Notes on *Semiparametric Models*

Chuhan Xie

June 18, 2022

Abstract

This is a lecture note on *Semiparametric Models* taught by Wang Miao.

Contents

 Email: ch_xie@pku.edu.cn.

Reference texts

1. Van der Vaart. (2000) Asymptotic Statistics, Chapter 25.

2. Tsiatis, A. (2006) Semiparametric Theory and Missing Data.
   (The first half of the book gives an accessible non-super technical introduction to semiparametric theory.)

3. Newey, W. (1990) Semiparametric Efficiency Bounds. Journal of Applied Econometrics, vol 5, 99-135.
   (This is a GREAT introductory paper on semiparametric theory.)

4. Van der Vaart. (2002) Semiparametric Statistics in Lectures on Probability and Statistics.
   (This is a monograph with material expanded a bit more than Chapter 25 of the asymptotic statistics book.)

5. Bickel, Klaassen, Ritov and Wellner. (1993) Efficient and Adaptive Inference in Semiparametric Models.
   (This book provides a rigorous treatment of semiparametric theory but it is hard to read.)

Abstract

1. Introduction to semiparametric and parametric models. Brief review on statistical inference, probability, and functional analysis.

2. Efficiency theory in parametric models.
   The Cramer-Rao bound, Regular parametric models, Regular estimators, Hajek representation theorem, Asymptotically linear estimators, Characterization of the influence functions of regular asymptotically linear estimators, The efficient influence function, The efficient score.

3. Efficiency theory in semiparametric models.
   Regular parametric submodels, Regular parameters, The semiparametric variance bound, Pathwise differentiable parameters, Gradients, The tangent space, The efficient influence function, Asymptotic efficiency in semiparametric models, The semiparametric efficient score, Convex models, Double-robust estimation, One-step estimation, Semiparametric maximum profile likelihood estimation.

4. Examples.
   (a) Estimands in the nonparametric model. (b) Parameters of models defined by conditional moment restrictions. (c) Parameters of partially parametric regression model. (d) Regression parameter in the proportional odds model and in the location shift model. (e) Parameters of missing data and causal inference models.

1 Introduction to Semiparametric Models

Modern epidemiological and clinical studies routinely collect, on each of n subjects, **high dimensional data** (often comprised of many baselines and time dependent variables). However, the scientific interest is often on a **low dimensional functional** $\beta(F)$ of the data distribution F, with little or no knowledge about F. The methods in this course will give valid inferences under non- or semi-parametric models that make minimal assumptions about the parts of the law of the data that are not of scientific interest. As such, they are protected from misspecification of models for these secondary parts of the law of the data.

For parametric models, data are drawn from a distribution F which is assumed to belong to the family $\mathcal{F} = \{F_\theta \colon \theta \in \Theta\}$, where $\Theta \in \mathbb{R}^p$ for some $p \in \mathbb{N}$. For semiparametric models, Θ can be a "massive" set, i.e., it is NOT a subset of any Euclidean space. We aim at estimating the value of $\beta(F)$ for some function $\beta \colon \mathcal{F} \to \mathbb{R}^k$. Oftentimes we use the notation $\beta(\theta) = \beta(F_\theta)$.

Remark 1.1. We always assume the distributions in $\mathcal{F}$ are dominated by a measure μ. We use f to denote the density of F w.r.t. μ, and use $\mathrm{d}x$ to denote $\mathrm{d}\mu(x)$.

1.1 Examples

Example 1.1 (Nonparametric model)**.** $\mathcal{F}$ is the collection of ALL probability distributions (on a given sample space). Here, Θ is the collection of all probability distributions, and each $\theta \in \Theta$ is a probability distribution. We might be interested in estimating its mean, i.e.,

$$\beta(F) = \int z f(z) \mathrm{d}z.$$

Example 1.2 (Symmetric distributions)**.** Suppose Z is a scalar continuous r.v. with distribution F assumed to have a density f satisfying

$$f(z) = g(z - \beta^*),$$

for some unknown $\beta^* \in \mathbb{R}$ and some unknown function $g(u)$ which is symmetric around 0, i.e.,

$$g(u) = g(-u) \text{ for all } u.$$

Then $\mathcal{F}$ can be indexed by $\theta = (\beta, g) \in \mathbb{R} \times \mathcal{G}$ where $\mathcal{G}$ is the set of positive real valued functions that are symmetric around 0 and integrate to 1. We might be interested in the center of location

$$\beta(F) = \beta^* \text{ such that } f(z) = g(z - \beta^*),$$

or the variance

$$\beta(F) = \mathrm{var}(Z) = \int z^2 g(z) \mathrm{d}z.$$

Example 1.3 (Conditional mean model)**.** $Z = (Y, X)$, Y is a response (assumed to be continuous), X is a vector of covariates. The model assumes that

$$\mathbb{E}(Y \mid X) = g(X; \beta^*),$$

where $g(X; \beta)$ is a known function of X and β, e.g., $g(X; \beta) = \exp(X^\top \beta)$, and $\beta^* \in \mathbb{R}^k$ is unknown. Other popular example is **linear model** and **logistic regression model**. No other assumptions are made.

Define $\varepsilon = Y - g(X; \beta^*)$. Then β and the joint distribution of (ε, X) determine the joint distribution of (Y, X). Specifically, an arbitrary distribution in the model has the density

$$\begin{aligned} f_{Y,X}(y, x; \beta, \eta_1, \eta_2) &= f_{Y|X}(y \mid x; \beta, \eta_1) f_X(x; \eta_2) \\ &= f_{\varepsilon|X}(y - g(x; \beta) \mid x; \eta_1)\eta_2(x) \\ &= \eta_1(y - g(x; \beta), x)\eta_2(x), \end{aligned}$$

where $\eta_1(\varepsilon, x)$ is the conditional density of ε given $X = x$ and $\eta_2(x)$ is the density of X. They are non-negative functions restricted only by

$$\int \eta_1(\varepsilon, x)\mathrm{d}\varepsilon = 1 \text{ and } \int \varepsilon\eta_1(\varepsilon, x)\mathrm{d}\varepsilon = 0 \text{ for all } x; \quad \int \eta_2(x)\mathrm{d}x = 1.$$

In this model $\mathcal{F}$ can be indexed by $\theta = (\beta, \eta_1, \eta_2) \in \mathbb{R}^k \times \boldsymbol{\eta_1} \times \boldsymbol{\eta_2}$.

Remark 1.2. In a parametric linear model, we often assume

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2), \quad X \sim P.$$

In its semiparametric counterpart, we only make minimal assumptions

$$Y = \beta_0 + \beta_1 X + \varepsilon, \quad \mathbb{E}(\varepsilon \mid X) = 0.$$

This avoids misspecification when ε is not Gaussian.

Example 1.4 (Cox proportional hazards model)**.** $Z = (T, X)$, T is the time to an event, X is a vector of covariates. This model only assumes that

$$\lambda(t \mid X) = \lambda_0(t) \exp(\beta^\top X),$$

where $\lambda(t \mid X)$ is the conditional hazard at time t, i.e.,

$$\lambda(t \mid X) = \lim_{h \to 0} \frac{1}{h} \mathbb{P}(t \le T < t + h \mid T \ge t, X),$$

$\lambda_0(t)$ is an unknown baseline intensity that varies over t, and the $\exp(\beta^\top X)$ term is the contribution of the covariates. If T is continuous, then $\lambda(t \mid X) = \frac{f(t|X)}{1-F(t|X)} = \frac{f(t|X)}{S(t|X)}$ can be interpreted as the "instantaneous probability" of experiencing an event at time t given you have not experienced an event before t.

An arbitrary distribution in this model has density

$$\begin{aligned} f_{T,X}(t,x;\beta,\lambda_0,\eta) &= f_{T|X}(t \mid x;\beta,\lambda_0) f_X(x;\eta) \\ &= \lambda(t \mid x;\beta,\lambda_0) \exp\left\{-\int_0^t \lambda(u \mid x;\beta,\lambda_0)\mathrm{d}u\right\} f_X(x;\eta) \\ &= \lambda_0(t)\exp(\beta^\top x)\exp\left\{-\int_0^t \lambda_0(u)\exp(\beta^\top x)\mathrm{d}u\right\}\eta(x), \end{aligned}$$

where we have used the fact that

$$S(t \mid X) = \exp\left\{-\int_0^t \lambda(u \mid X)\mathrm{d}u\right\} \text{ and } f(t \mid X) = \lambda(t \mid X)S(t \mid X).$$

In this model $\mathcal{F}$ can be indexed by $\theta = (\beta, \lambda_0, \eta) \in \mathbb{R}^k \times \mathbf{\Lambda} \times \boldsymbol{\eta}$, where $\mathbf{\Lambda}$ is the set of all non-negative functions of t and $\boldsymbol{\eta}$ is the set of all non-negative functions of x that integrates to 1.

Example 1.5 (Partially linear regression model)**.** $Z = (R, V, Y)$, Y is a real valued continuous response variable, R and V are vectors of covariates. The model assumes that

$$\mathbb{E}(Y \mid R, V) = h(V) + \beta^{*\top} R,$$

where $h(V)$ is an unknown function of V and β^* is an unknown vector. $\mathcal{F}$ can be indexed by $\theta = (\beta^*, h, \eta_1, \eta_2)$ where η_1 and η_2 are the distributions of (V, R) and $\varepsilon = Y - h(V) - \beta^{*\top} R$ given (V, R), respectively. Also called semiparametric regression with identity link function.

Example 1.6 (Generalized partially linear regression model)**.** $Z = (R, V, Y)$, Y is either binary or a count, R and V are vectors of covariates. The model assumes that

$$g\{\mathbb{E}(Y \mid R, V)\} = h(V) + \beta^{*\top} R,$$

where g is either a log or logit link function. Also called semiparametric regression with link function g.

Example 1.7 (Single index binary choice model)**.** $Z = (V, Y)$, Y is a binary variable, V are vectors of covariates. The model assumes that

$$\mathbb{E}(Y \mid V) = g(\beta^{*\top} V),$$

where g is an unknown c.d.f., β^* is an unknown vector. This model is like a generalized conditional mean model (Example 1.3) with the link function g being unrestricted. "Single index" means there is only one index $\beta^{*\top} V$. A multiple index version is

$$\mathbb{E}(Y \mid V) = g(\beta_1^\top V, \ldots, \beta_p^\top V).$$

Example 1.8 (Semiparametric additive instrumental variable (IV) model)**.** $Z = (V, R, A, Y)$, Y is a continuous outcome, V are vectors of covariates, R is a treatment variable, and A is an IV. The model assumes that

$$\mathbb{E}(Y - \beta^* R \mid A, V) = \mathbb{E}(Y - \beta^* R \mid V), \tag{1.1}$$

where β^* is an unknown treatment effect on the additive scale. **Identification** requires that $\mathbb{E}(R \mid A = a, V = v)$ depends on a for at least one value of v.

Note that β^* is not identified by standard regression of Y on R, whether or not one conditions on V and A, because the necessary condition for identification by such a "direct regression" approach requires $\mathbb{E}(Y - \beta^* R \mid R, A, V) = \mathbb{E}(Y - \beta^* R \mid A, V)$, which is not implied by (1.1). The treatment effect is said to be **confounded**.

(1.1) also has a more clear form

$$Y = \beta^* R + U + h(V), \quad \mathbb{E}(U \mid A, V) = \mathbb{E}(U \mid V), \tag{1.2}$$

i.e., the confounder U and the instrument A are correlated only through their correlations with the covariate V.

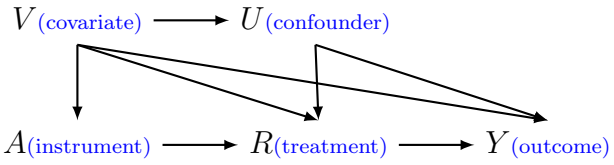

Figure 1.1: DAG for semiparametric additive IV model

Definition 1.1 (**Instrumental variable**)**.** An instrumental variable (IV) A satisfies:

1. A is not directly correlated to the outcome Y;
2. A has no direct causal relations with the confounder U;
3. A is correlated to the treatment R.

For example, "father's education" might be an IV when studying the effect of "education" on "wage".

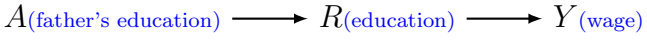

Figure 1.2: IV example

Example 1.9 (Missing data model)**.** Consider the full data restricted mean model

$$\mathbb{E}(Y \mid X) = g(X; \beta^*),$$

where $X = (X_1, X_2)$. We observe i.i.d. samples on (Y, R, RX_1, X_2) where $R = 1$ if X_1 is observed and $R = 0$ is X_1 is missing. This restricted mean model is NOT point identified without placing a restriction on the missing data mechanism.

A common assumption is **missing at random** (MAR), i.e., $R \perp X_1 \mid Y, X_2$. In this case, the semiparametric model on the observed data has likelihood for a single realization

$$\left\{ \int f(\varepsilon(\beta^*) \mid X; \eta_1) f(X; \eta_2) \mathrm{d}X_1 \right\}^{1-R} f(R \mid X_2, Y; \eta_3)$$
$$\times \{ f(\varepsilon(\beta^*) \mid X; \eta_1) f(X; \eta_2) \}^R,$$

which introduces a new nuisance parameter η_3 indexing the missing data mechanism.

1.2 Causal interpretation of two models

Here we discuss two models in the causal lens: semiparametric regression model (Example 1.5) and semiparametric additive IV model (Example 1.8).

Let Y_r denote the potential outcome one would observe if one would observe if one could intervene to set the treatment R to r. In a randomized experiment, randomization ensures such intervention can be done as the treatment is under control of the experimenter. In an observational study, we try to mimic a randomized trial by assuming that within levels of covariates, it is as if nature performed a randomized trial, i.e., there is **no unmeasured confounder** and thus $R \perp (Y_0, Y_1) \mid V$. Then

$$\mathbb{E}(Y_r \mid V) = \mathbb{E}(Y \mid R = r, V) = h(V) + \beta^* r,$$

implies constant average causal effect of the treatment R conditional on V,

$$\mathbb{E}(Y_1 - Y_0 \mid V) = \beta^*. \tag{1.3}$$

Here, **structural model** (1.3) and **no unmeasured confounder** imply **semiparametric regression model**.

Now, suppose one only has access to a randomized instrumental variable A, i.e., instead of $R \perp (Y_0, Y_1) \mid V$, one only has $A \perp (Y_0, Y_1) \mid V$ (one needs A and R to be correlated). Then the structural model is given by

$$\mathbb{E}(Y_r - Y_0 \mid V, A, R = r) = \beta^* r, \tag{1.4}$$

which implies that

$$\mathbb{E}(Y_R - \beta^* R \mid V, A) = \mathbb{E}(Y_0 \mid V, A) = \mathbb{E}(Y_0 \mid V).$$

Here, **structural model** (1.4) and **IV assumption** imply **semiparametric additive IV model**

$$\mathbb{E}(Y - \beta^* R \mid V, A) = \mathbb{E}(Y - \beta^* R \mid V).$$

1.3 Semiparametric estimators

A semiparametric estimator $\widehat{\beta}_n$ of $\beta(F)$ is one which satisfies

$$\sqrt{n}\left\{\widehat{\beta}_n - \beta(F)\right\} \xrightarrow{D(F)} \mathcal{N}(0, \Sigma(F)),$$

for all distributions $F \in \mathcal{F}$.

For example, the solution to the GEE equations

$$\sum_{i=1}^{n} d(X_i; \widehat{\beta}_n)\{Y_i - g(X_i; \widehat{\beta}_n)\} = 0, \tag{1.5}$$

is (under regularity conditions) a semiparametric estimator in the conditional mean model (Example 1.3). We consider linear model and logistic regression model.

For linear model

$$Y = X^\top \beta + \varepsilon, \quad \mathbb{E}(\varepsilon \mid X) = 0,$$

the well-known OLS estimator $\widehat{\beta}_{\text{blue}} = (X^\top X)^{-1} X^\top Y$ is the solution to (1.5) when $g(x;\beta) = x^\top \beta$ and $d(x;\beta) = x$. The GEE estimator is the solution to (1.5) when $g(x;\beta) = x^\top \beta$ and $d(x;\beta) = d(x)$. Which estimator is better depends on the **model assumption**. For example, under the Gauss-Markov condition, $\widehat{\beta}_{\text{blue}}$ is optimal; and if $\text{cov}(\varepsilon) = \Sigma$, then GLS is preferred.

For logistic regression model

$$\mathbb{P}(Y = 1 \mid X) = \frac{\exp(\beta_0 + \beta_1^\top X)}{1 + \exp(\beta_0 + \beta_1^\top X)} = \pi(X; \beta),$$

the score equation (based on MLE) is

$$\sum_{i=1}^{n} (\pi(X_i; \beta) - Y_i) X_i = 0,$$

which is of the form (1.5). Another estimator is the solution to

$$\sum_{i=1}^{n} \left\{1 - \frac{Y_i}{\pi(X_i; \beta)}\right\} X_i = 0.$$

The former estimator has a smaller variance, while the latter one is more robust to perturbations.

1.4 Questions of interest

1. For which semiparametric $\beta(F)$ can we hope to find semiparametric estimators?
2. When they exist, how do we construct them?
3. How do we define an analogous to the Cramer-Rao variance bound?

4. For which functionals are there "optimal" semiparametric estimators whose variance is the smallest possible, i.e., that they achieve the semiparametric C-R bound?

5. If an "optimal" estimator exists, how do we construct it?

2 Basic Notions of Hilbert Spaces

2.1 Hilbert spaces

A Hilbert space is a complete linear inner product space. Any Hilbert space V is closed; and any of its finite dimensional subspace is closed. We are interested in the spaces

$$\mathcal{L}_2(\theta) = \left\{ b(\cdot) \text{ real valued: } \int b(x)^2 f(x;\theta)\mathrm{d}x < \infty \right\},$$

and

$$\mathcal{L}_2^0(\theta) = \left\{ b(\cdot) \text{ real valued: } \int b(x)^2 f(x;\theta)\mathrm{d}x < \infty, \int b(x) f(x;\theta)\mathrm{d}x = 0 \right\}.$$

The spaces $\mathcal{L}_2(\theta)$ and $\mathcal{L}_2^0(\theta)$ are Hilbert spaces with inner product given by

$$\langle b_1(X), b_2(X)\rangle_\theta = \mathbb{E}_\theta\{b_1(X)b_2(X)\}.$$

Note that in $\mathcal{L}_2^0(\theta)$, $\|b(X)\|_\theta^2 = \mathrm{var}_\theta\{b(X)\}$.

Theorem 2.1 (Pythagorean theorem)**.** *Let $v_1, \ldots, v_k$ be mutually orthogonal vectors in V, then*

$$\left\| \sum_{i=1}^k v_i \right\|^2 = \sum_{i=1}^k \|v_i\|^2.$$

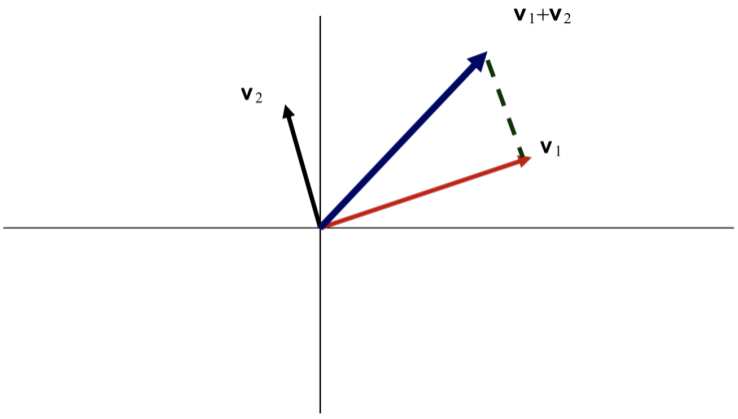

Figure 2.1: 2-dimensional case

Lemma 2.1 (Projection lemma 1)**.** *Let V be a Hilbert space and M be a closed linear subspace. Then*

1. *corresponding to any vector $x \in V$, there exists a unique vector $m_0 \in M$ such that*

$$\|x - m_0\| \leq \|x - m\| \text{ for all } m \in M; \tag{2.1}$$

2. m_0 *satisfies* (2.1) *iff* $x - m_0 \perp m$ *for all* $m \in M$.

Lemma 2.2 (Projection lemma 2)**.** *Let* V *be a inner product space,* M *be a subspace, and* x *be a vector in* V*. Then*

1. *if there exists* $m_0 \in M$ *such that*

$$\|x - m_0\| \leq \|x - m\| \text{ for all } m \in M, \tag{2.2}$$

then m_0 *is unique;*

2. m_0 *satisfies* (2.2) *iff* $x - m_0 \perp m$ *for all* $m \in M$.

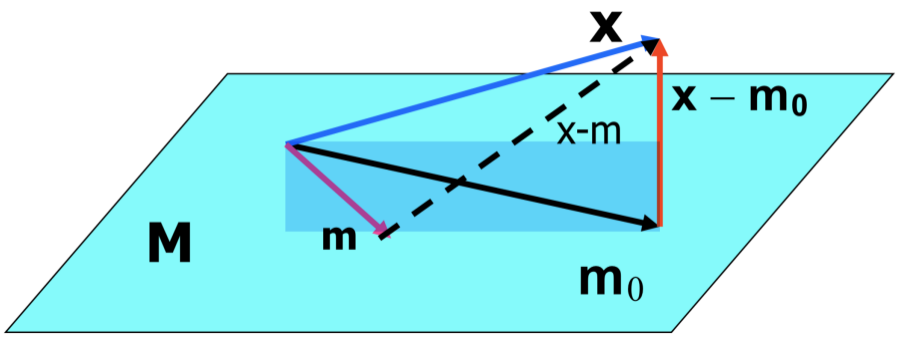

Figure 2.2: Projection lemma

Remark 2.1. We denote $\overline{B}$ as the closure of set B and $[B]$ as the linear span of B. For a subspace M, we denote the projection of $x \in V$ onto M as $\Pi[x \mid M]$ and its orthogonal space as $M^{\perp}$.

Proposition 2.1. *The following two conclusions hold.*

1. *Suppose* $\{M_a\}_{a \in \mathcal{A}}$ *is a collection of closed linear spaces. Let*

$$M = \overline{\left[\bigcup_{a \in \mathcal{A}} M_a\right]}.$$

If $\Pi[v \mid M_a] = 0$ *for all* $a \in \mathcal{A}$*, then* $\Pi[v \mid M] = 0$.

2. *If* M_1 *and* M_2 *are orthogonal closed linear spaces, then*

$$\Pi[v \mid M_1 \oplus M_2] = \Pi[v \mid M_1] + \Pi[v \mid M_2].$$

Remark 2.2. Proposition 2.1 is crucial, because (the tangent space of) a semiparametric model can be regarded as the union of (the tangent spaces of) its parametric submodels.

2.2 Calculation of the projection

Suppose M is a subspace of dimension p spanned by $v_1, \ldots, v_p$. The explicit form of $\Pi[x \mid M]$ is

$$\Pi[x \mid M] = \sum_{i=1}^{p} a_i v_i,$$

where $a_1, \ldots, a_p$ satisfy the **normal equation**

$$\underbrace{\begin{bmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \cdots & \langle v_1, v_p \rangle \\ \langle v_2, v_1 \rangle & \langle v_2, v_2 \rangle & \cdots & \langle v_2, v_p \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_p, v_1 \rangle & \langle v_p, v_2 \rangle & \cdots & \langle v_p, v_p \rangle \end{bmatrix}}_{\text{Gram matrix}} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_p \end{bmatrix} = \begin{bmatrix} \langle x, v_1 \rangle \\ \langle x, v_2 \rangle \\ \vdots \\ \langle x, v_p \rangle \end{bmatrix}.$$

In the statistical setting, V will be $\mathcal{L}_2^0(\theta)$ and $v_1, \ldots, v_p$ will be the **scores**, i.e.,

$$v_j = S_{\theta_j}(\theta), \quad S_{\theta_j}(\theta) = \frac{\partial \ln f(X;\theta)}{\partial \theta_j} \text{ for } j = 1, \ldots, p.$$

These vectors span the **tangent space $\Lambda(\theta)$ for the model at F_θ**, i.e.,

$$\Lambda(\theta) = \{a^\top S_\theta(\theta) \colon a \in \mathbb{R}^p\}.$$

The Gram matrix is the **Fisher information (FI) matrix** $I(\theta) = \mathbb{E}_\theta[S_\theta(\theta) S_\theta(\theta)^\top]$. Therefore, the projection of $b(X) \in \mathcal{L}_2^0(\theta)$ into $\Lambda(\theta)$ is

$$\Pi_\theta[b(X) \mid \Lambda(\theta)] = \mathbb{E}_\theta[b(X) S_\theta(\theta)^\top] I(\theta)^{-1} S_\theta(\theta).$$

Since a projection is contracting, we have

$$\begin{aligned} \operatorname{var}_\theta\{b(X)\} &= \|b(X)\|_\theta \\ &\geq \|\Pi_\theta[b(X) \mid \Lambda(\theta)]\|_\theta \\ &= \operatorname{var}_\theta \left\{\mathbb{E}_\theta[b(X) S_\theta(\theta)^\top] I(\theta)^{-1} S_\theta(\theta)\right\} \\ &= \mathbb{E}_\theta[b(X) S_\theta(\theta)^\top] I(\theta)^{-1} \mathbb{E}_\theta[S_\theta(\theta) b(X)], \end{aligned}$$

which provides a lower bound on the variance of an unbiased estimator $\widehat{\beta}$ by setting $b(X) = \widehat{\beta} - \beta(\theta)$.

Similar results hold in the multivariate case. Suppose $b_1(X), \ldots, b_k(X) \in \mathcal{L}_2^0(\theta)$. Define $\underline{b}(X) = (b_1(X), \ldots, b_k(X))^\top$ and

$$\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)] = \begin{bmatrix} \Pi_\theta[b_1(X) \mid \Lambda(\theta)] \\ \Pi_\theta[b_2(X) \mid \Lambda(\theta)] \\ \vdots \\ \Pi_\theta[b_k(X) \mid \Lambda(\theta)] \end{bmatrix}.$$

It can be shown that

$$\mathbb{E}\left\{(\underline{b}(X) - \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)])\,\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]^\top\right\} = \underbrace{0}_{k\times k},$$

and thus

$$\begin{aligned}
\underbrace{\mathrm{var}_\theta(\underline{b}(X))}_{k\times k} &= \mathrm{var}_\theta\left\{(\underline{b}(X) - \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]) + \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]\right\} \\
&= \mathrm{var}_\theta\left\{(\underline{b}(X) - \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)])\right\} + \mathrm{var}_\theta\left\{\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]\right\} \\
&\quad + \mathbb{E}_\theta\left\{(\underline{b}(X) - \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)])\,\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]^\top\right\} \\
&\quad + \mathbb{E}_\theta\left\{\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]\,(\underline{b}(X) - \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)])^\top\right\} \\
&= \underbrace{\mathrm{var}_\theta\left\{(\underline{b}(X) - \Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)])\right\}}_{\text{variance of the orthogonal component}} + \underbrace{\mathrm{var}_\theta\left\{\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)]\right\}}_{\text{variance of the projection}}.
\end{aligned}$$

Remembering that $\Pi_\theta[\underline{b}(X) \mid \Lambda(\theta)] = \mathbb{E}_\theta\left[\underline{b}(X)S_\theta(\theta)^\top\right] I(\theta)^{-1}S_\theta(\theta)$, we have

$$\begin{aligned}
\mathrm{var}_\theta\{\underline{b}(X)\} &\geq \mathrm{var}_\theta\left\{\mathbb{E}_\theta[\underline{b}(X)S_\theta(\theta)^\top]I(\theta)^{-1}S_\theta(\theta)\right\} \\
&= \mathbb{E}_\theta[\underline{b}(X)S_\theta(\theta)^\top]I(\theta)^{-1}\mathbb{E}_\theta[S_\theta(\theta)\underline{b}(X)^\top].
\end{aligned}$$

Here $A \geq B$ means $A - B$ is p.s.d.

3 Efficiency in Parametric Models

3.1 Motivating example: unbiased estimators

Consider i.i.d. $X_1, \ldots, X_n$ and let $\underline{X} = (X_1, \ldots, X_n)$. Let $S_\theta^{(n)}(\theta)$ and $\Lambda^{(n)}(\theta)$ be the score for θ and the tangent space at θ. Suppose $\widehat{\beta} = \widehat{\beta}(\underline{X})$ is an unbiased estimator of an $\mathbb{R}^k$-valued parameter $\beta(\theta)$. In Section 2.2 we have shown

$$\begin{aligned}\mathrm{var}_\theta\{\widehat{\beta}(\underline{X})\} &\geq \mathbb{E}_\theta\left\{\widehat{\beta}(\underline{X})S_\theta^{(n)}(\theta)^\top\right\}\mathrm{var}_\theta\left\{S_\theta^{(n)}(\theta)\right\}^{-1}\mathbb{E}_\theta\left\{S_\theta^{(n)}(\theta)\widehat{\beta}(\underline{X})^\top\right\}\\ &= n^{-1}\mathbb{E}_\theta\left\{\widehat{\beta}(\underline{X})S_\theta^{(n)}(\theta)^\top\right\}I(\theta)^{-1}\mathbb{E}_\theta\left\{S_\theta^{(n)}(\theta)\widehat{\beta}(\underline{X})^\top\right\},\end{aligned}$$

by setting $\underline{b}(\underline{X}) = \widehat{\beta}(\underline{X}) - \beta(\theta)$.

What is special for unbiased estimator is the fact that the projections of ALL unbiased estimators into the tangent space coincide. See Figure 3.1 for illustration.

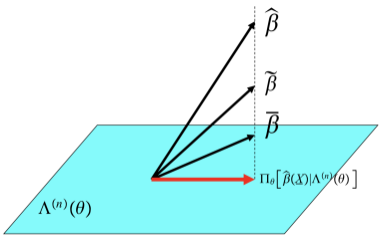

Figure 3.1: Projection of unbiased estimators

This property comes from the fact that

$$\frac{\partial\beta(\theta)}{\partial\theta^\top} = \mathbb{E}_\theta\left\{\widehat{\beta}(\underline{X})S_\theta^{(n)}(\theta)^\top\right\}. \tag{3.1}$$

If (3.1) holds, then the projection of any unbiased estimator is

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

where the last formula does not depend on the estimator $\widehat{\beta}(\underline{X})$.

The **Cramer-Rao bound** is defined as

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

In the unbiased estimator case, $\mathrm{var}_\theta\{\widehat{\beta}(\underline{X})\} \geq n^{-1}C_{\mathcal{F}}(\theta)$.

Informal proof of (3.1). Since for all θ, unbiasedness implies

$$\beta(\theta) = \int \widehat{\beta}(\underline{x}) f(\underline{x}; \theta) \mathrm{d}\underline{x}.$$

Assume exchange of differentiation and integration is possible. Then

$$\begin{aligned}
\frac{\partial \beta(\theta)}{\partial \theta^\top} &= \int \widehat{\beta}(\underline{x}) \frac{f(\underline{x}; \theta)}{\partial \theta^\top} \mathrm{d}\underline{x} \\
&= \int \widehat{\beta}(\underline{x}) S_\theta^{(n)}(\theta)^\top f(\underline{x}; \theta) \mathrm{d}\underline{x} \\
&= \mathbb{E}_\theta \left\{ \widehat{\beta}(\underline{X}) S_\theta^{(n)}(\theta)^\top \right\}.
\end{aligned}$$

□

Remark 3.1. Informal proof of (3.1) requires that

1. the derivative of $f(x; \theta)$ w.r.t. θ exists for all x;
2. exchange of differentiation and integration is possible.

The same result will be established under much weaker assumptions in the following sections.

3.2 Mean squared differentiability

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

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

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

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

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

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

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

Remark 3.2. Typically, whenever

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

as $h \to 0$, it holds that for μ-a.s. x,

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

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

In a lemma below we will state that mean squared differentiability holds under first order differentiability of $\sqrt{f(x;\theta)}$ as a function of θ for each x and continuity of the information matrix. However, mean squared differentiability can hold even without pointwise differentiability as the following example shows.

Example 3.1. Consider the double exponential density

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

Then for each x, the map $\theta \to \sqrt{f(x;\theta)}$ is not differentiable at $\theta = x$, yet the map $\theta \to \sqrt{f(\cdot;\theta)}$ is mean square differentiable at θ^* with $s_\theta(x;\theta^*) = \mathrm{sgn}(x-\theta^*)$.

In the above example, for each x, the double exponential density $f(x;\theta) = \frac{1}{2}\exp(-|x-\theta|)$ is infinitely differentiable with respect to θ at θ^* except for $x = \theta^*$. One may think that differentiability with respect to θ of $f(x;\theta)$ for μ-a.s. x, implies mean squared differentiability of $\theta \in \Theta \to \sqrt{f(\cdot;\theta)}$. This is however incorrect. The following provides a counterexample.

Example 3.2. Consider the uniform density

$$f(x;\theta) = \theta^{-1}\mathbb{I}_{(0,\theta)}(x), \quad \theta > 0.$$

$f(x;\theta)$ is differentiable with respect to θ for μ-a.s. x, but the map $\theta \to \sqrt{f(\cdot;\theta)}$ is NOT mean squared differentiable at any given θ^*.

Lemma 3.1. [1] *Suppose at each fixed x the map $\theta \to \sqrt{f(\cdot;\theta)}$ is continuously differentiable. Furthermore, suppose the elements of the matrix*

$$I(\theta) = \int \frac{\partial \log f(x;\theta)}{\partial \theta}\frac{\partial \log f(x;\theta)}{\partial \theta^\top} f(x;\theta)\mathrm{d}x$$

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

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

In other words, $\sqrt{f(x;\theta)}$ continuously differentiable (w.r.t. θ) + FI matrix continuous = mean squared differentiable.

Remark 3.3. The continuity of the FI matrix can be seen as a restriction on $f(x;\theta)$ on an average perspective, since the notion of mean squared differentiability is itself a property on average (integrating over x).

[1] Lemma 7.6 of [Van der Vaart, 2000].

Lemma 3.2. [2] *Suppose that Θ is an open subset of $\mathbb{R}^p$ and that the map $\theta \to \sqrt{f(x;\theta)}$ is mean squared differentiable at θ^*. Then, the score has mean zero at θ^*, i.e.,*

$$\int s_\theta(x;\theta^*) f(x;\theta^*) \mathrm{d}x = 0;$$

and all the entries of the FI matrix

$$I(\theta^*) = \int s_\theta(x;\theta^*) s_\theta(x;\theta^*)^\top f(x;\theta) \mathrm{d}x$$

exist. In other words, mean squared differentiability $\Rightarrow$ score has mean zero and finite variance.

The following lemma is analogous to (3.1) for a general estimator T. In addition, the assumptions on the density $f(x;\theta)$ is weakened.

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

1. *the map $\theta \to f(x;\theta)$ is continuous at θ^* for μ-a.s. x,*
2. *the map $\theta \to \sqrt{f(x;\theta)}$ is mean squared differentiable at θ^*,*
3. *the map $\theta \to \mathbb{E}_\theta T^2$ is continuous at θ^*.*

Then, the partial derivatives of the map $\theta \to \mathbb{E}_\theta T$ exist at θ^* and satisfy

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

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

Applying Lemma 3.3 with unbiased $T = \widehat{\beta}(\underline{X})$ recovers (3.1) under much weaker assumptions.

3.3 Regular parametric models

Definition 3.2 (**Regular parametric model**)**.** A parametric model is **regular** iff there exists a parameterization indexed by $\theta \in \Theta$ such that

1. Θ is an open set in $\mathbb{R}^p$,
2. for each $\theta^* \in \Theta$, the map $\theta \to f(x;\theta)$ is continuous at θ^* for μ-a.s. x,
3. the map $\theta \to \sqrt{f(\cdot;\theta)}$ is mean squared differentiable at all $\theta \in \Theta$,
4. the information matrix $I(\theta)$ is non-singular for all $\theta \in \Theta$.

[2] Lemma 7.2 of [Van der Vaart, 2000].

Example 3.3 (Weibull translation model)**.** The Weibull translation model has density

$$p(x;\theta) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \exp\left\{-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right\} \mathbb{I}_{(\gamma,+\infty)}(x),$$

with $\theta = (\alpha, \beta, \gamma) \in \Theta = \{\theta \colon \alpha > 0, \beta > 0, \gamma \in \mathbb{R}\}$. This model is not regular, but the model restricted on $\Theta_0 = \{\theta \colon \alpha > 0, \beta > 2, \gamma \in \mathbb{R}\} \subset \Theta$ is regular. The perturbation around $\beta = 1$ breaks the mean squared differentiability.

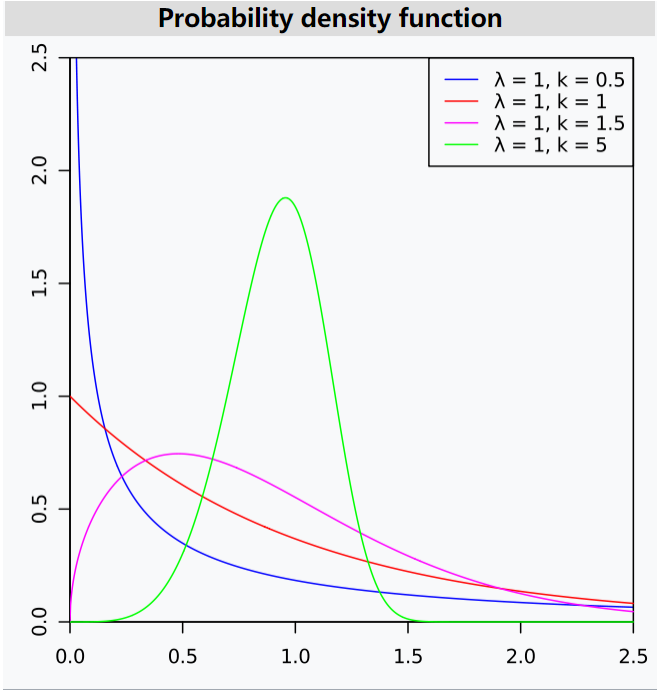

Figure 3.2: Weibull distribution

Example 3.4 (Three-parameter log-normal model)**.** The three-parameter log-normal model is

$$X = \gamma + \exp(Y), \quad Y \sim \mathcal{N}(\mu, \sigma^2),$$

so the density of X is

$$p(x;\theta) = \frac{1}{\sigma(x-\gamma)}\phi\left(\frac{\log(x-\gamma)-\mu}{\sigma}\right) \mathbb{I}_{(\gamma,+\infty)}(x),$$

with $\theta = (\mu, \sigma, \gamma) \in \Theta = \{\theta \colon \mu \in \mathbb{R}, \sigma^2 > 0, \gamma \in \mathbb{R}\}$. This model is also not regular.

