Biometrika (2017), **104**, 3, *pp.* 561–581
Printed in Great Britain
doi: 10.1093/biomet/asx027
Advance Access publication 15 June 2017

Multiple robustness in factorized likelihood models

By J. MOLINA

Instituto de Cálculo, Universidad de Buenos Aires, Intendente Guiraldes 2160, Pabellon II, Buenos Aires 1428, Argentina
julietechka@gmail.com

A. ROTNITZKY

Department of Economics, Di Tella University, Figueroa Alcorta 7350, Buenos Aires 1428, Argentina
arotnitzky@utdt.edu

M. SUED

Instituto de Cálculo, Universidad de Buenos Aires, Intendente Guiraldes 2160, Pabellon II, Buenos Aires 1428, Argentina
marielasued@gmail.com

and J. M. ROBINS

Department of Epidemiology, Harvard T. H. Chan School of Public Health, 655 Huntington Avenue, Boston, Massachusetts 02115, U.S.A.
robins@hsph.harvard.edu

Summary

We consider inference under a nonparametric or semiparametric model with likelihood that factorizes as the product of two or more variation-independent factors. We are interested in a finite-dimensional parameter that depends on only one of the likelihood factors and whose estimation requires the auxiliary estimation of one or several nuisance functions. We investigate general structures conducive to the construction of so-called multiply robust estimating functions, whose computation requires postulating several dimension-reducing models but which have mean zero at the true parameter value provided one of these models is correct.

Some key words: Causal inference; Estimating function; Missing data; Semiparametric model.

1. Introduction

The goal of this paper is to develop a general theory of multiply robust estimation functions for a parameter $\beta(p)$ in semiparametric or nonparametric models for the law p under which the likelihood p factorizes as $p = gh$ where g and h are variation-independent functions and $\beta(p)$ depends on p only through g. Leading examples of such models are ignorable missing data and causal inference models or combinations thereof where h is the likelihood factor corresponding to the missingness and/or treatment assignment mechanism and g is the

likelihood factor corresponding to the full and/or the counterfactual data. In the literature, an estimating function is called doubly robust if it has mean zero if either a model for h or a model for g is correct, giving the analyst two opportunities to obtain an unbiased estimating function for $\beta(p)$.

Often g and h factorize further as $g = g_1 \cdots g_{K+1}$ and $h = h_1 \cdots h_K$ ($K \geqslant 2$) with all functions g_k and h_k being variation-independent. For example, in a longitudinal study, h_k would correspond to the missingness mechanism and/or treatment assignment mechanism at occasion k and g_k to the likelihood contribution to the full and/or the counterfactual data at occasion k. In this setting, we still refer to an estimating function as doubly robust if it has mean zero when either a model for g is correct or a model for h is correct. A multiply robust estimating function is not only doubly robust but, in addition, has mean zero when a model for various strict subsets of the g_k and h_k are correct.

Many doubly and multiply robust procedures have been proposed to estimate parameters of models used in causal inference; see Robins (2000), Murphy et al. (2001), Lunceford & Davidian (2004), Bang & Robins (2005), Tan (2006, 2010a), Zhang et al. (2012), Orellana et al. (2010), Vansteelandt et al. (2008), Goetgeluk et al. (2008), Tchetgen Tchetgen & Shpitser (2012), van der Laan & Rubin (2006), and van der Laan & Gruber (2010). Doubly robust estimators have also been proposed to estimate parameters of models often assumed in studies with missing data, e.g., Scharfstein et al. (1999), Vansteelandt et al. (2007), Cao et al. (2009), Tsiatis et al. (2011), Tan (2010b, 2011), Rotnitzky et al. (2012) and Vermeulen & Vansteelandt (2015).

Recently, Han & Wang (2013), Chan & Yam (2014), and Han (2014, 2016) have described estimators that are consistent under correct specification of one of multiple working models for h or multiple working models for g. They refer to these estimators as multiply robust. By our definition, their estimators would be considered doubly robust but not multiply robust, because they still require that a model for h or a model for g be correct to achieve consistency. If one wanted to use the term multiply robust for their estimators, one would need a different description for the estimators of the present article. However, for compatibility with the terminology used in the literature prior to 2013, e.g., Vansteelandt et al. (2007), Vansteelandt et al. (2008), Tchetgen Tchetgen (2009), and Tchetgen Tchetgen & Shpitser (2012), we will use our above definition of multiple robustness.

The construction of doubly and multiply robust estimators appears to be ad hoc. With the exception of Robins et al. (2000) and Robins & Rotnitzky (2001) for doubly robust estimation and Robins et al. (2008) for doubly and multiply robust estimation, there is a paucity of theoretical results giving sufficient conditions under which one can hope to find multiply robust procedures. The goal of this article is to contribute to filling this gap.

We provide a general theoretical framework under which one can construct estimating equations that could be used to compute multiply robust estimators. Our results extend some of the theoretical results in Robins et al. (2000) and Robins & Rotnitzky (2001) from doubly to multiply robust estimation. An important contribution is to offer sufficient conditions under which estimation of finite-dimensional nuisance parameters indexing the nuisance functions is possible without the need to estimate new nuisance functions. Our framework assumes a model where the likelihood can be written as the product of two or more variation-independent, possibly infinite-dimensional, factors. Furthermore, the parameter $\beta(p)$ of interest is functionally dependent on only one of the likelihood factors and equals the solution of a population moment equation. Although not all problems where doubly or multiply robust estimators exist fit into this framework (e.g., Tchetgen Tchetgen & Shpitser, 2012; Zhang et al., 2012; Scharfstein et al., 1999; Vansteelandt et al., 2008) our theory is general enough to explain the existence of many available doubly and multiply robust estimators and sheds light on the

construction of multiply robust estimators in other interesting problems. Proofs are relegated to the Supplementary Material.

2. Motivating examples

2·1. *Example 1*: *mean estimation in longitudinal missing data*

Consider the estimation of the mean of a possibly unobserved outcome L_3^f at time t_3 in a longitudinal study that suffers from drop-out. Suppose that the study calls for measurements L_1^f and L_2^f to be made at times t_1 and t_2. Assume L_1^f is always observed. Let A_j denote the indicator of still being in the study at time t_{j+1} $(j = 1, 2)$. The data recorded on a given study participant are then the random vector $Z = (L_1, A_1, L_2, A_2, L_3)$ where $L_1 = L_1^f, L_2 = A_1 L_2^f$ and $L_3 = A_1 A_2 L_3^f$. Assume $\text{pr}(A_1 = 1 \mid L_1, L_3^f)$ and $\text{pr}(A_2 = 1 \mid A_1 = 1, L_1, L_2, L_3^f)$ do not depend on L_3^f, so L_3^f is missing at random. Under these assumptions the mean of L_3^f is the following function of the density p of Z (Robins et al., 1995):

$$\beta(p) \equiv E_p\left[E_p\left\{E_p\left(L_3 \mid A_2 = A_1 = 1, L_1, L_2\right) \mid A_1 = 1, L_1\right\}\right].$$

Here and in what follows $E_p(\cdot)$ denotes expectation computed under p. The missing at random assumption does not impose restrictions on p (Gill et al., 1997). Thus, formally, our goal is to estimate $\beta(p)$ based on a random sample of Z under a model $\mathcal{M}$ for the law p of Z that places no restrictions on p.

For any $z = (l_1, a_1, l_2, a_2, l_3)$ let

$$g_1(z) \equiv p_{L_1}(l_1), \quad g_2(z) \equiv p_{L_2|A_1,L_1}(l_2 \mid a_1, l_1), \quad g_3(z) \equiv p_{L_3|\bar{A}_2,\bar{L}_2}\left(l_3 \mid \bar{a}_2, \bar{l}_2\right), \tag{1}$$

$$h_1(z) \equiv p_{A_1|L_1}(a_1 \mid l_1), \quad h_2(z) \equiv p_{A_2|A_1,\bar{L}_2}\left(a_2 \mid a_1, \bar{l}_2\right), \tag{2}$$

where, throughout, an overbar over a variable subscripted by k denotes the history of the variable up to k, i.e., $\bar{L}_k = (L_1, \ldots, L_k)$. The mean $\beta(p)$ depends on p only through $g \equiv (g_1, g_2, g_3)$, so we denote it by $\beta(g)$. In fact, $\beta(g)$ depends just on g_1 and on

$$\begin{aligned} \rho_1(g) &\equiv E_{g_2}\left\{E_{g_3}\left(L_3 \mid A_2 = A_1 = 1, \bar{L}_2\right) \mid A_1 = 1, L_1\right\}, \\ \rho_2(g) &\equiv E_{g_3}\left(L_3 \mid A_2 = A_1 = 1, \bar{L}_2\right). \end{aligned} \tag{3}$$

Furthermore, $\beta(g)$ is the unique solution of the equation $E_p\{M(\beta, h)\} = 0$, where

$$M(\beta, h) \equiv \left[A_1 A_2 / \left\{E_{h_1}(A_1 \mid L_1) E_{h_2}\left(A_2 \mid A_1, \bar{L}_2\right)\right\}\right](L_3 - \beta),$$

$h \equiv (h_1, h_2)$, and, throughout, $E_q(\cdot \mid \cdot)$ stands for conditional expectation under q. The function $M(\beta, h)$ is a so-called inverse probability of censoring weighted estimating function. The function takes value zero for drop-outs. For non-drop-outs, $M(\beta, h)$ equals the residual $L_3 - \beta$ weighted by the inverse of the product of the probabilities of being observed at each occasion, conditional on past data.

When L_1 and/or L_2 are multivariate with continuous components, consistent estimation of $\beta(g)$ under model $\mathcal{M}$ is not feasible in practice due to the curse of dimensionality (see, e.g.,

Robins et al., 1994), Robins & Ritov (1997) proved that in this setting there are no uniformly consistent estimators of $\beta(g)$. Many dimension-reduction strategies exist for addressing the curse of dimensionality, including those that yield doubly robust estimators of $\beta(g)$. Such estimators are consistent when either working parametric models $\mathcal{H}_j \equiv \{h_{j,\alpha_j} : \alpha_j \in \Xi_j\}$ for h_j $(j = 1, 2)$ are correct, or working parametric models $\mathcal{R}_j \equiv \{\rho_{j,\tau_j} : \tau_j \in \Upsilon_j\}$ for $\rho_j(g)$ $(j = 1, 2)$ are correct (van der Laan & Robins, 2003; Tsiatis, 2006). Here Ξ_j and Υ_j $(j = 1, 2)$ are subsets of Euclidean spaces. Tchetgen Tchetgen (2009) recently showed that it is possible to construct a multiply robust estimator of $\beta(g)$ which confers further protection with respect to model misspecification. We will now review his estimator. Throughout, for conciseness, we write $p \in \mathcal{R}_j$ whenever $\rho_j(g) \in \mathcal{R}_j$, and we write $p \in \mathcal{H}_j$ whenever $h_j \in \mathcal{H}_j$.

