Introduction to Double Robust Methods for Incomplete Data

Shaun R. Seaman[1] **and Stijn Vansteelandt**[2,3]

[1] Medical Research Council Biostatistics Unit, University of Cambridge, Cambridge, UK.
shaun.seaman@mrc-bsu.cam.ac.uk.

[2] Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Gent, Belgium.

[3] Department of Epidemiology and Population Health, London School of Hygiene and Tropical Medicine, London, UK

Running title: Double Robust Methods

Abstract

Most methods for handling incomplete data can be broadly classified as inverse probability weighting (IPW) strategies or imputation strategies. The former model the occurrence of incomplete data; the latter, the distribution of the missing variables given observed variables in each missingness pattern. Imputation strategies are typically more efficient, but they can involve extrapolation, which is difficult to diagnose and can lead to large bias. Double robust (DR) methods combine the two approaches. They are typically more efficient than IPW and more robust to model misspecification than imputation. We give a formal introduction to DR estimation of the mean of a partially observed variable, before moving to more general incomplete-data scenarios. We review strategies to improve the performance of DR estimators under model misspecification, reveal connections between DR estimators for incomplete data and 'design-consistent' estimators used in sample surveys, and explain the value of double robustness when using flexible data-adaptive methods for IPW or imputation.

AMS 1991 subject classifications. primary – 62 Statistics; secondary – 62A01 Foundations and philosophical topics

Key words and phrases: augmented inverse probability weighting, calibration estimators, data-adaptive methods, doubly robust, empirical likelihood, imputation, inverse probability weighting, missing data, semiparametric methods

Grants and funding: SRS is funded by MRC Grant MC_U105260558.

1 Introduction

Statistical analysis of data is often complicated by the data being incomplete, e.g., due to individuals in a survey not answering a question, patients missing a clinic visit, or data simply being lost. The individuals on whom complete data are obtained (the 'complete cases') often constitute a non-representative subset of the sample. This makes an analysis that uses only this subset potentially biased, as well as being potentially inefficient because it discards the data available on the individuals with incomplete data (the 'incomplete cases'). More sophisticated approaches for analysing incomplete data are designed to reduce bias and/or increase efficiency. They can broadly be classified into imputation strategies and inverse probability weighting (IPW) approaches [17].

Imputation approaches involve specifying a model (the 'imputation model') for the partially observed variables given any fully observed variables. Missing values are then 'predicted' based on this model. Multiple imputation is the most popular such approach, and has close connections to maximum likelihood (ML) methods for incomplete data. The latter methods involve implicit imputation of the missing data. A drawback of imputation approaches is that they can involve much modelling of the incomplete data and there may be large bias when the imputation model is misspecified. This potential for bias is especially large when the distribution of observed data in individuals with a given missingness pattern is very different from that in the overall population. In that case, imputation involves extrapolation under the imputation model, so that even minor model misspecification over the range of the observed data may induce large bias. Additional concerns may arise from the difficulty of specifying the imputation model in a way that obeys the structure imposed by the model that will be used to analyse the imputed data (i.e. such that it is 'congenial with the analysis model' [19]).

IPW methods avoid these issues of extrapolation and uncongeniality by not using an imputation model. They instead rely on a missingness model, i.e. a model for the probability that an individual is a complete case given a set of predictors of missingness. The analysis model is then fitted to just the complete cases, inversely weighting each by its estimated probability of being complete given its missingness predictors. A drawback of IPW is that it can be very inefficient, because, like the complete-case analysis, it ignores potentially useful data on the incomplete cases. It can also be subject to large finite-sample bias. Recognition of these problems led to research on augmented IPW (AIPW) estimators. These, like imputation estimators, involve a model for the conditional distribution of the partially observed variables given fully observed variables. AIPW estimators are more efficient than (unaugmented) IPW estimators when this imputation model is correctly specified. Indeed, among all estimators that, like IPW estimators, are consistent whenever the missingness model is correctly specified, AIPW estimators with correctly specified imputation models are the most efficient.

In 1999, Scharfstein et al. [32] noted that an AIPW estimator previously developed by Robins et al. [28] for estimating the mean of a partially observed variable had the property of being consistent not only when the missingness model was correctly

specified, but also when an imputation model for the conditional distribution of this variable was correctly specified and the missingness model was misspecified. This property became known as 'double robustness' [26]. At about the same time, it was recognised [24] that Robins et al.'s estimator was closely related to a 'generalised regression' estimator first developed in the 1970's for improving the efficiency of an IPW estimator of a finite-sample population mean when sampling probabilities are known [6]. Since the double robust (DR) property was discovered, many estimators possessing this property have been developed. However, the DR property has also been criticised. Simulation studies which showed that minor misspecifications of both the imputation and missingness models can sometimes induce large bias and variance in the DR estimator led to a questioning of the practical usefulness of double robustness [12]. Such scepticism has been reinforced by the availability of imputation and IPW approaches based on flexible 'data-adaptive' methods for fitting the imputation and missingness model, respectively, which reduce the risk of model misspecification [16].

This article is an introduction to DR methodology for incomplete data. As in much of the literature on missing data (and DR estimators in particular), we shall assume that data are missing at random (MAR). Data are said to be MAR if the conditional probability that a particular missingness pattern occurs given the data does not depend on the missing values in that pattern [35]. In Section 2, we consider the problem of estimating the mean of a partially observed variable using fully observed auxiliary variables, using this example to contrast imputation with IPW and to present a DR AIPW estimator. In Section 3 we introduce more formality and give a review of the general semiparametric theory underlying DR estimation. This enables us to describe DR estimators for more general missing data problems. So-called 'standard' DR estimators use ML to estimate the parameters of the missingness and imputation models. In Section 4 we review more recently developed methods which seek to improve the performance of DR estimators (relative to standard DR estimators) under model misspecification by using alternative estimators of these parameters. In Section 5, we consider the use of data-adaptive methods (e.g. smoothing methods or regularisation methods) for the imputation or missingness model. We argue that there are advantages to using these methods in DR estimators (rather than in imputation or IPW estimators). Section 6 discusses the wide variety of statistical models for which DR estimators have been proposed, DR methods for non-monotone missing and missing not at random (MNAR) data (most work has been on monotone missing, MAR data), and some possible directions of future research. Implementation of DR estimators in standard statistical packages is described in the supplemental article [36].

2 IPW, RI and AIPW for a missing outcome

For pedagogic purposes, we first consider the problem of estimating the expectation $\beta = E(Y)$ of a partially observed random variable Y from a sample of size n when auxiliary variables $\boldsymbol{W}$ are observed on the whole sample. This has been the focus of much of the work on DR estimation. In Section 3, we discuss DR estimation for

more general missing data problems.

Let Y_i and $\boldsymbol{W}_i$ denote Y and $\boldsymbol{W}$ for the ith individual in the sample, and R_i be an indicator that Y_i is observed ($R_i = 1$ if Y_i is observed; $R_i = 0$ if missing). Individuals with $R_i = 1$ are 'complete cases'; those with $R_i = 0$ are 'incomplete cases'. Assume $(\boldsymbol{W}_1, Y_1, R_1), \ldots, (\boldsymbol{W}_n, Y_n, R_n)$ are independent and identically distributed. Henceforth, we omit subscripts i unless needed.

The full-data (or 'complete-data') estimator, $n^{-1}\sum_{i=1}^n Y_i$, for β is infeasible when Y can be missing. The complete-case estimator, $\sum_{i=1}^n R_i Y_i / \sum_{i=1}^n R_i$, is typically inconsistent unless R is independent of Y. The IPW, regression imputation (RI) and DR estimators described below are valid under the weaker assumption that R is independent of Y given $\boldsymbol{W}$, i.e. that $(\boldsymbol{W}_1, Y_1, R_1), \ldots, (\boldsymbol{W}_n, Y_n, R_n)$ are MAR.

In IPW, each complete case is weighted by $\pi(\boldsymbol{W})^{-1}$, where $\pi(\boldsymbol{W}) = P(R = 1 \mid \boldsymbol{W})$ is the probability that an individual with this value of $\boldsymbol{W}$ would be a complete case. Each complete case then represents $\pi(\boldsymbol{W})^{-1}$ individuals in the population, all with the same $\boldsymbol{W}$ value. One of these would have observed Y if sampled; the others would have missing Y. The weighted sample of complete cases therefore has (over repeated samples) the same distribution of $\boldsymbol{W}$ as the population, and by MAR, also the same distribution of Y as the population. This motivates the IPW estimators of β: $n^{-1}\sum_{i=1}^n R_i \pi(\boldsymbol{W}_i)^{-1} Y_i$ [11] and $\sum_{i=1}^n R_i \pi(\boldsymbol{W}_i)^{-1} Y_i \,/\, \sum_{i=1}^n R_i \pi(\boldsymbol{W}_i)^{-1}$. Since $\pi(\boldsymbol{W})$ is unknown unless data are missing by design, a model $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$, called the 'missingness model', is specified for it and an estimator $\hat{\boldsymbol{\alpha}}$ of $\boldsymbol{\alpha}$ calculated from data $(R_1, \boldsymbol{W}_1, \ldots, R_n, \boldsymbol{W}_n)$. E.g., one could use $\pi(\boldsymbol{W}; \boldsymbol{\alpha}) = \text{expit}(\boldsymbol{\alpha}^\top \boldsymbol{W})$, with $\boldsymbol{\alpha}$ estimated by ML. Let $\hat{\beta}_{\text{IPW}} = \hat{\beta}_{\text{IPW}}(\hat{\boldsymbol{\alpha}}) = n^{-1}\sum_{i=1}^n R_i \pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})^{-1} Y_i$ and $\hat{\beta}_{\text{IPW,B}} = \hat{\beta}_{\text{IPW,B}}(\hat{\boldsymbol{\alpha}}) = \sum_{i=1}^n R_i \pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})^{-1} Y_i \,/\, \sum_{i=1}^n R_i \pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})^{-1}$ denote the IPW estimators with estimated weights ('B' in the subscript 'IPW,B' stands for 'sample bounded': $\hat{\beta}_{\text{IPW,B}}$ is guaranteed to lie within the range of the observed Y values). If $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ is correctly specified and $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$, then $\hat{\beta}_{\text{IPW}}$ and $\hat{\beta}_{\text{IPW,B}}$ are consistent estimators of β, provided that there exists a $\delta > 0$ such that $P\{\pi(\boldsymbol{W}) \geq \delta\} = 1$ (this 'positivity' assumption rules out scenarios where individuals with certain values of $\boldsymbol{W}$ cannot be complete cases) and $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ is a sufficiently smooth function of $\boldsymbol{\alpha}$.

