partial \beta^T} E_{\theta^*}\{u(X;\beta)\}\right|_{\beta=\beta(\theta^*)}\right\}^{-1} E_{\theta^*}\left\{u\left(X;\beta\left(\theta^*\right)\right) S_\theta\left(\theta^*\right)^T\right\}. \end{aligned}$$

The last condition holds under additional "regularity conditions". Specifically, from Lemma 3.3, if the parametric model is regular and the map $\theta \to E_\theta\left\{u\left(X;\beta\left(\theta^*\right)\right)^T u\left(X;\beta\left(\theta^*\right)\right)\right\}$ is continuous at θ^*, then

$$E_{\theta^*}\left\{u\left(X;\beta\left(\theta^*\right)\right) S_\theta\left(\theta^*\right)^T\right\} = \left.\frac{\partial}{\partial \theta^T} E_\theta\left\{u\left(X;\beta\left(\theta^*\right)\right)\right\}\right|_{\theta=\theta^*}. \quad (3)$$

Furthermore, if the following map has continuous partial derivatives in an open neighborhood of $(\beta,\theta) = \left(\beta\left(\theta^*\right),\theta^*\right)$, $(\beta,\theta) \to E_\theta\{u(X;\beta)\}$, then, by the implicit function theorem, $\beta(\theta)$ is continuously differentiable at θ^* and the derivative of $\beta(\theta)$ at θ^* can be obtained by differentiating using the chain rule the identity $0 = E_\theta\{u(X;\beta(\theta))\}$ for all θ, thus obtaining

$$\begin{aligned} &\left.\frac{\partial}{\partial \theta^T} E_\theta\left\{u\left(X;\beta\left(\theta^*\right)\right)\right\}\right|_{\theta=\theta^*} \\ &= -\left\{\left.\frac{\partial}{\partial \beta^T} E_\theta\{u(X;\beta)\}\right|_{\beta=\beta(\theta^*)}\right\}\left\{\left.\frac{\partial}{\partial \theta^T}\beta(\theta)\right|_{\theta=\theta^*}\right\}. \end{aligned} \quad (4)$$

Displays (3) and (4) of the preceding slide imply that

$$\frac{\partial}{\partial\theta^T}\beta(\theta)\bigg|_{\theta=\theta^*}$$
$$= -\left\{\frac{\partial}{\partial\beta^T}E_{\theta^*}\{u(X;\beta)\}\bigg|_{\beta=\beta(\theta^*)}\right\}^{-1} E_{\theta^*}\left\{u\left(X;\beta\left(\theta^*\right)\right)S_\theta\left(\theta^*\right)^T\right\}.$$

So, we conclude that if:

(i) the conditions of Proposition 3.4 holds,

(ii) the map $\theta \to E_\theta\left\{u\left(X;\beta\left(\theta^*\right)\right)^T u\left(X;\beta\left(\theta^*\right)\right)\right\}$ is continuous at θ^* and,

(iii) the map

$$(\beta,\theta) \to E_\theta\{u(X;\beta)\}$$

has continuous partial derivatives in an open neighborhood of $\left(\beta\left(\theta^*\right),\theta^*\right)$, then the Z-estimator $\widehat{\beta}_n$ is not only asymptotically linear but also regular.

Example 3.3 (Continued). Let $\mathcal{F} = \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a regular parametric model. Let $\beta(F)$ be the median of F, i.e. solution to $F(\beta) = 1/2$ and $\widehat{\beta}_n$ the sample median, which satisfies $\sqrt{n}\mathbb{P}_n\left[u\left(X_i,\beta\right)\right] = o_p(1)$ with $u(X,\beta) \equiv 2\mathrm{I}(X \leq \beta) - 1$.
If the dominating measure in model $\mathcal{F}$ is the Lebesgue measure and the densities $f(x;\theta)$ are continous fcns of x, then we have already seen that $\widehat{\beta}_n$ is asymptotically linear.
Let us check what additional assumptions suffice to ensure that the sample median is regular.
We first note that the requirement that the map

$$\theta \to E_\theta\left\{u\left(X;\beta\left(\theta^*\right)\right)^T u\left(X;\beta\left(\theta^*\right)\right)\right\}$$

be continuous holds trivially because $u(X;\beta)^2 = 1$ for all x and β.