Let $\tau \equiv (\tau_1, \tau_2)$ and $\alpha \equiv (\alpha_1, \alpha_2)$. Consider the estimating functions

$$U(\beta, \tau, \alpha) \equiv \{A_2 A_1 / (\pi_{1,\alpha_1} \pi_{2,\alpha_2})\}(L_3 - \rho_{2,\tau_2}) + (A_1/\pi_{1,\alpha_1})(\rho_{2,\tau_2} - \rho_{1,\tau_1}) + (\rho_{1,\tau_1} - \beta),$$
$$T_1(\tau, \alpha) \equiv A_1 d(L_1; \tau, \alpha_2)\{(A_2/\pi_{2,\alpha_2})(L_3 - \rho_{2,\tau_2}) + (\rho_{2,\tau_2} - \rho_{1,\tau_1})\},$$

where $\pi_{1,\alpha_1} \equiv E_{h_{1,\alpha_1}}(A_1 \mid L_1)$ and $\pi_{2,\alpha_2} \equiv E_{h_{2,\alpha_2}}(A_2 \mid A_1, \bar{L}_2)$ and $d(L_1; \tau, \alpha_2)$ is a specified column vector-valued function of $(L_1; \tau, \alpha_2)$ with as many rows as the dimension of τ_1. The displayed functions satisfy:

(I) $E_p[U\{\beta(g), \tau, \alpha\}] = 0$, provided $p \in (\mathcal{H}_1 \cup \mathcal{R}_1) \cap (\mathcal{H}_2 \cup \mathcal{R}_2)$ and the parameters indexing the correctly specified working models are replaced by their true values, e.g., if $p \in \mathcal{R}_j$, τ_j is replaced by $\tau_j(g)$ satisfying $\rho_j(g) = \rho_{j,\tau_j(g)}$;

(II) $E_p[T_1\{\tau_1(g), \tau_2, \alpha\}] = 0$, provided $p \in \mathcal{R}_1$ and either $p \in \mathcal{R}_2$ and $\tau_2 = \tau_2(g)$, or $p \in \mathcal{H}_2$ and α_2 is such that $h_2 = h_{2,\alpha_2}$.

Facts (I) and (II), which were proved by Tchetgen Tchetgen (2009), are a consequence of the results presented in § 3, which imply that one can construct a multiply robust estimator $\hat{\beta}$, in the sense that it converges in probability to $\beta(g)$ whenever $p \in (\mathcal{H}_1 \cup \mathcal{R}_1) \cap (\mathcal{H}_2 \cup \mathcal{R}_2)$, as follows.

Let $\hat{\alpha}_j$ be the maximum likelihood estimator of α_j $(j = 1, 2)$. Let $\hat{\tau}_2$ be the possibly nonlinear least squares estimator of τ_2 in the regression of L_3 on $\bar{L}_2$ among units with $A_2 = A_1 = 1$, i.e., $\hat{\tau}_2$ solves $P_n\{T_2(\tau_2)\} = 0$, where $T_2(\tau_2) = A_1 A_2 (\partial \rho_{2,\tau_2}/\partial \tau_2)(L_3 - \rho_{2,\tau_2})$. Let $\hat{\tau}_1$ be the estimator of τ_1 solving $P_n\{T_1(\tau_1, \hat{\tau}_2, \hat{\alpha}_2)\} = 0$. Here and throughout, P_n is the empirical mean operator, i.e., $P_n(V) = n^{-1} \sum_{i=1}^n V_i$. The multiply robust estimator $\hat{\beta}$ solves $P_n\{U(\beta, \hat{\tau}_1, \hat{\tau}_2, \hat{\alpha})\} = 0$ and was derived by Tchetgen Tchetgen (2009).

2·2. *Example 2: partial linear regression with missing outcomes*

Consider a study that requires measurements of (L_1, A_1, L_2, L_3^f) on a random sample of subjects. Suppose that L_3^f is a scalar unbounded outcome, A_1 is a binary treatment indicator, L_1 is a vector of pre-treatment covariates and L_2 is a vector of post-treatment variables measured prior to L_3^f. Consider the estimation of $\beta^* \in \mathbb{R}$ under the partially linear regression model (Engle et al., 1986)

$$E(L_3^f \mid A_1, L_1) = \beta^* A_1 + v(L_1) \tag{4}$$

when (L_1, A_1, L_2) is always observed but L_3^f is missing in some subjects. In (4) $v(L_1)$ is an unknown function. Under the assumption of no unmeasured confounders, model (4) coincides

with a structural nested mean model for a point exposure (Robins, 2000) and β^* quantifies the average treatment effect of A_1 on the mean of L_3^f in subjects with covariates L_1. The variable L_2 is not included as a covariate of the semiparametric regression model because conditioning on post-treatment variables may bias the estimates of treatment effects.

Assume L_3^f is missing at random, that is, $\text{pr}(A_2 = 1 \mid L_2, A_1, L_1, L_3^f) = \text{pr}(A_2 = 1 \mid L_2, A_1, L_1)$, where A_2 denotes the binary indicator that L_3^f is not missing. Under this assumption β^* is a function $\beta(p)$ of the density p of the observed data vector $Z = (L_1, A_1, L_2, A_2, L_3)$, where $L_3 = A_2 L_3^f$. The parameter $\beta(p)$ satisfies

$$E_p\left\{E_p\left(L_3 \mid A_2 = 1, \bar{L}_2, A_1\right) \mid A_1, L_1\right\} = \beta(p) A_1 + v_p\left(L_1\right) \tag{5}$$

for some unknown function $v_p(L_1)$. Thus, formally, our problem is to estimate $\beta(p)$ from n independent and identically distributed copies of Z under a model $\mathcal{M}$ that imposes no restrictions on p other than (5).

Let g_j and h_j be defined as in (1) and (2), respectively. Observe that $\beta(p)$ and $v_p(\cdot)$ depend on p just through $g \equiv (g_1, g_2, g_3)$. This is because each conditional expectation in the left-hand side of (5) depends on p only through g. Throughout, we write $\beta(g)$ instead of $\beta(p)$ and $\rho_1(g)$ instead of $v_p(\cdot)$ to stress this point. The functional $\beta(g)$ can also be characterized as the unique solution of $E_p\{M(\beta, h)\} = 0$, where

$$M(\beta, h) \equiv \left\{A_2 / E_{h_2}(A_2 \mid \bar{L}_2, A_1)\right\}(L_3 - \beta A_1)\left\{A_1 - E_{h_1}(A_1 \mid L_1)\right\}. \tag{6}$$

To interpret $M(\beta, h)$ observe that if L_3 were always observed, i.e., if A_2 were always 1, (6) would reduce to $(L_3 - \beta A_1)\left\{A_1 - E_{h_1}(A_1 \mid L_1)\right\}$. This product has mean zero given L_1 at $\beta(p)$ when L_3 is always observed because in such a case, equation (5) is equivalent to the assertion that $E_p\{L_3 - \beta(p) A_1 \mid A_1, L_1\}$ does not depend on A_1. Thus, this product would be an unbiased estimating function for $\beta(p)$ if L_3 were never missing. The right-hand side of (6) is the inverse probability weighted version of this full-data unbiased estimating function.

Robins (2000) describes doubly robust estimators of $\beta(g)$ that are consistent and asymptotically normal so long as either some user-specified parametric models $\mathcal{H}_j \equiv \left\{h_{j,\alpha_j} : \alpha_j \in \Xi_j\right\}$ for h_j $(j = 1, 2)$ are correct or some parametric models $\mathcal{R}_j \equiv \left\{\rho_{j,\tau_j} : \tau_j \in \Upsilon_j\right\}$ for $\rho_j(g)$ $(j = 1, 2)$, where $\rho_2(g) \equiv E\left(L_3 \mid A_2 = 1, \bar{L}_2, A_1\right)$, are correct. Here we construct a multiply robust estimator of $\beta(g)$. To do so, we define π_{k,α_k} $(k = 1, 2)$ as in Example 1, and

$$U(\beta, \tau, \alpha) \equiv \left\{\left(A_2 / \pi_{2,\alpha_2}\right)\left(L_3 - \rho_{2,\tau_2}\right) + \rho_{2,\tau_2} - \beta A_1 - \rho_{1,\tau_1}\right\}\left(A_1 - \pi_{1,\alpha_1}\right),$$

$$T_1(\beta, \tau_1, \tau_2, \alpha) \equiv d(L_1, A_1; \tau, \alpha_2)\left\{\left(A_2 / \pi_{2,\alpha_2}\right)\left(L_3 - \rho_{2,\tau_2}\right) + \rho_{2,\tau_2} - \beta A_1 - \rho_{1,\tau_1}\right\}\left(A_1 - \pi_{1,\alpha_1}\right),$$

where $d(L_1, A_1; \tau, \alpha_2)$ is a specified column vector-valued function of $(L_1, A_1; \tau, \alpha_2)$ with as many rows as the dimension of τ_1.

The functions U and T_1 satisfy the properties (I) and (II) stated in Example 1, except that in condition (II) $T_1\{\tau_1(g), \tau_2, \alpha\}$ is replaced by $T_1\{\beta(g), \tau_1(g), \tau_2, \alpha\}$. These properties are easily checked directly, but also follow from the general results presented in § 3. They imply that the following estimator $\hat{\beta}$ is, under regularity conditions, multiply robust in the sense that it converges in probability to $\beta(g)$ whenever $p \in (\mathcal{H}_1 \cup \mathcal{R}_1) \cap (\mathcal{H}_2 \cup \mathcal{R}_2)$. Specifically, $\hat{\beta}$ solves $P_n\left[U\left\{\beta, \hat{\tau}_1(\beta), \hat{\tau}_2, \hat{\alpha}\right\}\right] = 0$ where for each fixed β, $\hat{\tau}_1(\beta)$ solves $P_n\left\{T_1\left(\beta, \tau_1, \hat{\tau}_2, \hat{\alpha}\right)\right\} = 0$. In these equations $\hat{\tau}_2$ is the least squares estimator of τ_2 in the regression of L_3 on $\left(\bar{L}_2, A_1\right)$ among

units with $A_2 = 1$ assuming $\rho_2(g) \in \mathcal{R}_2$, e.g., $\hat{\tau}_2$ solves $P_n\{T_2(\tau_2)\} = 0$ where $T_2(\tau_2) = A_2\left(\partial\rho_{2,\tau_2}/\partial\tau_2\right)\left(L_3 - \rho_{2,\tau_2}\right)$. Also, $\hat{\alpha}_j$ is the maximum likelihood estimator of α_j $(j = 1, 2)$ under $\mathcal{H}_j$.

3. MODELS RESTRICTING FUNCTIONALS OF CONDITIONAL DENSITIES

3·1. *Assumptions*

Let $Z = (L_1, A_1, \ldots, L_K, A_K, L_{K+1})$ be a random vector that takes values in some set $\mathcal{Z}$ and has density p. Assume $p(z) > 0$ for all $z \in \mathcal{Z}$. Let

$$g_k(z) \equiv p\left(l_k \mid \bar{l}_{k-1}, \bar{a}_{k-1}\right), \quad h_k(z) \equiv p\left(a_k \mid \bar{l}_k, \bar{a}_{k-1}\right)$$

denote the associated conditional densities. Let $g \equiv (g_1, \ldots, g_{K+1})$ and $h \equiv (h_1, \ldots, h_K)$ and let $\mathcal{M}$ be a large, i.e., semiparametric or nonparametric, model for p of strictly positive densities on $\mathcal{Z}$ that may restrict g to belong to some set $\mathcal{G}$ but does not restrict h. For instance, in Example 2, $\mathcal{G}$ is the set $\{g : g \text{ satisfies } (5)\}$. Assume there exists $M(\beta, h) \equiv m(Z; \beta, h)$, a known $\mathbb{R}^r$-valued function of $(Z; \beta, h)$ where β is a $r \times 1$ vector, such that for all p in $\mathcal{M}$ the equation

$$E_p\{M(\beta, h)\} = 0 \tag{7}$$

has a unique solution $\beta(p)$ that depends on p only through g and thus is hereafter denoted by $\beta(g)$. We study estimation of $\beta(g)$ based on n independent and identically distributed copies $Z_1, \ldots, Z_n$ of Z assuming, as in Examples 1 and 2, that due to the curse of dimensionality, consistent estimation of $\beta(g)$ under $\mathcal{M}$ is unfeasible in practice.

