Discussion of "The Blessings of Multiple Causes" by Wang and Blei*

Kosuke Imai† Zhichao Jiang‡

October 1, 2019

We begin by congratulating Yixin Wang and David Blei for their thought-provoking article that opens up a new research frontier in the field of causal inference. The authors directly tackle the challenging question of how to infer causal effects of many treatments in the presence of unmeasured confounding. We expect their article to have a major impact by further advancing our understanding of this important methodological problem. This commentary has two goals. We first critically review the deconfounder method and point out its advantages and limitations. We then briefly consider three possible ways to address some of the limitations of the deconfounder method.

1 The Advantages and Limitations of the Deconfounder Method

We first discuss several advantages offered by the deconfounder method. We then examine the assumptions required by the method and discuss its limitations.

1.1 The Deconfounder Method

Suppose that we have a simple random sample of n units from a population. We have a total of m treatments, represented by the m-dimensional vector, $\boldsymbol{A}_i = (A_{i1}, A_{i2}, \ldots, A_{im})^\top$, for unit i. For the sake of simplicity, we ignore the possible existence of observed confounders $\mathbf{X}_i$. But, all the arguments of this commentary are applicable, conditional on $\mathbf{X}_i$. The deconfounder method consists of the following simple two steps. The first step fits the following factor model to the observed

*We thank Naoki Egami, Connor Jerzak, Michael Li, and Linbo Wang for helpful discussions.

†Professor, Department of Government and Department of Statistics, Institute for Quantitative Social Science, Harvard University, Cambridge MA 02138. Phone: 617–384–6778, Email: Imai@Harvard.Edu, URL: https://imai.fas.harvard.edu

‡Assistant Professor, Department of Biostatistics and Epidemiology, University of Massachusetts – Amherst, Amherst MA 01002.

treatments,

$$p(A_{i1}, A_{i2}, \ldots, A_{im}) \;=\; \int p(\boldsymbol{Z}_i) \prod_{j=1}^{m} p(A_{ij} \mid \boldsymbol{Z}_i)\; d\boldsymbol{Z}_i, \tag{1}$$

where $\boldsymbol{Z}_i = (Z_{i1}, Z_{i2}, \ldots, Z_{ik})^\top$ represents the k-dimensional vector of latent factors.

Once the estimates of the factors $\widehat{\boldsymbol{Z}}_i$, which Wang and Blei call the *substitute confounders*, are obtained, the second step estimates the average causal effects of multiple treatments by adjusting for these substitute confounders as follows,

$$\tau(\boldsymbol{a}, \boldsymbol{a}') \;=\; \mathbb{E}\{Y_i(\boldsymbol{a}) - Y_i(\boldsymbol{a}')\} \;=\; \mathbb{E}\{\mathbb{E}(Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \widehat{\boldsymbol{Z}}_i) - \mathbb{E}(Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}', \widehat{\boldsymbol{Z}}_i)\}, \tag{2}$$

where $\boldsymbol{a} \in \mathcal{A}$ and $\boldsymbol{a}' \in \mathcal{A}$ are the vectors of selected treatment values with $\boldsymbol{a} \neq \boldsymbol{a}'$ and $\mathcal{A}$ represents the support of $\boldsymbol{A}_i$. In practice, a regression model may be used to adjust for the substitute confounders as demonstrated by Wang and Blei in their empirical application.

The deconfounder method is attractive to applied researchers for several reasons. First, it is a simple procedure based on two classes of familiar statistical models — factor models and regression models. Second, the method offers diagnostics in observational studies with unmeasured confounding. Specifically, researchers can check the conditional independence among the observed treatments given the estimated factors,

$$A_{ij} \perp\!\!\!\perp \boldsymbol{A}_{i,-j} \mid \widehat{\boldsymbol{Z}}_i \tag{3}$$

for any $j = 1, \ldots, m$ and $\boldsymbol{A}_{i,-j}$ represents all the treatments except A_{ij}. If this conditional independence does not hold, then there may exist unobserved confounders that affect both A_{ij} and some of $\boldsymbol{A}_{i,-j}$, yielding a biased causal estimate. As discussed below, however, the lack of conditional independence may also be due to the misspecification of factor model, which, for example, would be present if there are causal relationships among treatments.

In sum, the deconfounder method proposes a simple solution to a long-standing problem of inferring causal effects of multiple treatments in observational studies. Many analysts of observational studies rely upon the assumption that the treatments are unconfounded conditional on a set of observed pre-treatment covariates. And yet, it is often difficult to rule out the possible existence of unobserved confounders. The deconfounder method not only offers a new identification strategy in the presence of unobserved confounding, but also shows how to check the validity of the resulting estimates under certain assumptions.