In RI, a parametric model $m(\boldsymbol{W}; \boldsymbol{\gamma})$ for $E(Y \mid \boldsymbol{W})$ is specified. This is called the 'outcome model'. Let $\hat{\boldsymbol{\gamma}}$ be an estimator of $\boldsymbol{\gamma}$ (e.g. the ML estimator calculated using the complete cases). Parameter β is then estimated by $\hat{\beta}_{\text{RI}} = \hat{\beta}_{\text{RI}}(\hat{\boldsymbol{\gamma}}) = n^{-1}\sum_{i=1}^n m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}})$. If $m(\boldsymbol{W}; \boldsymbol{\gamma})$ is correctly specified and $\hat{\boldsymbol{\gamma}}$ is consistent, then $\hat{\beta}_{\text{RI}}$ is consistent. Moreover, if $\hat{\boldsymbol{\gamma}}$ is efficient, then so is $\hat{\beta}_{\text{RI}}$. Note that if $m(\boldsymbol{W}; \boldsymbol{\gamma})$ is a canonical generalised linear model that includes an intercept and $\hat{\boldsymbol{\gamma}}$ is the ML estimator, then $\sum_{i=1}^n R_i m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}}) = \sum_{i=1}^n R_i Y_i$ and so $\hat{\beta}_{\text{RI}}$ can be written as $n^{-1}\sum_{i=1}^n \{R_i Y_i + (1 - R_i) m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}})\}$. The RI estimator then equals the mean of Y after replacing missing values by imputed values $m(\boldsymbol{W}; \hat{\boldsymbol{\gamma}})$.

The efficiency of $\hat{\beta}_{\text{RI}}$ comes at the cost of assuming that model $m(\boldsymbol{W}; \boldsymbol{\gamma})$ is correctly specified. When there is little overlap between the distributions of $\boldsymbol{W}$ in complete and incomplete cases, the RI estimator works by extrapolating the relation between $\boldsymbol{W}$ and Y estimated from complete cases to regions of the $\boldsymbol{W}$ space

where incomplete cases but few (if any) complete cases lie. This extrapolation is potentially risky, because even models that fit the data on complete cases perfectly may give a poor approximation of $E(Y \mid \boldsymbol{W})$ in these regions [45]. This is illustrated by the following example.

Example 1: Let $P(W = 0) = P(W = 1) = P(W = 2) = 1/3$, logit $P(R = 1 \mid W) = 4 - 4W$, and either a) $Y \mid W \sim N(W, \sigma^2)$ or b) $Y \mid W \sim N(I(W \geq 1), \sigma^2)$, where $I(.)$ denotes the indicator function.

In case a), the RI estimator $\hat{\beta}_{\mathrm{RI}}$ based on linear regression model $m(W; \boldsymbol{\gamma}) = \gamma_1 + \gamma_2 W$ with ML estimator $\hat{\boldsymbol{\gamma}}$ is consistent; in case b), it is inconsistent ($\hat{\beta}_{\mathrm{RI}} \xrightarrow{p} 0.94$, whereas $E(Y) = 0.67$). This is a concern because, unless the sample size were very large, it would be difficult to decide on the basis of the observed data whether this linear regression model is correctly specified, as there would generally be few complete cases with $W = 2$. The IPW estimators $\hat{\beta}_{\mathrm{IPW}}$ and $\hat{\beta}_{\mathrm{IPW,B}}$ based on model logit $\pi(W; \boldsymbol{\alpha}) = \alpha_1 + \alpha_2 W$ with ML estimator $\hat{\boldsymbol{\alpha}}$ are consistent in both cases. While these also rely on a model (for $\pi(W)$) which may be misspecified, its goodness of fit is arguably easier to assess because this requires data only on R and W, which are fully observed, and there is no need for extrapolation outside the observed data range. The variances of both IPW estimators are larger than that of $\hat{\beta}_{\mathrm{RI}}$, because of the large weights attributed to the small proportion of complete cases with $W = 2$. For example, using simulation we estimated that, when $n = 100000$ and $\sigma^2 = 1$, the variances ($\times 10^5$) of $\hat{\beta}_{\mathrm{IPW}}$, $\hat{\beta}_{\mathrm{IPW,B}}$ and $\hat{\beta}_{\mathrm{RI}}$ are 51, 28 and 6.1, respectively. The relatively large variances of the IPW estimators can be seen as reflecting genuine uncertainty about β, in contrast to the variance of $\hat{\beta}_{\mathrm{RI}}$, which does not reflect model and extrapolation uncertainty about $E(Y \mid W = 2)$. This uncertainty could be accommodated by using more flexible outcome models, but this would drastically increase the variance of $\hat{\beta}_{\mathrm{RI}}$; indeed ultimately, if $\boldsymbol{W}$ is categorical and the missingness and outcome models are saturated, the IPW and RI estimators (and their variance estimators) are equivalent [21]. More generally, when the outcome model is not saturated and there is very little overlap between the distributions of $\boldsymbol{W}$ in complete and incomplete cases, it may be difficult to ensure that a flexible outcome model is sufficiently flexible outside the region of the $\boldsymbol{W}$ space where the complete cases lie.

The inefficiency of the IPW estimators is a serious drawback, but it can be reduced by making more use of the $\boldsymbol{W}$ data on the incomplete cases. In particular, the augmented IPW (AIPW) estimator

$$\begin{aligned} \hat{\beta}_{\mathrm{DR}} = \hat{\beta}_{\mathrm{DR}}(\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}}) &= \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})} Y_i + \frac{1}{n}\sum_{i=1}^{n} \left\{ 1 - \frac{R_i}{\pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})} \right\} m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}}) \quad (1) \\ &= \frac{1}{n}\sum_{i=1}^{n} m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}}) + \frac{1}{n}\sum_{i=1}^{n} \frac{R_i}{\pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})} \{ Y_i - m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}}) \}, \quad (2) \end{aligned}$$

of β, where $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ are estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$, is efficient relative to all estimators that rely solely on correct specification of the missingness model, provided that the outcome model is also correctly specified (see Section 3). The first term on the right-hand side of equation (1) is just $\hat{\beta}_{\mathrm{IPW}}$ and the second term is called the

augmentation term. This uses data on $\boldsymbol{W}$ on the incomplete cases to improve its efficiency. In the alternative (equivalent) expression for $\hat{\beta}_{\mathrm{DR}}$, equation (2), the first term on the right equals $\hat{\beta}_{\mathrm{RI}}$ and the second term can be viewed as a 'correction' factor: it uses IPW to estimate how much $\hat{\beta}_{\mathrm{RI}}$ overestimates (or underestimates) $E(Y)$ and then subtracts this. Estimator $\hat{\beta}_{\mathrm{DR}}$ is consistent and asymptotically normal distributed when either i) $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ is correctly specified and $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$, or ii) $m(\boldsymbol{W};\boldsymbol{\gamma})$ is correctly specified and $\hat{\boldsymbol{\gamma}}$ is a consistent estimator of $\boldsymbol{\gamma}$, a property known as 'double robustness'. A formal proof of this is given in the supplemental article [36], but essentially it is because: i) when $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ is correctly specified, the augmentation term converges to zero (because then $\hat{\boldsymbol{\alpha}}$ converges to the true value of $\boldsymbol{\alpha}$ and $E\{R/\pi(\boldsymbol{W};\boldsymbol{\alpha}) \mid \boldsymbol{W}\} = 1$ at this true value); and ii) when $m(\boldsymbol{W};\boldsymbol{\gamma})$ is correctly specified, the correction term converges to zero (because then $\hat{\boldsymbol{\gamma}}$ converges to the true value of $\boldsymbol{\gamma}$ and $E(Y \mid \boldsymbol{W}, R) = m(\boldsymbol{W};\boldsymbol{\gamma})$ at this true value).

The DR estimator $\hat{\beta}_{\mathrm{DR}}$ can be much more efficient than $\hat{\beta}_{\mathrm{IPW}}$ when both $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ and $m(\boldsymbol{W};\boldsymbol{\gamma})$ are correctly specified and $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ are consistent, especially when $\mathrm{Var}(Y \mid \boldsymbol{W})$ is small relative to $\mathrm{Var}(Y)$, i.e. when $\boldsymbol{W}$ is a strong predictor of Y [25]. This is because the correction term in equation (2) is then small relative to the first term, and so $\hat{\beta}_{\mathrm{DR}} \approx \hat{\beta}_{\mathrm{RI}}$. Indeed, when both $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ and $m(\boldsymbol{W};\boldsymbol{\gamma})$ are correctly specified, it can be shown (see [36]) that $n\mathrm{Var}(\hat{\beta}_{\mathrm{DR}}) \to \mathrm{Var}(Y) + E[\{1 - \pi(\boldsymbol{W})\}\pi(\boldsymbol{W})^{-1}\mathrm{Var}(Y \mid \boldsymbol{W})]$ as $n \to \infty$, which equals n times the variance of the (infeasible) full-data estimator $n^{-1}\sum_{i=1}^{n} Y_i$ when $\mathrm{Var}(Y \mid \boldsymbol{W}) = 0$. To illustrate this, we return to case a) of Example 1, where $Y \sim N(W, \sigma^2)$.

Example 1 continued: When $n = 100000$ and $\sigma^2 = 1$, the variances ($\times 10^5$) of $\hat{\beta}_{\mathrm{IPW}}$, $\hat{\beta}_{\mathrm{IPW,B}}$, $\hat{\beta}_{\mathrm{RI}}$, $\hat{\beta}_{\mathrm{DR}}$ and the full-data estimator are, respectively, 51, 28, 6.1, 20 and 1.7: the DR estimator is more efficient than the IPW estimators, though not as efficient as the RI estimator. When $n = 100000$ and $\sigma^2 = 0.01$, the variances ($\times 10^5$) are 31, 8.8, 0.72, 0.86 and 0.67: the DR, RI and full-data estimators are close to being equally efficient.

Example 2: Wirth et al. [48] used data from the National Family Health Survey 3 to estimate the percentage of sexually-active Indian men who had paid for sex in the past year. Of the 49700 men surveyed, 3% refused to answer the question about paying for sex; these were more likely to be young, unmarried, unemployed and to believe that a husband has the right to have sex with another woman. Among men who answered the question, the percentage reporting paying for sex was 0.9%. Wirth et al. built missingness and outcome models using 24 variables thought to be predictive of paying for sex and/or refusing to answer (e.g. age, education, marital status). The resulting DR estimate of the percentage paying for sex was 1.1%. Among unmarried men, 18% refused to answer the question, 6.9% of those who answered reporting paying for sex, and the DR estimate was 12.3%.

3 Semiparametric theory of DR estimators

DR estimators do not require correct specification of the entire data-generating distribution, and are semiparametric in this sense. Semiparametric efficiency was, and continues to be, very important in DR theory: the development of DR estimators by Robins and others, and of earlier related survey sampling estimators, was motivated by the goal of improving the efficiency of IPW estimators; only later was the DR property of these estimators recognised. In this section, we give an introduction to the semiparametric theory that underlies DR estimators and describe estimators for more general missing data problems than that discussed in Section 2. A more detailed account of semiparametric theory for DR estimators can be found in, e.g. [40] or [41].

3.1 Semiparametric models and m-estimators

Assume that random variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_n$ are independently and identically distributed with density $f(\boldsymbol{z})$. A semiparametric model is a model $\mathcal{M}$ for the density $f(\boldsymbol{z})$ of $\boldsymbol{Z}$ that parameterises one or more aspects of $f(\boldsymbol{z})$ in terms of an unknown finite-dimensional parameter $\boldsymbol{\beta}$ but leaves other aspects unrestricted.

An example is the model for $\boldsymbol{Z} = (\boldsymbol{Y}^\top, \boldsymbol{X}^\top, \boldsymbol{W}^\top)^\top$ that assumes

$$E(\boldsymbol{Y} \mid \boldsymbol{X}) = \boldsymbol{\mu}(\boldsymbol{X}; \boldsymbol{\beta}), \tag{3}$$

where $\boldsymbol{\mu}(\boldsymbol{X}; \boldsymbol{\beta})$ is a known vector function of $\boldsymbol{X}$ and $\boldsymbol{\beta}$, but which otherwise leaves $f(\boldsymbol{z})$ unrestricted. This is known as a restricted moment model and is usually fitted using generalised estimating equations [15]. A specific example of this model is the semiparametric regression model $E(Y \mid \boldsymbol{X}) = g(\boldsymbol{\beta}^\top \boldsymbol{X})$ for scalar outcome Y, covariates $\boldsymbol{X}$ and known link function $g(.)$. Other examples of semiparametric models are the Cox proportional hazards model, which restricts hazard ratios but otherwise leaves $f(\boldsymbol{z})$ unrestricted, and the nonparametric model, which places no restriction on $f(\boldsymbol{z})$. The parameter of interest in the nonparametric model could be, e.g., $\beta = E(Y) = \int Y f(\boldsymbol{z})\, d\boldsymbol{Z}$, where Y denotes an element of $\boldsymbol{Z}$; then the obvious estimator of β is $n^{-1} \sum_{i=1}^{n} Y_i$.

Example 3: Schnitzer et al. [33] used data from randomised trials of anti-HIV therapy. The semiparametric regression model logit $P(Y = 1 \mid X_1, X_2, X_3) = \beta_{\text{int}} + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ was used to predict occurrence of a clinical event in a patient within five years (Y) as a function of his/her baseline CD4 (X_1) and CD8 cell (X_2) counts and age (X_3) while he/she remained on assigned therapy.

Example 4: Seaman and Copas [34] used data from a different HIV trial. The binary outcome of interest Y_t ($t = 1, \ldots, T$) was whether HIV RNA was detectable in the patient at timepoint t (RNA was measured each 12 weeks for three years). Seaman and Copas estimated how the probability of detectable RNA changed over time in each of the three trial arms. They used the semiparametric regression model logit $P(Y_t = 1 \mid X_1, X_2, X_3) = \sum_{k=1}^{3} X_k(\beta_{\text{int},k} + \beta_{\text{slo},k} t)$, where binary $X_k = 1$ if the patient is in arm k and $\beta_{\text{slo},k}$ is the slope for arm k.

Example 5: Qi et al. [23] used Cox regression to model the dependence of the hazard of bone fracture on age and bone mineral density in a cohort of postmenopausal women. The semiparametric model was $h(t \mid X_1, X_2) = h_0(t)\exp(\beta_1 X_1 + \beta_2 X_2)$, where $h_0(t)$ is the baseline hazard at time t and $h(t \mid X_1, X_2)$ is the hazard given age (X_1) and mineral density (X_2).

Much of the focus of semiparametric theory has been on finding consistent estimators with the greatest asymptotic efficiency, i.e smallest asymptotic variance. This search has been restricted to estimators that are regular asymptotic linear (RAL) (see the supplemental article [36] for definition of RAL). If an estimator $\hat{\boldsymbol{\beta}}$ of parameter $\boldsymbol{\beta}$ in a semiparametric or parametric model $\mathcal{M}$ is RAL, then, for all densities $f(\boldsymbol{z})$ allowed by model $\mathcal{M}$, $\hat{\boldsymbol{\beta}}$ is consistent and asymptotically normally distributed (CAN). Therefore, in particular, $\hat{\boldsymbol{\beta}}$ converges to $\boldsymbol{\beta}$ and $n\text{Var}(\hat{\boldsymbol{\beta}})$ converges to a constant (which may depend on $f(\boldsymbol{z})$) as $n \to \infty$.

For most models $\mathcal{M}$, the task of identifying which of the RAL estimators of $\boldsymbol{\beta}$ is asymptotically the most efficient among all the RAL estimators under that model requires correct specification of restrictions on aspects of $f(\boldsymbol{z})$ beyond the restrictions already implied by model $\mathcal{M}$. E.g., $\mathcal{M}$ might impose restrictions only on conditional expectations of $\boldsymbol{Z}$, while identifying the most efficient RAL estimator under $\mathcal{M}$ might additionally require correct specification of conditional variances. When this RAL estimator is only the most asymptotically efficient when those further aspects of $f(\boldsymbol{z})$ are correctly modelled, it is called 'locally (semiparametric) efficient' under model $\mathcal{M}$; otherwise it is called 'globally efficient'. For many models $\mathcal{M}$, locally semiparametric efficient estimators are difficult to obtain. We therefore often content ourselves with finding the most asymptotically efficient among all the RAL estimators in a large subclass of RAL estimators. Such estimators are called 'locally (semiparametric) efficient' over the considered class. Local efficiency is important in DR theory, because most — if not all — DR RAL estimators are locally efficient over a large class of RAL estimators. This explains why the search for DR estimators is often helped by the search for efficient estimators, as we show in the next section.

Many RAL estimators for parametric and semiparametric models are m-estimators. We shall focus on these. An m-estimator $\hat{\boldsymbol{\beta}}$ is the solution to estimating equations of the form $\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}) = \boldsymbol{0}$ for some function $\boldsymbol{u}(\boldsymbol{Z}; \boldsymbol{\beta})$ of $\boldsymbol{Z}$ and $\boldsymbol{\beta}$ such that $E\{\boldsymbol{u}(\boldsymbol{Z}; \boldsymbol{\beta}_0)\} = \boldsymbol{0}$, where $\boldsymbol{\beta}_0$ denotes the true value of $\boldsymbol{\beta}$. Subject to regularity conditions [40], $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$ as $n \to \infty$. One example of an m-estimator is that using $u(\boldsymbol{Z}; \boldsymbol{\beta}) = Y - \beta$ to estimate $\beta = E(Y)$ in the nonparametric model. Solving $\sum_{i=1}^{n}(Y_i - \beta) = 0$ yields the estimator $\hat{\beta} = n^{-1}\sum_{i=1}^{n} Y_i$. All RAL estimators of β in this model are asymptotically equivalent to this $\hat{\beta}$ (which is therefore globally efficient over the class of all RAL estimators under this model). Another example is estimation of $\boldsymbol{\beta}$ in the restricted moment model (equation (3)). It can be shown that all RAL estimators of $\boldsymbol{\beta}$ in this model are asymptotically equivalent to an m-estimator with $\boldsymbol{u}(\boldsymbol{Z}, \boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{X})\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X}; \boldsymbol{\beta})\}$ for some conformable matrix $\boldsymbol{A}(\boldsymbol{X})$ of full rank, and conversely that all m-estimators of this form are RAL estimators of $\boldsymbol{\beta}$ in this model [40]. Over the class of all RAL estimators of $\boldsymbol{\beta}$ in this model, the locally efficient one at the true distribution of $\boldsymbol{Z}$ is that using

$\boldsymbol{A}(\boldsymbol{X}) = \boldsymbol{D}^\top(\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{X})$, where $\boldsymbol{D}(\boldsymbol{X}) = \partial\boldsymbol{\mu}(\boldsymbol{X},\boldsymbol{\beta})/\partial\boldsymbol{\beta}^\top$ evaluated at $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and $\boldsymbol{V}(\boldsymbol{X}) = \mathrm{Var}(\boldsymbol{Y} \mid \boldsymbol{X})$. A third example of an m-estimator is the ML estimator of $\boldsymbol{\beta}$ in a parametric model: here $\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta})$ is the score function.

3.2 Construction of DR estimators

Suppose $\boldsymbol{Z}$ is only partially observed. The aim is still to estimate $\boldsymbol{\beta}$ in the semiparametric model $\mathcal{M}$ for the full data $(\boldsymbol{Z}_1,\ldots,\boldsymbol{Z}_n)$, but incompleteness of the data makes use of the full-data m-estimator of Section 3.1 infeasible and we instead seek an estimator that uses only the observed data. Semiparametric theory shows how to convert a RAL m-estimator for full data into a RAL m-estimator for observed data. This is relatively straightforward when data $\boldsymbol{Z}$ are MAR and monotone missing, and we now show how to do this. Consider first the situation where there are only two missingness patterns. Here we can write $\boldsymbol{Z} = (\boldsymbol{Z}^{(1)\top}, \boldsymbol{Z}^{(2)\top})^\top$, where $\boldsymbol{Z}^{(1)}$ is observed on the whole sample and $\boldsymbol{Z}^{(2)}$ is observed on a subset of the sample. The latter could be, e.g. the outcome or a covariate in a restricted moment model. For each individual, let $R = 1$ if $\boldsymbol{Z}^{(2)}$ is observed and $R = 0$ if it is missing. Individuals with $R = 1$ are the complete cases. The observed data are $(\boldsymbol{Z}_1^{(1)}, R_1\boldsymbol{Z}_1^{(2)}, R_1, \ldots, \boldsymbol{Z}_n^{(1)}, R_n\boldsymbol{Z}_n^{(2)}, R_n)$.

The MAR assumption implies that $P(R = 1 \mid \boldsymbol{Z}) = P(R = 1 \mid \boldsymbol{Z}^{(1)})$. Let $\pi(\boldsymbol{Z}^{(1)}) = P(R = 1 \mid \boldsymbol{Z}^{(1)})$. Assume there exists a $\delta > 0$ such that $P\{\pi(\boldsymbol{Z}^{(1)}) \geq \delta\} = 1$. A parametric model $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ is specified for $\pi(\boldsymbol{Z}^{(1)})$, where $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ is a sufficiently smooth function of $\boldsymbol{\alpha}$. Denote by $\mathcal{M}_{\mathrm{miss}}$ the semiparametric model for $(\boldsymbol{Z}^{(1)}, R\boldsymbol{Z}^{(2)}, R)$ defined by model $\mathcal{M}$ for $\boldsymbol{Z}$, model $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ for R given $\boldsymbol{Z}^{(1)}$, and the MAR assumption. Suppose that the solution to the m-estimating equations $\sum_{i=1}^n \boldsymbol{u}(\boldsymbol{Z}_i;\boldsymbol{\beta}) = \boldsymbol{0}$ is a full-data RAL estimator for $\boldsymbol{\beta}$ under model $\mathcal{M}$. Then a corresponding observed-data estimator is the solution to the AIPW estimating equations

$$\sum_{i=1}^n \frac{R_i}{\pi(\boldsymbol{Z}_i^{(1)};\hat{\boldsymbol{\alpha}})}\boldsymbol{u}(\boldsymbol{Z}_i;\boldsymbol{\beta}) + \left\{1 - \frac{R_i}{\pi(\boldsymbol{Z}_i^{(1)};\hat{\boldsymbol{\alpha}})}\right\}\boldsymbol{\phi}(\boldsymbol{Z}_i^{(1)};\boldsymbol{\beta}) = \boldsymbol{0}, \tag{4}$$

where $\hat{\boldsymbol{\alpha}}$ is an estimator of $\boldsymbol{\alpha}$ based on data $(R_1, \boldsymbol{Z}_1^{(1)}, \ldots, R_n, \boldsymbol{Z}_n^{(1)})$, e.g. the ML estimator, and $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta})$ is some function of $\boldsymbol{Z}^{(1)}$ and $\boldsymbol{\beta}$. If $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ is correctly specified and $\hat{\boldsymbol{\alpha}}$ is a consistent estimator of $\boldsymbol{\alpha}$, then the solution to equation (4) is a RAL estimator for $\boldsymbol{\beta}$ under model $\mathcal{M}_{\mathrm{miss}}$. That is, it is CAN when models $\mathcal{M}$ and $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ are correctly specified and data $\boldsymbol{Z}$ are MAR. We prove this later, after introducing the DR estimator. For the restricted moment model in particular, all observed-data RAL estimators of $\boldsymbol{\beta}$ are asymptotically equivalent to an m-estimator of the form of equation (4) with $\boldsymbol{u}(\boldsymbol{Z},\boldsymbol{\beta}) = \boldsymbol{A}(\boldsymbol{X})\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X};\boldsymbol{\beta})\}$ for some conformable matrix $\boldsymbol{A}(\boldsymbol{X})$ of full rank.

If $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta})$ is chosen to be zero, equations (4) reduce to IPW estimating equations, which use only data on complete cases. Semiparametric theory shows that the optimally efficient choice of $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta})$ is $\boldsymbol{\phi}_{\mathrm{opt}}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta}) = E\{\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta}) \mid \boldsymbol{Z}^{(1)}, R = 1\}$. That is, the asymptotically most efficient RAL estimator of $\boldsymbol{\beta}$

among the class of estimators that solve equations (4) for a fixed choice of $\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta})$ is that which uses $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta}) = \boldsymbol{\phi}_{\text{opt}}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta})$. Put formally, $V(\boldsymbol{u},\boldsymbol{\phi}) - V(\boldsymbol{u},\boldsymbol{\phi}_{\text{opt}})$ is non-negative definite for any $\boldsymbol{\phi}(.)$, where $V(\boldsymbol{u},\boldsymbol{\phi})$ denotes the asymptotic variance of the estimator that uses $\boldsymbol{u}(.)$ and $\boldsymbol{\phi}(.)$.

In practice, $E\{\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta}) \mid \boldsymbol{Z}^{(1)}, R=1\}$ is unknown. So, a parametric imputation model $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta},\boldsymbol{\gamma})$ for $E\{\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta}) \mid \boldsymbol{Z}^{(1)}, R=1\}$ is specified. This model can be specified either directly, or indirectly by choosing a model $f(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}^{(1)}, R=1;\boldsymbol{\gamma})$ for $f(\boldsymbol{z}^{(2)} \mid \boldsymbol{Z}^{(1)}, R=1)$. Denote by $\mathcal{M}_{\text{imp}}$ the semiparametric model for $(\boldsymbol{Z}^{(1)}, R\boldsymbol{Z}^{(2)}, R)$ defined by models $\mathcal{M}$ and $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta},\boldsymbol{\gamma})$ and the MAR assumption. Let $\hat{\boldsymbol{\gamma}}$ denote an estimator of $\boldsymbol{\gamma}$ based on the complete cases (e.g. the ML estimator). Now, $\boldsymbol{\beta}$ can be estimated as the solution $\hat{\boldsymbol{\beta}}_{\text{DR}} = \hat{\boldsymbol{\beta}}_{\text{DR}}(\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$ to the DR (AIPW) estimating equations

$$\sum_{i=1}^{n} \boldsymbol{S}_{\boldsymbol{\beta},i}(\boldsymbol{\beta},\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \frac{R_i}{\pi(\boldsymbol{Z}_i^{(1)};\hat{\boldsymbol{\alpha}})}\boldsymbol{u}(\boldsymbol{Z}_i;\boldsymbol{\beta}) + \left\{1 - \frac{R_i}{\pi(\boldsymbol{Z}_i^{(1)};\hat{\boldsymbol{\alpha}})}\right\}\boldsymbol{\phi}(\boldsymbol{Z}_i^{(1)};\boldsymbol{\beta},\hat{\boldsymbol{\gamma}}) = \boldsymbol{0}. \tag{5}$$

This is a RAL estimator for $\boldsymbol{\beta}$ under model $\mathcal{M}_{\text{miss}}$ when $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ is correctly specified and $\hat{\boldsymbol{\alpha}}$ is consistent. Moreover, it turns out that $\hat{\boldsymbol{\beta}}_{\text{DR}}$ is also a RAL estimator for $\boldsymbol{\beta}$ under model $\mathcal{M}_{\text{imp}}$ when $\hat{\boldsymbol{\gamma}}$ is consistent. That is, $\hat{\boldsymbol{\beta}}_{\text{DR}}$ is CAN if model $\mathcal{M}$ is correctly specified, the data are MAR, and either i) $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ is correctly specified and $\hat{\boldsymbol{\alpha}}$ is consistent, or ii) $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\beta,\boldsymbol{\gamma})$ is correctly specified and $\hat{\boldsymbol{\gamma}}$ is consistent (or both) (see the supplemental article [36] for proof). For this reason, $\hat{\boldsymbol{\beta}}_{\text{DR}}$ is called 'double robust'. On the other hand, the IPW estimator which replaces $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta},\boldsymbol{\gamma})$ with zero is a RAL estimator of $\boldsymbol{\beta}$ only under model $\mathcal{M}_{\text{miss}}$. That is, it is CAN only if $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ is correctly specified.

Let us apply equation (5) to the missing outcome problem of Section 2. In this case, $Z^{(2)} = Y$, $\boldsymbol{Z}^{(1)} = \boldsymbol{W}$, $\boldsymbol{u}(\boldsymbol{Z};\beta) = Y - \beta$ and $\phi(\boldsymbol{Z}^{(1)};\beta,\hat{\boldsymbol{\gamma}}) = m(\boldsymbol{W};\hat{\boldsymbol{\gamma}}) - \beta$. Here $E\{u(Z^{(2)};\beta) \mid \boldsymbol{Z}^{(1)}, R=1\} = E(Y \mid \boldsymbol{W}, R=1) - \beta$, and so it suffices to specify a model $m(\boldsymbol{W};\boldsymbol{\gamma})$ for $E(Y \mid \boldsymbol{W}, R=1)$. It is easy to show that the solution $\hat{\beta}_{\text{DR}}$ to equation (5) is the same estimator $\hat{\beta}_{\text{DR}}$ that we met in Section 2. If, on the other hand, $\phi(\boldsymbol{Z}^{(1)};\beta,\hat{\boldsymbol{\gamma}})$ is set to zero, then the solution to equation (5) is $\hat{\beta}_{\text{IPW,B}}$.

When $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ and $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta},\boldsymbol{\gamma})$ are correctly specified, the solution $\hat{\boldsymbol{\beta}}_{\text{DR}}$ to equations (5) is locally efficient **over the class of estimators** that solve equations (4) for the given $\boldsymbol{u}(\boldsymbol{Z}_i;\boldsymbol{\beta})$ and arbitrary $\boldsymbol{\phi}(\boldsymbol{Z}_i^{(1)};\boldsymbol{\beta})$. More broadly, however, the efficiency of $\hat{\boldsymbol{\beta}}_{\text{DR}}$ also depends on the choice of function $\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta})$. The choice that maximises efficiency under model $\mathcal{M}_{\text{miss}}$ is generally difficult to find [40]. It is usually different from that which gives local efficiency under model $\mathcal{M}$. For example, we saw in Section 3.1 that for the restricted moment model, $\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta}) = \boldsymbol{D}^{\top}(\boldsymbol{X})\boldsymbol{V}^{-1}(\boldsymbol{X})\{\boldsymbol{Y} - \boldsymbol{\mu}(\boldsymbol{X};\boldsymbol{\beta})\}$ gives the locally efficient estimator under $\mathcal{M}$. This is not necessarily the efficient choice under $\mathcal{M}_{\text{miss}}$. An exception to this general rule is the missing outcome problem of Section 2, where $u(\boldsymbol{Z};\beta) = Y - \beta$ gives global efficiency under $\mathcal{M}$ and local efficiency under $\mathcal{M}_{\text{miss}}$.

So far, we have considered the case where there are only two missingness patterns. The general case of monotone missing data (e.g. longitudinal data with dropout) is treated in the supplemental article [36]. Here, the DR estimator is CAN if either

of two sets of models is correctly specified. The first set is for the conditional probability of dropout at each time point given the variables available at that time. The second set is for the conditional expectation of $\boldsymbol{u}(\boldsymbol{Z};\boldsymbol{\beta})$ given the variables available up to each time point and not dropping out before that time.

Example 3 continued: The difficulty faced by Schnitzer et al. in estimating the parameters of their prediction model was that many clinical events were censored, due to loss to follow-up or deviation from assigned therapy. To deal with this, they used a DR estimator. During follow-up, CD4-cell and HIV-RNA counts were measured at least every 16 weeks, and the dropout and conditional expectation models for each timepoint used the CD4 and RNA counts measured at the previous timepoint.

Example 4 continued: In the trial considered by Seaman and Copas, 16% of patients dropped out before the end. Dropout was higher among patients who were younger, injected drugs or were no longer on assigned therapy, making estimates based on complete cases potentially biased. So, Seaman and Copas used DR estimation. The dropout and conditional expectation models for each timepoint used treatment arm, injecting behaviour, and an indicator of being on assigned therapy, CD4 cell count and RNA count at the previous timepoint.

Example 5 continued: In the cohort used by Qi et al. bone mineral density was measured in less than 10% of women, making a complete-case analysis potentially inefficient. They instead used DR estimation to handle these missing covariate data. This required models for the probability that mineral density was observed and for the distribution of mineral density given that it was observed. The covariates in these models were the event/censoring time, the event indicator and age, and both models were estimated using kernel smoothers. By using the data on all the women, the precisions of the hazard ratio estimates were increased relative to complete-case estimates. A description of the method used can be found in the supplemental article [36].

3.3 Asymptotic distribution of DR estimators

The variance of $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ generally depends on the choice of estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$. Suppose these are obtained as the solutions to estimating equations $\sum_{i=1}^{n}\boldsymbol{S}_{\alpha,i}(\hat{\boldsymbol{\alpha}})=\boldsymbol{0}$ and $\sum_{i=1}^{n}\boldsymbol{S}_{\gamma,i}(\hat{\boldsymbol{\gamma}})=\boldsymbol{0}$, and let $\boldsymbol{\theta}=(\boldsymbol{\beta}^{\top},\boldsymbol{\alpha}^{\top},\boldsymbol{\gamma}^{\top})^{\top}$. **For example, if** $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})=\mathrm{expit}(\boldsymbol{\alpha}^{\top}\boldsymbol{Z}^{(1)})$**, then** $\boldsymbol{S}_{\alpha}(\boldsymbol{\alpha})=\boldsymbol{Z}^{(1)}\{R-\mathrm{expit}(\boldsymbol{\alpha}^{\top}\boldsymbol{Z}^{(1)})\}$**.** When $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ or $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\boldsymbol{\beta},\boldsymbol{\gamma})$ (or both) is correctly specified, the variance of $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ is consistently estimated by the sandwich estimator $n^{-1}\left\{\sum_{i=1}^{n}\partial\boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})/\partial\boldsymbol{\beta}^{\top}\right\}^{-1}\left\{\sum_{i=1}^{n}\boldsymbol{S}_{i}(\boldsymbol{\theta})\boldsymbol{S}_{i}(\boldsymbol{\theta})^{\top}\right\}\left\{\sum_{i=1}^{n}\partial\boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})^{\top}/\partial\boldsymbol{\beta}\right\}^{-1}$ evaluated at $\hat{\boldsymbol{\theta}}=(\hat{\boldsymbol{\beta}}_{\mathrm{DR}},\hat{\boldsymbol{\alpha}}^{\top},\hat{\boldsymbol{\gamma}}^{\top})^{\top}$, where

$$\begin{aligned}\boldsymbol{S}_{i}(\boldsymbol{\theta}) &= \boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})-\left\{\sum_{i=1}^{n}\frac{\partial\boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})}{\partial\boldsymbol{\alpha}^{\top}}\right\}\left\{\sum_{i=1}^{n}\frac{\partial\boldsymbol{S}_{\alpha,i}(\boldsymbol{\alpha})}{\partial\boldsymbol{\alpha}^{\top}}\right\}^{-1}\boldsymbol{S}_{\alpha,i}(\boldsymbol{\alpha})\\ &\quad-\left\{\sum_{i=1}^{n}\frac{\partial\boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})}{\partial\boldsymbol{\gamma}^{\top}}\right\}\left\{\sum_{i=1}^{n}\frac{\partial\boldsymbol{S}_{\gamma,i}(\boldsymbol{\gamma})}{\partial\boldsymbol{\gamma}^{\top}}\right\}^{-1}\boldsymbol{S}_{\gamma,i}(\boldsymbol{\gamma}).\end{aligned}$$

Here, the terms involving $\boldsymbol{S}_{\alpha,i}(\boldsymbol{\alpha})$ and $\boldsymbol{S}_{\gamma,i}(\boldsymbol{\gamma})$ can be viewed as accounting for the uncertainty in $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, respectively. An alternative to the sandwich estimator is nonparametric bootstrap. The latter is commonly used, possibly because the former may be negatively biased when the effective sample size is small or to construct confidence intervals that do not rely on a normal assumption [7].

When both $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ and $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\beta,\boldsymbol{\gamma})$ are correctly specified, $\boldsymbol{S}_i(\boldsymbol{\theta}) = \boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})$ (up to a term that converges to zero in probability — see proof of DR in [36]). An important implication of this is that the asymptotic variance of $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ does not depend on the choice of (consistent) estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ in that case, and in fact equals the asymptotic variance of the DR estimator $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}(\boldsymbol{\alpha},\boldsymbol{\gamma})$ that uses the true values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$. It is therefore tempting to replace $\boldsymbol{S}_i(\boldsymbol{\theta})$ by $\boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})$ in the sandwich variance estimator. We discourage this in general, because, although $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ is DR, inference for $\boldsymbol{\beta}$ is not DR when this is done, as consistency of the resulting variance estimator is no longer guaranteed as soon as one or both of $\pi(\boldsymbol{Z}^{(1)};\boldsymbol{\alpha})$ and $\boldsymbol{\phi}(\boldsymbol{Z}^{(1)};\beta,\boldsymbol{\gamma})$ is misspecified. Under such misspecification, or when the sample size is small, the choice of estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ can be very important. We return to this issue in the next section.

4 Improved double robust estimators

For simplicity, we concentrate in this section on the missing outcome problem of Section 2. The notation is the same as used there. Also, $\boldsymbol{\alpha}_0$ and $\boldsymbol{\gamma}_0$ denote the probability limits of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, i.e. $\hat{\boldsymbol{\alpha}} \xrightarrow{p} \boldsymbol{\alpha}_0$ and $\hat{\boldsymbol{\gamma}} \xrightarrow{p} \boldsymbol{\gamma}_0$. Much of the material in this section is adapted from Rotnitzky and Vansteelandt [31] and more details can be found there, including information on which methods have been extended to estimate the parameters of a semiparametric regression model with partially observed outcome and fully observed covariates or to handle longitudinal data with dropout.

4.1 Drawbacks of the standard DR estimator

Let $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ denote locally efficient semiparametric estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ under the missingness and outcome models, respectively. For example, $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ could be ML estimators in logistic and linear regression models, respectively. When $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ and $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ are used, the estimator $\hat{\beta}_{\mathrm{DR}}$ given by equation (1) or (equivalently) (2) is sometimes called the 'standard' DR estimator [5]. There are some issues with this estimator.

First, $\hat{\beta}_{\mathrm{DR}}$ may lie outside its parameter space (e.g. outside $[0,1]$ when Y is binary). Even when guaranteed to lie within its parameter space, it may not be within the range of the observed Y values. An estimate of $E(Y)$ that is less (more) than the minimum (maximum) observed value of Y may be difficult to defend [25].

Second, when model $m(\boldsymbol{W};\boldsymbol{\gamma})$ is misspecified, there is no guarantee that $\hat{\beta}_{\mathrm{DR}}$ will be at least as efficient as the IPW estimators $\hat{\beta}_{\mathrm{IPW}}$ and $\hat{\beta}_{\mathrm{IPW,B}}$.

Third, in practical applications, both models $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ and $m(\boldsymbol{W};\boldsymbol{\gamma})$ are likely to be at least mildly misspecified, so that neither of the conditions for consistency of $\hat{\beta}_{\mathrm{DR}}$ applies. The hope is that $\hat{\beta}_{\mathrm{DR}}$ will still perform well when at least one of these models is approximately correctly specified. However, Kang and Schafer [12] demonstrated that this is not necessarily the case. They gave an example of a data-generating mechanism for $(Y, \boldsymbol{W}, R)$ and two misspecified models $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ and $m(\boldsymbol{W};\boldsymbol{\gamma})$ and showed that the standard DR estimator has very large bias and variance in this example, even though the model misspecification is not easily detected from the observed data on a sample of moderate size. They also showed that the RI estimator $\hat{\beta}_{\mathrm{RI}}$ has relatively small bias and variance in this example. Robins et al. [25] examined Kang and Schafer's data-generating mechanism. They noted that the overlap between the distributions of $\boldsymbol{W}$ in the complete and incomplete cases was small. As discussed in Section 2, this means that $\hat{\beta}_{\mathrm{RI}}$ relies on potentially dangerous extrapolation, and thus that its good performance is partly a matter of luck. Indeed, Robins et al. [25] showed that if Kang and Schafer's missingness mechanism was altered by making complete cases into incomplete cases and vice versa (by replacing R by $1-R$), the performance of $\hat{\beta}_{\mathrm{RI}}$ became much worse than that of $\hat{\beta}_{\mathrm{DR}}$. Nevertheless, this example cast some doubt on the practical usefulness of the DR property of $\hat{\beta}_{\mathrm{DR}}$.

The response to these issues has been the development of improved DR estimators, which aim at greater efficiency and reduced bias relative to the standard DR estimator. These differ from that estimator in the way that $\boldsymbol{\alpha}$ and/or $\boldsymbol{\gamma}$ are estimated. As noted in Section 3, the choice of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ affects the asymptotic variance of $\hat{\beta}_{\mathrm{DR}}$ unless both $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ and $m(\boldsymbol{W};\boldsymbol{\gamma})$ are correctly specified, and affects its asymptotic bias when neither is correctly specified.

These improved estimators are not a panacea for scenarios where the population variance of the true weights $\pi(\boldsymbol{W})^{-1}$ is large. In this case, there is limited overlap between the distributions of $\boldsymbol{W}$ in complete and incomplete cases and, unless one is prepared to trust in extrapolation to incomplete cases of an outcome model fitted to complete cases, considerable uncertainty in the estimate of β is inevitable. However, the improved estimators go a long way to resolving the issues with the standard DR estimator listed above. First, most of them guarantee $\hat{\beta}$ lies within the parameter space of β. Some are even sample bounded. As well as avoiding implausible estimates, sample boundedness can reduce the variance of $\hat{\beta}_{\mathrm{DR}}$ when the weights are highly variable. Second, some of the improved estimators are asymptotically efficient over a class of estimators that includes the simple IPW estimators, provided that $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ is correctly specified, even when $m(\boldsymbol{W};\boldsymbol{\gamma})$ is potentially misspecified. Third, some more recent methods aim to improve performance when both $m(\boldsymbol{W};\boldsymbol{\gamma})$ and $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ may be misspecified or when the true weights are unstable. We now review these improved DR methods.

4.2 DR RI and DR sample-bounded IPW estimators

Several methods calculate $\hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$ and then estimate $\boldsymbol{\gamma}$ in such a way that ensures

$$\sum_{i=1}^{n} \frac{R_i}{\pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}}_{\mathrm{ML}})} \{Y_i - m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}})\} = 0. \tag{6}$$

As the left-hand side of equation (6) is the 'correction' term in equation (2), this ensures that $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ reduces to a RI estimator, i.e. $\hat{\beta}_{\mathrm{DR}} = n^{-1} \sum_{i=1}^{n} m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}})$. The advantage of this is that, if the range of $m(\boldsymbol{W}; \boldsymbol{\gamma})$ equals the parameter space of β, then $\hat{\beta}_{\mathrm{DR}}$ must lie within this space. Further, Gruber and van der Laan show how to ensure that the range of $m(\boldsymbol{W}; \boldsymbol{\gamma})$ equals the range of the observed Y values, making the resulting RI estimator sample-bounded [10].

When $m(\boldsymbol{W}; \boldsymbol{\gamma})$ is a generalised linear model with canonical link function, two ways to make equation (6) hold are: i) to estimate $\boldsymbol{\gamma}$ using the ML estimator with weights $\pi(\boldsymbol{W}; \hat{\boldsymbol{\alpha}}_{\mathrm{ML}})^{-1}$ [12]; or ii) to include $\pi(\boldsymbol{W}; \hat{\boldsymbol{\alpha}}_{\mathrm{ML}})^{-1}$ as an extra covariate in $m(\boldsymbol{W}; \boldsymbol{\gamma})$ and then estimate $\boldsymbol{\gamma}$ by ML [32] (the first way requires that $m(\boldsymbol{W}; \boldsymbol{\gamma})$ include an intercept term). In either case, equation (6) is one of the score equations for $\hat{\boldsymbol{\gamma}}$ (corresponding to the intercept in the first case and to the covariate $\pi(\boldsymbol{W}; \hat{\boldsymbol{\alpha}}_{\mathrm{ML}})^{-1}$ in the second case) and hence holds at $\hat{\boldsymbol{\gamma}}$. Note that if the original model for $E(Y \mid \boldsymbol{W})$ is correctly specified, then the extended model with covariate $\pi(\boldsymbol{W}; \hat{\boldsymbol{\alpha}}_{\mathrm{ML}})^{-1}$ added will still be correctly specified. When the original model for $E(Y \mid \boldsymbol{W})$ is misspecified, the first DR RI estimator usually has better performance than the second [31].

Robins et al. [25] proposed calculating $\hat{\boldsymbol{\gamma}}_{\mathrm{ML}}$ and then estimating $\boldsymbol{\alpha}$ in such a way that $\sum_{i=1}^{n} R_i \pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})^{-1} \{m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}}_{\mathrm{ML}}) - n^{-1} \sum_{j=1}^{n} m(\boldsymbol{W}_j; \hat{\boldsymbol{\gamma}}_{\mathrm{ML}})\} = 0$. The sample-bounded estimator $\hat{\beta}_{\mathrm{IPW,B}}(\hat{\boldsymbol{\alpha}})$ is then DR. This DR estimator is related to the minimum-discrepancy estimators discussed in [36]: they all calculate the weights in such a way that the weighted average of $m(\boldsymbol{W}; \hat{\boldsymbol{\gamma}}_{\mathrm{ML}})$ in the complete cases is equal to the corresponding unweighted average in the whole sample.

4.3 Efficient estimators over a class of estimators

All the improved estimators described so far suffer from the drawback that, if $m(\boldsymbol{W}; \boldsymbol{\gamma})$ is misspecified, they can potentially be less efficient than the IPW estimators $\hat{\beta}_{\mathrm{IPW}}$ and $\hat{\beta}_{\mathrm{IPW,B}}$. We now describe DR estimators that are, when $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ is correctly specified, guaranteed to be at least as asymptotically efficient as the IPW estimators that use the same model $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$.

Consider a correctly specified model $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ and a fixed choice of (possibly misspecified) model $m(\boldsymbol{W}; \boldsymbol{\gamma}) = h(\boldsymbol{\gamma}^{\top} \boldsymbol{W})$, where h is a known link function, and let $\hat{\boldsymbol{\alpha}} = \hat{\boldsymbol{\alpha}}_{\mathrm{ML}}$. Let $\hat{\beta}(\nu_1, \nu_2, \boldsymbol{\gamma})$, where ν_1 and ν_2 are real numbers, denote the estimator that solves equation (5) with $\boldsymbol{Z}_i^{(1)} = \boldsymbol{W}_i$, $\boldsymbol{u}(\boldsymbol{Z}_i; \boldsymbol{\beta}) = Y_i - \beta$ and $\boldsymbol{\phi}(\boldsymbol{Z}_i^{(1)}; \boldsymbol{\beta}, \hat{\boldsymbol{\gamma}})$ replaced by $\nu_1 + \nu_2 m(\boldsymbol{W}_i; \boldsymbol{\gamma}) - \beta$. So, in particular, $\hat{\beta}(0, 1, \hat{\boldsymbol{\gamma}}_{\mathrm{ML}})$ is the standard DR estimator and $\hat{\beta}(0, 0, \mathbf{0})$ and $\hat{\beta}(\beta, 0, \mathbf{0})$ are, respectively, $\hat{\beta}_{\mathrm{IPW}}$ and $\hat{\beta}_{\mathrm{IPW,B}}$.

Cao et al. [5] and Tan [38, 39] independently derived estimators that are asymptotically efficient over the set $\{\hat{\beta}(\nu_1, \nu_2, \boldsymbol{\gamma}) : -\infty < \nu_1, \nu_2 < \infty, \boldsymbol{\gamma} \in \Gamma\}$, where Γ is the parameter space of $\boldsymbol{\gamma}$. That is, their asymptotic variances cannot be greater than that of any AIPW estimator that uses in its augmentation term $\nu_1 + \nu_2 m(\boldsymbol{W}; \boldsymbol{\gamma})$ for any fixed ν_1, ν_2 and $\boldsymbol{\gamma}$. In particular, they cannot be greater than those of $\hat{\beta}_{\text{IPW}}$, $\hat{\beta}_{\text{IPW,B}}$ and the standard DR estimator (because the last has the same asymptotic variance as $\hat{\beta}(0, 1, \boldsymbol{\gamma}_0)$ when $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ is correctly specified — see proof of DR in [36]). **Cao et al.'s method is essentially obtained by choosing β as the value that minimises the empirical asymptotic variance of the DR estimator (as obtained via a sandwich estimator). Rotnitzky et al. [30] derived a DR RI estimator that is at least as asympotically efficient as both $\hat{\beta}(0, 1, \boldsymbol{\gamma})$ for any $\boldsymbol{\gamma} \in \Gamma$ and $\hat{\beta}_{\text{IPW,B}}$. If $m(\boldsymbol{W}; \boldsymbol{\gamma}) = 0$ for some $\boldsymbol{\gamma} \in \Gamma$, then $\hat{\beta}(0, 1, \boldsymbol{\gamma}) = \hat{\beta}(0, 0, \mathbf{0}) \equiv \hat{\beta}_{\text{IPW}}$ for this value of $\boldsymbol{\gamma}$, so that Rotnitzky et al.'s estimator is also at least as asympotically efficient as $\hat{\beta}_{\text{IPW}}$.**

Tan's [39] estimator (which builds on his earlier work [37]) has the advantage that it is sample bounded. Cao et al.'s [5] method (further developed by Tsiatis et al. [42]) and Rotnizky et al.'s [30] method have the advantage that they allow estimation of the parameters of a semiparametric regression model, even for longitudinal data with dropout. However, when $\boldsymbol{\beta}$ is a vector, Cao et al.'s estimator ensures asymptotic efficiency for only one specified element of $\boldsymbol{\beta}$. Rotnitzky et al.'s estimator ensures asymptotic efficiency for all elements of $\boldsymbol{\beta}$ (and indeed for a finite number of arbitrary scalar functions of $\boldsymbol{\beta}$).

4.4 Bias-reduced DR estimators

The methods listed in Section 4.3 minimise the asymptotic variance of $\hat{\beta}_{\text{DR}}$ over a class of AIPW estimators when $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ is correctly specified, but are not guaranteed to do so when it is misspecified. Vermeulen and Vansteelandt [46] took a different approach. Rather than seeking directly to minimise the asymptotic variance, their 'bias-reduced DR estimator' uses the estimators $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ obtained by locally minimising the squared asymptotic bias of $\hat{\beta}_{\text{DR}}$ when both models $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ and $m(\boldsymbol{W}; \boldsymbol{\gamma})$ are misspecified. This makes the bias-reduced DR estimator less sensitive than the standard DR estimator to mild model misspecification. This can be understood as follows. The asymptotic bias of $\hat{\beta}_{\text{DR}}$ equals $E[\{\pi(\boldsymbol{W}; \boldsymbol{\alpha}_0) - \pi(\boldsymbol{W})\}\{m(\boldsymbol{W}; \boldsymbol{\gamma}_0) - E(Y \mid \boldsymbol{W})\}\pi(\boldsymbol{W}; \boldsymbol{\alpha}_0)^{-1}]$ [46]. That is, it is the product of the degrees of misspecification of the two models inversely weighted by $\pi(\boldsymbol{W}; \boldsymbol{\alpha}_0)$. This weighting is concerning, because it is in the region where $\pi(\boldsymbol{W}; \boldsymbol{\alpha}_0)$ is small that few complete cases are observed, and so misspecification of $m(\boldsymbol{W}; \boldsymbol{\gamma})$ is most likely to remain undetected. Vermeulen and Vansteelandt's choice of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ makes the asymptotic bias reduce to $E[\{m(\boldsymbol{W}; \boldsymbol{\gamma}_0) - E(Y \mid \boldsymbol{W})\}\{1 - \pi(\boldsymbol{W})\}]$, hence avoiding this problem.

Bias-reduced DR estimation can be used for quite general semiparametric models, even when data are assumed to be MNAR. However, when $\boldsymbol{\beta}$ is a vector, the squared asymptotic bias is minimised only for one specified element of $\boldsymbol{\beta}$.

Estimating the variance of $\hat{\boldsymbol{\beta}}_{\mathrm{DR}}$ is straightforward for the bias-reduced estimator, because a fortunate effect of the way that $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$ are calculated is that uncertainty in these parameters can be ignored even when both models $\pi(\boldsymbol{W};\boldsymbol{\alpha})$ and $m(\boldsymbol{W};\boldsymbol{\gamma})$ are misspecified. The variance can thus be estimated as explained in Section 3.3, replacing $\boldsymbol{S}_i(\boldsymbol{\theta})$ by $\boldsymbol{S}_{\beta,i}(\boldsymbol{\theta})$. This may also explain why the bias-reduced estimator appears to have good efficiency in simulation studies [46].

Simulation studies that compare many of the improved DR methods discussed in Section 4.2–4.4 have been reported [39, 22, 46]. In these studies, the improved methods had less bias and greater efficiency than the standard DR estimator when the outcome model was misspecified; differences were less marked when only the missingness model was misspecified. The estimators of Sections 4.3 and 4.4 performed better than those of Section 4.2, but among the former group no method was uniformly best. The range of data-generating mechanisms considered in these studies was quite small, however, and more research would be welcome.

5 Data-adaptive methods

The increasing popularity and availability of data-adaptive statistical methods (e.g. kernel smoothing, penalised likelihood, ensemble learners) may lead the reader to wonder what is the use of DR estimators when RI estimators and IPW estimators can be based on outcome imputations and missingness probabilities, respectively, obtained via such flexible methods [16]. In this section, we provide insight into this matter, and argue that DR estimators are in fact especially useful when data-adaptive methods are used.

For simplicity, we return to the missing outcome problem of Section 2. Consider the RI estimator $\hat{\beta}_{\mathrm{RI}} = n^{-1}\sum_{i=1}^{n} m(\boldsymbol{W}_i;\hat{\boldsymbol{\gamma}})$, where $\hat{\boldsymbol{\gamma}}$ is an estimate obtained through some data-adaptive statistical method (e.g. standard variable selection). The estimator $\hat{\boldsymbol{\gamma}}$ will typically have a complicated finite-sample distribution [13] and non-uniform convergence of this distribution to a normal distribution, properties which the RI estimator $\hat{\beta}_{\mathrm{RI}}$ will usually inherit. The practical implication of this is that uniformly valid confidence intervals with nominal coverage for β based on $\hat{\beta}_{\mathrm{RI}}$ are difficult to obtain. **Confidence intervals that are not uniformly valid are not guaranteed to perform well, because, for any given n, no matter how large, there exist distributions of the full data for which their coverage is poor.** This problem is well known for, e.g., lasso-estimators $\hat{\boldsymbol{\gamma}}$, where a small change in the data-generating mechanism (e.g. an element of $\boldsymbol{\gamma}$ changing from 0 to $n^{-1/2}$) may lead to a relatively large change in the distribution of $\hat{\boldsymbol{\gamma}}$ even for large n, because it may lead to different variables being selected asymptotically [14, 13].

To develop more formal insight into this, we consider the difficulty that arises in the specific example of lasso or post-lasso (post-lasso is the procedure that uses lasso as a variable-selection procedure and then refits the selected model using a standard procedure (e.g. ML) to reduce shrinkage bias [2]). Similar problems arise with other data-adaptive methods. Let $\hat{\boldsymbol{\gamma}}$ be an estimator of $\boldsymbol{\gamma}$ obtained via lasso

or post-lasso. Then [3, 8, 9],

$$\begin{aligned}\sqrt{n}(\hat{\beta}_{\mathrm{RI}}-\beta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(\boldsymbol{W}_i;\hat{\boldsymbol{\gamma}})-\beta_0\}\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(\boldsymbol{W}_i;\boldsymbol{\gamma}_0)-\beta_0\}+\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(\boldsymbol{W}_i;\hat{\boldsymbol{\gamma}})-m(\boldsymbol{W}_i;\boldsymbol{\gamma}_0)\}\\ &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\{m(\boldsymbol{W}_i;\boldsymbol{\gamma}_0)-\beta_0\}+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial m}{\partial\boldsymbol{\gamma}}(\boldsymbol{W}_i;\boldsymbol{\gamma})\Big|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\sqrt{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)\\ &\quad +\sqrt{n}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2^2 O_p(1). \qquad (7)\end{aligned}$$

where $\|.\|_2$ denotes the Euclidean norm. Assuming that $m(\boldsymbol{W};\boldsymbol{\gamma})$ is correctly specified, the first term in the expansion (7) generally has an asymptotic normal mean-zero distribution and the remainder term $\sqrt{n}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2^2 O_p(1)$ tends to be of lower order than the other two terms. Although the term $n^{-1}\sum_{i=1}^{n}\partial m(\boldsymbol{W}_i;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\times\sqrt{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ does (for fixed $\boldsymbol{\gamma}_0$ and assuming regularity conditions) converge in distribution to a normal distribution, this convergence is generally not uniform. That is, for any n, no matter how large, there exist values of $\boldsymbol{\gamma}_0$ for which the distribution of $\sqrt{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ is far from its asymptotic distribution — and hence for which $\sqrt{n}(\hat{\beta}_{\mathrm{RI}}-\beta_0)$ is far from its asymptotic distribution.

An additional concern arises when p, the dimension of $\boldsymbol{\gamma}$, is large relative to n. Lasso and other penalised likelihood methods are commonly used in such settings. Large-sample behaviour of $\hat{\boldsymbol{\gamma}}$ as p increases with n is therefore of interest. When p increases with n, there is (in addition to the forementioned difficulty of obtaining uniformly valid confidence intervals) a problem that bias in $\hat{\beta}_{\mathrm{RI}}$ may vanish only slowly with increasing n unless the true data-generating mechanism shows sufficient sparsity, i.e. unless the rate at which s, the number of non-zero elements of $\boldsymbol{\gamma}_0$, increases as n increases is sufficiently small [2]. More specifically, it follows from [2] that, for lasso and post-lasso estimators, $\sqrt{n}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2^2 = O_p((s/\sqrt{n})\log(p\vee n))$, where $a\vee b$ denotes the maximum of a and b. When there is sufficient sparsity to ensure that $(s/\sqrt{n})\log(p\vee n)$ converges to zero, the second-order term $\sqrt{n}\|\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0\|_2^2$ converges to zero. However, greater sparsity is required to prevent the term $n^{-1}\sum_{i=1}^{n}\partial m(\boldsymbol{W}_i;\boldsymbol{\gamma})/\partial\boldsymbol{\gamma}|_{\boldsymbol{\gamma}=\boldsymbol{\gamma}_0}\times\sqrt{n}(\hat{\boldsymbol{\gamma}}-\boldsymbol{\gamma}_0)$ in equation (7) from diverging to infinity, and so to **ensure that bias in $\sqrt{n}(\hat{\beta}_{\mathrm{RI}}-\beta_0)$ vanishes as $n\to\infty$.**

The above concerns largely disappear when data-adaptive methods are combined with DR estimators, because DR estimators enjoy a small bias property [20, 8]. This means that their bias vanishes faster than the bias in the nuisance parameter estimator (e.g. $\hat{\boldsymbol{\gamma}}$) when the smoothing parameter (e.g. the bandwidth in a kernel estimator or the penalty parameter in a lasso-estimator) goes to zero. This property is important for ensuring correct inference when data-adaptive methods are used [3]. This can more formally be understood as follows. Consider again the estimator $\hat{\beta}_{\mathrm{DR}} = n^{-1}\sum_{i=1}^{n}\rho(R_i, R_iY_i, \boldsymbol{W}_i;\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}})$, where

$$\rho(R,RY,\boldsymbol{W};\hat{\boldsymbol{\alpha}},\hat{\boldsymbol{\gamma}}) = \frac{R}{\pi(\boldsymbol{W};\hat{\boldsymbol{\alpha}})}Y+\left\{1-\frac{R}{\pi(\boldsymbol{W};\hat{\boldsymbol{\alpha}})}\right\}m(\boldsymbol{W};\hat{\boldsymbol{\gamma}}),$$

with $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ obtained through some data-adaptive statistical method. Then upon repeating the expansion of equation (7) with $\hat{\beta}_{\mathrm{DR}}$ in place of $\hat{\beta}_{\mathrm{RI}}$, $\rho(R, RY, \boldsymbol{W}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$ in place of $m(\boldsymbol{W}; \hat{\boldsymbol{\gamma}})$, and $\boldsymbol{\theta} = (\boldsymbol{\alpha}^\top, \boldsymbol{\gamma}^\top)^\top$ in place of $\boldsymbol{\gamma}$, one can see that slow convergence of $\hat{\boldsymbol{\gamma}}$ and $\hat{\boldsymbol{\alpha}}$ does not necessarily induce erratic behaviour in $\hat{\beta}_{\mathrm{DR}}$. This is because, as noted in the proof of DR in [36], $\partial\rho(R, RY, \boldsymbol{W}; \boldsymbol{\alpha}, \boldsymbol{\gamma})/\partial\boldsymbol{\theta}$ has expectation zero at $(\boldsymbol{\alpha}_0, \boldsymbol{\gamma}_0)$ when $\pi(\boldsymbol{W}; \boldsymbol{\alpha})$ and $m(\boldsymbol{W}; \boldsymbol{\gamma})$ are correctly specified, and so slow convergence of the first-order term $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ in the expansion is not a problem (so long as $\sqrt{n}\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\|_2^2$ converges to zero).

Farrell [9] uses this idea to demonstrate that, under conditions that we specify next, $\hat{\beta}_{\mathrm{DR}}$ is asymptotically unbiased and uniformly valid 95% confidence regions for β can be straightforwardly calculated as $\hat{\beta}_{\mathrm{DR}} \pm 1.96\sqrt{\hat{\sigma}^2/n}$, where $\hat{\sigma}^2$ is the sample variance of $\rho(R, RY, \boldsymbol{W}; \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\gamma}})$. These conditions are that the empirical mean squared errors of $m(\boldsymbol{W}; \hat{\boldsymbol{\gamma}})$ and $\pi(\boldsymbol{W}; \hat{\boldsymbol{\alpha}})$ converge in probability to zero, and that their product converges at faster than n^{-1}-rate. This in particular allows slow convergence of $\hat{\boldsymbol{\alpha}}$, so long as $\hat{\boldsymbol{\gamma}}$ converges sufficiently fast, and vice versa.

The results of Farrell [9] apply to any data-adaptive method for estimating $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\gamma}}$, so long as it satisfies the aforementioned conditions. Targeted maximum likelihood estimation (TMLE), proposed by van der Laan and Rubin [44] and refined by Gruber and van der Laan [10], is one such procedure. It is designed to ensure that the DR estimator reduces to a RI estimator (or 'substitution estimator' in their terminology). It involves two steps. First, a preliminary estimate $m^{(0)}(\boldsymbol{W}; \hat{\boldsymbol{\gamma}})$ of $E(Y \mid \boldsymbol{W})$ based on a data-adaptive learning algorithm (e.g. an ensemble learner) is obtained, and a parametric missingness model is fitted to obtain $\hat{\boldsymbol{\alpha}}$. Second, a canonical generalised linear model for $E(Y \mid \boldsymbol{W})$ is fitted, with link function $h(.)$, offset term $h^{-1}\left\{m^{(0)}(\boldsymbol{W}; \hat{\boldsymbol{\gamma}})\right\}$ and the single covariate $R/\pi(\boldsymbol{W}; \hat{\boldsymbol{\alpha}})$. This covariate is chosen because ML estimation of its coefficient involves setting $\sum_{i=1}^n R_i \pi(\boldsymbol{W}_i; \hat{\boldsymbol{\alpha}})^{-1} \{Y_i - m(\boldsymbol{W}_i; \hat{\boldsymbol{\gamma}})\}$ to zero, thereby making the DR estimator equivalent to a RI estimator.

6 Discussion

Much research on DR estimators has been for the missing outcome problem of Section 2 and for restricted moment models with missing outcome or covariates (see Section 3 and [36]). Other applications have included, e.g., estimating the area under an ROC curve with missing outcome or predictor [18, 29]. The DR property is not unique to methods for incomplete data. The missing outcome problem of Section 2 is closely related to that of estimating an average causal effect, and essentially the same DR estimators appear in this literature (e.g. [1]). DR estimators have also been proposed for many other causal inference problems. Rotnitzky and Vansteelandt [31] list numerous examples of DR estimators, within and without the causal inference literature.

We have focussed on DR incomplete-data estimators for scenarios where a full-data m-estimator is available. In the supplemental article [36], we describe more general DR theory, and illustrate this using the Cox model with a partially observed

covariate. The usual full-data estimator for the Cox model is the solution to partial-likelihood estimating equations, which do not take the form $\sum_{i=1}^{n} \boldsymbol{u}(\boldsymbol{Z}_i; \hat{\boldsymbol{\beta}}) = \boldsymbol{0}$.

The AIPW estimator of Section 2 has close connections to sample survey estimators that pre-date the work of Robins et al. [28], and to DR empirical likelihood (EL) and generalised EL estimators. In the supplemental article [36], we describe these connections and provide an introduction to DR EL estimators.

In missing-data problems, DR estimators require correct specification of either a model for the missingness process (given the full data) or a model for (some functional of) the outcome distribution (given the missing data patterns). When the data are non-monotone missing, plausible models for the missingness process can be difficult to construct. This has hindered the development of DR estimators in such settings [27]. The development of DR estimators for non-monotone missing data constitutes one of the primary open problems in this domain.

The construction of DR estimators for MNAR data is complicated by the lack of factorisation of the likelihood, which makes it difficult to describe the model for the missingness process (given the full data) and the model for (some functional of) the outcome distribution (given the missing data patterns) using variation-independent parameters. Such variation-independent parameterisation is needed to ensure that consistent estimators of the missingness probabilities can be obtained even when the outcome model is misspecified, and vice versa. Nevertheless, some progress has been made. A common approach uses a 'tilt' function (e.g. [29]). A simple application of this approach to the missing outcome problem of Section 2 would assume that $P(R = 1 \mid \boldsymbol{W}, Y) = \text{expit}\{\omega Y + a(\boldsymbol{W})\}$, where $a(\boldsymbol{W})$ is some function of $\boldsymbol{W}$ and ω is a known parameter (here ωY is the 'tilt' function). This implies that $f(y \mid \boldsymbol{W}, R = 0) = f(y \mid \boldsymbol{W}, R = 1) \exp(-\omega Y) c(\boldsymbol{W})$, where $c(\boldsymbol{W})$ is a normalising constant. The DR estimator of β is consistent if either a model $a(\boldsymbol{W}; \boldsymbol{\alpha})$ for $a(\boldsymbol{W})$ or a model $b(\boldsymbol{W}; \boldsymbol{\gamma})$ for $f(y \mid \boldsymbol{W}, R = 1)$ is correctly specified.

Finally, although in Section 5 we considered the implications of using variable (or model) selection strategies for the missingness and/or imputation models, we did not discuss how such selection is best done. Just as the choice of estimators of the nuisance parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$) can have a major impact on the performance of the DR estimator when at least one of these models is misspecified, also the choice of selection strategy can be extremely influential. This is well known when instrumental variables are observed, i.e. variables that are predictive of missingness, but not of the partially observed variables themselves [4]. The selection of such variables in the missingness model can cause a major loss of efficiency, and can moreover drastically amplify biases, e.g. due to model misspecification.

The development of variable selection strategies that prevent selection of instrumental variables in the missingness model has been an area of vigorous recent research [43, 47]. One such approach is the 'collaborative TMLE' method [43]. In the context of the missing outcome problem of Section 2, this method selects, from a given number of TMLEs for a nested sequence of models for $\pi(\boldsymbol{W})$, the one which minimises a penalised log-likelihood criterion, e.g. the sum of the squared residuals from the fitted model for $E(Y \mid \boldsymbol{W})$ plus the mean-squared error of the

estimator of β estimated by cross-validation. Because selecting instrumental variables inflates the mean-squared error of the estimator of β without changing the sum of the squared residuals, such variables are unlikely to be selected. While targeted variable selection strategies like the above tend to bring major efficiency improvements relative to routine strategies, a concern is that all of them (directly or indirectly) involve jointly modelling the missingness process and the conditional distribution of partially observed variables. As such, they risk giving up on the DR property, since misspecification of one of these two models may then result in inconsistent estimation of the other model, even when it is correctly specified.

References

[1] H Bang and JM Robins. Doubly robust estimation in missing data and causal inference models. *Biometrics*, 61:962–972, 2005.

[2] A Belloni and V Chernozhukov. l_1-penalised quantile regression in high-dimensional sparse models. *Annals of Statistics*, 39:82–130, 2011.

[3] A Belloni, V Chernozhukov, and C Hansen. Lasso methods for Gaussian instrumental variables models. ArXiv, 2016.

[4] MA Brookhart and MJ Van der Laan. A semi-parametric model selection criterion with applications to the marginal structural model. *Computational Statistics and Data Analysis*, 50:475–498, 2006.

[5] W Cao, AA Tsiatis, and M Davidian. Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. *Biometrika*, 96:723–734, 2009.

[6] CM Cassel, CE Sarndal, and JH Wretman. Some results on generalized difference estimation and generalized regression estimation for finite populations. *Biometrika*, 63:615–620, 1976.

[7] G Cheng, Z Yu, and JZ Huang. The cluster bootstrap consistency in generalized estimating equations. *Journal of Multivariate Analysis*, 115:33–47, 2013.

[8] V Chernozhukov, JC Escanciano, H Ichimura, and WK Newey. Locally robust semiparametric estimation. ArXiv, 2016.

[9] MH Farrell. Robust inference on average treatment effects with possibly more covariates than observations. *Journal of Econometrics*, 189:1–23, 2015.

[10] S Gruber and MJ van der Laan. A targeted maximum likelihood estimator of a causal effect on a bounded continuous outcome. *International Journal of Biostatistics*, 6, 2010.

[11] DG Horvitz and DJ Thompson. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, 47:663–68, 1952.

[12] JDY Kang and JL Schafer. Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. *Statistical Science*, 22:523–539, 2007.

[13] H Leeb and BM Pötscher. Model selection and inference: Facts and fiction. *Econometric Theory*, 21:21–59, 2005.

[14] H Leeb and BM Pötscher. Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. *Econometric Theory*, 22:69–97, 2006.

[15] K-Y Liang and SL Zeger. Longitudinal data analysis using generalised linear models. *Biometrika*, 73:13–22, 1986.

[16] R Little and H An. Robust likelihood-based analysis of multivariate data with missing values. *Statistica Sinica*, 14:949–968, 2004.

[17] RJA Little and DB Rubin. *Statistical Analysis With Missing Data*. Wiley, New Jersey, 2002.

[18] Q Long, X Zhang, and BA Johnson. Robust estimation of area under ROC curve using auxiliary variables in the presence of missing biomarker values. *Biometrics*, 67:559–567, 2011.

[19] X-L Meng. Multiple-imputation inferences with uncongenial sources of input. *Statistical Science*, 9:538–573, 1994.

[20] WK Newey, F Hsieh, and JM Robins. Twicing kernels and a small bias property of semiparametric estimators. *Econometrica*, 72:947–962, 2004.

[21] MC Paik. The generalized estimating equations approach when data are not missing completely at random. *Journal of the American Statistical Association*, 92:1320–1329, 1997.

[22] KE Porter, S Gruber, MJ van der Laan, and JS Sekhon. The relative performance of targeted maximum likelihood estimators. *International Journal of Biostatistics*, 7, 2011.

[23] L Qi, CY Wang, and RL Prentice. Weighted estimators for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*, 100:1250–1263, 2005.

[24] J Robins and A Rotnitzky. Discussion on the paper by Firth and Bennett. *Journal of the Royal Statistical Society, Series B*, 60:51–52, 1998.

[25] J Robins, M Sued, Q Lei-Gomez, and A Rotnitzky. Comment: Performance of double-robust estimators when "inverse probability" weights are highly variable. *Statistical Science*, 22:544–559, 2007.

[26] JM Robins. Robust estimation in sequentially ignorable missing data and causal inference models. In *Proceedings of the American Statistical Association Section on Bayesian Statistical Science 1999*, pages 6–10, 2000.

[27] JM Robins and RD Gill. Non-response models for the analysis of non-monotone ignorable missing data. *Statistics in Medicine*, 16:39–56, 1997.

[28] JM Robins, A Rotnitzky, and LP Zhao. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89:846–866, 1994.

[29] A Rotnitzky, D Faraggi, and E Schisterman. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, 101:1276–1288, 2006.

[30] A Rotnitzky, QH Lei, M Sued, and JM Robins. Improved double-robust estimation in missing data and causal inference models. *Biometrika*, 99:439–456, 2012.

[31] A Rotnitzky and S Vansteelandt. Double-robust methods. In G Molenberghs, G Fitzmaurice, MG Kenward, A Tsiatis, and G Verbeke, editors, *Handbook of Missing Data Methodology*, chapter 9, pages 185–212. Chapman & Hall/CRC Press, 2014.

[32] DO Scharfstein, A Rotnitzky, and JM Robins. Adjusting for nonignorable drop-out using semiparametric nonresponse models: Rejoinder. *Journal of the American Statistical Association*, 94:11335–1146, 1999.

[33] ME Schnitzer, JJ Lok, and RJ Bosch. Double robust and efficient estimation of a prognostic model for events in the presence of dependent censoring. *Biostatistics*, 17:165–177, 2016.

[34] S Seaman and A Copas. Doubly robust generalized estimating equations for longitudinal data. *Statistics in Medicine*, 28:937–955, 2009.

[35] SR Seaman, J Galati, D Jackson, and Carlin J. What is meant by 'missing at random'? *Statistical Science*, 28:257–268, 2013.

[36] SR Seaman and S Vansteelandt. Supplement to "Introduction to double robust methods for incomplete data". doi:?, 2018.

[37] Z Tan. A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association*, 101:1619–1637, 2006.

[38] Z Tan. Comment: Improved local efficiency and double robustness. *International Journal of Biostatistics*, 4, 2008.

[39] Z Tan. Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika*, 97:661–682, 2010.

[40] AA Tsiatis. *Semiparametric Theory and Missing Data*. Springer, New York, 2006.

[41] AA Tsiatis and M Davidian. Missing data methods: A semi-parametric perspective. In G Molenberghs, G Fitzmaurice, MG Kenward, A Tsiatis, and G Verbeke, editors, *Handbook of Missing Data Methodology*, chapter 8. Chapman & Hall/CRC Press, 2014.

[42] AA Tsiatis, M Davidian, and W Cao. Improved double-robust estimation when data are monotone coursened, with application to longitudinal studies with dropout. *Biometrics*, 67:536–545, 2011.

[43] MJ van der Laan and S Gruber. Collaborative double robust targeted maximum likelihood estimation. *International Journal of Biostatistics*, 6, 2010.

[44] MJ van der Laan and DB Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics*, 2, 2006.

[45] S Vansteelandt, J Carpenter, and MG Kenward. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. *Methodology*, 6:37–48, 2015.

[46] K Vermeulen and S Vansteelandt. Bias-reduced doubly robust estimation. *Journal of the American Statistical Association*, 110:1024–1036, 2015.

[47] A Wilson and BJ Reich. Confounder selection via penalized credible regions. *Biometrics*, 70:852–861, 2014.

[48] KE Wirth, EJ Tchetgen Tchetgen, and M Murray. Adjustment for missing data in complex surveys using doubly robust estimation. *Epidemiology*, 21:863–871, 2010.