APPLICATION NOTE

Supplemental Material for “Identifying mediating variables with graphical models: an application to the study of causal pathways in people living with HIV”

Adrian Dobra[a], Katherine Buhikire[b] and Joachim G. Voss[b]

[a]Department of Statistics, University of Washington, Seattle, WA, USA; [b]Frances Payne Bolton School of Nursing, Case Western Reserve University, Cleveland, OH, USA

ARTICLE HISTORY
Compiled August 29, 2019

ABSTRACT
This document contains a simulation study and additional tables for the main article.

1. Simulation study

We developed our simulation study to match the categorical dataset we analyzed in the main article. The R code necessary to replicate our numerical experiments is provided at the end of this document. We simulate contingency tables with three variables which we call X1, X2 and IC with 4 categories, 4 categories and 2 categories, respectively. These tables are sampled from a multinomial distribution with 16 categories associated with the $4 \times 4 \times 2 = 32$ cells that is consistent with the graphical loglinear model [X1,X2][X2,IC]. Each table was generated by following these steps: (i) randomly sampling cell probabilities from a uniform distribution on $(0, 1)$, then normalizing them to make their sum equal to 1; (ii) sampling a table from the multinomial distribution associated with these cell probabilities; (iii) fitting the loglinear model [X1,X2][X2,IC] to the table sampled in the previous step to determine the expected cell probabilities under this model; and (iv) sample from the multinomial distribution with cell probabilities determined at step (iii). We sampled 1000 tables for each of the sample sizes $100, 150, \ldots, 1500$ which represents a total of 29000 tables.

The undirected graph associated with the loglinear model [X1,X2][X2,IC] is shown in Figure 1. This graph has two edges that correspond with the two pairwise interaction terms of this model. The graph in Figure 1 has a weak decomposition $(\{X1\}, \{X2\}, \{IC\})$ that implies that, under this model, variable X1 is conditional independent of variable IC given variable X2. Thus, if IC represents a treatment variable, its effect on X1 is mediated by X2 under the loglinear model [X1,X2][X2,IC]. The regression model for X1 induced by model [X1,X2][X2,IC] involves X2 as the only predictor. The regression model for X2 induced by [X1,X2][X2,IC] involves both X1 and IC as predictors. The regression model for IC induced by model [X1,X2][X2,IC] involves X2 as the only predictor. For each of the 29000 sampled $4 \times 4 \times 2$ contingency tables, we determined the loglinear model with the smallest AIC, as well as the

CONTACT A. Dobra. Email: adobra@uw.edu

regression models with the smallest AIC for X1, X2 and IC. There are 8 hierarchical loglinear models with three variables we considered as candidate models: [X1][X2][IC], [X1,X2][IC], [X1][X2,IC], [X1,IC][X2], [X1,X2][X1,IC], [X1,X2][X2,IC], [X1,IC][X2,IC] and [X1,X2][X1,IC][X2,IC]. We fitted four ordinal logistic regressions for X1 that involve no predictors, X2 as the only predictor, IC as the only predictor, and both X1 and IC as predictors. We fitted the corresponding set of four ordinal logistic regressions for X2. We fitted four logistic regressions for IC that involve no predictors, X1 as the only predictor, X2 as the only predictor, and both X1 and X2 as predictors.

Figure 2 gives the proportion of times the true loglinear model [X1,X2][X2,IC] has been identified as the model with the smallest AIC for the 1000 sampled tables with sample sizes $100, 150, \ldots, 1500$. Figure 3 shows the proportion of times the true regression model for X1 with X2 as the only predictor has been identified as the model with the smallest AIC. Figure 4 shows the proportion of times the true regression model for X2 with both X1 and IC as predictors has been identified as the model with the smallest AIC. Figure 5 shows the proportion of times the true regression model for IC with X2 as the only predictor has been identified as the model with the smallest AIC. Figure 6 shows the proportion of times the true regression models for X1, X2 and IC have been jointly identified as the models with the smallest AIC among their corresponding set of candidate models.

The loglinear model [X1,X2][X2,IC] implies regression models for each of X1, X2 and IC, hence the correct identification of this model is equivalent to correctly identifying the three implied regression models for each of the variables X1, X2 and IC. As Figures 2 and 6 show, for each sample size, the proportion of times the loglinear model [X1,X2][X2,IC] had the smallest AIC is at least 50% larger than the proportion of times the implied regression models for X1, X2 and IC had the smallest AIC for the same sampled table. This difference seems to be larger for smaller sample sizes. Therefore, this simulation study shows that determining candidate mediating variables based on multivariate conditional independence relationships is performed more efficiently by searching for graphical loglinear models that are best supported by the data than by searching multivariate regressions.

