Constructing deconvolved causal graphs from GWAS summary data

G Hemani et al

2017-02-28

```
## Warning: package 'knitr' was built under R version 3.3.2
## Warning: package 'tidyr' was built under R version 3.3.2
## Warning: package 'ggplot2' was built under R version 3.3.2
```

Code available at: https://scmv-ieugit.epi.bris.ac.uk/gh13047/graph_mr

Summary

Suppose we have complete summary statistics for an arbitrary number of traits, and each trait has valid instruments. Two-sample Mendelian randomisation can be used to calculate the causal relationships of $M \times M$ pairs of traits, thereby constructing a matrix of *total* effects of each trait on each of the other traits. This paper introduces a method to deconvolve this matrix to obtain the set of all direct causal effects between the traits without the use of individual level data. It goes on to show that the statistical power of identifying causal relationships between traits improves substantially by chaining together the direct effects on the deconvolved pathway, when compared to evaluating the total effect as is typical in Mendelian randomisation.

Introduction

Many methods exist that attempt to decompose correlation matrices into a set of terse direct correlations, with a view towards obtaining the minimum set of correlations that can explain the observed matrix. This process is known as deconvolution[1] [describe other methods in more detail]. A typical use case in biology is to calculate the correlation matrix of gene expression levels, and deconvolve to identify genes that drive networks. The application of matrix deconvolution to the context of causal inference in Mendelian randomisation (MR), as shown in Figre 1, could have the following attractive features:

- Identify direct pathways through which a particular exposure influences an outcome
- Identify instances of partial mediation, which might suggest that there are unknown variables that remain to be uncovered that mediate the path from exposure to outcome
- Potentially improve power in identifying causal relationships between traits that are on the same causal pathway

Growth in available GWAS summary data on phenotypes is on a steep trajectory, such that we may soon be asymptoting towards a situation where we can use two-sample MR to test 'everything against everything'. What this entails is that we could construct a pairwise causal relationship of all (available) traits (e.g. see Figure 1a,b). Each element in such a matrix would represent the *total* causal effect between the two traits. The purpose of this paper is to explore how to deconvolve this set of *total* effects into a terse set of *direct* effects.

Attempts at constructing networks using MR have been proposed[2], based on the idea of mediation by MR[3], though implementation of this method to more than three variables has not been demonstrated. While existing deconvolution methods can be applied to an arbitrary number of traits, they tend to require that the initial matrix is symmetrical, and they often do not incorporate information that disentangles correlation from causation. The setting for this paper assumes the contrary in both counts. The matrix of total MR effects is asymmetrical because the estimate of A on B (instrumented by SNPs that relate to A) is not the

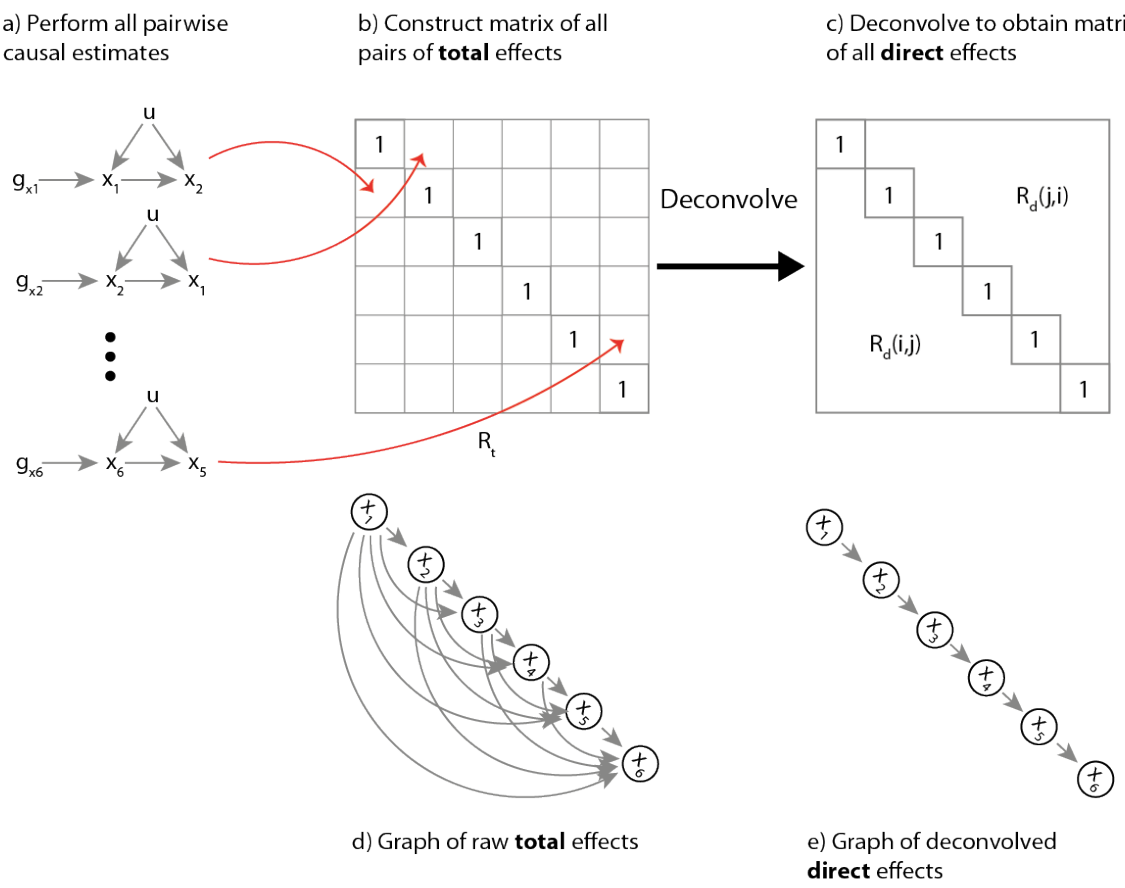

Figure 1: Deconvolving causal estimates from raw MR estimates. a) For each trait that can be instrumented, perform MR of its effect against every other trait. e.g. If there are six traits then there are 6^2 - 6 = 30 MR estimates to generate. b) Use these results to construct the matrix of total effects. c) Use a deconvolution method to obtain a set of direct causal effects. d and e) The true causal model is shown in (e), but if the raw causal effects are estimated using MR (i.e. the *total* effects) then a number of extra paths will be identified owing to transitive effects of the direct paths.

same as the estimate of B on A (instrumented by SNPs that relate to B). Further, each element in the matrix of total effects is assumed to be a causal estimate.

Methods

Assumptions

1. It is important to acknowledge that, though beyond the scope of this paper, the premise of obtaining the causal effect of 'everything on everything' requires methodological advancements to address many issues, including but not limited to multiple testing, consideration of phenotypic definitions, and incorporating temporal effects and critical effect periods.
2. We begin by assuming that the causal effect estimates of each element in the *total* effects matrix is unbiased (Figure 1a,b). In practice it is not yet clear how to ensure this, though methods are continually developing to improve the reliability of MR estimates to violations of its main assumptions[4–7].

Direct and total effects

The latter can be summarised as follows. Suppose there are six variables of interest, 1-6, and the causal relationships are

```
1 -> 2
2 -> 3
3 -> 4
4 -> 5
5 -> 6
```

This can be depicted in graph form as in Figure 1e. If, however, we performed MR of `1 -> 3`, `1 -> 4`, etc, we would identify associations because they exist indirectly. Hence, after testing everything against everything our graph would look like Figure 1d.

MR for mediation (AKA network MR) operates in the case where there are three phenotypic variables[8] as follows. The direct effect of trait 1 on trait 2, $\beta_{1\Rightarrow 2}$, is obtained from:

$$\beta_{1\Rightarrow 2} = \beta_{1\rightarrow 2} - \beta_{1\rightarrow 3}\beta_{3\rightarrow 2}$$

With four variables it looks like:

$$\begin{aligned}\beta_{1\Rightarrow 2} = \beta_{1\rightarrow 2} &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 4}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 2} \\ &- \beta_{1\rightarrow 4}\beta_{4\rightarrow 2}\end{aligned}$$

With five variables it looks like:

$$\begin{aligned}\beta_{1\Rightarrow 2} = \beta_{1\rightarrow 2} &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 4}\beta_{4\rightarrow 5}\beta_{5\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 4}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 5}\beta_{5\rightarrow 2} \\ &- \beta_{1\rightarrow 4}\beta_{4\rightarrow 5}\beta_{5\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 5}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 3}\beta_{3\rightarrow 2} \\ &- \beta_{1\rightarrow 4}\beta_{4\rightarrow 2} \\ &- \beta_{1\rightarrow 5}\beta_{5\rightarrow 2}\end{aligned}$$

and this is performed for each of the 5×5 possible pairwise combinations of variables, ultimately reducing a matrix of total effect relationships, R_t e.g.