Next, the map $(\beta,\theta) \to E_\theta\{u(X;\beta)\}$ is the map

$$(\beta,\theta) \to 2F_\theta(\beta) - 1$$

where $F_\theta(\cdot)$ is the CDF of $f(x;\theta)$. So,

$$\partial E_\theta[u(X;\beta)]/\partial\beta = 2f(\beta;\theta),$$

and the right hand side is a continuous function of θ because we have assumed that the densities in the model are continuous. On the other hand, by Lemma 3.3 (with T in that Lemma equal to $u(X;\beta)$ for a fixed β), we have

$$\begin{aligned}\partial E_\theta[u(X;\beta)]/\partial\theta &= E_\theta\left[u(X;\beta)s_\theta(X;\theta)\right] \\ &= 2E_\theta\left[\mathrm{I}(X \leq \beta)s_\theta(X;\theta)\right].\end{aligned}$$

So, if for all (β,θ) in an open neighborhood $\left(\beta\left(\theta^*\right),\theta^*\right)$ the map $\theta \to E_\theta\left[\mathrm{I}(X \leq \beta)s_\theta(X;\theta)\right]$ is continuous, then conditions (ii) and (iii) of slide 73 are met. Continuity of the last map holds, for instance, in exponential family models.

Lemma 3.5 has several important consequences which, for ease of reference, we summarize here and derive afterwards.

1) the projections of all influence functions of RAL estimators into the tangent space of the model coincide.

2) the variance of the projection of any inf. function of a RAL estimator into the tangent space is equal to the CR bound.

3) the set of all influence functions of RAL estimators of $\beta(\theta)$ at F_θ is equal to $\{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$.

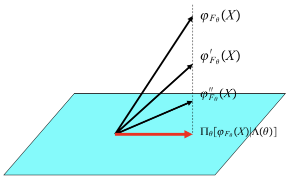
4) influence functions of RAL estimators are orthogonal, i.e. uncorrelated, with the "nuisance tangent space"(to be defined later).

5) The covariance of any influence function with the score for β is equal to the identity.

1) The projections of all influence functions of RAL estimators into the tangent space $\Lambda(\theta)$ coincide

This, in turn, implies that the variance of the UNIQUE projection is a lower bound for the variance of ANY RAL estimator of $\beta(\theta)$ at F_θ.

If $\varphi_{F_\theta}(X), \varphi'_{F_\theta}(X)$ and $\varphi''_{F_\theta}(X)$ are influence functions of RAL estimators of $\beta(\theta)$, then



The variance of $\Pi_\theta\left[\varphi_{F_\theta}(X) \mid \Lambda(\theta)\right]$ is $C_{\mathcal{F}}(\theta)$.

Proof that Lemma 3.5 implies that the projections of all influence functions of RAL estimators into the tangent space coincide.

$$\begin{aligned}\Pi_\theta\left[\varphi_{F_\theta}(X) \mid \Lambda(\theta)\right] =& E_\theta\left\{\varphi_{F_\theta}(X) S_\theta(\theta)^T\right\} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} S_\theta(\theta)\\ =& \frac{\partial\beta(\theta)}{\partial\theta^T} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} S_\theta(\theta)\\ =& E_\theta\left\{\varphi'_{F_\theta}(X) S_\theta(\theta)^T\right\} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} S_\theta(\theta)\\ =& \Pi_\theta\left[\varphi'_{F_\theta}(X) \mid \Lambda(\theta)\right].\end{aligned}$$

(The second and third equality are true by Lemma 3.5.)

