MAZE: Mediation Analysis for ZEro-inflated mediators

Meilin Jiang* Zhigang Li†

19 January 2023

1. Introduction

The causal mediation analysis is a statistical technique to investigate and identify relationships in a causal mechanism involving one or more intermediate variables (i.e., mediators) between an independent variable and an outcome. In addition to a better understanding of the causal pathways in proposed theoretical mechanisms, mediation analyses can help to confirm and refine treatments when it is not possible or ethical to intervene the independent variable.

However, challenges arise in mediation analyses in datasets with an excessive number of zero data point for mediators, especially in count data or non-negative measurements. The standard mediation analysis approaches may not be valid due to the violation of distributional assumptions. Moreover, the excessive zero mediator values could contain both true and false zeros. A true zero means that the measurement is truly zero, while a false zero means the measurement is positive but might be too small to be detected given the accuracy of devices used. Therefore, there is an unmet need for mediation analysis approaches to account for the zero-inflated structures of these mediators.

To address the difficulties, we proposed a novel mediation analysis approach to estimate and test direct and indirect effects to handle zero-inflated mediators that are non-negative. The zero-inflated log-normal (ZILoN), zero-inflated negative binomial (ZINB), and zero-inflated Poisson (ZIP) mediators were considered as the possible options of distributions for these mediators.

The R package `MAZE` implements the proposed causal mediation analysis approach for zero-inflated mediators in the corresponding paper to estimate and test natural indirect effect (NIE). Given the zero-inflated nature, the mediation effect (i.e., NIE) can be decomposed in to two components NIE_1 and NIE_2.

2. Model

For an independent variable X, a zero-inflated mediator M and a continuous outcome variable Y, the following regression equation is used to model the association between Y and (X, M):

$$Y_{xm1_{(m>0)}} = \beta_0 + \beta_1 m + \beta_2 1_{(m>0)} + \beta_3 x + \beta_4 x 1_{(m>0)} + \beta_5 xm + \epsilon, \tag{1}$$

where $Y_{xm1_{(m>0)}}$ is the potential outcome of Y when $(X, M, 1_{(M>0)})$ take the value of $(x, m, 1_{(m>0)})$, $1_{(\cdot)}$ is an indicator function. Equation (1) is an regression model where $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5$ are regression coefficients and ϵ is the random error following the normal distribution $N(0, \delta^2)$. Notice that interactions between X and the two mediators M and $1_{(M>0)}$ can be accommodated by the product terms $\beta_4 X 1_{(M>0)}$ and $\beta_5 XM$ in the model, which is an advantage of potential-outcomes mediation analysis approaches. Users can specify whether to include either one, both, or none of the two possible interactions using the argument `XMint`.

*University of Florida, meilin.jiang@ufl.edu
†University of Florida, zhigang.li@ufl.edu

2.1 Zero-inflated mediators

2.1.1 Zero-inflated log-normal (ZILoN) mediators For a ZILoN mediator, its two-part density function can be rewritten as:

$$f(m;\theta) = \begin{cases} \Delta, & m = 0 \\ (1-\Delta)\phi(m;\mu,\sigma), & m > 0 \end{cases}, \tag{2}$$

where $\phi(\cdot)$ is the density function of the log-normal distribution indexed by the parameters μ and σ which are the expected value and standard deviation, respectively, of the random variable after natural-log transformation.

The ZILoN mediator M depends on X through the following equations:

$$\mu = \alpha_0 + \alpha_1 X, \tag{3}$$

$$\log\left(\frac{\Delta}{1-\Delta}\right) = \gamma_0 + \gamma_1 X. \tag{4}$$

Equations (1), (3) and (4) together form the full mediation model for a ZILoN mediator and a continuous outcome.

