KATHOLIEKE UNIVERSITEIT
LEUVEN

Faculty of Business and Economics

On model selection and model misspecification in causal inference

Stijn Vansteelandt, Maarten Bekaert and Gerda Claeskens

DEPARTMENT OF DECISION SCIENCES AND INFORMATION MANAGEMENT (KBI)

KBI 1021

On model selection and model misspecification in causal inference

Stijn Vansteelandt, Maarten Bekaert
Department of Applied Mathematics and Computer Sciences
Ghent University, 281 (S9) Krijgslaan, 9000 Ghent, Belgium

and Gerda Claeskens
OR & Business Statistics, K.U. Leuven
Naamsestraat 69, 3000 Leuven, Belgium

Abstract. Standard variable-selection procedures, primarily developed for the construction of outcome prediction models, are routinely applied when assessing exposure effects in observational studies. We argue that this tradition is sub-optimal and prone to yield bias in exposure effect estimates as well as their corresponding uncertainty estimates. We weigh the pros and cons of confounder-selection procedures and propose a procedure directly targeting the quality of the exposure effect estimator. We further demonstrate that certain strategies for inferring causal effects have the desirable features (a) of producing (approximately) valid confidence intervals, even when the confounder-selection process is ignored, and (b) of being robust against certain forms of misspecification of the association of confounders with *both* exposure and outcome.

Keywords: Causal inference; Confounder selection; Double robustness; Influential weights; Model selection; Model uncertainty; Propensity score.

1 Introduction

The primary goal of most observational studies is to assess cause-effect relationships. Model-selection procedures - in particular variable-selection procedures - are routinely employed in this

process, but rarely with regard to the ultimate focus on causal effects[1,2]. In addition, the reliance on model-selection procedures is commonly ignored when causal inferences are ultimately drawn. We will reconsider principles of model-selection when the focus is on the estimation of causal effects. We give a brief outline below.

Decisions to exclude/include covariates in a regression model are commonly based on the strength of evidence for their (residual) association with the outcome. When the (causal) effect of a given exposure on the outcome is targeted, then this routine strategy is not ideal and may result in a potentially substantial bias in the exposure effect estimate. The decision to include covariates in a regression model must ideally be based on the strength of evidence for these covariates confounding the association between exposure and outcome. Since by definition, confounders are simultaneously associated with exposure and outcome, procedures that ignore the covariate-exposure association can be sub-optimal, especially for covariates that have strong associations with the exposure[3,4]. Causal inference procedures that naturally evaluate the strength of covariate-exposure associations (e.g. propensity score adjusted estimators[5]) may thus behave differently than standard (outcome-regression based) procedures, especially when combined with model-selection strategies.

In Section 2.1, we argue that the set of potential confounders amongst all measured covariates is often high-dimensional in practice and that there is some tension between the desire to acknowledge all of them through regularization methods, such as ridge regression, and the desire to reduce the covariate space through confounder-selection procedures. We discuss limitations of the most commonly adopted confounder-selection procedures in Section 2.3 and argue in Section 2.4 that ideally such procedures should directly target the quality of the exposure effect estimator. One proposal is worked out in detail for logistic regression models and applied in Section 2.5 to the analysis of an observational study for the effect of right-heart catherization

on 180-day mortality in critically ill patients. A limitation to the use of confounder-selection strategies is that they have a tendency to produce under-covering confidence intervals by not acknowledging model uncertainty. In Section 2.6 we focus on causal inference procedures that return consistent causal effect estimators when a model for the exposure distribution, given confounders, is correctly specified. We demonstrate that, surprisingly, these procedures remain (approximately) confidence valid in the presence of exposure model selection. For this and other reasons mentioned in the article, they thus succeed better than standard estimation procedures at quantifying the total degree of uncertainty.

In the remainder of the article, we focus on the broader problem of model building as opposed to variable-selection. We discuss principles of causal model building in Section 3.1 and examine the consequences of model misspecification in Section 3.2. In particular, we study misspecification bias affecting so-called doubly robust[6] estimation procedures which promise consistent estimation of causal effects when at least one of two (possibly overlapping) nuisance working models is correctly specified. This leads to estimation procedures that perform well under more global forms of working model misspecification, which are seen to substantially outperform more standard procedures in simulation studies reported in Section 3.3.

2 Confounder-selection

2.1 Confounder-selection versus regularization

Throughout - unless otherwise specified - we assume that a possibly high-dimensional collection of covariates is available, which includes all confounders for the effect of exposure A on outcome Y, and thus contains at least one subset of covariates that are sufficient to control for confounding[7]. Determining such subset is impossible in the absence of background knowledge on the causal data-generating mechanism[8]. This is largely because adjustment for covariates

that are affected by the exposure or the outcome can actually increase bias[9–11], which makes purely associational approaches to confounder selection fallible[12]. Causal diagrams[7,8,10] are very helpful to communicate and visualize the data-generating mechanism and, subsequently, to identify covariate sets that are sufficient for confounding control[7].

It is presumably true that in most realistic applications, all covariates in a sufficient covariate set will have some association with both the outcome and the exposure[13]. From that perspective, with concern for bias, it seems beneficial to adjust for all available covariates in the set[13–15]. This has the further advantage that, by acknowledging the uncertainty regarding all covariate effects, it returns a more honest reflection of the overall uncertainty regarding the exposure effect estimator. However, it has the disadvantage that it may induce a bias and inefficiency as a result of overfitting in the outcome regression model. To guard against this, one could use regularization methods such as ridge regression (see Greenland[13] and Budtz-Jorgensen et al.[16] for convincing examples). Alternatively, because propensity-score adjusted estimators can cope better with some overfitting in the propensity score[17,18], one could consider propensity-score adjustment based on a fitted propensity score model which includes all available covariates[14].

The folklore that conditioning on measured covariates reduces bias, must however be taken with caution. This is not only true because of the increased concerns of model misspecification, of possible shrinkage bias and of missing or mismeasured covariate data as more covariates are considered. More fundamentally, evidence is accruing that even adjustment for antecedents of the exposure may induce or aggravate selection bias. This may happen when, as in the causal diagram of Figure 1, non-causal relationships are observed between the confounders L and both exposure A and outcome Y. In that case, the adjustment for L induces a so-called M-bias[7,19–21] by connecting exposure A and outcome Y along the path $A \leftarrow U1 \rightarrow L \leftarrow U2 \rightarrow Y$. When the causal effects of L on exposure and outcome are weak, this bias may in principle

exceed the bias of an unadjusted analysis. In particular, when L affects neither exposure, nor outcome, then interestingly the unadjusted analysis, but not the adjusted analysis, would be valid. A further problem occurs when the association between A and Y is confounded through an unmeasured common cause (i.e., $U3$ in Figure 1). In that case, the bias of the unadjusted analysis may surprisingly be amplified upon adjusting for L, provided L is strongly correlated with the exposure[22,23] (see also Section 2.1).

Figure 1 about here.

In view of the concerns for M-bias and bias amplification, it may be advantageous to adjust for a strictly smaller subset of covariates that are minimally[24] sufficient to control for confounding (in the sense that, given these covariates, all remaining covariates are only associated with either the exposure or the outcome, but not both). Adjusting for a subset of available covariates may have the further advantage of yielding more efficient effect estimators. In particular, Hahn[25] elegantly shows that adjustment for covariates that have no (residual) association with the outcome can reduce the efficiency of nonparametric estimators of the marginal treatment effect, unlike adjustment for covariates that have no (residual) association with the exposure. In view of this and of the aforementioned concerns about bias amplification, it has been suggested that the selection of confounders should be based on their importance with respect to the outcome, rather than the exposure[23,26]. Whether such recommendation to reduce a sufficient set of confounders is successful, is arguable however. First, the results of Hahn[25] refer to settings where *a priori* knowledge is available that certain covariates have no residual association with the outcome. In practice, the selection of confounders is virtually always (at least partly) data-driven, but the ensuing uncertainty is most often ignored. Upon acknowledging the additional model uncertainty, one may well find effect estimators obtained after variable-selection being less efficient than those obtained from a full model which includes all available covariates[16,27].

Second, in the next section we will find that in well designed studies where efforts have been made to collect data on causal risk factors for the exposure that are also associated with the outcome, the concerns for M-bias and bias amplification may be more modest. Third, even when these concerns are justified, then as a result of multicollinearity, it would still be difficult to measure the importance of a covariate with respect to the outcome whenever that covariate is strongly correlated with the exposure. Standard variable-selection based on hypothesis testing in outcome regression models may therefore lack power to detect even relatively strong confounders.

Extensive simulation studies are needed, complementing the early work of Greenland and collaborators[28,29], to be able to gauge the relative importance of the aforementioned pros and cons of confounder-selection versus no selection. Making a choice between these strategies is further complicated by the fact that M-bias and bias amplification occur only in the presence of *unmeasured* common causes of exposure, outcome and confounders, so that one cannot protect against it or know to what extent these biases - which primarily affect strategies that avoid selection - are present. In the following section, which may be skipped by the less interested reader, we therefore attempt to develop insight into the extent to which the concerns for M-bias and bias amplification are justified in practical applications.

2.2 M-bias and bias amplification

We compute the magnitude of the biases of the unadjusted and adjusted analysis in the Appendix for multivariate normal variates following the path diagram of Figure 1, extending the work of Wooldridge[22] and Pearl[23]. Let ρ_1 denote the standardized path coefficients[30] between A and $U1$, L and $U1$, L and $U2$, or Y and $U2$ (which we assume to be equal for simplicity), ρ_2 denote the standardized path coefficients between A and $U3$, or Y and $U3$ (which we assume

to be equal for simplicity), ρ_{al} denote the correlation between A and L and ρ_{yl} denote the correlation between Y and L in the absence of an exposure effect (or upon setting A to a fixed value, uniformly in the population). Then the bias of the adjusted analysis (either based on standard regression adjustment or based on inverse probability weighting by $1/f(A|L)$; see Section 3.2) is

$$\frac{\rho_2^2 - \rho_1^4}{1 - \rho_{al}^2},$$

where the first term reflects bias due to the unmeasured common cause $U3$ of A and Y, and the second term reflects M-bias; that is, the two terms reflect spurious associations along the paths $A \leftarrow U3 \rightarrow Y$ and $A \leftarrow U1 \rightarrow L \leftarrow U2 \rightarrow Y$, respectively. The denominator suggests that strong correlations between exposure and measured confounders not only have a tendency to amplify bias resulting from unmeasured confounders $U3$, in line with the conclusions of others[22,23], but also M-bias. The bias of the unadjusted analysis is

$$\rho_{al}\rho_{yl} - \rho_1^4 + \rho_2^2,$$

which does not suffer this amplification. Here, the first two terms encode bias due to not adjusting for the measured confounder L and the last term measures bias due to the unmeasured common cause $U3$ of A and Y; that is, the three terms reflect spurious associations along the paths $A \leftarrow U1 \rightarrow L \rightarrow Y$, $A \leftarrow L \rightarrow Y$, $A \leftarrow L \leftarrow U2 \rightarrow Y$ and $A \leftarrow U3 \rightarrow Y$. We thus find that the adjusted analysis will have larger bias than the unadjusted analysis when the correlation between Y and L (other than through A) is sufficiently weak in the sense that

$$\rho_{yl} < \frac{\rho_{al}}{1 - \rho_{al}^2}(\rho_2^2 - \rho_1^4). \tag{1}$$

Even if the presence of an unmeasured common cause $U3$ of A and Y could be ruled out, the existence of unmeasured common causes such as $U1$ and $U2$ would be difficult to exclude in any given application. In particular when L is high-dimensional, it would be difficult to believe that

all of its components are only linked to A or Y by means of a causal effect. This suggests that M-bias is likely to arise in practice, although the fourth order terms express that its magnitude is likely going to be small. Similar findings were obtained by Greenland[19] in the all binary case. An exception occurs when the correlation between A and L is strong, for then even a modest degree of M-bias may in principle be amplified by a potentially important magnitude.

