Bayesian Networks

Albers, Casper J.

Link to publication in University of Groningen/UMCG research database

The SAGE Encyclopedia of Research Design

Bayesian Networks

Bayesian networks are probabilistic graphical models depicting the relations within a potentially large set of variables. Commonly, Bayesian networks show the conditional dependencies, information that cannot easily be derived from, for example, a correlation table. Thus, these networks are a versatile tool for exploring and studying relations in large data sets and are used in many branches of science.

Graphical Model

In its essence, Bayesian network models are graphical visualizations of variables and their relations, whereby the values of their relations depend on random variation. Just as in structural equation models and factor models, for example, these visualizations show each variable as an ellipse, called a node, and the relations are depicted by arrows, called edges or ties, between variables.

The graph satisfies the formal requirements of Bayesian networks if it is a so-called directed acyclic graph (DAG). In a directed graph, all connections are arrows—usually indicating causal relations—and there are no cycles in the graph. No cycles implies that if there is a directed path from node A to node B, there cannot be a path from node B to node A.

Many studies focus on correlational relations rather than causal relations. A variant of the Bayesian networks, where directed arrows are replaced by undirected lines, is useful in this context. Although technically not a Bayesian network, these graphs are also commonly referred to as Bayesian networks.

Figure 1 displays the Bayesian network for a fictitious example. For 300 primary school children, three variables are recorded: age ($X1$), their shoe size ($X2$), and their numeracy level, based on some arithmetic test ($X3$). These nodes are depicted as ellipses, with the relations as arrows. In this example, the direction of the causal arrows is clear: one's feet grow with age, and so does one's arithmetic skills.

Figure 1 Example of a Bayesian Network
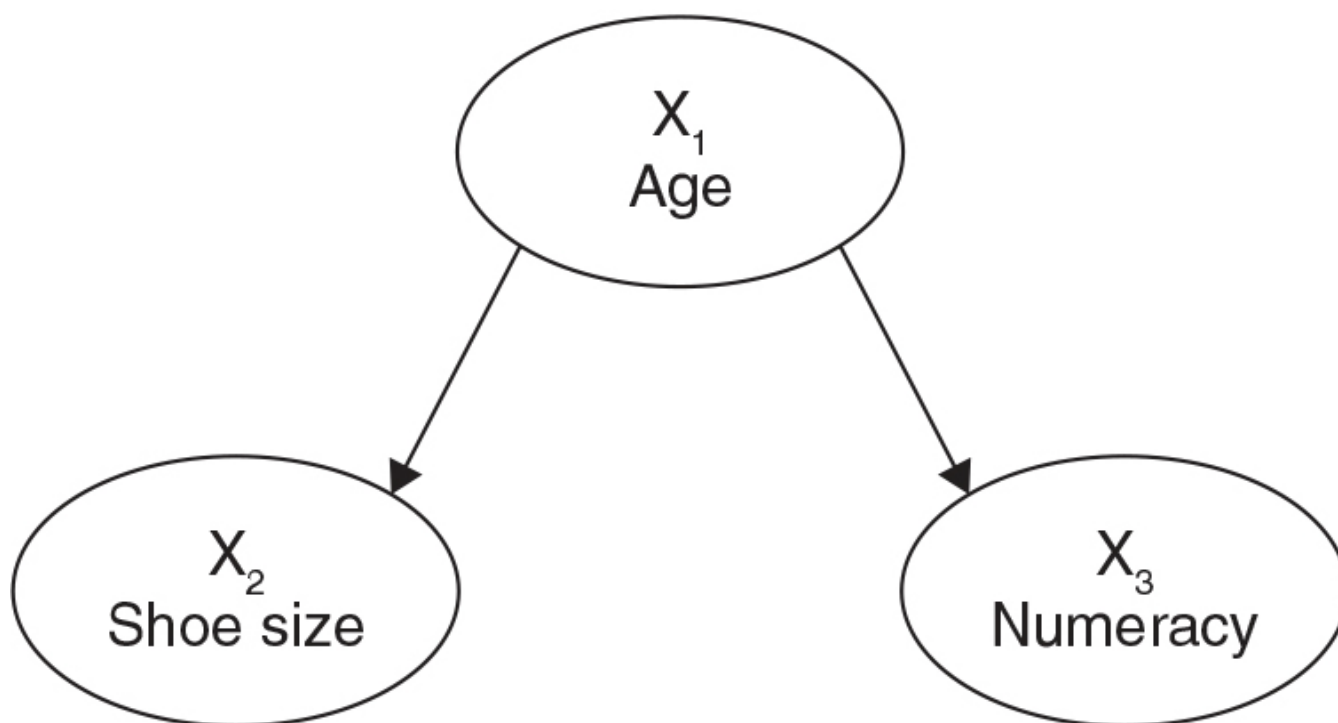
Perhaps the most interesting thing about Figure 1 is the lack of edge between X2 and X3. This does not imply that these two variables are unrelated. In fact, they are probably strongly correlated: kids with large feet tend

to also be the kids with better arithmetic skills, as these children are the older ones in the sample. Variables X2 and X3 have no edge as these are conditionally independent: given the value for X3, the two variables are unrelated. In other words, for children of the same age, there is no relation between shoe size and numeracy.

