

Pseudocode for MRPC Update

Jarred M. Kvamme
University of Idaho
Department of Statistical Science

February 4, 2022

1.1 - The Trio Specific Case:

In a network consisting of three nodes: $G(V_1, T_1, T_2)$ we can identify the 5 possible topologies laid out under MRPC using results of the coefficient tests from the pair of regressions on each non-instrumental variable (Please refer to the **Appendix** at the end for Tables, Figures, and mathematical definitions).

Step 1. Perform two regressions treating each non-instrumental variable as the response once:

$$T_1 = \beta_0 + \beta_{11}V_1 + \beta_{21}T_2 + \Gamma U + \epsilon \quad (1)$$

$$T_2 = \beta_0 + \beta_{12}V_1 + \beta_{22}T_1 + \Gamma U + \epsilon \quad (2)$$

and retain the p -values $p_{\beta_{11}}, p_{\beta_{21}}, p_{\beta_{12}}, p_{\beta_{22}}$ and t-statistics $t_{\text{obs}_{11}}, t_{\text{obs}_{21}}, t_{\text{obs}_{12}}, t_{\text{obs}_{22}}$ for tests on the coefficients $\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}$

Step 2. Calculate the frequency of the minor allele of the variant/instrumental variable as f_{minor} .

Step 2.1 - If the minor variant frequency f_{minor} (allele frequency or copy number variation) is less than the predetermined threshold γ , perform the permuted regressions described in **Section 1.2** else proceed to **Step 3**

Step 3. Obtain the vector of p -values for hypothesis tests on $\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}$ (in this order) as the vector \mathbf{p} . If permuted regression was performed we will use the nominal p -values p_{21}^*, p_{22}^* for β_{21}, β_{22} . Convert \mathbf{p} into the indicator vector \mathbf{x}_p where 1 denotes a significant p -value at threshold α and 0 denotes a nonsignificant p -value

Step 4. If \mathbf{x}_p matches the expected result for the M3 topology given in **Tables 1 and 2** allocate the trio to M3, return the correct adjacency matrix for M3, and **Stop**. Else, test the marginal associations between V_1 and the T-nodes: T_1 and T_2 . Convert the p -values for the marginal tests into the indicator vector \mathbf{r}_p where 1 denotes a significant p -value at threshold α and 0 denotes a nonsignificant p -value

Step 5. - Compare \mathbf{x}_p and \mathbf{r}_p with the expected results for the other four model topologies given in **Tables 1 and 2** in the **Appendix** and allocate the trio to model type for which it matches (note that models M0, M1, and M2 have two cases each depending on the directions of the edges). If no match is available allocate the trio the class "other". Return the correct adjacency matrix from the inferred model

1.2 - Permuted Regression for Rare Variants

- We will apply the permuted regression described by Yang et al., 2017 whenever the instrumental variable contains a rare count at frequency $< \gamma$. The permuted regression

is preformed to obtain a robust estimate of the mediation effect between the nodes T_i and T_j in $G(V_k, T_i, T_j)$ which may be masked in the standard regression when V_k contains few observations for the minor variant. The algorithm for preforming the permuted regression is as follows:

Step 1. - permute T_2 in (1) within the levels of V_1 denoted T_2^* . Similarly, permute T_1 in (2) within the levels of V_1 denoted T_1^* .

Step 1.1 - Next preform the regressions in **Section 1.1** using the permuted variables:

$$T_1 = \beta_0 + \beta_{11}V_1 + \beta_{21}^*T_2^* + \Gamma\mathbf{U} + \epsilon \quad (3)$$

$$T_2 = \beta_0 + \beta_{12}V_1 + \beta_{22}^*T_1^* + \Gamma\mathbf{U} + \epsilon \quad (4)$$

Step 1.2 - store the observed t -statistic for β_{21}^* in the vector Θ_{21} and the t -statistic for β_{22}^* in the vector Θ_{22}

Step 2. repeat **Steps 1 - 1.2** m times to obtain the $m \times 1$ vectors of observed t -statistics Θ_{21} and Θ_{22} .

