2
General Methodology

2.1 The General Model and Overview

Let X be the full data structure for one subject, and it is assumed that the full data distribution F_X is an element of a model $\mathcal{M}^F$. Let $Y = \Phi(X, C)$ be the observed data for one subject, where Φ is a known many to one mapping and C is a censoring variable. Typically, in most of our applications we parametrize the data structure such that C is always observed, but this is not required; obviously, this is always possible, since one can define $C = Y$, in particular. Let $G(\cdot \mid X)$ be the conditional distribution of C, given X, which is assumed to satisfy coarsening at random. The set of all conditional distributions satisfying coarsening at random will be denoted with $\mathcal{G}(CAR)$. Because of the curse of dimensionality, it typically will not suffice to assume only that $F_X \in \mathcal{M}^F$ and $G \in \mathcal{G}(CAR)$. In this chapter, we develop estimating functions and corresponding locally efficient estimators for two models. Firstly, given working models $\mathcal{M}^{F,w} \subset \mathcal{M}^F$ for F_X and $\mathcal{G} \subset \mathcal{G}(CAR)$ for G, we consider the following model for the distribution of Y:

$$\mathcal{M} = \{P_{F_X,G} : F_X \in \mathcal{M}^F, G \in \mathcal{G}\} \cup \{P_{F_X,G} : F_X \in \mathcal{M}^{F,w}, G \in \mathcal{G}(CAR)\}.$$

In other words, either F_X needs to be an element of $\mathcal{M}^{F,w}$ or G needs to be an element of $\mathcal{G}$. We will also consider the less nonparametric model

$$\mathcal{M}(\mathcal{G}) = \{P_{F_X,G} : F_X \in \mathcal{M}^F, G \in \mathcal{G}\},$$

which assumes a correctly specified model for the censoring mechanism. The data consist of n i.i.d. copies $Y_1, \ldots, Y_n$ of Y. Let $\mu = \Phi(F_X) \in \mathbb{R}^k$ be a k-dimensional Euclidean parameter of interest.

In this chapter, we propose general mappings from full data estimating functions into observed data estimating functions. In the next section, we study full data estimating functions for two general classes of full data models: multivariate generalized linear regression models and multiplicative intensity models. In Section 2.3 we propose methods for constructing mappings from full data estimating functions into observed data estimating functions for model $\mathcal{M}(\mathcal{G})$ and a doubly robust mapping for model $\mathcal{M}$. These doubly robust mappings are $\mathcal{G}$-orthogonalized initial mappings in the sense that they are defined as an initial mapping minus its projection onto a nuisance tangent space of G corresponding to a convex model $\mathcal{G}$. In Section 2.4 we define the optimal mapping (based on (1.52) in Theorem 1.3) from full data estimating functions into observed data estimating functions, which can be used for both models $\mathcal{M}$ and $\mathcal{M}(\mathcal{G})$. The optimal mapping is optimal in the sense that it is an $\mathcal{G}(CAR)$-orthogonalized initial mapping, and by Theorem 1.3 it covers all estimating functions, including the optimal one.

Since this optimal mapping does not always exist in closed form, the methods of Section 2.3 can be preferable and are therefore still very important as well. Section 2.5 defines the corresponding estimating equations and, in model $\mathcal{M}(\mathcal{G})$, we show how to adjust the estimating equation to obtain an estimator that is guaranteed more efficient than an initial estimator. Section 2.6 proposes confidence intervals, and Section 2.7 presents two asymptotic theorems for the one-step estimator based on the estimating equation of Section 2.5 in model $\mathcal{M}(\mathcal{G})$ and in model $\mathcal{M}$, respectively, which provide templates for proving local efficiency of the one-step estimator. Section 2.8 presents representations of the optimal index $h_{opt}(F_X, G)$ of the full data estimating functions. In particular, we prove a theorem for general censored data that provides a closed-form expression of the optimal index if the full data model is a multivariate generalized linear regression model with uncensored covariates. In Section 2.9 we derive a general reparametrization of h_{opt} and propose a corresponding substitution estimator h_n. Finally, in Section 2.10 we present a general locally efficient estimator based on the representation of the efficient influence curve in terms of score and information operators as presented in Bickel, Klaassen, Ritov and Wellner (1993).

2.2 Full Data Estimating Functions.

Given a full data model $\mathcal{M}^F$ and parameter $\mu = \Phi(F_X) \in \mathbb{R}^k$ of interest, finding the class of estimating functions requires finding the orthogonal

complement of the nuisance tangent space at F_X for each $F_X \in \mathcal{M}^F$. We refer the reader to Chapter 1 for an overview of the relevant efficiency and estimating functions theory. Here we will provide a short summary. Subsequently, we will derive the orthogonal complement of the nuisance tangent space in multivariate generalized linear regression models and multiplicative intensity models. These two general models form two of the most important full data models in the literature and will act as possible full data models in this book. Finally, we show how one links the orthogonal complement of the nuisance tangent space to a class of estimating functions.

It is assumed that the parameter $\mu = \Phi(F_X) \in \mathbb{R}^k$ of interest is pathwise differentiable in the full data model $\mathcal{M}^F$ with canonical gradient $S^{*F}_{eff}(\cdot \mid F_X) \in L^2_0(F_X)$ relative to a class of parametric submodels with tangent space $T^F(F_X)$. The canonical gradient is also called the efficient influence curve. The canonical gradient $S^{*F}_{eff}(X \mid F_X)$ is of great importance since the asymptotic variance of a regular asymptotically linear estimator of β at $P_{F_X,G}$ is bounded below by the variance of the canonical gradient and a regular estimator is efficient at $P_{F_X,G}$ if and only if it is asymptotically linear with influence curve equal to the canonical gradient (i.e., efficient influence curve) at $P_{F_X,G}$. Let $T^F_{nuis}(F_X) \subset L^2_0(F_X)$ be the nuisance tangent space in the full data model $\mathcal{M}^F$ (i.e., the closure of the linear span of all scores of 1-dimensional submodels F_ϵ through F at $\epsilon = 0$ for which $d/d\epsilon\mu(F_\epsilon)|_{\epsilon=0} = 0$). Let $T^{F,\perp}_{nuis}(F_X)$ be the orthogonal complement of $T^F_{nuis}(F_X) \subset L^2_0(F_X)$.

We will index each element of $T^{\perp}_{nuis}(F_X)$ with an index h running over an index set $\mathcal{H}^F(F_X)$. Specifically, assume we can represent

$$T^{F,\perp}_{nuis}(F_X) = \{D_h(X \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\}, \tag{2.1}$$

where $\rho(F_X)$ is a parameter defined on $\mathcal{M}^F$. Note that these index sets $\mathcal{H}^F(F_X)$ will typically depend on F_X. For example, in the multivariate generalized linear regression model $E(Z \mid X^*) = g(X^* \mid \beta)$ of Lemma 2.1 below, we have $T^{F,\perp}_{nuis}(F_X) = \{D_h(X \mid \beta) = h(X^*)\epsilon(\beta) : h \in \mathcal{H}^F(F_X)\}$ with $\mathcal{H}^F(F_X) = \{h(X^*) : E_{F_X}\{h(X^*)\epsilon(\beta)\}^2 < \infty\}$. In words, in this case the index h is allowed to be any function of X^* so that $h(X^*)\epsilon(\beta)$ has finite variance w.r.t. F_X. Let $h_{ind,F_X} : L^2_0(F_X) \to \mathcal{H}^F(F_X)$ be the index mapping defined by

$$\Pi(V \mid T^{F,\perp}_{nuis}(F_X)) = D_{h_{ind,F_X}(V)}(\cdot \mid \mu(F_X), \rho(F_X)). \tag{2.2}$$

Let $h_{eff}(F_X) = h_{ind,F_X}(S^{*F}_{eff}(\cdot \mid F_X))$ be the index of the full data canonical gradient. In general, the mapping h_{ind,F_X} is determined by the mapping $\Pi(\cdot \mid T^{\perp}_{nuis}(F_X))$ and the representation (2.1) in the same manner as above.

As an illustration we consider the generalized linear regression example. Lemma 2.1 teaches us that for $D \in L^2_0(F_X)$

$$\Pi(D \mid T^{F,\perp}_{nuis}) = E(D(X)\epsilon^\top \mid X^*)E(\epsilon\epsilon^\top \mid X^*)^{-1}\epsilon.$$

Thus, we have that the index mapping is given by

$$h_{ind,F_X}(D) = E(D(X)\epsilon^\top \mid X^*)E(\epsilon\epsilon^\top \mid X^*)^{-1}.$$

In particular, $h_{eff}(F_X) = h_{ind,F_X}(S^F_{eff})$, which can be simplified as in Lemma 2.1 below.

2.2.1 Orthogonal complement of the nuisance tangent space in the multivariate generalized linear regression model (MGLM)

The following lemma provides the orthogonal complement of the nuisance tangent space in multivariate generalized linear regression models $Z = g(X^* \mid \alpha) + \epsilon$, the projection onto this space, and the efficient score. We allow these models to model a user-supplied location parameter of the conditional error distribution by requiring that $E(K(\epsilon) \mid X^*) = 0$ for a user-supplied monotone function $K(\epsilon)$. For example, if $K(\epsilon) = \epsilon$, then the regression curve $g(X^* \mid \alpha)$ models the mean, if $K(\epsilon) = I(\epsilon > 0) - 1/2$, then it models the median, and, in general, if $K(\epsilon) = I(\epsilon > 0) - (1-p)$, then it models the pth quantile of the conditional error distribution of ϵ, given X^*. Allowing this flexibility is particularly crucial for the censored data models since estimation of a mean based on censored data might not be possible due to lack of data in the tails of the distribution, while the median might not be a smooth enough functional of the observed data (e.g., see Chapter 4). By truncating $K(\epsilon) = \epsilon$ for $\mid \epsilon \mid > M$ (i.e., set it equal to M), one obtains a truncated mean, and by setting $K(\epsilon)$ equal to a smooth approximation of $I(\epsilon > 0) - 1/2$, one obtains a smooth median.

Lemma 2.1 *Let Z be a p-dimensional vector of outcomes. Suppose that we observe n i.i.d. observations of $X = (Z, X^*)$ for some vector of covariates X^*. Consider the multivariate regression model of Z on X^*,*

$$Z = g(X^* \mid \alpha) + \epsilon, \;\; E(K(\epsilon) \mid X^*) = 0, \tag{2.3}$$

where $g = (g_1, \ldots, g_p)^\top$ is a p-dimensional vector of functions $g_j(X^* \mid \alpha)$, ϵ is a p-dimensional vector of residuals, K is a given real-valued monotone increasing function with $K(-\infty) < 0$ and $K(\infty) > 0$, and $\alpha = (\alpha_1, \ldots, \alpha_q)$ is a q-dimensional regression parameter. Here $K(\epsilon) = (K(\epsilon_1), \ldots, K(\epsilon_p))^\top$ is a p-dimensional vector.

The orthogonal complement of the nuisance tangent space at F_X is given by

$$T^{F,\perp}_{nuis}(F_X) = \{h(X^*)K(\epsilon) : h \; 1 \times p \; vector\}.$$

The projection $\Pi(V \mid T^{F,\perp}_{nuis})$ onto this subspace of the Hilbert space $L^2_0(F_X)$, endowed with inner product $\langle f, g\rangle_{F_X} = E_{F_X} f(X)g(X)$, is given by

$$E(\{V(X) - E(V \mid X^*)\}K(\epsilon)^\top \mid X^*)E(K(\epsilon)K(\epsilon)^\top \mid X^*)^{-1}K(\epsilon). \tag{2.4}$$

Assume that the conditional distribution of ϵ, given X^*, has a Lebesgue p-variate density $f(\epsilon \mid X^*)$. The score for α_j is given by

$$S_j(X) = -\frac{d}{d\alpha_j} g(X^* \mid \alpha)^\top \frac{f'(\epsilon \mid X^*)}{f(\epsilon \mid X^*)},$$

where $f'(\epsilon \mid X^*)$ is a p-dimensional vector containing the p partial derivatives w.r.t. $\epsilon_1, \ldots, \epsilon_p$. We can represent the q-dimensional score vector $S(X) = (S_1(X), \ldots, S_q(X))^\top$ as

$$S(X) = -\frac{d}{d\alpha} g(X^* \mid \alpha)^\top_{q\times p} \frac{f'(\epsilon \mid X^*)}{f(\epsilon \mid X^*)}.$$

The efficient score is given by

$$\begin{aligned} S^* &\equiv \Pi(S_j \mid T^{F,\perp}_{nuis}(F_X))^q_{j=1} \\ &= -\frac{d}{d\alpha} g(X^* \mid \alpha)^\top_{q\times p} A(X^*)_{p\times p} E(K(\epsilon)K(\epsilon)^\top \mid X^*)^{-1}_{p\times p} K(\epsilon), \end{aligned}$$

where $A(X^*) \equiv E\left(\frac{f'(\epsilon|X^*)}{f(\epsilon|X^*)} K(\epsilon)^\top \mid X^*\right)_{p\times p}$. If we assume that $f(\epsilon \mid X^*)$ equals zero at the end of its support and K is absolutely continuous w.r.t. the Lebesgue measure, then by integration by parts it follows that

$$A(X^*) = -\mathrm{diag}\left(E(K'(\epsilon) \mid X^*)\right)_{p\times p},$$

where $\mathrm{diag} E(K'(\epsilon) \mid X^*)$ *denotes the $p\times p$ diagonal matrix with jth diagonal element $E(K'(\epsilon_j) \mid X^*)$. As a consequence, under this assumption, we have that the efficient score vector is given by*

$$S^* = \frac{d}{d\alpha} g(X^* \mid \alpha)^\top_{q\times p} \mathrm{diag}\left(E(K'(\epsilon) \mid X^*)\right)_{p\times p} E(K(\epsilon)K(\epsilon)^\top \mid X^*)^{-1}_{p\times p} K(\epsilon).$$

For example, if $p = 1$ and $K(\epsilon) = (I(\epsilon > 0) - 1/2)$, which corresponds with median regression, then

$$S^*_j(X) = f_{\epsilon|X^*}(0 \mid X^*) \frac{d}{d\alpha_j} g(X^* \mid \alpha)^\top E(K^2(\epsilon) \mid X^*)^{-1} K(\epsilon).$$

Proof. The density of X can be written as

$$f_X(X) = f(Z \mid X^*) f_{X^*}(X^*) = f_{\epsilon|X^*}(\epsilon(\alpha) \mid X^*) f_{X^*}(X^*).$$

This density is indexed by the parameter of interest α, f_{X^*} and the conditional distribution of $\epsilon = \epsilon(\alpha)$, given X^*, where the latter ranges over all conditional distributions with conditional expectation of $K(\epsilon)$, given X^*, equal to zero. Here f_{X^*} and $f_{\epsilon|X^*}$ are the nuisance parameters.

Let α be fixed. For any uniformly bounded function $s(X^*)$ with $E(s(X^*)) = 0$ and uniformly bounded function $s(\epsilon \mid X^*)$ with $E(s(\epsilon \mid X^*) \mid X^*) = E(s(\epsilon \mid X^*)K(\epsilon) \mid X^*) = 0$, we have that

$$f_\delta(X) = (1 + \delta s(X^*)) f_{X^*}(X^*)(1 + \delta s(\epsilon \mid X^*)) f_{\epsilon|X^*}(\epsilon \mid X^*)$$

is a one-dimensional submodel of the full data model with parameter δ going through the truth f_X at $\delta = 0$. Notice that these one-dimensional models provide a rich class of fluctuations for our nuisance parameter. The nuisance tangent space is defined by the closure of the linear span of the scores of this class of one-dimensional submodels in the Hilbert space $L_0^2(F_X) = \{s(X) : Es(X) = 0, Es^2(X) < \infty\}$ endowed with the inner product $\langle f, g \rangle_{F_X} = E_{F_X} f(X)g(X)$. It is given by the orthogonal sum of the two spaces generated by the $s(X^*)$'s and $s(\epsilon \mid X^*)$'s,

$$T_{nuis}(F_X) = L_0^2(F_{X^*}) \oplus H,$$

where $H \subset L_0^2(F_X)$ is the Hilbert space of functions s satisfying $E(s(\epsilon \mid X^*) \mid X^*) = E(s(\epsilon \mid X^*)K(\epsilon) \mid X^*) = 0$.

We have that $\Pi(V \mid L_0^2(F_{X^*})) = E(V \mid X^*)$. Let $H^+ \supset H$ be the Hilbert space of functions s only satisfying $E(s(\epsilon \mid X^*) \mid X^*) = 0$. We have $\Pi(V \mid H^+) = V - E(V \mid X^*)$. Now note that H consists of the orthogonal complement of the p-dimensional space $\langle K(\epsilon_1), \ldots, K(\epsilon_p) \rangle$ in the world where X^* is fixed. Thus, the projection operator onto this space is the identity operator minus the projection onto $\langle K(\epsilon_1), \ldots, K(\epsilon_p) \rangle$. The projection onto a p-dimensional space of functions $(K(\epsilon_j) : j = 1, \ldots, p)$ is given by the formula

$$E(V(X)K(\epsilon)^\top)E(K(\epsilon)K(\epsilon)^\top)^{-1}K(\epsilon).$$

Now, we simply need to apply this formula in the world with X^* fixed, so we have for any function $\eta \in H^+$

$$\Pi(\eta \mid H) = \eta(X) - E(\eta(X)K(\epsilon) \mid X^*)^\top \left\{E(K(\epsilon)K(\epsilon)^\top \mid X^*)\right\}^{-1} K(\epsilon).$$

The rest of the proof is straightforward. □

2.2.2 Orthogonal complement of the nuisance tangent space in the multiplicative intensity model

Suppose that the full data $X = \bar{X}(T) = (X(t) : 0 \leq t \leq T)$ is a stochastic time-dependent process up to a possibly random time T. In addition, suppose that $X(t) = (N(t), V_1(t), V_2(t))$, where $N(t)$ is a counting process of interest and $V(t) = (V_1(t), V_2(t))$ is a time-dependent covariate process. Let $R(t) = I(T \leq t)$ be a component of $N(t)$ so that observing the process $X(t)$ up to T includes observing T itself. Let $Z(t) = (N(t), V_1(t))$. In these settings, there is often interest in modeling the intensity of $N(t)$ w.r.t. history $\bar{Z}(t-) = (Z(s) : s < t)$. The following lemmas provide us with the orthogonal complement of the nuisance tangent space for this multiplicative intensity model, the projection operator onto this space, and the efficient score.

Lemma 2.2 *Consider the setting above. Consider the model for the distribution of $X = \bar{X}(T)$ defined by the multiplicative intensity model (for a*

continuous counting process $N(t)$) assumption:

$$\lambda(t)dt \equiv E(dN(t) \mid \bar{Z}(t-)) = Y(t)\lambda_0(t)\exp(\beta W(t))dt,$$

where $Y(t)$ and the k-dimensional vector $W(t)$ are uniformly bounded functions of $\bar{Z}(t-)$. Here $Y(t)$ is the indicator that $N(t)$ is still at risk of jumping right before time t. Here $\beta \in \mathbb{R}^k$ and λ_0 are unspecified. Let β be the parameter of interest and let (λ_0, η) represent the nuisance parameter (so β, λ_0, η identify f_X). Let $dM(t) = dN(t) - E(dN(t) \mid \bar{Z}(t-))$. The orthogonal complement of the nuisance tangent space in the model in which λ_0 is known is given by

$$T_\eta^\perp \equiv \overline{\left\{ \int H(t, \bar{Z}(t-))dM(t) : H \right\}} \cap L_0^2(F_X).$$

The nuisance tangent space of $\Lambda_0 = \int_0^\cdot \lambda_0$ is given by

$$T_{\Lambda_0} \equiv \overline{\left\{ \int g(t)dM(t) : g \right\}} \cap L_0^2(F_X).$$

We have

$$\Pi\left(\int H(t, \bar{Z}(t-))dM(t) \mid T_{\Lambda_0} \right) = \int g(H)(t)dM(t),$$

where

$$g(H)(t) = \frac{E\left\{H(t, \bar{Z}(t-))Y(t)\exp(\beta W(t))\right\}}{E\{Y(t)\exp(\beta W(t))\}}. \tag{2.5}$$

Thus, the orthogonal complement of the nuisance tangent space of β $T_{nuis}^{F,\perp} = T_\eta^\perp \cap T_{\Lambda_0}^\perp$ is given by

$$T_{nuis}^{F,\perp} = \overline{\left\{ \int \{H(t, \bar{Z}(t-)) - g(H)(t)\}dM(t) : H \right\}} \cap L_0^2(F_X).$$

We have

$$\Pi\left(\int H(t, \bar{Z}(t-))dM(t) \mid T_{nuis}^{F,\perp} \right) = \int \{H(t, \bar{Z}(t-)) - g(H)\}dM(t).$$

The score for β is given by $S_\beta = \int W(t)dM(t)$. Thus the efficient score for β is given by

$$S_{eff}^F = \int \left\{ W(t) - \frac{E\left\{W(t)Y(t)\exp(\beta W(t))\right\}}{E\{Y(t)\exp(\beta W(t))\}} \right\} dM(t).$$

This efficient score formula is due to Ritov, and Wellner (1988). We want to have an expression for the projection onto $T_{nuis}^{F,\perp}$ of any function $D(X)$. The previous lemma provides the projection of full data functions of the form $\int H(t, \bar{Z}(t-))dM(t)$. In the next lemma, we establish the projection in the case where $N(t)$ can only jump at a given set of points, thereby avoiding

technical measurability conditions. Since continuous processes can be arbitrarily well-approximated by discrete processes, it will also provide us with a formula for the general projection operator onto $T_{nuis}^{F,\perp}$ for the continuous multiplicative intensity model. Note that the multiplicative intensity model makes only sense for discrete data on a relatively fine grid of points so that the modeled probabilities are bounded by 1.

Lemma 2.3 *Assume that the counting process $N(t)$ can only jump at given points t_j, $j = 1, \ldots, p$, and consider the multiplicative discrete intensity model $\lambda(t_j) = P(dN(t_j) = 1 \mid \bar{Z}(t_j-)) = Y(t_j)\lambda_0(t_j)\exp(\beta W(t_j))$, where $W(t_j)$ are uniformly bounded functions of $\bar{Z}(t_j)$, $j = 1, \ldots, p$. Let $dM(t) = dN(t) - \lambda(t)$ for $t \in \{t_1, \ldots, t_p\}$. Then, the statements in Lemma 2.2 hold and, in addition, we have that for any $D \in L^2(F_X)$*

$$\Pi(D \mid T_\eta^\perp) = \int H_D(t, \bar{Z}(t-))dM(t), \tag{2.6}$$

where

$$H_D(t, \bar{Z}(t-)) = E(D(X) \mid dN(t) = 1, \bar{Z}(t-)) - E(D(X) \mid dN(t) = 0, \bar{Z}(t-)).$$

Thus

$$\Pi(D \mid T_{nuis}^{F,\perp}) = \int \{H_D(t, \bar{Z}(t-)) - g(H_D)\}dM(t), \tag{2.7}$$

where the mapping $g(h)$ is defined in (2.5).

We conjecture that, given appropriate measurability conditions so that the conditional expectations are properly defined, this projection formula (2.6) and thereby (2.7) holds in the continuous setting of Lemma 2.2 as well.

A direct proof of the representation $T_{nuis}^{F,\perp}$ given in Lemma 2.2 is obtained by directly computing the nuisance tangent space from the likelihood $f(X) = f(X \mid Z)\prod_t f(Z(t) \mid \bar{Z}(t-))$ which can be further factorized by $f(Z(t) \mid \bar{Z}(t-)) = f(dN(t) \mid \bar{Z}(t-))f(Z(t) \mid N(t), \bar{Z}(t-))$. Here $\prod_t f(dN(t) \mid \bar{Z}(t-)) = \prod_t \lambda(t)^{dN(t)}(1 - \lambda(t))^{1-dN(t)}$ is the partial likelihood, where this product integral representation of the likelihood is formally defined in Andersen, Borgan, Gill and Keiding (1993). The proof Lemma 2.3 below provides an intuitive non-formal way of understanding Lemma 2.2 and provides a formal proof of Lemma 2.3 in which $N(t)$ is a discrete counting process.

Proof of Lemma 2.3. Let $\mathcal{M} = \{F_X : E(dN(t) \mid \bar{Z}(t-)) = Y(t)\lambda_0(t)\exp(\beta W(t))$ all $t\}$ be this model. Let (η, λ_0) represent the nuisance parameter of β. We have that the nuisance tangent space T_{η,λ_0} equals the sum of the nuisance tangent space T_η in the model with λ_0 known and the nuisance tangent space T_{λ_0} in the model with η known: $T_{\eta,\lambda_0} = \overline{T_\eta + T_{\lambda_0}}$. Thus $T_{\eta,\lambda_0}^\perp = T_\eta^\perp \cap T_{\lambda_0}^\perp$. It follows directly from differentiating the log-partial-likelihood

$\log \prod_t \lambda(t)^{dN(t)}(1-\lambda(t))^{1-dN(t)}$ along one-dimensional fluctuations $\lambda_0(\cdot)+\epsilon h(\cdot)$ of λ_0 that $T_{\lambda_0} = \overline{\{\int g(t)dM(t) : g\}}$.

We will now prove that $T_\eta^\perp = \overline{\{\int H(t,\bar{Z}(t-))dM(t) : H\}}$. Notice that $\mathcal{M} = \cap_{t=1}^p \mathcal{M}_t$, where $\mathcal{M}_t = \{F_X : \lambda(t) = Y(t)\lambda_0(t)\exp(\beta W(t))\}$. In other words, $\mathcal{M}$ can be viewed as an intersection of t-specific models only restricting the intensity $\lambda(t) = E(dN(t) \mid \bar{Z}(t-))$ at a fixed point t. The orthogonal complement of the nuisance tangent space of β in model $\mathcal{M}_t$ equals $\overline{\{H(\bar{Z}(t-))dM(t) : H\}}$. This is proved directly from the likelihood representation in the same manner as we proved that in the regression model $E(Z \mid X^*) = m(X^* \mid \beta)$ the orthogonal complement of the nuisance tangent space equals $\{H(X^*)(Z - m(X^* \mid \beta)) : H\}$. In fact, since N can only jump at predetermined grid points $\mathcal{M}_t$ can be viewed as a regression model of $Z = dN(t)$ on $X^* = \bar{Z}(t-)$ with $m(X^* \mid \beta) = Y(t)\lambda_0(t)\exp(\beta W(t))$. The orthogonal complement of the nuisance tangent space of the intersection of models $\mathcal{M}_t$ equals the sum (integral) of the orthogonal complements of the nuisance tangent spaces for the models $\mathcal{M}_t$, where the nuisance tangent space for $\mathcal{M}$ equals the intersection of the nuisance tangent spaces for model $\mathcal{M}_t$. Thus, the orthogonal complement of the nuisance tangent space in the model with λ_0 known equals

$$T_\eta^\perp = \overline{\left\{\int H(t,\bar{Z}(t-))dM(t) : H\right\} \cap L_0^2(F_X)}.$$

Therefore, we can conclude that

$$T_\eta^\perp \cap T_{\lambda_0}^\perp = \overline{\left\{\int (H(t,\bar{Z}(t-)) - g(H)(t))dM(t) : H\right\}},$$

where $\int g(H)(t)dM(t) = \Pi(\int HdM \mid T_{\lambda_0})$. It can be directly verified that $g(H)(t) = E\{H(t,\bar{Z}(t-))Y(t)\exp(\beta W(t))\}/E\{Y(t)\exp(\beta W(t))\}$.

We will now prove the projection formula (2.6). Firstly, we note that $T_\eta^\perp = \{\int H(t,\bar{Z}(t-))dM(t) : H\} = H_1 \oplus \ldots \oplus H_k$ is an orthogonal sum of subspaces $H_j \equiv \{H(\bar{Z}(t_j-))dM(t_j) : H\}$. Therefore, we have that $\Pi(D \mid T_\eta^\perp) = \sum_{j=1}^k \Pi(D \mid H_j)$. As explained above, we can apply Lemma 2.4 with $\epsilon = dM(t_j)$, $X^* = \bar{Z}(t_j-)$, and $K(\epsilon) = \epsilon$ to obtain that $\Pi(D \mid H_j)$ is given by

$$E(\{D(X) - E(D(X) \mid \bar{Z}(t_j-))\}dM(t_j) \mid \bar{Z}(t_j-))\times \frac{1}{E(dM(t_j)^2 \mid \bar{Z}(t_j-))}dM(t_j).$$

We have

$$E(D(X) \mid dN(t_j), \bar{Z}(t_j-)) - E(D(X) \mid \bar{Z}(t_j-)) = \left\{E(D(X) \mid dN(t_j) = 1, \bar{Z}(t_j-)) - E(D(X) \mid dN(t_j) = 0, \bar{Z}(t_j-))\right\} dM(t_j)$$

This proves that $\Pi(D \mid H_j)$ is given by

$$\left\{E(D(X) \mid dN(t_j) = 1, \bar{Z}(t_j-)) - E(D(X) \mid dN(t_j) = 0, \bar{Z}(t_j-))\right\} dM(t_j),$$

which proves Lemma 2.3. □

Suppose now that the counting process is discrete on a sparse set of points so that one might want to assume the logistic regression intensity model. The proof of the previous lemma proves, in particular, a simple representation of $T^{F,\perp}_{nuis}$ and the projection operator onto $T^{F,\perp}_{nuis}$ for parametric discrete intensity models such as the logistic regression intensity model. The results are stated in the following lemma.

