Semiparametric Statistics

Bodhisattva Sen

April 4, 2018

1 Introduction

By a **semiparametric model** we mean a *statistical model*[1] that involves both *parametric* and *nonparametric* (infinite-dimensional[2]) components. However, we are mostly interested in estimation and inference of a finite-dimensional parameter in the model.

Example 1.1 (Population mean)**.** Suppose that $X_1, \ldots, X_n$ are i.i.d. P belonging to the class $\mathcal{P}$ of distribution. Let $\psi(P) \equiv \mathbb{E}_P[X_1]$, the mean of the distribution, be the parameter of interest.
Question: Suppose that $\mathcal{P}$ is the class of all distributions that have a finite variance. What is the most **efficient** estimator of $\psi(P)$, i.e., what is the estimator with the best asymptotic performance?

Example 1.2 (Partial linear regression model)**.** Suppose that we observe i.i.d. data $\{X_i \equiv (Y_i, Z_i, V_i) : i = 1, \ldots, n\}$ from the following partial linear regression model:

$$Y_i = Z_i^\top \beta + g(V_i) + \epsilon_i, \tag{1}$$

where Y_i is the scalar response variable, Z_i and V_i are vectors of predictors, $g(\cdot)$ is the unknown (nonparametric) function, and ϵ_i is the unobserved error. For simplicity and to focus on the semiparametric nature of the problem, we assume that $(Z_i, V_i) \sim f(\cdot, \cdot)$, where we assume that the density $f(\cdot, \cdot)$ is known, is independent of $\epsilon_i \sim N(0, \sigma^2)$ (with σ^2 known). The model, under these assumptions, has a parametric component β and a

[1] A model $\mathcal{P}$ is simply a collection of probability distributions for the data we observe.

[2] By an infinite-dimensional linear space we mean a space which cannot be spanned by any finite set of elements in the set. An example of an infinite-dimensional linear space is the space of continuous functions defined on the real line.

nonparametric component $g(\cdot)$.
Question: How well can we estimate the parametric component β?

Example 1.3 (Symmetric location)**.** Suppose we have data $X_1, \ldots, X_n$ generated from a law which has a *symmetric* and smooth density and one is interested in estimating the center of symmetry of the density. This means that one can write the density of the data as $f(\cdot - \theta)$, for some even function f (i.e., $f(-x) = f(x)$) and some unknown parameter θ.

If one assumes that the shape f is known, then only θ is unknown and the model is parametric. But a much more flexible model is obtained by taking f also unknown. The corresponding model $\mathcal{P}$ is a *separated semiparametric*[3] model indexed by the pair (θ, f). Historically, this model is one of the first semiparametric models that have been considered. In the seminal 1956 paper, Charles Stein pointed out the surprising fact that, at least asymptotically (as $n \to \infty$), it is no harder to estimate θ when f is known than when it is not[4].

Example 1.4 (The simplest 'semiparametric' model)**.** Suppose that one observes a vector $X = (X_1, X_2)$ in $\mathbb{R}^2$ whose law belongs to the Gaussian family $\{N(\theta, \Sigma) : \theta \in \mathbb{R}^2\}$ and Σ is a positive definite matrix (which we denote $\Sigma > 0$). The matrix Σ is assumed to be known. Let us write the vector θ as $[\theta_1\ \theta_2]^\top$.

Goal: We are interested in estimation of the **first coordinate** $\mu := \theta_1$ from the single observation $X \sim N(\theta, \Sigma)$ (results with a similar interpretation can be derived for n observations). Consider the two following cases:

1. The second coordinate θ_2 is known.
2. The second coordinate θ_2 is unknown.

The natural questions are:

- Does it make a difference?
- In both cases, $\hat{\mu} = X_1$ seems to be a reasonable estimator. Is it optimal, say already among unbiased estimators when the quadratic risk is considered?

[3] We say that the model $\mathcal{P} = \{P_{\nu,\eta}\}$ is a *separated semiparametric* model, where ν is a Euclidean parameter and η runs through a nonparametric class of distributions (or some infinite-dimensional set). This gives a semiparametric model in the strict sense, in which we aim at estimating ν and consider η as a nuisance parameter.

[4] This is usually termed as **adaptive estimation**. Adaptive estimation refers to models where parameters of interest can be estimated equally well when the nonparametric part of the model is unknown as when it is known. Such models are those where the semiparametric bound is equal to the parametric bound that applies when the nonparametric part of the model is known.

- Does Σ play a role?

Consider the estimation of a parameter of interest $\nu = \nu(P)$, where the data has distribution $P \in \mathcal{P}$. Here are some frequent goals or questions:

(Q 1) How well can we estimate $\nu = \nu(P)$? What is our "gold standard"?

(Q 2) Can we compare absolute "in principle" standards for estimation of ν in a model $\mathcal{P}$ with estimation of ν in a submodel $\mathcal{P}_0 \subset \mathcal{P}$? What is the effect of not knowing η on estimation of ν when $\mathcal{P} = \{P_\theta : \theta \equiv (\nu, \eta) \in \Theta\}$?

(Q 3) How do we construct *efficient* estimators of $\nu(P)$?

Efficiency bounds (i.e., asymptotic lower bounds on variances of estimators of parameters) are of fundamental importance for semiparametric models. Such bounds quantify the efficiency loss that can result from a semiparametric, rather than parametric, approach. The extent of this loss is important for the decision to use semiparametric models. The bounds also provide a guide to estimation methods. They give a standard against which the asymptotic efficiency of any particular estimator can be measured.

Semiparametric efficiency bounds were introduced by Stein (1956), and developed by Koshevnik and Levit (1976), Pfanzagl (1982), Begun et al. (1983), and Bickel et al. (1993). The treatment in this course will closely follow the texts Tsiatis (2006), van der Vaart (1998, Chapter 25) and Bolthausen et al. (2002).

One could imagine that the data are generated by a parametric model that satisfies the semiparametric assumptions and contains the truth. Such a model is referred to as a **parametric submodel**, where the 'sub' prefix refers to the fact that it is a subset of the model consisting of all distributions satisfying the assumptions. One can obtain the classical Cramer-Rao lower bound for a parametric submodel. Any **semiparametric estimator**, i.e., one that is consistent and asymptotically normal under the semiparametric assumptions, has an asymptotic variance that is comparable to the Cramer-Rao lower bound of a semiparametric model, and therefore has an asymptotic variance no smaller than the bound for the submodel. Since this comparison holds for each parametric submodel that one could imagine, it follows that:

The asymptotic variance of any semiparametric estimator is no smaller than the supremum of the Cramer-Rao lower bounds for all parametric submodel.

2 Parametric theory: The classical approach

We consider a model $\mathcal{P}$ defined as the collection of probability measures $\{P_\theta : \theta \in \Theta\}$ on some measurable space $(\mathfrak{X}, \mathcal{A})$ where Θ is an open subset of $\mathbb{R}^k$. For each $\theta \in \Theta$, let p_θ be a density of P_θ with respect to some dominating σ-finite measure μ.

Question: Suppose that we are interested in estimating $\psi(\theta)$ based on the data $X \sim P_\theta$, where $\psi : \Theta \to \mathbb{R}$ is a known function. A natural question that arises in this regard is: What is the "gold-standard" for the performance of an estimator $T(X)$ for $\psi(\theta)$?".

The log-likelihood for one observation is denoted by

$$\ell_\theta(x) = \log p_\theta(x), \qquad \text{for } x \in \mathfrak{X}.$$

Suppose that $\theta \mapsto p_\theta(x)$ is *differentiable* at θ for all $x \in \mathfrak{X}$ (for μ-almost all x suffices); we denote this derivative by $\dot{\ell}_\theta(x)$ and call it the **score function**.

Definition 2.1 (Fisher information)**.** The **Fisher information** matrix at θ is defined as

$$I(\theta) \equiv I_\theta = \mathbb{E}[\dot{\ell}_\theta(X)\dot{\ell}_\theta(X)^\top], \qquad \text{where } X \sim P_\theta.$$

The best known lower bound on the performance of any unbiased estimator of $\psi(\theta)$ is the famous **Cramér-Rao inequality**.

Theorem 2.2 (Cramér-Rao inequality)**.** Suppose that:

(A1) $X \sim P_\theta$ on $(\mathfrak{X}, \mathcal{A})$ where $P_\theta \in \mathcal{P} := \{P_\theta : \theta \in \Theta\}$, Θ being an open subset of $\mathbb{R}^k$;

(A2) for each $\theta \in \Theta$, $p_\theta \equiv dP_\theta/d\mu$ exists where μ is a σ-finite measure;

(A3) $\theta \mapsto P_\theta$ is differentiable with respect to θ (for μ-almost all x), i.e., there exists a set B with $\mu(B) = 0$ such that, for $x \in B^c$, $\frac{\partial}{\partial\theta} p_\theta(x)$ exists for all $\theta \in \Theta$;

(A4) $A := \{x \in \mathfrak{X} : p_\theta(x) = 0\}$ does not depend on θ;

(A5) the $k \times k$ information matrix $I(\theta)$ is positive definite;

(A6) the map $\psi : \Theta \to \mathbb{R}$ is differentiable at $\theta \in \Theta$ with derivative $\dot{\psi}_\theta \equiv \nabla\psi(\theta)$ (we think of $\dot{\psi}_\theta$ as a row vector);

(A7) $T(X)$ is an unbiased estimator of $\psi(\theta)$, i.e., $b(\theta) := \mathbb{E}_\theta[T(X)] - \psi(\theta) = 0$.

(A8) $\int p_\theta(x)d\mu(x)$ and $\int T(x)p_\theta(x)d\mu(x)$ can both be differentiated with respect to θ under the integral sign, i.e.,

$$\frac{\partial}{\partial\theta}\int p_\theta(x)d\mu(x) = \int \frac{\partial}{\partial\theta}p_\theta(x)d\mu(x),$$

and

$$\frac{\partial}{\partial\theta}\int T(x)p_\theta(x)d\mu(x) = \int T(x)\frac{\partial}{\partial\theta}p_\theta(x)d\mu(x).$$

Then the variance of $T(X)$ at θ is bounded below by

$$\mathrm{Var}(T(X)) \geq \dot{\psi}_\theta I_\theta^{-1}\dot{\psi}_\theta^\top. \tag{2}$$

Proof. Exercise. □

Remark 2.1. Let us consider the case when $k = 1$. In this case, the above theorem states that for any unbiased estimator of $\psi(\theta)$, $\mathrm{Var}_\theta(T_n) \geq \dot{\psi}_\theta^2/I_\theta$.

Remark 2.2 (When does equality hold in the Cramer-Rao bound?)**.** Note that equality in the Cauchy-Schwarz inequality holds if and only if $\dot{\ell}_\theta(\cdot)$ and $T(\cdot)$ are linearly related, i.e., iff

$$\dot{\ell}_\theta(x) = A(\theta)[T(x) - \mathbb{E}_\theta(T(X))] \qquad \text{a.s. } P_\theta$$

for some constant $A(\theta)$. By (A2) this implies that this holds a.e. μ. Under further regularity conditions this holds if and only if P_θ is an exponential family; see e.g., Lehmann and Casella (1998, Theorem 5.12, page 121).

Remark 2.3. If, in addition to conditions (A1)-(A9), $\int p_\theta(x)d\mu(x)$ can be differentiated twice under the integral sign, then the Fisher information matrix can also be expressed as

$$I(\theta) = -\mathbb{E}_\theta[\ddot{\ell}_\theta(X)] = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_\theta(X)\right].$$

Remark 2.4 (I.i.d. case)**.** When $X = (X_1, \ldots, X_n)$ with the X_i's i.i.d. $P_\theta \in \mathcal{P}$ satisfying (A1)–(A9), then

$$\begin{aligned}\dot{\ell}_\theta(X) &= \sum_{i=1}^n \dot{\ell}_\theta(X_i),\\ I_n(\theta) &= nI_1(\theta) \equiv nI(\theta),\end{aligned}$$

and the conclusion can be written, for an unbiased estimator $T_n \equiv T(X_1, \ldots, X_n)$, as

$$\mathrm{Var}[\sqrt{n}(T_n - \psi(\theta))] \geq \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top. \tag{3}$$

We can now ask: "Can we find a function of the X_i's such that this lower bound, in the above display, is attained?". Observe that the function (for sample size $n = 1$) defined as

$$\tilde{\ell}_\psi(X_1) = \dot{\psi}_\theta I_\theta^{-1} \dot{\ell}_\theta(X_1)$$

satisfies the bound. Of course, $\tilde{\ell}_\psi(\cdot)$ is not an estimator (as it depends on the unknown parameters).

We will call $\tilde{\ell}_\psi(\cdot)$ the **efficient influence function** for estimation of $\psi(\theta)$, i.e., if T_n is an asymptotically efficient estimator of $\psi(\theta)$ then T_n is **asymptotically linear**[5] with **influence function**[6] exactly $\tilde{\ell}_\psi$:

$$\sqrt{n}(T_n - \psi(\theta)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{\ell}_\psi(X_i) + o_p(1) \xrightarrow{d} N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top). \tag{5}$$

Exercise: Consider $X_1, \ldots, X_n$ i.i.d. $N(\mu, \sigma^2)$. Let $\hat{\mu}_n$ and $\hat{\sigma}^2$ denote the maximum likelihood estimators of μ and σ^2, respectively. Prove that $\hat{\mu}_n$ and $\hat{\sigma}^2$ are asymptotically linear and show that their i-th influence functions are given by $(X_i - \mu)$ and $(X_i - \mu)^2 - \sigma^2$ respectively.

Most reasonable estimators for the parameter $\xi(\theta)$, in either parametric or semiparametric models, are asymptotically linear and can be uniquely characterized by the influence function of the estimator. The following result makes this rigorous (Exercise).

Lemma 2.3 (Exercise)**.** An asymptotically linear estimator has a unique (a.s.) influence function.

[5] Given i.i.d. data $X_1, \ldots, X_n \sim P$, an estimator $S_n \equiv S_n(X_1, \ldots, X_n)$ of $\xi(P)$ (where $\xi : \mathcal{P} \to \mathbb{R}^q$, $q \geq 1$) is *asymptotically linear* if there exists a random vector (i.e., a q-dimensional measurable random function) $\varphi_{q\times 1}(X)$, such that $\mathbb{E}[\varphi(X)] = 0$ and

$$\sqrt{n}(S_n - \xi(P)) = n^{-1/2} \sum_{i=1}^{n} \varphi(X_i) + o_p(1), \tag{4}$$

where $o_p(1)$ is a term that converges in probability to zero as n goes to infinity and $\mathbb{E}[\varphi(X_1)\varphi(X_1)^\top]$ is finite and nonsingular.

[6] The random vector $\varphi(X_i)$ in (4) is referred to as the i-th *influence function* of the estimator S_n or the influence function of the i-th observation of the estimator S_n. The term influence function comes from the robustness literature, where, to first order, $\varphi(X_i)$ is the influence of the i-th observation on S_n.

Remark 2.5. The above definitions should not depend on the parametrization of the model $\mathcal{P}$. Strictly speaking, we are interested in estimating an Euclidean parameter, say ν, defined on the regular parametric model $\mathcal{P}$. We can identify ν with the parametric function $\psi : \Theta \to \mathbb{R}$, defined by

$$\psi(\theta) = \nu(P_\theta), \qquad \text{for } P_\theta \in \mathcal{P}.$$

Fix $P = P_\theta$ and suppose that ψ has differential $\dot{\psi}$ at θ. Define the information bound for ν as

$$I^{-1}(P|\nu, \mathcal{P}) = \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top, \tag{6}$$

and the efficient influence function for ν as

$$\tilde{\ell}(\cdot, P|\nu, \mathcal{P}) = \dot{\psi}_\theta I_\theta^{-1} \dot{\ell}_\theta(\cdot). \tag{7}$$

As defined above, the information bound and influence function appear to depend on the parametrization $\theta \mapsto P_\theta$ of $\mathcal{P}$. However, as our notation indicates, they actually depend only on ν and $\mathcal{P}$. This is proved in the following proposition.

Proposition 2.4 (Exercise)**.** The information bound $I^{-1}(P|\nu, \mathcal{P})$ and the efficient influence function $\tilde{\ell}(\cdot, P|\nu, \mathcal{P})$ are invariant under smooth changes of parametrization.

Remark 2.6 (Optimality of maximum likelihood estimation)**.** Let $\hat{\theta}_n$ be the maximum likelihood estimator (MLE) of θ in the experiment: $X_1, \ldots, X_n$ i.i.d. P_θ where the model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ satisfies the assumptions in Theorem 2.2 (in fact, a lot less is required; more on this later). Then the MLE of $\psi(\theta)$ is $\psi(\hat{\theta}_n)$. According to the Cramér-Rao lower bound, the variance of an unbiased estimator $\psi(\theta)$ is at least $n^{-1}\dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top$. Thus, we could infer that the MLE is asymptotically uniformly minimum-variance unbiased, and in this sense optimal. We write "could" because the preceding reasoning is informal and unsatisfying. The asymptotic normality does not warrant any conclusion about the convergence of the moments $\mathbb{E}_\theta[\psi(\hat{\theta}_n)]$ and $\mathrm{Var}_\theta[\psi(\hat{\theta}_n)]$; we have not introduced an asymptotic version of the Cramér-Rao theorem; and the Cramér-Rao bound does not make any assertion concerning asymptotic normality. Moreover, the unbiasedness required by the Cramer-Rao theorem is restrictive and can be relaxed considerably in the asymptotic situation. We present a more satisfying discussion later.

Remark 2.7. When $\psi : \Theta \to \mathbb{R}^q$, $q > 1$, similar results can be derived, where now, the information bound is a $m \times m$ covariance matrix, and the efficient influence function is a vector-valued function. Here $\dot{\psi}_\theta$ is the $q \times k$ Jacobian matrix whose (i, j)-th element is given by $\partial \psi_i(\theta)/\partial \theta_j$.

2.1 Back to the simplest 'semiparametric' model

Let us try to answer the questions raised in Example 1.4. We denote the elements of Σ^{-1} by $((\delta_{ij}))$, i.e., $(\Sigma^{-1})_{i,j} = \delta_{ij}$.