2.3 Confounder-selection strategies

Amongst the various confounder-selection strategies that are routinely adopted in practice, backward elimination based on hypothesis tests in outcome regression models is the default strategy. It is not ideal, however, because it is based on accepting the null hypothesis when covariates are non-significantly associated with the outcome[13] and because it ignores the association between exposure and covariates when deciding whether a given covariate confounds the association between exposure and outcome[3]. As such, it has a tendency to under-select important confounders[31] by ignoring covariates that have relatively weak associations with the outcome (conditional on the exposure), but strong associations with the exposure[3,4]. Such covariates are typically dismissed because they induce problems of multicollinearity (arising from correlation between the exposure and covariates), thereby inflating the uncertainty on the estimated treatment effect. This uncertainty is often interpreted as a sign of inefficiency, which is justified in some cases but should more generally be viewed as a reflection of the lack of information about the exposure effect[32]. By eliminating these covariates, one thus risks not only to induce a bias in the estimated exposure effect, but also to understate the actual uncertainty. Precisely in settings where there is much separation in the covariate distributions of exposed and unexposed subjects, and therefore much uncertainty about the exposure effect, conventional backward elimination strategies will tend to remove covariates from the outcome regression model and, thereby, yield misleadingly precise exposure effect estimates. Similar concerns apply

to penalization methods such as the lasso or elastic nets[33,34] and certain confounder-selection methods based on identification results for minimally sufficient sets of confounders[24,35] because of their tendency to dismiss covariates that are strongly associated with the exposure.

In epidemiology, some of these concerns have contributed to the popularity of change-in-estimate procedures which tend to have better success[28,29,31] by directly evaluating the impact of confounder-selection on the magnitude of the exposure effect estimate. While these target more directly a reduction of confounding bias, also these approaches are not ideal because they ignore estimation uncertainty and may be inefficient by under-selecting covariates that are only predictive of the response[36]. Furthermore, apart from finite-sample imprecision and model misspecification, inclusion of a covariate in a regression model may induce a change in treatment effect estimate, even when that covariate is not a confounder of the exposure-outcome relation. This may happen as a result of non-collapsibility of association measures in nonlinear models[24,28,29], which may change in magnitude upon adjusting for a covariate that is solely associated with the outcome (but independent of the exposure). This may also happen as a result of M-bias or bias amplification in both linear and nonlinear models.

2.4 Focused confounder selection

We believe that an 'optimal' confounder-selection strategy should focus on the quality of the exposure effect estimator. We will therefore closely follow the idea of change-in-estimate procedures, but accommodate their limitations, albeit necessarily presupposing that there are no unmeasured confounders (i.e. in particular, that $U3$ and either $U1$ or $U2$ are absent in the causal diagram of Figure 1). Specifically, let τ^* denote the target effect parameter and $\hat{\tau}$ an estimator of it. Then we will focus confounder-selection on the precision of the exposure effect estimator, as measured through its mean squared error $E\{(\hat{\tau} - \tau^*)^2\}$. Our choice not to pursue

conventional confounder-selection procedures based on the likelihood function (e.g. based on the AIC or BIC), is further guided by the fact that, as shown in Section 3.2, standard maximum likelihood inference can be sub-optimal for the estimation of nuisance working models (e.g. for modeling the association of confounders with either the outcome or exposure). Mean squared error is also the focus of Claeskens and Hjort[2], whose focused information criterion (FIC) is based on exact or asymptotic calculations in parametric models, and of Brookhart and van der Laan[37] who use cross-validation instead. Alternatively, one could focus model/confounder selection on the (counterfactual) prediction error, as in Claeskens, Croux and Van Kerckhoven[38], who use a prediction-focused information criterion, and Mortimer et al.[39] and Haight et al.[40] who use cross-validation instead.

Given our focus on the mean squared error of the exposure effect estimator, an important consideration is whether the estimators $\hat{\tau}_S$ corresponding to different models S are all consistently estimating the same parameter τ^* under correct model specification. This is not usually the case for conditional exposure effects due to noncollapsibility of nonlinear association measures[24] and the possibility of effect modification. This makes approaches for model-selection focused on the mean squared error not entirely appropriate for estimating the usual conditional exposure effects. This problem can be overcome by targeting confounder-selection at the marginal or population-averaged exposure effect. For instance, let A be a dichotomous exposure (taking values 0 and 1) and consider the parameter β^* indexing $\text{logit}P(Y=1|A,L) = \omega(L;\gamma^*) + \beta^* A$, where is $\omega(L;\gamma)$ a known function, smooth in γ, and γ^* is an unknown finite-dimensional parameter. For instance, $\omega(L_i;\gamma) = \gamma_0 + \gamma_l L_i$ in the case of standard regression adjustment, or $\omega(L_i;\gamma) = \gamma_0 + \gamma_p \pi(L_i;\gamma)$ with $\pi(L_i;\gamma) = P(A_i = 1|L_i;\gamma) = \text{expit}(\gamma_1 + \gamma_l L_i)$ in the case of propensity score adjustment[5]. Then, with $Y(a)$ denoting the counterfactual outcome following exposure level a, the marginal causal odds ratio $\tau^* = \text{odds}\{Y(1)=1\}/\text{odds}\{Y(0)=1\}$ can, for given estimates $\hat{\gamma}$ of γ^* and $\hat{\beta}$

of β^*, be estimated as

$$\hat{\tau} = \frac{\sum_{i=1}^{n} \text{expit}\left\{\omega(L_i;\hat{\gamma}) + \hat{\beta}\right\} / \sum_{i=1}^{n} \text{expit}\left\{-\omega(L_i;\hat{\gamma}) - \hat{\beta}\right\}}{\sum_{i=1}^{n} \text{expit}\left\{\omega(L_i;\hat{\gamma})\right\} / \sum_{i=1}^{n} \text{expit}\left\{-\omega(L_i;\hat{\gamma})\right\}}. \tag{2}$$

Thus focussing on the marginal treatment effect τ^*, we propose the following focused confounder-selection procedure, which inherits from work by Claeskens et al.[38] and Crainiceanu et al.[3]. We divide the model space into $M+1$ orbits, where M is the number of potential covariates (i.e., confounders and/or functions of confounders, such as higher order terms or interactions) and where the jth orbit, $j = 1, \ldots, M+1$ comprises all models with $j-1$ covariates and an intercept. Within each orbit, we select the outcome regression model that minimizes the mean squared error of $\hat{\tau}$. This is done using the following stochastic search method, which is closely linked to that in Crainiceanu et al.[3]. Starting from a model in the $(j-1)$th orbit, we add the covariate that provides the largest reduction in mean squared error. The stochastic search then selects at random one covariate which is in the model and one which is not in the model, and constructs a new model by interchanging both covariates. The new model is accepted when $L_{\text{new}} < L_{\text{old}}$, where L_{old} and L_{new} are the mean squared errors of $\hat{\tau}$ under the old and new model, respectively. When $L_{\text{new}} > L_{\text{old}}$, the new model is accepted with probability $(L_{\text{old}}/L_{\text{new}})^{\alpha}$, where α is a user-selected tuning parameter. Alternatively, a deletion/substitution/addition algorithm[40] could be used, which involves exhaustive model search within model subclasses obtained by either deleting, substituting or adding one covariate to those already available in the model. In this process, the mean squared error can be estimated based on a cross-validation procedure where the data are partitioned into a training sample and validation sample V times. That is, the mean squared error of the estimator $\hat{\tau}$ can be approximated with $(1/V)\sum_{v=1}^{V}(\hat{\tau}_v - \hat{\tau}_0)^2$, where $\hat{\tau}_v$ is the estimator of τ^* as obtained under the considered model on the training sample, and where $\hat{\tau}_0$ is an estimator of τ^* as obtained under the full model on the validation sample. Minimization of this estimated loss function is then equivalent to minimization of the mean squared error when

the estimator $\hat{\tau}_0$ is unbiased[37]. Computing time can be drastically reduced through asymptotic approximations of the mean squared error, which can be made under a local misspecification assumption (see Section 2.6). A framework for this is developed in Hjort and Claeskens[41] for parametric models and adapted to our specific setting in the Appendix.

2.5 Application

We evaluate the proposed confounder-selection procedure in an observational study investigating the effect of right heart catherization (RHC) on 180 day mortality in 5735 critically ill patients[42]. For every patient, the exposure of interest A was coded 1 if RHC was used within 24 hours of admission and 0 otherwise. In total, 61 covariates (L) on the patients' underlying health condition within 24h of ICU admission (physiological status), on their underlying comorbidity and on demographic information were available for analysis. The original analysis[42] used logistic regression to develop an estimated propensity score for each patient, which was then used for matching RHC patients to non-RHC patients. In this Section, we will contrast different confounder-selection methods, including the one proposed in the previous section. R-code for the analyses can be found on http://users.ugent.be/~svsteela/Site/Publications.html.

Figure 2 about here.

Figure 2 (left) shows the mean squared error (MSE) for the best model within each orbit as obtained by minimizing the mean squared error of the marginal log odds ratio (MLOR) (or equivalently, by minimizing the focused information criterion (FIC) which measures the mean squared error of the MLOR up to an additive constant, see the Appendix) in the case of standard covariate adjustment (i.e., $\omega(L_i;\gamma) = \gamma_0 + \gamma_l L_i$ in Section 2.4). The MSE is largest for the narrow (due to large bias) and full model (due to large variance); minimal MSE is attained for simple models involving 2 covariates only. For illustrative purposes, Figure 2 (right) compares the

thus obtained estimates for the MLOR under standard covariate adjustment (solid black line) with the conditional log odds ratios (CLOR) corresponding to the same models (dotted line). It demonstrates the stability of the estimated MLOR over the different orbits, which is useful information in itself as the observed stability strengthens confidence in the analysis results. It suggests also increasing conditional treatment effect estimates over different orbits, which is due to noncollapsibility of the odds ratio. This underscores the potential limitations of variable-selection procedures that focus on conditional effect measures, which tend to mix confounding with non-collapsibility of association measures.

