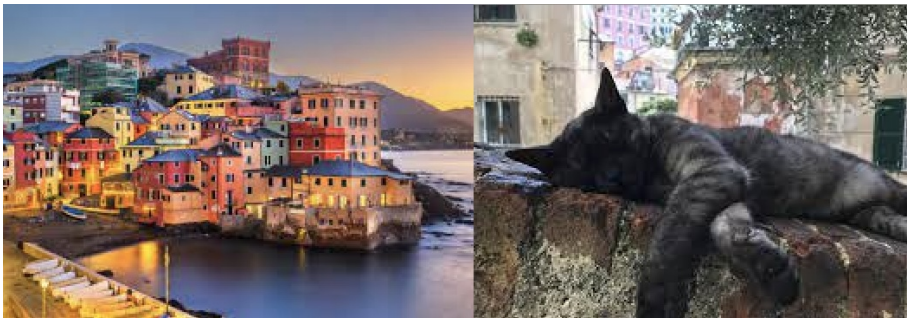
Trees4Cat Workshop 2024

Book of Abstracts

Department of Mathematics (DIMA), University of Genoa

21st - 23rd October 2024



The event is supported by:

Contents

Tutorials

Eliana Duarte (University of Porto)

Title: Algebraic Statistics of Tree Models

Abstract:
This tutorial introduces participants to key concepts in algebraic statistics and how these enable a better understanding of statistical properties of tree models. One of the main insights in algebraic statistics is that statistical models can often be defined by systems of polynomial equations. Understanding the algebraic properties of these equations reveals useful characteristics of the model. The tutorial will start by introducing algebraic concepts such as polynomial rings, ideals, and the zero sets that they define, followed by relevant examples in statistics. Next, these concepts will be applied to study probability tree models and their main results.

Manuele Leonelli (IE University)

Title: Software for Staged Tree Models

Abstract:
This tutorial will focus on the application of software for data analysis using staged tree models, a versatile class of statistical models designed to represent and reason about asymmetric dependence structures and sample spaces. Staged trees offer a flexible framework for modeling complex data dependencies, making them particularly valuable in various real-world scenarios. Participants will be guided through case studies that demonstrate the practical utility of staged tree models, showing how the software can be employed to uncover patterns and relationships in data. Through hands-on examples, attendees will learn how to build, analyze, and interpret staged tree models, enhancing their data analysis workflows.

Main Workshop

Anna Maria Bigatti (University of Genoa)

Title: Discovery of Statistical Equivalence Classes using Computer Algebra

Abstract:
Equivalent representations of statistical data as "staged trees" are useful for analysing the underlying statistical model with a view on putative causal interpretations.

Here we present the concept of "staged tree" from an algebraist's point of view. We give the definition and show their representation as a polynomial, then show how we designed and implement an algorithm for finding all equivalent staged trees within the CoCoA System (Computations in Commutative Algebra).

Hollie Callie (University of Exeter)

Title: Chain Event Graphs with Spatial Extensions: A Shiny App for Encoding Expert Judgement

Abstract:
In fields such as criminal justice and public health, domain experts can provide statisticians with descriptions of processes of events, pertaining to which events might happen, why they might happen, and the possible outcomes generated by certain sequences of events. Statisticians play a vital role in converting these verbal descriptions into statistical models, though translating this information into a coherent mathematical representation is no trivial task.
Bayesian models tend to encode these expert judgements in the form of prior distributions on specific variables within the model. However, to a non-statistician, this process can often be relatively opaque, leading to a lack of understanding regarding how the outputs of these models are produced (which in turn can call into question a model's overall reliability). It follows that in the age of AI and black box models such as Bayesian Neural Networks, the appeal for interpretable models for use in these situations is increasing. One candidate is the Chain Event Graph (CEG), which produces a graphical representation of a series of events along with its underlying statistical model, enhancing accessibility and comprehension.
That being said, it remains the case that popular methods for Bayesian inference are able to handle much more complex interactions between variables than CEGs. Notably, while several statisticians have focused on the extension of CEGs to model temporal interactions, no one (as of yet) has presented developments for spatial modelling. To this end, this talk introduces a versatile R Shiny application developed in the first year of my PhD, designed to support ongoing research into expanding the capabilities of CEGs by integrating spatial characteristics. This advancement would significantly broaden the scope of the types of situations CEGs can model. The app exhibits novel features, such as clickable colouring, and the ability to fully specify conditional dependencies without relying on algorithms like Agglomerative Hierarchical Clustering (AHC). As my PhD research progresses, further spatial extensions will be incorporated into the app, with these integrating with the existing temporal CEG theory.