Formally, variables A and B are independent conditional on variables C1, …, Ck if

$$P A \cap B C_1, \ \dots \ , \ C_k = P A C_1, \ \dots \ , \ C_k \times P B C_1, \ \dots \ , \ C_k.$$

As this notation relies on conditional probabilities, Bayes's theorem is required, hence the name Bayesian network. The notation for this is A||B|C1, …, Ck. The conditional independence property of DAGs is what makes the Bayesian network so useful in practice. This is especially the case when working with a large number of variables. By conditioning on certain nodes, sets of other nodes can become independent, which is of great help in interpreting the results.

Gaussian Graphical Model

A common type of Bayesian network model is the Gaussian graphical model (GGM). This model assumes that the joint distribution of all variables is multivariate normal: $X \sim N \mu, \Sigma$. In this case, the inverse of the variance/covariance matrix Σ immediately provides the values for the edges of the network. Where Σ itself is used to compute the correlations between variables, $\Sigma - 1$ is used to compute the so-called partial correlations. The normality assumption of the GGM can be checked in a similar way as checking this assumption for standard linear regression models.

There are two methods for visualizing the edges of a graph: (1) lines and arrows are either present or absent, denoting the presence or absence of a relation, or (2) lines and arrows vary in thickness, denoting the strength of the relation, with another property (different line color or using dashed lines) to indicate when relations are negative. As partial correlations in practice never are exactly zero, raw visualizations of a GGM can yield visual overload: when all edges are drawn, it is difficult to assess where the interesting edges are. A straightforward solution to this is called thresholding: visualize only those partial correlations that, in absolute value, exceed some threshold (e.g., 0.2). Correlations smaller than this threshold are deemed to be of too little interest. More sophisticated solutions, such as the graphical lasso, might provide a more elegant graph.

Bayesian Networks in the Social Sciences

Within the social sciences, Bayesian networks occur in two classes: social networks and psychological networks. In social network analysis, the nodes represent entities, such as Facebook users, and the edges represent 0/1 variables, such as indicating whether or not two Facebook users are (mutual) friends. The interest does not lie in specific individuals in the network but on the behavior of the network as a whole. Relevant concepts in this field include density (how many edges are there, relative to the possible number of edges) and centrality (which nodes are most connected, who are the key influencers).

Psychological network analysis has a different focus. Here, the nodes concern psychological variables, such as symptoms of anxiety disorder, and the edges represent the partial correlations. Thus, the values of the edges are now realizations of a random process, rather than observed 0/1 scores. One of the main aims in this branch of network analysis is to find pathways connecting one symptom to another. By intervening on one of the symptoms on this path, and thus keeping this constant, one hopes to avoid having other symptoms worsen.

Especially thanks to recent innovations in software for network analysis, the use of network models has increased rapidly since the first decade of the 21st century.

See also Bayes's Theorem; Network Analysis; Network Visualization

Casper J. Albers
http://dx.doi.org/10.4135/9781071812082.n44

10.4135/9781071812082.n44

Bhushan, N., Mohnert, F., Sloot, D., Jans, L., Albers, C. J., & Steg, L. (2019). Using a Gaussian graphical model to explore relationships between items and variables in environmental psychology research. Frontiers in Psychology, 10, 1050. doi:http://dx.doi.org/10.3389/fpsyg.2019.01050.

Bishop, C. M. (2006). Pattern recognition and machine learning. New York, NY: Springer.

Borsboom, D., & Cramer, A. O. J. (2013). Network analysis: An integrative approach to the structure of psychopathology. Annual Review of Clinical Psychology, 9, 91–121. doi:http://dx.doi.org/10.1146/annurev-clinpsy-050212-185608.

Epskamp, S., Cramer, A. O. J., Waldorp, L. J., Schmittmann, V. D., & Borsboom, D. (2012). qgraph: Network visualizations of relationships in psychometric data. Journal of Statistical Software, 48, 1–18. doi:http://dx.doi.org/10.18637/jss.v048.i04.

Jones, P. J., Mair, P., & McNally, R. J. (2018). Visualizing psychological networks: A tutorial in R. Frontiers in Psychology, 3389. doi:http://dx.doi.org/10.3389/fpsyg.2018.01742.

Lauritzen, S. L. (1996). Graphical models. Oxford, UK: Clarendon Press.

Pearl, J. (2000). Causality: Models, reasoning, and inference. Cambridge, UK: Cambridge University Press.

Van Duijn, M. A., & Vermunt, J. K. (2006). What is special about social network analysis? Methodology, 2(1), 2–6.