3.4 Super-efficient estimators

In Section 3.1 we have shown that the variance of an unbiased estimator of $\beta(\theta)$ is no smaller than the C-R bound. For general estimators $\widehat{\beta}_n$ such that $\widehat{\beta}_n - \beta(\theta)$ converges (in an appropriate rate) to a mean zero distribution, its variance is usually

no less than the C-R bound, but not always true. One can construct "super-efficient" estimators of $\beta(\theta)$ that have variance equal to the C-R bound at most values of θ, while having smaller variance at some other values of θ. These estimators have excellent performance (as measured by mean squared error) at a θ at the expense of very poor performance in a neighborhood of θ.

Example 3.5 (Hodges)**.** Consider i.i.d. data $Z_1, \ldots, Z_n \sim \mathcal{N}(\beta, 1)$, and the estimator

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

Then

$$\sqrt{n}(\widehat{\beta}_n - \beta) \xrightarrow[n\to\infty]{D(F_\beta)} \begin{cases} \mathcal{N}(0,1), & \text{if } \beta \neq 0, \\ 0, & \text{if } \beta = 0. \end{cases}$$

The above example shows that we need to focus on "shrinking neighborhoods" of a parameter value, whose length shrinks with n at an appropriate rate. This can be made rigorous by **local asymptotic normality**.

Proposition 3.1. [3] *Suppose that $\mathcal{F} = \{f(x;\theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\}$ is a regular parametric model, then for $h \in \mathbb{R}^p$,*

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

This means in regular parametric models, it should be as hard to estimate h from n i.i.d. observations drawn from a law in the local model

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

as it should be to estimate h from one observation drawn from the normal model

$$\mathcal{F}_{\text{normal}} = \{\mathcal{N}(h, I(\theta)^{-1}) \colon h \in \mathbb{R}\}.$$

3.5 Regular estimators

Definition 3.3 (**Regular estimators**)**.** Given a collection of densities $\mathcal{F} = \{f(x;\theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\}$ and n i.i.d. observations $X_1^{(n)}, \ldots, X_n^{(n)}$, an estimator sequence $\widehat{\beta}_n$ based on $X_1^{(n)}, \ldots, X_n^{(n)}$ is a **regular estimator** of $\beta(\theta)$ at the law F_θ iff there exists a law G_θ such that for all $h \in \mathbb{R}^p$,

$$\sqrt{n}\{\widehat{\beta}_n - \beta(\theta + h/\sqrt{n})\} \xrightarrow{D(F_{\theta + h/\sqrt{n}})} G_\theta.$$

In other words, a regular estimator should behave "uniformly well" in the local model:

1. the convergence holds for all $h \in \mathbb{R}^p$;
2. the convergence rate is of order $\sqrt{n}$.

[3] Theorem 7.2 of [Van der Vaart, 2000].

3.6 Efficient estimators in regular parametric models

Proposition 3.2 (**Hajek's convolution theorem**). *Let $\mathcal{F} = \{f(x;\theta)\colon \theta \in \Theta \subseteq \mathbb{R}^p\}$ be a regular parametric model and let $\beta(\theta)$ be an $\mathbb{R}^k$-valued parameter that is differentiable w.r.t. θ. If $\widehat{\beta}_n$ is a regular estimator of $\beta(\theta)$ at F_θ, then*

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

where $U^* \sim \mathcal{N}_k(0, C_{\mathcal{F}}(\theta))$ and U is independent of U^*. In other words, regular model + regular estimator = convolution.

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

$$\sqrt{n}\{\widehat{\beta}_n - \beta(\theta)\} \xrightarrow{D(F_\theta)} \mathcal{N}_k(0, C_{\mathcal{F}}(\theta)).$$

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

If the map $\theta \to f(x;\theta)$ satisfies a "smoothness" condition, then a globally asymptotically efficient estimator exists, and is the maximum likelihood estimator (MLE). However, the globally efficient estimator (MLE) is not always asymptotically "optimal" in the sense of mean squared error.

Example 3.6 (Stein's shrinkage estimator). Suppose we are given one draw Z from $N_p(h, I_p)$ and need to estimate $h \in \mathbb{R}^p$. If $p \geq 3$, define the **Stein's shrinkage estimator**

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

$\hat{h}$ is biased whenever $h \neq 0$. For all $h \in \mathbb{R}^p$,

$$\mathbb{E}_h\|\hat{h} - h\|^2 < \mathbb{E}_h\|Z - h\|^2 = p.$$

So, even though Z is minimax for the quadratic loss, the loss of $\hat{h}$ is always strictly smaller than that of Z. In the case of n draws, it can be proved that the Stein's shrinkage estimator is NOT regular (by contracting with Definition 3.3).

3.7 Asymptotically linear estimators

Definition 3.5 (**Asymptotically linear estimators**). A sequence of estimators $\widehat{\beta}_n = \widehat{\beta}_n(X_1, \ldots, X_n)$ of a parameter $\beta(\cdot)\colon \mathcal{F} \to \mathbb{R}^k$ is **asymptotically linear** iff there exists a random vector $\varphi_F(X)$ with mean zero and finite variance under F such that

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

The function $\varphi_F(\cdot)$ is called the **influence function** of the estimator $\widehat{\beta}_n$ at F.

If $\widehat{\beta}_n$ is an asymptotically linear estimator with influence function $\varphi_F(\cdot)$, then

$$\sqrt{n}\{\widehat{\beta}_n - \beta(F)\} \xrightarrow{D(F_\theta)} \mathcal{N}_k(0, \mathrm{var}\{\varphi_F(X)\}).$$

Lemma 3.4. [4] *Suppose that $\mathcal{F} = \{f(x;\theta)\colon \theta \in \Theta\}$ is a regular parametric model and $\beta(\theta)$ is an $\mathbb{R}^k$-valued parameter that is differentiable w.r.t. θ. Then, if $\widehat{\beta}_n$ is asymptotically efficient, it must be asymptotically linear. In other words, in regular parametric models, $\{$asymptotically efficient$\} \subseteq \{$asymptotically linear$\}$.*

3.8 Properties of RAL estimators

The next lemma provides a characterization of the set of influence functions of **regular and asymptotically linear** (RAL) estimators, and is an "asymptotic version" of Lemma 3.3. It is a corollary of Le-Cam's third lemma.

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

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

where $S_\theta(\theta^*) = s_\theta(X;\theta^*)$.

Remark 3.4. A direct comparison between (3.1), Lemma 3.3 and Lemma 3.5:

$$\begin{cases} \dfrac{\partial \beta(\theta)}{\partial \theta^\top} = \mathbb{E}_\theta\{\widehat{\beta}_n S_\theta(\theta)^\top\} \text{ unbiased estimator}, \\ \dfrac{\partial \mathbb{E}_\theta T}{\partial \theta^\top} = \mathbb{E}_\theta\{T S_\theta(\theta)^\top\} \text{ general estimator}, \\ \dfrac{\partial \beta(\theta)}{\partial \theta^\top} = \mathbb{E}_\theta\{\varphi_F(X) S_\theta(\theta)^\top\} \text{ RAL estimator}. \end{cases}$$

Remark 3.5. Lemma 3.5 has several important consequences:

1. the projections of all influence functions of RAL estimators into the tangent space of the model coincide;
2. the variance of the projection of any influence function of a RAL estimator into the tangent space is equal to the C-R bound;
3. the set of all influence functions of RAL estimators of $\beta(\theta)$ at F_θ is equal to $\{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$;
4. influence functions of RAL estimators are orthogonal, i.e. uncorrelated, with the "nuisance tangent space";

[4] A corollary of Lemma 8.14 of [Van der Vaart, 2000].

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

Proof of Remark 3.5. We prove the five claims.

1. By Lemma 3.5, we know that

$$\begin{aligned}\Pi_\theta\left[\varphi_{F_\theta}(X) \mid \Lambda(\theta)\right] &= \mathbb{E}_\theta\left\{\varphi_{F_\theta}(X) S_\theta(\theta)^\top\right\} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} S_\theta(\theta) \\ &= \frac{\partial \beta(\theta)}{\partial \theta^\top} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} S_\theta(\theta) \\ &= \mathbb{E}_\theta\left\{\varphi'_{F_\theta}(X) S_\theta(\theta)^\top\right\} \operatorname{var}_\theta\left(S_\theta(\theta)\right)^{-1} S_\theta(\theta) \\ &= \Pi_\theta\left[\varphi'_{F_\theta}(X) \mid \Lambda(\theta)\right].\end{aligned}$$

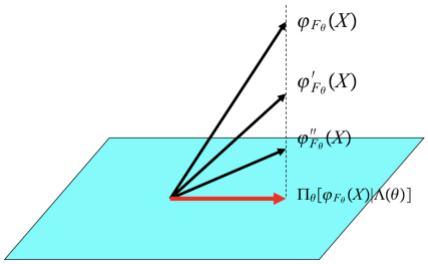

Figure 3.3: Projection into the tangent space

2. Direct calculation yields

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

3. Suppose $\varphi_{F_\theta}(X)$ and $\varphi'_{F_\theta}(X)$ are influence functions of two RAL estimators. Then

$$\mathbb{E}_\theta\left\{\left[\varphi_{F_\theta}(X) - \varphi'_{F_\theta}(X)\right] S_\theta(\theta)^\top\right\} = \frac{\partial \beta(\theta)}{\partial \theta^\top} - \frac{\partial \beta(\theta)}{\partial \theta^\top} = 0,$$

so we have $\varphi'_{F_\theta}(X) \in \{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$. On the other hand, take any random vector $\psi_{F_\theta}(X) \in \{\varphi_{F_\theta}(X)\} + \Lambda(\theta)^\perp$, which can be represented as

$$\psi_{F_\theta}(X) = \varphi_{F_\theta}(X) + \omega_{F_\theta}(X), \text{ for some } \omega_{F_\theta}(X) \in \Lambda(\theta)^\perp.$$

It can be shown (Section 3.3 of [Tsiatis, 2006]) that $\psi_{F_\theta}(X)$ is the influence function of an asymptotically linear estimator. To show that $\psi_{F_\theta}(X)$ is regular, it suffices to use Lemma 3.5 in the inverse direction.

4. Suppose the model is indexed by variational independent parameters β and η, i.e., $\theta^\top = (\beta^\top, \eta^\top)$. Define the **nuisance tangent space** as the tangent space where β is fixed and η is unknown,

$$\Lambda_{\mathcal{F},\text{nuis}}(\theta) = \{a^\top S_\eta(\theta) \colon a \in \mathbb{R}^q\}.$$

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

$$\mathbb{E}_\theta\{\varphi_{F_\theta}(X) S_\eta(\theta)^\top\} = \frac{\partial \beta(\theta)}{\partial \eta} = 0.$$

The second equality follows from the variational independence of β and η. Thus $\varphi_{F_\theta}(X)$ satisfies

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

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

$$\mathbb{E}_\theta\{\varphi_{F_\theta}(X) S_\beta(\theta)^\top\} = \frac{\partial \beta(\theta)}{\partial \beta} = I_k.$$

□

3.9 Example: asymptotic linearity and regularity of Z-estimators

Suppose that $\mathcal{F} = \{f(x;\theta) \colon \theta \in \Theta\}$ ia s parametric or semiparametric model and let $\beta(\cdot) \colon \mathcal{F} \to \mathbb{R}^k$. Suppose $u(X;\beta)$ is a known $\mathbb{R}^k$-valued vector function of X and β such that for each $\theta \in \Theta$ there exists a unique $\beta(\theta) = \beta(F_\theta)$ satisfying

$$\mathbb{E}_\theta[u(X;\beta(\theta))] = 0.$$

The Z-estimator for β is the estimator $\widehat{\beta}_n$ solving the equation

$$\sum_{i=1}^n u(X_i;\beta) = 0.$$

Proposition 3.3 (Asymptotic linearity of Z-estimators)**.** *If $\widehat{\beta}_n$ satisfies*

$$n^{-1/2} \sum_{i=1}^n u(X_i;\widehat{\beta}_n) = o_p(1) \tag{3.2}$$

and

1. $\left[\partial \mathbb{E}_F[u(X;\beta)]/\partial \beta^\top\right]\big|_{\beta=\beta(F)}$ *exists and is invertible,*
2. $\mathbb{E}_F\left[u(X;\beta(F))^\top u(X;\beta(F))\right] < \infty$*,*
3. *some other regularity conditions hold (Donsker property),*

then the estimator $\widehat{\beta}_n$ *satisfies*

$$\sqrt{n}\{\widehat{\beta}_n - \beta(F)\} = -\left\{\frac{\partial \mathbb{E}_F[u(X;\beta)]}{\partial \beta^\top}\bigg|_{\beta=\beta(F)}\right\}^{-1} \frac{1}{\sqrt{n}}\sum_{i=1}^n u(X_i;\beta(F)) + o_p(1).$$

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

$$\varphi_F(X) = -\left\{\frac{\partial \mathbb{E}_F[u(X;\beta)]}{\partial \beta^\top}\bigg|_{\beta=\beta(F)}\right\}^{-1} u(X;\beta(F)).$$

Remark 3.6. In the case of MLE, $u(X;\beta) = \nabla \log f(X;\beta) = s_\beta(X;\beta)$, so the influence function is $\varphi_F(X) = I(\beta)^{-1} s_\beta(X;\beta)$.

Proposition 3.4. *If the assumptions in Proposition 3.3 hold, and in addition,*

1. *the map* $\theta \to \mathbb{E}_\theta\{u(X;\beta(\theta^*))^\top u(X;\beta(\theta^*))\}$ *is continuous at* θ^*,
2. *the map* $(\beta,\theta) \to \mathbb{E}_\theta\{u(X;\beta)\}$ *has continuous partial derivatives in an open neighborhood of* $(\beta(\theta^*),\theta^*)$.

Then the Z-estimator $\widehat{\beta}_n$ *is not only asymptotically linear but also regular.*

Example 3.7 (Median)**.** Let $\mathcal{F}$ be the model for the law of X which only assumes that the law is absolutely continuous with respect to the Lebesgue measure and the density $f(x)$ is continuous. Let $\beta(F)$ be the median of F. Define the sample median $\widehat{\beta}_n$ as $X_{((n+1)/2)}$ if n is odd and as $(X_{(n/2)} + X_{(n/2+1)})/2$ if n is even.

We begin by verifying assumptions in Proposition 3.3. It is easy to see that $\widehat{\beta}_n$ solves

$$n^{-1/2}\sum_{i=1}^n u(X_i;\beta) = o_p(1)$$

with $u(X;\beta) = \operatorname{sgn}(\beta - X) = 2\mathbb{I}(X \le \beta) - 1$. In addition,

$$\frac{\partial \mathbb{E}_F[u(X;\beta)]}{\partial \beta}\bigg|_{\beta=\beta(F)} = \frac{\partial(2F(\beta)-1)}{\partial\beta}\bigg|_{\beta=\beta(F)} = 2f(\beta(F)).$$

By Proposition 3.3, $\widehat{\beta}_n$ is asymptotically linear with influence function

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

We then verify the assumptions in Proposition 3.4. It holds trivially that

$$\mathbb{E}_\theta\{u(X;\beta(\theta^*))^\top u(X;\beta(\theta^*))\} = 1.$$

In addition, since $\mathbb{E}_\theta\{u(X;\beta)\} = 2F_\theta(\beta) - 1$, we have

$$\frac{\partial \mathbb{E}_\theta\{u(X;\beta)\}}{\partial\beta} = 2f(\beta;\theta), \quad \frac{\partial \mathbb{E}_\theta\{u(X;\beta)\}}{\partial\theta} = \mathbb{E}_\theta\{u(X;\beta)s_\theta(X;\theta)\} = 2\mathbb{E}_\theta\{\mathbb{I}(X\le\beta)s_\theta(X;\theta)\}.$$

Assume the above two quantities is continuous in an open neighborhood of $(\beta(\theta^*)$. Then by Proposition 3.4, $\widehat{\beta}_n$ is also regular.

3.10 Efficient influence functions and efficient scores

In Remark 3.5, we know that the projection of any RAL estimator into the tangent space is the RAL estimator achieving the C-R bound. We call it the **efficient influence function**, i.e.,

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

Suppose the parametric model is indexed by $\theta = (\beta^\top, \eta^\top)^\top \in \mathbb{R}^p$ with $\beta \in \mathbb{R}^k$ and $\eta \in \mathbb{R}^q$ being variational independent. We want to calculate the C-R bound to find an additional useful geometric interpretation.

The C-R bound is

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

where

$$\underbrace{\frac{\partial\beta(\theta)}{\partial\theta^\top}}_{k\times p} = \begin{bmatrix} I_k & 0_{k\times q}\end{bmatrix},$$

and

$$\begin{aligned} I(\theta)^{-1} &= \mathbb{E}_\theta\left\{S_\theta(\theta)S_\theta(\theta)^\top\right\}^{-1} \\ &= \begin{bmatrix}\mathbb{E}_\theta[S_\beta(\theta)S_\beta(\theta)^\top] & \mathbb{E}_\theta[S_\beta(\theta)S_\eta(\theta)^\top] \\ \mathbb{E}_\theta[S_\eta(\theta)S_\beta(\theta)^\top] & \mathbb{E}_\theta[S_\eta(\theta)S_\eta(\theta)^\top]\end{bmatrix}^{-1} \\ &= \begin{bmatrix} I_{\beta,\beta}(\theta) & I_{\beta,\eta}(\theta) \\ I_{\eta,\beta}(\theta) & I_{\eta,\eta}(\theta)\end{bmatrix}^{-1} \\ &= \begin{bmatrix} I^{\beta,\beta}(\theta) & I^{\beta,\eta}(\theta) \\ I^{\eta,\beta}(\theta) & I^{\eta,\eta}(\theta)\end{bmatrix}. \end{aligned}$$

Thus,

$$C_{\mathcal{F}}(\theta) = \begin{bmatrix} I_k & 0_{k\times q}\end{bmatrix}\begin{bmatrix} I^{\beta,\beta}(\theta) & I^{\beta,\eta}(\theta) \\ I^{\eta,\beta}(\theta) & I^{\eta,\eta}(\theta)\end{bmatrix}\begin{bmatrix} I_k \\ 0_{q\times k}\end{bmatrix} = I^{\beta,\beta}(\theta).$$

From the formula of a partitioned matrix we have that

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

and replacing each term by their expressions we obtain

$$\begin{aligned} I^{\beta,\beta}(\theta) =& \left[\text{var}_\theta\left\{S_\beta(\theta)\right\} - \mathbb{E}_\theta\left\{S_\beta(\theta)S_\eta(\theta)^\top\right\}\text{var}_\theta\left\{S_\eta(\theta)\right\}^{-1}\mathbb{E}_\theta\left\{S_\eta(\theta)S_\beta(\theta)^\top\right\}\right]^{-1} \\ =& \left\{\text{var}_\theta\left[S_\beta(\theta) - \mathbb{E}_\theta\left\{S_\beta(\theta)S_\eta(\theta)^\top\right\}\text{var}_\theta\left\{S_\eta(\theta)\right\}^{-1}S_\eta(\theta)\right]\right\}^{-1} \\ =& \left\{\text{var}_\theta\left[S_\beta(\theta) - \Pi_\theta\left[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\text{nuis}}(\theta)\right]\right]\right\}^{-1}. \end{aligned}$$

If we define the **efficient score** as

$$S_{\beta,\text{eff}} = S_\beta(\theta) - \Pi_\theta[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\text{nuis}}(\theta)],$$

then

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

In addition, the efficient influence function and the efficient score are related as

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

This means we can obtain an efficient influence function once we have obtained an efficient score. It is useful since directly finding an influence function and projecting it into the tangent space will be more cumbersome.

When η is known, the inverse of the C-R bound is the FI matrix $\text{var}_\theta\{S_\beta(\theta)\}$. It is no smaller than $\text{var}_\theta\{S_{\beta,\text{eff}}(\theta)\}$ and the equality holds iff $S_\beta(\theta)$ and $S_\eta(\theta)$ are orthogonal, in which case there is no loss of information in not knowing η.

3.11 Summary

1. Let $\varphi_{F_\theta}(X)$ be the influence function of an asymptotically linear estimator. Then
$$\frac{\partial \beta(\theta)}{\partial \theta} = \mathbb{E}_\theta\{\varphi_{F_\theta}(X) S_\theta(\theta)^\top\}$$
iff the estimator is regular.

2. If $\varphi_{F_\theta}(X)$ and $\varphi'_{F_\theta}(X)$ are the influence functions of two RAL estimators, then
$$\Pi_\theta[\varphi_{F_\theta}(X) \mid \Lambda(\theta)] = \Pi_\theta[\varphi'_{F_\theta}(X) \mid \Lambda(\theta)].$$

3. The efficient influence function is defined as
$$\varphi_{F_\theta,\text{eff}} = \Pi_\theta[\varphi_{F_\theta}(X) \mid \Lambda(\theta)].$$

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

5. The C-R bound is equal to
$$C_{\mathcal{F}}(\theta) = \text{var}_\theta\{\varphi_{F_\theta}(X)\} = \text{var}_\theta\{\Pi_\theta[\varphi_{F_\theta}(X) \mid \Lambda(\theta)]\}.$$

6. If $\theta^\top = (\beta^\top, \eta^\top)$, then the set of influence functions of RAL estimators of β is equal to
$$\{\varphi_{F_\theta}(X) \in \mathcal{L}_2^0(\theta) \colon \Pi_\theta[\varphi_{F_\theta}(X) \mid \Lambda_{\mathcal{F},\text{nuis}}(\theta)] = 0 \text{ and } \mathbb{E}_\theta[\varphi_{F_\theta}(X) S_\beta(\theta)^\top] = I_k\}.$$

7. If $\theta^\top = (\beta^\top, \eta^\top)$, then the efficient score for β is defined as
$$S_{\beta,\text{eff}} = S_\beta(\theta) - \Pi_\theta[S_\beta(\theta) \mid \Lambda_{\mathcal{F},\text{nuis}}(\theta)].$$

8. If $\theta^\top = (\beta^\top, \eta^\top)$, then the efficient influence function is equal to
$$\varphi_{F_\theta,\text{eff}}(X) = \text{var}_\theta\{S_{\beta,\text{eff}}(\theta)\}^{-1} S_{\beta,\text{eff}}(\theta).$$

9. If $\theta^\top = (\beta^\top, \eta^\top)$, then the C-R bound for β is equal to
$$C_{\mathcal{F}}(\theta) = \text{var}_\theta\{S_{\beta,\text{eff}}(\theta)\}^{-1}.$$

The geometry of inference

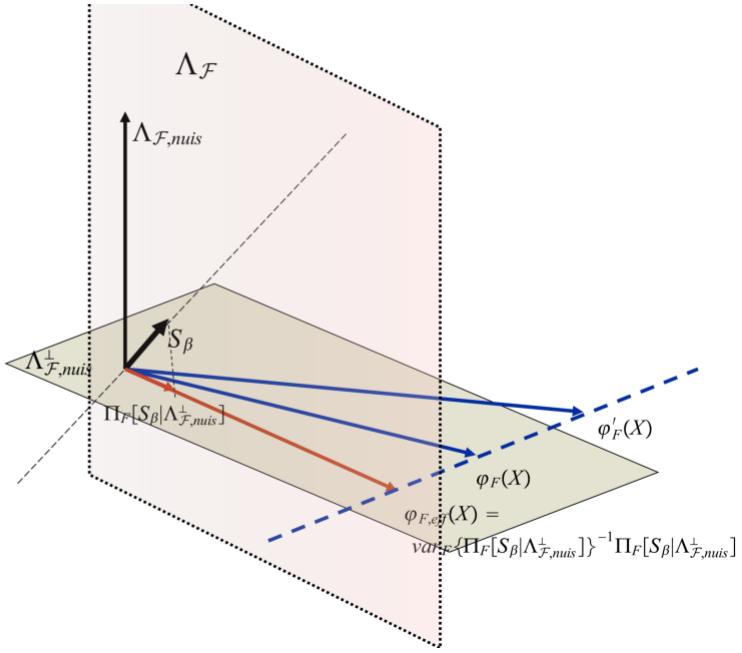

Figure 3.4: The geometry of inference

4 Efficiency in Semiparametric Models

4.1 Regular parametric submodels

Definition 4.1 (**Regular parametric submodel**)**.** Given a semiparametric model $\mathcal{F}$, a regular parametric model is a **submodel** of $\mathcal{F}$ at F iff

1. F is in the submodel;
2. the family of distributions allowed by the parametric model is included in $\mathcal{F}$;
3. the parametric model is regular.

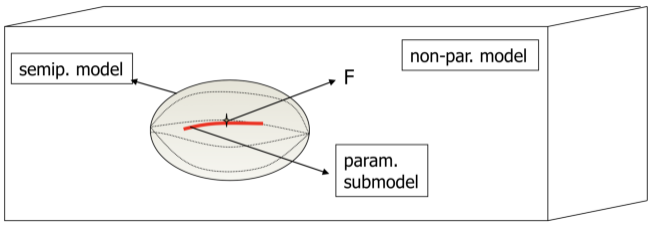

Figure 4.1: Parametric submodels

Example 4.1. Suppose $\mathcal{F}$ is the nonparametric model and F is the $\mathcal{N}(0,1)$ distribution. Then

$$\mathcal{F}_{\text{sub}} = \{\mathcal{N}(\theta, 1) \colon \theta \in \mathbb{R}\},$$

and

$$\mathcal{F}'_{\text{sub}} = \{\mathcal{N}(\theta_1, \theta_2^2) \colon \theta_1 \in \mathbb{R}, \theta_2 \in \mathbb{R}^+\},$$

are regular parametric submodels of $\mathcal{F}$ at F.

4.2 Tangent sets and tangent spaces

Let $\mathcal{A}$ be a collection of regular parametric submodels of a semiparametric model $\mathcal{F}$ through F, i.e.,

$$\mathcal{A} = \{\mathcal{F}_{\text{sub}} \colon \mathcal{F}_{\text{sub}} \text{ is a regular parametric submodel of } \mathcal{F} \text{ through } F\}.$$

The **tangent set** of the model $\mathcal{F}$ w.r.t. $\mathcal{A}$ at F is defined as

$$\bigcup_{\mathcal{F}_{\text{sub}} \colon \mathcal{F}_{\text{sub}} \in \mathcal{A}} \Lambda_{\mathcal{F}_{\text{sub}}}(F),$$

where $\Lambda_{\mathcal{F}_{\text{sub}}}(F)$ is the tangent space of $\mathcal{F}_{\text{sub}}$ at F. The **tangent space** of the model $\mathcal{F}$ w.r.t. $\mathcal{A}$ at F is defined as the **closure of the linear span** of the tangent set,

$$\Lambda_{\mathcal{F}}(F) = \overline{\left[\bigcup_{\mathcal{F}_{\text{sub}} \colon \mathcal{F}_{\text{sub}} \in \mathcal{A}} \Lambda_{\mathcal{F}_{\text{sub}}}(F)\right]}.$$

Remark 4.1. The notion of the tangent set/space of a semiparametric model $\mathcal{F}$ is w.r.t. a collection $\mathcal{A}$ of parametric submodels.

The calculation of a tangent space usually consists of three steps:

1. guess the form of the tangent space (by the restrictions the scores must satisfy);
2. show that the scores of regular submodels are in the postulated set (easy);
3. show that for any given element of the postulated set, a sequence a regular parametric submodels exists such that linear combinations of the scores in the sequence tend to the given element (difficult).

4.3 Example: maximal tangent space in the nonparametric model

Suppose that $\mathcal{A}$ is the class of all regular parametric submodels of the nonparametric model $\mathcal{F}$. Since the nonparametric model does not impose any restrictions on the distributions, the set of scores of all regular parametric submodels should not have any restriction beyond the requirement that they be mean zero and with finite variance. So, it is natural to conjecture that the tangent space at a given law F is equal to $\mathcal{L}_2^0(F)$.

On the one hand, $\Lambda_{\mathcal{F}}(F) \subseteq \mathcal{L}_2^0(F)$ by definition. On the other hand, for any $g(X) \in \mathcal{L}_2^0(F)$, define the one dimensional submodel

$$\mathcal{F}_{\text{sub},g} = \{f(x;\theta)\colon \theta \in (-\varepsilon, \varepsilon)\},$$

where $f(x;\theta) = f(x)k\{\theta g(x)\}c(\theta)$ with $f(x)$ the density of F, $k(u) = 2(1+e^{-2u})^{-1}$ and $c(\theta) = \left\{\int f(x)k(\theta g(x))\mathrm{d}x\right\}^{-1}$. Setting $\theta = 0$ recovers the law F. The score at $\theta = 0$ is

$$\begin{aligned} s_\theta(x;\theta)|_{\theta=0} &= \frac{\partial}{\partial\theta}\log f(x;\theta)\bigg|_{\theta=0} \\ &= \frac{f(x)g(x)k'\{\theta g(x)\}c(\theta) + f(x)k\{\theta g(x)\}c'(\theta)}{f(x)k\{\theta g(x)\}c(\theta)}\bigg|_{\theta=0} \\ &= \frac{f(x)g(x)k'(0)c(0) + f(x)k(0)c'(0)}{f(x)k(0)c(0)}. \end{aligned}$$

Note that $k(0) = k'(0) = 1$, and

$$c'(\theta) = -\left[\int f(x)k\{\theta g(x)\}\mathrm{d}x\right]^{-2}\int f(x)k'\{\theta g(x)\}g(x)\mathrm{d}x$$

implies $c'(0) = 0$ because $g(X)$ has mean zero and $k'(0) = 1$. Furthermore, $c(0) = 1$. Thus $s_\theta(x;\theta)|_{\theta=0} = g(x)$.

Remark 4.2 (**Two useful submodels**). The following two submodels are useful:

1. $f(x;\theta) = f(x)k\{\theta g(x)\}c(\theta)$ with $f(x)$ the density of F, $k(u) = 2(1+e^{-2u})^{-1}$ and $c(\theta) = \{\int f(x)k(\theta g(x))\mathrm{d}x\}^{-1}$;

2. $f(x;\theta) = f(x)\{1+\theta g(x)\}c(\theta)$ with $f(x)$ the density of F and $c(\theta) = \{\int f(x)(1+\theta g(x))\mathrm{d}x\}^{-1}$.

In the second one, $f(x;\theta)$ is a density as long as $g(x)$ is bounded with mean zero and θ is in a sufficiently small interval. For any unbounded function in $\mathcal{L}_2^0(F)$, it can be approximated by bounded functions in $\mathcal{L}_2^0(F)$, i.e., for any $h \in \mathcal{L}_2^0(F)$ there exist bounded $g_1, g_2, \ldots$ such that $\|g_n - h\|_{\mathcal{L}_2^0(F)} \to 0$, and proofs may be completed.

Remark 4.3. There are two important points:

1. different submodels can have the same score at F;

2. different collections of submodels can have the same tangent space at F.

4.4 Regular estimators

Definition 4.2 (**Regular estimators**)**.** Given a collection $\mathcal{A}$ of regular parametric submodels of a semiparametric model $\mathcal{F}$, an estimator $\widehat{\beta}_n$ is said to be a **regular estimator** of a parameter $\beta(F)$ in model $\mathcal{F}$ with respect to $\mathcal{A}$ at F, if it is a regular estimator at F under every parametric submodel in $\mathcal{A}$.

Example 4.2. Consider the nonparametric model $\mathcal{F}$ for X which is absolutely continuous w.r.t. Lebesgue measure. For a given function $b(\cdot)$, let $\beta(F) = \mathbb{E}_F[b(X)]$. Let $\mathcal{A}$ be the class of all regular parametric submodels such that for each model in $\mathcal{A}$, indexed by θ, through F^* at θ^*, the map $\theta \to \mathbb{E}_\theta[b(X)^2]$ is continuous in an open neighborhood of θ^*.

The empirical mean $\widehat{\beta}_n = n^{-1}\sum_{i=1}^n b(X_i)$ satisfies

$$\sqrt{n}\{\widehat{\beta}_n - \beta(F)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^n \{b(X_i) - \beta(F)\},$$

so it is asymptotically linear with influence function

$$\varphi_{F^*}(X) = b(X) - \mathbb{E}_{F^*}[b(X)].$$

In each regular parametric submodel, we have by Lemma 3.3,

$$\left.\frac{\partial\beta(\theta)}{\partial\theta^\top}\right|_{\theta=\theta^*} = \left.\frac{\partial\mathbb{E}_\theta[b(X)]}{\partial\theta^\top}\right|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}[b(X)s_\theta(X;\theta^*)^\top] = \mathbb{E}_{\theta^*}[\varphi_{F^*}(X)s_\theta(X;\theta^*)^\top].$$

Then by Lemma 3.5, the empirical mean $\widehat{\beta}_n$ is a regular estimator of $\beta(F)$ in the model $\mathcal{F}$ w.r.t. $\mathcal{A}$ at F^*.

Remark 4.4. In the above example, the requirement on the submodels in the class $\mathcal{A}$ of the continuity of the map $\theta \to \mathbb{E}_\theta[b(X)^2]$ is a technical requirement that we impose to ensure that the sample average is a "regular" estimator in the submodel. Although the class $\mathcal{A}$ is not the class of all regular parametric submodels of the non-parametric model $\mathcal{F}$, the tangent space w.r.t. $\mathcal{A}$ at F^* is indeed equal to the maximal tangent space of the non-parametric model, i.e, $\mathcal{L}_2^0(F^*)$. This will be an important point when we discuss the efficiency bounds. To show the tangent space w.r.t. $\mathcal{A}$ is $\mathcal{L}_2^0(F^*)$, it suffices to show that for any $g(X) \in \mathcal{L}_2^0(F^*)$, the map $\theta \to \mathbb{E}_\theta[b(X)^2]$ is continuous under the submodel

$$\mathcal{F}_{\text{sub},g} = \{f(x;\theta)\colon f(x;\theta) = f^*(x)c(\theta)k(\theta g(x)), \theta \in (-\varepsilon, \varepsilon)\},$$

which is defined in Section 4.3.

As in this example, in nearly all inferential problems, the "maximal class" of all regular parametric submodels of a given semiparametric model is too big to find an estimator that is regular w.r.t. the maximal submodel class. However, in the spirit of making inferences valid in the big semiparametric model, i.e., to ensure valid inference under a large set of data generating laws, we would like to find regular estimators w.r.t. a class $\mathcal{A}$ that is as big as possible. In many examples, we can choose the class $\mathcal{A}$ so that the tangent space associated with $\mathcal{A}$ coincides with the maximal tangent space of the assumed semiparametric model.

4.5 Pathwise differentiable parameters and semiparametric C-R bounds

Suppose that $\widehat{\beta}_n$ is a regular estimator of a scalar parameter $\beta(F)$ in a semiparametric model $\mathcal{F}$ w.r.t. $\mathcal{A}$ at F. Then, the asymptotic variance of $\sqrt{n}\{\widehat{\beta}_n - \beta(F)\}$ cannot be smaller than $C_{\mathcal{F}_{\text{sub}}}(F)$ for every $\mathcal{F}_{\text{sub}} \in \mathcal{A}$. Thus it cannot be smaller than

$$\sup_{\mathcal{F}_{\text{sub}}\colon \mathcal{F}_{\text{sub}} \in \mathcal{A}} C_{\mathcal{F}_{\text{sub}}}(F).$$

Therefore, a necessary condition for $\widehat{\beta}_n$ to be regular w.r.t. $\mathcal{A}$ is

$$\sup_{\mathcal{F}_{\text{sub}}\colon \mathcal{F}_{\text{sub}} \in \mathcal{A}} C_{\mathcal{F}_{\text{sub}}}(F) < \infty.$$

In this section, we will see that a sufficient condition on an estimand $\beta(F)$ for the above display to be true is that the functional $F \to \beta(F)$ be pathwise differentiable at F w.r.t. $\mathcal{A}$.