Danai Deligeorgaki (KTH)

Title: Marginal Independence Structure of DAG Models

Abstract:
We consider the problem of estimating the marginal independence structure of a DAG model from observational data. In order to so, we divide the space of directed acyclic graphs (DAGs) into certain equivalence classes, where each class can be represented by a unique undirected graph called the unconditional dependence graph. The unconditional dependence graphs satisfy certain graphical properties, namely having equal intersection and independence number. Using this observation, we can construct a Grobner basis for an associated toric ideal and define additional binomial relations to connect the space of unconditional dependence graphs. With these moves, we can implement a search algorithm, GrUES (Grobner-based Unconditional Equivalence Search), that estimates the conditional independence structure of the graphical model. The implementation shows that GrUES recovers the true marginal independence structure via a BIC-optimal or MAP estimate at a higher rate than simple independence tests while also yielding an estimate of the posterior. This is joint work with Alex Markham, Pratik Misra and Liam Solus.

Jane Hutton (University of Warwick)

Title: Evaluating Screening Policies Using Chain Event Graphs

Abstract:
Police officers have powers to stop and search people, to prevent or detect crime. The UK has at least 10 screening programmes, to detect people with a higher chance of disease. Most data is categorical, and the relevant questions rely on conditional probabilities.

Headlines such as "Ethnic minorities unfairly targeted by police" refers to differences in the proportion of people of category A in the population being stop and searched compare to people B. To preserve law and the King's peace, do we wish to stop the same proportions: do we wish to stop the same proportion of women and men, old and young, black and white? Or do we wish similar conditional probabilities for A and B stopped to be carrying an illegal article? Or do we wish similar outcomes for A and B who are carrying an illegal article?

Summaries of Metropolitan police stop and search data from October 2017 to March 2024 are presented. Preliminary results indicate that a quarter of those stopped are found to be carrying an illegal item. This rate is higher than positive predictive powers of 5% or more accepted in health. Decisions on the best rate depend on assumptions about the values of not stopping people and of preventing crime. The objects of searches differ by age, sex and ethnicity. The outcome of a search, adjusted for its purpose object of search, varies slightly between ethnic groups. Data appears not be to missing at random, with different patterns depending on age, year and ethnicity.

Maria Kateri (RWTH Aachen University)

Title: Statistical Information Theory meets Categorical Data Analysis

Abstract:
Standard statistical models for categorical data (e.g., models for contingency tables, binary regression, models for rank data) are redefined in a statistical information theoretical setup through their link to divergences. This way, new properties for these models are revealed that lead to a deeper understanding of their nature and provide new interpretation options. Furthermore, through families of divergences, that include the well-known Kullback–Leibler and Pearson χ^2 divergences as special cases, standard models are generalized to flexible families of models. The potential of such model families is demonstrated by examples.

Manuele Leonelli (IE University)

Title: Modeling and Visualizing Uncertainty in Context-Specific Models

Abstract:
Structural learning techniques for staged tree models have been introduced in both frequentist and Bayesian frameworks. However, limited attention has been given to the uncertainty surrounding the dependences learned from these models. In this talk, we present a fully Bayesian approach that addresses this gap by using a split-and-merge MCMC algorithm to explore the space of staged tree models. This approach introduces novel classes of prior distributions over the model space, compared to the uniform priors predominantly used in the literature. The output of the algorithm is a large collection of staged trees, raising the challenge of summarizing this information effectively. To address this, we borrow techniques from Bayesian clustering and construct a credible ball of staged trees, providing a visual and interpretable summary of the model uncertainty.

Monia Lupparelli (University of Florence)

Title: Path-Dependent Parametric Decompositions in Ising Models

Abstract:
The analysis of paths in undirected graph models can be used to quantify the relevance of the strength of association in multiple paths connecting a pair of vertices of the graph. Some results are available in multivariate Gaussian settings as the covariance of two variables can be decomposed into the sum of measures related to paths joining the variables of the underlying graph. This work studies the analysis of paths in undirected graph models for binary data, with special focus on Ising models, where the propagation of the variable status through multiple paths joining a pair of vertices is an aspect of interest. A novel logistic regression approach for baseline events in multi-way tables is proposed to show that a parameter of pairwise association can be computed by the sum of components related to paths. These components are based on products of odds ratios which are typically used to measure the dependence represented by the edges in Ising models. The results are illustrated through an application to cyber-security risk assessment in industrial networks.