Lemma 2.4 *Assume that the counting process $N(t)$ can only jump at given points t_j, $j = 1, \ldots, p$, and consider a discrete intensity model $\lambda(t_j) = P(dN(t_j) = 1 \mid \bar{Z}(t_j-)) = Y(t_j)m(W(t_j), t_j \mid \beta)$, where $m(W(t_j), t_j \mid \beta)$ is parametrized by a k-dimensional regression parameter β and is uniformly bounded: For example, $m(W(t), t \mid \beta) = 1/(1+\exp(-(\beta_0 + \beta_1 * t + \beta_2 W(t))$. Let $dM(t) = dN(t) - \lambda(t)$ for $t \in \{t_1, \ldots, t_p\}$. Then, the orthogonal complement of the nuisance tangent space at F_X is given by*

$$T^{F,\perp}_{nuis}(F_X) = \overline{\left\{ \int H(t, \bar{Z}(t-))dM(t) : H \right\}} \cap L^2_0(F_X).$$

In addition, we have that for any $D \in L^2(F_X)$

$$\Pi(D \mid T^{F,\perp}_{nuis}) = \int H_D(t, \bar{Z}(t-))dM(t), \tag{2.8}$$

where

$$H_D(t, \bar{Z}(t-)) = E(D(X) \mid dN(t) = 1, \bar{Z}(t-)) - E(D(X) \mid dN(t) = 0, \bar{Z}(t-)).$$

2.2.3 Linking the orthogonal complement of the nuisance tangent space to estimating functions

Consider the full data structure model $\mathcal{M}^F$ with parameter of interest $\mu = \mu(F_X)$. Given representations of $T^{F,\perp}_{nuis}(F_X) = \{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\}$ at all $F_X \in \mathcal{M}^F$, the goal is to define a class of full data estimating functions $\{(X, \mu, \rho) \to D_h(X \mid \mu, \rho) : h \in \mathcal{H}^F\}$ for μ with a (possibly different) nuisance parameter $\rho = \rho(F_X)$ and an index set $\mathcal{H}^F$ independent of F_X so that

$$\{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F\} \subset T^{F,\perp}_{nuis}(F_X) \text{ for all } F_X \in \mathcal{M}^F. \tag{2.9}$$

Recall that it yields estimating functions indexed by $\mathcal{H}^{Fk}$ for $\mu \in \mathbb{R}^k$ by defining $D_h = (D_{h_1}, \ldots, D_{h_k})$ for any $h = (h_1, \ldots, h_k) \in \mathcal{H}^{Fk}$.

In this subsection, we provide a template for deriving such a class of estimating functions from these representations of $T^{F,\perp}_{nuis}(F_X)$. Firstly, let $\mathcal{H}^F$ be an index set containing each $\mathcal{H}^F(F_X)$, $F_X \in \mathcal{M}^F$, and $(D^1_h : h \in \mathcal{H}^F)$ be a class of estimating functions $D^1_h : \mathcal{X} \times \{(\mu(F_X), \rho(F_X)) : F_X \in \mathcal{M}^F\} \to \mathbb{R}$ so that

$$\{D^1_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\} = T^{F,\perp}_{nuis}(F_X) \text{ for all } F_X \in \mathcal{M}^F.$$

For example, $\mathcal{H}^F = \cup_{F_X \in \mathcal{M}^F} \mathcal{H}^F(F_X)$. Since $T^{F,\perp}_{nuis}(F_X)$ is defined in $L^2_0(F_X)$, we mean that for each element in $T^{F,\perp}_{nuis}(F_X)$ there exists a function $D^1_h(X \mid \mu(F_X), \rho(F_X))$ that is equal to this element in $L^2_0(F_X)$. Now, $D^1_h(\cdot \mid \mu, \rho)$, $h \in \mathcal{H}^F$, is a class of (biased and unbiased) full data estimating functions.

Since $\mathcal{H}^F(F_X)$ possibly depends on unknown parameters of F_X, the membership indicator $I(h \in \mathcal{H}^F(F_X))$, which guarantees the unbiasedness, represents a nuisance parameter of the estimating function $D^1_h(\cdot \mid \mu, \rho)$. In order to acknowledge this fact, we reparametrize $D^1_h(\cdot \mid \mu, \rho)$ as follows. Let $\Pi(\cdot \mid \mathcal{H}^F(F_X))$ be a user-supplied mapping from $\mathcal{H}^F$ into $\mathcal{H}^F(F_X)$ satisfying $\Pi(h \mid \mathcal{H}^F(F_X)) = h$ if $h \in \mathcal{H}^F(F_X)$. We now redefine the class of full data estimating functions $\{D^1_h(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F\}$

$$\{D^2_h(\cdot \mid \mu, \rho') \equiv D_{\Pi(h \mid \mathcal{H}^F(F_X))}(\cdot \mid \mu, \rho) : h \in \mathcal{H}\}, \tag{2.10}$$

where ρ' denotes ρ augmented with the parameters indexing $\Pi(h \mid \mathcal{H}^F(F_X))$. For the sake of notational simplicity, we redefine $D^2_h(\cdot \mid \mu, \rho')$ as $D_h(\cdot \mid \mu, \rho)$ again. Note that we can now state

$$\{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F\} = T^{F,\perp}_{nuis}(F_X).$$

If $\Pi(\cdot \mid \mathcal{H}^F(F_X))$ were not required to be the identity on $\mathcal{H}^F(F_X)$, it might have a range that is a strict subset of $\mathcal{H}^F(F_X)$. Even so, we still would have

$$\{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F\} \subset T^{F,\perp}_{nuis}(F_X).$$

One might choose such a mapping $\Pi(\cdot \mid \mathcal{H}^F(F_X))$ to simplify the parametrization of the estimating function, where one now takes the risk of excluding (e.g.) the optimal estimating function.

As a side remark here, we mention that ρ plays the role of a nuisance parameter that will be estimated with external (relative to the estimating function) procedures. For example, in the full data world, we would be solving $0 = \sum_{i=1}^n D_h(X_i \mid \mu, \rho_n)$ for a given estimator ρ_n of ρ. Thus, one wants to choose ρ as variation-independent of μ as possible in order to maximize efficiency of the estimator of μ that solves the estimating equation. If $\rho = \rho(\mu, \eta)$ for two variation independent parameters μ, η, then one redefines the full data estimating functions as $D_h(\cdot \mid \mu, \eta) = D_h(\cdot \mid \mu, \rho(\mu, \eta))$. We conjecture that it essentially will always be possible to parametrize the estimating function so that μ and ρ are locally variation independent.

Such a collection $\{D_h : h \in \mathcal{H}^F\}$ represents a set of full data structure estimating functions. We want to choose the index set as large as possible in the sense that if $\rho = (\rho_1, \rho_2)$ with

$$E_{F_X} D_h(X \mid \mu(F_X), \rho_1, \rho_2(F_X)) \in T^{F,\perp}_{nuis}(F_X) \text{ for all possible } \rho_1$$

and $F_X \in \mathcal{M}^F$, then one should make ρ_1 a component of the index h.

In most full data models, the estimating functions D_h and index set $\mathcal{H}^F$ are naturally implied by the representation of $F_X \to T^{F,\perp}_{nuis}(F_X) = \{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\}$ and does not require much thinking.

Because it is of interest to be able to map a full data estimating function into its corresponding index, we also want to extend the index mapping h_{ind,F_X} (2.2) to be well-defined on pointwise well-defined functions of X. Let $\mathcal{D} = \{D_h(\cdot \mid \mu, \rho) : \mu, \rho, h \in \mathcal{H}^F\}$ be the set of full data functions. Let $\mathcal{L}(\mathcal{X})$ be the space of functions of X with finite supremum norm over a set K for which we know that $P(X \in K) = 1$ w.r.t. the true F_X. It will be assumed that $\mathcal{D} \subset \mathcal{L}(\mathcal{X})$. Let $h_{ind,F_X} : \mathcal{L}(\mathcal{X}) \to \mathcal{H}$ be an index mapping satisfying for any $D \in \mathcal{L}(\mathcal{X})$

$$D_{h_{ind,F_X}(D)}(\cdot \mid \mu(F_X), \rho(F_X)) = \Pi(D \mid T^{F,\perp}_{nuis}(F_X)),$$

where we formally mean that the equality holds in $L^2_0(F_X)$ (since the right-hand side is defined in $L^2_0(F_X)$).

Example 2.1 (Multivariate generalized linear regression; continuation of Example 2.1) In our multivariate generalized linear regression example, a natural candidate for the index set $\mathcal{H}^F$ is simply all functions of X^*:

$$\mathcal{H}^F \equiv \{h(X^*) : \text{ any h}\}.$$

A possible mapping into $\mathcal{H}^F(F_X) = \{h \in \mathcal{H}^F : E_{F_X}\{h(X^*)K(\epsilon(\alpha))\}^2 < \infty\}$ (here α is the true parameter value corresponding with F_X) is given by

$$\Pi(h \mid \mathcal{H}^F(F_X))(X^*) = \min(h(X^*), M),$$

where the truncation constant M is user-supplied. Notice that indeed the finite supremum norm of this index (say) h^* guarantees that $h^*(X^*)\epsilon(\alpha)$ has finite variance for any $F_X \in \mathcal{M}^F$. Note also that the range of this mapping does not necessarily cover $\mathcal{H}^F(F_X)$. This mapping has no unknown nuisance parameters. Thus, the corresponding set of full data estimating functions (2.10) is given by

$$\{(X, \alpha) \to \min(h(X^*), M)K(\epsilon(\alpha)) : h \in \mathcal{H}^F\}.$$

If the full data structure model $\mathcal{M}^F$ assumes that, for a specified k, $X^{*k}\epsilon(\beta)$ has finite variance, then one can define

$$\Pi(h \mid \mathcal{H}^F(F_X)) = \min(h(X^*), X^{*k}).$$

For any $D \in \mathcal{D} = \{\min(h(X^*), M)K(\epsilon(\alpha)) : h \in \mathcal{H}^F, \alpha\}$, we define

$$h_{ind,F_X}(D)(X^*) = E_{F_X}(D(X)K(\epsilon)^\top \mid X^*)E_{F_X}(K(\epsilon)K(\epsilon)^\top \mid X^*)^{-1},$$

where $\epsilon = \epsilon(\alpha(F_X))$. □

2.3 Mapping into Observed Data Estimating Functions

In this section, we provide a variety of methods to construct mappings from full data estimating functions to observed data estimating functions. The first five sections are focused on mappings that one can use to construct estimators in model $\mathcal{M}(\mathcal{G})$, and they can form the basis of an orthogonalized mapping such as the optimal mapping in the next section. In Subsection 2.3.6, we will show that making a given mapping orthogonal to the tangent space of G for a convex model containing $\mathcal{G}$ yields the double robustness property so that it can be used to obtain RAL estimators in model $\mathcal{M}$ as well.

2.3.1 Initial mappings and reparametrizing the full data estimating functions

Let $D_h \rightarrow IC_0(Y \mid Q_0, G, D_h)$ be a mapping from full data estimating functions $\{D_h : h \in \mathcal{H}^F\}$ into observed data estimating functions indexed by nuisance parameters $Q_0(F_X, G)$ and G. Let $\mathcal{Q}_0 \equiv \{Q_0(F_X, G) : F_X \in \mathcal{M}^{F,w}, G \in \mathcal{G}\}$ be the parameter space of this nuisance parameter Q_0, where F_X ranges over a submodel $\mathcal{M}^{F,w}$ of $\mathcal{M}^F$. For each possible parameter value $(\mu, \rho) \in \{(\mu(F_X), \rho(F_X)) : F_X \in \mathcal{M}^F\}$ and $G \in \mathcal{G}$, let $\mathcal{H}^F(\mu, \rho, \rho_1, G) \subset \mathcal{H}^F$ be a collection of h for which

$$E_G(IC_0(Y \mid Q, G, D_h(\cdot \mid \mu, \rho)) \mid X) = D_h(X \mid \mu, \rho) \ F_X\text{-a.e. for all } Q \in \mathcal{Q}^0 \tag{2.11}$$

and

$$\text{VAR}_{P_{F_X,G}} IC_0(Y \mid Q, G, D_h(\cdot \mid \mu(F_X), \rho(F_X))) < \infty \text{ for all } Q \in \mathcal{Q}^0. \tag{2.12}$$

Note that the statement F_X-a.e. in (2.11) also creates dependence on F_X. Since the latter restriction only affects the variance of the estimating function (it is unbiased by (2.11)) it can often be arranged by a simple truncation of h (see e.g., Example 2.1). Therefore, we suppressed the possible dependence of $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ on another nuisance parameter needed to guarantee (2.12). This dependence is expressed by the parameter ρ_1 of F_X.

It is also natural to make (2.12) a model assumption or, equivalently, a regularity condition in an asymptotics theorem. Note that, if we ignore the (2.12) constraint, then the maximal set $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ is given by

$$\left\{ h \in \mathcal{H}^F : \sup_{Q_0 \in \mathcal{Q}_0} \mid E_G(IC_0(Y \mid Q_0, G, D_h(\cdot \mid \mu, \rho) \mid X) - D_h(X \mid \mu, \rho) \mid = 0 \right\},$$

where the equality needs to hold F_X-a.e.

It will be convenient to also define the set of allowed full data estimating functions directly (instead of in terms of the index sets).

Definition 2.1 *Let* $\mathcal{D} = \{D_h(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F, (\mu, \rho)\}$ *with* (μ, ρ) *ranging over the parameter space* $\{\mu(F_X), \rho(F_X) : F_X \in \mathcal{M}^F\}$. *Let* $\mathcal{Q}_0 = \{Q_0(F_X, G) : F_X \in \mathcal{M}^F, G \in \mathcal{G}\}$. *For each* $G \in \mathcal{G}$ *and* $F_X \in \mathcal{M}^F$, *we define the set* $\mathcal{D}(\rho_1(F_X), G)$ *as*

$$\{D \in \mathcal{D} : E_G(IC_0(Y \mid Q, G, D) \mid X) = D(X) \ F_X\text{-a.e. for all } Q \in \mathcal{Q}_0. \tag{2.13}$$

Given an initial mapping IC_0, the dependence of $\mathcal{D}(\rho_1, G)$ on F_X, G typically has to do with the support of $D(X)$; that is the possible values of X at which $D(X)$ is non zero relative to the support of G. As a consequence, under strong conditions on G, one will typically have $\mathcal{D}(\rho_1, G) = \mathcal{D}$.

Example 2.2 (Right censored data structure with time-dependent covariates) Consider the right-censored data structure $Y = (\tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T), \bar{L}(\tilde{T}))$. For $D(X)$ we define $\Delta(D) = I(D(X)$ observed). There exists a real-valued random variable $V(D) \leq T$ so that $I(D(X)$*is observed*$) = I(C \geq V(D))$. We define

$$IC_0(Y \mid G, D) = \frac{D(X)\Delta(D)}{P_G(\Delta(D) = 1 \mid X)} = \frac{D(X)\Delta(D)}{\bar{G}(V(D) \mid X)},$$

where $\bar{G}(t \mid X) \equiv P(C \geq t \mid X)$. Note that if $\bar{G}(T \mid X) > \delta > 0$ F_X-a.e. for some $\delta > 0$, then one has $\mathcal{D}(\rho_1, G) = \mathcal{D}$, but this condition might not be necessary for identification of a particular parameter μ. □

Having identified an appropriate mapping IC_0, in many models it is indeed the case that $E_G(IC_0(Y \mid Q, G, D_h(\cdot \mid \mu(F_X), \rho(F_X))) \mid X) = D_h(X \mid \mu(F_X), \rho(F_X))$ only holds for $D_h(\cdot \mid \mu(F_X), \rho(F_X))$ ranging over a true (not even dense) subset of $T^{F,\perp}_{nuis}(F_X)$ (i.e., $\mathcal{H}^F(\mu(F_X), \rho(F_X), \rho_1(F_X), G) \subset \mathcal{H}^F$). In this case, there exist many full data structure estimating functions (i.e., full data structure model gradients) that cannot be mapped into an observed data estimating function (i.e., observed data model $\mathcal{M}(G)$ gradients).

Typical candidates of the mapping $D_h \rightarrow IC_0(Y \mid Q_0, G, D_h)$ are so-called inverse probability of censoring weighted mappings, as we provided in Chapter 1 and will provide below for various censored data structures. These mappings involve a censoring probability or density in the denominator. By assuming that this censoring probability is uniformly bounded away from zero, one will typically have $\mathcal{H}^F(\mu, \rho, \rho_1, G) = \mathcal{H}^F$ (i.e., all full data structure estimating functions satisfy (2.11)). On the other hand, a given censoring distribution not satisfying this property can still allow estimation of the particular parameter of interest or, equivalently, there is still a real subset $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ of $\mathcal{H}^F$ for which (2.11) holds. For example, in the current status data location (mean, smooth median, truncated mean)

regression model covered in Chapter 4, the class of allowed full data structure estimating functions is a function of the support of the monitoring mechanism and the support of the location parameter. The next example illustrates this for linear regression with a right-censored outcome.

Example 2.3 (Linear regression with right-censored outcome) Suppose that one is interested in estimating a linear regression parameter $\mu = \beta$ of log-survival on a treatment dose Z. We denote the log survival and log censoring time by T and C. Our model is $T = \beta_0 + \beta_1 Z + \epsilon$, $E(\epsilon \mid Z) = 0$. Let $X = (T, Z)$ be the full data structure. We assume that the right-censoring time C is conditionally independent of survival, given treatment dose Z. We observe n i.i.d. copies of $Y = (\tilde{T} = \min(T, C), \Delta = I(T \leq C), Z)$. A rich class of full data structure estimating functions is $\{D_h(T, Z \mid \beta) = \min(h(Z), M)\epsilon(\beta) : h\}$ for a bound $M < \infty$ to guarantee that all of these estimating functions have finite variance. Let $IC_0(Y \mid G, D_h) = \min(h(Z), M)\epsilon(\beta)\Delta/\bar{G}(T \mid Z)$ which equals zero if $\Delta = 0$, regardless of the denominator, where $\bar{G}(t \mid Z) = P(C \geq t \mid Z)$.

Suppose that T, given Z, has compact support $[\alpha_Z, \alpha^Z]$. For example, if ϵ has known support $[-\tau, \tau]$, then we have $\alpha^Z = \beta_0 + \beta_1 Z + \tau$. Note that

$$E\left(h(Z)\epsilon(\beta)I(T \leq C)/\bar{G}(T \mid Z) \mid T, Z\right) = h(Z)\epsilon(\beta)I(\bar{G}(T \mid Z) > 0).$$

Define for some fixed small $\delta > 0$

$$\mathcal{Z}(\beta, G) = \{z : \bar{G}(\alpha^Z \mid Z) > \delta > 0\}$$

as the set of treatment values z for which the conditional probability (given treatment) $\bar{G}(T \mid Z = z)$ that a subject's survival time is observed is bounded away from zero. It follows that for all functions $h(Z)$ that are zero for $Z \notin \mathcal{Z}(\beta, G)$

$$E_G(IC_0(Y \mid G, D_h(\cdot \mid \beta) \mid X) = D_h(X \mid \beta) \text{ for all } G \text{ and } \beta.$$

Thus, we can set

$$\begin{aligned} \mathcal{H}^F(\mu, \rho, \rho_1, G) &= \mathcal{H}(\beta, G) = \{h(Z) : h(Z) = h(Z)I(Z \in \mathcal{Z}(\beta, G))\}, \\ \mathcal{D}(\beta, G) &= \{h(Z)\epsilon(\beta) : h(Z) = 0 \text{ if } \bar{G}(\alpha^Z \mid Z) = 0\}. \end{aligned}$$

In other words, one can estimate β by simply throwing away all subjects with a treatment dose outside $\mathcal{Z}(\beta, G)$. This example can be directly generalized to the general location regression model: $T = \beta_0 + \beta_1 Z + \epsilon$, $E(K(\epsilon) \mid Z) = 0$ for a monotone function K that has derivative zero outside an interval, say $(-\tau, \tau)$. □

Reparametrizing the full data structure estimating functions

Since $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ possibly depends on the unknown parameters (μ, ρ, ρ_1, G), this makes the index h essentially a nuisance parameter of the estimating function $IC_0(Y \mid Q, G, D_h(\cdot \mid \mu, \rho))$. In other words, to estimate μ, we need to estimate the set $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ and try to make sure

that our choice h_n converges to an element in $\mathcal{H}(\mu, \rho, \rho_1, G)$. In order to acknowledge this fact, we reparametrize $D_h(\cdot \mid \mu, \rho)$ in the following manner, which is completely analogous to (2.10).

Let $\Pi(\cdot \mid \mathcal{H}^F(\mu, \rho, \rho_1, G))$ be a user-supplied mapping from $\mathcal{H}^F$ into $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ that only depends on the unknown (μ, ρ, ρ_1, G), which equals the identity on a rich subset (preferably all) of $\mathcal{H}^F(\mu, \rho, \rho_1, G)$. We now redefine the class of full data estimating functions $\{D_h(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F\}$ so that it is guaranteed to satisfy (2.11) for all $h \in \mathcal{H}^F$:

$$\{D_h^r(\cdot \mid \mu, \rho' = (\rho, \rho_1, G)) \equiv D_{\Pi(h|\mathcal{H}^F(\mu,\rho,\rho_1,G))}(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F\}. \quad (2.14)$$

For the sake of notational simplicity, we will denote the parameter ρ' with ρ, again, and we denote $D_h^r(\cdot \mid \mu, \rho')$ with $D_h(\cdot \mid \mu, \rho)$ again, but we now need to remind ourselves that ρ possibly also includes G as a component and that $h \to D_h$ can be a many-to-one mapping in the sense that many $h \in \mathcal{H}^F$ are mapped into the same full data structure estimating function. The reparametrized class of estimating functions are now elements of $T_{nuis}^{F,\perp}(F_X)$ *and* $\mathcal{D}(\rho_1(F_X), G)$ when evaluated at the true parameter values. Consequently, we now have for *all* $h \in \mathcal{H}^F$

$$E_G(IC_0(Y \mid Q, G, D_h(\cdot \mid \mu, \rho)) \mid X) = D_h(X \mid \mu, \rho) \text{ for all } Q \in \mathcal{Q}^0. \quad (2.15)$$

This and (2.12) imply that

$$\{IC_0(Y \mid Q, G, D_h(\cdot \mid \mu(F_X), \rho(F_X, G))) : h \in \mathcal{H}^F, Q \in \mathcal{Q}^0\} \subset T_{nuis}^{\perp}(\mathcal{M}(G));$$

that is, IC_0 maps full data structure estimating functions into observed data estimating functions that are orthogonal to the nuisance tangent space in the model with G known. Consequently, the estimating function still has the property that the first-order asymptotics of the locally (variation-independent of μ) F_X components of ρ_n will not affect the influence curve of the estimator μ_n solving $0 = \sum_i IC_0(Y_i \mid Q_n, G_n, D_h(\cdot \mid \mu, \rho_n))$.

Example 2.4 (Linear regression with right-censored outcome; continuation of Example 2.3) The class of full data structure estimating functions is $\{D_h(X \mid \beta) = \min(h(Z), M)\epsilon(\beta) : h \in \mathcal{H}^F\}$, where $\mathcal{H}^F$ denotes all functions of Z. We derived in the previous example a subclass $\mathcal{H}^F(\mu, \rho, \rho_1, G) = \mathcal{H}^F(\beta, G) = \{\min(h(Z), M) : h(Z) = h(Z)I(Z \in \mathcal{Z}(\beta, G))\}$ so that for all $h \in \mathcal{H}^F(\beta, G)$ $E_G(IC_0(Y \mid G, D_h(\cdot \mid \beta)) \mid X) = D_h(X \mid \beta)$. To reparametrize the estimating functions $\{IC_0(Y \mid G, D_h(\cdot \mid \beta)) : h \in \mathcal{H}^F(\beta, G)\}$ in terms of a class of unbiased estimating functions with an index h running over an index set independent of any unknown parameters, we define the mapping $\Pi(h \mid \mathcal{H}^F(\beta, G)) = h(Z)I(Z \in \mathcal{Z}(\beta, G))$. This yields the reparametrized full data structure estimating functions

$$D_h^r(X \mid \mu = \beta, G) = \min(h(Z), M)I(Z \in \mathcal{Z}(\beta, G))\epsilon(\beta).$$

For notational convenience, we denote this latter full data structure estimating function with $D_h(X \mid \beta, G)$. □

This mapping can be viewed as a mapping from full data estimating functions D_h for μ into observed data estimating functions $IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho))$ for μ indexed by an unknown nuisance parameter G and unknown (but) protected nuisance parameter Q_0. Therefore, it can be used to construct an initial estimator in the model $\mathcal{M}(\mathcal{G})$ in which we assume that $G \in \mathcal{G}$. For a given $h \in \mathcal{H}^{F^k}$, an estimator ρ_n of ρ, G_n of G and Q_{0n} of Q_0, let μ_n^0 be the solution of the estimating equation

$$0 = \sum_{i=1}^{n} IC_0(Y_i \mid Q_n^0, G_n, D_h(\cdot \mid \mu, \rho_n)). \tag{2.16}$$

Here, the G component of ρ is estimated with the same G_n. One can solve the estimating equation with the Newton-Raphson algorithm. Let μ_n^0 be an initial guess or estimator. Set $l = 0$. The first step of the Newton–Raphson procedure involves estimation of a derivative (matrix) w.r.t. μ at μ_n^l of the estimating equation. This derivative at $\mu = \mu_1$ is defined by

$$c(\mu_1) = c(h, \mu_1, \rho, Q_0, G, P) = \frac{d}{d\mu} PIC_0(Y \mid Q_0, G, D_h(\cdot \mid \mu, \rho))\Big|_{\mu=\mu_1},$$

where we used the notation $Pf \equiv \int f(y)dP(y)$. Note that $c(\mu)$ is a $k \times k$ matrix with $c_{ij}(\mu) = \frac{d}{d\mu_j} PIC_{0,i}(Y \mid Q_0, G, D_{h_n}(\mu, \rho))$. Its estimate at $\mu = \mu_1$ is given by

$$\begin{aligned} c_n(\mu_1) &\equiv c(h_n, \mu_1, \rho_n, Q_{0n}, G_n, P_n) \\ &= \frac{1}{n}\sum_{i=1}^{n} \frac{d}{d\mu} IC_0(Y_i \mid Q_{0n}, G_n, D_{h_n}(\mu, \rho_n))\Big|_{\mu=\mu_1}. \end{aligned}$$

If $IC_0(Y \mid Q_{0n}, G_n, D_{h_n}(\cdot \mid \mu, \rho_n))$ is not differentiable in μ, but the integral of $IC_0(Y)$ w.r.t. $P_{F_X,G}$ is differentiable w.r.t. μ, then we replace the analytical derivative $d/d\mu$ by a numerical derivative: for a given function $f : \mathbb{R} \to \mathbb{R}$, the numerical derivative w.r.t. x at $x = x_1$ is defined as

$$\frac{f(x_1 + \Delta_n) - f(x_1)}{\Delta_n} \text{ for a sequence } \Delta_n = O(n^{-1/2}).$$

The $(l+1)$th step of the Newton–Raphson procedure is given by

$$\mu_n^{l+1} = \mu_n^l - c_n(\mu_n^l)^{-1} \frac{1}{n}\sum_{i=1}^{n} IC(Y_i \mid Q_{0n}, G_n, D_{h_n}(\cdot \mid \mu_n^l, \rho_n)). \tag{2.17}$$

If μ_n^0 is a decent consistent estimator of μ (if a second-order Taylor expansion in μ exists, one needs $\| \mu_n^0 - \mu \| = o_P(n^{-1/4})$ and otherwise $\| \mu_n^0 - \mu \| = O_P(n^{-1/2})$ suffices), then further iteration beyond the one-step estimator μ_n^1 will not result in first-order improvements (i.e., μ_n^0 does now only affects the second-order asymptotics of μ_n^1). Therefore, in this case one can just use μ_n^1. If no consistent initial estimator is available, then one can repeat these updating steps until convergence is established. To guarantee

convergence, the following modification of the algorithm is often needed. For a given vector norm $\| \cdot \|$, we can define

$$l(\mu) \equiv \| \sum_{i=1}^{n} IC_0(Y_i \mid G_n, D_h(\cdot \mid \mu, \rho_n)) \| .$$

For example, we could use the Euclidean norm or average of absolute value norm. Now, if $l(\mu_n^{l+1}) < l(\mu_n^l)$, then we accept the update μ_n^{l+1}, but otherwise we take as update $\epsilon\mu_n^l + (1-\epsilon)\mu_n^{l+1}$ with ϵ chosen to be the minimizer of $\epsilon \to l(\epsilon\mu_n^l + (1-\epsilon)\mu_n^{l+1})$. It actually is not necessary to determine the exact minimizer, but one needs to find an ϵ that improves the update w.r.t. the criterion l. This minimization problem can be carried out with the S-plus function nlminb().

In this book, we will not be concerned with the existence of solutions and or multiple solutions o estimating equations, but we would like to make the following suggestions. The existence of solutions has been a non issue in our experience, but we have experienced cases where estimating equations had multiple solutions. In this case, it is very helpful if either a consistent initial estimator μ_n^0 is available so that the one-step estimator suffices or that an initial ad hoc guess is available so that certain solutions can be ruled out right away. A useful idea to deal with multiple solutions comes from noting that it is unlikely that the same wrong solution will consistently come up in different estimating equations. Therefore, solving a number of estimating equations can be a sensible approach to rule out certain solutions. More formally, a promising method is to solve a number of estimating equations simultaneously. In other words, let $U(\beta)$ be a stack (i.e., vector) of estimating equations, and we estimate β by minimizing $U(\beta)^\top E(U(\beta)U(\beta)^\top)^{-1}U(\beta)$ over β. By incorporating enough estimating equations, this method will often uniquely identify the true β. By Hansen (1982), the efficiency of the resulting estimator corresponds with the estimator solving the optimal k-dimensional linear combination of the components of $U(\beta)$, provided the number of components in $U(\beta)$ does not increase too quickly with sample size.

Example 2.5 (Linear regression with right-censored outcome; continuation of Example 2.3) Let β_n be the solution of

$$0 = \sum_{i=1}^{n} \min(h(Z_i), M) I(Z_i \in \mathcal{Z}(\beta, G_n)) \epsilon_i(\beta) \frac{\Delta_i}{\bar{G}_n(T_i \mid Z_i)},$$

where G_n is an estimator of the conditional distribution $G(t \mid Z) = P(C \geq t \mid Z)$. For example, we could assume the Cox proportional hazards model for $\lambda_C(t \mid Z)$ and estimate G accordingly. We have

$$\mathcal{Z}(\beta, G_n) = \{z : \bar{G}_n(\beta_0 + \beta_1 z + \tau \mid z) > \delta > 0\}.$$

Instead of enforcing $\bar{G}_n$ to be larger than $\delta > 0$, we could also just require positivity. In this case,

$$I(Z \in \mathcal{Z}(\beta, G)) = I(\beta_0 + \beta_1 Z + \tau < \alpha^Z).$$

We will now verify that, under some smoothness conditions, the changes of the order σ in β and α^Z only have an effect of order σ^2 on $E\left\{\min(h(Z), M)I(Z \in \mathcal{Z}(\beta, G))\epsilon(\beta)\frac{\Delta}{\bar{G}(T|Z)}\right\}$. Let us denote the set $\mathcal{Z}(\beta, G)$ with $\mathcal{Z}(\beta, \alpha^Z)$. Define the set $A(\sigma)$ as all elements $z \in \mathcal{Z}(\beta + \sigma, \alpha^Z + \sigma)$ that are not an element of $\mathcal{Z}(\beta, \alpha^Z)$, where $\vec{x} + \sigma$ denotes adding the constant σ to each component of $\vec{x}$. By noting that the conditional expectation, given $Z = z \in \mathcal{Z}(\beta, G)$, of the estimating function equals zero, it follows that

$$\begin{aligned}
&E\left(h(Z)\{I(Z \in \mathcal{Z}(\beta + \sigma, \alpha^Z + \sigma)) - I(Z \in \mathcal{Z}(\beta, \alpha^Z))\}\epsilon(\beta)\frac{I(T \leq C)}{\bar{G}(T \mid Z)}\right) \\
&\leq \int_{Z \in A(\sigma)} \left\{\int_T h(Z)(T - \beta Z)I(\bar{G}(T|Z) > 0)dF(T|Z)\right\} dF_Z(Z) \\
&\equiv \int_{Z \in A(\sigma)} g(Z)dF_Z(Z).
\end{aligned}$$

Now, note that $g(Z) = 0$ for $Z \in \mathcal{Z}(\beta, \alpha^Z)$. Thus, if g is a smooth function in Z, then $g(Z) = O(\sigma)$ for $Z \in \mathcal{Z}(\beta + \sigma, \alpha^Z + \sigma)$ and, in particular, for $Z \in A(\sigma)$. This shows that the last term equals $\int_{A(\sigma)} O(\sigma)dF_Z(z) = F_Z(A(\sigma))O(\sigma) = O(\sigma^2)$.