Figure 2 (right) also displays the results obtained upon applying the procedure advocated in Crainiceanu et al.[3]. This procedure involves first selecting covariates on the basis of their association with the exposure as measured in terms of the AIC, and subsequently selecting any remaining covariates on the basis of their residual association with the outcome, again measured in terms of the AIC. The estimates for the CLOR (which is the focus of that procedure) are initially very unstable as a result of selecting covariates that are strongly associated with the exposure, but not with the outcome. Stability in the estimates is attained only for very large orbits which, again, may be partly due to non-collapsibility of the odds ratio. Like standard model selection procedures, it thereby gives a somewhat misleading impression that the association between RHC and mortality is confounded by many of the measured covariates. The proposed procedure improves upon this (a) by focusing the model selection on a parameter which is identically defined over the different orbits, and (b) by selecting covariates on the basis of their potential to increase the precision of the treatment effect estimate, as well as their ability to reduce confounding in the treatment effect estimate.

Table 1 about here.

Table 1 reports estimates of the effect of RHC on mortality as obtained from these different

confounder-selection procedures. As shown in Figure 2 (left), minimal MSE is attained for simple models involving 2 covariates only (age and a covariate which indicates the presence of a Solid Tumor, Metastatic Disease, Chronic Leukemia/Myeloma, Acute Leukemia or Lymphoma) and results in a MOR of 1.33 (95% CI 1.21 to 1.46). The unadjusted analysis gave a MOR of 1.25 (95% CI 1.14 to 1.38) with an MSE of 0.0037, versus 0.0035 for the full model. In contrast, the 'optimal' model of Crainiceanu et al.[3] includes 36 predictors of right heart catherization, regardless of their association with the outcome, and 11 additional covariates on the basis of their residual association with the outcome. Also covariate adjustment and propensity score adjustment based on backward elimination (BE) strategies tend to select many more covariates at the expense of accuracy. They do so because the decision to enter covariates into the model is based on either their association with the outcome (as in standard covariate adjustment), or their association with the exposure (as in propensity score adjustment), but not on the basis of a more balanced evaluation in terms of the quality of the treatment effect estimate. Given the large number of patients in this study, many of these associations are strong in terms of the evidence provided by p-values, but not necessarily in terms of their potential to distort the treatment-outcome association by an important magnitude.

2.6 Model uncertainty

In small data sets where the variance of the exposure effect estimator is dominant, focused confounder-selection strategies might have a tendency to delete confounders when their adjustment causes a large variance inflation. While this may be beneficial to the overall accuracy of the exposure effect estimator, a concern is that it may come at the expense of confidence validity, considering that confidence intervals capture sampling variability, but not bias. Confidence validity may be further compromised by the fact that uncertainty resulting from the data-driven model building process is commonly ignored. Although the bootstrap or asymptotic approxi-

mations[41] could be used to acknowledge this, these are often not considered in practice. We will now argue that these concerns can be tempered to some extent by the use of propensity-score based estimators.

First, propensity-score based estimators which force important predictors of the exposure into the propensity score (e.g. the procedure advocated by Crainiceanu et al.[3]) are relatively less susceptible to bias resulting from insufficient confounding adjustment because the set of confounders forms a subset of the exposure predictors; the same is not true for approaches which rely on outcome predictors because the magnitude of the residual association between outcome and predictor, given the exposure, is difficult to assess for predictors that are strongly associated with the exposure. A drawback is that such propensity-score based procedures can be inefficient and more prone to bias amplification when they include predictors that have (almost) no residual association with the outcome[23,26].

Second, in the Appendix, we study the asymptotic behaviour of exposure effect estimators which solely rely on correct specification of a propensity score model, as obtained after model-selection. Examples are the G-estimator[17] and the inverse probability weighted estimator[43] of the average causal effect (see Section 3.2). Because the potential for model misspecification cannot be ignored in the presence of model-selection, the asymptotic behaviour of such estimators is examined within the local misspecification framework of Hjort and Claeskens[41]. More precisely, we assume that the true exposure data-generating mechanism is of the form $f(A|L) = f(A|L; \alpha_1^*, \alpha_2^* + \delta/\sqrt{n})$, with $f(A|L; \alpha_1, \alpha_2)$ a conditional density function of A, given L, which is smooth in α_1 and α_2, where α_1^* and δ are unknown finite-dimensional parameters and where α_2^* is a chosen finite-dimensional parameter (e.g. $\alpha_2^* = 0$). Here, α_1 encodes the unknown part of the parameter vector which is shared between all competing submodels. Each exposure working model, denoted S, thus assumes some of the components α_{-S} of α_2 to be known and

equal to the corresponding components of α_2^*, and assumes the remaining components α_S to be unknown. Note that the reason to allow for misspecification of the model parameters within a 1 over root-n distance is because in large samples standard model selection techniques would systematically choose the narrow model (which assumes α_2 equals α_2^*) when smaller misspecifications are considered, and systematically select the wide model (which assumes α_2 is unknown) when larger misspecifications are considered[41].

In the Appendix, we then show that interestingly a conservative asymptotic variance of the considered exposure effect estimators is obtained when imprecision due to estimation and model-selection on the propensity score is ignored, provided that an efficient estimator is used for the parameters indexing the propensity score model. This result is of importance as it suggests that the model/confounder-selection procedure can be ignored in inference about the exposure effect, provided that the local misspecification assumption holds. It does not immediately follow, however, that confidence intervals which ignore estimation and model uncertainty in the propensity score will attain the nominal coverage probability. This is because, as shown in the Appendix, the distribution of the exposure effect estimator in the presence of model-selection is not centered at zero, but follows a mixture distribution with bias components converging at root-n rate to zero. Preliminary simulation studies (not shown) confirmed that, nonetheless, close to nominal coverage levels are attained even when this is ignored.

3 Model building

3.1 Principles of causal model building

We will now broaden the focus from confounder-selection to model building. Though historically, the use of parametric models combined with maximum likelihood inference has been dominant (cfr. structural equation models (SEMs)), more recently - stimulated by pioneering

work of James Robins - a trend is now seen towards semi-parametric modeling of causal effects. Path diagrams, used by SEM practitioners as convenient representations of a multivariate normal model and as convenient tools for combining path-specific effects into exposure effects of interest, are substituted by 'non-parametric' causal diagrams[44]; these can be combined with semi-parametric models directly parameterizing the exposure effect of interest[45].

The appeal of semi-parametric inference for causal effects surmounts the usual concerns for model misspecification and limited flexibility in parametric inference. Parametric likelihood-based procedures explicitly ignore information on the exposure distribution which has nevertheless demonstrated to be relevant for confounder-selection in Sections 2.5 and 2.6. For instance, in the absence of an exposure effect, the common strategy of forcing the exposure into the model may lead one to systematically ascribe an effect of extraneous covariates to an exposure effect[46]. This can be overcome using propensity score methods which force the propensity score into the outcome regression model, irrespective of whether it is significantly associated with the outcome. Robins and Ritov[47] underscored more formally the importance of using information on the exposure distribution in causal inference by demonstrating that, due to the curse of dimensionality, likelihood-based procedures fail to estimate treatment effects in randomized experiments where randomization is conditional on a high-dimensional covariate; see also[48,49].

In further clarification of the philosophical principles behind semi-parametric modeling of causal effects, suppose that interest lies in the direct effect of A on Y which is not mediated by M in the causal diagram of Figure 3. SEM procedures would typically dismiss U from the path diagram and thereby arrive at biased causal effect estimates. Alternatively, they would include U, thus requiring models for the conditional densities $f(Y|A,M,U)$, $f(L|U)$ and $f(U)$, and subsequently yield causal effects conditional on U. These are not only difficult to specify, estimate and interpret by the fact that U is unmeasured, but additionally raise questions as to

whether the identification of the direct effect under the model comes from structural assumptions (e.g., assumptions about the absence of specific direct effects or common causes) alone, or from parametric assumptions (e.g. regarding the distribution of U) in addition. In the latter case, we say that the considered causal effect is not non-parametrically identified[50]. This is not ideal as it can make the results heavily sensitive to the chosen (semi-)parametric modeling assumptions (see e.g. Little[51] and Scharfstein, Rotnitzky and Robins[50]; see Vansteelandt[52] for an example illustrating the importance of nonparametric identification in a more general context). G-computation[53] enables identifying the counterfactual mean $E\{Y(a,m)\}$ corresponding to setting the exposure A at a and the mediator M at m as $\int E(Y|A=a,M=m,L)f(L|A=a)dL$, where the conditional mean $E(Y|A,M,L)$ and density $f(L|A)$ could be substituted with parametric likelihood-based estimators. Also this approach is not ideal as it does not directly parameterize the (controlled) direct effect[54] $E\{Y(a,m)-Y(a^*,m)\}$ of interest and henceforth does not enable researchers to express hypotheses of interest (e.g., that a direct effect of A on Y is not modified by M) in a parsimonious way. In addition, it is essentially impossible to postulate nonlinear models for $E(Y|A,M,L)$ and $f(L|A)$, which accommodate a dependence on A (as suggested by Figure 3), and are such that $\int E(Y|A=a,M=m,L)f(L|A=a)dL$ does not depend on a for all m. This is the root cause of the so-called null paradox[45] according to which G-computation based tests of the null hypothesis of no direct effect will with certainty be rejected in large samples. These subtleties underscore the importance of parameterizing the exposure effect of interest directly, which may be most naturally approached through the use of semi-parametric inference[55–57].

Figure 3 about here.

