09401 Abstracts Collection
Machine learning approaches to statistical dependences and causality
— Dagstuhl Seminar —

Dominik Janzing[1], Steffen Lauritzen[2] and Bernhard Schölkopf[3]

[1] MPI für biologische Kybernetik - Tübingen, D
dominik.janzing@tuebingen.mpg.de
[2] University of Oxford, GB
steffen@stats.ox.ac.uk
[3] MPI für biologische Kybernetik - Tübingen, D
bernhard.schoelkopf@tuebingen.mpg.de

Abstract. From 27.09.2009 to 02.10.2009, the Dagstuhl Seminar 09401 "Machine learning approaches to statistical dependences and causality" was held in Schloss Dagstuhl – Leibniz Center for Informatics. During the seminar, several participants presented their current research, and ongoing work and open problems were discussed. Abstracts of the presentations given during the seminar as well as abstracts of seminar results and ideas are put together in this paper. The first section describes the seminar topics and goals in general. Links to extended abstracts or full papers are provided, if available.

Keywords. Machine learning, statistical dependences, causality

09401 Executive Summary – Machine learning approaches to statistical dependences and causality

The 2009 Dagstuhl Seminar "Machine Learning approaches to Statistical Dependences and Causality", brought together 27 researchers from machine learning, statistics, and medicine.

Machine learning has traditionally been focused on prediction. Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. Statistics, in contrast, has traditionally focused on data modeling, i.e., on the estimation of a probability law that has generated the data.

During recent years, the boundaries between the two disciplines have become blurred and both communities have adopted methods from the other. However, it is probably fair to say that neither of them has yet fully embraced the field of causal modeling, i.e. the detection of causal structure underlying the data. This has different reasons.

Dagstuhl Seminar Proceedings 09401
Machine learning approaches to statistical dependences and causality
http://drops.dagstuhl.de/opus/volltexte/2010/2363

Many statisticians would still shun away from developing and discussing formal methods for inferring causal structure, other than through experimentation, as they would traditionally think of such questions as being outside statistical science and internal to any science where statistics is applied. Researchers in machine learning, on the other hand, have too long focused on a limited set of problems neglecting the mechanisms underlying the generation of the data, including issues like stochastic dependence and hypothesis testing – tools that are crucial to current methods for causal discovery.

Since the Eighties there has been a community of researchers from statistics, computer science, and philosophy, who in spite of the pertaining views described above have developed methods aiming at inferring causal relationships from observational data, building on the pioneering work of Glymour, Scheines, Spirtes, and Pearl. While this community has remained relatively small, it has recently been complemented by a number of researchers from machine learning. This introduces a new viewpoint on the issues at hand, as well as a new set of tools, such as nonlinear methods for testing statistical dependencies using reproducing kernel Hilbert spaces, and modern methods for independent component analysis.

The goal of the seminar was to discuss future strategies of causal learning, as well as the development of methods supporting existing causal inference algorithms, including recent developments lying on the border between machine learning and statistics such as novel tests for conditional statistical dependences.

The Seminar was divided into two blocks, where the main block was devoted to discussing state of the art and recent results in the field. The second block consisted of several parallel brainstorming sessions exploring potential future directions in the field. The main block contained 23 talks whose lengths varied between 1.5 hours and 10 minutes (depending on whether they were meant to be tutorials or more specific contributions).

Several groups presented recent approaches to causal discovery from non-interventional statistical data that significantly improve on state of the art methods. Some of them allow for better analysis of hidden common causes, others benefit from using methods from other branches of machine learning such as regression techniques, new independence tests, and independent component analysis. Scientists from medicine and brain research reported successful applications of causal inference methods in their fields as well as challenges for the future.

In the brainstorming sessions, the main questions were, among others, (1) formalizing causality (2) justifying concepts of simplicity in novel causal inference methods, (3) conditional independence testing for continuous domains.

Regarding (1), the question of an appropriate language for causality was crucial and involved generalizations of the standard DAG-based concept to chain-graphs, for instance. The session on item (2) addressed an important difference between causal learning to most of the other machine learning problems: Occam's Razor type arguments usually rely on the fact that simple hypotheses may perform better than complex ones even if the "real world" is complex because it prevents overfitting when only limited amount of data is present. The problem of causal learning, however, even remains in the infinite sample limit.