2.1.2 Zero-inflated negative binomial (ZINB) mediators The two-part density function for a ZINB mediator M is given by:

$$f(m;\theta) = \begin{cases} \Delta = \Delta^* + (1-\Delta^*)(\frac{r}{r+\mu})^r, & m = 0 \\ \\ (1-\Delta)\frac{\Gamma(r+m)}{\Gamma(r)m!}\frac{(\frac{\mu}{r+\mu})^m}{(\frac{r}{r+\mu})^{-r}-1}, & m = 1,2,\dots \end{cases}, \tag{5}$$

where the parameter vector $(\mu, r)^T$ controls the number of zeros generated from the NB distribution, $0 < \Delta^* < 1$ is the parameter controlling the number of excessive zeros (i.e., not generated from the NB distribution), r is the dispersion parameter, and μ is the expectation of the negative binomial distribution. The ZINB mediator M depends on X through the following equations:

$$\log(\mu) = \alpha_0 + \alpha_1 X, \tag{6}$$

$$\log\left(\frac{\Delta^*}{1-\Delta^*}\right) = \gamma_0 + \gamma_1 X. \tag{7}$$

Equations (1), (6) and (7) together form the full mediation model for a ZINB mediator and a continuous outcome.

2.1.3 Zero-inflated Poisson (ZIP) mediators The two-part density function for a ZIP mediator can be rewritten as:

$$f(m;\theta) = \begin{cases} \Delta = \Delta^* + (1-\Delta^*)\exp(-\lambda), & m = 0 \\ (1-\Delta)\frac{\lambda^m}{m!(\exp(\lambda)-1)}, & m = 1,2,\dots \end{cases}, \tag{8}$$

where $\lambda > 0$ is the mean of the Poisson distribution. λ controls the number of zeros generated by the data generating process underlying the Poisson distribution, while $0 < \Delta^* < 1$ controls the number of excessive zeros in addition to zeros from the Poisson distribution. The ZIP mediator M depends on X through the following equations:

$$\log(\lambda) = \alpha_0 + \alpha_1 X, \tag{9}$$

$$\log\left(\frac{\Delta^*}{1-\Delta^*}\right) = \gamma_0 + \gamma_1 X. \tag{10}$$

Equations (1), (9) and (10) together form the full mediation model for a ZIP mediator and a continuous outcome.

2.2 Probability mechanism for observing false zeros

It is common to observe two types of zeros for M in a data set with excessive zeros: true zeros and false zeros. We use M to denote the true value of the mediator and use M^* for the observed value of M. When the observed value of the mediator is positive (i.e., $M^* > 0$), we assume $M^* = M$. However, when $M^* = 0$, we don't know whether M is truly zero or M is positive but incorrectly observed as zero. We consider the following mechanism for observing a zero:

$$P(M^* = 0|M) = \begin{cases} \exp(-\eta^2 M), & M \leq B \\ 0, & M > B \end{cases}, \tag{11}$$

where the parameter η needs to be estimated, and $B > 0$ is a known constant. The value of B can be informed on the basis of the insights and judgements of professionals in the specific field from which the data arose.

2.3 Mediation effect and direct effect

The natural indirect effect (NIE), natural direct effects (NDE) and controlled direct effect (CDE) are derived for the proposed mediation model. The NIE is also called the mediation effect. The total effect of the independent variable X is equal to the summation of NIE and NDE. Let M_x denote the value of M if X is taking the value of x. Let $1_{(M_x>0)}$ denote the value of $1_{(M>0)}$ if X takes the value of x. The average NIE, NDE and CDE if X changes from x_1 to x_2 are given by:

$$\text{NIE} = E\big(Y_{x_2 M_{x_2} 1_{(M_{x_2}>0)}} - Y_{x_2 M_{x_1} 1_{(M_{x_1}>0)}}\big), \tag{12}$$

$$\text{NDE} = E\big(Y_{x_2 M_{x_1} 1_{(M_{x_1}>0)}} - Y_{x_1 M_{x_1} 1_{(M_{x_1}>0)}}\big), \tag{13}$$

$$\text{CDE} = E\big(Y_{x_2 m 1_{(m>0)}} - Y_{x_1 m 1_{(m>0)}}\big), \text{ for a fixed (i.e., controlled) value of } M, \tag{14}$$

