Published in final edited form as:

Biometrics. 2018 June ; 74(2): 403–406. doi:10.1111/biom.12784.

Discussion of “Data-driven confounder selection via Markov and Bayesian networks” by Häggström

Thomas S. Richardson^{1,*}, James M. Robins^{2,**}, and Linbo Wang^{2,***}

¹Department of Statistics, University of Washington, Seattle, WA 98195, U.S.A

²Department of Biostatistics, Harvard School of Public Health, 677 Huntington Ave., Boston, MA 02115, U.S.A

Keywords

Causal inference; Generalized back-door criterion; M-bias; Partial ancestral graph; Variable selection

We are grateful for the opportunity to discuss this article. Selecting the appropriate set of covariates to adjust for when estimating causal effects in observational studies is an important methodological problem. A number of criteria have been proposed and the paper makes an important contribution by comparing and contrasting these methods via theoretical results, simulation and a data analysis.

In what follows, we discuss an alternative way to represent counterfactual causal models with graphs, different approaches for selecting sets to adjust for and make some comments on the specific data example. We do not discuss the important problem of inference after model selection in our commentary.

1. Formalisms for causal inference: Combining graphs and counterfactuals

The author uses graphs that include potential outcome (aka counterfactual) variables, such as $Y(t)$ in Figure 1 and 2 (in the main paper). These DAGs serve as a means of representing the ignorability conditions. Though this follows from the approach in De Luna et al. (2011) it may not be immediately obvious to readers exactly what causal hypothesis such a graph represents. (To those unfamiliar with potential outcomes, the absence of an edge from T to $Y(t)$ might be seen (incorrectly) as representing the null hypothesis that T does not affect Y , so $T \perp\!\!\!\perp Y/X$, rather than $T \perp\!\!\!\perp Y(t) | X$.) Richardson and Robins (2013a,b) present a simple general graphical procedure for deriving such graphs, from an initial causal graph containing observed variables. The derived graphs are called Single World Intervention Graphs (SWIGs) and contain counterfactual random variables associated with a specific hypothetical intervention on one or more treatment variables.

The SWIG is constructed from the original DAG in three steps: (i) treatment nodes that are intervened upon are each split into a random piece and a fixed piece; (ii) the random piece

* thomasr@uw.edu. ** robins@hsph.harvard.edu. *** linbowang@g.harvard.edu.

inherits all incoming edges on the original graph and the fixed piece inherits all out-going edges; (iii) nodes that are descendants of the fixed portion are replaced with counterfactual nodes associated with this fixed value.

See Figure 1 for a simple example. Here the intervention fixes A to $\tilde{a}$, or equivalently $\text{do}(\tilde{a})$. In Steps (i) and (ii), the node A is split into the random (A), and the fixed ($\tilde{a}$). In step (iii) the node L is replaced with $L(\tilde{a})$.

Richardson and Robins (2013b) showed that under the (naturally associated) Finest Fully Randomized Causally Interpretable Structured Tree Graph (FFRCISTG) model (Robins, 1986), the distribution of the counterfactual variables on the SWIG factors according to the graph. The DAG and corresponding SWIG encode, respectively, the following factorizations:

$$p(A = a, L = l, Y = y) = p(L = l)p(A = a|L = l)p(Y = y|A = a, L = l); \quad (1)$$

$$p(A = a, L = l, Y(\tilde{a}) = y) = p(L = l)p(A = a|L = l)p(Y(\tilde{a}) = y|L = l) = p(L = l)p(A = a|L = l)p(Y = y|A = \tilde{a}, L = l).$$

(2)

The last equality here follows from consistency. The well-known backdoor formula for $P(Y(\tilde{a}))$ follows directly from (2).

Technical Note: *In fact (2) is also implied by a condition termed ‘modularity’ by Richardson and Robins (2013b), that is slightly weaker than consistency. Modularity requires that the conditional distribution of a variable given its parents in the SWIG is obtained from the corresponding distribution in the original DAG, by conditioning on the specified value of any variable that has been split. In this example, in Figure 1, $p(Y(\tilde{a})=y|L=l) = p(Y=y|L=l, A=\tilde{a})$.*

2. Variable Selection for Causal Inference

We consider two broad approaches to variable selection for causal inference:

- I. Approaches that assume adjusting for *all* pre-treatment variables is sufficient. Under this assumption, the goal of these approaches is to either increase efficiency without introducing bias or to minimize mean square error by appropriately trading off bias with variance. For a few examples see Shortreed and Ertefaie (2017), Schnitzer et al. (2016), De Luna et al. (2011), van der Laan and Gruber (2010), Claeskens and Hjort (2003) and Lepskii (1992). As these approaches do not attempt to learn the underlying causal structure, they cannot avoid M-bias, even given an infinite sample size.

