

The authors technical description is for estimation of regular 1-dimensional parameters, but they note that extensions to Banach-valued parameters follow as well. However, in considering Banach-valued parameters, there are several possible notions of regularity and asymptotic linearity of estimators; see for instance Definition 5.2.5 of BKRW regarding asymptotic linearity. By working exclusively with estimates of general parameters as collections of estimates of 1-dimensional parameters, it appears that the authors will adopt the weak versions of these definitions – weak regularity of parameters, weakly asymptotically linear estimators, and a resulting weak form of efficiency wherein an estimator T_n of a general parameter $b \in \mathcal{B}$ is efficient if the limiting distribution of $b^*(\sqrt{n}(T_n - b))$ is as concentrated as possible. This form of optimality does not, however, translate into the stronger conclusion that the asymptotic distribution of $\sqrt{n}(T_n - b)$ is optimal. It does not even imply $\{T_n\}$ is consistent. See the discussion of Example 1 in Section 4 of McNeney and Wellner (2000, p.464), for an example.

In their Example 1 discussing d -sample problems, the authors derive the same tangent space for the case of fixed covariates as in a randomized version that produces i.i.d. data, so that information bounds are the same in the two problems. Note that this is more generally true in that information bounds will be the same in two problems whenever the two tangent spaces are isomorphic as Hilbert spaces. See Corollary 4.4 of McNeney and Wellner (2000).

Department of Statistics and Actuarial Science, Simon Fraser University, 8888 University Drive, Burnaby, B.C. V5A 1S6, Canada.

E-mail: mcneney@stat.sfu.ca

Department of Statistics, University of Washington Box 354322, Seattle, WA 98195, U.S.A.

E-mail: jaw@stat.washington.edu

COMMENTS

James M. Robins¹ and Andrea Rotnitzky^{1,2}

¹*Harvard School of Public Health and* ²*Di Tella University*

We thank the editors for giving us this opportunity to discuss Bickel and Kwon's stimulating article and to give our perspective on the future of semiparametric inference. We found Bickel and Kwon's extension of methods for calculating information bounds to non i.i.d. models both enlightening and novel, and their list of five open questions relevant and challenging. In this discussion

we would like to pose a sixth question and to describe some initial attempts at an answer. Specifically we will discuss the question of how to approach the estimation of a finite dimensional parameter θ in very large semi parametric models, like those studied in Ritov and Bickel (1992), in which the semiparametric variance bound for $n^{1/2}$ -consistent estimators of θ is finite (i.e., θ is a regular parameter) and yet θ is not estimable at rate n^α for any $\alpha > 0$. Robins and Ritov (1997) have argued that the study of these models is of major importance because the asymptotic behavior of an estimator in these very large models accurately mimics the finite sample behavior of estimators in the high dimensional models typically used in biomedical applications. Here we argue, following Scharfstein, Rotnitzky and Robins (1999) and Robins Rotnitzky and Van der Laan (2000), that in such large models one promising partial answer to our question is to employ so called doubly robust (DR), equivalently doubly protected, estimators when such estimators exist. Section 1 of our discussion will serve as motivation for and an introduction to DR estimation. Section 2-4 will summarize the current state of knowledge. Section 3 also outlines an approach to DR estimation of non-regular parameters. A discussion and bibliographic history of DR estimation concludes.

1. Motivation

Consider a follow-up study with data on outcome Y , a binary treatment R , and a high-dimensional vector of potential confounding factors V , many of which are continuous, such as age, red blood count, white blood count, liver function tests and weight. In realistic epidemiologic studies it would not be unusual for the sample size n to be between 500 and 2000 and yet for V to be 50-100 dimensional. Because V is high-dimensional and continuous, neither nonparametric smoothing nor stratification can be used for confounder control. As a consequence, statistical models are required for dimension reduction. Typically this involves regressing the outcome on the treatment and the confounders using linear, logistic, or log linear models.

For example if Y were continuous, we might choose to fit by ordinary least squares (OLS) the linear outcome regression (OR) model

$$E(Y | R, V) = \beta_0 + \beta'V + \theta R$$

owing to the infeasibility of fitting the semiparametric regression (SR) model $E(Y | R, V) = \omega(V) + \theta R$ by multivariate non-parametric (e.g. kernel) smoothing, where $\omega(V)$ is an unknown arbitrary function. In the absence of measurement error or confounding by unmeasured factors, the parameter θ of the SR model represents the treatment effect. Even if, as we assume, the SR model assumption of no treatment-covariate interaction is correct, the OLS estimate $\hat{\theta}_{OR}$ from the OR model may be badly biased if $\omega(V)$ cannot be well approximated

by $\beta_0 + \beta'V$. In particular, if the nonlinear part of $\omega(V)$ is highly correlated with R and highly predictive of Y , then $\hat{\theta}_{OR}$ will be badly biased, even though the estimated regression function $\hat{\beta}_{OLS,0} + \hat{\beta}'_{OLS}V + \hat{\theta}_{OR}R$ may be highly predictive of the response Y and the power of standard global lack of fit tests may be small. Partially non/semiparametric dimension reducing techniques such as generalized additive models may improve somewhat upon a linear regression model but cannot solve the dimensionality problem. For example, GAM models ignore interactions among components of V .

Recently alternative methods of confounder control based on an estimated propensity score have been introduced. The propensity score $P \equiv pr(R = 1 | V)$ is the conditional probability of exposure given the covariates (Rosenbaum and Rubin (1983)). Because P is unknown, and the fitting of the nonparametric logistic regression model $\text{logit } pr(R = 1 | V) = \gamma(V)$, with $\gamma(V)$ an unknown unrestricted function is infeasible, we might choose to estimate P by the predicted value $\hat{P} = \text{expit}(\hat{\alpha}_0 + \hat{\alpha}'V)$ from the maximum likelihood fit of a linear logistic model

$$\text{logit } pr(R = 1 | V) = \alpha_0 + \alpha'V,$$

Here $\text{logit } x = \ln\{x/(1-x)\}$ and $\text{expit}(x) = \{1 + \exp(-x)\}^{-1}$. A suitable propensity score estimator $\hat{\theta}_P$ of θ turns out to be the estimator of θ in the OLS fit of the model $E(Y | R, V) = \beta_0 + \theta R + \zeta\hat{P}$ (Robins (2000)).

There has been considerable debate as to which approach to confounder control is to be preferred, as the first is biased if the outcome regression model is misspecified while the second approach is biased if the treatment regression, i.e., propensity, model is misspecified. This controversy could be resolved if an estimator were available that was guaranteed to be consistent for θ whenever at least one of the two models was correct under an asymptotic sequence in which the outcome and treatment regression models remain fixed as the sample size n increases to infinity. We refer to such combined methods as doubly-robust or doubly-protected as they can protect against misspecification of either the outcome or treatment model, although not against simultaneous misspecification of both.

A natural first guess that turns out to be correct is that the OLS estimator $\hat{\theta}_{DR}$ based on an expanded model $E(Y | R, V) = \beta_0 + \beta'V + \theta R + \zeta\hat{P}$ that adds \hat{P} as a regressor is doubly robust. One could wonder about the actual advantage of using DR estimators as, in practice, all models including the outcome and treatment regression models are misspecified and thus even the DR estimator of θ may be considerably biased. In our opinion, a DR estimator has the following advantage that argues for its routine use: if either the model for the outcome or the model for the propensity score is nearly correct, then the bias of a DR estimator of θ will be small. Thus, the DR estimator $\hat{\theta}_{DR}$, in contrast with

both the usual outcome regression estimator $\hat{\theta}_{OR}$ or the propensity estimator $\hat{\theta}_P$, gives the analyst two chances, instead of only one, to get nearly correct inference about the treatment effect. Of course, there can be an efficiency cost to using a DR estimator rather than the outcome regression estimator of θ : if the outcome regression model is correct both the $\hat{\theta}_{DR}$ and $\hat{\theta}_{OR}$ will be consistent but the DR estimator will be less efficient. However, in our opinion, we have already paid homage to the need for efficiency by using parametric, albeit high dimensional, models in the outcome and treatment regressions; at this juncture the hope to control bias due to model misspecification with DR estimators trumps further efficiency concerns.