As a consequence of this result, the asymptotics of β_n is not affected by the first-order behavior of $\mathcal{Z}(\beta_n, G_n)$. Thus, under weak conditions, β_n will be asymptotically equivalent with the estimator using $\mathcal{Z}(\beta, G)$ as given and known. This is a helpful insight for derivation of the influence curve of β_n. □

A general initial mapping only indexed by the censoring distribution.

Firstly, consider a censored data model for which

$$P(X \text{ is observed} \mid X = x) > 0 \text{ for almost all } x. \tag{2.18}$$

Given a $D(X)$, define a random variable that is 1 if $D(X)$ is observed and zero otherwise:

$$\Delta(D) = \begin{cases} 1 & \text{if } D(X) \text{ is observed.} \\ 0 & \text{if } D(X) \text{ is censored.} \end{cases}$$

Define $\Pi_{G,D}(x) = P(\Delta(D) = 1 \mid X = x)$. Now, define for $D \in L_0^2(F_X)$ the following inverse probability of censoring weighted mapping

$$IC_0(Y \mid G, D) = \frac{D(X)\Delta(D)}{\Pi_{G,D}(X)}.$$

Notice that indeed $E(IC_0(Y \mid G, D) \mid X) = D(X)$ for all $D(X)$. If (2.18) only holds on a subset of the support of X, then, as in our linear regression Example 2.3, $E_G(IC_0(Y \mid G, D) \mid X) = D(X)$ can still hold for a subset of D's covering full data structure estimating functions for a parameter of interest.

In censored data models in which X is never completely observed, such as in the current status data example below, it might not be so easy to find an initial mapping $IC_0(\cdot \mid G, D)$. In this case, the following theorem provides a general representation of an initial mapping IC_0.

Theorem 2.1 *Let $A_{F_X} : L_0^2(F_X) \rightarrow L_0^2(P_{F_X,G})$ be the nonparametric score operator for F_X:*

$$A_{F_X}(s)(Y) = E(s(X) \mid Y).$$

The adjoint $A_G^\top : L_0^2(P_{F_X,G}) \rightarrow L_0^2(F_X)$ of A_{F_X} is given by

$$A_G^\top(V)(X) = E(V(Y) \mid X).$$

Let $\mathbf{I}_{F_X,G} = A_G^\top A_{F_X} : L_0^2(F_X) \rightarrow L_0^2(F_X)$ which will be referred to as the nonparametric information operator.

Let F_1 be given. Let $(\mathcal{L}(\mathcal{X}), \|\cdot\|_\infty)$ be the space of all functions of X defined on set K with $P(X \in K) = 1$ with finite supremum norm over this set K. We have that $\mathbf{I}_{F_1,G} : (\mathcal{L}(\mathcal{X}), \|\cdot\|_\infty) \rightarrow (\mathcal{L}(\mathcal{X}), \|\cdot\|_\infty)$. Assume that $D \in \mathcal{L}(\mathcal{X})$; that is, D has finite supremum norm in X, and either (i) D lies in the range of $\mathbf{I}_{F_1,G} : (\mathcal{L}(\mathcal{X}), \|\cdot\|_\infty) \rightarrow (\mathcal{L}(\mathcal{X}), \|\cdot\|_\infty)$ or (ii) in the range of $\mathbf{I}_{F_1,G} : L_0^2(F_X) \rightarrow L_0^2(F_X)$. Then

$$IC_0(Y \mid G, D) \equiv A_{F_1}\mathbf{I}_{F_1,G}^-(D)(Y) \in L_0^2(P_{F_X,G})$$

satisfies $E(IC_0(Y \mid G, D) \mid X) = D(X)$ for all values of $X \in K$ (if (i) holds) or with probability one (if (ii) holds).

By Theorem 1.3 we also have

$$A_{F_1}\mathbf{I}_{F_1,G}^{-1}(D) = U_{F_1,G}(D) - \Pi_{F_1,G}(U_{F_1,G}(D) \mid T_{CAR}(P_{F_1,G}))$$

for any $U_{F_1,G}(D)$ satisfying $E(U_{F_1,G}(D)(Y) \mid X) = D(X)$.

Proof. Given assumption (i), for each $x \in K$ we have

$$E(A_{F_1}\mathbf{I}_{F_1,G}^-(D)(Y) \mid X = x) = \mathbf{I}_{F_1,G}\mathbf{I}_{F_1,G}^-(D)(x) = D(x).$$

Similarly, given assumption (ii), we prove this statement with probability one. □

Condition (i) is stronger than condition (ii), but the supremum norm invertibility condition (i) is needed to prove most asymptotic theorems for the estimator solving the corresponding estimating equation.

Example 2.6 (Current status data structure) Consider a carcinogenicity experiment in which the time T until onset of a tumor in a

mouse is the random variable of interest. Suppose that one collects time-independent covariates $L(0)$ and possibly time-dependent covariates (such as the weight of the mouse) $L(t)$ up to the sacrificing time C. Then, the full data structure is $X = (T, L(\cdot))$ and the observed data structure is $Y = (C, \Delta = I(T \leq C), \bar{L}(C))$.

To begin with, we will consider the current status data structure (C, Δ, L) with time-independent covariates L. In this case, $X = (T, L)$ and CAR is equivalent with assuming $G(\cdot \mid X) = G(\cdot \mid L)$. Below we will derive an explicit form of $IC_0(Y \mid G, D) = A_{F_1}\mathbf{I}^-_{F_1,G}(D)$ that will provide an $IC_0(Y \mid G, D)$ for the general data structure $(C, \Delta, \bar{L}(C))$ by simply replacing $G(\cdot \mid L)$ by the true G only satisfying CAR w.r.t. the general data structure. We actuall suggest this as a general method for finding such mappings IC_0; that is, first obtain a mapping for a marginal data structure (e.g., not involving covariates or not involving time-dependent covariates) and subsequently simply replace the censoring mechanism for the marginal data structure by the true censoring mechanism. The latter type of method will be discussed in more detail in the next subsection.

We have

$$A_{F_1}(h) = \frac{\int_0^C h(t, L)dF_1(t \mid L)}{F_1(C \mid L)}\Delta + \frac{\int_C^\infty h(t, L)dF_1(t \mid L)}{1 - F_1(C \mid L)}(1 - \Delta),$$

and its adjoint is given by

$$A^\top_G(V) = \int_0^T V(c, 0, L)dG(c \mid L) + \int_T^\infty V(c, 1, L)dG(c \mid L).$$

Thus

$$\begin{aligned}\mathbf{I}_{F_1,G}(h)(T, L) &= \int_0^T \frac{\int_0^c h(t, L)dF_1(t \mid L)}{F_1(c \mid L)}dG(c \mid L) \\ &\quad + \int_T^\infty \frac{\int_c^\infty h(t, L)dF_1(t \mid L)}{1 - F_1(c \mid L)}dG(c \mid L).\end{aligned}$$

Consider the equation $\mathbf{I}_{F_1,G}(h)(t, L) = D(t, L)$ for a D that is differentiable in the first coordinate. We assume that $dG(t \mid L) = g(t \mid L)dt$. Differentiation w.r.t. t yields

$$\frac{\int_0^t h(s, L)dF_1(s \mid L)}{F_1(t \mid L)} - \frac{\int_t^\infty h(s, L)dF_1(s \mid L)}{1 - F_1(t \mid L)} = \frac{D_1(t, L)}{g(t \mid L)},$$

where $D_1(t, L) = d/dtD(t, L)$. Now, we write $\int_0^t h(s, L)dF_1(s \mid L) = \int_0^\infty h(s, L)dF_1(s \mid L) - \int_t^\infty h(s, L)dF_1(s \mid L)$. Solving for $\int_t^\infty h(s, L)dF_1(s \mid L)$ in terms of D_1 and $\Phi_h(L) \equiv \int_0^\infty h(s, L)dF_1(s \mid L)$ now yields

$$\int_t^\infty h(s, L)dF_1(s \mid L) = \frac{D_1(t, L)}{g(t \mid L)}F_1(1 - F_1)(t \mid L) + \{1 - F_1(t \mid L)\}\Phi_h(L).$$

The last equality gives us also

$$-\int_0^t h(s,L)dF_1(s \mid L) = \frac{D_1(t,L)}{g(t \mid L)}F_1(1-F_1)(t \mid L) - F_1(t \mid L)\Phi_h(L).$$

Thus

$$A_{F_1}\mathbf{I}^-_{F_1,G}(D) = \frac{D_1(C,L)}{g(C \mid L)}\{F_1(C \mid L) - \Delta\} + \Phi_h(L).$$

Consider now the equation $\mathbf{I}_{F_1,G}(h)(\alpha_L, L) = D(\alpha_L, L)$, where α_L is the most leftmost point of the support of $g(\cdot \mid L)$. This equation reduces to

$$\Phi_h(L) = D(\alpha_L, L) - \int_0^\infty D_1(c,L)\{1 - F_1(c,L)\}dc.$$

We conclude that (here $\bar{F}_1 = 1 - F_1$):

$$\begin{aligned} IC_0(Y \mid G, D) &= \frac{D_1(C,L)}{g(C \mid L)}\{\bar{F}_1(C \mid L) - (1-\Delta)\} \\ &\quad - \int D_1(c,L)\bar{F}_1(c \mid L)dc + D(\alpha_L, L). \end{aligned} \tag{2.19}$$

Consider now the general data structure $(C, \Delta, \bar{L}(C))$. We still assume the full data estimating functions D to depend only on data $(T, W = L(0))$. For this data structure, $g(C \mid X)$ satisfies CAR if $g(C \mid X) = h(C, \bar{L}(C))$ for some measurable function h. In (2.19), replace $g(C \mid L)$ by $g(C \mid X)$, and we can replace $\bar{F}_1(C \mid L)$ by any function $\phi(C, \bar{L}(C))$. Assume that $D_1(c,L)I(T > c)/g(c \mid L) < \infty$ for all c, F_X-a.e. We will now verify that $IC_0(Y \mid G, D)$ indeed satisfies the desired property: for any $D(T, W = L(0))$, we have

$$\begin{aligned} E(IC_0(Y \mid G, D) \mid X) &= \int D_1(c,W)\phi(c, \bar{L}(c))dc + \int_{\alpha_L}^T D_1(c,W)dc \\ &\quad - \int D_1(c,W)\phi(c, \bar{L}(c))dc + D(\alpha_L, W) \\ &= D(T,W). \end{aligned}$$

By setting $\phi = 1$, we obtain the mapping

$$IC_0(Y \mid G, D) = \frac{D_1(C,L)(1-\Delta)}{g(C \mid L)} + D(\alpha_L, L).$$

One can also treat the conditional distribution $F_1(\cdot \mid L(0))$ as a nuisance parameter of the mapping IC_0 and thus define

$$\begin{aligned} IC_0(Y \mid F, G, D) &= \frac{D_1(C,L)}{g(C \mid X)}\{\bar{F}(C \mid L(0)) - (1-\Delta)\} \\ &\quad - \int D_1(c,L)\bar{F}(c \mid L(0))dc + D(\alpha_L, L). \end{aligned}$$

Since $E(IC_0(Y \mid F_1, G, D) \mid X) = D(X)$ for all F_1, the resulting estimator will remain CAN under misspecification of $F(\cdot \mid L(0))$. Therefore, the latter is an example of an initial mapping indexed by the censoring mechanism and a protected nuisance parameter Q_0. By Theorem 1.3, if there are no time-dependent covariates (i.e., $L = L(0)$), and $D = D_{opt}$ is the optimal full data estimating function choice, then $IC_0(Y \mid F, G, D)$ is the efficient influence curve. □

2.3.2 Initial mapping indexed by censoring and protected nuisance parameter

By Theorem 2.1, we have $I_{F,G} : (\mathcal{L}(\mathcal{X}), \| \cdot \|_\infty) \to (\mathcal{L}(\mathcal{X}), \| \cdot \|_\infty)$ for all F, G. Let $R^\infty(I_{F,G})$ denote its range. Consider as mapping

$$IC_0(Y \mid F_X, G, D) \equiv A_{F_X} \mathbf{I}^-_{F_X,G}(D)(Y), \tag{2.20}$$

which can also be represented as $IC_0(Y \mid G, D) - \Pi(IC_0 \mid T_{CAR})$ for any $IC_0(Y \mid G, D)$ satisfying $E(IC_0(Y \mid G, D) \mid X) = D(X)$ (Theorem 1.3 Chapter 1). Given a working model $\mathcal{M}^{F,w}$, suppose that

$$\mathcal{D}(\rho_1, G) = \{D \in \mathcal{D} : D \in R^\infty(I_{F_1,G}) \text{ for all } F_1 \in \mathcal{M}^{F,w}\}.$$

is non empty. By Theorem 2.1, for any $D \in \mathcal{D}(\rho_1, G)$ it satisfies $E(IC_0(Y \mid F, G, D) \mid X) = D(X)$ F_X-a.e. for all $F \in \mathcal{M}^{F,w}$. Thus, this mapping indeed satisfies (2.11) with $Q_0(F_X) = F_X$ for appropriately chosen $\mathcal{H}^F(\mu, \rho, \rho_1, G)$ (e.g., defined by the conditions of Theorem 2.1 at a fixed F_X, G).

Again, when applied to full data estimating functions for a particular parameter of interest, IC_0 yields a mapping from full data estimating functions $D_h(\cdot \mid \mu, \rho)$ for μ into observed data estimating functions $IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho))$ for μ indexed by unknown nuisance parameter $G \in \mathcal{G}$ and Q_0. As in the previous subsection, one needs 1) to identify a subset $\mathcal{H}^F(\mu, \rho, \rho_1, G) \subset \mathcal{H}^F$ so that for all $h \in \mathcal{H}^F(\mu, \rho, \rho_1, G)$ $E_G(IC_0(Y \mid F, G, D_h(\cdot \mid \mu, \rho)) \mid X) = D_h(X \mid \mu, \rho)$ F_X-a.e. (for all possible μ, ρ, G) and 2) to reparametrize this restricted class of full data structure estimating functions $\{D_h : h \in \mathcal{H}^F(\mu, \rho, \rho_1, G)\}$ as $\{D^r_h : h \in \mathcal{H}^F\}$ by incorporating the extra nuisance parameters ρ_1, G (needed to map any $h \in \mathcal{H}^F$ into $\mathcal{H}^F(\mu, \rho, \rho_1, G)$) in the nuisance parameter of D^r_h. Subsequently, we denote this reparametrized class of estimating functions with $\{D_h(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F\}$ again, where ρ includes the old ρ, ρ_1, and G.

This mapping is actually the optimal mapping of the next section, which can be used to construct locally efficient estimators of μ in model $\mathcal{M}$. We highlight this in this section as a special choice that one can use to construct estimators in $\mathcal{M}(\mathcal{G})$, where one might even consider extremely small parametric models $\mathcal{M}^{F,w}$ since one already assumed correct specification of G.

2.3.3 Extending a mapping for a restricted censoring model to a complete censoring model

The basic goal in this section is to find a mapping $IC_0(Y \mid Q_0, G, D)$ such that for a reasonable set of full data structure functions D (i.e., $\mathcal{D}(\rho_1, G)$ is rich enough) and all $G \in \mathcal{G}$

$$E_G(IC_0(Y \mid Q_0, G, D) \mid X) = D(X) \ F_X\text{-a.e. for all possible } Q_0. \quad (2.21)$$

For each particular full data structure model and parameter of interest, one still needs to specify the actual class $\mathcal{D}(\rho_1, G)$ of full data structure estimating functions for which (2.21) holds (see (2.11)) and specify the corresponding index sets $\mathcal{H}^F(\mu, \rho, \rho_1, G)$. Thus, (2.21) is not a formal property, but we want to separate the construction of sensible (i.e., in principle satisfying (2.21)) initial mappings from the verification of (2.21) for a particular set of full data estimating functions.

The mapping (2.20) is the optimal mapping defined in the next section, which might not always be easy to calculate. Therefore, we proceed with discussing various useful approaches for obtaining ad hoc mappings $IC_0(Y \mid Q_0, G, D)$ satisfying (2.21). Suppose that one has obtained a particular mapping satisfying (2.21) for G in a restricted censoring model $\mathcal{G}^* \subset \mathcal{G}$ of the true model $\mathcal{G}$: for a desired set of full data structure functions D

$$E_G(IC_0(Y \mid Q_0, G, D) \mid X) = D(X) \ F_X\text{-a.e. for all } Q_0,\ G \in \mathcal{G}^*.$$

For example, one might develop such a mapping under the assumption that censoring C is completely independent of X: in particular, one can set $IC_0(Y \mid Q_0, G, D)$ equal to the influence curve of an ad hoc RAL estimator under such an independent censoring model. In this case, the mapping $IC_0(Y \mid Q_0, G, D_h)$ straightforwardly extended to all $G \in \mathcal{G}$ typically satisfies (2.21) at all G. When formulating the extension, one might want to note that $IC_0(Y \mid Q_0, G, D_h)$ depends on (F_X, G) only through the law $P_{F_X,G} \in \mathcal{M}(\mathcal{G}^*)$ of the observed data when the conditional distribution of C, given X, is given by an element of $\mathcal{G}^*$. Thus, one needs to extend this mapping defined on $\mathcal{M}(\mathcal{G}^*)$ to $\mathcal{M}(\mathcal{G})$, but a straightforward ad hoc substitution typically works. This method provides a powerful way of obtaining initial mappings from full data estimating functions into observed data estimating functions since it only requires understanding a strongly simplified version (e.g., independent censoring) of the true data-generating experiment.

Example 2.7 (Right censored data structure: continuation of Example 2.2) Consider the right-censored data structure $(\tilde{T} = \min(T, C), \Delta, \bar{X}(\tilde{T}))$ and suppose that μ is a parameter of the marginal distribution F of T. Firstly, assume the independent censoring model $\mathcal{G}^* = \{G(\cdot \mid X) = G(\cdot)\}$, where C is independent of X under G^*. In this model, one can use the optimal mapping $A_F I^{-1}_{F,G}(D)$ for the *marginal*

right-censored data structure $(\tilde{T}, \Delta)$, which is given by

$$IC_0(Y \mid F, G, D) = D(T)\Delta/\bar{G}(T) + \int E(D(T) \mid T > u) dM_G(u)/\bar{G}(u),$$

where $dM_G(u) = I(\tilde{T} \in du, \Delta = 0) - I(\tilde{T} \geq u) dG(u)/\bar{G}(u)$ and $\bar{G}(u) \equiv P(C \geq u)$. Simply replacing the independent censoring $G \in \mathcal{G}^*$ by a $G \in \mathcal{G}(CAR)$ now yields an extension $IC_0(Y \mid F, G, D)$ with protected nuisance parameter F satisfying the desired property (2.21) for all $G \in \mathcal{G}(CAR)$ provided $D(T)/\bar{G}(T \mid X) < \infty$ F_X-a.e. □

Example 2.8 (Multivariate right censored data structure) Let (T_1, T_2) be a bivariate survival time of interest, and let μ be a parameter of the bivariate cumulative distribution function F of (T_1, T_2). Let $C = (C_1, C_2)$ be a bivariate censoring variable. Suppose that we observe $(\tilde{T}_j = \min(T_j, C_j), \Delta_j = I(T_j \leq C_j), \bar{L}_j(\tilde{T}_j))$, $j = 1, 2$, where $L_j(\cdot)$ are covariate processes. We have for full data $X = (T_1, \bar{L}_1(T_1), T_2, \bar{L}(T_2))$. The observed data distribution is indexed by the full data distribution F_X and the conditional bivariate distribution G of (C_1, C_2), given X. CAR is a complicated concept for this data structure, but nice rich submodels of CAR are provided in Chapter 6, where we study this data structure in detail.

Firstly, assume the independent censoring model $\mathcal{G}^* = \{G : G(\cdot \mid X) = G(\cdot)\}$. In this model, one can use the optimal mapping $A_F I_{F,G}^{-1}(D)$ for the *marginal* bivariate right-censored data structure $(\tilde{T}_j, \Delta_j)$, $j = 1, 2$. The inverse $I_{F,G}^{-1}(D) = \sum_{k=0}^{\infty} (I - I_{F,G})^k(D)$ can be represented by a Neumann series mapping, which has been implemented in (Quale, van der Laan and Robins, 2001, see also Chapter 5). Replacing the marginal G by $G \in \mathcal{G}(CAR)$ now yields a mapping $IC_0(Y \mid F, G, D) = A_F I_{F,G}^{-1}(D)$ from full data estimating functions into observed data estimating functions with protected nuisance parameter F (bivariate cumulative distribution function) satisfying the desired property (2.21) for all $G \in \mathcal{G}(CAR)$. □

2.3.4 Inverse weighting a mapping developed for a restricted censoring model

We will now provide an alternative to the previous method. Again, consider a particular mapping $IC_0(Y \mid Q_1, G, D_h)$ developed under a restricted censoring model $\mathcal{G}^* \subset \mathcal{G}$ of the true model $\mathcal{G}$. Thus, it satisfies (2.21) at $G \in \mathcal{G}^* \subset \mathcal{G}$. For each $G \in \mathcal{G}$, let $G^* = G^*(G) \in \mathcal{G}^*$ be an approximation of G defined by a mapping $\Pi : \mathcal{G} \to \mathcal{G}^*$. For example, if G is an element of a multiplicative intensity model, then Π might correspond with setting some or all of the regression coefficients equal to zero. Alternatively, Π can be an unknown mapping defined by the Kullback–Leibner projection of G onto $\mathcal{G}^*$: the latter would be estimated by maximizing the likelihood over the restricted model $\mathcal{G}^*$. In addition, suppose that for each $G \in \mathcal{G}$ it is known that the Radon–Nykodim derivative dG^*/dG exists and is uniformly

bounded:

$$G^*(\cdot \mid X) \ll G(\cdot \mid X) \ F_X\text{-a.e.}$$

In this case, the mapping

$$IC_0(Y \mid Q_0, G, D_h) \equiv IC_0(Y \mid Q_0, G^*, D_h)\frac{dG^*(Y \mid X)}{dG(Y \mid X)} \tag{2.22}$$

satisfies (2.21) at all $G \in \mathcal{G}$.

Example 2.9 (Marginal structural models, continued) Consider the previously covered example in Section 1.3 of Chapter 1. Thus, for each subject, we observe a realization (i.e., the data) of vector $\bar{A} = (A(1), \ldots, A(p))$ of exposures and treatments, a vector of outcomes $\bar{Z} = (Z(1), \ldots, Z(p))$, and the covariates $L(\cdot)$ (including many time-dependent covariates) of interest. For observed data we have

$$Y = (\bar{A}, \bar{Z}, \bar{L}),$$

which in terms of counterfactuals is represented as the missing data structure:

$$Y = (\bar{A}, \bar{X}_{\bar{A}}) = (\bar{A}, \bar{Z}_{\bar{A}}, \bar{L}_{\bar{A}}).$$

It is assumed that the missingness mechanism (i.e., the conditional distribution of $\bar{A}$, given X) satisfies the SRA,

$$g(\bar{A} \mid X) = \prod_t g(A(t) \mid \bar{A}(t-), X) = \prod_t g(A(t) \mid \bar{A}(t-), \bar{X}_{\bar{A}}(t)), \tag{2.23}$$

and we consider a marginal structural repeated measures regression model as the full data model,

$$E(Z_{\bar{a}}(t) \mid V) = g_t(\bar{a}, V \mid \beta),$$

where $g_t(\bar{a}, V \mid \beta_j)$ is some specified regression curve (e.g., linear or logistic regression) indexed by the unknown regression coefficient vector, β, and V is the set of adjustment covariates (i.e., variables not affected by $\bar{a}$ by which one wants to stratify). The goal is to estimate the causal parameter β based on the observed data Y.

Let $\mathcal{A}$ be the set of possible sample paths of $\bar{A}$, where we assume that $\mathcal{A}$ is finite. In Example 1.3, we showed that the set of full data estimating functions is given by

$$\left\{ D_h(X \mid \beta) = \sum_{\bar{a}} h(\bar{a}, V)\epsilon_{\bar{a}}(\beta) : h \right\}.$$

Let $\mathcal{G}^* = \{g : g(A(j) \mid \bar{A}(j-1), \bar{X}_A(j)) = g(A(j) \mid \bar{A}(j-1), V)\}$ assume SRA w.r.t. treatment past and V. Thus, for each $g \in \mathcal{G}^*$, we have $g(\bar{A} \mid X) = g(\bar{A} \mid V)$. We define $IC_0(Y \mid G, D_h) = \frac{h(\bar{A},V)}{g(\bar{A}|V)}\epsilon_A(\beta)$. Note that,

indeed, at any $G \in \mathcal{G}^*$ we have

$$E_G(IC_0(Y \mid G, D_h) \mid X) = \sum_{\bar{a}} h(\bar{a}, V)\epsilon_{\bar{a}}(\beta) = D_h(X).$$

Given $g \in \mathcal{G}(SRA)$, let $g^* = g^*(g) \in \mathcal{G}^*$ be its projection (in some sense) of g onto $\mathcal{G}^*$ satisfying $\max_{\bar{a}\in\mathcal{A}}\{h(\bar{a}, V)g^*(\bar{a} \mid V)\}/g(\bar{a} \mid X) < \infty$ F_X-a.e. Then

$$IC_0(Y \mid G, D_h) \equiv IC_0(Y \mid G^*, D_h)\frac{g^*(\bar{A} \mid X)}{g(\bar{A} \mid X)} = \frac{h(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)}$$

is the IPTW mapping presented in Example 1.3, which satisfies the desired property $E_G(IC_0(Y \mid G, D_h) \mid X) = D_h(X)$ F_X-a.e. and at each $G \in \mathcal{G}(SRA)$. Note that this condition on h defines the set of allowed indexes $\mathcal{H}(\mu, \rho, \rho_1, G)$ and the allowed set of full data functions $\mathcal{D}(\rho_1, G)$. Notice that we would have obtained the same IPTW mapping by simply extending $IC_0(Y \mid G, D_h) = \frac{h(\bar{A},V)}{g(\bar{A}|V)}\epsilon_A(\beta)$ to $\mathcal{G}(SRA)$. Thus, the methods of the previous subsection and this subsection yield identical results for this example.

In Chapter 6, we also apply this method to obtain the class of all estimating functions in marginal structural nested models, another class of causal inference models that allows one to estimate dynamic treatment-regime-specific outcome distributions. Other important applications of this method in causal inference are covered in Murphy, van der Laan and Robins (2001) and van der Laan, Murphy and Robins (2002). □

2.3.5 Beating a given RAL estimator

We will now show, given an RAL estimator of μ, how one can obtain a mapping $D_h \rightarrow IC_0(Y \mid F, G, D_h)$ from full data estimating functions into observed data estimating functions so that for a specified full data estimating function D_h it provides an estimating function that when evaluated at the true parameter values equals the influence curve of the given RAL estimator (and thus results in an estimator that is asymptotically equivalent with the given RAL estimator). Let μ_n be a given RAL estimator, and let $IC(Y \mid F_X, G)$ be its influence curve. Since $IC(Y \mid F_X, G)$ is a gradient of the pathwise derivative of μ, we have that $E_G(IC(Y \mid F, G) \mid X) \in T_{nuis}^{F,\perp,*}(F)$ for all $F \in \mathcal{M}^F$. Thus, by taking a conditional expectation, given X, $IC(Y \mid F, G)$ maps into a particular full data estimating function for μ. Let $h^* \equiv h_{ind,F_X}(E(IC(Y \mid F_X, G) \mid X))$ be the corresponding index of this estimating function:

$$D_{h^*}(X \mid \mu(F_X), \rho(F_X)) = E(IC(Y \mid F_X, G) \mid X).$$

For a multivariate full data function $D = (D_1, \ldots, D_k)$, one defines $h_{ind,F_X}(D) = (h_{ind,F_X}(D_1), \ldots, h_{ind,F_X}(D_k))$. Note that $h^* = h^*(F_X, G)$. Let $D_h \rightarrow IC_0(Y \mid G, D_h)$ be an initial mapping from full data estimating

functions into observed data estimating functions satisfying $E_G(IC_0(Y \mid G, D_{h^*(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X, G))) \mid X) = D_{h^*(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X, G))$ for all $F_X \in \mathcal{M}^F$ and $G \in \mathcal{G}$. We now define $IC_{CAR}(Y \mid F_X, G)$ as

$$IC(Y \mid F_X, G) - IC_0(Y \mid G, D_{h^*(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X))).$$

Note that $E(IC_{CAR}(Y \mid F, G) \mid X) = 0$ for all $F \in \mathcal{M}^F$.

We now define as mapping from full data estimating functions into observed data estimating functions

$$IC(Y \mid F_X, G, D_h) = IC_0(Y \mid G, D_h) + IC_{CAR}(F_X, G).$$

Note that it satisfies (2.21) and, in addition, $IC(Y \mid F_X, G, D_{h^*}) = IC(Y \mid F_X, G)$. Consequently, under the regularity conditions of our asymptotic Theorem 2.4, the estimating equation with index h^* (or a consistent estimator thereof) yields an estimator that is asymptotically equivalent with μ_n. Other choices of h might result in more efficient estimators than μ_n.

In the following example, we combine this method described above with the extension method of Subsection 2.3.3 into a powerful application for the bivariate right-censored data structure.

