ICML "Beyond Bayes" Workshop, 2022.

TECHNICAL REPORT
R-526
July 2022

On Quantifying Bias in Causal Effects When Data Are Non-IID

Chi Zhang[1] **Karthika Mohan**[2] **Judea Pearl**[1]

Abstract

Traditional causal inference techniques assume data are independent and identically distributed (IID) and thus ignore interactions among units. In this paper, we analyze the bias of causal identification techniques in linear models if IID is falsely assumed. Specifically, we discuss 1) when it is safe to apply traditional IID methods on non-IID data, 2) how large the bias is if IID methods are blindly applied, and 3) how to correct the bias. We present the results through a real-world example of vaccine efficacy.

1. Introduction

Majority of the existing machine learning and causal inference algorithms assume the data are independent and identically distributed (IID)(Rubin, 1978; Schölkopf, 2022). Unfortunately IID rarely holds true inreal-world datasets. Suppose we are interested in studying the effectiveness of Covid-19 vaccines. Specifically, we are interested in the causal effect of vaccine doses, V, on the severity of sickness S. A naive method would be building a causal model on V, S, and other related factors, and estimating the causal effect of V on S using available data. However, applying traditional causal methods that assume IID may result in biased estimation since units/samples/individuals are not isolated from each other in a pandemic setting. We exemplify below a few instances where IID is violated.

- **Case 1:** The vaccination V of a unit i, (V_i), decreases their viral load, L_i, which in turn decreases the transmission rate of the virus, and hence decreases the viral load of a contact j, (L_j), and hence make j less sick.
- **Case 2:** Exposure to high viral load L_j exacerbates the cormorbidities of i, C_i.
- **Case 3:** Certain hidden factor Z_i might affect V_i and S_j at the same time.

Such interactions between units plague both observational and experimental studies. If the latter is performed in a controlled environment where units are isolated from each other, the results would not be valid for the target environment, where units affect one another. Hence, blindly assuming IIDness might result in a biased outcome. The scenario exemplified above raises several questions regarding the computation of causal effects given non-IID data.

1. Under what conditions can we safely ignore unit interactions with the guarantee that assuming IID (and applying existing estimation techniques) will result in negligible bias?
2. How large is the bias if we assume IIDness on non-IID data?
3. If assuming IID would yield a significantly biased estimate, then how can we get rid of this bias?

We answer those questions through results from our full paper (Zhang et al., 2022), presented in the following sections.

2. Modeling Unit Interactions

2.1. Interference

To detect, quantify, and remove bias, we need to model the exact interacting patterns of the units. One of the most studied concepts related to interactions among units is interference (Cox, 1958). Interference is the phenomenon in which treatment of unit i (V_i) causally affects the outcome, S_j, of another unit j. In almost all existing literature this is interpreted as there existing a causal pathway from V_i to S_j. Case-1 above is a typical example. Clearly, ignoring unit interactions while computing causal effects would result in a biased estimate. However, we note that interference is not the only type of interaction between units that can yield biased estimates. For example, Case-3 above is a confounding path between V_i and S_j that is not classified as interference. As we will discuss in the next section, this also yields biased estimates. In addition, we might have two units i, j, where the treatment of i is correlated with its own outcome through j.

[1] Department of Computer Science, University of California, Los Angeles, USA [2] School of Electrical Engineering and Computer Science, Oregon State University, USA. Correspondence to: Chi Zhang <zccc@cs.ucla.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022. Copyright 2022 by the author(s).

2.2. Interaction Models

We need to model different types of interactions. A useful graphical tool is the interaction models. Interaction models are derived from the traditional causal models, except that the variables are replaced with "explicit variables" that are variables specific to units. For example, "sickness" is a variable in a traditional causal model, while "sickness of unit i" is an explicit variable.

We show an example of interaction models by modeling the vaccine-sickness example in the introduction. The three variables we consider are the vaccine doses (V), the viral load (L), and the sickness (S). The interaction model should include the "explicit" version of those variables for all units. We consider a dataset with 4 units, where the causal relationships among them are as follows. An interaction network can be constructed using information regarding these units and expert knowledge.