2) The variance of the projection of any inf. function of a RAL estimator into the tangent space is equal to the CR bound

Proof that Lemma 3.5 implies that the variance of the projection of any influence function into the tangent space is equal to the CR bound

$$\begin{aligned}C_{\mathcal{F}}(\theta) &= \frac{\partial\beta(\theta)}{\partial\theta^T} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} \frac{\partial\beta(\theta)}{\partial\theta}\\ &= E_\theta\left\{\varphi_{F_\theta}(X) S_\theta(\theta)^T\right\} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} E_\theta\left\{\varphi_{F_\theta}(X) S_\theta(\theta)^T\right\}^T\\ &= \operatorname{var}_\theta\left\{\Pi_\theta\left[\varphi_{F_\theta}(X) \mid \Lambda(\theta)\right]\right\}.\end{aligned}$$

(The second equality is true by Lemma 3.5)

The projection of any influence function into the tangent space is called the efficient influence function and denoted with $\varphi_{F_\theta,\text{eff}}(X)$, i.e.

$$\varphi_{F_\theta,\text{eff}}(X) = \Pi_\theta\left[\varphi_F(X) \mid \Lambda(\theta)\right].$$

Note that

$$\begin{aligned}\frac{\partial\beta(\theta)}{\partial\theta^T} &= E_\theta\left\{\varphi_{F_\theta}(X) S_\theta(\theta)^T\right\}\\ &= E_\theta\left\{\varphi_{F_\theta,\text{eff}}(X) S_\theta(\theta)^T\right\} + \underbrace{E_\theta\left\{\left[\varphi_{F_\theta}(X) - \Pi_\theta\left[\varphi_{F_\theta}(X) \mid \Lambda(\theta)\right]\right] S_\theta(\theta)^T\right\}}_{=0 \text{ because } \varphi_F(X)-\Pi_\theta[\varphi_F(X)|\Lambda(\theta)] \text{ is in } \Lambda(\theta)^\perp}\\ &= E_\theta\left\{\varphi_{F_\theta,\text{eff}}(X) S_\theta(\theta)^T\right\}.\end{aligned}$$

3) The set of all influence functions of RAL estimators of $\beta(\theta)$ at F_θ is equal to $\{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$

Suppose that $\varphi_{F_\theta}(X)$ and $\varphi'_{F_\theta}(X)$ are two influence functions of RAL estimators of an $\mathbb{R}^k$-valued $\beta(\theta)$. Then,

$$E_\theta\left\{\left[\varphi_{F_\theta}(X) - \varphi'_{F_\theta}(X)\right] S_\theta(\theta)^T\right\} = \frac{\partial\beta(\theta)}{\partial\theta^T} - \frac{\partial\beta(\theta)}{\partial\theta^T} = 0.$$

Thus, for any inf fcn $\varphi'_{F_\theta}(X)$ we have

$$\varphi'_{F_\theta}(X) = \varphi_{F_\theta}(X) + \left\{\varphi'_{F_\theta}(X) - \varphi_{F_\theta}(X)\right\},$$

and we have just shown that the r.vector in curly brackets has all its components orthogonal to the tangent space $\Lambda(\theta)$ so if, in an abuse of notation we let $\Lambda(\theta)^\perp$ denote the $\mathbb{R}^k$-valued random vectors with finite var. and mean zero under F_θ, we have

$$\varphi'_{F_\theta}(X) \in \{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp.$$

Next take any $\mathbb{R}^k$-valued random vector, say $\psi_{F_\theta}(X)$, in the set $\{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$. Then,

$$\psi_{F_\theta}(X) = \varphi_{F_\theta}(X) + \omega_{F_\theta}(X)$$

for some $\omega_{F_\theta}(X) \in \Lambda(\theta)^\perp$.