A further advantage of DR estimation is that comparison of the three estimators $\hat{\theta}_{DR}$, $\hat{\theta}_P$, and $\hat{\theta}_{OR}$ with one another serves as a useful informal goodness of fit test. Specifically if the DR estimator differs from both the propensity and outcome regression estimator by much more than can be explained by sampling variation (say, as evaluated using the bootstrap) then we can conclude that both the propensity and outcome regression model must have been badly misspecified and that all three estimators probably suffer from substantial bias. In that event the specification of both the propensity and outcome regression model should be modified, say by adding additional nonlinear and interaction terms to the model. If $\hat{\theta}_{DR}$ and $\hat{\theta}_P$ are close but differ greatly from $\hat{\theta}_{OR}$, one can take that as some evidence that the propensity model may be nearly correct, that the outcome regression model is probably badly misspecified, and that $\hat{\theta}_{DR}$ and $\hat{\theta}_P$ may suffer from only a small amount of bias. Similar remarks apply with the roles of $\hat{\theta}_{OR}$ and $\hat{\theta}_P$ reversed. This informal goodness of fit test is based directly on estimators of the parameter θ of interest and thus will presumably be both more sensitive and inferentially relevant than global goodness of fit tests of the outcome and propensity regression models themselves.

Doubly robust estimators do not always exist, and even when they do, their construction may not always be obvious. As an example, suppose that in our motivating problem, either the outcome Y is Bernoulli and we fit a linear logistic outcome model $\text{logit}E(Y | R, V) = \beta_0 + \beta'V + \theta R$ or the outcome Y is a count variable and we fit the log linear outcome regression model $\log E(Y | R, V) = \beta_0 + \beta'V + \theta R$. In both cases the iteratively reweighted least squares (IRLS) estimator of θ (i.e., the ML estimator under the Bernoulli and Poisson likelihoods respectively) obtained by adding the term $\varsigma \hat{P}$ to the model is, in contrast with the linear regression model, inconsistent whenever the outcome regression model is misspecified and the true value of θ is non-zero, even if the propensity model is correct. Indeed, as we discuss in Sections 2 and 3, (i) no DR estimator exists for the linear logistic model and (ii) a DR estimator exists in the log linear model but it is not constructed by adding functions of \hat{P} to a log linear regression model.

In Sections 2-4 we use the semiparametric theory developed by Bickel and others to provide some preliminary answers to the question of existence and construction of DR estimators.

2. The Formal Problem and Doubly Robust Estimating Functions

To formalize our problem we consider inference about a possibly vector valued functional $\theta \equiv \theta(\rho)$ under a model $M(\mathcal{R})$ indexed by an infinite dimensional parameters $\rho \in \mathcal{R}$ for n i.i.d. copies of a random vector X . We are interested in settings in which the parameter space \mathcal{R} is very large and inference about θ is practically unfeasible due to the curse of dimensionality. Specifically, following Ritov and Bickel (1992), Robins and Ritov (1997) and Robins, Rotnitzky and van der Laan (2000), we consider models $M(\mathcal{R})$ which have the following properties: (i) the semiparametric variance bound for $n^{1/2}$ -consistent estimators of θ is finite at all $\rho \in \mathcal{R}$, and yet, no estimator is consistent for θ uniformly over $\rho \in \mathcal{R}$, much less uniformly asymptotically normal (UAN); (ii) no estimator of θ attains a pointwise (i.e., non-uniform) rate of convergence of n^α at all $\rho \in \mathcal{R}$ for any $\alpha > 0$; (iii) there does not exist a regular asymptotically linear estimator (RAL) of θ at any $\rho \in \mathcal{R}$. In this setting in both theory and practice some method of dimension reduction is necessary by imposing additional modelling restrictions. One dimension reduction strategy often used in practice is to introduce a parametrization (κ, γ) , $\kappa \in \mathcal{K}$ and $\gamma \in \Gamma$, of ρ with κ and γ variation independent, i.e., $\mathcal{R} = \mathcal{K} \times \Gamma$ and replace model $M(\mathcal{R})$ by either a working submodel $M(\mathcal{K}_{sub} \times \Gamma)$ or a working submodel $M(\mathcal{K} \times \Gamma_{sub})$ where $\mathcal{K}_{sub} \subset \mathcal{K}$ and $\Gamma_{sub} \subset \Gamma$, and hope that RAL estimators can be found in one or both of the working submodels. However, because \mathcal{K}_{sub} and Γ_{sub} are only working submodels, it is unknown whether the true value of γ is in Γ_{sub} or the true value of κ is in \mathcal{K}_{sub} . Thus, the best that can be hoped for is an estimator that is RAL in the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ that assumes that the true value of ρ either lies in $\mathcal{K} \times \Gamma_{sub}$ or in $\mathcal{K}_{sub} \times \Gamma$. We refer to such an estimator as doubly robust or doubly protected under the parametrization $\rho = (\kappa, \gamma)$ and submodels Γ_{sub} and \mathcal{K}_{sub} .

The ultimate goal would be to characterize necessary and sufficient conditions for the existence of a DR estimators and, where they exist, provide constructive methods for finding them. In this discussion, we summarize current progress. Before studying the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ of interest, it will be advantageous to study unbiased estimating functions in the special case $M(\mathcal{K} \times \Gamma) \cup M(\mathcal{K} \times \gamma)$ in which \mathcal{K}_{sub} and Γ_{sub} are singletons.

Definition 1. We say that a function $U(\theta, \kappa, \gamma) \equiv u(X, \theta, \kappa, \gamma)$ is a DR estimating function for a, possibly vector valued, functional $\theta(\kappa, \gamma)$ in model

$M(\mathcal{K} \times \Gamma)$ under parametrization (κ, γ) if it is an unbiased estimating function in the union model $M(\kappa \times \Gamma) \cup M(\mathcal{K} \times \gamma)$. That is for all (κ, γ) and $(\kappa^*, \gamma^*) \in \mathcal{K} \times \Gamma$, $E_{\kappa^*, \gamma^*} [U(\theta(\kappa^*, \gamma^*), \kappa^*, \gamma)] = E_{\kappa^*, \gamma^*} [U(\theta(\kappa^*, \gamma^*), \kappa, \gamma^*)] = 0$ and $\partial E_{\kappa^*, \gamma^*} [U(\theta, \kappa, \gamma^*)] / \partial \theta \Big|_{\theta=\theta(\kappa^*, \gamma^*)}$ and $\partial E_{\kappa^*, \gamma^*} [U(\theta, \kappa^*, \gamma)] / \partial \theta \Big|_{\theta=\theta(\kappa^*, \gamma^*)}$ are invertible.

Henceforth, we always assume invertibility of $\partial E_{\kappa^*, \gamma^*} [U(\theta, \kappa, \gamma^*)] / \partial \theta \Big|_{\theta=\theta(\kappa^*, \gamma^*)}$ and $\partial E_{\kappa^*, \gamma^*} [U(\theta, \kappa^*, \gamma)] / \partial \theta \Big|_{\theta=\theta(\kappa^*, \gamma^*)}$. Clearly, a necessary (but not sufficient) condition for the existence of a doubly robust estimating function is that there is an unbiased estimating function $U_1(\theta, \gamma)$ for $\theta(\kappa, \gamma)$ in model $M(\mathcal{K} \times \gamma)$ and an unbiased estimating function $U_2(\theta, \kappa)$ for $\theta(\kappa, \gamma)$ in model $M(\kappa \times \Gamma)$. Further progress requires semiparametric theory definitions.