Shuhei Mano (The Institute of Statistical Mathematics)

Title: Rational MLE and Direct Sampling from Conditional Distributions

Abstract:
We can directly sample from the conditional distribution of any toric (log-affine) model without using the Metropolis algorithm (Electronic Journal of Statistics 11: 4452–4487, 2017; arXiv:2110.14922 for recent progress). The direct sampling algorithm is a Markov chain on a bounded integer lattice, where each element has a one-to-one correspondence to the $\mathcal{A}$-hypergeometric polynomial. The transition probability of the Markov chain is the ratio of two $\mathcal{A}$-hypergeometric polynomials. The algorithm's computational complexity critically depends on the computation of the transition probability. For log-linear models, the transition probability is the ratio of the MLE of the expected count to the sample size. In this respect, having the rational MLE, as in decomposable graphical models, is ideal since we can avoid the computational burden. In this presentation, we will see that the algorithm works efficiently for any log-linear model as having the rational MLE with the aid of the iterative proportional fitting.

Orlando Marigliano (University of Genoa)

Title: Maximum Likelihood Degree of Discrete Models

Abstract:
I introduce the concept of maximum likelihood (ML) degree in algebraic statistics and talk about the case where the ML degree is one. In that case we can write the ML estimate of a model as a rational function in the data. I show this function for the case of staged tree models, discuss more general discrete models, and illustrate the use of the ML degree to do likelihood estimation with numerical-algebraic techniques.

Alex Markham (University of Copenhagen)

Title: Scalable Structure Learning for Sparse Context-Specific Systems

Abstract:
Several approaches to graphically representing context-specific relations among jointly distributed categorical variables have been proposed, along with structure learning algorithms. While existing optimization-based methods have limited scalability due to the large number of context-specific models, the constraint-based methods are more prone to error than even constraint-based directed acyclic graph learning algorithms since more relations must be tested. We present an algorithm for learning context-specific models that scales to hundreds of variables. Scalable learning is achieved through a combination of an order-based Markov chain Monte-Carlo search and a novel, context-specific sparsity assumption that is analogous to those typically invoked for directed acyclic graphical models. Unlike previous Markov chain Monte-Carlo search methods, our Markov chain is guaranteed to have the true posterior of the variable orderings as the stationary distribution. To implement the method, we solve a first case of an open problem recently posed by Alon and Balogh. Future work solving increasingly general instances of this problem would allow our methods to learn increasingly dense models. The method is shown to perform well on synthetic data and real world examples, in terms of both accuracy and scalability.

Tamás Rudas (Eötvös Loránd University)

Title: On the analysis of data from incomplete experiments

Abstract:
Most of multivariate statistical analysis takes place in the Cartesian product of the ranges of the variables of interest. There are, however, experimental designs (and populations) that do not lead to such a data structure. This talk concentrates on data sets that have a product structure but the product is incomplete. Examples include, in addition to incomplete factorial designs, register data, and life tables. Classical independence and its generalizations are not always meaningful for such data, and one has to apply alternative simple structures in modeling. Some of the properties of estimates and test statistics are discussed. An important class of such experiments are sequential, where whether or not a further observation is made, depends on the outcome of a previous observation. Examples include, in addition to staged trees, also data about offsprings or testing the effects of interventions. The talk discusses the correct distributional assumptions and properties of the resulting estimates. This is joint work with Anna Klimova.

Liam Solus (KTH)

Title: Getting the Most from Your Data in Causal Structure Identification with the Help of Geometry

Abstract:
When modeling causal systems with directed acyclic graphs, unsupervised learning methods for recovering the graph of the causal system faces a natural issue: Without any additional modeling assumptions, it is typically unidentifiable from only observational data. On the other hand, structural identifiability is indeed possible under additional assumptions such as model parameter homogeneities or context-specific information. From an unsupervised perspective, rather than imposing such assumptions a priori, one can aim to learn these features of the data-generating distribution from the observational data and thereby solve the structural identifiability problem to the fullest extent possible; i.e., up to classical Markov equivalence or (ideally) beyond. To obtain the necessary theory supporting such algorithms a number of tasks lay before us: (1) We require a characterization of those submodels that are generically distinguishable provided with this additional information, and (2) we require algorithms for efficiently searching the space of corresponding submodels. In this talk, we will see examples of how geometric perspectives provide us with tools for addressing both of these problems. We will further discuss some resulting problems that are perhaps of specific interest to the workshop attendees.