Step 2.1 - Using Θ_{21} , Θ_{22} , $t_{\text{obs}_{21}}$, and $t_{\text{obs}_{22}}$, calculate the permutation nominal p-values for β_{21} and β_{22}

$$Z_{\text{obs}_{ij}} = \frac{T_{\text{obs}_{ij}} - \frac{\sum \Theta_{ij}}{m}}{SE(\Theta_{ij})}$$

$$p_{\beta_{ij}}^* = 2 \times P(Z > |Z_{\text{obs}_{ij}}|)$$

Step 3. - retain $p_{\beta_{21}}^*$ and $p_{\beta_{22}}^*$ and return to **Step 3** in **Section 1.1**

2. General Algorithm

Step 1. - In this step we use the precision matrix to infer the graph skeleton by filtering unnecessary edges

Step 1.1 Given a data matrix \mathbf{X} of q instrumental variables, p non-instrumental variables, and g confounders. Calculate the partial correlation matrix \mathbf{H} and extract the first $\lambda = \{1 : q + p\}$ rows and columns of \mathbf{H} which represents the partial correlations between all nodes in the graph $G(V_1, V_2, \dots V_q, T_1, T_2, \dots T_p)$.

Step 1.2 - perform a partial correlation test on all non-diagonal entries in $\mathbf{H}[\lambda, \lambda]$ to obtain the $(q + p) \times (q + p)$ matrix of p -values \mathbf{P} corresponding to the nodes in $G(V_1, V_2, \dots, V_q, T_1, T_2, \dots, T_p)$.

Step 1.3 - For each non-diagonal entry in \mathbf{P} replace significant p -values at threshold α with 1 and nonsignificant p -values with 0 to obtain the $(q + p) \times (q + p)$ adjacency matrix \mathbf{A} for the skeleton of $G(V_1, V_2, \dots, V_q, T_1, T_2, \dots, T_p)$.

Step 2. - This step is to infer the network of trios involving the variant(s)

Step 2.1 - For each instrumental variable V_i : Allocate all possible trios with V_i such that the row in \mathbf{A} representing V_i has at least one edge between V_i and one of the non-instrumental variables in each trio. Preallocate all trios into a list where each entry in the list is the $n \times (3 + g)$ data matrix $[V_i, T_j, T_k, \mathbf{U}]$.

Let p_i be the number of non-instrumental variables that have an edge with V_i (e.g $p_i = \sum_{j=q+1}^p \mathbf{a}_{ij}$) then there are $p_i \times (p - p_i) + \binom{p_i}{2}$ trios that will be formed with V_i

Step 2.2 For each trio: Determine the directed structure using the regressions and tests outlined in **Sections 1.1-1.2**, and use the inference returned to update/direct the appropriate entries in the overall adjacency matrix \mathbf{A}

Step 3. - This step is to infer trios among only T-nodes by utilizing the inference made for trios in **Step 2** i.e direct additional edges in the overall network

Step 3.1 - Form all possible trios among strictly T-nodes such that the sub-adjacency matrix from \mathbf{A} corresponding to the T-nodes T_i, T_j, T_k denoted $\mathbf{A}_{3 \times 3}^{\text{sub}}$ has exactly two edges: i.e $(\sum \sum \mathbf{A}_{3 \times 3}^{\text{sub}} = 2)$. Preallocate all trios formed into a list where each entry in the list is the $n \times (3 + g)$ data matrix for the trio $[T_i, T_j, T_k, \mathbf{U}]$

Step 3.2 - For each trio: Determine the directed structure and use the inference to update the appropriate entries in adjacency matrix \mathbf{A} :