- II. Approaches that do not assume that ignorability holds given all pre-treatment variables. VanderWeele and Shpitser (2011) propose a disjunctive criterion that results in ignorability only assuming that a sufficient set of observed covariates for confounder control exists. This disjunctive criterion relies entirely on expert knowledge to find the set of covariates to control for. Maathuis and Colombo (2015) propose an alternative generalized backdoor criterion and, as advocated in Häggstrom (2017), use a purely data-driven procedure to guide covariate selection under the assumption of faithfulness. In contrast to (I), these approaches avoid M-bias when their assumptions are satisfied.

Approaches in (I) have been studied more often in the literature. In the following, we discuss approaches in (II) and compare them to the method of Häggstrom (2017).

2.1 Disjunctive cause criterion

As discussed by Häggstrom (2017), VanderWeele and Shpitser (2011) introduce a disjunctive cause criterion for confounder selection. However, we note that the disjunctive cause criterion as stated by Häggstrom (2017) differs from the formulation of VanderWeele and Shpitser (2011) in one important respect. Under the original formulation, the adjustment set corresponds to all variables that are parents ('direct causes') or non-parental ancestors ('indirect causes') of T or Y . However, under the disjunctive cause criterion given by Häggstrom (2017), the variables to be adjusted for are those that are either *direct* causes of T or *direct* causes of Y . The DAG shown in Figure 2(a) makes this distinction clear: here the set $X_{\rightarrow T \rightarrow Y} = \{X_1, X_3, X_5\}$, while the set as originally defined in VanderWeele and Shpitser (2011) corresponds to $\{X_1, X_2, X_3, X_4, X_5\}$. In the absence of unobserved confounders, the set $X_{\rightarrow T \rightarrow Y}$ suffices to control confounding (as do $X_{\rightarrow T}$ and $X_{\rightarrow Y}$). However, in the setting of VanderWeele and Shpitser (2011), where unobserved variables are permitted, it is necessary to include indirect causes. To see this, consider the DAG in Figure 2(b), where U_1 and U_2 are unobserved. Here the set of observed direct causes is $\{X_3\}$, while the set given by disjunctive cause criterion is $\{X_2, X_3, X_4\}$; the latter, but not the former, suffices to control confounding.

The method of VanderWeele and Shpitser (2011) is often presented as a clear improvement over approach (I) that ignores the possibility of M-bias. However, this need not be the case. First, as VanderWeele and Shpitser (2011) show, if an analysis proceeds assuming that there exists a set sufficient to control confounding, and this is false, then adjusting for the set chosen by the disjunctive cause criterion may result in greater bias compared with adjusting for all pre-treatment variables (even if the analyst has correctly identified causes of treatment and outcome). Second, it is possible that adjustment for all pre-treatment variables suffices to control for confounding, yet an analyst using the disjunctive cause criterion may report a biased effect estimate due to lack of perfect substantive knowledge required for use of the disjunctive cause criterion. This greatly diminishes the appeal of the disjunctive cause criterion as a practical means of addressing M-bias.

On a historical note, we also point out that the disjunctive cause criterion of VanderWeele and Shpitser (2011) is a special case of the equivalence of (iv) and (v) in Theorem 4.2 of (Richardson and Spirtes, 2002); it is also a direct corollary of the result of Verma and Pearl

(1990), showing that in a DAG with latent variables there exists a set d -separating T and Y if and only if there is no ‘inducing path’ between T and Y . (An inducing path between T and Y is a path on which every observed variable is a collider, and every collider is an ancestor of T or Y . The result follows since it is not hard to show that a path that d -connects T and Y given the observed ancestors of T and Y forms an inducing path between them since, by the definition of d -connection, such a path contains no observed non-colliders, and every collider is an ancestor of T or Y .)

2.2 Generalized back-door criterion

In a recent work, Maathuis and Colombo (2015) present a ‘generalized backdoor criterion’, that may identify a sufficient set of covariates to adjust for under the assumption of faithfulness, but without any assumption concerning either the existence of a sufficient set, knowledge of the graphical structure, or assumption that all confounders are measured.

