Statistica Sinica **31** (2021), 1285-1307
doi:https://doi.org/10.5705/ss.202018.0334

DOUBLY ROBUST REGRESSION ANALYSIS FOR DATA FUSION

Katherine Evans, BaoLuo Sun, James Robins and Eric J. Tchetgen Tchetgen

Verily Life Sciences, National University of Singapore,
Harvard T.H. Chan School of Public Health and
The Wharton School of the University of Pennsylvania

Abstract: This study investigates parametric inferences for the regression of an outcome variable Y on covariates (V, L). Here, the data are fused from two separate sources, one of which contains information only on (V, Y), while the other contains information only on the covariates. This setting may be viewed as an extreme form of missing data in which the probability of observing complete data (V, L, Y) on any given subject is zero. We develop a large class of semiparametric estimators, including doubly robust estimators, of the regression coefficients in the fused data. The proposed method is doubly robust in that it is consistent and asymptotically normal if, in addition to the model of interest, we correctly specify a model for either the data source process under an ignorability assumption, or the distribution of the unobserved covariates. We evaluate the performance of our estimators using an extensive simulation study. Then, we apply the proposed methods to investigate the relationship between net asset value and total expenditure among U.S. households in 1998, while controlling for potential confounders, including income and other demographic variables.

Key words and phrases: Data fusion, doubly robust.

1. Introduction

Parametric likelihood-based inferences for regression analyses is a well-developed area of modern statistical theory. The theory on how to account for incomplete outcome or covariate information in a regression analysis is relatively well established, and includes methods such as inverse probability weighting (IPW) of complete cases and multiple imputation (Robins, Rotnitzky and Zhao (1994); Little and Rubin (2014)). Most missing data methods assume that the probability of observing a subject with complete data is bounded away from zero. Also known as the positivity assumption, this is often necessary to identify the full data law and its smooth functionals (Robins, Rotnitzky and Zhao (1994)). In

Corresponding author: Eric J. Tchetgen Tchetgen, Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: ett@wharton.upenn.edu.

this study, we consider a more extreme form of incomplete data, in which the positivity assumption does not hold; that is, the probability of observing complete data is zero for all units in the population.

This situation may arise, for instance, when two data sets from separate sources are fused together, such that no unit belongs to both sources, and some variables obtained from one source are not available in the other source. For instance, as in this paper, it may be that the outcome of interest Y is collected in the first data set, but not in the second; similarly, a subset of regressors L may be observed in the second data set, but not in the first. Both data sets contain information on the common variables V. This situation is common in main/validation study design of comparative effectiveness studies. Here, the main study sample in contains the outcome, treatment variable, and a relatively limited subset of confounders. Then, this sample is enriched with an external validation sample that contains extensive potential confounders, together with treatment information, but lacks outcome information (Stürmer et al. (2005)). The two data sets are then fused together in the hope that the information from the validation sample can somehow be leveraged to reduce the confounding bias.

Another example, somewhat related to using a meta-analysis to evaluate a prediction model (Riley, Lambert and Abo-Zaid (2010); Debray et al. (2013, 2017)), might involve enriching the data set of a clinical study to improve clinical risk prediction, using covariate information from a separate source, say a study containing socio-demographic or summary-level information, but no outcome data (Chen and Chen (2000); Chatterjee et al. (2016)). Clearly, in both examples, a regression model for the outcome on the combined set of covariates can be identified only under fairly stringent parametric assumptions and, as we discuss below, provided there is a nontrivial overlap in the amount of information available from both sources of data. We shall refer to this general framework as regression analysis for data fusion.

The missing data literature has previously described the data fusion problem as that of "statistical matching." D'Orazio, Di Zio and Scanu (2006) and Rässler (2012) provide extensive overviews of the state of the art for data fusion. D'Orazio, Di Zio and Scanu (2010)compare existing data-matching methods, as well as the assumptions needed to recover valid inferences using these methods. Much of this literature relies on the assumption of conditional independence between Y and L, given V, an assumption that is likely untenable in practice. This assumption is particularly problematic in the two settings described above, where a potential non-null association between Y and L, given V, is an important part of the scientific hypothesis under consideration. When the samples are drawn

from a finite population according to a complex survey design, concatenation (Rubin (1986)) and calibration (Renssen (1998); Wu (2004)) are two commonly used methods for statistical matching. Concatenation involves modifying the sample weights to obtain a unique sample given by the union of the original sample and the new weights that represent the population of interest. The new weights require computing the probability of the subjects in one sample under the survey design of the other sample, which requires detailed knowledge of the survey designs. Calibration preserves both samples and calibrates the two sets of survey weights. The method obtains a unique estimate of the distribution of the common variable, V, by combining the estimates of the distribution of V from both samples, and then calibrating the original sample weights to the obtained estimate. The weights are then used to estimate the distribution $f(L|V)$ in the sample with L, and the distribution $f(Y|V)$ in the sample with outcome Y. Wu (2004) suggests similar approaches with different constraints for the sample weights, such as forbidding negative weights. Conti, Marella and Scanu (2016) estimate the distribution function of variables not jointly observed in the presence of logical constraints, without necessarily imposing the conditional independence assumption, and the corresponding bounds for the matching error can be estimated from the sample data. Graham, Pinto and Egel (2016) propose a general framework for data combination under moment restrictions and estimators that are doubly robust (DR) only under a restricted model specification of the nuisance parameters.

Data fusion is also prominent in the literature on instrumental variable (IV) methods for causal inference. An IV is an exogenous variable known to be associated with a treatment or exposure variable of interest, and to be associated with an outcome of interest only through its association with the treatment. The IV approach can, under certain conditions, be used to recover an unbiased estimate of a causal effect in the presence of unmeasured confounding. The most common IV approach assumes a linear model relating the outcome to the exposure and the observed covariates, together with a linear model relating the exposure to the IV and the covariates. Angrist and Krueger (1992) examine estimation and inference related to the causal effect of the exposure under such linear models when the IV and exposure are available from one data source, while the outcome and IV are available in a separate data source. As such, no subject has data available on all three variables, namely, the IV, exposure, and outcome. These two-sample IV estimators deliver point identification and inference by explicitly leveraging parametric assumptions. Klevmarken (1982) proposed the two-sample two-stage least squares regression, which was later shown by Inoue and Solon (2010) to

be more efficient than the two-sample IV estimator. These methods assume that both samples are independent and identically distributed (i.i.d.) random samples from the same population with finite fourth moments. Pacini (2017) assumes the samples are independent,and uses the marginal distributions to characterize the identified set of the coefficients of interest when no assumption is imposed on the joint distribution of (Y, V, L).

Robins, Hsieh and Newey (1995) consider a missing data setting closely related to ours. The main contribution of their study is to characterize a large class of semiparametric estimators of a parametric conditional density of Y, given (L, V), when L is missing at random. They consider a general semiparametric missing data model in which the model for the full data constitutes the only restriction. Then, they derive the efficient influence function for the parameters of the parametric model, which is the solution to an integral equation that is not generally available in closed form. They also point out in a remark that Bickel et al. (1993) and Hasminskii and Ibragimov (1983) obtained results similar to theirs when Y and L are never observed together, which is the data fusion setting we address here.