Step 3.2.1 - If $\mathbf{A}_{3 \times 3}^{\text{sub}}$ contains a single directed edge after it has been updated by **Step 2**, determine if the trio is M2 or M1 by treating the parent node of the directed edge as the variant (see **Figure 7**):

if we have $T_1 \rightarrow T_2$ and $T_2 - T_3$ then we perform only a single regression from **Section 1.1, Step 1**:

$$T_2 = \beta_0 + \beta_{12}T_1 + \beta_{22}T_3 + \mathbf{\Gamma}\mathbf{U} + \epsilon$$

to determine the direction of $T_2 - T_3$

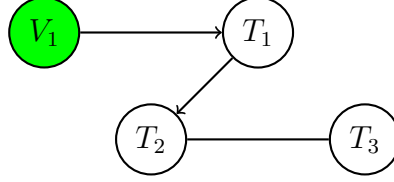


Figure 1: Example of a network among T-nodes where after **Step 2** in the **General Algorithm** we have one directed edge and one undirected edge in the graph $G(T_1, T_2, T_3)$. Utilizing this information we determine whether the edge $T_2 - T_3$ is $T_2 \rightarrow T_3$ (M1) or $T_2 \leftarrow T_3$ (M2)

Step 3.2.2 - If $\mathbf{A}_{3 \times 3}^{\text{sub}}$ has two undirected edges after updating, replace V_1 in **Section 1.1, Step 1** with one of the T-nodes:

$$T_1 = \beta_0 + \beta_{11}T_3 + \beta_{21}^*T_2^* + \mathbf{\Gamma}\mathbf{U} + \epsilon \quad (5)$$

$$T_2 = \beta_0 + \beta_{12}T_3 + \beta_{22}^*T_1^* + \mathbf{\Gamma}\mathbf{U} + \epsilon \quad (6)$$

and determine if the network is M2 using the regressions and tests from **Section 1.1**

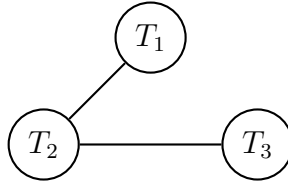


Figure 2: Example of the network among T-nodes where after **Step 2** in the **General Algorithm** we have two undirected edges in the graph $G(T_1, T_2, T_3)$. Without any additional information we can only uniquely infer if the trio is M2 (see **Figure 2**)

Return **A**

3. Simulation Strategy to Validate Section 1.1

To verify that the indicator vectors \mathbf{x}_p derived from the hypothesis tests in **Section 1.1** are sufficient in identifying the proposed model structures from MRPC given in **Tables 1 and 2**, we propose the following simulation methodology:

(1.) - Using the linear models for the 5 basic topologies described by Badsha and Fu 2019, we propose to simulate each trio topology under varying signal strengths (and possibly sample size) using the simulation functions in the *R* package *MRPC*.

(2.) - we then apply the algorithm proposed in **Section 1.1** to determine if the expected results given in **Tables 1 and 2** are identified. we compare the indicator vector \mathbf{x}_p to the expected indicator vectors for each topology and allocate the trio to one of the 5 topologies as laid out in **Section 1.1 Step 5**. Our goal is to determine if the inference for each trio is adequate for identifying the generating model or if unexpected indicator vectors arise.

(3.) - Investigate, if any, the indicator vectors allocated to the "Other" class.

Appendix

Definitions

V_i - The i^{th} instrumental variable when $q > 1$

T_i - a non-instrumental variable/node

p - the number of non-instrumental variables/nodes in a network

q - the number of instrumental variables

g - the number of confounding variables selected for a network

m - the number of permutations to preform in a permuted regression (mediation test)

n - the sample size of the data

\mathbf{U} - the $(n \times g)$ matrix whose columns represent confounding variables

\mathbf{X} - the $(n \times p + q + g)$ data matrix of all variables/nodes and all confounders

\mathbf{H} - the $(p + q + g \times p + q + g)$ precision matrix

A - a $(p + q \times p + q)$ adjacency matrix for the network

$G(A, B, C)$ - a graph with nodes A, B, and C

$F(\cdot)$ - the Fisher transformation function

f_{minor} - The frequency of the minor variant of V_i (when V represents a type of genetic variation)

γ - the threshold frequency of the minor variant for which we determine if a permuted regression is needed