Suppose that we wish to estimate the mean of a real valued function of X, i.e., $\beta(F) = \mathbb{E}_F[\psi(X)]$, for some known function $\psi(X)$ under a given semiparametric model $\mathcal{F}$ with the restriction

$$\text{var}_F[\psi(X)] < \infty, \quad \text{for all } F \in \mathcal{F},$$

and perhaps some other restrictions. For any $\mathcal{F}_{\text{sub}} \in \mathcal{A}$, due to results on unbiased estimators in Section 3.1, it holds that

$$C_{\mathcal{F}_{\text{sub}}}(F) = \text{var}_F\{\Pi_F[\psi(X) \mid \Lambda_{\mathcal{F}_{\text{sub}}}(F)]\}.$$

Road map to the answer

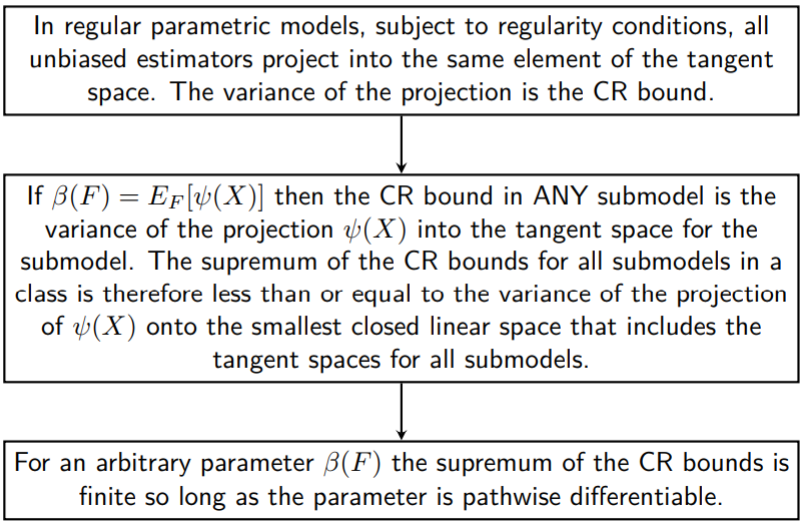

Figure 4.2: Roadmap to the answer

Taking supremum over $\mathcal{A}$, we get

$$\begin{aligned}\sup_{\mathcal{F}_{\text{sub}}:\, \mathcal{F}_{\text{sub}} \in \mathcal{A}} C_{\mathcal{F}_{\text{sub}}}(F) &= \sup_{\mathcal{F}_{\text{sub}}:\, \mathcal{F}_{\text{sub}} \in \mathcal{A}} \text{var}_F\{\Pi_F[\psi(X) \mid \Lambda_{\mathcal{F}_{\text{sub}}}(F)]\} \\ &\leq \text{var}_F\{\Pi_F[\psi(X) \mid \Lambda_{\mathcal{F}}(F)]\},\end{aligned}$$

where

$$\Lambda_{\mathcal{F}}(F) = \overline{\left[\bigcup_{\mathcal{F}_{\text{sub}}:\, \mathcal{F}_{\text{sub}} \in \mathcal{A}} \Lambda_{\mathcal{F}_{\text{sub}}}(F)\right]}.$$

Since $\Lambda_{\mathcal{F}}(F)$ is a closed linear subspace of $\mathcal{L}_2^0(F)$, we get

$$\text{var}_F\{\Pi_F[\psi(X) \mid \Lambda_{\mathcal{F}}(F)]\} < \infty.$$

In the above arguments, the key point is that

$$C_{\mathcal{F}_{\text{sub}}}(F) = \text{var}_F\{\Pi_F[\psi(X) \mid \Lambda_{\mathcal{F}_{\text{sub}}}(F)]\},$$

which indicates that the C-R bound for $\beta(F) = \mathbb{E}_F[\psi(X)]$ in ANY regular parametric submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}$ is the variance of the projection of the SAME r.v. $\psi(X)$ into the tangent space of the submodel. For a general parameter $\beta\colon \mathcal{F} \to \mathbb{R}$, we may hope that there exists a random variable $\psi_F(X)$ such that

$$C_{\mathcal{F}_{\text{sub}}}(F) = \text{var}_F\{\Pi_F[\psi_F(X) \mid \Lambda_{\mathcal{F}_{\text{sub}}}(F)]\}.$$

This prompt the definition of "pathwise differentiable" parameters.

Definition 4.3 (**Pathwise differentiable parameters**)**.** Given a semiparametric model $\mathcal{F}$, a law $F^* \in \mathcal{F}$, and a class $\mathcal{A}$ of regular parametric submodels of $\mathcal{F}$, a real valued functional $\beta\colon \mathcal{F} \to \mathbb{R}$ is said to be a **pathwise differentiable or regular** parameter at F^* w.r.t. $\mathcal{A}$ in the model $\mathcal{F}$ iff there exists $\psi_{F^*}(X) \in \mathcal{L}_2(F^*)$ such that for each submodel in $\mathcal{A}$, indexed by θ with $F^* = F_{\theta^*}$ and score $S_\theta(\theta^*) = s_\theta(X;\theta^*)$ at θ^*, it holds that

$$\left.\frac{\partial \beta(F_\theta)}{\partial \theta^\top}\right|_{\theta=\theta^*} = \mathbb{E}_{F^*}\left[\psi_{F^*}(X) S_\theta(\theta^*)^\top\right].$$

$\psi_{F^*}(X)$ is called a **gradient** of β at F^* w.r.t. $\mathcal{A}$. If in addition $\psi_{F^*}(X)$ is mean zero, then it is commonly referred to as an **influence function**.

Now suppose that β is a real-valued pathwise differentiable parameter w.r.t. a class $\mathcal{A}$ of regular parametric submodels through F at θ. If $\psi_F(\cdot)$ is a gradient of β, then for any submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}$,

$$\begin{aligned} C_{\mathcal{F}_{\text{sub}}}(\theta) &= \left.\frac{\partial \beta(\theta')}{\partial \theta'^\top}\right|_{\theta'=\theta} \operatorname{var}_\theta\{S_\theta(\theta)\}^{-1} \left.\frac{\partial \beta(\theta')}{\partial \theta'}\right|_{\theta'=\theta} \\ &= \mathbb{E}_\theta\left\{\psi_F(X) S_\theta(\theta)^\top\right\} \operatorname{var}_\theta\{S_\theta(\theta)\}^{-1} \mathbb{E}_\theta\left\{\psi_F(X) S_\theta(\theta)^\top\right\}^\top \\ &= \operatorname{var}_\theta\{\Pi_\theta[\psi_F(X) \mid \Lambda_{\mathcal{F}_{\text{sub}}}(\theta)]\}. \end{aligned}$$

Then

$$\sup_{\mathcal{F}_{\text{sub}}\colon \mathcal{F}_{\text{sub}} \in \mathcal{A}} C_{\mathcal{F}_{\text{sub}}}(F) \le \operatorname{var}_F\{\Pi_F[\psi_F(X) \mid \Lambda_{\mathcal{F}}(F)]\} < \infty.$$

Remark 4.5. The supremum of the C-R bounds over all regular parametric submodels in a class $\mathcal{A}$ for a scalar pathwise differentiable parameter w.r.t. $\mathcal{A}$ is finite and it is less than or equal to the variance of the projection of ANY gradient for the parameter into the tangent space of the model w.r.t. $\mathcal{A}$. The key point is that $\psi_{F^*}(X)$ is the same over all choices of the submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}$. Under pathwise differentiability, the existence of the C-R bound is ensured, but it remains unclear whether there exists an estimator achieving this bound (usually it is true).

Definition 4.4 (**Semiparametric C-R bounds**)**.** Given a class $\mathcal{A}$ of parametric submodels of a semiparametric model $\mathcal{F}$, the semiparametric C-R bound w.r.t. $\mathcal{A}$ of a regular $\mathbb{R}^k$-valued parameter $\beta(\cdot)$ at F with gradient $\psi_F(X)$ is defined as

$$C_{\mathcal{F}}(F) = \operatorname{var}_F\{\Pi_F[\psi_F(X) \mid \Lambda_{\mathcal{F}}(F)]\},$$

where $\Lambda_{\mathcal{F}}(F)$ is the tangent space of the model $\mathcal{F}$ w.r.t. the class $\mathcal{A}$ at F.

4.6 Examples: calculation of gradients and C-R bounds

Example 4.3 (Mean functional)**.** Let $\mathcal{F}$ be the nonparametric model for a random variable X restricted only by the condition that $\mathbb{E}_F[b(X)^2] < \infty$ for all $F \in \mathcal{F}$, where $b(\cdot)$ is a given real valued function. Let $\beta(F) = \mathbb{E}_F[b(X)]$. Let $\mathcal{A}$ be the class of all regular parametric submodels such that for each submodel, indexed by θ with

$F^* = F_{\theta^*}$, the map $\theta \to \mathbb{E}_\theta[b(X)^2]$ is continuous in an open neighborhood of θ^*. The setting is the same as in Example 4.2.

We will compute the gradient of $\beta(F)$ w.r.t. $\mathcal{A}$ at F^*. By Lemma 3.3, we have that for any submodel in the class $\mathcal{A}$ with score $S_\theta(\theta^*) = s_\theta(X;\theta^*)$,

$$\left.\frac{\partial \beta(F_\theta)}{\partial \theta^\top}\right|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}\left[b(X)S_\theta(\theta^*)^\top\right].$$

So we have that $b(X)$ is a gradient and $\psi_{F^*}(X) = b(X) - \mathbb{E}_{F^*}[b(X)]$ is a mean zero gradient.

Note that by Remark 4.4, the tangent space w.r.t. $\mathcal{A}$ is $\Lambda_{\mathcal{F}}(F^*) = \mathcal{L}_2^0(F^*)$. So the C-R bound is

$$C_{\mathcal{F}}(F^*) = \mathrm{var}_{F^*}\left\{\Pi_{F^*}\left[b(X) - \mathbb{E}_{F^*}[b(X)] \mid \mathcal{L}_2^0(F^*)\right]\right\} = \mathrm{var}_{F^*}[b(X)].$$

Remark 4.6. Suppose F^* is the $\mathcal{N}(\mu^*, \sigma^{*2})$ distribution and $b(X) = X$. The C-R bound for estimating $\beta(F) = \mathbb{E}_F(X)$ in the nonparametric model at F^* is $\mathrm{var}_{F^*}(X) = \sigma^{*2}$, which is the same as the C-R bound for estimating the mean under the parametric normal submodel

$$\mathcal{F}^*_{\mathrm{sub}} = \left\{\mathcal{N}(\mu, \sigma^2) \colon \mu \in \mathbb{R}, \sigma > 0\right\}.$$

Then, the normal submodel $\mathcal{F}^*_{\mathrm{sub}}$ is a least favorable submodel.

On the other hand, suppose F^* is the $\log\mathcal{N}(\mu^*, \sigma^{*2})$ distribution and $b(X) = X$. The C-R bound for estimating $\beta(F) = \mathbb{E}_F(X)$ in the nonparametric model at F^* is $\mathrm{var}_{F^*}(X) = \{\exp(\sigma^{*2}) - 1\}\exp(2\mu^{*2} + \sigma^{*2})$, which is strictly larger than the C-R bound for estimating the mean under the parametric log-normal submodel

$$\mathcal{F}^*_{\mathrm{sub}} = \left\{\log\mathcal{N}(\mu, \sigma^2) \colon \mu \in \mathbb{R}, \sigma > 0\right\}.$$

Then, the log-normal submodel $\mathcal{F}^*_{\mathrm{sub}}$ is not a least favorable submodel.

Example 4.4 (Median functional)**.** Let $\mathcal{F}$ be the nonparametric model for a random variable X absolutely continuous with respect to the Lebesgue measure, restricted only by the requirements that the density $f(x)$ be continuous and satisfy $f(\beta(F)) \neq 0$, where $\beta(F)$ is the median functional, i.e., $\beta(F)$ solves $F(\beta) = 1/2$. Let $\mathcal{A}$ be the class of all regular parametric submodels indexed by θ with $F^* = F_{\theta^*}$, and such that the map $(\theta, u) \to F_\theta(u)$ has continuous partial derivatives in an open neighborhood of (θ^*, u^*) with $u = \beta(F_{\theta^*})$.

The assumptions that the map $(\theta, u) \to F_\theta(u)$ is continuously differentiable and that $f(\beta(F)) \neq 0$ implies, by the implicit function theorem, that the map $\theta \to \beta(F_\theta)$ is differentiable at θ^* and it holds that

$$\begin{aligned}
0 &= \left.\frac{\mathrm{d}(1/2)}{\mathrm{d}\theta}\right|_{\theta=\theta^*} \\
&= \left.\frac{\mathrm{d}F_\theta(\beta(F_\theta))}{\mathrm{d}\theta}\right|_{\theta=\theta^*} \\
&= \left.\frac{\mathrm{d}F_\theta(\beta(F_{\theta^*}))}{\mathrm{d}\theta}\right|_{\theta=\theta^*} + \left.\frac{\mathrm{d}F_{\theta^*}(\beta)}{\mathrm{d}\beta}\right|_{\beta=\beta(\theta^*)} \left.\frac{\mathrm{d}\beta(F_\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta^*},
\end{aligned}$$

so

$$\left.\frac{\mathrm{d}\beta(\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta^*} = -\left\{\left.\frac{\mathrm{d}F_{\theta^*}(\beta)}{\mathrm{d}\beta}\right|_{\beta=\beta(\theta^*)}\right\}^{-1}\left.\frac{\mathrm{d}F_\theta(\beta(\theta^*))}{\mathrm{d}\theta}\right|_{\theta=\theta^*}.$$

Let $b(X) = \mathbb{I}(X \le \beta(F_{\theta^*}))$. By definition, $\mathbb{E}_\theta[b(X)] = F_\theta(\beta(F_{\theta^*}))$. Since the map $\theta \to \mathbb{E}_\theta[b(X)^2]$ is continuous, by Lemma 3.3,

$$\left.\frac{\mathrm{d}F_\theta(\beta(F_{\theta^*}))}{\mathrm{d}\theta}\right|_{\theta=\theta^*} = \left.\frac{\mathrm{d}\mathbb{E}_\theta[b(X)]}{\mathrm{d}\theta}\right|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}[b(X)S_\theta(\theta^*)].$$

Together with the fact that

$$\left.\frac{\mathrm{d}F_{\theta^*}(\beta)}{\mathrm{d}\beta}\right|_{\beta=\beta(F_{\theta^*})} = f^*(\beta(F_{\theta^*})),$$

we get

$$\left.\frac{\mathrm{d}\beta(\theta)}{\mathrm{d}\theta}\right|_{\theta=\theta^*} = -\mathbb{E}_{\theta^*}\left[\frac{\mathbb{I}(X \le \beta(F_{\theta^*}))}{f^*(\beta(F_{\theta^*}))}S_\theta(\theta^*)\right].$$

Therefore, $-\frac{\mathbb{I}(X\le\beta(F^*))}{f^*(\beta(F^*))}$ is a gradient at F^* and

$$\psi_{F^*}(X) = -\frac{\mathbb{I}(X \le \beta(F^*)) - 1/2}{f^*(\beta(F^*))}$$

is a mean zero gradient at F^*. This gradient $\psi_{F^*}(X)$ coincides with the influence function in Example 3.7.

Again, it is easy to use the first submodel in Remark 4.2 to show the tangent space w.r.t. $\mathcal{A}$ is $\mathcal{L}_2^0(F^*)$. Thus the semiparametric C-R bound is

$$\begin{aligned} C_{\mathcal{F}}(F^*) &= \mathrm{var}_{F^*}\left[\frac{\mathbb{I}(X \le \beta(F_{\theta^*})) - 1/2}{f^*(\beta(F_{\theta^*}))}\right] \\ &= \frac{\mathbb{E}_{F^*}[\mathbb{I}(X \le \beta(F^*))]\{1 - \mathbb{E}_{F^*}[\mathbb{I}(X \le \beta(F^*))]\}}{f^*(\beta(F_{\theta^*}))^2} \\ &= \frac{1/4}{f^*(\beta(F_{\theta^*}))^2}. \end{aligned}$$

Both the mean and median functional solve moment equations and the C-R bounds for them can be seen as special cases of the C-R bound for functionals that solve moment equations derived in the next example.

Example 4.5 (Moment equations)**.** Let $(X, \beta) \to u(X, \beta)$ be an $\mathbb{R}^k$-valued map where $\beta \in \mathbb{R}^k$ and X is a random vector. Consider the model $\mathcal{F}$ comprised of all laws F such that the equation in β

$$\mathbb{E}_F[u(X; \beta)] = 0$$

has a unique solution, denoted with $\beta(F)$ and such that

1. the map $\beta \to \mathbb{E}_F[u(X;\beta)]$ has continuous partial derivatives in an open neighborhood of $\beta(F)$,
2. $\left[\partial \mathbb{E}_F[u(X;\beta)]/\partial \beta^\top\right]\big|_{\beta=\beta(F)}$ is non-singular,
3. $\mathbb{E}_F\left[u(X;\beta(F))^\top u(X;\beta(F))\right] < \infty$.

Note that these assumptions are the same as in Proposition 3.3 (asymptotic linearity of Z-estimators). Setting $u(X;\beta) = X - \beta$ recovers Example 4.3, and setting $u(X;\beta) = 2\mathbb{I}(X \le \beta) - 1$ recovers Example 4.4.

Let $\mathcal{A}$ be the class of all regular parametric submodels through F^*, indexed by θ with $F^* = F_{\theta^*}$, such that

(a) the map $\theta \to \mathbb{E}_\theta\{u(X;\beta(\theta^*))^\top u(X;\beta(\theta^*))\}$ is continuous at θ^*,

(b) the map $(\beta,\theta) \to \mathbb{E}_\theta\{u(X;\beta)\}$ has continuous partial derivatives in an open neighborhood of $(\beta(\theta^*),\theta^*)$.

Note that these assumptions are the same as in Proposition 3.4 (regularity of Z-estimators). Under condition (b) and the assumption of non-singularity of the Jacobian matrix $\left[\partial \mathbb{E}_F[u(X;\beta)]/\partial \beta^\top\right]\big|_{\beta=\beta(F)}$, the implicit function theorem implies that in a neighborhood of θ^*, the map $\theta \to \beta(\theta) = \beta(F_\theta)$ is differentiable and

$$\begin{aligned} 0 &= \left.\frac{\partial \mathbb{E}_\theta[u(X;\beta(\theta))]}{\partial \theta^\top}\right|_{\theta=\theta^*} \\ &= \left.\frac{\partial \mathbb{E}_\theta[u(X;\beta(\theta^*))]}{\partial \theta^\top}\right|_{\theta=\theta^*} + \left.\frac{\partial \mathbb{E}_{\theta^*}[u(X;\beta)]}{\partial \beta^\top}\right|_{\beta=\beta(\theta^*)} \left.\frac{\partial \beta(\theta)}{\partial \theta^\top}\right|_{\theta=\theta^*}. \end{aligned}$$

Under the condition (a), by Lemma 3.3 we have

$$\left.\frac{\partial \mathbb{E}_\theta[u(X;\beta(\theta^*))]}{\partial \theta^\top}\right|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}[u(X;\beta(\theta^*))S_\theta(\theta^*)^\top],$$

from which we conclude that

$$\left.\frac{\partial \beta(\theta)}{\partial \theta^\top}\right|_{\theta=\theta^*} = -\left\{\left.\frac{\partial \mathbb{E}_{\theta^*}[u(X;\beta)]}{\partial \beta^\top}\right|_{\beta=\beta(\theta^*)}\right\}^{-1} \mathbb{E}_{\theta^*}[u(X;\beta(\theta^*))S_\theta(\theta^*)^\top].$$

Therefore,

$$\psi_{F^*}(X) = -\left\{\left.\frac{\partial \mathbb{E}_{F^*}[u(X;\beta)]}{\partial \beta^\top}\right|_{\beta=\beta(F^*)}\right\}^{-1} u(X;\beta(F^*)) \tag{4.1}$$

is a mean zero gradient w.r.t. $\mathcal{A}$ at F^*. The tangent space w.r.t. $\mathcal{A}$ is $\mathcal{L}_2^0(F^*)$, so the semiparametric C-R bound is

$$\begin{aligned} C_{\mathcal{F}}(F^*) &= \mathrm{var}_{F^*}\{\psi_{F^*}(X)\} \\ &= \left\{\left.\frac{\partial \mathbb{E}_{F^*}[u(X;\beta)]}{\partial \beta^\top}\right|_{\beta=\beta(F^*)}\right\}^{-1} \mathrm{var}_{F^*}[u(X;\beta(F^*))] \left\{\left.\frac{\partial \mathbb{E}_{F^*}[u(X;\beta)]}{\partial \beta}\right|_{\beta=\beta(F^*)}\right\}^{-1}. \end{aligned}$$

Remark 4.7. Note that there are differences between the mean zero gradient and the influence function. The mean zero gradient is an intrinsic quantity of a pathwise differentiable parameter $\beta(F)$ in a semiparametric model $\mathcal{F}$ (w.r.t. a class $\mathcal{A}$ of regular parametric submodels at F), while the influence function is an intrinsic quantity of a RAL estimator $\widehat{\beta}_n$ in a regular parametric model (usually, a submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}$).

Remark 4.8. By Proposition 3.3, Proposition 3.4 and Example 4.5, we know that the RAL estimator $\widehat{\beta}_n$ solving the moment equation (3.2) achieves the semiparametric C-R bound, and its influence function coincides with the mean zero gradient $\psi_{F^*}(X)$ in (4.1) of β.

Remark 4.9. Example 4.5 explains why the Polyak-Ruppert averaged estimator [Polyak and Juditsky, 1992] in SGD achieves the C-R bound. Let $x^*\colon F \to \mathbb{R}^k$ be a functional solving

$$\mathbb{E}_F \nabla f(x^*(F), \xi) = 0,$$

where $\xi \sim F$. Suppose the true distribution of ξ is F^* and $x^*(F^*) = x^*$. Consider the nonparametric model $\mathcal{F}$ satisfying all regularity conditions in Example 4.5 and the class $\mathcal{A}$ of all regular parametric submodels through F^*, indexed by θ with $F^* = F_{\theta^*}$ and satisfying all regularity conditions in Example 4.5. Then, the semiparametric C-R bound is

$$C_{\mathcal{F}}(F^*) = \left\{\nabla^2 f(x^*)\right\}^{-1} \operatorname{var}_{F^*}\{\nabla f(x^*, \xi)\} \left\{\nabla^2 f(x^*)\right\}^{-1},$$

which is achieved by the Polyak-Ruppert averaged estimator.

4.7 Representation of the set of gradients

The following lemma states that, except when the tangent set for the model is $\mathcal{L}_2^0(F)$, there are infinitely many mean zero gradients.

Lemma 4.1. *Let $\beta(\cdot)$ be an $\mathbb{R}^k$-valued pathwise differentiable parameter in the model $\mathcal{F}$ at F and let $\psi_F(X)$ be a gradient of $\beta(\cdot)$ at F w.r.t. $\mathcal{A}$. Then*

$$\psi'_F(X) = \psi_F(X) + b_F(X)$$

is a gradient of $\beta(F)$ at F w.r.t. $\mathcal{A}$ iff

$$b_F(X) \in \Lambda_{\mathcal{F}}(F)^{\perp}.$$

Remark 4.10. We denote the set of all mean zero gradients at F for a pathwise differentiable parameter $\beta(F)$ in the model $\mathcal{F}$ as $\mathrm{IF}^0_{\mathcal{F}}(F)$. From Lemma 4.1 we know that $\mathrm{IF}^0_{\mathcal{F}}(F) = \{\psi_F(X)\} + \Lambda_{\mathcal{F}}(F)^{\perp}$. In the following (Section 4.8) we will show that all influence functions of RAL estimators are in $\mathrm{IF}^0_{\mathcal{F}}(F)$. However, there may exists a mean zero gradient that is not an influence function of any RAL estimator. In other words, {all influence functions of RAL estimators} $\subseteq \mathrm{IF}^0_{\mathcal{F}}(F)$.

Definition 4.5 (**Efficient influence functions**)**.** The projection of any gradient $\psi_F(X)$ into the tangent space $\Lambda_{\mathcal{F}}(F)$ for the model $\mathcal{F}$ w.r.t. $\mathcal{A}$ is called the **efficient influence function** for the parameter $\beta(F)$ in the semiparametric model $\mathcal{F}$ w.r.t. $\mathcal{A}$ at F. That is,

$$\psi_{F,\text{eff}}(X) = \Pi_F[\psi_F(X) \mid \Lambda_{\mathcal{F}}(F)].$$

Lemma 4.2. *$\psi_{F,eff}(X)$ is a gradient.*

Corollary 4.1. $\text{IF}^0_{\mathcal{F}}(F) = \{\psi_{F,eff}(X)\} \oplus \Lambda_{\mathcal{F}}(F)^{\perp}$. *In the nonparametric model the maximal tangent space is $\Lambda_{\mathcal{F}}(F) = \mathcal{L}^0_2(F)$, so in this case $\text{IF}^0_{\mathcal{F}}(F) = \{\psi_{F,eff}(X)\}$.*

Remark 4.11. Consider two classes of regular parametric submodels $\mathcal{A}$ and $\mathcal{A}'$ with the same tangent space. To ensure the sets of their gradients coincide, one needs that the map from $\mathcal{L}_2(\mu)$ to $\mathbb{R}$:

$$\sqrt{f} \to \beta(\sqrt{f}),$$

is Frechet differentiable, i.e., there exists a continuous linear map $\dot{B}\colon \mathcal{L}_2(F^*) \to \mathbb{R}$ such that

$$\beta(\sqrt{f}) - \beta(\sqrt{f^*}) = \dot{B}(\sqrt{f} - \sqrt{f^*}) + o\left(\left\|\sqrt{f} - \sqrt{f^*}\right\|^2_{\mathcal{L}_2(F^*)}\right).$$

This is stronger than pathwise differentiability. In the following, we will make this assumption unless particularly mentioned.

4.8 Representation of the set of influence functions of RAL estimators

The following lemma states that the set of influence functions of RAL estimators for a pathwise differentiable parameter is included in the set of mean zero gradients for the parameter.

Lemma 4.3. [5] *Let $\mathcal{F}$ be a semiparametric model. Suppose that $\widehat{\beta}_n$ is an asymptotically linear estimator at F of a parameter $\beta\colon \mathcal{F} \to \mathbb{R}^k$, with influence function $\varphi_F(X)$. Suppose that for every regular parametric submodel in a class $\mathcal{A}$ indexed by θ that goes through F at θ^*, $\beta(\theta)$ is differentiable at θ^*. Then $\widehat{\beta}_n$ is a regular estimator if and only if for all regular parametric submodels in $\mathcal{A}$ with score $S_\theta(\theta^*)$ at θ^*,*

$$\left.\frac{\partial \beta(\theta)}{\partial \theta^\top}\right|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}\left[\varphi_F(X) S_\theta(\theta^*)^\top\right].$$

Remark 4.12. The above lemma has several implications:

1. for any RAL estimator $\widehat{\beta}_n$, the asymptotic variance of $\sqrt{n}\{\widehat{\beta}_n - \beta(F)\}$ is no less than $\text{var}_F\{\psi_{F,\text{eff}}(X)\}$, since $\varphi_F(X)$ is a gradient by the lemma and projection contracts the variance;

2. since in the nonparametric model there is a unique mean zero gradient $\psi_{F,\text{eff}}(X)$, all RAL estimators w.r.t. $\mathcal{A}$ have the same influence function, and thus are asymptotically equivalent.

[5] Lemma 25.23 of [Van der Vaart, 2000].

4.9 The convolution theorem

This section is a semiparametric analogy of Section 3.6.

Proposition 4.1 (**The convolution theorem**). [6] *Let $\mathcal{F}$ be a semiparametric model. Let $\beta(\cdot)\colon \mathcal{F} \to \mathbb{R}^k$ be a pathwise differentiable parameter at F with efficient influence function $\psi_{F,eff}(X)$ w.r.t. a class $\mathcal{A}$. If $\widehat{\beta}_n$ is regular at F and the tangent set w.r.t $\mathcal{A}$ at F is convex, then*

$$\sqrt{n}\{\widehat{\beta}_n - \beta(F)\} \xrightarrow{D(F)} U + U^*,$$

where $U^* \sim \mathcal{N}_k(0, C_{\mathcal{F}}(F))$, $C_{\mathcal{F}}(F) = \mathrm{var}_F\{\psi_{F,eff}(X)\}$ and U is independent of U^*.

Definition 4.6 (**Asymptotically efficient estimators**). Locally/Globally asymptotically efficient estimators are defined as follows.

1. Let $\beta(F)$ be a parameter such that at each $F \in \mathcal{F}$ it is pathwise differentiable w.r.t. to a class $\mathcal{A}_F$. An estimator $\widehat{\beta}_n$ is a **locally asymptotically efficient** estimator of $\beta(F)$ in $\mathcal{F}$ at F^* w.r.t. $\mathcal{A}_{F^*}$, if

   (a) it is a regular estimator of $\beta(\cdot)$ in $\mathcal{F}$ at every $F \in \mathcal{F}$ w.r.t. $\mathcal{A}_F$,

   (b) it satisfies
   $$\sqrt{n}\{\widehat{\beta}_n - \beta(F^*)\} \xrightarrow{D(F^*)} \mathcal{N}_k(0, C_{\mathcal{F}}(F^*)),$$
   where $C_{\mathcal{F}}(F^*)$ is the semiparametric C-R bound.

2. Let $\mathcal{F}^* \subseteq \mathcal{F}$. An estimator $\widehat{\beta}_n$ is a **locally asymptotically efficient** estimator of $\beta(F)$ in $\mathcal{F}$ at $\mathcal{F}^*$ w.r.t. $\{\mathcal{A}_F\colon F \in \mathcal{F}^*\}$, if it is regular and asymptotically efficient in $\mathcal{F}$ at F w.r.t. $\mathcal{A}_F$ for every $F \in \mathcal{F}^*$.

3. If $\mathcal{F}^* = \mathcal{F}$, then $\widehat{\beta}_n$ is called **globally asymptotically efficient** (e.g., sample mean).

Locally, much less globally, efficient estimators of a pathwise differentiable parameter do not always exist. However, if an efficient estimator exists, then the next lemma states that it must be RAL and have influence function equal to the efficient influence function.

Note that this result states, just as in the parametric case, that as far as efficiency is concerned, we don't lose anything by restricting attention to regular and asymptotically linear estimators.

Lemma 4.4. *Suppose that $\mathcal{F}$ is a semiparametric model and $\beta(\cdot)\colon \mathcal{F} \to \mathbb{R}^k$, is a pathwise differentiable parameter at F with efficient influence function $\psi_{F,eff}(X)$ w.r.t. $\mathcal{A}$ with a convex tangent set. The estimator $\widehat{\beta}_n$ is locally asymptotically efficient at F w.r.t. $\mathcal{A}$ if and only if*

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

[6] Theorem 25.20 of [Van der Vaart, 2000].

5 Semiparametric Models

In this section we will consider semiparametric models that can be represented using the class of densities

$$\mathcal{M}_{F_{\theta^*}} = \{f(X;\theta)\colon \beta \in \mathbb{R}^s, \eta \in \Delta\},$$

where $\theta = (\beta, \eta)$, and the parameter of interest β is variational independent of the infinite dimensional parameter η.

5.1 Semiparametric restricted mean model

Example 5.1 (Restricted mean model)**.** Consider the restricted mean model (RMM)

$$\begin{aligned} Y &= m(Z;\beta) + \varepsilon, \\ \mathbb{E}\{\varepsilon \mid Z\} &= 0, \end{aligned}$$

where $m(Z;\beta)$ is a known function of Z up to β. The likelihood for one observation is

$$\begin{aligned} f_{Y,Z}(y,z;\beta,\eta) &= f_{\varepsilon,Z}(y - m(Z;\beta), z; \beta, \eta) \\ &= \eta_1(\varepsilon, z)\eta_2(z), \end{aligned}$$

where $\eta_1(\varepsilon, z)$ is the conditional density of ε given z and $\eta_2(z)$ is the density of z. They are restricted only by

$$\begin{aligned} \int \eta_1(\varepsilon, z)\mathrm{d}\mu(\varepsilon) &= 1, \\ \int \varepsilon\eta_1(\varepsilon, z)\mathrm{d}\mu(\varepsilon) &= 0, \\ \int \eta_2(z)\mathrm{d}\mu(z) &= 1. \end{aligned}$$

Result 5.1. The following four results hold.

1. The nuisance tangent space of RMM is given by

$$\begin{aligned} \Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2} \subset \mathcal{L}_2^0, \\ \Lambda_{\text{nuis},1} &= \left\{a_1(\varepsilon, Z)\colon \mathbb{E}\{a_1(\varepsilon, Z) \mid Z\} = 0,\ \mathbb{E}\{\varepsilon a_1(\varepsilon, Z)^\top \mid Z\} = 0\right\} \cap \mathcal{L}_2^0, \\ \Lambda_{\text{nuis},2} &= \{a_2(Z)\colon \mathbb{E}\{a_2(Z)\} = 0\} \cap \mathcal{L}_2^0. \end{aligned}$$

2. The orthogonal complement to the nuisance tangent space is

$$\begin{aligned} \Lambda_{\text{nuis}}^\perp &= \Lambda_{\text{nuis},1}^\perp \cap \Lambda_{\text{nuis},2}^\perp \cap \mathcal{L}_2^0 \\ &= \{h(Z)\varepsilon(\beta^*)\colon h\} \cap \mathcal{L}_2^0. \end{aligned}$$

3. The set of influence functions is

$$\left\{\mathbb{E}\left\{h(Z)\varepsilon(\beta^*)S_\beta^\top\right\}^{-1} h(Z)\varepsilon(\beta^*)\colon h\right\} \cap \mathcal{L}_2^0,$$

where S_β is the score of β under the RMM.

4. The efficient score for β is

$$\begin{aligned} S_{\text{eff}}(Z;\beta^*,\eta^*) &= \Pi[S_\beta \mid \Lambda_{\text{nuis}}^\perp] = h^{\text{opt}}(Z)\varepsilon(\beta^*), \\ h^{\text{opt}}(Z) &= \mathbb{E}\left\{S_\beta\varepsilon(\beta^*)^\top \mid Z\right\}\mathbb{E}\left\{\varepsilon(\beta^*)\varepsilon(\beta^*)^\top \mid Z\right\}^{-1} \end{aligned}$$

Proof of Result 5.1. We prove the above four results.

1. Consider a regular parametric submodel $\eta_t(\varepsilon, z) = \eta_{1,t}(\varepsilon, z)\eta_{2,t}(z)$ with unknown parameter $t \in (-\varepsilon, \varepsilon)$ such that $\eta_{t=0}(\varepsilon, z) = f(\varepsilon, z)$. It is easy to verify that

$$\frac{\partial \log \eta_t(\varepsilon, z)}{\partial t} = a_1(\varepsilon, z) + a_2(z),$$

where $\mathbb{E}\{a_1(\varepsilon, Z) \mid Z\} = \mathbb{E}\{a_2(Z)\} = 0$ and thus $\mathbb{E}\{a_1(\varepsilon, Z)a_2(Z)\} = 0$. The first term a_1 is a score for the conditional density of $\varepsilon \mid Z$, while the second term a_2 is a score for the density of Z. Furthermore, we have for all $t \in (-\varepsilon, \varepsilon)$,

$$\int \eta_{1,t}(\varepsilon, z)\varepsilon^\top \mathrm{d}\mu(\varepsilon) = 0,$$

so

$$\int \left.\frac{\partial \eta_{1,t}(\varepsilon, z)}{\partial t}\right|_{t=0} \varepsilon^\top \mathrm{d}\mu(\varepsilon) = 0,$$

which implies $\mathbb{E}\{a_1(\varepsilon, Z)\varepsilon^\top \mid Z\} = 0$. From these results we have

$$\begin{aligned} \Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2} \subset \mathcal{L}_2^0, \\ \Lambda_{\text{nuis},1} &= \left\{a_1(\varepsilon, Z)\colon \mathbb{E}\{a_1(\varepsilon, Z) \mid Z\} = 0,\ \mathbb{E}\{\varepsilon a_1(\varepsilon, Z)^\top \mid Z\} = 0\right\} \cap \mathcal{L}_2^0, \\ \Lambda_{\text{nuis},2} &= \{a_2(Z)\colon \mathbb{E}\{a_2(Z)\} = 0\} \cap \mathcal{L}_2^0. \end{aligned}$$

Technically, for each element of Λ_{nuis} we must also exhibit parametric submodels with corresponding score equal to the given element. This is easy as we can apply two useful submodels in Remark 4.2. See Section 4.5 of [Tsiatis, 2006].

2. The orthogonal complement to the nuisance tangent space is

$$\Lambda_{\text{nuis}}^\perp = (\Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2})^\perp = \Lambda_{\text{nuis},1}^\perp \cap \Lambda_{\text{nuis},2}^\perp \cap \mathcal{L}_2^0.$$

Note that any function $b(\varepsilon, Z) \in \Lambda_{\text{nuis},2}^\perp \cap \mathcal{L}_2^0$ must satisfy

$$\mathbb{E}(b(\varepsilon, Z)a_2(Z)) = 0$$

for all $a_2(Z) \in \Lambda_{\text{nuis},2} \cap \mathcal{L}_2^0$, so $\mathbb{E}(b(\varepsilon, Z) \mid Z) = 0$. Therefore, to characterize $\Lambda_{\text{nuis}}^\perp$ entails finding all $b(\varepsilon, Z) \in \Lambda_{\text{nuis},2} \cap \mathcal{L}_2^0$ such that $\mathbb{E}(b(\varepsilon, Z)a_1(\varepsilon, Z)) = 0$

for all $a_1(\varepsilon, Z) \in \Lambda_{\text{nuis},1}$. Since $a_1(\varepsilon, Z) \in \Lambda_{\text{nuis},1}$ implies $\mathbb{E}\{a_1(\varepsilon, Z)\varepsilon \mid Z\} = 0$, a natural candidate for $\Lambda_{\text{nuis}}^{\perp}$ is

$$\{b(\varepsilon, Z) = h(Z)\varepsilon \colon h\} \cap \mathcal{L}_2^0.$$

Note that the subspace of $\Lambda_{\text{nuis},2}^{\perp}$ orthogonal to ε given Z is given by

$$\left\{c(\varepsilon, Z) - \mathbb{E}(C\varepsilon^\top \mid Z)\mathbb{E}(\varepsilon\varepsilon^\top \mid Z)^{-1}\varepsilon \colon \mathbb{E}(C \mid Z) = 0\right\} \cap \mathcal{L}_2^0,$$

and this is also an alternative characterization of $\Lambda_{\text{nuis},1}$, thus concluding the claim.