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]   -1    1    0    0    0
## [3,]    2   -2    1    0    0
## [4,]    2   -2    1    1    0
## [5,]    6   -6    3    3    1
```

into a matrix of direct relationships, R_d e.g.

```
##      [,1] [,2] [,3] [,4] [,5]
## [1,]    1    0    0    0    0
## [2,]   -1    1    0    0    0
## [3,]    0   -2    1    0    0
## [4,]    0    0    1    1    0
```

```
## [5,]    0    0    0    3    1
```

There are two potential drawbacks with the MR for mediation approach. First, the combinatorial increase in the number of terms that are required for calculating the direct effects gets large very quickly. For example, for a graph with 10 variables there are 52 unique paths for each of the 100 elements in the matrix, and identifying those paths itself is a computationally slow process. Second, perhaps more importantly, this method may not actually generalise beyond the three-trait case.

Simulations

Assume that there are M variables measured in N samples represented in a $N \times M$ matrix P. Further, each variable has a valid instrument, hence there are M instruments also, represented in a $N \times M$ matrix G. As stated above, in this analysis I am assuming that every causal estimate made by MR is reliable.

DAGs are simulated such that the M variables are related to each other by random causal effects. Cycles are avoided. Following on, two-stage least squares is used to calculate all pairwise causal relationships e.g.

$$R_t(1 \rightarrow 2) = \frac{cov(P_{,2}, G_{,1} cov(P_{,1}, G_{,1}/var(G_{,1})))}{var(G_{,1} cov(P_{,1}, G_{,1}/var(G_{,1})))}$$

Three methods are then used to try to deconvolve the graph R_t into R_d.

Method 1 - mediation by MR

This is as described earlier, adapting from [Relton2012] and [Burgess2015a].

Method 2 - inversion

Simply a method to orthogonalise the matrix by

$$R_d = R_t^{-1}$$

If R_t were a variance covariance matrix then this method is known as obtaining the precision matrix.

Method 3 - Feizi deconvolution

In[1] a method is outlined for network deconvolution that is primarily aimed at correlation matrices (i.e. symmetric, non-causal versions of R_t). The method is:

$$R_d = R_t(I + R_t)^{-1}$$

Standard errors