- For each unit $i = 1, 2, 3, 4$, the vaccine dose of i affects the viral load of i which then affects the sickness of i ($V_i \to L_i \to S_i$).
- Units 1 and 2 live together. At a specific time stamp, 2 has higher viral load, so the viral load of 2 affects the viral load of 1 ($L_2 \to L_1$).
- Units 1 and 2's viral loads cause their own and each others' comorbidities, and in turn cause the sickness ($L_1 \to S_2, L_2 \to S_1$). Comorbidities are not explicitly portrayed.
- Units 2 and 3 are friends. A hidden factor Z_2 (e.g., Unit 2's mental condition) affects both Unit 2's vaccination decision and Unit 3's sickness ($V_2 \leftarrow Z_2 \to S_3$).
- A hidden factor W_4 (e.g., Unit 4's wealthiness) affects the sickness of both Unit 4 and Unit 3 ($S_4 \leftarrow W_4 \to S_3$).

In this way we construct the interaction network in Figure 1.

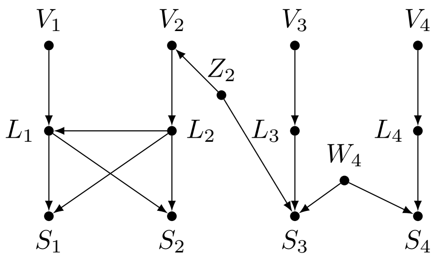

Figure 1: The interaction network for the vaccine-sickness example. Z_1, Z_3, Z_4 and W_1, W_2, W_3 are omitted from the graph.

3. Interaction Bias

Almost all machine learning algorithms including those that employ causal techniques assume that data are IID ((Schölkopf, 2022), section 3). In other words, the theoretical and performance guarantees of these algorithms are based on data being IID. As such it would be useful to determine conditions under which an algorithm meant for IID data can be applied on non-IID data with the certainty that the resulting *bias* would be negligible. The interaction bias is defined as the bias obtained by blindly assuming IID and applying IID methods to estimate the query Q.

In this work, we are primarily interested in the case where Q is the "non-IID version" of the ACE, which we name as *true average causal effects (TACE)*. Assuming IIDness, the ACE of V on S is identified as β_{SV}, the linear regression coefficient of S on V, if there is no non-causal path between V and S (Pearl et al., 2016; Pearl, 2017). TACE is the average unit effect through each unit itself, but not through the interactions with other individuals. In other words, TACE is ACE as if all the units were isolated. For example, in Figure 1, TACE of V on S is the average path-specific effects through $V_i \to L_i \to S_i$. We are interested in analyzing the difference between the TACE and the estimation obtained by incorrectly assuming IIDness ($\hat{\beta}_{SV}$).

4. Quantifying and Detecting Interaction Bias

In this section, we analyze how to quantify the interaction bias for an interaction model, and how to detect interaction bias given an interaction network. We make a few symmetric assumptions on the interaction models that include 1) the treatment is IID, 2) the "unit model" (with only explicit variables of this unit, and with interacting components removed) for each unit is the same, and 3) the treatment and outcome of a unit is not confounded by itself. Elements 1) and 2) of this assumption is still weaker than the traditional IID assumption, since we still allow unit interactions. Element 3) assumes no confounding within a unit, but allows a unit's treatment and outcome to be confounded by another individual.

4.1. Detecting Bias

There are two main types of problematic graphical structures in a linear interaction network that introduces bias in the estimation of TACE of V on S.

1. **Deflecting bias structure**: an open path between V_j and S_i for any $i \neq j$.
2. **Reflecting bias structure**: an open path between V_i and S_i through some explicit variable W_j, $i \neq j$.

For example, in Figure 1, examples of deflecting bias structures include $V_1 \rightarrow L_1 \rightarrow S_2$, $V_2 \leftarrow Z_2 \rightarrow S_3$, etc. $V_2 \rightarrow L_2 \rightarrow L_1 \rightarrow S_2$ is a reflecting bias structure.