1.2 Assumptions

What assumptions does the deconfounder method require? Wang and Blei uses a graphical model to represent the conditional dependencies required by the deconfounder method. Here, we reproduce

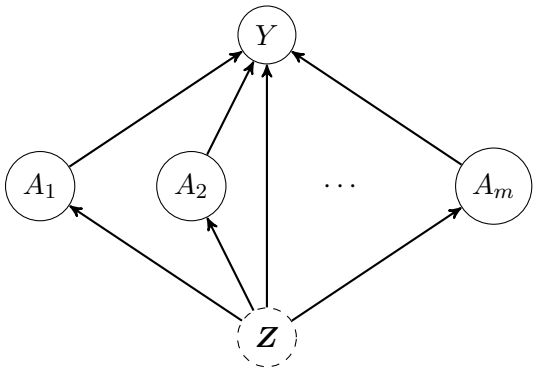


Figure 1: Directed Acyclic Graph for the Deconfounder Method.

the graphical model using the directed acyclic graph (DAG) in Figure 1. In addition to the SUTVA (Rubin, 1990), this DAG implies several key assumptions. First, the unobserved confounders $\boldsymbol{Z}$ should represent all confounding variables such that the treatments are ignorable given $\boldsymbol{Z}$,

$$Y_i(\boldsymbol{a}) \perp\!\!\!\perp \boldsymbol{A}_i \mid \boldsymbol{Z}_i \tag{4}$$

for any $\boldsymbol{a} \in \mathcal{A}$. The assumption implies that the multi-cause confounder $\boldsymbol{Z}_i$ suffices to adjust for the treatment-outcome confounding.

Second, the DAG also implies the following conditional independence assumption,

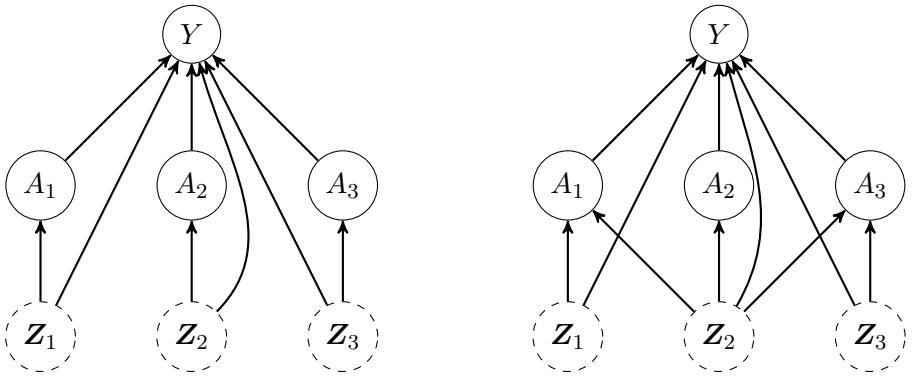
$$A_{ij} \perp\!\!\!\perp \boldsymbol{A}_{i,-j} \mid \boldsymbol{Z}_i \tag{5}$$

for any $j = 1, 2, \ldots, m$. The assumption justifies the factor model in equation (1). This assumption is violated if, for example, there exists a causal relationship among treatments. In the movie revenue application considered in the original article, the assumption is violated if the choice of actor for the main role (e.g., Sean Connery in a James Bond movie) influences the selection of actor for another role (e.g., Bernard Lee as the character of M). This is an important limitation of the deconfounder method as the problem may be common in applied research with multiple treatments.

In addition, according to Wang and Blei, the deconfounder method also requires the following overlap assumption that is not explicitly represented in the DAG,

$$p(\boldsymbol{A}_i \in \mathcal{A}^* \mid \boldsymbol{Z}_i) \; > \; 0 \tag{6}$$

for all sets $\mathcal{A}^* \subset \mathcal{A}$ with $p(\boldsymbol{A}_i \in \mathcal{A}^*) > 0$. The assumption implies that the choice of treatment values $\boldsymbol{a}$ may be constrained when estimating $\mathbb{E}\{Y_i(\boldsymbol{a})\}$. If the selected value of $\boldsymbol{a}$ does not belong to $\mathcal{A}^*$, then the resulting causal inference will be based on extrapolation.



(a) only unobserved single-cause confounders exist

(b) both unobserved single-cause and multiple-cause confounders exist

Figure 2: Examples of Unobserved Single-cause Confounders.