Bootstrapping is used to obtain the standard errors of the direct effects for the inversion method. Each element in R_t is resampled with $R_t(i,j)* \sim N(R_t(i,j), se(R_t(i,j)))$. The inversion method is then applied to the resampled matrix R_t* and the results R_d* are stored. This is performed 1000 times to obtain a distribution of effects for each element of the R_d matrix. The standard deviation of the distribution from each element is taken to be the standard error of that direct effect estimate.

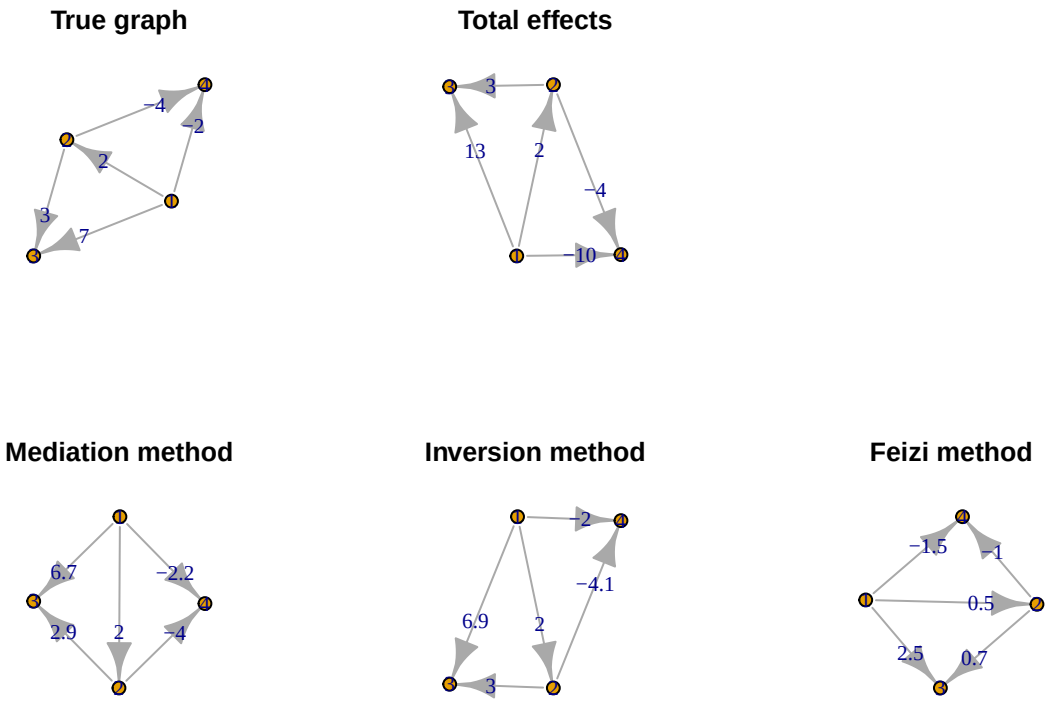

Figure 2: Graphs depicting the simulated (true) causal graph, and the results from the different methods of estimation

Results

The analysis is divided into two sections. First, a demonstration that in most cases the inversion method is rapid and returns the true direct effects; and second, an evaluation of statistical power, particularly in comparing the ability to detect a causal relationships via the total effect vs constructing a chain of intermediary effects.

Comparison of deconvolution methods

Here will be presented analysis of four different causal networks. In all cases, 500000 samples are simulated to have the requisite phenotypes, and each phenotype is simulated to have a valid instrument. These data are used to construct the matrix of *total* causal effects, as shown in Figure 1a,b. The network is then deconvolved each of the three methods into R_d, and the estimates of the direct causal estimates are compared to the true direct estimates that were simulated initially.

Illustrative example

In each of the following cases large samples (n=500000) are simulated and causal effect sizes are large, to ensure that power is not influencing the fidelity of matrix deconvolution by the different methods. An illustrative example with the causal structure between four traits is shown in Figure 2. In this simple case, all methods resolve the same graph, however only the mediation and inversion methods obtain good agreement between the deconvolved direct effect estimates and the simulated direct effects (Figure 3).

The Feizi method appears to generate estimates proportional to the raw total effects, but on a scale that is shrunk towards 0.

Further simulations

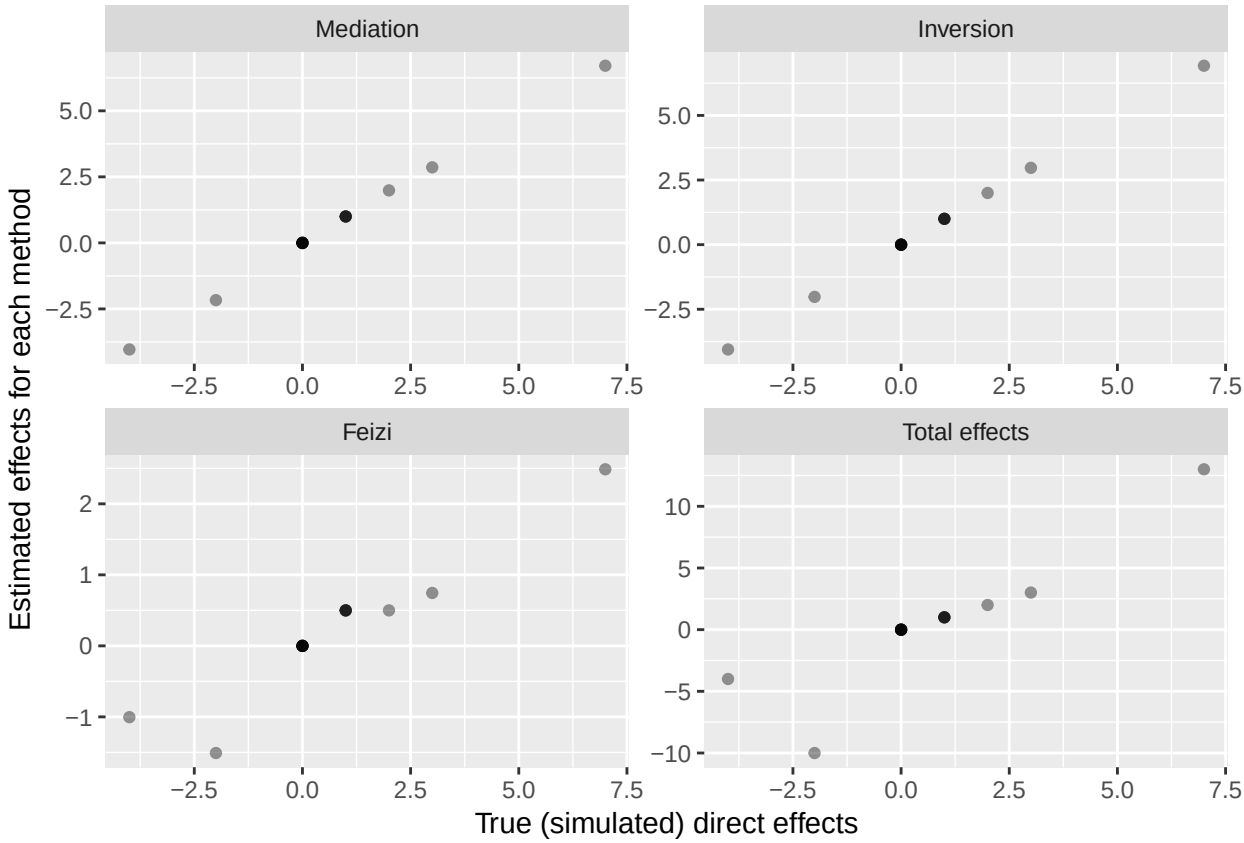

Figure 3: Comparison of deconvolved causal effect estimates against the simulated causal effects, using the elements of the Rd and Rt matrices.

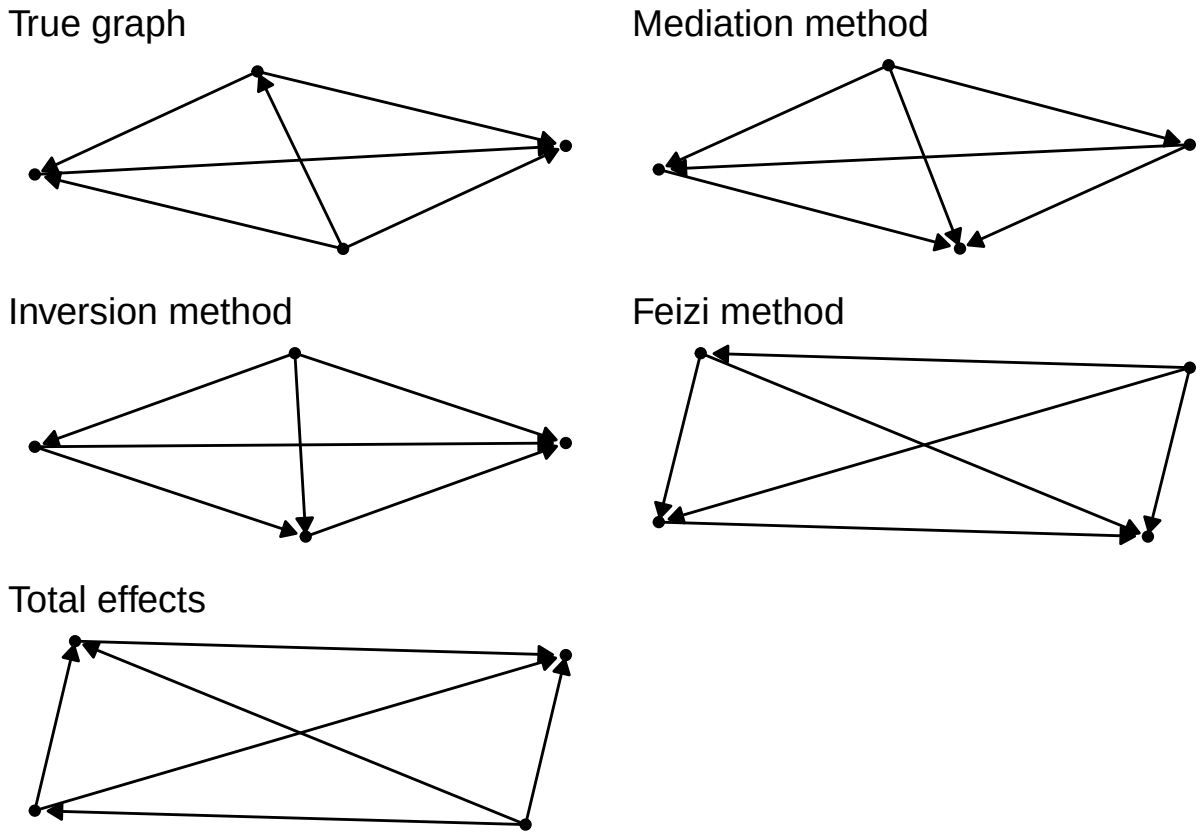

Figure 4: Graphs for '4-variables, comples' model