Absence of bias structure implies no bias. If there is no bias structure in an interaction network, then $\hat{\beta}_{SV}$ would be an unbiased estimation of TACE of V on S. In this case, we can simply assume IIDness to estimate TACE. Note that no bias structure does not imply IIDness. In Figure 1, S_3 and S_4 are dependent, and hence non-IID, but this interaction does not constitute a bias structure.

4.2. Quantifying Bias

We can quantify the interaction bias created when blindly assuming IIDness in the estimation of ACE.

Theorem 4.1 ((Zhang et al., 2022)). *Let $M^*(G^*, S^*)$ be an interaction model with the symmetrical assumptions satisfied. Let D be the available data generated by M^* and let $G^\dagger$ be the approximate graph constructed using D. Let $TACE_{VS}$ be identifiable in $G^\dagger$ and be given by β_{SV}, the regression coefficient of S on V. Let α denote the true value of $TACE_{VS}$ in M^*. The interaction bias is given by,*

$$\left|E[\hat{\beta}_{SV}] - \alpha\right| = \Bigg|\frac{1}{n}\sum_{1\le i\le n}\sum_{p\in P[iji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_V} - \frac{1}{n(n-1)}\sum_{1\le i\le n}\sum_{p\in P[ji]} Val(p)\frac{\sigma^2_{R_p}}{\sigma^2_V}\Bigg|, \quad (1)$$

where $P[iji]$ is the set of reflecting bias structures between V_i and S_i through any explicit variable W_j of unit j with $i \neq j$, $P[ji]$ is the set of deflecting bias structures between V_j and S_i with $i \neq j$, and R_p is the root of path p.

It follows that the reflecting and deflecting structures are the only two structures that will bias the estimation of $TACE$. Note that although the definition of interaction bias on TACE is for any unbiased estimator for ACE, we focus only on the ordinary least squares estimator in this paper. This is because among the class of unbiased linear estimators, the OLS estimator has the minimum variance (Johnson et al., 2014).

We exemplify the above theorem.

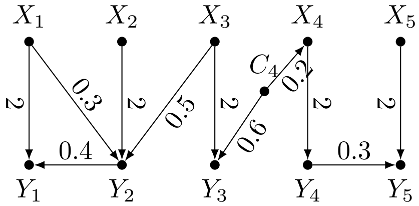

Figure 2: Interaction network with 4 units. The numbers represent edge coefficients. (C_1, C_2, C_3, C_5 are omitted)

Example 4.2. Figure 2 shows an example of an interaction model with 4 units where $X_1, \ldots, X_5$ are the treatments, and $Y_1, \ldots, Y_5$ the outcomes. The numbers on the edges are the edge coefficients. C_i for $i = 1, 2, 3, 5$ are omitted from the graph for simplicity.

Suppose we want to estimate $TACE_{XY}$, the ACE of X on Y as if the units were isolated:

- **Input:** the interaction network G^* as shown in Figure 2 (no parameter i.e., S^* is not an input),
- **Output:** the $TACE_{XY}$ (should equal to 2).

If we estimate ACE_{XY} ignoring the connections between units, our estimator will be $\hat{\beta}_{YX}$, with $Y = \{Y_1, \ldots, Y_5\}$ and $X = \{X_1, \ldots, X_5\}$. This is because ignoring the connections, the graph becomes $X_i \rightarrow Y_i$ separated for $i = 1, \ldots, 5$, so is essentially $X \rightarrow Y$ (Pearl, 2009). However, by Equation (1),

$$\begin{aligned}&\beta_{YX}\\ =&2 + \frac{0.3 \cdot 0.4}{4} - \frac{1}{12}\cdot 0.5 - \frac{1}{12}\frac{0.6 \cdot 0.2\sigma_C^2}{\sigma_X^2} - \frac{1}{12}\cdot 2 \cdot 0.3\\ \neq&2.\end{aligned}$$

Hence, the result is biased, and does not give us what we want.