Finally, the key identification condition of the deconfounder method is the assumption of "no unobserved single-cause confounder." Wang and Blei formalize this assumption as the following set of conditional independence assumptions (see Definition 4 of the original article),

$$Y_i(\boldsymbol{a}) \quad \perp\!\!\!\perp \quad A_{ij} \mid \mathbf{V}_{ij} \tag{7}$$

$$A_{ij} \quad \perp\!\!\!\perp \quad \boldsymbol{A}_{i,-j} \mid \mathbf{V}_{ij} \tag{8}$$

for any $j = 1, 2, \ldots, m$, $\boldsymbol{a} \in \mathcal{A}$, and some random variable $\mathbf{V}_{ij}$. In addition, the authors require that these conditional independence relations do not hold when conditioning on any proper subset of the sigma algebra of $\mathbf{V}_{ij}$.

Unfortunately, these conditional independence assumptions are not sufficient to eliminate the possible existence of unobserved single-cause confounders. Figure 2 presents two examples, in which single-cause confounders exist, but equations (7) and (8) still hold. In addition, both cases can be reduced to the DAG in Figure 1 where no single-cause unobserved confounder exists by defining the unobserved multi-cause confounder as $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \boldsymbol{Z}_3)$. The examples demonstrate that a single multi-cause confounder can be decomposed into multiple single-cause confounders, and that several single-cause confounders can be combined into a single multi-cause confounder. Therefore, it is difficult to distinguish between single-cause and multiple-cause confounders without the knowledge of causal relationships among the variables.

We believe that it is important to develop the precise formal statement of the no unobserved single-cause confounder assumption. Such formalization allows us to understand how this assumption enables the identification of causal effects. In addition, our discussion implies that assessing the credibility of the assumption requires the scientific knowledge about the underlying causal structure involving unobserved confounders.

1.3 Nonparametric Identification

Wang and Blei establish the nonparametric identification of the average treatment effect given in equation (2) under the aforementioned assumptions in two steps. First, they show that a factor model of the observed treatments can be used to consistently estimate the substitute confounder. Second, they show that given the substitute confounder, the average treatment effects can be nonparametrically identified using equation (2) above.

In an insightful paper, D'Amour (2019) demonstrates that this two-step proof strategy leads to two problems for the deconfounder method. First, there may be more than one factor model that is compatible with the distribution of the observed treatments. He provides an example where different factor models that are compatible with the distribution of the observed treatments under the structure of Figure 1 yield different causal estimates. Second, D'Amour shows that even if a factor model is uniquely identified, the nonparametric identification is in general impossible.

Moving beyond the counterexamples, we consider the identification assumption for the factor model, discuss the role of the substitute confounder, and assess the overlap assumption required by the deconfounder method.

With respect to the identifiability of factor models, Kruskal (1977) and Allman, Matias and Rhodes (2009) give the general identification assumptions when observed variables are discrete. In this case, a crucial assumption is that the latent factor is correlated with the observed variables. In our context, this means that $\boldsymbol{Z}$ must causally affect each treatment A_j. In the causal inference literature, this assumption is known as faithfulness (Spirtes et al., 2000), which states that there exists conditional independence among variables in the population distribution if and only if it is entailed in the corresponding DAG. Thus, although Wang and Blei only discuss a set of conditional independence assumptions, the deconfounder method requires the faithfulness assumption in order to ensure the identifiability of factor model.

Next, we discuss the role of the substitute confounder. In the proof of the deconfounder method, Wang and Blei not only assume that the true unobserved confounder $\boldsymbol{Z}_i$ can be consistently estimated, but also treat the estimated substitute confounder $\widehat{\boldsymbol{Z}}_i$ as its true counterpart. This proof strategy ignores the crucial fact that the (estimated) substitute confounder is a function of observed treatments $\widehat{\boldsymbol{Z}}_i = \widehat{h}_M(\boldsymbol{A}_i) = \mathbb{E}_M(\boldsymbol{Z}_i \mid \boldsymbol{A}_i)$, where $\hat{h}_M$ indicates the fact that the substitute confounder is estimated from the data and depends on the choice of factor model and $\mathbb{E}_M$ represents the expectation with respect to the fitted factor model. We emphasize that the substitute confounder $\widehat{\boldsymbol{Z}}_i$ does not converge in probability to the true confounder $\boldsymbol{Z}_i$, which in itself is a random variable. Rather,

the substitute confounder converges to a function of observed treatments. Yet, this consistency result is required for the key results of the paper (i.e., Theorems 6–8).