Now, $\varphi_{F_\theta}(X)$ is an influence function of a RAL estimator, say $\widehat{\beta}_n$, so

$$\widetilde{\beta}_n \equiv \widehat{\beta}_n + n^{-1}\sum_{i=1}^{n} \omega_{F_\theta}(X_i)$$

is an estimator of $\beta(\cdot)$ which is asymptotically linear at F_θ and whose inf. fcn at F_θ is $\varphi_{F_\theta}(X) + \omega_{F_\theta}(X)$. This shows that $\psi_{F_\theta}(X)$ is the inf fcn of some a. linear estimator of $\beta(\cdot)$ at F_θ. To show that $\psi_{F_\theta}(X)$ is the inf fcn of a regular estimator, by Lemma 3.5 it suffices to show that

$$\partial\beta(\theta)/\partial\theta = E_{F_\theta}\left[\psi_{F_\theta}(X) S_\theta(\theta)^T\right].$$

This holds because

$$\begin{aligned} E_{F_\theta}\left[\psi_{F_\theta}(X)S_\theta(\theta)^T\right] &= E_{F_\theta}\left[\varphi_{F_\theta}(X)S_\theta(\theta)^T\right] + E_{F_\theta}\left[\omega_{F_\theta}(X)S_\theta(\theta)^T\right] \\ &= E_{F_\theta}\left[\varphi_{F\theta}(X)S_\theta(\theta)^T\right] + 0, \text{ because } \omega_{F_\theta}(X) \in \Lambda(\theta)^\perp \\ &= \partial\beta(\theta)/\partial\theta, \text{ because } \varphi_{F_\theta}(X) \text{ is an inf fcn.} \end{aligned}$$

This proves that $\psi_{F_\theta}(X)$ is an inf fcn of a RAL estimator.

Remark: From the lemma 3.5 we can immediately deduce (even without invoking the convolution theorem) that the CR Bound is a lower bound for the a. variance of any RAL estimator, since

a) the a. variance of a RAL estimator with inf fcn $\varphi_F(X)$ is $\mathrm{var}_F\{\varphi_F(X)\}$,

b) by the previous argument $C_{\mathcal{F}}(\theta) = \mathrm{var}_\theta\{\Pi_\theta[\varphi_{F_\theta}(X) \mid \Lambda(\theta)]\}$, and

c) projections "contract"the length of vectors.

4) Influence functions of RAL estimators are orthogonal, i.e. uncorrelated, with the "nuisance tangent space"

Suppose that the model is indexed by variation independent parameters β and η. That is, $\Theta = \mathrm{E} \times \mathrm{B}$ and

$$\underbrace{\theta^T}_{1\times p} = (\underbrace{\beta^T}_{1\times k}, \underbrace{\eta^T}_{1\times q}), \beta \text{ in B and } \eta \text{ in E}.$$

Let $S_\eta(\theta)$ be the score for η in the model in which β is fixed and known. Define the nuisance tangent space as the tangent space in the model in which β is fixed and known and η is unknown,

$$\Lambda_{\mathcal{F},\mathsf{nuis}}(\theta) \equiv \left\{a^T S_\eta(\theta) : a \in \mathbb{R}^q\right\}.$$

Lemma 3.5 implies that if $\varphi_{F_\theta}(X)$ is an influence function of a RAL estimator of β, then

$$\frac{\partial \beta(\eta)}{\partial \eta^T} = E\left\{\varphi_{F_\theta}(X) S_\eta(\theta)^T\right\}.$$

On the other hand, β does not change with η, so

$$\frac{\partial \beta(\eta)}{\partial \eta^T} = 0,$$

then

$$0 = E\left\{\varphi_{F_\theta}(X) S_\eta(\theta)^T\right\}.$$

Thus, inf. fcns $\varphi_{F_\theta}(X)$ of RAL estimators of β satisfy

$$\Pi_\theta\left[\varphi_{F_\theta}(X) \mid \Lambda_{\mathcal{F},\text{nuis}}(\theta)\right] = 0.$$

5) The covariance of any influence function with the score for β is equal to the identity