Example 2.10 (Multivariate right-censored data structure; continuation of Example 2.8) We refer to Example 2.8. Thus, we observe $(\tilde{T}_j = \min(T_j, C_j), \Delta_j = I(T_j \leq C_j), \bar{L}_j(\tilde{T}_j))$, $j = 1, 2$, where $L_j(\cdot)$ are covariate processes. The parameter of interest is $\mu = S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. Let F be the bivariate cumulative distribution of (T_1, T_2). For full data, we have $X = (T_1, \bar{L}_1(T_1), T_2, \bar{L}(T_2))$. The observed data distribution is indexed by the full data distribution F_X and the conditional bivariate distribution G of (C_1, C_2), given X, which is assumed to be modeled with some submodel $\mathcal{G}$ of CAR, as provided in Chapter 6. Let the full data model be nonparametric so that the only full data estimating function is $D(X \mid \mu) = I_{(t_1,\infty)\times(t_2,\infty)}(T_1, T_2) - \mu$. Firstly, assume the independent censoring model $\mathcal{G}^* = \{G(\cdot \mid X) = G(\cdot)\}$.

A well-known estimator of $\mu = S(t_1, t_2)$ based on marginal bivariate right-censored data in the independent censoring model is the Dabrowska estimator (Dabrowska, 1988,1989). We want to apply the method above to find an estimating function for μ that yields an estimator that is asymptotically equivalent with Dabrowska's estimator. Subsequently, extending this mapping to $G \in \mathcal{G}$ yields an estimating function for our extended bivariate data structure only assuming our posed model $\mathcal{G}$ for G. The influence curve $IC_{Dab}(Y \mid F, G, (t_1, t_2))$ of Dabrowska's estimator is derived in Gill, van der Laan and Wellner (1995) and van der Laan (1990) and is given by

$$\begin{aligned} IC(Y) \quad = \quad & \bar{F}(t_1, t_2) \left\{ - \int_0^{t_1} \frac{I(\tilde{T}_1 \in du, \Delta_1 = 1) - I(\tilde{T}_1 \geq u)\Lambda_1(du)}{P_{F,G}(\tilde{T}_1 \geq u)} \right. \\ & \left. - \int_0^{t_2} \frac{I(\tilde{T}_2 \in du, \Delta_2 = 1) - I(\tilde{T}_2 \geq u)\Lambda_2(du)}{P_{F,G}(\tilde{T}_2 \geq u)} \right. \end{aligned}$$

$$
\begin{aligned}
&+\int_0^{t_1}\int_0^{t_2} \frac{I(\tilde{T}_1 \in du, \tilde{T}_2 \in dv, \Delta_1 = 1, \Delta_2 = 1)}{P_{F,G}(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)} \\
&-\int_0^{t_1}\int_0^{t_2} \frac{I(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)\Lambda_{11}(du, dv)}{P_{F,G}(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)} \\
&-\int_0^{t_1}\int_0^{t_2} \frac{I(\tilde{T}_1 \in du, \tilde{T}_2 \geq v, \Delta_1 = 1)\Lambda_{01}(dv, u)}{P_{F,G}(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)} \\
&+\int_0^{t_1}\int_0^{t_2} \frac{I(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)\Lambda_{10}(du, v)\Lambda_{01}(dv, u)}{P_{F,G}(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)} \\
&-\int_0^{t_1}\int_0^{t_2} \frac{I(\tilde{T}_1 \geq u, \tilde{T}_2 \in dv, \Delta_2 = 1)\Lambda_{10}(du, v)}{P_{F,G}(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)} \\
&+\int_0^{t_1}\int_0^{t_2} \frac{I(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)\Lambda_{10}(du, v)\Lambda_{01}(dv, u)}{P_{F,G}(\tilde{T}_1 \geq u, \tilde{T}_2 \geq v)} \Bigg\},
\end{aligned}
$$

where $\Lambda_j(du) = P(T_j \in du \mid T_j \geq u)$, $j = 1, 2$, $\Lambda_{10}(du \mid v) = P(T_1 \in du \mid T_1 \geq u, T_2 \geq v)$, $\Lambda_{01}(dv, u) = P(T_2 \in dv \mid T_1 \geq u, T_2 \geq v)$, and $\Lambda_{11}(du, dv) = P(T_1 \in du, T_2 \in dv \mid T_1 \geq u, T_2 \geq v)$. Here $P_{F,G}(\tilde{T}_1 > s, \tilde{T}_2 > t) = S(s,t)\bar{G}(s,t)$. Firstly, we note that it is straightforward to verify that, if $\bar{G}(t_1, t_2) > 0$, then $E_G(IC_{Dab}(Y \mid F, G, (t_1, t_2)) \mid X) = D(X \mid \mu)$ for all $G \in \mathcal{G}^*$ satisfying independent censoring. In addition, if we replace G by any $G \in \mathcal{G}(CAR)$ satisfying CAR, then we still have

$$E_G(IC_{Dab}(Y \mid F, G, (t_1, t_2)) \mid X) = I_{(t_1,\infty)\times(t_2,\infty)}(T_1, T_2) - \mu$$

for all bivariate distributions F, as predicted in Subsection 2.3.3.

Let $IC_0(Y \mid G, D) = D(X)\Delta_1\Delta_2/\bar{G}(T_1, T_2 \mid X)$. We now define

$$IC_{CAR}(Y \mid F, G) \equiv IC_{Dab}(Y \mid F, G) - IC_0(Y \mid G, D(\cdot \mid \mu(F))).$$

Note that $E_G(IC_{CAR}(Y \mid F, G) \mid X) = 0$ for all bivariate distributions F and all $G \in \mathcal{G}(CAR)$.

We now define an observed data estimating function for μ indexed by the true censoring mechanism G and a bivariate distribution F:

$$IC(Y \mid F, G, D(\cdot \mid \mu)) = IC_0(Y \mid G, D(\cdot \mid \mu)) + IC_{CAR}(F, G). \tag{2.24}$$

It follows that this estimating function for μ satisfies $E_G(IC(Y \mid F, G, D(\cdot \mid \mu) \mid X) = D(X \mid \mu)$ for all $G \in \mathcal{G}(CAR)$ and, at the true μ and F, it reduces to Dabrowska's influence curve. Given consistent estimators F_n of F and G_n of G according to model $\mathcal{G}$, let μ_n be the solution of

$$0 = \frac{1}{n}\sum_{i=1}^{n} IC(Y_i \mid F_n, G_n, D(\cdot \mid \mu)).$$

Under the regularity conditions of Theorem 2.4, μ_n is asymptotically linear with influence curve $IC(Y) \equiv \Pi(IC(\cdot \mid F, G, D(\cdot \mid \mu)) \mid T_2(P_{F_X,G}))$, where $T_2(P_{F_X,G}) \subset T_{CAR}$ is the observed data tangent space of G under

the posed model $\mathcal{G}$. Firstly, assume that $\mathcal{G}^*$ is the independent censoring model. Since $IC(Y \mid F, G, D(\cdot \mid \mu))$ is already orthogonal to the tangent space $T(\mathcal{G}^*)$ of G for the independent censoring model $\mathcal{G}^*$, we have that $IC(Y) = IC_{Dab}(Y)$. Secondly, if the tangent space $T_2(P_{F_X,G})$ contains scores that are not in $T(\mathcal{G}^*)$, then it will result in an estimator more efficient than Dabrowska's estimator, even when (C_1, C_2) is independent of X. In Chapter 5, we provide a simulation study comparing this estimator with Dabrowska's estimator and further improve on this estimating function by orthogonalizing w.r.t. a tangent space of a rich submodel $\mathcal{G}$ of $\mathcal{G}(CAR)$ for G.

Note that the method used in this example can be used, in general, to generalize an estimator for a marginal data structure into an estimator for an extended data structure. □

2.3.6 Orthogonalizing an initial mapping w.r.t. G: Double robustness

Consider the following class of parametric submodels through the censoring mechanism $G_{Y|X}$:

$$\{(1 + \epsilon V(y))dG(y|x) : V \in L_0^2(P_{F_X,G}), E(V(Y) \mid X) = 0\}.$$

It is straightforward to show that the tangent space of G in the model $\mathcal{M}(\mathcal{G}_{CAR})$ generated by this class of parametric submodels is given by

$$T_{CAR}(P_{F_X,G}) = N(A_G^\top) = \{v \in L_0^2(P_{F_X,G}) : E(v(Y) \mid X) = 0\}.$$

An initial mapping can be orthogonalized w.r.t. $T_{CAR}(P_{F_X,G})$ itself, resulting in the optimal mapping in the next section, or w.r.t. subspaces of $T_{CAR}(P_{F_X,G})$ as in this subsection.

We will now present a general way of obtaining a mapping of estimating functions $IC_0(Y \mid Q_0, G, D)$ indexed by nuisance parameters Q_0, G with a double robustness property. Let $H(P_{F_X,G})$ be a tangent space of G according to some submodel of $\mathcal{G}(CAR)$. Thus $H(P_{F_X,G}) \subset T_{CAR}(P_{F_X,G})$ is a subspace of $T_{CAR}(P_{F_X,G})$ for all $P_{F_X,G} \in \mathcal{M}$. For example, one can set $H(P_{F_X,G})$ equal to the observed data tangent space $T_2(P_{F_X,G})$ of G in model $\mathcal{M}(\mathcal{G})$.

Since $T_{CAR}(P_{F_X,G}) = \{V(Y) \in L_0^2(P_{F_X,G}) : E_G(V(Y) \mid X) = 0\}$ only depends on F_X to make sure that the elements have finite variance w.r.t. $P_{F_X,G}$, it is always possible to find a rich common subset of $T_{CAR}(P_{F_X,G})$ only depending on G. By the same argument, one will also always be able to choose a rich common subset $H(G)$ of $H(P_{F_X,G})$. Let $\mathcal{Q}_0$ be an index set (independent of (F_X, G)) for $H(G)$ so that

$$H(G) \equiv \{IC_{nu}(\cdot \mid Q_0, G) : Q_0 \in \mathcal{Q}_0\} \subset H(P_{F_X,G}) \text{ for all } G \in \mathcal{G},$$

where $IC_{nu}(\cdot \mid Q_0, G)$ denotes the Q_0 indexed element of $H(G)$. It will be possible to define $(Q_0, G) \to (Y \to IC_{nu}(Y \mid Q_0, G))$ as a mapping from

$\mathcal{Q}_0 \times \mathcal{G}$ into pointwise well-defined functions of Y, which we will need in order to define the estimating function $IC_0(Y \mid Q_0, G, D)$ below.

We can now make a mapping $IC_0(Y \mid G, D)$ satisfying (2.21) orthogonal (at the truth) to $H(P_{F_X,G})$ by introducing another nuisance parameter $Q_0 = Q_0(F_X, G)$ as follows:

$$IC_0(Y \mid Q_0, G, D) = IC_0(Y \mid G, D) - IC_{nu}(Y \mid Q_0, G),$$

where the unknown parameter $Q_0(F_X, G)$ is defined by

$$IC_{nu}(\cdot \mid Q_0(F_X, G), G) = \Pi_{F_X,G}\left(IC_0(\cdot \mid G, D) \mid H(P_{F_X,G})\right)$$

and the equality holds in $L_0^2(P_{F_X,G})$.

This mapping $D \rightarrow IC_0(Y \mid Q_0, G, D)$ maps full data estimating functions $D_h(\cdot \mid \mu, \rho)$ for μ into observed data estimating functions $IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho))$ for μ with unknown nuisance parameters $Q_0 = Q_0(F_X, G)$ and G. It has the property $E_G(IC_0(Y \mid Q_0, G, D) \mid X) = D(X)$ for all possible Q_0, and thus it remains unbiased when Q_0 is misspecified. Therefore, it can be used to construct an initial estimator in the model $\mathcal{M}(\mathcal{G})$ in which we assume that $G \in \mathcal{G}$: for a given $h \in \mathcal{H}$, an estimator ρ_n of ρ, Q_{0n} of Q_0, G_n of G, let μ_n^0 be the solution of the estimating equation

$$0 = \sum_{i=1}^{n} IC_0(Y_i \mid Q_{0n}, G_n, D_h(\cdot \mid \mu, \rho_n)). \tag{2.25}$$

If Q_{0n} converges to some Q_0 not necessarily equal to $Q_0(F_X, G)$, then application of Theorem 2.4 yields, under regularity conditions, that the estimator is asymptotically linear with influence curve $-\Pi(c^{-1}IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho)) \mid T_2(P_{F_X,G})^{\perp})$, where c is the derivative matrix w.r.t. μ of the expectation of IC_0. In particular, if $Q_0 = Q_0(F_X, G)$ and $H(P_{F_X,G}) = T_2(P_{F_X,G})$, then this influence curve equals $-c^{-1}IC_0(\cdot \mid Q_0(F_X, G), G, D, D_h(\cdot \mid \mu, \rho))$.

Finally, we note that the mapping $IC_0(Y \mid F, G, D) = A_F I^-_{F,G}(D)$ considered in the previous subsection is also of this type since

$$A_{F_X} I^-_{F_X,G}(D) = IC_0(Y \mid G, D) - \Pi_{F_X,G}(IC_0(Y \mid G, D) \mid T_{CAR})$$

for any initial $IC_0(Y \mid G, D)$ satisfying $E(IC_0(Y \mid G, D) \mid X) = D(X)$. In other words, if we set $H(P_{F_X,G}) = T_{CAR}(P_{F_X,G})$, then $IC_0(Y \mid Q_1, G, D)$ reduces to $IC_0(Y \mid F, G, D) = A_F I^-_{F,G}(D)$.

Example 2.11 (Right-censored data structure; continuation of Example 2.2) Consider the right-censored data structure $Y = (\tilde{T} = \min(T, C), \Delta = I(\tilde{T} = T), \bar{L}(\tilde{T}))$. Suppose that $H(P_{F_X,G})$ is the tangent space of G in the independent censoring model. We have that $H(P_{F_X,G}) = \overline{\{\int H(u) dM_G(u) : H\}} \cap L_0^2(P_{F_X,G})$, where $dM_G(u) = I(\tilde{T} \in du, \Delta = 0) - I(\tilde{T} \geq u)\Lambda_C(du)$. If $\int \mid dM_G(u) \mid < \infty$, then one can choose $H(G) = \{\int H(u) dM_G(u) :\parallel H \parallel_\infty < \infty\}$. Let $IC_0(Y \mid G, D) = \frac{D(X)\Delta(D)}{G(V(D)|X)}$,

where $\Delta(D)$ is the indicator that $D(X)$ is observed, $V(D)$ is the minimum time at which D is fully observed, and $\bar{G}(V(D) \mid X)$ is the probability that $\Delta(D) = 1$, given X.

Application of Lemma 3.2, formula (3.17) from the next chapter, yields

$$\Pi_{F_X,G}(IC_0 \mid H(P_{F_X,G})) = -\int \frac{E_{F_X,G}\left(\frac{D(X)\Delta(D)}{\bar{G}(V(D)|X)} I(T \geq u)\right)}{P(\tilde{T} \geq u)} dM_G(u)$$

so that, for any function $Q_0(u)$, we can define

$$IC_0(Y \mid Q_0, G, D) = \frac{D(X)\Delta(D)}{\bar{G}(V(D) \mid X)} + \int Q_0(u) dM_G(u),$$

where $Q_0(F_X, G)(u) = E_{F_X,G}(D(X)\Delta(D)I(T \geq u)/\bar{G}(V(D) \mid X))/P(\tilde{T} \geq u) = E_{F_X}(D(X) \mid T > u)/\bar{G}(u)$.

Let us now consider the special case in which $\mu = F(t) = P(T \leq t)$, $D(X) = I(T \leq t) - F(t)$, and we assume the independent censoring model $G(\cdot \mid X) = G(\cdot)$ for G. Then $IC_0(Y \mid Q_0(F_X, G), G, D)$ equals the influence curve IC_{KM} of the Kaplan–Meier estimator. The corresponding estimating equation results in an estimator μ_n that is asymptotically equivalent with the Kaplan–Meier estimator. If one assumes the Cox proportional hazards model for G with covariates extracted from the observed past, then μ_n is asymptotically linear with influence curve $IC_{KM} - \Pi(IC_{KM} \mid T_2(P_{F_X,G}))$, where $T_2(P_{F_X,G})$ denotes the tangent space of the Cox proportional hazards model. Thus, in the last case, μ_n will be more efficient than the Kaplan–Meier estimator. □

Double protection (robustness) when orthogonalizing w.r.t. convex censoring models

If $H(P_{F_X,G})$ is the tangent space of G for a *convex* model $\mathcal{G}(conv) \subset \mathcal{G}(CAR)$ *containing* $\mathcal{G}$, then the mapping $D \to IC_0(\cdot \mid Q_0, G, D)$ from full data estimating functions to observed data estimating functions satisfies a double protection property against misspecification of G and F_X defined by (2.27) below. This follows from Theorem 1.6 and Lemma 1.9. As a consequence, in this case it actually yields estimating functions in model $\mathcal{M}$. A special case is $H(P_{F_X,G}) = T_{CAR}(P_{F_X,G})$, where $T_{CAR}(P_{F_X,G})$ is the tangent space of G for the model $\mathcal{G}(CAR)$, which makes IC_0 corresponded with the optimal mapping $A_F I^-_{F,G}(D)$ as introduced above. The latter mapping will actually be the mapping proposed in the next section and applied in all subsequent chapters, which will allow locally efficient estimation. For some data structures, the projection on $T_{CAR}(P_{F_X,G})$ does not exist in closed form. In that case, the estimating function $IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho))$ with $H(P_{F_X,G}) \subset T_{CAR}(P_{F_X,G})$ chosen so that the projection operator on $H(P_{F_X,G})$ exists in closed form provides an interesting alternative. Such a mapping is used in Chapter 5 to provide estimators for the extended bivariate right-censored data structure and in Chapter 6 to identify causal and

non causal parameters in complex longitudinal data structures involving censoring and time-dependent informative treatment assignments.

Let us prove the double protection property (2.27) to make this section self-contained. By definition, $IC_0(\cdot \mid Q_0(F_X, G), G, D)$ is actually orthogonal to the tangent space $H(P_{F_X,G})$. By the convexity of $\mathcal{G}(conv)$, we know that for all $G_1 \in \mathcal{G}(conv)$ with $dG_1/dG < \infty$, $\alpha G_1 + (1-\alpha)G$ is a submodel of $\mathcal{G}(conv)$. Consequently, the line $dP_{F_X,\alpha G_1+(1-\alpha)G}$ is a submodel of $\mathcal{M}(\mathcal{G}(conv))$ that has score (by linearity of $G \to P_{F_X,G}$) $(dP_{F_X,G_1} - dP_{F_X,G})/dP_{F_X,G}$. Thus, the latter score is an element of $H(P_{F_X,G})$. Thus, the orthogonality of $IC_0(Y \mid Q_0(F_X, G), G, D)$ to $(dP_{F_X,G_1} - dP_{F_X,G})/dP_{F_X,G}$ now yields

$$\begin{aligned} 0 &= E_{P_{F_X,G}} IC_0(Y \mid Q_0(F_X, G), G, D) \frac{dP_{F_X,G_1} - dP_{F_X,G}}{dP_{F_X,G}}(Y) \\ &= E_{P_{F_X,G_1} - P_{F_X,G}} IC_0(Y \mid Q_0(F_X, G), G, D) \\ &= E_{P_{F_X,G_1}} IC_0(Y \mid Q_0(F_X, G), G, D) \text{ if } D \in \mathcal{D}(\rho_1(F_X), G) \end{aligned}$$

and $E_{F_X} D(X) = 0$. Here we used that $D \in \mathcal{D}(\rho_1(F_X), G)$ guarantees that $E_{P_{F_X,G}} IC_0(Y \mid Q(F_X, G), G, D) = E_{F_X} D(X) = 0$. Exchanging the role of G and G_1 proves the following result: for all pairs $G, G_1 \in \mathcal{G}(conv)$ with $dG/dG_1 < \infty$, we have for all $D \in \mathcal{D}(\rho_1(F_X), G_1)$ with $E_{F_X} D(X) = 0$

$$0 = E_{P_{F_X,G}} IC_0(Y \mid Q_0(F_X, G_1), G_1, D).$$

We note that this provides a sufficient, but not necessary condition. For example, if the identity holds at $G_1 = G_{1m}$ for a sequence at G_{1m}, $m = 1, \ldots$, which approximates a G^* in the sense that $E_{P_{F_X,G}} IC_0(Y \mid Q_0(F_X, G_{1m}), G_{1m}, D) \to E_{P_{F_X,G}} IC_0(Y \mid Q_0(F_X, G^*), G^*, D)$, then it follows that the identity also holds at $G_1 = G^*$. Therefore, it is not surprising that in many applications the identity also holds for pairs G_1, G not satisfying $dG/dG_1 < \infty$. This identity gives us protection against misspecification of G when the Q_0 component of IC is correctly estimated in the sense that if G_n converges to some $G_1 \in \mathcal{G}(conv)$ with $dG/dG_1 < \infty$, then μ_n^0 (2.25) will still be consistent.

Given $D \in \mathcal{D}(\rho_1(F_X), G)$, the conditional expectation of $IC_0(Y \mid Q_0, G, D)$, given X, equals $D(X)$, which proves that for any $Q_0 \in \mathcal{Q}_0$ $E_{F_X,G} IC(Y \mid Q_0, G, D) = 0$. This gives us protection against misspecification of $Q_0(F_X, G)$ when G is correctly estimated. To summarize, our definition of the mapping $IC_0(\cdot \mid Q_0, G, D)$ of full data estimating functions to observed data estimating function depends on the unknown $Q_0(F_X, G)$ and G, but it is protected by misspecification of either F_X or G in the following sense.

Theorem 2.2 *We have*

$$\begin{aligned} E_{F_X,G} IC_0(Y \mid Q_0(F_X, G_1), G_1, D) &= E_{F_X} D(X) \text{ for } D \in \mathcal{D}(\rho_1(F_X), G_1) \\ &\text{and all } G_1 \in \mathcal{G}(conv) \text{ with } G \ll G_1. \end{aligned} \tag{2.26}$$

$$E_{F_X,G} IC_0(Y \mid Q_0, G, D) = E_{F_X} D(X) \quad \text{for all } Q_0 \in \mathcal{Q}_0 \text{ and } D \in \mathcal{D}(\rho_1(F_X), G). \tag{2.27}$$

Note that the protection against misspecification of G can be exploited by estimating Q_0 and the nuisance parameter $\rho(F_X, G)$ in $D_h(\cdot \mid \mu, \rho)$ with substitution estimators $Q_0(F_n, G_n)$ and $\rho(F_n, G_n)$.

2.3.7 Ignoring information on the censoring mechanism improves efficiency

Let $T_2(P_{F_X,G})$ be the tangent space of G in model $\mathcal{M}(\mathcal{G})$. Application of Theorem 2.4 below shows that, under regularity conditions, μ_n^0 (2.25) is asymptotically linear with influence curve

$$\Pi(c^{-1} IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho)) \mid T_2^{\perp}(P_{F_X,G})), \tag{2.28}$$

where $c = d/d\mu E IC_0(Y \mid Q_0, G, D_h(\cdot \mid \mu, \rho))$ and Q_0 is the limit of Q_{0n}. In particular, this teaches us that μ_n^0 will become more efficient if one estimates G more nonparametrically. Thus, if G is known and one sets $G_n = G$ in the estimating equation (2.25), then μ_n^0 is asymptotically linear with influence curve $c^{-1} IC_0(\cdot \mid G, c(\mu), D_h(\cdot \mid \mu, \rho))$, which can have much larger variance than the influence curve (2.28) for a reasonable size model $\mathcal{G}$.

To understand this feature of the estimator, we prove the following general result. The proof of this theorem actually shows, in general, that optimal estimation of an orthogonal nuisance parameter leads to an asymptotic improvement of the estimator. Application of this theorem with $\mu_n(G_n) = \mu_n^0$ and $\mu_n(G)$ being the solution of the estimating equation (2.25) with $G_n = G$ known explains the result (2.28).

Theorem 2.3 *Let $\mathcal{M}(G) = \{P_{F_X,G} : F_X \in \mathcal{M}^F\}$ be the model $\mathcal{M}$ with G known. Let $\mu_n(G)$ be a regular asymptotically linear estimator of μ in the model $\mathcal{M}(G)$ with G known with influence curve $IC_0(Y \mid F_X, G)$. Assume now that for an estimator G_n*

$$\mu_n(G_n) - \mu = \mu_n(G) - \mu + \Phi(G_n) - \Phi(G) + o_P(1/\sqrt{n})$$

for some functional Φ of G_n. Assume that $\Phi(G_n)$ is an asymptotically efficient estimator of $\Phi(G)$ for the model $\mathcal{M}(\mathcal{G})$ with tangent space generated by G given by $T_2(P_{F_X,G})$. Then $\mu_n(G_n)$ is regular asymptotically linear with influence curve

$$IC_1(F_X, G) = \Pi(IC_0(F_X, G) \mid T_2(P_{F_X,G})^{\perp}).$$

Proof. We decompose $L_0^2(P_{F_X,G})$ orthogonally in $T_1(P_{F_X,G}) + T_2(P_{F_X,G}) + T^{\perp}(P_{F_X,G})$, where $T^{\perp}(P_{F_X,G})$ is the orthogonal complement of $T_1 + T_2$, and $T_1 = T_1(P_{F_X,G})$ and $T_2(P_{F_X,G})$ are the tangent spaces corresponding to F_X and G, respectively. The assumptions in the lemma imply that $\mu_n(G_n)$ is

asymptotically linear with influence curve $IC = IC_0 + IC_{nu}$, where IC_{nu} is an influence curve corresponding with an estimator of the nuisance parameter $\Phi(G)$ estimated under the model with nuisance tangent space T_2. Let $IC_0 = a_0 + b_0 + c_0$ and $IC_{nu} = a_{nu} + b_{nu} + c_{nu}$ according to the orthogonal decomposition of $L_0^2(P_{F_X,G})$ above. From now on, the proof uses the following two general facts about influence curves of regular asymptotically linear estimators (see Bickel, Klaassen, Ritov and Wellner, 1993): an influence curve is orthogonal to the nuisance tangent space, and the efficient influence curve lies in the tangent space. Since IC_{nu} is an influence curve of $\Phi(G)$ in the model where nothing is assumed on F_X it is orthogonal to T_1; that is, $a_{nu} = 0$. Since $\Phi(G_n)$ is efficient, IC_{nu} lies in the tangent space T_2 and hence $c_{nu} = 0$ as well. We also have that $IC_0 + IC_{nu}$ is an influence curve for an estimator of μ and hence is orthogonal to T_2, so $b_0 + b_{nu} = 0$. Consequently, we have that

$$IC_1 = IC_0 + IC_{nu} = a_0 + c_0 = \Pi(IC_0 \mid T_2^{\perp}).$$

This completes the proof. □

Example 2.12 (Marginal right-censored data) Suppose that we observe n i.i.d. observations of $Y = (\tilde{T} = T \wedge C, \Delta = I(\tilde{T} = T) = I(C \geq T))$. Let F be the cumulative distribution function of the full data T, and let G be the conditional distribution of C, given T, satisfying CAR. In this case, CAR is equivalent with assuming that the censoring hazard $\lambda_{C|T}(t \mid T)$ only depends on $\bar{X}(t)$, where $X(t) = I(T \leq t)$.

Let us first consider the observed data model with G known. In that model, we could estimate $\mu = F(t)$ with the inverse probability of censoring weighted estimator

$$\mu_n(G) = \frac{1}{n} \sum_{i=1}^{n} I(T_i \leq t) \frac{\Delta_i}{\bar{G}(T_i)},$$

where $\bar{G}(t) = P(C \geq t)$. We have that $\mu_n(G)$ is regular and asymptotically linear with influence curve $IC_0(Y \mid G, \mu) = I(T \leq t)\Delta/\bar{G}(T) - \mu$. Consider now the model where we only assume CAR on G. Let G_n be the Kaplan–Meier estimator of G based on the n censored observations $(\tilde{T}, 1 - \Delta)$. It is well-known that G_n is an efficient estimator of G in the model $\mathcal{M}(CAR)$. Application of Lemma 2.3 yields that $\mu_n(G_n)$ is a regular and asymptotically linear estimator with influence curve $IC_0(Y \mid G, \mu)$ minus its projection on the tangent space $T_{CAR}(P_{F_X,G})$ for G when only assuming CAR.

Let $A_F : L_0^2(F) \to L_0^2(P_{F,G})$ $A_F(h)(Y) = E_F(h(T) \mid Y)$ be the nonparametric score operator, and let $A_G(V)(X) = E_G(V(Y) \mid X)$ be its adjoint. Since the full data model is nonparametric, we actually have that the closure of the range of A_F is the tangent space of F in the model $\mathcal{M}(CAR)$. We have that $T_{CAR}^{\perp} = N(A_G^{\top})^{\perp} = \overline{R(A_F)}$. This proves that the influence curve of $\mu_n(G_n)$ is an element of the tangent space $\overline{R(A_F)}$ and thus

must equal the efficient influence curve; Here, we use the fact that the efficient influence curve (which equals the canonical gradient of the pathwise derivative) is the only influence curve which is an element of the tangent space. This proves that $\mu_n(G_n)$ is an efficient estimator of μ, while $\mu_n(G)$ is far from efficient. In this particular example, we have the remarkable coincidence that $\mu_n(G_n)$ equals the Kaplan–Meier estimator algebraically, assuming that we define the Kaplan-Meier estimator to be zero after the last uncensored observation. □

2.4 Optimal Mapping into Observed Data Estimating Functions

Let $D_h \rightarrow IC_0(Y \mid Q_0, G, D_h)$ be an initial mapping from full data estimating functions into observed data estimating functions satisfying (2.21). Let $IC_{CAR}(\cdot \mid Q, G)$ with Q ranging over a parameter space $\mathcal{Q}$ be pointwise well defined functions of Y satisfying

$$\{IC_{CAR}(\cdot \mid Q, G) : Q \in \mathcal{Q}\} \subset T_{CAR}(P_{F_X,G}) \text{ for all } P_{F_X,G} \in \mathcal{M}. \quad (2.29)$$

Let $IC_{CAR}(\cdot \mid Q(F_X, G), G, D)$ be a pointwise well-defined function of Y that equals the projection

$$\Pi_{F_X,G}(IC_0(\cdot \mid Q_0, G, D) \mid T_{CAR}(P_{F_X,G}))$$

of $IC_0(\cdot \mid Q_0, G, D)$ onto $T_{CAR}(P_{F_X,G})$ in the Hilbert space $L^2(P_{F_X,G})$.