3. An influence function $\varphi_F(\varepsilon, Z)$ must satisfy $\Pi[\varphi_F(\varepsilon, Z) \mid \Lambda_{\text{nuis}}] = 0$ as well as $\mathbb{E}[\varphi_F(\varepsilon, Z)S_\beta^\top] = I_{\dim(\beta)}$, so it has the form

$$\varphi_F(\varepsilon, Z) = \mathbb{E}\{h(Z)\varepsilon(\beta^*)S_\beta^\top\}^{-1}h(Z)\varepsilon(\beta^*).$$

4. This follows directly from the definition of the projection,

$$\begin{aligned} S_{\text{eff}}(Z; \beta^*, \eta^*) &= \Pi(S_\beta \mid \Lambda_{\text{nuis}}^{\perp}) \\ &= \Pi\left(S_\beta \mid \{b(\varepsilon, Z) = h(Z)\varepsilon \colon h\} \cap \mathcal{L}_2^0\right) \\ &= \mathbb{E}\{S_\beta\varepsilon(\beta^*)^\top \mid Z\}\mathbb{E}\{\varepsilon(\beta^*)\varepsilon(\beta^*)^\top \mid Z\}^{-1}\varepsilon(\beta^*). \end{aligned}$$

□

Remark 5.1. Note that for all submodels,

$$\int \eta_1(\varepsilon(\beta), z)\varepsilon(\beta)^\top \mathrm{d}\mu(\varepsilon) = 0$$

for all β. Therefore,

$$0 = \int \frac{\partial \eta_1(\varepsilon(\beta), z)}{\partial \beta}\bigg|_{\beta=\beta^*} \varepsilon(\beta^*)^\top \mathrm{d}\mu(\varepsilon) - \int \eta_1(\varepsilon(\beta^*), z)\frac{\partial m(Z; \beta)^\top}{\partial \beta}\bigg|_{\beta=\beta^*} \mathrm{d}\mu(\varepsilon),$$

which implies that

$$\mathbb{E}\{S_\beta\varepsilon(\beta^*)^\top \mid Z\} = \frac{\partial m(Z; \beta)^\top}{\partial \beta}\bigg|_{\beta=\beta^*}.$$

We conclude that

$$S_{\text{eff}}(Z; \beta^*, \eta^*) = D(Z)^\top V(Z)^{-1}\varepsilon(\beta^*),$$

where

$$D(Z) = \frac{\partial m(Z; \beta)}{\partial \beta}\bigg|_{\beta=\beta^*}, \quad V(Z) = \mathbb{E}\{\varepsilon(\beta^*)\varepsilon(\beta^*)^\top \mid Z\}.$$

Also, the efficient influence function is

$$\varphi_{F,\text{eff}}(\varepsilon, Z) = \mathbb{E}\{D(Z)^\top V(Z)^{-1}D(Z)\}^{-1}D(Z)^\top V(Z)^{-1}\varepsilon,$$

and the semiparametric C-R bound is given by

$$\mathbb{E}\{D(Z)^\top V(Z)^{-1}D(Z)\}^{-1}.$$

Adaptive semiparametric estimators The class of influence functions is particularly useful as it provides very good candidate estimating equations. Specifically, let

$$u_h(X;\beta) = U_h(\beta) = h(Z)(Y - m(Z;\beta)) = h(Z)\varepsilon(\beta)$$

with $h \in \mathbb{R}^{\dim(\beta)\times\dim(\varepsilon)}$. Note that $\{U_h(\beta)\colon h\} \cap \mathcal{L}_2^0 = \Lambda_{\text{nuis}}^{\perp}$. As shown in Example 4.5, under weak regularity conditions, the solution to the equation

$$\sum_{i=1}^{n} U_{h,i}(\widehat{\beta}_h) = 0. \tag{5.1}$$

for a user-specified h is a RAL estimator with influence function given by

$$\mathbb{E}\{h(Z)D(Z)\}^{-1}h(Z)\varepsilon(\beta).$$

The estimating equations (5.1) is an example of the so-called generalized estimating equations. It is reasonable to restrict attention to the semiparametric efficient estimator which is given by

$$\begin{aligned} h^{\text{opt}}(Z) &= D(Z)^{\top}V(Z)^{-1}\varepsilon(\beta), \\ D(Z) &= \left.\frac{\partial m(Z;\beta)}{\partial\beta^{\top}}\right|_{\beta=\beta^*}, \\ V(Z) &= \mathbb{E}\left\{\varepsilon(\beta^*)\varepsilon(\beta^*)^{\top} \mid Z\right\}. \end{aligned}$$

This estimator is infeasible as it requires $D(Z)$ and $V(Z) = \text{var}(Y \mid Z)$. The dependence on β^* of $D(Z)$ is not much an issue, since we can substitute β^* with a primary estimator $\widehat{\beta}_I$ by solving the GEE with a user-specified choice of $h_I(Z)$. The dependence on $\text{var}(Y \mid Z)$ is more challenging. One way is to estimate it nonparametrically (e.g. kernel smoothing), but such methods are unstable particularly if Z is multi-dimensional. An alternative approach is to specify a working model for $V(Z)$, say $V(Z;\beta,\gamma)$ with a finite dimensional parameter γ. It may be estimated by solving the estimating equation

$$0 = \sum_{i=1}^{n} T(Z_i,\widehat{\beta}_I,\gamma)\left\{\left(Y_i - m(Z;\widehat{\beta}_I)\right)^2 - V(Z;\widehat{\beta}_I,\gamma)\right\}.$$

Replacing $V(Z;\widehat{\beta}_I,\widehat{\gamma})$ for $V(Z)$ into $h^{\text{opt}}(Z)$ will result in a RAL estimator for β with influence function

$$\mathbb{E}\left\{D(Z)^{\top}V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}^{-1}D(Z)^{\top}V(Z;\beta^*,\gamma^*)^{-1}\varepsilon,$$

where γ^* is the probability limit of $\widehat{\gamma}$. Consequently, solutions to such estimating equations where one uses a working variance model is called a locally efficient estimator, because the resulting estimator is RAL under the semiparametric model irrespective of whether the variance model is correct or not, but it only achieves the semiparametric efficiency bound if the working variance model is correctly specified.

Remark 5.2. One must take care when estimating the asymptotic variance of the locally efficient estimator. Its asymptotic variance is given by the variance of its influence function, and equals to

$$\begin{aligned}&\mathbb{E}\left\{D(Z)^\top V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}^{-1}\cdot\\&\mathbb{E}\left\{D(Z)^\top V(Z;\beta^*,\gamma^*)^{-1}V(Z)V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}\cdot\\&\mathbb{E}\left\{D(Z)^\top V(Z;\beta^*,\gamma^*)^{-1}D(Z)\right\}^{-1}.\end{aligned}$$

This simplifies to the semiparametric C-R bound only if the variance model is correct.

5.2 Semiparametric location-shift model

Example 5.2 (Location-shift model)**.** Suppose we observe i.i.d. data (Y, Z). Y is continuous that follows a semiparametric location-shift model (LSM)

$$Y = m(Z;\beta^*) + \varepsilon,$$

where $m(Z;\cdot)$ is known, $\beta^* \in \mathbb{R}^s$ and $\varepsilon \perp Z$, i.e.,

$$\eta(\varepsilon, Z) = \eta_1(\varepsilon)\eta_2(Z),$$

with unrestricted η_1 and η_2. We aim to make inference about the finite dimensional parameter β^*. For identification, we assume $m(0;\beta^*) = m(Z;0) = 0$.

Result 5.2. The following four results hold.

1. The nuisance tangent space of LSM is given by

$$\begin{aligned}\Lambda_{\text{nuis}} &= \Lambda_{\text{nuis},1} \oplus \Lambda_{\text{nuis},2} \subset \mathcal{L}_2^0,\\ \Lambda_{\text{nuis},1} &= \{a_1(\varepsilon)\colon \mathbb{E}\{a_1(\varepsilon)\} = 0\} \cap \mathcal{L}_2^0,\\ \Lambda_{\text{nuis},2} &= \{a_2(Z)\colon \mathbb{E}\{a_2(Z)\} = 0\} \cap \mathcal{L}_2^0.\end{aligned}$$

2. The orthogonal complement to the nuisance tangent space is

$$\begin{aligned}\Lambda_{\text{nuis}}^\perp &= \Lambda_{\text{nuis},1}^\perp \cap \Lambda_{\text{nuis},2}^\perp \cap \mathcal{L}_2^0\\ &= \{h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon) \mid Z\} - \mathbb{E}\{h(Z,\varepsilon) \mid \varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\}\colon h\} \cap \mathcal{L}_2^0.\end{aligned}$$

3. the set of influence function is

$$\left\{ \begin{aligned} &\mathbb{E}\left\{\tilde{h}(Z,\varepsilon)S_\beta^\top\right\}^{-1}\tilde{h}(Z,\varepsilon),\\ &\tilde{h}(Z,\varepsilon) = h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon) \mid Z\} - \mathbb{E}\{h(Z,\varepsilon) \mid \varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \end{aligned} : h \right\} \cap L_2^0,$$

   where S_β is the score of β under the LSM.

4. The efficient score for β is

$$\begin{aligned} S_{\text{eff}}(X;\beta^*,\eta^*) &= \Pi\left(S_\beta \mid \Lambda_{\text{nuis}}^\perp\right) = \tilde{h}^{\text{opt}}(Z,\varepsilon), \\ h^{\text{opt}}(Z,\varepsilon) &= \dot{m}(Z;\beta^*)\kappa(\varepsilon), \\ \dot{m}(Z;\beta^*) &= \left.\frac{\partial m(Z;\beta^*)}{\partial\beta}\right|_{\beta=\beta^*} \\ \kappa(\varepsilon) &= -\frac{\partial \log \eta_1(\varepsilon)}{\partial\varepsilon}. \end{aligned}$$

Because $\mathbb{E}\kappa(\varepsilon) = 0$, we have that

$$S_{\text{eff}}(Z;\beta^*,\eta^*) = \{\dot{m}(Z;\beta^*) - \mathbb{E}\dot{m}(Z;\beta^*)\}\,\kappa(\varepsilon).$$

Proof of Result 5.2. We prove the above four results.

1. Consider a regular parametric submodel $\eta_t(\varepsilon,z) = \eta_{1,t}(\varepsilon)\eta_{2,t}(z)$ with unknown parameter $t \in (-\varepsilon,\varepsilon)$ such that $\eta_{t=0}(\varepsilon,z) = f(\varepsilon,z)$. It is easy to verify that

$$\frac{\partial \log \eta_t(\varepsilon,z)}{\partial t} = a_1(\varepsilon) + a_2(z),$$

where $\mathbb{E}\{a_1(\varepsilon)\} = \mathbb{E}\{a_2(Z)\} = 0$ and $\mathbb{E}\{a_1(\varepsilon)a_2(Z)\} = 0$ by independence.

2. For any $b(Z,\varepsilon) \in \Lambda_{\text{nuis}}^\perp$, it must satisfy $\mathbb{E}(b(Z,\varepsilon)a_1(\varepsilon)) = \mathbb{E}(b(Z,\varepsilon)a_2(Z)) = 0$. We conjecture that

$$\Lambda_{\text{nuis}}^\perp = \{h(Z,\varepsilon) - \mathbb{E}\{h(Z,\varepsilon) \mid Z\} - \mathbb{E}\{h(Z,\varepsilon) \mid \varepsilon\} + \mathbb{E}\{h(Z,\varepsilon)\} \colon h\} \cap \mathcal{L}_2^0.$$

It is easy to verify this conjecture.

3. This follows from the fact that an influence function $\varphi_F(Z,\varepsilon)$ must satisfy $\Pi[\varphi_F(\varepsilon,Z) \mid \Lambda_{\text{nuis}}] = 0$ as well as $\mathbb{E}[\varphi_F(\varepsilon,Z)S_\beta^\top] = I_{\dim(\beta)}$.

4. We have

$$S_{\text{eff}}(X;\beta^*,\eta^*) = \Pi[S_\beta \mid \Lambda_{\text{nuis}}^\perp] = S_\beta - \mathbb{E}\{S_\beta \mid Z\} - \mathbb{E}\{S_\beta \mid \varepsilon\} + \mathbb{E}\{S_\beta\}.$$

Note that

$$\begin{aligned} S_\beta &= \frac{\partial \log \eta_1(\varepsilon(\beta))}{\partial\beta} \\ &= -\frac{\partial \log \eta_1(\varepsilon(\beta))}{\partial\varepsilon}\dot{m}(Z;\beta^*) \\ &= \dot{m}(Z;\beta^*)\kappa(\varepsilon), \end{aligned}$$

so we conclude that

$$\begin{aligned} S_{\text{eff}}(X;\beta^*,\eta^*) &= \dot{m}(Z;\beta^*)\kappa(\varepsilon) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\kappa(\varepsilon) \mid Z\right\} \\ &\quad - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\kappa(\varepsilon) \mid \varepsilon\right\} + \mathbb{E}\left\{\dot{m}(Z;\beta^*)\kappa(\varepsilon)\right\} \\ &= \left\{\dot{m}(Z;\beta^*) - \mathbb{E}\left\{\dot{m}(Z;\beta^*)\right\}\right\}\kappa(\varepsilon) \end{aligned}$$

by independence and the fact that $\mathbb{E}\{\kappa(\varepsilon)\} = 0$.

□

Remark 5.3. Note that $\mathbb{E}_\beta\{\tilde{h}(Z,\varepsilon(\beta))\} = 0$ for all β, taking differentiation yields

$$\mathbb{E}\left\{\tilde{h}(Z,\varepsilon)S_\beta^\top\right\} = -\mathbb{E}\left\{\frac{\partial\tilde{h}(Z,\varepsilon)}{\partial\beta^\top}\right\}.$$

Thus the influence functions are of the form

$$-\mathbb{E}\left\{\frac{\partial\tilde{h}(Z,\varepsilon)}{\partial\beta^\top}\right\}\tilde{h}(Z,\varepsilon).$$

The efficient influence function is given by

$$\varphi_{F,\text{eff}} = \mathbb{E}\left\{\{\dot{m}(Z;\beta^*) - \mathbb{E}\{\dot{m}(Z;\beta^*)\}\}^{\otimes 2}\kappa(\varepsilon)^2\right\}^{-1} \cdot \{\dot{m}(Z;\beta^*) - \mathbb{E}\{\dot{m}(Z;\beta^*)\}\}\kappa(\varepsilon),$$

and the semiparametric C-R bound is given by

$$\mathbb{E}\left\{\kappa(\varepsilon)^2\right\}^{-1}\left[\mathbb{E}\left\{\{\dot{m}(Z;\beta^*) - \mathbb{E}\{\dot{m}(Z;\beta^*)\}\}^{\otimes 2}\right\}\right]^{-1}.$$

Adaptive semiparametric estimators Note that constructing a semiparametric estimator based on the efficient score requires knowing $\kappa(\varepsilon)$. Consider a feasible semiparametric efficient estimator $\widehat{\beta}_{\widehat{h}^{\text{opt}}}$ which solves

$$\mathbb{P}_n\{\dot{m}(Z;\beta) - \mathbb{P}_n\{\dot{m}(Z;\beta)\}\}\widehat{\kappa}(Y - m(Z;\beta)) = 0,$$

where we have replaced $\mathbb{E}\{\dot{m}(Z;\beta)\}$ with its sample version $\mathbb{P}_n\{\dot{m}(Z;\beta)\}$ and κ with an estimator $\widehat{\kappa}$. One way is to nonparametrically estimate $\kappa(\varepsilon)$ (or $\eta_1(\varepsilon)$), which is likely to be very unstable unless sample size is very large. Another way is to specify a working model for $\eta_1(\varepsilon)$, say $\eta_1(\varepsilon;\gamma)$ with a finite dimensional parameter γ, which may be estimated by standard maximum likelihood, i.e., by solving

$$\widehat{\gamma} = \arg\max_\gamma \mathbb{P}_n \log\eta_1(\varepsilon(\widehat{\beta}_I);\gamma),$$

where $\widehat{\beta}_I$ is an inefficient initial estimator of β solving a user-specified estimating equation, e.g.,

$$\mathbb{P}_n\left[\tilde{h}(\varepsilon(\beta),Z)\right] = 0$$

with $h(\varepsilon(\beta),Z) = \varepsilon(\beta)\dot{m}(Z;\beta)$.

5.3 Partially linear regression model

Consider the partially linear regression model given in Example 1.5

$$\mathbb{E}(Y \mid X,Z) = \beta^\top X + \eta(Z).$$

The parameter space is $\theta = (\beta,\eta,g_1,g_2)$, where $g_1(x,z)$ is the density of (X,Z), and $g_2(\varepsilon;x,z)$ is the density of $\varepsilon = Y - \beta^\top X - \eta(Z)$ given (X,Z). Let $\sigma^2(x,z) = \mathbb{E}[\varepsilon^2 \mid X = x, Z = z]$ be the variance of the (heterogeneous) noise.

Result 5.3. The following results hold.

1. The score of β is

$$S_\beta(\theta^*) = -\frac{g_2'(Y - \beta^{*\top} X - \eta^*(Z); X, Z)}{g_2(Y - \beta^{*\top} X - \eta^*(Z); X, Z)} X.$$

2. The nuisance tangent space is

$$\Lambda_{\text{nuis}} = \Bigg\{ a_1(X, Z) + a_2(\varepsilon, X, Z) + \frac{g_2'(\varepsilon; X, Z)}{g_2(\varepsilon; X, Z)} a_3(Z) : \\ \mathbb{E}[a_1(X, Z)] = 0, \mathbb{E}[a_2(\varepsilon, X, Z) \mid X, Z] = 0, \mathbb{E}[\varepsilon a_2(\varepsilon, X, Z) \mid X, Z] = 0 \Bigg\},$$

   and its orthogonal complement is

$$\Lambda_{\text{nuis}}^\perp = \{ h(X, Z)\varepsilon : \mathbb{E}[h(X, Z) \mid Z] = 0 \}.$$

3. The efficient score is

$$\begin{aligned} S_{\beta,\text{eff}}(\theta^*) &= \Pi[S_\beta(\theta^*) \mid \Lambda_{\text{nuis}}^\perp] \\ &= \left\{ \frac{X}{\sigma^2(X, Z)} - \frac{\mathbb{E}\left(\frac{X}{\sigma^2(X,Z)} \mid Z \right)}{\mathbb{E}\left(\frac{1}{\sigma^2(X,Z)} \mid Z \right)} \frac{1}{\sigma^2(X, Z)} \right\} \varepsilon, \end{aligned}$$

   so the efficient influence function is $\left\{ \mathbb{E}[S_{\beta,\text{eff}}(\theta^*) S_{\beta,\text{eff}}(\theta^*)^\top] \right\}^{-1} S_{\beta,\text{eff}}(\theta^*)$ and the C-R bound is $\left\{ \mathbb{E}[S_{\beta,\text{eff}}(\theta^*) S_{\beta,\text{eff}}(\theta^*)^\top] \right\}^{-1}$.

Remark 5.4. The following equality is frequently used in the proof of Result 5.3:

$$\mathbb{E}\left\{ -\frac{g_2'(\varepsilon; X, Z)}{g_2(\varepsilon; X, Z)} \varepsilon \right\} = 1,$$

which is derived by differentiating both sides of the following equation w.r.t. β

$$\int \varepsilon(\beta) g_2(\varepsilon(\beta); x, z) \mathrm{d}\varepsilon = 0.$$

Such step is common and useful as it also appears in Remarks 5.1 and 5.3.

Remark 5.5. Partially linear regression (PLR) is a special case in the **conditional moment framework**

$$\mathbb{E}[\rho(x, y, \theta_0, h_0(x_2)) \mid x] = 0,$$

where ρ is the moment restriction, $x = (x_1^\top, x_2^\top)^\top$, θ_0 is the true value of a finite dimensional parameter θ, and h_0 is the true value of an unknown function h. Setting $\rho(x, y, \theta, \tau) = y - \theta^\top x_1 - \tau$ recovers PLR. [Chamberlain, 1992] derives the C-R bound

of the semiparametric model with the conditional moment restriction only (with a different approach). The inverse of the C-R bound is given by

$$\mathcal{V}^{-1} = \mathbb{E}\left\{\mathbb{E}\left(D_0^\top \Sigma_0^{-1} D_0 \mid x_2\right) - \mathbb{E}\left(D_0^\top \Sigma_0^{-1} H_0 \mid x_2\right)\left[\mathbb{E}\left(H_0^\top \Sigma_0^{-1} H_0 \mid x_2\right)\right]^{-1} \mathbb{E}\left(H_0^\top \Sigma_0^{-1} D_0 \mid x_2\right)\right\},$$

where

$$\begin{aligned} D_0(x) &= \mathbb{E}[\partial \rho(x, y, \theta_0, h_0(x_2))/\partial \theta^\top \mid x], \\ H_0(x) &= \mathbb{E}[\partial \rho(x, y, \theta_0, h_0(x_2))/\partial \tau^\top \mid x], \\ \Sigma_0(x) &= \mathbb{E}[\rho(x, y, \theta_0, h_0(x_2))\rho(x, y, \theta_0, h_0(x_2))^\top \mid x]. \end{aligned}$$

I guess the efficient score is

$$\begin{aligned} S_{\theta,\text{eff}} = {} & D_0^\top \Sigma_0^{-1} \rho(x, y, \theta_0, h_0(x_2)) \\ & - \mathbb{E}\left(D_0^\top \Sigma_0^{-1} H_0 \mid x_2\right)\left[\mathbb{E}\left(H_0^\top \Sigma_0^{-1} H_0 \mid x_2\right)\right]^{-1} H_0 \Sigma_0^{-1} \rho(x, y, \theta_0, h_0(x_2)). \end{aligned}$$

5.4 Cox proportional hazards model with censored data

5.5 Randomized trials with baseline covariates

Consider a double-blind placebo controlled randomized trial to evaluate the average causal effect (ATE) of an intervention A, which is binary, on an outcome Y. By randomization, the average causal effect on the additive scale is given by

$$\beta = \mathbb{E}(Y \mid A = 1) - \mathbb{E}(Y \mid A = 0).$$

We assume $\mathbb{P}(A = 1) = 1/2$. We also observe a large collection of baseline characteristics L, supposed to be independent of A, i.e., $A \perp L$. Let $\mathcal{M}$ be the set of all regular law $f_{Y,A,L}(y, a, l)$ such that

$$f_{Y,A,L}(y, a, l) = f_{Y|A,L}(y \mid a, l) f_{A|L}(a) f_L(l).$$

Result 5.4. The following results hold.

1. The tangent space is
$$\Lambda = \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3,$$
where
$$\begin{aligned} \Lambda_1 &= \{b_1(Y, A, L) \colon \mathbb{E}\left(B_1 \mid A, L\right) = 0\} \cap \mathcal{L}_2^0, \\ \Lambda_2 &= \{b_2(A) \colon \mathbb{E}\left(B_2 \mid L\right) = \mathbb{E}\left(B_2\right) = 0\} \cap \mathcal{L}_2^0, \\ \Lambda_3 &= \{b_3(L) \colon \mathbb{E}\left(B_3\right) = 0\} \cap \mathcal{L}_2^0. \end{aligned}$$

2. The orthogonal complement of the tangent space is
$$\begin{aligned} \Lambda^\perp &= \Lambda_1^\perp \cap \Lambda_2^\perp \cap \Lambda_3^\perp \\ &= \left\{\left(A - \frac{1}{2}\right)\left(b(L) - \mathbb{E}(b(L))\right) \colon b(L)\right\} \cap \mathcal{L}_2^0. \end{aligned}$$

3. The efficient influence function in the nonparametric model $\mathcal{M}_{\text{np}}$ is

$$\text{IF}_{\beta,\text{np}} = 2\mathbb{I}(A=1)(Y-\mu_1) - 2\mathbb{I}(A=0)(Y-\mu_0),$$

where $\mu_1 = \mathbb{E}(Y \mid A=1)$ and $\mu_0 = \mathbb{E}(Y \mid A=0)$.

4. The set of influence function in the semiparametric $\mathcal{M}$ is

$$\text{IF}_{\beta,\text{np}} + \Lambda^{\perp} = \left\{ \begin{array}{c} 2\mathbb{I}(A=1)(Y-\mu_1) - 2\mathbb{I}(A=0)(Y-\mu_0) \\ + \left(A - \frac{1}{2}\right)(b(L) - \mathbb{E}(b(L))) \end{array} : b(L) \right\} \cap \mathcal{L}_2^0.$$

5. The efficient influence function in $\mathcal{M}$ is

$$\begin{aligned}
\text{IF}_{\beta}^{\text{eff}} &= \Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda\right) \\
&= \Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_1 \oplus \Lambda_2 \oplus \Lambda_3\right) \\
&= \Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_1\right) \oplus \Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_2\right) \oplus \Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_3\right) \\
&= \underbrace{2\mathbb{I}(A=1)(Y-\mathbb{E}(Y \mid A=1,L)) - 2\mathbb{I}(A=0)(Y-\mathbb{E}(Y \mid A=0,L))}_{\Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_1\right)} \\
&\quad + \underbrace{0}_{\Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_2\right)} + \underbrace{\mathbb{E}(Y \mid A=1,L) - \mu_1 - \mathbb{E}(Y \mid A=0,L) + \mu_0}_{\Pi\left(\text{IF}_{\beta,\text{np}} \mid \Lambda_3\right)} \\
&= 2\mathbb{I}(A=1)(Y-\mu_1) - 2\mathbb{I}(A=0)(Y-\mu_0) \\
&\quad - (2\mathbb{I}(A=1)-1)(\mathbb{E}(Y \mid A=1,L) - \mathbb{E}(Y \mid A=0,L) - \mu_1 + \mu_0).
\end{aligned}$$

So it is in $\text{IF}_{\beta,\text{np}} + \Lambda^{\perp}$ with

$$b(L) = b^{\text{opt}}(L) = -2\{\mathbb{E}(Y \mid A=1,L) - \mathbb{E}(Y \mid A=0,L)\}.$$

Remark 5.6. It is straightforward to establish that $\text{IF}_{\beta,\text{np}}$ is the influence function of the nonparametric estimator of β given by

$$\widehat{\beta} = \frac{\mathbb{P}_n(YA)}{\mathbb{P}_n(A)} - \frac{\mathbb{P}_n(Y(1-A))}{\mathbb{P}_n((1-A))}.$$

However, as we have shown, $\text{IF}_{\beta,\text{np}}$ is not the efficient influence function in the semiparametric model $\mathcal{M}$. As a consequence, we have established that the difference in sample means (which is the standard to analyze a randomized trial) is inefficient when baseline characteristics are available in a randomized trial.

Adaptive semiparametric estimators A semiparametric locally efficient estimator can be obtained by using the efficient influence function as an estimating equation

$$\begin{aligned}
0 = \sum_{i=1}^{n} \big\{ &2\mathbb{I}(A_i=1)(Y_i - \mathbb{E}(Y \mid A=1,L)) - 2\mathbb{I}(A_i=0)(Y_i - \mathbb{E}(Y \mid A=0,L)) \\
&+ \mathbb{E}(Y \mid A=1,L) - \mathbb{E}(Y \mid A=0,L) - \beta \big\},
\end{aligned}$$

which depends on the unknown conditional mean function $\mathbb{E}(Y \mid A, L)$. We can construct a simple parametric working model

$$b(A, L; \eta) = \left(1, A, AL^\top, L^\top\right) \eta$$

and estimate η using OLS. Then the estimator $\widehat{\beta}$ of the previous estimating equation is consistent and asymptotically normal, and achieves the semiparametric C-R bound if the working model for $\mathbb{E}(Y \mid A, L)$ is correctly specified.

5.6 Causal effects of a point exposure

Suppose randomization in Section 5.5 no longer holds, i.e., we do not have $(Y, L) \perp A$. Instead, suppose that the "no unmeasured confounding assumption" (NUCA) holds, i.e., $\{Y_0, Y_1\} \perp A \mid L$. The intuition is that, we have measured enough covariates L so that within levels of L, the data mimics a randomized trial.

In causal literature, there are several quantities of great interest:

- individual causal effect (ICE): $\text{ICE} = Y_{i1} - Y_{i0}$;
- average causal effect (ACE): $\text{ACE} = \mathbb{E}(\text{ICE}) = \mathbb{E}Y_1 - \mathbb{E}Y_0$;
- conditional average treatment effect (CATE): $\text{CATE} = \mathbb{E}(Y_1 - Y_0 \mid V)$, with V contained in L;
- average treatment effect on the treated (ATT): $\text{ATT} = \mathbb{E}(Y_1 - Y_0 \mid A = 1)$;
- ratio: $\mathbb{E}Y_1 / \mathbb{E}Y_0$.

We show that NUCA is sufficient to identify $\{\mathbb{E}Y_a \colon a\}$ and thus $\psi = \mathbb{E}Y_1 - \mathbb{E}Y_0$. We have

$$\begin{aligned}
\mathbb{E}Y_a &= \mathbb{E}[\mathbb{E}(Y_a \mid L)] \\
&= \sum_l \mathbb{E}(Y_a \mid L = l) f_L(l) \\
&\overset{\text{NUCA}}{=\!=} \sum_l \mathbb{E}(Y_a \mid A = a, L = l) f_L(l) \\
&\overset{\text{CA}}{=\!=} \sum_l \mathbb{E}(Y \mid A = a, L = l) f_L(l) \\
&= g(a).
\end{aligned}$$

$g(a)$ is the **direct standardization** of $\mathbb{E}(Y \mid A = a, L)$, and is a special case of Robin's G-formula. Consequently,

$$\psi = g(1) - g(0) = \sum_l \{\mathbb{E}(Y \mid A = 1, L = l) - \mathbb{E}(Y \mid A = 0, L = l)\} f_L(l) \tag{5.2}$$

is the **standard risk difference**. Note that crude association $\neq$ causation, as

$$\sum_l \mathbb{E}(Y_a \mid A=a, L=l) f_L(l) = \mathbb{E}(Y_a)$$
$$\neq \mathbb{E}(Y \mid A=a) = \sum_l \mathbb{E}(Y \mid A=a, L=l) f_L(l \mid A=a).$$

However, if NUCA holds and either of the following conditions holds:

$$Y \perp L \mid A \text{ or } A \perp L,$$

then $\mathbb{E}(Y_a) = \mathbb{E}(Y \mid A=a)$, and RA actually holds.

5.6.1 G-formula

As shown previously, when estimating ACE/ATE, the corresponding G-formula is the so-called point exposure G-formula

$$g(a) = \mathbb{E}Y_a = \int \mathbb{E}(Y \mid A=a, L=l) \mathrm{d}F(l). \tag{5.3}$$

The G-formula is not restricted to the mean. For example, when estimating CATE, the G-formula becomes

$$g(a,v) = \mathbb{E}(Y_a \mid V=v) = \int \mathbb{E}(Y \mid A=a, L=l) \mathrm{d}F(l \mid V=v).$$

5.6.2 G-computation

Given the observed data $O_i = (Y_i, A_i, L_i)$, G-computation refers to nonparametric inference on the G-formula (5.3). A natural nonparametric estimator of $g(a)$ is given by the nonparametric plug-in estimator, which requires nonparametric estimates $\widehat{b}(a,l)$ of $\mathbb{E}(Y \mid A=a, L=l) = b(a,l)$ and $\widehat{F}_L(l)$ of $F_L(l)$.

Sample average estimator. Assume A and L are both categorical variables with moderate number of levels, so that $\widehat{b}(a,l)$ and $\widehat{F}_L(l)$ are given by **stratified sample average**:

$$\widehat{b}(a,l) = \frac{\sum_{i=1}^n \mathbb{I}(A_i=a, L_i=l) Y_i}{\sum_{i=1}^n \mathbb{I}(A_i=a, L_i=l)},$$
$$\widehat{f}_L(l) = n^{-1} \sum_{i=1}^n \mathbb{I}(L_i=l).$$

Then the nonparametric estimator of the G-formula is given by

$$\widehat{g}(a) = \sum_l \widehat{b}(a,l) \widehat{f}_L(l) = n^{-1} \sum_l \widehat{b}(a,l) \sum_{i=1}^n \mathbb{I}(L_i=l)$$
$$= n^{-1} \sum_{i=1}^n \sum_l \widehat{b}(a,l) \mathbb{I}(L_i=l) = n^{-1} \sum_{i=1}^n \widehat{b}(a, L_i).$$

Asymptotic variance. The additive terms $\widehat{b}(a, L_i)$ and $\widehat{b}(a, L_j)$ are highly dependent, so the standard i.i.d. CLT does not apply. However, it is still possible to derive the asymptotic variance of $\widehat{g}(a)$:

$$
\begin{aligned}
n^{1/2}\widehat{g}(a) &= n^{-1/2}\sum_l \sum_{i=1}^n \widehat{b}(a,l)\mathbb{I}(L_i = l) \\
&= n^{-1/2}\sum_l \sum_{i=1}^n \mathbb{I}(L_i = l)\left\{\frac{\sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)Y_s}{\sum_{j=1}^n \mathbb{I}(A_j = a, L_j = l)}\right\} \\
&= n^{-1/2}\sum_l \sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)Y_s\left\{\frac{\sum_{i=1}^n \mathbb{I}(L_i = l)}{\sum_{j=1}^n \mathbb{I}(A_j = a, L_j = l)}\right\} \\
&= n^{-1/2}\sum_l \sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)(Y_s - b(a,l))\left\{\frac{n^{-1}\sum_{i=1}^n \mathbb{I}(L_i = l)}{n^{-1}\sum_{j=1}^n \mathbb{I}(A_j = a, L_j = l)}\right\} \\
&\quad + n^{-1/2}\sum_l b(a,l)\left\{\sum_{i=1}^n \mathbb{I}(L_i = l)\right\} \\
&= n^{-1/2}\sum_l \sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)(Y_s - b(a,l))f_{A|L}^{-1}(a \mid l) \\
&\quad + n^{-1/2}\sum_l b(a,l)\left\{\sum_{i=1}^n \mathbb{I}(L_i = l)\right\} + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^n \left\{\frac{\mathbb{I}(A_i = a)}{f_{A|L}(A_i \mid L_i)}(Y_i - b(A_i, L_i)) + b(a, L_i)\right\} + o_p(1).
\end{aligned}
$$

So we finally obtain

$$
\begin{aligned}
n^{1/2}(\widehat{g}(a) - g(a)) &= n^{-1/2}\sum_{i=1}^n \left\{\frac{\mathbb{I}(A_i = a)}{f_{A|L}(A_i \mid L_i)}(Y_i - b(A_i, L_i)) + b(a, L_i) - g(a)\right\} + o_p(1) \\
&= n^{-1/2}\sum_{i=1}^n \mathrm{IF}_i(a) + o_p(1) \xrightarrow{d} \mathcal{N}(0, \mathbb{E}(\mathrm{IF}(a)^2)).
\end{aligned}
$$

Statistical inference. Based on the estimator $\widehat{g}(a)$ and its influence function $\mathrm{IF}(a)$, we can construct a 95% Wald-type CI for ψ in (5.2):

$$
\widehat{\psi} \pm 1.96\sqrt{\widehat{\mathrm{var}}(\widehat{\psi})},
$$

where $\widehat{\psi} = \widehat{g}(1) - \widehat{g}(0)$ and $\widehat{\mathrm{var}}(\widehat{\psi})$ is a consistent estimator of the variance of $\widehat{\psi}$. Specifically, the CI is given by

$$
\widehat{g}(1) - \widehat{g}(0) \pm 1.96\sqrt{n^{-1}\sum_{i=1}^n \left(\widehat{\mathrm{IF}}_i(1) - \widehat{\mathrm{IF}}_i(0)\right)^2},
$$

where

$$\widehat{\mathrm{IF}}_i(a) = \frac{\mathbb{I}(A_i = a)}{\widehat{f}_{A|L}(A_i \mid L_i)}(Y_i - \widehat{b}(A_i, L_i)) + \widehat{b}(a, L_i) - \widehat{g}(a),$$

and $\widehat{f}_{A|L}(a \mid l) = \frac{\sum_{i=1}^n \mathbb{I}(A_i=a, L_i=l)}{\sum_{i=1}^n \mathbb{I}(L_i=l)}$ is the nonparametric estimator of $f_{A|L}(a \mid l)$.

Remark 5.7. There are several notes to mention:

1. in the form of the influence function $\mathrm{IF}(a)$, the term $b(a, L) - g(a)$ reflects the variability due to the estimation of $F_L(l)$, whereas $\frac{\mathbb{I}(A=a)}{f_{A|L}(A|L)}(Y - b(A, L))$ captures the variability due to the estimation of $b(a, L)$;

2. the influence function and the variance of $\widehat{\psi}$ depends on the treatment process $f_{A|L}(a \mid l)$, even though the pluggin estimator appears not to.

The IPTW estimator. We have shown that the influence function of the plug-in estimator $\widehat{g}(a)$ depends on the propensity score $f_{A|L}(a \mid l)$, though the estimator itself does not contain such a component. Actually, $\widehat{g}(a)$ has a dual representation characterizing the role of the propensity score:

$$\widehat{g}(a) = \sum_l \widehat{b}(a, l)\widehat{f}_L(l) = \frac{\sum_{s=1}^n \mathbb{I}(A_s = a)Y_s \widehat{f}^{-1}_{A|L}(A_s \mid L_s)}{\sum_{s=1}^n \mathbb{I}(A_s = a)\widehat{f}^{-1}_{A|L}(A_s \mid L_s)}.$$

This is because

$$\begin{aligned}
\widehat{g}(a) &= \sum_l \widehat{b}(a, l)\widehat{f}_L(l) \\
&= n^{-1} \sum_l \sum_{i=1}^n \mathbb{I}(L_i = l) \left\{ \frac{\sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)Y_s}{\sum_{j=1}^n \mathbb{I}(A_j = a, L_j = l)} \right\} \\
&= n^{-1} \sum_l \sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)Y_s \left\{ \frac{\sum_{i=1}^n \mathbb{I}(L_i = l)}{\sum_{j=1}^n \mathbb{I}(A_j = a, L_j = l)} \right\} \\
&= n^{-1} \sum_l \sum_{s=1}^n \mathbb{I}(A_s = a, L_s = l)Y_s \widehat{f}^{-1}_{A|L}(A_s \mid L_s) \\
&= n^{-1} \sum_{s=1}^n \mathbb{I}(A_s = a)Y_s \widehat{f}^{-1}_{A|L}(A_s \mid L_s) \\
&= \frac{\sum_{s=1}^n \mathbb{I}(A_s = a)Y_s \widehat{f}^{-1}_{A|L}(A_s \mid L_s)}{\sum_{s=1}^n \mathbb{I}(A_s = a)\widehat{f}^{-1}_{A|L}(A_s \mid L_s)},
\end{aligned}$$

where the last equality holds because

$$\begin{aligned}\sum_{s=1}^{n}\mathbb{I}(A_s=a)\widehat{f}^{-1}_{A|L}(A_s\mid L_s) &= \sum_{s=1}^{n}\mathbb{I}(A_s=a)\frac{\sum_{k=1}^{n}\mathbb{I}(L_k=L_s)}{\sum_{j=1}^{n}\mathbb{I}(A_j=A_s,L_j=L_s)}\\ &= \sum_{s=1}^{n}\sum_{k=1}^{n}\mathbb{I}(A_s=a)\frac{\mathbb{I}(L_k=L_s)}{\sum_{j=1}^{n}\mathbb{I}(A_j=a,L_j=L_k)}\\ &= \sum_{k=1}^{n}\frac{\sum_{s=1}^{n}\mathbb{I}(A_s=a,L_s=L_k)}{\sum_{j=1}^{n}\mathbb{I}(A_j=a,L_j=L_k)}=n.\end{aligned}$$

Consequently, $\widehat{g}(a)$ is also called the **inverse probability of treatment weighted (IPTW)** estimator.