Assume as in point (4) that the model is indexed by variation independent parameters β and η. Let $S_\beta(\theta)$ be the score for β in the model in which η is fixed and known and β is unknown. If $\varphi_{F_\theta}(X)$ is the inf fcn of a RAL estimator of β, then

$$\frac{\partial \beta(\beta)}{\partial \beta^T} = E\left\{\varphi_{F_\theta}(X) S_\beta(\theta)^T\right\},$$

but $\frac{\partial \beta(\beta)}{\partial \beta^T} = Id_{k\times k}$, so,

$$Id_{k\times k} = E\left\{\varphi_{F_\theta}(X) S_\beta(\theta)^T\right\}.$$

Proof of Lemma 3.5

Lemma 3.5 is a corollary of a key result in asymptotic statistics, known as Le-Cam's third Lemma which, for completeness, I state here (in fact, I state a simplified version of the Lemma which is just what we need to prove Lemma 3.5). For a proof of Le Cam's Lemma see ch. 6 of van der Vaart, 2000.

Le-Cam's third Lemma: Suppose that $T_n = t_n(X_1, \ldots, X_n)$ be an $\mathbb{R}^k$ valued of iid random vectors $X_1^{(n)}, \ldots, X_n^{(n)}$. Let $f_n(x)$ and $g_n(x), n = 1, 2, \ldots$ be two sequences of possible marginal densities for $X_i^{(n)}$. Let

$$L_n \equiv \log \prod_{i=1}^n \left\{ g_n\left(X_i^{(n)}\right) / f_n\left(X_i^{(n)}\right)\right\}.$$

Suppose

$$\begin{pmatrix} T_n \\ L_n \end{pmatrix} \overset{f_n}{\to} N_{k+1}\left(\begin{pmatrix} \Delta \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Omega & \Upsilon \\ \Upsilon^T & \sigma^2 \end{pmatrix}\right),$$

then

$$T_n \overset{g_n}{\to} N_k(\Delta + \Upsilon, \Omega).$$

By assumption $\widehat{\beta}_n$ is asymp. linear with inf. fcn. $\varphi_{F^*}(X)$ at $F^* = F_{\theta^*}$
On the other hand, $\mathcal{F} \equiv \{f(x;\theta) : \theta \in \Theta \subseteq \mathbb{R}^p\}$ is a regular model, so by proposition 3.1,

$$\begin{pmatrix} \sqrt{n}\left\{\widehat{\beta}_n - \beta\left(\theta^*\right)\right\} \\ \log \prod_{i=1}^n \left\{\frac{f(X_i;\theta^* + h/\sqrt{n})}{f(X_i;\theta^*)}\right\} \end{pmatrix} = \frac{1}{\sqrt{n}} \sum_{i=1}^n \begin{pmatrix} \varphi_{F_{\theta^*}}(X_i) \\ h^T s_\theta(X_i;\theta^*) \end{pmatrix} - \begin{pmatrix} 0 \\ \frac{1}{2} h^T I(\theta^*) h \end{pmatrix} + o_{p_{\theta^*}}(1) \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$\overset{\to}{f(x;\theta^*)} N_{k+1}\left(\begin{pmatrix} 0 \\ -\frac{1}{2} h^T I(\theta^*) h \end{pmatrix}, \begin{pmatrix} \operatorname{var}_{\theta^*}\left(\varphi_{F_{\theta^*}}(X)\right) & E\left(\varphi_{F_{\theta^*}}(X) s_\theta(X_i;\theta^*)^T h\right) \\ E\left(\varphi_{F_{\theta^*}}(X) s_\theta(X_i;\theta^*)^T h\right)^T & h^T I(\theta^*) h \end{pmatrix}\right),$$

Applying Le Cam's third lemma with $f_n(x) = f\left(x;\theta^*\right)$ and $g_n(x) = f\left(x;\theta^* + h/\sqrt{n}\right)$, we conclude that

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta\left(\theta^*\right)\right\} \overset{f\left(x;\theta^* + h/\sqrt{n}\right)}{\to}$$
$$N_k\left(E\left(\varphi_{F_{\theta^*}}(X) s_\theta\left(X_i;\theta^*\right)^T h\right), \operatorname{var}_{\theta^*}\left(\varphi_{F_{\theta^*}}(X)\right)\right).$$