Case 1. In this case θ_2 is known so it can be considered fixed, say $\theta_2 = b$, and the statistical model consists of the distributions $\{N([\theta_1,\ b]^\top, \Sigma) : \theta_1 \in \mathbb{R}\}$. We have

$$\ell_{\theta_1}(X) = -\log(2\pi|\Sigma|^{1/2}) - \frac{1}{2}[X_1 - \theta_1,\ X_2 - b]\ \Sigma^{-1}\ [X_1 - \theta_1,\ X_2 - b]^\top.$$

From this expression we deduce that

$$\begin{aligned} \frac{\partial}{\partial\theta_1}\ell_{\theta_1}(X) &= [1,\ 0]\Sigma^{-1}[X_1 - \theta_1,\ X_2 - b]^\top, \\ I_{\theta_1} &= [1,\ 0]\Sigma^{-1}[1,\ 0]^\top. \end{aligned}$$

From the 1-dimensional version of the Cramer-Rao lemma, we deduce that any estimator $\hat{\mu}$ of $\mu = \theta_1 = \psi(\theta_1)$ in experiment 1 satisfies

$$\mathbb{E}_\theta[(\hat{\mu} - \theta_1)^2] \geq I_{\theta_1}^{-1} = \delta_{1,1}^{-1} = (\Sigma^{-1})_{1,1}^{-1}.$$

Case 2. In this case both θ_1 and θ_2 are unknown and the statistical model consists of the 2-dimensional distributions $\{N([\theta_1,\ \theta_1]^\top, \Sigma) : \theta = (\theta_1, \theta_2) \in \mathbb{R}^2\}$. We are interested in the functional $\psi(\theta_1, \theta_2) = \theta_1$ which has a derivative (gradient) equal to $[1,\ 0]$. We have

$$\ell_\theta(X) = -\log(2\pi|\Sigma|^{1/2}) - \frac{1}{2}(X - \theta)\ \Sigma^{-1}\ (X - \theta)^\top.$$

Thus,

$$\begin{aligned} \frac{\partial}{\partial\theta}\ell_\theta(X) &= \Sigma^{-1}(X - \theta), \\ I_\theta &= \mathbb{E}_\theta[\nabla\ell_\theta \nabla\ell_\theta^\top] = \mathbb{E}_\theta[\Sigma^{-1}(X - \theta)(X - \theta)^\top\Sigma^{-1}] = \Sigma^{-1}. \end{aligned}$$

From the 2-dimensional Cramer-Rao lemma, we deduce that any unbiased estimator $\hat{\mu}$ of $\mu = \theta_1 = \psi(\theta_1, \theta_2)$ in experiment 2 satisfies

$$\mathbb{E}_\theta[(\hat{\mu} - \theta_1)^2] \geq \dot{\psi}_\theta I_{\theta_1}^{-1}\dot{\psi}_\theta^\top = [1,\ 0]\ \Sigma\ [1,\ 0]^\top = \Sigma_{1,1}.$$

We can now answer the questions. It is a general fact that $(\Sigma^{-1})_{1,1}^{-1} \leq \Sigma_{1,1}$ (Prove this).

Thus we have established, for unbiased estimators, the intuitively clear fact ("information of a sub-model should be larger") that the best possible variance in the case where θ_2 is known is always smaller or equal to the best possible variance when θ_2 is unknown. Moreover, those bounds are achieved for the estimators

$$\hat{\mu}^{(1)} = X_1 - (\Sigma_{1,2}/\Sigma_{2,2})(X_2 - \theta_2), \qquad \text{and} \qquad \hat{\mu}^{(2)} = X_1.$$

Notice that $\hat{\mu}^{(1)}$ is not an estimator in the second experiment since it depends on the unknown quantity θ_2. Finally, we note that the information bounds are the same in both experiments if and only if the two coordinates of the observed Gaussian vector are independent (i.e., if Σ is diagonal).

2.2 Hodges' estimator

Characterizing asymptotic optimality (efficiency) in general is not as straightforward as it might seem. In fact, it is not enough to rank all consistent, asymptotically normal estimators by asymptotic variance. The **Hodges' super-efficient estimator** (given below) shows that there exists estimators with an asymptotic variance less than that of the MLE for some true parameter values:

Suppose that $X_1, \ldots, X_n$ are i.i.d. $N(\mu, 1)$, where $\mu \in \mathbb{R}$ is unknown. Of course, the sample mean $\bar{X}_n$ is the MLE here. Consider the following estimator:

$$\hat{\mu}_n = \begin{cases} \bar{X}_n, & \text{if } |\bar{X}_n| > n^{-1/4}, \\ 0 & \text{if } |\bar{X}_n| \leq n^{-1/4}. \end{cases}$$

Then $\hat{\mu}_n$ is equal to $\bar{X}_n$ with probability approaching 1 if $\mu \neq 0$ and is equal to zero with probability approaching 1 if $\mu = 0$ (Exercise: Show this). Thus, the asymptotic distribution of $\hat{\mu}_n$ is the same as the sample mean if $\mu \neq 0$ but has asymptotic variance zero at $\mu = 0$. At first sight, $\hat{\mu}_n$ is an improvement of $\bar{X}_n$. For every $\mu \neq 0$, the estimators behave the same, while for $\mu = 0$, the sequence $\hat{\mu}_n$ has an "arbitrary fast" rate of convergence. However, this reasoning is a bad use of asymptotics.

Figure 1 shows why $\hat{\mu}_n$ is no improvement. It shows the graph of the risk function $\mu \mapsto \mathbb{E}_\mu[(\hat{\mu}_n - \mu)^2]$ for three different sample sizes (n). These functions are close to 1 on most of the domain but possess peaks close to zero. As $n \to \infty$, the locations and widths of the peaks converge to zero but their heights go to infinity. The conclusion is that $\hat{\mu}_n$ "buys" its better asymptotic behavior at $\mu = 0$ at the expense of erratic behavior close to zero.

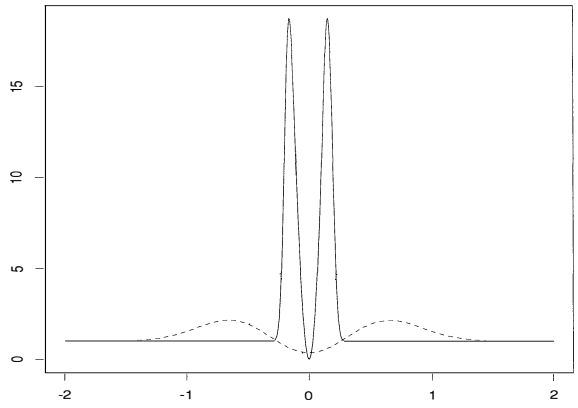

Figure 1: Quadratic risk function of the Hodges' estimator based on the means of samples of size 10 (dashed) and 1000 (solid) observations from the $N(\mu, 1)$ distribution.

Because the values of μ at which $\hat{\mu}_n$ is bad differ from n to n, the erratic behavior is not visible in the pointwise limit distributions under fixed μ.

2.3 Convolution theorems

Consider the estimation of $\psi(\theta)$, where $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ and the model satisfies assumptions (A1)-(A2). Here $\psi : \Theta \to \mathbb{R}^q$, $q \geq 1$, is a known function. In the following theorems we take an asymptotic approach and prove in a variety of ways that the best possible limit distribution for any estimator of $\psi(\theta)$ is the $N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top)$-distribution.

It is certainly impossible to give a nontrivial lower bound on the limit distribution of a standardized estimator $\sqrt{n}(T_n - \psi(\theta))$ for a single θ. Hodges' example shows that it is not even enough to consider the behavior under every θ, pointwise for all θ. Different values of the parameters must be taken into account simultaneously when taking the limit as $n \to \infty$. We shall do this by studying the performance of estimators under parameters in a "shrinking" neighborhood of a fixed θ (see Definition 2.7).

To carry out this exercise, we need (i) some "smoothness" conditions on the family of distributions $\mathcal{P}$ (see the notion 2.5 described below); (ii) some regularity on the estimators considered for $\psi(\theta)$ as described in 2.7 (which rules out examples like the Hodges' estimator); (iii) differentiability of the functional $\psi(\theta)$ (cf. the regularity conditions needed for the Cramér-Rao inequality to hold).

We start with the concept of **differentiability in quadratic mean** which leads to a fruitful analysis of the parametric model under minimal assumptions.

Definition 2.5 (Differentiable in quadratic mean)**.** The (parametric) statistical model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ is called *differentiable in quadratic mean* (DQM) at θ if there exists a vector of measurable functions $\dot{\ell}_\theta : \mathfrak{X} \to \mathbb{R}^k$ such that

$$\int \left[\sqrt{p_{\theta+h}(x)} - \sqrt{p_\theta(x)} - \frac{1}{2} h^\top \dot{\ell}_\theta(x) \sqrt{p_\theta(x)} \right]^2 d\mu(x) = o(\|h\|^2), \qquad h \to 0. \tag{8}$$

Remark 2.8. Usually $\frac{1}{2}\dot{\ell}_\theta(x)\sqrt{p_\theta(x)}$ is the derivative of the map $h \mapsto \sqrt{p_{\theta+h}(x)}$ at $h = 0$ for almost every x. In this case,

$$\dot{\ell}_\theta(x) = 2 \frac{1}{\sqrt{p_\theta(x)}} \left[\frac{\partial}{\partial \theta} \sqrt{p_\theta(x)} \right] = \frac{\partial}{\partial \theta} \log p_\theta(x).$$

Condition (39) does not require differentiability of the map $\theta \mapsto p_\theta(x)$ for any single x, but rather differentiability in (quadratic) mean.

Definition 2.6 (Local data generating process (LDGP))**.** We consider a triangular array of random variables $\{X_{in} : i = 1, \ldots, n\}$ which are i.i.d. P_{θ_n}, where $\sqrt{n}(\theta_n - \theta) \to h \in \mathbb{R}^k$ as $n \to \infty$ (i.e., θ_n is close to some fixed parameter θ). This data generating process is usually referred to as a LDGP.

Definition 2.7 (Regular estimator)**.** An estimator T_n (more specifically $T_n(X_{1n}, \ldots, X_{nn})$) is called **regular** at θ for estimating $\psi(\theta)$ if, for every $h \in \mathbb{R}^k$,

$$\sqrt{n}(T_n - \psi(\theta + h/\sqrt{n})) \overset{\theta+h/\sqrt{n}}{\to_d} L_\theta, \tag{9}$$

where L_θ is an arbitrary probability measure that does not depend on h. Informally, T_n is regular if its limiting distribution (after appropriate normalization) does not change with small perturbation of the true parameter θ.

A regular estimator sequence attains its limit distribution in a "locally uniform" manner. This type of regularity is common and is often considered desirable: A disappearing small change should not change the (limit) distribution at all. However, some estimator sequences of interest, such as shrinkage estimators, are not regular. The following convolution theorem designates a best estimator sequence among the regular estimator sequences.

Theorem 2.8 (Convolution theorem)**.** Assume that the experiment $\{P_\theta : \theta \in \Theta\}$ is DQM at the point θ with nonsingular Fisher information matrix I_θ. Let $\psi(\theta)$ be differentiable at θ with derivative $\dot{\psi}_\theta$. Let T_n be a regular estimator sequence at θ in the experiments $\{P_\theta^n : \theta \in \Theta\}$ with limit distribution L_θ. Then there exists a probability measure M_θ such

that

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top) \star M_\theta.$$

In particular, if L_θ has covariance matrix Σ_θ, then the matrix $\Sigma_\theta - \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top$ is nonnegative definite.

Proof. To be given later. □

The above result imposes an a priori restriction on the set of permitted estimator sequences. The following almost-everywhere convolution theorem imposes no (serious) restriction but yields no information about some parameters, albeit a null set of parameters.

Theorem 2.9 (Almost-everywhere convolution theorem)**.** Assume that the experiment $\{P_\theta : \theta \in \Theta\}$ is DQM at every θ with nonsingular Fisher information matrix I_θ. Let $\psi(\theta)$ be differentiable at every θ. Let T_n be an estimator sequence in the experiments $\{P_\theta^n : \theta \in \Theta\}$ such that $\sqrt{n}(T_n - \psi(\theta))$ converges to a limit distribution L_θ under every θ. Then there exist probability distributions M_θ such that for Lebesgue almost every θ,

$$L_\theta = N(0, \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top) \star M_\theta.$$

In particular, if L_θ has covariance matrix Σ_θ, then the matrix $\Sigma_\theta - \dot{\psi}_\theta I_\theta^{-1} \dot{\psi}_\theta^\top$ is nonnegative definite for Lebesgue almost every θ.

2.4 Contiguity

The proof of Theorem 2.9 relies on the notion of **contiguity**, which we define below.

Definition 2.10 (Contiguity)**.** Let $(\Omega_n, \mathcal{A}_n)$ be measurable spaces, each equipped with a pair of probability measures P_n and Q_n; $n \geq 1$.

The sequence $\{Q_n\}$ is **contiguous** with respect to the sequence $\{P_n\}$ if

$$P_n(A_n) \to 0 \qquad \text{implies} \qquad Q_n(A_n) \to 0 \tag{10}$$

for every sequence of measurable sets A_n. This is denoted as $Q_n \triangleleft P_n$.

"Contiguity"[7] can be thought of as "asymptotic absolute continuity". Contiguity arguments are a technique to obtain the limit distribution of a sequence of statistics under laws Q_n from a limiting distribution under laws P_n. The following result illustrates this point.

[7] The concept and theory of contiguity was developed by Le Cam in Le Cam (1960).

Lemma 2.11 (Le Cam's first lemma)**.** Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$. Further assume that P_n and Q_n have densities p_n and q_n with respect to a measure μ. Then the following statements are equivalent:

(i) $Q_n \triangleleft P_n$.

(ii) If $p_n/q_n \overset{Q_n}{\to}_d U$ along a subsequence, then $\mathbb{P}(U > 0) = 1$.

(iii) If $q_n/p_n \overset{P_n}{\to}_d V$ along a subsequence, then $\mathbb{E}[V] = 1$.

(iv) For statistics $T_n : \Omega_n \to \mathbb{R}^s$ $(s \geq 1)$: If $T_n \overset{P_n}{\to} 0$, then $T_n \overset{Q_n}{\to} 0$.

Proof. See van der Vaart (1998, Lemma 6.4). □

Note that the equivalence of (i) and (iv) follows directly from the definition of contiguity: Given statistics T_n, consider the sets $A_n = \{\|T_n\| > \epsilon\}$; given sets A_n, consider the statistics $T_n = \mathbf{1}_{A_n}$.

The above lemma gives us equivalent conditions under which a random variable which is $o_p(1)$ under P_n, is also $o_p(1)$ under Q_n. The following result allows one to find the exact weak limit of a random variable under Q_n, if we know its limit under P_n.

Lemma 2.12 (Le Cam's third lemma)**.** Let P_n and Q_n be sequences of probability measures on measurable spaces $(\Omega_n, \mathcal{A}_n)$. Further assume that P_n and Q_n have densities p_n and q_n with respect to a measure μ. Let $T_n : \Omega_n \to \mathbb{R}^s$ be a sequence of random variables (vectors). Suppose that $Q_n \triangleleft P_n$ and

$$\left(T_n, \frac{q_n}{p_n}\right) \overset{P_n}{\to}_d (T, V).$$

Then $L(B) := \mathbb{E}[\mathbf{1}_B(T)V]$ defines a probability measure, and $T_n \overset{Q_n}{\to}_d L$.

An useful consequence of Le Cam's third lemma is the following example.

Exercise 2.13 (Show this)**.** If

$$\left(T_n, \log \frac{q_n}{p_n}\right) \overset{P_n}{\to}_d N_{s+1}\left(\begin{bmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{bmatrix}, \begin{bmatrix} \Sigma & \tau \\ \tau^\top & \sigma^2 \end{bmatrix}\right).$$

Then $T_n \overset{Q_n}{\to}_d N_s(\mu + \tau, \Sigma)$.

2.5 Local Asymptotic Normality

Recall the setting of Section 2.

Definition 2.14 (Local Asymptotic Normality)**.** We say that the model $\{P_\theta^n : \theta \in \Theta\}$[8] is *locally asymptotic normal* (LAN) at the point θ if the following expansion holds for any $h \in \mathbb{R}^k$ (as $n \to \infty$):

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = \frac{1}{\sqrt{n}} h^\top \sum_{i=1}^n \dot{\ell}_\theta(X_i) - \frac{1}{2} h^\top I_\theta h + o_{P_\theta}(1). \tag{11}$$

Local asymptotic normality[9] is a property of a sequence of statistical models, which allows this sequence to be asymptotically approximated by a *normal location model*, after a rescaling of the parameter, i.e., for large n, the experiments

$$\left\{P^n_{\theta+h/\sqrt{n}} : h \in \mathbb{R}^k\right\} \quad \text{and} \quad \left\{N(h, I_\theta^{-1}) : h \in \mathbb{R}^k\right\}$$

are similar[10] in statistical properties, whenever the original experiments $\theta \mapsto P_\theta$ are "smooth" in the parameter. The second experiment consists of observing a single observation from a normal distribution with mean h and known covariance matrix (equal to the inverse of the Fisher information matrix). This is a simple experiment, which is easy to analyze, whence the approximation yields much information about the asymptotic properties of the original experiments.

As a consequence of (11), for every $h \in \mathbb{R}^k$,

$$\log \prod_{i=1}^n \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) \to_d N\left(-\frac{1}{2} h^\top I_\theta h, h^\top I_\theta h\right).$$

An important example when the local asymptotic normality holds is in the case of i.i.d. sampling from a regular parametric model, as shown below.

Theorem 2.15 (DQM implies LAN)**.** Suppose that Θ is an open subset of $\mathbb{R}^k$ and that the model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ. Then $P_\theta \dot{\ell}_\theta = 0$ and the Fisher information matrix $I_\theta = P_\theta \dot{\ell}_\theta \dot{\ell}_\theta^\top$ exists. Furthermore, for every $h \in \mathbb{R}^k$, as $n \to \infty$, (11) holds.

[8] Here P_θ^n denotes the joint distribution of $(X_1, \ldots, X_n)$ where X_i's are i.i.d. P_θ.
[9] The notion of local asymptotic normality was introduced by Le Cam (1960).
[10] Exercise: Find the likelihood ratio of the normal location model and compare with (11).

Proof. See van der Vaart (1998, Theorem 7.2). □

Exercise: Show that the sequences of distributions $P^n_{\theta+h/\sqrt{n}}$ and P^n_θ are contiguous.

3 Influence Functions for Parametric Models

As before, we borrow the notation and setup of Section 2. To keep the presentation simple, we only consider *regular* (see Definition (2.7)) and *asymptotically linear* (see (4)) estimators of $\psi(\theta)$. Note that most reasonable estimators are indeed regular and asymptotically linear (RAL). In this section we study the geometry of influence functions. To do this, we need some background on Hilbert spaces, introduced below.

3.1 Preliminaries: Hilbert spaces

Recall that a Hilbert space, denoted by $\mathcal{H}$, is a complete normed linear vector space equipped with an inner product (say $\langle \cdot, \cdot \rangle$). The following is an important result.