Given an arbitrary semiparametric model $M(\Psi_1 \times \Psi_2)$ indexed by variation independent, possibly infinite dimensional parameters, ψ_1 and ψ_2 , and a, possibly p -dimensional vector valued, functional $\theta(\psi)$ where $\psi = (\psi_1, \psi_2)$, let $\mathcal{L}_2^0(\psi)$ be the Hilbert space of random vectors of the dimension of θ with mean zero and covariance inner product under ψ . Let $\Lambda_{\Psi_j}(\psi) \subset \mathcal{L}_2^0(\psi)$ and $\Lambda_{\Psi}(\psi) \subset \mathcal{L}_2^0(\psi)$ be the tangent spaces (i.e., closed linear span of scores) for $\psi_j, j = 1, 2$, and for ψ , respectively, when the data is generated under ψ and let $\Lambda_{\Psi_j}^\perp(\psi), j = 1, 2$, and $\Lambda_{\Psi}^\perp(\psi)$ denote their orthogonal complements in $\mathcal{L}_2^0(\psi)$. Finally, in any semiparametric model $M(\Psi)$ indexed by ψ , let $IF(\psi)$ denote the influence function space for θ at ψ . That is, $IF(\psi)$ is the direct sum of $\Lambda_{\Psi}^\perp(\psi)$ and the linear space spanned by the efficient influence function EIF (ψ) for θ . In many models $IF(\psi)$ is called the orthogonal complement to the nuisance tangent space for θ . To make our discussion concrete we use two models $M(\mathcal{K} \times \Gamma)$ to illustrate our results. In the first, $\theta(\kappa, \gamma)$ may be a function of both κ and γ . In the second, $\theta(\kappa, \gamma) = \theta(\kappa)$ only depends on κ . When $\theta(\kappa, \gamma) = \theta(\kappa)$, we define $IF(\kappa, \gamma)$ in model $M(\kappa \times \Gamma)$ to be $\Lambda_{\Gamma}^\perp(\kappa, \gamma)$.

Example 1. A Partially Missing Response Model

Suppose we have a model with underlying full data (R, Y, V) , Y and R Bernoulli and V highly multivariate and continuous. The parameter of interest θ is the mean of Y . However, if $R = 1$ then Y is not observed. Thus $X = (R, V, RY)$ is observed. Scharfstein, Rotnitzky and Robins (1999) consider the model $M(\mathcal{K} \times \Gamma)$ for X that imposes the sole assumption that

$$pr(R = 1|Y, V; \gamma) = \phi\{\gamma(V) + \alpha Y\} \equiv \Phi(\gamma) \tag{1}$$

where α is a known selection bias parameter, $\phi(\cdot)$ is a known differentiable, strictly increasing, cumulative distribution function with support on $(-\infty, \infty)$, and $\Gamma = \{\gamma = \gamma(\cdot)\}$ is the set of all functions of V . When $\alpha = 0$, the

data are said to be coarsened at random (CAR) and the missingness is said to be ignorable. When $\alpha \neq 0$, missingness is said to be nonignorable. Write $pr(Y = 1|V, R = 1; \omega) = \omega(V)$ and let $\Omega = \{\omega = \omega(\cdot); 0 < \omega(V) < 1\}$ be the set of all integrable functions taking values in $(0, 1)$, so ω is the conditional mean function of Y given V in the subpopulation of units with Y observed. At $\alpha = 0$, ω is also the conditional mean function for the entire population. Let $\mathcal{N} = \{\eta = \eta(\cdot)\}$ be the set of all densities for V and let $\mathcal{K} = \mathcal{N} \times \Omega$. Robins and Rotnitzky (RR) (2001a) show that κ and γ are variation independent, i.e., their joint parameter space is the product space $\mathcal{K} \times \Gamma$. The individual likelihood contribution is $\mathcal{L}(\kappa, \gamma) = \mathcal{L}_1(\kappa)\mathcal{L}_2(\kappa, \gamma)$ where $\mathcal{L}_1(\kappa) = \eta(V)[\omega(V)^Y\{1 - \omega(V)\}^{1-Y}]^R$ and $\mathcal{L}_2(\kappa, \gamma) = E_{\omega, \gamma}(R|V)^R\{1 - E_{\omega, \gamma}(R|V)\}^{1-R}$. RR (2001a) show that $E_{\omega, \gamma}(R|V) = E_{\omega}\{\Phi(\gamma)^{-1}|V, R = 1\}^{-1}$. Under CAR, i.e., under $\alpha = 0$, $\theta(\kappa, \gamma) = \theta(\kappa)$, $E_{\omega, \gamma}(R|V) = E_{\gamma}(R|V)$ does not depend on ω and $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$ factors into a function of κ only and a function of γ only. Rotnitzky, Robins and Scharfstein (1998) showed that model $M(\mathcal{K} \times \Gamma)$ is a non-parametric model for the law F_X of the observed data and that the joint law $F_{Y, R, V}$ is identified. In particular, the marginal mean of Y , $E_{\kappa, \gamma}(Y)$ is $\theta(\kappa, \gamma) = E_{\eta}[E_{\omega}\{Y/\Phi(\gamma)|V, R = 1\}/E_{\omega}\{1/\Phi(\gamma)|V, R = 1\}]$.

RR (2001a) show that the efficient influence function for $\theta(\kappa, \gamma)$ is $S_{eff}(\kappa, \gamma, \theta(\kappa, \gamma))$ where

$$S_{eff}(\kappa, \gamma, \theta) = \frac{R}{\Phi(\gamma)} \left(Y - \frac{E_{\omega}\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}Y|R=1, V\right\}}{E_{\omega}\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}|R=1, V\right\}} \right) + \frac{E_{\omega}\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}Y|R=1, V\right\}}{E_{\omega}\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}|R=1, V\right\}} - \theta,$$

and the influence function space for θ in model $M(\mathcal{K} \times \Gamma)$ is thus $\{cS_{eff}(\kappa, \gamma, \theta(\kappa, \gamma)); c \in \mathbf{R}\}$.

We later use Lemma 1 below to show that a DR estimating function exists for the parametrization (κ, γ) if and only if the ratio $B(\omega, \gamma) = E_{\omega}\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}Y|R=1, V\right\}/E_{\omega}\left\{\frac{\Phi'(\gamma)}{\Phi(\gamma)^2}|R=1, V\right\}$ is free of γ . RR (2001a) proved that this ratio is free of γ if and only either (i) $\alpha = 0$ and thus there is CAR or, (ii) the known CDF $\phi(\cdot)$ satisfies

$$\phi(x) = \frac{\exp\{k\alpha[x/\alpha] + q(x - \alpha[x/\alpha])\}}{1 + \exp\{k\alpha[x/\alpha] + q(x - \alpha[x/\alpha])\}}, \quad (2)$$

where k is any positive constant, $[x]$ is the largest integer less than or equal to x , and $q(u)$ is any increasing and differentiable function on $[0, \alpha]$ such that $k\alpha[x/\alpha] + q(x - \alpha[x/\alpha])$ is differentiable. The choice $q(u) = u$ and $k = 1$ gives the logistic cumulative distribution function $\phi(x) = \exp(x) / \{1 + \exp(x)\}$. However the logistic CDF is not the only possible choice for $q(u)$. For example, $q(u) = I_{[0, 1/2)}(u)2u^2 + I_{[1/2, 1]}(u)\{1 - 2(u - 1)^2\}$ is a valid choice for $\alpha = 1$ and $k = 1$.

When (i) or (ii) hold, it is easy to check that $B(\omega, \gamma) = E_\omega\{e^{-k\alpha Y}Y|R = 1, V\}/E_\omega\{e^{-k\alpha Y}|R = 1, V\}$ and $\{cS_{eff}(\kappa, \gamma, \theta); c \in \mathbf{R}\}$ is a set of DR estimating functions. Robins and Rotnitzky (2001a) use Lemma 1 below to prove that this set contains all DR estimating functions.

A necessary condition for the existence of a DR estimating function is the existence of both an unbiased estimating function for θ with κ known and another with γ known. A natural question is whether there exist simple primitive sufficient conditions for the existence of these estimating functions. Example 1 suggests not as the sufficient condition (ii) is very complex and nonintuitive.

Example 2. Semiparametric Regression

Model $M(\mathcal{K} \times \Gamma)$ for $X = (R, V, Y)$ imposes the sole assumption that $\phi^{-1}\{E(Y|R, V)\} = \theta R + \omega(V)$, with Y and R being continuous, count, or dichotomous outcome and treatment variables, V a highly multivariate continuous random vector with support in \mathcal{V} , $\phi^{-1}(x)$ a known 1-1 link function with range $(-\infty, \infty)$, $\Omega = \{\omega = \omega(\cdot)\}$ the set of all functions of V . To avoid distracting technicalities, we will additionally impose the assumption that $f(V)$ is known. This we do without loss of generality because the influence function space for θ is the same whether $f(V)$ is known or unknown. Let $\epsilon(\theta, \omega)$ denote $Y - \phi\{\theta R + \omega(V)\} \equiv Y - \Phi\{\theta, \omega\}$. Then $\mathcal{K} = \{\kappa = (\theta, \omega, \eta); \theta \in \mathbf{R}^1, \omega \in \Omega, \eta \in \mathcal{N}(\theta, \omega)\}$, where $\mathcal{N}(\theta, \omega)$ is the set of all mean zero conditional densities for $\epsilon(\theta, \omega)$ given V (except when Y is binary in which case $\kappa = (\theta, \omega)$). We define $\Gamma = \{\gamma \equiv \gamma(R|V)\}$ to be the set of all conditional densities for R given V . The individual likelihood contribution factors as $\mathcal{L}(\kappa, \gamma) = \mathcal{L}_1(\kappa)\mathcal{L}_2(\gamma)$, where $\mathcal{L}_1(\kappa) = \eta(\epsilon(\theta, \omega)|V)$ and $\mathcal{L}_2(\gamma) = \gamma(R|V)$.