3.2 Model misspecification

Many semi-parametric procedures for causal effects separate the modeling of confounders from the modeling of the causal effects of interest. The use of complex confounder models thus need not complicate the interpretation of results; however, their misspecification may induce a bias in the exposure effect estimator. To enrich our understanding, we study the impact of working model misspecification in more detail for so-called G-estimators[17] and inverse probability weighted (IPW) estimators[43] under the assumption that

$$E(Y|A,L) = \omega_0^*(L) + \tau^* A$$

for some unknown function $\omega_0^*(L)$ of L, where τ^* encodes the exposure effect. For simplicity of exposition, we assume that A is a dichotomous exposure, taking values 0 and 1, and that Y is a continuous outcome. The G-estimator[17] is obtained as the solution to an estimating equation of the form

$$0 = \sum_{i=1}^{n} \{A_i - \hat{\pi}(L_i)\} \{Y_i - \tau A_i - \phi\hat{\omega}_0(L_i)\},$$

where $\hat{\omega}_0(L)$ and $\hat{\pi}(L)$ are estimates of $\omega_0^*(L)$ and the propensity score $\pi^*(L) = P(A = 1|L)$, respectively, based on possibly misspecified models. Further, ϕ is a user-specified constant. If set to 0, it yields the so-called G-estimator which is a consistent and asymptotically normal (CAN) estimator of τ^* if $\hat{\pi}(L)$ is a consistent estimator of $\pi^*(L)$ for all L. In linear models, this estimator is equivalent with the ordinary least squares estimator obtained via regression adjustment for the propensity score[5]. If set to 1, it yields the so-called doubly-robust G-estimator which is a CAN estimator of τ^* if for each L, either $\hat{\pi}(L)$ is a consistent estimator of $\pi^*(L)$ or $\hat{\omega}_0(L)$ is a consistent estimator of $\omega_0^*(L)$. Here, $\hat{\omega}_0(L)$ may be obtained via a standard (linear) regression model for $E(Y|A,L)$; $\hat{\pi}(L)$ is typically obtained via a standard logistic regression model. The

resulting (doubly-robust) G-estimator can be calculated as

$$\hat{\tau}_G(\phi) = \frac{\sum_{i=1}^{n} \{A_i - \hat{\pi}(L_i)\} \{Y_i - \phi\hat{\omega}_0(L_i)\}}{\sum_{i=1}^{n} \{1 - \hat{\pi}(L_i)\} A_i}.$$

The (doubly robust) inverse probability weighted (IPW) estimator[43] is obtained as

$$\hat{\tau}_{IPW}(\phi) = \sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(L_i)} \{Y_i - \phi\hat{\omega}_1(L_i)\} - \frac{1 - A_i}{1 - \hat{\pi}(L_i)} \{Y_i - \phi\hat{\omega}_0(L_i)\} + \phi \{\hat{\omega}_1(L_i) - \hat{\omega}_0(L_i)\},$$

where $\hat{\omega}_1(L)$ and $\hat{\omega}_0(L)$ are estimates of $E(Y|A = 1, L)$ and $E(Y|A = 0, L)$, respectively, based on possibly misspecified models. Again, ϕ is a user-specified constant. If set to 0, it yields the so-called IPW-estimator which is a CAN estimator of τ^* if $\hat{\pi}(L)$ is a consistent estimator of $\pi^*(L)$ for all L. If set to 1, it yields the so-called doubly-robust IPW-estimator which is a CAN estimator of τ^* if either $\hat{\pi}(L)$ is a consistent estimator of $\pi^*(L)$ for each L or $\hat{\omega}_j(L), j = 0, 1$ is a consistent estimator of $E(Y|A = j, L)$ for each L[58].

Over the past decade, much attention has been given to the development of doubly robust estimation procedures[6]. Facing the truth that in practice 'all' models are misspecified, the practical benefit of such doubly robust procedures has been questioned and concerns have been raised that such procedures may be very sensitive to misspecification affecting both nuisance working models[59]. We therefore evaluate the asymptotic bias (i.e., mean difference between the estimator and estimand) of the suggested G-estimators and IPW estimators under misspecification occurring in all nuisance working models. Upon using that the asymptotic bias of a root-n (asymptotically linear) estimator of τ^* with estimating function $U(\tau)$ equals $E\{\partial U(\tau^*)/\partial\tau\}^{-1} E\{U(\tau^*)\}$, we obtain asymptotic biases of

$$\frac{E[\{\pi^*(L) - \pi(L)\} \{\omega_0^*(L) - \phi\omega_0(L)\}]}{E[\{1 - \pi(L)\} \pi^*(L)]} \tag{3}$$

for the G-estimator, and

$$E\left[\left\{\frac{\pi^*(L)}{\pi(L)} - 1\right\} \{\omega_0^*(L) + \tau^* - \phi\omega_1(L)\} - \left\{\frac{1 - \pi^*(L)}{1 - \pi(L)} - 1\right\} \{\omega_0^*(L) - \phi\omega_0(L)\}\right], \tag{4}$$

for the IPW-estimator. Here, $\pi(L), \omega_1(L)$ and $\omega_0(L)$ are the probability limits of $\hat{\pi}(L), \hat{\omega}_1(L)$ and $\hat{\omega}_0(L)$, respectively.

We will first focus on the estimators that set ϕ equal to zero. These are consistent estimators of τ^* under correct specification of the propensity score, but not necessarily otherwise. It is then seen that any degree of model misspecification in the propensity score of magnitude $\delta(L) = \pi^*(L) - \pi(L)$ at a given L, yields a contribution to the bias of the G-estimator of magnitude

$$\frac{\delta(L)\omega_0^*(L)}{E\left[\{1-\pi(L)\}\,\pi^*(L)\right]} \tag{5}$$

and to the bias of the IPW-estimator of magnitude

$$\delta(L)\left[\frac{\omega_0^*(L)}{\{1-\pi(L)\}\,\pi(L)} + \frac{\tau^*}{\pi(L)}\right]. \tag{6}$$

In the absence of an exposure effect (i.e. $\tau^* = 0$), the bias contribution of the IPW-estimator is thus

$$\frac{E\left[\{1-\pi(L)\}\,\pi^*(L)\right]}{\{1-\pi(L)\}\,\pi(L)}$$

times that of the G-estimator. This ratio can be substantial within L-regions corresponding to propensity score values close to 0 or 1. Considering that such regions are typically located in the tails of the data distribution where model misspecification is more likely, we conclude that the IPW-estimator will generally be much more vulnerable than the G-estimator to misspecification of the propensity score.

Interestingly, the G-estimator can be consistent under misspecification of the propensity score model. This would happen for instance if the propensity score model were of the form $\pi(L) = \text{expit}(\alpha^{*'}L)$, α^* were estimated using a maximum likelihood procedure and $\omega_0^*(L)$ happened to be linear in L. In that case, the fitted propensity score would satisfy $0 =$

$E\left[\{A-\pi(L)\}\,\omega_0^*(L)\right] = E\left\{\delta(L)\omega_0^*(L)\right\}$, thus giving the estimating functions for the G-estimator, and in particular the bias term (5), mean zero. This is not the case for the IPW estimator when the propensity score is fitted using maximum likelihood inference as it follows from (6) that its bias due to propensity score misspecification depends on the magnitude of the exposure effect. Any cancelation of the bias of the IPW estimator under propensity score misspecification must thus be accidental in the sense of occurring only at one specific exposure effect size τ^*. Following a Bayesian argument (with an absolutely continuous prior density on τ), such cancelation occurs with zero probability. In view of this, it can be desirable to estimate the propensity score in such a way that consistency of the IPW-estimator is attained within a larger class of data-generating distributions than those that correspond to a correctly specified propensity score model. In particular, we recommend calculating the IPW-estimator as

$$\hat{\tau}_{IPW}(\phi) = \sum_{i=1}^{n} \frac{A_i}{\hat{\pi}_1(L_i)}\left\{Y_i - \phi\hat{\omega}_1(L_i)\right\} - \frac{1-A_i}{1-\hat{\pi}_0(L_i)}\left\{Y_i - \phi\hat{\omega}_0(L_i)\right\} + \phi\left\{\hat{\omega}_1(L_i) - \hat{\omega}_0(L_i)\right\},$$

where ϕ is as before, $\hat{\pi}_1(L_i)$ is a consistent estimator of $\pi(L_i)$ obtained by solving an estimating equation of the form

$$0 = \sum_{i=1}^{n}\left(\frac{A_i}{\pi(L_i)} - 1\right)\varphi(L_i), \tag{7}$$

and $\hat{\pi}_0(L_i)$ is a consistent estimator of $\pi(L_i)$ obtained by solving an estimating equation of the form

$$0 = \sum_{i=1}^{n}\left(\frac{1-A_i}{1-\pi(L_i)} - 1\right)\varphi(L_i), \tag{8}$$

where $\varphi(L_i)$ is an arbitrary index function of the dimension of α^*. Note that we use the same propensity score model, but different consistent estimators for the probability of exposure versus no exposure. Cancelation of the asymptotic bias may now occur when $\varphi(L)$ includes the constant 1 and $\omega_0^*(L)$ happens to be a linear combination of the components in the vector $\varphi(L)$. This can be seen from (4) with $\phi = 0$ upon noting that (7) and (8) then imply

$E\left[\{\pi^*(L)/\pi(L)-1\}\,\omega_0^*(L)\right] = E\left[\{(1-\pi^*(L))/(1-\pi(L))-1\}\,\omega_0^*(L)\right] = 0$. This would happen for instance if $\varphi(L) = (1, L)'$ and $\omega^*(L)$ happened to be linear in L. Note also that when $\varphi(L)$ includes the constant 1, then the asymptotic bias of the IPW-estimator is no longer dependent upon τ^* because it follows from equation (7) that $E\left\{\pi^*(L)/\pi(L)\right\}$ equals 1 in that case. In addition, the fitted propensity scores $\hat{\pi}_1(L_i)$ $(\hat{\pi}_0(L_i))$ are then such that the sum of the weights $1/\hat{\pi}_1(L_i)$ $(1/\{1-\hat{\pi}_0(L_i)\})$ in the exposed (unexposed) subjects equals the total sample size. We will therefore refer to estimation of the propensity scores following (7) and (8) as stabilized estimation. With a different attainment goal in mind, namely improving the stability of inverse weighting procedures, Cao, Tsiatis and Davidian[60] make a related, although different proposal in a missing data context.

We will now focus on doubly-robust estimators obtained by setting $\phi = 1$. It is easily seen from both bias expressions (3) and (4) that any degree of model misspecification in the propensity score of magnitude $\delta(L) = \pi^*(L) - \pi(L)$ at a given L, and in the outcome regression models of magnitudes $\Delta_1(L) = \omega_0^*(L) + \tau^* - \omega_1(L)$ and $\Delta_0(L) = \omega_0^*(L) - \omega_0(L)$ yields a contribution to the bias of the doubly-robust G-estimator of magnitude

$$\frac{\delta(L)\Delta_0(L)}{E\left[\{1-\pi(L)\}\,\pi^*(L)\right]} \tag{9}$$

and to the bias of the doubly-robust IPW-estimator of magnitude

$$\delta(L)\left[\frac{\Delta_1(L)}{\pi(L)} - \frac{\Delta_0(L)}{1-\pi(L)}\right]. \tag{10}$$

It is immediate from these expressions that the doubly-robust G- and IPW-estimator have mean zero under misspecification of one, but not both nuisance working models. These estimators will typically also have smaller bias under propensity score misspecification than the previously considered G-estimator and IPW-estimator because any misspecification of magnitude $\delta(L)$ now gets inflated only proportional to the degree of misspecification in the outcome regression model.

