Abstract Title Page

Not included in page count.

Title: The Threshold of Embedded M Collider Bias and Confounding Bias

Author(s): Benjamin Kelcey (ben.kelcey@gmail.com), Joanne Carlisle (jfcarl@umich.edu)

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Despite the theoretical effectiveness of stratification methods such as those based on the propensity score (PS), the opacity and uniqueness of most enacted treatment selection mechanisms in nonexperimental social science research make it difficult to know a priori the appropriate covariates on which to stratify. Yet, because the plausibility of strong ignorability of the treatment assignment and corresponding inferences are highly dependent upon the selected covariates, a central concern in pretreatment stratification methods is how to which covariates to stratify on (e.g. Cook, Steiner & Pohl, 2009). In particular, stratifications that are too coarse (e.g. too few relevant covariates) likely gives rise to biased treatment estimates (e.g. Smith & Todd, 2005). Conversely, the inclusion of bias amplifying or extraneous covariates has also been shown to import extra bias and degrade the efficiency of the treatment effect estimator (e.g. Brookhart et al., 2006; Pearl, 2010; Wooldridge, 2009). For instance, stratifying on instrumental variables and covariates with colliding paths potentially increases bias and variance over unstratified estimates. Of relevance to this study is collider bias originating from stratification on pretreatment variables thought to form an M structural design (Figure 1). Given a treatment, Z, an observed pretreatment covariate, X, two unobserved and independent pretreatment covariates, U_1 and U_2, and an outcome, Y, the covariate X is a collider and may amplify bias when assessing the effect of the treatment on the outcome. That is, when two variables (e.g. U_1 and U_2) share a common effect (e.g. X), stratification on that effected variable (e.g. X) induces a statistical relation between otherwise independent factors (e.g. U_1 and U_2) (Figure 2). In turn, because these unobserved independent covariates are also causes of the treatment and outcome, stratifying on only X further induces a spurious relation between the treatment and outcome beyond the true treatment effect (i.e. collider bias). However, because the observed covariate, X, is hypothesized to be a confounder (e.g. Figure 3), concerns about confounding bias frequently dominate the potential for collider bias from unobserved bias amplifying covariates (e.g. Greenland, 2003). As a result, modal advice has been to stratify along a rich combination of observed covariates (e.g. Rubin & Thomas, 1996; Stuart & Rubin, 2007; Stuart, 2010). Yet, recent empirical investigations have demonstrated sizeable bias potentially corresponding to such collider bias especially with saturated stratifications (Whitcomb, Schisterman, Perkins & Platt, 2009; Steiner, Cook, Shadish & Clark, 2010). Such applications indicate the complexity of applying principles and suggest that there is much more to bias reduction than simply stratifying on many covariates.

Purpose / Objective / Research Question / Focus of Study:

Of particular import to this study, is collider bias originating from stratification on pretreatment variables forming an embedded M or bowtie structural design (Figure 4). That is, rather than assume an M structural design which suggests that X is a collider but not a confounder, we adopt what we consider to be a more reasonable position and that is X is both a collider and confounder. Accordingly, in this study we examined the extent to which confounder induced bias exceeds collider induced bias. To inform this tradeoff, we quantified the bias from two simple linear model estimators which are asymptotically equivalent to stratification and matching on these variables (alone or with the propensity score) (e.g. Pearl, 2009). More specifically, we examined this tradeoff by quantifying the net bias induced from adjusting for X versus the net bias from ignoring it. As a result, stratifying on X removes confounding bias but induces collider bias whereas ignoring X alleviates collider bias but invokes confounding bias.

For that reason, this study quantified the threshold by which collider and confounder bias due to a hypothesized confounder (e.g. *X*) is equal. The intention is to provide pragmatic guidance as to the consequences of and the decision to stratify on covariates hypothesized to be confounders and/or colliders.