Further simulations of more complex graphs are now performed. Beyond 7 variables the mediation method is computationally intractible using the algorithm as implemented here, but regardless it can be seen that any models more complex than those shown in Figure 2 lead to the mediation method exhibiting departures from the simulated effects.

The models tested are shown in Figures 4-8. There is a mixture of different numbers of variables, different model complexities, and simulating effect sizes from non-guassian distributions. A comparison of the deconvolved effect estimates from each of these models against the true simulated effects is shown in Figure 9.

The inversion method performs reliably in most cases, however occasionally in more complex models there will appear to be some direct effects that have non-zero estimates where they should have had no effect. This may be an indication of saturation.

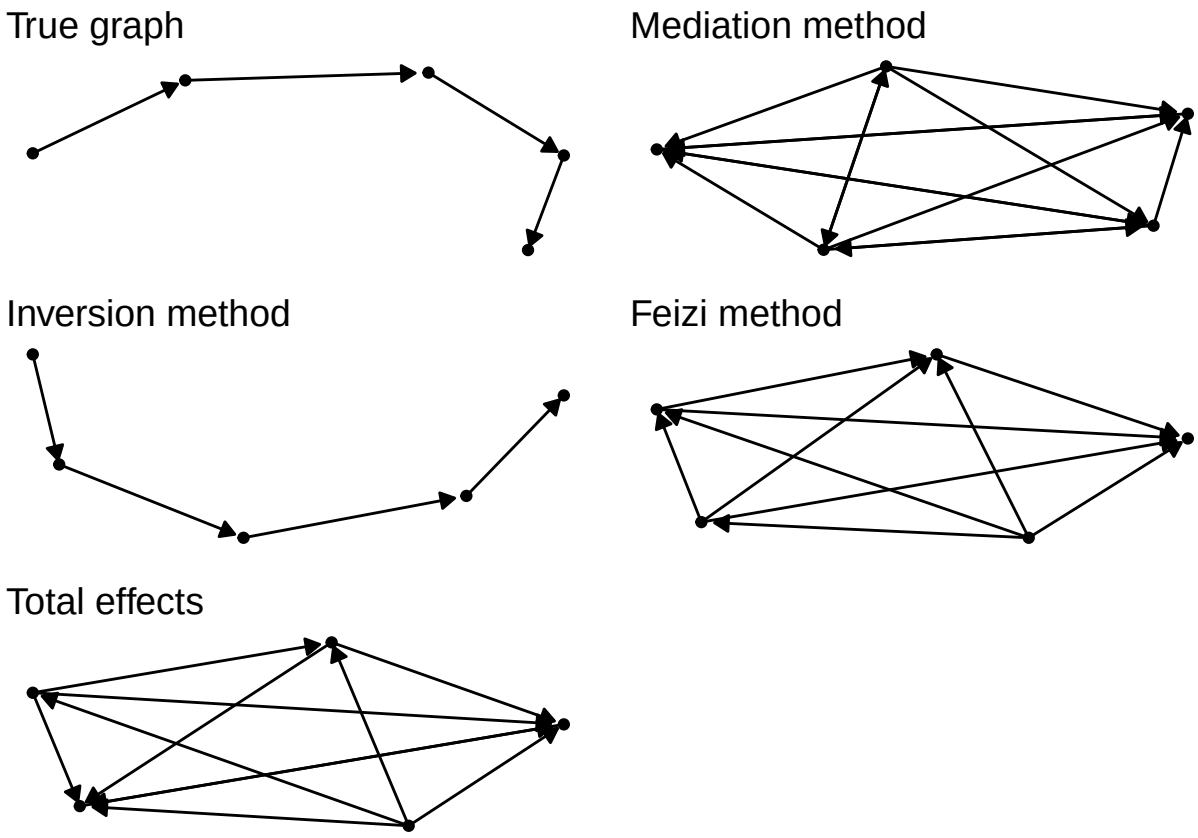

Figure 5: Graphs for '5-variables, causal chain' model

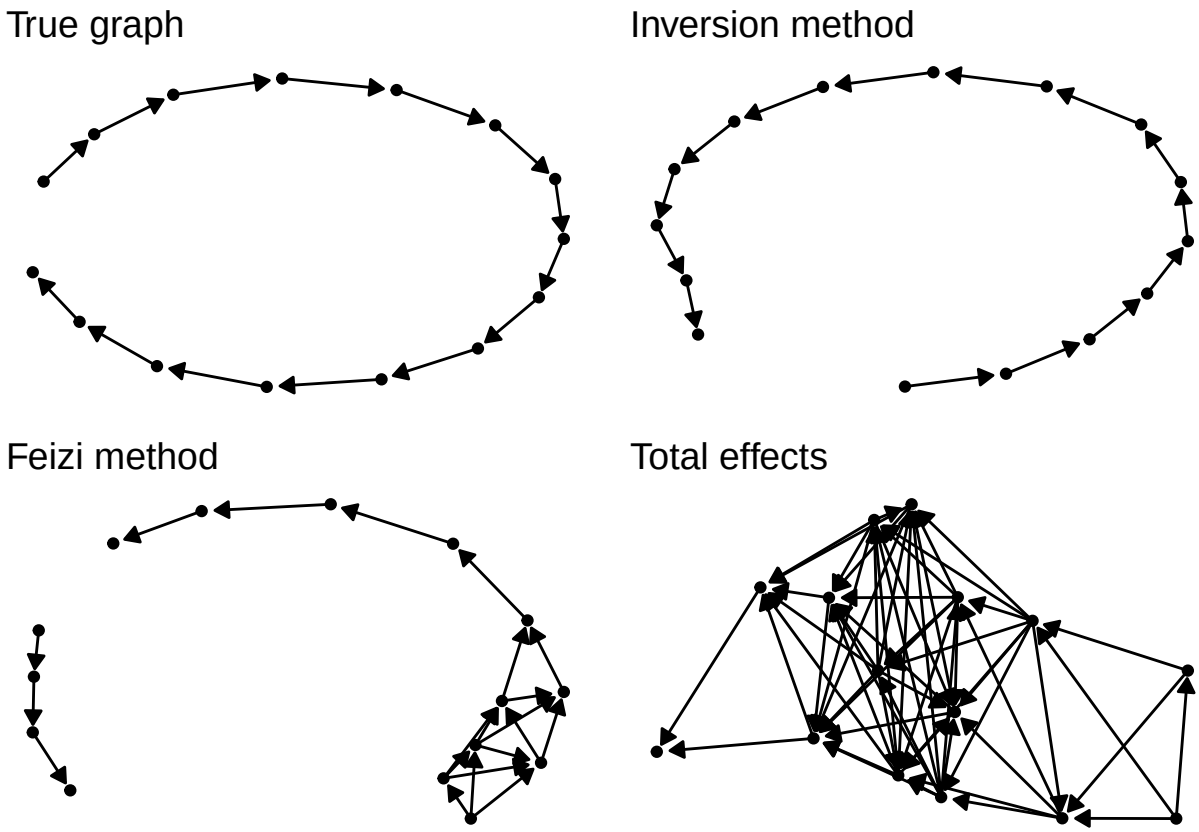

Figure 6: Graphs for '15-variables, causal chain' model

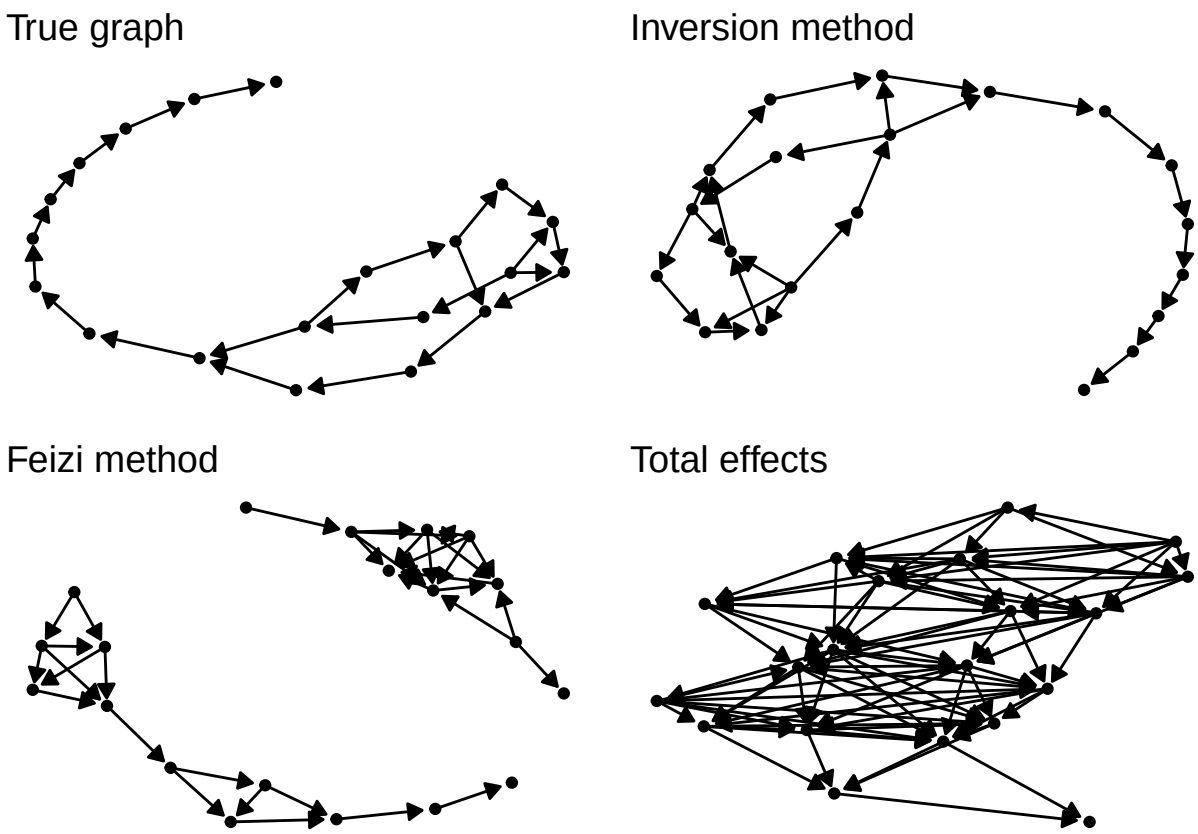

Figure 7: Graphs of '20-variables, complex' model

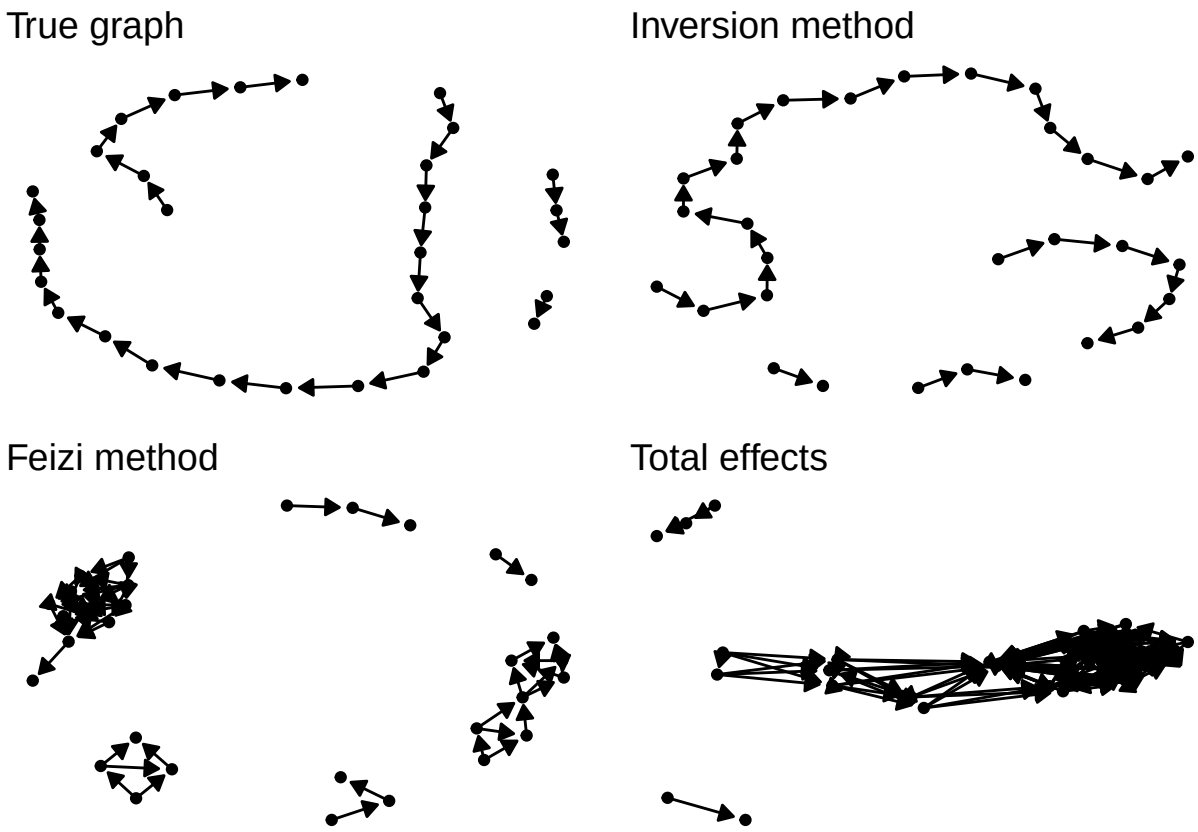

Figure 8: Graphs of '30-variables, non-gaussian' model

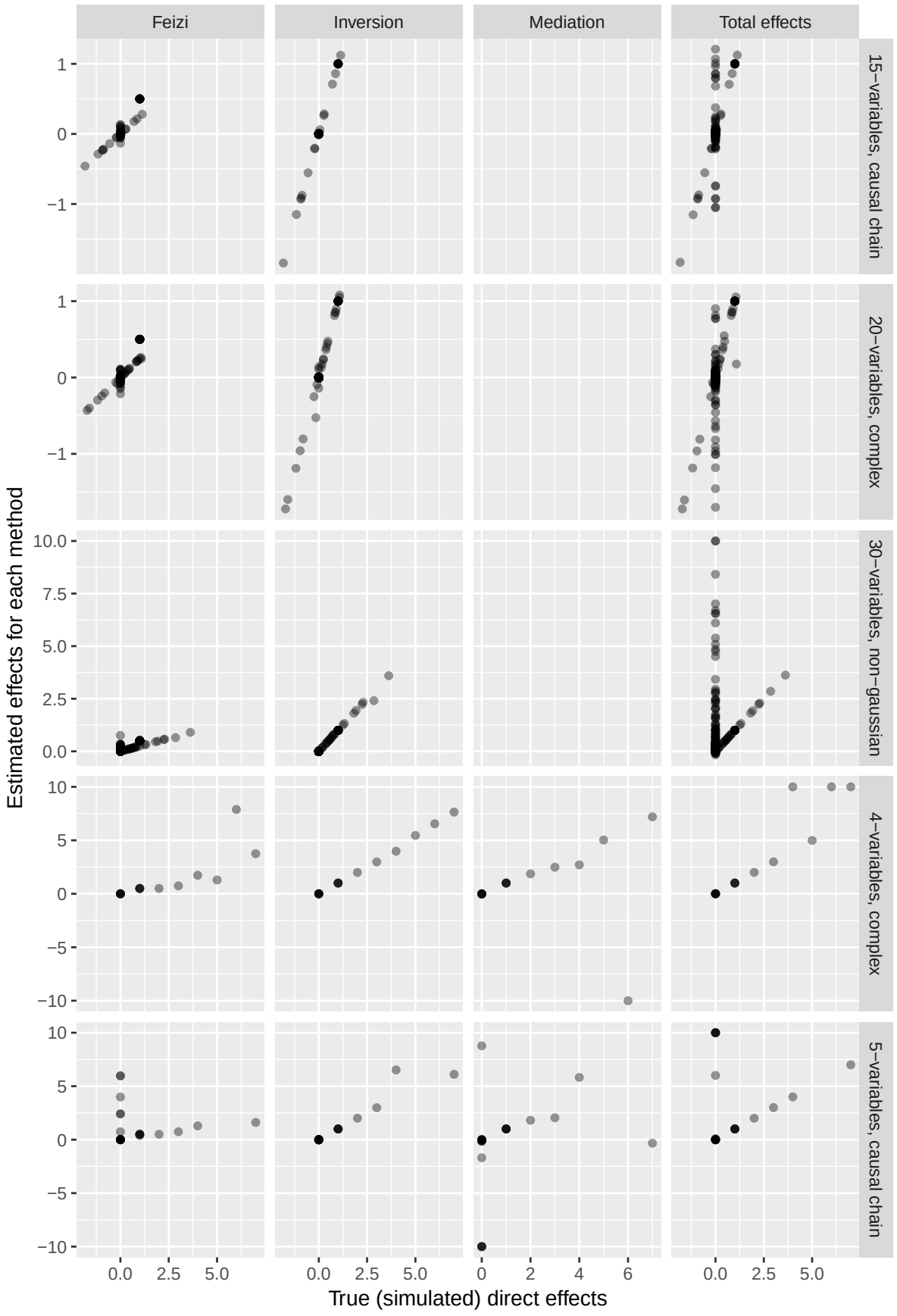

Figure 9: Comparison of deconvolved direct effect estimates from different methods against the true simulated direct effects.

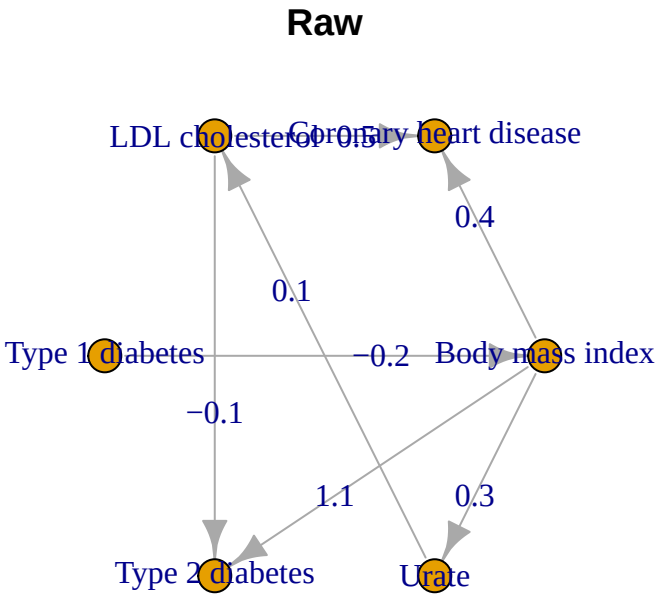

Figure 10: Empirical results, raw graph

Empirical example

As an illustration, the inversion method was performed on the following variables using the data in MR-Base:

- Body mass index
- Coronary heart disease
- Type 1 diabetes
- Type 2 diabetes
- Serum urate
- LDL cholesterol

Instruments were extracted for each trait (297 in total), and each of those instruments was extracted from each trait, thus enabling a 2-sample MR analysis of each trait against each of the other traits.

In order to automatically choose the most reliable MR estimate for each of the 6×6 analyses, the meta-analytic framework described by[9] and adapted to MR by[10] was used. Briefly, this entails:

- No heterogeneity, no directional pleiotropy - **IVW (fixed effects)**
- Heterogeneity, no directional pleiotropy - **IVW (multiplicative random effects)**
- No heterogeneity, directional pleiotropy - **MR Egger (fixed effects)**