Theorem 3.1 (Projection theorem)**.** Let $(\mathcal{H}, \langle \cdot, \cdot \rangle)$ be a Hilbert space and let $\mathcal{U} \subset \mathcal{H}$ be a closed linear subspace. For any $h \in \mathcal{H}$, there exists a unique $u_0 \in \mathcal{U}$ that is closest to h, i.e.,

$$u_0 = \arg\min_{u \in \mathcal{U}} \|h - u\|.$$

Furthermore, u_0 is characterized by the fact that $h - u_0$ is orthogonal to $\mathcal{U}$, i.e.,

$$\langle h - u_0, u \rangle = 0, \qquad \text{for all } u \in \mathcal{U}.$$

Example 3.2 (q-dimensional random functions)**.** Let $X \sim P$ be a random variable taking values in $(\mathfrak{X}, \mathcal{A})$. Let $\mathcal{H}$ be the Hilbert space of *mean-zero* q-dimensional measurable functions of X (i.e., $h(X)$), with finite second moments equipped with the inner product

$$\langle h_1, h_2 \rangle := \mathbb{E}[h_1(X)^\top h_2(X)].$$

Let $v(X) \equiv (v_1(X), \ldots, v_r(X))$ be an r-dimensional random function with mean zero and $\mathbb{E}[v(X)^\top v(X)] < \infty$. Consider the linear subspace $\mathcal{U}$ spanned by $v(X)$, i.e.,

$$\mathcal{U} = \{B_{q \times r} v(X) : \text{where } B \in \mathbb{R}^{q \times r} \text{ is any arbitrary matrix}\}. \tag{12}$$

The linear subspace $\mathcal{U}$ defined above is a finite-dimensional linear subspace contained in the infinite-dimensional Hilbert space $\mathcal{H}$. If the elements $v_1(X), \ldots, v_r(X)$ are linearly independent, then the dimension of $\mathcal{U}$ is $q \times r$ (Exercise: Show this).

We now consider the problem of finding the projection of an arbitrary element $h(X) \in \mathcal{H}$

onto $\mathcal{U}$. By Theorem 3.1, such a projection $B_0 v(X)$ ($B_0 \in \mathbb{R}^{q\times r}$) is unique and must satisfy

$$\langle h - B_0 v, Bv\rangle = \mathbb{E}[\{h(X) - B_0 v(X)\}^\top Bv(X)] = 0 \qquad \text{for all } B \in \mathbb{R}^{q\times r}. \tag{13}$$

The above statement being true for all $B \in \mathbb{R}^{q\times r}$ is equivalent to (Exercise: Show this):

$$\mathbb{E}[\{h(X) - B_0 v(X)\}v(X)^\top] = 0 \qquad \Leftrightarrow \qquad B_0 \mathbb{E}[v(X)v(X)^\top] = \mathbb{E}[h(X)v(X)^\top].$$

Therefore, assuming that $\mathbb{E}[v(X)v(X)^\top]$ is nonsingular (i.e., positive definite),

$$B_0 = \mathbb{E}[h(X)v(X)^\top]\ \{\mathbb{E}[v(X)v(X)^\top]\}^{-1}.$$

Hence, the unique projection of $h(X) \in \mathcal{H}$ onto $\mathcal{U}$ is

$$\Pi(h|\mathcal{U}) = \mathbb{E}[h(X)v(X)^\top]\ \{\mathbb{E}[v(X)v(X)^\top]\}^{-1} v. \tag{14}$$

Remark 3.1. Let $\mathcal{H}^{(1)}$ be the Hilbert space of one-dimensional mean-zero random functions of X (with finite variance), where we use the superscript (1) to emphasize one-dimensional random functions. If h_1 and h_2 are elements of $\mathcal{H}^{(1)}$ that are orthogonal to each other, then, by the Pythagorean theorem, we know that

$$\mathrm{Var}(h_1 + h_2) = \mathrm{Var}(h_1) + \mathrm{Var}(h_2),$$

making it clear that $\mathrm{Var}(h_1 + h_2)$ is greater than or equal to $\mathrm{Var}(h_1)$ or $\mathrm{Var}(h_2)$.

Unfortunately, when $\mathcal{H}$ consists of q-dimensional mean-zero random functions, there is no such general relationship with regard to the variance matrices. However, there is an important special case when this does occur, which we now discuss.

Definition 3.3 (q-replicating linear space)**.** A linear subspace $\mathcal{U} \subset \mathcal{H}$ is a **q-replicating linear space** if $\mathcal{U}$ is of the form $\mathcal{U}^{(1)} \times \cdots \times \mathcal{U}^{(1)} \equiv [\mathcal{U}^{(1)}]^q$, where $\mathcal{U}^{(1)}$ denotes a linear subspace in $\mathcal{H}^{(1)}$. Note that $[\mathcal{U}^{(1)}]^q \subset \mathcal{H}$ represents the linear subspace in $\mathcal{H}$ that consists of elements $h = (h^{(1)}, \cdots, h^{(q)})^\top$ such that $h^{(j)} \in \mathcal{U}^{(1)}$ for all $j = 1, \ldots, q$; i.e., $[\mathcal{U}^{(1)}]^q$ consists of q-dimensional random functions, where each element in the vector is an element of $\mathcal{U}^{(1)}$, or the space $\mathcal{U}^{(1)}$ stacked up on itself q times.

Remark 3.2. The linear subspace spanned by an r-dimensional vector of mean zero finite variance random functions $v_{r\times 1}(X)$, discussed in Example 3.2 is such a subspace. This is easily seen by defining $\mathcal{U}^{(1)}$ to be the space $\{b^\top v(X) : b \in \mathbb{R}^{r\times 1}\}$.

Theorem 3.4 (Multivariate Pythagorean theorem)**.** If $h \in \mathcal{H}$ and is an element of a q-replicating linear space $\mathcal{U}$, and $t \in \mathcal{H}$ is orthogonal to $\mathcal{U}$, then

$$\mathrm{Var}(t+h) = \mathrm{Var}(t) + \mathrm{Var}(h),$$

where $\mathrm{Var}(h) := \mathbb{E}(hh^\top)$. As a consequence, we obtain a multivariate version of the Pythagorean theorem; namely, for any $h^* \in \mathcal{H}$,

$$\mathrm{Var}(h^*) = \mathrm{Var}(\Pi[h^*|\mathcal{U}]) + \mathrm{Var}(h^* - \Pi[h^*|\mathcal{U}]).$$

Proof. Exercise. □

3.2 Geometry of Influence Functions

Recall the notation in the beginning of Section 2. Let $\theta_0 \in \Theta$ be the true value of the parameter. We have the following important result.

Theorem 3.5. Assume that the experiment $\{P_\theta : \theta \in \Theta\}$ is DQM at the point θ_0 with nonsingular Fisher information matrix I_{θ_0}. Let the parameter of interest be $\psi(\theta)$, a q-dimensional function of the k-dimensional parameter θ ($q < k$) such that

$$\frac{\partial}{\partial\theta}\psi(\theta) \equiv \dot{\psi}_\theta,$$

the $q \times k$-dimensional matrix of partial derivatives (i.e., $\dot{\psi}_\theta = (\partial\psi_i(\theta)/\partial\theta_j)_{1\le i\le q}^{1\le j\le k}$) exists at θ_0. Also let T_n be an asymptotically linear estimator with influence function $\varphi(X)$ such that $\mathbb{E}_\theta[\varphi(X)^\top\varphi(X)]$ exists at θ_0. Then, if T_n is regular, this will imply that

$$\mathbb{E}[\varphi(X)\dot{\ell}_{\theta_0}(X)^\top] = \dot{\psi}_{\theta_0}. \tag{15}$$

Proof. Given in class. This follows from using the results on contiguity, the regularity and asymptotically linearity of T_n. □

Remark 3.3. Although influence functions of RAL estimators for $\psi(\theta)$ must satisfy (15) of Theorem 3.5, a natural question is whether the converse is true, i.e., for any element of the Hilbert space satisfying (15), does there exist an RAL estimator for $\psi(\theta)$ with that influence function? Indeed this is true.

To prove this in full generality, especially later when we consider infinite-dimensional nuisance parameters, is difficult and requires that some careful technical regularity conditions

hold. We will come back to this later on.

Remark 3.4. Note that RAL estimators are asymptotically normally distributed:

$$\sqrt{n}(T_n - \psi(\theta)) \xrightarrow{d} N(0, \mathbb{E}[\varphi(X)\varphi(X)^\top]).$$

Because of this, we can compare competing RAL estimators for $\psi(\theta)$ by looking at the asymptotic variance, where clearly the better estimator is the one with smaller asymptotic variance. We argued earlier, however, that the asymptotic variance of an RAL estimator is the variance of its influence function. Therefore, it suffices to consider the variance of influence functions. We already illustrated that influence functions can be viewed as elements in a subspace of a Hilbert space. Moreover, in this Hilbert space the distance to the origin (squared) of any element (random function) is the variance of the element. Consequently, the search for the best estimator (i.e., the one with the smallest asymptotic variance) is equivalent to the search for the element in the subspace of influence functions that has the shortest distance to the origin.

Now we specialize slightly: suppose that $\theta \equiv (\psi, \eta)$ where $\psi \in \mathcal{S} \subset \mathbb{R}^q$, $\eta \in \mathcal{N} \subset \mathbb{R}^{k-q}$; here ψ is the parameter of interest and η is the nuisance parameter. We can think of this as $\psi(\theta) \equiv \psi$ so that $\Gamma(\theta) \equiv \dot{\psi}_\theta = (I_q, 0_{q\times(k-q)})$ is a $q \times k$ matrix; here I_q is the $q \times q$ identity matrix and $0_{q\times(k-q)}$ denotes the $q \times (k-q)$ matrix of all zeros. We decompose $\dot{\ell}_\theta = (\dot{\ell}_\theta^{(1)}, \dot{\ell}_\theta^{(2)})$ where $\dot{\ell}_\theta^{(1)} \equiv \partial\ell_\theta/\partial\psi \in \mathbb{R}^q$ and $\dot{\ell}_\theta^{(2)} \equiv \partial\ell_\theta/\partial\eta \in \mathbb{R}^{k-q}$. We immediately have the following corollary.

Corollary 3.6. Under the assumptions of Theorem 3.5,

$$\mathbb{E}[\varphi(X)\dot{\ell}_{\theta_0}^{(1)}(X)^\top] = I_q \tag{16}$$

and

$$\mathbb{E}[\varphi(X)\dot{\ell}_{\theta_0}^{(2)}(X)^\top] = 0_{q\times(k-q)}. \tag{17}$$

Although influence functions of RAL estimators for ψ must satisfy conditions (16) and (17) of Corollary 3.6, a natural question is whether the converse is true; that is, for any element of the Hilbert space satisfying conditions (16) and (16), does there exist an RAL estimator for ψ with that influence function? We address this below. But before we do this let us digress and introduce empirical process theory which will give us many tools to construct and study "complicated" estimators.

3.3 Digression: Empirical process theory

Suppose now that $X_1, \ldots, X_n$ are i.i.d. P on $(\mathfrak{X}, \mathcal{A})$. Then the *empirical measure* $\mathbb{P}_n$ is defined by

$$\mathbb{P}_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i},$$

where δ_x denotes the Dirac measure at x. For each $n \geq 1$, $\mathbb{P}_n$ denotes the random discrete probability measure[11] which puts mass $1/n$ at each of the n points $X_1, \ldots, X_n$. For a real-valued function f on $\mathfrak{X}$, we write

$$\mathbb{P}_n[f] := \int f d\mathbb{P}_n = \frac{1}{n} \sum_{i=1}^{n} f(X_i).$$

If $\mathcal{F}$ is a collection of real-valued functions defined on $\mathfrak{X}$, then $\{\mathbb{P}_n(f) : f \in \mathcal{F}\}$ is the *empirical measure* indexed by $\mathcal{F}$. Let us assume that[12]

$$P[f] := \int f dP$$

exists for each $f \in \mathcal{F}$. The *empirical process* $\mathbb{G}_n$ is defined by

$$\mathbb{G}_n := \sqrt{n}(\mathbb{P}_n - P),$$

and the collection of random variables $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ as f varies over $\mathcal{F}$ is called the *empirical process*[13] indexed by $\mathcal{F}$. The goal of empirical process theory is to study the properties of the approximation of Pf by $\mathbb{P}_n f$, *uniformly in* $\mathcal{F}$. Mainly, we would be concerned with probability estimates of the random quantity

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - Pf| \tag{18}$$

In particular, we will find appropriate conditions to answer the following two questions:

1. **Glivenko-Cantelli**: Under what conditions on $\mathcal{F}$ does $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ converge to zero almost surely (or in probability)? If this convergence holds, then we say that $\mathcal{F}$ is a

[11] Thus, for any Borel set $A \subset \mathfrak{X}$, $\mathbb{P}_n(A) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_A(X_i) = \frac{\#\{i \leq n : X_i \in A\}}{n}$.

[12] We will use the this *operator* notation for the integral of any function f with respect to P. Note that such a notation is helpful (and preferable over the expectation notation) as then we can even treat *random* (data dependent) functions.

[13] Note that the classical empirical process for real-valued random variables can be viewed as the special case of the general theory for which $\mathfrak{X} = \mathbb{R}$, $\mathcal{F} = \{\mathbf{1}_{(-\infty, x]}(\cdot) : x \in \mathbb{R}\}$.

P-Glivenko-Cantelli class of functions. More generally, given a function class $\mathcal{F}$, we are interested in tight bounds on the tail probability $\mathbb{P}(\|\mathbb{P}_n - P\|_{\mathcal{F}} > \epsilon)$, for $\epsilon > 0$.

2. **Donsker**: Under what conditions on $\mathcal{F}$ does $\{\mathbb{G}_n(f) : f \in \mathcal{F}\}$ converges as a process to some limiting object as $n \to \infty$.

   If this convergence holds, then we say that $\mathcal{F}$ is a *P-Donsker* class of functions.

Our main findings reveal that the answers (to the two above questions and more) depend crucially on the *complexity*[14] or *size* of the underlying function class $\mathcal{F}$. However, the scope of empirical process theory is much beyond answering the above two questions[15]. The following section introduces the topic of M-estimation (also known as *empirical risk minimization*), a field that naturally relies on the study of empirical processes.

3.3.1 M-estimation (or empirical risk minimization)

Many problems in statistics and machine learning are concerned with estimators of the form

$$\hat{\theta}_n := \arg\max_{\theta \in \Theta} \mathbb{P}_n[m_\theta] = \arg\max_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i). \tag{19}$$

where $X, X_1, \ldots, X_n$ denote (i.i.d.) observations from P taking values in a space $\mathfrak{X}$. Here Θ denotes the parameter space and, for each $\theta \in \Theta$, m_θ denotes the a real-valued (loss-) function on $\mathfrak{X}$. Such a quantity $\hat{\theta}_n$ is called an *M-estimator* as it is obtained by maximizing (or minimizing) an objective function. The map

$$\theta \mapsto -\mathbb{P}_n m_\theta = -\frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i)$$

can be thought of as the "empirical risk" and $\hat{\theta}_n$ denotes the empirical risk minimizer over $\theta \in \Theta$. Here are some examples:

1. **Maximum likelihood estimators**: These correspond to $m_\theta(x) = \log p_\theta(x)$.

[14] We will consider different geometric (packing and covering numbers) and combinatorial (shattering and combinatorial dimension) notions of complexity.

[15] In the last 20 years there has been enormous interest in understanding the *concentration* properties of $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ about its mean. In particular, one may ask if we can obtain exponential inequalities for the difference $\|\mathbb{P}_n - P\|_{\mathcal{F}} - \mathbb{E}\|\mathbb{P}_n - P\|_{\mathcal{F}}$ (when $\mathcal{F}$ is uniformly bounded). Talagrand's inequality (Talagrand (1996)) gives an affirmative answer to this question; a result that is considered to be one of the most important and powerful results in the theory of empirical processes in the last 30 years. We will cover this topic towards the end of the course (if time permits).

2. **Location estimators**:

   (a) **Median**: corresponds to $m_\theta(x) = |x - \theta|$.

   (b) **Mode**: may correspond to $m_\theta(x) = \mathbf{1}\{|x - \theta| \leq 1\}$.

3. **Nonparametric maximum likelihood**: Suppose $X_1, \ldots, X_n$ are i.i.d. from a density θ on $[0, \infty)$ that is known to be non-increasing. Then take Θ to be the collection of all non-increasing densities on $[0, \infty)$ and $m_\theta(x) = \log \theta(x)$. The corresponding M-estimator is the MLE over all non-increasing densities. It can be shown that $\hat{\theta}_n$ exists and is unique; $\hat{\theta}_n$ is usually known as the Grenander estimator.

4. **Regression estimators**: Let $\{X_i = (Z_i, Y_i)\}_{i=1}^n$ denote i.i.d. from a regression model and let

$$m_\theta(x) = m_\theta(z, y) := -(y - \theta(z))^2,$$

   for a class $\theta \in \Theta$ of real-valued functions from the domain of Z[16]. This gives the usual least squares estimator over the class Θ. The choice $m_\theta(z, y) = -|y - \theta(z)|$ gives the least absolute deviation estimator over Θ.

In these problems, the parameter of interest is

$$\theta_0 := \arg\max_{\theta \in \Theta} P[m_\theta].$$

Perhaps the simplest general way to address this problem is to reason as follows. By the law of large numbers, we can approximate the 'risk' for a fixed parameter θ by the empirical risk which depends only on the data, i.e.,

$$P[m_\theta] \approx \mathbb{P}_n[m_\theta].$$

If $\mathbb{P}_n[m_\theta]$ and $P[m_\theta]$ are *uniformly* close, then maybe their argmax's $\hat{\theta}_n$ and θ_0 are close. The problem is now to quantify how close $\hat{\theta}_n$ is to θ_0 as a function of the number of samples n, the dimension of the parameter space Θ, the dimension of the space $\mathfrak{X}$, etc. The resolution of this question leads naturally to the investigation of quantities such as the uniform deviation

$$\sup_{\theta \in \Theta} |(\mathbb{P}_n - P)[m_\theta]|.$$

[16] In the simplest setting we could parametrize $\theta(\cdot)$ as $\theta_\beta(z) := \beta^\top z$, for $\beta \in \mathbb{R}^d$, in which case $\Theta = \{\theta_\beta(\cdot) : \beta \in \mathbb{R}^d\}$.

Closely related to M-estimators are Z-estimators, which are defined as solutions to a system of equations of the form $\sum_{i=1}^{n} m_\theta(X_i) = 0$ for $\theta \in \Theta$, an appropriate function class.

We will learn how to establish *consistency*, *rates of convergence* and the *limiting distribution* for M and Z-estimators; see van der Vaart and Wellner (1996, Chapters 3.1-3.4) for more details.