Setting:
(May not be applicable for Methods submissions)

Population / Participants / Subjects:
(May not be applicable for Methods submissions)

Intervention / Program / Practice:
(May not be applicable for Methods submissions)

Significance / Novelty of study:

To a large extent there are competing theories and evidence as to which set of variables one should stratify on to approximate the strong ignorability of treatment assignment (e.g. Rubin, 2009; 2001; Pearl, 2010; Steiner et al., 2010). One (experimentalist) perspective has primarily suggested stratifying along a rich set of variables to produce covariate balance across treatment groups on all observed variables. In opposition, other (structuralist) perspectives are particularly concerned with the fundamental structure of germane (un)observed variables (e.g. Pearl, 2010). The surrounding empirical literature has demonstrated support for both sides (e.g. Steiner et al., 2010; Rubin, 2001). In this study, we take on explication of the conditions under which confounding bias dominates collider bias. In particular, this study develops an approach to quantify the conditions under which the net bias (confounding plus collider) is reduced through stratification on a confounder/collider.

Statistical, Measurement, or Econometric Model:

In assessing the unique relationship between *Z* and *Y* given in the directed acyclic graph in Figure 4, we may choose to stratify on *X* or not. As *X* is both a collider and confounder, either approach will address one form of bias but induce another. In order to assess this exchange, we might quantify the change in bias by identifying the threshold by which the potential collider bias introduced by including X exceeds the observed confounding bias induced by omitting *X*. That is, we might construct and stratify on a propensity score with or without variable *X* or with asymptotic equivalence (e.g. Pearl, 2009) we might consider the equations

$$Y_i = \beta_0 + \beta_1 X_i + \widehat{\delta}_{\Omega} Z_i + \varepsilon_i \tag{1.1}$$

$$Y_i = \beta_0 + \widehat{\delta}_{\omega} Z_i + \varepsilon_i \tag{1.2}$$

Given the variable relationships in Figure 4, the estimator $\widehat{\delta}_{\Omega}$ in equation (1.1) addresses the confounding bias brought about by *X*, but induces collider stratification bias as a result of the conditional relationships with the unobserved variables. In contrast, the estimator $\widehat{\delta}_{\omega}$ in equation (1.2) neglects the confounding bias but circumvents the collider bias. Because in practice we have not measured the unobserved variables, we cannot stratify on the unobserved variables and *X* to address both confounding and collider bias. However, because the introduction of collider bias is limited by the observed relationships of *X* with *Z* and *Y*, we can assess the change in net bias. Upon stratifying on *X*, confounding bias has been eliminated and the remaining bias is that of the collider

$$Bias(\hat{\delta}_{\Omega}) = \hat{\delta}_{\Omega} - \delta \tag{1.3}$$

Similarly, bias from the unstratified estimator is a result of confounding bias only

$$Bias(\hat{\delta}_{\omega}) = \hat{\delta}_{\omega} - \delta \tag{1.4}$$

The difference between collider and confounding bias between the estimators is

$$| \, Bias(\hat{\delta}_{\Omega}) | - | \, Bias(\hat{\delta}_{\omega}) | = | \, [E(\hat{\delta}_{\Omega}) - \delta] | - [E(\hat{\delta}_{\omega}) - \delta] | \tag{1.5}$$

This difference (1.5) can take on the following four situations

$$\begin{aligned}
&(1)\hat{\delta}_{\Omega} > \delta, \hat{\delta}_{\omega} > \delta : [E(\hat{\delta}_{\Omega}) - \delta] - [E(\hat{\delta}_{\omega}) - \delta] = E(\hat{\delta}_{\Omega}) - E(\hat{\delta}_{\omega}) \\
&(2)\hat{\delta}_{\Omega} < \delta, \hat{\delta}_{\omega} > \delta : [\delta - E(\hat{\delta}_{\Omega})] - [E(\hat{\delta}_{\omega}) - \delta] = 2\delta - E(\hat{\delta}_{\Omega}) - E(\hat{\delta}_{\omega}) \\
&(3)\hat{\delta}_{\Omega} > \delta, \hat{\delta}_{\omega} < \delta : [E(\hat{\delta}_{\Omega}) - \delta] - [\delta - E(\hat{\delta}_{\omega})] = E(\hat{\delta}_{\Omega}) - E(\hat{\delta}_{\omega}) - 2\delta \\
&(4)\hat{\delta}_{\Omega} < \delta, \hat{\delta}_{\omega} < \delta : [\delta - E(\hat{\delta}_{\Omega})] - [\delta - E(\hat{\delta}_{\omega})] = E(\hat{\delta}_{\omega}) - E(\hat{\delta}_{\Omega})
\end{aligned} \tag{1.6}$$