Interestingly, the doubly-robust G-estimator not only is consistent under the union model which correctly specifies either the propensity score or the outcome regression, but also under certain data-generating mechanisms corresponding to misspecification affecting both nuisance working models. This would occur, for instance, if the misspecified propensity score model were of the form $\pi(L) = \text{expit}(\alpha_0 + \alpha_1 L + \alpha_2 L^2)$ and fitted using maximum likelihood inference, the fitted outcome regression model were of the form $\omega(L) = \gamma_0 + \gamma_1 L$ and $\omega^*(L)$ happened to be linear in L and L^2. Indeed, in that case the fitted propensity score model would satisfy $E\left[\{A - \pi(L)\}\{\omega_0^*(L) - \omega_0(L)\}\right] = E\left\{\delta(L)\Delta_0(L)\right\} = 0$. The doubly-robust IPW-estimator with propensity scores fitted through maximum likelihood inference does not satisfy a similar property. In addition, it follows from (10) that misspecification in the regression model for $E(Y|A = 1, L)$ (or $E(Y|A = 0, L)$) can get dramatically inflated in L-regions where data on exposed subjects (on unexposed subjects) are relatively scarce. These are regions where model misspecification is also most likely, suggesting that doubly robust IPW-estimators may in fact exacerbate the extrapolation problem in view of which propensity-score adjusted estimators were designed. As a way of improving the performance of doubly robust estimators in the presence of influential weights, Robins et al.[61] proposed fitting the outcome regression models $\omega_1(L)$ and $\omega_0(L)$, respectively, via standard weighted regression in the exposed and unexposed subjects, with weights $1/\pi(L)$ and $1/\{1 - \pi(L)\}$, respectively:

$$0 = \sum_{i=1}^{n} \frac{A_i}{\hat{\pi}(L_i)} \{Y_i - \hat{\omega}_1(L_i)\} \varphi_1(L_i) \tag{11}$$

$$0 = \sum_{i=1}^{n} \left\{ \frac{1 - A_i}{1 - \hat{\pi}(L_i)} \right\} \{Y_i - \hat{\omega}_0(L_i)\} \varphi_0(L_i), \tag{12}$$

where $\varphi_1(L_i)$ and $\varphi_0(L_i)$ are arbitrary vector functions of the dimension of the unknown parameters indexing $\omega_1(L_i)$ and $\omega_0(L_i)$, and including the constant 1. They refer to the resulting doubly robust estimator of the exposure effect as a regression doubly robust estimator. The advantage of

this is clear from the fact that the above equations imply that $E\{(1+\delta(L)/\pi_1(L))\Delta_1(L)\}=0$ and $E\{(1-\delta(L)/(1-\pi_0(L)))\Delta_0(L)\}=0$ so that the asymptotic bias of the doubly-robust IPW-estimator becomes

$$\delta(L)\left[\frac{\Delta_1(L)}{\pi(L)}-\frac{\Delta_0(L)}{1-\pi(L)}\right]=-E\{\Delta_1(L)-\Delta_0(L)\}.$$

Bias due to model misspecification in the tails of the data distribution is thereby no longer inflated. Further robustness against model misspecification is attained by fitting the propensity score through equations (7) and (8), for then bias due to model misspecification cancels whenever $\Delta_1(L)$ and $\Delta_0(L)$ happen to be linear combinations of the components of $\varphi(L)$. This would occur, for instance, if the misspecified propensity score model were of the form $\pi(L)=\text{expit}(\alpha_0+\alpha_1L+\alpha_2L^2)$, $\varphi(L)=(1,L,L^2)$, the fitted outcome regression model were of the form $\omega(L)=\gamma_0+\gamma_1L$ and $\omega_0^*(L)$ happened to be linear in L and L^2.

3.3 Simulation study

In this section, we illustrate the impact of global misspecification of the nuisance working models in G-estimators and IPW-estimators through a small simulation study. In each of 5000 simulation runs, a data set of 500 independent samples was generated with L a standard normal variate. In the first experiment, $Y=-2+A+2L+N(0,1)$ and $\pi^*(L)=\text{expit}(-3+L)$. In the next 4 experiments, $Y=-2+A+2L-L^2+N(0,1)$, with $\pi^*(L)=\text{expit}(-4+1.5\sqrt{|L|}+0.75L+0.5|L|^{1.5})$ in the second experiment, $\pi^*(L)=\text{expit}(-2+2\sin(2L))$ in the third experiment, and $\pi^*(L)=\text{expit}(-0.5+\sin(2L)-0.5\cos(3L)-0.25L^2)$ in the fourth and fifth experiment. In all experiments, linear outcome working models were used. Second and third order logistic propensity score working models were used in the first three and last two experiments, respectively. Table 2 shows the bias and empirical standard deviation of the ordinary least squares estimates with (OLS-A) and without (OLS-U) adjustment for L, the G-estimator (G), the IPW-estimator

with (IPW-S) and without (IPW) stabilized estimation of the propensity score, the regression doubly robust IPW estimator with (RDR-S) and without (RDR) stabilized estimation of the propensity score and the doubly robust IPW estimator with maximum likelihood estimation of the outcome working model and with (DR-S) and without (DR) stabilized estimation of the propensity score. The results demonstrate that in the absence of model misspecification (i.e. simulation experiment 1), stabilized estimation of the propensity score improves the finite-sample bias of the IPW estimator and yields a minor efficieny gain, although an efficiency loss for the doubly robust estimators. In the presence of model misspecification, major improvements in both the bias and precision of (doubly robust) IPW estimators are observed. In particular, for the considered data-generating mechanisms, no bias was observed despite all working models being misspecified. The fourth and fifth experiment used the same data generating models, but $\varphi(L) = (1, L, |L|^{1.5}, L^2)$ in (7) and (8) in the fourth experiment and $\varphi(L) = (1, L, |L|^2, L^3)$ in the fifth experiment.

Table 2 about here.

4 Discussion

Modern procedures for marginal causal effects (see e.g. Section 3.2) require working models for the outcome and/or exposure, but their complexity does not affect the interpretability of the final effect estimand. The desire to use parsimonious models is therefore not so much stimulated by the need for obtaining interpretable results, but rather by concerns of bias and inefficiency which may result from overfitting. Two caveats are in place, however. First, while the possibility of bias resulting from overfitting is well understood for conditional effects (cfr. the Neyman-Scott paradox), to the best of our knowledge, the extent to which it affects the estimation of marginal effects remains to be evaluated. Second, it has been documented that efficiency gains

may be realized when a priori knowledge is available that given covariates are only associated with the exposure, but have no residual association with outcome[25]. However, in practice, such a priori information is rarely, if ever, available. Without such information, data-driven decisions must be made to exclude covariates from the analysis and it is unclear under what conditions the additional uncertainty induced by these selection approaches still enables a meaningful efficiency gain.

Most strategies used by practitioners to select confounders are based on excluding potential confounders from the analysis when they are non-significantly associated with the outcome conditional on the exposure; some focus on associations with the exposure instead. Such strategies are sub-optimal for various reasons. First, since confounders are by definition jointly associated with exposure and outcome, the importance of a variable as a confounder must ideally be judged against criteria that involve both associations. Second, even when for a given variable both associations are assessed, their significance is not directly informative about the extent to which adjusting for this variable will reduce confounding bias and, ultimately, improve the quality of the exposure effect estimator. Third, even when a more rigorous confounder-selection process is adopted, it remains difficult to acknowledge the uncertainty resulting from the selection process into the final inference. By ignoring this, one risks to obtain under-covering confidence intervals.

We have attempted to shed light on these issues and proposed a focused confounder-selection strategy which aims at minimum mean squared error of the exposure effect estimator. This strategy is closely linked to one recommended in Brookhart and van der Laan[37], but computationally more attractive by avoiding the use of cross-validation. Its application overcomes the aforementioned first two concerns. In particular, when applied to estimators that are consistent under correct specification of a propensity score model, we expect it will overcome the usual difficulties[26] in selecting confounders in the propensity score model as the selection is made

in terms of an 'optimal' trade-off between bias and efficiency of the exposure effect estimate and thus will have a tendency to 'automatically' exclude covariates that are solely associated with the exposure and include covariates that are solely associated with the outcome. For such estimators, as shown in Section 2.6, it also roughly overcomes the third concern in the sense of retaining confidence validity even when the confounder-selection process is ignored. In spite of these attractions of focused confounder-selection based on propensity-score adjusted estimators, several limitations remain and warrant further study. First, the calculation of the mean squared error relies on estimates obtained from a full model which involves all potential confounders. Simulation studies are needed to evaluate finite-sample performance when these estimates are inefficient or biased as a result of overfitting. Second, in small samples, the procedure may choose to exclude potentially important confounders in order to reduce mean squared error at the expense of a bias, whose magnitude is difficult to assess. Stability plots like Figure 2 may help detect whether this occurs; one may use them, for instance, to restrict the procedure to all submodels that do not generate a bias exceeding a scientifically meaningful magnitude.

Given the aforementioned caveats and limitations of variable-selection, we see much value in the idea of avoiding confounder-selection by using regularization techniques such as ridge regression instead. This idea has been much advocated by Sander Greenland[13,15]. Further research is needed to evaluate these contrasting viewpoints in realistic settings involving unmeasured confounding, missing confounder data and large separation in the confounder distributions of exposed and unexposed subjects. Perhaps the ideal future lies in an approach whereby the nuisance parameters indexing the working models for the association between covariates on the one hand, and exposure and outcome on the other hand, are estimated as those values that minimize the mean squared error of the exposure effect estimator. Such approach would combine the benefits of focused confounder-selection and regularization approaches that do not involve selection, and might improve upon them in various ways. In comparison with confounder-selection

approaches, it would further lower the mean squared error by not being restricted to specific submodels and, by avoiding repeated model fitting, might enable a more easy assessment of the overall uncertainty. In comparison with approaches that involve no selection, it would have the advantage of directly targeting minimal mean squared error of the exposure effect estimator. It is unclear at present whether such approach is attainable.

Acknowledgements

We are very grateful to Miguel Hernán for stimulating us to write on this topic, and to two anonymous referees for very helpful comments. The authors acknowledge support from IAP research network grant nr. P06/03 from the Belgian government (Belgian Science Policy). The second author acknowledges support from the Institute for the Promotion of Innovation through Science and Technology in Flanders (IWT-Vlaanderen).

References

[1] Hand DJ, Vinciotti V. Local versus global models for classification problems: fitting models where it matters. The American Statistician. 2003;57:124–131.

[2] Claeskens G, Hjort NL. The focused information criterion. Journal of the American Statistical Association. 2003;98:900–916. With discussion and a rejoinder by the authors.

[3] Crainiceanu CM, Dominici F, Parmigiani G. Adjustment uncertainty in effect estimation. Biometrika. 2008;95:635–651.

[4] Rubin DB. Estimating causal effects from large data sets using propensity scores. Annals of Internal Medicine. 1997;127:757–763.

[5] Rosenbaum PR, Rubin DB. The central role of the propensity score in observational studies for causal effects. Biometrika. 1983;70:41–55.

[6] Robins JM, Rotnitzky A. Comment on a paper by P. Bickel and J. Kwon. Statistica Sinica. 2001;11:920–936.

[7] Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. Epidemiology. 1999;10:37–48.

[8] Robins JM. Data, design, and background knowledge in etiologic inference. Epidemiology. 2001;12:313–320.

[9] Rosenbaum PR. The consequences of adjustment for a concomitant variable that has been affected by the treatment. Journal of the Royal Statistical Society, Series A. 1984;147:656–666.

[10] Pearl J. Causality: Models, Reasoning and Inference. Cambridge: Cambridge University Press; 2009.