An important contribution of our study is to show that a large class of influence functions for the parameters of the conditional density $f(Y|L, V)$ is available in closed form in a missing data model that is otherwise unrestricted; therefore, these functions are convenient candidates as estimating functions. The proposed semiparametric estimating functions include DR estimating functions that yield estimators that are consistent and asymptotic normal if, in addition to the outcome model of interest, we correctly specify the model for the data source process or the distribution of the unobserved covariates. Importantly, unlike Graham, Pinto and Egel (2016), we do not restrict the specification of the nuisance models to belong to a certain class of models. For example, their DR result holds only if the missing data model is specified as a certain logistic regression model. In addition, we show that the efficient influence function for the parameters of the conditional density is available in closed form in the special case where the outcome is polytomous.

In Section 2, we lay out notation and assumptions. In Section 3, we develop the general class of estimators, as well as a new semiparametric DR method. In Section 4, we discuss the implementation. In Section 5, we discuss the local efficiency in the special case of a binary outcome, although the result readily generalizes to a polytomous outcome. Here, we also provide approximately efficient influence functions in the case of a continuous outcome. We evaluate the finite-sample performance of the DR approach in an extensive simulation study

summarized in Section 6. Then, in Section 7, we demonstrate the proposed methods using fused data provided by the U.S. Bureau of Labor Statistics' Consumer Expenditure Survey (CEX) and the Federal Reserve Board's Survey of Consumer Finances (SCF). Section 8 concludes the paper.. All proofs and derivations can be found in the online Supplementary Material.

2. Notation and Assumptions

Let R be an indicator that a subject is observed in data source $\mathcal{A}$ ($R = 1$) or in data source $\mathcal{B}$ ($R = 0$). Let V denote the covariates observed in both sources, Y denote the outcome observed only in source $\mathcal{A}$, and L denote the covariates observed only in source $\mathcal{B}$. The full data (Y, L, V) are i.i.d. realizations from a common law $f(Y, L, V)$. Let $f(Y|V, L)$ denote the true conditional distribution of Y, given (V, L). Let $\pi(V) = \Pr(R = 1|V)$ be the probability that a subject is in data source $\mathcal{A}$. Throughout, we make the following assumptions:

A1 Correct outcome model: $f(Y|V, L; \theta)$ is correctly specified, such that $f(Y|V, L; \theta^\dagger) = f(Y|V, L)$, for some value $\theta^\dagger$;

A2 Positivity: $\delta < \pi(V) < 1 - \delta$ almost surely, for a fixed positive constant δ;

A3 Ignorability: $R \perp (Y, L)|V$.

In addition, we let $\mathcal{M}$ denote the set of models that satisfy (A1–3). Assumption (A1) requires that the outcome model proposed for f is correctly specified. The positivity assumption (A2) states that the probability of observing a subject in either data source is bounded away from both zero and one. Note that (A2) is strictly weaker than the positivity assumption typically assumed in missing data problems, which requires a positive probability of observing complete data for each subject. Assumption (A3) states that the probability that a unit is observed in either data source depends only on V, and thus does not depend on Y or L. This assumption is akin to missing at random, and is imposed on the data source process, which is technically a nuisance parameter not of primary scientific interest. In contrast, the conditional independence assumption $Y \perp L|V$ is imposed on the full data law of primary interest by some existing methods, such as matching (D'Orazio, Di Zio and Scanu (2010)).

Assumption (A3) is satisfied for a practically relevant type of stratified sampling in which either or both of the two samples use sampling rates that vary with some of the fully observed baseline variables. For example, household surveys commonly use different sampling rates depending on demographic variables

such as age and race. In our empirical study on fused data from the U.S. Bureau of Labor Statistics' CEX and the Federal Reserve Board's SCF, the former oversamples relatively wealthy families from the population. Another example concerns Mendelian randomization (MR) studies, in which one aims to establish a causal relationship between the exposure and the outcome by leveraging one or more genetic markers defining the IV (Davey Smith and Ebrahim (2003); Lawlor et al. (2008); Burgess, Small and Thompson (2017)). It is common for the sampling mechanisms of studies to differ. For example, the US-based Health and Retirement Study has oversampled residents of Florida. As a result, the data vary with fully observed demographic variables and ancestry, which may correlate with the instruments. An important difference between assumption (A3) and the conditional independence assumption is that, while the former can be guaranteed to hold if the investigator allow the sampling mechanism to depend only on V in the design stage, the latter relates to the full data law of primary interest and, thus, is beyond the control of the investigator.

3. Estimating Functions

In this section, we describe a large class of IPW estimating functions for θ under various sets of modeling assumptions of nuisance parameters. Let $\pi(V;\eta) = P(R=1|V;\eta)$ denote a parametric model for the data source process indexed by a finite-dimensional parameter η. We make the following assumption:

A4 $\pi(V;\eta)$ is correctly specified, such that $\pi(V;\eta^*) = \pi(V)$, for some value η^*.

Let $\mathcal{M}_\pi = \mathcal{M} \cap \{\pi(V;\eta) : \eta\}$. For a user-specified function $g(Y,V)$ of (Y,V), let

$$U_g(\theta;\eta) = \frac{R}{\pi(V;\eta)} g(Y,V) - \frac{1-R}{1-\pi(V;\eta)} E_\theta[g(Y,V)|V,L]. \tag{3.1}$$

Below, we discuss the assumptions $g(Y,V)$ must satisfy to ensure identification.

Result 1. Under $\mathcal{M}_\pi$,

$$E_{\eta^*}\left[U_g(\theta^\dagger;\eta^*)\right] = 0. \tag{3.2}$$

The parallel IPW estimating function given in (3.1) assigns to every subject the inverse probability of observing the subject from the data source in which he or she was indeed observed. Interestingly, this general class of estimating functions includes a large set of DR estimating functions. Suppose that one has specified a parametric model $t(L,V;\alpha)$ for the density $f(L|V)$ of L, given V.

A5 $t(L,V;\alpha)$ is correctly specified, such that $t(L,V;\alpha^\ddagger) = t$, for some value $\alpha^\ddagger$.

Let $\mathcal{M}_t = \mathcal{M} \cap \{t(L, V; \alpha) : \alpha\}$. Then, let

$$U_g^{DR}(\theta; \eta, \alpha) = \frac{R}{\pi(V; \eta)} \{g(Y, V) - E_{\theta,\alpha}[g(Y, V)|V]\} + \frac{1 - R}{1 - \pi(V; \eta)} \{E_{\theta,\alpha}[g(Y, V)|V] - E_\theta[g(Y, V)|V, L]\}, \quad (3.3)$$

where we note the dependence of

$$E[g(Y, V)|V = v] = \int \int g(y, v) f(y|v, l; \theta) t(l, v; \alpha) dy dl$$

on (θ, α).

Result 2. Under the union model $\mathcal{M}_{\pi \cup t} = \mathcal{M}_\pi \cup \mathcal{M}_t$,

$$E_{\eta^*, \alpha^\ddagger}\left[U_g^{DR}(\theta^\dagger; \eta, \alpha)\right] = 0, \quad (3.4)$$

if either $\eta = \eta^*$ or $\alpha = \alpha^\ddagger$, but not necessarily both.

The estimating function (3.3) is said to be DR for θ in that estimators based on (3.3) are consistent for $\theta^\dagger$, provided that we correctly specify a model for $t(V; \alpha)$ or $\pi(V; \eta)$, but not necessarily both. Additionally, when both models are correctly specified, the estimator for θ based on $U_g^{DR}(\theta; \eta, \alpha)$ is the most efficient (for a fixed choice of g) in $\mathcal{M}_{\pi \cup t}$.