Remark 5.8. Note that we only use NUCA in deriving $\mathbb{E}(Y_a)=\mathbb{E}[\mathbb{E}(Y\mid A=a,L)]$, which enables us to express the parameter of interest as a function of the observed data distribution. Specifically, Y_a is not observed in every sample, so its expectation should be taken under the full data distribution; but we observe the distribution of $Y\mid(A=a,L)$, as well as the distribution of L.

5.6.3 An efficiency paradox

We revisit the setting where A is randomized, i.e., $\{Y_0,Y_1,L\}\perp A$. Denote by L a pretreatment categorical covariate that is a strong predictor of the outcome Y. As argued before, the crude estimator

$$\widetilde{\psi}=\frac{\sum_{s=1}^{n}\mathbb{I}(A_s=1)Y_s}{\sum_{s=1}^{n}\mathbb{I}(A_s=1)}-\frac{\sum_{s=1}^{n}\mathbb{I}(A_s=0)Y_s}{\sum_{s=1}^{n}\mathbb{I}(A_s=0)}$$

is a valid estimator of ψ. Note that $\widetilde{\psi}$ can be seen as an IPTW estimator with known weights $f^{-1}_{A|L}(A_s\mid L_s)=(1/2)^{-1}$, and its asymptotic variance is given by

$$\operatorname{avar}(\sqrt{n}(\widetilde{\psi}-\psi))=2\mathbb{E}\left((Y-g(1))^2\mid A=1\right)+2\mathbb{E}\left((Y-g(0))^2\mid A=0\right).$$

If we ignore the randomization assumption and only assume NUCA, i.e., $\{Y_0,Y_1\}\perp A\mid L$, then the IPTW estimator is given using G-computation as before:

$$\widehat{\psi}=\frac{\sum_{s=1}^{n}\mathbb{I}(A_s=1)Y_s\widehat{f}^{-1}_{A|L}(A_s\mid L_s)}{\sum_{s=1}^{n}\mathbb{I}(A_s=1)\widehat{f}^{-1}_{A|L}(A_s\mid L_s)}-\frac{\sum_{s=1}^{n}\mathbb{I}(A_s=0)Y_s\widehat{f}^{-1}_{A|L}(A_s\mid L_s)}{\sum_{s=1}^{n}\mathbb{I}(A_s=0)\widehat{f}^{-1}_{A|L}(A_s\mid L_s)}.$$

In fact, we can show $\operatorname{var}(\widehat{\psi})\le\operatorname{var}(\widetilde{\psi})$ (we verify this in Appendix A). That means the estimator becomes less efficient if we uses available information on the treatment mechanism!

However, it will be clear from Section 6 that the treatment process is ancillary, i.e., the efficiency bound for estimating ψ is the same whether or not one knows the treatment mechanism.

The paradox is resolved by realizing that $\widetilde{\psi}$ uses the available information, but in a very inefficient way as it ignores the fact that L is a strong correlate of Y; in contrast, the G-formula and the IPTW with nonparametrically estimated weights correctly incorporates this information.

6 Factorized Likelihood Models

This section discusses a special but useful semiparametric models, called "factorized likelihood models".

6.1 The parametric case

First, consider a parametric model with factorized likelihood, i.e.,

$$\mathcal{F}_{\text{par}} = \{f(z;\theta) = g_1(z;\theta_1)g_2(z;\theta_2) \colon (\theta_1,\theta_2) \in \Theta_1 \times \Theta_2 \subseteq \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}\},$$

where g_1 and g_2 are known functions. A simple example is the linear regression

$$y = X\beta + \varepsilon,$$

where we assume X is independent of ε, and the density of ε is known. The overall density is

$$f(x,y) = f_\varepsilon(y - X\beta) f_X(x;\eta).$$

Suppose enough regularity conditions hold. Then the score is

$$\underbrace{s_\theta(Z;\theta)}_{(p_1+p_2)\times 1} = \begin{bmatrix} s_{\theta_1}(Z;\theta) \\ s_{\theta_2}(Z;\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\log f(z;\theta) \\ \frac{\partial}{\partial\theta_2}\log f(z;\theta) \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial\theta_1}\log g_1(z;\theta_1) \\ \frac{\partial}{\partial\theta_2}\log g_2(z;\theta_2) \end{bmatrix},$$

and the FI matrix is

$$\underbrace{I(\theta)}_{(p_1+p_2)\times(p_1+p_2)} = -\mathbb{E}_\theta\left[\frac{\partial^2}{\partial\theta\partial\theta^\top}\log f(z;\theta)\right] = -\mathbb{E}_\theta\begin{bmatrix} \frac{\partial^2}{\partial\theta_1\partial\theta_1^\top}\log g_1(z;\theta_1) & 0 \\ 0 & \frac{\partial^2}{\partial\theta_2\partial\theta_2^\top}\log g_2(z;\theta_2) \end{bmatrix},$$

where we used

$$\frac{\partial^2}{\partial\theta_1\partial\theta_2^\top}\log f(z;\theta) = \frac{\partial}{\partial\theta_1} s_{\theta_2}(Z;\theta)^\top = 0.$$

Thus $\mathbb{E}_\theta\left[s_{\theta_1}(Z;\theta)s_{\theta_2}(Z;\theta)^\top\right] = 0$. So the C-R bound for estimation of θ_1 in the model $\mathcal{F}_{\text{par}}$ is the same as the C-R bound for estimation of θ_1 in the parametric model in which θ_2 is known, i.e.,

$$\mathcal{F}_{\text{par},1,\theta_2^*} = \{f(z;\theta) = g_1(z;\theta_1)g_2(z;\theta_2^*) \colon \theta_1 \in \Theta_1 \subseteq \mathbb{R}^{p_1}\}$$

for any θ_2^*. Also, if $\beta(\theta)$ depends on θ only through θ_1, then the C-R bound for $\beta(\theta)$ at any $F^* = F_{\theta^*}$ is the same in models $\mathcal{F}_{\text{par},1,\theta_2^*}$ and $\mathcal{F}_{\text{par}}$.

In other words, in a factorized likelihood model, the efficiency (with which we can estimate with large samples) of a parameter that depends only on one of the factors is the same regardless of whether the remaining factor is known or unknown.

The above result also implies a geometric result. The tangent space in $\mathcal{F}_{\text{par},1,\theta_2^*}$ at $F^* = F_{\theta^*}$ is equal to

$$\Lambda_{\mathcal{F}_{\text{par},1,\theta_2^*}}(F^*) = \{a_1^\top s_{\theta_1}(Z;\theta^*) \colon a_1\mathbb{R}^{p_1}\},$$

and because $s_{\theta_1}(Z;\theta^*)$ is the same for all θ_2^*, the space $\Lambda_{\mathcal{F}_{\text{par},1,\theta_2^*}}(F^*)$ does not change with θ_2^*. Likewise $\Lambda_{\mathcal{F}_{\text{par},2,\theta_1^*}}(F^*)$, the tangent space of $\mathcal{F}_{\text{par},2,\theta_1^*}$ at F^*, does not depend on θ_1^*. Now, because the tangent space $\Lambda_{\mathcal{F}_{\text{par}}}$ is comprised of all linear combinations of the components of $s_{\theta_1}(Z;\theta^*)$ and $s_{\theta_2}(Z;\theta^*)$, the orthogonality of $s_{\theta_1}(Z;\theta^*)$ and $s_{\theta_2}(Z;\theta^*)$ implies that

$$\Lambda_{\mathcal{F}_{\text{par}}}(F^*) = \Lambda_{\mathcal{F}_{\text{par},1,\theta_2^*}}(F^*) \oplus \Lambda_{\mathcal{F}_{\text{par},2,\theta_1^*}}(F^*).$$

Thus we conclude that, in factorized likelihood parametric models with variational independent parameters, the tangent space is the direct sum of two orthogonal tangent spaces, each tangent space being the tangent space of the model where one of the factors in the likelihood is known.

6.2 The semiparametric case

Now we extend the preceding result to semiparametric models. A factorized likelihood semiparametric model is of the form

$$\mathcal{F} = \{f(z) = g_1(z)g_2(z) \colon g_1 \in G_1, g_2 \in G_2\}, \tag{6.1}$$

where G_1 and G_2 are sets of functions.

Example 6.1. An example of a factorized likelihood model is one in which $Z = (Z_1^\top, Z_2^\top)^\top$ and

$$g_1(z) = f_{Z_1}(z), \quad g_2(z) = f_{Z_2|Z_1}(z_2 \mid z_1).$$

If $\mathcal{F}$ is the nonparametric model, then G_1 is the set of all possible densities of Z_1 and G_2 is the set of all possible conditional densities of Z_2 given Z_1. If instead, G_2 is the set of all densities $f_{Z_2|Z_1}$ of $Z_2 \mid Z_1$ such that there exists β verifying $\mathbb{E}_{F_{Z_2|Z_1}}(Z_2 \mid Z_1) = \beta^\top Z_1$, then $\mathcal{F}$ is a strictly semiparametric model.

For given $g_1^*(z)$ and $g_2^*(z)$, we define

$$\mathcal{F}_1 = \{f(z) = g_1(z)g_2^*(z) \colon g_1 \in G_1\}, \tag{6.2}$$
$$\mathcal{F}_2 = \{f(z) = g_1^*(z)g_2(z) \colon g_2 \in G_2\}, \tag{6.3}$$

i.e., two semiparametric submodels of $\mathcal{F}$ such that g_2 (resp. g_1) is known and equal to g_2^* (resp. g_1^*). In addition, we define a class $\mathcal{A}$ of regular parametric models

$$\mathcal{F}_{\text{sub}} = \{f(z;\theta) = g_1(z;\theta)g_2(z;\theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p, g_j(z;\theta) \in G_j, j = 1,2\} \tag{6.4}$$

through F^*, such that $g_1(z) = g_1(z;\theta^*)$ and $g_2^*(z) = g_2(z;\theta^*)$. For each $\mathcal{F}_{\text{sub}} \in \mathcal{A}$ and each $\theta' \in \Theta$, the submodels

$$\mathcal{F}_{\text{sub},1,\theta'} = \{f(z;\theta) = g_1(z;\theta)g_2(z;\theta') \colon \theta \in \Theta \subseteq \mathbb{R}^p\}, \tag{6.5}$$
$$\mathcal{F}_{\text{sub},2,\theta'} = \{f(z;\theta) = g_1(z;\theta')g_2(z;\theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\}, \tag{6.6}$$

Note that $\mathcal{F}_{\text{sub},1,\theta'}$ is a parametric submodel of $\mathcal{F}_1$ in which $g_2(z)$ is known and equal to $g_2(z;\theta')$, and likewise for $\mathcal{F}_{\text{sub},2,\theta'}$ with roles of the subscripts 1 and 2 interchanged. For a given θ^* we also define

$$\mathcal{A}_1 = \{\mathcal{F}_{\text{sub},1,\theta^*} : \mathcal{F}_{\text{sub}} \in \mathcal{A}\}, \tag{6.7}$$
$$\mathcal{A}_2 = \{\mathcal{F}_{\text{sub},2,\theta^*} : \mathcal{F}_{\text{sub}} \in \mathcal{A}\}. \tag{6.8}$$

If the models in $\mathcal{A}_1$ are regular at $F^* = F_{\theta^*}$, we let $\Lambda_{\mathcal{F}_1}(F^*)$ denote the tangent space at F^* w.r.t. $\mathcal{A}_1$ in $\mathcal{F}_1$, and define $\Lambda_{\mathcal{F}_2}(F^*)$ likewise.

Lemma 6.1. *Suppose that*

1. *for each $\mathcal{F}_{sub} \in \mathcal{A}$,*
   (a) *$\mathcal{F}_{sub,1,\theta'}$ and $\mathcal{F}_{sub,2,\theta'}$ are regular parametric models for any $\theta' \in \Theta$, with scores at θ^* that do not depend on θ', and are denoted as $s_1(Z;\theta^*)$ and $s_2(Z;\theta^*)$;*
   (b) *$\mathcal{F}_{sub,1,\theta^*}$ and $\mathcal{F}_{sub,2,\theta^*}$ are submodels in the class $\mathcal{A}$;*
2. *for each $\mathcal{F}_{sub} \in \mathcal{A}$, the score $s_\theta(Z;\theta^*)$ for θ at θ^* in the model $\mathcal{F}_{sub}$ can be decomposed as*
$$s_\theta(Z;\theta^*) = s_1(Z;\theta^*) + s_2(Z;\theta^*);$$
3. *the following map is continuous at $\theta = \theta^*$:*
$$\theta \to \mathbb{E}_\theta \left[s_1(Z;\theta^*)^2 \right],$$
where $\mathbb{E}_\theta(\cdot)$ is the expectation under $f(z;\theta)$ in the model $\mathcal{F}_{sub,2,\theta^*}$, i.e.,
$$\mathbb{E}_\theta \left[s_1(Z;\theta^*)^2 \right] = \int s_1(z;\theta^*)^2 g_1(z;\theta^*) g_2(z;\theta) \mathrm{d}z.$$
Then,
$$\Lambda_{\mathcal{F}}(F^*) = \Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*).$$

Lemma 6.2. *Suppose the assumptions in Lemma 6.1 hold. Suppose a parameter $\beta(F)$ of interest is pathwise differentiable w.r.t. $\mathcal{A}$ at F^* and depends on F only through g_1, i.e.,*
$$f = g_1 g_2 \text{ and } f' = g_1 g_2' \Rightarrow \beta(F) = \beta(F').$$
Then,

1. *for any gradient $\psi_{F^*}(Z)$ of $\beta(F)$ w.r.t. $\mathcal{A}$ in $\mathcal{F}$ at F^* it holds that*
$$\psi_{F^*}(Z) \perp \Lambda_{\mathcal{F}_2}(F^*);$$
2. *the efficient influence function $\psi_{F^*,eff}$ of $\beta(F)$ w.r.t. $\mathcal{A}$ in $\mathcal{F}$ at F^* satisfies*
$$\psi_{F^*,eff}(Z) \in \Lambda_{\mathcal{F}_1}(F^*);$$

3. *if* $\psi_{1,F^*,\mathit{eff}}(Z)$ *denotes the efficient influence function of* $\beta(F)$ *w.r.t.* $\mathcal{A}$ *in* $\mathcal{F}_1$ *at* F^**, then*

$$\psi_{F^*,\mathit{eff}}(Z) = \psi_{1,F^*,\mathit{eff}}(Z),$$

and consequently, the C-R bound for $\beta(F)$ *at* F^* *w.r.t.* $\mathcal{A}$ *in* $\mathcal{F}_1$ *is the same as the C-R bound at* F^* *w.r.t.* $\mathcal{A}$ *in* $\mathcal{F}$*.*

Example 6.1 (continued)**.** Recall in this example, $Z = (Z_1^\top, Z_2^\top)^\top$ and

$$g_1(z) = f_{Z_1}(z), \quad g_2(z) = f_{Z_2|Z_1}(z_2 \mid z_1).$$

The assumptions of Lemma 6.1 boil down to assumptions about a class $\mathcal{A}$ of regular parametric submodels of the form

$$\mathcal{F}_{\text{sub}} = \{f(z_1, z_2; \theta) = f_{Z_1}(z_1; \theta) f_{Z_2|Z_1}(z_2 \mid z_1; \theta) \colon \theta \in \mathbb{R}^p\}.$$

1. The first assumption requires that for each $\mathcal{F}_{\text{sub}} \in \mathcal{A}$,

   (a) two models

   $$\mathcal{F}_{\text{sub},1,\theta'} = \{f(z_1, z_2; \theta) = f_{Z_1}(z_1; \theta) f_{Z_2|Z_1}(z_2 \mid z_1; \theta') \colon \theta \in \mathbb{R}^p\},$$
   $$\mathcal{F}_{\text{sub},2,\theta'} = \{f(z_1, z_2; \theta) = f_{Z_1}(z_1; \theta') f_{Z_2|Z_1}(z_2 \mid z_1; \theta) \colon \theta \in \mathbb{R}^p\},$$

   are regular parametric models for any $\theta' \in \Theta$, with scores at θ^* that do not depend on θ', and are denoted as $s_1(Z; \theta^*)$ and $s_2(Z; \theta^*)$; [Note that this is not at all restrictive, since most "well behaved" submodels $\mathcal{F}_{\text{sub}}$ will satisfy this condition]

   (b) $\mathcal{F}_{\text{sub},1,\theta^*}$ and $\mathcal{F}_{\text{sub},2,\theta^*}$ are submodels in the class $\mathcal{A}$. [Note that this can always be achieved by enlarging the class $\mathcal{A}$ to include the submodels $\mathcal{F}_{\text{sub},1,\theta^*}$ and $\mathcal{F}_{\text{sub},2,\theta^*}$]

2. The second assumption requires that for each $\mathcal{F}_{\text{sub}} \in \mathcal{A}$, the score $s_\theta(Z; \theta^*)$ for θ at θ^* in the model $\mathcal{F}_{\text{sub}}$ can be decomposed as

   $$s_\theta(Z; \theta^*) = s_1(Z; \theta^*) + s_2(Z; \theta^*).$$

   Note this assumption holds trivially if scores are derivatives of the log-likelihood, i.e., if

   $$s_\theta(Z; \theta^*) = \frac{\partial}{\partial \theta^\top} \log\{f_{Z_1}(z_1; \theta) f_{Z_2|Z_1}(z_2 \mid z_1; \theta)\}\Big|_{\theta=\theta^*},$$
   $$s_1(Z; \theta^*) = \frac{\partial}{\partial \theta^\top} \log\{f_{Z_1}(z_1; \theta) f_{Z_2|Z_1}(z_2 \mid z_1; \theta')\}\Big|_{\theta=\theta^*},$$
   $$s_2(Z; \theta^*) = \frac{\partial}{\partial \theta^\top} \log\{f_{Z_1}(z_1; \theta') f_{Z_2|Z_1}(z_2 \mid z_1; \theta)\}\Big|_{\theta=\theta^*}.$$

3. The third assumption holds trivially because the map

$$\theta \to \mathbb{E}_\theta \left[s_1(Z;\theta^*)^2 \right]$$

is constant, since

$$\begin{aligned}
&\int s_1(z;\theta^*)^2 g_1(z;\theta^*) g_2(z;\theta) \mathrm{d}z \\
&= \iint s_1(z_1;\theta^*)^2 f_{Z_1}(z_1;\theta^*) f_{Z_2|Z_1}(z_2 \mid z_1;\theta) \mathrm{d}z_1 \mathrm{d}z_2 \\
&= \int s_1(z_1;\theta^*)^2 f_{Z_1}(z_1;\theta^*) \mathrm{d}z_1 \underbrace{\int f_{Z_2|Z_1}(z_2 \mid z_1;\theta) \mathrm{d}z_2}_{=1} \\
&= \int s_1(z_1;\theta^*)^2 f_{Z_1}(z_1;\theta^*) \mathrm{d}z_1.
\end{aligned}$$

Proof of Lemma 6.1. Because $s_1(Z;\theta^*)$ is a score at $\theta = \theta^*$ in $\mathcal{F}_{\text{sub},1,\theta'}$ for all $\theta' \in \Theta$, we have

$$\int s_1(Z;\theta^*) g_1(z;\theta^*) g_2(z;\theta') \mathrm{d}z = 0$$

for all $\theta' \in \Theta$. Then

$$0 = \frac{\partial}{\partial \theta^\top} \left[\int s_1(z;\theta^*) g_1(z;\theta^*) g_2(z;\theta) \mathrm{d}z \right] \Bigg|_{\theta=\theta^*}.$$

Since by assumption, the model

$$\mathcal{F}_{\text{sub},2,\theta^*} = \{ f(z;\theta) = g_1(z;\theta^*) g_2(z;\theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p \}$$

is regular and the map

$$\theta \to \mathbb{E}_\theta \left[s_1(Z;\theta^*)^2 \right] = \int s_1(z;\theta^*)^2 g_1(z;\theta^*) g_2(z;\theta) \mathrm{d}z$$

is continuous at $\theta = \theta^*$, then by Lemma 3.3 with $T = s_1(Z;\theta^*)$ under the model $\mathcal{F}_{\text{sub},2,\theta^*}$,

$$\frac{\partial}{\partial \theta^\top} \left[\int s_1(z;\theta^*) g_1(z;\theta^*) g_2(z;\theta) \mathrm{d}z \right] \Bigg|_{\theta=\theta^*} = \int s_1(z;\theta^*) s_2(z;\theta^*)^\top f(z;\theta^*) \mathrm{d}z,$$

which shows that $\Lambda_{\mathcal{F}_1}(F^*)$ and $\Lambda_{\mathcal{F}_2}(F^*)$ are orthogonal.

It remains to show $\Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*) = \Lambda_{\mathcal{F}}(F^*)$. By assumption (1.b), $s_1(Z;\theta^*)$ is a score at F^* in $\mathcal{F}_{\text{sub},1,\theta^*} \in \mathcal{A}$ so $\Lambda_{\mathcal{F}_1}(F^*) \subseteq \Lambda_{\mathcal{F}}(F^*)$. Likewise, $\Lambda_{\mathcal{F}_2}(F^*) \subseteq \Lambda_{\mathcal{F}}(F^*)$. Consequently, $\Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*) \subseteq \Lambda_{\mathcal{F}}(F^*)$. On the other hand, by assumption (2), for any $a \in \mathbb{R}^p$ and any submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}$ with score $s_\theta(Z;\theta^*)$ it holds that

$$a^\top s_\theta(Z;\theta^*) \in \Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*).$$

Therefore,

$$\Lambda_{\mathcal{F}}(F^*) \subseteq \overline{[\Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*)]}.$$

But $\Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*) = \overline{[\Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*)]}$ since $\Lambda_{\mathcal{F}_j}(F^*) = \overline{[\Lambda_{\mathcal{F}_j}(F^*)]}$, $j = 1, 2$, and $\Lambda_{\mathcal{F}_1}(F^*) \perp \Lambda_{\mathcal{F}_2}(F^*)$. Thus $\Lambda_{\mathcal{F}}(F^*) \subseteq \Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*)$. □

Proof of Lemma 6.2. By assumption (1.b),

$$\mathcal{F}_{\text{sub},2,\theta^*} = \{f(z;\theta) = g_1(z;\theta^*)g_2(z;\theta) \colon \theta \subseteq \Theta \subseteq \mathbb{R}^p\}$$

is a submodel in the class $\mathcal{A}$. Since $\beta(F)$ does not depend on g_2, $\beta(F_\theta)$ is constant over all $f(x;\theta) \in \mathcal{F}_{\text{sub},2,\theta^*}$. Therefore, since $\beta(F)$ is a pathwise differentiable parameter w.r.t. $\mathcal{A}$ at F^*, we have that with F_θ varying over the model $\mathcal{F}_{\text{sub},2,\theta^*}$,

$$0 = \frac{\partial \beta(F_\theta)}{\partial \theta}\bigg|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}\left[\psi_{F^*}(Z)s_2(Z;\theta^*)^\top\right],$$

where $\psi_{F^*}(Z)$ is any gradient of $\beta(F)$. This shows part (1) of Lemma 6.2 that $\psi_{F^*}(Z) \perp \Lambda_{\mathcal{F}_2}(F^*)$.

The proof of part (2) follows from

$$\begin{aligned}
\psi_{F^*,\text{eff}}(Z) &= \Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}}(F^*)] \\
&= \Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}_1}(F^*) \oplus \Lambda_{\mathcal{F}_2}(F^*)] \\
&= \Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}_1}(F^*)] + \underbrace{\Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}_2}(F^*)]}_{=0 \text{ bc } \psi_{F^*}(Z) \perp \Lambda_{\mathcal{F}_2}(F^*)} \\
&= \Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}_1}(F^*)],
\end{aligned}$$

which shows that $\psi_{F^*,\text{eff}}(Z) \in \Lambda_{\mathcal{F}_1}(F^*)$.

Finally, to prove part (3), we first note that $\psi_{F^*}(Z)$ is a gradient of $\beta(F)$ w.r.t. $\mathcal{A}_1$ at F^* in $\mathcal{F}_1$ because by assumption (1.b), $\mathcal{A}_1 \subseteq \mathcal{A}$. Then, the efficient influence function of $\beta(F)$ w.r.t. $\mathcal{A}_1$ at F^* in $\mathcal{F}_1$ satisfies

$$\psi_{1,F^*,\text{eff}}(Z) = \Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}_1}(F^*)].$$

This together with $\psi_{F^*,\text{eff}}(Z) = \Pi[\psi_{F^*}(Z) \mid \Lambda_{\mathcal{F}_1}(F^*)]$ (which we have just proved) concludes the proof. □

Remark 6.1. Part (1) of Lemma 6.2 is a semiparametric analogy to claim (4) of Remark 3.5, and their proofs are almost the same. In Lemma 6.2, the parameter g_2 can be regarded as a nuisance parameter, and $\Lambda_{\mathcal{F}_2}(F^*)$ is the corresponding nuisance tangent space.

Example 6.1 (continued)**.** Suppose that model F for $Z = (Z_1^\top, Z_2^\top)^\top$ is defined by the sole restriction that for each $F \in \mathcal{F}$ there exists $\beta(F) \in \mathbb{R}^k$ such that

$$\mathbb{E}_F(Z_2 \mid Z_1) = \beta(F)^\top Z_1.$$

Then, Lemma 6.2 implies that the C-R bound for $\beta(F)$ in a model that does not impose restrictions on the marginal law of the covariate Z_1 is the same as in a model in which the law of Z_1 is known (provided the class $\mathcal{A}$ satisfies the restrictions of Lemma 6.1 which are not at all restrictive).

6.3 Example: potential outcome mean estimation under MAR

Let F be the c.d.f. of a random vector $Z = (Y, R, X^\top)^\top$, where R is a binary r.v., Y is a scalar r.v. (discrete or continuous), and X is a random vector with discrete and/or continuous components.

We will study inference about the parameter

$$\beta(F) = \mathbb{E}_F[\mathbb{E}_F(Y \mid R = 1, X)]$$

under three different models, namely:

1. $\mathcal{F}_{\text{np}} = \{F \colon \mathbb{E}_F[\{\mathbb{E}_F(Y \mid R = 1, X)\}^2] < \infty, \mathbb{P}_F(R = 1 \mid X) > \sigma_F > 0\}$,
2. $\mathcal{F}_{\text{sem,fixed}} = \{F \in \mathcal{F}_{\text{np}} \colon \mathbb{P}_F(R = 1 \mid X) = \pi^*(X)\}$ where $\pi^*(x)$ is known,
3. $\mathcal{F}_{\text{sem,par}} = \{F \in \mathcal{F}_{\text{np}} \colon \mathbb{P}_F(R = 1 \mid X) = \pi(X; \alpha), a \in \Xi \subseteq \mathbb{R}^r\}$ where $\pi(x; \alpha)$ is a known function of x and α, which is differentiable w.r.t. α at every x.

The motivation (in the missing data context) for $\beta(F)$ is as follows. Suppose that on a random sample of units from a population of interest, we always measure the random vector X, but we measure the vector Y^f only on a subsample, i.e., Y^f is missing in some study units.

Suppose that R is the missing data indicator, i.e.,

$$R = \begin{cases} 1 & \text{if } Y^f \text{ is observed} \\ 0 & \text{if } Y^f \text{ is missing} \end{cases}$$

and

$$Y = \begin{cases} Y^f & \text{if } Y^f \text{ is observed} \\ \text{NA} & \text{if } Y^f \text{ is missing} \end{cases}$$

Then on each unit we observe $(Y, R, X^\top)^\top$.

The outcome Y^f is said to be **missing at random** (MAR) iff

$$\mathbb{P}(R = 1 \mid Y^f, X) = \mathbb{P}(R = 1 \mid X).$$

There are some implications of the MAR assumption:

1. the MAR assumption essentially postulates that X contains all the predictors of the outcome Y^f that are associated with non-response;
2. the MAR assumption is untestable, i.e., it does not impose any restriction on the law of the observed data;
3. the MAR assumption holds in two-stage study designs, where at the first stage, a cheap surrogate X for the outcome Y^f is measured on all study units and at the second stage, an expensive outcome Y^f is measured in a subsample selected with known probability that may depend on X.

Under the MAR assumption, the mean of Y^f happens to be equal to $\beta(F)$. To see this, let F^f denote the law of the "full" data vector $(Y^f, R, X^\top)^\top$ and let F denote the law of $(Y, R, X^\top)^\top$ implied by F^f. Then

$$\begin{aligned}\mathbb{E}_{F^f}(Y^f) &= \mathbb{E}_{F^f}[\mathbb{E}_{F^f}(Y^f \mid X)] \\ &= \mathbb{E}_{F^f}[\mathbb{E}_{F^f}(Y^f \mid R=1, X)] \quad \text{(by MAR)} \\ &= \mathbb{E}_F[\mathbb{E}_F(Y \mid R=1, X)] \quad \text{(by the definition of } Y) \\ &= \beta(F).\end{aligned}$$

In the following, we will investigate on the semiparametric inference in the three mentioned models $\mathcal{F}_{\text{np}}$, $\mathcal{F}_{\text{sem,fixed}}$ and $\mathcal{F}_{\text{sem,par}}$.

Characterization of the class $\mathcal{A}_{\text{np}}$. First, we introduce the class of regular parametric submodels $\mathcal{A}_{\text{np}}$ of the nonparametric model $\mathcal{F}_{\text{np}}$ with respect to which we will compute the efficient influence function and the C-R bound. Consider an arbitrary parametric submodel

$$\mathcal{F}_{\text{sub}} = \{f_{Y,R,X}(y, r, x; \theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\}$$

through F^* at θ^*, i.e., $F^* = F_{\theta^*}$. Writing

$$f_{Y,R,X}(y, r, x; \theta) = f_X(x; \theta) f_{R|X}(r \mid x; \theta) f_{Y|R,X}(y \mid r, x; \theta),$$

We see that $\mathcal{F}_{\text{sub}}$ implies a parametric submodel for each of the following:

1. the marginal law of X,

$$\mathcal{F}_{X,\text{sub}} = \{f_X(x; \theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\};$$

2. for each x, the conditional probability of R given $X = x$,

$$\mathcal{F}_{R|X=x,\text{sub}} = \{f_{R|X}(r \mid x; \theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\};$$

3. for each $R = r$ and $X = x$, the law of Y given $(R = r, X = x)$,

$$\mathcal{F}_{Y|R,X,\text{sub}} = \{f_{Y|R,X}(y \mid r, x; \theta) \colon \theta \in \Theta \subseteq \mathbb{R}^p\}.$$

Suppose $\mathcal{F}_{\text{sub}}$ is regular and goes through F^* at θ^*, i.e., $F^* = F_{\theta^*}$. Its score at θ^* is denoted as $s_\theta(Y, R, X; \theta^*) = S_\theta(\theta^*)$. Assumption (1.a) requires that $\mathcal{A}_{\text{np}}$ be comprised of submodels $\mathcal{F}_{\text{sub}}$ such that the implied submodels $\mathcal{F}_{X,\text{sub}}$, $\mathcal{F}_{R|X=x,\text{sub}}$ for all x, and $\mathcal{F}_{Y|R=r,X=x,\text{sub}}$ for all (r, x) are all regular. We denote the scores at θ^* in each submodel $\mathcal{F}_{X,\text{sub}}$, $\mathcal{F}_{R|X=x,\text{sub}}$ and $\mathcal{F}_{Y|R=r,X=x,\text{sub}}$ respectively with

$$s_{X,\theta}(X; \theta^*), \quad s_{R|X,\theta}(R, x; \theta^*), \quad s_{Y|R,X,\theta}(Y, r, x; \theta^*).$$

Note that by the mean zero property of scores we have that

$$
\begin{aligned}
0 &= \mathbb{E}_{F^*_X}\left[s_{X,\theta}(X;\theta^*)\right], \\
0 &= \mathbb{E}_{F^*_{R|X}}\left[s_{R|X,\theta}(R,X;\theta^*) \mid X = x\right], \\
0 &= \mathbb{E}_{F^*_{Y|R,X}}\left[s_{Y|R,X,\theta}(Y,R,X;\theta^*) \mid R = r, X = x\right].
\end{aligned}
$$

In addition, each parametric submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}_{\text{np}}$ needs to satisfy the following conditions:

1. the following map is continuous at θ^*

$$\theta \to \mathbb{E}_\theta\left[\mathbb{E}_{F^*}(Y \mid R = 1, X)^2\right];$$

2. for each x, the following map is continuous at θ^*

$$\theta \to \mathbb{E}_\theta\left(Y^2 \mid R = 1, X = x\right);$$

3. the derivative in the left-hand side below exists and equals the expression in the right-hand side

$$\frac{\partial}{\partial \theta^\top}\mathbb{E}_{\theta^*}\left\{\mathbb{E}_\theta(Y \mid R = 1, X)\right\}\Big|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}\left\{\frac{\partial}{\partial \theta^\top}\mathbb{E}_\theta(Y \mid R = 1, X)\Big|_{\theta=\theta^*}\right\};$$

4. the scores $s_\theta(Y,R,X;\theta^*)$, $s_{X,\theta}(X;\theta^*)$, $s_{R|X,\theta}(R,x;\theta^*)$ and $s_{Y|R,X,\theta}(Y,r,x;\theta^*)$ satisfy

$$s_\theta(Y,R,X;\theta^*) = s_{X,\theta}(X;\theta^*) + s_{R|X,\theta}(R,X;\theta^*) + s_{Y|R,X,\theta}(Y,R,X;\theta^*).$$

(Note that this last condition holds if scores are computed as derivatives of the log-likelihood.)

