Journal of
Data Mining in Genomics & Proteomics

Yazdani et al., J Data Mining Genomics Proteomics 2016, 7:2
DOI: 10.4172/2153-0602.1000188

Research Article Open Access

Conceptual Aspects of Causal Networks in an Applied Context

Azam Yazdani*, Akram Yazdani, Eric Boerwinkle

Human Genetics Center, UT Health School of Public Health, 1200 Pressler Street, Suite E-447, Houston, Texas, USA

Abstract

Making causal inference is conceptually straightforward in the setting of a randomized intervention, such as a clinical trial. However, in observational studies, which represent the majority of most large-scale epidemiologic studies, causal inference is complicated by confounding and lack of clear directionality underlying an observed association. In most large scale biomedical applications, causal inference is embodied in Directed Acyclic Graphs (DAG), which is an illustration of causal relationships (i.e., arrows) among the variables (i.e., nodes). A key concept for making causal inference in the context of observational studies is the assignment mechanism, whereby some individuals are treated and some are not. This perspective provides a structure for thinking about causal networks in the context of the assignment mechanism (*AM*). Estimation of effect sizes of the observed directed relationships is presented and discussed.

Keywords: Causal inference; Assignment mechanism; Confounder; Causal network; Markov condition; Causal effect

Introduction

Inferring cause-effect relationships among variables is of primary importance in many sciences and is growing in importance as a result of very large datasets in health and genomics. There are several statistical frameworks underlying causal inference, such as those of Rubin's potential outcome framework [1,2]. Pearl's structural equation modeling framework [3] and Dawid's regime indicator framework [4], that have been established for making causal inference. These frameworks are hardly known to most biomedical researchers or biostatisticians who could by applying them to address real world problems. Large segments of the statistical community and decision makers find it hard to benefit from causal analyses. The main reason, we believe, is not a philosophical barrier about data analysis establishing causality, but rather lack of familiarity with the vocabulary and methods in the field. Undertaking statistical causal inference requires systematic extensions to the standard language of statistics, and this perspective provides a step toward this end.

Among available statistical causal inference frameworks, Pearl's causal networks, which are compatible with structural equation models (SEM) [3], can be seen as a pragmatic approach to solving real world problems, especially in the age of large data sets. [5] Has critiqued Pearl's framework and suggests that it requires additional explicit, methodological and philosophical justifications. The concept of the assignment mechanism developed by Rubin [2] describes the circumstances by which some individuals are exposed to a treatment of interest and some are not. In this perspective, we first connect causal networks to the concept of the assignment mechanism (AM). Then, we formalize the causal network parameterization using the AM notation. After discussing the concept and notations of causal networks and the AM, we present effect/causal effect estimation.

Overview of the Assignment Mechanism

The questions that motivate most studies in the health, economics, social and behavioral sciences are causal relationships and not only associations, such as the efficacy of a given drug in a given population. The classical approach for determining such relationships uses randomized experiments where single or a few variables are intervened on. Such intervention experiments, however, are expensive, unethical or even infeasible in many of the cases. Hence, it is desirable to infer causal effects from so-called observational data obtained by observing a system without subjecting it to interventions. Then, to estimate the effect of a treatment on a response, we need to know how different values of the treatment are assigned. The circumstances by which some individuals are exposed to a treatment of interest and some are not is called the assignment mechanism (*AM*).

To achieve causal inference, the important data elements include not only the value of the observations but also the reason why one of the possible exposures or treatments has been realized and not others. The notation *AM* (K_R) is introduced as the third element (in addition to treatment and response value) and is called the causal element [6]. The practitioners need to understand the underlying mechanisms by which some individuals have a certain exposure level and some do not. The knowledge related to response is represented by K_R and is required to identify the *AM*. In a randomized clinical trial the *AM* is straight forward (i.e., the treatment assignment mechanism is unrelated to response). In an observational study, many covariates may influence the *AM* but only some of them are related to response. Variables / covariates that influence both the outcome and the *AM* are termed confounders [7]. The aim of considering the *AM* is to identify individuals with similar confounder distributions as if there were a randomization. In an epidemiologic study, this is similar to matching [8]. In a data analysis setting, this is equivalent to SEM [3] where the *AM* is understood and modeled. Formalizing the *AM* in the context of causal networks compatible with the SEM is more practical in the age of big data. Therefore, in this study, we formalize the *AM* within the context of statistical causal networks.