Based on the sequential order of the two mediators M and $1_{(M>0)}$, NIE can be further decomposed:

$$\begin{aligned} \text{NIE} &= E\big(Y_{x_2 M_{x_2} 1_{(M_{x_2}>0)}} - Y_{x_2 M_{x_1} 1_{(M_{x_1}>0)}}\big) \\ &= E\big(Y_{x_2 M_{x_2} 1_{(M_{x_2}>0)}} - Y_{x_2 M_{x_1} 1_{(M_{x_2}>0)}}\big) + E\big(Y_{x_2 M_{x_1} 1_{(M_{x_2}>0)}} - Y_{x_2 M_{x_1} 1_{(M_{x_1}>0)}}\big) \\ &:= \text{NIE}_1 + \text{NIE}_2, \end{aligned} \tag{15}$$

where NIE_1 is the mediation effect through M summing the two causal pathways $X \to M \to Y$ and $X \to M \to 1_{(M>0)} \to Y$, and NIE_2 is the mediation effect through only $1_{(M>0)}$ on the causal pathway $X \to 1_{(M>0)} \to Y$.

2.3.1 Effects for ZILoN mediators

$$\begin{aligned} \text{NIE}_1 &= (\beta_1 + \beta_5 x_2)\left[(1-\Delta_{x_2})\exp\left(\mu_{x_2} + \frac{\sigma^2}{2}\right) - (1-\Delta_{x_1})\exp\left(\mu_{x_1} + \frac{\sigma^2}{2}\right)\right], \\ \text{NIE}_2 &= (\beta_2 + \beta_4 x_2)(\Delta_{x_1} - \Delta_{x_2}), \\ \text{NDE} &= (x_2 - x_1)\left\{\beta_3 + (1-\Delta_{x_1})\left[\beta_4 + \beta_5 \exp\left(\mu_{x_1} + \frac{\sigma^2}{2}\right)\right]\right\}, \\ \text{CDE} &= (x_2 - x_1)\big(\beta_3 + \beta_4 1_{(m>0)} + \beta_5 m\big). \end{aligned}$$

2.3.2 Effects for ZINB mediators

$$\text{NIE}_1 = (\beta_1 + \beta_5 x_2)\Big[(1-\Delta^*_{x_2})\mu_{x_2} - (1-\Delta^*_{x_1})\mu_{x_1}\Big],$$

$$\text{NIE}_2 = (\beta_2 + \beta_4 x_2)\Big\{\big(1-\Delta^*_{x_2}\big)\Big[1-(\frac{r}{r+\mu_{x_2}})^r\Big] - \big(1-\Delta^*_{x_1}\big)\Big[1-(\frac{r}{r+\mu_{x_1}})^r\Big]\Big\},$$

$$\text{NDE} = (x_2 - x_1)\Big\{\beta_3 + (1-\Delta^*_{x_1})\Big[\beta_4\Big(1-(\frac{r}{r+\mu_{x_1}})^r\Big) + \beta_5\mu_{x_1}\Big]\Big\},$$

$$\text{CDE} = (x_2 - x_1)\big(\beta_3 + \beta_4 1_{(m>0)} + \beta_5 m\big).$$

2.3.3 Effects for ZIP mediators

$$\text{NIE}_1 = (\beta_1 + \beta_5 x_2)\Big[(1-\Delta^*_{x_2})\lambda_{x_2} - (1-\Delta^*_{x_1})\lambda_{x_1}\Big],$$

$$\text{NIE}_2 = (\beta_2 + \beta_4 x_2)\Big\{\big(1-\Delta^*_{x_2}\big)\Big[1-\exp\left(-\lambda_{x_2}\right)\Big] - \big(1-\Delta^*_{x_1}\big)\Big[1-\exp\left(-\lambda_{x_1}\right)\Big]\Big\},$$

$$\text{NDE} = (x_2 - x_1)\Big\{\beta_3 + (1-\Delta^*_{x_1})\Big[\beta_4\Big(1-\exp\left(-\lambda_{x_1}\right)\Big) + \beta_5\lambda_{x_1}\Big]\Big\},$$

$$\text{CDE} = (x_2 - x_1)\big(\beta_3 + \beta_4 1_{(m>0)} + \beta_5 m\big).$$

3. Installing the R package

The R package `MAZE` can be installed from the Github webpage.

```
require(devtools)
devtools::install_github("https://github.com/meilinjiang/MAZE", build_vignettes = TRUE)
```

4. Main function `MAZE()`

To estimate and test NIE, NIE_1, and NIE_2, NDE, and CDE, the R function `MAZE` is used to implement the proposed mediation analysis approach for zero-inflated mediators.