Bickel, Klaassen, Ritov and Wellner (1993) and RR (2001b) show that the influence function space for θ is

$$IF(\kappa, \gamma) = \{U(\kappa, \gamma, g, \phi) = \epsilon(\theta(\kappa), \omega(\kappa))\{g(R, V) - M(\gamma, \kappa, g, \phi)\}; g \in \mathcal{G}\},$$

with \mathcal{G} the set of all functions of (R, V) , and $M(\gamma, \kappa, g, \phi) = E_\gamma\{g(R, V)|V\}$ if ϕ^{-1} is the identity link, $M(\gamma, \kappa, g, \phi) = E_\gamma\{g(R, V)e^{\theta(\kappa)R}|V\}/E_\gamma\{e^{\theta(\kappa)R}|V\}$ if ϕ^{-1} is the log link and $M(\gamma, \kappa, g, \phi) = \frac{E_\gamma[g(R, V)\Phi\{\theta(\kappa), \omega(\kappa)\}][1 - \Phi\{\theta(\kappa), \omega(\kappa)\}]|V}{E_\gamma[\Phi\{\theta(\kappa), \omega(\kappa)\}][1 - \Phi\{\theta(\kappa), \omega(\kappa)\}]|V}$ if ϕ^{-1} is the logit link. Throughout, we write $\theta(\kappa)$ and $\omega(\kappa)$ when we wish to emphasize that θ and ω are formally functions of κ . Note that $IF(\kappa, \gamma)$ does not depend on η , so without loss of generality we write it as $IF(\theta, \omega, \gamma)$. RR (2001b) have proved that the only links for which $M(\gamma, \kappa, g, \phi)$ depends on κ only through $\theta(\kappa)$ for all $g \in \mathcal{G}$ are the identity and exponential. It is straightforward to check that, for ϕ^{-1} the identity or the log-link, the set $\mathcal{D} = \{U(\theta, \kappa, \gamma, g, \phi) = \epsilon(\theta, \omega(\kappa))\{g(R, V) - M_{est}(\gamma, \theta, g, \phi)\}; g \in \mathcal{G}\}$ is comprised of DR estimating functions, where $M_{est}(\gamma, \theta, g, \phi) = E_\gamma\{g(R, V)|V\}$ if ϕ^{-1} is

the identity link and $M_{est}(\gamma, \theta, g, \phi) = E_\gamma\{g(R, V)e^{\theta R}|V\}/E_\gamma\{e^{\theta R}|V\}$ if ϕ^{-1} is the log link. Note $U(\theta, \kappa, \gamma, g, \phi)$ is $U(\kappa, \gamma, g, \phi)$ with $\theta(\kappa)$ replaced by the free parameter θ .

We next summarize results in RR (2001a) and apply them to prove that the set \mathcal{D} contains all DR estimating functions that depend on κ only through $\omega(\kappa)$ when ϕ^{-1} is either the identity or the log link, and that no DR estimating function exists when Y is binary and ϕ^{-1} is the logit link.

- A necessary condition for $U(\theta, \kappa, \gamma)$ to be DR is that $U(\theta(\kappa, \gamma), \kappa, \gamma)$ must be an element of the influence function space $IF(\kappa, \gamma)$ in the unrestricted model $M(\mathcal{K} \times \Gamma)$ at each (κ, γ) .
- Suppose that $\theta(\kappa, \gamma) = \theta(\kappa)$ is a function of κ alone. Then (i) our model $M(\mathcal{R})$ will admit parameterizations $\mathcal{R} = \mathcal{K} \times \Gamma$ with $\mathcal{K} = \{\kappa = (\theta, \delta) : \theta \in \Theta, \theta(\kappa) = \theta \text{ and } \delta \in \Delta(\theta)\}$, where $\Delta(\theta)$ is a set that can possibly depend on θ , (ii) $IF(\kappa, \gamma)$ can be expressed as the set $IF(\theta, \delta, \gamma) = \{\tilde{V}(\theta, \delta, \gamma) \equiv \tilde{U}((\theta, \delta), \gamma); \tilde{U}(\kappa, \gamma) \in IF(\kappa, \gamma)\}$ of functions of (θ, δ, γ) , and (iii), by the previous remark, a necessary condition for an estimating function $U(\theta, \kappa, \gamma) = U(\theta, \delta, \gamma)$ that depends on κ only through δ to be DR is that it be an element of $IF(\theta, \delta, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$.

Example 2. (continuation). In Example 2, take $\delta = (\omega, \eta)$ and $\Delta(\theta) = \{(\omega, \eta) : \omega \in \Omega, \eta \in \mathcal{N}(\theta, \omega)\}$. Then since $\kappa = (\theta, \delta)$ with $\theta(\kappa, \gamma) = \theta(\kappa) = \theta$, any DR estimating function $U(\theta, \omega, \gamma)$ must be in the set $IF(\theta, \delta, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$. But as pointed out above, $IF(\theta, \delta, \gamma) = IF(\theta, \omega, \gamma)$ does not depend on η . Now, when ϕ^{-1} is the logit link, RR (2001a) showed that $E_{\theta, \omega, \gamma^*}[U(\theta, \omega, \gamma)] \neq 0$ for all $U(\theta, \omega, \gamma) \in IF(\theta, \omega, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$. Thus, no DR estimating function can exist when ϕ^{-1} is the logit link. When ϕ^{-1} is the identity or log link, we noted above that all elements $U(\theta, \omega, \gamma)$ of $IF(\theta, \omega, \gamma)$ are doubly robust. This proves that $IF(\theta, \omega, \gamma)$ is the set of all DR estimating functions $U(\theta, \kappa, \gamma) = U(\theta, \omega, \gamma)$ that depends on κ only through ω for ϕ^{-1} the identity or the log link.

When $\theta(\kappa, \gamma)$ depends on κ and γ the above strategy is not available. The following Lemma provides a sometimes useful way to prove the absence of DR estimating functions in this case.

Lemma 1. *A necessary condition for the existence of a doubly robust estimating function is that, for each θ in $\Theta(\gamma) \equiv \{\theta(\kappa, \gamma) : \kappa \in \mathcal{K}\}$, we have $\cap_{\kappa^* : \theta(\kappa^*, \gamma) = \theta} IF(\kappa^*, \gamma) \neq \emptyset$ in model $M(\mathcal{K} \times \gamma)$, and for each θ in $\Theta(\kappa) \equiv \{\theta(\kappa, \gamma) : \gamma \in \Gamma\}$, we have $\cap_{\gamma^* : \theta(\kappa, \gamma^*) = \theta} IF(\kappa, \gamma^*) \neq \emptyset$ in model $M(\kappa \times \Gamma)$.*

Informally, we interpret Lemma 1 as saying that in the model $M(\mathcal{K} \times \gamma)$ there exists an element of $IF(\kappa^*, \gamma)$ that depends on κ^* only through $\theta(\kappa^*, \gamma)$,

and in model $M(\kappa \times \Gamma)$ there is an element of $IF(\kappa, \gamma^*)$ that depends on γ^* only through $\theta(\kappa, \gamma^*)$.

Example 1. (continuation). In model $M(\kappa \times \Gamma)$, RR (2001a) prove that $IF(\kappa, \gamma) = \{S_{eff}(\kappa, \gamma, \theta(\kappa, \gamma)) + a(X); a \in \mathcal{A}(\kappa)\}$, where

$$\mathcal{A}(\kappa) = \left\{ a(X) = Rs(Y, V) + (1 - R)E_\omega[s(Y, V)|R=1, V] - \varphi(s; \kappa); s \text{ unrestricted} \right\},$$

and $\varphi(s; \kappa) = E_\eta\{E_\omega[s(Y, V)|R = 1, V]\}$. RR (2001a) show that in order for $\cap_{\gamma: \theta(\kappa, \gamma) = \theta} IF(\kappa, \gamma) \neq \emptyset$ it must be that $B(\omega, \gamma)$ does not depend on γ .