- Heterogeneity, directional pleiotropy - **MR Egger (multiplicative random effects)**

This was applied without close scrutiny for illustrative purposes, but further refinement to this procedure is possible.

Showing associations with $p < 0.05$, the graph in Figure 10 shows the nominally significant edges for the raw associations. The deconvolved graph is shown in Figure 11.

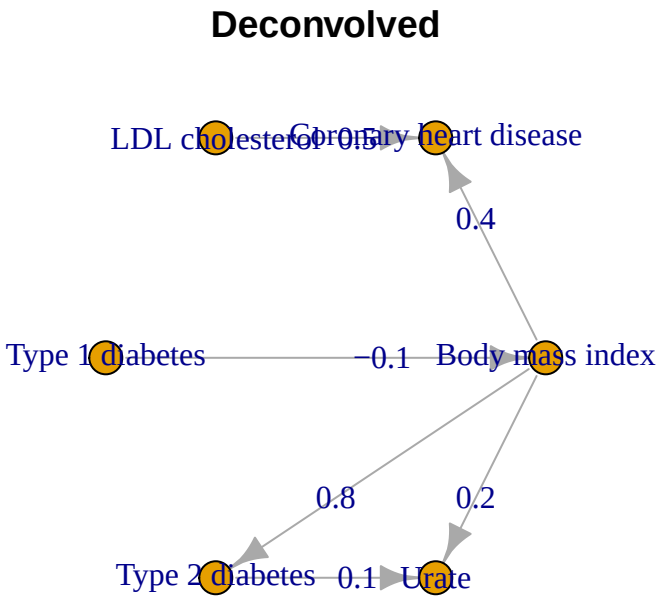

Figure 11: Empirical results, deconvolved graph

Statistical power

Basic simulations

As an illustration, in the following analyses a 6-trait network is created as in figure 1. Is it statistically more efficient to

1. estimate the total effect of trait 1 on trait 6 using standard MR, or
2. obtain the direct effects of the 6-trait network, and estimate the causal chain from trait `1 -> 2 -> 3 -> 4 -> 5 -> 6`?

Strategy (2) entails finding a 'significant' p-value at each of the 5 direct relationships that map trait 1 to trait 6, whereas strategy (1) simply requires that the total effect of trait 1 on trait 6 is 'significant'. For the purposes of evaluating power we are determining 'significance' at a $p < 0.01$ threshold, but such practice is not considered reliable when applied to real data[11].

A significant path is found for a hypothesised relationsip of 1 causing 6 as follows. A directed path is searched for that links the two nodes along edges that all have $p < 0.01$. If any edge doesn't satisfy this threshold then another path needs to be found. This uses a graph traversal algorithm (breadth first search of Dijkstra's algorithm). This means that some paths from 1 to 6 might be found *through the wrong pathway*. But the extent to which this is happening is evaluated in the false discovery rate.

The false discovery rate is estimated as follows. For a graph where a path of 1 causing 6 is simulated, paths are searched for that link **6 causing 5**, i.e. the reverse, for which there should be no paths. This is attempting to be conservative because a) it allows for strong edges to exist in the null model, b) the target node is known to be influenced by at least one other node already.

The simulation was performed using 5000 samples, and the simulated direct effect sizes for each chain were sampled from $\beta \sim N(0, 1)$. The results are shown in Figure 12.