The generalized backdoor criterion proceeds by first learning a ‘partial ancestral graph’ (PAG) that represents a class of DAGs with latent variables that are Markov equivalent over the observed variables. Under the assumption of faithfulness, and an oracle method to test for conditional independence the PAG may be recovered from the distribution of the observables. We illustrate this approach in the context of the DAGs in Figures 1 and 2 of Häggström (2017). The corresponding partial ancestral graphs are illustrated, respectively, in Figures 3 and 4. Given these PAGs, the generalized backdoor criterion implies that the sets $\{X_1, X_2, X_7\}$ and $\{X_1, X_2, X_8\}$ are sufficient to control for confounding in the first case, while $\{X_1, X_2, X_4, X_7\}$ and $\{X_1, X_2, X_4, X_8\}$ suffice in the second case. Without going into the full details, this can be seen intuitively from the PAGs by seeing that these vertices ‘block’ all ‘backdoor’ paths between T and Y , and do not include any colliders. Thus this approach avoids the problem of M-bias.

Technical details on PAGs: A pair of vertices in a PAG are adjacent if they are connected by an inducing path in the underlying graph. There is an ‘arrowhead’ on an edge ($V \prec - * W$) if in every DAG in the Markov equivalence class (over the observed variables) V is not an ancestor of W . Conversely there is a ‘tail’ on an edge ($V - - * W$) if in every DAG in the Markov equivalence class V is an ancestor of W . A circle ($V \circ - * W$) indicates that there exist DAGs in which V is and is not an ancestor of W . See Maathuis and Colombo (2015); Richardson and Spirtes (2002) for more details.

A key distinction between the approach of Maathuis and Colombo (2015) and all others considered here is that the generalized backdoor criterion may conclude that there is no set of observed variables that is certain to control for confounding. Of course, as an alternative one can use a sensitivity analysis to acknowledge that the empirically untestable assumptions needed to identify the causal effect under any of the other approaches are unlikely to hold.

We note that even if the causal effect were identified had the causal graph been known *a priori*, such a causal effect may not be identifiable if the causal graph is unknown. For example, given a distribution $p(X_2, X_3, X_4, T, Y)$ faithful to the graph in Figure 2(b), the generalized backdoor criterion would not be able to infer that $\{X_2, X_3, X_4\}$ is sufficient.

This is because there are other DAGs with latent variables in which this set is not sufficient, but which imply the same Markov structure (i.e. conditional independence relations) on $\{X_2, X_3, X_4, T, Y\}$.

The main limitation of the generalized backdoor formula is that in practical settings, conditional independence tests may have low power, and thus often result in erroneously concluding conditional independences. In this case, the generalized backdoor method may incorrectly conclude that a causal effect is identified and output a biased effect estimate. Moreover, methods that assume faithfulness are not uniformly consistent even when the assumption of faithfulness is correct; see Robins et al. (2003); Uhler et al. (2013); Spirtes and Zhang (2014) for more discussion.

2.3 Discussion of Häggstrom's method

We now place Häggstrom's method in the context of our discussion.

Though Häggstrom (2017) does *not* assume the structure is known, her theoretical approach (in §4.4) does presuppose that there are no unobserved covariates. However, in this case the disjunctive cause, the backdoor-path and the pre-treatment criterion will all yield consistent estimates, if bias is not introduced subsequently while attempting to decrease variance. In the absence of bias, choosing amongst them is solely a matter of efficiency.

The approach of Häggstrom (2017) applies either MMHC or MMPC to learn the structure of a graph. The assumption of faithfulness is then used in order to infer from the graph the conditional independence required to justify reducing the set of covariates to adjust for. However, if one is prepared to assume faithfulness then the method of Maathuis and Colombo (2015) has the advantage that unlike the method of Häggstrom (2017), it does not preclude the possibility of M-bias.

3. Comments on the Data Example

We now comment on the analysis of the effect of C-section on the risk of Asthma. The use of record-linkage data to obtain a full set of data is impressive.

Definition of the population and exposure variable

The study considers births at 37 or more full-weeks of gestation. It would be interesting to see whether the effect is moderated by the number of weeks. Similarly, if it is possible to distinguish elective versus non-elective C-sections, this might also be an important moderator.

Definition of the outcome variable

The prevalence rate of wheeze or asthma in both the C-section and vaginal delivery groups seem high relative to other rates that we found reported in the literature (e.g. Gibson et al., 2013). Since asthma medications are also used to treat other acute bronchial conditions it seems possible that the use of asthma drugs may be higher than the prevalence of asthma.