We have described ways to rule out DR estimating functions for $\theta(\kappa, \gamma)$ by checking necessary conditions for their existence. We now explore ways to rule in DR estimating function by finding further sufficient conditions for their existence. Consider the following.

Condition 1. $IF(\kappa^*, \gamma)$ in model $M(\mathcal{K} \times \gamma)$ depends on κ^* only through $\theta(\kappa^*, \gamma)$, and $IF(\kappa, \gamma^*)$ in model $M(\kappa \times \Gamma)$ depends on γ^* only through $\theta(\kappa, \gamma^*)$.

Condition 2. $\theta(\kappa, \gamma)$ is function of κ alone and $\mathcal{K} = \{\kappa = (\theta, \delta) : \theta \in \Theta, \theta(\kappa, \gamma) = \theta, \text{ and } \delta \in \Delta(\theta)\}$.

Condition 3. There exists a function $q(\cdot)$ such that $q(\theta)$ is a linear functional of both the law indexed by κ (with γ fixed) and the law indexed by γ (with κ fixed), in the sense that for some $U_1(\gamma)$ and $U_2(\kappa)$, $E_{\kappa\gamma}\{U_1(\gamma)\} = E_{\kappa\gamma}\{U_2(\kappa)\} = q\{\theta(\kappa, \gamma)\}$. The following Theorem, proved in RR (2001a), states that when Condition 1 and one of Conditions 2 or 3 hold, then there exist DR estimating equations. Indeed, the theorem shows how to construct them.

Theorem 1. (a) If Conditions 1 and 2 hold, then all elements $U(\theta, \delta, \gamma)$ of $IF(\theta, \delta, \gamma) = IF(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$ are doubly robust; (b) if Conditions 1 and 3 hold, $IF(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$ has the form $IF(\kappa, \gamma) = \{\tilde{U}(\gamma) - q[\theta(\kappa, \gamma)]; \tilde{U}(\gamma) \in \tilde{\mathcal{U}}(\gamma)\}$, where $\tilde{\mathcal{U}}(\gamma)$ is the set of all random variables $\tilde{U}(\gamma)$ for which $\tilde{U}(\gamma) - q[\theta(\kappa, \gamma)] \in IF(\kappa, \gamma)$ in both $M(\mathcal{K} \times \gamma)$ and $M(\kappa \times \Gamma)$. Further the set $\{\tilde{U}(\gamma) - q[\theta]; \tilde{U}(\gamma) \in \tilde{\mathcal{U}}(\gamma)\}$ consists of DR estimating functions.

A necessary and sufficient condition for Condition 1 to hold is that in model $M(\kappa \times \Gamma)$ there exists an unbiased estimating function $U_2(\theta, \kappa)$ for θ and the orthogonal complement $\Lambda_\Gamma^\perp(\kappa, \gamma) = \Lambda_\Gamma^\perp(\kappa)$ to the tangent space for γ is the same for all (i.e., does not depend on) $\gamma \in \Gamma$, and in model $M(\mathcal{K} \times \gamma)$ there exists an unbiased estimating function $U_1(\theta, \gamma)$ for θ and $\Lambda_\mathcal{K}^\perp(\kappa, \gamma) = \Lambda_\mathcal{K}^\perp(\gamma)$ is the same for all $\kappa \in \mathcal{K}$. This raises the question of when one might expect $\Lambda_\mathcal{K}^\perp(\kappa, \gamma) = \Lambda_\mathcal{K}^\perp(\gamma)$ and/or $\Lambda_\Gamma^\perp(\kappa, \gamma) = \Lambda_\Gamma^\perp(\kappa)$. RR (2001a) used ideas from Bickel (1982) on convex models to show that these identities hold if model $M(\kappa \times \Gamma)$ is convex in its parameter γ and $M(\mathcal{K} \times \gamma)$ is convex in κ . A model $M(\Psi)$ is convex if for all ψ^* ,

$\psi \in \Psi$, any mixture of the laws governed by ψ^* and ψ lies in the model. We say that κ and γ are mutually convex in model $M(\mathcal{K} \times \Gamma)$ if both model $M(\mathcal{K} \times \gamma)$ and model $M(\kappa \times \Gamma)$ are convex. The models of both Example 1 and Example 2 have κ and γ mutually convex.

Example 1.(continuation). Model $M(\mathcal{K} \times \gamma)$ is convex in κ . Thus $\Lambda_{\mathcal{K}}^{\perp}(\kappa^*, \gamma) = \{(\frac{R}{\Phi(\gamma)} - 1)g(V); g \in \mathcal{G}\}$ does not depend on κ^* . Model $M(\kappa \times \Gamma)$ is convex in γ . Thus $\Lambda_{\Gamma}^{\perp}(\kappa, \gamma^*) = \mathcal{A}(\kappa)$ does not depend on γ^* .

Robins and Rotnitzky (2001a) provide an example that shows that condition 1 alone is not sufficient for the existence of doubly robust estimating functions.

Theorem 1 provides sufficient conditions for the existence of DR estimating functions only in models under the quite strong Condition 1. The following two theorems provide sufficient conditions for the existence of DR estimating functions in models that need not satisfy this condition. The first theorem considers the case where $\theta(\kappa, \gamma) = \theta(\kappa)$. It strengthens the suppositions of Theorem 1a by assuming that the likelihood factors into a κ -part and a γ -part. It relaxes the suppositions of Theorem 1a by no longer assuming either that model $M(\kappa \times \Gamma)$ admits an unbiased estimating function for θ or that $\Lambda_{\mathcal{K}}^{\perp}(\kappa, \gamma)$ in model $M(\mathcal{K} \times \gamma)$ depends only on γ .

Theorem 2. (Robins, Rotnitzky and Van der Laan, (2000)). *Suppose in model $M(\mathcal{K} \times \Gamma)$, the parameter $\theta(\kappa, \gamma) = \theta(\kappa)$ depends only on κ , the likelihood $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$ factors, $\Lambda_{\Gamma}^{\perp}(\kappa, \gamma) = \Lambda_{\Gamma}^{\perp}(\kappa)$ in model $M(\kappa \times \Gamma)$ depends only on κ , and there exist an unbiased estimating function $\tilde{U}(\theta, \gamma)$ in model $M(\mathcal{K} \times \gamma)$, i.e., $E_{\kappa^*, \gamma}\{\tilde{U}(\theta(\kappa^*), \gamma)\} = 0$ for all (κ^*, γ) . Then $U(\theta, \kappa, \gamma) = \tilde{U}(\theta, \gamma) - \Pi_{\kappa, \gamma}[\tilde{U}(\theta, \gamma) | \Lambda_{\Gamma}(\gamma)]$ is a DR estimating equation where $\Pi_{\kappa, \gamma}[A | \mathcal{B}]$ is the projection of the random variable A on the closed linear space \mathcal{B} , and $\Lambda_{\Gamma}(\kappa, \gamma) = \Lambda_{\Gamma}(\gamma)$ by the factorization of the likelihood.*

Example 2. (continuation). Robins and Rotnitzky (2001b) proved that for any choice of ϕ^{-1} other than the identity or the log link and for R continuous or discrete, model $M(\mathcal{K} \times \gamma)$ was not convex in κ and $\Lambda_{\mathcal{K}}^{\perp}(\kappa, \gamma)$ varies with κ . Thus Theorem 1 cannot be used to guarantee the existence of a DR estimating function. However it is clear that the suppositions of Theorem 2 hold for any choice of ϕ , except perhaps the existence of an unbiased estimating function $\tilde{U}(\theta, \gamma)$ in model $M(\mathcal{K} \times \gamma)$. Further for R continuous and $\phi(x) = x^3$, Robins and Rotnitzky (2001b) proved that $\tilde{U}(\theta, \gamma) \equiv \epsilon(\theta)h(R, V, \gamma)$ is an unbiased estimating function, where $\epsilon(\theta) = Y - \theta^3 R^3$ and $h(R, V, \gamma) = R^3 - E_{\gamma}(R^3 B | V)\{E_{\gamma}(BB^T | V)\}^{-1}B$ is the residual from the population conditional least squares regression of R^3 on $B = (1, R, R^2)^T$ provided $h(R, V, \gamma) = 0$ wp1 is false.