Here it is apparent that as the causal chain grows larger the statistical power of the second strategy, which uses the direct effects estimated from the graph, becomes substantially more powerful than the standard approach of identifying the total causal effect. The false discovery rate of the second strategy appears grow as the graph size grows also though.

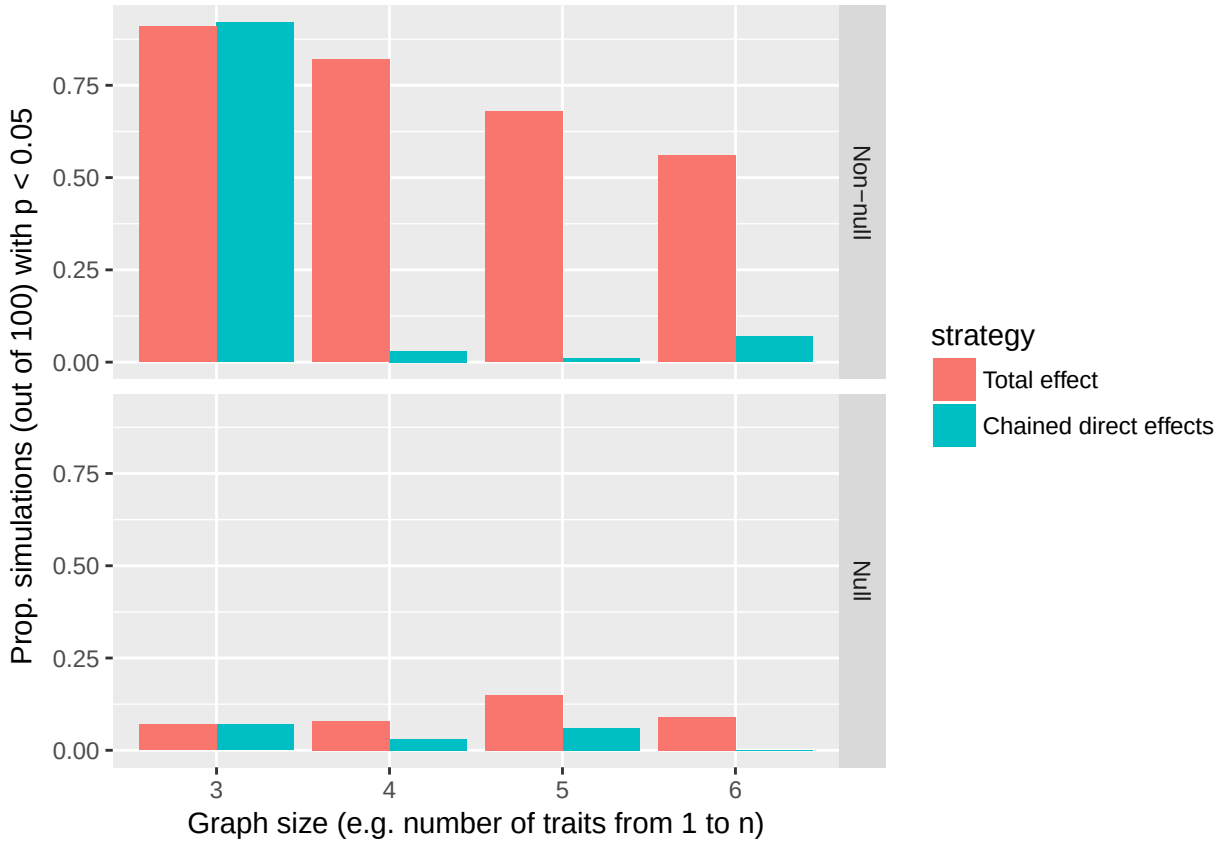

Figure 12: Statistical power comparison between standard MR and graph MR using inversion method. Top graph shows the results from simulations where causal influences were simulated with non-zero effect sizes. Bottom graph shows depicts the false discovery rate, i.e. the same simulation but where all effect sizes are 0.

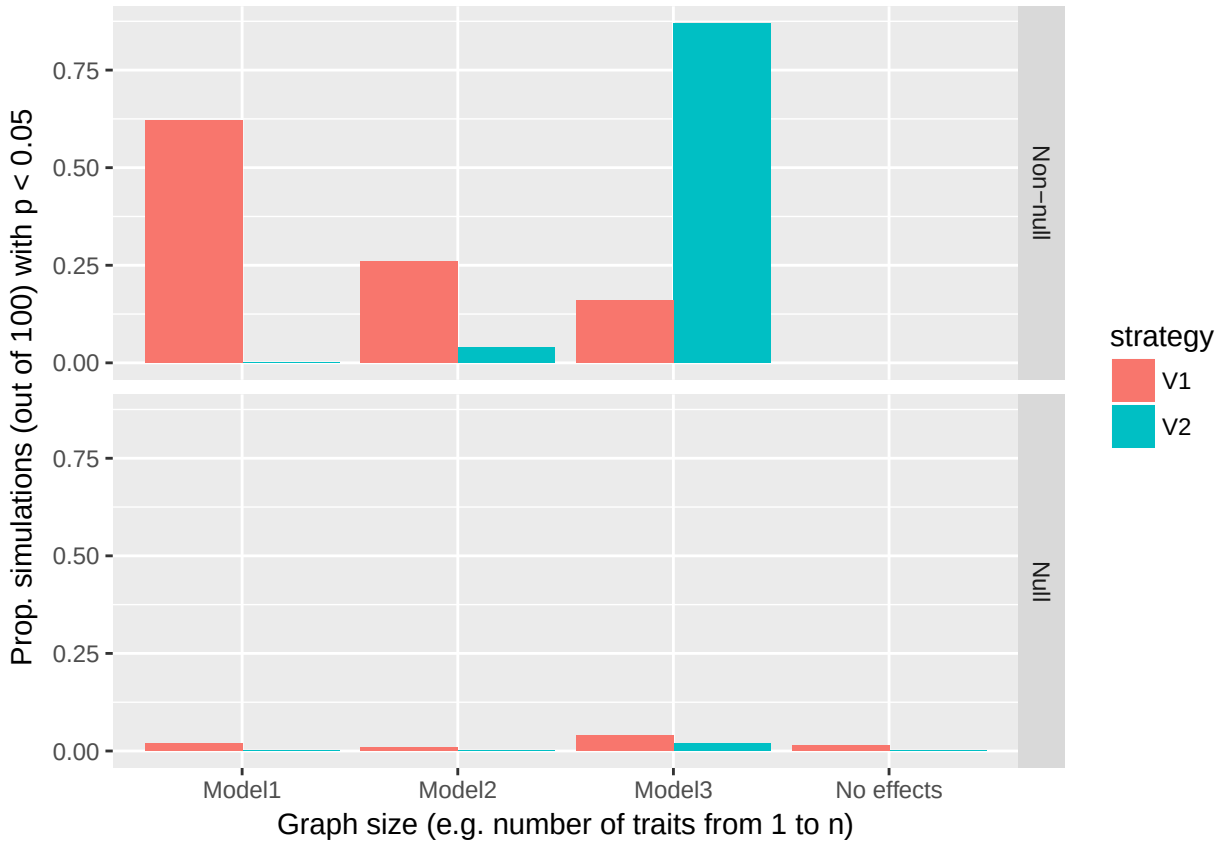

Figure 13: Statistical power comparison between standard MR and graph MR using inversion method. Top graph shows the results from simulations where causal influences were simulated with non-zero effect sizes. Bottom graph shows depicts the false discovery rate, i.e. the same simulation but where all effect sizes are 0.

Exploring causal graphs with multiple pathways

There are three causal models being evaluated.

- Model 1: One path from 1 to 6
  - `1 -> 2 -> 3 -> 4 -> 5 -> 6`
- Model 2: Two paths from 1 to 6
  - `1 -> 2 -> 3 -> 6`
  - `1 -> 4 -> 5 -> 6`
- Model 3: Three paths from 1 to 6
  - `1 -> 2 -> 6`
  - `1 -> 3 -> 6`
  - `1 -> 4 -> 5 -> 6`
- Null model: No causal effects simulated

The simulation was performed using 5000 samples, and the simulated direct effect sizes for each chain were sampled from $\beta \sim N(0, 1)$.

To generate a basic FDR, a different dataset was generated with no simulated causal paths. A graph traversal algorithm is applied to find a path of significant links to make a chain from 1 to 6. This is not constrained to go through 2,3,4,5, it can go through any path to get from 1 to 6. Results from these extended simulations are shown in Figure 13.

False discovery rates

The number of possible k length paths between two nodes in a graph of size n is given as

$$p_n(k) = \frac{(n-2)!}{(n-k)!}$$

and the probability of a false discovery for a path of length k is α^k where α here is 0.05. Hence, the probability of finding a path between two nodes is

$$FDR_n = \alpha + \sum_{k=2}^{n} \alpha^k \frac{(n-2)!}{(n-k)!}$$

this can be simplified to be computationally tractable for large graphs. The FDR for a particular path length K is