[11] Schisterman EF, Cole SR, Platt RW. Overadjustment Bias and Unnecessary Adjustment in Epidemiologic Studies. Epidemiology. 2009;20:488–495.

[12] Hernan MA, Hernandez-Diaz S, Werler MM, Mitchell AA. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. American Journal of Epidemiology. 2002;155:176–184.

[13] Greenland S. Variable Selection versus Shrinkage in the Control of Multiple Confounders. American Journal of Epidemiology. 2008;167:523–529.

[14] D'Agostino R Jr. Propensity score methods for bias reduction in the comparison treatment to a non-randomized control group. Statistics in Medicine. 1998;17:2265–2281.

[15] Greenland S. Bayesian perspectives for epidemiological research. II. Regression analysis. International Journal of Epidemiology. 2007;36:195–202.

[16] Budtz-Jorgensen E, Keiding N, Grandjean P, Weihe P. Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure. Annals of Epidemiology. 2007;17:27–35.

[17] Robins JM, Mark SD, K NW. Estimating exposure effects by modelling the expectation of exposure conditional on confounders. Biometrics. 1992;48:479–495.

[18] Vansteelandt S, Carpenter J, Kenward MG. Analysis of incomplete data using inverse probability weighting and doubly robust estimators. Methodology. 2010;6:37–48.

[19] Greenland S. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. Epidemiology. 2003;13:300–306.

[20] Pearl J. Remarks on the method of propensity score. Statistics in Medicine. 2009;28:1415–1416.

[21] Sjolander A. Propensity scores and M-structures. Statistics in Medicine. 2009;28:1416–1420.

[22] Wooldridge J. Should instrumental variables be used as matching variables? Michigan State University; 2009.

[23] Pearl J. On a Class of Bias-Amplifying Covariates that Endanger Effect Estimates. In: Grunwald P, Spirtes P, editors. Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence;. .

[24] Greenland S, Robins JM, Pearl J. Confounding and collapsibility in causal inference. Statistical Science. 1999;14:29–46.

[25] Hahn J. Functional Restriction and Efficiency in Causal Inference. The Review of Economics and Statistics. 2004;86:73–76.

[26] Brookhart MA, Schneeweiss S, Rothman KJ, Glynn RJ, Avorn J, Sturmer T. Variable selection for propensity score models. American Journal of Epidemiology. 2006;163:1149–1156.

[27] Claeskens G, Hjort NL. Model Selection and Model Averaging. Cambridge: Cambridge University Press; 2008.

[28] Mickey RM, Greenland S. The impact of confounder selection criteria on effect estimation. American Journal of Epidemiology. 1989;129:125–137.

[29] Maldonado G, Greenland S. Simulation study of confounder-selection strategies. American Journal of Epidemiology. 1993;138:923–936.

[30] Wright S. The method of path coefficients. Annals of Mathematical Statistics. 1934;5:161–215.

[31] Greenland S, Neutra RR. Control of confounding in the assessment of medical technology. International Journal of Epidemiology. 1980;9:361–367.

[32] Tan Z. Understanding OR, PS, and DR. Statistical Science. 2008;22:560–568.

[33] Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996;58:267–288.

[34] Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, Series B. 2005;67:301–320.

[35] De Luna X, Richardson TS, Waernbaum I. Covariate selection for the non-parametric estimation of an average treatment effect. Umea University; 2010.

[36] Greenland S. Modeling and Variable Selection in Epidemiologic Analysis. American Journal of Public Health. 1986;79:340–349.

[37] Brookhart MA, van der Laan MJ. A semiparametric model selection criterion with applications to the marginal structural model. Computational Statistics & Data Analysis. 2006;50(2):475–498.

[38] Claeskens G, Croux C, Van Kerckhoven J. Variable selection for logistic regression using a prediction focussed information criterion. Biometrics. 2006;62:972–979.

[39] Mortimer KM, Neugebauer R, van der Laan M, Tager IB. An application of model-fitting procedures for marginal structural models. American Journal of Epidemiology. 2005;162:382–388.

[40] Haight TJ, Wang Y, van der Laan MJ, Tager IB. A cross-validation-deletion,-substitution,-addition model selection algorithm: Application to marginal structural models. Computational Statistics and Data Analysis. 2010;In press.

[41] Hjort NL, Claeskens G. Frequentist model average estimators. Journal of the American Statistical Association. 2003;98:879–899. With discussion and a rejoinder by the authors.

[42] Connors AF, Speroff T, Dawson NV, Thomas C, Harrell F, Wagner D, et al. The Effectiveness of Right Heart Catheterization in the Initial Care of Critically Ill Patients. Journal of the American Medical Association. 1996;11:889–897.

[43] Robins JM, Hernan MA, Brumback B. Marginal structural models and causal inference in epidemiology. Epidemiology. 2000;11:550–560.

[44] Pearl J. Graphs, Causality, and Structural Equation Models. Sociological Methods and Research. 1998;27:226–284.

[45] Robins JM. Causal inference from complex longitudinal data. In: Latent variable modeling and applications to causality (Los Angeles, CA, 1994). vol. 120 of Lecture Notes in Statist. New York: Springer; 1997. p. 69–117.

[46] Robins JM, Greenland S. The role of model selection in causal inference from nonexperimental data. American Journal of Epidemiology. 1986;123:392–402.

[47] Robins JM, Ritov Y. Toward a curse of dimensionality appropriate (CODA) asymptotic theory for semi-parametric models. Statistics in Medicine. 1997;16:285–319.

[48] Vansteelandt S, VanderWeele T, Tchetgen EJ, Robins JM. Multiply robust inference for statistical interactions. Journal of the American Statistical Association. 2008;103:1693–1704.

[49] Rosenblum M, van der Laan MJ. Using Regression Models to Analyze Randomized Trials: Asymptotically Valid Hypothesis Tests Despite Incorrectly Specified Models. Biometrics. 2009;65:937–945.

[50] Scharfstein DO, Rotnitzky A, Robins JM. Adjusting for non-ignorable drop-out using semiparametric non-response models. Journal of the American Statistical Association. 1999;94:1121–1146. With discussion and a rejoinder by the authors.

[51] Little RJ. A Note About Models for Selectivity Bias. Econometrica. 1985;53:1469–1474.

[52] Vansteelandt S. Discussion on 'Identifiability and Estimation of Causal Effects in Randomized Trials with Noncompliance and Completely Non-ignorable Missing-Data'. Biometrics. 2009;65:686–689.

[53] Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical

Modelling. 1986;7:1393–1512. Mathematical models in medicine: diseases and epidemics, Part 2.

[54] Robins JM, Greenland S. Identifiability and exchangeability for direct and indirect effects. Epidemiology. 1992;3:143–155.

[55] Robins JM. Testing and estimation of direct effects by reparameterizing directed acyclic graphs with structural nested models. In: Computation, causation, and discovery. Menlo Park, CA: AAAI Press; 1999. p. 349–405.

[56] VanderWeele TJ. Marginal structural models for the estimation of direct and indirect effects. Epidemiology. 2009;20:18–26.

[57] Vansteelandt S. Estimating direct effects in cohort and case-control studies. Epidemiology. 2009;20:851–860.

[58] Leon S, Tsiatis AA, Davidian M. Semiparametric estimation of treatment effect in a pretest-posttest study. Biometrics. 2003;59:1046–1055.

[59] Kang JDY, Schafer JL. Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. Statistical Science. 2008;22:523–539.

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

[61] Robins JM, Sued M, Lei-Gomez Q, Rotnitzky A. Performance of double-robust estimators when 'inverse probability' weights are highly variable. Statistical Science. 2008;22:544–559.

[62] Tsiatis AA. Semiparametric Theory and Missing Data. New York: Springer; 2006.

Appendix

Assessment of M-bias and bias amplification

Consider the path diagram in Figure 1. Let Y^*, A^* and L^* denote standardized[30] variables corresponding to Y, A and L, respectively. Assume that $E(Y^*|A^*, L^*, U_2, U_3) = cA^* + bL^* + c_{2y}U_2 + c_{3y}U_3$, then $E(Y^*|A^*, L^*) = cA^* + bL^* + c_{2y}E(U_2|A^*, L^*) + c_{3y}E(U_3|A^*, L^*)$. Let $E(U_2|A^*, L^*) = \alpha_2 L^* + \beta_2 A^*$, $E(L^*|U_1, U_2) = c_{1l}U_1 + c_{2l}U_2$ and $E(A^*|L^*, U_1, U_3) = c_{1a}U_1 + c_{3a}U_3 + aL^*$ Then, proceeding as in Pearl[23], we have $E(U_2L^*) \equiv c_{2l} = \alpha_2 + \beta_2\rho_{al}$ and $E(U_2A^*) \equiv c_{2l}a = \alpha_2\rho_{al} + \beta_2$, where $\rho_{al} = a + c_{1a}c_{1l}$, from which

$$\alpha_2 = c_{2l}\frac{(1 - \rho_{al}^2 + c_{1a}c_{1l}\rho_{al})}{1 - \rho_{al}^2} \ , \ \beta_2 = -c_{2l}\frac{c_{1a}c_{1l}}{1 - \rho_{al}^2}.$$

Likewise, $E(U_3|A^*, L^*) = -c_{3a}\rho_{al}/(1 - \rho_{al}^2)L^* + c_{3a}/(1 - \rho_{al}^2)A^*$. It follows that

$$\begin{aligned} E(Y^*|A^*, L^*) &= \left(b + c_{2y}c_{2l}\frac{(1 - \rho_{al}^2 + c_{1a}c_{1l}\rho_{al})}{1 - \rho_{al}^2} - \frac{c_{3a}c_{3y}\rho_{al}}{1 - \rho_{al}^2}\right)L^* \\ &\quad + \left(c - c_{2y}c_{2l}\frac{c_{1a}c_{1l}}{1 - \rho_{al}^2} + \frac{c_{3y}c_{3a}}{1 - \rho_{al}^2}\right)A^*, \end{aligned}$$

and $E(Y^*|A^*) = \{(b + c_{2y}c_{2l})\rho_{al} + c - c_{2y}c_{2l}c_{1a}c_{1l} + c_{3y}c_{3a}\}A^*$. The bias reported in the main text is the difference between the coefficient joining A^* in the above expressions, and the population causal effect c. It is easy to demonstrate that inverse weighting by $1/f(A^*|L^*)$ yields an exposure-outcome covariance equal to

$$\int YAf(Y|A, L)f(L)dYdAdL = \left(c - c_{2y}c_{2l}\frac{c_{1a}c_{1l}}{1 - \rho_{al}^2} + \frac{c_{3y}c_{3a}}{1 - \rho_{al}^2}\right).$$

In Figure 4, we develop a better understanding of the magnitude of these biases under the assumption that ρ_2, as defined in the main text, is at most ρ_{al}. We make this assumption to respect that, arguably, U_3 will have weaker correlations with exposure and outcome than L when the focus of the study is on assessing the effect of A on Y, as efforts have then been targeted

at collecting data on common causes of exposure and outcome. We make a similar assumption for ρ_1 to respect the fact that ρ_1 indirectly contributes to the magnitude of ρ_{al}. The solid line in Figure 4 displays the upper bound (1) in a setting where $\rho_1 = \rho_{al}/2$ and $\rho_2 = \rho_{al}/3$. It shows that the adjusted analysis will only be more biased than the unadjusted analysis when the correlation between A and L is extremely large and the correlation between Y and L is extremely small. We believe this is unlikely to occur in practice. The figure further suggests that, under the considered scenario, the impact of M-bias (see bottom line in Figure 2) is not much less sizeable than that of unmeasured confounding (see top line in Figure 4), although only of importance for exposure-confounder correlations exceeding 0.5.