For brevity we focus on the most common situation where there is a positive treatment effect and the confounding variables are positively correlated with the treatment and outcome such that (1.1) underestimates and (1.2) overestimates the treatment effect as summarized in (2) in (1.6). Rewriting (2) in (1.6) as the least squares estimators using correlation coefficients

$$Change\, in\, Bias = 2\delta - [\frac{\sigma_y}{\sigma_z}(\frac{\rho_{yz} - \rho_{yx}\rho_{xz}}{1 - \rho_{xz}^2})] - [\frac{\sigma_y}{\sigma_z}\rho_{yz}] \tag{1.7}$$

where *ρ* and *σ* indicate the appropriate correlation and standard deviation. This equation expresses the change in net bias from both colliding and confounding. Setting the bias terms equal to each other, we can obtain a threshold by which colliding and confounding bias are similar:

$$[\frac{\sigma_y}{\sigma_z}\rho_{yz} - \delta] = [\delta - \frac{\sigma_y}{\sigma_z}(\frac{\rho_{yz} - \rho_{yx}\rho_{xz}}{1 - \rho_{xz}^2})] \tag{1.8}$$

Equation (1.8) depicts when one form of bias dominates the other. For instance, when

$$[\frac{\sigma_y}{\sigma_z}\rho_{yz} - \delta] > [\delta - \frac{\sigma_y}{\sigma_z}(\frac{\rho_{yz} - \rho_{yx}\rho_{xz}}{1 - \rho_{xz}^2})] \tag{1.9}$$

the net bias from confounding will exceed the net bias from colliding and we should stratify on *X*. Similar derivations can establish a threshold with respect to evaluative measures which further incorporate the variability of the estimator such as the mean-squared error. More specifically,

$$MSE(\hat{\delta}) = (\text{bias}(\hat{\delta}))^2 + \text{var}(\hat{\delta}) \tag{1.10}$$

Rewriting (1.10) for both estimators using correlations, we have the change in MSE as

$$\left[\frac{\sigma_y}{\sigma_z}\rho_{yz}-\delta\right]^2+\left[\left(\frac{\sigma_{\mathrm{y}}^2}{\sigma_{\mathrm{z}}^2}\right)*\frac{1-\rho_{\mathrm{yz}}^2}{n\text{ - }q\text{ - }1}\right]=$$

$$\left[\frac{\sigma_y}{\sigma_z}\left(\frac{\rho_{yz}-\rho_{yx}\rho_{xz}}{1-\rho_{xz}^2}\right)-\delta\right]^2+\left[\left(\frac{\sigma_{\mathrm{y}}^2}{\sigma_{\mathrm{z}}^2}\right)*\frac{1-\left(\frac{\rho_{\mathrm{yz}}^2+\rho_{\mathrm{yx}}^2-2\rho_{yz}\rho_{yx}\rho_{xz}}{1-\rho_{xz}^2}\right)}{n\text{ - }q\text{ - }1}*\frac{1}{1-\rho_{\mathrm{xz}}^2}\right] \tag{1.11}$$

where similar comparisons and thresholds can be made.