$$\begin{aligned} FDR_n(K) &= \alpha^K \prod_{k=2}^{K} n-k-1 \\ &= \alpha^2 \prod_{k=2}^{K} \alpha(n-k-1) \end{aligned}$$

Hence, the total FDR across all path lengths is

$$FDR_n = \alpha + \alpha^2 + \sum_{K=3}^{N} \left(\alpha^2 \prod_{k=2}^{K} \alpha(n-k-1) \right)$$

The effect of graph size on the FDR is shown in the top plot of Figure 14. This is what happens if any path size is allowed to be searched for. If the maximum path size is limited to 6 (bottom graph) then the saturation is lower.

What does this saturation look like e.g. for 50 trait graph with $\alpha = 0.05$? This is shown in Figure 15.

An alternative way to check significance of a path is to permute. Here the off-diagonal elements of the matrix of p-values are permuted. What this entails is that the same number of strong edges are retained, and then the question is 'with this number of strong edges amongst all nodes, how likely is it to find a path from A to B?' Results are shown in Figure 16.

This looks to be a much more conservative way to test the significance of the link between two nodes.

Cycles (non-DAGs)

This looks problematic

Discussion

The inversion method appears to be reasonably accurate in a range of scenarios, though not perfect when graphs become larger and relationships more complex. Departures from the true direct effect model are typically small, but exploration of what leads to this are warranted.

Issues:

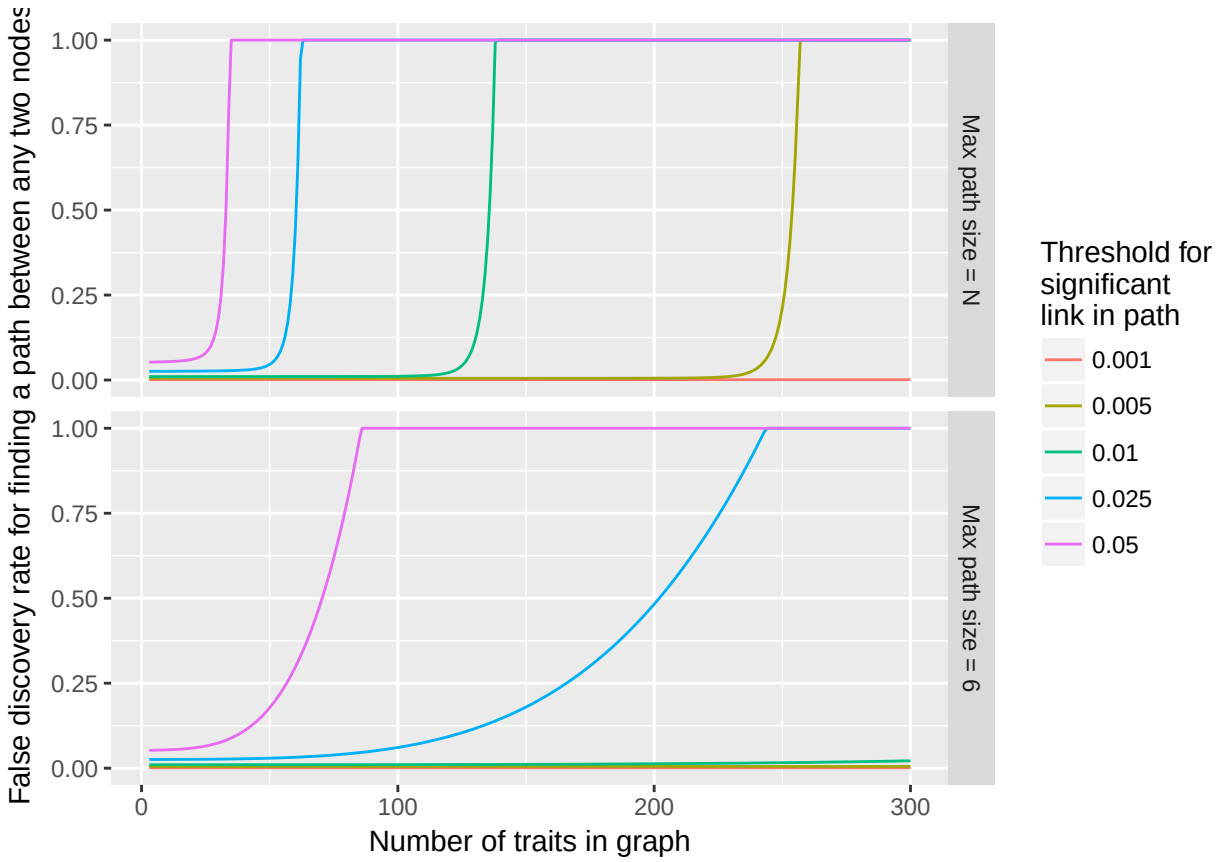

Figure 14: Relationship between graph size, alpha value and FDR

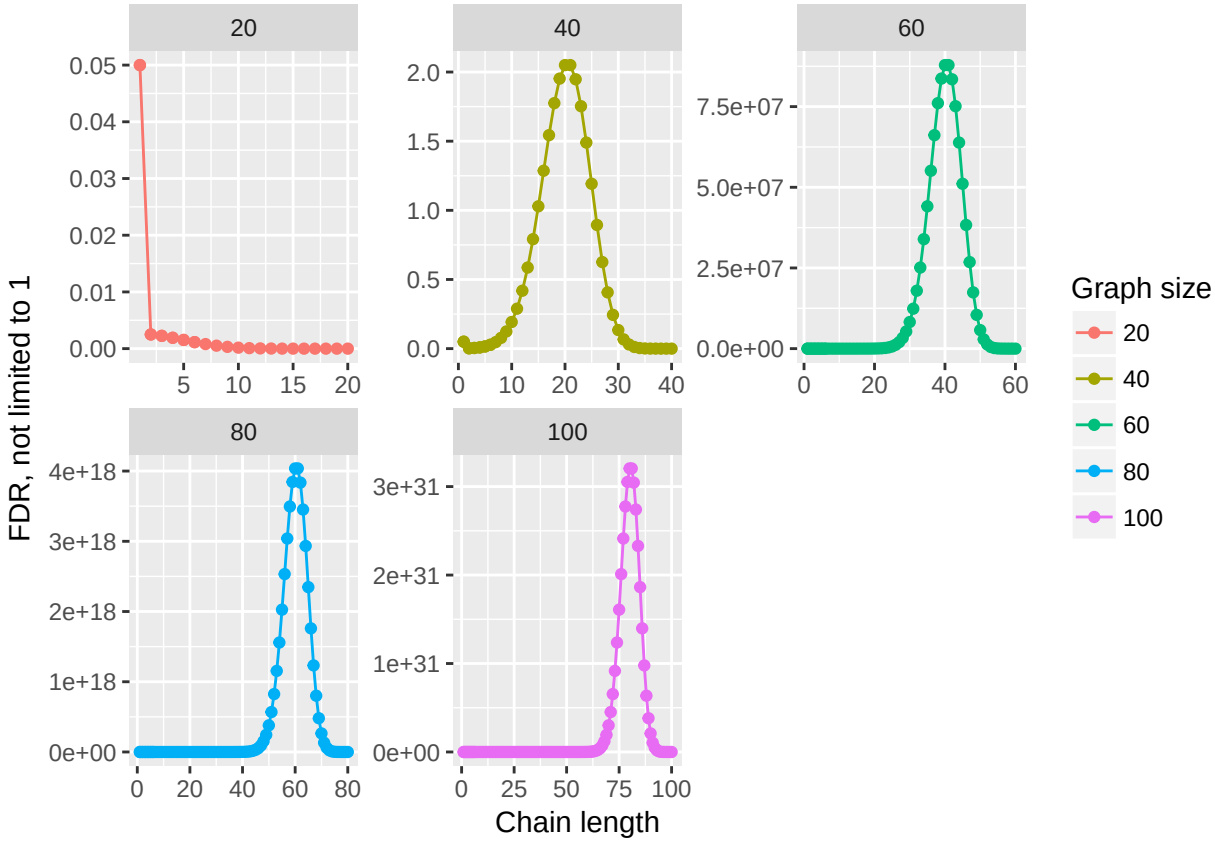

Figure 15: Identifying which path lengths contribute most to elevating false discovery rates

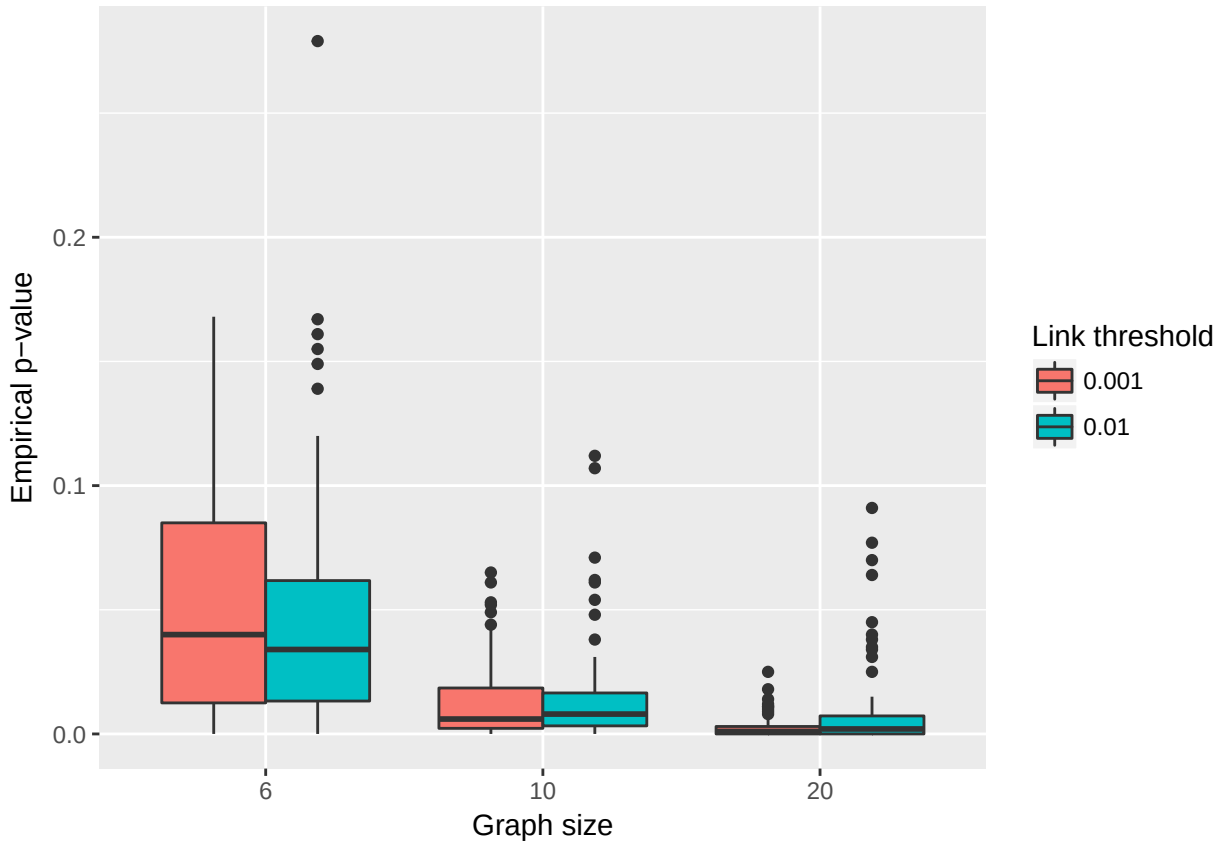

Figure 16: P-value distributions from permutation analysis

- The mediation method might be interpreted too simplisticly here. I think there needs to be recursion beyond the 3 variable example - i.e. the direct effects are a function of indirect effects at the moment, but these indirect effects probably need to be reduced to direct effects also. This would require identifying a path through which to traverse the graph and recursively estimate the direct effects
- These simulations only looked at relatively sparse graphs
- Need to evaluate much larger graphs, e.g. thousands of traits
- Why isn't the Feizi method working? Need to make sure it works for symmetric correlation graphs