The following is an alternative to Theorem 1b that does not require mutual convexity in κ and γ .

Theorem 3. *Suppose there exists a function $q(\cdot)$ such that $q(\theta)$ is a linear functional of both the law indexed by κ (with γ fixed) and the law indexed by γ (with κ fixed) in the sense that for some $U_1(\gamma)$ and $U_2(\kappa)$, $E_{\kappa\gamma}\{U_1(\gamma)\} = E_{\kappa\gamma}\{U_2(\kappa)\} = q(\theta(\kappa, \gamma))$ and the model $M(\mathcal{K} \times \Gamma) = M(\mathcal{R})$ is convex in $\rho = (\kappa, \gamma)$. Then there exists a DR estimating function. Specifically for any $S(\kappa, \gamma) \in IF(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$, $U(\kappa, \gamma) - q(\theta) = S(\kappa, \gamma) + q(\theta(\kappa, \gamma)) - q(\theta)$ is doubly robust.*

Example 4. Suppose we have a nonparametric i.i.d. model for densities of X absolutely continuous variable w.r.t. Lebesgue measure parameterized by $\kappa = Var(X)$, $\kappa \in K = \{\kappa; \kappa > 0\}$, and $\gamma = (\mu, \eta(\cdot))$, where $\mu = E(X)$ and $\eta(\cdot)$ is the density of $(X - \mu)/\kappa^{1/2}$, with $\gamma \in \Gamma = \{\mu, \eta(\cdot); \mu \in R^1, \eta(\cdot) \text{ a density with mean 0 and variance 1}\}$. Let $\theta(\kappa, \gamma) = \mu\kappa$. The space $\Lambda_{\mathcal{K}}^{\perp}(\kappa, \gamma)$ varies with κ and thus $M(\mathcal{K} \times \Gamma)$ is not convex in κ so the suppositions of Theorem 1 do not hold. Nonetheless, by the convexity of the entire nonparametric model $M(\mathcal{K} \times \Gamma)$ and $E_{\kappa, \gamma^*}[X\kappa] = \theta(\kappa, \gamma^*)$ and $E_{\kappa^*, \gamma}[\mu(X^2 - \mu^2)] = \theta(\kappa^*, \gamma)$, a DR estimating function exists by Theorem 3. Indeed, since $IF(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$ is the span of $S(\kappa, \mu) = (X - \mu)(\kappa - 2\mu^2) - \mu(X^2 - \mu^2 - \kappa)$, the function $S(\kappa, \mu) + \mu\kappa - \theta$ is a DR estimating function. Of course, this example does not suffer from the curse of dimensionality, but it does confirm that Theorem 3 can be applied in settings in which Theorem 1b does not apply.

The suppositions of Theorem 1b do not imply those of Theorem 3 because, as can be demonstrated with the semiparametric regression model of Example 2 for the identity link, mutual convexity in κ and γ does not imply convexity in $\rho = (\kappa, \gamma)$. Conversely Example 4 demonstrates that convexity in $\rho = (\kappa, \gamma)$ does not imply mutual convexity in κ and γ .

3. Estimation When Γ_{sub} or \mathcal{K}_{sub} Are not Singletons

We now consider inference when Γ_{sub} or \mathcal{K}_{sub} are not singletons, this being the problem of practical interest.

Definition. An estimator $\hat{\theta}$ is a doubly robust * estimator for $\theta(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$ with respect to the parametrization (κ, γ) and submodels $\Gamma_{sub} \subset \Gamma$ and $\mathcal{K}_{sub} \subset \mathcal{K}$ if $\hat{\theta}$ is a * estimator for $\theta(\kappa, \gamma)$ in the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$, where * can represent any property of interest such as consistent, RAL, etc.

RR (2001a) show that when the true value of (κ, γ) lies in the intersection submodel $M(\mathcal{K}_{sub} \times \Gamma_{sub})$, the influence function of any RAL DR estimator must be an element of the influence function space $IF(\kappa, \gamma)$ of the full model $M(\mathcal{K} \times \Gamma)$. Following the introduction of some notation we discuss a number of settings where DR estimators exist.

Throughout, for a given random function $U(\theta, \kappa, \gamma)$, we write $U(\theta, \kappa, \gamma) = \tilde{U}(\theta, k, j)$ where $k = k(\kappa)$ and $j = j(\gamma)$ are maximal coarsening functions of κ and γ with respect to the function $U(\theta, \kappa, \gamma)$, in the sense that if $U(\theta, \kappa_1, \gamma_1) = U(\theta, \kappa_2, \gamma_2)$ then $k(\kappa_1) = k(\kappa_2)$ and $j(\gamma_1) = j(\gamma_2)$.

Example 1. (continuation). $S_{eff}(\kappa, \gamma, \theta) \equiv \tilde{S}_{eff}(\theta, k(\kappa), j(\gamma))$ with $k(\kappa) = \omega$ and $j(\gamma) = \gamma$.

One setting where DR RAL estimators exist is when (i) $U(\theta, \kappa, \gamma)$ is a DR estimating function for $\theta(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$ and (ii) given $U(\theta, \kappa, \gamma) = \tilde{U}(\theta, k(\kappa), j(\gamma))$, one can construct a consistent estimator \hat{j} of $j(\gamma)$ in model $M(\mathcal{K} \times \Gamma_{sub})$ and a consistent estimator \hat{k} of $k(\kappa)$ in model $M(\mathcal{K}_{sub} \times \Gamma)$. Under (i) and (ii), the estimator $\hat{\theta}(\hat{k}, \hat{j})$ solving $P_n[\tilde{U}(\theta, \hat{k}, \hat{j})] = 0$ will be a DR RAL estimator under regularity conditions provided, as we assume, that the size of Γ_{sub} and \mathcal{K}_{sub} are chosen small enough so that \hat{j} converges to $j(\gamma)$, $\gamma \in \Gamma_{sub}$ under (κ, γ) and \hat{k} converges to $k(\kappa)$, $\kappa \in \mathcal{K}_{sub}$ under (κ, γ) , at sufficiently fast rates. P_n is the empirical distribution expectation operator.

Supposition (ii) holds when the likelihood factors as $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$ in model $M(\mathcal{K}_{sub} \times \Gamma_{sub})$, since then the scores $S_\gamma(\gamma)$ and $S_\kappa(\kappa)$ can be used as unbiased estimating functions if Γ_{sub} and \mathcal{K}_{sub} are finite dimensional. More generally (ii) holds when there exists a possibly expanded model $M(\mathcal{K}_{exp} \times \Gamma_{exp})$ with $\mathcal{K} \subseteq \mathcal{K}_{exp}$, $\Gamma \subseteq \Gamma_{exp}$ such that $M(\mathcal{K}_{exp} \times \Gamma_{exp})$ is mutually convex in κ and γ . This is so because, in model $M(\mathcal{K}_{exp} \times \Gamma)$, $\Lambda_{\mathcal{K}_{exp}}^\perp(\kappa, \gamma) = \Lambda_{\mathcal{K}_{exp}}^\perp(\gamma)$ does not depend on $\kappa \in \mathcal{K}_{exp}$ so that elements $U(\gamma)$ of $\Lambda_{\mathcal{K}_{exp}}^\perp(\gamma)$ can be used as unbiased estimating functions for $\gamma \in \Gamma_{sub}$.

In this section we assume condition (ii) holds.

Example 1. (continuation). To construct $\tilde{S}_{eff}(\theta, \hat{\omega}, \hat{\gamma})$, suppose $\mathcal{K}_{sub} = \Omega_{sub} \times N$ where Γ_{sub} and Ω_{sub} are q_γ - and q_ω -dimensional parametric models. Then, although the likelihood does not factor as $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$, nonetheless $M(\mathcal{K} \times \Gamma)$ is mutually convex in κ and γ for any choice of ϕ . In particular, we can find $\hat{\gamma} = \hat{\gamma}(g) \in \Gamma_{sub}$ as the solution to a q_γ -dimensional estimating equation $P_n[(\frac{R}{\phi(\gamma)} - 1)g(V)] = 0$, each component of which is in $\Lambda_{\mathcal{K}}^\perp(\gamma)$. Similarly we can find $\hat{\omega} = \hat{\omega}(s) \in \Omega_{sub}$ as the solution to $P_n[R\{Y - \omega(V)\}s(V)] = 0$, each component of which is in $\Lambda_{\Gamma}^\perp(\kappa)$. Recall that even though we can find consistent estimators of ω , and γ , $\tilde{S}_{eff}(\theta, \omega, \gamma)$ is DR only if condition (i) or (ii) of page 926 holds.