Note that owing to the DR property of the estimating function given in (3.3), its unbiasedness still holds for any choice of $\pi(V)$ if the conditional density $t(L, V; \alpha)$ is correctly specified. Heuristically, the resulting estimator works by correctly imputing the missing values in L conditional on V. For a user-specified function $g(Y, V)$, let

$$U_g^{imp}(\theta; \alpha) = U_g^{DR}(\theta^\dagger; \pi = 0.5, \alpha) \quad (3.5)$$

$$\propto R\{g(Y, V) - E_{\theta,\alpha}[g(Y, V)|V]\} + \quad (3.6)$$

$$(1 - R)\{E_{\theta,\alpha}[g(Y, V)|V] - E_\theta[g(Y, V)|V, L]\}. \quad (3.7)$$

Corollary 1. *Under* $\mathcal{M}_t$,

$$E_{\alpha^\ddagger}\left[U_g^{imp}(\theta^\dagger; \alpha^\ddagger)\right] = 0. \quad (3.8)$$

In the next section, we construct feasible IPW, imputation (IMP), and DR estimators as solutions to the empirical versions of (3.2), (3.4), and (3.8), respectively. We also describe the large-sample behavior of the resulting estimators of θ.

4. Implementation of Estimators

Deriving feasible IPW, IMP, and DR estimators involves a first-stage estimation of the nuisance parameters η and α. We propose the following estimator for η, which maximizes the log-likelihood,

$$\hat{\eta} = \arg\max \sum_i \{R_i \log \pi(V_i;\eta) + (1-R_i)\log[1-\pi(V_i;\eta)]\}. \tag{4.1}$$

By ignorability assumption (A3), α can be estimated using the likelihood maximization restricted to sample $\mathbb{B}$. That is,

$$\hat{\alpha} = \arg\max \left\{ \sum_i (1-R_i)\log t(L_i, V_i;\alpha) \right\}. \tag{4.2}$$

As pointed out by a reviewer, although the components $t(L,V)$ and $\pi(V)$ of the observed data law are variationally independent, in general it is more challenging to model the former, especially when L or V is of moderate dimension and consists of a mixture of discrete and continuous variables. Let $W_d = (L_d, V_d)$ and $W_c = (L_c, V_c)$ denote the discrete and continuous components of (L,V), respectively. One strategy is to adopt the general location model (Olkin and Tate (1961)), defining the joint density $f(L,V)$ in terms of the marginal distribution of W_d and the conditional distribution of W_c, given W_d. The former is described by a multinomial distribution on the cell counts x, $x|\rho \sim M(n,\rho)$, where ρ is an array of cell probabilities of the same dimension as the number of possible values in the support of W_d. Then W_c is modeled as conditionally multivariate normal, given W_d. The general location model is also amenable to restrictions on the parameter space when the number of possible values in the support of W_d is large relative to the sample size and can be estimated using maximum likelihood methods (Little and Schluchter (1985)). Finally, $t(L,V) = f(L,V)/f(V)$, where $f(V)$ is obtained by marginalizing out L; closed-form expressions for the conditional density can be found in the missing data literature on building predictive distributions for multiple imputation under the general location model (Schafer (1997)).

Let $\mathbb{P}_n$ denote the empirical mean operator $\mathbb{P}_n f(O) = n^{-1}\sum_i f(O_i)$. Then, the IPW, IMP, and DR estimates of θ are solutions to the estimating functions $\mathbb{P}_n\{U_g(\theta;\hat{\eta})\} = 0$, $\mathbb{P}_n\left\{U_g^{imp}(\theta;\hat{\alpha})\right\} = 0$, and $\mathbb{P}_n\left\{U_g^{DR}(\theta;\hat{\eta},\hat{\alpha})\right\} = 0$, respectively. Under the standard regularity conditions given in Theorem 2.6 of Newey and McFadden (1994), the resulting IPW estimator of θ is consistent if $\pi(V;\eta)$ is correctly specified, and the DR estimator is consistent if either $\pi(V;\eta)$ or $t(L,V;\alpha)$, but not necessarily both, is correctly specified.

To illustrate, suppose that we have univariate Y, p-dimensional L, and q-dimensional V, which are all continuous, with a constant term embedded in V. Let A^T denote the transpose of A. The IPW estimation proceeds by first obtaining $\hat{\eta}$. For example, assuming a logistic model $\pi(V;\eta) = (1+\exp^{-V^T\eta})^{-1}$, we then solve (4.1) by fitting a logistic regression to the observed data (R,V). The DR estimation also requires that we estimate $\hat{\alpha}$. Suppose the conditional density of L, given V, is multivariate normal $\mathcal{N}(\alpha^T V,\Sigma)$, where the errors in Σ may be correlated, but do not vary among observations. The $q\times p$ estimate $\hat{\alpha}$ can be computed using least squares estimation $\hat{\alpha} = \left(L, V_{\mathbb{B}}^T V_{\mathbb{B}}\right)^{-1} V_{\mathbb{B}}^T L_{\mathbb{B}}$, where $(L, V_{\mathbb{B}}, L_{\mathbb{B}})$ is the $n\times(p+q)$ covariate matrix from data source $\mathbb{B}$ with n observations. Finally, we assume that $Y|V,L$ is normally distributed as $\mathcal{N}(\beta^T(V^T,L^T)^T,\Sigma)$, $\theta=(\beta,\Sigma)$. If we are primarily interested in the mean parameters β, and not the variance component Σ, then a convenient choice for $g(Y,V)$ is given by $Yg(V)$, where $g(V)$ is of the same dimension as β. Then, we have the following set of estimating functions:

$$U_g(\theta;\eta) = g(V)\left\{\frac{R}{\pi(V;\eta)}Y - \frac{1-R}{1-\pi(V;\eta)}E_\theta[Y|V,L]\right\}, \tag{4.3}$$

$$\begin{aligned} U_g^{DR}(\theta;\eta,\alpha) = g(V)\Bigg\{&\frac{R}{\pi(V;\eta)}\left\{Y - E_{\theta,\alpha}\left[Y|V\right]\right\} \\ &+ \frac{1-R}{1-\pi(V;\eta)}\left\{E_{\theta,\alpha}\left[Y|V\right] - E_\theta[Y|V,L]\right\}\Bigg\}, \end{aligned} \tag{4.4}$$

$$\begin{aligned} U_g^{imp}(\theta;\alpha) = g(V)\Big\{&R\left\{Y - E_{\theta,\alpha}\left[Y|V\right]\right\} \\ &+(1-R)\left\{E_{\theta,\alpha}\left[Y|V\right] - E_\theta[Y|V,L]\right\}\Big\}, \end{aligned} \tag{4.5}$$

where $E_\theta[Y|V,L] = \beta^T(V^T,L^T)^T$ and $E_{\theta,\alpha}[Y|V] = \beta^T(V^T,V^T\alpha)^T$.