Calculation of the gradients in $\mathcal{F}_{\text{np}}$. Now we can derive the gradients of $\beta(F)$ w.r.t. the class $\mathcal{A}_{\text{np}}$ at a given F^*. Let $\mathcal{F}_{\text{sub}} = \{F_\theta \colon \theta \in \Theta\}$ be a submodel in $\mathcal{A}_{\text{np}}$. We have

$$
\begin{aligned}
\frac{\partial \beta(\theta)}{\partial \theta^\top}\Big|_{\theta=\theta^*} &= \frac{\partial}{\partial \theta^\top}\mathbb{E}_\theta\left\{\mathbb{E}_\theta(Y \mid R = 1, X)\right\}\Big|_{\theta=\theta^*} \\
&= \frac{\partial}{\partial \theta^\top}\mathbb{E}_\theta\left\{\mathbb{E}_{\theta^*}(Y \mid R = 1, X)\right\}\Big|_{\theta=\theta^*} + \frac{\partial}{\partial \theta^\top}\mathbb{E}_{\theta^*}\left\{\mathbb{E}_\theta(Y \mid R = 1, X)\right\}\Big|_{\theta=\theta^*} \\
&= \mathbb{E}_{\theta^*}\left\{\mathbb{E}_{\theta^*}(Y \mid R = 1, X) s_{X,\theta}(X;\theta^*)^\top\right\} + \mathbb{E}_{\theta^*}\left\{\frac{\partial}{\partial \theta^\top}\mathbb{E}_\theta(Y \mid R = 1, X)\right\}\Big|_{\theta=\theta^*} \\
&= \mathbb{E}_{\theta^*}\left\{\mathbb{E}_{\theta^*}(Y \mid R = 1, X) s_{X,\theta}(X;\theta^*)^\top\right\} \\
&\quad + \mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left\{Y\, s_{Y|R=1,X;\theta}(Y, R = 1, X;\theta^*)^\top \mid R = 1, X\right\}\right].
\end{aligned}
$$

We need to write the above two terms in the form of $\mathbb{E}_{\theta^*}\left[\psi_{F^*}(Y,R,X)S_\theta(\theta^*)^\top\right]$. Letting

$$s_{Y,R|X,\theta}(Y,R,X;\theta^*) = s_{Y|R,X,\theta}(Y,R,X;\theta^*) + s_{R|X,\theta}(R,X;\theta^*),$$

we have

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left\{\mathbb{E}_{\theta^*}(Y \mid R=1,X)s_{Y,R|X,\theta}(Y,R,X;\theta^*)^\top \mid X\right\}\\ =\ &\mathbb{E}_{\theta^*}(Y \mid R=1,X)\underbrace{\mathbb{E}_{\theta^*}\left\{s_{Y,R|X,\theta}(Y,R,X;\theta^*)^\top \mid X\right\}}_{=0}.\end{aligned}$$

Therefore,

$$\mathbb{E}_{\theta^*}\left\{\mathbb{E}_{\theta^*}(Y \mid R=1,X)s_{X,\theta}(X;\theta^*)^\top\right\} = \mathbb{E}_{\theta^*}\left\{\mathbb{E}_{\theta^*}(Y \mid R=1,X)S_\theta(\theta^*)^\top\right\}, \tag{6.9}$$

where $S_\theta(\theta^*) = s_{X,\theta}(X;\theta^*) + s_{Y,R|X,\theta}(Y,R,X;\theta^*)$ is the score at θ^* in the submodel $\mathcal{F}_{\text{sub}}$. On the other hand,

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left\{Y s_{Y|R=1,X,\theta}(Y,R=1,X;\theta^*)^\top \mid R=1,X\right\}\right]\\ \overset{(1)}{=}\ &\mathbb{E}_{\theta^*}\left[\mathbb{E}_{\theta^*}\left\{\frac{RY}{\mathbb{P}_{\theta^*}(R=1 \mid X)}s_{Y|R=1,X,\theta}(Y,R=1,X;\theta^*)^\top \mid X\right\}\right]\\ \overset{(2)}{=}\ &\mathbb{E}_{\theta^*}\left[\frac{RY}{\mathbb{P}_{\theta^*}(R=1 \mid X)}s_{Y|R,X,\theta}(Y,R,X;\theta^*)^\top\right]\\ \overset{(3)}{=}\ &\mathbb{E}_{\theta^*}\left[\frac{R\{Y-\mathbb{E}_{\theta^*}(Y \mid R,X)\}}{\mathbb{P}_{\theta^*}(R=1 \mid X)}s_{Y|R,X,\theta}(Y,R,X;\theta^*)^\top\right]\\ \overset{(4)}{=}\ &\mathbb{E}_{\theta^*}\left[\frac{R\{Y-\mathbb{E}_{\theta^*}(Y \mid R,X)\}}{\mathbb{P}_{\theta^*}(R=1 \mid X)}\left[s_{Y|R,X,\theta}(Y,R,X;\theta^*)^\top + s_{R,X,\theta}(R,X;\theta^*)^\top\right]\right]\\ =\ &\mathbb{E}_{\theta^*}\left[\frac{R\{Y-\mathbb{E}_{\theta^*}(Y \mid R=1,X)\}}{\mathbb{P}_{\theta^*}(R=1 \mid X)}S_\theta(\theta^*)^\top\right],\end{aligned} \tag{6.10}$$

where (1) holds because for any W, $\mathbb{E}[W \mid R=1,X] = \mathbb{E}[RW \mid X]/\mathbb{P}(R=1 \mid X)$; (2) holds because $Rs_{Y|R=1,X,\theta}(Y,R=1,X;\theta^*) = Rs_{Y|R,X,\theta}(Y,R,X;\theta^*)$; (3) holds because

$$\begin{aligned}&\mathbb{E}_{\theta^*}\left[\frac{R\mathbb{E}_{\theta^*}(Y \mid R,X)}{\mathbb{P}_{\theta^*}(R=1 \mid X)}s_{Y|R,X,\theta}(Y,R,X;\theta^*)^\top \mid R,X\right]\\ =&\frac{R\mathbb{E}_{\theta^*}(Y \mid R,X)}{\mathbb{P}_{\theta^*}(R=1 \mid X)}\underbrace{\mathbb{E}_{\theta^*}\left[s_{Y|R,X,\theta}(Y,R,X;\theta^*)^\top \mid R,X\right]}_{=0} = 0;\end{aligned}$$

(4) holds because $\mathbb{E}_{\theta^*}[q(R,X)\{Y-\mathbb{E}_{\theta^*}(Y \mid R,X)\}] = 0$ for any $q(R,X)$, in particular, for $q(R,X) = \frac{R}{\mathbb{P}_{\theta^*}(R=1|X)}s_{R,X,\theta}(R,X;\theta^*)^\top$ where $s_{R,X,\theta}(R,X;\theta^*) = s_{X,\theta}(X;\theta^*) + s_{R|X,\theta}(R \mid X;\theta^*)$. Combining (6.9) and (6.10) yields

$$\left.\frac{\partial\beta(\theta)}{\partial\theta^\top}\right|_{\theta=\theta^*} = \mathbb{E}_{\theta^*}\left\{\left[\mathbb{E}_{\theta^*}(Y \mid R=1,X) + \frac{R\{Y-\mathbb{E}_{\theta^*}(Y \mid R=1,X)\}}{\mathbb{P}_{\theta^*}(R=1 \mid X)}\right]S_\theta(\theta^*)^\top\right\}.$$

We therefore conclude that

$$\mathbb{E}_{F^*}(Y \mid R=1,X) + \frac{R\{Y-\mathbb{E}_{F^*}(Y \mid R=1,X)\}}{\mathbb{P}_{F^*}(R=1 \mid X)}$$

is a gradient of $\beta(F)$ at F^* w.r.t. $\mathcal{A}_{\text{np}}$ in $\mathcal{F}_{\text{np}}$. Furthermore, because

$$\begin{aligned}
&\mathbb{E}_{F^*}\left[\frac{R\{Y-\mathbb{E}_{F^*}(Y\mid R=1,X)\}}{\mathbb{P}_{\theta^*}(R=1\mid X)}\right]\\
&=\mathbb{E}_{F^*}\left\{\mathbb{E}_{F^*}\left[\frac{R\{Y-\mathbb{E}_{F^*}(Y\mid R=1,X)\}}{\mathbb{P}_{\theta^*}(R=1\mid X)}\mid R,X\right]\right\}\\
&=\mathbb{E}_{F^*}\left\{\frac{R\{\mathbb{E}_{F^*}(Y\mid R,X)-\mathbb{E}_{F^*}(Y\mid R=1,X)\}}{\mathbb{P}_{\theta^*}(R=1\mid X)}\right\}\\
&=\mathbb{E}_{F^*}\left\{\frac{R\{\mathbb{E}_{F^*}(Y\mid R=1,X)-\mathbb{E}_{F^*}(Y\mid R=1,X)\}}{\mathbb{P}_{\theta^*}(R=1\mid X)}\right\}\\
&=0,
\end{aligned}$$

and $\mathbb{E}_{F^*}[\mathbb{E}_{F^*}(Y\mid R=1,X)]=\beta(F^*)$, we conclude that a mean zero gradient of $\beta(F)$ at F^* w.r.t. $\mathcal{A}_{\text{np}}$ in $\mathcal{F}_{\text{np}}$ is

$$\psi_{F^*}(Y,R,X)=\mathbb{E}_{F^*}(Y\mid R=1,X)+\frac{R\{Y-\mathbb{E}_{F^*}(Y\mid R=1,X)\}}{\mathbb{P}_{F^*}(R=1\mid X)}-\beta(F^*).$$

Remark 6.2. Techniques used in deriving (6.10) are of much importance. The appearance of the inverse propensity score $\mathbb{P}_{\theta^*}(R=1\mid X)$ at step (1) aims at getting rid of conditioning on $R=1$, and the subtraction of the mean $\mathbb{E}_{\theta^*}(Y\mid R,X)$ at step (3) enables adding the rest of the scores $s_{R,X,\theta}(R,X;\theta^*)$.

Calculation of the tangent space of $\mathcal{A}_{\text{np}}$ in $\mathcal{F}_{\text{np}}$. The tangent space of $\mathcal{A}_{\text{np}}$ is indeed $\mathcal{L}_2^0(F^*)$. To see this, given any $s(Y,R,X)\in\mathcal{L}_2^0(F^*)$ we decompose it as

$$\begin{aligned}
s_1(X)&=\mathbb{E}_{F^*}[s(Y,R,X)\mid X],\\
s_2(R,X)&=\mathbb{E}_{F^*}[s(Y,R,X)\mid R,X]-\mathbb{E}_{F^*}[s(Y,R,X)\mid X],\\
s_3(Y,R,X)&=s(Y,R,X)-\mathbb{E}_{F^*}[s(Y,R,X)\mid R,X].
\end{aligned}$$

We can construct a submodel

$$\mathcal{F}_{\text{sub}}=\left\{f_{Y,R,X}(y,r,x;\theta)=f_X(x;\theta_1)f_{R|X}(r\mid x;\theta_2)f_{Y|R,X}(y\mid r,x;\theta_3)\colon \theta_j\in\mathbb{R}\right\}$$

where f_X, $f_{R|X}$ and $f_{Y|R,X}$ are of the form as in submodels in Remark 4.2. By Remark 4.2, the score at θ^* in $\mathcal{F}_{\text{sub}}$ is $s(Y,R,X)$. And we can easily verify that $\mathcal{F}_{\text{sub}}\in\mathcal{A}_{\text{np}}$ (by verifying some trivial regularity conditions).

Remark 6.3. There is a unique influence function in the nonparametric model, which is also the efficient influence function

$$\psi_{F^*,\text{eff,np}}(Y,R,X)=\mathbb{E}_{F^*}(Y\mid R=1,X)+\frac{R\{Y-\mathbb{E}_{F^*}(Y\mid R=1,X)\}}{\mathbb{P}_{F^*}(R=1\mid X)}-\beta(F^*).\tag{6.11}$$

The assumption $\mathbb{P}_{F^*}(R=1\mid X)>\sigma_F>0$ ensures that $\psi_{F^*,\text{eff,np}}(Y,R,X)$ has finite variance under F^*, i.e., $\psi_{F^*,\text{eff,np}}(Y,R,X)\in\mathcal{L}_2^0(F^*)$. Note also that for any $s_2(R,X)$ satisfying $\mathbb{E}_{F^*}[s_2(R,X)\mid X]=0$, it holds that

$$\mathbb{E}_{F^*}[\psi_{F^*,\text{eff,np}}(Y,R,X)s_2(R,X)]=0.$$

This is reasonable since the definition of $\beta(F)$ does not depend on the propensity score $\mathbb{P}_{F^*}(R = 1 \mid X)$, whose tangent space consists of these $s_2(R, X)$. In other words, our example is a factorized likelihood model with $\beta(F)$ not depending on the propensity score.

Calculation of the efficient influence function in $\mathcal{F}_{\text{sem,fixed}}$. Recall this model is restricted only by the condition that

$$\mathbb{P}_{F^*}(R = 1 \mid X = x) = \pi^*(x)$$

for a known function $\pi^*(x)$. Writing

$$\begin{aligned} g_1(y, r, x) &= f_X(x) f_{Y|R,X}(y \mid r, x), \\ g_2(y, r, x) &= f_{R|X}(r \mid x), \end{aligned}$$

we have that $f_{Y,R,X}(y, r, x) = g_1(y, r, x) g_2(y, r, x)$. Consequently, $\mathcal{F}_{\text{np}}$ is a factorized likelihood model and $\mathcal{F}_{\text{sem,fixed}}$ is like the model $\mathcal{F}_1$ in Section 6.2, in which g_2 is known and equals to

$$g_2^*(y, r, x) = \pi^*(x)^r \left\{1 - \pi^*(x)\right\}^{1-r}.$$

Furthermore,

$$\beta(F) = \mathbb{E}_F \left[\mathbb{E}_F(Y \mid R = 1, X)\right] = \int f_X(x) \int y f_{Y|R=1,X}(y \mid r = 1, x) \mathrm{d}y \mathrm{d}x$$

depends only on $g_1(y, r, x)$. To use results for factorized likelihood models, we require that submodels in $\mathcal{A}_{\text{np}}$ also satisfy the assumptions in Lemma 6.1. It can be checked that the tangent space w.r.t. this newly defined $\mathcal{A}_{\text{np}}$ remains equal to $\mathcal{L}_2^0(F^*)$, and the efficient influence function remains to be $\psi_{F^*,\text{eff,np}}(Y, R, X)$ in (6.11). Next, for each submodel in $\mathcal{A}_{\text{np}}$

$$\mathcal{F}_{\text{sub}} = \left\{f_{Y,R,X}(y, r, x; \theta) = f_X(x; \theta) f_{R|X}(r \mid x; \theta) f_{Y|R,X}(y \mid r, x; \theta) \colon \theta \in \mathbb{R}^p\right\},$$

define the submodel

$$\mathcal{F}_{\text{sub,fixed}} = \left\{f_{Y,R,X}(y, r, x; \theta) = f_X(x; \theta) f_{R|X}^*(r \mid x) f_{Y|R,X}(y \mid r, x; \theta) \colon \theta \in \mathbb{R}^p\right\},$$

where $f_{R|X}^*(r \mid x) = \pi^*(x)^r \left\{1 - \pi^*(x)\right\}^{1-r}$ is fixed and known. Also, define the class

$$\mathcal{A}_{\text{sem,fixed}} = \left\{\mathcal{F}_{\text{sub,fixed}} \colon \mathcal{F}_{\text{sub}} \in \mathcal{A}_{\text{np}}\right\}.$$

From Lemma 6.2 we now conclude that $\psi_{F^*,\text{eff,np}}(Y, R, X)$ is also the efficient influence function of $\beta(F)$ at F^* w.r.t. $\mathcal{A}_{\text{sem,fixed}}$ in the semiparametric model $\mathcal{F}_{\text{sem,fixed}}$ (to see this, let $\mathcal{A}_{\text{sem,fixed}}$ play the role of $\mathcal{A}_1$). And the C-R bound of $\beta(F)$ at F^* is the same in $\mathcal{F}_{\text{np}}$ w.r.t. $\mathcal{A}_{\text{np}}$ as in $\mathcal{F}_{\text{sem,fixed}}$ w.r.t. $\mathcal{A}_{\text{sem,fixed}}$.

Calculation of the efficient influence function in $\mathcal{F}_{\text{sem,par}}$. Finally, consider the semiparametric model $\mathcal{F}_{\text{sem,par}}$. It is also a factorized likelihood model with

$$\begin{aligned} f(y,r,x) &= g_1(y,r,x)g_2(y,r,x;\alpha), \\ g_1(y,r,x) &= f_X(x;\theta)f_{Y|R,X}(y \mid r,x;\theta), \\ g_2(y,r,x;\alpha) &= \pi(x;\alpha)^r\{1-\pi(x;\alpha)\}^{1-r}. \end{aligned}$$

Next, for each submodel in $\mathcal{A}_{\text{np}}$

$$\mathcal{F}_{\text{sub}} = \left\{ f_{Y,R,X}(y,r,x;\theta) = f_X(x;\theta)f_{R|X}(r \mid x;\theta)f_{Y|R,X}(y \mid r,x;\theta) \colon \theta \in \mathbb{R}^p \right\},$$

define the submodel

$$\mathcal{F}_{\text{sub,par}} = \left\{ f_{Y,R,X}(y,r,x;\theta) = f_X(x;\theta)f_{R|X}(r \mid x;\alpha)f_{Y|R,X}(y \mid r,x;\theta) \colon \theta \in \mathbb{R}^p, \alpha \in \mathbb{R}^r \right\}.$$

Also, define the class

$$\mathcal{A}_{\text{sem,par}} = \{ \mathcal{F}_{\text{sub,par}} \colon \mathcal{F}_{\text{sub}} \in \mathcal{A}_{\text{np}} \}.$$

From Lemma 6.2 with $\mathcal{A}_{\text{sem,par}}$ and $\mathcal{F}_{\text{sem,par}}$ playing the role of $\mathcal{A}$ and $\mathcal{F}$ in that lemma, and $\mathcal{A}_{\text{sem,fixed}}$ and $\mathcal{F}_{\text{sem,fixed}}$ playing the role of $\mathcal{A}_1$ and $\mathcal{F}_1$ in that lemma, we conclude that the efficient influence functions for $\beta(F)$ in both models are the same. Since we have just shown that the efficient influence function w.r.t. $\mathcal{A}_{\text{sem,fixed}}$ is $\psi_{F^*,\text{eff,np}}(Y,R,X)$, it is also the efficient influence function w.r.t. $\mathcal{A}_{\text{sem,par}}$ in the semiparametric model $\mathcal{F}_{\text{sem,fixed}}$.

Remark 6.4. We thus conclude that: the semiparametric C-R bound of $\beta(F)$ is the same regardless of whether the missingness probabilities are known, modeled, or fully unspecified.

An application of the preceding results is that examination of the C-R bound allows us to quantify the information loss for estimating $\mathbb{E}_{F^f}(Y^f)$ due to Y^f missing in some study subjects and the value of measuring the predictor X for reducing the information loss.

Rewrite the efficient influence function as

$$\begin{aligned} &\psi_{F^*,\text{eff,np}}(Y,R,X) \\ &= \mathbb{E}_{F^*}(Y \mid R=1,X) + \frac{R\{Y - \mathbb{E}_{F^*}(Y \mid R=1,X)\}}{\mathbb{P}_{F^*}(R=1 \mid X)} - \beta(F^*) \\ &= \underbrace{\frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)}}_{\text{IPW}} Y - \underbrace{\left\{ \frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)} - 1 \right\} \mathbb{E}_{F^*}(Y \mid R=1,X)}_{\text{augmentation based on the conditional outcome mean}} - \beta(F^*). \end{aligned}$$

Since by definition, $Y = Y^f$ if $R = 1$, so

$$\begin{aligned} &\psi_{F^*,\text{eff,np}}(Y,R,X) \\ &= \frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)} Y^f - \left\{ \frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)} - 1 \right\} \mathbb{E}_{F^*}(Y \mid R=1,X) - \beta(F^*) \\ &= \underbrace{\{Y^f - \beta(F^*)\}}_{\text{efficient influence function if data is not missing}} + \left\{ \frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)} - 1 \right\} \{Y^f - \mathbb{E}_{F^*}(Y \mid R=1,X)\}. \end{aligned}$$

Thus the C-R bound is

$$\begin{aligned}
&\operatorname{var}_{F^*}\left[\psi_{F^*,\text{eff,np}}(Y,R,X)\right] \\
&= \operatorname{var}_{F^{*f}}\left[\left\{Y^f - \beta(F^*)\right\} + \left\{\frac{R}{\mathbb{P}_{F^{*f}}(R=1 \mid X)} - 1\right\}\left\{Y^f - \mathbb{E}_{F^{*f}}(Y \mid R=1, X)\right\}\right] \\
&= \underbrace{\operatorname{var}_{F^{*f}}\left(Y^f\right)}_{\text{C-R bound if data is not missing}} + \underbrace{\operatorname{var}_{F^{*f}}\left[\left\{\frac{R}{\mathbb{P}_{F^{*f}}(R=1 \mid X)} - 1\right\}\left\{Y^f - \mathbb{E}_{F^{*f}}(Y \mid R=1, X)\right\}\right]}_{\text{penalty for not observing } Y^f \text{ under MAR}}
\end{aligned}$$

The penalty term can be further simplified as

$$\mathbb{E}_{F^{*f}}\left[\frac{\left\{1 - \mathbb{P}_{F^{*f}}(R=1 \mid X)\right\}}{\mathbb{P}_{F^{*f}}(R=1 \mid X)} \operatorname{var}_{F^{*f}}\left(Y^f \mid X\right)\right]$$

Note that if X is a perfect predictor of Y^f, then $\operatorname{var}_{F^{*f}}\left(Y^f \mid X\right) = 0$ and the penalty is 0. If no X is observed, then the penalty is $\mathbb{E}_{F^{*f}}\left[\frac{\left\{1-\mathbb{P}_{F^{*f}}(R=1|X)\right\}}{\mathbb{P}_{F^{*f}}(R=1|X)} \operatorname{var}_{F^{*f}}\left(Y^f\right)\right]$.

In the following, we derive the set of mean zero gradients of $\beta(F)$ w.r.t. $\mathcal{A}_{\text{sem,fixed}}$ in $\mathcal{F}_{\text{sem,fixed}}$ and w.r.t. $\mathcal{A}_{\text{sem,par}}$ in $\mathcal{F}_{\text{sem,par}}$. For each set, it suffices to characterize the orthogonal complement of the tangent space.

Lemma 6.3. *Suppose that H_1, H_2 and H_3 are three subspaces of a Hilbert space such that*

$$H_1 = H_2 \oplus H_3,$$

then

$$H_2^\perp = H_1^\perp \oplus H_3.$$

In particular, if

$$\mathcal{L}_2^0(F^*) = H_2 \oplus H_3,$$

then

$$H_2^\perp = H_3.$$

To compute $\Lambda_{\mathcal{F}_{\text{sem,fixed}}}(F^*)$, we recall that since $\mathcal{F}_{\text{np}}$ is a factorized likelihood model, the class $\mathcal{A}_{\text{np}}$ verifies the assumptions of Lemma 6.1 and $\mathcal{A}_{\text{sem,fixed}}$ is defined as $\mathcal{A}_1$ in that lemma, then we can apply the conclusions of Lemma 6.1.

The analogous of $\mathcal{F}_2$ in Lemma 6.1 corresponds to

$$\mathcal{F}_{\text{sem,miss}} = \left\{f_{Y,R,X}(y,r,x) = f_X^*(x) f_{R|X}(r \mid x) f_{Y|R,X}^*(y \mid r, x) \colon f_{R|X}(1 \mid x) > \sigma_F > 0\right\}.$$

To define the analogy of $\mathcal{A}_2$, for every submodel $\mathcal{F}_{\text{sub}} \in \mathcal{A}_{\text{np}}$ we consider the submodel

$$\mathcal{F}_{\text{sub,miss}} = \left\{f_{Y,R,X}(y,r,x;\theta) = f_X^*(x) f_{R|X}(r \mid x;\theta) f_{Y|R,X}^*(y \mid r, x) \colon \theta \in \mathbb{R}^p\right\}$$

and then define (the analogy of $\mathcal{A}_2$)

$$\mathcal{A}_{\text{sem,miss}} = \left\{\mathcal{F}_{\text{sub,miss}} \colon \mathcal{F}_{\text{sub}} \in \mathcal{A}_{\text{np}}\right\}.$$

Once again, we can show the tangent space $\Lambda_{\mathcal{F}_{\text{sem,miss}}}(F^*)$ is given by

$$\Lambda_{\mathcal{F}_{\text{sem,miss}}}(F^*) = \left\{ s(R,X) \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}_{F^*}[s(R,X) \mid X] = 0 \right\}.$$

By Lemmas 6.1 and 6.3, we conclude that $\Lambda_{\mathcal{F}_{\text{sem,fixed}}}(F^*)^\perp = \Lambda_{\mathcal{F}_{\text{sem,miss}}}(F^*)$. Note that $\mathbb{E}_{F^*}[s(R,X) \mid X] = 0$ iff $s(R,X) = d(X)\,[R - \mathbb{P}_{F^*}(R=1 \mid X)]$ for some $d(X)$, and the efficient influence function can be written as

$$\begin{aligned}
\psi_{F^*,\text{eff,np}}(Y,R,X) &= \mathbb{E}_{F^*}(Y \mid R=1,X) + \frac{R\{Y - \mathbb{E}_{F^*}(Y \mid R=1,X)\}}{\mathbb{P}_{F^*}(R=1 \mid X)} - \beta(F^*) \\
&= \frac{RY}{\mathbb{P}_{F^*}(R=1 \mid X)} - \left\{ \frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)} - 1 \right\} \mathbb{E}_{F^*}(Y \mid R=1,X) - \beta(F^*) \\
&= \frac{R\{Y - \beta(F^*)\}}{\mathbb{P}_{F^*}(R=1 \mid X)} - \{R - \mathbb{P}_{F^*}(R=1 \mid X)\} \underbrace{\frac{\{\mathbb{E}_{F^*}(Y \mid R=1,X) - \beta(F^*)\}}{\mathbb{P}_{F^*}(R=1 \mid X)}}_{\text{a function of } X \text{ only}},
\end{aligned}$$

so the mean zero gradients of $\beta(F)$ at F^* w.r.t. $\mathcal{A}_{\text{sem,fixed}}$ in $\mathcal{F}_{\text{sem,fixed}}$ are of the form

$$\psi_{F^*}(Y,R,X) = \frac{R}{\mathbb{P}_{F^*}(R=1 \mid X)} \{Y - \beta(F^*)\} + d(X)\{R - \mathbb{P}_{F^*}(R=1 \mid X)\}.$$

We now derive the set of mean zero gradients of $\beta(F)$ at F^* w.r.t. $\mathcal{A}_{\text{sem,par}}$ in $\mathcal{F}_{\text{sem,par}}$. It is easy to show the tangent space $\Lambda_{\mathcal{F}_{\text{sem,par}}}(F^*)$ is given by

$$\Lambda_{\mathcal{F}_{\text{sem,par}}}(F^*) = \Omega_1(F^*) \oplus \Omega_3(F^*) \oplus \Omega_{2,\text{par}}(F^*),$$

where

$$\begin{aligned}
\Omega_1(F^*) &= \left\{ s_1(X) \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}_{F^*}[s_1(X)] = 0 \right\}, \\
\Omega_3(F^*) &= \left\{ s_3(Y,R,X) \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}_{F^*}[s_3(Y,R,X) \mid R,X] = 0 \right\}, \\
\Omega_{2,\text{par}}(F^*) &= \left\{ a^\top s_\alpha(R,X;\alpha^*) \colon a \in \mathbb{R}^r \right\},
\end{aligned}$$

and

$$s_\alpha(R,X;\alpha^*) = \frac{\mathrm{d}}{\mathrm{d}\alpha} \log\left[\pi(X;\alpha)^R \{1 - \pi(X;\alpha)\}^{1-R} \right] \Big|_{\alpha=\alpha^*}.$$

Since we have $\mathcal{L}_2^0(F^*) = \Omega_1(F^*) \oplus \Omega_3(F^*) \oplus \Omega_2(F^*)$, where

$$\Omega_2(F^*) = \left\{ s_2(R,X) \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}_{F^*}[s_2(R,X) \mid X] = 0 \right\},$$

and $\Omega_2(F^*) \subseteq \Omega_{2,\text{par}}(F^*)$, we have

$$\Lambda_{\mathcal{F}_{\text{sem,par}}}(F^*)^\perp = \Omega_{2,\text{resid}}(F^*) = \{ s_2(R,X) - \Pi[s_2(R,X) \mid \Omega_{2,\text{par}}(F^*)] : s_2(R,X) \in \Omega_2(F^*) \}.$$

The mean zero gradients of $\beta(F)$ at F^* w.r.t. $\mathcal{A}_{\text{sem,par}}$ in $\mathcal{F}_{\text{sem,par}}$ are of the form

$$\begin{aligned}
\psi_{F^*}(X) = &\frac{R[Y - \beta(F^*)]}{\mathbb{P}_{F^*}(R=1 \mid X)} + d(X)[R - \mathbb{P}_{F^*}(R=1 \mid X)] \\
&- \Pi_{F^*}\left[\frac{R[Y - \beta(F^*)]}{\mathbb{P}_{F^*}(R=1 \mid X)} + d(X)[R - \mathbb{P}_{F^*}(R=1 \mid X)] \mid \Omega_{2,\text{par}}(F^*) \right].
\end{aligned}$$

Other semiparametric models. In this section, we have discussed three semiparametric models $\mathcal{F}_{\text{np}}$, $\mathcal{F}_{\text{sem,fixed}}$ and $\mathcal{F}_{\text{sem,par}}$, in which the propensity scores are modeled differently. In Appendix B.1, we will consider another two semiparametric models, in which the outcome models $\mathbb{E}(Y \mid X, R = 1)$ are different. We will see in this case that the efficient influence functions and the semiparametric C-R bounds are different in the nonparametric model, the fixed outcome model, and the parametric outcome model. We will also consider an easy MNAR case in Section B.2, in which the propensity score has a logistic form and the interaction function between X and Y is known.

7 Asymptotic Theory for the Semiparametric One-Step Estimator

Consider a semiparametric model for the law F of a random vector Z,

$$\mathcal{F} = \{F_{\eta,\vartheta} \colon \eta \in \Xi, \vartheta \in O\},$$

where both η and ϑ can be infinite dimensional. Suppose for each $(\eta, \vartheta) \in \Xi \times O$, $\beta(F)$ is a pathwise differentiable parameter at $F_{\eta,\vartheta}$ w.r.t. $\mathcal{A}$ in $\mathcal{F}$, and let $\psi_{F_{\eta,\vartheta}}(Z)$ be a mean zero gradient of $\beta(F_{\eta,\vartheta})$ at $F_{\eta,\vartheta}$. In addition, suppose that both $\beta(F_{\eta,\vartheta})$ and $\psi_{F_{\eta,\vartheta}}(Z)$ depend on (η, ϑ) only through η, so that they can be written as $\beta(\eta)$ and $\psi(Z; \eta)$ for short.

Example 7.1 (Missing data example in Section 6.3)**.** In this example, $Z = (Y, R, X^\top)^\top$ where R is a binary r.v., Y is a continuous scalar r.v., and X is a random vector. The parameter of interest is

$$\beta(F) = \mathbb{E}_F[\mathbb{E}_F(Y \mid R = 1, X)].$$

Note that this parameter depends only on

$$b_F(\cdot) = \mathbb{E}_F(Y \mid R = 1, X = \cdot) \quad \text{and} \quad f_X(\cdot).$$

In Section 6.3 we have shown that in three semiparametric models $\mathcal{F}_{\text{np}}$, $\mathcal{F}_{\text{sem,fixed}}$ and $\mathcal{F}_{\text{sem,par}}$, the efficient influence function at F is always

$$\psi_F(Y, R, X) = \mathbb{E}_F(Y \mid R = 1, X) + \frac{R\{Y - \mathbb{E}_F(Y \mid R = 1, X)\}}{\mathbb{P}_F(R = 1 \mid X)} - \beta(F). \tag{7.1}$$

Note that $\psi_F(Y, R, X)$ depends on F only through

$$b_F(\cdot) = \mathbb{E}_F(Y \mid R = 1, X = \cdot) \quad \text{and} \quad \pi_F(\cdot) = \mathbb{P}_F(R = 1 \mid X = \cdot).$$

Thus we can define

$$\eta = (b_F, \pi_F, F_X) \quad \text{and} \quad \vartheta = (F_{\varepsilon|R=1,X}),$$

where $\varepsilon = Y - \mathbb{E}_F(Y \mid R = 1, X)$, and the previous assumptions hold for the model $\mathcal{F}_{\text{np}}$, the parameter $\beta(F)$ and the gradient $\psi_F(Y, R, X)$.

In the following, we study a general estimation strategy for $\beta(\eta)$. We will evaluate in particular, a set of conditions under which our strategy yields a RAL estimator of $\beta(\eta)$ and in particular, a set of conditions under which the influence function is equal to a given gradient of $\beta(\eta)$, say, $\psi(Z; \eta)$.

7.1 The one-step estimator

Sample splitting. Given i.i.d. data $Z_1, \ldots, Z_n \sim F_{\eta,\vartheta}$, we define $m = \lfloor n/2 \rfloor$ and divide the data into two groups, and estimate η based on the two groups of data, i.e.,

$$\widetilde{\eta}_1 = \widetilde{\eta}_m(Z_1, \ldots, Z_m), \quad \widetilde{\eta}_2 = \widetilde{\eta}_{n-m}(Z_{m+1}, \ldots, Z_n).$$

Now, define

$$\widehat{\beta}_1 = \beta(\widetilde{\eta}_2) + \frac{1}{m}\sum_{i=1}^{m} \psi(Z_i; \widetilde{\eta}_2), \quad \widehat{\beta}_2 = \beta(\widetilde{\eta}_1) + \frac{1}{n-m}\sum_{i=m+1}^{n} \psi(Z_i; \widetilde{\eta}_1).$$

Here, in the definition of $\widehat{\beta}_1$, the term $\beta(\widetilde{\eta}_2)$ is a preliminary plug-in estimator based on the second half of the data $Z_{m+1}, \ldots, Z_n$ (training sample), and the term $\frac{1}{m}\sum_{i=1}^{m} \psi(Z_i; \widetilde{\eta}_2)$ is an estimator of $e(\widetilde{\eta}_2)$ based on the first half of the data $Z_1, \ldots, Z_m$ (validation sample), where for any η',

$$e(\eta') = \mathbb{E}_{\eta,\vartheta}[\psi(Z; \eta')] = \int \psi(z; \eta') f_{\eta,\vartheta}(z) \mathrm{d}z. \tag{7.2}$$

Clearly, $e(\eta) = 0$. In the definition of $\widehat{\beta}_2$, two groups of the data switch their roles.

Averaging. The final one-step estimator is defined by averaging $\widehat{\beta}_1$ and $\widehat{\beta}_2$:

$$\widehat{\beta} = \frac{m}{n}\widehat{\beta}_1 + \frac{n-m}{n}\widehat{\beta}_2.$$

Remark 7.1. Often in research, we can divide the data into K groups $\{\mathbb{I}_k\}_{k=1}^K$ of equal size $m = n/K$, and construct K estimators $\{\widetilde{\eta}_k\}_{k=1}^K$ with each estimator $\widetilde{\eta}_k$ based only on the data in $\mathbb{I}_k^c$. Such K estimators for η result in K estimators $\{\widehat{\beta}_k\}_{k=1}^K$ for $\beta(\eta)$, define as

$$\widehat{\beta}_k = \beta(\widetilde{\eta}_k) + \frac{1}{m}\sum_{i \in \mathbb{I}_k} \psi(Z_i; \widetilde{\eta}_k),$$

and the final one-step estimator is given by

$$\widehat{\beta} = \frac{1}{K}\sum_{k=1}^{K} \widehat{\beta}_k.$$

Example 7.1 (continued)**.** In the missing data example, we consider the nonparametric model $\mathcal{F}_{\text{np}}$, but nevertheless, with the purpose of estimating η, we specify "parametric working models" for b and π. Suppose $\widetilde{\eta}_2 = (\widetilde{b}_2, \widetilde{\pi}_2, \widetilde{F}_{2,X})$, where

- $\widetilde{F}_{2,X}$ is the empirical marginal distribution of X in the training sample;
- $\widetilde{b}_2(X) = \widetilde{\gamma}^\top \widetilde{X}$ where $\widetilde{X} = (1, X^\top)^\top$ and $\widetilde{\gamma}_2$ solves the least squares equation

$$\sum_{i=m+1}^{n} R_i \widetilde{X}_i \left[Y_i - \widetilde{X}_i^\top \gamma \right] = 0;$$

- $\widetilde{\pi}_2(X) = \pi(X; \widetilde{\alpha}_2)$ where $\widetilde{\alpha}_2$ is the MLE of the logistic regression model

$$\pi(X; \alpha) = \frac{e^{\alpha^\top \widetilde{X}}}{1 + e^{\alpha^\top \widetilde{X}}},$$

  i.e., solving

$$\sum_{i=m+1}^{n} \widetilde{X}_i \left[R_i - \frac{e^{\alpha^\top \widetilde{X}}}{1 + e^{\alpha^\top \widetilde{X}}} \right] = 0.$$

Then, with $\psi(Z; \eta)$ being the unique mean zero gradient for $\beta(F)$ in $\mathcal{F}_{\text{np}}$, we obtain

$$\begin{aligned}
\widehat{\beta}_1 &= \beta(\widetilde{\eta}_2) + \frac{1}{m} \sum_{i=1}^{m} \psi(Z_i; \widetilde{\eta}_2) \\
&= \beta(\widetilde{\eta}_2) + \frac{1}{m} \sum_{i=1}^{m} \left[\widetilde{b}_2(X_i) + \frac{R_i\{Y_i - \widetilde{b}_2(X_i)\}}{\widetilde{\pi}_2(X_i)} - \beta(\widetilde{\eta}_2) \right] \\
&= \frac{1}{m} \sum_{i=1}^{m} \left[\widetilde{b}_2(X_i) + \frac{R_i\{Y_i - \widetilde{b}_2(X_i)\}}{\widetilde{\pi}_2(X_i)} \right].
\end{aligned}$$

7.2 Asymptotic decomposition of $\sqrt{m}\{\widehat{\beta}_1 - \beta(\eta)\}$

Now we want to know the asymptotic properties of $\sqrt{n}\{\widehat{\beta} - \beta(\eta)\}$. Actually, it suffices to focus on $\widehat{\beta}_1$. Suppose $\widetilde{\eta}_2 \xrightarrow{p} \eta^*$ (in a sense to be defined later), which is not necessarily the true η. We have

$$\begin{aligned}
\sqrt{m}\{\widehat{\beta}_1 - \beta(\eta)\} = & \underbrace{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi(Z_i; \widetilde{\eta}_2) - e(\widetilde{\eta}_2)] - \frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi(Z_i; \eta^*) - e(\eta^*)]}_{A_m} \\
& + \underbrace{\frac{1}{\sqrt{m}} \sum_{i=1}^{m} [\psi(Z_i; \eta^*) - e(\eta^*)]}_{B_m} \\
& + \underbrace{\sqrt{m}\{\beta(\widetilde{\eta}_2) - \beta(\eta)\} + \sqrt{m} e(\widetilde{\eta}_2)}_{C_m},
\end{aligned}$$

where $e(\eta')$ is defined as (7.2).

The term A_m. We have assumed that $\widetilde{\eta}_2 \xrightarrow{p} \eta^*$ in the sense that

$$\int \|\psi(z; \widetilde{\eta}_2) - \psi(z; \eta^*)\|^2 f(z; \eta, \vartheta) \mathrm{d}z \xrightarrow[m \to \infty]{P(F_{\eta,\vartheta})} 0, \tag{7.3}$$

and we will show $A_m \xrightarrow[m \to \infty]{P(F_{\eta,\vartheta})} 0$ with this assumption. Recall that $\widetilde{\eta}_2$ depends on the training data $Z_{m+1}, \ldots, Z_n$ which is independent of the validation data $Z_1, \ldots, Z_m$.

Therefore, we have

$$\begin{aligned}
&\mathbb{E}_{\eta,\vartheta}\left[A_m \mid Z_{m+1},\ldots,Z_n\right] \\
&= \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\left\{\mathbb{E}_{\eta,\vartheta}\left[\psi(Z_i;\widetilde{\eta}_2)-\psi(Z_i;\eta^*)\mid Z_{m+1},\ldots,Z_n\right]-\left[e(\widetilde{\eta}_2)-e(\eta^*)\right]\right\} \\
&= \sqrt{m}\left\{\mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\widetilde{\eta}_2)-\psi(Z;\eta^*)\mid Z_{m+1},\ldots,Z_n\right]-\left[e(\widetilde{\eta}_2)-e(\eta^*)\right]\right\} \\
&= 0,
\end{aligned}$$

and

$$\begin{aligned}
\mathrm{var}_{\eta,\vartheta}\left[A_{j,m}\mid Z_{m+1},\ldots,Z_n\right] &= \mathrm{var}_{\eta,\vartheta}\left[\psi_j(Z;\widetilde{\eta}_2)-\psi_j(Z;\eta^*)\mid Z_{m+1},\ldots,Z_n\right] \\
&\le \mathbb{E}_{\eta,\vartheta}\left[\left\{\psi_j(Z;\tilde{\eta}_2)-\psi_j(Z;\eta^*)\right\}^2\mid Z_{m+1},\ldots,Z_n\right] \\
&= \int\left\{\psi_j(z;\widetilde{\eta}_2)-\psi_j(z;\eta^*)\right\}^2 f(z;\eta,\vartheta)\mathrm{d}z \\
&\le \int\left\|\psi(z;\widetilde{\eta}_2)-\psi(z;\eta^*)\right\|^2 f(z;\eta,\vartheta)\mathrm{d}z \xrightarrow[m\to\infty]{P(F_{\eta,\vartheta})} 0.
\end{aligned}$$

Consequently,

$$\mathbb{E}_{\eta,\vartheta}\left[A_{j,m}^2\mid Z_{m+1},\ldots,Z_n\right] = \mathrm{var}_{\eta,\vartheta}\left[A_{j,m}\mid Z_{m+1},\ldots,Z_n\right] \xrightarrow[m\to\infty]{P(F_{\eta,\vartheta})} 0.$$

By Chebyshev's inequality, for any $\delta>0$,

$$Q_m = \mathbb{P}_{\eta,\vartheta}\left[|A_{j,m}|>\delta\mid Z_{m+1},\ldots,Z_n\right] \le \frac{\mathbb{E}_{\eta,\vartheta}\left[A_{j,m}^2\mid Z_{m+1},\ldots,Z_n\right]}{\delta^2} \xrightarrow[m\to\infty]{P(F_{\eta,\vartheta})} 0.$$

Since $|Q_m|\le 1$, Q_m is a bounded sequence that converges to 0 in probability. By the bounded convergence theorem we have $\mathbb{E}_{\eta,\vartheta}[Q_m] \xrightarrow[m\to\infty]{} 0$, which implies that $\mathbb{P}_{\eta,\vartheta}[|A_{j,m}|>\delta] \xrightarrow[m\to\infty]{} 0$. This holds true for every component $j=1,\ldots,k$, so we conclude $A_m \xrightarrow[m\to\infty]{P(F_{\eta,\vartheta})} 0$.