Example 2. (continuation). Returning to the set-up of Section 1, for the identity or log link, we obtain a DR estimator of θ by solving $P_n[\epsilon(\theta, \hat{\omega})\{g(R, V) - M_{est}(\hat{\gamma}, \theta, g, \phi)\}] = 0$ for $g \in \mathcal{G}$, where $\hat{\omega}$ is the IRLS estimator of ω under the model \mathcal{K}_{sub} for which $\omega \in \Omega_{sub} = \{\omega; \omega(V) = \beta_0 + \beta'V\}$ and $\hat{\gamma}$ is the MLE in the model $\Gamma_{sub} = \{\gamma; \log it \gamma(V) = \alpha_0 + \alpha'V\}$. Furthermore, if we take

$\Omega_{sub} = \{\omega; \omega(V) = \beta_0 + \beta'V + \varsigma M_{est}(\hat{\gamma}, \theta, g, \phi)\}$ and $g(R, V) = R$, we also obtain a DR estimator which for the identity link is algebraically equivalent to the OLS DR estimator of Section 1. RR (2001a) show there is no DR estimator for the logit link.

Again suppose (i) and (ii) are true. In this setting, an estimation strategy we do not recommend is the following. One first performs a global lack of fit test for model $M(\mathcal{K} \times \Gamma_{sub})$; then if the test rejects, one estimates $\theta(\kappa, \gamma)$ assuming model $M(\mathcal{K}_{sub} \times \Gamma)$ is true; if it accepts then one tests the fit of $M(\mathcal{K}_{sub} \times \Gamma)$ and if it rejects, one estimates $\theta(\kappa, \gamma)$ assuming model $M(\mathcal{K} \times \Gamma_{sub})$ is true. If neither test rejects one uses the doubly robust estimation strategies described above. This preliminary test strategy can result in a RAL estimator in the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ provided that, in order to insure regularity, the lack of fit tests have power zero against all parametric Pitman alternatives. However, in our view, we do not recommend this strategy because it takes too seriously the truth of the union model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ which is really only a working model that is used because the model $M(\mathcal{K} \times \Gamma)$ is too large. We would therefore recommend that if a lack of fit test rejects either model $M(\mathcal{K} \times \Gamma_{sub})$ or $M(\mathcal{K}_{sub} \times \Gamma)$, one enlarges Γ_{sub} or \mathcal{K}_{sub} (until neither lack of fit test rejects) and then uses the above DR estimation strategy.

We now turn to the question of how we might obtain DR estimators when no DR estimating function $U(\theta, \kappa, \gamma)$ for θ exists. We do so by extending to doubly robust estimation several well-known approaches to estimation of a parameter θ in a model where no unbiased estimating function for θ exists.

We first consider how to construct DR estimators of certain non-regular parameters (i.e., parameters that do not have finite semiparametric information bounds). Our strategy is to approximate the non-regular parameters by a regular parameter, as in Van der Laan and Robins (1998) and Bickel and Ritov (2000), because non-regular parameters do not admit unbiased estimating functions. Let $\theta(\kappa, \gamma)$ be a nonregular parameter. Suppose that $\theta_\delta(\kappa, \gamma)$ is a regular parameter such that $\theta_\delta(\kappa, \gamma)$ converges to $\theta(\kappa, \gamma)$ as $\delta \downarrow 0$. Suppose a DR estimating function $U_\delta(\theta, \kappa, \gamma) = \tilde{U}_\delta(\theta, k(\kappa), j(\gamma))$ exists for $\theta_\delta(\kappa, \gamma)$. Then, in general, the estimator $\hat{\theta}_{\delta(n)}(\hat{k}, \hat{j})$ solving $P_n[\tilde{U}_{\delta(n)}(\theta, \hat{k}, \hat{j})] = 0$ will under regularity conditions be a DR consistent estimator for $\theta(\kappa, \gamma)$ if $\delta(n) \downarrow 0$ as $n \uparrow \infty$ at an appropriate rate.

Example 1. (continuation). Suppose now that Y is a continuous variable with a twice differentiable density w.r.t. Lebesgue measure and ϕ is logistic. Let $\theta(\kappa, \gamma) = f(y; \kappa, \gamma)$ be the density of Y at y and let $\theta_\delta(\kappa, \gamma) = E_{\kappa, \gamma}\{W(\delta)\}$ where $W(\delta) = \{w((Y - y)/\delta)\}/\delta$, $w(\cdot)$ is a mean zero smooth positive kernel function and δ a suitable bandwidth. Then

$$U_\delta(\kappa, \gamma, \theta) = \tilde{U}_\delta(\omega, \gamma, \theta) = R\Phi^{-1}(\gamma)\{W(\delta) - \theta\}$$

$-\{R\Phi^{-1}(\gamma)-1\}E_{\omega}[\Phi'(\gamma)\Phi(\gamma)^{-2}\{W(\delta)-\theta\}|R=1, V]/E_{\omega}\{\Phi'(\gamma)\Phi(\gamma)^{-2}|R=1, V\}$ is a DR estimating function for $\theta_{\delta}(\kappa, \gamma)$. Under suitable regularity conditions, and with $\delta(n) = n^{-1/5}$, $\hat{\theta}_{\delta(n)}(\hat{\omega}, \hat{\gamma})$ solving $P_n[\tilde{U}_{\delta(n)}(\hat{\omega}, \hat{\gamma}, \theta)] = 0$ will be $n^{2/5}$ -consistent for $\theta(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$.

Suppose next $\theta(\kappa, \gamma)$ is a regular parameter for which no DR estimating function exists, but there exists a possibly non-regular parameter $\psi(\kappa, \gamma)$ and a function $U(\theta, \psi, \kappa, \gamma) = \tilde{U}(\theta, \psi, k(\kappa), j(\gamma))$ that is a DR estimating function for $\theta(\kappa, \gamma)$ with $\psi(\kappa, \gamma)$ known. That is $E_{\kappa^*, \gamma}[U(\theta(\kappa^*), \gamma), \psi(\kappa^*), \kappa, \gamma)] = E_{\kappa, \gamma^*}[U(\theta(\kappa, \gamma^*), \psi(\kappa, \gamma^*), \kappa, \gamma)] = 0$. Suppose further there exists a DR $n^{1/4}$ -consistent estimator $\hat{\psi} = \hat{\psi}(\theta(\kappa, \gamma))$ for $\psi(\kappa, \gamma)$ when $\theta(\kappa, \gamma)$ is known. Then subject to regularity conditions the estimator $\hat{\theta}(\hat{\psi}, \hat{k}, \hat{j})$ solving $P_n[\tilde{U}(\theta, \hat{\psi}(\theta), \hat{k}, \hat{j})] = 0$ will be a doubly robust RAL estimator. RR (2001a) provide a concrete example.

Finally, suppose $\theta(\kappa, \gamma)$ is a regular parameter; no DR estimating function for $\theta(\kappa, \gamma)$ exists even with some other parameter known; but $\theta(\kappa, \gamma)$ is known function $b(\zeta(\kappa, \gamma), \tau(\kappa, \gamma))$ of parameters $\zeta(\kappa, \gamma)$ and $\tau(\kappa, \gamma)$ that admit RAL DR estimators. In this setting we can obtain a RAL DR estimator of $\theta(\kappa, \gamma)$ by evaluating $b(\cdot, \cdot)$ at RAL DR estimators of $\zeta(\kappa, \gamma)$ and $\tau(\kappa, \gamma)$. RR (2001a) provide a concrete example. Note that the existence of DR estimating functions for $\zeta(\kappa, \gamma)$ and $\tau(\kappa, \gamma)$ does not imply that $b(\zeta(\kappa, \gamma), \tau(\kappa, \gamma))$ has a DR estimating function.