Then, for any $D \in \mathcal{D}, Q_0 \in \mathcal{Q}_0, Q \in \mathcal{Q}, F_X \in \mathcal{M}^F, G \in \mathcal{G}$,

$$IC(Y \mid Q_0, Q, G, D) \equiv IC_0(Y \mid Q_0, G, D) - IC_{CAR}(Y \mid Q, G, D)$$

is a pointwise well-defined function of Y. Note that if $IC_0(Y \mid Q_0, G, D) = IC_0(Y \mid G, D) + IC_{nu}(Y \mid Q_0, D)$ with $IC_{nu}(Y \mid Q_0, G, D) \in T_{CAR}$, such as the orthogonalized mapping of the previous subsection, then $IC(Y \mid Q_0, Q, G, D)$ does not depend on Q_0. For simplicity, let Q include Q_0 if needed so that we can denote $IC(Y \mid Q_0, Q, G, D)$ with $IC(Y \mid Q, G, D)$.

Again, this mapping can be viewed as a mapping from full data estimating functions $D_h(\cdot \mid \mu, \rho)$ for μ into observed data estimating functions $IC(\cdot \mid Q, G, D_h(\cdot \mid \mu, \rho))$ for μ, indexed by unknown nuisance parameters $Q(F_X, G)$ and G. Theorem 1.3 proves that, if the set $\{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(\mu(F_X), \rho(F_X), \rho_1(F_X), G)\}$ of full data functions satisfying $E_G(IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu(F_X), \rho(F_X))) \mid X) = D_h(X \mid \mu(F_X), \rho(F_X))$ equals $T^{F,\perp}_{nuis}(F_X)$, then $\{IC(Y \mid Q(F_X, G), G, D) : D \in T^{F,\perp}_{nuis}(F_X)\}$ equals the orthogonal complement of the nuisance tangent space $T^{\perp}_{nuis}(P_{F_X,G})$ in model $\mathcal{M}(CAR)$, which includes the efficient influence curve $IC(Y \mid Q(F_X, G), G, D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X, G)))$. This mapping generates a class of estimating functions for the model $\mathcal{M}$ because

of the double protection property proved above:

$$E_{F_X,G}IC(Y \mid Q(F_X, G_1), G_1, D) = E_{F_X}D(X)$$
$$\text{for all } D \in \mathcal{D}(\rho_1(F_X), G_1) \text{ and } G_1 \in \mathcal{G} \text{ with } dG/dG_1 < \infty. \quad (2.30)$$
$$E_{F_X,G}IC(Y \mid Q, G, D) = E_{F_X}D(X)$$
$$\text{for all } Q \in \mathcal{Q},\ D \in \mathcal{D}(\rho_1(F_X), G). \quad (2.31)$$

Implications of protection property (2.30) for estimation of Q and ρ.

In model $\mathcal{M}(\mathcal{G})$, one only needs to rely on (2.31), which allows one to estimate $Q(F_X, G)$ with any estimator Q_n, and one needs to estimate $\rho = \rho(F_X, G)$ with a consistent estimator. However, in model $\mathcal{M}$, one needs to exploit (2.30). Let F_n be an estimator of F_X according to a working model $\mathcal{M}^w$, and assume that we use a substitution estimator $Q_n = Q(F_n, G_n)$, and $\rho_n = \rho(F_n, G_n)$ of $\rho(F_X, G)$. Consider the situation that $F_n \to F_X$ but $G_n \to G_1$ for a possibly wrong G_1. Then (2.30) teaches us that we need $Q_n \to Q(F_X, G_1)$, which naturally will hold, and $D \in \mathcal{D}(\rho_1(F_X), G_1)$. We will now explain why the latter condition can also be expected to hold. Recall that the nuisance parameter ρ in the full data structure estimating function $D_h(X \mid \mu, \rho)$ includes G as a component, which we needed to make sure that the estimating function at the true parameter values is an element of $\mathcal{D}(\rho_1(F_X), G)$. Since we estimate G with G_n, $G_n \to G_1$, and $F_n \to F_X$, we would precisely obtain that

$$D_h(\cdot \mid \mu_n, \rho_n) \to D_h(\cdot \mid \mu(F_X), \rho(F_X, G_1)) \in \mathcal{D}(\rho_1(F_X), G_1),$$

in the limit, as required.

Note also, as stressed in the discussion after Theorem 1.6, one does not need that $D_h(\cdot \mid \mu(F_X), \rho(F_X, G)) \in \mathcal{D}(\rho_1(F_X), G))$ necessarily; that is, $IC_0(Y \mid G, D(\cdot \mid \mu(F_X), \rho(F_X, G)))$ is allowed to be biased under the true data generating distribution $P_{F_X,G}$, but we do need that $IC_0(Y \mid G_1, D(\cdot \mid \mu(F_X), \rho(F_X, G_1)))$ needs to be unbiased under the possibly misspecified P_{F_X,G_1}, which holds if $D_h(\cdot \mid \mu(F_X), \rho(F_X, G_1)) \in \mathcal{D}(\rho_1(F_X), G_1)$.

A score operator representation

Let $A_{F_X} : L_0^2(F_X) \to L_0^2(P_{F_X,G})$ be the nonparametric score operator for F_X:

$$A_{F_X}(s)(Y) = E(s(X) \mid Y).$$

The adjoint $A_G^\top : L_0^2(P_{F_X,G}) \to L_0^2(F_X)$ of A_{F_X} is given by

$$A_G^\top(V)(X) = E(V(Y) \mid X).$$

Let $\mathbf{I}_{F_X,G} = A_{F_X}A_G^\top : L_0^2(F_X) \to L_0^2(F_X)$ which will be referred to as the nonparametric information operator. As shown in Theorem 1.3 (Chapter 1), for $D \in R(\mathbf{I}_{F_X,G})$ we have

$$IC(Y \mid Q(F_X, G), G, D) = A_{F_X}\mathbf{I}_{F_X,G}^{-1}(D). \quad (2.32)$$

Double protection property

The least squares representation $D \rightarrow IC(\cdot \mid Q(F_X, G), G, D) = IC(\cdot \mid F_X, G, D) \equiv A_{F_X} \mathbf{I}^-_{F_X,G}(D)$ from full data estimating functions to observed data estimating equations indexed by nuisance parameters F_X, G satisfies the double protection property (see Theorem 1.7 for the fact that we do not need the condition $dG/dG_1 < \infty$):

$$\begin{aligned} E_{F_X,G} IC(Y \mid F_X, G_1, D) &= E_{F_X} D(X) \text{ for } D \in \mathcal{D}(F_X, G_1) \\ &\quad \text{for all } G_1 \in \mathcal{G}(CAR), \\ E_{F_X,G} IC(Y \mid F_{X1}, G, D) &= E_{F_X} D(X) \\ &\quad \text{for } F_{X1} \in \mathcal{M}^F,\ D \in \mathcal{D}(F_X, G), \end{aligned}$$

where

$$\mathcal{D}(F_X, G) = \{D \in \mathcal{D} : D \in R(I_{F_1,G}) \text{ for all } F_1 \in \mathcal{M}^F\}$$

plays the role of $\mathcal{D}(\rho_1(F_X), G)$. Here $D \in R(\mathbf{I}_{F_1,G})$ denotes that D is an element of the range of the information operator $\mathbf{I}_{F_1,G} : L_0^2(F_X) \rightarrow L_0^2(F_X)$. Alternatively, we could require $D \in R_\infty(I_{F_1,G})$ for all $F_1 \in \mathcal{M}^F$, as defined in Theorem 2.1.

2.4.1 The corresponding estimating equation

Consider the optimal mapping $IC(Y \mid Q, G, D)$ from full data structure estimating functions $\{D_h : h \in \mathcal{H}^F\}$ into observed data estimating functions. As described in detail in the previous section, given such a mapping, one first needs to identify the index set $\mathcal{H}^F(\mu, \rho, \rho_1, G) \subset \mathcal{H}^F$ so that

$$\{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F((\mu, \rho, \rho_1)(F_X), G)\} \subset \mathcal{D}(\rho_1(F_X), G).$$

Subsequently, one reparametrizes the allowed full data structure estimating functions $\{D_h : h \in \mathcal{H}^F(\mu, \rho, \rho_1, G)\}$ as $\{D_h^r : h \in \mathcal{H}^F\}$ by including the membership $I(D_h(\cdot \mid \mu(F_X), \rho(F_X)) \in \mathcal{D}(\rho_1(F_X), G))$ as an additional nuisance parameter as in (2.14). In this manner, one obtains a set of full data estimating functions that satisfy $E_G(IC(Y \mid Q(F_X, G), G, D_h^r) \mid X) = D_h^r(X)$ F_X-a.e. when D_h^r is evaluated at the true parameter values. As we mentioned, for notational convenience, this reparametrized set of allowed full data structure estimating functions is denoted with $D_h(\cdot \mid \mu, \rho)$, where ρ now also includes G as a component.

Consider estimators G_n and Q_n. In model $\mathcal{M}$, we assume that $Q_n = Q(F_n, G_n)$ is a substitution estimator, where F_n is an estimator of F_X that is consistent at $F_X \in \mathcal{M}^{F,w}$ so that either G_n or F_n will be consistent. Assume also that we have available an estimator ρ_n that is consistent for $\rho(F_X, G_1$, where G_1 is the limit of G_n; that is, ρ_n is consistent for $\rho(F_X, G)$ in model $\mathcal{M}(\mathcal{G})$, and consistent for $\rho(F_X, G_1)$ in model $\mathcal{M}$. Thus, in model $\mathcal{M}$ one should use a substitution estimator $\rho_n = \rho(F_n, G_n)$. In the next

paragraph we provide a general strategy for providing such a estimator ρ_n in model $\mathcal{M}$. Note that in essence we require the existence of a doubly robust estimator of the F_X-parameter $\rho(F_X, G_1)$, which might itself require the doubly robust estimation methodology we present for estimation of μ.

Remark: Doubly robust estimation of nuisance parameter ρ in model $\mathcal{M}$.

For simplicity, consider the case where the nuisance parameter $\rho = \rho(F_X)$ does not have a G-component. As pointed out above, if ρ includes a G-component and G_n converges to a G_1, then we need that ρ_n converges to $\rho = \rho(F_X, G_1)$. As a consequence, in this case one just applies the following to the F_X-parameter $\rho_1(F_X) = \rho(F_X, G_1)$, where G_1 is estimated with G_n. We can obtain consistent estimator (i.e. doubly robust estimator) ρ_n of ρ under Model M, even when ρ is a non-regular parameter (i.e., a parameter for which the semiparametric information bound is 0) in model M^F. To do so, we express the non-regular parameter ρ as the limit of a regular parameter ρ_σ as $\sigma \to 0$, where σ is a bandwidth or other regularization parameter. See van der Laan and Robins (1998) for an example, and van der Laan, van der Vaart (2002). Because ρ_σ is a regular parameter of F_X we can often construct consistent estimators of $\rho(\sigma)$ in Model M (i.e., doubly robust estimators of ρ_σ). Let σ_n be an appropriate bandwidth or regularization parameter corresponding to sample size n and let $\hat{\rho}_{\sigma_n}$ be the corresponding doubly robust estimator of ρ_{σ_n}. Then, as in Robins and Rotnitzky (2001), we obtain a doubly robust estimator of μ using the approach discussed above by using $\hat{\rho}_{\sigma_n}$ as an estimator of ρ

To achieve higher efficiency, it makes sense to use a data-dependent index h_n. We use as the estimating equation for μ

$$0 = \frac{1}{n}\sum_{i=1}^{n} IC(Y_i \mid Q_n, G_n, D_{h_n}(\cdot \mid \mu, \rho_n)). \tag{2.33}$$

Note that h indexes a whole class of estimating functions. In Section 2.8, we identify the optimal index $h_{opt}(F_X, G)$, which yields the optimal estimating function, given that we know the true (F_X, G). In model $\mathcal{M}$, one estimates $h_{opt}(F_X, G)$ with a plug-in estimator $h_n = h_{opt}(F_{X,n}, G_n)$, assuming our working models $\mathcal{M}^{F,w}$ and $\mathcal{G}$. This estimator h_n will be consistent for $h_{opt}(F_X, G)$ if both working models $\mathcal{M}^{F,w}$ and $\mathcal{G}$ are correctly specified.

One can solve the estimating equation (2.33) with the Newton–Raphson procedure described in Section 2.3. Let μ_n^0 be an initial estimator or guess. The first step of the Newton–Raphson procedure involves estimation of a derivative (matrix) w.r.t. μ_n^0. This derivative is defined by

$$c(\mu) = c(h, \mu, \rho, Q, G, P) = \frac{d}{d\mu} PIC(Y \mid Q, G, D_{h_n}(\cdot\mu, \rho)),$$

where we used the notation $Pf \equiv \int f(y)dP(y)$. Let

$$IC(Y \mid Q, G, c, D_h(\cdot\mu, \rho)) = c^{-1}IC(Y \mid Q, G, D_h(\cdot\mu, \rho)) \tag{2.34}$$

be the standardized estimating function in the sense that it has the derivative minus the identity.

Note that $c(\mu)$ is a $k \times k$ matrix with $c_{ij}(\mu) = P\frac{d}{d\mu_j}IC_i(Y \mid Q, G, D_{h_n}(\mu, \rho))$. Its estimate $c_n(\mu_n^0)$ is given by

$$c(h_n, \mu_n^0, \rho_n, Q_n, G_n, P_n) = \frac{1}{n}\sum_{i=1}^{n}\frac{d}{d\mu}IC(Y_i \mid Q_n, G_n, D_{h_n}(\mu, \rho_n))\Big|_{\mu=\mu_n^0}.$$

If $IC(Y \mid Q_n, G_n, D_{h_n}(\cdot \mid \mu, \rho_n))$ is not differentiable in μ, but the integral of $IC(Y)$ w.r.t. $P_{F_X,G}$ is differentiable w.r.t. μ, then the derivative $d/d\mu$ is defined as a numerical derivative. The first step of the Newton–Raphson procedure is now defined by

$$\mu_n^1 = \mu_n^0 - c_n(\mu_n^0)^{-1}\frac{1}{n}\sum_{i=1}^{n}IC(Y_i \mid Q_n, G_n, D_{h_n}(\cdot \mid \mu_n^0, \rho_n)). \tag{2.35}$$

If one has a decent initial estimator μ_n^0 available, then one can use this one-step estimator μ_n^1. Otherwise, one iterates until convergence is established and we possibly need to use the line-search modification as provided in Section 2.3.

2.4.2 Discussion of ingredients of a one-step estimator

At this stage, it is appropriate to discuss the ingredients of our proposed estimator (2.35). To begin with, let us discuss estimation of $Q = Q(F_X, G)$ and the censoring mechanism G. Under coarsening at random, the likelihood of Y actually factorizes in a likelihood parametrized by F_X and a likelihood parametrized by G. In model $\mathcal{M}$, one assumes that the user has supplied a lower-dimensional model $\mathcal{M}^{F,w}$ for F_X and a lower-dimensional model $\mathcal{G}$ for G. If these models are of low enough dimension, then one can estimate F_X and G by maximizing their corresponding likelihoods. In all applications covered in this book, we can estimate G with maximum likelihood methods. For example, in the right-censored data structures, we estimate G with the maximum partial likelihood estimator for the multiplicative intensity (i.e., Cox proportional hazards) model. However, since one does not need to estimate the whole full data distribution F_X but just the component $Q(F_X, G_n)$, other direct methods for estimation of $Q(F_X, G_n)$ are often available. We provide such methods in Chapters 3 and 6.

In model $\mathcal{M}(\mathcal{G})$, one does not need to estimate Q by substituting estimators F_n of F_X and G_n of G. Instead, one can often estimate Q directly with standard software, which will be illustrated in our examples.

Another issue is the choice h of the full data estimating function. The efficiency of the proposed estimator depends on this choice. In particular, in Section 2.8 we provide a choice $h_{opt} = h_{opt}(F_X, G)$, which makes the estimating function optimal. In model $\mathcal{M}$, one estimates h_{opt} with $h_n = h_{opt}(F_n, G_n)$, while in model $\mathcal{M}(\mathcal{G})$, other direct methods are often available. In particular, Theorem 2.8 establishes such a direct easy-to-estimate representation of h_{opt} for the multivariate generalized linear regression full data model. If the full data model is locally saturated then the optimal choice of full data structure estimating function is actually the optimal estimating function one would use in the full data model. In that case, $D_{h_{opt}}(X \mid \mu(F_X), \rho(F_X))$ equals the efficient score or efficient influence function $S^{*F}_{eff}(X \mid F_X)$ of μ in the full data structure model.

In general, evaluating $D_{h_{opt}}$ requires inverting a linear Hilbert space operator (i.e., a possibly infinite-dimensional system of linear equations). In Section 2.8, we provide a Neumann series algorithm for evaluating $D_{h_{opt}}$ for a given (F_X, G) and useful characterizations of the inversion problem that have resulted in closed-form solutions in many of our examples.

2.5 Guaranteed Improvement Relative to an Initial Estimating Function

If one assumes model $\mathcal{M}(\mathcal{G})$, then it will be possible to construct an estimating function that yields estimators at least as efficient as a given initial estimator.

Define

$$IC(\cdot \mid Q_0, Q, G, c_{nu}, D) = IC_0(\cdot \mid Q_0, G, D) - c_{nu} IC_{nu}(\cdot \mid Q, G, D), \quad (2.36)$$

where $IC_{nu}(Y \mid Q(F_X, G), G, D)$ parametrizes the projection of $IC_0(Y \mid Q_0, G, D)$ onto a subspace $H(P_{F_X,G})$ of T_{CAR}, as in Section 2.3. Given functions $IC_0(Y)$ and $IC_{nu}(Y)$, we define $c_{nu} = c_{nu}(IC_0, IC_{nu}, P_{F_X,G})$ as the projection matrix

$$E_{P_{F_X,G}}(IC_0(Y)IC_{nu}^{\top}(Y))E_{P_{F_X,G}}\{IC_{nu}(Y)IC_{nu}^{\top}(Y)\}^{-1}$$

so that $c_{nu}IC_{nu} = \Pi(IC_0 \mid \langle IC_{nu} \rangle)$. Note that the j-th component of $c_{nu}IC_{nu}$ equals the projection of IC_{0j} on the space $\langle IC_{nu,l}, l = 1, \ldots, k \rangle_{j=1}^{k}$ spanned by $IC_{nu,l}$, $l = 1, \ldots, k$. Note also that if IC_{nu} equals the projection of IC_0 onto a subspace of $L_0^2(P_{F_X,G})$, then $c_{nu} = I$, where I denotes the identity matrix. Estimation of c_{nu} only involves taking empirical expectations of the already estimated IC_0, IC_{nu}, and it will guarantee that the estimating function is more efficient than the estimating function $IC_0(\cdot \mid Q_0, G, D_h(\cdot \mid \mu, \rho))$, even when Q_n is an inconsistent estimator of

$Q(F_X, G)$. We estimate c_{nu} with

$$c_{nu,n} = \left[1/n \sum_{i=1}^{n} \widehat{IC_0}(Y_i)\widehat{IC}_{nu}^{\top}(Y_i)\right]\left[1/n \sum_{i=1}^{n} \widehat{IC}_{nu}(Y_i)\widehat{IC}_{nu}(Y_i)^{\top}\right]^{-1},$$

where $\widehat{IC}_0(Y) = IC_0(Y \mid Q_{0n}, G_n, D_h(\mu, \rho_n))$ and similarly we define $\widehat{IC}_{nu}$. With this c_{nu} extension, the estimating equation (2.33) for μ becomes

$$0 = \frac{1}{n} \sum_{i=1}^{n} IC(Y_i \mid Q_{0n}, Q_n, G_n, c_{nu,n}, D_{h_n}(\cdot \mid \mu, \rho_n)). \tag{2.37}$$

If the parameter $Q_0(F_X, G)$ of the initial estimating function $IC_0(\cdot \mid Q_0, G, D(\cdot \mid \mu, \rho))$ is already easy to estimate in the model $\mathcal{M}^F$, then we recommend estimating $Q_0(F_X, G)$ consistently. In that case, the proposed one-step estimator μ_n^1 corresponding with our estimating function (2.37) is asymptotically linear with an influence curve with smaller variance than $c^{-1}IC_0(\cdot \mid Q_0(F_X, G), G, D_h(\cdot \mid \mu, \rho))$, even when Q_n is inconsistent. Note that $IC_0(\cdot \mid Q_0(F_X, G), G, D_h(\cdot \mid \mu, \rho))$ can be chosen to represent the influence curve of a good initial estimator μ_n^0 (e.g., inverse probability of censoring weighted estimator estimating G according to a model $\mathcal{G}$) so that μ_n^1 is guaranteed to be more efficient than μ_n^0. It is also of interest to note that inspecting $c_{nu,n}$ for a number of fits $\widehat{IC}_{nu}$ can provide insight into which fit Q_n results in the best approximation of $\Pi(\widehat{IC}_0 \mid H(P_{F_X,G}))$; that is, one selects the fit which makes $c_{nu,n}$ closest to the identity matrix.

If one assumes the more nonparametric model $\mathcal{M}$, then the c_{nu} extension is not a good idea since it will destroy the protection w.r.t. misspecification of G at correctly specified guessed full data structure model $\mathcal{M}^{F,w}$.

Example 2.13 (Multivariate right censored data structure; continuation of Example 2.8) Let (T_1, T_2) be a bivariate survival time of interest, and let $\mu = S(t_1, t_2) = P(T_1 > t_1, T_2 > t_2)$. Let F be the bivariate cumulative distribution function of (T_1, T_2). Let $C = (C_1, C_2)$ be a bivariate censoring variable. Suppose that we observe $(\tilde{T}_j = \min(T_j, C_j), \Delta_j = I(T_j \leq C_j), \bar{L}_j(\tilde{T}_j))$, $j = 1, 2$, where $L_j(\cdot)$ are covariate processes. We have for full data $X = (T_1, \bar{L}_1(T_1), T_2, \bar{L}(T_2))$. The observed data distribution is indexed by the full data distribution F_X and the conditional bivariate distribution G of (C_1, C_2), given X, which is assumed to be modeled with some submodel of CAR (e.g., see Chapter 6). Let the full data model be nonparametric so that the only full data estimating function is $D(X \mid \mu) = I_{(t_1,\infty)\times(t_2,\infty)}(T_1, T_2) - \mu$.

In the previous coverage of this example, we derived an estimating function $IC(Y \mid F, G, D(\cdot \mid \mu))$ (2.24) for μ indexed by a marginal bivariate distribution F and the censoring mechanism G, which yields an influence curve equal to or better than the influence curve of Dabrowska's estimator. We propose now to use this as the initial influence curve IC_0 in the

estimating function $IC_0 - c_{nu}IC_{nu}$, where IC_{nu} denotes the projection of IC_0 onto a subspace of T_{CAR}.

In this multivariate right-censored data model (no common censoring time), the CAR model $\mathcal{G}(CAR)$ for the censoring mechanism is hard to understand and, in particular, projections on its tangent space $T_{CAR}(P_{F_X,G})$ do not exist in closed form. However, in Chapter 6, we consider an interesting submodel of CAR, only assuming that censoring actions at time t are sequentially randomized w.r.t. the observed past. This submodel of $\mathcal{G}(CAR)$ has a closed-form observed data tangent space $T_{SRA}(P_{F_X,G}) \subset T_{CAR}$. In Chapter 6, we also propose a semiparametric multiplicative intensity model $\mathcal{G}$ of this SRA model that yields an estimator of G with standard software. A closed-form representation of its tangent space T_{SRA} and the projection operator on this tangent space are provided in Chapter 6. Thus $IC_{nu} = \Pi(IC_0 \mid T_{SRA})$ exists in closed form. The estimating function $IC_0 - c_{nu}IC_{nu}$ now yields an RAL estimator of μ based on the extended bivariate right-censored data structure in model $\mathcal{M}(\mathcal{G})$, which is guaranteed to be more efficient than the Dabrowska estimator under independent censoring, even for the marginal data structure in which data on the covariate processes $L_j(\cdot)$ are not available. □

2.6 Construction of Confidence Intervals

Firstly, consider model $\mathcal{M}(\mathcal{G})$; that is, we are willing to assume that the censoring mechanism is correctly specified so that $\mathcal{G}$ contains the true G. Let $T_2(P_{F_X,G})$ denote the corresponding tangent space generated by G. Given an initial estimator μ_n^0 that converges to μ at an appropriate rate, we consider the one-step estimator (2.35) that is given by $\mu_n^0 + 1/n \sum_{i=1}^n \widehat{IC}(Y_i)$, where $\widehat{IC}(Y) \equiv IC(Y \mid Q_n, G_n, c_{nu,n}, c_n, D_{h_n}(\cdot \mid \mu_n^0, \rho_n))$ defined by (2.34). Under the conditions of Theorem 2.4, μ_n^1 is asymptotically linear with influence curve $IC_1(Y) - \Pi(IC_1 \mid T_2)$, where $IC_1(Y) = IC(Y \mid Q^1, G, c, D_h(\cdot \mid \mu, \rho))$ represents the limit for $n \to \infty$ of $\widehat{IC}(Y)$. Thus, estimation of the influence curve requires computing an expression for the projection formula of IC_1 onto the tangent space generated by the censoring mechanism in the observed data model $\mathcal{M}(\mathcal{G})$. We provide this projection formula in Lemma 3.2 for the case where C is identified with a counting process and one uses a multiplicative intensity model to model the intensity of this counting process w.r.t. the observed past. Alternatively, one can note that this influence curve has variance smaller than or equal to the variance of $IC_1(Y) \equiv IC(Y \mid Q^1, G, c_{nu}, c, D_h(\cdot \mid \mu, \rho))$ and use a conservative estimate of the asymptotic covariance matrix of μ_n^1:

$$\widehat{\Sigma} = \frac{1}{n} \sum_{i=1}^n \widehat{IC}(Y_i) \widehat{IC}(Y_i)^\top.$$

This can be used to construct a conservative 95% confidence interval for μ,

$$\mu_{nj}^1 \pm 1.96\frac{\widehat{\Sigma}_{jj}}{\sqrt{n}}. \tag{2.38}$$

This confidence interval is asymptotically correct if one consistently estimates $Q(F_X, G)$, and it is asymptotically conservative otherwise. However, if one uses the $c_{nu,n}$ adjustment and guarantees that Q_{0n} is consistent, then it is always less conservative than using as the influence curve $c^{-1}IC_0(Y \mid Q_0(F_X, G), G, D_h(\mu, \rho))$; see Section 2.5. This confidence interval (2.38) is practical since one gets it for free after having computed the estimator μ_n^1.

Consider now the more nonparametric model $\mathcal{M}$ only assuming that either the censoring mechanism model $\mathcal{G}$ or the $\mathcal{M}^{F,w}$ is correctly specified. Under the conditions of Theorem 2.5, μ_n^1 is asymptotically linear with an influence curve equal to a sum of two components of which one is consistently estimated by $\widehat{IC}$. The other component of this influence curve will depend on the linear expansion of the estimators of a smooth functional of the unknown parameters in model $\mathcal{M}^{F,w}$. If one wants to avoid the $\mathcal{M}^{F,w}$-specific technical exercise and wants a confidence interval that is also correct when G is misspecified (and $\mathcal{M}^{F,w}$ is correctly specified), then we recommend using the semiparametric or nonparametric bootstrap (e.g., Gill, 1989; Efron, 1990; Gine and Zinn, 1990; Efron and Tibshirani, 1993; van der Vaart and Wellner; 1996).

2.7 Asymptotics of the One-Step Estimator

An estimator μ_n of μ is asymptotically linear at $P_{F_X,G}$ with influence curve $IC(Y \mid F_X, G)$ if $\mu_n - \mu = n^{-1}\sum_{i=1}^n IC(Y_i \mid F_X, G) + o_P(n^{-1/2})$. From Bickel, Klaassen, Ritov and Wellner (1993), we have that an estimator is asymptotically efficient if it is asymptotically linear with the influence curve the so-called efficient influence curve, $IC^*(Y \mid F_X, G)$. The efficient influence curve is also called the canonical gradient and $IC^*(Y \mid F_X, G) = IC(Y \mid Q(F_X, G), G, D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X)))$ for a specified (next section) index $h_{opt}(F_X, G)$.

We prove two asymptotics theorems for the one-step estimator μ_n^1 corresponding with the estimating equation (2.37), one for model $\mathcal{M}(G)$ and one for model $\mathcal{M}$. We note that the estimating function $IC(Y \mid Q_0, G, c_{nu}, D_h) = IC_0(Q_0, G, D_h) - c_{nu}IC_{nu}(Q, G, D_h)$ captures all proposed estimating functions in the previous sections, where one can set $c_{nu} = I$ and/or $IC_{nu} = 0$ and/or IC_0 equal to the optimal mapping to obtain the various proposed estimating functions. In particular, in model $\mathcal{M}$ we set $c_{nu} = I$. Theorem 2.4 for $\mathcal{M}(\mathcal{G})$ below assumes consistent estimation of the censoring mechanism. Theorem 2.5 for $\mathcal{M}$ assumes *either*

consistent estimation of the censoring mechanism *or* consistent estimation of F_X. This does not require choosing which of the two quantaties are consistently estimated. Obviously, the last theorem provides the most nonparametric consistency and asymptotic normality result, but the price one has to pay is that one cannot use the conservative confidence interval (2.38). Because of this and the fact that for many censored data structures it is easier to estimate the censoring mechanism than it is to estimate the full data distribution, we feel that the theorem for $\mathcal{M}(\mathcal{G})$ deserves a separate treatment.

We note that Theorem 2.4 can be applied to any one-step estimator corresponding with the non optimal estimating equations $0 = \sum_i IC_0(Y_i \mid Q_{1n}, G_n, D_{h_n}(\cdot \mid \mu, \rho_n))$ provided in Section 2.3. Similarly, Theorem 2.5 can be applied to any IC_0 that is orthogonalized w.r.t. the tangent space $T_2(P_{F_X,G})$ of G so that it satisfies (2.30).

2.7.1 Asymptotics assuming consistent estimation of the censoring mechanism

The following theorem provides a template for proving asymptotic linearity with specified influence curve of this one-step estimator μ_n^1 (2.35) (i.e., set $c_{nu,n} = c_{nu} = 1$) or, if one uses the adjustment constant $c_{nu,n}$, then it is the one-step estimator corresponding with (2.37). Recall the following Hilbert space terminology: $L_0^2(P_{F_X,G})$ is the Hilbert space of functions of Y with finite variance and mean zero endowed with the covariance inner product $< v_1, v_2 >_{P_{F_X,G}} \equiv \sqrt{\int v_1 v_2 dP_{F_X,G}}$. The tangent space $T_2 = T_2(P_{F_X,G})$ for the parameter G is the closure of the linear extension in $L_0^2(P_{F_X,G})$ of the scores at $P_{F_X,G}$ from all correctly specified parametric submodels (i.e., submodels of the assumed semiparametric model $\mathcal{G}$) for the distribution G.