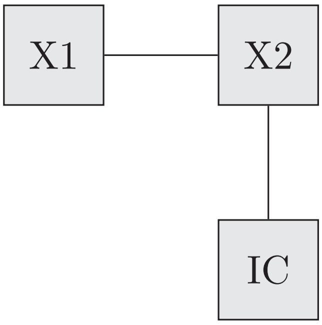

Figure 1. Interaction graph for the simulation study. This is the graphical loglinear model with minimal sufficient statistics [X1,X2][X2,IC]. Each vertex of this graph corresponds with an observed variable. Each edge of this graph corresponds with a pairwise interaction term.

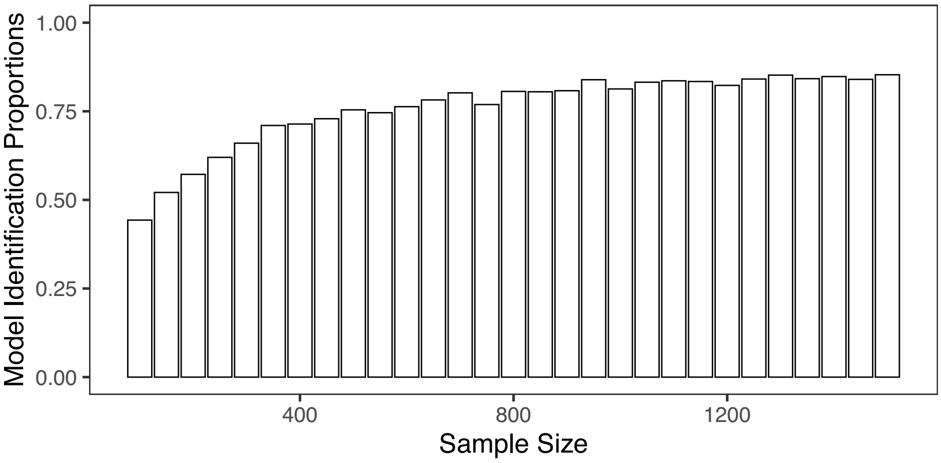

Figure 2. Graphical loglinear model identification in the simulation study. Each bar represents the proportion of times the true loglinear model [X1,X2][X2,IC] has been identified as the model with the smallest AIC.

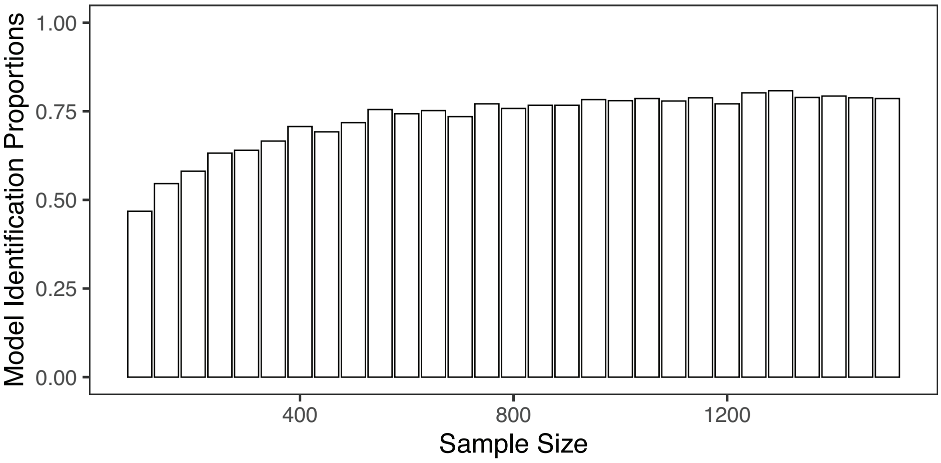

Figure 3. Identification of the regression for X1 in the simulation study. Each bar represents the proportion of times X2 was identified as the the only predictor of X1.