Causal networks are illustrations of the *AM*, the data generating process underlying the study observations, and provide a pragmatic approach to distinguish confounders of the *AM* from among the covariates, and allows one to analyze observational data as if an

***Corresponding author:** Azam Mandana Yazdani, University of Texas Health Science Center, Houston-1200, Herman Pressler, Houston, Texas, United States, Tel: 713-500-9808; E-mail: azam.yazdani@uth.tmc.edu

Received January 14, 2016; **Accepted** February 10, 2016; **Published** February 17, 2016

Citation: Yazdani A, Yazdani A, Boerwinkle E (2016) Conceptual Aspects of Causal Networks in an Applied Context. J Data Mining Genomics Proteomics 7: 188. doi:10.4172/2153-0602.1000188

intervention was carried out. It is important to understand and take into account that any model in a causal setting is conditioned explicitly or implicitly on illumination of assignment mechanism.

Assume the assignment mechanism over p variables $Y_1, ..., Y_p$ is formalized by a network, here a Directed Acyclic Graph (DAG). The distribution P over these variables is:

$$P(Y_1, ..., Y_p) = \prod_{j=1}^{p} P(Y_j \mid Y_{pa(j)}) \quad (1)$$

Where $pa\ (j)$ denotes the set of predecessors of node j and are directly connected to j in the network, called parents of node j. For $i \in pa(j)$, there is $i \to j$ in the DAG. Note that the formula in (1) represents the Markov properties over these set of variables compatible with the underlying DAG, an illustration of the assignment mechanism that governs over this set of variables. This is a strong assumption in application of DAGs and can be represented in (1) as

$$P(Y_1, ..., Y_p \mid AM(K_R)) = \prod_{j=1}^{p} P(Y_j \mid Y_{pa(j)})$$

By conditioning factorized distributions on the causal element $AM\ (K_R)$, we explicitly represent that the AM is taken into account and the work can, therefore, be considered to be within a causal setting.

Assume the AM over four variables X, Y, Z and H is illustrated in Figure 1. The variable of interest is H, and we are typically investigating the influence of the other variables on H. To factorize the joint distribution over these 4 variables, we first identify potential confounders.

Variables X, Y, and Z are all called covariates. However, the effect of X reaches to H only through Z and Y. Therefore, X is not a confounder of the value of H. The set of confounders for variable H is $C\ (H) = \{Y,Z\}$. The interested reader is referred to the backdoor criterion in [3] for further information. The joint probability over these variables are then factorized as

$$P(H, Y, Z, X \mid AM(K\ )) = P(H \mid Y, Z)P(Z \mid X)P(Y \mid X)P(X$$

Without conditioning on the causal element, $AM(K_R)$, such a unique factorization is not possible [9,10].

Formally Representation of Causal Networks

Assume a DAG D D = (v,ε), where v is a set of nodes with p elements corresponds a set of p random variables with joint Gaussian distribution and ε is a set of edges which connect the nodes and represent the conditional dependencies between two corresponding variables. The existence of a directed edge between two nodes shows the direction of effect (the flow of information) between the correspondent variables. The concept of a DAG D = (v,ε), depends on the nodes in v and edges in ε and any inference depends on the set D=(v,ε). Assume P is a joint probability distribution over variables $Y_1,...,Y_p$ corresponding with nodes in DAG D = (v,ε). D and P must satisfy the Markov condition, the strong assumption in causal inference using networks. These variables have a joint distribution which satisfies the Markov property with respect to the DAG D and all marginal and conditional independencies can be directly obtained from the graph D: every variable Y_i, $i \in v$, is independent of any subset of its predecessors conditioned on the set of its direct or immediate causes of Y_i, corresponding with parents of i,

$$Y_i \perp \{Y_k ; i \ \&\ k \in v \setminus pa(i)\} \mid (Y_{pa(i)}, AM(K_R))$$

Where Y_k occurs before Y_i and parental set $pa(i)$ denotes the set of directly connected nodes to i relatives to AM formalized by DAG D(= (v,ε).