In general, if we are interested in estimating the full set of parameters θ that indexes the assumed parametric model $f(Y|V,L;\theta)$, the choice of $g(Y,V)$ should be such that it is of at least the same dimension as θ, where $E\left[U_g^T(\theta)U_g(\theta)\right] < \infty$ and $E\left[\partial U_g(\theta,\eta)/\partial\theta\right]$ is nonsingular. Because

$$\begin{aligned} &E\left[\frac{\partial}{\partial\theta}U_g(\theta,\eta)\right] \\ &= E\left\{E[g(Y,V)S^T(Y|V,L;\theta)|V,L]\right\} \\ &= E\left\{\frac{R}{\pi(V;\eta)}g(Y,V)E[S^T(Y|V,L;\theta)|Y,V]\right\} \\ &= E\left\{g(Y,V)S^T(Y|V,L;\theta)\right\}, \end{aligned}$$

where $S(Y|V,L;\theta)$ is the full data score vector associated with the conditional

density $f(Y|V,L;\theta)$, a sufficient rank condition for a local identification of $\theta^\dagger$ is that the matrix

$$\Omega \equiv E\left\{g(Y,V)S^T(Y|V,L;\theta^\dagger)\right\} \quad \text{is nonsingular} \tag{4.6}$$

(Newey and McFadden (1994)). Similar derivations for the IMP and DR estimating functions show that the form of Ω is the same. We provide two examples below.

Example 1. Suppose Y is binary and (V,L) are two scalar continuous random variables. Then, assuming the logistic model $f(Y=1|V,L;\theta) = \{1+\exp[-\theta_0 + \theta_1 V + \theta_2 L]\}^{-1}$, we have $S(Y|V,L;\theta) = [Y - f(Y=1|V,L;\theta)](1,V,L)^T$. For the choice $g(Y,V) = Y[1,V,h(V)]^T$, $\Omega = E\{\text{var}(Y|V,L;\theta^\dagger)[1,V,h(V)]^T(1,V,L)\}$; thus, local identification requires a choice of function $h(V)$ such that $h(V)$ is correlated with L. A simple choice is $h(V) = V^2$. Identification would then fail if L and V^2 are uncorrelated.

Example 2. If we are only interested in the conditional mean of $f(Y|V,L)$, we can specify a parametric model $E(Y|V,L;\theta)$ directly. For example, suppose (Y,V,L) are three scalar continuous random variables, and we specify $E(Y|V,L;\theta) = \theta_0 + \theta_1 V + \theta_2 L$, $g(Y,V) = Y[1,V,h(V)]^T$ for the user-specified nonlinear function $h(\cdot)$. Then,

$$\frac{\partial}{\partial\theta} E\left\{U_g(\theta,\eta)\right\} = E\left\{[1,V,h(V)]^T(1,V,L)\right\},$$

such that identification requires that $h(V)$ be correlated with L.

Note that the generalized method of moments (GMM) approach can be adopted to allow $g(Y,V)$ to have a larger dimension than that of θ.

Let ϕ denote the set of nuisance parameters, that is, $\phi = \eta$, $\phi = (\eta,\alpha)$, and $\phi = \alpha$ for the IPW, DR, and imputation estimations, respectively, and let ϕ^* denote the probability limit of $\hat{\phi}$. The scores for the nuisance parameters are

$$S_\eta = \frac{d}{d\eta}\log\left\{\pi(V;\eta)^R[1-\pi(V;\eta)]^{1-R}\right\}$$
$$S_\alpha = \frac{d}{d\alpha}\log\left\{t(L|V;\alpha)^{1-R}\right\}.$$

Let $S_\phi = S_\eta$, $S_\phi = \left(S_\eta^T, S_\alpha^T\right)^T$, and $S_\phi = S_\alpha$ for the IPW, DR, and imputation-based estimations, respectively, and let

$$U_{\theta,\phi} = \begin{cases} \left(U_g^T(\theta;\eta), S_\phi^T\right)^T, & \text{for IPW} \\ \left(U_g^{DR,T}(\theta;\eta,\alpha), S_\phi^T\right)^T, & \text{for DR estimation} \\ \left(U_g^{imp,T}(\theta;\alpha), S_\phi^T\right)^T, & \text{for imputation-based estimation.} \end{cases}$$

In addition, let

$$\begin{aligned} G_\theta &= E\left[\frac{\partial}{\partial\theta} U_{\theta^\dagger,\phi^*}\right] \\ G_\phi &= E\left[\frac{\partial}{\partial\phi} U_{\theta^\dagger,\phi^*}\right] \\ M &= E\left[\frac{\partial}{\partial\phi} S_{\phi^*}\right] \\ \Psi &= -M^{-1} S_{\phi^*}, \end{aligned}$$

where all the expectations are evaluated at the true parameter values. Then, under the standard regularity conditions given in Theorem 6.1 of Newey and McFadden (1994),

$$\sqrt{n}\left(\hat{\theta} - \theta^\dagger\right) \xrightarrow{d} \mathcal{N}(0, \Sigma_\theta), \tag{4.7}$$

where

$$\Sigma_\theta = G_\theta^{-1} E\left\{\left[U_{\theta^\dagger,\phi^*} + G_\phi\Psi\right]\left[U_{\theta^\dagger,\phi^*} + G_\phi\Psi\right]^T\right\} G_\theta^{-1,T}. \tag{4.8}$$

For inference purposes, a consistent estimator $\hat{\Sigma}_\theta$ of the asymptotic covariance matrix given in (4.8) can be constructed by replacing all expected values with the empirical averages evaluated at $(\hat{\theta}, \hat{\phi})$. Then, a 95% Wald confidence interval for θ_j is found by calculating $\hat{\theta}_j \pm 1.96\hat{\sigma}_j$, where $\hat{\sigma}_j$ is the square root of the j^{th} component of the diagonal of $n^{-1}\hat{\Sigma}_\theta$. Alternatively, a nonparametric bootstrap can be performed to obtain estimates of the variance.

In general, the choice of the function $g(Y, V)$ used to index various estimating equations affects the efficiency, but not the consistency of the resulting estimators (provided it satisfies the identification condition (4.6)). Modern semiparametric efficiency theory may be used to identify an optimal choice for such an index function that minimizes the first-order asymptotic variance of the resulting estimator (Bickel et al. (1993)). The optimal choice of $g(Y, V)$ should, in fact, ensure identification whenever the model is identified, without necessarily having to consider a large number of candidate choices for such a function. When Y contains continuous components, the optimal choice of $g(Y, V)$ is, in general, not available in closed form. Section 5 provides the results for the optimal index when

Y is categorical, as well as methods to construct approximately locally efficient estimators when Y is continuous.

5. Local Efficiency

For binary Y, any function $g(\cdot)$ of Y and V can be expressed as $g(Y,V) = Yg_1(V) + g_0(V)$, where $g_1(\cdot)$ and $g_0(\cdot)$ are arbitrary functions of V. Therefore, the class of DR estimating functions in (3.3) is equivalently given by

$$\mathcal{L}_{DR} = \{g_1(V)M(\theta) : g_1(\cdot) \text{ arbitrary}\},$$

where

$$M(\theta) = \frac{R}{\pi(V;\eta)}\{Y - E_{\theta,\alpha}[Y|V]\} + \frac{1-R}{1-\pi(V;\eta)}\{E_{\theta,\alpha}[Y|V] - E_{\theta}[Y|V,L]\}.$$

We have the following result.