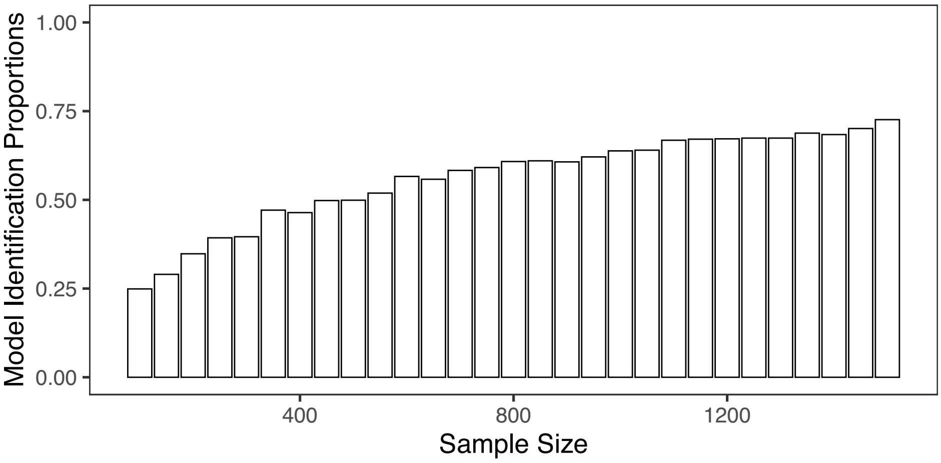

Figure 4. Identification of the regression for X2 in the simulation study. Each bar represents the proportion of times X1 and IC were identified as the two predictors X2.

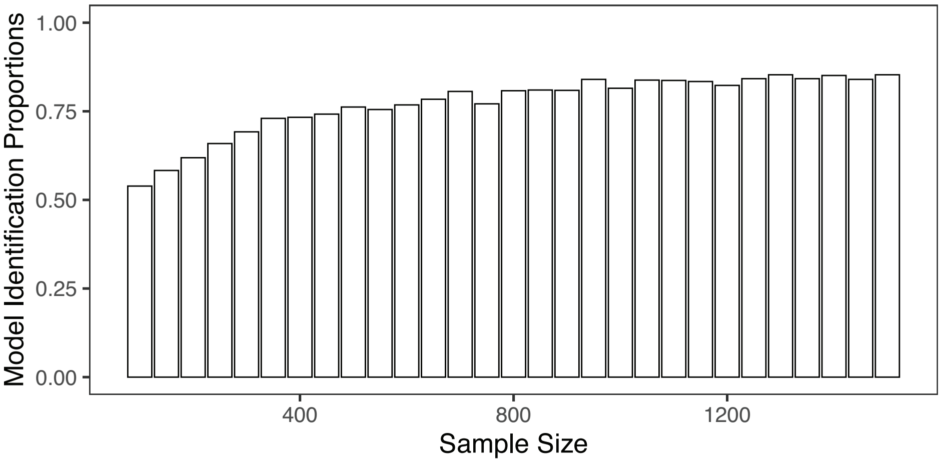

Figure 5. Identification of the regression for IC in the simulation study. Each bar represents the proportion of times X2 was identified as the only predictor of IC.