Research Design:
(May not be applicable for Methods submissions)

Data Collection and Analysis:
(May not be applicable for Methods submissions)

Findings / Results:

Figure 5 graphically displays the change in bias ($\left| Bias(\hat{\delta}_{\Omega})\right| - \left| Bias(\hat{\delta}_{\omega})\right|$) as a function of the treatment-confounder (*Z-X*) correlation for several outcome-confounder (*Y-X*) correlations using a medium treatment effect size of 0.5 and standardized variables. More specifically, negative 'Change in Bias' values indicate situations where it is better to stratify on *X* because estimators based on this stratification tend to offer a less biased estimate than those which exclude *X*. Evident from the example depicted in Figure 5, it is often more important to address confounding bias by stratifying on *X* than collider bias by excluding *X*. The exception comes when outcome-confounder (*Y-X*) correlation exceeds that of the outcome-treatment (*Y-Z*) and the treatment-confounder (*Z-X*) correlation is very high (>0.90). Similar plots of different effect sizes indicated that the outcome-confounder tends to need correlations similar or greater than the outcome-treatment correlations for collider bias to be of practical concern in structural systems that contain embedded M-relationships.

Usefulness / Applicability of Method:

To ground the relevance of this approach, we discuss a simplified application assessing the effect of teacher instructional practice in reading on student reading achievement while adjusting for teachers' reading knowledge. We very briefly frame the study and describe the potential for both confounder and collider bias. With renewed emphasis on observation of enacted classroom process (e.g. teaching) as a central feature of research designs, there has been substantial development of a diverse set of standardized classroom observations systems focusing on direct assessments (e.g., Cameron, Connor, & Morrison, 2005). The expectation is that such systems will help uncover reliable evidence concerning the processes which drive teachers' contribution to students' growth. To address variation across and within teachers in the nature their instruction, the Assessment of Pedagogical Knowledge of Teachers of Reading study (APK) sought to investigate what instructional quality is and the extent to which it actually matters. In particular, the study focused on measuring instruction in first through third grade teachers from urban school districts in Michigan using several observation methods. Further, the study centered on identifying and summarizing those instructional practices that are associated with students’ gains in reading over the course of a year in early literacy instruction. To capture the content, style, and delivery of each lesson, the study developed an observation system which had trained observers record observable instructional actions (IAs) within individual lessons. IAs included lesson features like giving directions, assessing student work, providing opportunities for students to participate or teacher soliciting student participation. Further, because teachers had undergone extensive professional development, the study assessed teachers' reading knowledge with emphasis on the knowledge about reading teachers draw on to teach early reading.

The part we focus on is the association of teachers' instructional reading practice on students' reading achievement when adjusting for teachers' reading knowledge. In particular, because we suspected that teachers' knowledge impacts students' achievement both through instructional reading practice as well as through other classroom instructional domains, teachers' reading knowledge may represent an important confounding variable. However, teachers' reading knowledge may also represent a colliding variable. For example, teachers' reading knowledge and instructional reading practice may have a common cause such as repeated exposure to professional development in reading. Similarly, teachers' reading knowledge and students' reading achievement may also share a common source such as teachers' general knowledge or ability to communicate effectively. Because it is suspected that teachers' reading knowledge informs instructional practice in reading as well as students reading outcomes through other channels, there is a potential for confounding bias if teachers' reading knowledge is omitted. Similarly, because we measured teachers' reading knowledge but did not appropriately measure professional development and general knowledge, there is a potential for collider bias when controlling for teachers' reading knowledge. This hypothesized relationship is depicted in Figure 6 such that it forms an embedded M or bowtie structure. To assess the potential tradeoff between confounder and collider bias, we can apply the above derivations to the empirical data. The correlation between the outcome, the Iowa Test of Basic Skills-Reading Comprehension, and the measure of instructional reading practice, was approximately 0.2. Similarly, the correlation between the outcome and our measure of teachers' reading knowledge was about 0.1 whereas the teachers' reading knowledge was correlated with practice at 0.4. Applying the above thresholds, the observed data strongly suggests we should make adjustments for teachers' reading knowledge as the potential reduction in bias from its adjustment likely exceeds the potential collider bias introduced by not adjusting for it (Figure 7). More specifically, in absence of the true treatment effect we graphed the change in bias curve for effect sizes of 0.1, 0.2 and 0.4. Our results indicated that the treatment effect size would have to exceed 0.45 in order for collider bias to be of more concern than confounder bias. Given a zero order correlation between the treatment and outcome of 0.20, it seems highly unlikely that the effect size would be of such magnitude.