3.3.2 Asymptotic equicontinuity: a further motivation to study empirical processes

A commonly recurring theme in statistics is that we want to prove consistency or asymptotic normality of some statistic which is not a sum of independent random variables, but can be related to some natural sum of random functions indexed by a parameter in a suitable (metric) space. The following example illustrates the basic idea.

Example 3.7. Suppose that $X, X_1, \ldots, X_n, \ldots$ are i.i.d. P with c.d.f. G, having a Lebesgue density g, and $\mathbb{E}(X^2) < \infty$. Let $\mu = \mathbb{E}(X)$. Consider the absolute deviations about the sample mean,

$$M_n := \mathbb{P}_n|X - \bar{X}_n| = \frac{1}{n}\sum_{i=1}^{n} |X_i - \bar{X}_n|,$$

as an estimate of scale. This is an average of the dependent random variables $|X_i - \bar{X}_n|$. Suppose that we want to find the almost sure (a.s.) limit and the asymptotic distribution[17] of M_n (properly normalized).

There are several routes available for showing that $M_n \overset{a.s.}{\to} M := \mathbb{E}|X - \mu|$, but the methods we will develop in this section proceeds as follows. Since $\bar{X}_n \overset{a.s.}{\to} \mu$, we know that for any $\delta > 0$ we have $\bar{X}_n \in [\mu - \delta, \mu + \delta]$ for all sufficiently large n almost surely. Let us define, for $\delta > 0$, the random functions

$$\mathbb{M}_n(t) = \mathbb{P}_n|X - t|, \qquad \text{for } |t - \mu| \leq \delta.$$

This is just the empirical measure indexed by the collection of functions

$$\mathcal{F}_\delta := \{f_t : |t - \mu| \leq \delta\}, \qquad \text{where} \quad f_t(x) := |x - t|.$$

[17] This example was one of the illustrative examples considered by Pollard (1989).

Note that $M_n \equiv \mathbb{M}_n(\bar{X}_n)$. To show that $M_n \overset{a.s.}{\to} M := \mathbb{E}|X-\mu|$, we write

$$\begin{aligned} M_n - M &= \mathbb{P}_n(f_{\bar{X}_n}) - P(f_\mu) \\ &= (\mathbb{P}_n - P)(f_{\bar{X}_n}) + \left[P(f_{\bar{X}_n}) - P(f_\mu)\right] \\ &= I_n + II_n. \end{aligned}$$

Note that,

$$|I_n| \leq \sup_{f \in \mathcal{F}_\delta} |(\mathbb{P}_n - P)(f)| \overset{a.s.}{\to} 0, \tag{20}$$

if $\mathcal{F}_\delta$ is P-Glivenko-Cantelli. As we will see, this collection of functions $\mathcal{F}_\delta$ is *a VC subgraph class of functions*[18] with an integrable *envelope*[19] function, and hence empirical process theory can be used to establish the desired convergence.

The convergence of the second term in II_n is easy: by the triangle inequality

$$|II_n| = \left|P(f_{\bar{X}_n}) - P(f_\mu)\right| \leq P|\bar{X}_n - \mu| = |\bar{X}_n - \mu| \overset{a.s.}{\to} 0.$$

Exercise: Give an alternate direct (rigorous) proof of the above result (i.e., $M_n \overset{a.s.}{\to} M := \mathbb{E}|X-\mu|$).

The corresponding central limit theorem is trickier. Can we show that $\sqrt{n}(M_n - M)$ converges to a normal distribution? This may still not be unreasonable to expect. After all if $\bar{X}_n$ were replaced by μ in the definition of M_n this would be an outcome of the CLT (assuming a finite variance for the X_i's) and $\bar{X}_n$ is the natural estimate of μ. Note that

$$\begin{aligned} \sqrt{n}(M_n - M) &= \sqrt{n}(\mathbb{P}_n f_{\bar{X}_n} - P f_\mu) \\ &= \sqrt{n}(\mathbb{P}_n - P) f_\mu + \sqrt{n}(\mathbb{P}_n f_{\bar{X}_n} - \mathbb{P}_n f_\mu) \\ &= \mathbb{G}_n f_\mu + \mathbb{G}_n(f_{\bar{X}_n} - f_\mu) + \sqrt{n}(\psi(\bar{X}_n) - \psi(\mu)) \\ &= A_n + B_n + C_n \ \text{(say)}, \end{aligned}$$

where $\psi(t) := P(f_t) = \mathbb{E}|X-t|$. We will argue later that B_n is *asymptotically negligible*

[18] We will formally define VC classes of functions later. Intuitively, these classes of functions have simple combinatorial properties.

[19] An envelope function of a class $\mathcal{F}$ is any function $x \mapsto F(x)$ such that $|f(x)| \leq F(x)$, for every $x \in \mathfrak{X}$ and $f \in \mathcal{F}$.

using an *equicontinuity* argument. Let us consider $A_n + C_n$. It can be easily shown that

$$\psi(t) = \mu - 2\int_{-\infty}^{t} xg(x)dx - t + 2tG(t), \qquad \text{and} \qquad \psi'(t) = 2G(t) - 1.$$

The delta method now yields:

$$A_n + C_n = \mathbb{G}_n f_\mu + \sqrt{n}(\bar{X}_n - \mu)\psi'(\mu) + o_p(1) = \mathbb{G}_n[f_\mu + X\psi'(\mu)] + o_{\mathbb{P}}(1).$$

The usual CLT now gives the limit distribution of $A_n + C_n$.

Exercise: Complete the details and derive the exact form of the limiting distribution.

Definition 3.8. Let $\{Z_n(f) : f \in \mathcal{F}\}$ be a stochastic process indexed by a class $\mathcal{F}$ equipped with a semi-metric[20] $d(\cdot, \cdot)$. Call $\{Z_n\}_{n\geq 1}$ to be *asymptotically* (or stochastically) *equicontinuous* at f_0 if for each $\eta > 0$ and $\epsilon > 0$ there exists a neighborhood V of f_0 for which[21]

$$\limsup_{n\to\infty} \mathbb{P}\left(\sup_{f\in V} |Z_n(f) - Z_n(f_0)| > \eta\right) < \epsilon.$$

Exercise: Show that if $\{\hat{f}_n\}_{n\geq 1}$ is a sequence of (random) elements of $\mathcal{F}$ that converge in probability to f_0 (i.e.,$d(\hat{f}_n, f_0) \xrightarrow{P} 0$), and $\{Z_n(f) : f \in \mathcal{F}\}$ is asymptotically equicontinuous at f_0, then $Z_n(\hat{f}_n) - Z_n(f_0) = o_{\mathbb{P}}(1)$. [Hint: Note that with probability tending to 1, $\hat{f}_n$ will belong to each V.]

Empirical process theory offers very efficient methods for establishing the asymptotic equicontinuity of $\mathbb{G}_n$ over a class of functions $\mathcal{F}$. The fact that $\mathcal{F}$ is a *VC class* of functions with square-integrable *envelope* function will suffice to show the desired asymptotic equicontinuity.

[20] A semi-metric has all the properties of a metric except that $d(s,t) = 0$ need not imply that $s = t$.

[21] There might be measure theoretical difficulties related to taking a supremum over an uncountable set of f values, but we shall ignore these for the time being.

3.4 Constructing estimators

Let $\varphi(X)$ be a q-dimensional measurable function with zero mean and finite variance that satisfies conditions (16) and (17). Define

$$m_{\psi,\eta}(x) = \varphi(x) - \mathbb{E}_{\psi,\eta}[\varphi(X)]. \tag{21}$$

We assume that we can find a root-n consistent estimator $\hat{\eta}_n$ for the nuisance parameter η_0 (i.e., $\sqrt{n}(\hat{\eta}_n - \eta_0)$ is bounded in probability). In many cases the estimator $\hat{\eta}_n$ will be ψ-dependent (i.e., $\hat{\eta}_n(\psi)$). For example, we might use the MLE for $\hat{\eta}_n$, or the restricted MLE for η, fixing the value of ψ.

We will now argue that the solution to the equation

$$\frac{1}{n}\sum_{i=1}^{n} m_{\psi,\hat{\eta}_n(\psi)}(X_i) = 0 \tag{22}$$

which we denote by $\hat{\psi}_n$, will be an asymptotically linear estimator with influence function $\varphi(X)$. The above equation shows that $\hat{\psi}_n$ is a Z-estimator. Using empirical process notation, we have $\mathbb{P}_n[m_{\hat{\psi}_n,\hat{\eta}_n}] = 0$. In general, there are many results that give sufficient conditions under which a (finite-dimensional) Z-estimator will be $\sqrt{n}$-consistent and asymptotically normal. The following is one such result; see van der Vaart (1998, Theorem 5.21).

Theorem 3.9. Suppose that $X_1, \ldots, X_n$ are i.i.d. P on $(\mathfrak{X}, \mathcal{A})$. For each β in an open subset of Euclidean space, let $x \mapsto g_\beta(x)$ be a measurable vector-valued function such that $\mathbb{G}_n[g_\beta] \equiv \sqrt{n}(\mathbb{P}_n - P)[g_\beta]$ is asymptotically equicontinuous at $\beta = \beta_0$[22], i.e., $\mathbb{G}_n[g_{\tilde{\beta}_n} - g_{\beta_0}] = o_p(1)$ if $\tilde{\beta}_n \xrightarrow{P} \beta_0$. Assume that $P[g_{\beta_0}^\top g_{\beta_0}] < \infty$ and that the map $\beta \mapsto P[g_\beta]$ is differentiable at a zero β_0 (i.e., $P[g_{\beta_0}] = 0$) with nonsingular derivative matrix V_{β_0}. If $\mathbb{P}_n[g_{\hat{\beta}_n}] = o_p(n^{-1/2})$, and $\hat{\beta}_n \xrightarrow{P} \beta_0$, then

$$\sqrt{n}(\hat{\beta}_n - \beta_0) = -V_{\beta_0}^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n} g_{\beta_0}(X_i) + o_P(1).$$

In particular, the sequence $\sqrt{n}(\hat{\beta}_n - \beta_0)$ is asymptotically normal with mean zero and

[22]For example, such asymptotically equicontinuity holds if there exists a measurable function $L : \mathfrak{X} \to \mathbb{R}$ with $P[L^2] < \infty$ such that for every β_1 and β_2 in a neighborhood of β_0,

$$\|g_{\beta_1}(x) - g_{\beta_2}(x)\| \le L(x)\|\beta_1 - \beta_2\|.$$

covariance matrix $V_{\beta_0}^{-1} P[g_{\beta_0} g_{\beta_0}^\top](V_{\beta_0}^{-1})^\top$.

If $m_{\psi,\eta}$ is a "nice" class of functions (indexed by (ψ, η)), and $(\hat{\psi}_n, \hat{\eta}_n)$ is a consistent estimator of $\theta_0 \equiv (\psi_0, \eta_0)$ we can expect, by asymptotic equicontinuity,

$$\mathbb{G}_n[m_{\hat{\psi}_n, \hat{\eta}_n} - m_{\psi_0,\eta_0}] = o_P(1).$$

However,

$$\mathbb{G}_n[m_{\hat{\psi}_n, \hat{\eta}_n} - m_{\psi_0,\eta_0}] \quad = \quad \sqrt{n}\mathbb{P}_n[m_{\hat{\psi}_n, \hat{\eta}_n}] - \sqrt{n}\mathbb{P}_n[m_{\psi_0,\eta_0}] - \sqrt{n}P[m_{\hat{\psi}_n, \hat{\eta}_n} - m_{\psi_0,\eta_0}].$$

Observe that the first term on the right side is 0 (by definition); the second term is asymptotically normal by the CLT; the third term can be handled by using DQM of the parametric model at (ψ_0, η_0) (Exercise: Show this.).

3.5 Tangent spaces

We first note that the score vector $\dot{\ell}_{\theta_0}(X)$, under suitable regularity conditions (e.g., DQM of the parametric model at θ_0), has mean zero (i.e., $\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}(X)] = 0_{k\times 1}$).

Definition 3.10 (Tangent space)**.** We can define the finite-dimensional linear subspace $\mathcal{T} \subset \mathcal{H}$ spanned by the k-dimensional score vector $\dot{\ell}_{\theta_0}(X)$ (similar to Example 3.2) as the set of all q-dimensional mean-zero random vectors consisting of $B_{q\times k}\dot{\ell}_{\theta_0}(X)$, i.e.,

$$\mathcal{T} := \{B\dot{\ell}_{\theta_0}(X) : \text{where } B \in \mathbb{R}^{q\times k} \text{ is any arbitrary matrix}\}. \tag{23}$$

The linear subspace $\mathcal{T}$ is referred to as the **tangent space**.

Definition 3.11 (Nuisance tangent space)**.** In the case where θ can be partitioned as (ψ, η), consider the linear subspace spanned by the **nuisance score vector** $\dot{\ell}^{(2)}_{\theta_0}(X)$, i.e.,

$$\Lambda := \{B\dot{\ell}^{(2)}_{\theta_0}(X) : \text{where } B \in \mathbb{R}^{q\times(k-q)} \text{ is any arbitrary matrix}\}. \tag{24}$$

This space is referred to as the **nuisance tangent space**[23] and will be denoted by Λ.

[23] Since tangent spaces and nuisance tangent spaces are linear subspaces spanned by score vectors, these are examples of q-replicating linear spaces.

We note that by (17) of Corollary 3.6 this is equivalent to saying that the q-dimensional influence function $\varphi(X)$ of T_n is *orthogonal* to the nuisance tangent space Λ.

3.6 Efficient Influence Function

We will show how the geometry of Hilbert spaces will allow us to identify the **efficient influence function** (i.e., the influence function with the smallest variance). First, however, we give some additional notation and definitions regarding operations on linear subspaces that will be needed shortly.

Definition 3.12 (Direct sum)**.** We say that $M \oplus N$ is a **direct sum** of two linear subspaces $M \subset \mathcal{H}$ and $N \subset \mathcal{H}$ if $M \oplus N$ is a linear subspace in $\mathcal{H}$ and if every element $x \in M \oplus N$ has a unique representation of the form $x = m + n$, where $m \in M$ and $n \in N$.

Definition 3.13 (Orthogonal complement)**.** The set of elements of a Hilbert space that are **orthogonal** to a linear subspace M is denoted by $M^{\perp}$. The space $M^{\perp}$ is also a linear subspace, referred to as the **orthogonal complement** of M. Moreover, if M is closed (note that all finite-dimensional linear spaces are closed), the entire Hilbert space can be written as

$$\mathcal{H} = M \oplus M^{\perp}.$$

Condition (17) of Corollary 3.6 can now be stated as follows: If $\varphi(X)$ is an influence function of an RAL estimator, then $\varphi(X) \in \Lambda^{\perp}$, where Λ denotes the nuisance tangent space defined by (24).

For any arbitrary element $h(X) \in \mathcal{H}$, by the projection theorem $\Pi(h|\Lambda)$, referred to as the projection of $h(X)$ onto the space Λ, is the unique element (in Λ) such that

$$\langle h - \Pi(h|\Lambda), a \rangle = 0 \qquad \text{for all } a \in \Lambda.$$

The element with the minimum norm, $h - \Pi(h|\Lambda)$, is sometimes referred to as the residual of h after projecting onto Λ, and it is easy to show that

$$h - \Pi(h|\Lambda) = \Pi(h|\Lambda^{\perp}).$$

Also, observe that $\Pi(h|\Lambda)$ has an exact expression as given in (14).

We need the following definition of a **linear variety** (sometimes also called an affine space).

Definition 3.14 (Linear variety)**.** A *linear variety* is the translation of a linear subspace away from the origin; i.e., a linear variety V can be written as $V = x_0 + M$, where $x_0 \in \mathcal{H}$ and $x_0 (\neq 0) \notin M$, and M is a linear subspace.

Theorem 3.15. The set of all influence functions, namely the elements of $\mathcal{H}$ that satisfy condition (15) of Theorem 3.5, is the linear variety $\varphi^*(X) + \mathcal{T}^\perp$, where $\varphi^*(X)$ is any influence function and $\mathcal{T}^\perp$ is the space perpendicular to the tangent space.

Proof. Let $\varphi^*(X)$ be any influence function. Note that any element $t(X) \in \mathcal{T}^\perp$ must satisfy

$$\mathbb{E}[t(X)\dot{\ell}_{\theta_0}(X)^\top] = 0_{q\times k}.$$

Let $\varphi : \mathfrak{X} \to \mathbb{R}^q$ be defined as $\varphi = \varphi^* + t$, where $\varphi^*(\cdot)$ is any influence function. Then

$$\begin{aligned}
\mathbb{E}[\varphi(X)\dot{\ell}_{\theta_0}(X)^\top] &= \mathbb{E}[\{\varphi^*(X) + t(X)\}\dot{\ell}_{\theta_0}(X)^\top] \\
&= \mathbb{E}[\varphi^*(X)\dot{\ell}_{\theta_0}(X)^\top] + \mathbb{E}[t(X)\dot{\ell}_{\theta_0}(X)^\top] \\
&= \dot{\psi}_{\theta_0} + 0_{q\times k}.
\end{aligned}$$

Hence, $\varphi(X)$ is an influence function satisfying condition (15) of Theorem 3.5.

Conversely, if $\varphi(X)$ is an influence function satisfying (15) of Theorem 3.5, then

$$\varphi(X) = \varphi^*(X) + \{\varphi(X) - \varphi^*(X)\}.$$

It is a simple exercise to verify that $\{\varphi(X) - \varphi^*(X)\} \in \mathcal{T}^\perp$. □

Definition 3.16 (Efficient influence function)**.** The **efficient influence function** $\varphi^{\text{eff}}(X)$, if it exists, is the influence function with the smallest variance matrix; i.e., for any influence function $\varphi(X) \neq \varphi^{\text{eff}}(X)$, $\text{Var}[\varphi^{\text{eff}}(X)] - \text{Var}[\varphi(X)]$ is nonpositive definite.

That an efficient influence function exists and is unique is now easy to see from the geometry of the problem and is explained below.