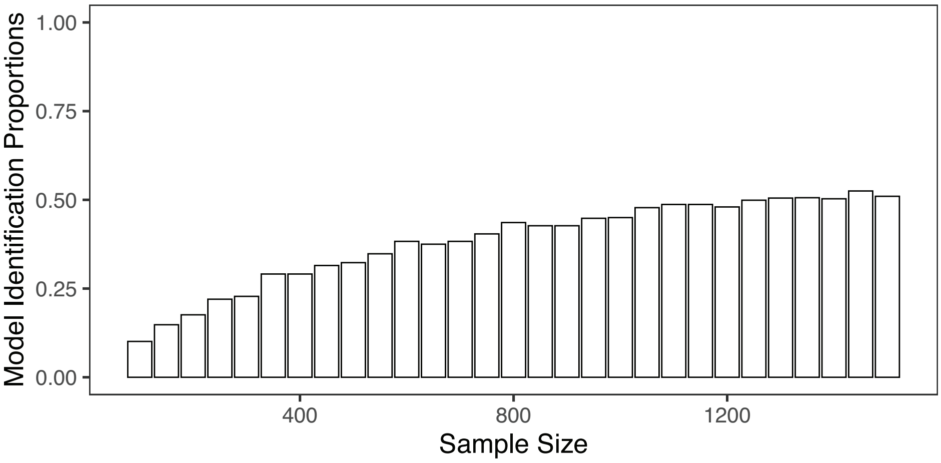

Figure 6. Joint identification of the regression models for X1, X2 and IC. Each bar represents the proportion of times the true sets of predictors for X1, X2 and IC were identified together.

```r
library(MASS)
#for polr
library(gRim)

tableDimens = c(4,4,2)

sampleSize = seq(from=100,to=1500,by=50)
nReplicates = 1000
variableNames = list()
variableNames[["X1"]] = seq(from=0,to=tableDimens[1]-1,by=1)
variableNames[["X2"]] = seq(from=0,to=tableDimens[2]-1,by=1)
variableNames[["IC"]] = seq(from=0,to=tableDimens[3]-1,by=1)

targetModelMargins = list(c(1,2),c(2,3))
targetModel = "~ X1:X2 + X2:IC"

AllModels = c("~ X1 + X2 + IC",
              "~ X1:X2 + IC",
              "~ X1 + X2:IC",
              "~ X1:IC + X2",
              "~ X1:X2 + X2:IC",
              "~ X1:X2 + X1:IC",
              "~ X1:IC + X2:IC",
              "~ X1:X2 + X2:IC + X1:IC")
regsX1 = c("X1 ~ 1",
           "X1 ~ X2",
           "X1 ~ IC",
           "X1 ~ X2 + IC")
targetModelX1 = regsX1[2]

regsX2 = c("X2 ~ 1",
           "X2 ~ X1",
           "X2 ~ IC",
           "X2 ~ X1 + IC")
targetModelX2 = regsX2[4]

regsIC = c("IC ~ 1",
           "IC ~ X1",
           "IC ~ X2",
           "IC ~ X1 + X2")
targetModelIC = regsIC[3]

simulateTable = function(tableDimens,sampleSize,targetModelMargins)
{
  p = runif(prod(tableDimens))
  tableProbs = p/sum(p)
  tableCounts = rmultinom(n=1,size = sampleSize,prob = tableProbs)
  rawTable = as.table(array(tableCounts, dim=tableDimens, dimnames=variableNames))
  rawLoglin = loglin(table = rawTable,margin = targetModelMargins,fit = TRUE,print = FALSE)
  tableCounts = rmultinom(n=1,size = sampleSize,prob = as.data.frame(rawLoglin$fit)[,"Freq"]/sampleSize)
  simTable = as.table(array(tableCounts, dim=tableDimens, dimnames=variableNames))
  return(simTable)
}

correctLoglin = numeric(length = length(sampleSize))
correctX1 = numeric(length = length(sampleSize))
correctX2 = numeric(length = length(sampleSize))
correctIC = numeric(length = length(sampleSize))
correctX1X2IC = numeric(length = length(sampleSize))
```

```
for(i in seq_len(length(sampleSize)))
{
  cat("Working ",i," sample size ",sampleSize[i],"\n")
  for(arep in seq(from=1,to=nReplicates,by=1))
  {
    simTable = simulateTable(tableDimens,sampleSize[i],targetModelMargins)
    loglinSelection = sapply(AllModels,function(x) { AIC(dmod(formula = formula
        (x),data = simTable))})
    if(names(which.min(loglinSelection))==targetModel)
    {
      correctLoglin[i] = correctLoglin[i]+1
    }

    regX1Selection = sapply(regsX1,function(x) { AIC(polr(formula = formula(x),
        data = as.data.frame(simTable),weights = Freq)) })
    if(names(which.min(regX1Selection))==targetModelX1)
    {
      correctX1[i] = correctX1[i]+1
    }

    regX2Selection = sapply(regsX2,function(x) { AIC(polr(formula = formula(x),
        data = as.data.frame(simTable),weights = Freq)) })
    if(names(which.min(regX2Selection))==targetModelX2)
    {
      correctX2[i] = correctX2[i]+1
    }

    regICSelection = sapply(regsIC,function(x) { AIC(glm(formula = formula(x),
        family = binomial(link = "logit"),data = as.data.frame(simTable),weights
         = Freq)) })
    if(names(which.min(regICSelection))==targetModelIC)
    {
      correctIC[i] = correctIC[i]+1
    }

    if((names(which.min(regX1Selection))==targetModelX1)&
       (names(which.min(regX2Selection))==targetModelX2)&
       (names(which.min(regICSelection))==targetModelIC)
      )
    {
      correctX1X2IC[i] = correctX1X2IC[i] + 1
    }
  }
}

correctLoglin = correctLoglin/nReplicates
correctX1 = correctX1/nReplicates
correctX2 = correctX2/nReplicates
correctIC = correctIC/nReplicates
correctX1X2IC = correctX1X2IC/nReplicates
Results = cbind(correctLoglin,correctX1,correctX2,correctIC,correctX1X2IC)
colnames(Results) = c("Loglin","X1","X2","IC","X1X2IC")
rownames(Results) = sampleSize

save(Results,file="simulationResults.Rdata")
```

Listing 1 Code needed to replicate the results in the simulation study from Section 1.