Investigating the combined application of Mendelian Randomization and constraint-based causal discovery methods

Mihály Vetró, Márton Bendegúz Bankó, Gábor Hullám
Budapest University of Technology and Economics
Department of Measurement and Information Systems
Budapest, Hungary
Email: {vetro, bankom}@edu.bme.hu, gabor.hullam@mit.bme.hu

***Abstract*—Mendelian randomization (MR) is often used in medical studies and biostatistics, to reveal direct causation effects between exposures and diseases, typically the effect of some exposure (like chemicals, habits and other factors) to a known disease or disorder. However, this procedure has some strict prerequisites, which often do not comply with the known variables, or the exact causal structure of the variables is not known in advance. In this study, we investigate the use of constraint-based causal discovery algorithms (PC, FCI and RFCI) to produce a sufficient causal structure from the known observations, to aid us in finding variable triplets, upon which MR can be performed. In addition, we show that the validity of MR cannot always be determined based on its results alone. Finally, we investigate the application of the MR principle to determine the direction of causality between variable-pairs, which is a problem most constraint-based causal discovery methods struggle with.**

***Index Terms*—Mendelian Randomization, Bayesian networks, constraint-based causal discovery, causal effect strength, biostatistics**

I. Introduction

In this study, we investigate three constraint-based causal discovery methods, and Mendelian Randomization (MR), which is a well-known method for causal effect estimation in biostatistics and medical studies, and then we examine two different approaches to combine them. A similar study including genetic anchors has been performed by Howey et al. [1], in which they investigated some simple causal structures regarding MR. Here, we take a more general approach with regards to the size and number of the investigated causal graphs, and we also examine the usability of the MR principle to help determine the direction of causality in uncertain cases.

II. Mendelian randomization

MR can be classified as a local causal discovery method applied in the field of genetic studies. However, while observational data based general causal discovery methods learn the causal structure from data with no prior assumptions regarding the structure, MR methods are based on a predefined causal structure relying on a set of assumptions which need to be fulfilled [2]. The MR model (causal structure) is a causal chain formed by a triplet of variables ($G \rightarrow E \rightarrow D$) which consists of the following elements:

- Genetic variant (G): the gene whose effect is being studied.
- Exposure factor (E): an event, occurrence or influencing factor to which susceptibility is influenced by a genetic variant, and that factor has an effect on the disease.
- Disease (D): the diagnosis itself, which may be influenced by the previous factors.
- Confounding factor (U): additional variable that is not part of the chain but may affect the exposure and disease variables.

MR methods use the effect size (e.g. log odds ratio in case of categorical variables) between the gene - exposure (β_{GE}) and the gene - disease (β_{GD}) variables to infer the magnitude of the effect between the exposure and the disease (β_{ED}) as: $\beta_{ED} = \frac{\beta_{GD}}{\beta_{GE}}$, which can be treated as a Wald ratio and its significance can be determined accordingly [3]. A significant ratio can be considered as an indication that the causal effect between the exposure and the disease is significant and that the causal relationship $E \rightarrow D$ exist.

The assumed MR structure (displayed in Fig. 1) encodes the following assumptions:

- A1: The association between the genetic variant and the exposure factor should be strong. In its absence, the strength of the MR is reduced and bias may occur.
- A2: The genetic variant is independent of the confounding factor. Otherwise, the confounding factor would affect both the disease and the genetic variant, which may imply that the gene-disease effect detected by the method is only indicative of the difference due to the confounding factor.
- A3: The disease is conditionally independent of the genetic variant given the exposure factor. It follows that the gene does not directly influence the presence of the disease, instead it only has a mediated effect.

The main question of MR methods is how to ensure this specific structure shown in Fig.1. In general, it requires

This research was supported by the ÚNKP-21-5-BME-362 New National Excellence Program of the Ministry for Innovation and Technology from the source of the National Research, Development and Innovation Fund, and the János Bolyai Research Scholarship.

