Pseudocode for MRPC Update

Jarred M. Kvamme
University of Idaho
Department of Statistical Science

January 12, 2022

1. - Details On Statistical Methods

1.1 - The Trio Specific Case:

In a network consisting of three nodes: $G(V_k, T_i, T_j)$ 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 as the response once and we include the set of selected confounding variables/Principle components **U**. This produces the pair of linear models:

$$T_i = \beta_0 + \beta_{1i}V_k + \beta_{2i}T_j + \mathbf{\Gamma U} + \epsilon \tag{1}$$

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

we also test the marginal relationship between V_k and T_j. Combining the results of the tests for all β_{ij} regression coefficients from the models above and the marginal test between V_k and T_j, we can determine the topology of the trio (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 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. - Let f_{minor} be the frequency of the minor variant of V_k. If $f_{\text{minor}} < \gamma$ preform the permuted regression(s) else preform the standard regression(s).

step 2. - 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 preform the regressions using the permuted variables:

$$T_i = \beta_0 + \beta_{1i}V_k + \beta_{2i}^*T_j^* + \mathbf{\Gamma U} + \epsilon \tag{3}$$

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

Step 3. - Let $\boldsymbol{\Theta}_{2i}$ and $\boldsymbol{\Theta}_{2j}$ denote $(m \times 1)$ 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:

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

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

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

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

and

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

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

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

where we

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

1.3 - Inferring Trios Without Variants

We can infer the graph skeleton for any 3-node network $G(T_i, T_j, T_k)$ using the tests on the coefficients from the linear system obtained from regressing each node on the other nodes and confounders (see **Table 2**):

$$T_i = \beta_0 + \beta_{1i}T_j + \beta_{2i}T_k + \boldsymbol{\Gamma}\mathbf{U} + \epsilon$$

$$T_j = \beta_0 + \beta_{1j}T_i + \beta_{2j}T_k + \boldsymbol{\Gamma}\mathbf{U} + \epsilon$$

$$T_k = \beta_0 + \beta_{1k}T_i + \beta_{2k}T_j + \boldsymbol{\Gamma}\mathbf{U} + \epsilon$$

this is equivalent to the partial correlations between each pair of nodes adjusted for the other node and confounders i.e $\hat{\rho}_{\mathbf{T_i},\mathbf{T_j} \cdot \mathbf{T_k},\mathbf{U}}$ for any $i \neq j \neq k$

2. General Algorithm

Step 1. - Starting with a fully connected graph of $p+q$ nodes and g selected confounding variables/PC's, compute the precision matrix of the data:

Assuming $\mathbf{X}$ is centered:

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

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 ith 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 ith 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

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 $q \times \binom{p}{2}$ 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 all possible 3-node networks involving two non-instrumental variables and a single instrumental variable i.e all $G(V_k, T_i, T_j)$ $k \in \{1 : q\}; i, j \in \{1 : p\}; \forall i \neq j$. Then update the enties $a_{k,i}, a_{k,j}, a_{i,j}$ in the symmetric adjacency matrix $\mathbf{A}$.

> **Step 2.1** - (Specific to Genomics) if the minor variant frequency f_{minor} (allele frequency or copy number variation) of the instrumental variable V_i is less than the predetermined threshold γ, preform the permuted regression described in **Section 1.2** on all 3-node networks involving V_i. Repeat for each V_i in which $f_{\text{minor}} < \gamma$.

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 1.3**. This step is to find edges that may be explained away when conditioning on other non-instrumental variable nodes in the network. i.e we infer all possible $G(T_i, T_j, T_k)$ $i, j, k \in \{1 : p\}; \forall i \neq j \neq k$. Then update the enties $a_{k,i}, a_{k,j}, a_{i,j}$ in the symmetric adjacency matrix $\mathbf{A}$.

Appendix

Definitions

V_i - The ith 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

$\rho_{\mathbf{x_i}, \mathbf{x_j} \cdot \mathbf{x}_{-(\mathbf{i}, \mathbf{j})}}$ - the partial correlation between the ith 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	$\beta_{\mathbf{1i}}$	$\beta_{\mathbf{2i}}$	$\beta_{\mathbf{1j}}$	$\beta_{\mathbf{2j}}$	$V_i \perp\!\!\!\perp T_j$
M0	$\neq 0$	$= 0$	$= 0$	$= 0$	Yes
	$= 0$	$= 0$	$\neq 0$	$= 0$	No
M1	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	No
M2	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	Yes
M3	$\neq 0$	$= 0$	$\neq 0$	$= 0$	No
M4	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	No
Conditionally: $Y \sim$	$V_i \vert T_j, \mathbf{U}$	$T_j \vert V_i, \mathbf{U}$	$V_i \vert T_i, \mathbf{U}$	$T_j \vert V_i, \mathbf{U}$	

Table 2: - 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)

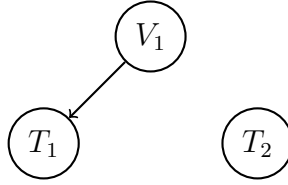| **Response**	T_i	T_j	T_k
T_i	0	$H_0 : T_i \perp\!\!\!\perp T_j|T_k, \mathbf{U}$	$T_i \perp\!\!\!\perp T_k|T_j, \mathbf{U}$
T_j	$H_0 : T_j \perp\!\!\!\perp T_i|T_k, \mathbf{U}$	0	$T_j \perp\!\!\!\perp T_k|T_i, \mathbf{U}$
T_k	$H_0 : T_k \perp\!\!\!\perp T_i|T_j, \mathbf{U}$	$H_0 : T_k \perp\!\!\!\perp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

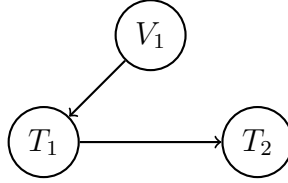Figure 1: M0 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2$

Figure 2: M1 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V_1 \perp\!\!\!\perp T_2|T_1$
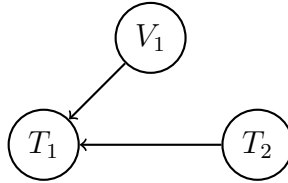

Figure 3: M2 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V_1 \not\perp\!\!\!\perp T_2|T_1$
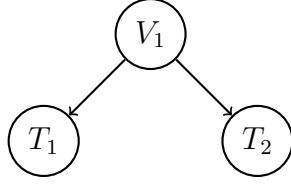
Figure 4: fig: M3 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2 | V_1$
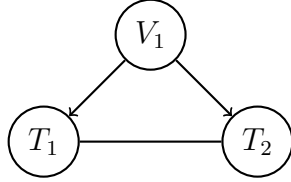
Figure 5: fig: M4 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2 | V_1$