4. Generalized Double Robustness

In this section we discuss settings in which (i) exact DR estimators are difficult to compute, or (ii) no exact DR estimators exist. For such situations we propose the use of “generalized” DR estimators. Generalized DR estimators are those which have small asymptotic bias if either one of two (possibly incompatible) lower dimensional models Γ_{sub} or K_{sub} is approximately correct. Thus, a generalized DR estimator shares with a true DR estimator the crucial property of giving the analyst two chances for approximately correct inference about θ . We will illustrate a generalized DR estimator in setting (ii). RR (2001a) provide an example of a generalized estimator in setting (i).

Example 1. (continuation). Consider Example 1 with ϕ logistic, $\alpha \neq 0$, except with Y continuous. Suppose we wish to estimate the parameter $\theta(\kappa, \gamma)$ of a given marginal parametric model $f(Y; \theta)$. For concreteness, we use a normal model with mean θ_1 and variance θ_2 . As noted earlier the parameters κ and γ , defined as before, determine the marginal law of Y . However, in contrast to our previous discussion of Example 1, the model is no longer non-parametric, and when $\alpha \neq 0$ the set of parameters (κ, γ) compatible with the parametric model $f(Y; \theta)$ is no longer a product space. We can make \mathcal{R} a product space by choosing the following new parameterization. We let $\Gamma = \{\gamma = \gamma(\cdot)\}$ remain unchanged, but now let κ

parametrize the law $f(Y, V)$ rather than the laws $f(Y|V, R = 1)$ and $f(V)$. Thus, we now take \mathcal{K} to be $\mathcal{K} = \{\kappa = (\theta, \omega); \omega \in \Omega, \theta = (\theta_1, \theta_2), \theta_1 \in R^1, \theta_2 \in (0, \infty)\}$, where Ω is the set of all densities for V given Y . In model $M(\mathcal{K} \times \Gamma)$, the IF space for θ is $\{U(\theta(\kappa), \kappa, \gamma, c); c \in \mathcal{C}\}$, with \mathcal{C} the set of all functions Y , and

$$U(\theta, \kappa, \gamma, c) = \frac{R\tilde{C}(\theta)}{\Phi(\gamma)} - \left\{ \frac{R}{\Phi(\gamma)} - 1 \right\} \frac{E_{\kappa, \gamma} [e^{-\alpha Y} \tilde{C}(\theta) | R = 1, V]}{E_{\kappa, \gamma} [e^{-\alpha Y} | R = 1, V]},$$

$$\tilde{C}(\theta) = c(Y) - \int c(Y) f(Y; \theta) dY.$$

Note that because of the redefinition of κ , $\theta(\kappa, \gamma) = \theta(\kappa)$, and $f(Y|R = 1, V; \kappa, \gamma)$ is now a function of both γ and κ . RR (2001a) show that no DR estimating function for $\theta(\kappa)$ exists in model $M(\mathcal{K} \times \Gamma)$ with respect to the new parametrization (κ, γ) for any submodels $\Gamma_{sub} \subset \Gamma$ and $\mathcal{K}_{sub} \subset \mathcal{K}$ when $\alpha \neq 0$. However a “generalized” DR estimator $\hat{\theta}(\hat{\tau}, \hat{\gamma}, c)$ is obtained by solving $P_n[\tilde{U}(\theta, \hat{\tau}(\theta), \hat{\gamma}, c)] = 0$ with $\tilde{U}(\theta, \hat{\tau}(\theta), \hat{\gamma}, c) = R\tilde{C}(\theta)\Phi(\hat{\gamma})^{-1} - \{R\Phi(\hat{\gamma})^{-1} - 1\}b(V; \hat{\tau}(\theta))$, where $b(V; \tau(\theta))$ is a user specified model for the ratio $E_{\kappa, \gamma}[e^{-\alpha Y} \tilde{C}(\theta) | R = 1, V] / E_{\kappa, \gamma}[e^{-\alpha Y} | R = 1, V]$ indexed by a finite dimensional parameter τ , and $\hat{\tau}(\theta)$ is the $e^{-\alpha Y}$ -weighted non linear least squares regression estimator of τ solving $P_n[e^{-\alpha Y} R(\tilde{C}(\theta) - b(V; \tau)) \partial b(V; \tau) / \partial \tau] = 0$. The theoretical difficulty with this approach is that the model $b(V; \tau)$ for $E_{\kappa, \gamma}[e^{-\alpha Y} \tilde{C}(\theta) | R = 1, V] / E_{\kappa, \gamma}[e^{-\alpha Y} | R = 1, V]$ will often be incompatible with the model $M(\mathcal{K} \times \Gamma)$, in the sense that there does not exist a joint distribution that satisfies both. In such case, $\hat{\theta}(\hat{\tau}, \hat{\gamma}, c)$ of course cannot be a DR RAL estimator in model $M(\mathcal{K} \times \Gamma)$. However, this theoretical difficulty does not seem to us to be a practical difficulty. After all, as discussed in Section 1, even for models that admit DR estimators, the chosen low dimensional models \mathcal{K}_{sub} and Γ_{sub} are practically (although not logically) certain to be misspecified; thus our best hope is that one of the two submodels is nearly correct, so the bias of the DR estimator will be small. In precise analogy if either model Γ_{sub} for γ or model $b(V; \tau(\theta))$ for $E_{\kappa, \gamma}[e^{-\alpha Y} \tilde{C}(\theta) | R = 1, V] / E_{\kappa, \gamma}[e^{-\alpha Y} | R = 1, V]$ is nearly correct, the bias of $\hat{\theta}(\hat{\tau}, \hat{\gamma}, c)$ for $\theta(\kappa)$ will be small.

Discussion: Heretofore we have been studying union models $M(\mathcal{K} \times \Gamma_{sub}) \cup M(\mathcal{K}_{sub} \times \Gamma)$ that possess a non-empty intersection submodel $M(\mathcal{K}_{sub} \times \Gamma_{sub})$. A consequence of this fact is that, by Theorem 1, any unbiased estimating function $U(\theta, \kappa, \gamma)$ for θ in $M(\mathcal{K} \times \gamma) \cup M(\kappa \times \Gamma)$ must satisfy $U(\theta(\kappa, \gamma), \kappa, \gamma) \in IF(\kappa, \gamma)$ in model $M(\mathcal{K} \times \Gamma)$. It is this consequence that underlies many of the results presented in this discussion. RR (2001a) discuss double robustness in union models with empty intersections.

As far as we are aware Brillinger (1983) was the first to call attention to and provide examples of DR-like estimators. Other examples are given by Ruud (1983, 1986), Duan and Li (1987, 1991), Newey (1990), Robins, Mark and Newey (1992), Ritov and Robins (1997), Lipsitz and Ibrahim (1999). All these examples have $\theta(\kappa, \gamma) = \theta(\kappa)$ and likelihood factorization $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$; thus they are all special cases of the general model treated in Theorem 2 above. Scharfstein et al. (1999) and Robins (2000) went beyond individual examples to provide a general theory of double robustness in missing data and counterfactual causal inference models in which the data was coarsened at random (CAR). Robins et al. (2000) extended these latter results to cover all models with $\theta(\kappa, \gamma) = \theta(\kappa)$ and likelihood factorization $L(\kappa, \gamma) = L_1(\kappa)L_2(\gamma)$; they stated and proved Theorem 2. Scharfstein et al. (1999) treated Example 1 which is the only previous example we have found in the literature in which $\theta(\kappa, \gamma)$ depends on both (κ, γ) and the likelihood does not factor. At present, Theorem 2 seems to be our most significant practical result in the sense that the set of models that are known to admit DR estimators, but that do not satisfy the suppositions of Theorem 2, is still quite small.

Acknowledgement

This work was partially completed while Andrea Rotnitzky was visiting the Department of Economics at Di Tella University, Buenos Aires. James Robins and Andrea Rotnitzky were partially funded by grants from the National Institutes of Health.

Department of Epidemiology and Biostatistics, Harvard School of Public Health, 655 Huntington Ave. Boston, MA 02115, U.S.A.

E-mail: robins@hsph.harvard.edu

Department of Economics, Di Tella University, Minones 2159. Bouenos Aires, Argentina.

E-mail: andrea@hsph.harvard.edu

COMMENTS

Xiaotong Shen and Bing Li

The Ohio State University and The Penn State University

Bickel and Kwon are to be congratulated for their insights into many important issues in semiparametric and nonparametric inferences, and for sharing