Consider parametric working models

$$\mathcal{H}_k = \left\{h_{k,\alpha_k} : \alpha_k \in \Xi_k\right\} \ (k \in [K]), \quad \mathcal{P}_k = \left\{g_{k,\varsigma_k} : \varsigma_k \in \Gamma_k\right\} \ (k \in [K+1]) \tag{8}$$

for h_k and g_k respectively, where Ξ_k and Γ_k lie in some Euclidean spaces and, throughout, for any k, $[k] \equiv \{1, \ldots, k\}$. In the following $p = gh$ stands for $p = g_1 \cdots g_{K+1} h_1 \cdots h_K$ and, for conciseness, we write $p \in \mathcal{H}_j$ if $h_j \in \mathcal{H}_j$ and $p \in \mathcal{P}_k$ if $g_k \in \mathcal{P}_k$. Let $\hat{h} \equiv \left(h_{1,\hat{\alpha}_1}, \ldots, h_{K,\hat{\alpha}_K}\right)$ and $\hat{g} \equiv \left(g_{1,\hat{\varsigma}_1}, \ldots, g_{K+1,\hat{\varsigma}_{K+1}}\right)$, where $\hat{\alpha}_k$ and $\hat{\varsigma}_k$ are the maximum likelihood estimators of α_k and ς_k. Under regularity conditions, the estimator $\hat{\beta}_{\hat{h}}$ solving $P_n\left\{M\left(\beta, \hat{h}\right)\right\} = 0$ converges in probability to $\beta(g)$ if $p \in \mathcal{H}_k$ for all $k \in [K]$. Likewise, the substitution-type estimator $\beta(\hat{g})$ converges to $\beta(g)$ if $p \in \mathcal{P}_k$ for all $k \in [K+1]$. However, convergence of $\hat{\beta}_{\hat{h}}$ to $\beta(g)$ is not guaranteed if $p \notin \mathcal{H}_k$ for some k, and likewise convergence of $\beta(\hat{g})$ to $\beta(g)$ is not ensured if $p \notin \mathcal{P}_k$ for some k.

3·2. *Multiple robustness*

Robins et al. (2000) noted that under our assumptions and regularity conditions, it is possible to construct an estimator $\tilde{\beta}$ that is doubly robust in that $\tilde{\beta}$ converges in probability to $\beta(g)$ if (a) $p \in \bigcap_{k=2}^{K+1} \mathcal{P}_k$ or (b) $p \in \bigcap_{k=1}^{K} \mathcal{H}_k$. The estimator $\tilde{\beta}$ solves $P_n\left\{Q\left(\beta, \hat{h}, \hat{g}\right)\right\} = 0$, where for any $p = gh$,

$$Q(\beta, h, g) \equiv M(\beta, h) - \sum\nolimits_{k=1}^{K} \Pi_k\left[M(\beta, h); p\right] \tag{9}$$

and

$$\Pi_k[W;p] \equiv E_p(W \mid \overline{A}_k, \overline{L}_k) - E_p(W \mid \overline{A}_{k-1}, \overline{L}_k). \tag{10}$$

When W is a real-valued random variable, the operator $\Pi_k[\cdot;p]$ maps W into the closest element of

$$\Lambda_k(p) \equiv \left\{B = b\left(\overline{A}_k, \overline{L}_k\right) \text{ real-valued} : E_p\left(B \mid \overline{A}_{k-1}, \overline{L}_k\right) = 0, E_p\left(B^2\right) < \infty\right\}. \tag{11}$$

That is, $\Pi_k[W;p]$ is in $\Lambda_k(p)$ and satisfies $E_p\left\{(W - \Pi_k[W;p])^2\right\} \leqslant E_p\left\{(W - B)^2\right\}$ for all B in $\Lambda_k(p)$. The projection $\Pi_k[W;p]$ is the unique element in $\Lambda_k(p)$ such that $W - \Pi_k[W;p]$ is orthogonal to, i.e., uncorrelated with, all elements of $\Lambda_k(p)$ under p. When W is multivariate, $\Pi_k[W;p]$ stands for the vector formed by the projections into $\Lambda_k(p)$ of the coordinates of W. The space $\Lambda_k(p)$ is indeed the tangent space in the submodel of $\mathcal{M}$ that fixes all conditional densities but h_k, i.e., the space of scores in submodels of model $\mathcal{M}$ that parameterize h_k and leave all other conditional densities fixed. The spaces $\Lambda_k(p)$ ($k \in [K]$) are mutually orthogonal. See § 4.

In fact $\tilde{\beta}$ converges, under regularity conditions, to $\beta(g)$ not only under conditions (a) and (b) above but also if (c) for some $k \in [K-1], p \in \left(\bigcap_{j=1}^{k} \mathcal{H}_j\right) \cap \left(\bigcap_{j=k+2}^{K+1} \mathcal{P}_j\right)$. We refer to an estimator as a $(K+1)$-multiply robust estimator of $\beta(g)$ if, like $\tilde{\beta}$, it converges in probability to $\beta(g)$ whenever conditions (a) or (b) of the previous paragraph hold or condition (c) holds. To the best of our knowledge, the $(K+1)$-multiple robustness property of $\tilde{\beta}$ has not been reported elsewhere. Aside from regularity conditions, this property is a consequence of Proposition 2, which is a corollary to the following key result.

THEOREM 1. *Given $p^* = gh^*$ and $p = gh$ such that $p(Z)/p^*(Z) \in L_2(p^*)$ and any random variable $D = d(Z)$ such that $E_{p^*}\left(D^2\right) < \infty$,*

$$E_p\left(D - \sum\nolimits_{k=1}^{K} \Pi_k[D;p^*]\right) = E_{p^*}(D). \tag{12}$$

Below we provide two important results, stated in Propositions 1 and 2 which follow from Theorem 1 after we observe that by the definition of Π_j, for any $p = gh$ and $p^{**} = g^*h^*$,

$$E_p\left(\Pi_j[D;p^{**}] \mid \overline{A}_{j-1}, \overline{L}_j\right) = E_{h_j}\left\{E_{p^{**}}\left(D \mid \overline{A}_j, \overline{L}_j\right) \mid \overline{A}_{j-1}, \overline{L}_j\right\} - E_{h_j^*}\left\{E_{p^{**}}\left(D \mid \overline{A}_j, \overline{L}_j\right) \mid \overline{A}_{j-1}, \overline{L}_j\right\}.$$

Then, for any two densities p and p^{**} that agree on h_j, it follows that the right-hand side of the previous equation is 0. Consequently, if $p = gh$, $p^* = gh^*$ and $p^{**} = g^*h^*$, then

$$h_j = h_j^* \Rightarrow E_p\left(\Pi_j[D;p^*]\right) = E_p\left(\Pi_j[D;p^{**}]\right) = 0. \tag{13}$$

Now, suppose that for some set $C \subseteq [K]$, (i) $h_j = h_j^*$ for all $j \in C$ and (ii) $\Pi_k[D;p^{**}] = \Pi_k[D;p^*]$ for all $k \in \overline{C}$ where throughout $\overline{C} \equiv [K] - C$. Then

$$E_p\left(D - \sum\nolimits_{k=1}^{K} \Pi_k[D;p^{**}]\right) = E_p\left(D - \sum\nolimits_{k=1}^{K} \Pi_k[D;p^*]\right)$$

because $E_p\left(\sum_{k\in C}\Pi_k[D;p^*]\right) = 0 = E_p\left(\sum_{k\in C}\Pi_k[D;p^{**}]\right)$ by (i) and (13) and because $\sum_{k\in\overline{C}}\Pi_k[D;p^*] = \sum_{k\in\overline{C}}\Pi_k[D;p^{**}]$ by (ii). Theorem 1 then implies that

$$E_p\left(D - \sum\nolimits_{k=1}^{K}\Pi_k[D;p^{**}]\right) = E_{p^*}(D).$$

This result states that we can recover from the law p the mean of any random variable D under another law p^* so long as (i) p and p^* agree on the factors h_j with indexes j in some set C and (ii) for those h_j factors for which they don't agree we can compute the correct projections into the corresponding tangent spaces. Specifically, under (i) and (ii) the mean under p^* of D agrees with the mean under p of D minus the sum of the projections of D into the K tangent spaces.

Upon substituting $M\{\beta(g), h^*\}$ for D, the preceding argument is a proof of the following proposition.

PROPOSITION 1. *Let $p^{**} = g^*h^*$, $p^* = gh^*$ and $p = gh$ be such that $p(Z)/p^*(Z) \in L_2(p^*)$ and $p(Z)/p^{**}(Z) \in L_2(p^{**})$. Suppose that g satisfies the restrictions of model $\mathcal{M}$ and for some $C \subseteq [K]$,* (i) *$h_j = h_j^*$ for all $j \in C$ and* (ii) *$\Pi_k[M(\beta,h^*);p^{**}] = \Pi_k[M(\beta,h^*);p^*]$ for all $k \in \overline{C}$ and $\beta \in \mathbb{R}^r$. Then*

$$E_p\left(M\{\beta(g),h^*\} - \sum\nolimits_{k=1}^{K}\Pi_k\left[M\{\beta(g),h^*\};p^{**}\right]\right) = E_{p^*}\left[M\{\beta(g),h^*\}\right] = 0. \tag{14}$$

We are now ready to argue why the doubly robust estimator $\tilde{\beta}$ is actually a $(K+1)$-multiply robust estimator. First, notice that because $\Pi_k[D;p]$ depends on g only through $\underline{g}_{k+1} \equiv (g_{k+1},\ldots,g_{K+1})$, we have that $\Pi_k[D;p^{**}] = \Pi_k[D;p^*]$ for all $k \in \{l+1,\ldots,K\}$ when $g_j = g_j^*$ for $j \in \{l+2,\ldots,K+1\}$. Thus, given $l \in \{0,\ldots,K\}$, the assumptions (i) and (ii) of Proposition 1 hold for $C = \{1,\ldots,l\}$ if (a) $h_j = h_j^*$ for all $j \in \{1,\ldots,l\}$ and (b) $g_j = g_j^*$ for $j \in \{l+2,\ldots,K+1\}$ where $\{1,\ldots,l\} \equiv \emptyset$ if $l = 0$ and $\{l+2,\ldots,K+1\} \equiv \emptyset$ if $l = K$. Consequently, we have the following corollary of Proposition 1.