We also closely examine the identification formula given in equation (2) by explicitly writing out the conditional expectation,

$$\mathbb{E}\{\mathbb{E}(Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \widehat{\boldsymbol{Z}}_i)\} = \int \mathbb{E}(Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \widehat{\boldsymbol{Z}}_i) p(\widehat{\boldsymbol{Z}}_i) d\widehat{\boldsymbol{Z}}_i \tag{9}$$

Notice that equation (9) does not follow unless the support of $p(\widehat{\boldsymbol{Z}}_i \mid \boldsymbol{A}_i = \boldsymbol{a})$ is identical to the support of $p(\widehat{\boldsymbol{Z}}_i)$ for any given $\boldsymbol{a} \in \mathcal{A}$. Unfortunately, since the substitute confounder is estimated using the observed treatments, $p(\widehat{\boldsymbol{Z}}_i \mid \boldsymbol{A}_i = \boldsymbol{a})$ is in general degenerate. The overlap assumption given in equation (6) is not applicable because the assumption is about the (true) unobserved confounders $\boldsymbol{Z}_i$ rather than the (estimated) substitute confounders, $\widehat{\boldsymbol{Z}}_i$. This means that we can only identify $\mathbb{E}(Y_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \widehat{\boldsymbol{Z}}_i = \boldsymbol{z}) = \mathbb{E}(Y_i \mid \boldsymbol{A}_i = \boldsymbol{a})$ for the values of $\boldsymbol{z}$ with $\boldsymbol{z} = \widehat{h}_M(\boldsymbol{a})$, implying that only a certain set of causal effects are identifiable.

In Theorem 6 of the original paper, Wang and Blei address this problem by imposing two additional restrictions. First, it is assumed that the outcome is separable in the following sense,

$$\mathbb{E}\{Y_i(\boldsymbol{a}) \mid \widehat{\boldsymbol{Z}}_i\} = f_1(\boldsymbol{a}) + f_2(\widehat{\boldsymbol{Z}}_i), \tag{10}$$

$$\mathbb{E}(Y_i \mid \boldsymbol{A}_i, \widehat{\boldsymbol{Z}}_i) = f_3(\boldsymbol{A}_i) + f_4(\widehat{\boldsymbol{Z}}_i), \tag{11}$$

where we use $\widehat{\boldsymbol{Z}}_i$ instead of $\boldsymbol{Z}_i$ to emphasize the fact that the substitute confounder is estimated. Although equation (10) allows us to write the average treatment effect as a function of treatment values alone, i.e., $\mathbb{E}\{Y_i(\boldsymbol{a}) - Y_i(\boldsymbol{a}')\} = f_1(\boldsymbol{a}) - f_1(\boldsymbol{a}')$, this assumption is not particularly helpful for identification since conditioning on $\widehat{\boldsymbol{Z}}_i$ is still required to identify the mean potential outcomes. In addition, equation (11) can be rewritten as $\mathbb{E}(Y_i \mid \boldsymbol{A}_i) = f_3(\boldsymbol{A}_i) + f_4(\hat{h}_M(\boldsymbol{A}_i))$ because $\widehat{\boldsymbol{Z}}_i$ is a deterministic function of $\boldsymbol{A}_i$. This suggests that the validity of this restriction about the outcome model critically depends on the choice of factor model.

The second restriction is that when the treatments are continuous, the substitute confounder is a piece-wise constant function, i.e., $\nabla_{\boldsymbol{a}} f_{\boldsymbol{\theta}}(\boldsymbol{a}) = 0$ where a parametric model is assumed for $p(\widehat{\boldsymbol{Z}}_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{\theta}) = \delta_{f_{\boldsymbol{\theta}}(\boldsymbol{a})}$ with a vector of parameters $\boldsymbol{\theta}$. A similar restriction is proposed for the case of discrete treatments. Since $p(\widehat{\boldsymbol{Z}}_i \mid \boldsymbol{A}_i = \boldsymbol{a}, \boldsymbol{\theta}) = \delta_{\hat{h}_M(\boldsymbol{a})}$ automatically holds, the assumption is valid if $\hat{h}_M(\boldsymbol{a})$ is a piece-wise constant function. Thus, this second restriction also suggests that the choice of factor model is critical for the validity of the deconfounder method.