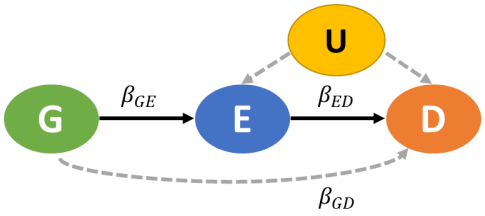

Fig. 1. The assumed MR model.

considerable background knowledge to exclude variables from the dataset that do not satisfy the validity assumptions. The main weakness of the MR method is its simplistic, rigid model. Although it can be efficient when the investigated relationships are simple, i.e. there are no confounding or interacting factors, in more complex cases however, the assumptions of the MR model are unrealistic. This either leads to the inability to use MR methods in several real-word scenarios or to an inappropriate application of MR disregarding some of the validity assumptions. To address this issue, the basic model has been extended in several ways to make the method more robust: MR Egger [4], MR-link [5]. However, the main feature of causal structure learning algorithms, i.e. the structure is learned from the data, is still missing from MR methods.

III. Causal structure learning methods

We selected three widely-known constraint-based methods to investigate their integrated application with a MR method: the Peter-Clark (or PC) algorithm, Fast Causal Inference (FCI) and *Really* Fast Causal Inference (RFCI).

A. The Peter-Clark algorithm

The Peter-Clark (PC) algorithm is a local method [7], relying on examining variable pairs to determine if they are (conditionally) dependent, and variable triplets - or more precisely, chain structures containing exactly 3 variables - to determine the direction of causality between the dependent variables. The former step leads to a skeleton, i.e. an undirected graph of dependency relationships which can be facilitated by applying conditional independence tests on variables. The latter step requires the detection of triplet-based uniquely identifiable dependency structures, called V-structures [6] ($X \rightarrow Z \leftarrow Y$), whose edges can be unambigously directed. The second step leaves those edges undirected that are not part of a V-structure, which may include a significant number of edges. Additional steps using various heuristics may be applied to orient these undirected edges.

B. Fast Causal Inference

While the PC algorithm is built to reconstruct the full causal graph of the variables, the Fast Causal Inference (FCI) [8] and its more efficient version, the *Really* Fast Causal Inference (RFCI) algorithm [9] both aim to reconstruct the equivalence class of the original causal structure, represented by its essential graph, which is a partially directed acyclic graph (PDAG).

IV. Approach No. 1: Augmenting MR with causal structure learning

In our first approach, we investigated the usefulness of the causal discovery methods described in section III. to determine if MR is applicable for a given Gene-Exposure-Disease variable triplet. In our methods, if the three variables form a directed chain in the same order ($G \rightarrow E \rightarrow D$), then it is considered as a valid candidate for MR, and considered invalid otherwise.

First, we examined some simple causal structures, shown in Fig. 2. For these models, the PC, FCI and RFCI algorithms were all capable to reliably reconstruct the original causal graph from at least 1000 samples. This is the expected result, because almost all of the edges are part of at least one V-structure, apart from the edge ($3 \rightarrow 4$) in Model 1 and the edges ($3 \rightarrow 4$) and ($3 \rightarrow 5$) in Model 3. In our tests, all the variables were binary, which represents a discrete variable case of MR. Note that it is also possible to apply MR for continuous disease score and exposure variables. The conditional probabilities were sampled randomly from a uniform distribution. From the simpler graphs, the invalid triplets (where at least one of the $G \rightarrow E$ and $E \rightarrow D$ edges were missing) produced similar Wald-ratios (which are estimations[1] for β_{ED} given by $\beta_{ED} = \frac{\beta_{GD}}{\beta_{GE}}$) to the valid triplets. This suggests that there are certain cases, in which the applicability of MR cannot be determined by the estimation on β_{ED} alone.

To investigate this more generally, we made 50 randomly generated causal graphs, using a simple stochastic algorithm, which iteratively generated random parent-sets for every node (selected from the previously visited nodes), thus creating a guaranteed DAG. Out of the 50 models 25 models had 5 Gene, 3 Exposure and 2 Disease variables (10 in total), while the other 25 had 15 Gene, 10 Exposure and 5 Disease variables (30 in total). The 25 smaller graphs had 6.6 valid and 23.4 invalid paths on average (30 possible paths in total), while the 25 larger models had 20 valid and 730 invalid paths on average (750 possible paths in total). This level of sparsity is roughly representative of the true causal structures of real-world datasets containing Genetic, Exposure and Disease variables. To examine the results of MR, we are only concerned with the estimated strength of causal effect between the Exposure and Disease variables. This value is higher, if β_{ED} is far from 0 in any direction (positive or negative), therefore it is appropriate to use the absolute value of β as a measure for the strength of causal effect. The resulting $|\beta|$ values for the randomly generated partitioned graphs are presented in Table I. From these results, it is evident that the expected value ($\mathbb{E}$) and standard error (σ) of $|\beta_{ED}|$ is significantly larger for invalid triplets, compared to the valid ones.