PROPOSITION 2. *Let $p^{**} = g^*h^*$, $p^* = gh^*$ and $p = gh$ be such that $p(Z)/p^*(Z) \in L_2(p^*)$ and $p(Z)/p^{**}(Z) \in L_2(p^{**})$. Suppose that g satisfies the restrictions of model $\mathcal{M}$ and one of the following holds:* (a) *$h_k = h_k^*$ for all $k \in [K]$,* (b) *$g_k = g_k^*$ for all $k \in \{2,\ldots,K+1\}$ or* (c) *there exists $k \in [K-1]$ such that $h_j = h_j^*$ for $j \in \{1,\ldots,k\}$ and $g_j = g_j^*$ for $j \in \{k+2,\ldots,K+1\}$. Then equality (14) holds.*

The $(K+1)$-multiple robustness of the estimator $\tilde{\beta}$ follows from Proposition 2 under regularity conditions. Specifically, suppose that $p = gh$ belongs either to $\bigcap_{j=1}^{K}\mathcal{H}_j$ or to $\bigcap_{j=2}^{K+1}\mathcal{P}_j$ or to $\left(\bigcap_{j=1}^{k}\mathcal{H}_j\right)\cap\left(\bigcap_{j=k+2}^{K+1}\mathcal{P}_j\right)$ for some $k \in \{1,\ldots,K-1\}$. Then, under p and regularity conditions, $\hat{h}$ converges to some h^* and $\hat{g}$ converges to some g^* where h^* and g^* satisfy one of the conditions (a), (b) or (c) of Proposition 2. Thus, $Q\left(\beta,\hat{h},\hat{g}\right)$ converges to $Q(\beta,h^*,g^*) = M(\beta,h^*) - \sum_{k=1}^{K}\Pi_k[M(\beta,h^*);p^{**}]$ where $p^{**} = g^*h^*$, which, by the conclusion (14) of Proposition 2, is an unbiased estimating function for $\beta(g)$.

In computing $\tilde{\beta}$ one specifies fully parametric models for each of the components g_k of g. This modelling strategy yields $\tilde{\beta}$ that is generally inconsistent if for some $k \in [K]$: (A) the model for h_k is incorrect and (B) the parametric model for any single one of the components of $\underline{g}_{k+1} \equiv (g_{k+1},\ldots,g_{K+1})$ is incorrect. This is so because letting $p^{**} = g^*h^*$ with g^* and h^* the

probability limits of $\hat{g}$ and $\hat{h}$, (A) and (B) imply that there can be no subset C of $[K]$ satisfying conditions (i) and (ii) of Proposition 1 and thus $Q\left(\beta, \hat{h}, \hat{g}\right)$ is not guaranteed to converge to an unbiased estimating function for $\beta(g)$. Notice that no such subset C can exist, because if it existed, then k would have to belong to either C or $\overline{C}$. However, k cannot belong to C because $h_k \neq h_k^*$, and k cannot belong to $\overline{C}$ because $\Pi_k\left[M\left(\beta, h^*\right); p^{**}\right] \neq \Pi_k\left[M\left(\beta, h^*\right); p^*\right]$ for $p^* = h^* g$ as $\underline{g}_{k+1} \neq \underline{g}_{k+1}^*$ and $\Pi_k\left[M\left(\beta, h^*\right); p^{**}\right]$ and $\Pi_k\left[M\left(\beta, h^*\right); p^*\right]$ depend on $\underline{g}_{k+1}^*$ and $\underline{g}_{k+1}$ respectively.

In the next section we will argue that parametric models for the conditional densities g_k ($k \in [K]$) are unnecessarily restrictive in the sense that consistent estimators of $\beta(g)$ exist under models that parameterize just some components of g_k ($k \in [K]$) but not the entire conditional laws.

3·3. *Multiply robust estimating functions for the parameter of interest*

Proposition 1 and the analysis of the reasons for inconsistency of $\tilde{\beta}$ when (A) and (B) hold invite one to explore the existence of functionals of $\underline{g}_{k+1}$ that are non-injective and do not depend on h, such that $E_p\left\{M\left(\beta, h\right) \mid \overline{A}_k, \overline{L}_k\right\}$ or $E_p\left[M\left\{\beta(g), h\right\} \mid \overline{A}_k, \overline{L}_k\right]$, and consequently $\Pi_k\left[M\left(\beta, h\right); p\right]$ or $\Pi_k\left[M\left\{\beta(g), h\right\}; p\right]$, depend on g only through those functionals. The hope is that, by modelling such functionals for all $k \in [K]$ instead of the entire g, one could perhaps find an estimator of $\beta(g)$ that confers even more robustness with respect to model misspecification than $\tilde{\beta}$.

Such non-injective functionals do indeed exist in Examples 1 and 2. This point is seen in Example 1 after noticing that the conditional expectation of $M\left(\beta, h\right)$ can be written as

$$E_p\left\{M\left(\beta, h\right) \mid \overline{A}_k, \overline{L}_k\right\} = \prod_{j=1}^{k}\left\{A_j / E_{h_j}\left(A_j \mid \overline{A}_{j-1}, \overline{L}_j\right)\right\}\left\{\rho_k\left(g\right) - \beta\right\}, \tag{15}$$

where $\rho_k\left(g\right)$ $(k = 1, 2)$ are defined in (3). Thus, we see that the conditional expectation in (15) depends on g through the non-injective functionals $\rho_k\left(g\right)$ $(k = 1, 2)$. These functionals were precisely the ones modelled to arrive at the multiply robust estimator of Example 1.

In Example 2, the conditional expectations of $M\left(\beta, h\right)$ are

$$E_p\left\{M\left(\beta, h\right) \mid \overline{A}_2, \overline{L}_2\right\} = \left\{\rho_2(g) - \beta A_1\right\}\left\{A_1 - E_{h_1}\left(A_1 \mid L_1\right)\right\} A_2 / E_{h_2}(A_2 \mid A_1, \overline{L}_2), \tag{16}$$

$$E_p\left\{M\left(\beta, h\right) \mid A_1, L_1\right\} = \left[\rho_1(g) - \left\{\beta - \beta(g)\right\} A_1\right]\left\{A_1 - E_{h_1}\left(A_1 \mid L_1\right)\right\}, \tag{17}$$

where $\rho_2\left(g\right) = E\left(L_3 \mid A_2 = 1, \overline{L}_2, A_1\right)$ and $\rho_1\left(g\right)$ is the nonparametric component $\nu_p\left(\cdot\right)$ in model (5) . Observe that $E_p\left\{M\left(\beta, h\right) \mid \overline{A}_2, \overline{L}_2\right\}$ in (16) depends on g only through $\rho_2(g)$ and $E_p\left\{M\left(\beta, h\right) \mid A_1, L_1\right\}$ in (17) evaluated at $\beta = \beta(g)$ depends on g only through $\rho_1(g)$. The functionals $\rho_1(g)$ and $\rho_2(g)$ were precisely the ones modelled to construct the multiply robust estimator of Example 2.

We then see that we can arrive at the hoped-for functionals by exploring the form of $E_p\left\{M\left(\beta, h\right) \mid \overline{A}_k, \overline{L}_k\right\}$ $(k = 1, 2)$. Interestingly, in (17) the mean of $M\left(\beta, h\right)$ depends on g through $\rho_1(g)$ and also through the unknown $\beta(g)$,whereas the mean of $M\left\{\beta(g), h\right\}$ depends on g only through $\rho_1(g)$. In contrast, in equation (16), the mean of $M\left(\beta, h\right)$ does not depend on $\beta(g)$ but the mean of $M\left\{\beta(g), h\right\}$ does. These examples illustrate that we can eliminate the dependence on the unknown $\beta(g)$ by modelling sometimes $E_p\left\{M\left(\beta, h\right) \mid \overline{A}_k, \overline{L}_k\right\}$, i.e., the mean of $M\left(\beta, h\right)$ with β unspecified, and sometimes $E_p\left[M\left\{\beta(g), h\right\} \mid \overline{A}_k, \overline{L}_k\right]$, i.e., the mean of $M\left(\beta, h\right)$ with β fixed at $\beta(g)$.

In formulating the general theory that we will develop next, we will assume that there exist functionals $\eta_k(g)$ $(k \in [K])$, independent of β and h, such that $E_p\left\{M(\beta,h) \mid \overline{A}_k, \overline{L}_k\right\}$ depends on g only through $\eta_k(g)$. Whereas we could write this assumption as $E_p\left\{M(\beta,h) \mid \overline{A}_k, \overline{L}_k\right\} = e_k\{\eta_k(g), \beta, h\}$ for some e_k, we will instead write it as

$$E_p\left\{M(\beta,h) \mid \overline{A}_k, \overline{L}_k\right\} = e_k[\psi_k\{\eta_k(g), \beta\}, h] \tag{18}$$

for some e_k and ψ_k. This is without loss of generality, as ψ_k may be taken to be the identity, but writing the assumption as (18) will allow us to distinguish situations in which we model the dependence on g of the conditional expectation of $M(\beta,h)$ with β fixed at $\beta(g)$, corresponding to modelling $\psi_k\{\eta_k(g), \beta(g)\}$, from situations in which we model the dependence on g of the expectation of $M(\beta,h)$ with β unspecified, corresponding to modelling $\eta_k(g)$. While our results will be valid whether or not $\eta_k(g)$ is an injective functional of g, they will only be useful for coming up with estimators that confer more robustness to model misspecification than the $(K+1)$-multiply robust estimator $\tilde{\beta}$ when, for some $k \in [K]$, $\eta_k(g)$ is non-injective.

To clarify the meaning of η_k and ψ_k, consider again Examples 1 and 2. Examining the identity (15), we see that (18) holds in Example 1 if we define $\eta_k(g) \equiv \rho_k(g)$ and $\psi_k(\cdot,\cdot)$ to be the identity function $(k = 1, 2)$. The models adopted to construct the multiply robust estimator of Example 1 are models for $\eta_k(g)$ $(k = 1, 2)$, i.e., for the dependence of $\Pi_k[M(\beta,h); p]$ on g for any β. On the other hand, upon examining identity (16) we see that (18) holds if we define $\eta_2(g) \equiv \rho_2(g)$ and $\psi_2(\cdot,\cdot)$ to be the identity function, whereas by examining identity (17) we see that (18) holds if we define $\eta_1(g) \equiv (\rho_1(g), \beta(g))$ and $\psi_1\{\eta_1(g), \beta\} \equiv \rho_1(g) - \{\beta(g) - \beta\}A_1$. Note that $\psi_1\{\eta_1(g), \beta(g)\} \equiv \rho_1(g)$ is the functional that governs the dependence on g of $E_p\{M(\beta,h) \mid A_1, L_1\}$ with β fixed at $\beta(g)$. This is precisely the functional modelled to construct the multiply robust estimator of Example 2.

Recall that in Example 2, $\mathcal{G}$ is the set $\{g : g \text{ satisfies } (5)\}$. However, in this example, $\eta_2(g) = \rho_2(g) = E_g\left(L_3 \mid A_2 = 1, \overline{L}_2, A_1\right)$ is indeed defined on the larger set $\mathcal{G}_2 = \{g : g \text{ unrestricted}\}$. Furthermore,

$$e_2[\psi_2\{\eta_2(g), \beta\}, h] \equiv \left\{A_2/E_{h_2}(A_2 \mid A_1, \overline{L}_2)\right\}\{\rho_2(g) - \beta A_1\}\left\{A_1 - E_{h_1}(A_1 \mid L_1)\right\}$$

satisfies (18) for $g \in \mathcal{G}$ but is also well defined for all g in $\mathcal{G}_2$ and is a random variable that depends only on $\left(\overline{A}_2, \overline{L}_2\right)$. This, in turn, implies that for $k = 2$,

$$h_k = h_k^* \Rightarrow E_{gh}\left(\pi_k\left[\psi_k\{\eta_k(g), \beta\}, h^*\right]\right) = 0 \text{ for any } g \text{ in } \mathcal{G}_k \tag{19}$$

where

$$\pi_k[\psi_k\{\eta_k(g), \beta\}, h] \equiv e_k[\psi_k\{\eta_k(g), \beta\}, h] - E_{h_k}\left(e_k[\psi_k\{\eta_k(g), \beta\}, h] \mid \overline{A}_{k-1}, \overline{L}_k\right). \tag{20}$$

That property (19) holds on a set $\mathcal{G}_k$ that possibly includes $\mathcal{G}$ strictly is important to ensure the multiple robustness of the estimator $\hat{\beta}$ in Example 2 even when the model for $\rho_2(g)$ is incompatible with the model (5). For instance, in Example 2, even in the extreme case in which one were to specify a model $\mathcal{R}_2 = \left\{\rho_{2,\tau_2} = \tau_2 L_1 A_1 : \tau_2 > 0\right\}$ for $\rho_2(g) = E\left(L_3 \mid A_2 = 1, \overline{L}_2, A_1\right)$, which cannot be true for any g satisfying (5), the estimator $\hat{\beta}$ would be consistent for $\beta(g)$ provided model $\mathcal{H}_2$ for h_2 is correct and either model $\mathcal{R}_1$ for g or model $\mathcal{H}_1$ for h_1 is correct.