In sum, we conclude that the nonparametric identification is generally difficult to obtain under the deconfounder method. Because the substitute confounder is a function of observed treatments, it

leads to the violation of the overlap assumption. Wang and Blei introduce two additional restrictions to address this problem. However, these assumptions impose severe constraints on the choice of factor model as well as that of outcome model. As a consequence, they may significantly limit the practical applicability of the deconfounder method. Even when researchers carefully choose a factor model that satisfies these restrictions, they may obtain causal effects only for a restricted range of treatment values.

2 Alternative Approaches

We next consider three alternative approaches to the important question of identifying the causal effects of multiple treatments in the presence of unobserved confounders. The approaches in this section will be based on equation (4). Unlike the deconfounder method, however, we will directly consider the identification of the probability distributions involving the (true) unobserved confounder $p(\boldsymbol{A}_i, \boldsymbol{Z}_i)$ and $p(Y_i \mid \boldsymbol{A}_i, \boldsymbol{Z}_i)$ rather than adopting Wang and Blei's two-step proof strategy.

2.1 Parametric Approach

Wang and Blei use parametric models in their empirical applications. Here, we consider a more general parametric approach. A primary advantage of the parametric approach is simplicity, whereas its major limitation is the required modeling assumptions that may not be credible in practice.

Suppose that there exists a uniquely identifiable factor model for the treatments, and that the joint distribution of $(\boldsymbol{A}, \boldsymbol{Z})$ is also identifiable. We assume the following additive model for the outcome variable,

$$\mathbb{E}\{Y_i(\boldsymbol{a}) \mid \boldsymbol{Z}_i\} \;=\; \sum_{j=1}^{m} \beta_j b_j(a_j) + \sigma g(\boldsymbol{Z}_i),$$

where $b_j(\cdot)$ and $g(\cdot)$ are pre-specified functions. Under this setting, it can be shown that if σ is known, then the average treatment effect is identifiable so long as $(b_1(A_{i1}), \ldots, b_m(A_{im}))$ is linearly independent. In contrast, if σ is unknown, then the average treatment effect is identifiable if $(b_1(A_{i1}), \ldots, b_m(A_{im}), \mathbb{E}\{g(\boldsymbol{Z}_i) \mid \boldsymbol{A}_i\})$ is linearly independent. This linear independence assumption is analogous to the overlap assumption discussed earlier, but the assumption can be tested using the observed data.

To illustrate this parametric approach, consider an example, in which we have three binary treatments $m = 3$ and one binary latent factor Z_i. Further assume that we have the following

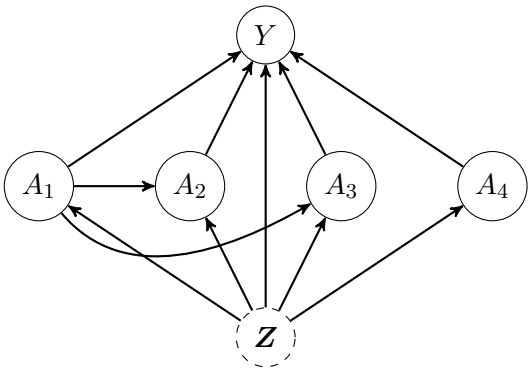


Figure 3: Directed Acyclic Graph in the Presence of Causal Relations among Treatments.

outcome model,

$$\mathbb{E}\{Y_i(\boldsymbol{a}) \mid Z_i\} \;=\; \beta_0 + \sum_{j=1}^{3} \beta_j A_{ij} + \sigma Z_i.$$

Now, consider a scenario, under which A_{ij}'s are mutually independent of one another given Z_i. Then, the joint distribution $p(A_{i1}, A_{i2}, A_{i3}, Z_i) = p(Z_i) \prod_{j=1}^{3} p(A_{ij} \mid Z_i)$ is identifiable based on the joint distribution of (A_{i1}, A_{i2}, A_{i3}) up to label switching (see Kruskal, 1977). Note that the average treatment effects are invariant to label switching. Thus, under this condition, even if σ is unknown, β_j's are identifiable so long as $\mathbb{E}(Z_i \mid A_{i1}, A_{i2}, A_{i3})$ is not linear in (A_{i1}, A_{i2}, A_{i3}).