We then conclude that

$$\sqrt{n}\left\{\widehat{\beta}_n-\beta\left(\theta^*+h/\sqrt{n}\right)\right\}=\sqrt{n}\left\{\widehat{\beta}_n-\beta\left(\theta^*\right)\right\}-\sqrt{n}\left\{\beta\left(\theta^*+h/\sqrt{n}\right)-\beta\left(\theta^*\right)\right\}$$
$$\stackrel{f(x;\theta^*+h/\sqrt{n})}{\rightarrow} N_k\left(E\left(\varphi_{F_{\theta^*}}(X)s_\theta\left(X_i;\theta^*\right)^T h\right),\operatorname{var}_{\theta^*}\left(\varphi_{F_{\theta^*}}(X)\right)\right)-\frac{\partial\beta(\theta)}{\partial\theta^T}h.$$

Now, $\widehat{\beta}_n$ is regular if and only if the limiting distribution of

$$\sqrt{n}\left\{\widehat{\beta}_n-\beta\left(\theta^*+h/\sqrt{n}\right)\right\}$$

under $f\left(x;\theta^*+h/\sqrt{n}\right)$ does not depend on h and in view of the last display, if and only if

$$\left[E\left(\varphi_{F_{\theta^*}}(X)s_\theta\left(X_i;\theta^*\right)^T\right)-\frac{\partial\beta(\theta)}{\partial\theta^T}\right]h$$

is not a function of h. The last display is not a function of h if and only if the matrix in squared brackets is 0. This concludes the proof.

Suppose that in a parametric model the index θ is

$$\theta=\left(\beta^T,\eta^T\right)^T,$$

Where β and θ are variation independent. We will now see that the CR bound for β has one additional useful geometric interpretation which we will later extend to the semiparametric setting.

Recall that the CR bound is equal to

$$\underbrace{C_{\mathcal{F}}(\theta)}_{k\times k}=\underbrace{\frac{\partial\beta(\theta)}{\partial\theta^T}}_{k\times p}\underbrace{I(\theta)}_{p\times p}^{-1}\underbrace{\frac{\partial\beta(\theta)}{\partial\theta}}_{p\times k}.$$

Now, when

$$\underbrace{\theta^T}_{1\times p}=(\underbrace{\beta^T}_{1\times k},\underbrace{\eta^T}_{1\times q}),$$

and $\beta(\theta)=\beta$ then

$$\underbrace{\frac{\partial\beta(\theta)}{\partial\theta^T}}_{k\times p}=\left[\begin{array}{cc} Id_{k\times k} & \mathbf{0}_{k\times q}\end{array}\right],$$

where $Id_{k\times k}$ is the $k\times k$ identity matrix.

Furthermore,

$$\begin{aligned} I(\theta)^{-1} &= E_\theta\left\{S_\theta(\theta)S_\theta(\theta)^T\right\}^{-1}\\ &= \left[\begin{array}{cc} E\left[S_\beta(\theta)S_\beta(\theta)^T\right] & E\left[S_\beta(\theta)S_\eta(\theta)^T\right]\\ E\left[S_\eta(\theta)S_\beta(\theta)^T\right] & E\left[S_\eta(\theta)S_\eta(\theta)^T\right]\end{array}\right]^{-1}\\ &\equiv \left[\begin{array}{cc} I_{\beta,\beta}(\theta) & I_{\beta,\eta}(\theta)\\ I_{\eta,\beta}(\theta) & I_{\eta,\eta}(\theta)\end{array}\right]^{-1}\\ &= \left[\begin{array}{cc} I^{\beta,\beta}(\theta) & I^{\beta,\eta}(\theta)\\ I^{\eta,\beta}(\theta) & I^{\eta,\eta}(\theta)\end{array}\right].\end{aligned}$$