Returning to the formulation of the general theory, define $\pi_k[\psi_k\{\eta_k(g), \beta\}, h]$ as in (20). With this definition, for any $g \in \mathcal{G}$ and $p = gh$, one has that

$$\Pi_k[M(\beta,h); p] = \pi_k[\psi_k\{\eta_k(g), \beta\}, h]. \tag{21}$$

Examples 1 and 2 illustrate that given the decomposition (18) under our general setting, for each k one may choose to model either $\psi_k\{\eta_k(g), \beta(g)\}$ or $\eta_k(g)$. That is, one may choose to model $\psi_k\{\eta_k(g), \beta(g)\}$ for all k in some subset E of $[K]$ and model $\eta_k(g)$ for all k not in E, i.e., to model functionals defined as

$$\rho_k(g) \equiv \begin{cases} \psi_k\{\eta_k(g), \beta(g)\}, & k \in E, \\ \eta_k(g), & k \notin E. \end{cases} \tag{22}$$

Until § 4, we will assume that the decomposition (18) holds for some e_k, ψ_k and η_k and all g in a set $\mathcal{G}_k$, possibly larger than $\mathcal{G}$, and $\rho_k(g)$ is defined as in (22). Having defined the target functional $\rho_k(g)$ for each $k \in [K]$, we now define a parametric model

$$\mathcal{R}_k \equiv \left\{\rho_{k,\tau_k} : \tau_k \in \Upsilon_k\right\},$$

where Υ_k lies in some Euclidean space. We assume that $\mathcal{R}_k \subset \{\rho_k(g) : g \in \mathcal{G}_k\}$; the inclusion ensures that $e_k\left(\rho_{k,\tau_k}, h\right)$ $(k \in E)$ and $e_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h\right\}$ $(k \notin E)$ are functions of $\left(\overline{A}_k, \overline{L}_k\right)$ for all $\tau_k \in \Upsilon_k$ and all $\beta \in \mathbb{R}^r$. Defining

$$\begin{aligned} \pi_k\left(\rho_{k,\tau_k}, h\right) &\equiv e_k\left(\rho_{k,\tau_k}, h\right) - E_{h_k}\left\{e_k\left(\rho_{k,\tau_k}, h\right) \mid \overline{A}_{k-1}, \overline{L}_k\right\}, && k \in E, \\ \pi_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h\right\} &\equiv e_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h\right\} - E_{h_k}\left[e_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h\right\} \mid \overline{A}_{k-1}, \overline{L}_k\right], && k \notin E, \end{aligned}$$

we conclude that the following condition holds, where $\Lambda_k(p)$ is defined in (11).

Condition 1. There exist maps π_k $(k \in [K])$ satisfying (21) such that for any $\tau_k \in \Upsilon_k$, any $\beta \in \mathbb{R}^r$ and any $p = gh$, (i) $\pi_k\left(\rho_{k,\tau_k}, h\right) \in \Lambda_k(p)$ for $k \in E$, and (ii) $\pi_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h\right\} \in \Lambda_k(p)$ for $k \notin E$.

Observing that the elements of $\Lambda_k(p)$ have mean zero under p, we conclude that for $p = gh$ and $p^* = gh^*$ such that $p(Z)/p^*(Z) \in L_2(p^*)$,

$$h_k = h_k^* \Rightarrow \begin{cases} E_p\left\{\pi_k\left(\rho_{k,\tau_k}, h^*\right)\right\} = 0, & k \in E, \\ E_p\left[\pi_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h^*\right\}\right] = 0, & k \notin E; \end{cases} \tag{23}$$

this restates (19) in terms of the parameterization of model $\mathcal{R}_k$ for $\rho_k(g)$.

Next, suppose that in addition to models $\mathcal{R}_k$ for $\rho_k(g)$, we postulate models $\mathcal{H}_k$ as in (8) for h_k, $k \in [K]$. Define

$$U(\beta, \tau, \alpha) \equiv M(\beta, h_\alpha) - \sum\nolimits_{j \in E} \pi_j\left(\rho_{j,\tau_j}, h_\alpha\right) - \sum\nolimits_{j \notin E} \pi_j\left\{\psi_j\left(\rho_{j,\tau_j}, \beta\right), h_\alpha\right\}, \tag{24}$$

where $h_\alpha \equiv \left(h_{1,\alpha_1}, \ldots, h_{K,\alpha_K}\right)$. The right-hand side of (24) yields the estimating function U of Examples 1 and 2, taking $E = \emptyset$ in Example 1 and $E = \{1\}$ in Example 2.

We can now invoke Theorem 1 and (23) to argue that, as in Examples 1 and 2, $U(\beta, \tau, \alpha)$ must satisfy $E_p[U\{\beta(g), \tau, \alpha\}] = 0$ whenever $\rho_j(g) \in \mathcal{R}_j$ or $h_j \in \mathcal{H}_j$ for each $j \in [K]$, for $p = gh$, and the parameters indexing the correctly specified dimension-reducing models are replaced by their true values. To give a precise statement we write, as in § 2, $p \in \mathcal{R}_k$ if $\rho_k(g) \in \mathcal{R}_k$ and for such g we define $\tau_k(g)$ such that $\rho_{k,\tau_k(g)} = \rho_k(g)$; likewise, if $p \in \mathcal{H}_j$ we define $\alpha_j\left(h_j\right)$ such that $h_{j,\alpha_j(h_j)} = h_j$, and give the following definition.

DEFINITION 1. *We say that $U(\beta,\tau,\alpha)$ is a multiply robust estimating function for $\beta(g)$ in the intersection-union model $\bigcap_{k=1}^{K}(\mathcal{H}_k \cup \mathcal{R}_k)$ if for any $C \subseteq [K]$ and*

$$p \in \mathcal{M}(C) \equiv \left\{p = gh \in \mathcal{M} : p \in \mathcal{H}_l \text{ for } l \in C \text{ and } p \in \mathcal{R}_s \text{ for } s \in \overline{C}\right\},$$

we have that $E_p\{U(\beta,\tau,\alpha)\} = 0$ at $\beta = \beta(g), \alpha_l = \alpha_l(h_l)$, and $\tau_s = \tau_s(g)$, for $l \in C$ and $s \in \overline{C}$.

We then have the following result.

LEMMA 1. *Suppose that for all $p \equiv gh \in \mathcal{M}$ and all α, it is true that $p(Z)/p_\alpha(Z) \in L_2(p_\alpha)$, where $p_\alpha \equiv gh_\alpha$. Then $U(\beta,\tau,\alpha)$ defined as in (24) is multiply robust for $\beta(g)$ in model $\bigcap_{k=1}^{K}(\mathcal{H}_k \cup \mathcal{R}_k)$.*

Lemma 1 is stronger than Proposition 2 because it establishes that by modelling functions of $\underline{g}_{j+1}$ and not necessarily the entire $\underline{g}_{j+1}$, one can obtain estimators that confer more protection against model misspecification than the $(K+1)$-multiply robust estimators in § 3·2.

We now to highlight a few points about the structure of model $\bigcap_{k=1}^{K}(\mathcal{H}_k \cup \mathcal{R}_k)$. Observe that this model is the same as $\bigcup_{C:C\subset[K]}\mathcal{M}(C)$. When no $\mathcal{M}(C)$ is included in another $\mathcal{M}(C')$ and no $\mathcal{M}(C)$ is empty, the union comprises 2^K distinct models. In such a case, we would hope that an estimation strategy exploiting $U(\beta,\tau,\alpha)$ would confer 2^K modelling options for arriving at valid inference about $\beta(g)$. Our formulation, however, allows the possibility that $\mathcal{M}(C) \subset \mathcal{M}(C')$ for some $C \neq C'$ or that some $\mathcal{M}(C)$ is empty. The latter may occur either because for some $k \in \overline{C}$ there is no p in $\mathcal{M} \cap \mathcal{R}_k$, or because for some $k, k' \in \overline{C}$ models $\mathcal{R}_k$ and $\mathcal{R}_{k'}$ are incompatible in the sense that there is no p in both $\mathcal{R}_k$ and $\mathcal{R}_{k'}$. Yet, even under such scenarios, model $\bigcap_{k=1}^{K}(\mathcal{H}_k \cup \mathcal{R}_k)$ is never empty, because it includes $\bigcap_{k=1}^{K}\mathcal{H}_k$, which is non-empty since models for different h'_k are always compatible. Thus, even under a worst model incompatibility scenario, a multiply robust estimating function is always unbiased for $\beta(g)$ in model $\bigcap_{k=1}^{K}\mathcal{H}_k$.

3·4. *Multiply robust estimating functions for nuisance parameters*

We now construct K different estimating functions $T_k(\beta,\tau,\alpha)$ $(k \in [K])$, distinct from $U(\beta,\tau,\alpha)$, such that, as in Examples 1 and 2, each T_k is multiply robust for $\{\beta(g), \tau_k(g)\}$ in $\bigcap_{j=1,j\neq k}^{K}(\mathcal{H}_j \cup \mathcal{R}_j)$. More precisely, $T_k(\beta,\tau,\alpha)$ $(k \in [K])$ is a system of K nuisance multiply robust estimating functions as defined next.

DEFINITION 2. *A collection $\{T_k(\beta,\tau,\alpha) : k \in [K]\}$ is a set of nuisance multiply robust estimating functions in model $\bigcap_{k=1}^{K}(\mathcal{H}_k \cup \mathcal{R}_k)$ if for any $C \subseteq [K]$, any $p \in \mathcal{M}(C)$ and any $k \in \overline{C}$, i.e., any k such that model $\mathcal{R}_k$ is correctly specified under p, we have that $E_p\{T_k(\beta,\tau,\alpha)\} = 0$ at $(\beta,\tau_k) = \{\beta(g), \tau_k(g)\}$, $\alpha_l = \alpha_l(h_l)$ and $\tau_s = \tau_s(g)$ for $l \in C$ and $s \in \overline{C}$.*

The usefulness of nuisance multiply robust estimating functions is that, as Examples 1 and 2 illustrate, we can use them to form a system of estimating equations in (β,τ),

$$P_n\left\{U\left(\beta,\tau,\hat{\alpha}\right)\right\} = 0, \quad P_n\left\{T_k\left(\beta,\tau,\hat{\alpha}\right)\right\} = 0 \quad (k \in [K]).$$

Under regularity conditions, such a system should have a solution $\hat{\tau} = \left(\hat{\tau}_1, \ldots, \hat{\tau}_K\right)$ for τ such that $\hat{\tau}_k$ converges in probability to $\tau_k(g)$ whenever $p \in \mathcal{M}(C)$, provided $k \in \overline{C}$. The solution

$\hat{\beta}$ of the system would then be multiply robust for $\beta(g)$ in the sense that it would converge in probability to $\beta(g)$ whenever $p \in \mathcal{M}(C)$ for some $C \subseteq [K]$.