Theorem 2.4 *Consider the observed data model $\mathcal{M}(\mathcal{G})$. Let $Y_1, \ldots, Y_n$ be n i.i.d. observations of $Y \sim P_{F_X,G} \in \mathcal{M}(\mathcal{G})$. Consider a one-step estimator of the parameter $\mu \in \mathbb{R}^k$ of the form $\mu_n^1 = \mu_n^0 - P_n IC(\cdot \mid Q_n, G_n, c_{nu,n}, c_n, D_{h_n}(\mu_n^0, \rho_n))$ corresponding with (2.37). Assume that the limit of $IC(Q_n, G_n, c_{nu,n}, D_{h_n}(\mu_n^0, \rho_n))$ specified in (ii) below satisfies*

$$E_G(IC(Y \mid Q^1, G, c_{nu}, D_h(\cdot \mid \mu, \rho)) \mid X) = D_h(X \mid \mu, \rho) \; F_X\text{-a.e} \quad (2.39)$$

$$D_h(\cdot \mid \mu, \rho) \in T_{nuis}^{F,\perp}(F_X). \quad (2.40)$$

Let $f_n(\mu) \equiv P_n IC(\cdot \mid Q_n, G_n, c_{nu,n}, D_{h_n}(\mu, \rho_n))$. Assume (we write $f \approx g$ for $f = g + o_P(1/\sqrt{n})$)

$$c_n^{-1}\{f_n(\mu_n^0) - f_n(\mu)\} \approx \mu_n^0 - \mu. \quad (2.41)$$

and

$$E_{P_{F_X,G}} IC(Y \mid Q_n, G, c_{nu,n}, D_{h_n}(\mu, \rho_n)) = o_P(1/\sqrt{n}). \quad (2.42)$$

where the G component of ρ_n is set equal to G as well.

In addition, assume

(i) $IC(\cdot \mid Q_n, G_n, c_{nu,n}, c_n, D_{h_n}(\cdot \mid \mu_n^0, \rho_n))$ falls in a $P_{F_X,G}$-Donsker class with probability tending to 1.

(ii) Let $IC_n(\cdot) = IC(\cdot \mid Q_n, G_n, c_{nu,n}, c_n, D_{h_n}(\cdot \mid \mu_n^0, \rho_n))$. For some (h, Q^1), we have

$$\| IC_n(\cdot) - IC(\cdot \mid Q^1, G, c_{nu}, c, D_h(\cdot \mid \mu, \rho)) \|_{P_{F_X,G}} \to 0,$$

where the convergence is in probability. Here (suppressing the dependence of the estimating functions on parameters) $c_{nu} = \langle IC_0, IC_{nu}^\top \rangle \langle IC_{nu}, IC_{nu}^\top \rangle^{-1}$ is such that $c_{nu} IC_{nu}$ equals the projection of IC_0 onto the k-dimensional space $< IC_{nu,j}, j = 1, \ldots, k >$ in $L_0^2(P_{F_X,G})$.

(iii) Define for a G_1

$$\Phi(G_1) = P_{F_X,G} IC(\cdot \mid Q^1, G_1, c_{nu}, c, D_h(\mu, \rho)).$$

For notational convenience, let

$$\begin{aligned} IC_n(G) &\equiv IC(\cdot \mid Q_n, G, c_{nu,n}, c_n, D_{h_n}(\mu, \rho_n)), \\ IC(G) &\equiv IC(\cdot \mid Q^1, G, c_{nu}, c, D_h(\mu, \rho)). \end{aligned}$$

Assume

$$P_{F_X,G}\{IC_n(G_n) - IC_n(G)\} \approx \Phi(G_n) - \Phi(G).$$

(iv) $\Phi(G_n)$ is an asymptotically efficient estimator of $\Phi(G)$ for the CAR model $\mathcal{G}$ containing the true G with tangent space $T_2(P_{F_X,G}) \subset T_{CAR}(P_{F_X,G})$.

Then μ_n^1 is asymptotically linear with influence curve given by

$$IC \equiv \Pi(IC(\cdot \mid Q^1, G, c_{nu}, c, D_h(\cdot \mid \mu, \rho)) \mid T_2^\perp(P_{F_X,G})).$$

If $Q^1 = Q(F_X, G)$ and $IC(Y \mid Q(F_X, G), G, c_{nu}, D_h(\cdot \mid \mu, \rho)) \perp T_2(P_{F_X,G})$, then this influence curve equals $IC(\cdot \mid Q(F_X, G), G, c_{nu} = 1, c, D_h(\mu, \rho))$. In particular, if $h = h_{opt}$ so that $IC(Y \mid Q(F_X, G), G, D_{h_{opt}}(\cdot \mid \mu(F_X), \rho(F_X, G)))$ equals the efficient influence curve $S_{eff}^{*F}(Y \mid F_X, G)$, then μ_n^1 is asymptotically efficient.

Discussion of asymptotic linearity Theorem 2.4

We will discuss the assumptions of Theorem 2.4 and illustrate that the assumptions are natural. Firstly, note that the structural conditions (2.39) and (2.40) hold for our estimating functions by (2.31) and the fact that we choose the full data estimating functions to be elements of $T_{nuis}^{F,\perp}(F_X)$ at the true parameter values. Note also that these conditions imply that the estimating function is orthogonal to all F_X nuisance parameters in the sense that it is an element of $T_{nuis}^\perp(\mathcal{M}(G))$ in the model with G known at any Q_1. This explains why condition (2.42) is a natural condition; see Subsection 1.4.3.

Condition (2.41) is a natural condition as well, which is illustrated as follows. Define $f_n(\mu) \equiv P_n IC(\cdot \mid Q_n, G_n, D_{h_n}(\cdot \mid \mu, \rho_n))$. By definition, we have

$$c_n = f_n'(\mu_n^0) \equiv d/d\mu f_n(\mu)|_{\mu=\mu_n^0}.$$

In this notational setting, condition (2.41) translates to

$$\{f_n'(\mu_n^0)\}^{-1}\{f_n(\mu_n^0) - f_n(\mu)\} = \mu - \mu_n^0 + o_P(1/\sqrt{n}). \tag{2.43}$$

Under regularity conditions (e.g., a Taylor expansion of $f_n(\mu)$ at μ_n^0), one expects to have a first-order expansion

$$f_n(\mu) - f_n(\mu_n^0) = f_n'(\mu_n^0)(\mu_n^0 - \mu) + o(\mid \mu_n^0 - \mu \mid). \tag{2.44}$$

If the second-order term is $o_P(n^{-1/2})$ and the determinant $f_n'(\mu_n^0)$ is bounded away from zero uniformly in n, then this expansion proves (2.43) and thus (2.41). If $f_n(\mu)$ is continuously differentiable in μ with bounded derivative, then the second order term is $o_P(\mid \mu_n^0 - \mu \mid)$, while if it is twice continuously differentiable, then the second-order term is $O(\mid \mu_n^0 - \mu \mid^2)$. Consider now the case where $f(\mu) = P_{F_X,G} IC(\cdot \mid Q, G, D_h(\cdot \mid \mu, \rho))$ is continuously differentiable but $f_n(\mu)$ is not differentiable. In that case, one still expects this expansion (2.44) to hold with $f_n'(\mu_n^0)$ now being a numerical derivative. In many of our censored data models with a nonparametric full data model, we actually have that $f_n(\mu)$ is linear in μ so that (2.41) holds with remainder zero. However, in general, we can conclude that condition (2.41) typically requires a convergence rate of the initial estimator μ_n^0.

Condition (i) can often be arranged by choosing truly lower-dimensional working models $\mathcal{G}$ and $\mathcal{M}^{F,w}$ when estimating G and F_X. This condition formally represents the "asymptotic curse of dimensionality" since if one uses as working models $\mathcal{M}^{F,w} = \mathcal{M}^F$ and $\mathcal{G} = \mathcal{G}(CAR)$, then the class of functions of $IC(\cdot \mid Q_n, G_n, D_{h_n}(\cdot \mid \mu_n^0, \rho_n))$ of Y generated by varying Q_n, G_n over all possible parameter values will typically be very large (meaning that for finite samples the first-order asymptotics is irrelevant) or not even be a Donsker class. Condition (ii) is a weak consistency condition requiring that G_n be consistent and Q_n converge to something. This condition will hold if our model $\mathcal{M}(\mathcal{G})$ contains the true $P_{F_X,G}$. Regarding condition (iii), we have

$$\begin{aligned} P_{F_X,G} IC_n(G_n) - IC_n(G) &\approx P_{F_X,G} IC_n(G_n) \text{ by (2.42)} \\ &\approx P_{F_X,G-G_n} IC_n(G_n), \end{aligned}$$

where the latter approximation is expected to hold because of the protection (2.31) against misspecification of Q with the data-generating distribution being P_{F_X,G_n}. Thus, condition (iii) requires that second-order terms involving integrals of differences $(G_n - G)(Q_n - Q^1)$, $(G_n - G)(\rho_n - \rho)$ be $o_P(1/\sqrt{n})$. Thus, if G_n converges to G at a rate $n^{-1/2}$, then this condition typically only requires consistency of the other nuisance parameter

estimates Q_n, ρ_n, μ_n^0, but if G_n converges at a very low rate to G, then the other nuisance parameter estimates will have to compensate for this by converging at an appropriate rate.

Condition (iv) just requires that one estimates G with an efficient procedure such as maximum likelihood estimation. Condition (iv) is not needed to establish that μ_n^1 is RAL, but it is needed to obtain the elegant formula of its influence curve.

This finishes our discussion of the assumptions. Let us now consider the conclusions of Theorem 2.4. It is interesting to consider what the limit distribution of μ_n^1 would be when $G(\cdot \mid x)$ is known and is used in the one-step estimator. In that case, the nuisance tangent space T_2 is empty. Thus, by Theorem 2.4, the influence curve of μ_n^1 is then given by $IC(\cdot \mid Q^1, G, c_{nu}, c, D_h(\cdot \mid \mu, \rho))$, which has variance greater than or equal to that of the influence curve IC based on an efficient estimator of $G(\cdot \mid X)$ according to a (any) model for G. Lemma 2.3 provides the general understanding of the fact that efficient estimation of a known orthogonal nuisance parameter (such as G) improves efficiency of estimation of a parameter μ of the distribution of the full data structure X.

We also note that, due to the c_{nu}-adjustment, $IC(\cdot \mid Q^1, G, c_{nu}, c, D_h(\cdot \mid \mu, \rho))$ has variance smaller than or equal to the variance of $IC_0(\cdot \mid Q_0, G, c, D_h(\cdot \mid \mu, \rho))$. Thus, by choosing $IC_0(\cdot \mid Q_0(F_X, G), G, c, D_h(\cdot \mid \mu, \rho))$ equal to an influence curve of a given estimator, the one-step estimator will always be asymptotically more efficient than this estimator. Therefore, the inclusion of c_{nu} in the definition of the one-step estimator is only really useful if one sets $IC_0(\cdot \mid Q_0, G, c, D_h(\mu, \rho))$ equal to a challenging influence curve. Note that one can always make the choice IC_0 more challenging by redefining a new IC_0 as the old IC_0 minus the projection of the old IC_0 onto any given subset of scores in T_{CAR}.

Finally, we make some comments about the efficiency condition. We have that $h_{opt} = h_{opt}(F_X, G)$ is a functional of the true (F_X, G). In some examples, one has available a closed-form representation of h_{opt} that will imply natural methods of estimation. In general, we provide a Neumann series algorithm for calculating $h_{opt}(F_X, G)$ for a given (F_X, G). Let $F_{X,n}$ and G_n be the estimates of F_X and G assuming the lower dimensional working models $\mathcal{M}^{F,w} \subset \mathcal{M}^F$ and $\mathcal{G} \subset \mathcal{G}(CAR)$. Then, one can estimate h_{opt} with the plug-in method:

$$h_n = h_{opt}(F_{X,n}, G_n).$$

If the working model contains the truth, then (h_n, Q_n) consistently estimates $(h_{opt}, Q(F_X, G))$ so that under the "regularity" conditions (i)–(iv) μ_n^1 is asymptotically efficient. Otherwise, (h_n, Q_n) will still converge to some (h, Q^1) so that μ_n^1 will still be consistent and asymptotically linear.

2.7.2 Proof of Theorem 2.4

For notational convenience, we give the proof for $c_{nu,n} = 1$ and use obvious short-hand notation. We have

$$\begin{aligned}\mu_n^1 &= \mu_n^0 + c_n^{-1} P_n \left\{ IC(Q_n, G_n, D_{h_n}(\mu_n^0, \rho_n)) - IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)) \right\} \\ &\quad + c_n^{-1} P_n IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)).\end{aligned}$$

By condition (2.41), the difference on the right-hand side equals $\mu - \mu_n^0 + o_P(1/\sqrt{n})$. Thus, we have

$$\begin{aligned}\mu_n^1 - \mu &= (P_n - P) c_n^{-1} IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)) \\ &\quad + c_n^{-1} P IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)).\end{aligned}$$

For empirical process theory, we refer to van der Vaart and Wellner (1996). Conditions (i) and (ii) in the theorem imply that the empirical process term on the right-hand side is asymptotically equivalent with $(P_n - P_{F_X,G}) c^{-1} IC(\cdot \mid Q^1, G^1, D_h(\mu, \rho))$, so it remains to analyze the term

$$c_n^{-1} P IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)).$$

Now, we write this term as a sum of two terms $A + B$, where

$$\begin{aligned}A &= c_n^{-1} P \left\{ IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)) - IC(Q^1, G, D_h(\mu, \rho)) \right\}, \\ B &= c_n^{-1} P IC(Q^1, G, D_h(\mu, \rho)).\end{aligned}$$

By (2.39) and (2.40), we have $B = 0$. As in the theorem, let

$$\begin{aligned}IC_n(G) &\equiv IC(\cdot \mid Q_n, G, D_{h_n}(\mu, \rho_n(G))), \\ IC(G) &\equiv IC(\cdot \mid Q^1, G, D_h(\mu, \rho)).\end{aligned}$$

We decompose $A = A_1 + A_2$ as follows:

$$\begin{aligned}A &= P_{F_X,G}\{IC_n(G_n) - IC(G)\} \\ &= P_{F_X,G}\{IC_n(G) - IC(G)\} + P_{F_X,G}\{IC_n(G_n) - IC_n(G)\}.\end{aligned}$$

By assumption (2.42), we have $A_1 = o_P(1/\sqrt{n})$. By assumption (iii),

$$A_2 = \Phi_2(G_n) - \Phi_2(G) + o_P(1/\sqrt{n}).$$

By assumption (iv), we can conclude that μ_n^1 is asymptotically linear with influence curve $IC(\cdot \mid Q^1, G, c, c_{nu}, D_h(\mu, \rho)) + IC_{nuis}$, where IC_{nuis} is the influence curve of $\Phi_2(G_n)$. Now, the same argument as given in the proof of Theorem 2.3 proves that this influence curve of μ_n^1 is given by

$$\Pi(IC(\cdot \mid Q^1, G, c, c_{nu}, D_h(\mu, \rho)) \mid T_2^{\perp}).$$

This completes the proof. □

2.7.3 Asymptotics assuming that either the censoring mechanism or the full data distribution is estimated consistently

If one is only willing to assume that either the censoring mechanism or the full data distribution is modeled correctly (but not necessarily both), then one can apply the following asymptotic theorems.

Theorem 2.5 *Consider the observed data model $\mathcal{M}$. Let $Y_1, \ldots, Y_n$ be n i.i.d. observations of $Y \sim P_{F_X,G} \in \mathcal{M}$. Consider a one-step estimator $\mu_n^1 = \mu_n^0 + P_n IC(Y_i \mid Q_n, G_n, c_n, D_{h_n}(\mu_n^0, \rho_n))$ (e.g., (2.35) or the one-step estimator corresponding with estimating equation (2.37) with $c_{nu} = I$) of the parameter $\mu \in \mathbb{R}^k$. Assume that the limit of $IC(\cdot \mid Q^1, G^1, c, D_h(\mu, \rho(F_X, G^1)))$ in (ii) satisfies*

$$P_{F_X,G} IC(Y \mid Q^1, G^1, c_n, D_{h_n}(\mu(F_X), \rho(F_X, G^1))) = 0. \tag{2.45}$$

Let $f_n(\mu) = P_n \{IC(\cdot \mid Q_n, G_n, D_{h_n}(\mu, \rho_n))\}$. Assume that

$$c_n^{-1}\{f_n(\mu_n^0) - f_n(\mu)\} = \mu_n^0 - \mu + o_P(1/\sqrt{n}). \tag{2.46}$$

In addition, assume that
(i) $IC(\cdot \mid Q_n, G_n, c_n, D_{h_n}(\mu_n^0, \rho_n))$ falls in a $P_{F_X,G}$-Donsker class with probability tending to 1.
(ii) Let $IC_n(\cdot) = IC(\cdot \mid Q_n, G_n, c_n, D_{h_n}(\cdot \mid \mu_n^0, \rho_n))$. For some (h, Q^1, G^1) with either $Q^1 = Q(F_X, G^1)$ or $G^1 = G$, we have

$$\| IC_n(\cdot) - IC(\cdot \mid Q^1, G^1, c, D_h(\cdot \mid \mu, \rho(F_X, G^1))) \|_{P_{F_X,G}} \to 0,$$

where the convergence is in probability.
(iii) Let $\rho = \rho(F_X, G) = (\rho^*, G)$ for a F_X-parameter $\rho^*(F_X)$. Define

$$\begin{aligned}
\Phi_1(Q) &= P_{F_X,G} IC(\cdot \mid Q, G^1, c, D_h(\cdot \mid \mu, (\rho^*, G^1))), \\
\Phi_2(G') &= P_{F_X,G} IC(\cdot \mid Q^1, G', c, D_h(\cdot \mid \mu, (\rho^*, G'))), \\
\Phi_3(\rho^*) &= P_{F_X,G} IC(\cdot \mid Q^1, G^1, c, D_h(\cdot \mid \mu, (\rho^*, G^1))).
\end{aligned}$$

Assume that

$$P_{F_X,G}\left\{IC(\cdot \mid Q_n, G_n, c_n, D_{h_n}(\mu, \rho_n)) - IC(Q^1, G^1, c_n, D_{h_n}(\mu, \rho^*, G^1))\right\}$$

$$= \Phi_1(Q_n) - \Phi_1(Q^1) + \Phi_2(G_n) - \Phi_2(G^1) + \Phi_3(\rho_n^*) - \Phi_3(\rho^*) + o_P(1/\sqrt{n}).$$

(iv) Assume that $\Phi_1(Q_n)$ is a regular asymptotically linear estimator at $P_{F_X,G}$ of $\Phi_1(Q^1)$ with influence curve $IC_1(Y \mid F_X, G)$, $\Phi_2(G_n)$ is a regular asymptotically linear estimator at $P_{F_X,G}$ of $\Phi_2(G^1)$ with influence curve $IC_2(Y \mid F_X, G)$, and $\Phi_3(\rho_n^*)$ is a regular asymptotically linear estimator at $P_{F_X,G}$ of $\Phi_3(\rho^*(F_X))$ with influence curve $IC_3(Y \mid F_X, G)$.

Then μ_n^1 is a regular asymptotically linear estimator with influence curve

$$IC \equiv IC(\cdot \mid Q^1, G^1, c, D_h(\cdot \mid \mu, \rho)) + (IC_1 + IC_2 + IC_3)(Y \mid F_X, G).$$

Now, also assume

$$\Phi_1(Q_n) = o_P(1/\sqrt{n}) \text{ at } G^1 = G, \tag{2.47}$$
$$\Phi_2(G_n) = o_P(1/\sqrt{n}) \text{ at } Q^1 = Q(F_X, G^1), \tag{2.48}$$
$$\Phi_3(\rho_n^*) = o_P(1/\sqrt{n}) \text{ at } G^1 = G. \tag{2.49}$$

If $G^1 = G$*, then* $IC_1 = IC_3 = 0$*. If* $Q^1 = Q(F_X, G)$*, then* $IC_2 = 0$*. If* $G^1 = G$ *and* $Q^1 = Q(F_X, G)$*, then* $IC = IC(\cdot \mid Q(F_X, G), G, c, D_h(\cdot \mid \mu, \rho))$*. In particular, if also* $h = h_{opt}$ *so that* $IC(Y \mid Q(F_X, G), G, D_{h_{opt}}(\cdot \mid \mu(F_X), \rho(F_X, G))) = S^*_{eff}(Y \mid F_X, G)$*, then* μ_n^1 *is asymptotically efficient.*

Note that condition (2.45) relies on the double robustness of the estimating function. One expects (2.47) and (2.48) to hold by the protection (2.31) against misspecification of Q and protection (2.30) against misspecification of G, respectively. In addition, one often expects (2.49) to hold since the estimating function at G is orthogonal to all F_X nuisance parameters in the sense that it is an element of $T^{\perp}_{nuis}(\mathcal{M}(G))$ in the model with G known at any Q_1. Specifically we would expect (2.49) to hold when regular estimators of μ can be constructed based on full data structure X, and consistent estimators of $\rho(F_X, G^1)$ can be constructed in model $\mathcal{M}$ using the approach described in the remark in Section 2.4.1. This shows that all structural conditions in this theorem are natural.

2.7.4 Proof of Theorem 2.5

We have

$$\begin{aligned}\mu_n^1 &= \mu_n^0 + c_n^{-1} P_n IC(Q_n, G_n, D_{h_n}(\mu_n^0, \rho_n)) - IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)) \\ &\quad + c_n^{-1} P_n IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)).\end{aligned}$$

By condition (2.41), the difference on the right-hand side equals $\mu - \mu_n^0 + o_P(1/\sqrt{n})$. Thus, we have

$$\begin{aligned}\mu_n^1 - \mu &= (P_n - P) c_n^{-1} IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)) \\ &\quad + c_n^{-1} P IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)).\end{aligned}$$

For empirical process theory, we refer to van der Vaart and Wellner (1996). Conditions (ii) and (iii) in the theorem imply that the empirical process term on the right-hand side is asymptotically equivalent with $(P_n - P_{F_X,G}) IC(\cdot \mid Q^1, G^1, D_h(\mu, \rho))$, so it remains to analyze the term

$$c_n^{-1} P IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)).$$

Now, we write this term as a sum of two terms $A + B$, where

$$\begin{aligned}A &= c_n^{-1} P IC(Q_n, G_n, D_{h_n}(\mu, \rho_n)) - IC(Q^1, G^1, D_{h_n}(\mu, (\rho^*, G^1))), \\ B &= c_n^{-1} P IC(Q^1, G^1, D_{h_n}(\mu, (\rho^*, G^1))).\end{aligned}$$

We have $B = 0$ by (2.45). By conditions (iii) and (iv), we have that A equals in first order $(P_n - P)\{IC(Q^1, G^1, D_h(\mu, \rho(G^1))) + IC_1 + IC_2 + IC_3\}$. The other statements are true by assumption. □

2.8 The Optimal Index

Consider representations $T^{F,\perp}_{nuis}(F_X) = \{D_h(X \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\}$ of the orthogonal complement of the full data nuisance tangent space for all $F_X \in \mathcal{M}^F$. Let $h_{ind,F_X} : L^2_0(F_X) \to \mathcal{H}^F(F_X)$ be the index mapping (2.2) from $L^2_0(F_X)$ to the index set $\mathcal{H}^F(F_X)$. Let $h_{eff}(F_X) = h_{ind,F_X}(S^F_{eff}(\cdot \mid F_X))$ be the index of the full data canonical gradient.

Consider the optimal mapping $D_h \to IC(Y \mid Q(F_X, G), G, D_h)$ characterized by the conditions $E_G(IC(Y \mid Q, G, D_h) \mid X) = D_h(X)$ F_X-a.e. and $IC(Y \mid Q(F_X, G), G, D_h) \perp T_{CAR}(P_{F_X,G})$. For simplicity, we will here assume that this mapping satisfies these conditions for all $D \in T^{F,\perp}_{nuis}(F_X)$. Due to its double robustness property, Theorem 2.5 for model $\mathcal{M}$ shows, under regularity conditions, that if both working models are correctly specified, then our proposed estimator μ^1_n is regular and asymptotically linear at $P_{F_X,G} \in \mathcal{M}$ with influence curve $IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu(F_X), \rho(F_X)))$. The following corollary of Theorem 1.3 in Chapter 1 shows that, given any influence curve $IC(Y \mid F_X, G)$ of a regular asymptotically linear estimator of μ at $P_{F_X,G}$, we can choose $D_{h(F_X,G)}(\cdot \mid \mu, \rho) \in T^{F,\perp}_{nuis}(F_X)$ in such a way that $IC(Y \mid Q(F_X, G), G, D_{h(F_X,G)}(\cdot \mid \mu, \rho)) = IC(Y \mid F_X, G)$. In particular, it shows that for an appropriate choice $D_{h_{opt}(F_X,G)}(\cdot \mid \mu, \rho) \in T^{F,\perp}_{nuis}(F_X)$ we have that $IC(Y \mid Q(F_X, G), G, D_{h_{opt}}(F_X, G)(\cdot \mid \mu, \rho))$ equals the efficient influence curve $S^*_{eff}(Y \mid F_X, G)$ of μ at $P_{F_X,G}$.

Theorem 2.6 *Consider the model $\mathcal{M}(CAR)$. Let $IC(Y \mid Q(F_X, G), G, D) \perp T_{CAR}$ and $E(IC(Y \mid Q(F_X, G), G, D) \mid X) = D(X)$ F_X-a.e. for all $D \in T^{F,\perp}_{nuis}$. Let $IC(Y \mid F_X, G)$ be a gradient of μ in the model $\mathcal{M}(CAR)$. We have that*

$$D(X) \equiv E(IC(Y \mid F_X, G) \mid X) \in T^{F,\perp}_{nuis}(F_X)$$

and

$$IC(Y \mid Q(F_X, G), G, D) = IC(Y \mid F_X, G).$$

In particular, if $S^*_{eff}(Y \mid F_X, G)$ is the canonical gradient (i.e., the efficient influence curve) of μ at $P_{F_X,G}$, then

$$D_{opt}(X) = E(S^*_{eff}(Y \mid F_X, G) \mid X) \in T^{F,\perp}_{nuis}(F_X)$$

and

$$IC(Y \mid Q(F_X, G), G, D_{opt}) = S^*_{eff}(Y \mid F_X, G).$$

Equivalently, but in terms of indexes, if $h(F_X, G) \equiv h_{ind,F_X}(E(IC(Y \mid F_X, G) \mid X))$*, then*

$$IC(Y \mid Q(F_X, G), G, D_{h(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X))) = IC(Y \mid F_X, G),$$

and if $h_{opt}(F_X, G) = h_{ind,F_X}(E(S^*_{eff}(Y \mid F_X, G) \mid X))$*, then*

$$IC(Y \mid Q(F_X, G), G, D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X))) = S^*_{eff}(Y \mid F_X, G).$$

Thus, the optimal index $h_{opt}(F_X, G)$ is uniquely identified as the index $h \in \mathcal{H}^F(F_X)$, for which

$$D_h(X \mid \mu(F_X), \rho(F_X)) = E(S^*_{eff}(Y \mid F_X, G) \mid X). \tag{2.50}$$

If the full data model is locally saturated, then $T^{F,\perp *}_{nuis}(F_X) = \{S^{*F}_{eff}(X \mid F_X)\}$ so that the right-hand side of (2.50) equals $S^{*F}_{eff}(X \mid F_X)$ and $h_{opt} = h_{eff}(F_X)$ is the index of the full data canonical gradient.

Since S^*_{eff} is not always trivially computed, we will now provide algorithms for determining the optimal index $h_{opt}(F_X, G)$.

Theorem 2.7 *In this theorem, we will suppress in the notation the dependence on* F_X, G *of Hilbert space operators, Hilbert spaces, index sets, and indexes (such as* h_{eff} *and* h_{opt}*). Let* $A(s) = E(s(X) \mid Y)$ *and* $A^\top(V) = E(V(Y) \mid X)$*. Let* $\mathbf{I} = A^\top A : L^2_0(F_X) \to L^2_0(F_X)$*. Let* $\mathbf{I}^* \equiv \Pi_{T^F}\mathbf{I} : T^F \to T^F$*, where* Π_{T^F} *is the projection operator onto the full data tangent space* T^F *in the Hilbert space* $L^2_0(F_X)$*. It is assumed that both operators* $\mathbf{I}^*$ *and* $\mathbf{I}$ *are 1-1. Assume that the efficient score* $S^F_{eff} = D_{h_{eff}} \in R(\mathbf{I}^*)$*. Let* $h_{ind} : L^2_0(F_X) \to \mathcal{H}^F(F_X)$ *be the index mapping. Then, we have the following representations of* h_{opt}*.*

- *(Robins and Rotnitzky, 1992) Consider the mapping* $B : T^{F,\perp}_{nuis} \cap R(\mathbf{I}) \to T^{F,\perp}_{nuis}$ *defined by*

  $$B(D) = \Pi(\mathbf{I}^{-1}(D) \mid T^{F,\perp}_{nuis}).$$

  Then

  $$h_{opt} = h_{ind}B^{-1}D_{h_{eff}}.$$

  An alternative way to define h_{opt} *is the following. Define* $B' : \mathcal{H}^F(F_X) \to \mathcal{H}^F(F_X)$ *by*

  $$B'(h) = h_{ind}\mathbf{I}^{-1}D_h.$$

  Then

  $$h_{opt} = B'^{-1}h_{eff}.$$

- *(van der Vaart, 1991) We have*

$$h_{opt} = h_{ind}\mathbf{II}^{*-1}D_{h_{eff}}.$$

In full notation, this theorem provides us with the following representations of h_{opt}:

$$h_{opt}(F_X, G) = h_{ind,F_X}\left\{B^{-1}_{F_X,G}D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X), \rho(F_X))\right\}, \quad (2.51)$$

$$h_{opt}(F_X, G) = B'^{-1}_{F_X,G}(h_{eff}(F_X)), \quad (2.52)$$

$$h_{opt}(F_X, G) = h_{ind,F_X}\mathbf{I}_{F_X,G}\mathbf{I}^{*-}_{F_X,G}D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X), \rho(F_X)). \quad (2.53)$$

Thus, we have two mappings, $B^{-1}_{F_X,G}$ and $\mathbf{I}_{F_X,G}\mathbf{I}^{*-1}_{F_X,G}$, mapping $D_{h_{eff}(F_X)}$ into the optimal full data function $D_{h_{opt}(F_X,G)}$. Although, by definition $B^{-1}_{F_X,G}$ maps $T^{F,\perp}_{nuis}(F_X)$ into itself, it might be less obvious to see that $\mathbf{I}_{F_X,G}\mathbf{I}^{*-1}_{F_X,G}$ maps $D_{h_{eff}}$ into an element of $T^{F,\perp}_{nuis}$. We will now give a proof of this fact. We parametrize the projection operator on the full data tangent space as

$$\Pi(D \mid T^F) = \Pi(D \mid \langle D_{h_{eff}}\rangle) + D - \Pi(D \mid T^{F,\perp}_{nuis}).$$

Note that, for any $D \in T^{F,\perp}_{nuis}$, we have

$$D - \Pi(D \mid T^F) \in T^{F,\perp}_{nuis}.$$

Define $D = \mathbf{I}_{F_X,G}\mathbf{I}^{*-}_{F_X,G}(D_{h_{eff}})$. Now write

$$D = \{D - \Pi(D \mid T^F)\} + \Pi(D \mid T^F),$$

and note that the second term equals $D_{h_{eff}} \in T^{F,\perp}_{nuis}$. This proves that $\mathbf{I}_{F_X,G}\mathbf{I}^{*-1}_{F_X,G}D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X), \rho(F_X)) \in T^{F,\perp}_{nuis}(F_X)$. Since $T^{F,\perp}_{nuis}(F_X) = \{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\}$, we have that $\mathbf{I}_{F_X,G}\mathbf{I}^{*-1}_{F_X,G}D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X), \rho(F_X)) = D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X))$ for some $h_{opt}(F_X, G) \in \mathcal{H}^F(F_X)$.