- Haven't evaluated the influence of cycles in the graph - this is hard to simulate, though if a temporal component were introduced then there will be no cycles.
- Graphical lasso may be useful to make the matrix sparse

References

1. Feizi, S., Marbach, D., Médard, M. & Kellis, M. Network deconvolution as a general method to distinguish direct dependencies in networks. *Nature Biotechnology* **31,** 726–733 (2013).

2. Burgess, S., Daniel, R. M., Butterworth, A. S. & Thompson, S. G. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International journal of epidemiology* **44,** 484–95 (2015).

3. Relton, C. L. & Davey Smith, G. Two-step epigenetic Mendelian randomization: a strategy for establishing the causal role of epigenetic processes in pathways to disease. *International journal of epidemiology* **41,** 161–76 (2012).

4. Bowden, J., Davey Smith, G. & Burgess, S. Mendelian randomization with invalid instruments: effect estimation and bias detection through Egger regression. *International Journal of Epidemiology* **44,** 512–25 (2015).

5. Bowden, J. *et al.* Assessing the suitability of summary data for two-sample Mendelian randomization analyses using MR-Egger regression: the role of the I2 statistic. *International Journal of Epidemiology* dyw220 (2016). doi:10.1093/ije/dyw220

6. Bowden, J., Davey Smith, G., Haycock, P. C. & Burgess, S. Consistent Estimation in Mendelian Randomization with Some Invalid Instruments Using a Weighted Median Estimator. *Genetic Epidemiology* **40,** 304–314 (2016).

7. Hemani, G. *et al.* MR-Base: a platform for systematic causal inference across the phenome using billions of genetic associations. *BioRxiv* **10.1101/07,** (2016).

8. Burgess, S., Daniel, R. M., Butterworth, A. S. & Thompson, S. G. Network Mendelian randomization: using genetic variants as instrumental variables to investigate mediation in causal pathways. *International journal of epidemiology* **44,** 484–95 (2015).

9. Rucker, G., Schwarzer, G., Carpenter, J. R., Binder, H. & Schumacher, M. Treatment-effect estimates adjusted for small-study effects via a limit meta-analysis. *Biostatistics* **12,** 122–142 (2011).

10. Bowden, J. *et al.* A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization. *Statistics in Medicine* (2017). doi:10.1002/sim.7221

11. Sterne, J. A. C. & Smith, G. D. Sifting the evidence—what's wrong with significance tests? *BMJ* **322,** 226–231 (2001).