Under the assumptions of this section it is indeed possible to come up with nuisance multiply robust estimating functions. To construct $T_k(\beta, \tau, \alpha)$ we first exhibit an $r_k \times 1$ random function $M_k(\beta, \tau_k, h)$ of (β, τ_k), with $r_k \equiv \dim(\tau_k)$, such that when $p \in \mathcal{M}$ and $p \in \mathcal{R}_k$, $\{\beta(g), \tau_k(g)\}$ solves the equation $E_p\{M_k(\beta, \tau_k, h)\} = 0$. Given $U(\beta, \tau, \alpha)$ defined in (24) for a choice of E, one such function is

$$M_k(\beta, \tau_k, h) \equiv \begin{cases} d_k\left(\overline{A}_k, \overline{L}_k\right)\left\{M(\beta, h) - e_k\left(\rho_{k,\tau_k}, h\right)\right\}, & k \in E, \\ d_k\left(\overline{A}_k, \overline{L}_k\right)\left[M(\beta, h) - e_k\left\{\psi_k\left(\rho_{k,\tau_k}, \beta\right), h\right\}\right], & k \notin E, \end{cases} \tag{25}$$

where $d_k\left(\overline{A}_k, \overline{L}_k\right)$ is a given $r_k \times r$ matrix-valued function of $\left(\overline{A}_k, \overline{L}_k\right)$. Next, just as earlier starting from $M(\beta, h)$ we constructed a multiply robust estimating function $U(\beta, \tau, \alpha)$ for $\beta(g)$, now starting from $M_k(\beta, \tau_k, h)$ we construct a multiply robust estimating function $T_k(\beta, \tau, \alpha)$ for $\{\beta(g), \tau_k(g)\}$. We first note that if condition (18) holds then

$$\begin{array}{ll} \Pi_j[M_k\{\beta(g), \tau_k(g), h\}; p] = 0, & j \in [k], \\ \Pi_j[M_k(\beta, \tau_k, h); p] = d_k\left(\overline{A}_k, \overline{L}_k\right)\pi_j[\psi_j\left\{\eta_j(g), \beta\right\}, h], & j \in [K] - [k]. \end{array} \tag{26}$$

We thus note that the right-hand sides in (25) and (26) depend on g at most through $\eta_k(g)$, i.e., no new functionals of g emerge, so we can reproduce the construction in § 3·3 without having to model new functionals of g. Specifically, having already constructed a function U from M as in (24) using a given set E, we now use the same recipe to construct a function T_k from M_k, choosing to model $E\left[M_k\{\beta(g), \tau_k(g), h\} \mid \overline{A}_j, \overline{L}_j\right]$ for all j in $E^* \equiv \{1, \ldots, k\} \cup E$. This yields

$$\begin{aligned} T_k(\beta, \tau, \alpha) \equiv\ & M_k(\beta, \tau_k, h_\alpha) \\ & - d_k\left(\overline{A}_k, \overline{L}_k\right)\left[\sum\nolimits_{j \in E, j > k} \pi_j\left(\rho_{j,\tau_j}, h_\alpha\right) + \sum\nolimits_{j \notin E, j > k} \pi_j\left\{\psi_j\left(\rho_{j,\tau_j}, \beta\right), h_\alpha\right\}\right]. \end{aligned} \tag{27}$$

Recall that E is the set of indexes $j \in [K]$ for which a model for $E\left\{M(\beta, h) \mid \overline{A}_j, \overline{L}_j\right\}$ with β fixed at $\beta(g)$, rather than unspecified, is posited in constructing $U(\beta, \tau, \alpha)$ in (24). Because of (26), choosing to model $E\left\{M_k(\beta, \tau_k, h) \mid \overline{A}_j, \overline{L}_j\right\}$ with (β, τ_k) fixed at $\{\beta(g), \tau_k(g)\}$ not only for j in E but also for j in $\{1, \ldots, k\}$, i.e., for all $j \in E^*$, makes the right-hand side of (27) not involve a summation from 1 to k. In fact, T_k depends on $\tau = (\tau_1, \ldots, \tau_K)$ only through $(\tau_k, \ldots, \tau_K)$, but we do not make this explicit in the argument of T_k to simplify notation. The following result mimics Lemma 1.

LEMMA 2. *Suppose that for all $p \equiv gh \in \mathcal{M}$ and all α, $p(Z)/p_\alpha(Z) \in L_2(p_\alpha)$ where $p_\alpha \equiv gh_\alpha$. Then $\{T_k(\beta, \tau, \alpha) : k \in [K]\}$ is a set of nuisance multiply robust estimating functions for $\beta(g)$.*

The right-hand side of (27), with specific choices of d_k that possibly depend also on α and τ, recovers the functions T_k $(k = 1, 2)$ of Examples 1 and 2.

4. FACTORIZED LIKELIHOOD MODELS

4·1. *Main results*

In this section we exhibit conditions, more general than those of § 3, under which it is possible to extend several of the results of that section. Our goal is to derive a theory that explains, for instance, the existence of doubly and multiply robust estimators in settings with nonmonotone, i.e., intermittent, missing-at-random data. In such settings the density of the observed data can be factorized as $p = gh$ where only g depends on the distribution of the full, i.e., intended, data. Consequently, inference about parameters of the full data distribution is tantamount to inference about parameters $\beta(g)$ that, as in § 3, depend on p only through g. However, unlike § 3, with nonmonotone missing-at-random data, neither g nor h are marginal or conditional densities of the observed data. Thus, the theory of § 3 does not apply. In the Supplementary Material we provide a motivating example.

Suppose that Z is a random element taking values in some set $\mathcal{Z}$ whose density p belongs to a specified set $\mathcal{M}$, and that all the densities in $\mathcal{M}$ are strictly positive on $\mathcal{Z}$. Assume also that there exist functionals $H_1, \ldots, H_K$ and G on $\mathcal{M}$ such that for any $p \in \mathcal{M}$,

$$p = G(p) \times H_1(p) \times \cdots \times H_K(p),$$

where $g \equiv G(p)$ and $h_k \equiv H_k(p)$ $(k \in [K])$ are real-valued, positive functions on the sample space of Z which vary independently on $\mathcal{M}$, i.e., for any set of densities $p_0, p_1, \ldots, p_K$ in $\mathcal{M}$, the product $G(p_0) \times H_1(p_1) \times \cdots \times H_K(p_K)$ is also in $\mathcal{M}$. As in § 3, in the following $p = gh$ stands for $p = g \times h_1 \times \cdots \times h_K$.

For given $k \in [K]$, g and $\{h_j\}_{j:j \neq k}$ consider the submodel of $\mathcal{M}$ in which g and $\{h_j\}_{j:j \neq k}$ are fixed, i.e.,

$$\mathcal{M}_k \equiv \{p^* \in \mathcal{M} : G(p^*) = g,\ H_j(p^*) = h_j,\ j \neq k\}.$$

Given $p \in \mathcal{M}_k$ we now let $\Lambda_k(p)$ denote the maximal tangent space of submodel $\mathcal{M}_k$ at p, i.e., $\Lambda_k(p)$ is the $L_2(p)$-closed linear span of scores at $t = 0$ for regular one-dimensional parametric submodels

$$\mathcal{M}_{k,\text{par}} \equiv \{p_k(\cdot; t) \equiv g(\cdot)\, h_1(\cdot) \cdots h_{k-1}(\cdot)\, h_{k+1}(\cdot) \cdots h_K(\cdot)\, h_k(\cdot; t) : t \in [0, \varepsilon)\}, \tag{28}$$

with $p(z) = p_k(z; 0)$ (van der Vaart 2000, §25.3). In addition, for any $W \equiv w(Z) \in L_2(p)$ we let $\Pi_k[W; p]$ denote the $L_2(p)$-projection of W into $\Lambda_k(p)$. Also, $\Pi_k[W; p]$ stands for $(\Pi_k[W_1; p], \ldots, \Pi_k[W_r; p])^{\mathrm{T}}$ if $W = (W_1, \ldots, W_r)^{\mathrm{T}}$.

The setting of §3 is a special case of the present one, with

$$G(p) \equiv G_1(p) \times \cdots \times G_{K+1}(p), \tag{29}$$

where $G_k(p)(z) \equiv p\left(l_k \mid \bar{l}_{k-1}, \bar{a}_{k-1}\right)$ and $H_k(p)(z) \equiv p\left(a_k \mid \bar{l}_k, \bar{a}_{k-1}\right)$. In the setting of §3, for each k the tangent space $\Lambda_k(p)$ is equal to the set of square-integrable measurable real-valued functions of $\left(\bar{A}_k, \bar{L}_k\right)$ with conditional mean zero given $\left(\bar{A}_{k-1}, \bar{L}_k\right)$ under h_k, i.e., precisely the set defined in (11). In addition, the projection $\Pi_k[D; p]$ is given by (10).

Let us now discuss an extension of Theorem 1 to the present framework. Inspection of its proof reveals that the following condition suffices to arrive at the identity (12) .

Condition 2. The densities $p = gh$ and $p^* = gh^*$ are in $\mathcal{M}$ and satisfy: (i) $p(Z)/p^*(Z) \in L_2(p^*)$ and (ii) there exist s_k^* ($k \in [K]$) satisfying:

$$(\mathrm{C}\cdot 0) \quad p/p^* - 1 = \sum\nolimits_{k=1}^{K} s_k^*, \tag{30}$$

(C·1) $S_k^* \equiv s_k^*(Z) \in \Lambda_k(p^*)$ and (C·2) S_k^* is uncorrelated under p^* with all the elements of $\Lambda_j(p^*)$ ($j \neq k$).

We then have the following extension of Theorem 1.

THEOREM 2. *Suppose that p and p^* satisfy Condition 2. Then, for any random variable $D = d(Z)$ such that $E_{p^*}(D^2) < \infty$, equality (12) holds.*

The preceding theorem raises the question of when Condition 2 will hold. Now, $p(z)/p^*(z) - 1$ is the derivative at $t = 0$ of

$$t \to \log\{tp(z) + (1-t)p^*(z)\}, \tag{31}$$

which maps each $t \in [0,1]$ to the logarithm of the mixture of p and p^* with mixing probability t. This suggests that Condition 2 would hold whenever the submodel of $\mathcal{M}$ that keeps g fixed is convex, i.e., whenever for any $p = gh$ and $p^* = gh^*$ as in the assumptions of Lemma 1 and any $t \in [0,1]$, there exists $h_k(\cdot;t)$ in the range of H_k ($k \in [K]$) such that

$$tp(z) + (1-t)p^*(z) = g(z)\prod\nolimits_{k=1}^{K} h_k(z;t). \tag{32}$$

An informal justification for this is as follows. Differentiating the map (31) with respect to t at $t = 0$, we should obtain from the identity (32) that requirement (C·0) of Condition 2 holds with $s_k^*(z) = \mathrm{d}\log h_k(z;t)/\mathrm{d}t|_{t=0}$. Next, assuming that s_k^* is the score of the parametric submodel $\mathcal{M}_{k,\mathrm{par}}^* \equiv \{p_k^*(\cdot;t) : t \in [0,\varepsilon)\}$ defined as in (28) but with h_j replaced by h_j^* for $j \in [K] - \{k\}$ and $h_k(\cdot;t)$ as in (32), we would conclude that s_k^* ($k \in [K]$) satisfies requirement (C·1) of Condition 2. Finally, since scores corresponding to models that parameterize separate factors of the likelihood are often orthogonal, we would arrive at the conclusion that s_k^* ($k \in [K]$) satisfies requirement (C·2) of Condition 2.

