Pseudocode for MRPC Update

Jarred M. Kvamme
University of Idaho
Department of Statistical Science

January 7, 2022

1.1 The Trio Specific Case:

In a network consisting of three nodes: $G(V, T_1, T_2)$ 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).

we preform two regressions using each non-intrumental variable node T_i as the response once to yield the pair of linear models:

$$T_1 = \beta_0 + \beta_{11}V + \beta_{21}T_2 + \mathbf{\Gamma U} + \epsilon \tag{1}$$

$$T_2 = \beta_0 + \beta_{21}V + \beta_{22}T_1 + \mathbf{\Gamma U} + \epsilon \tag{2}$$

we also estimate and test the marginal relationship between V and T_2. Let

$$\hat{r} = \frac{\text{cov}(V, T_2)}{\hat{\sigma}_V \hat{\sigma}_{T_2}}$$

we can obtain the hypothesis test

$$H_0 : r = 0 \quad H_A : r \neq 0$$

which after use of the Fisher transformation

$$F(\hat{r}) = \frac{1}{2} \ln \left(\frac{1 + \hat{r}}{1 - \hat{r}} \right)$$

approximates a normal distribution asymptotically in the sample size with parameters

$$\text{mean} = F(r) \text{ and SE} = \frac{1}{\sqrt{n-3}}$$

therefore by standardizing we have

$$Z_{\text{obs}} = \frac{F(\hat{r}) - F(r_0)}{SE} = \frac{\sqrt{n-3}}{2} \ln \left(\frac{1 + \hat{r}}{1 - \hat{r}} \right) \approx N(0, 1)$$

where we

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

where r_0 denotes the correlation under the null hypothesis and $F(r_0)$ is zero.

Combining the results of the tests for all β_{ij} regression coefficients from the models above and the marginal test defined above, we can determine the topology of the network (see **Table 1** in **Appendix**)

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 T_i and T_j in $G(V, T_i, T_j)$ which may be masked in the standard regression when V contains few observations for the minor variant. The algorithm for preforming the permuted regression is as follows:

step 1. - Let f_{minor} be the frequency of the minor variant of V. If $f_{\text{minor}} < \gamma$ preform the permuted regression(s) else preform the standard regression(s).

step 2. - repeat m times: permute T_2 in (**1**) within the levels of V denoted T_2^*. Similarly, permute T_1 in (**2**) within the levels of V denoted T_1^*. Next preform the regressions using the permuted variables:

$$T_1 = \beta_0 + \beta_{11}V + \beta_{21}^* T_2^* + \mathbf{\Gamma U} + \epsilon \tag{3}$$

$$T_2 = \beta_0 + \beta_{21}V + \beta_{22}^* T_1^* + \mathbf{\Gamma U} + \epsilon$$

Step 3. - Let $\mathbf{\Theta_{21}}$ and $\mathbf{\Theta_{22}}$ denote $(m \times 1)$ vectors representing the collection of T statistics from the wald tests on β_{21}^* and β_{22}^* coefficients (respectively) from the permuted regressions in **Step 2.**. such that:

$$\mathbf{\Theta}_{21} = \left[T_{21}^{*(1)},\ T_{21}^{*(2)},\ T_{21}^{*(3)}, \cdots \right]$$

$$\mathbf{\Theta}_{22} = \left[T_{22}^{*(1)},\ T_{22}^{*(2)},\ T_{22}^{*(3)}, \cdots \right]$$

We next test the association between T_1 and T_2 using the nominal test defined by Yang et. al., 2017. Let $T_{\text{obs}_{21}}$ be the observed wald statistic from (**1**) and $T_{\text{obs}_{22}}$ be the observed wald statistic from (**2**). We formulate the testable hypotheses:

$$H_0 : T_{\text{obs}_{21}} = \mu_{21}^*, \quad H_A : T_{\text{obs}_{21}} \neq \mu_{21}^*$$

and

$$H_0 : T_{\text{obs}_{22}} = \mu_{22}^*, \quad H_A : T_{\text{obs}_{22}} \neq \mu_{22}^*$$

where μ_{21}^* and μ_{22}^* denote the centers of the non-central T distributions of $\mathbf{\Theta_{21}}$ and $\mathbf{\Theta_{22}}$ respectively. Therefore the mediation test statistic is:

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

where we

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

2. General Algorithm

Step 1. - Starting with a fully connected graph of $p+1$ nodes, compute the precision matrix of the data:
Assuming $\mathbf{X}$ is centered:

$$X \sim N_k(\mathbf{0}, \mathbf{\Sigma}) \quad \text{for } \mathbf{k} = \mathbf{p} + \mathbf{q} + \mathbf{1}$$

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

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

$\mathbf{H}$ 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:

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

we can obtain the $(p+1 \times p+1)$ adjacency matrix $\mathbf{A}$ for the network skeleton

The use of the partial correlations matrix has been shown to be applicable for the construction of an undirected acyclic graph (insert references here). We suggest here to use the partial correlation network obtained from **Step 1.** to construct the graph skeleton

Step 2. - Determine the directed structure of all $\binom{p}{2}$ possible 3-node networks involving the instrumental variable(s) using the regressions and tests outlined in **Section 1.**. Here we are breaking up the structure of the larger network into trios of nodes involving the instrumental variable i.e all $G(V, T_i, T_j)\ \forall i \neq j$

> **Step 2.1** - (Specific to Genomics) if the minor variant frequency (allele frequency or copy number variation) of the instrumental variable V is less than the predetermined threshold $i.e < \gamma$, preform the permuted regression described in **Section 1.2**.

Step 3. - Determine the directed structure of all $\binom{p}{3}$ possible 3-node networks involving only the non-instrumental variable nodes using the regressions and tests outlined in **Section 1.**. This step is to find edges that may be explained away when conditioning on other nodes in the network. i.e we infer all possible $G(T_i, T_j, T_k)\ \forall i \neq j \neq k$

Appendix

Definitions
V - The instrumental variable
T_i - a non-instrumental variable/node
p - the number of non-instrumental variables/nodes in a network
q - 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 $(n \times q)$ matrix whose columns represent confounding variables
$\mathbf{X}$ - the $(n \times p+q+1)$ data matrix of all variables/nodes and all confounders
$\mathbf{H}$ - the $(p+q+1 \times p+q+1)$ precision matrix
A - a $(p+1 \times p+1)$ adjacency matrix for the network
$G(A, B, C)$ - a graph with nodes A, B, and C
$F(\cdot)$ - the Fisher transformation function
γ - the minimum frequency for which we would preform 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

Model	$\beta_{\mathbf{11}}$	$\beta_{\mathbf{21}}$	$\beta_{\mathbf{12}}$	$\beta_{\mathbf{22}}$	$V \perp\!\!\!\perp T_2$
M0	$= 0$	$\neq 0$	$= 0$	$= 0$	Yes
	$= 0$	$= 0$	$= 0$	$\neq 0$	No
M1	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	No
M2	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	Yes
M3	$= 0$	$\neq 0$	$= 0$	$\neq 0$	Yes
M4	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	No

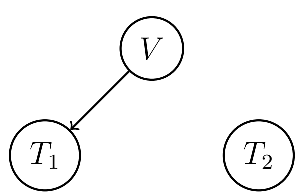

Figure 1: M0 - $V \not\perp\!\!\!\perp T_1; V \perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2$

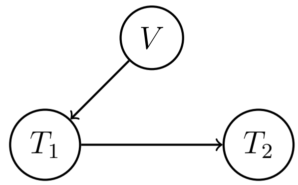

Figure 2: M1 - $V \not\perp\!\!\!\perp T_1; V \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V \perp\!\!\!\perp T_2 | T_1$

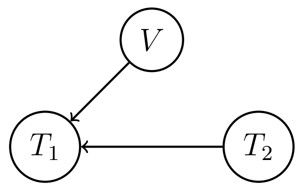

Figure 3: M2 - $V \not\perp\!\!\!\perp T_1; V \perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V \not\perp\!\!\!\perp T_2 | T_1$

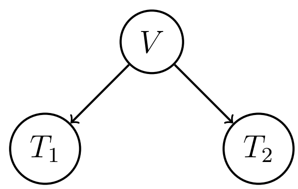

Figure 4: fig: M3 - $V \not\perp\!\!\!\perp T_1; V \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2 | V$

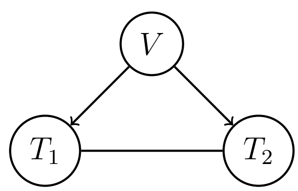

Figure 5: fig: M4 - $V \not\perp\!\!\!\perp T_1; V \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2 | V$