Next, consider a different case shown as the DAG in Figure 3, in which one treatment causally affects other treatments. In this case, we may focus on estimating the causal effects of (A_2, A_3, A_4) conditional on A_1. We assume the following model for the outcome variable,

$$\mathbb{E}\{Y_i(\boldsymbol{a}) \mid Z_i\} \;=\; \beta_0 + \sum_{j=1}^{4} \beta_j A_{ij} + \sigma Z_i.$$

The joint distribution of $\boldsymbol{A}_i$ and Z_i under Figure 3 is given by $p(Z_i)p(A_{i1} \mid Z_i)p(A_{i2} \mid A_{i1}, Z)p(A_{i3} \mid A_{i1}, Z_i)p(A_{i4} \mid Z_i)$. This factorization is identifiable from the observed data (Allman, Matias and Rhodes, 2009). Then, even when σ is unknown, we can identify the parameters in the outcome model so long as $\mathbb{E}(Z_i \mid A_{i1}, A_{i2}, A_{i3}, A_{i4})$ is not linear in $(A_{i1}, A_{i2}, A_{i3}, A_{i4})$. Using these estimated parameters, we can obtain the estimates for the causal effects.

2.2 Nonparametric Approach

In the causal inference literature, many scholars first consider the problem of nonparametric identification by asking whether or not causal effects can be identified without making any modeling assumption. Only after the nonparametric identification of causal effects is established, researchers

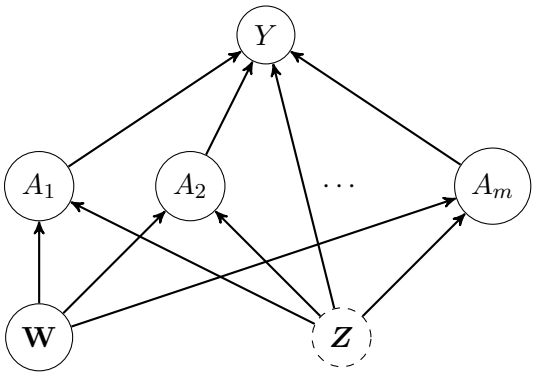


Figure 4: Directed Acyclic Graph for the Instrumental Variable Approach.

proceed to their estimation and inference. Cox and Donnelly (2011) regard this approach as a general principle of applied statistics. They state,

> *If an issue can be addressed nonparametrically then it will often be better to tackle it parametrically; however, if it cannot be resolved nonparametrically then it is usually dangerous to resolve it parametrically.* (p. 96)

To enable the general nonparametric identification of causal effects in the current setting, we must introduce auxiliary variables. D'Amour (2019) considers the use of proxy variables. Here, we examine an approach based on instrumental variables. Figure 4 presents the DAG for this approach where $\mathbf{W}$ represents a set of instrumental variables. Instrumental variables have the property that they are not affected by the unobserved confounders $\boldsymbol{Z}$ and influence the outcome Y only through the treatments $\boldsymbol{A}$.

For the sake of simplicity, we begin by considering the following separable model for the outcome,

$$\mathbb{E}\{Y_i(\boldsymbol{a}) \mid \boldsymbol{Z}_i\} \;=\; q(\boldsymbol{a}) + r(\boldsymbol{Z}_i),$$

where $\mathbb{E}\{r(\boldsymbol{Z}_i)\} = 0$ without loss of generality. Since the instrumental variables satisfy $\mathbb{E}\{r(\boldsymbol{Z}_i) \mid \mathbf{W}_i\} = \mathbb{E}\{r(\boldsymbol{Z}_i)\} = 0$, we obtain,

$$\mathbb{E}(Y_i \mid \mathbf{W}_i) \;=\; \mathbb{E}\{q(\boldsymbol{A}_i) \mid \mathbf{W}_i\} \;=\; \sum_{\boldsymbol{a} \in \mathcal{A}} q(\boldsymbol{A}_i = \boldsymbol{a}) p(\boldsymbol{A}_i = \boldsymbol{a} \mid \mathbf{W}_i). \tag{12}$$

Since we can identify $\mathbb{E}(Y_i \mid \mathbf{W}_i)$ and $p(\boldsymbol{A}_i \mid \mathbf{W}_i)$ from the observed data, the causal effects are identifiable if we can uniquely solve $q(\cdot)$ using equation (12). Suppose that all the treatments are binary and the instrumental variable is discrete with L levels. Since there are 2^m parameters in $q(\boldsymbol{a})$, equation (12) implies that the identification requires the $2^m \times L$ matrix $\{p(\boldsymbol{A}_i \mid \mathbf{W}_i)\}$ to be full-rank. This condition is analogous to the overlap assumption discussed earlier and can be checked using the

observed data. The proposed approach here, however, requires the instrumental variables to have more than 2^m levels. When m is large, it may be difficult to find instrumental variables that satisfy this condition.