The term B_m. This term is well-behaved since

$$B_m = \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varphi(Z_i;\eta),$$

where

$$\varphi(Z;\eta) = \psi(Z;\eta^*) - \int\psi(z;\eta^*)f(z;\eta,\vartheta)\mathrm{d}z.$$

The term C_m. $C_m = \sqrt{m}\{\beta(\widetilde{\eta}_2)-\beta(\eta)\}+\sqrt{m}e(\widetilde{\eta}_2)$. If we define for any η',

$$\chi(\eta') = \beta(\eta')-\beta(\eta)+\mathbb{E}_{\eta,\vartheta}[\psi(Z;\eta')], \tag{7.4}$$

then $\chi(\eta) = 0$ and C_m can be written as

$$C_m = \sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta)\}.$$

The term C_m must be analyzed individually in each estimation problem and its asymptotic behavior will depend on the nature of the estimator $\widetilde{\eta}_2$. There are some implications about different asymptotic behaviors of C_m:

1. if $\sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta)\}$ diverges as $m \to \infty$, then so is $\sqrt{m}\{\widehat{\beta}_1 - \beta(\eta)\}$;

2. if instead, $\chi(\widetilde{\eta}_2)$ is an asymptotically linear estimator of $\chi(\eta)$ with influence function $\phi(Z;\eta)$, i.e.,

$$\sqrt{n-m}\{\chi(\widetilde{\eta}_2) - \chi(\eta)\} = \frac{1}{n-m}\sum_{i=m+1}^{n}\phi(Z_i;\eta) + o_p(1),$$

then

$$\begin{aligned}
\sqrt{m}\{\widehat{\beta}_1 - \beta(\eta)\} &= A_m + B_m + C_m + o_p(1) \\
&= o_p(1) + \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varphi(Z_i;\eta) + \sqrt{m}\,\{\chi(\widetilde{\eta}_2) - \chi(\eta)\} \\
&= o_p(1) + \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varphi(Z_i;\eta) + \underbrace{\left(\frac{\sqrt{m}}{\sqrt{n-m}}\right)}_{1+o(1)}\underbrace{\sqrt{n-m}\,\{\chi(\widetilde{\eta}_2) - \chi(\eta)\}}_{O(1)} \\
&= o_p(1) + \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varphi(Z_i;\eta) + \sqrt{n-m}\,\{\chi(\widetilde{\eta}_2) - \chi(\eta)\} \\
&= o_p(1) + \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varphi(Z_i;\eta) + \frac{1}{\sqrt{n-m}}\sum_{i=m+1}^{n}\phi(Z_i;\eta),
\end{aligned}$$

which in turn implies

$$\sqrt{n}\{\widehat{\beta} - \beta(\eta)\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}[\varphi(Z_i;\eta) + \phi(Z_i;\eta)] + o_p(1),$$

i.e., $\widehat{\beta}$ is an asymptotically linear estimator of $\beta(\eta)$ at $F_{\eta,\vartheta}$ with influence function

$$\varphi(Z;\eta) + \phi(Z;\eta) = \psi(Z;\eta^*) - \mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\eta^*)\right] + \phi(Z;\eta);$$

3. if $\sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta)\} = o_p(1)$ and $\eta^* = \eta$ (which is typically true), then $\widehat{\beta}$ is an asymptotically linear estimator of $\beta(\eta)$ at $F_{\eta,\vartheta}$ with influence function $\psi(Z;\eta)$.

7.3 Convergence rate of $\sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta)\}$

We first show how $\chi(\eta') - \chi(\eta)$ scales with $\|\eta' - \eta\|$, where $\chi(\cdot)$ is defined as (7.4). For any parametric submodel $\mathcal{F}_{\text{sub}} = \{F_{\eta_\theta, \vartheta_\theta} : \theta \in \Theta\}$ such that $\eta_{\theta^*} = \eta$, we have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\theta}\chi(\eta_\theta)\bigg|_{\theta=\theta^*} &= \frac{\mathrm{d}}{\mathrm{d}\theta}\beta(\eta_\theta)\bigg|_{\theta=\theta^*} + \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\eta\left[\psi(Z;\eta_\theta)\right]\bigg|_{\theta=\theta^*} \\
&= \mathbb{E}_\eta\left[\psi(Z;\eta)S_\theta(\theta^*)\right] + \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\eta\left[\psi(Z;\eta_\theta)\right]\bigg|_{\theta=\theta^*} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_{\eta_\theta}\left[\psi(Z;\eta)\right]\bigg|_{\theta=\theta^*} + \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_\eta\left[\psi(Z;\eta_\theta)\right]\bigg|_{\theta=\theta^*} \\
&= \frac{\mathrm{d}}{\mathrm{d}\theta}\mathbb{E}_{\eta_\theta}\left[\psi(Z;\eta_\theta)\right]\bigg|_{\theta=\theta^*} = 0.
\end{aligned}$$

This implies $\chi(\eta') - \chi(\eta)$ is roughly $O(\|\eta' - \eta\|^2)$.

Remark 7.2. More formally, if Ξ is a normed space and if the map $\eta' \to \chi(\eta')$ admits the expansion

$$\chi(\eta') = \chi(\eta) + \chi_\eta(\eta' - \eta) + O(\|\eta' - \eta\|^2),$$

where $\chi_\eta(\cdot)$ is the Frechet derivative of $\chi(\cdot)$ and the class $\mathcal{A}$ has the maximal tangent space of $\mathcal{F}$ at $F_{\eta,\vartheta}$, then $\chi_\eta(\eta' - \eta) = 0$ for all η'. This accords with the Neyman orthogonality property [Chernozhukov et al., 2018], except that the latter is in the Gateaux derivative sense.

Remark 7.3. In many cases when η involves two functions, i.e., $\eta = (\nu, \kappa)$, it often holds that

$$\chi(\widetilde{\eta}_2) - \chi(\eta) = \beta(\widetilde{\eta}_2) - \beta(\eta) + \mathbb{E}_\eta\left[\psi(Z;\widetilde{\eta}_2)\right] = O(\|(\widetilde{\nu}_2 - \nu)\,(\widetilde{\kappa}_2 - \kappa)\|).$$

Then we can discuss the asymptotic properties depending on the consistency and different convergence rates of $\widetilde{\nu}_2$ and $\widetilde{\kappa}_2$, just as discussed in Section 7.2, the paragraph regarding C_m.

Remark 7.4. If $\widetilde{\eta}_2$ involves an ordinary smoothing estimator of a regression function and X is d-dimensional, then typically

$$\|\widetilde{\eta}_2 - \eta\| = O_p\left(m^{-\frac{\delta/d}{1+2\delta/d}}\right),$$

where δ is the number of derivatives (i.e., smoothness) of the true regression function η. So if $\delta/d > 1/2$, then $\frac{\delta/d}{1+2\delta/d} > 1/4$ and $\sqrt{m}\|\widetilde{\eta}_2 - \eta\|^2 = o_p(1)$. In this way, $C_m = o_p(1)$ and $\widehat{\beta}$ is an asymptotically linear estimator with influence function $\psi(Z;\eta)$. However, if $\delta/d < 1/2$, then $\sqrt{m}\|\widetilde{\eta}_2 - \eta\|^2$ diverges and thus $\widehat{\beta}$ is not $\sqrt{n}$-consistent.

7.4 The missing data example; double-robustness

We go back to Example 7.1. Recall that $\widetilde{\eta}_2 = (\widetilde{b}_2, \widetilde{\pi}_2, \widetilde{F}_{2,X})$, and $\psi(Z;\eta)$ is the unique mean zero gradient for $\beta(F)$ in $\mathcal{F}_{\text{np}}$. We have obtained

$$\widehat{\beta}_1 = \frac{1}{m}\sum_{i=1}^{m}\left[\widetilde{b}_2(X_i) + \frac{R_i\{Y_i - \widetilde{b}_2(X_i)\}}{\widetilde{\pi}_2(X_i)}\right].$$

In this example, for $\eta' = (b', \pi', F'_X)$, the form of $\chi(\eta')$ is

$$\begin{aligned}
\chi(\eta') &= \beta(\eta') - \beta(\eta) + \mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\eta')\right] \\
&= \beta(\eta') - \mathbb{E}_{\eta,\vartheta}\left[b_{F_{\eta,\vartheta}}(X)\right] + \mathbb{E}_{\eta,\vartheta}\left[b'(X) + \frac{R\left\{Y - b'(X)\right\}}{\pi'(X)} - \beta(\eta')\right] \\
&= \mathbb{E}_{\eta,\vartheta}\left[\left\{b'(X) - b_{F_{\eta,\vartheta}}(X)\right\} + \frac{R\left\{Y - b'(X)\right\}]}{\pi'(X)}\right] \\
&= \mathbb{E}_{\eta,\vartheta}\left[\left\{b'(X) - b_{F_{\eta,\vartheta}}(X)\right\} + \frac{R\left\{b_{F_{\eta,\vartheta}}(X) - b'(X)\right\}}{\pi'(X)}\right] \\
&= \mathbb{E}_{\eta,\vartheta}\left[\left\{b'(X) - b_{F_{\eta,\vartheta}}(X)\right\}\left\{1 - \frac{\pi_{F_{\eta,\vartheta}}(X)}{\pi'(X)}\right\}\right].
\end{aligned}$$

Now, suppose we use the parametric working models $\widetilde{b}_2(X) = \widetilde{\gamma}^\top \widetilde{X}$ and $\widetilde{\pi}_2 = \pi(X;\widetilde{\alpha}_2)$ as described before. Regardless of whether or not the models are correct, suppose

$$\mathbb{E}_{\eta,\vartheta}\left[u(Z;\gamma)\right] = \mathbb{E}_{\eta,\vartheta}\left[R\widetilde{X}\left(Y - \widetilde{X}^\top\gamma\right)\right] = 0$$

has a unique solution $\gamma(\eta)$, and

$$\mathbb{E}_{\eta,\vartheta}\left[S_\alpha(Z;\alpha)\right] = \mathbb{E}_{\eta,\vartheta}\left[\widetilde{X}\left(R - \frac{e^{\alpha^\top\widetilde{X}}}{1 + e^{\alpha^\top\widetilde{X}}}\right)\right] = 0$$

has a unique solution $\alpha(\eta)$. Then, under regularity conditions, it follows from the asymptotic theory for Z-estimators in Section 3.9 that

$$\sqrt{n-m}\left\{\begin{bmatrix}\widetilde{\gamma}_2\\ \widetilde{\alpha}_2\end{bmatrix} - \begin{bmatrix}\gamma(\eta)\\ \alpha(\eta)\end{bmatrix}\right\} = -\frac{1}{\sqrt{n-m}}\sum_{i=m+1}^{n}\begin{bmatrix}\phi^\gamma(Z_i;\eta)\\ \phi^\alpha(Z_i;\eta)\end{bmatrix} + o_p(1),$$

where

$$\begin{aligned}
\phi^\gamma(Z;\eta) &= -\left\{\partial\mathbb{E}_{\eta,\vartheta}[u(Z;\gamma)]/\,\partial\gamma^\top\big|_{\gamma=\gamma(\eta)}\right\}^{-1} u(Z;\gamma(\eta)), \\
\phi^\alpha(Z;\eta) &= -\left\{\partial\mathbb{E}_{\eta,\vartheta}\left[S_\alpha(Z;\alpha)\right]/\,\partial\alpha^\top\big|_{\alpha=\alpha(\eta)}\right\}^{-1} S_\alpha(Z;\alpha(\eta)).
\end{aligned}$$

In particular, $\widetilde{\gamma}_2 \xrightarrow{P(F_{\eta,\vartheta})} \gamma(\eta)$ and $\widetilde{\alpha}_2 \xrightarrow{P(F_{\eta,\vartheta})} \alpha(\eta)$. Without much effort we can verify the convergence assumption (7.3) with $\eta^* = (b(\cdot;\gamma(\eta)), \pi(\cdot;\alpha(\eta)), F_{X,\eta,\vartheta})$, and by the delta method, we have

$$\begin{aligned}
\sqrt{m}\left\{\chi(\widetilde{\eta}_2) - \chi(\eta^*)\right\} &= \sqrt{n-m}\left\{\tau(\widetilde{\gamma}_2,\widetilde{\alpha}_2) - \tau(\gamma(\eta),\alpha(\eta))\right\} + o_p(1) \\
&= \frac{1}{\sqrt{n-m}}\sum_{i=m+1}^{n}\phi(Z_i;\eta) + o_p(1), \qquad (7.5)
\end{aligned}$$

where $\tau(\gamma,\alpha) = \chi(b(\cdot;\gamma), \pi(\cdot;\alpha))$ and

$$\phi(Z;\eta) = \frac{\partial\tau(\gamma,\alpha)}{\partial(\gamma^\top,\alpha^\top)}\bigg|_{(\gamma,\alpha)=(\gamma(\eta),\alpha(\eta))}\begin{bmatrix}\phi^\gamma(Z;\eta)\\ \phi^\alpha(Z;\eta)\end{bmatrix}. \qquad (7.6)$$

Now we are equipped to obtain the asymptotic property of C_m and thus $\widehat{\beta}_1$ and $\widehat{\beta}$. We consider four possible scenarios:

1. both models for b and π are incorrect;
2. both models for b and π are correct;
3. the model for π is correct but the model for b is incorrect;
4. the model for b is correct but the model for π is incorrect.

Scenario 1. In this case, we have

$$b(\cdot, \gamma(\eta)) \neq b_{F_{\eta,\vartheta}}(\cdot) \quad \text{and} \quad \pi(\cdot, \alpha(\eta)) \neq \pi_{F_{\eta,\vartheta}}(\cdot),$$

so

$$\chi(\eta^*) = \mathbb{E}_{\eta,\vartheta}\left[\left\{b(X, \gamma(\eta)) - b_{F_{\eta,\vartheta}}(X)\right\}\left\{1 - \frac{\pi_{F_{\eta,\vartheta}}(X)}{\pi(X, \alpha(\eta))}\right\}\right] \neq 0.$$

Therefore,

$$\begin{aligned}\sqrt{m}\{\chi(\widetilde{\eta}_2) - \underbrace{\chi(\eta)}_{=0}\} &= \sqrt{m}\left\{\chi(\widetilde{\eta}_2) - \chi(\eta^*)\right\} + \sqrt{m}\chi(\eta^*) \\ &= \frac{1}{\sqrt{n-m}}\sum_{i=m+1}^{n}\phi(Z_i;\eta) + o_p(1) + \sqrt{m}\chi(\eta^*) \to \infty.\end{aligned}$$

The estimator is not even consistent, since

$$\widehat{\beta}_1 - \beta(\eta) = o_p(1) + \chi(\widetilde{\eta}_2) \xrightarrow{P(F_{\eta,\vartheta})} \chi(\eta^*) \neq 0.$$

Preparations for Scenarios 2-4. In these cases, we have either

$$b(\cdot, \gamma(\eta)) = b_{F_{\eta,\vartheta}}(\cdot) \quad \text{or} \quad \pi(\cdot, \alpha(\eta)) = \pi_{F_{\eta,\vartheta}}(\cdot),$$

so $\chi(\eta^*) = 0$. From results in Section 7.2, we conclude that $\widehat{\beta}_1$ (and consequently $\widehat{\beta}$) is consistent and asymptotically normal for $\beta(\eta)$.

Definition 7.1 (**Double-robustness**)**.** Given a semiparametric model $\mathcal{F}$ and two of its submodels $\mathcal{F}_{\text{sub},1}$ and $\mathcal{F}_{\text{sub},2}$, an estimator $\widehat{\beta}$ is said to be

1. **double-robust consistent** for $\beta(F)$ in the union model $\mathcal{F}_{\text{sub},1} \cup \mathcal{F}_{\text{sub},2}$ if $\widehat{\beta}$ converges in probability to $\beta(F)$ under any $F \in \mathcal{F}_{\text{sub},1} \cup \mathcal{F}_{\text{sub},2}$;
2. **double-robust asymptotically normal and unbiased** for $\beta(F)$ in the union model $\mathcal{F}_{\text{sub},1} \cup \mathcal{F}_{\text{sub},2}$ if $\sqrt{n}\{\widehat{\beta} - \beta(F)\}$ converges to a mean zero normal distribution under any $F \in \mathcal{F}_{\text{sub},1} \cup \mathcal{F}_{\text{sub},2}$.

In the missing data example, let

$$\mathcal{F}_{\text{sub},1} = \{F \in \mathcal{F}_{\text{np}} \colon \text{the working model for } b \text{ holds}\},$$
$$\mathcal{F}_{\text{sub},2} = \{F \in \mathcal{F}_{\text{np}} \colon \text{the working model for } \pi \text{ holds}\}.$$

Then $\widehat{\beta}$ is a double-robust estimator.

Return to the asymptotically linear form of $\sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta^*)\}$, i.e., (7.5)-(7.6). To calculate $\frac{\partial \tau(\gamma,\alpha)}{\partial(\gamma^\top,\alpha^\top)}\Big|_{(\gamma,\alpha)=(\gamma(\eta),\alpha(\eta))}$ in (7.6), we decompose η as $\eta = (\lambda, \kappa)$, where

$$\lambda = \left(b_{F_{\eta,\vartheta}}, F_{X,\eta,\vartheta}\right) \quad \text{and} \quad \kappa = \pi_{F_{\eta,\vartheta}}.$$

In addition, we let

$$\lambda^* = (b(\cdot;\gamma(\eta)), F_{X,\eta,\vartheta}), \ \kappa^* = \pi(\cdot;\alpha(\eta)), \ \lambda_\gamma = (b(\cdot;\gamma), \ F_{X,\eta,\vartheta}), \ \kappa_\alpha = \pi(\cdot;\alpha).$$

Then

$$\frac{\partial \tau(\gamma,\alpha)}{\partial(\gamma,\alpha)}\Big|_{(\gamma,\alpha)=(\gamma(\eta),\alpha(\eta))} = \begin{bmatrix} \frac{\partial}{\partial\gamma}\chi(\lambda_\gamma,\kappa^*)\Big|_{\gamma=\gamma(\eta)} \\ \frac{\partial}{\partial\alpha}\chi(\lambda^*,\kappa_\alpha)\Big|_{\alpha=\alpha(\eta)} \end{bmatrix}.$$

Fact 7.1. Note that three facts hold in the missing data example:

1. the assumed model $\mathcal{F}$, namely $\mathcal{F}_{\text{np}}$, $\mathcal{F}_{\text{sem,fixed}}$ or $\mathcal{F}_{\text{sem,par}}$, is a factorized model where the first factor of the likelihood depends on λ and the second factor of the likelihood depends on κ;

2. $\beta(\eta)$ depends on η only through λ;

3. we have

$$\chi(\lambda,\kappa') = \beta(\lambda) - \beta(\lambda) + \mathbb{E}_{\lambda,\kappa,\vartheta}[\psi(Z;\lambda,\kappa')] = 0 \text{ for all } \kappa',$$
$$\chi(\lambda',\kappa) = \beta(\lambda') - \beta(\lambda) + \mathbb{E}_{\lambda,\kappa,\vartheta}[\psi(Z;\lambda',\kappa)] = 0 \text{ for all } \lambda'.$$

The last fact holds because $\chi(\eta') = \mathbb{E}_{\eta,\vartheta}\left[\left\{b'(X) - b_{F_{\eta,\vartheta}}(X)\right\}\left\{1 - \frac{\pi_{F_{\eta,\vartheta}}(X)}{\pi'(X)}\right\}\right]$.

Fact 7.1 is crucial for the double-robustness property of the one-step estimator. In general cases, if the model is factorizable and the parameter $\beta(\eta)$ only depends on one of the components in $\eta = (\lambda, \kappa)$, then the one-step estimator often has the double-robustness property.

Scenario 2. In this case, we have $\lambda^* = \lambda$ and $\kappa^* = \kappa$. Since by Fact 7.1, $\chi(\lambda',\kappa) = \chi(\lambda,\kappa') = 0$ for all λ' and κ', their derivatives are also zero, i.e.,

$$\frac{\partial \tau(\gamma,\alpha)}{\partial(\gamma,\alpha)}\Big|_{(\gamma,\alpha)=(\gamma(\eta),\alpha(\eta))} = \begin{bmatrix} \frac{\partial}{\partial\gamma}\chi(\lambda_\gamma,\kappa^*)\Big|_{\gamma=\gamma(\eta)} \\ \frac{\partial}{\partial\alpha}\chi(\lambda^*,\kappa_\alpha)\Big|_{\alpha=\alpha(\eta)} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Therefore,

$$C_m = \sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta)\} = \sqrt{m}\{\chi(\widetilde{\eta}_2) - \chi(\eta^*)\} = o_p(1),$$

and consequently,

$$\sqrt{m}\{\widehat{\beta}_1 - \beta(\eta)\} = \frac{1}{m}\sum_{i=1}^{m} \psi(Z_i;\eta) + o_p(1),$$
$$\sqrt{n}\{\widehat{\beta} - \beta(\eta)\} = \frac{1}{n}\sum_{i=1}^{n} \psi(Z_i;\eta) + o_p(1).$$

We thus conclude that if the working models for b and π are both correct, then the one-step estimator $\widehat{\beta}$ is asymptotically linear with influence function $\psi(Z;\eta)$ defined in (7.1) (which is the efficient one for $\beta(\eta)$ in $\mathcal{F}_{\text{np}}$, $\mathcal{F}_{\text{sem,fixed}}$ and $\mathcal{F}_{\text{sem,par}}$).

Scenario 3. The model for π is correct but the model for b is incorrect. In this case, we have $\kappa^* = \kappa$, so $\chi(\lambda_\gamma, \kappa^*) = 0$ for all γ. Therefore,

$$\left.\frac{\partial\tau(\gamma,\alpha)}{\partial(\gamma,\alpha)}\right|_{(\gamma,\alpha)=(\gamma(\eta),\alpha(\eta))} = \begin{bmatrix} \left.\frac{\partial}{\partial\gamma}\chi(\lambda_\gamma,\kappa^*)\right|_{\gamma=\gamma(\eta)} \\ \left.\frac{\partial}{\partial\alpha}\chi(\lambda^*,\kappa_\alpha)\right|_{\alpha=\alpha(\eta)} \end{bmatrix} = \begin{bmatrix} 0 \\ \left.\frac{\partial}{\partial\alpha}\chi(\lambda^*,\kappa_\alpha)\right|_{\alpha=\alpha(\eta)} \end{bmatrix},$$

and

$$\phi(Z;\eta) = \left.\frac{\partial}{\partial\alpha^\top}\chi(\lambda^*,\kappa_\alpha)\right|_{\alpha=\alpha(\eta)} \phi^\alpha(Z;\eta).$$

To calculate $\phi(Z;\eta)$, we note that by Fact 7.1, $\beta(\lambda^*) - \beta(\lambda) + \mathbb{E}_{\lambda,\kappa_\alpha,\vartheta}[\psi(Z;\lambda^*,\kappa_\alpha)] = 0$ for all α. Differentiating both sides yields

$$\begin{aligned} 0 &= \left.\frac{\partial}{\partial\alpha}\mathbb{E}_{\lambda,\kappa_\alpha,\vartheta}\left[\psi(Z;\lambda^*,\kappa_\alpha)\right]\right|_{\alpha=\alpha(\eta)} \\ &= \left.\frac{\partial}{\partial\alpha}\mathbb{E}_{\lambda,\kappa_\alpha,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right]\right|_{\alpha=\alpha(\eta)} + \left.\frac{\partial}{\partial\alpha}\mathbb{E}_{\lambda,\kappa,,\vartheta}\left[\psi(Z;\lambda^*,\kappa_\alpha)\right]\right|_{\alpha=\alpha(\eta)} \\ &= \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda^*,\kappa)S_\alpha(Z;\alpha(\eta))\right] + \left.\frac{\partial\tau(\gamma(\eta),\alpha)}{\partial\alpha}\right|_{\alpha=\alpha(\eta)}, \end{aligned}$$

so

$$\left.\frac{\partial}{\partial\alpha^\top}\chi(\lambda^*,\kappa_\alpha)\right|_{\alpha=\alpha(\eta)} = \left.\frac{\partial\tau(\gamma(\eta),\alpha)}{\partial\alpha^\top}\right|_{\alpha=\alpha(\eta)} = -\mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda^*,\kappa)S_\alpha(Z;\alpha(\eta))^\top\right].$$

On the other hand, because $\widetilde{\alpha}_2$ is the MLE of α under the logistic regression model for $\mathbb{P}(R=1 \mid X)$, its influence function is

$$\begin{aligned} \phi^\alpha(Z;\eta) &= -\left\{\partial\mathbb{E}_{\eta,\vartheta}\left[S_\alpha(Z;\alpha)\right] / \partial\alpha^\top\big|_{\alpha=\alpha(\eta)}\right\}^{-1} S_\alpha(Z;\alpha(\eta)) \\ &= \left\{\mathbb{E}_{\eta,\vartheta}\left[S_\alpha(Z;\alpha(\eta))S_\alpha(Z;\alpha(\eta))^\top\right]\right\}^{-1} S_\alpha(Z;\alpha(\eta)). \end{aligned}$$

Finally, we obtain

$$\begin{aligned}\phi(Z;\eta) &= \left.\frac{\partial\tau(\gamma(\eta),\alpha)}{\partial\alpha^\top}\right|_{\alpha=\alpha(\eta)} \phi^\alpha(Z;\eta) \\ &= -\mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda^*,\kappa)S_\alpha(Z;\alpha(\eta))^\top\right]\left\{\mathbb{E}_{\lambda,\kappa,\vartheta}\left[S_\alpha(Z;\alpha(\eta))S_\alpha(Z;\alpha(\eta))^\top\right]\right\}^{-1} S_\alpha(Z;\alpha(\eta)) \\ &= -\Pi\left[\psi(Z;\lambda^*,\kappa) \mid S_\alpha(Z;\alpha(\eta))\right].\end{aligned}$$

We then arrive at

$$\begin{aligned}\sqrt{n}\{\widehat{\beta}-\beta(\eta)\} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\varphi(Z_i;\eta)+\phi(Z_i;\eta)\}+o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\{\psi(Z_i;\lambda^*,\kappa)-\mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right]+\phi(Z_i;\eta)\}+o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n\big\{\psi(Z_i;\lambda^*,\kappa)-\mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right] \\ &\quad -\Pi\left[\psi(Z_i;\lambda^*,\kappa)-\mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right] \mid S_\alpha(Z_i;\alpha(\eta))\right]\big\}+o_p(1) \\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \mathrm{RESID}\left\{\psi(Z_i;\lambda^*,\kappa)-\mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right]\right\}+o_p(1),\end{aligned}$$

where

$$\mathrm{RESID}(W) = W - \Pi\left[W \mid S_\alpha(Z;\alpha(\eta))\right].$$

Back to the missing data example, recall that

$$\psi(Z;\lambda^*,\kappa) = b(X;\gamma(\eta)) + \frac{R}{\pi_{F_{\eta,\vartheta}}(X)}(Y-b(X;\gamma(\eta))) - \mathbb{E}_{\eta,\vartheta}[b(X;\gamma(\eta))],$$

and thus

$$\mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right] = \beta(\eta) - \mathbb{E}_{\eta,\vartheta}[b(X;\gamma(\eta))].$$

To summarize, the influence function of $\widehat{\beta}$ is

$$\begin{aligned}&\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}\{Y-\beta(\eta)\} - \left\{\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}-1\right\}\{b(X;\gamma(\eta))-\beta(\eta)\} \\ &-\Pi\left[\left\{\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}\{Y-\beta(\eta)\} - \left\{\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}-1\right\}\{b(X;\gamma(\eta))-\beta(\eta)\}\right\} \mid S_\alpha(Z;\alpha(\eta))\right].\end{aligned}$$

Remark 7.5. The above result seems a bit counter-intuitive. Suppose two investigators will analyze the same data from the following two stage study design. At the first stage of the study X was measured on a random sample and at the second stage a random subsample was selected with $\pi^*(X)$ and Y was measured on this subsample.

1. The first investigator will compute the one step estimator of $\beta(\eta) = \mathbb{E}_\eta[\mathbb{E}_\eta(Y \mid R=1,X)]$ assuming an incorrect model for b and a model for π that uniquely contains the ground truth $\pi^*(X)$. Denote this investigator's estimator as $\widehat{\beta}_{\text{fixed}}$.

2. The second investigator will compute the one step estimator of $\beta(\eta) = \mathbb{E}_\eta[\mathbb{E}_\eta(Y \mid R = 1, X)]$ assuming the same incorrect model for b but will assume a correctly specified parametric model for π such that $\pi(X;\alpha) = \pi^*(X)$. Denote this investigator's estimator as $\widehat{\beta}_{\text{par}}$.

Comparing the asymptotic variance of two estimators, we have

$$\begin{aligned}
V_{\text{par}}(\eta) &= \text{var}_{F_{\eta,\vartheta}}\left[\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}\{Y-\beta(\eta)\} - \left\{\frac{R}{\pi_{F_{\eta,\vartheta}}(X)} - 1\right\}\{b(X;\gamma(\eta)) - \beta(\eta)\}\right. \\
&\quad \left. - \Pi\left[\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}\{Y-\beta(\eta)\} - \left\{\frac{R}{\pi_{F_{\eta,\vartheta}}(X)} - 1\right\}\{b(X;\gamma(\eta)) - \beta(\eta)\} \mid S_\alpha(Z;\alpha(\eta))\right]\right] \\
&\leq \text{var}_{F_{\eta,\vartheta}}\left[\frac{R}{\pi_{F_{\eta,\vartheta}}(X)}\{Y-\beta(\eta)\} - \left\{\frac{R}{\pi_{F_{\eta,\vartheta}}(X)} - 1\right\}\{b(X;\gamma(\eta)) - \beta(\eta)\}\right] \\
&= V_{\text{fixed}}(\eta).
\end{aligned}$$

So, estimating the missingness probability $\pi^*(X)$ even when it is known, can never decrease the asymptotic precision with which one estimates $\beta(\eta)$!

Actually, this counter-intuitive result is justified by the following remarks.

1. The general belief that estimation of nuisance parameters can not decrease the variance with which one estimates a parameter of interest, is correct ONLY when one compares the asymptotic variance of **efficient estimators** of the parameter of interest under models that assume the nuisance parameter is known or unknown.

2. Neither is $\widehat{\beta}_{\text{fixed}}$ efficient in $\mathcal{F}_{\text{fixed}}$, nor is $\widehat{\beta}_{\text{par}}$ efficient in model $\mathcal{F}_{\text{par}}$.

3. Note that if the model for b is correct, $\widehat{\beta}_{\text{par}}$ is not asymptotically more precise than $\widehat{\beta}_{\text{fixed}}$. In fact, both are RAL with efficient influence functions, and the efficient influence functions are the same in $\mathcal{F}_{\text{par}}$ and $\mathcal{F}_{\text{fixed}}$.

Scenario 4. The model for b is correct but the model for π is incorrect. In this case, we have $\lambda^* = \lambda$, so $\chi(\lambda^*, \kappa_\alpha) = 0$ for all α. Therefore,

$$\left.\frac{\partial\tau(\gamma,\alpha)}{\partial(\gamma,\alpha)}\right|_{(\gamma,\alpha)=(\gamma(\eta),\alpha(\eta))} = \begin{bmatrix} \left.\frac{\partial}{\partial\gamma}\chi(\lambda_\gamma,\kappa^*)\right|_{\gamma=\gamma(\eta)} \\ \left.\frac{\partial}{\partial\alpha}\chi(\lambda^*,\kappa_\alpha)\right|_{\alpha=\alpha(\eta)} \end{bmatrix} = \begin{bmatrix} \left.\frac{\partial}{\partial\gamma}\chi(\lambda_\gamma,\kappa^*)\right|_{\gamma=\gamma(\eta)} \\ 0 \end{bmatrix},$$

and

$$\phi(Z;\eta) = \left.\frac{\partial}{\partial\gamma^\top}\chi(\lambda_\gamma,\kappa^*)\right|_{\gamma=\gamma(\eta)} \phi^\gamma(Z;\eta).$$

We thus conclude that

$$\begin{aligned}
\sqrt{n}\{\widehat{\beta} - \beta(\eta)\} &= \frac{1}{\sqrt{n}}\sum_{i=1}^n \left\{\psi(Z_i;\lambda^*,\kappa) - \mathbb{E}_{\eta,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right] + \left.\frac{\partial}{\partial\gamma^\top}\chi(\lambda_\gamma,\kappa^*)\right|_{\gamma=\gamma(\eta)} \phi^\gamma(Z_i;\eta)\right\} \\
&\quad + o_p(1).
\end{aligned}$$

7.5 A general result for one-step estimators

Proposition 7.1. *Suppose the following assumptions hold:*

1. $\mathcal{F} = \{f(z;\eta,\vartheta) = g_1(z;\lambda,\vartheta)g_2(z;\kappa)\colon \eta = (\lambda,\kappa), \lambda \in \mathbb{L}, \kappa \in \mathbb{K}, \vartheta \in O\}$;
2. *the parameter of interest* $\beta(F_{\eta,\vartheta})$ *depends on* (η,ϑ) *only through* λ, *so we write it as* $\beta(\lambda)$;
3. *a gradient* $\psi_{F_{\eta,\vartheta}}(Z)$ *of* $\beta(F_{\eta,\vartheta})$ *depends on* (η,ϑ) *only through* η, *so we write it as* $\psi(Z;\lambda,\kappa)$;
4. $\chi(\lambda',\kappa') = \beta(\lambda') - \beta(\lambda) + \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda',\kappa')\right]$ *satisfies*
$$\chi(\lambda,\kappa') = \beta(\lambda) - \beta(\lambda) + \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda,\kappa')\right] = 0 \text{ for all } \kappa',$$
$$\chi(\lambda',\kappa) = \beta(\lambda') - \beta(\lambda) + \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda',\kappa)\right] = 0 \text{ for all } \lambda';$$
5. $\widetilde{\eta} = (\lambda_{\widetilde{\gamma}}, \kappa_{\widetilde{\alpha}})$, *where* $\widetilde{\gamma}$ *is an estimator of* γ *indexing a parametric model* λ_γ *for* λ, *and* $\widetilde{\alpha}$ *is the MLE of* α *indexing a parametric model* κ_α *for* κ;
6. $\widetilde{\gamma}$ *is asymptotically linear under* $F_{\eta,\vartheta}$.

Let $\widehat{\beta}$ *be the one-step sample splitting estimator, i.e., with* $m = \lfloor n/2 \rfloor$,

$$\widehat{\beta} = \frac{m}{n}\left[\beta(\lambda_{\widetilde{\gamma}_2}) + \frac{1}{m}\sum_{i=1}^{m}\psi(Z_i;\lambda_{\widetilde{\gamma}_2},\kappa_{\widetilde{\alpha}_2})\right] + \frac{n-m}{n}\left[\beta(\lambda_{\widetilde{\gamma}_1}) + \frac{1}{n-m}\sum_{i=m+1}^{n}\psi(Z_i;\lambda_{\widetilde{\gamma}_1},\kappa_{\widetilde{\alpha}_1})\right].$$

Then, it holds that

1. *if the model* λ_γ *and* κ_α *are both correct under* $F_{\eta,\vartheta}$, *then*
$$\sqrt{n}\left\{\widehat{\beta} - \beta(\lambda)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\psi(Z_i;\lambda,\kappa) + o_p(1);$$
2. *if the model* κ_α *is correct but the model* λ_γ *is incorrect, then with* $\alpha(\eta) = \operatorname{plim}\widetilde{\alpha}$, $\gamma(\eta) = \operatorname{plim}\widetilde{\gamma}$, $\lambda^* = \lambda_{\gamma(\eta)}$, *and* $S_\alpha(Z;\alpha(\eta))$ *denoting the score for* α *at* $\alpha(\eta)$,
$$\sqrt{n}\left\{\widehat{\beta} - \beta(\lambda)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\Big[\left\{\psi(Z_i;\lambda^*,\kappa) - \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right]\right\} - \Pi\left[\left\{\psi(Z_i;\lambda^*,\kappa) - \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda^*,\kappa)\right]\right\} \mid S_\alpha(Z;\alpha(\eta))\right]\Big] + o_p(1);$$
3. *if the model* λ_γ *is correct but the model* κ_α *is incorrect, then with* $\alpha(\eta) = \operatorname{plim}\widetilde{\alpha}$, $\gamma(\eta) = \operatorname{plim}\widetilde{\gamma}$, $\kappa^* = \kappa_{\alpha(\eta)}$, *and* $\phi^\gamma(Z;\eta)$ *denoting the influence function of* $\widetilde{\gamma}$,
$$\sqrt{n}\left\{\widehat{\beta} - \beta(\lambda)\right\} = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left[\left\{\psi(Z_i;\lambda,\kappa^*) - \mathbb{E}_{\lambda,\kappa,\vartheta}\left[\psi(Z;\lambda,\kappa^*)\right]\right\} + \frac{\partial}{\partial\gamma^\top}\chi(\lambda_\gamma,\kappa^*)\Big|_{\gamma=\gamma(\eta)}\right] + o_p(1).$$

8 Nonparametric Estimators

Please refer to Chapter 1 in [Tsybakov, 2009] for this part.