The proof in the preceding paragraph actually shows which arguments in

$$\mathbf{I}_{F_X,G}\mathbf{I}^{*-1}_{F_X,G}D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X), \rho(F_X)) = D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X)) \quad (2.54)$$

determine that the left-hand side of (2.54) is an element of $T^{F,\perp}_{nuis}(F_X)$, and which arguments determine h_{opt}. The $\mu(F_X), \rho(F_X)$ in $D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X), \rho(F_X))$ and in $T^{F,\perp}_{nuis}(F_X) = \{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}^F(F_X)\}$ determine that the left-hand side of (2.54) is an element of $T^{F,\perp}_{nuis}(F_X)$ while 1) (F_X, G) in $\mathbf{I}_{F_X,G}$, 2) F_X in $h_{eff}(F_X)$ and 3) F_X in h_{ind,F_X} determine the index of $\Pi_{F_X}(\cdot \mid T^{\perp}_{nuis}(F_X))$ and the index $h_{opt}(F_X, G)$.

Proof of Theorem 2.7. If $S^{*F}_{eff} \in R(\mathbf{I}^*)$, then by van der Vaart (1991) $S^*_{eff} = A\mathbf{I}^{*-1}(S^{*F}_{eff})$. This can be written as

$$S^*_{eff} = A\mathbf{I}^{-1}(\mathbf{II}^{*-1})(S^{*F}_{eff}),$$

which proves by Theorem 1.3 that $D_{opt} = \mathbf{II}^{*-}(S^{*F}_{eff})$ and proves the second expression for $h_{opt}(F_X, G)$.

We will now prove the first statement. Since $\mathbf{I}^{*-1}(S^{*F}_{eff}) \in \langle S^{*F}_{eff} \oplus T^F_{nuis}(F_X)$, it follows that $\Pi(\mathbf{I}^{*-1}(S^{*F}_{eff}) \mid T^{F,\perp}_{nuis}(F_X)) = S^{*F}_{eff}$, and thus that $D_{opt} = \mathbf{II}^{*-}(S^{*F}_{eff})$ solves

$$\Pi(\mathbf{I}^{-1}(D) \mid T^{\perp}_{nuis}) = S^{*F}_{eff}. \tag{2.55}$$

In addition, by definition, it is an element of $R(\mathbf{I})$ and $\Pi_{T^F} D_{opt} = S^{*F}_{eff}$ so that it is indeed an element of $R(\mathbf{I}) \cap T^{F,\perp}_{nuis}$. We will now show that if $D \in T^{F,\perp}_{nuis} \cap R(\mathbf{I})$ and solves (2.55), then $A\mathbf{I}^{-1}(D) = S^*_{eff}$ and thus $D = D_{opt}$.

Firstly, if $D \in T^{F,\perp *}_{nuis}$, then it follows that $A\mathbf{I}^{-1}(D) \in T^{\perp *}_{nuis}$. Consider the following representation:

$$\mathbf{I}^{-1}(D) = \Pi(\mathbf{I}^{-1}(D) \mid T^F) - \Pi(\mathbf{I}^{-1}(D) \mid S^F_{eff}) + \Pi(\mathbf{I}^{-1}(D) \mid T^{F,\perp}_{nuis}).$$

Thus, (2.55) teaches us that $\mathbf{I}^{-1}(D) \in T^F$. Therefore $A\mathbf{I}^{-1}(D)$ is an element of the tangent space T in $\mathcal{M}(G)$. It is well known that if a gradient is an element of the tangent space, then it equals the canonical gradient. This proves that if $D \in T^{F,\perp *}_{nuis}$ and solves (2.55), then $A\mathbf{I}^{-1}(D) = S^*_{eff}$, which proves the first statement. □

We will provide two examples in which we solve for $h_{opt}(F_X, G)$ using representation (2.51).

Example 2.14 (Current status data structure; continuation of Example 2.6) Consider the current status data structure $Y = (C, \Delta = I(T \leq C), \bar{L}(C))$, where we have for the full data $X = (T, \bar{L})$. Assume as the full data model the univariate linear regression model $T = \beta Z + \epsilon$, where it is assumed that, for a given monotone function K, $E(K(\epsilon(\beta)) \mid Z) = 0$. Lemma 2.1 tells us $T^{F,\perp}_{nuis}(F_X) = \{h(Z)K(\epsilon(\beta)) : h\}$ and

$$\Pi(\mathbf{I}^-_{F_X,G}(D) \mid T^{F,\perp}_{nuis}(F_X)) = \frac{E(\mathbf{I}^{-1}_{F_X,G}(D)(X)K(\epsilon) \mid Z)}{E(K(\epsilon)^2 \mid Z)} K(\epsilon).$$

By (2.51), we have that $h_{opt}(Z)$ is the solution $h(Z)$ of

$$\frac{E(\mathbf{I}^{-1}_{F_X,G}(hK)(X)K(\epsilon) \mid Z)}{E(K(\epsilon)^2 \mid Z)} K(\epsilon) = h_{eff}(Z)K(\epsilon),$$

where $h_{eff}(Z) = Z/E(K(\epsilon)^2 \mid Z)$ is the index of the efficient score S^F_{eff} of β. Since Z is always observed, we have that $\mathbf{I}^-_{F_X,G}(hK)(X) = h(Z)\mathbf{I}^-_{F_X,G}(K)(X)$. Thus, this proves the following representation of h_{opt}:

$$h_{opt}(Z) = \frac{h_{eff}(Z)}{E(\mathbf{I}^{-1}_{F_X,G}(K)(X)K(\epsilon) \mid Z)} E(K(\epsilon)^2 \mid Z).$$

If L is time-independent, then $\mathbf{I}^-_{F_X,G}$ has a simple closed-form expression derived in Example 2.6. However, if L is time-dependent, then this inverse is very involved. Therefore, the following derivation is very useful. Let $\langle f, g\rangle_Z = E(f(X)g(X) \mid Z)$. Using that $A_G^\top$ is also the adjoint of A_{F_X} in the world where one conditions on Z, it follows that

$$\begin{aligned}
\langle \mathbf{I}^-_{F_X,G}(K), K\rangle_Z &= \langle \mathbf{I}^-_{F_X,G}(K), \mathbf{I}_{F_X,G}\mathbf{I}^-_{F_X,G}(K)\rangle_Z \\
&= \langle A_{F_X}\mathbf{I}^-_{F_X,G}(K), A_{F_X}\mathbf{I}^-_{F_X,G}(K)\rangle_Z \\
&= \langle IC(\cdot \mid Q(F_X,G), G, K), IC(\cdot \mid Q(F_X,G), G, K)\rangle_Z \\
&= E\left\{IC(Y \mid Q(F_X,G), G, K)^2 \mid Z\right\}.
\end{aligned}$$

This proves the representation of h_{opt}

$$h_{opt}(Z) = \frac{ZE(K(\epsilon)^2 \mid Z)}{E\left\{IC(Y \mid Q(F_X,G), G, K)^2 \mid Z\right\}},$$

where the denominator is actually straightforward to estimate by regressing $IC(Y_i \mid Q_n, G_n, K)$ onto Z_i, $i = 1, \ldots, n$. □

Example 2.15 (Generalized linear regression with missing covariate; continuation of Example 1.1) We refer to Example 1.1 in Chapter 1 for a description of this example and the optimal mapping (1.24) $IC(\cdot \mid Q(F_X), G, D) = \frac{D(X)\Delta}{\Pi(W)} - (\Delta - \Pi(W))\frac{E(D(X)|W)}{\Pi(W)}$. Thus, we have for the observed data $Y = (\Delta, \Delta X + (1-\Delta W))$, $W \subset X$, and we consider a univariate generalized linear regression model $Z = g(X^* \mid \beta) + \epsilon$, where $E(K(\epsilon) \mid X^*) = 0$. Here $Z \subset W$ is always observed, while $X^* \subset X$ has missing components. We have $T^{F,\perp}_{nuis}(F_X) = \{h(X^*)K(\epsilon) : h\}$ and $\Pi(D \mid T^{F,\perp}_{nuis}(F_X)) = \frac{E(D(X)K(\epsilon)|X^*)}{E(K^2(\epsilon)|X^*)}K(\epsilon)$. Our first goal is to determine a closed-form expression for $\mathbf{I}^{-1}_{F_X,G} : L^2_0(F_X) \to L^2_0(F_X)$. By Theorem 1.3, we have

$$\begin{aligned}
A_{F_X}\mathbf{I}^{-1}_{F_X,G}(D) &= IC(\cdot \mid Q(F_X), G, D) \\
&= \frac{D(X)\Delta}{\Pi(W)} - (\Delta - \Pi(W))\frac{E(D(X) \mid W)}{\Pi(W)}.
\end{aligned}$$

By definition of A_{F_X}, we have

$$A_{F_X}\mathbf{I}^{-1}_{F_X,G}(D) = \mathbf{I}^-_{F_X,G}(D)\Delta + E(\mathbf{I}^-_{F_X,G}(D)(X) \mid W)(1-\Delta).$$

Combining these two identities yields

$$\mathbf{I}^-_{F_X,G}(D) = \frac{D(X)}{\Pi(W)} + \left(1 - \frac{1}{\Pi(W)}\right)E(D(X) \mid W).$$

Thus $\Pi(\mathbf{I}^{-}_{F_X,G}(D_{h_{opt}}) \mid T^{F,\perp}_{nuis}(F_X)) = D_{h_{eff}}$ translates into

$$\frac{E(\{D_{h_{opt}}(X)/\Pi(W) + E(D_{h_{opt}}(X) \mid W) - 1\}K(\epsilon) \mid X^*)}{E(K^2(\epsilon) \mid X^*)}K(\epsilon)$$

$$= h_{eff}(X^*)K(\epsilon).$$

Since $D_{h_{opt}}(X) = h_{opt}(X^*)K(\epsilon)$, this reduces to

$$h_{opt}(X^*)E(K^2(\epsilon)/\Pi(W) \mid X^*) + E\left(E(h_{opt}(X^*)K(\epsilon) \mid W)K(\epsilon) \mid X^*\right)$$

$$= h_{eff}(X^*)E(K^2(\epsilon) \mid X^*) + E(K(\epsilon) \mid X^*).$$

Thus, the function $x^* \to h_{opt}(x^*)$ is the solution of an integral equation, first derived in Robins, Rotnitzky, Zhao (1994). □

In the current status data example above, we made use of the following general lemma for any censored data model.

Lemma 2.5 *Let $Y = \Phi(X, C)$, $X \sim F_X$, $C \mid X \sim G(\cdot \mid X)$ and assume that the conditional distribution G satisfies CAR. Consider the nonparametric information operator $I_{F_X,G} : L^2_0(F_X) \to L^2_0(F_X)$ defined by $I_{F_X,G}(s) = A^{\top}_G A_{F_X}(s) = E_G E_{F_X}(s(X) \mid Y) \mid X)$. Let $X^* \subset X$ and $X^* \subset Y$ (i.e., X^* is part of the full data structure and is always observed). Then, for any pair of functions $D_1, D_2 \in L^2_0(F_X)$ in the range of $I_{F_X,G}$, we have*

$$E(I^{-1}_{F_X,G}(D_1)D_2 \mid X^*) = E(A_{F_X} I^{-1}_{F_X,G}(D_1) A_{F_X} I^{-1}_{F_X,G}(D_2) \mid X^*).$$

Proof. In the Hilbert space with X^* fixed, we need to prove

$$\langle I^{-1}(D_1), D_2\rangle = \langle AI^{-1}(D_1), AI^{-1}(D_2)\rangle.$$

Since $A^{\top}$ is still the adjoint of A conditional on X^*, moving the first A to the other side and noting that $A^{\top}AI^{-1}$ is the identity operator gives the desired result. □

The closed-form representation of the optimal full data index h_{opt} in the current status example above can be generalized to general censored data structures with the full data model being the multivariate generalized linear regression model with covariates always observed.

Theorem 2.8 *Let $Y = \Phi(X, C)$, $X \sim F_X$, $C \mid X \sim G(\cdot \mid X)$ and assume that the conditional distribution G satisfies CAR. Let the full data model be the p-variate generalized regression model $Z = g(X^* \mid \beta) + \epsilon$, $\beta \in \mathbb{R}^k$, $E(K(\epsilon) \mid X^*) = 0$, and $K(\epsilon) = (K(\epsilon_1), \ldots, K(\epsilon_p))$. We refer to Lemma 2.1 for 1) the orthogonal complement of the nuisance tangent space in the full data model given by $\{D_h(X) = h(X^*)K(\epsilon) : h\ 1 \times p$ vector$\}$, 2) the index h_{eff} ($k \times p$ matrix) so that $h_{eff}(X^*)K(\epsilon)$ is the efficient score of β, and*

3) the index mapping given, for a function $D \in L_0^2(F_X)^k$*, by*

$$h_{ind,F_X}(D) \equiv E(\{D(X)-E(D(X) \mid X^*)\}K(\epsilon)^\top \mid X^*)E(K(\epsilon)K(\epsilon)^\top \mid X^*)^{-1}.$$

Note that $h_{ind,F_X}(D)$ *is an* $k \times p$ *matrix function of* X^**.*

Assume that X^* *is always observed (i.e.,* X^* *is a component of the full data* X*, which is also a function of* Y*). Consider the nonparametric information operator* $I_{F_X,G} : L_0^2(F_X) \to L_0^2(F_X)$ *defined by* $I_{F_X,G}(s) = A_G^\top A(s) = E_G E_{F_X}(s(X) \mid Y) \mid X)$*. Define* $IC(Y \mid F_X, G, D) = A_{F_X} I_{F_X,G}^{-1}(D)$*. We have that the* $p \times p$ *matrix* $\left\{h_{ind,F_X} I_{F_X,G}^{-1} K\right\}$ *is given by*

$$E(IC(Y \mid F_X, G, K)IC(Y \mid F_X, G, K)^\top \mid X^*)E(K(\epsilon)K(\epsilon)^\top \mid X^*)^{-1}.$$

If

$$h_{opt}(X^*) \equiv h_{eff}(X^*) \left\{h_{ind,F_X} I_{F_X,G}^{-1} K\right\}_{p\times p}^{-1}, \tag{2.56}$$

where we assume that the inverse exists a.e., then $IC(Y \mid F_X, G, D_{h_{opt}})$ *equals the efficient influence curve for* β*.*

This is clearly an important result since it allows us to compute closed form locally efficient estimators of regression parameters in univariate and multivariate generalized linear regression models under any type of censoring of the full data structure as long as the covariates X^* in the regression model are always observed.

2.8.1 Finding the optimal estimating function among a given class of estimating functions

Consider a class of estimating functions $IC(Y \mid Q, G, D_h(\cdot \mid \mu, \rho))$, $h \in \mathcal{H}^F$. If for all $F_X \in \mathcal{M}^F$ and $G \in \mathcal{G}$ $\{IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu(F_X), \rho(F_X, G))) : h \in \mathcal{H}^F\} \subset T_{nuis}^\perp(P_{F_X,G})$ in model $\mathcal{M}(\mathcal{G})$, then $c^{-1}IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu, \rho))$ denotes the influence curve corresponding with the estimating function $IC(Y \mid Q, G, D_h(\cdot \mid \mu, \rho))$, assuming correct specification of the nuisance parameters Q, G. Theorem 2.9 below, based on Newey and McFadden (1994), provides us with a formula identifying the estimating function whose corresponding influence curve has minimal variance among the class of estimating functions.

Theorem 2.9 *Consider the censored data structure* $Y = \Phi(C, X)$*,* $X \sim F_X \in \mathcal{M}^F$*, and* $C \mid X \sim G(\cdot \mid X) \in \mathcal{G} \subset \mathcal{G}(CAR)$*. Let* μ *be a* k*-dimensional real-valued parameter of* F_X*. Consider a class of* k*-dimensional full data structure estimating functions* $D_h(\cdot \mid \mu, \rho)$ *for* μ *indexed by* h *ranging over a set* $\mathcal{H}^k$ *that satisfies for all* $F_X \in \mathcal{M}^F$ *and all* $G \in \mathcal{G}$ $T_{nuis}^{F,\perp}(F_X) \supset \{D_h(\cdot \mid \mu(F_X), \rho(F_X, G)) : h \in \mathcal{H}\}$*. Let* $(H, \langle \cdot, \cdot \rangle_H)$ *be a Hilbert space defined by*

the closure of the linear span of $\mathcal{H}$, which is assumed to be embedded in a Hilbert space with inner product $\langle \cdot, \cdot \rangle_H$.

Consider a class of k-dimensional observed data estimating functions $IC(Y \mid Q, G, D_h)$ indexed by $h \in \mathcal{H}^k$, where $IC(Y \mid Q, G, D_h) = (IC(Y \mid Q, G, D_{h_1}), \ldots, IC(Y \mid Q, G, D_{h_k}))$. Here $Q = Q(F_X, G)$ and G denote the true parameter values for Q and G. Define for $h \in \mathcal{H}^k$

$$\kappa(h) = -\frac{d}{d\mu} E_{P_{F_X,G}} IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu, \rho(F_X, G)))\Big|_{\mu=\mu(F_X)},$$

where we assume that this $k \times k$ derivative matrix is well-defined. Assume that $\kappa(\cdot)$ is bounded and linear, so that (by the Riesz Representation theorem) there exists $h^* \in H^k$ so that for all $h \in \mathcal{H}^k$

$$\kappa(h)_{ij} = \langle h_i^*, h_j \rangle_H, (i, j) \in \{1, \ldots, k\}^2.$$

For notational convenience, define $\tilde{A} : (H, \langle \cdot, \cdot \rangle_H) \to L_0^2(P_{F_X,G})$ by $\tilde{A}(h) \equiv IC(\cdot \mid Q(F_X, G), G, D_h)$. Let $\tilde{A}^\top : L_0^2(P_{F_X,G}) \to (H, \langle \cdot, \cdot \rangle_H)$ be the adjoint of $\tilde{A}$. By applying these operators to each component of a multivariate function, we can also define these operators on multivariate functions so that $\tilde{A} : H^k \to L_0^2(P_{F_X,G})^k$ and $\tilde{A}^\top : L_0^2(P_{F_X,G})^k \to H^k$. Define

$$\Sigma(h) \equiv E(\kappa(h)^{-1}\tilde{A}(h)(Y)(\kappa(h)^{-1}\tilde{A}(h)(Y))^\top).$$

Assume that h^* is an element of the range of $\tilde{A}^\top\tilde{A} : H^k \to H^k$ and that $(\tilde{A}^\top\tilde{A}) : H \to H$ is 1-1. Then

$$h_{opt} \equiv \min^{-1}{}_{h \in H^k} c^\top \Sigma(h) c \text{ for all } c \in \mathbb{R}^k$$

exists, and is given by

$$h_{opt} = (\tilde{A}^\top\tilde{A})^-(h^*).$$

Thus

$$\tilde{A}(h_{opt}) = \tilde{A}(\tilde{A}^\top\tilde{A})^{-1}h^*.$$

Proof. Firstly, note that

$$\kappa(h) = \langle h^*, h \rangle_H = \langle \tilde{A}(\tilde{A}^\top\tilde{A})^{-1}(h^*), \tilde{A}(h) \rangle_{P_{F_X,G}},$$

where the inner product $\langle h_1, h_2 \rangle_H$ is defined as the matrix with (i, j)th element $\langle h_{1i}, h_{2j} \rangle_H$. Using this notation for inner products of vectors, it follows that

$$\begin{aligned}
&c^\top E(\kappa(h)^{-1}\tilde{A}(h)(Y)(\kappa(h)^{-1}\tilde{A}(h)(Y))^\top)c \\
&= c^\top \kappa(h)^{-1}\langle \tilde{A}(h), \tilde{A}(h) \rangle_{P_{F_X,G}} \kappa(h)^{-1\top} c \\
&= c^\top D(\tilde{A}(h))\langle \tilde{A}(h), \tilde{A}(h) \rangle_{P_{F_X,G}} D(\tilde{A}(h))^\top c,
\end{aligned}$$

where $D(\tilde{A}(h)) = \langle \tilde{A}(\tilde{A}^\top \tilde{A})^{-1}(h^*), \tilde{A}(h)\rangle^{-1}_{P_{F_X,G}}$. By the Cauchy–Schwarz inequality, this expression in $\tilde{A}(h)$ is minimized by $\tilde{A}(h_{opt}) = \tilde{A}(\tilde{A}^\top \tilde{A})^{-1}h^*$. □

A special case of Theorem 2.9 is obtained by letting the index set $\mathcal{H}$ be the class of full data estimating functions itself. Application of this theorem to our T_{CAR}-orthogonalized mapping from full data structure estimating functions to observed data estimating functions results in the formula for the efficient influence curve (and efficient score) as presented in Theorem 1.3 and originally derived in Robins, and Rotnitzky (1992). However, note the next theorem is more general since it can be applied to any mapping from full-data structure estimating functions to observed data estimating functions.

Theorem 2.10 *Consider the censored data structure $Y = \Phi(C, X)$, $X \sim F_X \in \mathcal{M}^F$, and $C \mid X \sim G(\cdot \mid X) \in \mathcal{G} \subset \mathcal{G}(CAR)$. Let μ be a k-dimensional real-valued parameter of F_X. Consider a class of k-dimensional full data structure estimating functions $D_h(\cdot \mid \mu, \rho)$ for μ indexed by h ranging over a set $\mathcal{H}^k$ that satisfies for all $F_X \in \mathcal{M}^F$ and $G \in \mathcal{G}$ $T^{F,\perp}_{nuis}(F_X) \supset \{D_h(\cdot \mid \mu(F_X), \rho(F_X, G)) : h \in \mathcal{H}\}$. Let $(H, \langle\cdot,\cdot\rangle_{F_X}) \subset L^2_0(F_X)$ be the sub-Hilbert space of $L^2_0(F_X)$ defined by the closure of the linear span of $\{D_h(\cdot \mid \mu(F_X), \rho(F_X)) : h \in \mathcal{H}\}$. Consider a class (not necessarily T_{CAR}-orthogonalized) of k-dimensional observed data estimating functions $IC(Y \mid Q, G, D_h(\cdot \mid \mu, \rho))$ indexed by $h \in \mathcal{H}^k$ with nuisance parameters $Q(F_X, G)$, G and $\rho(F_X, G)$.*

Define for $h \in \mathcal{H}^k$

$$\kappa(D_h) = -\frac{d}{d\mu} E_{P_{F_X,G}} IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu, \rho(F_X, G)))\Big|_{\mu=\mu(F_X)},$$

where we assume that this $k \times k$ derivative matrix is well-defined. Assume that $\kappa : (H, \langle\cdot,\cdot\rangle_{F_X}) \to \mathbb{R}$ is bounded and linear, so that (by the Riesz Representation theorem) there exists $D^* \in H^k$ so that for all $h \in \mathcal{H}^k$

$$\kappa(D_h) = \langle D_h, D^*\rangle_{F_X} \equiv E(D_h(X)D^{*\top}(X)).$$

For notational convenience, define $\tilde{A} : (H, \langle\cdot,\cdot\rangle_H) \to L^2_0(P_{F_X,G})$ by $\tilde{A}(D) \equiv IC(\cdot \mid Q(F_X, G), G, D)$. Let $\tilde{A}^\top : L^2_0(P_{F_X,G}) \to (H, \langle\cdot,\cdot\rangle_H)$ be the adjoint of A. By applying these operators to each component of a multivariate function, we can also define these operators on multivariate functions so that $\tilde{A} : H^k \to L^2_0(P_{F_X,G})^k$ and $\tilde{A}^\top : L^2_0(P_{F_X,G})^k \to H^k$. Define the covariance matrix

$$\Sigma(D) \equiv E(\kappa(D)^{-1}A(D)(Y)(\kappa(D)^{-1}A(D)(Y))^\top).$$

Assume that D^* is an element of the range of $\tilde{A}^\top\tilde{A} : H^k \to H^k$ and that $\tilde{A}^\top\tilde{A} : H \to H$ is 1-1. Then

$$D_{opt} \equiv \min_{D \in H^k}{}^{-1} c^\top \Sigma(D) c \text{ for all } c \in \mathbb{R}^k$$

exists, and is given by

$$D_{opt} = (\tilde{A}^\top \tilde{A})^{-1}(D^*).$$

Thus

$$\tilde{A}(D_{opt}) = \tilde{A}(\tilde{A}^\top \tilde{A})^{-1} D^*.$$

Remark:

Suppose that we apply this theorem to the T_{CAR}-optimal mapping into observed data estimating functions satisfying $IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu(F_X), \rho(F_X, G))$ satisfies for all D_h $E_G(IC(Y \mid Q(F_X, G), G, D_h) \mid X) = D_h(X)$, $IC(Y \mid Q(F_X, G), G, D_h(\cdot \mid \mu(F_X), \rho(F_X, G)) \perp T_{CAR}(P_{F_X,G})$, and $\{D_h(\cdot \mid \mu(F_X), \rho(F_X, G)) : h \in \mathcal{H}\} = T_{nuis}^{F,\perp}(F_X)$. Then $\tilde{A}(\tilde{A}^\top \tilde{A})^{-1}(D^*)$ has to equal the efficient influence curve $S_{eff}^*(\cdot \mid P_{F_X,G}) = A_{F_X} I_{F_X,G}^{*-1}(S_{eff}^{F*})$ for μ at $P_{F_X,G}$, Off course, this identity is a consequence of (by Theorem 1.3) the fact that in this case $\tilde{A}(D_h) = AI^{-1}D_h$. It follows that $D^* = S_{eff}^{*F}$ and $D_{opt} = (\tilde{A}^\top \tilde{A})^{-1}(D^*)$ corresponds with the optimal full data function defined by $D_{opt} = I_{F_X,G} I_{F_X,G}^{*-1}(S_{eff}^{*F})$ or the solution in $T_{nuis}^{F,\perp}$ satisfying $\Pi(I_{F_X,G}^{-1}(D) \mid T_{nuis}^{F,\perp}) = S_{eff}^{*F}$.

Closed-form optimal estimating functions for MGLM

We will now apply Theorem 2.9 to obtain a closed-form representation of the optimal estimating function among a general class (not necessarily the class of all estimating functions including the efficient influence curve as in Theorem 2.8) of estimating functions in a multivariate generalized linear regression model with covariates always observed. In such models, classes of estimating functions obtained by mapping full data estimating functions $h(X^*)K(\epsilon(\beta))$ into observed data estimating functions $IC(Y \mid Q, G, D_h)$ will have the property that $IC(Y \mid Q, G, D_h) = h(X^*)IC(Y \mid Q, G, K_\beta)$, where $K_\beta(X) = K(\epsilon(\beta))$. This special property combined with Theorem 2.9 results in a closed-form representation of the optimal estimating function. The result is a generalization of Theorem 2.8 since it specifies the optimal estimating function among any given class of estimating functions, not necessarily the class including the efficient influence curve and/or a T_{CAR}-orthogonalized class of estimating functions. For example, if one uses non optimal mappings only orthogonalizing w.r.t. a subspace of T_{CAR}, as provided in this chapter, then Theorem 2.11 below can be used to find the optimal choice.

Theorem 2.11 *Let $Y = \Phi(X, C)$, $X \sim F_X$, $C \mid X \sim G(\cdot \mid X)$, and assume that the conditional distribution G satisfies CAR. Let the full data model be the p-variate generalized regression model $Z = g(X^* \mid \beta) + \epsilon$, $\beta \in \mathbb{R}^k$, $E(K(\epsilon) \mid X^*) = 0$, and $K(\epsilon) = (K(\epsilon_1), \ldots, K(\epsilon_p))$. Assume that $g(X^* \mid \beta)$ is differentiable in β for each possible X^* and that K is differentiable. Consider the full data estimating functions $\{D_h(X \mid \beta) =$*

$h(X^*)^\top K(\epsilon(\beta)) : h \in \mathcal{H}\}$, where $\mathcal{H}$ is an index set of $1 \times p$ vector-valued functions of X^*. Consider a class of k-dimensional observed data estimating functions $IC(Y \mid Q, G, D_h(\cdot \mid \mu, \rho))$ with unknown nuisance parameters $Q(F_X, G), G, \rho(F_X, G)$ indexed by $h \in \mathcal{H}^k$, where h denotes a $k \times p$ matrix valued function of X^*. Suppose that for each $k \times p$ matrix-valued function $h \in \mathcal{H}^k$ of X^*,

$$IC(Y \mid Q, G, D_h(\cdot \mid \beta)) = h(X^*)IC(Y \mid Q, G, K_\beta).$$

Here $K_\beta(X) = K(\epsilon(\beta))$. For each $h \in \mathcal{H}^k$, we define the $k \times k$ derivative matrix $\kappa(h)$:

$$\kappa(h) = -E(d/d\beta IC(Y \mid Q, G, D_h)) = -E(h(X^*)d/d\beta IC(Y \mid Q, G, K_\beta)).$$

For each $h \in \mathcal{H}^k$, let $\Sigma(h) = E(\kappa(h)^{-1}IC(Y \mid Q, G, D_h)(\kappa(h)^{-1}IC(Y \mid Q, G, D_h)^\top)$ be the covariance matrix of $\kappa(h)^{-1}_{k\times k}IC(Y \mid Q, G, D_h)$. Let H be the sub-Hilbert space defined by the closure of the linear span of $\mathcal{H}$ w.r.t. the norm in $L_0^2(F_X^*)$. Consider

$$h^* = -E(d/d\beta IC(Y \mid K_\beta) \mid X^*)_{k\times p}E(IC(Y \mid K_\beta)IC(Y \mid K_\beta)^\top \mid X^*)^{-1},$$

where we used shorthand notation $IC(Y \mid K_\beta) = IC(Y \mid Q, G, K_\beta)$. Assume h^* is well defined (i.e., derivative and inverse exist) and $h^* \in H^k$. Then

$$h_{opt} \equiv \min_{h\in H^k}{}^{-1} c^\top \Sigma(h)c \text{ for all } c \in \mathbb{R}^k$$

exists, and is given by $h_{opt} = h^*$.