The deconfounder method is closely related to the control function methods developed in the econometrics literature. The control function is a variable that, when adjusted for, renders an otherwise endogenous treatment variable exogenous (see e.g., Wooldridge, 2015). Imbens and Newey (2009) consider the nonparametric identification of the following nonseparable triangular system of equations (as before, we omit observed pre-treatment confounding variables for simplicity),

$$Y_i = s_1(A_i, Z_i), \tag{13}$$

$$A_i = s_2(W_i, U_i) \tag{14}$$

where Z_i and U_i are unobserved, A_i is the endogenous treatment variable of interest, W_i is the instrumental variable with $W_i \perp\!\!\!\perp (Z_i, U_i)$, and $s_2(\cdot, \cdot)$ is a strictly monotonic function of U_i. When A_i is a vector and $U_i = Z_i$, equations (13) and (14) become identical to the setting of the deconfounder method. Imbens and Newey show that the control function C_i is given by the cumulative distribution function of A_i given W_i, i.e., $C_i = F_{A|W}(A_i, W_i)$. Like the substitue confounder, the control function unconfounds the treatment variable, i.e., $Y_i(a) \perp\!\!\!\perp A_i \mid C_i$. This is because C_i is a one-to-one function of U_i, and A_i depends only on W_i conditional on U_i.

It is important to emphasize that the control function methodology requires the overlap assumption that the support of the marginal distribution of the control function, i.e., $p(C_i)$, is the same as the support of the conditional distribution, i.e., $p(C_i \mid A_i)$. However, unlike the case of the deconfounder method, the control function is not a function of the treatment variable, making this overlap assumption more likely to be satisfied.

In sum, the nonparametric identification of causal effects in the current settings requires the existence of auxiliary variables. Here, we consider an approach based on instrumental variables. Even when such instrumental variables are available, certain overlap assumptions are needed. This point is also clearly shown for the control function methods that are closely related to the deconfounder method. As we discussed, the overlap assumptions required for these instrumental variable methods are less stringent than those required for the deconfounder method.

2.3 Stochastic Intervention Approach

Our discussion has identified the overlap assumption as a main methodological challenge for the deconfounder method. Because the estimated substitute confounder itself is a function of treatment

variables, conditioning on the particular treatment values alters the support of its distribution. The parametric and nonparametric approaches introduced above address this problem through the reliance on modeling assumptions and the use of instrumental variables, respectively.

The final approach we consider is to change the causal quantities of interest using the idea of stochastic intervention. Instead of comparing two sets of fixed treatment values, we propose to contrast the two different distributions of treatments. In the movie application of the original article, one may be interested in comparing the revenue of a film featuring a typical cast for action movies with that featuring common actors for Sci-Fi movies. Stochastic intervention is a useful approach especially in the settings where inferring the average outcome under the fixed treatment values is difficult. For example, Geneletti (2007) applies it to mediation analysis, while Hudgens and Halloran (2008) propose an experimental design with stochastic intervention to identify spillover effects. More recently, Kennedy (2019) considers the incremental interventions that shift propensity score values to avoid overlap assumption.

Specifically, we focus on the average causal effects of distributions of treatments rather than the effects of treatments themselves.

$$\delta(p_1, p_0) \;=\; \mathbb{E}\left\{\int Y_i(\boldsymbol{a}) p_1(\boldsymbol{A}_i = \boldsymbol{a}) d\boldsymbol{a} - \int Y_i(\boldsymbol{a}) p_0(\boldsymbol{A}_i = \boldsymbol{a}) d\boldsymbol{a}\right\} \tag{15}$$

where p_1 and p_0 are the pre-specified distributions of treatments to be compared. Various distributions can be selected for comparison. For example, we may compare the conditional distributions of treatments given the different values of observed covariates, i.e., $p_1(\boldsymbol{A}_i \mid \mathbf{X}_i = \mathbf{x}_1)$ and $p_0(\boldsymbol{A}_i \mid \mathbf{X}_i = \mathbf{x}_2)$. Moreover, if factors are interpretable, then we may choose the conditional distributions given some specific values of the factors, i.e., $p_1(\boldsymbol{A}_i \mid \boldsymbol{Z}_i = \boldsymbol{z}_1)$ and $p_0(\boldsymbol{A}_i \mid \boldsymbol{Z}_i = \boldsymbol{z}_2)$. Topic models in the analysis of texts and ideal point models in the analysis of roll calls are good examples of interpretable factor models (Blei, Ng and Jordan, 2003; Clinton, Jackman and Rivers, 2004).

