Pseudocode for MRPC Update

Jarred M. Kvamme
University of Idaho
Department of Statistical Science

January 12, 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. Calculate the frequency of the minor variant f_{minor}. If V_1 is an eQTL, then f_{minor} is the frequency of the minor allele. If V_1 is not and eQTL, then calculate f_{minor} as the frequency of the least common count or class.

Step 2. Preform two regressions treating each non-instumental variable as the response once:

$$T_1 = \beta_0 + \beta_{11} V_1 + \beta_{21} T_2 + \mathbf{\Gamma U} + \epsilon \tag{1}$$

$$T_2 = \beta_0 + \beta_{12} V_1 + \beta_{22} T_1 + \mathbf{\Gamma U} + \epsilon \tag{2}$$

> **Step 2.2** - If the minor variant frequency f_{minor} (allele frequency or copy number variation) from **Step 1.** is less than the predetermined threshold γ, preform the permuted regressions described in **Section 1.2**.

Step 3. calculate the correlation matrix for $[V_1, T_1, T_2]$ and preform hypothesis testing on r_{V_1,T_2} and r_{V_1,T_1} i.e the marginal relationships between V_1, T_2 and V_1, T_1

Step 4. following **Steps 1 - 3** obtain the vector of p-values for hypothesis tests on $\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, r_{V_1,T_2}$, and r_{V_1,T_1} (in this order) as the vector $\mathbf{p}$

> **Step 4.1** Convert the vector of p-values $\mathbf{p}$ into the indicator vector $\mathbf{I}_p$ where 1 denotes a significant p-value at threshold α and 0 denotes a nonsignificant p-value

Step 5. - Compare $\mathbf{I}_p$ with the expected results for each model topology 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".

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_1. If $f_{\text{minor}} < \gamma$ proceed to **Step 2.**

Step 2. - Preform the standard regressions in given in **(1) and (2)** and retain the observed t-statistics for the tests on β_{21} and β_{22} (which we will denote as $t_{\text{obs}_{21}}$ and $t_{\text{obs}_{22}}$, respectively)

> **Step 2.2** - Similarly, retain the observed p-values for the tests on β_{11} and β_{12} (which we will denote $p_{\beta_{11}}$ and $p_{\beta_{12}}$ respectively)

Step 3. - 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 3.1** - Next preform the regressions in **Section 1.1** using the permuted variables:

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

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

> **Step 3.2** - store the observed t-statistic for β_{21}^* in the vector $\mathbf{\Theta}_{21}$ and the t-statistic for β_{22}^* in the vector $\mathbf{\Theta}_{22}$

Step 4. repeat **Steps 3 - 3.2** m times to obtain the $m \times 1$ vectors of observed t-statistics $\mathbf{\Theta}_{21}$ and $\mathbf{\Theta}_{22}$.

> **Step 4.1** - Using $\mathbf{\Theta}_{21}$, $\mathbf{\Theta}_{22}$, $t_{\text{obs}_{21}}$, and $t_{\text{obs}_{21}}$, calculate the permutation nominal p-values for β_{21} and β_{22} (which we will denote $p_{\beta_{21}}^*$ and $p_{\beta_{22}}^*$ respectively)

Step 5. calculate the correlation matrix for $[V_1, T_1, T_2]$ and preform hypothesis testing on r_{V_1,T_2} and r_{V_1,T_1} i.e the marginal relationships between V_1, T_2 and V_1, T_1

Step 6. - allocate the vector of p-values for hypothesis tests on $\beta_{11}, \beta_{21}, \beta_{12}, \beta_{22}, r_{V_1,T_2}$, and r_{V_1,T_1} (in this order) as the vector $\mathbf{p}$ (using the nominal p-values for β_{21} and β_{22}) and proceed to **Step 4.1** in **Section 1.1**

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**):

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

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

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

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. - 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 $\lambda = \{1 : p + q\}$ rows and columns of $\mathbf{H}$ which represents the partial correlations between all nodes in the graph $G(V_1, V_2, ...V_q, T_1, T_2, ...T_p)$.

> **Step 1.1** - preform a partial correlation test on all non-diagonal entries in $\mathbf{H}[\lambda, \lambda]$ to obtain the $(p + q \times p + q)$ matrix of p-values $\mathbf{P}$.
>
> **Step 1.2** -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 $(p + q \times p + q)$ adjacency matrix $\mathbf{A}$ for the skeleton of $G(V_1, V_2, ...V_q, T_1, T_2, ...T_p)$

Step 2. - pre-allocate the data matrices for all $q \times \binom{p}{2}$ possible 3-node networks involving the instrumental variable(s) into a list. Each entry in the list is the $(n \times g + 3)$ data matrix $[V_k, T_i, T_j, \mathbf{U}]$

> **Step 2.1** - Similarly, construct the list of all $\binom{p}{3}$ possible 3-node networks involving only the non-instrumental variable nodes. Each entry in the list is the $(n \times g + 3)$ data matrix $[T_k, T_i, T_j, \mathbf{U}]$