Figure 4 about here.

FIC-based confounder selection

We consider the marginal log odds ratio as a focus parameter, which we define as

$$\tau^* = \log \frac{\mu_1(1-\mu_0)}{\mu_0(1-\mu_1)}$$

where $\mu_a = E\left[\text{expit}\left\{\omega(L;\gamma^*) + \beta_0^* + \beta_a^* a\right\}\right]$ for $a = 0, 1$. Denote furthermore

$$\hat{\mu}_a = n^{-1} \sum_{i=1}^{n} \left[\text{expit}\left\{\omega(L_i;\hat{\gamma}) + \hat{\beta}_0 + \hat{\beta}_a a\right\}\right]$$

for $a = 0, 1$. Assume, as in Claeskens and Hjort[2], that the true data density $f(Y, A|L)$ is indexed by a parameter $\beta^* = (\beta_0^*, \beta_a^*)'$, which is shared between all models, and $\gamma^* + \delta/\sqrt{n}$, where the term $\delta/\sqrt{n}$ encodes local model misspecification (see Section 2.6) and γ^* is the vector of values to which the nuisance parameter γ is set in the narrow model (that is, typically $\gamma^* = 0$). Let further $\theta_S \equiv (\beta, \gamma_S)'$ and $\hat{\theta}_S \equiv (\hat{\beta}_S, \hat{\gamma}_S)'$. Then we have that for any submodel S, the corresponding estimator $\hat{\mu}_{Sa}$ of μ_a^* (which is defined like $\hat{\mu}_a$, but with $\hat{\gamma}_S$ and $\hat{\beta}_S$ replacing $\hat{\gamma}$ and $\hat{\beta}$, respectively)

satisfies

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{expit}\left\{\omega(L_i;\hat{\gamma}_S,\gamma^*_{-S})+\hat{\beta}_{S0}+\hat{\beta}_{Sa}a\right\}-\sqrt{n}\hat{\mu}_{Sa} \\
&= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*a\right\}-\sqrt{n}\mu_a^* \\
&\quad +E\left[\frac{\partial}{\partial\theta_S}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*a\right\}\right]\sqrt{n}\left(\hat{\theta}_S-\theta_S^*\right) \\
&\quad -E\left[\frac{\partial}{\partial\gamma}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*a\right\}\right]\delta-\sqrt{n}(\hat{\mu}_{Sa}-\mu_a^*)+o_p(1)
\end{aligned}$$

from which

$$\begin{aligned}
\sqrt{n}(\hat{\mu}_{Sa}-\mu_a^*) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*a\right\}-\mu_a^* \\
&\quad +E\left[\frac{\partial}{\partial\theta_S}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*a\right\}\right]\sqrt{n}\left(\hat{\theta}_S-\theta_S^*\right) \\
&\quad -E\left[\frac{\partial}{\partial\gamma}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*a\right\}\right]\delta+o_p(1).
\end{aligned}$$

It follows from the Delta method that the influence function[62] for $\hat{\tau}_S$ is $D_\mu + d_\beta\sqrt{n}(\hat{\beta}_S-\beta^*)+ d_{\gamma_S}\sqrt{n}(\hat{\gamma}_S-\gamma_S^*)-d_\gamma\delta$, where

$$\begin{aligned}
d_\gamma &= \frac{1}{\mu_1^*(1-\mu_1^*)}E\left[\frac{\partial}{\partial\gamma}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*\right\}\right] \\
&\quad -\frac{1}{\mu_0^*(1-\mu_0^*)}E\left[\frac{\partial}{\partial\gamma}\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*\right\}\right] \\
D_\mu &= \frac{1}{\mu_1^*(1-\mu_1^*)}\left[\text{expit}\left\{\omega(L_i;\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*+\beta_a^*\right\}-\mu_1^*\right] \\
&\quad -\frac{1}{\mu_0^*(1-\mu_0^*)}\left[\text{expit}\left\{\omega(L_i;\gamma^*+\delta/\sqrt{n})+\beta_0^*\right\}-\mu_0^*\right],
\end{aligned}$$

and where d_β and d_{γ_S} are defined like d_γ, but with derivatives taken w.r.t. β and γ_S, respectively, rather than γ. Using Lemmas 3.2 and 3.3 in Hjort and Claeskens[41], it can be shown that

$\sqrt{n}(\hat{\tau}_S - \tau^*)$ converges in distribution to $\Lambda_0 + \omega'(\delta - G_S D)$, where

$$\begin{aligned}
\Lambda_0 &= d_\beta J_{00}^{-1} M' + D_\mu \\
D &= \delta + Q(N' - J_{10} J_{00}^{-1} M') \\
Q &= \left(J_{11} - J_{10} J_{00}^{-1} J_{01}\right)^{-1} \\
\omega &= J_{10} J_{00}^{-1} d_\beta - d_\gamma \\
G_S &= \pi_S \left\{\pi_S' Q^{-1} \pi_S\right\}^{-1} \pi_S Q^{-1}
\end{aligned}$$

with π_S the projection matrix for submodel S (i.e., a matrix of zeros with as many rows and columns as the dimensions of γ_S and γ, respectively, and with a 1 on each row in the column representing the corresponding component of γ_S), (M', N') following a mean zero normal distribution with covariance matrix

$$J = \begin{pmatrix} J_{00} & J_{01} \\ J_{10} & J_{11} \end{pmatrix},$$

which is 1 over n times the inverse of the asymptotic covariance matrix of $\hat{\theta} = (\hat{\beta}, \hat{\gamma})'$. These expressions rely on $\hat{\theta}$ being a maximum likelihood estimator. Further, Λ_0 can be shown to be uncorrelated with D because M is independent of D by Lemma 3.3[41], and because D_μ is uncorrelated with D by the fact that (a) $Q(N' - J_{10} J_{00}^{-1} M')$ is the asymptotic distribution of $\sqrt{n}(\hat{\gamma} - \gamma^* - \delta/\sqrt{n})$; and that (b) the influence functions of $\hat{\gamma}$ are uncorrelated with D_μ by the fact that the former have mean zero conditional on L, whilst the latter are functions of L. It now follows that $\sqrt{n}(\hat{\tau}_S - \tau^*)$ has limiting mean squared error given by $\mathrm{Var}(\Lambda_0) + \omega' G_S Q G_S' \omega + \omega'(I - G_S)\delta\delta'(I - G_S)'\omega$. Upon substituting[27] $\delta\delta'$ with $\max(0, D_n D_n' - \hat{Q})$, where $D_n = \sqrt{n}(\hat{\gamma} - \gamma^*)$, we obtain

$$\mathrm{Var}(\Lambda_0) - \omega' Q \omega + 2\omega' G_S Q G_S' \omega + \omega'(I - G_S) D_n D_n' (I - G_S)' \omega.$$

Because the first two terms are common to all models, we employ the remaining terms as a Focused Information Criterion[2,27], upon substituting all population values with consistent estimates.

Model uncertainty

Let $U(\tau, \alpha)$ be the estimating function for τ and $S_S(\alpha) = \partial \log f(A|L;\alpha)/\partial \alpha_S$, where α_S is the subvector of α which is free under model S and α_{-S} is the remaining part. Let $\hat{\tau}_S$ denote the estimator of τ^* as obtained under model S, and $\sum_{S\in\mathcal{A}} c(S|D_n)\hat{\tau}_S$ denote the estimator of τ^* obtained under model selection, where the weight $c(S|D_n)$ assigns 1 to the selected model and 0 to all other models and where $\mathcal{A}$ denotes the model space. Under the local misspecification assumption, we have that[2]

$$\sqrt{n}\left\{\sum_{S\in\mathcal{A}} c(S|D_n)\hat{\tau}_S - \tau^*\right\} \xrightarrow{d} \sum_{S\in\mathcal{A}} c(S|D)\Lambda_S,$$

where Λ_S is the limit distribution of $\sqrt{n}\,(\hat{\tau}_S - \tau^*)$. Under this assumption, a Taylor series expansion shows that

$$\begin{aligned}
0 &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\hat{\tau}_S, \hat{\alpha}_S, \alpha^*_{-S}) \\
&= \left\{\frac{1}{\sqrt{n}}\sum_{i=1}^{n} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right\} + E\left(\frac{\partial}{\partial\tau}U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)\sqrt{n}\,(\hat{\tau}_S - \tau^*) \\
&\quad + E\left(\frac{\partial}{\partial\alpha_S}U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)\sqrt{n}\,(\hat{\alpha}_S - \alpha^*_{Sn}) - E\left(\frac{\partial}{\partial\alpha_{-S}}U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)\delta_{-S} + o_p(1),
\end{aligned}$$

where δ_{-S} is the subvector of δ corresponding to α_{-S} and $\alpha_{Sn} = \alpha_S + \delta_S/\sqrt{n}$ and $\alpha_{-Sn} = \alpha_{-S} + \delta_{-S}/\sqrt{n}$. Likewise, we have that

$$\begin{aligned} 0 &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{Si}(\hat{\alpha}_S, \alpha^*_{-S}) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^{n} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn}) + E\left(\frac{\partial}{\partial \alpha_S} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right) \sqrt{n}\,(\hat{\alpha}_S - \alpha^*_{Sn}) \\ &\quad -E\left(\frac{\partial}{\partial \alpha_{-S}} S_{Si}(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right) \delta_{-S} + o_p(1), \end{aligned}$$

from which

$$\begin{aligned} \sqrt{n}\,(\hat{\tau}_S - \tau^*) &= E\left(\frac{\partial}{\partial \tau} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)^{-1} \left[-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right.\right. \\ &\quad -E\left(\frac{\partial}{\partial \alpha_S} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right) E\left(\frac{\partial}{\partial \alpha_S} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right)^{-1} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn}) \Big\} \\ &\quad +\delta_{-S} \left\{E\left(\frac{\partial}{\partial \alpha_{-S}} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right) - E\left(\frac{\partial}{\partial \alpha_S} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)\right. \\ &\quad \left.\left. \times E\left(\frac{\partial}{\partial \alpha_S} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right)^{-1} E\left(\frac{\partial}{\partial \alpha_{-S}} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right) \right\}\right] + o_p(1). \end{aligned}$$