Identification of causal effects with missing data

The author conducted a complete-case analysis that leaves out 12.4% of observations due to missing values in covariates. However, in this study the missing pattern of key confounders such as income, education and benefits may depend on the missing values themselves. With non-ignorable missingness of this type, such a complete-case analysis may be biased and inefficient.

To deal with the missing data here, one may consider the assumption that missingness of confounders is independent of the outcome conditional on (possibly missing) confounder values and the treatment variable, as these confounders are measured at baseline. This pattern is known as instrumental missingness (Yang et al., 2017).

Acknowledgments

This research was supported by U.S. National Institutes of Health grant R01 AI032475, AI113251 and U.S. Office of Naval Research grant N00014-15-1-2672. The authors thank Dr. Martin C. Cahn M.D. for helpful conversations.

References

- Claeskens G, Hjort NL. The focused information criterion. *Journal of the American Statistical Association*. 2003; 98:900–916.
- De Luna X, Waernbaum I, Richardson TS. Covariate selection for the nonparametric estimation of an average treatment effect. *Biometrika*. 2011; 98:861–875.
- Gibson, J.Loddenkemper, R.Sibille, Y., Lundbäck, B., editors. European Lung Whitebook. European Respiratory Society; 2013.
- Häggstrom J. Data-driven confounder selection via Markov and Bayesian networks. *Biometrics*. 2017; doi: 10.1111/biom.12788
- Lepskii O. Asymptotically minimax adaptive estimation. i: Upper bounds. optimally adaptive estimates. *Theory of Probability & Its Applications*. 1992; 36:682–697.
- Maathuis MH, Colombo D. A generalized back-door criterion. *Ann Statist*. 2015; 43:1060–1088.
- Richardson TS, Robins JM. Single world intervention graphs: A primer. *Second UAI Workshop on Causal Structure Learning*. 2013a
- Richardson TS, Robins JM. Single world intervention graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Technical report, Center for the Statistics and the Social Sciences, University of Washington Series. 2013bWorking Paper 128
- Richardson TS, Spirtes P. Ancestral graph Markov models. *Ann Statist*. 2002; 30:962–1030.
- 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.
- Robins JM, Scheines R, Spirtes P, Wasserman L. Uniform consistency in causal inference. *Biometrika*. 2003; 90:491–515.
- Schnitzer ME, Lok JJ, Gruber S. Variable selection for confounder control, flexible modeling and collaborative targeted minimum loss-based estimation in causal inference. *The International Journal of Biostatistics*. 2016; 12:97–115. [PubMed: 26226129]
- Shortreed SM, Ertefaie A. Outcome-adaptive lasso: Variable selection for causal inference. *Biometrics*. 2017:??–??.
- Spirtes P, Zhang J. A uniformly consistent estimator of causal effects under the k -triangle-faithfulness assumption. *Statist Sci*. 2014; 29:662–678.
- Uhler C, Raskutti G, Bühlmann P, Yu B. Geometry of the faithfulness assumption in causal inference. *Annals of Statistics*. 2013; 41:436–463.

- van der Laan MJ, Gruber S. Collaborative double robust targeted maximum likelihood estimation. *International Journal of Biostatistics*. 2010; 6 Article 17.
- VanderWeele TJ, Shpitser I. A new criterion for confounder selection. *Biometrics*. 2011; 67:1406–1413. [PubMed: 21627630]
- Verma, T., Pearl, J. Equivalence and synthesis of causal models. In: Henrion, M. Shachter, R. Kanal, L., Lemmer, J., editors. *Uncertainty in Artificial Intelligence: Proceedings of the 6th Conference*; Mountain View, CA. 1990. p. 220-227. Association for Uncertainty in AI
- Yang S, Wang L, Ding P. Nonparametric identification of causal effects with confounders subject to instrumental missingness. *arXiv preprint arXiv:1702.03951*. 2017

Figure 1.
Derivation of a DAG involving counterfactuals from a DAG for the observed distribution.

Figure 2. DAGs illustrating the difference between the formulation of the disjunctive cause criterion by Häggström (2017) and that of VanderWeele and Shpitser (2011).

Figure 3.
Partial Ancestral Graph (PAG) corresponding to Figure 1 in Häggstrom (2017).

Figure 4.
Partial Ancestral Graph (PAG) corresponding to Figure 2 in Häggstrom (2017).