[1]To estimate the β (effect size) between supposed cause-effect variable pairs, we used logistic regression with 35 steps and the Newton-Raphson optimizer, yielding logarithmic odds-ratios ($log(OR)$).

This can be explained by a simple phenomenon: if we take four variables: $\{X, Y, Z, W\}$, where $(X \to Y \to Z)$ form a valid triplet, so Y is strongly dependent on X, and Z is strongly dependent on Y, but W has no causal connection to any of the other three variables. Therefore, the value of $|\beta_{XZ}|$ will be high (because X affects Z through Y), but the value of β_{XW} will be close to 0, because they are independent. As a result, if we (wrongly) perform MR on the $\{X, W, Z\}$ invalid triplet, then we will get a high value for $|\beta_{WZ}| = |\beta_{XZ}/\beta_{XW}|$. Because of this, we can get a significantly higher $|\beta|$ value for invalid triplets even compared to the valid ones, therefore the use of causal discovery methods are well justified to rule out the invalid cases.

In terms of predictive performance, all three methods were able to find on average 50% of the valid triplets in our 50 partitioned models at 15.000 samples, with a precision of 98%, which means, that 98% of the predicted triplets were correct.

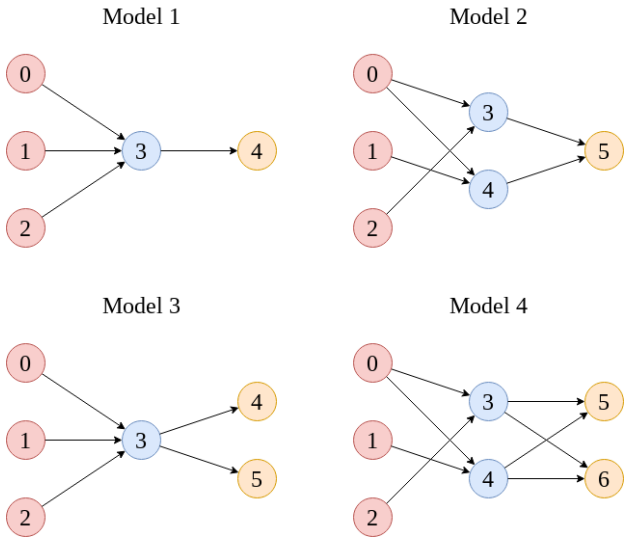

Fig. 2. Models used to demonstrate typical MR β-values. The gene, exposure and disease variables are marked accordingly with red, blue and yellow colors.

V. Approach No. 2: Orienting undirected edges using the MR principle

As we have discussed before, the FCI and RFCI algorithms produce a partially directed graph, where the undirected edges are assumed to be undirected – by the algorithm – in the original essential graph, which belongs to the equivalence class of the original causal structure. In other words, the algorithm assumes that these edges cannot be directed given the known data. Although, within real-world datasets, the number of known samples are finite, and often susceptible to noise. Because of this, we assume that the predicted essential graph will not be perfect, therefore in some cases, the edges that are left undirected by the FCI and RFCI algorithms can be directed by investigating the possible candidate triplets for MR, which the edge in question is a part of. To examine this theory, we propose a method, that consists of the following steps:

1) Acquire a partially directed acyclic graph $\mathcal{G}$ from the known samples using an arbitrary constraint-based structure learning algorithm.
2) For every undirected $(X - Y)$ edge in $\mathcal{G}$, search for all the possible genetic variables, which are not already invalidated by the known directed edges. This includes all the neighbors of X and Y, which are either their parents or they are connected to either of them by an undirected edge. Let's mark the set of these candidate variables by $\boldsymbol{C}_X$ and $\boldsymbol{C}_Y$ for the neighbors of X (not including Y) and the neighbors of Y (not including X) respectively.
3) For every undirected $(X - Y)$ edge in G, find the best candidates for genetic variables G_X and G_Y (in terms of both directions), which are given by:

$$G_X = \underset{G_X \in \boldsymbol{C}_X}{\arg\max} \left| \frac{\beta_{G_X Y}}{\beta_{G_X X}} \right| \quad G_Y = \underset{G_Y \in \boldsymbol{C}_Y}{\arg\max} \left| \frac{\beta_{G_Y X}}{\beta_{G_Y Y}} \right| \quad (1)$$

4) Use G_X to calculate β_{XY} and G_Y to calculate β_{YX}. If $\beta_{XY} > \beta_{YX}$ then orient the edge as $X \to Y$, otherwise orient the edge as $X \leftarrow Y$

This method basically finds the best possible MR triplet for both directions, and orients the edge at the direction determined by the triplet with the highest $|\beta_{ED}|$ score. While the PC method does not give a direct estimation to the essential graph of the original causal structure, it does not provide a direction for most of the edges in the skeleton which are not part of an V-structure. Therefore, we will also examine the applicability of the above described method for the edges that are left undirected by the PC algorithm.

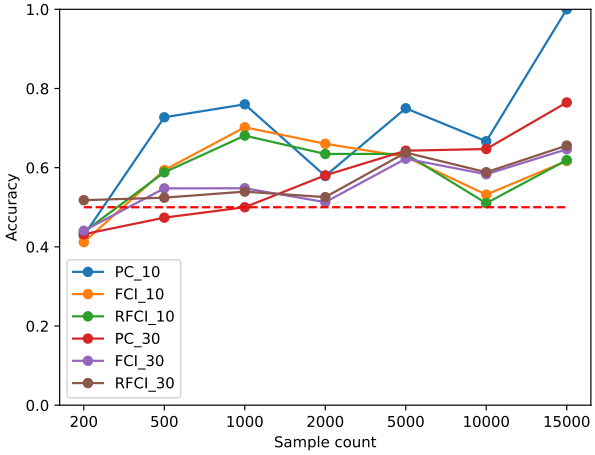

Fig. 3. Edge orientation accuracy (on undirected edges) by the MR β metric, across multiple causal discovery algorithms on partitioned causal graphs with 10 and 30 nodes, with $log(OR)$ beta values. Note, that in case of the unoriented edges of the PC algorithm for the 10-node partitioned graphs, a large number of edges could not be oriented by MR, because they had at least one 0 value in their contingency table, therefore its results are not significant for these graphs. This also explains the outlying accuracy numbers of PC_10 compared to FCI_10 and RFCI_10.

TABLE I
ABSOLUTE β-SCORES ON VALID AND INVALID EXPOSURE-DISEASE TRIPLETS

	Variable count			Log Odds-Ratio			
				Valid		Invalid	
Sample count	**G**	**E**	**D**	$\mathbb{E}(\lvert\beta\rvert)$	$\sigma(\lvert\beta\rvert)$	$\mathbb{E}(\lvert\beta\rvert)$	$\sigma(\lvert\beta\rvert)$
500	5	3	2	0.84	4.21	5.05	36.39
1000	5	3	2	0.69	2.43	5.16	18.57
5000	5	3	2	0.31	0.51	14.29	143.59
10000	5	3	2	0.32	0.98	8.89	37.49
15000	5	3	2	0.32	0.74	10.81	47.06
500	15	10	5	0.66	1.57	4.25	23.81
1000	15	10	5	0.91	4.76	7.08	181.81
5000	15	10	5	0.70	5.56	10.05	286.69
10000	15	10	5	0.45	4.20	6.96	60.45
15000	15	10	5	0.23	0.34	8.64	124.34