Theorem 3.17. The efficient influence function is given by

$$\varphi^{\text{eff}}(X) = \varphi^*(X) - \Pi(\varphi^*(X)|\mathcal{T}^\perp) = \Pi(\varphi^*(X)|\mathcal{T}), \tag{25}$$

where $\varphi^*(X)$ is an arbitrary influence function and can explicitly be written as

$$\varphi^{\text{eff}}(X) = \dot{\psi}_{\theta_0} I_{\theta_0}^{-1} \dot{\ell}_{\theta_0}(X). \tag{26}$$

Proof. By Theorem 3.15, the class of influence functions is a linear variety, $\varphi^*(X) + \mathcal{T}^\perp$, where $\varphi^*(X)$ is an arbitrary influence function. Let

$$\varphi^{\text{eff}} := \varphi^* - \Pi(\varphi^*|\mathcal{T}^\perp) = \Pi(\varphi^*|\mathcal{T}).$$

As $\Pi(\varphi^*|\mathcal{T}^\perp) \in \mathcal{T}^\perp$, this implies that φ^{eff} is an influence function. Moreover, φ^{eff} is orthogonal to $\mathcal{T}^\perp$. Consequently, any other influence function can be written as $\varphi = \varphi^{\text{eff}} + t$, with $t \in \mathcal{T}^\perp$. The tangent space $\mathcal{T}$ is an example of a q-replicating linear space as defined by Definition 3.3. As $\varphi^{\text{eff}} \in \mathcal{T}$ and $t \in \mathcal{T}^\perp$, by Theorem 3.4 we obtain

$$\text{Var}[\varphi(X)] = \text{Var}[\varphi^{\text{eff}}(X)] + \text{Var}[t(X)] \geq \text{Var}[\varphi^{\text{eff}}(X)],$$

which demonstrates that φ_{eff}, constructed as above, is an efficient influence function.

We deduce from the argument above that an efficient influence function for $\psi(\theta_0)$ is $\varphi^{\text{eff}} = \Pi(\varphi^*|\mathcal{T})$ is an element of the tangent space $\mathcal{T}$ and hence can be expressed as $\varphi^{\text{eff}}(X) = B_{\text{eff}}\dot{\ell}_{\theta_0}(X)$ for some constant matrix $B_{\text{eff}} \in \mathbb{R}^{q\times k}$. Since $\varphi^{\text{eff}}(X)$ is an influence function, it must also satisfy relationship (15), i.e.,

$$B^{\text{eff}}\mathbb{E}[\dot{\ell}_{\theta_0}(X)\dot{\ell}_{\theta_0}(X)^\top] = \dot{\psi}_{\theta_0} \qquad \Rightarrow \qquad B^{\text{eff}} = \dot{\psi}_{\theta_0} I_{\theta_0}^{-1}.$$

Consequently, the unique efficient influence function is given by $\varphi^{\text{eff}}(X) = \dot{\psi}_{\theta_0} I_{\theta_0}^{-1}\dot{\ell}_{\theta_0}(X)$. □

It is instructive to consider the special case of a separable semiparametric model, i.e., $\theta = (\psi, \eta)$. We first define the important notion of an **efficient score vector** and then show the relationship of the efficient score to the efficient influence function.

Definition 3.18 (Efficient score)**.** The *efficient score* is the residual of the score vector with respect to the parameter of interest after projecting it onto the nuisance tangent space, i.e.,

$$\dot{\ell}_{\theta_0}^{\text{eff}} := \dot{\ell}_{\theta_0}^{(1)} - \Pi(\dot{\ell}_{\theta_0}^{(1)}|\Lambda).$$

Note that by (14), we have

$$\Pi(\dot{\ell}_{\theta_0}^{(1)}|\Lambda) = \mathbb{E}[\dot{\ell}_{\theta_0}^{(1)}(X)\dot{\ell}_{\theta_0}^{(2)}(X)^\top]\{\mathbb{E}[\dot{\ell}_{\theta_0}^{(2)}(X)\dot{\ell}_{\theta_0}^{(2)}(X)^\top]\}^{-1}\dot{\ell}_{\theta_0}^{(2)}.$$

Corollary 3.19. When the parameter θ can be partitioned as (ψ, η), where ψ is the parameter of interest and η is the nuisance parameter, then the efficient influence function

(at θ_0) can be written as

$$\varphi^{\text{eff}} = \{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{\text{eff}}_{\theta_0}(X)^\top]\}^{-1}\dot{\ell}^{\text{eff}}_{\theta_0}. \tag{27}$$

Proof. By construction, the efficient score vector is orthogonal to the nuisance tangent space, i.e., it satisfies condition (17) for any influence function.

By appropriately scaling the efficient score, we can construct an influence function, which we will show is the efficient influence function. We first note that

$$\mathbb{E}_{\theta_0}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{(1)}_{\theta_0}(X)^\top] = \mathbb{E}_{\theta_0}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{\text{eff}}_{\theta_0}(X)^\top].$$

This follows as

$$\mathbb{E}_{\theta_0}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{(1)}_{\theta_0}(X)^\top] = \mathbb{E}_{\theta_0}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{\text{eff}}_{\theta_0}(X)^\top] + \mathbb{E}_{\theta_0}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\Pi(\dot{\ell}^{(1)}_{\theta_0}|\Lambda)^\top],$$

where the second term on the right side equals $0_{q\times q}$ as $\ell^{\text{eff}}_{\theta_0} \perp \Lambda$. Therefore, if we define $\varphi^* := \{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{\text{eff}}_{\theta_0}(X)^\top]\}^{-1}\ell^{\text{eff}}_{\theta_0}$, then φ^* is an influence function as it satisfies the two conditions in Corollary 3.6.

As argued above, the efficient influence function is the unique influence function belonging to the tangent space $\mathcal{T}$. Since both $\dot{\ell}^{(1)}_{\theta_0}$ and $\Pi(\dot{\ell}^{(1)}_{\theta_0}|\Lambda)$ are elements of $\mathcal{T}$, so is

$$\varphi^{\text{eff}} = \varphi^* = \{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{\text{eff}}_{\theta_0}(X)^\top]\}^{-1}[\dot{\ell}^{(1)}_{\theta_0} - \Pi(\dot{\ell}^{(1)}_{\theta_0}|\Lambda)]$$

thus demonstrating that (27) is the efficient influence function for RAL estimators of ψ. □

Remark 3.5. The variance of the efficient influence function is φ^{eff} is $\{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\theta_0}(X)\dot{\ell}^{\text{eff}}_{\theta_0}(X)^\top]\}^{-1}$, the inverse of the variance matrix of the efficient score. If we define

$$I_{11} := \mathbb{E}[\dot{\ell}^{(1)}_{\theta_0}(X)\dot{\ell}^{(1)}_{\theta_0}(X)^\top], \quad I_{22} := \mathbb{E}[\dot{\ell}^{(2)}_{\theta_0}(X)\dot{\ell}^{(2)}_{\theta_0}(X)^\top], \quad \text{and} \quad I_{12} := \mathbb{E}[\dot{\ell}^{(1)}_{\theta_0}(X)\dot{\ell}^{(2)}_{\theta_0}(X)^\top],$$

then we obtain the well-known result that the minimum variance for an efficient RAL estimator is

$$[I_{11} - I_{12}I_{22}^{-1}I_{21}]^{-1}$$

where I_{11}, I_{12}, I_{22} are elements of the information matrix used in likelihood theory.

Exercise: Compare the above variance with the limiting variance of an efficient RAL estimator of ψ if $\eta \equiv \eta_0$ where known.

4 Semiparametric models

4.1 Separated semiparametric models

We assume for most of the following discussion that the data $X_1, \ldots, X_n$ are i.i.d. random variables (vectors) taking values in $(\mathfrak{X}, \mathcal{A})$ with density belonging to the class

$$\mathcal{P} := \left\{ p_{\psi,\eta}(\cdot) : \text{where } \psi \text{ is } q\text{-dimensional and } \eta \text{ is infinite-dimensional} \right\}$$

with respect to some dominating measure μ. Thus we assume a separated semiparametric model[24]. We will denote the "truth" (i.e., the density that generated the data) by $p_0 \equiv p_{\psi_0,\eta_0} \in \mathcal{P}$.

Question: What is a non-trivial lower bound on the variance of any "reasonable" estimator of ψ in the semiparametric model $\mathcal{P}$?

As is often the case in mathematics, infinite-dimensional problems are tackled by first working with a finite-dimensional problem as an approximation and then taking limits to infinity. Therefore, the first step in dealing with a semiparametric model is to consider a simpler finite-dimensional parametric model contained within the semiparametric model and use the theory and methods developed in the previous sections. Towards that end, we define a parametric submodel.

Definition 4.1 (Parametric submodel)**.** A **parametric submodel**, which we will denote by $\mathcal{P}_{\psi,\gamma} = \{p_{\psi,\gamma}(\cdot)\}$, is a class of densities characterized by a finite-dimensional parameter $(\psi, \gamma) \in \Omega_{\psi,\gamma} \subset \mathbb{R}^{q+r}$ ($\Omega_{\psi,\gamma}$ is an open set) such that

(i) $\mathcal{P}_{\psi,\gamma} \subset \mathcal{P}$ (i.e., every density in $\mathcal{P}_{\psi,\gamma}$ belongs to the semiparametric model $\mathcal{P}$), and

(ii) $p_0 \equiv p_{\psi_0,\gamma_0} \in \mathcal{P}_{\psi,\gamma}$ (i.e., the parametric submodel contains the truth).

We further assume that $\mathcal{P}_{\psi,\gamma}$ is DQM at p_0.

Example 4.2 (Cox proportional hazards model)**.** In the proportional hazards model, we assume that

$$\lambda(t|Z) = \lambda(t) \exp(\psi^\top Z),$$

[24] In a *separated* semiparametric model, ψ, the parameter of interest, is finite-dimensional (q-dimensional) and η, the nuisance parameter, is infinite-dimensional, and ψ and η are variationally independent — i.e., any choice of ψ and η in a neighborhood about the true ψ_0 and η_0 would result in a density $p_{\psi,\eta}(\cdot)$ in the semiparametric model. This will allow us, for example, to explicitly define partial derivatives $\partial p_{\psi,\eta_0}(x)/\partial\psi|_{\psi=\psi_0}$.

where $Z = (Z_1, \ldots, Z_q)$ denotes a q-dimensional vector of covariates, $\lambda(t)$ is some arbitrary hazard function[25] (for the response Y) that is left unspecified and hence is infinite-dimensional (whose true value is denoted by λ_0), and ψ is the q-dimensional parameter of interest.

An example of a parametric submodel is as follows. Let $h_1(\cdot), \ldots, h_r(\cdot)$ be r different functions of time that are specified by the data analyst (any smooth function will do). Consider the model

$$\mathcal{P}_{\psi,\gamma} = \left\{ \text{class of densities with hazard function } \lambda(t|Z) = \lambda_0(t) \exp\left[\sum_{j=1}^{r} \gamma_j h_j(t) \right] \exp(\psi^\top Z) \right\}$$

where $\gamma = (\gamma_1, \ldots, \gamma_r) \in \mathbb{R}^r$. Note that is indeed a parametric submodel and the truth is obtained by setting $\psi = \psi_0$ and $\gamma = 0$.

Question: What are "reasonable" semiparametric estimators?

Definition 4.3 (Semiparametric RAL estimator)**.** An estimator for ψ is a RAL estimator for a semiparametric model (at (ψ_0, η_0)) if it is an AL estimator at (ψ_0, η_0) and a regular estimator for every parametric submodel.

Therefore, any influence function of an RAL estimator in a semiparametric model must be an influence function of an RAL estimator within a parametric submodel, i.e.,

$$\begin{aligned} & \{\text{class of influence functions of semiparametric RAL estimators of } \psi \text{ for } \mathcal{P}\} \\ \subset\ & \{\text{class of influence functions of RAL estimators of } \psi \text{ for } \mathcal{P}_{\psi,\gamma}\}\,. \end{aligned}$$

Consequently, the class of semiparametric RAL estimators must be contained within the class of RAL estimators for a parametric submodel. Therefore:

[25] Suppose that Y is a random variable with c.d.f. F. An alternative characterization of the distribution of Y is given by the **hazard function**, or instantaneous rate of occurrence of the event, defined as

$$\lambda(t) = \lim_{dt \to 0} \frac{\mathbb{P}(t \le Y < t + dt \mid Y \ge t)}{dt}.$$

The numerator of this expression is the conditional probability that the event (if Y denotes time of occurrence of an event) will occur in the interval $[t, t + dt)$ given that it has not occurred before, and the denominator is the width of the interval; dividing one by the other we obtain a rate of event occurrence per unit of time. Taking the limit as the width of the interval goes down to zero, we obtain an instantaneous rate of occurrence. If F has a density f and survival function $S \equiv 1 - F$, then the hazard function reduces to $\lambda(t) = f(t)/S(t)$. Moreover, if $Y \ge 0$, we can expresses the survival function in terms of the hazard function as $S(t) = \exp\{-\int_0^t \lambda(x)dx\}$.

- Any influence function of an RAL semiparametric estimator for ψ must be **orthogonal** to all parametric submodel nuisance tangent spaces.

- The variance of any semiparametric RAL influence function must be **greater** than or equal to
$$\left\{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)^\top]\right\}^{-1}$$
for all parametric submodels $\mathcal{P}_{\psi,\gamma}$, where $\dot{\ell}^{\text{eff}}_{\psi,\gamma}$ is the efficient score (at (ψ_0,η_0)) for ψ_0 for the parametric submodel $\mathcal{P}_{\psi,\gamma}$ (note the slight change in our notation for the efficient score). Recall that
$$\dot{\ell}^{\text{eff}}_{\psi,\gamma} = \dot{\ell}^{(1)}_{\psi,\gamma} - \Pi(\dot{\ell}^{(1)}_{\psi,\gamma}|\Lambda_\gamma), \tag{28}$$
where by $\dot{\ell}^{(1)}_{\psi,\gamma}$ we now mean the score vector for the finite-dimensional parameter ψ for the parametric submodel $\mathcal{P}_{\psi,\gamma}$ at the point (ψ_0,η_0), i.e.,
$$\dot{\ell}^{(1)}_{\psi,\gamma}(x) := \frac{\partial}{\partial\psi}\log p_{\psi,\gamma}(x)\Big|_{\psi=\psi_0,\gamma=0} = \frac{\partial}{\partial\psi}\log p_{\psi,\eta_0}(x)\Big|_{\psi=\psi_0}, \qquad \text{for all } x\in\mathfrak{X},$$
(here we have assumed that $\gamma = 0$ gives us η_0; we want the score function to be evaluated at the truth (ψ_0,η_0)) and
$$\Lambda_\gamma := \{B\dot{\ell}^{(2)}_{\psi,\gamma}(X) : B\in\mathbb{R}^{q\times r}\},$$
and $\dot{\ell}^{(2)}_{\psi,\gamma}$ is the score sub-vector for the nuisance parameter γ for the parametric submodel $\mathcal{P}_{\psi,\gamma}$ at the point (ψ_0,η_0). Note that in this new notation, $\dot{\ell}^{(1)}_{\psi,\gamma}$ is the same as $\dot{\ell}^{(1)}_{\psi_0,\eta_0}$ in our previous notation.

Hence, the variance of the influence function for any semiparametric estimator for ψ must be greater than or equal to
$$\left\{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)^\top]\right\}^{-1}. \tag{29}$$
for all parametric submodels $\mathcal{P}_{\psi,\gamma}$.

Definition 4.4 (Locally efficient semiparametric estimator)**.** A semiparametric RAL estimator T_n with asymptotic variance matrix $\mathcal{V}\in\mathbb{R}^{q\times q}$ is said to be **locally efficient** at p_0 if
$$\sup_{\left\{\text{all para. submodels } \mathcal{P}_{\psi,\gamma}\right\}} a^\top\left\{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)^\top]\right\}^{-1}a = a^\top\mathcal{V}a, \qquad \text{for all } a\in\mathbb{R}^q. \tag{30}$$

Definition 4.5 (Semiparametric efficiency bound)**.** The matrix V for which (30) holds is known as the **semiparametric efficiency bound**.

If the same estimator T_n is semiparametric efficient regardless of $p_0 \in \mathcal{P}$, then we say that such an estimator is **globally semiparametric efficient**.

4.2 Semiparametric Nuisance tangent set

Definition 4.6 (Nuisance tangent set for a semiparametric model)**.** The **nuisance tangent space** for a semiparametric model, denoted by Λ, is defined as the mean-square closure[26] of all parametric submodel nuisance tangent spaces Λ_γ.

Specifically, the mean-square closure of the spaces above is defined as the space $\Lambda \subset \mathcal{H}$ (where $\mathcal{H}$ consists of all measurable q-dimensional functions of X, i.e., $h(X)$, with $\mathbb{E}_{\psi_0,\eta_0}[h(X)] = 0$ and $\mathbb{E}_{\psi_0,\eta_0}[h(X)^\top h(X)] < \infty$) such that there exists a sequence $\{B_j \dot{\ell}_j^{(2)}(X)\}_{j\geq 1}$ such that

$$\|h - B_j \dot{\ell}_j^{(2)}\|^2 \to 0, \qquad \text{as } j \to \infty,$$

for a sequence of parametric submodels indexed by j (note that $\|h\|^2 = \mathbb{E}[h(X)^\top h(X)]$). Here by $\dot{\ell}_j^{(2)}(\cdot)$ we mean the the score sub-vector for the nuisance parameter γ for a parametric submodel indexed by j.

Remark 4.1. If we denote by $\mathcal{S}$ the union of all parametric submodel nuisance tangent spaces, then $\Lambda = \bar{\mathcal{S}}$ is the semiparametric nuisance tangent set (here we are equipping the Hilbert space $\mathcal{H}$ with the metric $d(h_1, h_2) = \|h_1 - h_2\|$).

Remark 4.2. Although the set Λ is closed, it may not necessarily be a linear space. However, in most applications it will be a linear space. In fact, Λ is always a cone, i.e., if $h \in \Lambda$ then $\alpha h \in \Lambda$, for any $\alpha > 0$ (Exercise: Show this).

For the rest of this section we assume that Λ, the nuisance tangent set, is a linear subspace.

Before deriving the semiparametric efficient influence function, we first define the semiparametric efficient score vector and give some results regarding the semiparametric efficiency bound.

[26] The closure of a set $\mathcal{S}$ in a metric space is defined as the smallest closed set that contains $\mathcal{S}$, or equivalently, as the set of all elements in $\mathcal{S}$ together with all the limit points of $\mathcal{S}$. The closure of $\mathcal{S}$ is denoted by $\bar{\mathcal{S}}$.