Result 3. Suppose $\hat{\theta}_h$ is a regular and asymptotically linear (RAL) estimator of θ in the semiparametric model $\mathcal{M}_{\pi\cup t}$. Then,

$$\sqrt{n}\left(\hat{\theta}_h - \theta^{\dagger}\right) \xrightarrow{D}$$
$$\mathcal{N}\left(0, E\left[h(V)\nabla_\theta M(\theta)\right]^{-1} E\left\{M^2(\theta^{\dagger})h(V)h(V)^T\right\} E\left[h(V)\nabla_\theta M(\theta)\right]^{-1T}\right),$$

for some $h(V)M(\theta) \in \mathcal{L}_{DR}$. Here, $\hat{\theta}_{\hat{h}}$ achieves the semiparametric efficiency bound for $\mathcal{M}_{\pi\cup t}$ at the intersection submodel $\mathcal{M}_{\pi} \cap \mathcal{M}_t$ if $\hat{h}$ converges in probability to

$$h^{opt}(V) = -E\left[\nabla_\theta M(\theta)|V\right] E\left[M^2(\theta)|V\right]^{-1}.$$

Using a similar approach, Result 3 can easily be extended to polytomous Y, with $s > 2$ levels, by noting that $g(Y,V) = \sum_{k=1}^{s-1} I(Y = y_k)g_k(V) + g_0(V)$ and, therefore,

$$\mathcal{L}^s_{DR} = \left\{\sum_{k=1}^{s-1} g_k(V)M_k(\theta) : g_k(\cdot) \text{ arbitrary for } k = 1,2,\ldots,s-1\right\},$$

where

$$\begin{aligned} M_k(\theta) =& \frac{R}{\pi(V;\eta)}\{I(Y = y_k) - P(Y = y_k|V;\theta,\alpha)\} \\ &+ \frac{1-R}{1-\pi(V;\eta)}\{P(Y = y_k|V;\theta,\alpha) - P(Y = y_k|V,L;\theta)\}, \\ k =& 1,2,\ldots,s-1. \end{aligned}$$

When Y contains continuous components, the semiparametric efficient influence function for θ is, in general, not available in closed form, in the sense that it cannot be expressed explicitly as a function of the true distribution (Robins, Hsieh and Newey (1995)). Let $L_2 \equiv L_2(F)$ denote the Hilbert space of zero-mean functions of p dimensions, $Z \equiv z(V,Y)$, with inner product $E_F\left(Z_1^T Z_2\right) = E\left(Z_1^T Z_2\right)$, and the corresponding squared norm $||Z||^2 = E\left(Z_1^T Z_2\right)$, where F is the distribution function that generated the data. We adopt the general strategy proposed in Newey (1993) (see also Tchetgen Tchetgen, Robins and Rotnitzky (2009)) to obtain an approximately locally efficient estimator. As such, we take a basis system $\psi_j(Y,V)$ $(j = 1,\ldots)$ of functions dense in L_2, such as the tensor products of trigonometric, wavelet, or polynomial bases for the controls V and Y. For approximate efficiency, in practice, we let the p-dimensional $g_K(Y,V) = \tau\Psi_K$, where $\tau \in \mathbb{R}^{p\times K}$ is a constant matrix, and $\Psi_K = \{\psi_1, \psi_2, \ldots, \psi_K\}^T$, for some finite $K > p$.

To derive an approximately locally efficient estimator for θ, let $\mathcal{K}$ denote the linear operator

$$\mathcal{K}(\cdot) = \frac{R}{\pi(V;\eta)}\left\{\cdot - E_{\theta,\alpha}\left[\cdot|V\right]\right\} + \frac{1-R}{1-\pi(V;\eta)}\left\{E_{\theta,\alpha}\left[\cdot|V\right] - E_\theta[\cdot|V,L]\right\},$$

defined over the space of arbitrary functions of Y and V in L_2. Consider the class of influence functions of the form

$$\mathcal{L}_{\Psi_K} = \left\{\tau\mathcal{K}(\Psi_K) = \tau\left[\mathcal{K}(\psi_1), \mathcal{K}(\psi_2), \ldots, \mathcal{K}(\psi_K)\right]^T : \tau \in \mathbb{R}^{p\times K}\right\}.$$

Analogous to Result 3, it can be shown based on Theorem 5.3 in Newey and McFadden (1994) that the efficient estimator of all estimators with influence functions of the form in $\mathcal{L}_{\Psi_K}$ is indexed by the constant matrix

$$\tau^{opt} = -E\left[\nabla_\theta\mathcal{K}(\Psi_K)\right] E\left[\mathcal{K}(\Psi)\mathcal{K}^T(\Psi_K)\right]^{-1}.$$

In particular, the inverse of the asymptotic variance of the estimator indexed by τ^{opt} is

$$\begin{aligned}\Omega_K &= E\left\{\nabla_\theta\mathcal{K}(\Psi_K)\right\}^T E\left\{\mathcal{K}(\Psi_K)\mathcal{K}^T(\Psi_K)\right\}^{-1} E\left\{\nabla_\theta\mathcal{K}(\Psi_K)\right\}\\ &= E\left\{S_\theta\mathcal{K}^T(\Psi_K)\right\} E\left\{\mathcal{K}(\Psi_K)\mathcal{K}^T(\Psi_K)\right\}^{-1} E\left\{S_\theta\mathcal{K}^T(\Psi_K)\right\}^T,\end{aligned}$$

evaluated at $\theta = \theta^\dagger$, and S_θ is the score vector with respect to θ. Thus, Ω_K is the variance of the population least squares regression of S_θ on the linear span of $\mathcal{K}(\Psi_K)$. Because Ψ_K is dense in L_2, as the dimension $K \to \infty$, the linear span

of $\mathcal{K}(\Psi_K)$ recovers the subspace in the orthocomplement nuisance tangent space $\Lambda^{\perp}$, which contains the efficient score $S_{\theta,\text{eff}}$. As a result, $\Omega_K \to ||\Pi\left(S_\theta|\Lambda^{\perp}\right)||^2 = \text{var}\left(S_{\theta,\text{eff}}\right)$, the semiparametric information bound for estimating $\theta^\dagger$ in the union model $\mathcal{M}_{\pi\cup t}$.

6. Simulation Study

In this section, we report a simulation study evaluating the finite-sample performance of our proposed estimators involving i.i.d. realizations of $(R, RY, (1-R)L, V)$. For each of the sample sizes $n = 500, 2000$, we simulated 1,000 data sets, as follows:

$$\begin{aligned} C \sim \mathcal{N}(0, 0.5^2), \quad A|C \sim \mathcal{N}\left(\lambda_0 + \lambda_1 C, \sigma_A^2\right), \quad V = (A, C) \\ L|V \sim \mathcal{N}(\alpha_0 + \alpha_1 A + \alpha_2 C + \alpha_3 AC, \sigma_L^2) \\ R|V \sim \text{Bernoulli}\left\{\pi(V;\eta)\right\}, \quad \pi(V;\eta) = \left(1 + \exp^{-\eta_0 - \eta_1 A + \eta_2 C}\right)^{-1} \\ Y|V, L \sim \mathcal{N}(\beta_0 + \beta_1 A + \beta_2 C + \beta_3 L, \sigma_Y^2), \end{aligned}$$

with $(\lambda_0, \lambda_1, \sigma_A) = (0.5, 0.5, 0.3)$, $(\alpha_0, \alpha_1, \alpha_2, \alpha_3, \sigma_L) = (-0.5, 1.5, 1.0, 2.0, 0.3)$, $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma_Y) = (0.5, -0.5, 1.0, 1.5, 0.4)$, and $(\eta_0, \eta_1, \eta_2) = (0.5, -0.75, -0.75)$, such that, marginally, $\Pr(R = 1) \approx 0.5$. Our aim is to estimate the conditional mean parameters $\beta = (\beta_0, \beta_1, \beta_2, \beta_3)$, based on the observed data, by solving empirical versions of (4.3–4.5) for the IPW, DR, and imputation-based estimations, respectively, with $g(V) = (1, A, C, AC)^T$, using the R package "BB" (Varadhan et al. (2009)). In each simulated sample, we estimated the proposed estimators' asymptotic variance given by (4.8); the Wald 95% confidence interval coverage rates were computed across the 1,000 simulations.