Then,

$$\underbrace{C_{\mathcal{F}}(\theta)}_{k\times k} = \underbrace{\frac{\partial\beta(\theta)}{\partial\theta^T}}_{k\times p}\underbrace{I(\theta)^{-1}}_{p\times p}\underbrace{\frac{\partial\beta(\theta)}{\partial\theta}}_{p\times k}$$
$$= \left[\begin{array}{cc} Id_{k\times k} & \mathbf{0}_{k\times q}\end{array}\right]\left[\begin{array}{cc} I^{\beta,\beta}(\theta) & I^{\beta,\eta}(\theta) \\ I^{\eta,\beta}(\theta) & I^{\eta,\eta}(\theta)\end{array}\right]\left[\begin{array}{c} Id_{k\times k} \\ \mathbf{0}_{q\times k}\end{array}\right]$$
$$= I^{\beta,\beta}(\theta).$$

From the formula of a partitioned matrix we have that

$$I^{\beta,\beta}(\theta) = \left[I_{\beta,\beta}(\theta) - I_{\beta,\eta}(\theta)I_{\eta,\eta}(\theta)^{-1}I_{\eta,\beta}(\theta)\right]^{-1},$$

and replacing each term by their expressions we obtain

$$I^{\beta,\beta}(\theta)$$
$$= \left[\mathrm{var}_\theta\left\{S_\beta(\theta)\right\} - E\left\{S_\beta(\theta)S_\eta(\theta)^T\right\}\mathrm{var}_\theta\left\{S_\eta(\theta)\right\}^{-1}E\left\{S_\eta(\theta)S_\beta(\theta)^T\right\}\right]^{-1}$$
$$= \left\{\mathrm{var}_\theta\left[S_\beta(\theta) - E\left\{S_\beta(\theta)S_\eta(\theta)^T\right\}\mathrm{var}_\theta\left\{S_\eta(\theta)\right\}^{-1}S_\eta(\theta)\right]\right\}^{-1}$$
$$= \left\{\mathrm{var}_\theta\left[S_\beta(\theta) - \Pi_\theta\left[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\mathsf{nuis}}(\theta)\right]\right]\right\}^{-1}.$$

In conclusion:

$$C_{\mathcal{F}}(\theta) = \mathrm{var}_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)\right\}^{-1},$$

where

$$S_{\beta,\mathsf{eff}}(\theta) \equiv S_\beta(\theta) - \Pi_\theta\left[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\mathsf{nuis}}(\theta)\right].$$

The random variable $S_{\beta,\mathsf{eff}}(\theta)$ is called the efficient score for β in model $\mathcal{F}$ at F_θ.

Note that

$$E_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)S_\beta(\theta)^T\right\}$$
$$= E_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)S_{\beta,\mathsf{eff}}(\theta)^T\right\} + \underbrace{E_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)\Pi\left[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\mathsf{nuis}}(\theta)\right]^T\right\}}_{=0 \text{ because } S_{\beta,\mathsf{eff}}(\theta)\in\Lambda_{\mathcal{F},\mathsf{nuis}}(\theta)^\perp}$$
$$= E_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)S_{\beta,\mathsf{eff}}(\theta)^T\right\},$$

and that

$$\mathrm{var}_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)\right\}^{-1} = \mathrm{var}_\theta\left[\mathrm{var}_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)\right\}^{-1}S_{\beta,\mathsf{eff}}(\theta)\right].$$

Indeed, you may check as a hmw that the efficient Influence function for β satisfies

$$\varphi_{F_\theta,\mathsf{eff}}(X) = \mathrm{var}_\theta\left\{S_{\beta,\mathsf{eff}}(\theta)\right\}^{-1}S_{\beta,\mathsf{eff}}(\theta).$$

The inverse of the CR bound for β is the information (per sample unit) for β in model $\mathcal{F}$. We shall denote it with $I_{\mathcal{F},\beta}(\theta)$. Let's compare $I_{\mathcal{F},\beta}(\theta)$ with the information about β in the parametric model $\mathcal{F}^*$ in which η is known.

