Parameter identification in graphical models

organized by
Mathias Drton and Seth Sullivant

Workshop Summary

Background on the Problem

A graphical statistical model is a family of probability distributions associated with a graph. The connection between graph and statistical model is made by identifying the nodes of the graph with random variables, and by requiring that the variables' joint distribution exhibit certain independence relations. These independence relations are associated to the non-edges in the graph. This connection has been formalized for different types of graphs including as the basic types undirected graphs and directed acyclic graphs.

Most commonly, three cases are considered:

(1) In the Gaussian case one assumes the joint distribution of all involved variables to be a multivariate normal distribution. For both undirected graphs and digraphs the associated Gaussian graphical models require the covariance matrix of the normal distribution to be of a certain form. In fact, the models correspond to polynomially parametrized subsets of the cone of positive definite matrices.
(2) The discrete case concerns a collection of random variables taking their values in finite sets. In this case the graphical models correspond to subsets of a probability simplex. Again, these subsets are given in parametric form as the image of a polynomial map.
(3) In the general non-parametric setting, we consider arbitrary noisy functional relationships between the random variables.

The Gaussian case, in particular, includes various forms of recursive linear regressions. Much of the original development of these types of models occurred in economics, where they are called structural equation models. SEMs continue to play an important role in modern econometrics. An important issue in SEMs is that of hidden variables, which are variables for which no data are available. In applications in the social sciences, they typically correspond to intrinsically unobservable quantities or theorized confounding variables that might lead to a correlation between observed random variables.

For discrete random variables, graphical models play an increasingly important role in computational biology and phylogenetics. Phylogenetic models are graphical models whose underlying graph is a tree. Each vertex in the tree corresponds to a different species. These models usually involve many hidden variables because the internal nodes in the tree correspond to species that are extinct. Modern analysis of phylogenetic data now involves mixtures of tree models, which are ever more complicated hidden variable models obtained from these tree models.

A fundamental problem in the area of graphical models is to determine which model parameters can be identified when the available information is partial, that is, when some

variables are hidden and the available data is restricted to a subset of the considered random variables. Considering a subset of the random variables amounts to a projection. In the Gaussian case the relevant projection consists of passing to a principal submatrix, and in the discrete case one forms sums over certain subsets of coordinates. In each case the projection preserves the polynomial structure of the parametrization map. Hence, the mathematical problem consists of studying when the concatenation of model parametrization and projection is bijective and exactly how bijectivity may fail.

Due to its importance to many applied fields there exists a considerable literature on the problem. In statistics and computer science, an active community working in an area referred to as "causal inference" publishes formulas identifying a parameter as a function of a given (partial) covariance matrix/probability vector. In economics, a special identification problem known as the instrumental variables problem has received a lot of attention. However, despite its algebraic and combinatorial nature it is only recently that the problem has attracted the attention of mathematicians.

The main goal of this workshop was to bring together key researchers working in statistics, computer science and discrete mathematics to formulate precise open problems and discuss approaches to their resolution using the machinery of algebraic geometry, commutative algebra, combinatorics, and symbolic computation. These tools come into play because the underlying maps are given by polynomials parametrizations that are combinatorially defined in terms of the underlying graphs.

Workshop Activities and Outcomes

The workshop followed the standard AIM format of having two introductory lectures in the morning of each day, Monday through Friday, and breaking into groups to work on open problems in the afternoons. Since the group consisted of a mix of participants with very different backgrounds (some from statistics and computer science, and some from algebra, combinatorics, and symbolic computation), the introductory lectures were essential for introducing the techniques and problems for the different areas.

The workshop opened with a general background lecture by Jin Tian who reviewed the different types of models mentioned in the above 'Background' section. The other introductory lectures treated algebraic methods for identification in Gaussian models (Luis Garcia), identifiability of discrete graphical models (John Rhodes and Elena Stanghellini), identification of causal effects in non-parametric models (Marco Valtorta and Ilya Shpitser), trek separation in Gaussian models (Kelli Talaska), the use of tetrad constraints in statistical inference (Peter Spirtes), and new statistical approaches and problems in non-parametric modelling (Jamie Robins and Thomas Richardson).