In SEM and under the assumption of a Gaussian distribution, we can write

$$Y_i \mid AM(K_R) = \sum_{j=1}^{i-1} \lambda_{ij} Y_j + U_i \quad (2)$$

where U_i is distributed normally and is independent of the Y_j is in the right side of the model. $\lambda_{ij} \neq 0$ is equivalent with an edge $j \to i$ in DAG D which is due to compatibility of SEM and the AM formalized as the DAG D. SEM is a deterministic form of probability models where all uncertainties are confined in the variable U.

Estimation of Causal Effect and Association

Given a causal network structure, the goal in this section is to discuss effect/causal effect estimation and distinguish it from mere association. To estimate the effect of Y on Z, we consider the causal element $AM\ (K_R)$ embodied in the DAG in Figure 2, which illustrates the assignment mechanism behind the observed variables Y and Z. To obtain the effect of Y on Z, variable X in the path $Y \leftarrow X \rightarrow Z$ is called a confounder. In other words, X confounds the assigning mechanism Y on Z since X influences both Y and Z. Recall that the causal network structure in Figure 2 is an illustration of the assignment mechanism over these three variables and all discussions and equations for the effect measurement is given the assignment mechanism.

To find the effect of Y (and not X) on Z and under Gaussian assumption, we first adjust for the effect of X on Y by

$$Y \mid AM(K_R) = \alpha_{yx} + \beta_{yx} X + e_{yx} \quad (3)$$

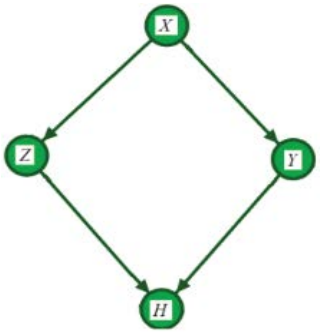

Figure 1: The illustration of the assignment mechanism of variable H formalized as a DAG over four variables.

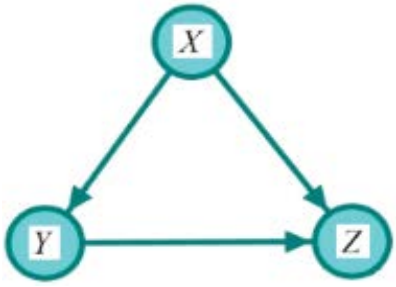

Figure 2: Illumination of the assignment mechanism for the variable of interest Z. In this structure, variable X is a confounder to measure the effect of Y on Z.

and then find the effect of variations in Y on Z by

$$Z \mid AM(K_R) = \alpha_z + \gamma e_{yx} + e_z \quad (4)$$

Equation (4) represents the degree to which variable Y is responsible for the variation in Z, excluding the effect of X. Therefore, the coefficient γ is interpreted as a causal effect. However, in the regression of Z on Y as

$$Z = \alpha^* + \lambda Y + e^* \quad (5)$$

the coefficient λ shows only association between Y and Z since some of the variations in Z attributed to Y is due to the confounder X.

We estimate the effect of X on Z excluding Y by:

$$e_z \mid AM(K_R) = \alpha + \beta X + e$$

where e_z is the residual Z after removing the effect of Y on Z. The coefficient β is interpreted as the effect of X on Z excluding the effect of Y. A mediator effect has not been discussed in this section and interested readers are referred to [11-13].

A Numerical Example for Effect and Association Estimation

To illustrate the above principles, we simulated three variables based on the underlying network in Figure 2 with the primary interest in the effect of Y on Z excluding any effect of X. We simulated 50 set of data and average of estimated effects and average of degree of association over 50 sets are tabulated in Table 1 for three different values of true effects. The standard deviations are presented in parenthesis in the Table 1. The degrees of associations are measured by regressing Z on Y.

