MY RESEARCH: GRAPHICAL MODELS

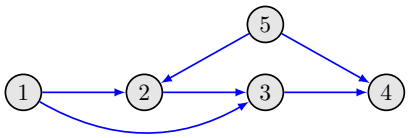
MATHIAS DRTON

Graphical models emerged in the 1980s on the interface between Statistics and Artificial Intelligence. This was a time when probabilistic approaches to model larger systems still faced some skepticism: How to compute with 2^{100} probabilities when studying a system with 100 on/off switches? This skepticism quickly disappeared as graphical models provided an abstract framework in which algorithms for efficient computation with structured probability distributions could be developed at a general level. Since these early days, the field has seen tremendous growth during which graphical models have found wide-spread applications and have also become an important tool for rigorous statistical investigation of causality.[1]

What is a graphical model? Let $(X_i)_{i\in V}$ be a finite collection of random variables. In a statistical context, the joint probability distribution of the variables is unknown and to be inferred from data. A statistical model is a set of candidates for the unknown distribution. In graphical modeling, each model is induced by a graph whose vertices correspond to the variables X_i. The graph's edges then encode structure that the model assumes of the unknown distribution. How this works precisely depends on the type of graph. Here I will focus on directed graphs, which naturally capture cause-effect relationships among the variables.

Let $G = (V, \mathcal{E})$ be a directed graph with vertex set V and edge set $\mathcal{E} \subseteq V \times V$, without self-loops, so $(i, i) \notin \mathcal{E}$ for all $i \in V$. For each vertex i, define a set of parents $\mathrm{pa}(i) = \{j \in V : (j, i) \in \mathcal{E}\}$. The model induced by the graph G hypothesizes that each variable X_i is a function of its parents X_j, $j \in \mathrm{pa}(i)$, and a stochastic noise term ϵ_i. Let us consider the case where the functional relationships are linear, as frequently assumed in practice.

Example. Take G to be the following graph (known in the field as the Verma graph):



The model induced by G is comprised of the joint distributions of random vectors $(X_1, \ldots, X_5)$ that solve the equation system

$$\begin{aligned} X_1 &= \epsilon_1, \\ X_2 &= \lambda_{12} X_1 + \lambda_{52} X_5 + \epsilon_2, \\ X_3 &= \lambda_{13} X_1 + \lambda_{23} X_2 + \epsilon_3, \\ X_4 &= \lambda_{34} X_3 + \lambda_{35} X_5 + \epsilon_4, \\ X_5 &= \epsilon_5, \end{aligned}$$

for a choice of real coefficients λ_{ij} and independent random variables $\epsilon_1, \ldots, \epsilon_5$. As G is acyclic, the equation system is triangular and always has a unique solution.

The equations write each variable as a (linear) function of other variables and noise. Taking the functional relations seriously and thinking of them as making an assignment of values is the basis for the model's causal interpretation, i.e., for letting the model also make statements about different experimental settings. To explain briefly, imagine an experimenter

[1] For a summary of the state-of-the-art, see the "Handbook of Graphical Models" which just appeared (co-edited by Marloes Maathuis, Steffen Lauritzen, Martin Wainwright and myself).

is able to "turn off" variable X_3 without changing how our hypothetical 5-variable system behaves. Then a model for this new experimental intervention is obtained by replacing the third equation by $X_3 = 0$ but leaving all other equations unchanged.

Research interests. My interest in graphical models is broad. On one hand, new applications continually call for refinements of models and methods. For instance, assumptions such as linearity or errors with Gaussian distributions may be inappropriate, or careful statistical considerations may be required to accurately estimate low-dimensional structure from high-dimensional data. On the other hand, there remain challenging open problems about fundamental properties of basic models, such as the exemplified models based on linear relations. I would like to elaborate on this latter point here.

Take up the above example. When the errors ϵ_i are Gaussian, all information about the underlying graph is captured by the (positive definite) covariance matrix of the variables X_i. In this sense, the model corresponds to a set of positive definite matrices. According to the "trek-rule," the covariances have beautiful combinatorial structure, which has driven graphical solutions to many statistical problems. For instance, in the above example, the covariance between X_2 and X_3 is

$$\mathrm{Cov}[X_2, X_3] \;=\; \lambda_{12}\lambda_{13}\omega_1 + \lambda_{12}^2\lambda_{23}\omega_1 + \lambda_{23}\omega_2 + \lambda_{23}\lambda_{52}^2\omega_5,$$

with ω_i being the variance of noise term ϵ_i. Observe how each one of the four summands corresponds to a particular path (or rather walk) between nodes 2 and 3 in the graph.

A lot is known about the set of covariance matrices when the considered graph $G = (V, \mathcal{E})$ is an acyclic digraph:

a) Its dimension is simply $|V| + |E|$, the count of vertices plus edges.
b) A simple rational expression recovers each coefficient λ_{ij} from the covariance matrix.
c) Each set of covariance matrices is a smooth manifold cut out by conditional independence relations (algebraically, these are special subdeterminants).
d) We understand precisely and can check efficiently if two graphs induce the same model/set of covariance matrices.

These facts have a natural generalization for models obtained from possibly non-linear functional relations, and they provide the basis for effective statistical methodology that learns from data the graph underlying a model as well as all unknown model parameters.

However, many practical problems bring about complications. For instance, often not all relevant variables can be measured (i.e., some variables are latent/unobserved) or a system may contain feedback loops (i.e., the graph may contain directed cycles). Taking up our running example, suppose we only observe $(X_1, \ldots, X_4)$. Then there are no conditional independence relations holding among these variables alone and the theory outlined above yields no useful information. Now this example is simple enough to derive that a positive definite 4×4 matrix $\Sigma = (\sigma_{ij})$ is the covariance matrix of (X_1, X_2, X_3, X_4) under our graphical model if and only if Σ satisfies a polynomial constraint of degree 4 with 8 terms. The constraint can be written, e.g., as the determinant of a matrix of 2×2 minors:

$$f_{\text{Verma}} \;=\; \begin{vmatrix} |\Sigma_{12,12}| & |\Sigma_{12,13}| \\ |\Sigma_{34,12}| & |\Sigma_{34,13}| \end{vmatrix} \;=\; 0.$$

However, for more complicated graphs it may be unclear what relations among observable covariances are when there are latent variables/feedback loops. It may also be unclear what the dimension of such a model is, or whether other graphs yield the same model for the observed variables. These types of questions have driven a significant part of my recent research,[2] and although much progress has been made, many questions still await an answer.[3]

[2] `https://arxiv.org/find/math,stat/1/au:+Drton_M/0/1/0/all/0/1`

[3] A relatively recent review can be found in `https://arxiv.org/abs/1612.05994`.