4.1 Input arguments

The input arguments to the function are

- `data`: a data frame containing variables X, M, Y, and Z (if any)
- `distM`: an optional character value for distribution to be used for the mediator. Possible choices are 'zilonm', 'zinbm', or 'zipm' for zero-inflated log-normal, negative binomial, or Poisson mediators respectively. By default, all three distributions will be fitted and the final mediation model is selected by AIC
- `K`: a user supplied sequence for the number of component K in the zero-inflated mixture mediators. Default is K = 1 for zero-inflated non-mixture mediators
- `selection`: model selection criterion to be used when more than one model is fitted. Possible choices are 'AIC' and 'BIC'. Default is 'AIC'

- `X`: variable name of the independent variable
- `M`: variable name of the mediator variable
- `Y`: variable name of the outcome variable (continuous)
- `Z`: name(s) of confounder variables
- `XMint`: a logical vector length 2 indicating whether to include the interaction terms between (i) X and 1(M>0) and (ii) X and M. Default is `c(TRUE, FALSE)`
- `x1`: the first value of independent variable of interest
- `x2`: the second value of independent variable of interest
- `zval`: the value of confounders to be conditional on in estimating effects
- `B`: the upper bound value B to be used in the probability mechanism of observing false zeros
- `seed`: an optional seed number to control randomness for reproducibility. The default is 1
- `ncore`: number of cores available for parallel computing

4.2 Outputs

A list object containing

- `results_effects`: a data frame for the results of estimated mediation effect
- `results_parameters`: a data frame for the results of model parameters
- `selected_model_name`: a string for the distribution of M and number of components K selected in the final mediation model
- `BIC`: a numeric value for the BIC of the final mediation model
- `AIC`: a numeric value for the AIC of the final mediation model
- `models`: a list with all fitted models
- `analysis2_out`: a list with output from analysis2() function (used for internal check)

5. Example

The `MAZE` package contains an example dataset `zinb10` that was generated using the proposed model with a zero-inflated negative binomial mediator ($K = 1$). It is a data frame with 100 observations and 3 variables: a continuous independent variable `X`, a continuous outcome `Y`, and a count mediator variable `Mobs`. The mediator variable contains 10% zero values in which half are false zeros.

```
library(MAZE)
# load the example dataset "zinb10"
data(zinb10)
# call MAZE() to perform mediation analysis
maze_out <- MAZE(data=zinb10,
                 distM=c('zilonm', 'zinbm', 'zipm'),
                 K = 1,
                 selection = "AIC",
                 X='X', M='Mobs', Y='Y', Z=NULL,
```

```
                XMint = c(TRUE, FALSE),
                x1=0, x2=1, zval = NULL,
                B=20, seed=1)

## results of selected mediation model
maze_out$results_effects # NIE1, NIE2, and NIE
maze_out$results_parameters # model parameters
maze_out$BIC; maze_out$AIC # BIC and AIC of the selected mediation model
maze_out$selected_disM # distribution of the mediator in the selected mediation model
```

Session Info

```
sessionInfo()
#> R version 4.2.2 (2022-10-31)
#> Platform: x86_64-apple-darwin17.0 (64-bit)
#> Running under: macOS Big Sur ... 10.16
#>
#> Matrix products: default
#> BLAS:   /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRblas.0.dylib
#> LAPACK: /Library/Frameworks/R.framework/Versions/4.2/Resources/lib/libRlapack.dylib
#>
#> locale:
#> [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
#>
#> attached base packages:
#> [1] stats     graphics  grDevices utils     datasets  methods   base
#>
#> loaded via a namespace (and not attached):
#>  [1] digest_0.6.31   lifecycle_1.0.3 magrittr_2.0.3  evaluate_0.19
#>  [5] rlang_1.0.6     stringi_1.7.8   cli_3.5.0       rstudioapi_0.14
#>  [9] vctrs_0.5.1     rmarkdown_2.19  tools_4.2.2     stringr_1.5.0
#> [13] glue_1.6.2      xfun_0.36       yaml_2.3.6      fastmap_1.1.0
#> [17] compiler_4.2.2  htmltools_0.5.4 knitr_1.41
```