The preceding informal justification is not rigorous for the following reasons. The condition $p(Z)/p^*(Z) \in L_2(p^*)$ suffices to ensure that the model $\{tp + (1-t)p^* : t \in [0,\varepsilon)\}$ is regular and its score at $t = 0$ is $p/p^* - 1$, as proved in Lemma 2 of the Supplementary Material. However, these conditions do not ensure that for $k \in [K]$ the model $\mathcal{M}_{k,\mathrm{par}}^*$ is regular; even if this model were regular, the conditions would not ensure that its score at $t = 0$ would be computed with the derivative of the logarithm of $p_k^*(z;t)$. Observe, for instance, that these conditions do not even imply the differentiability of the map $t \to \log h_k(z;t)$. Lemma 7·6 of van der Vaart (2000) gives conditions for the model $\mathcal{M}_{k,\mathrm{par}}^*$ to be regular with score $s_k^*(z)$ at $t = 0$ equal to $\mathrm{d}\log h_k(z;t)/\mathrm{d}t|_{t=0}$.

In § 4·4 we define the property of sequentially strong convexity of a model that implies, but is not implied by, the submodel convexity as defined in (32). We prove that under this property, the conditions of Lemma 7.6 of van der Vaart (2000) are satisfied and thus the decomposition (30) holds with s_k^* fulfilling (C·1) of Condition 2. Furthermore, we provide regularity conditions that additionally ensure (C·2), thus implying Condition 2. In § 4·4 we also establish that the model of § 3 is sequentially strongly convex.

Next, suppose that there exists $M(\beta, h) \equiv m(Z; \beta, h)$, a known $\mathbb{R}^r$-valued function of $(Z; \beta, h)$ with $\beta \in \mathbb{R}^r$ and each h_k in the range of H_k on $\mathcal{M}$, such that for all $p = gh$ in $\mathcal{M}$ the equation (7) has a unique solution $\beta(g)$ that depends on p only through g. We now wish to extend Proposition 1, which, as indicated in § 3·2, follows immediately from Theorem 1 and the assertion in (13). Having extended Theorem 1, in order to extend Proposition 1 we now consider the following condition which ensures that (13) holds.

Condition 3. For any p and p^* such that $p(Z)/p^*(Z) \in L_2(p^*)$, the elements of $\Lambda_k(p^*)$ have mean zero under p whenever $h_k = h_k^*$.

The validity of Condition 3 under the present general formulation holds, for instance, if p/p^* is bounded; see the Supplementary Material, Corollary 1.

An argument identical to that preceding Proposition 1 provides a proof of the following extension of that proposition to the present general setting.

PROPOSITION 3. *Let $p^{**} = g^*h^*, p^* = gh^*$ and $p = gh$ lie in $\mathcal{M}$. Suppose that conditions* (i) *and* (ii) *of Proposition 1 hold for some $C \subseteq [K]$. Suppose also that p and p^* satisfy Condition 2 and p and p^{**} satisfy Condition 3 with p^{**} in the role of p^*. Then* (14) *holds.*

Remark 1. When Proposition 3 is combined with Corollary 2 stated in the next section, it yields a result that generalizes Lemma 1 of Robins et al. (2000) from $K = 1$ to $K > 1$.

Given a representation of the projection of the form (21), we now follow the reasoning of §3·3. Specifically, we postulate parametric models $\mathcal{H}_k$ for h_k indexed by α_k and models $\mathcal{R}_k$ indexed by τ_k for functionals $\rho_k(g)$ defined as in (22), with $\mathcal{G}$ being the range of G on $\mathcal{M}$ and $\mathcal{G}_k$ either equal to $\mathcal{G}$ or a properly defined set larger than $\mathcal{G}$. The function $U(\beta, \tau, \alpha)$ defined in (24) now satisfies the following extension of Lemma 1. Its proof is identical to the proof of Lemma 1.

LEMMA 3. *Assume that Condition 1 holds and that for any α, $p = gh$ and $p_\alpha \equiv gh_\alpha$ satisfy Conditions 2 and 3 with p_α in the role of p^*. Then $U(\beta, \tau, \alpha)$ defined as in* (24) *is a multiply robust estimating function for $\beta(g)$ in model $\bigcap_{k=1}^{K}(\mathcal{H}_k \cup \mathcal{R}_k)$.*

Unfortunately, under the present level of generality, we cannot extend Lemma 3 to provide conditions for the construction of multiply robust nuisance estimating functions. In fact, it may well happen that for the chosen models $\mathcal{R}_k$ ($k \in [K]$), such estimating functions do not exist as consistent estimation of $\tau_k(g)$ may require the estimation of high-dimensional functionals other than those in $\{\rho_j(g) : j \in [K]\}$. We can nevertheless exploit Lemma 3 to come up with doubly and $(K+1)$-multiply robust estimators of $\beta(g)$, as discussed in the following two subsections.

4·2. *Double robustness*

Suppose $K = 1$. In the decomposition (21) we can always take $\eta_1(g) = g$. In such a case, we can define $\rho_1(g) = g$ and consider a parametric model $\mathcal{R}$ for g. Next, we can compute $\hat{g}$, the maximum likelihood estimator of g under $\mathcal{R}$, and $\hat{h}$, the maximum likelihood estimator of h under a parametric model $\mathcal{H}$ for h. Under the assumptions of Lemma 3 and regularity conditions, the equation $P_n\left\{Q\left(\beta, \hat{h}, \hat{g}\right)\right\} = 0$, where $Q(\beta, h, g)$ is as defined in (9) with $K = 1$, has a solution $\overline{\beta}$ that converges in probability to $\beta(g)$ so long as one of the following holds: $p \in \mathcal{R}$ or $p \in \mathcal{H}$. This result, for instance, recovers the double robustness of locally efficient estimators in models for nonmonotone missing data discussed in Chapter 10 of Tsiatis (2006). Such models are an instance in which g is not a marginal density. For constructing the doubly robust locally

efficient estimators with nonmonotone missing data, there does not exist a closed-form analytic expression for $\Pi_1[M(\beta,h);p]$. However, in that case, one can use a successive approximations algorithm to obtain a surrogate $\tilde{\Pi}_1[M(\beta,h);p]$ close to $\Pi_1[M(\beta,h);p]$ in $L_2(p)$ (see, e.g., Lemma 10·5 of van der Laan & Robins, 2003).

4·3. $(K+1)$-*multiple robustness*

Inspection of the argument leading to Proposition 2 reveals that under the present general setting the proposition remains valid if, in addition to Conditions 2 and 3, the following condition holds.

Condition 4. G satisfies (29). In addition, $\Pi_k[W;p]$ depends on p only though h and $\underline{g}_{k+1} \equiv (g_{k+1},\ldots,g_{K+1})$ where $g_k \equiv G_k(p)$ for $k \in [K+1]$.

We therefore have the following extension of Proposition 2 to the present setting.

PROPOSITION 4. *With the definitions of the present section, Proposition 2 is valid if Conditions 2, 3 and 4 hold.*

Under the assumptions of Proposition 4 and regularity conditions, mimicking the argument as in § 3·2, we conclude that $\tilde{\beta}$ solving $P_n\left\{Q\left(\beta,\hat{h},\hat{g}\right)\right\}=0$, where $\hat{g}_k$ and $\hat{h}_k$ are the maximum likelihood estimators of g_k and h_k under parametric models $\mathcal{P}_k$ for g_k $(k\in[K+1])$ and $\mathcal{H}_k$ for h_k $(k\in[K])$, is $(K+1)$-multiply robust. See the Supplementary Material for an example.

4·4. *Sequentially strong convexity*

In this section we define a notion of convexity under which decomposition (30) with s_k^* $(k\in[K])$ satisfying requirement (C·1) of Condition 2 holds. We also discuss regularity conditions that ensure requirement (C·2) of that condition.

DEFINITION 3. *Model $\mathcal{M}$ is said to be sequentially strongly convex according to the order $1,\ldots,K$ if for any $p,p^*\in\mathcal{M}$ such that $G(p)=G(p^*)$ and any $t\in[0,1]$ there exists $p_t\in\mathcal{M}$ such that for all $k\in[K]$,*

$$tH_1(p)\times\cdots\times H_k(p)+(1-t)\,H_1(p^*)\times\cdots\times H_k(p^*)=H_1(p_t)\times\cdots\times H_k(p_t). \tag{33}$$

By definition, a model $\mathcal{M}$ that is sequentially strongly convex satisfies the convexity condition (32). This is seen by taking $k=K$ and multiplying both sides of (33) by g. Additionally, the submodels of a sequentially strongly convex $\mathcal{M}$ that fix g and $h_{k+1},\ldots,h_K$ are also convex for all $k\in[K]$. This follows upon multiplying both sides of (33) by g and $h_{k+1}\cdots h_K$. This is the reason for the designation sequentially in Definition 3. The convexity of these submodels ensures that for each k, there exists a p_t^k satisfying (33) with $H_j(p_t)$ replaced by $H_j\left(p_t^k\right)$. This is not enough to satisfy the definition of strong sequential convexity because convexity does not ensure that the p_t^k are the same for all $k\in[K]$. The appellative strongly in Definition 3 is a reminder that the property requires that p_t be the same for all $k\in[K]$. In the Supplementary Material we provide an example of a convex model that is not sequentially strongly convex. Furthermore, we show that if a model is sequentially strongly convex according to an order $1,\ldots,K$, it cannot be sequentially strongly convex according to the order $\pi(1),\ldots,\pi(K)$ for any permutation of $(1,\ldots,K)$.

An instance in which sequential strong convexity holds is precisely the setting of § 3, as the following lemma establishes. Nevertheless, this is not the only setting where the property holds. The motivating example of this section, discussed in the Supplementary Material, gives another instance.

LEMMA 4. *The conditional density model $\mathcal{M}$ of § 3 is sequentially strongly convex with* $H_k(p)(z) \equiv p\left(a_k \mid \bar{l}_k, \bar{a}_{k-1}\right)$.

In the remainder of this section we will derive a number of results that imply that, under regularity conditions, when model $\mathcal{M}$ is sequentially strongly convex, Condition 2 holds. The rigorous result is established in Corollary 2 at the end of this section. To start, given p and p^* in $\mathcal{M}$ and p_t as in (33), we define for each k the submodel

$$\left\{p_k^*(\cdot;t) \equiv g(\cdot)\, h_1^*(\cdot)\cdots h_{k-1}^*(\cdot)\, h_{k+1}^*(\cdot)\cdots h_K^*(\cdot)\, H_k(p_t)(\cdot) : t \in [0,\varepsilon)\right\}. \tag{34}$$

Next, in Proposition 5 we establish that these submodels are regular and have scores s_k^* at $t = 0$ ($k \in [K]$) that satisfy the decomposition (30). Subsequently, we provide two results, stated in Propositions 6 and 7, from where we deduce that boundedness of the product of ratios $\prod_{s=1}^{k}\left(h_s/h_s^*\right)$ ($k \in [K]$) implies requirement (C·2) of Condition 2 so long as bounded scores of regular parametric submodels are dense in $\Lambda_j(p^*)$ for all $j \in [K]$.