We estimated the effects and the degrees of association using equation 3 through 5 while substituting estimate of e_{yx} in equation 4.

Conclusion

We have provided a short and selective perspective of causal inference, including network analysis, the concept of the assignment mechanism, and effect size estimation. A unique aspect of causal inference compared to traditional applied statistics is captured in the concept of the assignment mechanism. To achieve causal inference, the assignment mechanism must be understood and requires close collaboration between analysts and other biomedical scientists. Taking the *AM* into account, we are able to identify confounders and distinguish the effect from association. The assignment mechanism, here formalized in a DAG, can be either known a priori or estimated by an algorithm for directed structures. In this perspective, we assumed that the assignment mechanism is known. In the case of known *AM*, confounders can be identified from the *AM* and the measurements remain in a causal setting.

In most of the cases, the *AM* is not known and needs to be estimated. An ambitious approach is data integration. We have introduced an algorithm called granularity DAG (GDAG), which generates causal networks using data integration [14]. In an application, genomic information is extracted from SNPs scattered across genome by first selecting a subset of informative SNPS using hierarchical clustering and linkage disequilibrium [15] and second principal component analysis. The extracted genome information is used to generate a causal network over phenotypic variables (e.g. body mass index and blood cholesterol levels) of interest.

True effect	Estimated effect	Degree of association
2.000	$2.005_{(0.03)}$	$1.596_{(0.01)}$
3.000	$2.993_{(0.03)}$	$1.729_{(0.02)}$
4.000	$3.995_{(0.03)}$	$1.853_{(0.03)}$

Table 1: Average of estimated effects and degree of associations for three different true effects of Y on Z over 50 replication sets with standard deviation in subscript.

Acknowledgement

This work is supported by a training fellowship from the Keck Center for Interdisciplinary Bioscience Training of the Gulf Coast Consortia (Grant No. RP140113). The corresponding author appreciates Dr. Eric Boerwinkle's editions.

References

1. Rubin DB (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. Journal of Educational Psychology 66: 688-701.
2. Rubin DB (2005) Causal inference Using Potential Outcomes: Design, Modeling, Decisions. Journal of the American Statistical Association 100: 322-331.
3. Pearl J (2009) Causality: Models, Reasoning, and Inference. Cambridge University Press, New York.
4. Dawid P (2012) The Decision-Theoretic Approach to Causal Inference. John Wiley and Sons, Ltd: 25-42.
5. Dawid AP (2010) Beware of the DAG! NIPS Causality: Objectives and Assessment. JMLR: Workshop and Conference Proceedings 6: 59-86.
6. Yazdani A, Boerwinkle E (2014) Causal inference at the population level. International Journal of Research in Medical Sciences 2: 1368-1370.
7. Yazdani A, Boerwinkle E (2015) Causal Inference in the Age of Decision Medicine. Journal of data mining in genomics and proteomics 6: 163.
8. Rosenbaum P (2009) Design of Observational Studies. Springer Series in Statistics.
9. Dawid AP (2007) Fundamentals of statistical causality. Research Report 279, Department of Statistical Science, University College London.
10. Lauritzen SL, Dawid AP, Larsen BN, Leimer HG (1990) Independence Properties of Directed Markov Fields. Networks 20: 491-505.
11. Sobel M (2008) Identification of Causal Parameters in Randomized Studies with Mediating Variables. Journal of Educational and Behavioral Statistics 33: 230-231.
12. Pearl J (2010) An Introduction to Causal Inference. The International Journal of Biostatistics 6: 2.
13. Pearl J (2011) The Causal Foundations of Structural Equation Modelling. Handbook of Structural Equation Modeling. Guilford Press, New York.
14. Yazdani A, Yazdani A, Samiei A, Boerwinkle E (2016) Generating a Robust Statistical Causal Structure Over 13 Cardiovascular Disease Risk Factors by Data Integration. Journal of Biomedical Informatics.
15. Yazdani A, Dunson D (2015) A hybrid Bayesian approach for genome-wide association studies on related individuals. Bioinformatics 31: 3890-3886.