Conclusions:

There are clear potential benefits of empirically appraising the collider-confounder bias exchange. At a minimum, it helps to understand and bound the extent to which covariates' may serve to amplify or reduce bias and, in turn, mount a more informed evidentiary basis. Such appraisals also serve to shift issues surrounding stratification on colliders from a theoretical exercise to an empirical one. For instance, in the given example, rather than (not) stratify on teachers' reading knowledge solely on the theoretical basis that it is (not) a collider, such analyses allow empirical assessment of the variable's impact if it were a collider.

Appendices

Not included in page count.

Appendix A. References

Brookhart, M., Schneeweiss, S., Rothman, K., Glynn, R., Avorn, J., & Sturmer, T. (2006). Variable selection for propensity score models. *Practice of Epidemiology*, vol. 163, 12, pp. 1149-1156.

Cameron, C. E., Connor, C. M., & Morrison, F. J. (2005). Effects of variation in teacher organization on classroom functioning. *Journal of School Psychology, 43,* 61–85.

Cook, T., Steiner, P., & Pohl, S. (2009). How bias reduction is affected by covariate choice, unreliability and mode of data analysis: Results from two types of within-study comparisons, *Multivariate Behavioral Research,* 44, pp. 828-47.

Greenland, S. (2003). Quantifying biases in causal models: classical confounding vs collider stratification bias. *Epidemiology*, 14, 3, pp. 300-306.

Pearl, J. (2010). On a class of bias-amplifying covariates that endanger effect estimates. Technical Report R-356.

Pearl, J. (2010). *Causality: Models, Reasoning and Inference*. 2nd ed. Cambridge University Press, New York.

Rubin, D. (2009). Author's reply: Should observational studies be designed to allow lack of balance in covariate distributions across treatment groups? *Statistics in Medicine,* 28, pp. 1415-1424.

Rubin, D. (2001). Using propensity scores to help design observational studies: application to the tobacco litigation, *Health Services and Outcomes Research Methodology,* 2, pp. 169-188.

Rubin, D. & Thomas, N., (1996). Matching using estimated propensity score: relating theory to practice. *Biometrics*, vol. 52, pp 249-64.

Stuart, E., & Rubin, D. (2007). Best practices in quasi-experimental designs: matching methods for causal inference. *Best Practices in Quantitative Methods*. Sage Publications: New York, pp. 155-176.

Stuart, E. (2010). Matching methods for causal inference: a review and a look forward. *Statistical Science*, 25, 1, pp. 1-21.

Smith, J., & Todd, P. (2005). Does matching overcome LaLonde's critique? *Journal of Econometrics,* 125, pp. 305-353.

Steiner, P., Cook, T., Shadish, W., & Clark, M. (2010). The importance of covariate selection in controlling for selection bias in observational studies, *Psychological Methods*, 15, 3, pp. 250-267.

Whitcomb, B., Schisterman, E., Perkins N., & Platt, R. (2009). Quantification of collider-stratificatino bias and the birthweight paradox. *Paedriatric and Perinatal Epidemiology*, 23, pp. 394-402.

Wooldridge, J. (2009). Should instrumental variables be used as matching variables? Michigan State University Technical Report.

Appendix B. Tables and Figures

Figure 1: Variables forming an M structural relationship with a treatment, *Z*, an observed pretreatment covariate, *X,* two unobserved and independent pretreatment covariates, U_1 and U_2, and an outcome, *Y*. The pretreatment covariate *X* is a collider and may amplify bias when assessing the effect of the treatment on the outcome.