We also evaluated the performance of the proposed estimators in situations where some models may be misspecified. Let the superscript § denote the probability limits from fitting the misspecified models. The data source model was misspecified as $\tilde{\pi}$ by dropping C from the logistic model; that is, $\tilde{\pi}(V;\eta^{\S}) = (1 + \exp^{-\eta_0^{\S} - \eta_1^{\S} A})^{-1}$. The density of $L|V$ was misspecified as $\tilde{t}$ by fitting a standard linear regression using only (C, C^2) as regressors; that is, $E[L|V;\alpha^{\S}] = \alpha_0^{\S} + \alpha_1^{\S} C + \alpha_2^{\S} C^2$. We explored four scenarios: (i) correct models π and t; (ii) correct t, but incorrect model $\tilde{\pi}$; (iii) correct π, but incorrect model $\tilde{t}$; and (iv) incorrect models $\tilde{\pi}$ and $\tilde{t}$. Figure 1 presents the estimation results for the regression coefficient β_3 and Table 1 shows the corresponding empirical coverage rates; the results for the remaining regression coefficients $(\beta_0, \beta_1, \beta_2)$ are qualitatively similar and, therefore, relegated to the appendix.

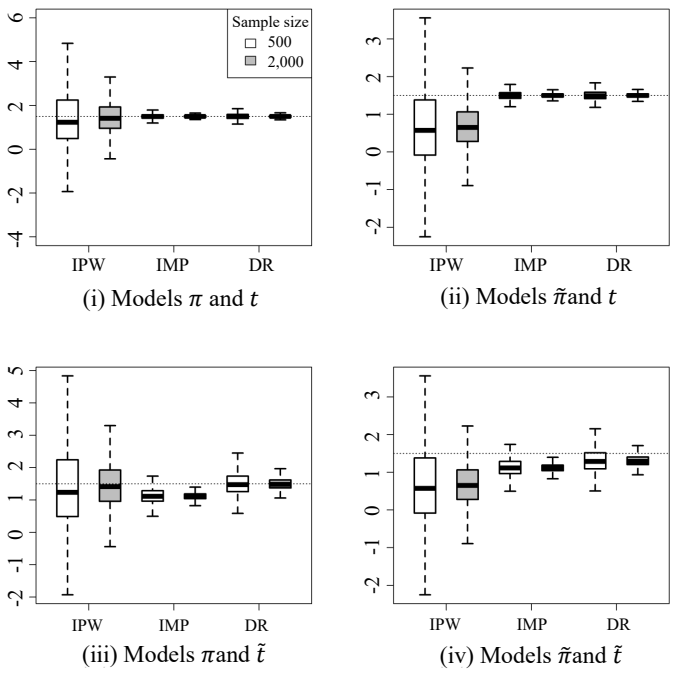

Figure 1. Box plots of inverse probability weighted (IPW), imputation-based (IMP), and doubly robust (DR) estimators of the regression coefficient β_3, which has a true value of 1.5, as marked by the horizontal line, when $\alpha_3 = 2$.

Table 1. Empirical coverage rates based on 95% Wald confidence intervals, as well as the accuracy of the standard deviation estimator, under four scenarios: (i) correct π and t; (ii) correct t, but incorrect $\tilde{\pi}$; (iii) correct π, but incorrect $\tilde{t}$; and (iv) incorrect $\tilde{\pi}$ and $\tilde{t}$. In each scenario, the first row presents the results for $n = 500$, and the second row shows those for $n = 2{,}000$.

	Coverage			SD ratio†		
	IPW	IMP	DR	IPW	IMP	DR
(i)	0.916	0.935	0.926	1.178	0.955	0.899
	0.948	0.939	0.938	1.141	0.972	0.958
(ii)	0.801	0.935	0.923	1.164	0.955	0.913
	0.681	0.939	0.941	1.125	0.972	0.958
(iii)	0.916	0.553	0.888	1.178	0.939	0.876
	0.948	0.139	0.938	1.141	1.038	0.998
(iv)	0.801	0.553	0.740	1.164	0.939	0.894
	0.681	0.139	0.634	1.125	1.038	1.016

† : Estimated SD/Monte Carlo SD

Under the correct model specifications (i), the IPW estimator has a small bias at $n = 500$, which diminishes with an increase in the sample size, while the DR and imputation-based estimators have negligible bias. Supporting our theoretical results, the IPW estimator is significantly biased in scenarios (ii) and (iv) where the data source process is incorrectly modeled as $\tilde{\pi}$. The DR estimator shows negligible bias across scenarios (i)–(iii), and only exhibits a significant bias in scenario (iv), where both models are misspecified as $\tilde{\pi}$ and $\tilde{t}$. The imputation-based estimator shows little bias in scenarios (i) and (ii), but exhibits significant bias in scenarios (iii) and (iv), with misspecified $\tilde{t}$. Under the data-generating mechanism considered in this simulation study, the imputation-based estimator is more efficient than the DR estimator, which is, in turn, more efficient than the IPW estimator across all scenarios considered. The efficiency of the DR estimator is reduced to a greater extent by the misspecification of t than it is by that of π. In the scenarios where the IPW, DR, and imputation-based estimators are unbiased, the empirical coverage rates are slightly lower than 0.95 at $n = 500$, but approach the nominal rate as the sample size increases.

In the Supplementary Material, we provide a second set of simulations in which the coefficient for the interaction between A and C in the model for generating L is reduced to $\alpha_3 = 0.5$, with all other parameters unchanged. When the effect of the (A, C) interaction in the model that generates L is weak, using AC in $g(V)$ leads to an increase in the finite-sample bias and a decrease in efficiency for all estimators considered here.

7. Application

As an empirical illustration, we apply the proposed methods to investigate the relationship between asset value (L) and consumption (Y), while controlling for potential confounders, including income and other demographic variables (V). Previous research by Bostic, Gabriel and Painter (2009) leverages fused data from the U.S. Bureau of Labor Statistics' CEX which contains detailed U.S. household expenditure information Y, and the Federal Reserve Board's SCF which provides detailed information on household assets and liabilities L, housing, and other demographic characteristics. For this application, the model of substantive interest is $E(Y|V, L) = (V^T, L)\beta$. We perform the proposed IPW and DR estimations for β based on the household expenditure and net worth data obtained from the CEX's 1997 fourth-quarter survey and the 1998 SCF, respectively, along with demographic information recorded in both surveys. The variables considered in this analysis are presented in Table 2.

Table 2. U.S. household (HH) variables used in the analysis.