Step 3. - Determine the directed structure of each 3-node network involving the instrumental variable(s) from **Step 2** using the regressions and tests outlined in **Section 1.1**. Here we are breaking up the structure of the larger network by inferring the graph for 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 3.1 - (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 4. - 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 $G(T_i, T_j, T_k) \quad i, j, k \in \{1 : p\}; \forall i \neq j \neq k$. 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 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

$\rho_{\mathbf{x_i},\mathbf{x_j}\cdot\mathbf{x}_{-(\mathbf{i},\mathbf{j})}}$ - 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	$\beta_{\mathbf{11}}$	$\beta_{\mathbf{21}}$	$\beta_{\mathbf{12}}$	$\beta_{\mathbf{22}}$	$V_1 \perp\!\!\!\perp T_2$	$V_1 \perp\!\!\!\perp T_1$
M0	$\neq 0$	$= 0$	$= 0$	$= 0$	Yes	
	$= 0$	$= 0$	$\neq 0$	$= 0$		Yes
M1	$\neq 0$	$\neq 0$	$\neq 0$	$= 0$	No	
	$\neq 0$	$= 0$	$\neq 0$	$\neq 0$		No
M2	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	Yes	
	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$		Yes
M3	$\neq 0$	$= 0$	$\neq 0$	$= 0$	No	
M4	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	No	
Conditionally: $Y \sim$	$V_i \mid T_j, \mathbf{U}$	$T_j \mid V_i, \mathbf{U}$	$V_i \mid T_i, \mathbf{U}$	$T_j \mid V_i, \mathbf{U}$		

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 a rejection of H_0 and 0 indicates a failure to reject

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

Table 3: - The set up for the adjacency matrix for each 3-node network denoted by the graph $G(T_i, T_j, T_k) \quad i, j, k \in \{1 : p\}; \forall i \neq j \neq k$ 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	$H_0 : T_i \perp\!\!\!\perp T_j \mid T_k, \mathbf{U}$	$T_i \perp\!\!\!\perp T_k \mid T_j, \mathbf{U}$
T_j	$H_0 : T_j \perp\!\!\!\perp T_i \mid T_k, \mathbf{U}$	0	$T_j \perp\!\!\!\perp T_k \mid T_i, \mathbf{U}$
T_k	$H_0 : T_k \perp\!\!\!\perp T_i \mid T_j, \mathbf{U}$	$H_0 : T_k \perp\!\!\!\perp T_j \mid T_i, \mathbf{U}$	0

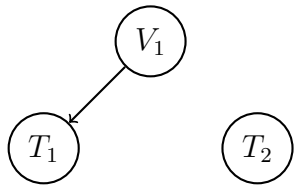

Figure 1: M0 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2$

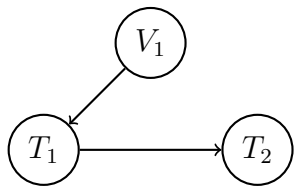

Figure 2: M1 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V_1 \perp\!\!\!\perp T_2 | T_1$

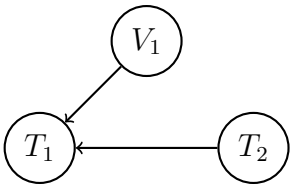

Figure 3: M2 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; V_1 \not\perp\!\!\!\perp T_2|T_1$

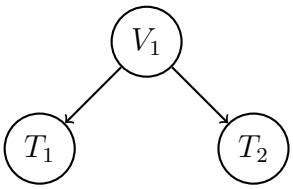

Figure 4: fig: M3 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \perp\!\!\!\perp T_2|V_1$

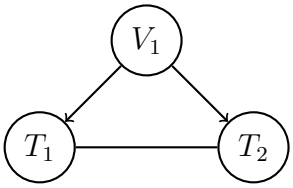

Figure 5: fig: M4 - $V_1 \not\perp\!\!\!\perp T_1; V_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2; T_1 \not\perp\!\!\!\perp T_2|V_1$

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:

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

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 i^{th} row and j^{th} column of $\mathbf{H}$:

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

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:

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

where null and alternative hypotheses are

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

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

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

Permutated 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 preform the regressions using the permuted variables:

$$T_i = \beta_0 + \beta_{1i} V_k + \beta_{2i}^* T_j^* + \mathbf{\Gamma U} + \epsilon \tag{5}$$

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

Let $\mathbf{\Theta}_{2i}$ and $\mathbf{\Theta}_{2j}$ 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:

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

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

We next test the conditional association between T_i and T_j using the nominal test defined by Yang et. al., 2017. Let $T_{\text{obs}_{2i}}$ 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 $\boldsymbol{\Theta}_{2i}$ and $\boldsymbol{\Theta}_{2j}$ respectively. Therefore the mediation test statistic is:

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

where we

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