PROPOSITION 5. *Let $p = gh$ and $p^* = gh^*$ be in $\mathcal{M}$. Then $p/p^* - 1 = \sum_{k=1}^{K} s_k^*$ where $s_k^* \equiv \left(\prod_{j=1}^{k-1} h_j/h_j^*\right)\left(h_k/h_k^* - 1\right)$. If, in addition, $s_k^*(Z) \in L_2(p^*)$ ($k \in [K]$) and for any $t \in [0,1]$ there exists $p_t \in \mathcal{M}$ such that (33) holds for all $k \in [K]$, then model (34) is regular at $t = 0$ with score s_k^* ($k \in [K]$). Consequently, $s_k^*(Z) \in \Lambda_k(p^*)$ ($k \in [K]$).*

COROLLARY 1. *Assume that the conditions of Proposition 5 hold. Then the decomposition (30) holds with s_k^* ($k \in [K]$) satisfying requirement (C·1) of Condition 2.*

Next, we state two results from which sufficient regularity conditions for the orthogonality of $s_k^*(Z)$ with $\Lambda_j(p^*)$ ($j \neq k$), for s_k^* given by Proposition 5, easily follow. In what follows assume that given p^* and $\tilde{p}_t$ in $\mathcal{M}$ ($t \in [0,1)$) such that $p^* = \tilde{p}_0$, the submodel of $\mathcal{M}$,

$$\left\{\tilde{p}_{k,t} \equiv g h_1^* \cdots h_{k-1}^* \tilde{h}_{k,t} h_{k+1}^* \cdots h_K^* : t \in [0,1)\right\}$$

where $\tilde{h}_{k,t} \equiv H_k(\tilde{p}_t)$, $g \equiv G(p^*)$ and $h_j^* \equiv H_j(p^*)$, is a regular parametric submodel of $\mathcal{M}$ through $\tilde{p}_{k,t=0} = p^*$ with score $\tilde{s}_k$ at $t = 0$.

PROPOSITION 6. *Let j and k be two distinct fixed elements of $[K]$. Assume that:* (i) *for each $t \in [0,1)$, there exists a real-valued constant σ_t such that $\tilde{h}_{k,t}(z)/h_k^*(z) < \sigma_t$ for all $z \in \mathcal{Z}$; and* (ii) *the elements of $\Lambda_j(p^*)$ that are bounded scores of regular parametric submodels are dense in $\Lambda_j(p^*)$. Then $\tilde{s}_k(Z)$ is uncorrelated under p^* with the elements of $\Lambda_j(p^*)$.*

PROPOSITION 7. *Assume that $\mathcal{M}$ is sequentially strongly convex. Let $p = gh$ and $p^* = gh^*$ be in $\mathcal{M}$ and let p_t satisfy (33). Suppose that $\prod_{j=1}^{k}\left\{h_j(z)/h_j^*(z)\right\} < \sigma$ for all $k \in [K]$ and some $\sigma > 0$. Then, for each $t \in [0,1)$ and each $k \in [K]$, there exists σ_t such that $H_k(p_t)(z)/h_k^*(z) < \sigma_t$ for all $z \in \mathcal{Z}$.*

Finally, upon combining Propositions 5, 6 and 7 we arrive at the corollary announced earlier that gives precise regularity conditions under which Condition 2 holds for p and p^* in a sequentially strongly convex model.

COROLLARY 2. *Let $p = gh$ and $p^* = gh^*$ belong to a sequentially strongly convex model $\mathcal{M}$. Suppose that* (i) *there exists $\sigma > 0$ such that $\prod_{j=1}^{k} \left\{ h_j(z)/h_j^*(z) \right\} < \sigma$ for all $k \in [K]$ and all $z \in \mathcal{Z}$, and* (ii) *the elements of $\Lambda_j(p^*)$ that are bounded scores of regular parametric submodels are dense in $\Lambda_j(p^*)$. Then p and p^* satisfy Condition 2.*

5. CONCLUDING REMARKS

In § 3 we contemplated models defined by restrictions on a set of conditional densities and considered inference for a parameter that solves a population moment equation and that depends on the data-generating law solely through these conditional densities. For such a parameter we showed that the usual doubly robust estimators are in fact $(K+1)$-multiply robust. We also constructed a set of estimating equations which, under regularity conditions, yield solutions that are more than $(K+1)$-multiply robust. In § 4 we examined the possibility of extending the theory in § 3 to a general factorized likelihood model and a parameter of interest that solves a population moment equation and that is a function of just some factors of the likelihood. We defined the notion of a sequentially strongly convex model and showed that when models satisfy this condition it is generally possible to construct multiply robust estimating functions for the parameter of interest. The results of our theory, for instance, explain the existence of the doubly robust estimators of Tsiatis (2006) in models for ignorable nonmonotone missing data; further, we obtain $(K+1)$-multiply robust estimators in this setting.

In practice, it may happen that all the working models are incorrect. In that case even multiply robust estimators will fail to be asymptotically unbiased. In unpublished work we show that the bias of multiply robust estimators is often less than that of either doubly robust or non-doubly robust estimators, which further strengthens the case for multiply robust estimators.

Finally, an important open problem is the derivation, under the general framework of § 4, of sufficient conditions that ensure that estimation of finite-dimensional parameters indexing semiparametric models for high-dimensional nuisance functionals is possible without the need for estimating new high-dimensional functionals. The existence of such estimators would likely enable the construction of multiply robust estimators of the parameter of interest in intersection-union models that confer even more robustness against model misspecification than doubly or $(K+1)$-multiply robust estimators.

ACKNOWLEDGEMENT

Rotnitzky and Robins were funded by the U.S. National Institutes of Health. Molina and Sued were funded by the Agencia de Promocion Cientifica y Tecnica de Argentina and the University of Buenos Aires. Rotnitzky and Sued hold affiliations with the Consejo Nacional de Investigaciones Cientificas y Tecnicas de Argentina. Rotnitzky is also affiliated with the Harvard T. H. Chan School of Public Health. The authors wish to thank the reviewers for helpful comments.

Supplementary material

Supplementary material available at *Biometrika* online includes proofs of the theoretical results and the motivating example of § 4.

References

Bang, H. & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61**, 962–73.

Cao, W., Tsiatis, A. A. & Davidian, M. (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723–34.

Chan, K. C. G. & Yam, S. C. P. (2014). Oracle, multiple robust and multipurpose calibration in a missing response problem. *Statist. Sci.* **29**, 380–96.

Engle, R. F., Granger, C. W. J., Rice, J. & Weiss, A. (1986). Semiparametric estimates of the relation between weather and electricity sales. *J. Am. Statist. Assoc.* **81**, 310–20.

Gill, R. D., van der Laan, M. J. & Robins, J. M. (1997). Coarsening at random: Characterizations, conjectures, counter-examples. In *Proc. 1st Seattle Symp. Biostatistics*. D. Y. Lin and T. Fleming, eds. Lecture Notes in Statistics. New York: Springer, pp. 255–94.

Goetgeluk, S., Vansteelandt, S. & Goetghebeur, E. (2008). Estimation of controlled direct effects. *J. R. Statist. Soc.* B **70**, 1049–66.

Han, P. (2014). Multiply robust estimation in regression analysis with missing data. *J. Am. Statist. Assoc.* **109**, 1159–73.

Han, P. (2016). Intrinsic efficiency and multiple robustness in longitudinal studies with drop-out. *Biometrika* **103**, 683–700.

Han, P. & Wang, L. (2013). Estimation with missing data: Beyond double robustness. *Biometrika* **100**, 417–30.

Lunceford, J. K. & Davidian, M. (2004). Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study. *Statist. Med.* **23**, 2937–60.

Murphy, S. A., van der Laan, M. J. & Robins, J. M. (2001). Marginal mean models for dynamic regimes. *J. Am. Statist. Assoc.* **96**, 1410–23.

Orellana, L., Rotnitzky, A. & Robins, J. M. (2010). Dynamic regime marginal structural mean models for estimation of optimal dynamic treatment regimes, Part I: Main content. *Int. J. Biostatist.* **6**.

Robins, J. M., Li, L., Tchetgen, E. & van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman*, vol. 2. D. Nolan and T. Speed, eds. Beachwood: Institute of Mathematical Statistics Collections, pp. 335–421.

Robins, J. M. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proc. Am. Statist. Assoc.* **1999**, 6–10.

Robins, J. M. & Ritov, Y. (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statist. Med.* **16**, 285–319.

Robins, J. M. & Rotnitzky, A. (2001). Inference for semiparametric models: some questions and an answer: Comment. *Statist. Sinica* **11**, 863–85.

Robins, J. M., Rotnitzky, A. & van der Laan, M. J. (2000). On profile likelihood: Comment. *J. Am. Statist. Assoc.* **95**, 477–82.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Am. Statist. Assoc.* **89**, 846–66.

Robins, J. M., Rotnitzky, A. & Zhao, L. P. (1995). Analysis of semiparametric regression models for repeated outcomes in the presence of missing data. *J. Am. Statist. Assoc.* **90**, 106–21.

Rotnitzky, A., Lei, Q., Sued, M. & Robins, J. M. (2012). Improved double-robust estimation in missing data and causal inference models. *Biometrika* **99**, 439–56.

Scharfstein, D. O., Rotnitzky, A. & Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models: Rejoinder. *J. Am. Statist. Assoc.* **94**, 1135–46.

Tan, Z. (2006). Regression and weighting methods for causal inference using instrumental variables. *J. Am. Statist. Assoc.* **101**, 1607–18.

Tan, Z. (2010a). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97**, 661–82.

Tan, Z. (2010b). Nonparametric likelihood and doubly robust estimating equations for marginal and nested structural models. *Can. J. Statist.* **38**, 609–32.

Tan, Z. (2011). Efficient restricted estimators for conditional mean models with missing data. *Biometrika* **98**, 663–84.

Tchetgen Tchetgen, E. (2009). A commentary on G. Molenberghs's review of missing data methods. *Drug Info. J.* **43**, 433–5.

Tchetgen Tchetgen, E. & Shpitser, I. (2012). Semiparametric theory for causal mediation analysis: Efficiency bounds, multiple robustness, and sensitivity analysis. *Ann. Statist.* **40**, 1816–45.

Tsiatis, A. A. (2006). *Semiparametric Theory and Missing Data*. New York: Springer.

Tsiatis, A. A., Davidian, M. & Cao, W. (2011). Improved doubly robust estimation when data are monotonely coarsened, with application to longitudinal studies with dropout. *Biometrics* **67**, 536–45.
van der Laan, M. J. & Gruber, S. (2010). Collaborative double robust targeted maximum likelihood estimation. *Int. J. Biostatistics* **6**.
van der Laan, M. J. & Robins, J. M. (2003). *Unified Methods for Censored Longitudinal Data and Causality*. New York: Springer.
van der Laan, M. J. & Rubin, D. (2006). Targeted maximum likelihood learning. *Int. J. Biostatist.* **2**.
van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.
Vansteelandt, S., Rotnitzky, A. & Robins, J. M. (2007). Estimation of regression models for the mean of repeated outcomes under nonignorable nonmonotone nonresponse. *Biometrika* **94**, 841–60.
Vansteelandt, S., Vanderweele, T. J., Tchetgen, E. J. & Robins, J. M. (2008). Multiply robust inference for statistical interactions. *J. Am. Statist. Assoc.* **103**, 1693–704.
Vermeulen, K. & Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *J. Am. Statist. Assoc.* **110**, 1024–36.
Zhang, Z., Chen, Z., Troendle, J. F. & Zhang, J. (2012). Causal inference on quantiles with an obstetric application. *Biometrics* **68**, 697–706.

[*Received on* 9 *December* 2015. *Editorial decision on* 18 *March* 2017]