Definition 4.7 (Semiparametric efficient score)**.** The **semiparametric efficient score** for ψ at (ψ_0, η_0) is defined as

$$\dot{\ell}^{\text{eff}} := \dot{\ell}^{(1)}_{\psi_0,\eta_0} - \Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda), \tag{31}$$

where $\dot{\ell}^{(1)}_{\psi_0,\eta_0}$ is the score vector for the finite-dimensional parameter ψ at the point (ψ_0, η_0), i.e.,

$$\dot{\ell}^{(1)}_{\psi_0,\eta_0}(x) = \frac{\partial}{\partial \psi} \log p_{\psi,\eta_0}(x)\Big|_{\psi=\psi_0}, \qquad \text{for all } x \in \mathfrak{X}.$$

Theorem 4.8. Suppose that Λ, the nuisance tangent set, is a linear subspace. Then, the semiparametric efficiency bound, defined by (30), is equal to the inverse of the variance matrix of the semiparametric efficient score; i.e.,

$$\sup_{\left\{\text{all parametric submodels } \mathcal{P}_{\psi,\gamma}\right\}} a^\top \left\{\mathbb{E}[\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)^\top]\right\}^{-1} a = a^\top \{\mathbb{E}[\dot{\ell}^{\text{eff}}(X)\dot{\ell}^{\text{eff}}(X)^\top]\}^{-1} a,$$

for all $a \in \mathbb{R}^q$.

Proof. For simplicity, we take ψ to be a scalar (i.e., $q = 1$). In this case we denote by $\mathcal{V}$ the semiparametric efficiency bound, i.e.,

$$\mathcal{V} := \sup_{\left\{\text{all parametric submodels } \mathcal{P}_{\psi,\gamma}\right\}} \|\dot{\ell}^{\text{eff}}_{\psi,\gamma}\|^{-2}.$$

Since $\Lambda_\gamma \subset \Lambda$, from the definition of $\dot{\ell}^{\text{eff}}_{\psi,\gamma}$ in (28) and $\dot{\ell}^{\text{eff}}$ in (31), it follows that $\|\dot{\ell}^{\text{eff}}(X)\| \leq \|\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)\|$ for all parametric submodels $\mathcal{P}_{\psi,\gamma}$. Hence,

$$\|\dot{\ell}^{\text{eff}}(X)\|^{-2} \geq \sup_{\left\{\text{all parametric submodels } \mathcal{P}_{\psi,\gamma}\right\}} \|\dot{\ell}^{\text{eff}}_{\psi,\gamma}(X)\|^{-2} = \mathcal{V}.$$

To complete the proof of the theorem, we need to show that $\|\dot{\ell}^{\text{eff}}(X)\|^{-2}$ is also less than or equal to $\mathcal{V}$. As $\Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda)$ this means that there exists a sequence of parametric submodels $\mathcal{P}_{\psi,\gamma_j}$ with nuisance score vectors $\dot{\ell}^{(2)}_j$ such that

$$\|\Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda) - B_j\dot{\ell}^{(2)}_j\| \to 0, \qquad \text{as } j \to \infty,$$

for matrices $B_j \in \mathbb{R}^{q \times r_j}$. Therefore,

$$\begin{aligned} \mathcal{V}^{-1} \leq \|\dot{\ell}^{\text{eff}}_{\psi,\gamma_j}\|^2 &= \|\dot{\ell}^{(1)}_{\psi,\gamma_j} - \Pi(\dot{\ell}^{(1)}_{\psi,\gamma}|\Lambda_{\gamma_j})\|^2 \leq \|\dot{\ell}^{(1)}_{\psi,\gamma_j} - B_j\dot{\ell}^{(2)}_j\|^2 \\ &= \|\dot{\ell}^{(1)}_{\psi_0,\eta_0} - \Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda)\|^2 + \|B_j\dot{\ell}^{(2)}_j - \Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda)\|^2 \end{aligned}$$

where the last equality follows from the Pythagorean theorem as $\dot{\ell}^{(1)}_{\psi_0,\eta_0} - \Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda)$ is orthogonal to Λ and $B_j\dot{\ell}^{(2)}_j - \Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda)$ is an element of Λ. Taking $j \to \infty$ implies that

$$\|\dot{\ell}^{(1)}_{\psi_0,\eta_0}(X) - \Pi(\dot{\ell}^{(1)}_{\psi_0,\eta_0}|\Lambda)(X)\|^2 = \|\dot{\ell}^{\text{eff}}(X)\|^2 \geq \mathcal{V}^{-1},$$

which completes the proof. □

Exercise: Prove this result for dimension $q > 1$ using the generalization of the Pythagorean theorem.

Definition 4.9 (Efficient influence function)**.** The **efficient influence function** is defined as the influence function of a semiparametric RAL estimator that achieves the semiparametric efficiency bound (if it exists[27]).

Theorem 4.10. Suppose that there exists a semiparametric RAL estimator for ψ at (ψ_0, η_0). Then any semiparametric RAL estimator for ψ at (ψ_0, η_0) must have an influence function $\varphi(X)$ that satisfies

$$\mathbb{E}[\varphi(X)\dot{\ell}^{(1)}_{\psi_0,\eta_0}(X)^\top] = \mathbb{E}[\varphi(X)\dot{\ell}^{\text{eff}}(X)^\top] = I_{q\times q}, \tag{32}$$

and

$$\Pi[\varphi|\Lambda] = 0, \tag{33}$$

i.e., $\varphi(X)$ is orthogonal to the nuisance tangent set.

Further, in this situation, the efficient influence function is now defined as the unique element satisfying conditions (32) and (33) whose variance matrix equals the efficiency bound and is equal to

$$\varphi^{\text{eff}} = \{\mathbb{E}[\dot{\ell}^{\text{eff}}(X)\dot{\ell}^{\text{eff}}(X)^\top]\}^{-1}\dot{\ell}^{\text{eff}}. \tag{34}$$

Proof. We first prove condition (33). To show that $\varphi(X)$ is orthogonal to Λ, we must prove that $\langle\varphi, h\rangle = 0$ for all $h \in \Lambda$. Given any $h \in \Lambda$, there exists a sequence $B_j\dot{\ell}^{(2)}_j(X)$,

[27] There is no guarantee that an semiparametric RAL estimator can be derived.

parametric submodels Λ_j indexed by j, such that

$$\|h - B_j \dot{\ell}_j^{(2)}\| \to 0, \qquad \text{as } j \to \infty.$$

Hence,

$$\langle \varphi, h \rangle = \langle \varphi, B_j \dot{\ell}_j^{(2)} \rangle + \langle \varphi, h - B_j \dot{\ell}_j^{(2)} \rangle = 0 + \langle \varphi, h - B_j \dot{\ell}_j^{(2)} \rangle,$$

as any influence function of a semiparametric RAL estimator for ψ must be an influence function for an RAL estimator in a parametric submodel, and thus by (17), φ is orthogonal to Λ_j. By the Cauchy-Schwartz inequality, we obtain

$$|\langle \varphi, h \rangle| \leq \|\varphi\| \|h - B_j \dot{\ell}_j^{(2)}\|.$$

Taking limits as $j \to \infty$ gives us the desired result.

To prove (32), we note that by (16) that $\varphi(X)$ must satisfy $\mathbb{E}[\varphi(X) \dot{\ell}^{(1)}_{\psi_0,\eta_0}(X)^\top] = I_{q\times q}$. The second equality in (32) can also be shown. Further it can be shown that the left side of (34) is an influence function whose variance matches the semiparametric efficiency bound (Exercise: Prove these two statements). □

4.3 Non-separated semiparametric model

Suppose now that the data $X_1, \ldots X_n$ are i.i.d. p_θ, where $\theta \in \Theta$, Θ being an infinite dimensional space. The interest is on estimating $\psi : \theta \mapsto \mathbb{R}^q$, where $\psi(\cdot)$ is a "smooth"[28] q-dimensional function of θ.

Definition 4.11 (Semiparametric tangent set)**.** The **semiparametric tangent set** is defined as the mean-square closure of all parametric submodel tangent spaces.

The following result characterizes all semiparametric influence functions of $\psi(\theta)$ and the efficient influence function.

Theorem 4.12. If a semiparametric RAL estimator for $\psi(\theta)$ exists, then the influence function of this estimator must belong to the space of all influence functions, the linear variety $\varphi(X) + \mathcal{T}^\perp$, where $\varphi(X)$ is the influence function of any semiparametric RAL estimator for $\psi(\theta)$ and $\mathcal{T}$ is the semiparametric tangent space. Moreover, if a RAL estimator for $\psi(\theta)$ exists that achieves the semiparametric efficiency bound (i.e., a semiparametric

[28] This notion will be made precise later.

efficient estimator), then the influence function of this estimator must be the unique and well-defined element

$$\varphi^{\text{eff}} := \varphi - \Pi(\varphi|\mathcal{T}^{\perp}) = \Pi(\varphi|\mathcal{T}).$$

Proof. Exercise: Show this. □

Remark 4.3. What is not clear is whether there exist semiparametric estimators that will have influence functions corresponding to the elements of the Hilbert space satisfying conditions (32) and (33) of Theorem 4.10 or Theorem 4.12 (although we might expect that arguments similar to those in Section 3.4, used to construct estimators for finite-dimensional parametric models, will extend to semiparametric models as well).

In many cases, deriving the space of influence functions, or even the space orthogonal to the nuisance tangent space, for semiparametric models, will suggest how semiparametric estimators may be constructed and even how to find locally or globally efficient semiparametric estimators.

4.4 Tangent space for nonparametric models

Suppose we are interested in estimating some q-dimensional parameter ψ for a nonparametric model. That is, let $X_1, \ldots, X_n$ be i.i.d. random variables (taking values in $(\mathfrak{X}, \mathcal{A})$) with arbitrary density $p(\cdot)$ with respect to a dominating measure μ, where the only restriction on $p(\cdot)$ is that

$$p(x) \geq 0 \quad \text{for all } x \in \mathfrak{X}, \qquad \text{and} \qquad \int p(x) d\mu(x) = 1.$$

Theorem 4.13. The tangent space (i.e., the mean-square closure of all parametric submodel tangent spaces) is the entire Hilbert space $\mathcal{H}$.

4.5 Semiparametric restricted moment model

A common statistical problem is to model the relationship of a response variable Y as a function of a vector of covariates X. The following example is taken from Tsiatis (2006, Section 4.5), where complete proofs of many of the results in this section are provided.

Suppose that we have i.i.d. data $\{(Y_i, X_i)\}_{i=1}^n$ where

$$Y_i = \mu(X_i, \beta) + \epsilon_i, \qquad \mathbb{E}[\epsilon_i|X_i] = 0, \tag{35}$$

or equivalently

$$\mathbb{E}[Y_i|X_i] = \mu(X_i, \beta).$$

Here $\mu(X, \beta)$ is a known function of $X \in \mathfrak{X} \subset \mathbb{R}^q$ and the unknown q-dimensional parameter β. The function $\mu(X, \beta)$ may be linear or nonlinear in β, and it is assumed that β is finite-dimensional. For example, we might consider a linear model where $\mu(X, \beta) = \beta^\top X$ or a nonlinear model, such as a log-linear model, where $\mu(X, \beta) = \exp(\beta^\top X)$. No other assumptions will be made on the class of probability distributions other than the constraint given by the conditional expectation of Y given X stated above.

The density of a single observation, denoted by $p(\cdot)$, belongs to the semiparametric model

$$\mathcal{P} = \{p_{\beta,\eta}(\cdot)\}$$

defined with respect to the dominating measure $\lambda \times \nu_X$. As there is a one-to-one transformation of (Y, X) and (ϵ, X), we can express the density

$$p_{Y,X}(y, x) = p_{\epsilon,X}(y - \mu(x, \beta), x),$$

where $p_{\epsilon,X}$ is a density with respect to the dominating measure $\lambda \times \nu_X$. The density of (ϵ, X) can be expressed as

$$p_{\epsilon,X}(e, x) = \eta_1(e, x)\ \eta_2(x),$$

where $\eta_1(e, x) = p_{\epsilon|X}(e|x)$ is any nonnegative function such that

$$\int \eta_1(e, x)de = 1, \ \text{ for all } x \in \mathfrak{X}, \qquad \int e\eta_1(e, x)de = 0, \ \text{ for all } x \in \mathfrak{X} \tag{36}$$

and $p_X(x) = \eta_2(x)$ is a nonnegative function of x such that $\int \eta_2(x)d\nu_X(x) = 1$. Suppose that the true density generating the data is denoted by

$$p_0(y, x) = \eta_{10}(y - \mu(x, \beta_0), x)\ \eta_{20}(x).$$

To develop the semiparametric theory and define the semiparametric nuisance tangent space, we first consider parametric submodels. We will consider parametric submodels

$$p_{\epsilon|X}(e|x, \gamma_1) \qquad \text{and} \qquad p_X(x, \gamma_2),$$

where γ_1 is an r_1-dimensional vector and γ_2 is an r_2-dimensional vector. Thus $\gamma = (\gamma_1, \gamma_2)$

is an r-dimensional vector, $r = r_1 + r_2$. This parametric submodel is given as

$$\mathcal{P}_{\beta,\gamma} = \left\{ p_{\beta,\gamma_1,\gamma_2}(y,x) = p_{\epsilon|X}(e|x,\gamma_1) p_X(x,\gamma_2),\ (y,x) \in \mathbb{R} \times \mathfrak{X},\ \text{ for } \ (\beta,\gamma_1,\gamma_2) \in \Omega_{\beta,\gamma} \subset \mathbb{R}^{q+r} \right\}.$$

It can be easily seen that the score vector for the nuisance parameters, at the truth $(\beta_0, \gamma_{10}, \gamma_{20})$, equals

$$\dot{\ell}_{\gamma_1}(y,x) = \frac{\partial \log p_{\epsilon|X}(y - \mu(x,\beta)|x,\gamma_{10})}{\partial \gamma_1}\Big|_{\beta=\beta_0,\gamma=\gamma_0}, \quad \text{and} \quad \dot{\ell}_{\gamma_2}(y,x) = \frac{\partial \log p_X(x,\gamma_{20})}{\partial \gamma_2}$$

where $\gamma = (\gamma_1, \gamma_2)$. Since we are taking derivatives with respect to γ_1 and γ_2 and leaving β fixed for the time being at β_0 we will use the simplifying notation that $\epsilon = Y - \mu(X, \beta_0)$.

A typical element in the parametric submodel nuisance tangent space is given by

$$B_{q\times r}\dot{\ell}_{\gamma}(\epsilon, X) = B^{(1)}_{q\times r_1}\dot{\ell}_{\gamma_1}(\epsilon, X) + B^{(2)}_{q\times r_2}\dot{\ell}_{\gamma_2}(X).$$

Therefore, the parametric submodel nuisance tangent space

$$\Lambda_\gamma = \{B\dot{\ell}_\gamma(\epsilon, X) :\ B \in \mathbb{R}^{q\times r}\}$$

can be written as the direct sum $\Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2}$, where

$$\Lambda_{\gamma_1} = \{B\dot{\ell}_{\gamma_1}(\epsilon, X) :\ B \in \mathbb{R}^{q\times r_1}\} \quad \text{and} \quad \Lambda_{\gamma_2} = \{B\dot{\ell}_{\gamma_2}(X) :\ B \in \mathbb{R}^{q\times r_2}\}.$$

It is easy to show that the space Λ_{γ_1} is orthogonal to the space Λ_{γ_2}, as we demonstrate in the following lemma.

Lemma 4.14. The space Λ_{γ_1} is orthogonal to the space Λ_{γ_2}.

Proof. Exercise: Show this. □

Thus, the semiparametric nuisance tangent space

$$\Lambda = \{\text{mean-square closure of } \Lambda_{\gamma_1} \oplus \Lambda_{\gamma_2},\ \text{ over all parametric submodels}\}.$$

As γ_1, γ_2 are variationally independent — i.e., proper densities in the parametric submodel can be defined by considering any combination of γ_1 and γ_2 — this implies that $\Lambda = \Lambda_{1s} \oplus \Lambda_{2s}$, where

$$\Lambda_{1s} = \{\text{mean-square closure of all } \Lambda_{\gamma_1}\} \quad \text{and} \quad \Lambda_{2s} = \{\text{mean-square closure of all } \Lambda_{\gamma_2}\}.$$

4.5.1 The Space Λ_{2s}

Since here we are considering marginal distributions of X with no restrictions, finding the space Λ_{2s} is similar to finding the nuisance tangent space for the nonparametric model given in Section 4.4.

Theorem 4.15. The space Λ_{2s} consists of all q-dimensional mean-zero functions of X with finite variance.

4.5.2 The Space Λ_{1s}

Theorem 4.16. The space Λ_{2s} is the space of all q-dimensional random functions $a(\epsilon, X)$ that satisfy

$$\mathbb{E}[a(\epsilon, X)|X] = 0_{q\times 1}, \tag{37}$$

and

$$\mathbb{E}[a(\epsilon, X)\epsilon] = 0_{q\times 1}, \tag{38}$$

Thus, Λ_{1s} is the intersection of two linear subspaces, i.e.,

$$\Lambda_{1s} = \Lambda_{1sa} \cap \Lambda_{1sb}$$

where

$$\Lambda_{1sa} := \left\{ a_{q\times 1}(\epsilon, X) : \mathbb{E}[a(\epsilon, X)|X] = 0_{q\times 1} \right\}$$

and

$$\Lambda_{1sb} := \left\{ a_{q\times 1}(\epsilon, X) : \mathbb{E}[a(\epsilon, X)\epsilon|X] = 0_{q\times 1} \right\}$$

One can also demonstrate easily that $\Lambda_{1s} \perp \Lambda_{2s}$; see Tsiatis (2006, Lemma 4.2).

Thus, the nuisance tangent space $\Lambda = (\Lambda_{1sa} \cap \Lambda_{1sb}) \oplus \Lambda_{2s}$. The following relationships hold among the above defined subspaces.

Theorem 4.17. We have

$$\Lambda_{1sa} = \Lambda_{2s}^{\perp}, \qquad \Lambda_{2s} \subset \Lambda_{1sb}$$

and thus,

$$\Lambda = (\Lambda_{1sa} \cap \Lambda_{1sb}) \oplus \Lambda_{2s} = \Lambda_{1sb}.$$

4.5.3 Influence Functions

The key to deriving the space of influence functions is first to identify elements of the Hilbert space that are orthogonal to Λ. Equivalently, the space $\Lambda^{\perp}$ is the linear space of residuals

$$h(\epsilon, X) - \Pi(h|\Lambda)(\epsilon, X)$$

for all $h(\epsilon, X) \in \mathcal{H}$.