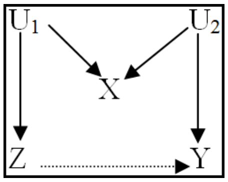


Figure 2: Stratification on *X* in Figure 1 produces a spurious relation between *Z* and *Y* beyond their true relation since U_1 and U_2 both effect *X*.

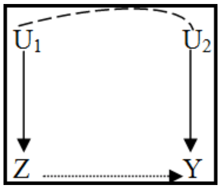


Figure 3: *X* as a confounder of the relationship between *Z* and *Y.*

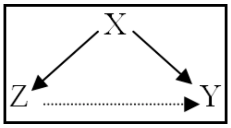


Figure 4: Embedded M or bowtie structural relations. Here *X* is both a confounder and collider.

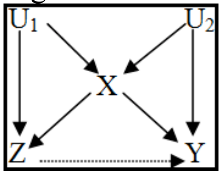


Figure 5: Change in bias ($|Bias(\hat{\delta}_{\Omega})| - |Bias(\hat{\delta}_{\omega})|$) as a function of the treatment-confounder (*Z*-*X*) correlation for several outcome-confounder (*Y*-*X*) correlations using a 0.5 effect size. When 'Change in Bias' is less than zero, it is bias is reduced by stratifying on *X*.

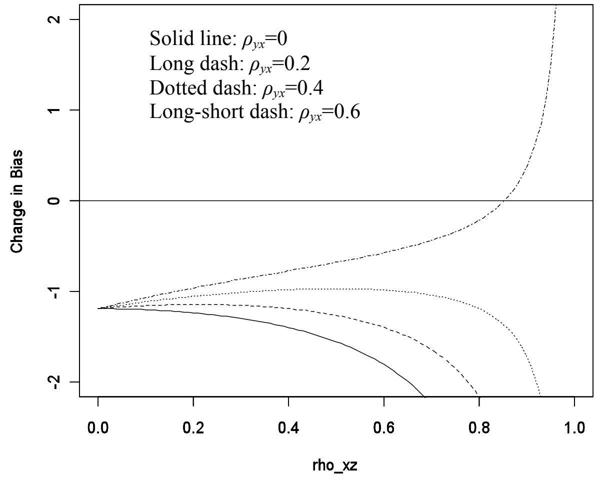


Figure 6: Structural relations among variables given in practical example. We would like to assess the association of practice, *Prac*, on students' reading comprehension achievement, *RC*, where teachers' reading knowledge, *TK*, is both a confounder and collider as professional development, *PD*, and teachers' general knowledge, *GK*, are both unobserved.

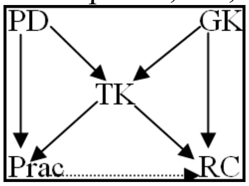
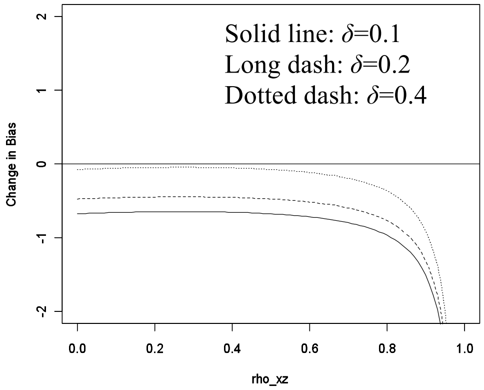


Figure 7: Change in bias ($|Bias(\hat{\delta}_{\Omega})| - |Bias(\hat{\delta}_{\omega})|$) as a function of the treatment-confounder (*Z*-*X*) correlation for several effect sizes for instructional practice example. When 'Change in Bias' is less than zero, it suggests that bias is reduced by stratifying on teachers' reading knowledge.


Solid line: δ=0.1
Long dash: δ=0.2
Dotted dash: δ=0.4
Change in Bias
rho_xz