In the current setting, we may use the following estimator,

$$\hat{\delta}(p_1, p_0) \;=\; \sum_{i=1}^{n} Y_i \frac{p_1(\boldsymbol{A}_i) - p_0(\boldsymbol{A}_i)}{\hat{p}(\boldsymbol{A}_i \mid \boldsymbol{Z}_i)} \tag{16}$$

where $\hat{p}(\boldsymbol{A}_i \mid \boldsymbol{Z}_i)$ is the estimated factor model. For this estimator, the required overlap assumption is that the support of $p_j(\boldsymbol{A}_i)$ is a subset of the support of $p(\boldsymbol{A}_i \mid \boldsymbol{Z}_i)$ for $j = 0, 1$. Researchers can choose $p_1(\boldsymbol{A}_i)$ and $p_0(\boldsymbol{A}_i)$ so that this overlap assumption is satisfied. Furthermore, although the deconfounder method is not applicable when one treatment causally affects another, under the stochastic intervention approach one could model causal relationships among treatments by

specifying $p(\boldsymbol{A}_i \mid \boldsymbol{Z}_i)$ provided that the model is identifiable. An example of such case is given in Figure 3.

3 Concluding Remarks

The article by Wang and Blei is an important contribution to the causal inference literature because it opens up a new research frontier. The authors study a relatively unexplored question of how to infer the causal effects of many treatments in the presence of unobserved confounders. The deconfounder method provides a novel and yet intuitive approach using familiar statistical models. A key insight is that under certain assumptions, the factorization of treatments can yield a substitute confounder as well as a practically useful diagnostic tool for checking the validity of the resulting substitute confounder.

Although the deconfounder method has advantages, as first pointed out by D'Amour (2019) and further elaborated in this commentary, the method is not free of limitations. In particular, it cannot achieve nonparametric identification without additional restrictions. We emphasized the violation of the overlap assumption due to the fact that the estimated substitute confounder is a function of observed treatments. Wang and Blei consider some restrictions on the outcome model that may overcome this limitation and enable identification. However, such restrictions may severely limit the applicability of the deconfounder method. More research is needed in order to investigate the consequences of these restrictions in practical settings.

We discussed three alternative approaches to the methodological problems of the deconfounder method. The first approach is based on parametric assumptions and extend the data analysis conducted in the original article. The second approach relies upon the use of instrumental variables and is related to the control function literature in econometrics. The final approach considers an alternative causal estimand based on stochastic intervention, which is particularly useful in the settings with high-dimensional treatments. We expect and hope that many researchers will follow up on the work of Wang and Blei and develop new methods for estimating the causal effects of multiple treatments in observational studies.

References

Allman, Elizabeth S, Catherine Matias and John A Rhodes. 2009. "Identifiability of parameters in latent structure models with many observed variables." *The Annals of Statistics* 37:3099–3132.

Blei, David M., Andrew Y. Ng and Michael I. Jordan. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research* 3:993–1022.

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98:355–370.

Cox, D. R. and Christi A. Donnelly. 2011. *Principles of Applied Statistics.* Cambridge: Cambridge University Press.

D'Amour, Alexander. 2019. "On Multi-Cause Causal Inference with Unobserved Confounding: Counterexamples, Impossibility, and Alternatives." *arXiv preprint arXiv:1902.10286.*

Geneletti, Sara. 2007. "Identifying direct and indirect effects in a non-counterfactual framework." *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 69:199–215.

Hudgens, Michael G. and Elizabeth Halloran. 2008. "Toward Causal Inference with Interference." *Journal of the American Statistical Association* 103:832–842.

Imbens, Guido W and Whitney K Newey. 2009. "Identification and estimation of triangular simultaneous equations models without additivity." *Econometrica* 77:1481–1512.

Kennedy, Edward H. 2019. "Nonparametric causal effects based on incremental propensity score interventions." *Journal of the American Statistical Association* 114:645–656.

Kruskal, Joseph B. 1977. "Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics." *Linear algebra and its applications* 18:95–138.

Rubin, Donald B. 1990. "Comments on "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9" by J. Splawa-Neyman translated from the Polish and edited by D. M. Dabrowska and T. P. Speed." *Statistical Science* 5:472–480.

Spirtes, Peter, Clark N Glymour, Richard Scheines, David Heckerman, Christopher Meek, Gregory Cooper and Thomas Richardson. 2000. *Causation, prediction, and search.* MIT press.

Wooldridge, Jeffrey M. 2015. "Control function methods in applied econometrics." *Journal of Human Resources* 50:420–445.