	Variable	Description
R		Data source indicator for CEX ($R = 1$) or SCF ($R = 0$)
Y	log(expd)	Log of total HH expenditures in fourth quarter of 1997
L	log(netw)	Log of HH total net worth in 1997
V	sex	Sex of HH head (male=0, female=1)
	age	Age of HH head
	single	Marital status of HH head (married=0, single=1)
	edu1	HH head with high school diploma or GED (no=0, yes=1)
	edu2	HH head with some college or Associate degree (no=0, yes=1)
	edu3	HH head with Bachelors degree or higher (no=0, yes=1)
	white	White HH head (no=0, yes=1)
	black	Black/African American HH head (no=0, yes=1)
	log(income)	Log of total HH income before taxes in 1997

Although the data source process is largely administrative, the 1998 SCF oversamples relatively wealthy families based on an index created by grossing up capital income flows observed in the tax data (Kennickell (1998)). For the IPW estimation, the data source model $\pi(V)$ is specified as a logistic regression with the main effects for the binary variables, and up to quadratic terms for age and log(income). In particular, total household income before taxes in 1997 is included in V, which may serve as a good proxy for the wealth index in the SCF's sampling design. For the DR estimation, we additionally specify $E[L|V]$ as a linear model involving the main effects for the binary variables, and up to quadratic terms for age and log(income) in V. We solve the empirical versions of (4.3–4.5) for the IPW, DR, and imputation-based estimations, respectively, with $g(V)$ specified as a vector that includes the main effects of the variables in V and the variable $\log(\text{income})^2$. Based on example 2, the parameter of interest β is identified if $\log(\text{income})^2$ is correlated with household net worth, which is only recorded in the SCF. We restrict the sample to household heads between 25 and 65 years of age to mitigate heterogenous consumption effects during college-age years and retirement. Furthermore, we truncate the SCF sample at the 90^{th} percentiles of observed total household income and net worth, owing to the oversampling of wealthy households in the SCF (Bostic, Gabriel and Painter (2009)). The final data set consists of $n = 5{,}919$ households (3,388 from the CEX and 2,531 from the SCF). Owing to missing values in the original survey data, the publicly available microdata from both the CEX and the SCF consist of five imputed replicates. An estimation is performed for each replicate, and the pooled results using Rubin's rule (Rubin (2004)) are presented in Table 3.

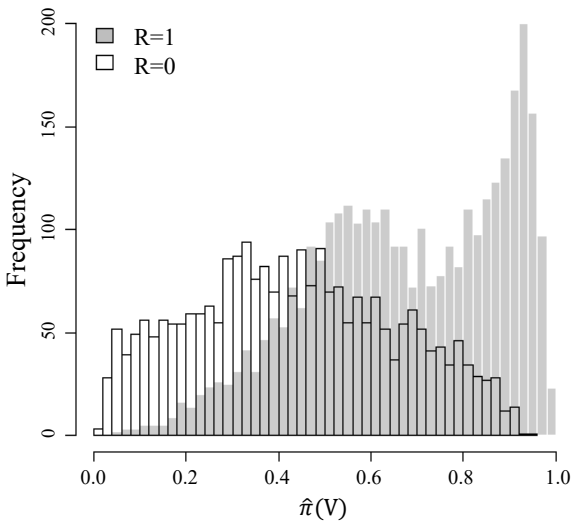

Figure 2. Histogram of fitted data source propensity scores.

The DR and imputation-based standard errors are smaller than those from the IPW, supporting our theoretical and simulation results. The IPW results suggest that, in general, households with married heads have greater total expenditure, holding the remaining variables at fixed values. Higher levels of education for the household head are associated with progressively greater total expenditure. Finally, after controlling for income and other demographic variables, the results from the IPW suggest there is a negative association between household net worth and total expenditure, although this is not statistically significant at the 0.05 level. To assess any practical violations of the positivity assumption A2, we plot histograms of the fitted data source propensity scores $\hat{\pi}(V) = \pi(V; \hat{\eta})$ for the $R = 1$ and $R = 0$ groups, separately, in Figure 2. While very few fitted scores are near zero in the $R = 1$ sample, about 6% of the scores in the $R = 0$ sample are greater than 0.8, corresponding to small values of $1 - \hat{\pi}(V)$, with the largest contributing weight equal to 23.3. This might explain the fairly large standard errors of the IPW estimates compared with those of the DR and imputation-based estimates.

In general, the DR and imputation-based estimates agree with each other. The statistically significant relationships include an inverse association between age and total expenditure, as well as a positive association between household net worth and total expenditure. Note that both associations agree qualitatively

Table 3. Estimates of the conditional mean parameters β for log(total household expenditure). Pooled standard errors are given in parentheses, and asterisks denote significance at the 0.05 level.

Variable	IPW		IMP		DR	
sex	2.338*	(0.284)	0.048	(0.067)	0.030	(0.058)
age	0.399	(0.247)	-0.264*	(0.054)	-0.160*	(0.042)
single	-4.109*	(0.367)	0.048	(0.055)	0.023	(0.042)
edu1	0.491	(0.254)	0.016	(0.081)	0.083	(0.079)
edu2	0.886*	(0.358)	0.038	(0.094)	0.081	(0.098)
edu3	1.373*	(0.460)	-0.001	(0.113)	0.035	(0.123)
white	0.580*	(0.229)	-0.094	(0.086)	-0.052	(0.083)
black	0.237	(0.269)	0.134	(0.096)	0.002	(0.104)
log(income)	0.537	(0.432)	-0.095	(0.096)	0.085	(0.066)
log(netw)	-0.620	(0.417)	0.499*	(0.089)	0.346*	(0.066)

with the findings of Bostic, Gabriel and Painter (2009). The similarity between the DR and imputation estimates suggests that the conditional model $E[L|V]$ may be specified nearly correctly (Robins and Rotnitzky (2001)). Here, Tchetgen Tchetgen and Robins (2010) describe a formal specification test to detect which of the two baseline models $\pi(V)$ and $t(L, V)$ is correct under the union model $\mathcal{M}_{\pi \cup t}$. Based on this and the DR property, it may be that the data source model in this illustrative analysis for the IPW is misspecified. As such, the results from the DR estimation may be more meaningful, given its additional protection against misspecifications of the data source model.

8. Discussion

Traditional regression models break down when two data sources are fused together such that no subject has complete data. Investigators often consider parametric models for a given outcome regressed on a number of independent variables. However, current parametric models do not adequately deal with the missing data structure that arises from data fusion. In this study, we have developed a general class of semiparametric parallel IPW estimating functions, the resulting estimators of which are consistent if the outcome regression and the data source process are correctly specified. This general class of estimating functions includes a large set of DR estimating functions, which require an additional model for the missing covariates. An estimator in this class is DR in that it is consistent and asymptotically normal if we correctly specify a model for either the data source process or the distribution of the unobserved covariates, but not

necessarily both.

There are several areas for additional research on this topic, notably, the open question of how to generalize this method to other settings. A clear extension is the setting of fusing more than two data sets together. Consider m data sources, with V observed for all, and each of $(L_1, L_2, \ldots, L_{m-1}, Y)$ observed in only one source, with respective indicators of observation $(R_1, R_2, \ldots, R_{m-1}, R_m)$ and inclusion probabilities $(\pi_1, \pi_2, \ldots, \pi_{m-1}, \pi_m)$. Therefore, the observed data are $O = (V, R_1L_1, R_2L_2, \ldots, R_{m-1}L_{m-1}, R_mY)$. Then, for example, it is easy to extend (4.3) for linear models to

$$U_g^m(\beta) = g(V)\left\{\frac{R_m}{\pi_m}Y - \left[\beta_0 + \frac{R_1}{\pi_1}\beta_1^T L_1 + \frac{R_2}{\pi_2}\beta_2^T L_2 + \cdots + \frac{R_{m-1}}{\pi_{m-1}}\beta_{m-1}^T L_{m-1} + \beta_m^T V\right]\right\},$$

provided V is rich enough for identification.