This theorem is a straightforward consequence of Theorem 2.9. In a typical application of this theorem one would have that $H = L_0^2(F_X^*)$ so that the last condition holds.

Example 2.16 (Optimal IPTW estimating function) Let $A(t)$ represent a time-dependent treatment process that potentially changes value at a finite prespecified set of points. Let $\mathcal{A}$ be the set of possible sample paths of A, where we assume that $\mathcal{A}$ is finite. For each possible treatment regime $\bar{a}$, we define $X_{\bar{a}}(t)$ as the data that one would observe on the subject if, possibly contrary to the fact, the subject had followed treatment regime $\bar{a}$. It is natural to assume that $X_{\bar{a}}(t) = X_{\bar{a}(t-)}$ (i.e., the counterfactual outcome at time t is not affected by treatment given after time t). One refers to $X_{\bar{a}} = (X_{\bar{a}}(t) : t)$ as counterfactual. Suppose that $X_{\bar{a}} = (Z_{\bar{a}}, L_{\bar{a}})$ consists of an outcome process $Z_{\bar{a}}$ and covariate process $L_{\bar{a}}$. The baseline covariates are included in $L_{\bar{a}}(0) = L(0)$. Let $X = (X_{\bar{a}}, \bar{a} \in \mathcal{A})$ be the full data structure, and the observed data structure is given by

$$Y = (\bar{A}, X_{\bar{A}}) = (\bar{A}, Z_{\bar{A}}, L_{\bar{A}}).$$

Let $Y_{\bar{a}}^*$ be a counterfactual outcome of interest such as $Z_{\bar{a}}(\tau)$ at an endpoint τ. Consider a marginal structural generalized linear regression model:

$$Y_{\bar{a}}^* = m(\bar{a}, V \mid \beta) + \epsilon_{\bar{a}}, \text{ where } E(\epsilon_{\bar{a}} \mid V) = 0,$$

where V is a subset of the baseline covariates and $m(\bar{a}, V \mid \beta)$ denotes a parametrization of the conditional mean of $Y^*_{\bar{a}}$, given V, parametrized by the parameter of interest β. Assume that $g(\bar{a} \mid X)$ satisfies the SRA (i.e., $P(A(t) = a(t) \mid \bar{A}(t-), X) = P(A(t) = a(t) \mid \bar{A}(t-), \bar{X}_{\bar{a}}(t))$). Consider the class of IPTW estimating functions

$$\left\{ IC(Y \mid G, D_h) = \frac{h(\bar{A}, V)\epsilon(\beta)}{g(\bar{A} \mid X)} : h \right\}$$

indexed by real-valued functions h of $\bar{A}, V$, where $\epsilon(\beta) = Y^* - m(\bar{A}, V \mid \beta)$ is the observed residual. An interesting choice of h is the h that gives an optimal covariance matrix $E(IC(Y \mid G, D_h)IC(Y \mid G, D_h)^\top)$ when g is known. In the same way as one proves Theorem 2.11, application of Theorem 2.9 teaches us that this optimal index is given by

$$h^*(\bar{A}, V) = \frac{d}{d\beta} m(\bar{A}, V \mid \beta) \frac{E(1/g(\bar{A} \mid X) \mid \bar{A}, V)}{E(\epsilon^2(\beta)/g^2(\bar{A} \mid X) \mid \bar{A}, V)}.$$

It is interesting to compare this choice with the computationally simple choice recommended in Robins (1999), given by

$$h(\bar{A}, V) = \frac{g^*(\bar{A} \mid V) d/d\beta m(\bar{A}, V \mid \beta)}{E(\epsilon^2(\beta) \mid \bar{A}, V)},$$

where $g^*(\bar{A} \mid V)$ is the conditional density of $\bar{A}$, given V. Note that both choices reduce to the optimal weighted least squares estimating function that is optimal among all estimating functions whose weights only depend on $\bar{A}, V$ in the situation where $g(\bar{A} \mid X) = g(\bar{A} \mid V)$ is only a function of V. Robins (1999, Section 4.1) provides the efficient choice $h_{opt}(\bar{A}, V)$ for this model in the special case where A is time independent. However, $h_{opt}(\bar{A}, V)$ is the solution to a Fredholm integral equation of the second kind which does not admit a closed form solution. Thus an easily computed alternative, although less efficient, is useful.

□

An algorithm for evaluating the representations of the optimal full data structure function

Consider the representations $\mathbf{I}_{F_X,G}\mathbf{I}^{*-}_{F_X,G}(D_{h_{eff}})$ and $B^{-1}_{F_X,G}(D_{h_{eff}})$ for $D_{h_{opt}}$. The following two lemmas prove that under reasonable conditions these inverses exist and provide general simple algorithms for determining them.

Lemma 2.6 *For notational convenience, in this lemma we suppress the dependence on (F_X, G) of the Hilbert space operators. Let $\mathbf{I}^* = \Pi_{T^F}\mathbf{I} : T^F \to T^F$ be the information operator. Assume that for all $h \in T^F$ with $\| h \|_{F_X} > 0$ we have $\| A(h) \|_{P_{F_X,G}} > 0$. Then $\mathbf{I}^*$ is 1–1.*

Suppose that for some $\delta > 0$

$$\inf_{\|h\|_{F_X}=1, h \in T^F} \| A(h) \|^2_{P_{F_X,G}} \geq \delta. \tag{2.57}$$

Then $\inf_{\|h\|_F=1} \| \mathbf{I}^*(h) \|_{F_X} \geq \delta$, $(I - \mathbf{I}^*)$ *has operator norm bounded by* $1 - \delta$, $\mathbf{I}^*$ *is onto and has bounded inverse with operator norm smaller than or equal to* $1/\delta$, *and its inverse is given by*

$$\mathbf{I}^{*-1} = \sum_{i=0}^{\infty} (I - \mathbf{I}^*)^i.$$

In addition, the following algorithm converges to $\mathbf{I}^{*-1}(f)$*: Set* $k = 0$

$$\begin{aligned} h^0 &= 0, \\ h^{k+1} &= f - \mathbf{I}^*(h^k) + h^k, \end{aligned}$$

and iterate until convergence. The convergence rate is bounded by

$$\| h^k - \mathbf{I}^{*-}(f) \|_{F_X} \leq \frac{(1-\delta)^k}{\delta} \| f \|_{F_X} .$$

If $\Delta = I(X$ *observed*$)$ *and* $\inf_x P(\Delta = 1 \mid X = x) > 0$, *then the condition for bounded invertibility above holds with* $\delta \geq \inf_x P(\Delta = 1 \mid X = x)$. *Finally, we note that*

$$\begin{aligned} \mathbf{I}^*(h) &= \Pi(\mathbf{I}(h) \mid < S^F_{eff} > \oplus T^F_{nuis}) \\ &= \Pi(\mathbf{I}(h) \mid < S^F_{eff} >) + \{\mathbf{I}(h) - \Pi(\mathbf{I}(h) \mid T^{F,\perp}_{nuis})\}. \end{aligned}$$

The condition $\inf_x P(\Delta = 1 \mid X = x) > 0$ is not a necessary condition for the bounded invertibility of $\mathbf{I}^*$. This lemma does not prove that the Neumann series converges if $\| A(h) \| > 0$ for all $h \neq 0$ (thus I^* is 1-1), but $\inf_{\|h\|=1} \| A(h) \|^2 = 0$. We conjecture that if $D_{h_{eff}} \in R(\mathbf{I}^*)$ and I^* is 1-1, then the Neumann series applied to $D_{h_{eff}}$ will converge to $\mathbf{I}^{*-}(S^{*F}_{eff})$. We feel (based on our empirical findings) comfortable recommending this algorithm in practice.

Proof. This lemma can be found in van der Laan (1998) except the actual convergence rate. The convergence rate is proved as follows. Firstly, we have that if $\| \mathbf{I}^*(h) \|_F \geq \delta \| h \|_F$, then

$$\| \mathbf{I}^*(h^k - \mathbf{I}^{*-}(f)) \| \geq \delta \| h^k - \mathbf{I}^{*-}(f) \| .$$

This proves that

$$\| h^k - \mathbf{I}^{*-}(f) \| \leq 1/\delta \| f - \mathbf{I}^*(h^k) \| .$$

Now, note that $f - \mathbf{I}^*(h^k) = h^{k+1} - h^k$. We have

$$h^{k+1} - h^k = (I - \mathbf{I}^*)(h^k - h^{k-1}) = (I - \mathbf{I}^*)^k (h^1 - h^0) = (I - \mathbf{I}^*)^k (f).$$

Since $I - \mathbf{I}^*$ is a contraction with operator norm $1 - \delta$, this shows that

$$\| h^k - \mathbf{I}^{*-}(f) \| \leq \frac{1}{\delta}(1-\delta)^k \| f \| .\square$$

Similarly, one proves the following algorithm for inverting $B_{F_X,G}$.

Lemma 2.7 *Consider the operator $B_{F_X,G} : T^{\perp}_{nuis}(F_X) \to T^{\perp}_{nuis}(F_X)$ defined by $B(D) = \Pi(\mathbf{I}^{-}_{F_X,G}(D) \mid T^{\perp}_{nuis}(F_X))$. We have that with $D \in T^{F,\perp}_{nuis}(F_X)$ and $D_1 \equiv \mathbf{I}^{-}_{F_X,G}(D)$,*

$$\begin{aligned} B_{F_X,G}(D) &= \Pi(\mathbf{I}^{-}(D) \mid T^{F,\perp}_{nuis}(F_X)) \\ &= D_1 - \Pi(D_1 - \mathbf{I}_{F_X,G}(D_1) \mid T_{nuis}(F_X)) \\ &\equiv D_1 - B_{1,F_X,G}(D_1). \end{aligned}$$

Thus $B^{-1}_{F_X,G}(f) = \mathbf{I}_{F_X,G}(I - B_{1,F_X,G})^{-1}(f)$. If $\mathbf{I}_{F_X,G} : L^2_0(F_X) \to L^2_0(F_X)$ is 1–1, then for all $D_1 \in L^2_0(F_X)$ $\| B_{1,F_X,G}(D_1) \|_{F_X} < \| D_1 \|_{F_X}$. If $B_{1,F_X,G} : L^2_0(F_X) \to L^2_0(F_X)$ has operator norm strictly smaller than 1, for example (2.57) holds, then

$$(I - B_{1,F_X,G})^{-1} = \sum_{k=0}^{\infty} B^k_{1,F_X,G}$$

and $(I - B_{1,F_X,G})(D_1) = f$ can be solved by successive substitution:

$$D_1^{k+1} = f + B_{1,F_X,G}(D_1^k).$$

In many examples, it is possible to invert $B_{F_X,G} : T^{F,\perp}_{nuis}(F_X) \to T^{F,\perp}_{nuis}(F_X)$ explicitly. In other examples, solving $B_{F_X,G}(D) = D_{h_{eff}}$ results in integral equations for which particular algorithms are available. Lemma 2.7 above shows, in particular, that the operator $B_{F_X,G}$ is invertible. Therefore, another sensible strategy is to approximate $B'_{F_X,G} : \mathcal{H}^F \to \mathcal{H}^F$ (2.53) by a square matrix mapping index vectors (identifying the index h of the element in $T^{\perp}_{nuis}(F_X)$) into index vectors and invert this matrix using matrix inversion routines. The advantage of the general algorithm of Lemma 2.6 is that it never requires more than being able to apply and store one matrix identifying $I^*_{F_X,G}$. The same remark holds for the algorithm of Lemma 2.7. The algorithms for $\mathbf{I}_{F_X,G}\mathbf{I}^{*-}_{F_X,G}(D_{h_{eff}})$ and $B^{-1}_{F_X,G}(D_{h_{eff}})$ are of similar complexity so that there is little reason to prefer one above the other.

2.9 Estimation of the Optimal Index

Given the estimates $Q_n, G_n, \mu^0_n, \rho_n$, the best estimating function for μ is $IC(Y \mid Q_n, G_n, D_{h_{opt}(F_X,G)}(\cdot \mid \mu, \rho_n))$, where we provided representations of the optimal index $h_{opt}(F_X, G)$ in the preceding sections. In this section, we propose a representation of h_{opt} which naturally provides an estimator h_n of h_{opt}.

In the previous section, we obtained representations h_{opt} of the form h_{ind,F_X} (the index mapping) applied to the optimal full data estimating function $D_{opt}(X \mid F_X, G)$ as a function of F_X and G. However, we know that $D_{opt}(X \mid F_X, G) = D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X))$. Therefore, we are now concerned with establishing a parametrization of D_{opt} in terms of its index h and (μ, ρ) so that we can estimate it by substitution.

In general, an estimator h_n of h_{opt} proceeds as follows. Firstly, we construct an estimate $D_n(X)$ of the optimal full data function $D_{h_{opt}(F_X,G)}(X \mid \mu, \rho)$. If D_n is of the form $D_{h_n}(\cdot \mid \mu_n^0, \rho_n)$, then we also have obtained as estimate h_n of h_{opt}. However, if D_n is based on an implicit representation of D_{opt} involving an inverse B^{-1} or $\mathbf{I}^{*-}$, then estimation of this inverse using a truncated Neumann series representation can result in an estimate $D_n \notin \mathcal{D}(\mu_n^0, \rho_n)$. In that case we need, as discussed below, to project D_n into $\mathcal{D}(\mu_n^0, \rho_n)$ to obtain an estimate h_n of h_{opt}.

To be specific, let $\mathcal{L}(\mathcal{X})$ be a subspace of all pointwise well-defined functions of X with finite supremum norm. We call any function $h : \mathcal{L}(\mathcal{X}) \to \mathcal{H}^{\mathcal{F}}$ an index mapping, and $h_{ind,F_X} : \mathcal{L}(\mathcal{X}) \to \mathcal{H}^{\mathcal{F}}$ is the true index mapping defined by

$$D_{h_{ind,F_X}(D)}(\cdot \mid \mu(F_X), \rho(F_X)) = \Pi_{F_X}(D \mid T_{nuis}^{F,\perp}(F_X)).$$

Let $h_{ind,n}$ be an estimator of h_{ind,F_X}. Then, if $D_n \notin \mathcal{D}(\mu_n^0, \rho_n)$ we estimate $h_{ind,F_X}(D_n)$, and we denote the resulting estimator of h_{opt} with h_n:

$$h_n \equiv h_{ind,n}(D_n).$$

In the next subsections, we reparametrize the representations (2.51) and (2.53) of $D_{opt}(X \mid F_X, G)$ in terms of μ, ρ and the parameters identifying the optimal index h_{opt} and estimate h_{opt}, and D_{opt} accordingly.

2.9.1 Reparametrizing the representations of the optimal full data function

If we state an inverse of an operator applied to a function, then it will be implicitly assumed that this operator is $1-1$ and that this function is in the range of the operator. Firstly, consider the representation (2.51) of $D_{opt} = B_{F_X,G}^{-1}(S_{eff}^F)$. Recall the definition $\mathcal{D}(\mu, \rho) = \{D_h(\cdot \mid \mu, \rho) : h \in \mathcal{H}^F\}$. For each (h_{ind}, μ, ρ) (h_{ind} being an index mapping), let $\Pi_{h_{ind},\mu,\rho} : \mathcal{L}(\mathcal{X}) \to \mathcal{D}(\mu, \rho)$ be an operator such that for all $F_X \in \mathcal{M}^F$ and $D \in \mathcal{L}(\mathcal{X})$

$$\begin{aligned} \Pi_{h_{ind,F_X},\mu(F_X),\rho(F_X)}(D) &= \Pi(D \mid T_{nuis}^{\perp}(F_X)) \\ &= D_{h_{ind,F_X}(D)}(\cdot \mid \mu(F_X), \rho(F_X)). \end{aligned}$$

Thus $\Pi_{h,\mu,\rho}$ parametrizes the projection operator onto $T_{nuis}^{F,\perp}$ in terms of an index mapping h and the parameter values μ, ρ. Let $\mathbf{I}_{F_1,G_1} : \mathcal{L}(\mathcal{X}) \to \mathcal{L}(\mathcal{X})$

and denote its range with $R_\infty(\mathbf{I}_{F_1,G_1})$. For any $D \in R_\infty(\mathbf{I}_{F_1,G_1})$, define

$$B_{F_1,G_1,h_{ind},\mu,\rho}(D) = \Pi_{h_{ind},\mu,\rho}\mathbf{I}^{-1}_{F_1,G_1}(D).$$

Consider now the following representation of D_{opt}:

$$D_{opt}(F_1,G_1,h_{ind},h_{eff},\mu,\rho) \equiv B^{-1}_{F_1,G_1,h_{ind},\mu,\rho}D_{h_{eff}}(\cdot \mid \mu,\rho).$$

We can now show the following result.

Lemma 2.8 *Let (μ,ρ), $(F_1,G_1), h_{ind}(\cdot)$ be given. Assume $\mathcal{D}(\mu,\rho) \subset R_\infty(\mathbf{I}_{F_1,G_1})$. We have that $B_{F_1,G_1,h_{ind},\mu,\rho} : \mathcal{D}(\mu,\rho) \to \mathcal{D}(\mu,\rho)$. In particular, for any $h_{eff} \in \mathcal{H}^F$*

$$D_{opt}(F_1,G_1,h_{ind},h_{eff},\mu,\rho) \equiv B^{-1}_{F_1,G_1,h_{ind},\mu,\rho}D_{h_{eff}}(\cdot \mid \mu,\rho) \in \mathcal{D}(\mu,\rho). \tag{2.58}$$

This shows that given (μ,ρ) we end up in $\mathcal{D}(\mu,\rho)$ regardless of our choice of F_1, G_1, h_{ind}, and h_{eff}. Thus

$$D_{opt}(F_1,G_1,h_{ind},h_{eff},\mu,\rho) = D_{h_{opt}(F_1,G_1,h_{ind},h_{eff},\mu,\rho)}(\cdot \mid \mu,\rho)$$

for some mapping $(F_1,G_1,h_{ind},h_{eff},\mu,\rho) \to h_{opt}(F_1,G_1,h_{ind},h_{eff},\mu,\rho) \in \mathcal{H}^F$. Thus, estimation of F, G, h_{ind}, h_{eff} only affects the estimator $h_{opt,n}$ of the optimal index h_{opt} in $D_{h_{opt,n}}(\cdot \mid \mu_n^0, \rho_n)$. Since we prove a similar lemma for the representation (2.53) below, we will omit the proof of Lemma 2.8 here.

Consider now the representation (2.53) of $D_{opt}(F_X,G) = \mathbf{I}_{F_X,G}\left\{\Pi_{T^F}\mathbf{I}_{F_X,G}\right\}^{-1}(S^F_{eff})$. Let us start by parametrizing the projection operator Π_{T^F} onto the full data tangent space. We have

$$\Pi_{F_X}(D \mid T^F(F_X)) = \Pi_{F_X}(D \mid< S^F_{eff}(\cdot \mid F_X) >) + D - \Pi_{F_X}(D \mid T^{F,\perp}_{nuis}(F_X)).$$

The first projection operator is given by

$$\Pi_{F_X}(D \mid< S^F_{eff}(\cdot \mid F_X) >) = c_{F_X}(D)D_{h_{eff}(F_X)}(\cdot \mid \mu(F_X),\rho(F_X)),$$

where $c_{F_X}(D) = \langle D, S^{F\top}_{eff}\rangle_{F_X}\langle S^F_{eff}, S^{F\top}_{eff}\rangle^{-1}_{F_X}$, and $S^F_{eff} = D_{h_{eff}}(\cdot \mid \mu(F_X),\rho(F_X))$. Above, we reparametrized the projection operator onto $T^{F,\perp}_{nuis}$ as $\Pi_{h_{ind},\mu,\rho}$ in terms of $(h_{ind,F_X},\mu(F_X),\rho(F_X))$. This suggests the following parametrization of the projection operator $\Pi_{F_X}(D \mid T^F(F_X))$: suppose that for every $(c,h_{ind},h_{eff},\mu,\rho) \in \{(c_{F_X},h_{ind,F_X},h_{eff}(F_X),\mu(F_X),\rho(F_X)) : F_X \in \mathcal{M}^F\}$

$$\Pi_{c,h_{ind},h_{eff},\mu,\rho}(D) \equiv c(D)D_{h_{eff}}(\cdot \mid \mu,\rho) + D - D_{h_{ind}(D)}(\cdot \mid \mu,\rho)$$

is a well-defined operator from $\mathcal{L}(\mathcal{X}) \to \mathcal{L}(\mathcal{X})$ satisfying

$$\Pi_{c_{F_X},h_{ind,F_X},h_{eff}(F_X),\mu(F_X),\rho(F_X)} = \Pi_{F_X}(\cdot \mid T^F(F_X)).$$

The unknown index mapping parameter h_{ind} ranges over all index mappings $h : \mathcal{L}(\mathcal{X}) \to \mathcal{H}^{\mathcal{F}}$.

This suggests the following parametrization $D_{opt}(F_1, G_1, c, h_{ind}, h_{eff}, \mu, \rho)$ of $D_{opt}(F_X, G)$ (2.53):

$$\mathbf{I}_{F_1,G_1}\{\Pi_{c,h_{ind},h_{eff},\mu,\rho}\mathbf{I}_{F_1,G_1}\}^{-1}D_{h_{eff}}(\cdot \mid \mu, \rho).$$

Here, we implicitly assumed that this is an element of $\mathcal{L}(\mathcal{X})$. Notice that indeed $D_{opt}(F_X, G, c_{F_X}, h_{ind,F_X}, h_{eff}(F_X), \mu(F_X), \rho(F_X)) = D_{h_{opt}(F_X,G)}(\cdot \mid \mu(F_X), \rho(F_X))$.

We can now prove the following lemma.

Lemma 2.9 *For notational convenience, let* $\Pi_{1,\mu,\rho} = \Pi_{c,h_{ind},h_{eff},\mu,\rho}$. *Assume that* $h_{ind} : \mathcal{L}(\mathcal{X}) \to \mathcal{H}^{\mathcal{F}}$. *For any* F_1, G_1 *and* $h_{eff} \in \mathcal{H}^F$ *(so that* $D_{h_{eff}}(\cdot \mid \mu, \rho) \in \mathcal{D}(\mu, \rho)$*) satisfying* $\{\Pi_{1,\mu,\rho}\mathbf{I}_{F_1,G_1}\}^{-1}(D_{h_{eff}}) \in \mathcal{L}(\mathcal{X})$, *we have*

$$\mathbf{I}_{F_1,G_1}\{\Pi_{1,\mu,\rho}\mathbf{I}_{F_1,G_1}\}^{-1}D_{h_{eff}}(\cdot \mid \mu, \rho) \in \mathcal{D}(\mu, \rho). \tag{2.59}$$

Proof. We will show that for any possible value of the parameters we have $D_{opt}(F_1, G_1, c, h_{ind}, h_{eff}, \mu, \rho) \in \mathcal{D}(\mu, \rho)$. The proof is almost a direct consequence of the fact that for any $D \in \mathcal{L}(\mathcal{X})$ and $h_{eff} \in \mathcal{H}^F$

$$D - \Pi_{c,h_{ind},h_{eff},\mu,\rho}(D) = D_{h(D)}(\cdot \mid \mu, \rho) - c(D)D_{h_{eff}}(\cdot \mid \mu, \rho) \in \mathcal{D}(\mu, \rho) \tag{2.60}$$

and that

$$\Pi_{c,h_{ind},h_{eff},\mu,\rho}D_{opt}(F_1, G_1, c, h_{ind}, h_{eff}, \mu, \rho) = D_{h_{eff}}(\cdot \mid \mu, \rho). \tag{2.61}$$

Using short-hand notation $D_{opt} = D_{opt}(F_1, G_1, c, h_{ind}, h_{eff}, \mu, \rho)$, $\Pi_{1,\mu,\rho} = \Pi_{c,h_{ind},h_{eff},\mu,\rho}$, we have by (2.60) and (2.61)

$$\begin{aligned} D_{opt} &= \{D_{opt} - \Pi_{1,\mu,\rho}D_{opt}\} + \Pi_{1,\mu,\rho}D_{opt} \\ &= D_{h(D_{opt})}(\cdot \mid \mu, \rho) - c(D_{opt})D_{h_{eff}}(\cdot \mid \mu, \rho) + D_{h_{eff}}(\cdot \mid \mu, \rho). \end{aligned}$$

Since $D_{opt} \in \mathcal{L}(\mathcal{X})$ and $h : \mathcal{L}(\mathcal{X}) \to \mathcal{H}^{\mathcal{F}}$, this proves that $D_{opt} \in \mathcal{D}(\mu, \rho)$. □

2.9.2 Estimation of the optimal full data structure estimating function

Consider the representation (2.58) of the optimal full data structure function $D_{opt}(F_X, G)$. Estimation of this representation involves estimation of the components (F_X, G) identifying the nonparametric information operator, h_{ind,F_X} identifying the index mapping of the projection operator onto the orthogonal complement of the nuisance tangent space, and $h_{eff}(F_X)$ identifying the index of the full data canonical gradient $S^F_{eff}(\cdot \mid F_X)$ and $(\mu(F_X), \rho(F_X))$. Substitution of estimators for each of these components yields an estimator of D_{opt}:

$$\begin{aligned} D_n &= D_{opt}(F_n, G_n, h_n, h_{eff,n}, \mu^0_n, \rho_n) \\ &= \left\{\Pi_{h_n,\mu^0_n,\rho_n}\mathbf{I}^-_{F_n,G_n}\right\}^{-1} D_{h_{eff,n}}(\cdot \mid \mu^0_n, \rho_n). \end{aligned}$$

Similarly, an estimator based on representation (2.59) is given by

$$\begin{aligned} D_n &= D_{opt}(F_n, G_n, c_n, h_n, h_{eff,n}, \mu_n^0, \rho_n) \\ &= I_{F_n,G_n}\{\Pi_{c_n,h_n,h_{eff,n},\mu_n^0,\rho_n}\mathbf{I}_{F_n,G_n}\}^- D_{h_{eff,n}}(\cdot \mid \mu_n^0, \rho_n). \end{aligned}$$

Lemmas 2.8 and 2.9 show that the right-hand side is indeed an element of $\mathcal{D}(\mu_n^0, \rho_n)$ so that

$$D_n = D_{h_{opt,n}}(\cdot \mid \mu_n^0, \rho_n) \text{ for some } h_{opt,n} \in \mathcal{H}^F.$$

Thus $F_n, G_n, c_n, h_n, h_{eff,n}$ only affect the index estimate $h_{opt,n}$. If one uses approximations $\tilde{D}_n$ of these estimators D_n that are not necessarily elements of $\mathcal{D}(\mu_n^0, \rho_n)$, then, as mentioned in the previous subsection, one should use

$$h_{opt,n} = h_{ind,n}(\tilde{D}_n).$$

If in model $\mathcal{M}$ the full data working model $\mathcal{M}^w$ is such that it yields an (e.g., maximum likelihood) estimator F_n of the full data distribution F_X itself, then one could decide to use a substitution estimator for each of the full data distribution parameters:

$$D_n(\cdot \mid \mu_n^0, \rho_n) = I_{F_n,G_n}\mathbf{I}^{*-1}_{F_n,G_n} D_{h_{eff}(F_n)}(\cdot \mid \mu_n^0, \rho_n), \tag{2.62}$$

where

$$I^{*-1}_{F_n,G_n} = \Pi_{c_{F_n},h_{F_n},h_{eff}(F_n),\mu_n^0,\rho_n}\mathbf{I}_{F_n,G_n}.$$

2.10 Locally Efficient Estimation with Score-Operator Representation

Recall that at the true F_X, G

$$IC(\cdot \mid Q(F_X, G), G, D_h(\cdot \mid \mu, \rho)) = A_{F_X}\mathbf{I}^-_{F_X,G} D_h(\cdot \mid \mu, \rho).$$

Suppose that we actually use this representation in terms of F_X, G to define our estimating function; that is, just parametrize IC in terms of F_X and G:

$$IC(\cdot \mid F_X, G, D_h(\cdot \mid \mu, \rho)) = A_{F_X}\mathbf{I}^-_{F_X,G} D_h(\cdot \mid \mu, \rho).$$

Let F_n be an estimator of F_X according to the working model $\mathcal{M}^{F,w}$ and let G_n be an estimator of G according to the working model $\mathcal{G}$. In addition, assume that we estimate $D_{h_{opt}}$ with $D_{h_{opt,n}} = \mathbf{I}_{F_n,G_n}\mathbf{I}^{*-1}_{F_n,G_n} D_{h_{eff}(F_n)}(\cdot \mid \mu, \rho_n)$ as defined by (2.62) using F_n and G_n. Then, the resulting estimating function for μ is given by

$$IC(\cdot \mid F_n, G_n, D_{h_{opt,n}}(\cdot \mid \mu, \rho_n)) = A_{F_n}\mathbf{I}^{*-}_{F_n,G_n} D_{h_{eff}(F_n)}(\cdot \mid \mu, \rho_n).$$

Since this is just a special application of the one-step estimator (2.35), we can apply Theorem 2.5, which shows, under regularity conditions, that

the resulting one-step estimator is consistent and asymptotically normal if either $\mathcal{M}^{F,w}$ is correctly specified or $\mathcal{G}$ is correctly specified and that it is efficient if both are correctly specified. The one-step estimator can be computed by inverting $\mathbf{I}^*_{F_n,G_n}$ with the successive substitution method of Lemma 2.6.

Note that, given estimators F_n, G_n, this provides us with a completely automated method for locally efficient estimation in any censored data model. Off course, here one also uses a substitution estimator $\rho_n = \rho(F_n, G_n)$.