To get the two groups interacting with each other early on, on Monday afternoon Luis Garcia lead an activity on using computer algebra packages to try to solve identifiability problems. This involved having the group break in to pairs, with one person from the Statistics/CS side, and one from the Algebra side. On Monday, we also had a session led by Peter Spirtes to solicit problems to work on for the remainder of the week. The afternoons of Tuesday through Thursday were then devoted to working on specific problems that were identified to be of mutual interest to the participants.

Each afternoon the participants split up into three or four groups. The following specific group topics were worked on:

Combining Methods for Discrete Graphical Models This group formed on Tuesday afternoon after the talks of John Rhodes and Elena Stanghellini. They realized that the techniques that each was employing to prove identifiability, could be combined to get stronger results and worked throughout the week with Marco Valtorta to try to develop their ideas further. Joined by John Rhodes' collaborator, Elizabeth Allman, the group of four will continue to pursue the identifiability problem regarding discrete graphical models with latent variables after the workshop. They are in fact preparing a SQUARES application to continue the collaboration at AIM.

Proving Generic Nonidentifiability While it is easy to prove that a given model is not identifiable (just produce two choices of parameters that map to the same probability distribution), proving that a model is not generically identifiable seems to be much more difficult, and currently only Gröbner basis techniques give a general purpose certificate for generic non-identifiability. The goal of this group was to develop methods for proving that a model is not generically identifiable. It focused on two problems: 1) to understand what the fibers of the parametrization are like in the not generically identifiable case and 2) to understand when the Jacobian of the parametrization drops rank. Some time was spent with computational experiments in the Gaussian case, where interesting factorizations of subdeterminants of the Jacobian were observed but more work remains to be done to obtain a better understanding of generic nonidentifiability.

The G-criterion Carlos Brito developed a method for proving generic identifiability in the case of linear structural equation models. However, his papers do not have complete proofs of the results. A group of participants worked to try to fill in possible holes in the proof. One of the difficulties (the nonvanishing of a certain determinant) was resolved using trek separation results that Kelli Talaska lectured on on Thursday morning. The group is continuing work on this problem after the workshop.

Converting Non-parametric Constraints to Gaussian Constraints The search for formulas for causal quantities in the general nonparametric setting leads to interesting non-parametric constraints on probability densities, which are not completely understood. These formulas involve a combination of marginalization and conditioning, and products and quotients of the resulting densities. The goal of this group was to try to understand how to convert these constraints into polynomial constraints on the covariance matrix in the Gaussian case. The basic problem for the so-called Verma graph was resolved during an afternoon session, but interesting open problems for future work remain in the general case.

Non-parametric Algebraic Statistics The existing non-parametric constraints and identification formulae are derived by careful manual simplification of factorizations of a joint probability density function that is being integrated with respect to a subset of variables. (The variables integrated out correspond to the hidden variables in the problem.) The main operations used in these derivations are: (i) replace one conditional density by another based on consideration of conditional independence and (ii) use that probability densities integrate to one to eliminate variables. This group discussed how one might develop a suitable algebraic framework for developing computer algorithms that find general nonparametric constraints and identification formulae. This goal is clearly too ambitious for resolution at the workshop but the discussion between the different camps of participants was stimulating.

Linearity Below the Choke Point Peter Spirtes suggested the very concrete problem of determining when the trek separation result can be extended to situations where there are not necessarily linear relationships between the random variables. This is of relevance for inference in partially linear hidden variable models and presents a step towards generalizing past work of Silva, Scheines, Glymour, and Spirtes. This problem was solved by the group. Further work is required, however, to allow the result to be used for statistical inference.

Deep Belief Networks Jason Morton described a specific family of graphical models, known as deep belief networks, which he wants to know identifiability of. The literature provides only partial results in the simplest case (one level of hidden variables). This group worked to try to extend results to higher levels of deep belief networks, and focused on computational aspects and understanding the geometry of these models.

The final session of the workshop, on Friday afternoon, consisted of two group discussions. The first was designed to identify major open problems in the area (summarized in an open problem list). The second was focused on activities that people would be pursuing immediately as a consequence of the workshop. This session provided all participants with a last opportunity to describe which of the various projects that arose from the workshop they would like to pursue further.