$\rho_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}$ - the partial correlation between the i^{th} and j^{th} columns/variables of \mathbf{X}

Table 1: - Expected results for the tests on the regression coefficients under each model scenario (trios with variants only).

| Model | β_{11} | β_{21} | β_{12} | β_{22} | $V_1 \perp\!\!\!\perp T_2$ | $V_1 \perp\!\!\!\perp T_1$ |
|---|-----------------------|-----------------------|-----------------------|-----------------------|----------------------------|----------------------------|
| M0 | $\neq 0$ | $= 0$ | $= 0$ | $= 0$ | | |
| | $= 0$ | $= 0$ | $\neq 0$ | $= 0$ | | |
| M1 | $\neq 0$ | $\neq 0$ | $= 0$ | $\neq 0$ | | |
| | $= 0$ | $\neq 0$ | $\neq 0$ | $\neq 0$ | | |
| M2 | $\neq 0$ | $\neq 0$ | $\neq 0$ | $\neq 0$ | Yes | |
| | $\neq 0$ | $\neq 0$ | $\neq 0$ | $\neq 0$ | | Yes |
| M3 | $\neq 0$ | $= 0$ | $\neq 0$ | $= 0$ | – | – |
| M4 | $\neq 0$ | $\neq 0$ | $\neq 0$ | $\neq 0$ | No | |
| Conditionally: $Y \sim$ | $V_i T_j, \mathbf{U}$ | $T_j V_i, \mathbf{U}$ | $V_i T_i, \mathbf{U}$ | $T_j V_i, \mathbf{U}$ | | |

$$T_1 = \beta_0 + \beta_{11}V_1 + \beta_{21}T_2 + \mathbf{\Gamma U} + \epsilon \quad (7)$$

$$T_2 = \beta_0 + \beta_{12}V_1 + \beta_{22}T_1 + \mathbf{\Gamma U} + \epsilon \quad (8)$$

Table 2: - Expected indicator table for the tests on the regression coefficients under each model scenario (trios with variants only). Note that 1 indicates a rejection of H_0 and 0 indicates a failure to reject

| Model | $H_0 : \beta_{11} = 0$ | $H_0 : \beta_{21} = 0$ | $H_0 : \beta_{12} = 0$ | $H_0 : \beta_{22} = 0$ | $H_0 : V_1 \perp\!\!\!\perp T_2$ | $H_0 : V_1 \perp\!\!\!\perp T_1$ |
|-----------|------------------------|------------------------|------------------------|------------------------|----------------------------------|----------------------------------|
| M0 | 1 | 0 | 0 | 0 | 0 | 1 |
| | 0 | 0 | 1 | 0 | 1 | 0 |
| M1 | 1 | 1 | 0 | 1 | 1 | 1 |
| | 0 | 1 | 1 | 1 | 1 | 1 |
| M2 | 1 | 1 | 1 | 1 | 0 | 1 |
| | 1 | 1 | 1 | 1 | 1 | 0 |
| M3 | 1 | 0 | 1 | 0 | 1 | 1 |
| M4 | 1 | 1 | 1 | 1 | 1 | 1 |

$$\mathbf{x}_p = \begin{cases} 1 & \text{if } \beta_{ij} \neq 0 \\ 0 & \text{else} \end{cases}$$

$$\mathbf{r}_p = \begin{cases} 1 & \text{if } V_i \not\perp\!\!\!\perp T_j \\ 0 & \text{else} \end{cases}$$