Theorem 4.18. The space orthogonal to the nuisance tangent space, $\Lambda^{\perp}$, or equivalently $\Lambda_{1sb}^{\perp}$, is

$$\big\{A_{q\times 1}(X)\epsilon : A(X) \text{ is a vector of arbitrary } q\text{-dimensional function of } X\big\}.$$

Moreover, the projection of any arbitrary element $h(\epsilon, X) \in \mathcal{H}$ onto Λ_{1sb} satisfies satisfies

$$h(\epsilon, X) - \Pi(h|\Lambda)(\epsilon, X) = g_{q\times 1}(X)\epsilon,$$

where

$$g(X) := \mathbb{E}[h(\epsilon, X)\epsilon|X]\{\mathbb{E}[\epsilon^2|X]\}^{-1}.$$

We have thus demonstrated that, for the semiparametric restricted moment model, any element of the Hilbert space perpendicular to the nuisance tangent space is given by

$$A(X)\epsilon \qquad \text{or} \qquad A(X)\{Y - \mu(X, \beta_0)\}.$$

Influence functions of RAL estimators for β (i.e., $\varphi(\epsilon, X)$) are normalized versions of elements perpendicular to the nuisance tangent space. That is, the space of influence functions, as well as being orthogonal to the nuisance tangent space, must also satisfy condition 32, namely,

$$\mathbb{E}[\varphi(\epsilon, X)\dot{\ell}_{\beta_0}(\epsilon, X)^{\top}] = I_q,$$

where $\dot{\ell}_{\beta_0}$ is the score vector with respect to the parameter β at the true parameter value (β_0).

If we start with any $A(X)$, and define $\varphi(\epsilon, X) = C_{q\times q}A(X)\epsilon$, where $C_{q\times q}$ is a $q \times q$ constant matrix (i.e., normalization factor), then the above condition implies

$$\mathbb{E}[CA(X)\epsilon\dot{\ell}_{\beta_0}(\epsilon, X)^{\top}] = I_q \qquad \Rightarrow \qquad C = \{\mathbb{E}[A(X)\epsilon\dot{\ell}_{\beta_0}(\epsilon, X)^{\top}]\}^{-1}.$$

Since a typical element orthogonal to the nuisance tangent space is given by $A(X)\{Y - \mu(X, \beta_0)\}$ and since a typical influence function is given by $CA(X)\{Y - \mu(X, \beta_0)\}$, where C is defined above, this motivates us to consider a Z-estimator for β of the form

$$\sum_{i=1}^{n} CA(X_i)\{Y_i - \mu(X_i, \beta)\} = 0.$$

Because C is a multiplicative constant matrix, then, as long as C is invertible, this is equivalent to solving the equation

$$\sum_{i=1}^{n} A(X_i)\{Y_i - \mu(X_i, \beta)\} = 0.$$

This logic suggests that estimators can often be motivated by identifying elements orthogonal to the nuisance tangent space, a theme that will be used frequently throughout the course.

4.5.4 The Efficient Influence Function

To derive an efficient semiparametric estimator, we must find the efficient influence function. For this, we need to derive the efficient score (i.e., the residual after projecting the score vector with respect to β onto the nuisance tangent space Λ). we can show that

$$\dot{\ell}^{\text{eff}} := \dot{\ell}_{\beta_0} - \Pi(\dot{\ell}_{\beta_0}|\Lambda) = \mathbb{E}[\dot{\ell}_{\beta_0}(\epsilon, X)\epsilon]V(X)^{-1}\epsilon.$$

We can further show that (Exercise) the efficient score is given by

$$\dot{\ell}^{\text{eff}} = D(X)^{\top}V(X)^{-1}\epsilon,$$

where

$$V(x) := \mathbb{E}[\epsilon^2|X = x] \qquad \text{and} \qquad D(x) = \frac{\partial \mu(x, \beta_0)}{\partial \beta}.$$

Further, the optimal estimator is obtained by solving the estimating equation (Exercise)

$$\sum_{i=1}^{n} D(X_i)^{\top}V(X_i)^{-1}\{Y_i - \mu(X_i, \beta)\} = 0.$$

Thus, the semiparametric efficiency bound is given as

$$\mathcal{V} = \{\mathbb{E}[\dot{\ell}^{\text{eff}}(\epsilon, X)\dot{\ell}^{\text{eff}}(\epsilon, X)^\top]\}^{-1} = \{\mathbb{E}[D(X)^\top V(X)^{-1} D(X)]\}^{-1}.$$

5 Semiparametric theory

Suppose that we observe a random sample $X_1, \ldots, X_n$ from a distribution P that is known to belong to a set $\mathcal{P}$ of probability measures on the sample space $(\mathfrak{X}, \mathcal{A})$. The goal is to estimate the value of $\psi(P)$ where $\psi : \mathcal{P} \to \mathbb{R}^q$ is a functional. In this section we develop a notion of information for estimating $\psi(P)$. For simplicity we assume that all measures are dominated by a common σ-finite measure μ.

To estimate the parameter $\psi(P)$ given the model $\mathcal{P}$ is certainly harder than to estimate this parameter given that P belongs to a submodel $\mathcal{P}_0 \subset P$. For every smooth parametric submodel $\mathcal{P}_0 = \{P_t : t \in \mathbb{R}\} \subset \mathcal{P}$, such that $P_0 \equiv P$, we can calculate the Fisher information for estimating $\psi(P)$. Then the information for estimating $\psi(P)$ in the whole model is certainly not bigger than the infimum of the informations over all submodels. We shall simply define the information for the whole model as this infimum. A submodel for which the infimum is attained (if there is one such) is called *least favorable* or a "hardest" submodel. In most situations it suffices to consider one-dimensional submodels $\mathcal{P}_0$.

5.1 Tangent sets

Definition 5.1 (Path)**.** A *path* (P_t) at P within the model $\mathcal{P}$ is a mapping where $t \mapsto P_t \in \mathcal{P}$, for $t \in [0, +\infty)$, with $P_0 = P$.

Definition 5.2 (Differentiable path)**.** A *differentiable path* (P_t) at P is a path at P such that there exists $g : \mathfrak{X} \to \mathbb{R}$ measurable such that, if p_t and p denote the respective densities of P_t and P with respect to μ, as $t \to 0$,

$$\int \left[\frac{\sqrt{p_t} - \sqrt{p}}{t} - \frac{1}{2} g \sqrt{p} \right]^2 d\mu \to 0. \tag{39}$$

The function g is called the *score* function of the path (P_t) at P. One also says that the model $\mathcal{P}$ is differentiable in quadratic mean (DQM) at P along the submodel $t \mapsto P_t$ with score function g.

Remark 5.1. One can check that this definition does not depend on the choice of the dominating measure. If there is no fixed dominating measure μ, one can extend the definition by checking it along all sequences $t_n \to 0$ and for each particular sequence (t_n) work with a convex combination of the countably many measures $P_{t_n} + P$ as the dominating measure.

In all cases, one sometimes adopts the following shorthand notation for (39):

$$\int \left[\frac{\sqrt{dP_t} - \sqrt{dP}}{t} - \frac{1}{2} g \sqrt{dP} \right]^2 \to 0. \tag{40}$$

Lemma 5.3. Assume that the path $t \mapsto P_t$ in $\mathcal{P}$ satisfies (40) with score function g. Then g belongs to $L_2(P)$ and $Pg = 0$.

Each particular differentiable path through the model leads to a score function g. One can then consider the set of all scores.

Definition 5.4 (Tangent set)**.** A *tangent set* of the model $\mathcal{P}$ at P is a collection of score functions of paths at P. The tangent set is denoted by $\dot{\mathcal{P}}_P$.

Definition 5.5 (Tangent space)**.** When the tangent set is, in fact, a linear space it is called a *tangent space*.

Remark 5.2. A tangent set can be identified with a subset of $L_2(P)$. Sometimes, but not always, one looks for a 'maximal tangent set', which is defined as the set of all possible score functions arising from differentiable paths. Note that the maximal tangent set is always a cone.

Remark 5.3. Geometrically, we may visualize the model $\mathcal{P}$, or rather the corresponding set of "square roots of measures" $\sqrt{p}$, as a subset of the unit ball of $L_2(\mu)$, and $\dot{\mathcal{P}}_P$, or rather the set of all objects $\frac{1}{2} g \sqrt{p}$, as its tangent set.

Remark 5.4. Usually, we construct the submodels $t \mapsto P_t$ such that, for every x,

$$g(x) = \frac{\partial}{\partial t} \log p_t(x) \Big|_{t=0}. \tag{41}$$

5.1.1 LAN and DQM property

As in a parametric model, the existence of a differentiable path P_t (with score g) implies that the likelihood ratio along the path is (locally) asymptotically quadratic (with curvature Pg^2) with a scaled normal distribution.

Theorem 5.6. Assume that the path $t \to P_t$ in $\mathcal{P}$ satisfies (40), then, for

$$\log \prod_{i=1}^{n} \frac{dP_{t/\sqrt{n}}}{dP}(X_i) = \frac{t}{\sqrt{n}} \sum_{i=1}^{n} g(X_i) - \frac{t^2}{2} Pg^2 + o_p(1). \tag{42}$$

In turn, the DQM property is valid if the model is smooth enough. Existence of one continuous derivative plus a continuity property suffices. For the next lemma, one supposes that paths are defined on an open neighborhood $t \in (-\epsilon, \epsilon)$ of $t = 0$ (this requires to extend the path to the left of zero, which is rarely a problem in practice).

Lemma 5.7. Let p_t be, for any t, a probability density relative to a fixed measure μ. Suppose that $t \mapsto \sqrt{p_t(x)}$ is continuously differentiable in an open neighborhood of $t = 0$ for all $x \in \mathfrak{X}$, and that $t \mapsto \int \dot{p}_t^2/p_t d\mu$ is finite and continuous in this neighborhood. Then the map $t \mapsto \sqrt{p_t}$ satisfies (40) with $p = p_0$ and score $g = \dot{p}_0/p_0$.

5.2 Information lower bounds and efficient influence function

Next, to define the information for estimating $\psi(P)$, we need a notion of "smoothness" for the functional ψ, which is introduced next.

Definition 5.8 (Differentiable functional)**.** We say that $\psi : \mathcal{P} \to \mathbb{R}^p$ is *differentiable* at P relative to a given tangent set $\dot{\mathcal{P}}_P$ if there exists a *continuous linear map* $\dot{\psi}_P : L_2(P) \to \mathbb{R}^p$ such that for every $g \in \dot{\mathcal{P}}_P$ and a submodel $t \mapsto P_t$ with score function g,

$$\frac{\psi(P_t) - \psi(P)}{t} \to \dot{\psi}_P g. \tag{43}$$

By the Riesz representation theorem[29], the map $\dot{\psi}_P$ can always be written in the form of an inner product with a fixed vector-valued, measurable function $\tilde{\psi}_P : \mathfrak{X} \to \mathbb{R}^p$, i.e.,

$$\dot{\psi}_P g = \langle \tilde{\psi}_P, g \rangle_P = \int \tilde{\psi}_P g \, dP. \tag{44}$$

Definition 5.9 (Efficient influence function)**.** The unique[30] function $\tilde{\psi}_P$ which satisfies (44) and whose coordinate functions are contained in $\overline{\text{lin}}\,\dot{\mathcal{P}}_P$ (the closure of the linear span of the tangent set) is called the *efficient influence* function.

[29] Let $(H, \langle \cdot, \cdot \rangle)$ be a Hilbert space, and let H^* denote its dual space, consisting of all continuous linear functionals from H into the field $\mathbb{R}$. If x is an element of H, then the function $\varphi_x : H \to \mathbb{R}$ defined by $\varphi_x(y) = \langle y, x \rangle$, is an element of H^*. The Riesz representation theorem states that every element of H^* can be written uniquely in this form.

[30] The function $\tilde{\psi}_P$, defined in (44), is not uniquely defined by the functional ψ and the model $\mathcal{P}$, because only inner products of $\tilde{\psi}_P$ with elements of the tangent set are specified, and the tangent set does not span all of $L_2(P)$. However, it is always possible to find a candidate $\tilde{\psi}_P$ whose coordinate functions are contained in $\overline{\text{lin}}\,\dot{\mathcal{P}}_P$; it can be found as the projection of any other "influence function" onto the closed linear span of the tangent set.

Remark 5.5. In the preceding set-up the tangent sets $\dot{\mathcal{P}}_P$ are made to depend both on the model $\mathcal{P}$ and the functional ψ. We do not always want to use the "maximal tangent set", which is the set of all score functions of differentiable submodels $t \mapsto P_t$, because the parameter ψ may not be differentiable relative to it (can we find one such example?). We consider every subset of a tangent set a tangent set itself.

Remark 5.6 (Parametric model)**.** Consider a parametric model $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ with parameter θ ranging over an open subset Θ of $\mathbb{R}^k$ given by densities p_θ with respect to some dominating measure μ. Suppose that there exists a vector-valued measurable map $\dot{\ell}_\theta$ such that (39) holds (i.e., the family is DQM at θ). Then a tangent set at P_θ is given by the linear space $\{h^\top \dot{\ell}_\theta : h \in \mathbb{R}^k\}$ spanned by the score functions for the coordinates of the parameter θ.

If the Fisher information matrix $I_\theta = P_\theta[\dot{\ell}_\theta \dot{\ell}_\theta^\top]$ is invertible, then every map $\chi : \Theta \to \mathbb{R}^p$ that is differentiable in the ordinary sense as a map between Euclidean spaces is differentiable as a map $\psi(P_\theta) = \chi(\theta)$ on the model relative to the given tangent space. This follows because the submodel $t \mapsto P_{\theta+th}$ has score $h^\top \dot{\ell}_\theta$ and

$$\frac{\partial}{\partial t}\chi(\theta + th)\Big|_{t=0} = \dot{\chi}_\theta h = P_\theta[(\dot{\chi}_\theta I_\theta^{-1}\dot{\ell}_\theta)h^\top \dot{\ell}_\theta],$$

where $\dot{\chi}_\theta$ is the Jacobian matrix at θ of the map $\chi(\cdot)$. The above display shows that the function $\tilde{\psi}_{P_\theta} = \dot{\chi}_\theta I_\theta^{-1}\dot{\ell}_\theta$ is the efficient influence function.

Remark 5.7 (Nonparametric models)**.** Suppose that $\mathcal{P}$ consists of all probability laws on the sample space. Then a tangent set at P consists of

$$\text{all measurable functions } g \text{ satisfying } \int g dP = 0 \text{ and } \int g^2 dP < \infty.$$

Because a score function necessarily has mean zero, this is the maximal tangent set. It suffices to exhibit suitable one-dimensional submodels. For a bounded function g, consider for instance the exponential family $dP_t(x) = c(t)\exp(tg(x))dP(x)$ or, alternatively, the model $dP_t(x) = (1 + tg(x))dP(x)$. Both models have the property that (41) holds, for every x. By a direct calculation we can show that both models also have score function g at $t = 0$. For an unbounded function g, these submodels are not necessarily well-defined. However, the models have the common structure $dP_t(x) = c(t)\xi(tg(x))dP(x)$ for a nonnegative function ξ with $\xi(0) = \xi'(0) = 1$. The function $\xi(x) = 2(1 + e^{-2x})^{-1}$ is bounded and can be used with any g.

5.2.1 Information

To motivate the definition of *information*, assume for simplicity that the parameter $\psi(P)$ is one-dimensional. The Fisher information about t in a submodel $t \mapsto P_t$ with score function g at $t = 0$ is Pg^2. Thus, the "optimal asymptotic variance" for estimating $\psi(P)$ is the Cramér-Rao bound (see (2)):

$$\frac{(\partial\psi(P_t)/dt)^2}{Pg^2} = \frac{\langle\tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P}.$$

The supremum of this expression over all submodels, equivalently over all elements of the tangent set, is a lower bound for estimating $\psi(P)$ given the model $\mathcal{P}$, when the "true measure" is P. This supremum can be expressed in the norm of the efficient influence function $\tilde{\psi}_P$.

Lemma 5.10. Suppose that the functional $\psi : \mathcal{P} \to \mathbb{R}$ is differentiable at P relative to the tangent set $\dot{\mathcal{P}}_P$. Then

$$\sup_{g \in \operatorname{lin}\dot{\mathcal{P}}_P} \frac{\langle\tilde{\psi}_P, g\rangle_P^2}{\langle g, g\rangle_P} = P[\tilde{\psi}_P^2].$$

Proof. The above result is a consequence of the Cauchy-Schwarz inequality $(P[\tilde{\psi}_P g])^2 \leq (P\tilde{\psi}_P^2)(Pg^2)$ and the fact that, by definition, the efficient influence function $\tilde{\psi}_P$ is contained in the closure of $\operatorname{lin}\dot{\mathcal{P}}_P$. To conclude take an approximating sequence of g's in the former linear span. □

Thus, the squared norm $P[\tilde{\psi}_P^2]$ of the efficient influence function plays the role of an "optimal asymptotic variance", just as does the expression $\dot{\psi}_\theta I_\theta^{-1}\dot{\psi}_\theta^\top$ in Section 2. Similar considerations (take linear combinations) show that the "optimal asymptotic covariance" for estimating a higher-dimensional parameter $\psi : \mathcal{P} \to \mathbb{R}^p$ is given by the covariance matrix $P[\tilde{\psi}_P\tilde{\psi}_P^\top]$ of the efficient influence function.

Example 5.11 (Example of a specific functional)**.** Suppose one wants to estimate the linear functional $\psi(P) = \int a(u)dP(u) =: Pa$, for some function $a \in L_2(P)$. Consider the paths P_t with density $p_t(x) = 1 + tg(x)$ with respect to P and bounded score function g. The tangent set is not maximal but its closure in $L_2(P)$ is the maximal tangent set

$$\dot{\mathcal{P}}_P^{NP} = \left\{ g : \mathfrak{X} \to \mathbb{R} \mid g \text{ measurable}, \int gdP = 0, \int g^2 dP < \infty \right\},$$

so we can work with these simple paths. For any g,

$$\frac{\psi(P_t) - \psi(P)}{t} = \int a(u)g(u)dP(u) = \langle a, g\rangle_P.$$

We need to find $\tilde{\psi}_P \in \dot{\mathcal{P}}_P^{NP}$ such that $\langle a, g\rangle_P = \langle \tilde{\psi}_P, g\rangle_P$, for any $g \in \dot{\mathcal{P}}_P^{NP}$. The previous identity implies that $\tilde{\psi}_P$ is the orthogonal projection onto the $\dot{\mathcal{P}}_P^{NP}$. Clearly then $\tilde{\psi}_P = a - Pa$ and is differentiable. The corresponding information bound is $P\tilde{\psi}_P^2$. Notice that this bound is 'attained' by the empirical distribution $\mathbb{P}_n[a] = n^{-1}\sum_{i=1}^n a(X_i)$, in the sense that, as $n \to \infty$, under P,

$$\sqrt{n}(\mathbb{P}_n a - Pa) \xrightarrow{d} N(0, P\tilde{\psi}_P^2).$$

Definition 5.12 (Regular estimator)**.** For every g in a given tangent set $\dot{\mathcal{P}}_P$, write $(P_{t,g})$ for a submodel with score function g along which the function ψ is differentiable. As usual, an estimator T_n is a measurable function $T_n(X_1, \ldots, X_n)$ of the observations. An estimator sequence T_n is called *regular* at P for estimating $\psi(P)$ (relative to $\dot{\mathcal{P}}_P$) if there exists a probability measure L such that

$$\sqrt{n}\left(T_n - \psi(P_{1/\sqrt{n},g})\right) \overset{P_{1/\sqrt{n},g}}{\to_d} L, \qquad \text{for every } g \in \dot{\mathcal{P}}_P.$$

Theorem 5.13 (Convolution)**.** Let $\psi : \mathcal{P} \to \mathbb{R}^p$ be differentiable at P relative to the tangent cone $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. Then the asymptotic covariance matrix of every regular sequence of estimators is bounded below by $P[\tilde{\psi}_P\tilde{\psi}_P^\top]$. Furthermore, if $\dot{\mathcal{P}}_P$ is a convex cone, then every limit distribution L of a regular sequence of estimators can be written $L = N(0, P[\tilde{\psi}_P\tilde{\psi}_P^\top]) \star M$ for some probability distribution M.