The CR bound for β in $\mathcal{F}^*$ is

$$C_{\mathcal{F}^*}(\theta) = \mathrm{var}_\theta \left\{ S_\beta(\theta) \right\}^{-1},$$

so

$$I_{\mathcal{F}^*,\beta}(\theta) = \mathrm{var}_\theta \left\{ S_\beta(\theta) \right\}.$$

$$\begin{aligned} I_{\mathcal{F},\beta}(\theta) &= \mathrm{var}_\theta \left\{ S_{\beta,\mathsf{eff}}(\theta) \right\} \\ &= \left\| S_\beta(\theta) - \Pi_\theta \left[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\mathsf{nuis}}(\theta) \right] \right\|_\theta^2 \\ &\leq \left\| S_\beta(\theta) \right\|_\theta^2 \\ &= I_{\mathcal{F}^*,\beta}(\theta). \end{aligned}$$

Thus, there is no loss of information in not knowing η if and only if $S_\beta(\theta)$ is orthogonal to the nuisance tangent space $\Lambda_{\mathcal{F},\mathsf{nuis}}(\theta)$, i.e. if and only if $S_\beta(\theta)$ and $S_\eta(\theta)$ are orthogonal.

1) Let $\varphi_{F_\theta}(X)$ be the inf. func. of an a. linear estimator. Then $\frac{\partial \beta(\theta)}{\partial \theta^T} = E_\theta \left\{ \varphi_{F_\theta}(X) S_\theta(\theta)^T \right\}$ iff the estimator is regular.

2) if $\varphi_{F_\theta}(X)$ and $\varphi'_{F_\theta}(X)$ are the infl. functions of two RAL estimators, then

$$\Pi_\theta \left[\varphi_{F_\theta}(X) \mid \Lambda(\theta) \right] = \Pi_\theta \left[\varphi'_{F_\theta}(X) \mid \Lambda(\theta) \right].$$

3) The efficient infl. fon is defined as

$$\varphi_{F_\theta,\mathsf{eff}}(X) \equiv \Pi_\theta \left[\varphi_{F_\theta}(X) \mid \Lambda(\theta) \right].$$

4) The set of influence functions is equal to $\{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$ and to $\{\varphi_{F_\theta,\mathsf{eff}}(X)\} \oplus \Lambda(\theta)^\perp$.

5) The CR bound is equal to

$$C_{\mathcal{F}}(\theta) = \mathrm{var}_\theta \left\{ \varphi_{F_\theta,\mathsf{eff}}(X) \right\} = \mathrm{var}_\theta \left\{ \Pi_\theta \left[\varphi_{F_\theta}(X) \mid \Lambda(\theta) \right] \right\}.$$

If $\underbrace{\theta^T}_{1\times p} = (\underbrace{\beta^T}_{1\times k}, \underbrace{\eta^T}_{1\times q})$ then,

6) The set of inf. functions of RAL estimators of β is equal to

$$\left\{ \varphi_{F_\theta}(X) \in \mathcal{L}_2^0(\theta) : \Pi_\theta \left[\varphi_{F_\theta}(X) \mid \Lambda_{\mathcal{F},\mathsf{nuis}}(\theta) \right] = 0 \ \& \ E\left[\varphi_{F_\theta}(X) S_\beta(\theta)^T \right] = Id \right\}.$$

7) The efficient score for β is defined as

$$S_{\beta,\mathsf{eff}}(\theta) \equiv S_\beta(\theta) - \Pi_\theta \left[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\mathsf{nuis}}(\theta) \right].$$

8) The efficient influence function is equal to

$$\varphi_{F_\theta,\mathsf{eff}}(X) = \mathrm{var}_\theta \left\{ S_{\beta,\mathsf{eff}}(\theta) \right\}^{-1} S_{\beta,\mathsf{eff}}(\theta).$$

9) The CR bound for β is equal to

$$C_{\mathcal{F}}(\theta) = \mathrm{var}_\theta \left\{ S_{\beta,\mathsf{eff}}(\theta) \right\}^{-1}.$$

The geometry of inference