Table 3: - The set up for the adjacency matrix for each 3-node network denoted by the graph $G(T_i, T_j, T_k)$ $i, j, k \in \{1 : p\}; \forall i \neq j \neq k$ found using the regressions outlined in **Section 1.3**. Each entry in the table shows the null hypothesis used for testing the edge between the node in the row and node in the column. An entry with 1 means we keep the edge between the nodes (i.e we reject H_0) and a 0 means we remove the edge between the nodes (i.e we fail to reject H_0)

| Response | T_i | T_j | T_k |
|----------|--|--|--|
| T_i | 0 | $H_0 : T_i \perp\!\!\!\perp T_j T_k, \mathbf{U}$ | $T_i \perp\!\!\!\perp T_k T_j, \mathbf{U}$ |
| T_j | $H_0 : T_j \perp\!\!\!\perp T_i T_k, \mathbf{U}$ | 0 | $T_j \perp\!\!\!\perp T_k T_i, \mathbf{U}$ |
| T_k | $H_0 : T_k \perp\!\!\!\perp T_i T_j, \mathbf{U}$ | $H_0 : T_k \perp\!\!\!\perp T_j T_i, \mathbf{U}$ | 0 |

The 5 MRPC Topologies

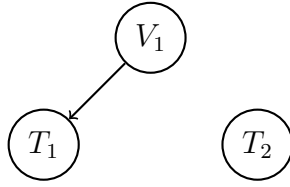


Figure 3: M0 - $V_1 \not\perp T_1; V_1 \perp T_2; T_1 \perp T_2$

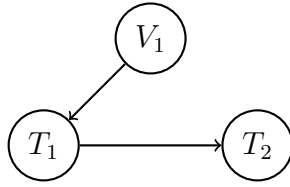


Figure 4: M1 - $V_1 \not\perp T_1; V_1 \not\perp T_2; T_1 \not\perp T_2; V_1 \perp T_2|T_1$

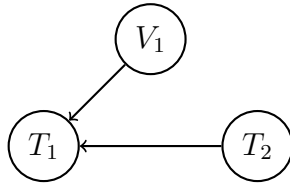


Figure 5: M2 - $V_1 \not\perp T_1; V_1 \perp T_2; T_1 \not\perp T_2; V_1 \not\perp T_2|T_1$

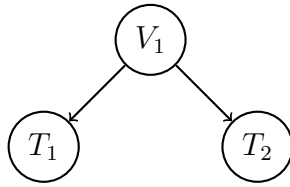


Figure 6: fig: M3 - $V_1 \not\perp T_1; V_1 \not\perp T_2; T_1 \not\perp T_2; T_1 \perp T_2|V_1$

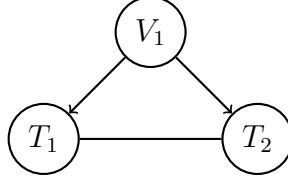


Figure 7: fig: M4 - $V_1 \not\perp\!\!\!\perp T_1$; $V_1 \not\perp\!\!\!\perp T_2$; $T_1 \not\perp\!\!\!\perp T_2$; $T_1 \not\perp\!\!\!\perp T_2|V_1$

Inferring the Network Among Non-Instrumental Variables

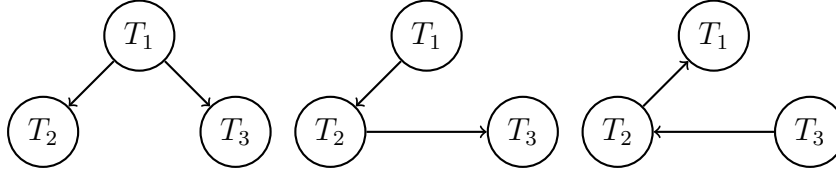


Figure 8: The three graphs above are Markov equivalent meaning they share the same conditional and marginal independence relations Markov: $T_1 \perp\!\!\!\perp T_3|T_2$ Minimality: $T_1 \not\perp\!\!\!\perp T_2$; $T_2 \not\perp\!\!\!\perp T_3$ faithfulness: $T_1 \not\perp\!\!\!\perp T_3$ and are therefore indistinguishable from each other

Details On Statistical Methods

Calculating Precision

Given a data matrix \mathbf{X} of q instrumental variables, p non-instrumental variables, and g confounders:

Assuming \mathbf{X} is centered:

$$X \sim N_k(\mathbf{0}, \Sigma) \quad \text{for } k = p + q + g$$

Then the precision matrix of \mathbf{X} is defined as

$$\mathbf{H} = \Sigma^{-1}$$

\mathbf{H} can be scaled to the partial correlation matrix for the entries in \mathbf{X} . Given the entry in the i^{th} row and j^{th} column of \mathbf{H} :

$$\mathbf{x}_i, \mathbf{x}_j | \mathbf{x}_{-(i,j)} = -\frac{h_{ij}}{\sqrt{h_{ii}}\sqrt{h_{jj}}} = \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}$$

which is a measure of the association between the i^{th} and j^{th} columns/variables in \mathbf{X} conditioned on all other variables.

The Fisher transformation can be used to formulate a test for each partial correlation coefficient of interest:

$$\frac{\sqrt{n - |\mathbf{x}_{-i,j}|} - 3}{2} \ln \left(\frac{1 + \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}}{1 - \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}}} \right) \approx N(0, 1)$$

where null and alternative hypotheses are

$$H_0 : \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}} = 0 \quad H_A : \hat{\rho}_{\mathbf{x}_i, \mathbf{x}_j \cdot \mathbf{x}_{-(i,j)}} \neq 0$$

$$\text{reject } H_0 \text{ if } |Z_{\text{obs}}| > Z_{1-\alpha/2}$$

by applying the cases:

$$a_{i,j} = \begin{cases} 1 & \text{if } 2 \times P(Z > |Z_{\text{obs}}|) < \alpha \\ 0 & \text{else} \end{cases} \quad \forall i, j \in \{1 : p + q\}$$

we can obtain the $(p + q \times p + q)$ adjacency matrix \mathbf{A} for the network skeleton

Permuted Regression mediation test

repeat m times: permute T_j in (1) within the levels of V_i denoted T_j^* . Similarly, permute T_i in (2) within the levels of V_k denoted T_i^* . Next perform the regressions using the permuted variables:

$$T_i = \beta_0 + \beta_{1i}V_k + \beta_{2i}^*T_j^* + \mathbf{\Gamma}\mathbf{U} + \epsilon \quad (9)$$

$$T_j = \beta_0 + \beta_{1j}V_k + \beta_{2j}^*T_i^* + \mathbf{\Gamma}\mathbf{U} + \epsilon$$

Let Θ_{2i} and Θ_{2j} denote the $(m \times 1)$ vectors representing the collection of t statistics from the wald tests on β_{2i}^* and β_{2j}^* coefficients (respectively) from the permuted regressions in **Step 2.** such that:

$$\Theta_{2i} = \left[T_{2i}^{*(1)}, T_{2i}^{*(2)}, T_{2i}^{*(3)}, \dots \right]$$

$$\Theta_{2j} = \left[T_{2j}^{*(1)}, T_{2j}^{*(2)}, T_{2j}^{*(3)}, \dots \right]$$

We next test the conditional association between T_i and T_j using the nominal test defined by Yang et. al., 2017. Let $T_{\text{obs}_{2i}}$ be the observed wald statistic from (1) and $T_{\text{obs}_{2j}}$ be the observed wald statistic from (2). We formulate the testable hypotheses:

$$H_0 : T_{\text{obs}_{2i}} = \mu_{2i}^*, \quad H_A : T_{\text{obs}_{2i}} \neq \mu_{2i}^*$$

and

$$H_0 : T_{\text{obs}_{2j}} = \mu_{2j}^*, \quad H_A : T_{\text{obs}_{2j}} \neq \mu_{2j}^*$$

where μ_{2i}^* and μ_{2j}^* denote the centers of the non-central t -distributions of Θ_{2i} and Θ_{2j} respectively. Therefore the mediation test statistic is:

$$Z_{\text{obs}_{ij}} = \frac{T_{\text{obs}_{ij}} - \frac{\sum \Theta_{ij}}{m}}{SE(\Theta_{ij})}$$

where we

$$\text{reject } H_0 \text{ if } 2 \times P(Z > |Z_{\text{obs}_{ij}}|) < \alpha$$