5. Removing Interaction Bias

We present a method for computing an unbiased estimate of TACE in cases where Equation (1) predicts significant bias. It proceeds by applying linear regression on a set of samples B that satisfy the condition that no bias inducing structures exist between any two distinct units i and j. In particular, a subset of samples/units B is termed as a **bias-free subset** for $TACE_{VS}$ if no reflecting bias structures exist for any $i \in B$ and no deflecting bias structures exist in G^*_B where G^*_B is the latent projection of G^* on B (Definition 2.6.1, (Pearl, 2009)).

For example in figure 1, B comprises of units 1 & 3 and G^*_B is $V_1 \rightarrow L_1 \rightarrow S_1$ $V_4 \rightarrow L_4 \rightarrow S_4$. However, B *is not unique* for a given interaction network. Another candidates for B are units 1 & 3, or units 3 & 4. A possible algorithm for constructing B starts by randomly initializing B with a sample. Then it goes through the rest of the samples and adds a sample to B if it does not have a reflecting bias structure, and the inclusion does not create deflecting bias structures in the resultant graph, G^*_B. Once we have B, TACE can be unbiasedly estimated by $\hat{\beta}_{SV}$ using only the data in B.

Note that bias-free subset of samples B is not necessarily IID. While no reflecting or deflecting bias structures exist in G^*_B, there is no restriction on other forms of interactions among these samples. For example, G^*_B can be $V_1 \rightarrow S_1 \leftarrow C_2 \rightarrow S_2 \leftarrow V_2$ where S_1 is caused by C_2. In this case S is not IID and hence B does not constitute an IID dataset.

Also note that to compute an unbiased estimate, we have at our disposal a smaller set of samples; so the variance of estimation will be larger. There is a trade off between ignoring interaction (large bias, small variance), and using this debias method (no bias, large variance). It remains future work to quantify the variance of the estimator in this debias method for different interaction models.

Applicability of bias quantification results to real world problems: A natural question that arises at this juncture is whether we need an entire interaction network to apply these results to real world problems. Theorem 4.1 quantifies bias and in doing so reveals to us if and how various factors such as sample size and strength of connections (value of path coefficients) influence bias. This in turn allows us to use available information about the problem from prior experience, domain knowledge or external sources to determine if bias would be negligible or not. Specifically, bias is inversely proportional to sample size; in fact the quadratic term $n(n-1)$ in the denominator of deflecting bias shows that it diminishes at a fast rate as sample size increases. It is also evident that if the values of path coefficients are high, $Val(p)$ would be high and this will result in increased bias.

Finally, if the interaction connections are sparse (fewer edges between units), the reduction in the total number of paths could potentially lower bias but more importantly the number of samples in the bias-free set B used in the debias method will tend to be larger, which in turn will help in computing better quality estimates.

6. Conclusion

In this work, we analyzed the bias induced from blindly assuming IID in causal effect estimation using non-IID data, based on the results presented in (Zhang et al., 2022). We showed that incorrectly assuming IID induces bias if certain interacting patterns exist, and we quantified the bias given graphical models of interaction. We further presented a debiasing method which allows applying IID methods to non-IID data while guaranteeing minimal bias.

References

Cox, D. *Planning of Experiments*. Wiley Series in Probability and Statistics - Applied Probability and Statistics Section. Wiley, 1958. ISBN 9780471181835.

Johnson, R. A., Wichern, D. W., et al. *Applied multivariate statistical analysis*, volume 6. Pearson London, UK:, 2014.

Pearl, J. *Causality*. Cambridge university press, 2009.

Pearl, J. A linear "microscope" for interventions and counterfactuals. *Journal of causal inference*, 5(1), 2017.

Pearl, J., Glymour, M., and Jewell, N. P. *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

Rubin, D. B. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pp. 34–58, 1978.

Schölkopf, B. Causality for machine learning. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, pp. 765–804. 2022.

Zhang, C., Mohan, K., and Pearl, J. Causal inference with non-iid data using linear graphical models. Technical Report R-514, Department of Computer Science, University of California, Los Angeles, CA, 2022.