Further, $E\left\{U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right\} = 0$ implies that $E\left\{\partial U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})/\partial \alpha_S\right\}$ equals $-E\left\{U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn}) S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right\}$ and likewise for $S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})$. We thus find that

$$\begin{aligned} \sqrt{n}\,(\hat{\tau}_S - \tau^*) &= E\left(\frac{\partial}{\partial \tau} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)^{-1} \left[-\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left\{U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right.\right. \\ &\quad -E\left\{U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn}) S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right\} E\left\{S^{\otimes 2}_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right\}^{-1} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn}) \Big\} \\ &\quad +\delta_{-S} \left\{E\left(\frac{\partial}{\partial \alpha_{-S}} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right) - E\left(\frac{\partial}{\partial \alpha_S} U_i(\tau^*, \alpha^*_{Sn}, \alpha^*_{-Sn})\right)\right. \\ &\quad \left.\left. \times E\left(\frac{\partial}{\partial \alpha_S} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right)^{-1} E\left(\frac{\partial}{\partial \alpha_{-S}} S_{Si}(\alpha^*_{Sn}, \alpha^*_{-Sn})\right) \right\}\right] + o_p(1), \end{aligned}$$

where for an arbitrary matrix, $A^{\otimes 2} \equiv AA'$. It then follows that $\sqrt{n}\left\{\sum_{S\in\mathcal{A}} c(S|D_n)\hat{\tau}_S - \tau^*\right\}$ is

$$E\left(\frac{\partial}{\partial\tau}U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn})\right)^{-1}\left[-\sum_{S\in\mathcal{A}} c(S|D)\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left\{U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn})\right.\right.$$
$$\left.-E\left\{U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn})S_{Si}(\alpha^*_{Sn},\alpha^*_{-Sn})\right\}E\left\{S^{\otimes 2}_{Si}(\alpha^*_{Sn},\alpha^*_{-Sn})\right\}^{-1}S_{Si}(\alpha^*_{Sn},\alpha^*_{-Sn})\right\}$$
$$+\sum_{S\in\mathcal{A}} c(S|D)\delta_{-S}\left\{E\left(\frac{\partial}{\partial\alpha_{-S}}U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn})\right)-E\left(\frac{\partial}{\partial\alpha_S}U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn})\right)\right.$$
$$\left.\left.\times E\left(\frac{\partial}{\partial\alpha_S}S_{Si}(\alpha^*_{Sn},\alpha^*_{-Sn})\right)^{-1}E\left(\frac{\partial}{\partial\alpha_{-S}}S_{Si}(\alpha^*_{Sn},\alpha^*_{-Sn})\right)\right\}\right]+o_p(1).$$

It now follows by the Cauchy-Schwarz inequality that an upper bound to the asymptotic variance of $\sqrt{n}\left\{\sum_{S\in\mathcal{A}} c(S|D_n)\hat{\tau}_S - \tau^*\right\}$ is the variance of

$$E\left(\frac{\partial}{\partial\tau}U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn})\right)^{-1}\frac{1}{\sqrt{n}}\sum_{i=1}^{n}U_i(\tau^*,\alpha^*_{Sn},\alpha^*_{-Sn}).$$

It does not immediately follow that standard confidence intervals based on this conservative variance estimate will themselves be conservative. This is because $\sum_{S\in\mathcal{A}} c(S|D_n)\hat{\tau}_S$ follows a mixture distribution with bias components converging at root-n rate to zero. Because misspecifications δ of the order 1 over root-n are not consistently estimable[2], there is little room for further correcting this, unless for instance a doubly robust estimator with correctly specified nuisance outcome working model happens to be used, in which case uncertainty in the propensity score model does not affect inferences for τ^*.

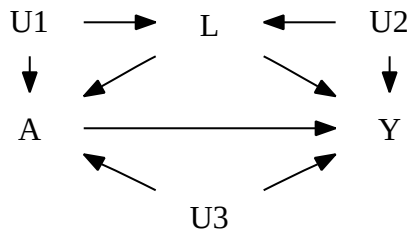

Figure 1: Causal diagram with measured variables A, L and Y, and with $U1, U2$ and $U3$ unmeasured variables.

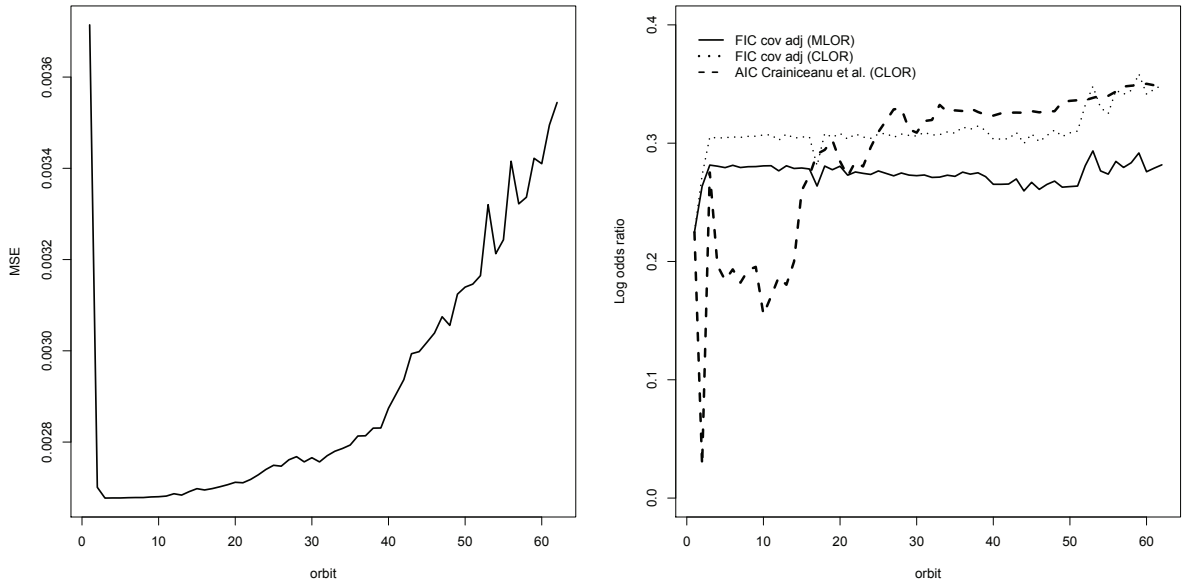

Figure 2: Left: Mean squared error (MSE) of the best model within each orbit which is obtained by minimizing the mean squared error of the marginal log odds ratio (MLOR) Right: Estimates of the marginal and conditional log odds ratio as obtained through FIC-based covariate adjustment, and through AIC-based selection as in Crainiceanu et al.[3].

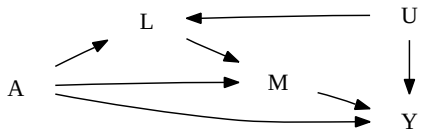

Figure 3: Causal diagram with measured variables A, L, M and Y, and with U an unmeasured confounder of the L-Y relationship.

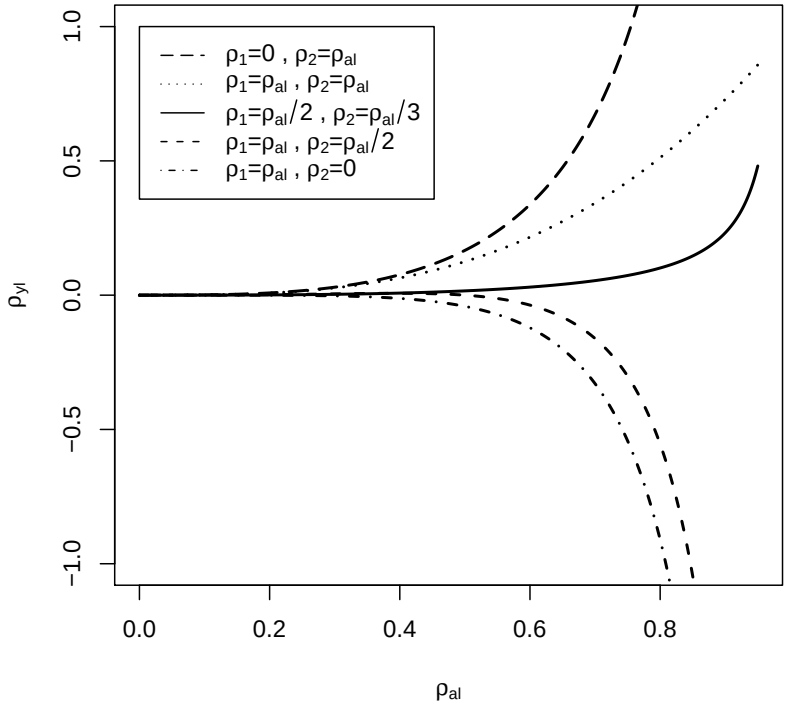

Figure 4: Values of ρ_{yl} below which the adjusted analysis has larger bias than the unadjusted analysis.

Table 1: Estimates of the effect of RHC on mortality, as obtained using different confounder-selection techniques and reported in terms of the conditional odds ratio (COR), the marginal odds ratio (MOR) with 95% confidence interval, MSE (mean squared error) and the FIC (focused information criterion).

Model selection technique	# covariates	COR	MOR	95% CI	MSE	FIC
Unadjusted analysis	0	1.25	1.25	[1.14 to 1.37]	0.0037	5.99
Full model	61	1.42	1.32	[1.18 to 1.49]	0.0035	5.01
BE covariate adjustment	15	1.39	1.31	[1.17 to 1.46]	0.0033	3.87
AIC (Crainiceanu et al.)	47	1.42	1.33	[1.18 to 1.49]	0.0035	4.94
FIC covariate adjustment	2	1.36	1.33	[1.21 to 1.46]	0.0027	0.04

Table 2: Simulation results: empirical bias and standard deviation in 5 simulation experiments.

	Exp 1		Exp 2		Exp 3		Exp 4		Exp 5	
Estimator	Bias	SD	Bias	SD	Bias	SD	Bias	SD	Bias	SD
OLS-U	1.856	0.39	0.503	0.37	0.821	0.34	0.974	0.23	0.974	0.23
OLS-A	-0.002	0.18	-1.685	0.32	0.000	0.22	0.308	0.16	0.308	0.16
G	-0.009	0.2	0.002	0.16	-0.154	0.23	-0.063	0.11	-0.063	0.11
IPW	0.151	0.45	-0.285	0.32	-1.716	1.86	-0.123	0.82	-0.123	0.82
RDR	-0.005	0.28	-0.126	0.24	-0.363	0.20	0.014	0.30	0.014	0.30
DR	-0.006	0.29	-0.115	0.86	-4.863	122.51	-1.879	43.72	-1.879	43.72
IPW-S	0.028	0.42	0.003	0.19	0.361	0.50	0.027	0.13	0.043	0.15
RDR-S	-0.001	0.37	-0.027	0.21	-0.019	0.19	0.023	0.12	-0.004	0.12
DR-S	-0.001	0.37	-0.027	0.21	-0.044	0.18	0.022	0.12	0.000	0.12