In terms of results, the orientation accuracy of our method on the edges left undirected by the PC, FCI and RFCI algorithms on the 50 randomly generated partitioned graphs can be seen in Fig. 3. These results indicate, that our method can predict the orientation of the undirected edges at above-chance levels, from at least 1000 samples. Note, that the partitioned nature of the original causal graph is not assumed by the causal discovery algorithms, and neither by our method, because of which most of the edges oriented by our method are not actually between supposed Exposure-Disease variables. If that were the case, the accuracy would be significantly higher. However, this is an assumption which we cannot make without prior knowledge about the variables, which is not always available. In the partitioned graphs with 10 nodes and 15.000 samples, on average 34% of the edges predicted by FCI and RFCI were undirected, while this ratio rose to 41% in the partitioned graphs with 30 nodes with both algorithms. The PC algorithm left 44% of the edges undirected in the partitioned graphs with 10 nodes, and this ratio fell to 38% for the partitioned graphs with 30 nodes. If we orient these edges randomly (with an even distribution), then in the partitioned graphs with 10 nodes, the FCI and RFCI algorithms oriented 74% of *all* predicted edges correctly, while this accuracy raises to 77% on average with both algorithms, if we used our method to orient the undirected edges. In case of the unoriented edges of the PC algorithm for the 10-node partitioned graphs, a large number of edges could not be oriented by MR, because they had at least one 0 value in their contingency table, therefore the results were not significant. This also explains the outlying metrics of PC_10 in Fig. 3. In case of the partitioned graphs with 30 nodes and also 15.000 samples, MR improved the orientation accuracy of the FCI and RFCI methods from 67% (with random edge orientation) to 72% (with MR edge orientation). However, on these 30-node graphs, the edge orientation accuracy of PC only marginally improved to 74% with MR, compared to the 73% that the algorithm would produce with random undirected edge orientation.

Finally, for the sake of completeness, we also examined the edge orientation performance of this method on completely random directed acyclic graphs (which are therefore not partitioned). Unsurprisingly, it did not produce the same above-chance accuracy values seen on partitioned graphs, further supporting our belief that it only works on the second edge of valid Gene-Exposure-Disease triplets.

VI. CONCLUSION

In this study, we showed that the validity of MR cannot necessarily be determined based on its result, therefore it is advisable to use causal discovery methods for this purpose. We also showed, that MR (for a certain class of directed acyclic causal structures) can improve the edge-orientation capability of the PC, FCI and RFCI methods in terms of the edges that are left unoriented by the original algorithm. As further research, we plan to investigate the integration of MR into other types of causal discovery algorithms, like score-based methods.

REFERENCES

[1] Howey, R., Shin, S. Y., Relton, C., Davey Smith, G. and Cordell, H. J. (2020). Bayesian network analysis incorporating genetic anchors complements conventional MR approaches for exploratory analysis of causal relationships in complex data. PLoS genetics, 16(3), e1008198.

[2] Burgess, S., Foley, C.N., Allara, E., Staley, J.R. and Howson, J.M., (2020). A robust and efficient method for Mendelian randomization with hundreds of genetic variants. Nature communications, 11(1), pp.1-11.

[3] Teumer A. (2018). Common Methods for Performing MR. Front. Cardiovasc. Med, Volume 5, 10.3389/fcvm.2018.00051.

[4] Bowden, J., Davey Smith, G., Haycock, P.C. and Burgess, S., 2016. Consistent estimation in MR with some invalid instruments using a weighted median estimator. Genetic epidemiology, 40(4), pp.304-314.

[5] van der Graaf, A., Claringbould, A., Rimbert, A., Westra, H.J., Li, Y., Wijmenga, C. and Sanna, S., 2020. Mendelian randomization while jointly modeling cis genetics identifies causal relationships between gene expression and lipids. Nature communications, 11(1), pp.1-12.

[6] Pearl, J. (2000). Causality: Models, Reasoning and Inference, C. U. P.

[7] Spirtes, P. and Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. Social Science Computer Review - SOC SCI COMPUT REV. 9. 62-72. 10.1177/089443939100900106.

[8] Spirtes, P. (2001). An Anytime Algorithm for Causal Inference. Proc. of the 8th Int. Workshop on AI and Statistics, PMLR R3:278-285, 2001.

[9] Colombo, D., Maathuis, H. M., Kalisch, M. and Richardson, S. T. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. The Annals of Statistics, 10.1214/11-aos940.