Identifying Potential Mediating Variables Using Variable Importance Measures

Brian G. Vegetabile
RAND Corporation
bvegetab@rand.org

Donna L. Coffman
Temple University
dcoffman@temple.edu

Daniel F. McCaffrey
Educational Testing Service
dmccaffrey@ets.org

October 1, 2019

1 Background

Mediation analysis has a long history in the social sciences as a mechanism for providing *explanations* for observed effects [7]. The goal in a mediation analysis is to decompose the effect of a treatment that is observed into a direct effect of the treatment on the outcome and an indirect effect through a mediator that is itself directly related to the outcome. Recently the *causal mediation framework* literature has extended the concepts of total, direct, and indirect effects within the context of the potential outcome framework [1, 7]. However, the causal mediation framework, in attempting to quantify effects, downplays the goal of many applied researchers which is primarily to know if there is evidence that putative mediators may be the mechanism for the observed effect.

In this paper we develop criteria for a variable to be a potential mediator within the causal mediation framework. We then propose to evaluate these criteria using variable importance measures through procedures that compare the relative loss of models trained using data sets with permutations of key variables. We demonstrate the proposed method using simulated data and data from an application.

2 Causal Mediation Framework

Causal mediation analysis [4, 3, 1, 7] extends the traditional potential outcome framework [6, 5, 2] by introducing additional sets of potential outcomes that result from the intervention on an exposure variable A_i, for each individual i. The observed outcome is potentially a function of both the treatment/intervention and mediator variables, M_i, i.e., there exists a potential outcome for any vector pair $(a, m) \in \mathcal{A} \times \mathcal{M}$ denoted $Y_i(a, m)$. It is apparent that the variable

M_i is also affected by intervention so there exists an $M_i(a)$ for all $a \in \mathcal{A}$. By these definitions, the effect of the treatment equals $\tau_i = Y_i(a', M(a')) - Y_i(a, M(a)) \equiv Y_i(a') - Y_i(a)$.

Further decomposition requires counter-factual mediator variables: the systemic treatment-mediator potential outcome $Y_i(a, M(a))$ and a counter-factual mediator that would have occurred under an alternative exposure a', i.e., $Y_i(a, M(a'))$. To decompose the effect of the treatment A_i, we add and subtract the counterfactual outcome $Y_i(a, M(a'))$ and observe that

$$\begin{aligned}
\tau_i &= Y_i(a', M(a')) - Y_i(a, M(a)) && (1)\\
&= \big[Y_i(a', M(a')) - Y_i(a', M(a))\big] + \big[Y_i(a', M(a)) - Y_i(a, M(a))\big] && (2)\\
&\equiv \big[\text{“Natural Indirect Effect”}\big] + \big[\text{“Natural Direct Effect”}\big] && (3)\\
&\equiv NIE_i + NDE_i && (4)
\end{aligned}$$

The decomposition above is often referred to as the “natural” decomposition and the direct and indirect effects are referred to as a “natural” direct and indirect effects respectively.

2.1 Implications of the Framework

By considering the expected value for potential outcomes ($Y_i(a, M(a))$, $Y_i(a, M(a'))$, $Y_i(a', M(a'))$, $Y_i(a', M(a))$), and the two extreme cases 1) where there are no indirect effects – i.e., the effect arises from only a direct effect and not through any of the hypothesized mediators – or 2) where there are no direct effects – the effect is entirely through the hypothesized mediators, we derive conditions for the existence of indirect and direct effects. For example, one of two conditions is sufficient for there to be no indirect effects:

1. The distribution of the mediator is the same under the two levels of exposure, that is treatment has no effect on the distribution of the mediator.

2. The expected value of the potential outcome is equal for every value of the mediator, that is the outcome does not depend on the value of the mediator.

Thus, for M to be a mediator, it must be affected by treatment and the potential outcome must vary with values of M. These conditions also arise in traditional linear model- based mediation analyses, but we show that the linear model constraints are not necessary to derive the conditions.

Similarly, we show that a sufficient condition for there to be no direct effect is that the expected value of the potential outcome at every value of M is equal for the two values of treatment, $E[Y(a, m)] = E[Y(a', m)]$ for every value of m in the support of M.

3 Methods: An Analysis Plan for Identifying Potential Mediators

Based on results in the previous section, we propose the following plan to identify the potential mediators and the existence of direct effects.

Step	Assessment	Conclusion
1	Assess the effect of A on Y	Decide if mediation analysis is warranted to explain observed effect
2	Assess if A is important in predicting Y	If not, any effect is mostly indirect
3	Within the exposure subset $A = a'$, assess if M is important in predicting Y	If not, any effect is mostly direct
4	Assess if A is important in predicting M	If not, the effect is mostly direct or not through M. Otherwise, there is a partial indirect effect through M.

Table 1: Outline of analysis plan

To implement this plan we propose using variable importance methods to conduct the assessments in Steps 2 to 4. In this method, we fit a nonparametric model for the outcome as a function of treatment assignment, the mediators, and covariates. We then permute values of the mediators (one at a time) of treatment assignment and calculate the loss in the model fit from permuting each variable. We use the relative size of loss in the model fit due to permuting different variables to determine whether or not the conditions needed are met for M to potentially be a mediator. We repeat this process by fitting a model for the mediator as a function of the treatment and covariates and going through similar permutation analyses.

4 Demonstration

We demonstrate the method using a small simulation study. In the study we simulate data under different scenarios for whether or not there are direct effects and whether or not there are indirect effects. The study shows that the proposed method worked well in identifying the presence of direct effects and a potential mediator. The presentation will include full details of the simulation and additional explorations of the method through an expanded simulation study. We will also present results of an application of the method.

References

[1] K. Imai, L. Keele, and D. Tingley. A general approach to causal mediation analysis. *Psychological methods*, 15(4):309, 2010.

[2] G. W. Imbens and D. B. Rubin. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, 2015.

[3] J. Pearl. Direct and indirect effects. In *Proceedings of the Seventeeth Conference on Uncertainty in Artificial Intelligence*, pages 411–20, June 2001.

[4] J. M. Robins and S. Greenland. Identifiability and exchangeability for direct and indirect effects. *Epidemiology*, 3(2):143–155, 1992.

[5] D. B. Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688, 1974.

[6] J. Splawa-Neyman. On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statistical Science*, 5(4):465–472, 1990. Translated and edited by D.M. Dabrowska and T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczych Tom X (1923) 1-51 (Annals of Agricultural Sciences).

[7] T. VanderWeele. *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.