Supplementary Material

The online Supplementary Material contains proofs of the results, as well as additional simulation results.

Acknowledgments

BaoLuo Sun's work was supported by the National University of Singapore start-up grant R-155-000-203-133. Eric Tchetgen Tchetgen and James Robins were supported by NIH grant R01AI127271.

References

Angrist, J. D. and Krueger, A. B. (1992). The effect of age at school entry on educational attainment: An application of instrumental variables with moments from two samples. *Journal of the American Statistical Association* **87**, 328–336.

Bickel, P. J., Klaassen, C. A., Bickel, P. J., Ritov, Y., Klaassen, J., Wellner, J. A. and Ritov, Y. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.

Bostic, R., Gabriel, S. and Painter, G. (2009). Housing wealth, financial wealth, and consumption: New evidence from micro data. *Regional Science and Urban Economics* **39**, 79–89.

Burgess, S., Small, D. S. and Thompson, S. G. (2017). A review of instrumental variable estimators for mendelian randomization. *Statistical Methods in Medical Research* **26**, 2333–2355.

Chatterjee, N., Chen, Y.-H., Maas, P. and Carroll, R. J. (2016). Constrained maximum likelihood estimation for model calibration using summary-level information from external big data sources. *Journal of the American Statistical Association* **111**, 107–117.

Chen, Y.-H. and Chen, H. (2000). A unified approach to regression analysis under double-sampling designs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **62**, 449–460.

Conti, P. L., Marella, D. and Scanu, M. (2016). Statistical matching analysis for complex survey data with applications. *Journal of the American Statistical Association* **111**, 1715–1725.

Davey Smith, G. and Ebrahim, S. (2003). 'Mendelian randomization': Can genetic epidemiology contribute to understanding environmental determinants of disease? *International Journal of Epidemiology* **32**, 1–22.

Debray, T. P., Damen, J. A., Snell, K. I., Ensor, J., Hooft, L., Reitsma, J. B., Riley, R. D. and Moons, K. G. (2017). A guide to systematic review and meta-analysis of prediction model performance. *BMJ* **356**, i6460.

Debray, T. P., Moons, K. G., Ahmed, I., Koffijberg, H. and Riley, R. D. (2013). A framework for developing, implementing, and evaluating clinical prediction models in an individual participant data meta-analysis. *Statistics in Medicine* **32**, 3158–3180.

D'Orazio, M., Di Zio, M. and Scanu, M. (2006). *Statistical Matching: Theory and Practice*. John Wiley & Sons.

D'Orazio, M., Di Zio, M. and Scanu, M. (2010). Old and new approaches in statistical matching when samples are drawn with complex survey designs. *Proceedings of the 45th "Riunione Scientifica della Societa'Italiana di Statistica", Padova*, 16–18.

Graham, B. S., Pinto, C. C. d. X. and Egel, D. (2016). Efficient estimation of data combination models by the method of auxiliary-to-study tilting (AST). *Journal of Business & Economic Statistics* **34**, 288–301.

Hasminskii, R. and Ibragimov, I. (1983). On asymptotic efficiency in the presence of an infinited-imensional nuisance parameter. In *Probability Theory and Mathematical Statistics* (Edited by S. Watanabe and Y. V. Prokhorov), 195–229. Springer-Verlag.

Inoue, A. and Solon, G. (2010). Two-sample instrumental variables estimators. *The Review of Economics and Statistics* **92**, 557–561.

Kennickell, A. B. (1998). List sample design for the 1998 survey of consumer finances. *Federal Reserve Board Mimeo*.

Klevmarken, A. (1982). *Missing Variables and Two-Stage Least-Squares Estimation from More than One Data Set*. Working Paper Series 62. Research Institute of Industrial Economics.

Lawlor, D. A., Harbord, R. M., Sterne, J. A., Timpson, N. and Davey Smith, G. (2008). Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology. *Statistics in Medicine* **27**, 1133–1163.

Little, R. J. and Rubin, D. B. (2014). *Statistical Analysis with Missing Data*. 2nd Edition. John Wiley & Sons, New Jersey.

Little, R. J. and Schluchter, M. D. (1985). Maximum likelihood estimation for mixed continuous and categorical data with missing values. *Biometrika* **72**, 497–512.

Newey, W. K. (1993). Efficient estimation of models with conditional moment restrictions. In *Handbook of Statistics 11: Econometrics*, 419–454. Elsevier.

Newey, W. K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics* **4**, 2111–2245.

Olkin, I. and Tate, R. F. (1961). Multivariate correlation models with mixed discrete and continuous variables. *The Annals of Mathematical Statistics* **32**, 448–465.

Pacini, D. (2017). Two-sample least squares projection. *Econometric Reviews* **38**, 1–29.

Rässler, S. (2012). *Statistical Matching: A Frequentist Theory, Practical Applications, and Al-*

ternative Bayesian Approaches. Volume 168. Springer Science & Business Media, Berlin.

Renssen, R. H. (1998). Use of statistical matching techniques in calibration estimation. *Survey Methodology* **24**, 171–184.

Riley, R. D., Lambert, P. C. and Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *BMJ* **340**, c221.

Robins, J. M., Hsieh, F. and Newey, W. (1995). Semiparametric efficient estimation of a conditional density with missing or mismeasured covariates. *Journal of the Royal Statistical Society: Series B (Statistical Methodological)* **57**, 409–424.

Robins, J. M. and Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer". *Statistica Sinica* **11**, 920–936.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American statistical Association* **89**, 846–866.

Rubin, D. B. (1986). Statistical matching using file concatenation with adjusted weights and multiple imputations. *Journal of Business & Economic Statistics* **4**, 87–94.

Rubin, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. John Wiley & Sons, New Jersey.

Schafer, J. L. (1997). *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC, Boca Raton.

Stürmer, T., Schneeweiss, S., Avorn, J. and Glynn, R. J. (2005). Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration. *American journal of Epidemiology* **162**, 279–289.

Tchetgen Tchetgen, E. J. and Robins, J. (2010). The semiparametric case-only estimator. *Biometrics* **66**, 1138–1144.

Tchetgen Tchetgen, E. J., Robins, J. M. and Rotnitzky, A. (2009). On doubly robust estimation in a semiparametric odds ratio model. *Biometrika* **97**, 171–180.

Varadhan, R., Gilbert, P. et al. (2009). Bb: An r package for solving a large system of nonlinear equations and for optimizing a high-dimensional nonlinear objective function. *Journal of Statistical Software* **32**, 1–26.

Wu, C. (2004). Combining information from multiple surveys through the empirical likelihood method. *Canadian Journal of Statistics* **32**, 15–26.

Katherine Evans

38 Joe Shuster Way #1629, Toronto, ON M6K 0A5, Canada.

E-mail: causalkathy@gmail.com

BaoLuo Sun

Department of Statistics and Applied Probability, National University of Singapore, Singapore 119077.

E-mail: stasb@nus.edu.sg

James Robins

Departments of Epidemiology and Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA 02115, USA.

E-mail: robins@hsph.harvard.edu

Eric J. Tchetgen Tchetgen

Department of Statistics, The Wharton School of the University of Pennsylvania, Philadelphia, PA 19104, USA.

E-mail: ett@wharton.upenn.edu

(Received August 2018; accepted October 2019)