A Proofs in Section 5.6.3: An Efficiency Paradox

We have already shown that the nonparametric estimator $\widehat{g}(a) = \sum_l \widehat{b}(a,l)\widehat{f}_L(l)$ is RAL with efficient influence function under NUCA

$$\varphi_{\text{eff}}(Y, A, L; a) = \frac{\mathbb{I}(A = a)}{f_{A|L}(A \mid L)}(Y - b(A, L)) + b(a, L) - g(a).$$

Therefore, the efficient influence function for estimating $\psi = g(1) - g(0) = \mathbb{E}Y_1 - \mathbb{E}Y_0$ is

$$\varphi_{\text{eff}}(Y, A, L; \psi) = \frac{A(Y - b(1, L))}{f_{A|L}(1 \mid L)} - \frac{(1 - A)(Y - b(0, L))}{f_{A|L}(0 \mid L)} + b(1, L) - b(0, L) - \psi.$$

Assume that under the true distribution F^*, we additionally have $A \perp L$ and that $\mathbb{P}(A = 1) = \mathbb{P}(A = 0) = 1/2$, i.e., it is a randomized trial. Then the semiparametric C-R bound is

$$\begin{aligned}
\text{var}_{F^*}(\varphi_{\text{eff}}) &= \text{var}_{F^*}\{2A(Y - b(1, L))\} + \text{var}_{F^*}\{2(1 - A)(Y - b(0, L))\} \\
&\quad + \text{var}_{F^*}\{b(1, L) - b(0, L)\} \\
&= 4\mathbb{E}_{F^*}\{A(Y - b(1, L))^2\} + 4\mathbb{E}_{F^*}\{(1 - A)(Y - b(0, L))^2\} \\
&\quad + \text{var}_{F^*}\{b(1, L) - b(0, L)\} \\
&= 2\mathbb{E}_{F^*}\{(Y - b(1, L))^2 \mid A = 1\} + 2\mathbb{E}_{F^*}\{(Y - b(0, L))^2 \mid A = 0\} \\
&\quad + \text{var}_{F^*}\{b(1, L) - b(0, L)\}.
\end{aligned}$$

Note that if there is no covariate L, then the bound degenerate to

$$2\mathbb{E}_{F^*}\{(Y - g(1))^2 \mid A = 1\} + 2\mathbb{E}_{F^*}\{(Y - g(0))^2 \mid A = 0\}.$$

The nonparametric estimator $\widehat{\psi}$ achieves the previous semiparametric C-R bound $\text{var}_{F^*}(\varphi_{\text{eff}})$.

Now we focus on the crude estimator

$$\widetilde{\psi} = \frac{\sum_{s=1}^n \mathbb{I}(A_s = 1)Y_s}{\sum_{s=1}^n \mathbb{I}(A_s = 1)} - \frac{\sum_{s=1}^n \mathbb{I}(A_s = 0)Y_s}{\sum_{s=1}^n \mathbb{I}(A_s = 0)}.$$

We have

$$\sqrt{n}(\widetilde{\psi} - \psi) = \frac{\frac{1}{\sqrt{n}}\sum_{s=1}^n \mathbb{I}(A_s = 1)(Y_s - g(1))}{\frac{1}{n}\sum_{s=1}^n \mathbb{I}(A_s = 1)} - \frac{\frac{1}{\sqrt{n}}\sum_{s=1}^n \mathbb{I}(A_s = 0)(Y_s - g(0))}{\frac{1}{n}\sum_{s=1}^n \mathbb{I}(A_s = 0)}.$$

Since $\frac{1}{n}\sum_{s=1}^n \mathbb{I}(A_s = 1) = \frac{1}{2} + o_p(1)$ and $\frac{1}{n}\sum_{s=1}^n \mathbb{I}(A_s = 0) = \frac{1}{2} + o_p(1)$, we have

$$\begin{aligned}
\text{avar}_{F^*}\{\sqrt{n}(\widetilde{\psi} - \psi)\} &= \text{var}_{F^*}\{2A(Y - g(1)) - 2(1 - A)(Y - g(0))\} \\
&= 4\text{var}_{F^*}\{A(Y - g(1))\} + 4\text{var}_{F^*}\{(1 - A)(Y - g(0))\} \\
&= 2\mathbb{E}_{F^*}\{(Y - g(1))^2 \mid A = 1\} + 2\mathbb{E}_{F^*}\{(Y - g(0))^2 \mid A = 0\}.
\end{aligned}$$

Note that this asymptotic variance is larger than the semiparametric C-R bound, and achieves it iff $Y \perp L$, since

$$\begin{aligned}\mathrm{avar}_{F^*}\{\sqrt{n}(\widetilde{\psi}-\psi)\}-\mathrm{var}_{F^*}(\varphi_{\mathrm{eff}}) &= 2\mathrm{var}_{F^*}\{b(1,L)\}+2\mathrm{var}_{F^*}\{b(0,L)\}\\ &\quad -\mathrm{var}_{F^*}\{b(1,L)-b(0,L)\}\\ &= \mathrm{var}_{F^*}\{b(1,L)+b(0,L)\}\geq 0.\end{aligned}$$

One may wonder whether such a paradox contradicts with the results in [Chernozhukov et al., 2018]. The fact is that [Chernozhukov et al., 2018] requires two key ingredients, **Neyman Orthogonality** and **Sample Splitting**, while in our example these two are both violated. On the one hand, the nonparametric estimator $\widehat{\psi}$ is of the form

$$\widehat{\psi}=\frac{\sum_{s=1}^n \mathbb{I}(A_s=1)Y_s\widehat{f}_{A|L}^{-1}(A_s\mid L_s)}{\sum_{s=1}^n \mathbb{I}(A_s=1)\widehat{f}_{A|L}^{-1}(A_s\mid L_s)}-\frac{\sum_{s=1}^n \mathbb{I}(A_s=0)Y_s\widehat{f}_{A|L}^{-1}(A_s\mid L_s)}{\sum_{s=1}^n \mathbb{I}(A_s=0)\widehat{f}_{A|L}^{-1}(A_s\mid L_s)},$$

which is fairly different from the estimator solving the estimating equation induced by the efficient influence function, i.e.,

$$\check{\psi}=\frac{1}{n}\sum_{s=1}^n\left\{\frac{A_s(Y_s-\widehat{b}(1,L_s))}{\widehat{f}_{A|L}(1\mid L_s)}-\frac{(1-A_s)(Y_s-\widehat{b}(0,L_s))}{\widehat{f}_{A|L}(0\mid L_s)}+\widehat{b}(1,L_s)-\widehat{b}(0,L_s)\right\},$$

where the nuisance parameters are estimated using sample splitting. Nevertheless, they are asymptotically equivalent, i.e., sharing the same (efficient) influence function. The estimator $\widehat{\psi}$ depends only on one nuisance parameter $f_{A|L}(a\mid l)$ and does not satisfy Neyman orthogonality; while the estimator $\check{\psi}$ depends on two nuisance parameters $f_{A|L}(a\mid l)$ and $b(a,l)$, and both of them satisfy Neyman orthogonality (one can check this by calculating the Gateaux derivative at $f_{A|L}$ and b along any direction, which equals zero). On the other hand, the estimator $\widehat{\psi}$ is constructed without sample splitting, so the results in [Chernozhukov et al., 2018] do not apply directly. In conclusion, the efficiency of estimators in our example can be summarized in Table A.1.

Estimator type	Sample usage	Nuis. est.	Neyman orthogonal?	Efficient?
Estimating eq.	Split sample	$\widehat{f}_{A\|L}(a,l)$	Yes	Yes ($\check{\psi}$)
		$1/2$	Yes	Yes
	Full sample	$\widehat{f}_{A\|L}(a,l)$	Yes	?
		$1/2$	Yes	?
G-computation	Split sample	$\widehat{f}_{A\|L}(a,l)$	No	?
		$1/2$	No	?
	Full sample	$\widehat{f}_{A\|L}(a,l)$	No	Yes ($\widehat{\psi}$)
		$1/2$	No	No ($\widetilde{\psi}$)

Table A.1: Efficiency of estimators

B Extended Example in Section 6.3

B.1 MAR with different outcome models

We still consider the MAR case. Let $\mathcal{F}_{fixed}$ denote the semiparametric model where the outcome mean $\mathbb{E}(Y \mid X, R=1)$ is known to be the ground truth, and let $\mathcal{F}_{para}$ denote the semiparametric model where $\mathbb{E}(Y \mid X, R=1) = \mathbb{E}(Y \mid X, R=1; \gamma)$ follows a parametric model. We perform semiparametric inference on such two models for estimating $\mu = \mathbb{E}Y$.

Semiparametric inference in $\mathcal{F}_{fixed}$. The parameter of interest is $\beta(F) = \mu = \mathbb{E}_F(Y) = \mathbb{E}_F[\mathbb{E}_F(Y \mid X, R=1)]$ by the MAR assumption. Denote the true distribution as F^*, and let $\pi(X) = \mathbb{P}_F(R=1 \mid X)$ and $b(X) = \mathbb{E}_F(Y \mid X, R=1)$.

As derived in class,

$$\begin{aligned}\psi(Y,R,X) &= \mathbb{E}_{F^*}(Y \mid X, R=1) + \frac{R\{Y - \mathbb{E}_{F^*}(Y \mid X, R=1)\}}{\mathbb{P}_{F^*}(R=1 \mid X)} - \beta(F^*) \\ &= b^*(X) + \frac{R\{Y - b^*(X)\}}{\pi^*(X)} - \beta(F^*)\end{aligned}$$

is a mean zero gradient in $\mathcal{F}_{np}$. Therefore, it is also a mean zero gradient in $\mathcal{F}_{fixed}$. It suffices to calculate the tangent space of $\mathcal{A}_{fixed}$, containing all regular parametric submodels of $\mathcal{F}_{fixed}$. Let $\varepsilon = Y - b^*(X)$. We claim that

$$\begin{aligned}\Lambda_{\mathcal{F}_{fixed}} = \{&s_1(R,X) + Rs_2(\varepsilon,R,X) + (1-R)s_3(Y,R,X) \in \mathcal{L}_2^0(F^*): \\ &\mathbb{E}_{F^*}[s_1(R,X)] = 0, \mathbb{E}_{F^*}[s_2(\varepsilon,R,X) \mid R=1, X] = 0, \\ &\mathbb{E}_{F^*}[\varepsilon s_2(\varepsilon,R,X) \mid R=1, X] = 0, \mathbb{E}_{F^*}[s_3(Y,R,X) \mid R=0, X] = 0\},\end{aligned}$$

and thus (by similar arguments in the restricted mean model)

$$\Lambda_{\mathcal{F}_{fixed}}^{\perp} = \{R\varepsilon s(R,X) \in \mathcal{L}_2^0(F^*)\}.$$

On the one hand, for any regular parametric model

$$\mathcal{F}_{sub,fixed} = \{f_{Y,R,X}(y,r,x;\theta) = f_X(x;\theta) f_{R|X}(r \mid x;\theta) f_\varepsilon(\varepsilon \mid 1, x;\theta)^r f_{Y|R,X}(y \mid 0, x;\theta)^{1-r} : \theta \in \mathbb{R}\},$$

where $\varepsilon = Y - b^*(X)$, the score at F^* is of the form

$$s_{Y,R,X}(Y,R,X) = s_X(X) + s_{R|X}(R,X) + Rs_\varepsilon(\varepsilon,R,X) + (1-R)s_{Y|R,X}(Y,R,X),$$

satisfying $\mathbb{E}_{F^*}[s_X(X)] = 0$, $\mathbb{E}_{F^*}[s_{R|X}(R,X) \mid X] = 0$, $\mathbb{E}_{F^*}[s_\varepsilon(\varepsilon,R,X) \mid R=1, X] = 0$, $\mathbb{E}_{F^*}[\varepsilon s_\varepsilon(\varepsilon,R,X) \mid R=1, X] = 0$, and $\mathbb{E}_{F^*}[s_{Y|R,X}(Y,R,X) \mid R=0, X] = 0$, so all scores are in $\Lambda_{\mathcal{F}_{fixed}}$. On the other hand, for any $s(Y,R,X) \in \Lambda_{\mathcal{F}_{fixed}}$, we can construct a regular parametric model of the form $f(z;\theta) = f^*(z)\{1 + \theta s(z)\}c(\theta)$ where $c(\theta)$ is the normalizing constant depending on θ. Specifically, we first divide $s(Y,R,X)$ into four parts

$$s(Y,R,X) = s_1(X) + s_2(R,X) + Rs_3(\varepsilon,R,X) + (1-R)s_4(Y,R,X),$$

satisfying $\mathbb{E}_{F^*}[s_1(X)] = 0$, $\mathbb{E}_{F^*}[s_2(R,X) \mid X] = 0$, $\mathbb{E}_{F^*}[s_3(\varepsilon,R,X) \mid R=1,X] = 0$, $\mathbb{E}_{F^*}[\varepsilon s_3(\varepsilon,R,X) \mid R=1,X] = 0$, and $\mathbb{E}_{F^*}[s_4(Y,R,X) \mid R=0,X] = 0$. Then we replace z with x, $(r \mid x)$, $(\varepsilon \mid 1,x)$ and $(y \mid 0,x)$ respectively to obtain the desired result if the corresponding score is bounded. For unbounded scores, we construct a sequence of bounded scores approaching it, and obtain the same result.

Therefore, the efficient influence function in $\mathcal{F}_{fixed}$ is

$$\psi_{fixed}(Y,R,X) = \Pi[\psi(Y,R,X) \mid \Lambda_{\mathcal{F}_{fixed}}] = \mathbb{E}_{F^*}(Y \mid X, R=1) - \beta(F^*) = b^*(X) - \beta(F^*),$$

and the efficiency bound is

$$\mathcal{V}_{fixed} = \mathrm{var}_{F^*}\{\psi_{fixed}(Y,R,X)\} = \mathrm{var}_{F^*}\{b^*(X)\}.$$

Semiparametric inference in $\mathcal{F}_{para}$. For the semiparametric model $\mathcal{F}_{para}$, consider any regular parametric submodel

$$\mathcal{F}_{sub,para} = \{f_{Y,R,X}(y,r,x;\theta,\gamma) = f_X(x;\theta) f_{R|X}(r \mid x;\theta) f_\varepsilon(\varepsilon \mid 1,x;\theta)^r f_{Y|R,X}(y \mid 0,x;\theta)^{1-r} : \\ \varepsilon = Y - b(X;\gamma), \theta \in \mathbb{R}, \gamma \in \mathbb{R}\},$$

The score at F^* is of the form

$$s_{Y,R,X}(Y,R,X) = \begin{bmatrix} s_X(X) + s_{R|X}(R,X) + R s_\varepsilon(\varepsilon,R,X) + (1-R) s_{Y|R,X}(Y,R,X) \\ -R\dot{b}(X;\gamma^*) \frac{\partial \log f_\varepsilon(\varepsilon|1,X;\theta^*)}{\partial \varepsilon} \end{bmatrix},$$

where the first row is the score w.r.t. θ, which is the same as in the previous case $\mathcal{F}_{fixed}$; and the second row is the score w.r.t. γ. The tangent space $\Lambda_{\mathcal{F}_{para}}$ consists of all scores of the above form. Now, we argue that

$$\Lambda^\perp_{\mathcal{F}_{para}} = \{R\varepsilon s(X) \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}_{F^*}[s(X)\dot{b}(X;\gamma^*)\pi^*(X)] = 0\}.$$

Firstly, $\Lambda^\perp_{\mathcal{F}_{para}} \subseteq \Lambda^\perp_{\mathcal{F}_{fixed}}$ since $\Lambda_{\mathcal{F}_{fixed}} \subseteq \Lambda_{\mathcal{F}_{para}}$. For $R\varepsilon s(X) \in \Lambda^\perp_{\mathcal{F}_{para}}$, it must satisfy

$$\mathbb{E}_{F^*}\left[R\varepsilon s(X) \cdot R\dot{b}(X;\gamma^*) \frac{\partial \log f_\varepsilon(\varepsilon \mid 1,X;\theta^*)}{\partial \varepsilon}\right] = 0.$$

Since $\mathbb{E}_{F^*}\left[\varepsilon \frac{\partial \log f_\varepsilon(\varepsilon|1,X;\theta^*)}{\partial \varepsilon} \mid R=1,X\right] = -1$, it holds that

$$\mathbb{E}_{F^*}[s(X)\dot{b}(X;\gamma^*)\pi^*(X)] = 0.$$

As argued before, $\psi(Y,R,X)$ is still a mean zero gradient in $\mathcal{F}_{para}$. Suppose that $\Pi[\psi(Y,R,X) \mid \Lambda^\perp_{\mathcal{F}_{para}}] = R\varepsilon s_0(X)$ for some $s_0(X)$. Then it must satisfy

$$\mathbb{E}_{F^*}[\{\psi(Y,R,X) - R\varepsilon s_0(X)\} R\varepsilon s(X)] = 0 \Rightarrow \mathbb{E}_{F^*}\left[\sigma(X)^2 \pi^*(X) s(X) s_0(X)\right] = \mathbb{E}_{F^*}\left[\sigma(X)^2 s(X)\right],$$

for all $s(X)$ satisfying $\mathbb{E}_{F^*}[s(X)\dot{b}(X;\gamma^*)\pi^*(X)] = 0$, where $\sigma(X)^2 = \mathrm{var}_{F^*}(\varepsilon \mid R = 1, X)$. Therefore, $s_0(X)$ is of the form

$$s_0(X) = \frac{1}{\pi^*(X)} + v^\top \frac{\dot{b}(X;\gamma^*)}{\sigma(X)^2}$$

for some $v \in \mathbb{R}^{\dim(\gamma)}$. Together with the fact that $\mathbb{E}_{F^*}[s_0(X)\dot{b}(X;\gamma^*)\pi^*(X)] = 0$, we conclude that

$$s_0(X) = \frac{1}{\pi^*(X)} - \mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)^\top\right]\left\{\mathbb{E}_{F^*}\left[\frac{\dot{b}(X;\gamma^*)\dot{b}(X;\gamma^*)^\top\pi^*(X)}{\sigma(X)^2}\right]\right\}^{-1}\frac{\dot{b}(X;\gamma^*)}{\sigma(X)^2}.$$

In conclusion, the efficient influence function in $\mathcal{F}_{para}$ is

$$\begin{aligned}
\psi_{para}(Y,R,X) &= \psi(Y,R,X) - \Pi[\psi(Y,R,X) \mid \Lambda_{\mathcal{F}_{para}}^{\perp}] \\
&= b^*(X) + \frac{R\varepsilon}{\pi^*(X)} - \beta(F^*) - R\varepsilon s_0(X) \\
&= b^*(X) - \beta(F^*) \\
&\quad + \mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)^\top\right]\left\{\mathbb{E}_{F^*}\left[\frac{\dot{b}(X;\gamma^*)\dot{b}(X;\gamma^*)^\top\pi^*(X)}{\sigma(X)^2}\right]\right\}^{-1}\frac{\dot{b}(X;\gamma^*)}{\sigma(X)^2}R\varepsilon,
\end{aligned}$$

and the efficiency bound is

$$\begin{aligned}
\mathcal{V}_{para} &= \mathrm{var}_{F^*}\{\psi_{para}(Y,R,X)\} \\
&= \mathrm{var}_{F^*}\{b^*(X)\} + \mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)^\top\right]\left\{\mathbb{E}_{F^*}\left[\frac{\dot{b}(X;\gamma^*)\dot{b}(X;\gamma^*)^\top\pi^*(X)}{\sigma(X)^2}\right]\right\}^{-1}\mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)\right].
\end{aligned}$$

Comparison between C-R bounds. The difference between the above two bounds is

$$\mathcal{V}_{para} - \mathcal{V}_{fixed} = \mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)^\top\right]\left\{\mathbb{E}_{F^*}\left[\frac{\dot{b}(X;\gamma^*)\dot{b}(X;\gamma^*)^\top\pi^*(X)}{\sigma(X)^2}\right]\right\}^{-1}\mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)\right],$$

which is caused by knowing less information about $b^*(X)$. As is known, the efficiency bound for $\mathcal{F}_{np}$ is

$$\mathcal{V}_{np} = \mathrm{var}_{F^*}\{b^*(X)\} + \mathbb{E}_{F^*}[\mathrm{var}_{F^*}(\varepsilon \mid R = 1, X)] = \mathrm{var}_{F^*}\{b^*(X)\} + \mathbb{E}_{F^*}\left[\frac{\sigma(X)^2}{\pi^*(X)}\right].$$

The differences with this efficiency bound are

$$\mathcal{V}_{np} - \mathcal{V}_{fixed} = \mathbb{E}_{F^*}\left[\frac{\sigma(X)^2}{\pi^*(X)}\right],$$

$$\mathcal{V}_{np} - \mathcal{V}_{para} = \mathbb{E}_{F^*}\left[\frac{\sigma(X)^2}{\pi^*(X)}\right] - \mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)^\top\right]\left\{\mathbb{E}_{F^*}\left[\frac{\dot{b}(X;\gamma^*)\dot{b}(X;\gamma^*)^\top\pi^*(X)}{\sigma(X)^2}\right]\right\}^{-1}\mathbb{E}_{F^*}\left[\dot{b}(X;\gamma^*)\right].$$

B.2 MNAR with a logistic propensity score and a known interaction function

Suppose the missingness is not at random (MNAR), i.e., $R \not\perp Y \mid X$. Suppose the propensity score is

$$f(R=1 \mid X,Y) = \text{expit}\{\alpha(X) + \gamma(X,Y)\}, \quad \gamma(X, Y=0) = 0.$$

Consider the semiparametric model where $\gamma(X,Y)$ is known and the remaining part of the joint distribution is unrestricted. We aim at estimating $\mu = \mathbb{E}Y$.

Calculation of the tangent space. The data distribution is

$$f(y,r,x) = f(x)\{f(y \mid x)f(r=1 \mid x,y)\}^r \left\{\mathbb{I}(y=\text{NA}) \int f(y \mid x)f(r=0 \mid x,y)\mathrm{d}y\right\}^{1-r},$$

where $f(r=1 \mid x,y) = \text{expit}\{\alpha(x) + \gamma(x,y)\}$. Consider any regular parametric submodel

$$\mathcal{F}_{sub} = \{f(y,r,x;\theta) \text{ of the above form with } f(x;\theta), f(y \mid x;\theta), \alpha(x;\theta) \colon \theta \in \mathbb{R}\}.$$

The score at $f^*(y,r,x)$ is

$$\begin{aligned}
s(Y,R,X) = \frac{\partial}{\partial\theta}\Bigg\{ & \log f_\theta(X) + R\left[\log f_\theta(Y \mid X) + \log \frac{e^{\alpha_\theta(X)+\gamma(X,Y)}}{1+e^{\alpha_\theta(X)+\gamma(X,Y)}}\right] \\
& + (1-R)\left[\log \int f_\theta(y \mid X)\frac{1}{1+e^{\alpha_\theta(X)+\gamma(X,y)}}\mathrm{d}y\right]\Bigg\}\Bigg|_{\theta=\theta^*} \\
= {} & s_X(X) + R\left[s_{Y|X}(Y \mid X) + \frac{\dot\alpha(X;\theta^*)}{1+e^{\alpha^*(X)+\gamma(X,Y)}}\right] \\
& + (1-R)\frac{\int\left[s_{Y|X}(y \mid X) - \frac{e^{\alpha^*(X)+\gamma(X,y)}\dot\alpha(X;\theta^*)}{1+e^{\alpha^*(X)+\gamma(X,y)}}\right] f^*(y \mid X)\frac{1}{1+e^{\alpha^*(X)+\gamma(X,y)}}\mathrm{d}y}{\int f^*(y \mid X)\frac{1}{1+e^{\alpha^*(X)+\gamma(X,y)}}\mathrm{d}y}.
\end{aligned}$$

Now we want to characterize the tangent space consisting of all scores of the above form. Suppose a score is $s(Y,R,X) = g_1(X) + Rg_2(X,Y) + (1-R)g_3(X)$. Then it must hold that

$$\dot\alpha(X;\theta^*)\mathbb{P}(R=0 \mid X) = \mathbb{E}[g_2(X,Y) \mid X] \Rightarrow \dot\alpha(X;\theta^*) = \frac{\mathbb{E}[g_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)},$$

and

$$s_{Y|X}(Y \mid X) = g_2(X,Y) - \frac{\mathbb{E}[g_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)}\mathbb{P}(R=0 \mid X,Y).$$

Plugging into the form of $g_3(X)$, we get

$$\begin{aligned}
g_3(X) &= \mathbb{E}[s_{Y|X}(Y \mid X) - \dot\alpha(X;\theta^*)\mathbb{P}(R=1 \mid X,Y) \mid X, R=0] \\
&= \mathbb{E}[g_2(X,Y) \mid X, R=0] - \frac{\mathbb{E}[g_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)} \\
&= -\frac{\mathbb{E}[Rg_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)}.
\end{aligned}$$

On the other hand, for any $g_1(X), g_2(X,Y), g_3(X)$ satisfying the above restrictions, we can construct a regular parametric submodel such that

$$s_X(X) = g_1(X), \quad \dot{\alpha}(X;\theta^*) = \frac{\mathbb{E}[g_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)},$$
$$s_{Y|X}(Y \mid X) = g_2(X,Y) - \frac{\mathbb{E}[g_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)}\mathbb{P}(R=0 \mid X,Y).$$

To conclude, the tangent space in the nonparametric model $\mathcal{F}$ is

$$\Lambda_{\mathcal{F}} = \left\{ s_1(X) + Rs_2(X,Y) - (1-R)\frac{\mathbb{E}[Rs_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)} \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}[s_1(X)] = 0 \right\}.$$

Calculation of a gradient Next we calculate a gradient. Note that

$$\begin{aligned}
\frac{\mathbb{E}[Y\mathbb{I}(R=0) \mid X=x]}{\mathbb{E}[\mathbb{I}(R=0) \mid X=x]} &= \frac{\int y f(x) f(y \mid x) f(r=0 \mid x,y) \mathrm{d}y}{\int f(x) f(y \mid x) f(r=0 \mid x,y) \mathrm{d}y} \\
&= \frac{\int y f(x) f(y \mid x) \frac{1}{1+e^{\alpha(x)+\gamma(x,y)}} \mathrm{d}y}{\int f(x) f(y \mid x) \frac{1}{1+e^{\alpha(x)+\gamma(x,y)}} \mathrm{d}y} \\
&= \frac{\int \frac{y}{e^{\gamma(x,y)}} f(x) f(y \mid x) \frac{e^{\alpha(x)+\gamma(x,y)}}{1+e^{\alpha(x)+\gamma(x,y)}} \mathrm{d}y}{\int \frac{1}{e^{\gamma(x,y)}} f(x) f(y \mid x) \frac{e^{\alpha(x)+\gamma(x,y)}}{1+e^{\alpha(x)+\gamma(x,y)}} \mathrm{d}y} \\
&= \frac{\int \frac{y}{e^{\gamma(x,y)}} f(x) f(y \mid x) f(r=1 \mid x,y) \mathrm{d}y}{\int \frac{1}{e^{\gamma(x,y)}} f(x) f(y \mid x) f(r=1 \mid x,y) \mathrm{d}y} \\
&= \frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mathbb{I}(R=1) \mid X=x\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mathbb{I}(R=1) \mid X=x\right]},
\end{aligned}$$

so

$$\begin{aligned}
\beta(F) = \mu &= \mathbb{E}(Y) \\
&= \mathbb{E}\{\mathbb{E}[Y\mathbb{I}(R=1) \mid X] + \mathbb{E}[Y\mathbb{I}(R=0) \mid X]\} \\
&= \mathbb{E}\left\{ \mathbb{E}[Y\mathbb{I}(R=1) \mid X] + \frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mathbb{I}(R=1) \mid X\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mathbb{I}(R=1) \mid X\right]} \mathbb{E}[\mathbb{I}(R=0) \mid X] \right\} \\
&= \mathbb{E}\left\{ \mathbb{E}[Y \mid X, R=1]\mathbb{P}(R=1 \mid X) + \frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]} \mathbb{P}(R=0 \mid X) \right\}.
\end{aligned}$$

For any regular parametric submodel

$$\mathcal{F}_{sub} = \{f_\theta(x) f_\theta(r \mid x) f_\theta(y \mid r,x) \colon \theta \in \mathbb{R}\},$$

denote the corresponding scores as $s_X(X)$, $s_{R|X}(R,X)$ and $s_{Y|R,X}(Y,R,X)$. Then

$$\begin{aligned}
\left.\frac{\partial\beta(\theta)}{\partial\theta}\right|_{\theta=\theta^*} &= \frac{\partial}{\partial\theta}\mathbb{E}_\theta\left\{\mathbb{E}_\theta[Y \mid X, R=1]\mathbb{P}_\theta(R=1 \mid X) + \frac{\mathbb{E}_\theta\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\mathbb{E}_\theta\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]}\mathbb{P}_\theta(R=0 \mid X)\right\} \\
&= \mathbb{E}\left\{\left[\mathbb{E}[Y \mid X, R=1]\mathbb{P}(R=1 \mid X) + \frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]}\mathbb{P}(R=0 \mid X)\right] s_X(X)\right\} \\
&\quad + \mathbb{E}\{\mathbb{E}[Y s_{Y|R,X}(Y,R,X) \mid X, R=1]\mathbb{P}(R=1 \mid X)\} \\
&\quad + \mathbb{E}\{\mathbb{E}[Y \mid X, R=1]\mathbb{P}(R=1 \mid X) s_{R|X}(R=1,X)\} \\
&\quad + \mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} s_{Y|R,X}(Y,R,X) \mid X, R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]}\mathbb{P}(R=0 \mid X)\right\} \\
&\quad - \mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\left\{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]\right\}^2}\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} s_{Y|R,X}(Y,R,X) \mid X, R=1\right]\mathbb{P}(R=0 \mid X)\right\} \\
&\quad + \mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]}\mathbb{P}(R=0 \mid X) s_{R|X}(R=0,X)\right\}.
\end{aligned}$$

We analyze the six parts respectively.

1. The first part is

$$\begin{aligned}
&\mathbb{E}\left\{\left[\mathbb{E}[Y \mid X, R=1]\mathbb{P}(R=1 \mid X) + \frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]}\mathbb{P}(R=0 \mid X)\right] s_X(X)\right\} \\
&= \mathbb{E}\left\{\left[\mathbb{E}[Y \mid X, R=1]\mathbb{P}(R=1 \mid X) + \frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right]}\mathbb{P}(R=0 \mid X)\right] s(Y,R,X)\right\}.
\end{aligned}$$

2. The second part is

$$\begin{aligned}
&\mathbb{E}\{\mathbb{E}[Y s_{Y|R,X}(Y,R,X) \mid X, R=1]\mathbb{P}(R=1 \mid X)\} \\
&= \mathbb{E}\{\mathbb{E}\{RY s_{Y|R,X}(Y,R,X) \mid X\}\} \\
&= \mathbb{E}\{R\{Y - \mathbb{E}(Y \mid R, X)\} s_{Y|R,X}(Y,R,X)\} \\
&= \mathbb{E}\{R\{Y - \mathbb{E}(Y \mid R, X)\} s(Y,R,X)\} \\
&= \mathbb{E}\{R\{Y - \mathbb{E}(Y \mid R=1, X)\} s(Y,R,X)\}.
\end{aligned}$$

3. The third part is

$$\begin{aligned}
&\mathbb{E}\{\mathbb{E}[Y \mid X, R=1]\mathbb{P}(R=1 \mid X) s_{R|X}(R=1,X)\} \\
&= \mathbb{E}\{R\mathbb{E}\{Y \mid R, X\} s_{R|X}(R,X)\} \\
&= \mathbb{E}\{\{\mathbb{E}[RY \mid R, X] - \mathbb{E}[RY \mid X]\} s(Y,R,X)\} \\
&= \mathbb{E}\{\{R\mathbb{E}[Y \mid R=1, X] - \mathbb{E}[RY \mid X]\} s(Y,R,X)\}.
\end{aligned}$$

4. The fourth part is

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}s_{Y|R,X}(Y,R,X)\mid X,R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\mathbb{P}(R=0\mid X)\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{RY}{e^{\gamma(X,Y)}}s_{Y|R,X}(Y,R,X)\mid X\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}\right\}\\
&=\mathbb{E}\left\{\frac{\frac{RY}{e^{\gamma(X,Y)}}}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}s_{Y|R,X}(Y,R,X)\right\}\\
&=\mathbb{E}\left\{\frac{R\left\{\frac{Y}{e^{\gamma(X,Y)}}-\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}s_{Y|R,X}(Y,R,X)\right\}\\
&=\mathbb{E}\left\{\frac{R\left\{\frac{Y}{e^{\gamma(X,Y)}}-\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}s(Y,R,X)\right\}.
\end{aligned}
$$

5. The fifth part is

$$
\begin{aligned}
&-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\left\{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}^2}\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}s_{Y|R,X}(Y,R,X)\mid X,R=1\right]\mathbb{P}(R=0\mid X)\right\}\\
&=-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\left\{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}^2}\mathbb{E}\left[\frac{R}{e^{\gamma(X,Y)}}s_{Y|R,X}(Y,R,X)\mid X\right]\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}\right\}\\
&=-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\left\{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}^2}\frac{R}{e^{\gamma(X,Y)}}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}s_{Y|R,X}(Y,R,X)\right\}\\
&=-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\left\{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}^2}R\left\{\frac{1}{e^{\gamma(X,Y)}}-\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}s_{Y|R,X}(Y,R,X)\right\}\\
&=-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\left\{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}^2}R\left\{\frac{1}{e^{\gamma(X,Y)}}-\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]\right\}\frac{\mathbb{P}(R=0\mid X)}{\mathbb{P}(R=1\mid X)}s(Y,R,X)\right\}.
\end{aligned}
$$

6. The sixth part is

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\mathbb{P}(R=0\mid X)s_{R|X}(R=0,X)\right\}\\
&=\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}(1-R)s_{R|X}(R,X)\right\}\\
&=-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\{R-\mathbb{P}(R=1\mid X)\}s_{R|X}(R,X)\right\}\\
&=-\mathbb{E}\left\{\frac{\mathbb{E}\left[\frac{Y}{e^{\gamma(X,Y)}}\mid X,R=1\right]}{\mathbb{E}\left[\frac{1}{e^{\gamma(X,Y)}}\mid X,R=1\right]}\{R-\mathbb{P}(R=1\mid X)\}s(Y,R,X)\right\}.
\end{aligned}
$$

Putting these together, we can show that

$$\tilde{\psi}(Y,R,X) = RY + (1-R)\frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}} + \frac{R}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\left[\frac{Y}{e^{\gamma(X,Y)}} - \frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\frac{1}{e^{\gamma(X,Y)}}\right]\frac{\mathbb{P}(R=0 \mid X)}{\mathbb{P}(R=1 \mid X)}$$

is a gradient at for $\beta(F)$ at F^* in the nonparametric model $\mathcal{F}$. When $\gamma(X,Y) \equiv 0$, which is the MAR case, this gradient $\tilde{\psi}(Y,R,X)$ degenerates to

$$\begin{aligned} &RY + (1-R)\mathbb{E}(Y \mid X, R=1) + R\{Y - \mathbb{E}(Y \mid X, R=1)\}\frac{\mathbb{P}(R=0 \mid X)}{\mathbb{P}(R=1 \mid X)} \\ &= \mathbb{E}(Y \mid X, R=1) + \frac{R\{Y - \mathbb{E}(Y \mid X, R=1)\}}{\mathbb{P}(R=1 \mid X)}, \end{aligned}$$

verifying the results in the MAR case. Since $\mathbb{E}\tilde{\psi}(Y,R,X) = \beta(F^*)$, a mean zero gradient is given by

$$\psi(Y,R,X) = -\beta(F^*) + RY + (1-R)\frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}} + \frac{R}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\left[\frac{Y}{e^{\gamma(X,Y)}} - \frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\frac{1}{e^{\gamma(X,Y)}}\right]\frac{\mathbb{P}(R=0 \mid X)}{\mathbb{P}(R=1 \mid X)}.$$

The efficient influence function and the efficiency bound We now argue that $\psi(Y,R,X)$ is in the tangent space

$$\Lambda_{\mathcal{F}} = \left\{ s_1(X) + Rs_2(X,Y) - (1-R)\frac{\mathbb{E}[Rs_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)} \in \mathcal{L}_2^0(F^*) \colon \mathbb{E}[s_1(X)] = 0 \right\},$$

so that it is also the efficient influence function. Actually, letting

$$\begin{aligned} s_1(X) &= \mathbb{E}[RY \mid X] + \frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\mathbb{P}(R=0 \mid X) - \beta(F^*), \\ s_2(X,Y) &= Y + \frac{1}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\left[\frac{Y}{e^{\gamma(X,Y)}} - \frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\frac{1}{e^{\gamma(X,Y)}}\right]\frac{\mathbb{P}(R=0 \mid X)}{\mathbb{P}(R=1 \mid X)} \\ &\quad - \mathbb{E}[RY \mid X] - \frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\mathbb{P}(R=0 \mid X), \end{aligned}$$

yields

$$\psi(Y,R,X) = s_1(X) + Rs_2(X,Y) - (1-R)\frac{\mathbb{E}[Rs_2(X,Y) \mid X]}{\mathbb{P}(R=0 \mid X)} \in \Lambda_{\mathcal{F}}.$$

Finally, the efficiency bound is given by

$$\begin{aligned}\mathcal{V} &= \mathrm{var}_{F^*}\{\psi(Y,R,X)\} \\ &= \mathrm{var}_{F^*}(Y) + \mathrm{var}_{F^*}\left\{\left[Y - \frac{\mathbb{E}\left\{\frac{Y}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\right]\left[\frac{\frac{R}{e^{\gamma(X,Y)}}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\frac{\mathbb{P}(R=0 \mid X)}{\mathbb{P}(R=1 \mid X)} - (1-R)\right]\right\},\end{aligned}$$

where we have used the fact that

$$\mathbb{E}\left[\frac{\frac{R}{e^{\gamma(X,Y)}}}{\mathbb{E}\left\{\frac{1}{e^{\gamma(X,Y)}} \mid X, R=1\right\}}\frac{\mathbb{P}(R=0 \mid X)}{\mathbb{P}(R=1 \mid X)} - (1-R) \mid X, Y\right] = 0.$$

References

[Chamberlain, 1992] Chamberlain, G. (1992). Efficiency bounds for semiparametric regression. *Econometrica: Journal of the Econometric Society*, pages 567–596.

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

[Polyak and Juditsky, 1992] Polyak, B. T. and Juditsky, A. B. (1992). Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855.

[Tsiatis, 2006] Tsiatis, A. A. (2006). *Semiparametric theory and missing data*. Springer.

[Tsybakov, 2009] Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer series in statistics. Springer, Dordrecht.

[Van der Vaart, 2000] Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.