Definition 5.14 (Asymptotically efficient estimator)**.** We shall say that an estimator sequence is *asymptotically efficient* at P, if it is regular at P with limit distribution $L = N(0, [\tilde{\psi}_P\tilde{\psi}_P^\top])$.

Lemma 5.15. Let the function $\psi : \mathcal{P} \to \mathbb{R}^p$ be differentiable at P relative to the tangent cone $\dot{\mathcal{P}}_P$ with efficient influence function $\tilde{\psi}_P$. A sequence of estimators T_n is regular at P with limiting distribution $N(0, P[\tilde{\psi}_P\tilde{\psi}_P^\top])$ if and only if it satisfies[31]

$$\sqrt{n}(T_n - \psi(P)) = \frac{1}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}_P(X_i) + o_p(1).$$

[31]The efficient influence function $\tilde{\psi}_P$ plays the same role as the normalized score function $I_\theta^{-1}\dot{\ell}_\theta$ in parametric models.

5.3 Efficient score function

A function $\psi(P)$ of particular interest is the parameter θ in a semiparametric model $\{P_{\theta,\eta} : \theta \in \Theta, \eta \in H\}$, where Θ is an open subset of $\mathbb{R}^k$ ($k \geq 1$) and H is an arbitrary set, typically infinite dimensional. The information bound for the functional of interest $\psi(P_{\theta,\eta}) = \theta$ can be conveniently expressed through the "efficient score function", which we define below.

Definition 5.16 (Nuisance tangent set)**.** As submodels, we use paths of the form $t \mapsto P_{\theta+ta,\eta_t}$, for given paths $t \mapsto \eta_t$ in the parameter set H. The score functions for such submodels (if they exist) will typically have the form of a sum of "partial derivatives" with respect to θ and η. If $\dot{\ell}_{\theta,\eta}$ is the ordinary score function for θ in the model where η is fixed, then we expect

$$\frac{\partial}{\partial t} \log dP_{\theta+t,\eta_t}(x)\Big|_{t=0} = a^\top \dot{\ell}_{\theta,\eta} + g.$$

The function g has the interpretation of a score function for η when θ is fixed, and will run through (possibly) an infinite-dimensional set if we are concerned with a "true" semiparametric model. We refer to this set as the *tangent set* for η (or the **nuisance tangent set**), and denote it by ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$.

The parameter $\psi(P_{\theta+ta,\eta_t}) = \theta + ta$ is certainly differentiable with respect to t in the ordinary sense, but is, by definition, *differentiable* as a parameter at $P_{\theta,\eta}$ if and only if there exists a function $\tilde{\psi}_{\theta,\eta}$ (the efficient influence function) such that

$$a = \frac{\partial}{\partial t}\psi(P_{\theta+ta,\eta_t})\Big|_{t=0} = \langle \tilde{\psi}_{\theta,\eta}, a^\top \dot{\ell}_{\theta,\eta} + g \rangle_{P_{\theta,\eta}}, \qquad \text{for all } a \in \mathbb{R}^k, g \in {}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}.$$

Setting $a = 0$, we see that $\tilde{\psi}_{\theta,\eta}$ must be orthogonal to the nuisance tangent set ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$.

Definition 5.17 (Efficient score function)**.** Define $\Pi_{\theta,\eta}$ as the *orthogonal projection* onto the closure of the linear span of ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$. For a vector $w \in L_2(P_{\theta,\eta})^k$, we define $\Pi_{\theta,\eta}w$ as the vector of the projections of each coordinate.

The **efficient score function** for θ is

$$\tilde{\ell}_{\theta,\eta} := \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}.$$

The **efficient information matrix** for θ is $\tilde{I}_{\theta,\eta} := P[\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^\top]$.

Lemma 5.18. Suppose that for every $a \in \mathbb{R}^k$ and every $g \in {}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ there exists a path

$t \mapsto \eta_t$ in H such that

$$\int \left[\frac{\sqrt{dP_{\theta+ta,\eta_t}} - \sqrt{dP_{\theta,\eta}}}{t} - \frac{1}{2}(a^\top \dot{\ell}_{\theta,\eta} + g)\sqrt{dP}_{\theta,\eta} \right]^2 \to 0, \qquad \text{as } t \to 0. \tag{45}$$

If $\tilde{I}_{\theta,\eta}$ is nonsingular, then the functional $\psi(P_{\theta,\eta}) = \theta$ is differentiable at $P_{\theta,\eta}$ relative to the tangent set $\dot{\mathcal{P}}_{P_{\theta,\eta}} = \operatorname{lin} \dot{\ell}_{\theta,\eta} + {}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ with efficient influence function $\tilde{\psi}_{\theta,\eta} = \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}$.

Proof. The given set $\dot{\mathcal{P}}_{P_{\theta,\eta}}$ is a tangent set by assumption. The vector $\tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}$ has coordinates within $\overline{\operatorname{lin}}\, \dot{\mathcal{P}}_{P_{\theta,\eta}}$. The function ψ is differentiable with respect to this tangent set since

$$\langle \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}, a^\top \dot{\ell}_{\theta,\eta} + g \rangle_{P_{\theta,\eta}} = \tilde{I}_{\theta,\eta}^{-1}\langle \tilde{\ell}_{\theta,\eta}, \dot{\ell}_{\theta,\eta}^\top \rangle_{P_{\theta,\eta}} a = a.$$

The last equality follows, because the inner product of a function and its orthogonal projection is equal to the square length of the projection. Thus, we may replace $\dot{\ell}_{\theta,\eta}$ by $\tilde{\ell}_{\theta,\eta}$. □

Remark 5.8. Consequently, an estimator sequence T_n is *asymptotically efficient* for estimating θ if

$$\sqrt{n}(T_n - \psi(P)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_{\theta,\eta}^{-1}\tilde{\ell}_{\theta,\eta}(X_i) + o_{P_{\theta,\eta}}(1).$$

This is very similar to the situation for efficient estimators in parametric models. The only difference is that the ordinary score function $\dot{\ell}_{\theta,\eta}$ is replaced by the efficient score function (and similarly for the informations). The intuitive explanation is that a part of the score function for θ can also be accounted for by score functions for the nuisance parameter η. When the nuisance parameter is unknown, a part of the information for θ is "lost", and this corresponds to a "loss" of a part of the score function.

5.3.1 Efficient information as minimal information along paths

Let us consider, assuming (45), a path $P_{\theta+ta,\eta_t}$ at $P_{\theta,\eta}$ with score function $\mathcal{L}_{a,g} := a^\top \dot{\ell}_{\theta,\eta} + g$. The information $\mathcal{I}_{a,g}$ along this path is $P_{\theta,\eta}\mathcal{L}_{a,g}^2$. Indeed, it is the quadratic term appearing in the LAN expansion (42), which can be interpreted as the 'curvature' of the model along this path. The following lemma is close in spirit to Lemma 5.10.

Lemma 5.19. Under (45) with scores $a^\top \ell_{\theta,\eta} + g$, suppose that the tangent set of $\mathcal{P}$ at $P_{\theta,\eta}$ is linear. Then for any $a \in \mathbb{R}^k$,

$$\inf_{g \in {}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}} \mathcal{I}_{a,g} = a^\top P_{\theta,\eta}[\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^\top]a = a^\top \tilde{I}_{\theta,\eta} a. \tag{46}$$

Moreover, if the infimum in the last display is attained for $a^\top \tilde{\Gamma}_{\theta,\eta}$ (for all $a \in \mathbb{R}^k$), for some vector $\tilde{\Gamma}_{\theta,\eta}$ with coordinates in ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$, then it holds $\tilde{\Gamma}_{\theta,\eta} = -\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$ ($P_{\theta,\eta}$-almost everywhere), i.e.,

$$\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} + \tilde{\Gamma}_{\theta,\eta}.$$

Definition 5.20 (Least favorable direction)**.** We call $\tilde{\Gamma}_{\theta,\eta}$ the **least favorable direction** and a path with score $a^\top\dot{\ell}_{\theta,\eta}+a^\top\tilde{\Gamma}_{\theta,\eta} = a^\top\tilde{\ell}_{\theta,\eta}$ (when it exists) is then called **least favorable submodel** along the vector a.

Proof. Let us express the information $\mathcal{I}_{a,g}$ with the help of $\tilde{\ell}_{\theta,\eta}$. For any $a \in \mathbb{R}^k$,

$$\begin{aligned} P_{\theta,\eta}\mathcal{L}_{a,g}^2 &= \mathbb{E}\left[a^\top(\dot{\ell}_{\theta,\eta} - \tilde{\ell}_{\theta,\eta} + \tilde{\ell}_{\theta,\eta}) + g\right]^2 \\ &= a^\top\mathbb{E}\left[\tilde{\ell}_{\theta,\eta}\tilde{\ell}_{\theta,\eta}^\top\right]a + 2\mathbb{E}\left[\{a^\top(\dot{\ell}_{\theta,\eta} - \tilde{\ell}_{\theta,\eta}) + g\}\tilde{\ell}_{\theta,\eta}^\top\right]a + \mathbb{E}\left[a^\top(\dot{\ell}_{\theta,\eta} - \tilde{\ell}_{\theta,\eta}) + g\right]^2 \end{aligned}$$

By definition, the efficient score is orthogonal to the nuisance score space so the cross-product term is zero. Also, since the tangent set is linear by assumption, there exists a sequence of scores $a^\top\dot{\ell}_{\theta,\eta} + g_n$ which converges in $L_2(P_{\theta,\eta})$ to $a^\top\tilde{\ell}_{\theta,\eta}$. Along this sequence, the last term in the above display tends to 0. Thus for any fixed a, (46) follows.

Moreover, the previous reasoning also shows that if the infimum in (46) can be written $a^\top\tilde{\Gamma}_{\theta,\eta}$, for some vector $\tilde{\Gamma}_{\theta,\eta}$ with coordinates in ${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}}$ then it holds ($P_{\theta,\eta}$-almost everywhere) $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} + \tilde{\Gamma}_{\theta,\eta}$.

Indeed, the third term in the display giving $\mathcal{I}_{a,g}$ is nonnegative and must be zero if $g = a^\top\tilde{\Gamma}_{\theta,\eta}$ by definition of the infimum. □

5.3.2 Efficient scores: Examples

Example 5.21 (Symmetric location)**.** Suppose that the model consists of all densities $x \to \eta(x-\theta)$ with $\theta \in \mathbb{R}$ and the "shape" η symmetric about 0 with finite Fisher information for location $I_\theta := \int \dot{\eta}^2/\eta$. Thus, the observations are sampled from a density that is symmetric about θ.

Suppose that 'true' η is also C^1 and positive on $\mathbb{R}$. Below we show that here $\dot{\ell}_{\theta,\eta}(x) = -(\eta'/\eta)(x-\theta)$ and that a tangent set in the nuisance part η is the (closed) linear space (Exercise: Show this)

$${}_\eta\dot{\mathcal{P}}_{P_{\theta,\eta}} = \left\{g(\cdot-\theta) : g(-x) = g(x), \int g^2\eta < \infty, \int g\eta = 0\right\}.$$

Any function in this space is symmetric about θ. On the other hand η' is the derivative of an even function so is odd. For any g in the tangent for nuisances,

$$\langle \eta'/\eta(\cdot-\theta), g(\cdot-\theta)\rangle_{L_2(P)} = \int \eta'(x-\theta)g(x-\theta)dx = 0.$$

Thus, scores for θ and for η are orthogonal. Therefore, $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta}$ and $\tilde{I}_{\theta,\eta} = I_\theta = \int \dot{\eta}^2/\eta$ and there is no loss of information! This fact was discovered by Charles Stein in 1956 and came a bit as a surprise. Finding the center of symmetry of a completely unknown symmetric density seems at first sight more complicated than if the shape is known. If one is able to find a semiparametric estimator that achieves the bound, which we will do later in this course, we will have shown that there is indeed asymptotically no loss of information.

Obtaining tangent sets:

Approach 1: We construct a tangent set by identifying scores with derivatives of the log-likelihood. Let g a symmetric, C^1 function with g, g' bounded over $\mathbb{R}$. Consider the paths

$$t \mapsto (\theta + ta, \eta_t) \quad \text{where} \quad \eta_t = \eta(1+tg).$$

Note that this is a well-defined path through the model for t small enough. Indeed, the functions η_t are symmetric, nonnegative for t small enough because g is bounded, and integrate to 1. Then,

$$\frac{\partial}{\partial t}\log p_{\theta_t,\eta_t}(x)\Big|_{t=0} = g(x-\theta) + a\left[-\frac{\eta'(x-\theta)}{\eta(x-\theta)}\right].$$

One can then check that Lemma 5.7 applies. With these paths we obtain the following tangent set

$$\dot{\mathcal{P}}_P = \text{lin}\left[-\frac{\eta'(x-\theta)}{\eta(x-\theta)}\right] + \left\{g(\cdot-\theta) : g(-x) = g(x), g, g' \text{ bounded on } \mathbb{R}, \int g\eta = 0\right\}.$$

Approach 2: We use exactly the same paths but check instead directly differentiability in quadratic mean of the model using (40). In particular show that the tangent set can be taken equal to (Exercise)

$$\dot{\mathcal{P}}_P^{(2)} = \text{lin}\left[-\frac{\eta'(x-\theta)}{\eta(x-\theta)}\right] + \left\{g(\cdot-\theta) : g(-x) = g(x), \int g\eta = 0, \int g^2\eta < \infty\right\}.$$

Notice that in both cases the scores for the nuisance part form a linear space with the same

closure in $L_2(P)$, so considering one or the other tangent set does not matter for efficient influence functions.

Example 5.22 (Restricted moment model regression)**.**

Example 5.23 (Cox proportional hazards model)**.** In the Cox model, *without censoring*, a typical observation is a pair $X = (T, Z)$ of a "survival time" $T \geq 0$ and a covariate $Z \in \mathbb{R}^k$. It is best described in terms of the conditional hazard function of T given Z. Recall that the hazard function λ corresponding to a probability density f is the function

$$\lambda(t) = \frac{f(t)}{1 - F(t)}, \qquad \text{for } \ t \geq 0,$$

where F is the distribution function corresponding to f. Simple algebra shows that, for $t \geq 0$,

$$\bar{F}(t) := 1 - F(t) = e^{-\Lambda(t)} \quad \text{and hence} \quad f(t) = \lambda(t) e^{-\Lambda(t)}$$

(here $\Lambda(t) = \int_0^t \lambda(u) du$ is the cumulative hazard function), so that the relationship between f and Λ is one-to-one.

In the Cox model the distribution of Z is arbitrary and the conditional hazard function of T given Z is postulated to be of the form

$$\lambda_{T|Z}(t|z) = e^{\theta^\top z} \lambda(t), \qquad \text{for } \ t \geq 0, \ z \in \mathbb{R}^k,$$

for $\theta \in \mathbb{R}^k$ and $\lambda(\cdot)$ being a completely unknown hazard function. The parameter θ has an interesting interpretation in terms of a ratio of hazards. For instance, if the i'th coordinate Z_i of the covariate is a $0-1$ variable then e^{θ_i} is the ratio of the hazards of two individuals whose covariates are $Z_i = 1$ and $Z_i = 0$, respectively, and whose covariates are identical otherwise. This is another reason for the popularity of the model: the model gives a better fit to data than a parametric model (obtained for instance by assuming that the baseline hazard function is of Weibull form), but its parameters are still easy to interpret. A third reason for its popularity is that statistical procedures for estimating the parameters take a simple form.

The density of an observation in the Cox model takes the form

$$(t, z) \mapsto e^{-e^{\theta^\top z} \Lambda(t)} \lambda(t) e^{\theta^\top z} p_Z(z).$$

Differentiating the logarithm of this expression with respect to θ gives the score function

for θ, with $x = (t, z)$,

$$\dot{\ell}_{\theta,\lambda} = z - ze^{\theta^\top z}\Lambda(t).$$

We can also insert appropriate parametric models $s \mapsto \lambda_s$ and differentiate with respect to s. If b is the derivative of $\log \lambda_s$ at $s = 0$, then the corresponding score for the model for the observation is $B_{\theta,\lambda}b$ where

$$[B_{\theta,\lambda}b](t, z) = b(t) - e^{\theta^\top z} \int_0^t b(u)d\Lambda(u).$$

References

Begun, J. M., Hall, W. J., Huang, W.-M., and Wellner, J. A. (1983). Information and asymptotic efficiency in parametric–nonparametric models. *Ann. Statist.*, 11(2):432–452.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD.

Bolthausen, E., Perkins, E., and van der Vaart, A. (2002). *Lectures on probability theory and statistics*, volume 1781 of *Lecture Notes in Mathematics*. Springer-Verlag, Berlin. Lectures from the 29th Summer School on Probability Theory held in Saint-Flour, July 8–24, 1999, Edited by Pierre Bernard.

Koshevnik, J. A. and Levit, B. J. (1976). On a nonparametric analogue of the information matrix. *Teor. Verojatnost. i Primenen.*, 21(4):759–774.

Le Cam, L. (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics*, 3:37–98.

Lehmann, E. L. and Casella, G. (1998). *Theory of point estimation*. Springer Texts in Statistics. Springer-Verlag, New York, second edition.

Pfanzagl, J. (1982). *Contributions to a general asymptotic statistical theory*, volume 13. Springer Science & Business Media.

Pollard, D. (1989). Asymptotics via empirical processes. *Statist. Sci.*, 4(4):341–366. With comments and a rejoinder by the author.

Stein, C. (1956). Efficient nonparametric testing and estimation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I*, pages 187–195. University of California Press, Berkeley and Los Angeles.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Invent. Math.*, 126(3):505–563.

Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer Series in Statistics. Springer, New York.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics.