View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by LSHTM Research Online

Improving the robustness and efficiency of covariate adjusted linear instrumental variable estimators

STIJN VANSTEELANDT

Department of Applied Mathematics, Computer Sciences and Statistics, Ghent University
and Centre for Statistical Methodology, London School of Hygiene and Tropical Medicine

and VANESSA DIDELEZ

Leibniz Institute for Prevention Research and Epidemiology - BIPS, and Department of Mathematics, University of Bremen

Two-stage least squares (TSLS) estimators and variants thereof are widely used to infer the effect of an exposure on an outcome using instrumental variables (IVs). TSLS estimators enjoy greater robustness to model misspecification than other two-stage estimators, but can be inefficient when the exposure is non-linearly related to the IV (or covariates). Locally efficient double-robust estimators overcome this concern. These make use of a possibly non-linear model for the exposure to increase efficiency, but remain consistent when that model is misspecified, so long as either a model for the IV or for the outcome model is correctly specified. However, their finite sample performance can be poor when the models for the IV, exposure and/or outcome are misspecified. We therefore develop double-robust procedures with improved efficiency and robustness properties under misspecification of some or even all working models. Simulation studies and a data analysis demonstrate remarkable improvements.

Key-words: bias; confounding; double-robustness; instrumental variable; model misspecification; semi-parametric efficiency.

Running title: Robust and efficient IV estimators

Funding: Fund for Scientific Research, Belgium; grant number G016116N

1 Introduction

An enormous body of research has developed in the econometrics and biostatistics literatures on how to assess the causal effect of an exposure X on an outcome Y in the presence of confounding by unobserved variables U, when a *vector of instrumental variables* Z (IVs) is available (see e.g. Bowden and Turkington, 1985; Robins, 1994; Angrist et al., 1996; Greenland, 2000; Wooldridge, 2002; Hernán and Robins, 2006; Didelez and Sheehan, 2007). It is therefore not surprising that a variety of competing approaches have been put forward. A simple and popular method is two-stage least squares (TSLS) estimation where, in the first stage, the endogenous variables (e.g. exposure as well as interactions between exposure and covariates) are predicted based on an ordinary least squares regression of the exposure on the IVs and the covariates; in the second stage, the outcome is regressed on the predicted exposure and covariates via ordinary least squares regression, and the exposure coefficient is taken as the final IV estimator of the desired causal effect. The simplicity of this approach has encouraged the development of other two-stage estimators, which are obtained along the same lines, but employ possibly non-linear regressions in the first or second stage (see e.g. Mullahy, 1997; or the review in Didelez et al., 2010).

The TSLS estimator has attractive properties that justify its widespread use. We provide an extensive review in Appendix S1 of the Supplemental Materials, along with extensions to more general two-stage estimators. A concise summary follows. The TSLS

estimator is a consistent estimator of the exposure effect on the outcome as soon as the second stage model is correctly specified, even when the first stage model is misspecified (Robins, 2000; Wooldridge, 2002, Theorem 5.1). Consistency is maintained, even when the outcome's dependence on covariates is misspecified in the second stage model, so long as the IVs have an expectation that is linear in those covariates (and thus in particular when the IVs are independent of covariates). Finally, the TSLS estimator of a constant linear exposure effect (i.e., that does not depend on covariates) is semi-parametric efficient under the model defined by additive effects of exposure and covariates on the outcome, and additive effects of instrument and covariates on the exposure, when a homoscedasticity assumption is satisfied (see Section 1.3 of the Supplemental Materials for specific detail).

In spite of this, there are important limitations. The TSLS estimator can be greatly inefficient under misspecification of the first and/or second stage model, to the extent that it may fail to be $\sqrt{n}$-consistent, even when the semi-parametric variance bound is finite. Moreover, the TSLS estimator's robustness against misspecification of the covariate effects on the outcome is lost when the IV depends non-linearly on covariates, as is for instance likely the case when the IV is dichotomous. Locally efficient double-robust IV-estimators (Robins, 1994; Okui et al., 2012) overcome these concerns. They have additional robustness against model misspecification: they are consistent if either a model for the main effect of covariates on the outcome or a model for the distribution of the IV, given covariates, is correctly specified, but not necessarily both (Okui et al., 2012). Moreover, since also the TSLS estimator is a double-robust IV estimator when the IV is linear in the covariates, locally efficient double-robust IV-estimators are at least as efficient as the TSLS estimator in that case, when models for the IV, exposure and outcome are correctly specified.

In simulation studies later in Section 6, the locally efficient double-robust IV estimator

is observed to outperform the TSLS estimator when based on correctly specified exposure and outcome models, but to perform sometimes much worse otherwise, making it unfit for general purpose use. In view of this, we develop two adaptive estimation procedures. The first makes use of empirical efficiency maximisation (Rubin and van der Laan, 2008) which is designed to maximise precision even under misspecification of the exposure and outcome model, and may result in drastic efficiency gains so long as the model for the distribution of the IV, given covariates, is correctly specified. The second makes use of bias-reduced double-robust estimation (Vermeulen and Vansteelandt, 2015), which is designed to prevent bias amplification under additional local misspecification of the IV distribution. Numerical results confirm the bias reduction and moreover demonstrate favourable performance regarding efficiency.

2 Linear instrumental variable models

Let Z be a vector of IVs for the effect of a scalar exposure X on a scalar outcome Y, conditional on a vector of observed covariates C. The literature provides different versions of the defining properties for Z to be a valid IV, e.g. in econometrics these are often stated in the context of linear structural equation models, while in biostatistics semiparametric structural models and potential outcomes are preferred. As we aim to investigate properties of related estimators from different such traditions, we briefly address how the assumptions relate to each other (for more details see e.g. Hernan and Robins (2006) and Didelez et al. (2010)).

In analogy to Didelez and Sheehan (2007) but extending the definition to account for covariates (see also Pearl, 2009, p.248), we formalise this by the following assumptions. Let U be a (set of) latent variable(s) such that (U, C) would be sufficient to control for confounding of the effect of X on Y were U observable; this formally means that

$Y_x \perp\!\!\!\perp X \mid (U, C)$ (Robins, 1994), with Y_x denoting the potential outcome that would be observed when setting X to x. We then have the IV assumptions that

(a) Z is associated with X conditional on C,

(b) $Z \perp\!\!\!\perp Y \mid (X, U, C)$ and

(c) $Z \perp\!\!\!\perp U \mid C$.

This formalisation of an IV is close to, but allows for greater flexibility than that in the econometric literature on IVs in the context of linear structural equation models (Wooldridge, 2002), where assumptions are usually in terms of no correlation instead of independence. In the causal inference literature, conditions (b) and (c) are often alternatively formalised as (and can be shown to imply) the assumption that (Robins, 1994) for all x

(b') $Y_x \perp\!\!\!\perp Z \mid C$.

The latter formulation avoids explicit reference to any specific unobserved confounders U.

We start by briefly addressing the relationship of different formulations of linear IV models, i.e. models where the target causal parameter relates to the linear effect of X on the expected outcome. Ultimately, we will assume the linear or additive structural mean model (Robins, 1994) given below in equation (3), but point out how this is implied by other typical models. Consider first the following assumption on the conditional mean of the outcome:

$$E(Y \mid X, Z, U, C) = \omega(C, U) + m(C; \psi^*)X. \tag{1}$$

Here, $\omega(C, U)$ is an unknown (i.e., unspecified) function of measured and unmeasured covariates. The term $m(C; \psi)$ is a known function of observed covariates, smooth in ψ, and ψ^* is an unknown finite-dimensional parameter, e.g. $m(C; \psi) = \psi$ or $m(C; \psi) = \psi^T C$,

where with a slight abuse of notation, the vector C includes 1 to allow for a main effect. When $m(C;\psi)$ is parameterised such that $m(C;\psi) = 0$ when $\psi = 0$, as in the previous examples, we have that $m(C;\psi^*)$ captures the *linear* effect of X, which, crucially, does not depend on U. In particular,

$$\begin{aligned} m(C;\psi^*)X &= E\left(Y \mid X, Z, U, C\right) - E\left(Y \mid X = 0, Z, U, C\right) \\ &= E\left(Y \mid X, Z, U, C\right) - E\left(Y_0 \mid X = 0, Z, U, C\right) \\ &= E\left(Y \mid X, Z, U, C\right) - E\left(Y_0 \mid X, Z, U, C\right) \\ &= E\left(Y \mid X, Z, C\right) - E\left(Y_0 \mid X, Z, C\right). \end{aligned} \tag{2}$$

This encodes the additive effect on the outcome of setting the exposure to zero in a subgroup of individuals with exposure X, IV Z and covariates C, which we consider the exposure effect of interest. In the above derivation, the second equality follows by assumption (b) and the consistency assumption that $Y = Y_0$ in subjects with $X = 0$, the third from the fact that U and C are sufficient to control for confounding of the effect of X on Y, and the fourth by averaging (conditional on X, Z and C) and the fact that the left-hand side does not involve U. Note that the linear structural equation version implies the above (1) and hence (2) but assumes more restrictively that $\omega(C, U)$ equals $\beta^{*T}C + U$ and $m(C;\psi) = \psi$ or $m(C;\psi) = \psi^T C$.

The model defined by restriction (2), i.e.

$$E(Y \mid X, Z, C) = E(Y_0 \mid X, Z, C) + m(C;\psi^*)X, \tag{3}$$

is called a linear or additive structural mean model (Robins, 1994). Together with the IV assumptions (a) and (b') it can be regarded as the substantive model of interest, as it merely parameterizes the exposure effect of interest. Note that $Y_0 \perp\!\!\!\perp Z \mid C$ as special case of (b') is sufficient. That the exposure effect is not modified by Z (or equivalently,

that Z does not appear on the right-hand side of (1), but is included on the left hand side) is known as 'no effect modification' by Z (Hernan and Robins, 2006; Clarke and Windmeier, 2010). Although structural mean models can be formulated that do not make this assumption, it is imposed here to enable identification of ψ^*. While it can be motivated by the additivity in (1), it is often made in its own right avoiding explicit reference to U and hence allowing greater generality.

Model (1) along with assumptions (a), (b) and (c) differs from model (3) along with assumptions (a) and (b') in its assumptions on unobservables, but both models impose the same restrictions on the observed data law (see Appendix S2 of the Supplemental Materials), namely that

$$E(Y - m(C;\psi^*)X \mid Z, C) = E(Y - m(C;\psi^*)X \mid C), \tag{4}$$

i.e. the left-hand side does not depend on Z, given C. We will therefore focus on inference under model $\mathcal{M}$ defined by (4) throughout, supposing that a sample of i.i.d. data (Y_i, X_i, Z_i, C_i) for $i = 1, ..., n$ is available.

3 Two-stage estimation

Two-stage approaches for fitting model $\mathcal{M}$ are based on rewriting it as:

$$E\left(Y \mid Z, C\right) = \omega(C) + m(C;\psi^*)E(X \mid Z, C), \tag{5}$$

for $\omega(C) \equiv E(Y - m(C;\psi^*)X \mid C)$. When C is high-dimensional (e.g. continuous or discrete with several components), the above cannot be fitted non-parametrically and additional modelling assumptions are needed to obtain estimators of ψ^* with adequate performance in moderate sample sizes. Equation (5) suggests postulating two additional models, one for $E(X \mid Z, C)$ and one for $\omega(C)$, and thereby lays the basis of two-stage estimation procedures.

In the first stage, a parametric model $\mathcal{A}_x$ is postulated for the exposure, i.e.

$$E\left(X \mid Z, C\right) = m_x(Z, C; \alpha^*), \tag{6}$$

where $m_x(Z, C; \alpha)$ is a known function of instruments and covariates, smooth in α, and α^* is an unknown finite-dimensional parameter. An obvious choice would be a linear or logistic regression model (e.g., $m_x(Z, C; \alpha) = \text{expit}(\alpha_z^T Z + \alpha_c^T C)$). The second stage model supplements the structural model $\mathcal{M}$ with a parametric model $\mathcal{A}_y$ for the main effects of covariates on the outcome:

$$\omega(C) = m_y(C; \beta^*), \tag{7}$$

where $m_y(C; \beta)$ is a known function of covariates, smooth in β and β^* is an unknown finite-dimensional parameter. A general two-stage procedure is now obtained by fixing α^* at some estimate $\hat{\alpha}$ obtained from fitting model (6) and then fitting model (5) with $E(X \mid Z, C)$ substituted by $m_x(Z, C; \hat{\alpha})$, using standard regression techniques at each stage.

When $\mathcal{A}_x$ and $\mathcal{A}_y$ are chosen to be linear with the same covariates, then we use the notation $\mathcal{A}_x^{\text{lin}}, \mathcal{A}_y^{\text{lin}}$ to make this explicit.

3.1 Two-stage least squares (TSLS)

Among the above two-stage methods, TSLS takes a prominent place (Wooldridge, 2002). The principle of TSLS is that all 'endogenous' exposures (those that are confounded, i.e. dependent on U) are replaced by their linear projections on all 'exogenous' variables (these are the IVs, covariates, and possible other unconfounded exposures in the outcome model). As the name suggests, TSLS is equivalent to explicit two-stage estimation because the linear projections are equivalent to fitting a linear first stage model with ordinary least squares, and these can then be plugged into the second stage model, again fitted

by least-squares (for details see Wooldridge, 2002; Section 5). For the equivalence it is, however, important to use the *implied* first stage model, i.e. a linear model for X (and other endogenous variables) given *all* exogenous variables as determined by the choice of IVs and $\mathcal{A}_y$. A formal definition of TSLS is given in Appendix S1 of the Supplemental Materials.

We illustrate this aspect of TSLS with an example. Consider the case where $C = (1, V)^T$ with V a scalar, $m(C;\psi)X = \psi_1 X + \psi_2 XV$ and $m_y(C;\beta) = \beta_0 + \beta_1 V + \beta_2 V^2$. There are two 'endogenous' variables, X and XV, as these both depend on U. For identification it is necessary that there are at least as many instruments as endogenous variables; hence, two instruments are needed, which could be Z and ZV. The linear projections would be of X and XV each on all of Z, ZV, V and V^2. It follows that the implied first stage models are $E(X \mid Z, C) = \alpha_0 + \alpha_1 Z + \alpha_2 ZV + \alpha_3 V + \alpha_4 V^2$ and $E(XV \mid Z, C) = \alpha_0' + \alpha_1' Z + \alpha_2' ZV + \alpha_3' V + \alpha_4' V^2$ where it is assumed that the coefficients of the instruments in the projections are non-zero (more precisely, that the matrix with first row α_1, α_2 and second row α_1', α_2' has full rank); the latter is a more specific version of assumption (a).

3.2 TSLS versus general two-stage methods

When model $\mathcal{M} \cap \mathcal{A}_x \cap \mathcal{A}_y$ is correctly specified, general two-stage IV estimators are consistent (see Appendix S1) but not necessarily efficient, in part because they are based on *separately* fitting the exposure model and the outcome model. It is therefore somewhat surprising that the TSLS estimator of ψ^* is semi-parametric (locally) efficient in linear exposure and outcome models, i.e. under $\mathcal{M} \cap \mathcal{A}_x^{\text{lin}} \cap \mathcal{A}_y^{\text{lin}}$, when $m(C;\psi) = \psi$ and a homoscedasticity assumption is satisfied (for details see Section 1.3 of the Supplemental Materials). General two-stage estimators (including TSLS estimators) can however be

inefficient when the exposure effect depends on covariates (i.e. when $m(C;\psi^*) = \psi^{*T}C$), even when the exposure and outcome obey simple linear models. For TSLS estimators, this can be intuitively seen by considering the following example with $m(C;\psi^*) = \psi_0^* + \psi_1^* V$ for $C = (1, V)^T$. Then TSLS is based on separate least squares regressions of X and XV on Z and V, without taking into account that the model for X implies the model for XV, and without considering that the postulated models may be incompatible (e.g., even when the model for X includes a main effect of V, the model for XV may not allow for a main effect of V^2). Moreover, two-stage estimators (including TSLS estimators) are generally inefficient when the true exposure relation is nonlinear in Z or C (e.g. because it includes an interaction between Z and components of C, or because it is of the logistic form), or when the outcome is dichotomous so that there is heteroscedasticity. In particular, it may happen under certain data laws that the TSLS estimator does not exist (more precisely, is not $\sqrt{n}$-consistent), even though other two-stage estimators with small variance exist. This is for instance the case when $E(X \mid Z, C) = Z - ZV$ for a scalar variate $V \in C$ which takes the values 0 and 1 with probability 1/2, independently of Z, and when furthermore $m(C;\psi) = \psi$ and $m_y(C;\beta) = \beta_0 + \beta_1 V$. In that case, the implied first stage model would ignore the interaction between Z and V and thus result in the population linear projection of X on (Z, C) equalling 0, thereby violating the necessary rank condition for TSLS estimators. In those cases, a two-stage estimator based on a first stage model that includes main effects of Z, V and their interaction, is indicated.

Besides being locally efficient, TSLS estimators also enjoy greater robustness compared to more general two-stage estimators. The former are consistent for the exposure effects on the outcome as soon as the second stage model is correctly specified, even when the first stage model is misspecified (Robins, 2000; Wooldridge, 2002, Theorem 5.1). Consistency is retained even when the outcome's dependence on covariates is misspecified in

the second stage model, so long as the IVs have an expectation that is linear in those covariates (and thus in particular when the IVs are independent of covariates) and $m(C;\psi)$ is linear in C (Okui et al., 2012). Thus, when $E(Z \mid C)$ is linear in V and V^2 (with $C = (1, V)^T$), then the TSLS estimator under model $m(C;\psi) = \psi$ will be robust against outcome model misspecification when the outcome model includes the term V^2 (regardless of whether it is associated with the outcome). We show in Appendix S2 (Proposition 5) of the Supplemental Materials that this property in fact still holds for the slightly more general case where the first and second stage are fitted by least squares, but the first stage model used is not the one implied by the second stage. Unfortunately, this robustness of the TSLS estimator against misspecification of the second stage model does not extend to general IVs, e.g. dichotomous IVs that obey a logistic regression model with main covariate effect C, nor to general two-stage estimators that involve nonlinear exposure models or effect heterogeneity (i.e. $m(C;\psi^*)$ depending on C).

4 Double-robust estimation and TSLS

A general approach to robustness against misspecification of the second stage model is double-robust estimation (Robins, 1994; Okui et al., 2012); we review this here with focus on its relation to covariate adjusted TSLS. Double-robust estimation makes use of an additional parametric model for the conditional IV distribution given the covariates, $\mathcal{A}_z$ defined by

$$f(Z \mid C) = f(Z \mid C;\gamma^*),$$

where $f(Z \mid C;\gamma)$ is a known density function, smooth in γ, and γ^* is an unknown finite-dimensional parameter, which we will estimate by maximum likelihood (however, see Section 5.2 for an alternative strategy). For instance, when Z is binary, we may assume that $P(Z = 1 \mid C) = \text{expit}(\gamma^{*T} C)$ and use standard logistic regression to estimate γ^*.

Let $\hat{\gamma}$ be the corresponding maximum likelihood estimator. Further, let $\hat{\beta}$ be a consistent estimator of β^* in $\mathcal{A}_y$ as obtained in the previous section. Then a consistent asymptotically normal (CAN) estimator of ψ^* under model $\mathcal{M} \cap (\mathcal{A}_y \cup \mathcal{A}_z)$ can be obtained by solving

$$0 = \sum_{i=1}^{n} \left[e(Z_i, C_i) - E\left\{ e(Z_i, C_i) \mid C_i; \hat{\gamma} \right\} \right] \left\{ Y_i - m_y(C_i; \hat{\beta}) - m(C_i; \psi) X_i \right\}, \tag{8}$$

for some conformable (i.e., of appropriate dimension) vector function $e(Z, C)$. Because the solution to (8) is a CAN estimator of ψ^* when either working model $\mathcal{A}_z$ or $\mathcal{A}_y$ holds, in addition to the linear IV model $\mathcal{M}$, it is called double-robust (Robins and Rotnitzky, 2001). Note that it follows from the remarks at the end of Section 3.2 that TSLS is double-robust for particular choices of $\mathcal{A}_y$ and $\mathcal{A}_z$.

Double-robust estimators are especially attractive in studies where the law of Z given C is (partially) known as this may guarantee robustness against misspecification of $\mathcal{A}_y$, while typical two-stage estimators fail to exploit this. Such knowledge, leading to correct specification of $\mathcal{A}_z$, is for instance given in randomized experiments where Z denotes randomization, or in Mendelian randomization studies where the genetic instrument is often known to be independent of covariates C, in which case $E\{e(Z, c) \mid C = c\}$ can be consistently estimated as $n^{-1} \sum_{i=1}^{n} e(Z_i, c)$.

In fact, the special case where it is known that $Z \perp\!\!\!\perp C$ is of particular interest; if additionally we have $m(C; \psi) = \psi$, one has the choice of whether to adjust for C at all. We therefore consider the questions of whether it is worthwhile, i.e. more efficient, to include covariates at all when there is the choice. To address this, we first recall how a a semi-parametric (locally) efficient estimator of ψ^* under model $\mathcal{M} \cap \mathcal{A}_z$ is constructed. It follows from Robins (1994) (see also Okui et al., 2012) that such estimator is obtained by choosing $e(Z, C)$ in (8) equal to

$$e_{\text{opt}}(Z, C) = \sigma^{-2}(Z, C) \frac{\partial m(C; \psi^*)}{\partial \psi} \left[E(X \mid Z, C) - \frac{E\left\{ \sigma^{-2}(Z, C) E(X \mid Z, C) \mid C \right\}}{E\left\{ \sigma^{-2}(Z, C) \mid C \right\}} \right] \tag{9}$$

with $\sigma^2(Z,C) \equiv \text{Var}\{Y - m(C;\psi^*)X \mid Z,C\}$. Since model $\mathcal{M} \cap (\mathcal{A}_z \cup \mathcal{A}_y)$ is less restrictive, this is also delivering the (locally) efficient estimator of ψ^* in model $\mathcal{M} \cap (\mathcal{A}_z \cup \mathcal{A}_y)$. For instance, assuming that $E(X \mid Z,C) = \alpha_1^{*T}C + \alpha_2^{*T}ZC$ for scalar Z and C, $m_y(C;\beta) = \beta^T C$ and $\sigma^2(Z,C) = \sigma^2$ for unknown parameters α_1^*, α_2^* and β^*, we have

$$e_{\text{opt}}(Z,C) = \sigma^{-2}\alpha_2^{*T}C\{Z - E(Z \mid C)\}. \tag{10}$$

A locally efficient estimator may now be obtained by substituting α_2^* by the ordinary least squares estimator in the above expression, setting σ^2 to 1 (as it is just a proportionality constant), and next solving (8) for the resulting choice of $e(Z,C) = e_{\text{opt}}(Z,C)$. These expressions suggest a way to optimally include covariates and, in a similar vein, to optimally combine multiple instruments (see e.g. Bowden and Vansteelandt, 2011). Since $e_{\text{opt}}(Z,C)$ as well as $m_y(C) = E(Y - \psi^*X \mid C)$ are generally dependent on the covariate data C, this suggests in particular that the covariate-adjusted analysis will be at least as efficient in large samples as the unadjusted analysis, provided the working models for $E(Y - \psi^*X \mid C)$, $E(X \mid Z,C)$ and $\sigma^2(Z,C)$ are correctly specified. While an efficiency gain is not generally guaranteed when these working models are misspecified, the following proposition demonstrates that covariate adjustment is guaranteed not to increase the asymptotic variance of the TSLS estimator under certain conditions.

Proposition 1 Efficiency of covariate adjusted TSLS estimators

When $Z \perp\!\!\!\perp C$ and $m(C;\psi) = \psi$ under model $\mathcal{M}$, if $Y - \psi^*X$ is conditionally independent of Z given C, then adjustment for C does not increase the asymptotic variance of the TSLS estimator of ψ^*; it decreases it when $Y - \psi^*X$ depends on C.

Proof: see Appendix S3 of the Supplemental Materials. □

Fisher and Goetghebeur (1999) also observed that covariate adjustment is typically beneficial in a linear IV context; however, their results are specific to the case of partial

compliance with full compliance in the control arm, where by design there is a corresponding interaction in the exposure model and where the IV model is specific to the treatment arm.

5 Improved double-robust estimation

Consistency of the double-robust estimator of ψ^* demands correct specification of either the outcome model $\mathcal{A}_y$ or the IV model $\mathcal{A}_z$; local efficiency demands correct specification of both these models, and additionally of models for the exposure distribution and conditional outcome variance. In practice all these models are typically somewhat misspecified. In Section 5.1, we therefore propose a strategy to guarantee efficiency within a subclass of double-robust estimators as soon as the IV model $\mathcal{A}_z$ is correctly specified. In Section 5.2, we propose strategies that aim to minimise locally the bias of the double-robust estimator when both the outcome model $\mathcal{A}_y$ and the IV model $\mathcal{A}_z$ are misspecified. Throughout these sections, results are confined to the main effect structural model $\mathcal{M}$ with $m(C;\psi) = \psi$.

5.1 Empirical efficiency maximisation

The semi-parametric efficient estimator of ψ^*, obtained by substituting the conditional expectations in (9) by estimates under parametric models, is not guaranteed to outperform simpler CAN estimators (e.g. obtained by solving (8) for $e(Z,C) = Z$ or by ignoring covariate information) under model misspecification, as we will see in the simulation study of Section 6. Okui et al. (2012) proposed regression double-robust estimators that have an asymptotic variance no larger than a given double-robust estimator, even under model misspecification. In this Section, we generalise their results with the potential for bigger efficiency gains in return. We will realise this by building on and extending the ideas

behind empirical efficiency maximisation, a procedure originally proposed by Rubin and van der Laan (2008) and Cao, Tsiatis and Davidian (2009) in the missing data literature. In this subsection, we assume that model $\mathcal{A}_z$ is correctly specified.

Let $\hat{\psi}(\alpha, \beta)$ be the double-robust estimator of ψ^* obtained by solving estimating equation (8) for a user-specified parameterisation $e(Z, C; \alpha)$ of $e(Z, C)$, evaluated at the given values α (and β indexing $m_y(C; \beta)$). This parameterisation may, but need not be guided by the form of the efficient index function given in (9). For instance, for a scalar Z, one may postulate that $e(Z, C)$ is of the form $\alpha^T CZ$ for some α. When the law of Z given C is known, then the asymptotic variance of $\hat{\psi}(\alpha, \beta)$ under model $\mathcal{M} \cap \mathcal{A}_z$ equals

$$\frac{\text{Var}\left(\left[e(Z, C; \alpha) - E\left\{e(Z, C; \alpha) \mid C\right\}\right]\left\{Y - m_y(C; \beta) - \psi^* X\right\}\right)}{nE\left(\left[e(Z, C; \alpha) - E\left\{e(Z, C; \alpha) \mid C\right\}\right] X\right)^2}. \tag{11}$$

Let $\tilde{\alpha}$ and $\tilde{\beta}$ be the values of α and β, respectively, that minimise the empirical analog of (11) with ψ^* substituted by a preliminary consistent estimator under model $\mathcal{M} \cap \mathcal{A}_z$, e.g. a double-robust estimator based on the choices $e(Z, C) = Z$ and model $\mathcal{A}_y^{\text{lin}}$. The proposition below then shows that $\hat{\psi}(\tilde{\alpha}, \tilde{\beta})$ is a double-robust estimator which is (asymptotically) at least as efficient as $\hat{\psi}(\alpha, \beta)$ for arbitrary α and β. Key properties that underlie the validity of the proposition are (a) that $\tilde{\beta}$ is CAN for β^* under model $\mathcal{A}_y$; and (b) that $\hat{\psi}(\tilde{\alpha}, \tilde{\beta})$ and $\hat{\psi}(\tilde{\alpha}^*, \tilde{\beta}^*)$ have the same asymptotic variance under model $\mathcal{M} \cap \mathcal{A}_z$, with $\tilde{\alpha}^*$ and $\tilde{\beta}^*$ being the probability limits of $\tilde{\alpha}$ and $\tilde{\beta}$ (provided $\tilde{\alpha}$ and $\tilde{\beta}$ converge at faster than $n^{1/4}$ rate).

Proposition 2 Efficiency within a subclass of double-robust estimators

Let $\tilde{\alpha}$ and $\tilde{\beta}$ minimise the empirical analog of (11). Then the estimator $\hat{\psi}(\tilde{\alpha}, \tilde{\beta})$ solving (8) is CAN under model $\mathcal{M} \cap (\mathcal{A}_y \cup \mathcal{A}_z)$.

Moreover, when the law of Z given C is known*, then we have that for all α and β*

$$\lim_{n\to\infty} \text{Var}\left[\sqrt{n}\left\{\hat{\psi}(\tilde{\alpha}, \tilde{\beta}) - \psi^*\right\}\right] \leq \lim_{n\to\infty} \text{Var}\left[\sqrt{n}\left\{\hat{\psi}(\alpha, \beta) - \psi^*\right\}\right].$$

Proof: see Appendix S3 of the Supplemental Materials. □

In Appendix S3 of the Supplemental Materials, we further discuss the case where the law of Z given C is known only up to a finite-dimensional parameter.

Consider for instance the choices $e(Z, C; \alpha) = \alpha^T CZ$ and $m_y(C; \beta) = \beta^T C$. Then by construction, $\hat{\psi}(\tilde{\alpha}, \tilde{\beta})$ is at least as efficient as the estimator obtained by solving (8) for the simple choices $e(Z, C) = Z$ and $m_y(C) = 0$, i.e. the estimator which ignores covariates. Hence, when $Z \perp\!\!\!\perp C$, then the resulting approach will deliver a covariate adjustment strategy that is guaranteed to be at least as efficient as an unadjusted analysis. More generally, efficiency is - by construction - always attained within the subclass of estimators allowed by varying α and β in the models for $e(Z, C)$ and $m_y(C)$. However, semi-parametric efficiency under model $\mathcal{M} \cap (\mathcal{A}_z \cup \mathcal{A}_y)$ is only attained when the efficient index function (9) happens to equal $e(Z, C; \alpha)$ for some α and when $E(Y - \psi^* X \mid C)$ equals $m_y(C; \beta)$ for some β.

Minimising the empirical analog of (11) can generally be done numerically, but in special cases also by suitably modified regression techniques. For example, we show in Appendix S3 of the Supplemental Materials that when $e(Z, C) = \alpha^T CZ$, then under certain assumptions minimising (11) w.r.t. α can be done by letting $\tilde{\alpha}$ be the ordinary least squares estimator of α in the regression model $E(X \mid Z, C) = \alpha^T C \{Z - E(Z \mid C)\}$. Minimising (11) w.r.t. β is possible by letting $\tilde{\beta}$ be the weighted least squares estimator of β in the regression model $E(Y - \psi^* X \mid C) = \beta^T C$ using weights $(\tilde{\alpha}^T C)^2 \{Z - E(Z \mid C)\}^2$. The above procedure needs some modification when the law of Z given C is unknown and the model $\mathcal{A}_y$ is (possibly) misspecified (see Appendix S3 of the Supplemental Materials for detail).

The regression double-robust estimator of Okui et al. (2012) may be viewed as a special case of the above proposal. It fixes α at some given value (which may not minimise

the asymptotic variance) and chooses $m_y(C;\beta) = \beta m_y(C)$ for some given $m_y(C)$.

5.2 Bias-reduced double-robust estimation

The efficiency results of Section 5.1 are especially attractive when model $\mathcal{A}_z$ is known to hold, as is the case in certain study designs. In other cases, bias becomes, arguably, a more dominant concern. Although there seems little hope that one can avoid bias in the estimation of ψ^* when both working models $\mathcal{A}_z$ and $\mathcal{A}_y$ are misspecified, Vermeulen and Vansteelandt (2015) found that for quite a general class of double-robust estimators, surprisingly, the nuisance parameters indexing $\mathcal{A}_z$ and $\mathcal{A}_y$ can be estimated so as to target bias reduction. Briefly, they note that the asymptotic bias (Stefanski and Boos, 2002) of an estimator for ψ^*, evaluated at fixed nuisance parameters β and γ, equals the expected value of its influence function $U(\psi^*,\beta,\gamma)$; for given α, this is here:

$$U(\psi,\beta,\gamma) = \frac{[e(Z,C;\alpha) - E\{e(Z,C;\alpha) \mid C;\gamma\}]\{Y - m_y(C;\beta) - \psi X\}}{E([e(Z,C;\alpha) - E\{e(Z,C;\alpha) \mid C;\gamma\}]X\}}.$$

Minimising the squared bias in the direction of β thus amounts to setting the gradient

$$2E\{U(\psi^*,\beta,\gamma)\}E\left\{\frac{\partial U}{\partial \beta}(\psi^*,\beta,\gamma)\right\}$$

to zero. Although the first component cannot generally be made zero without knowing aspects of the data-generating law, interestingly, the second component delivers an unbiased estimating function for γ (Vermeulen and Vansteelandt, 2015). This is so because, by the double-robustness, $U(\psi^*,\beta,\gamma)$ is mean zero for all β at γ^* when model $\mathcal{A}_z$ holds. The second component can thus be made zero empirically, by using it as a basis for estimation.

We will illustrate this for the case where the instrument Z is dichotomous with working model $P(Z=1 \mid C;\gamma) = \text{expit}(\gamma^T C)$ and where $m_y(C;\beta) = \beta^T C$. Further, let the index function $e(Z,C;\alpha)$ be of the form $Ze(C;\alpha)$ for some $e(C;\alpha)$ (as is the case for the efficient score for ψ^* under model $\mathcal{M} \cap (\mathcal{A}_z \cup \mathcal{A}_y)$ when $E(X \mid Z,C)$ is linear in Z and

$\text{Var}(Y \mid Z, C)$ does not depend on Z). Taking the gradient of $U(\psi, \beta, \gamma)$ with respect to β then results in estimating equations

$$0 \;=\; \sum_{i=1}^{n} \frac{\partial U_i(\gamma, \beta)}{\partial \beta} = \sum_{i=1}^{n} e(C_i; \alpha) \left\{ Z_i - P(Z_i = 1 \mid C_i; \gamma) \right\} C_i, \tag{12}$$

which are unbiased for γ. Since γ and β are of the same dimension, γ can thus be estimated as the solution to this equation. Solving equation (12) ensures that

$$\sum_{i=1}^{n} e(C_i; \alpha) \left\{ Z_i - P(Z_i = 1 \mid C_i; \gamma) \right\} m_y(C_i; \beta) = 0$$

so that the estimating equation for ψ reduces to

$$\begin{aligned} 0 \;&=\; \sum_{i=1}^{n} e(C_i; \alpha) \left\{ Z_i - P(Z_i = 1 \mid C_i; \gamma) \right\} \left\{ Y_i - m_y(C_i; \beta) - m(C_i; \psi^*) X_i \right\} \\ &=\; \sum_{i=1}^{n} e(C_i; \alpha) \left\{ Z_i - P(Z_i = 1 \mid C_i; \gamma) \right\} \left\{ Y_i - m(C_i; \psi^*) X_i \right\} \end{aligned}$$

which no longer involves β. The considered choice of estimator of γ thus overcomes the need to estimate β.

Solving (12) may not be straightforward for certain data sets. We therefore extend the logistic regression model for Z to

$$P(Z = 1 \mid C; \gamma) = \text{expit} \left\{ \gamma^T C + \theta^T C e(C; \alpha) \right\}.$$

This model contains the original working model $\mathcal{A}_z$ (corresponding to $\theta = 0$). Moreover, fitting this model using the default maximum likelihood procedure has the effect of making the identity (12) hold, as the latter corresponds with the score for the coefficient of $e(C; \alpha)C$. The resulting procedure will be referred to as BR_γ.

In Appendix S4 of the Supplemental Materials, we show that the above procedure reduces the order of the asymptotic bias of the double-robust estimator when model $\mathcal{A}_z$ is grossly misspecified and model $\mathcal{A}_y$ is locally misspecified (local in the sense of resulting in

a bias in $\hat{\beta}$ of the order $n^{-1/2}$). This is important because it prevents bias in the nuisance parameter estimators from propagating into the estimator of the target parameter. In particular, it suggests that the considered bias-reduced double-robust estimator will likely have little bias when model $\mathcal{A}_z$ is grossly misspecified, so long as model $\mathcal{A}_y$ is only mildly misspecified. Under gross misspecification of both working models, one can obviously not exclude that other nuisance parameter estimators happen to deliver less biased effect estimators under some data-generating mechanisms.

When using the procedure BR_γ, we continue to estimate α indexing $e(C;\alpha)$ as explained in Section 5.1. Although now, we no longer assume that model $\mathcal{A}_z$ is correctly specified, estimating α in this manner still has the effect of minimising the asymptotic variance of the double-robust estimator across all values of α. This is because the procedure BR_γ sets the gradient of the influence function w.r.t. β equal to zero, so that there is no need to account for the estimation of β in the calculation of the asymptotic variance (see Vermeulen and Vansteelandt, 2015; see also Appendix S4).

We also considered a related approach whereby we estimated γ using maximum likelihood and β by setting the gradient of the influence function $U(\gamma,\beta)$ with respect to γ to zero. This results in the following unbiased estimating equations for β:

$$0 \;=\; \sum_{i=1}^{n} \frac{\partial U_i(\psi^*,\gamma,\beta)}{\partial \gamma} = \sum_{i=1}^{n} e(C_i;\alpha)\left\{Y_i - m_y(C_i;\beta) - \psi^* X_i\right\}\Gamma_i \tag{13}$$

for

$$\begin{aligned}\Gamma_i \;=\;& \left\{Z_i - P(Z_i=1 \mid C_i;\gamma)\right\} E\left[e(C_i;\alpha)P(Z_i=1 \mid C_i;\gamma)P(Z_i=0 \mid C_i;\gamma)C_i X_i\right] \\ & -P(Z_i=1 \mid C_i;\gamma)P(Z_i=0 \mid C_i;\gamma)C_i E\left[e(C_i;\alpha)\left\{Z_i - P(Z_i=1 \mid C_i;\gamma)\right\}X_i\right]\end{aligned}$$

It can be verified that the effect of the factor Γ_i is to eliminate ψ^* from the estimating equation so that knowledge of the true ψ^* is not needed for estimating β. This approach

is designed to locally minimise the bias of the double-robust estimator in the direction of γ, at the maximum likelihood estimate $\hat{\gamma}$. In Appendix S4 of the Supplemental Materials, we show in particular that the above procedure, which will be referred to as BR_β, reduces the order of the asymptotic bias of the double-robust estimator when model $\mathcal{A}_y$ is grossly misspecified and model $\mathcal{A}_z$ is locally misspecified (local in the sense of resulting in a bias in $\hat{\gamma}$ of the order $n^{-1/2}$). To solve (13), we jointly fit an extended linear model for the outcome $Y - \psi^* X$ with covariates C and $e(C;\alpha)P(Z = 1 \mid C;\hat{\gamma})P(Z = 0 \mid C;\hat{\gamma})C$ using ordinary least squares (where, again, the choice of ψ^* does not affect results), and the (double-robust) estimating equation for ψ. This has the effect of making the identity (13) hold. Indeed, ordinary least squares estimation of the extended linear model ensures the following restrictions:

$$0 \;=\; \sum_{i=1}^{n} e(C_i;\alpha)P(Z_i = 1 \mid C_i;\hat{\gamma})P(Z_i = 0 \mid C_i;\hat{\gamma})C_i \left\{ Y_i - \psi^* X_i - \hat{\beta}^T C_i \right\},$$

which amounts to setting one component of (13) to zero; the remaining component is proportional to the (double-robust) estimating equation for ψ, which is made zero in the estimation process.

The original proposal of bias-reduced double-robust estimation (Vermeulen and Vansteelandt, 2015), which we refer to as BR, amounts to estimating γ and β by jointly solving (12) and (13). This seems preferable, in that it reduces the order of the asymptotic bias of the double-robust estimator when one working model is grossly misspecified and the other is locally misspecified, regardless of which. However, we did not pursue this approach in the numerical evaluations in Section 6 because of the difficulty in solving equations (12) and (13).

5.3 Standard errors

For all considered double-robust estimators $\hat{\psi}$, it follows by standard M-estimation arguments that the asymptotic variance can be straightforwardly estimated as 1 over n times the sample variance of the influence function

$$\left[\sum_{i=1}^{n}\left[e(Z_i,C_i)-E\left\{e(Z_i,C_i)\mid C_i;\hat{\gamma}\right\}\right]\left\{\partial m(C_i;\hat{\psi})/\partial\psi\right\}X_i\right]^{-1}$$
$$\times\left[e(Z_i,C_i)-E\left\{e(Z_i,C_i)\mid C_i;\hat{\gamma}\right\}\right]\left\{Y_i-m_y(C_i;\hat{\beta})-m(C_i;\hat{\psi})X_i\right\},$$

when both working models $\mathcal{A}_z$ and $\mathcal{A}_y$ are correctly specified. Under misspecification of at least one of these models, the above variance estimator must be corrected for the uncertainty in the nuisance parameter estimators as detailed in Vermeulen and Vansteelandt (2015), with the exception of the procedure BR. For the procedure BR_γ, the above variance calculation delivers a conservative estimator of the asymptotic variance of $\hat{\psi}$ when model $\mathcal{A}_y$ is misspecified (Rotnitzky, Li and Li, 2010), and an asymptotically unbiased estimator otherwise. The degree of conservatism of this procedure is influenced by the choice of $\hat{\beta}$, even though the bias-reduced double-robust estimator does not make use of an estimator of β. We recommend basing the calculation of the asymptotic variance on the ordinary least squares estimator of β in a regression of $Y-\hat{\psi}X$ on C, as we did in the simulation study of Section 6. Alternatively, robust sandwich standard errors can be calculated (Vermeulen and Vansteelandt, 2015), or the bootstrap can be used. For the procedure BR_β, the above variance calculation can be both liberal and conservative as a result of ignoring the uncertainty in $\hat{\beta}$. When this procedure is used, we therefore recommend the bootstrap.

6 Simulation study

We conducted a simulation experiment with $n = 500$ independent measurements on mutually independent and standard normal covariates U and V, Z dichotomous with $P(Z = 1 \mid V) = \text{expit}(-1+V/2+\lambda_z V^2/3)$, X normal with mean $Z+U+V-ZV+\lambda_x V^2$ and Y normal with mean $X - U - V + \lambda_y V^2$. Assuming a linear IV model with $m(C; \psi) = \psi$, we then evaluated the following estimators:

1. TSLS: the TSLS estimator using $(Z, VZ)^T$ as IV vector, based on a linear model for the exposure, involving main effects of V, Z and their interaction, and a linear model for the outcome involving main effects of V and the fitted value from the first stage regression. Including the VZ interaction in the first stage model ensures a fairer comparison with the subsequent estimators so that for all estimators misspecification of the exposure model is only due to omitting V^2.

2. Loc Eff: the locally efficient double-robust estimator (assuming homoscedasticity) based on a logistic model for Z with a main effect of V, a linear model for X with a main effect of Z and V and their interaction, and a linear outcome model (i.e., $m_y(C) = \beta^T C$ with $C = (1, V)^T$), all fitted using maximum likelihood.

3. Emp Eff: the locally efficient double-robust estimator using the same fitted model for Z as before, but using working models $e(Z, C) = \alpha^T CZ$ and $m_y(C) = \beta^T C$ fitted using empirical efficiency maximization (ignoring estimation of the model for Z, which is suboptimal when the outcome model is misspecified; see Appendix S3 in the Supplemental Materials).

4. BR_β, BR_γ: the double-robust estimator with α^* estimated using empirical efficiency maximization, but with either the outcome model or the IV model fitted using bias-

reduced estimation.

To obtain the estimators Loc Eff and Emp Eff, the TSLS estimator was used as a starting value; the obtained estimate was then updated a single time.

Table 1 and Figures 1 and 4 show the simulation results based on 1000 simulations. When all working models are correctly specified (i.e. $\lambda_z = \lambda_x = \lambda_y = 0$), then all estimators have nearly identical performance to TSLS. This is theoretically expected, because they are based on correctly specified working models in the calculation of the efficient score and are therefore asymptotically equivalent. When only the outcome model is misspecified (i.e. $\lambda_z = \lambda_x = 0, \lambda_y \neq 0$), then the TSLS estimator is biased (as the instrument expectation is not linear in the covariates) with larger standard deviation than the double-robust estimators, which were all unbiased. Bias-reduced estimation of the outcome model (BR_β) resulted in major efficiency gains in this case. This is likely the result of the orthogonality to (some of) the nuisance parameter estimators, brought about by this strategy (see also Vermeulen and Vansteelandt (2015)). When only the exposure model was misspecified (i.e. $\lambda_z = \lambda_y = 0, \lambda_x \neq 0$), then as theoretically predicted, the TSLS estimator and the double-robust estimators continue to be unbiased, but the performance of the locally efficient double-robust estimator was sometimes very poor because its efficiency is only attained at a correctly specified model for the exposure. In this case, drastic improvements were obtained via empirical efficiency maximization, because this strategy guarantees efficiency within a subclass of estimators, regardless of correct specification of an exposure model. The efficiency of the resulting double-robust estimator was sometimes better, sometimes worse than that of TSLS. When only the IV model was misspecified (i.e. $\lambda_x = \lambda_y = 0, \lambda_z \neq 0$), then all estimators were unbiased because of the double-robustness of the estimators and the fact that the TSLS estimator does not rely on correct specification of an IV model; all estimators had nearly

identical performance in this case. When both the exposure and outcome model are misspecified (i.e. $\lambda_z = 0, \lambda_x \neq 0, \lambda_y \neq 0$), then again TSLS is biased, unlike the double-robust estimators. The locally efficient estimator behaved poorly in this case and is greatly outperformed by empirical efficiency maximisation, which again performs best in combination with bias-reduced estimation of the outcome model. When all models were misspecified, then also the double-robust estimators were subject to bias. However, bias-reduced estimation of either the outcome model or the IV model resulted in bias reductions and efficiency gains. This is not surprising for BR_γ because the extended IV model happened to contain the truth: indeed, the inclusion of the covariate $e(C;\alpha)^T C$ in the instrument model was tantamount to the inclusion of V^2. For BR_β, where the extended outcome model did not contain the truth, this confirms that the proposed procedure reduces bias under model misspecification.

Table 2 evaluates the performance of the proposed sandwich standard error estimators, along with the coverage of 95% Wald confidence intervals. As predicted by the theory in Section 5.2, by ignoring nuisance parameter estimation, these intervals are slightly conservative in the case of BR_γ, but may undercover in the case of BR_β.

To further evaluate the bias-reduced estimation strategy, we additionally ran simulations under extreme misspecifications, such that both extended outcome and IV models did not contain the truth. In particular, we generated $n = 500$ independent measurements on mutually independent and standard normal covariates U and V, Z dichotomous with $P(Z = 1 \mid V) = 1 - \exp\{-\exp(-1 + V/2 - V^2/2 + \lambda_z V^3/8)\}$, X normal with mean $Z + U + V - ZV + 2V^2 + 2ZV^2 + 2\lambda_x V^3$ and Y normal with mean $X - U - V - 2V^2 + 2\lambda_y V^3$.

The working models were the same as before. The results are visualised in Figure 3 for all combinations of λ_x, λ_y and λ_z in $\{-1, 1\}$, and confirm the previous findings (see also Figure 4 which zooms in on the bias-reduced double-robust estimators). The

locally efficient double-robust estimator had very poor performance and, while empirical efficiency maximization resulted in major efficiency gains, it was still much worse than TSLS estimation. For instance, in the setting of Figure 3 (top, left), the locally efficient double-robust estimator had bias and standard deviation of -32.7 and 450, as opposed to -0.61 and 4.1 with empirical efficiency maximization, and -0.54 and 3.1 with TSLS. In combination with bias-reduced estimation, most of the bias disappeared and variance was often greatly reduced (see Figure 3 and 4).

All tables and figures about here.

7 Illustration

We illustrate the proposed methodology on a sample of 3010 working men aged between 24 and 34 who were part of the 1976 wave of the US National Longitudinal Survey of Young Men (Card, 1995). In particular, we will estimate the effect of years of education on the log of hourly wages in 1976 (Y). Following Card (1995), we use as an IV an indicator if the individual lived close to a college that offered 4 year courses in 1966 (Z). All reported analyses are adjusted for covariates (C) years of labour market experience and its square, marital status, an indicator if the individual is black, as well as various measures of geographical location in 1966 and 1976. Twelve years of education was most common (33%) in this study and was therefore used as a reference class by defining X to be the difference between the years of education and 12.

The log of hourly wages is reasonably normally distributed with mean 6.3 (SD 0.44), and is on average 0.075 (95% CI 0.068 to 0.082) higher per extra year of education, after linear regression adjustment for years of labour market experience, marital status, race and geographical location in 1966 and 1976. The partial correlation between education and the IV is 0.066. Below, we will report the results from IV analysis with 95% percentile-

based confidence intervals based on the nonparametric bootstrap with 1000 resamples.

TSLS analysis yields an education effect of 0.13 (SE 0.067, 95% CI 0.029 to 0.28) on the average log of the hourly wage, corresponding with a one-year increase in education. Because the instrument is dichotomous and strongly associated with covariates, its expectation is likely nonlinear in the covariates. The TSLS estimator is therefore sensitive to correct specification of the role of covariates in the outcome model. We thus evaluate the double-robust estimators based on a logistic regression model for the IV. The locally efficient double-robust estimator equals 0.10 (SE 0.044, 95% CI 0.025 to 0.18). Like the double-robust estimator based on empirical efficiency maximization (0.088, SE 0.045, 95% CI 0.0063 to 0.18), it is much more efficient than the TSLS estimator. Further, more minor efficiency gains are obtained through the proposed bias reduction strategies. In particular, we find that BR_γ equals 0.092 (SE 0.041, 95% CI 0.010 to 0.18) and BR_β equals 0.095 (SE 0.043, 95% CI 0.0063 to 0.19).

8 Discussion

In this article, we have argued that TSLS estimation, unlike many variations of the two-stage approach to estimation with an IV, is often robust against misspecification of the working models for the exposure and outcome. However, this robustness may come at the expense of a loss of precision, which can be considerable when, for instance, the exposure mean is nonlinear in the instrument and/or covariates, e.g. when the exposure is binary, multinomial or count data. Moreover, the suggested robustness of the TSLS estimators is limited to specific data-generating mechanisms: robustness against misspecification of the outcome model is for instance lost in TSLS estimators when the IV is nonlinear in covariates. We also demonstrated that another strength of TSLS, not generally shared by other two-stage estimators, is that including covariates will asymptotically not reduce, and

typically improve, efficiency when instrument and covariates are known to be independent and in the absence of effect modification.

In contrast, locally efficient double-robust IV estimators confer robustness against model misspecification in a wider class of data generating mechanisms. For instance, an attractive alternative, when instruments and covariates are known to be independent, is the estimator obtained by empirical efficiency maximisation: it is guaranteed consistent and efficient relative to a subclass of all CAN estimators. In other situations one should arguably worry more about bias than efficiency. We have shown that major improvements can be achieved by combining empirical efficiency maximisation with bias-reduced double-robust estimation. The resulting estimators have a very stable performance with considerable robustness against misspecification of all models for the instrument, exposure and outcome; their standard errors can be computed relatively easily using sandwich estimators. We are hopeful that by extending these results to double-robust estimators in nonlinear IV models (Robins, 1994; Vansteelandt et al., 2010), we will be able to improve the performance of IV estimators in these more complex settings where difficulties of estimation are common (Vansteelandt et al., 2011; Burgess et al., 2014). R-code for the considered estimators is given in Appendix S5 of the Supplemental Materials.

There are some limitations to our work. Our results are asymptotic and do not take into account the problem of 'weak instrument / small sample' bias (Bound et al., 1995). This may in practice exacerbate the problem of bias due to model misspecification. There are a number of variations on two-stage estimators that are designed to address this problem, such as e.g. limited information maximum likelihood (Anderson, 2005), but these will not generally exhibit comparable robustness towards model misspecification. It would be an important area for future research to tackle both sources of bias simultaneously. Related to this, although the results on empirical efficiency maximisation appear to suggest

that it is beneficial to adjust for all available covariates C when $Z \perp\!\!\!\perp C$, the performance of the resulting estimators may be affected in the presence of high-dimensional covariates. Whether and how to best select covariates in such cases, as well as in settings where it is not known whether $Z \perp\!\!\!\perp C$, constitutes an important area for future research.

References

Anderson, T.W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *J. Econometrics* **127**, 1-16.

Angrist, J.D., Imbens, G.W. & Rubin, D.B. (1996). Identification of causal effects using instrumental variables. *J. Am. Statist. Assoc.* **91**, 444-455.

Bound, J., Jaeger, D.A. & Baker, R.M. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous variable is weak. *J. Am. Statist. Assoc.* **90**, 443-50.

Bowden, J. & Vansteelandt, S. (2011). Mendelian randomisation analysis of case-control data using Structural Mean Models. *Statist. Med.* **30**, 678-694.

Bowden, R.J. & Turkington, D.A. (1985). *Instrumental Variables.* Cambridge University Press.

Burgess, S., Granell, R., Palmer, T.M., Sterne, J.A.C. & Didelez, V. (2014). Lack of identification in semi-parametric instrumental variable models with binary outcomes. *Am. J. Epidem.* **180**, 111-119.

Card, D. (1995). *Using geographic variation in college proximity to estimate the return to schooling.* In L. Christophides, E. Grant and R. Swidinsky (eds.), Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp.

Cao, W.H., Tsiatis, A.A. & Davidian, M. (2009). Improving efficiency and robustness of the double-robust estimator for a population mean with incomplete data. *Biometrika* **96**, 723-734.

Clarke, P.S. & Windmeijer, F. (2010). Identification of causal effects on binary outcomes using structural mean models. *Biostatistics* **11**, 756-770.

Didelez, V. & Sheehan, N.A. (2007). Mendelian randomization as an instrumental variable approach to causal inference. *Statist. Meth. Med. Res.* **16**, 309-330.

Didelez, V., Meng, S. & Sheehan, N.A. (2010). Assumptions of IV methods for observational epidemiology. *Statist. Sci.* **25**, 22-40.

Fischer, K. & Goetghebeur, E. (1999). Practical properties of some structural mean analyses of the effect of compliance in randomized trials. *Controlled Clinical Trials* **20**, 531-546.

Greenland, S. (2000). An introduction to instrumental variables for epidemiologists. *Int. J. Epidemiol.* **29**, 722-729.

Hernán, M.A. & Robins, J.M. (2006). Instruments for causal inference - An epidemiologist's dream? *Epidemiol.* **17**, 360-372.

Imbens, G.W. (2014). Instrumental Variables: An Econometricians Perspective. *Statist. Sci.* **29**, 323-358.

Mullahy, J. (1997). Instrumental variable estimation of count data models: Application to models of cigarette smoking behaviour. *Review Econom. Statist.* **79**, 586-593.

Okui, R., Small, D.S., Tan, Z.Q. & Robins, J.M. (2012). Doubly robust instrumental variable regression. *Statist. Sinica*, **22**, 173-205.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference.* Cambridge University Press, New York, NY, USA, 2nd edition.

Robins, J.M. (1994). Correcting for non-nompliance in randomized trials using structural nested mean models. *Comm. Statist. Theory Methods* **23**, 2379-2412.

Robins JM. (2000). Robust estimation in sequentially ignorable missing data and causal inference models. *Proceedings of the American Statistical Association*, Section on Bayesian Statistical Science 1999, pp. 6-10.

Robins, J.M. & Rotnitzky, A. (2001). Comment on "Inference for semiparametric models: Some questions and an answer," by P.J. Bickel and J. Kwon. *Statist. Sinica* **11**, 920-936.

Rotnitzky, A., Li, L.L. & Li, X.C. (2010). A note on overadjustment in inverse probability weighted estimation. *Biometrika* **97**, 997-1001.

Rubin, D. & van der Laan, M.J. (2008). Empirical efficiency maximization: improved locally efficient covariate adjustment in randomized experiments and survival analysis. *Int. J. Biostatistics* **4**, 1-40.

Stefanski, L.A. & Boos, D.D. (2002). The calculus of M-estimation. *Am. Statist* **56**, 29-38.

Tchetgen Tchetgen, E.J., Walter, S., Vansteelandt, S., Martinussen, T. & Glymour, M. (2015). Instrumental variable estimation in a survival context. *Epidemiology* **26**, 402-410.

Vansteelandt, S., Bowden, J., Babanezhad, M. & Goetghebeur, E. (2011). On instrumental variables estimation of causal odds ratios. *Statist. Sci.* **26**, 403-422.

Vermeulen, K. & Vansteelandt, S. (2015). Bias-reduced doubly robust estimation. *J. Am. Statist. Assoc.* **110**, 1024-1036.

Wooldridge, J. (2002). *Econometric Analysis of Cross Section and Panel Data*, MIT Press, Cambridge, MA.

Corresponding author:

Stijn Vansteelandt

Ghent University

Department of Applied Mathematics, Computer Science and Statistics

Krijgslaan 281, S9

9000 Gent, Belgium

stijn.vansteelandt@ugent.be

Table 1: Empirical bias and standard deviation of the two-stage estimator (TS), the locally efficient double-robust estimator (Loc Eff), the double-robust estimator based on empirical efficiency maximization (EEM) and these same estimators that employ bias-reduced nuisance parameter estimators (BR_γ and BR_β). The superscript number between brackets refers to the number of severely outlying estimates that were eliminated in the calculation of bias and empirical standard deviation; $\lambda_x \neq 0, \lambda_y \neq 0$ and $\lambda_z \neq 0$ refer to specific misspecifications of the exposure, outcome and instrument model, respectively.

	λ_x	λ_y	λ_z	TS	Loc Eff	EEM	BR_β	BR_γ
Bias	0	0	0	+0.0033	+0.0043	+0.0044	+0.0041	+0.0042
	0	1	0	−0.55	−0.0092	−0.035	+0.0024	−0.017
	0	-1	0	+0.56	+0.018	+0.044	+0.0059	+0.026
	1	0	0	+0.000073	+0.013	+0.0058	+0.0037	+0.0046
	-1	0	0	+0.0013	+0.0074	+0.0043	+0.0048	+0.0044
	0	0	1	−0.00057	−0.00033	+0.00009	−0.00009	+0.00029
	0	0	-1	+0.0053	+0.0055	+0.0095	+0.0050	+0.0045
	1	1	0	+0.15	$+0.11^{(2)}$	−0.040	+0.0039	−0.019
	-1	1	0	−0.41	+0.012	−0.021	+0.0016	−0.013
	1	-1	0	−0.15	$-0.095^{(4)}$	+0.051	+0.0038	+0.028
	-1	-1	0	+0.41	+0.0030	+0.030	+0.0079	+0.022
	1	1	1	+0.34	+0.36	+0.11	+0.021	−0.00028
	-1	1	1	−0.35	$-14^{(2)}$	−0.40	+0.024	+0.00073
	1	-1	1	−0.34	−0.36	−0.11	−0.023	−0.00059
	-1	-1	1	+0.35	$+15^{(2)}$	+0.86	−0.023	+0.0015
	1	1	-1	−0.94	$+0.59^{(1)}$	−0.48	+0.019	+0.0057
	-1	1	-1	−0.36	−0.084	−0.085	+0.018	+0.0048
	1	-1	-1	+0.94	$-0.20^{(2)}$	+0.50	−0.0081	+0.0039
	-1	-1	-1	+0.37	+0.10	+0.10	−0.0086	+0.0036
SD	0	0	0	0.11	0.11	0.11	0.11	0.11
	0	1	0	0.24	0.18	0.17	0.12	0.17
	0	-1	0	0.28	0.19	0.19	0.12	0.18
	1	0	0	0.18	$0.82^{(4)}$	0.12	0.12	0.12
	-1	0	0	0.068	0.12	0.11	0.11	0.11
	0	0	1	0.094	0.097	0.11	0.097	0.097
	0	0	-1	0.13	0.13	0.14	0.13	0.13
	1	1	0	0.46	$1.9^{(2)}$	0.19	0.12	0.17
	-1	1	0	0.11	0.23	0.17	0.12	0.17
	1	-1	0	0.48	$1.2^{(4)}$	0.20	0.12	0.18
	-1	-1	0	0.14	0.22	0.17	0.12	0.17
	1	1	1	0.15	0.12	0.16	0.10	0.14
	-1	1	1	0.15	$120^{(2)}$	8.30	0.11	0.14
	1	-1	1	0.14	0.10	0.16	0.099	0.13
	-1	-1	1	0.16	$140^{(2)}$	12	0.10	0.13
	1	1	-1	1.3	$11^{(1)}32$	0.85	0.13	0.17
	-1	1	-1	0.098	0.17	0.18	0.14	0.17
	1	-1	-1	1.5	$5.6^{(2)}$	0.98	0.13	0.17
	-1	-1	-1	0.12	0.17	0.18	0.13	0.16

Table 2: Empirical standard deviation (ESD), average of the (naïve) sandwich standard errors (ESE) and coverage (Cov) of 95% Wald intervals for the bias-reduced double-robust estimators. $\lambda_x \neq 0, \lambda_y \neq 0$ and $\lambda_z \neq 0$ refer to specific misspecifications of the exposure, outcome and instrument model, respectively.

λ_x	λ_y	λ_z	BR_β			BR_γ		
			ESD	ESE	Cov	ESD	ESE	Cov
0	0	0	0.11	0.12	96.1	0.11	0.12	96.7
0	1	0	0.12	0.12	96.1	0.17	0.22	97.7
0	-1	0	0.12	0.12	95.4	0.18	0.22	98.5
1	0	0	0.12	0.12	95.3	0.12	0.12	96.9
-1	0	0	0.11	0.12	95.0	0.11	0.12	97.7
0	0	1	0.097	0.12	98.5	0.097	0.10	95.9
0	0	-1	0.13	0.10	88.6	0.13	0.14	96.0
1	1	0	0.12	0.12	95.4	0.17	0.22	98.6
-1	1	0	0.12	0.12	95.0	0.17	0.22	95.3
1	-1	0	0.12	0.12	94.8	0.18	0.22	97.8
-1	-1	0	0.12	0.12	94.6	0.17	0.21	96.3
1	1	1	0.10	0.094	91.1	0.14	0.18	97.7
-1	1	1	0.11	0.42	99.7	0.14	0.19	98.3
1	-1	1	0.099	0.093	90.7	0.13	0.19	97.5
-1	-1	1	0.10	0.42	99.6	0.13	0.18	98.5
1	1	-1	0.13	0.12	89.8	0.17	0.20	97.5
-1	1	-1	0.14	0.099	85.3	0.17	0.20	96.4
1	-1	-1	0.13	0.12	88.8	0.17	0.20	97.0
-1	-1	-1	0.13	0.097	84.9	0.16	0.20	98.1

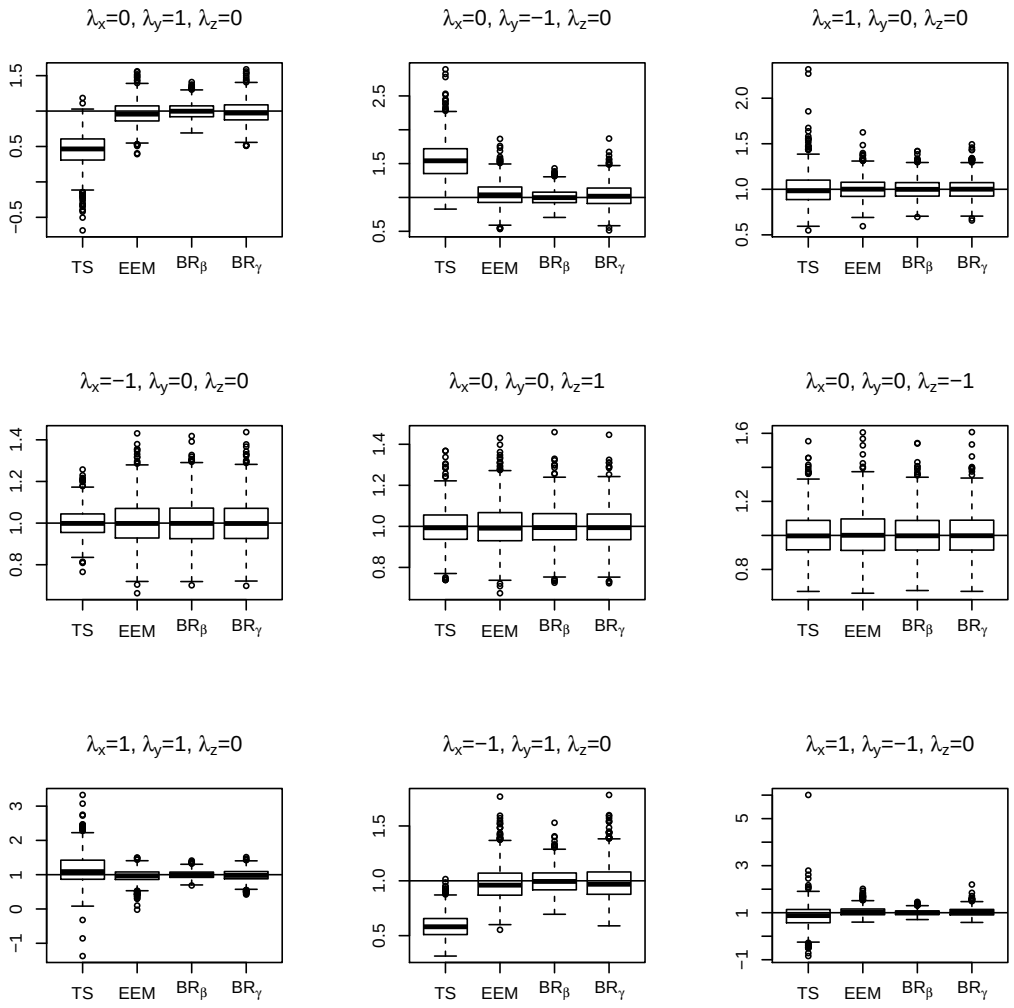

Figure 1: Boxplots of the two-stage estimator (TS), the double-robust estimator based on empirical efficiency maximization (EEM) and bias-reduced nuisance parameter estimators (BR_β and BR_γ) under the model misspecifications considered in Table 1. $\lambda_x \neq 0, \lambda_y \neq 0$ and $\lambda_z \neq 0$ refer to specific misspecifications of the exposure, outcome and instrument model, respectively.

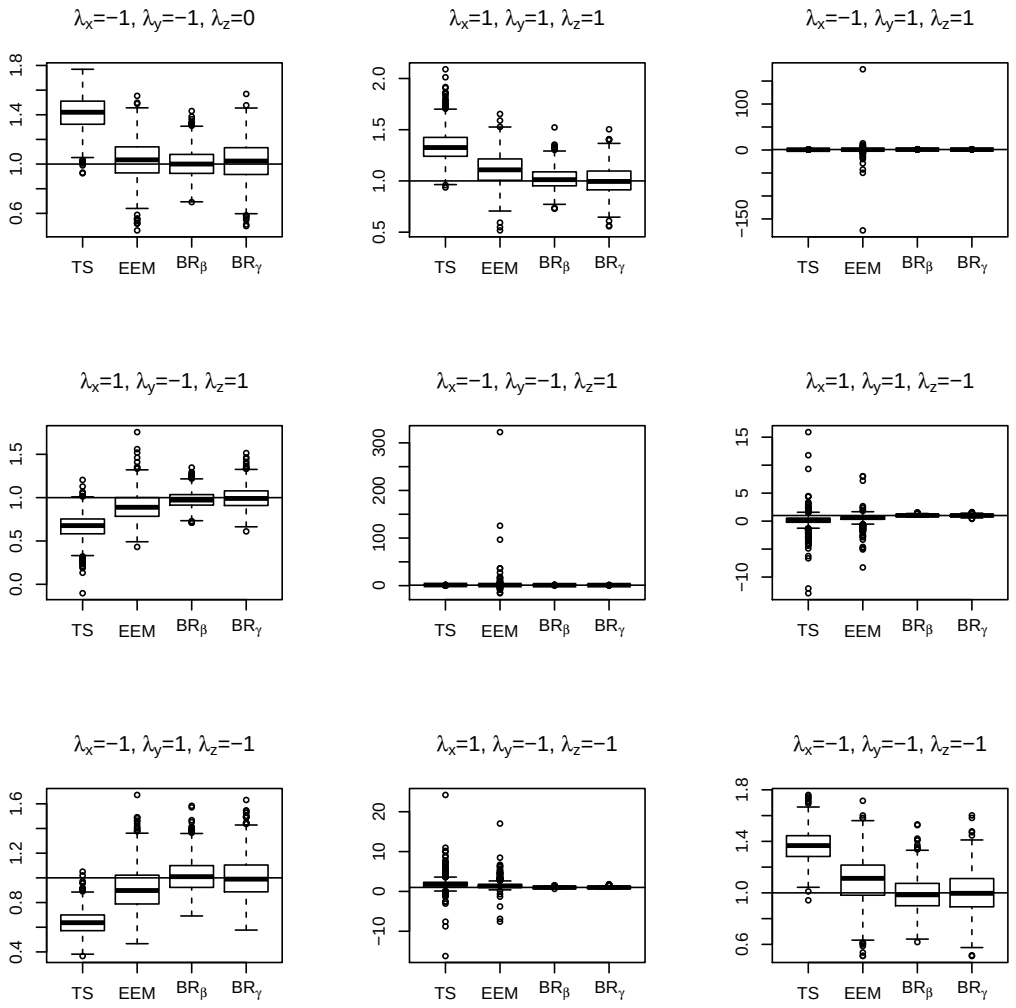

Figure 2: Boxplots of the two-stage estimator (TS), the double-robust estimator based on empirical efficiency maximization (EEM) and bias-reduced nuisance parameter estimators (BR_β and BR_γ) under the model misspecifications considered in Table 1. $\lambda_x \neq 0, \lambda_y \neq 0$ and $\lambda_z \neq 0$ refer to specific misspecifications of the exposure, outcome and instrument model, respectively.

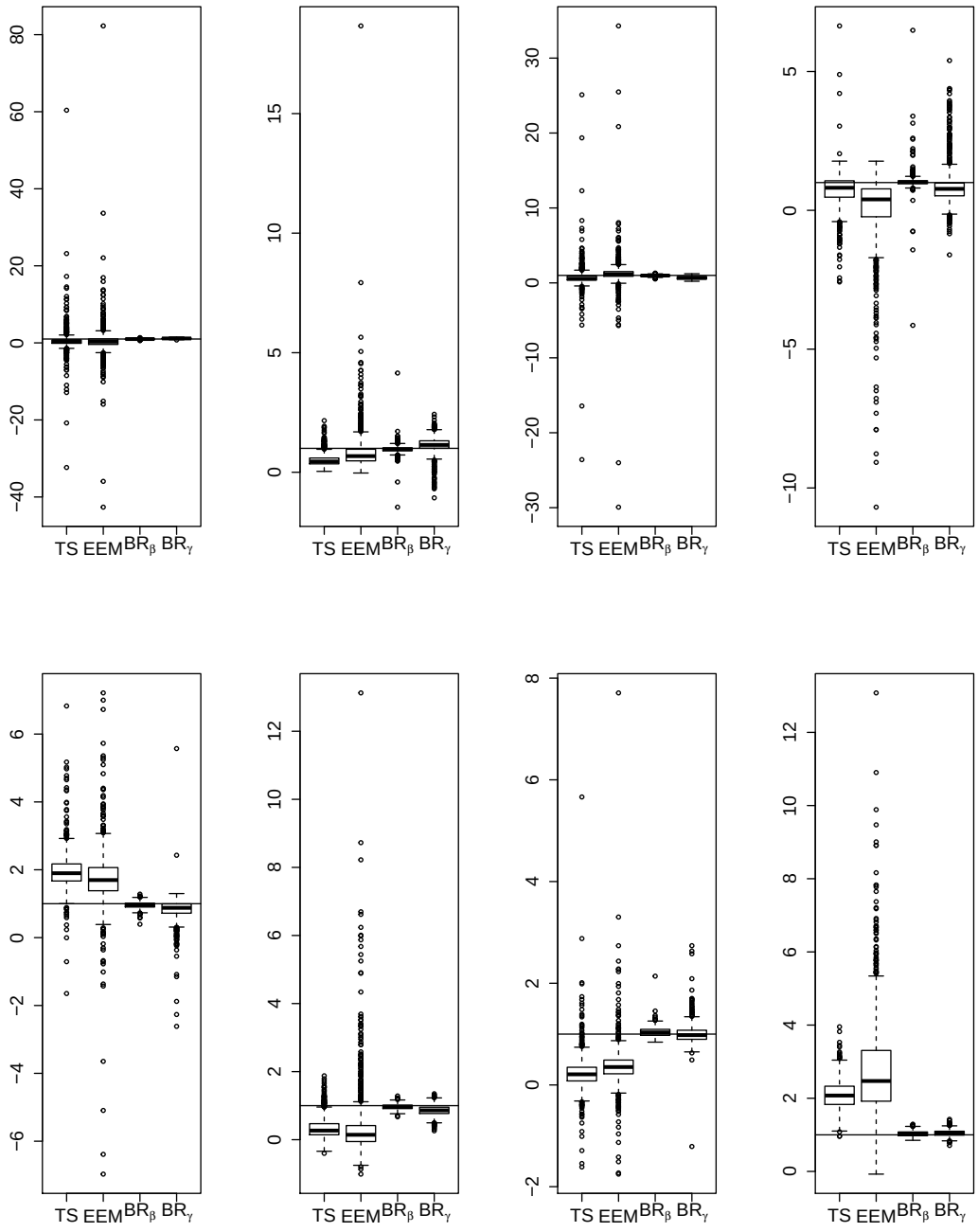

Figure 3: Boxplots of the two-stage estimator (TS), the double-robust estimator based on empirical efficiency maximization (EEM) and bias-reduced nuisance parameter estimators (BR_β and BR_γ) under extreme model misspecification.

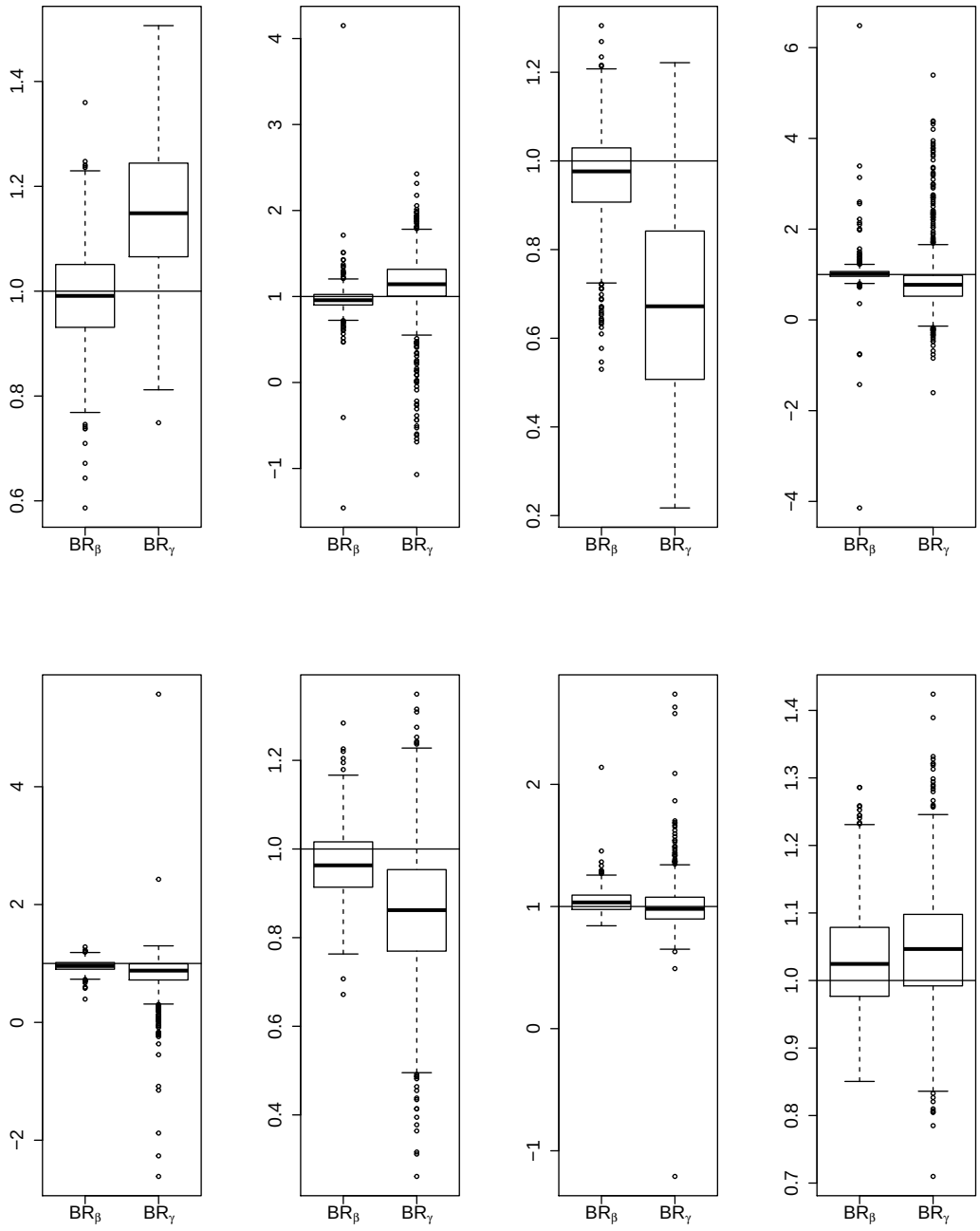

Figure 4: Boxplots of the bias-reduced nuisance parameter estimators (BR_β and BR_γ) under the same settings with extreme model misspecification as considered in Figure 3.