The discussion on conditional independence testing (3) focused on improving recent kernel-based methods.

Keywords: Causality, statistical dependences, machine learning

Joint work of: Janzing, Dominik; Lauritzen, Steffen; Schölkopf, Bernhard

No correlation without causation

Nihat Ay (MPI für Mathematik in den Naturwissenschaften, DE)

Mutual information, well-known within information theory, quantifies the extent to which two random variables are stochastically dependent. It is defined as the relative entropy distance of the joint distribution of these variables from the product of the corresponding marginals. The natural extension of this quantity to more than two variables, referred to as multi-information, is straightforward. The talk aims at relating this associational measure to the underlying cause-effect relations in terms of two results. The first result states that the multi-information of observed variables is upper bounded by the sum of causal information flows through essential channels in the underlying network, including the unobserved variables. The second result of the talk is more structural in nature. It relates the entropy of common roots, generalizations of common causes, to a slightly modified version of multi-information. This result allows to infer higher order common roots of observed variables based on a correspondingly high stochastic dependence.

Keywords: Common cause principle, mutual information, multi-information, causality, information flows

Joint work of: Ay, Nihat; Steudel, Bastian

See also: N. Ay. A Refinement of the Common Cause Principle. Discrete Applied Mathematics 157 (2009) 2439-2457.

A brief overview of causal and decisional states and their applications

Nicolas Brodu (LTSI Université de Rennes, FR)

This article introduces the decisional states of system, and provides a practical algorithm for computing them. The decisional states are defined as the internal states of a system that lead to the same decision, based on a user-provided utility or cost function. The transitions between these decisional states correspond to events that lead to a change of decision. The utility function encodes some a priori knowledge external to the system. The decisional states then take in account both the intrinsic underlying structure of the system (the epsilon-machine) and that external information. An algorithm is provided so as to estimate the states and their transitions from data. Application examples are given for edge-emitting hidden Markov model reconstruction, cellular automata filtering, and edge detection in images.

Keywords: Epsilon-machines, decisional states

Instrumental variables for causal inference in epidemiology

Vanessa Didelez (University of Bristol, GB)

In epidemiology we often want to estimate the causal effect of an exposure on a health outcome based on observational data, where the possibility of unobserved confounding cannot be excluded. To deal with this problem, it has recently become popular to use a technique called Mendelian randomization, where it is exploited that the exposure is associated with a genetic variant, which can be assumed to be unaffected by the same confounding factors and which makes it suitable as a so-called instrumental variable. In my talk, this technique is illustrated with various examples, in particular with the effect of alcohol consumption on blood pressure / hypertension. Different methods of using an instrumental variable to estimate the causal effect on a binary outcome are compared based on their theoretical properties as well as by simulation.

Keywords: Graphical models

Full Paper:
http://www.maths.bris.ac.uk/~maxvd/smmr_mendel_print.pdf

Nonparametric tests of independence and conditional independence with positive definite kernels

Kenji Fukumizu (Institute of Statistical Mathematics - Tokyo, JP)

A kernel nonparametric methodology for independence and conditional independence tests is presented. The methodology uses positive definite kernels or reproducing kernel Hilbert spaces (RKHS) for capturing nonlinearity or higher order moments of random variables. The main properties are (i) it uses high-dimensional feature spaces (RKHS) while keeping feasible computational costs; (ii) any types of random variables including continuous, discrete and structured variables can be handled seamlessly once positive definite kernels are defined. The latter property is in particular desirable for causal learning when various types of variables are seen in given data.

The covariance of the feature vectors mapped into the RKHS's can characterize the independence of the original random variables on the assumption that the RKHS's are rich enough to represent all the moments of variables. This gives a

method of nonparametric test of independence. The Hilbert-Schmidt norm of the covariance operator gives the test statistic, for which the asymptotic properties can be derived.

The conditional covariance or partial correlation defined on the RKHS's can characterize the conditional independence, which derives a method of conditional independence tests. This can be applied for the constraint approach to causal leaning, which involves conditional independence tests with various types of variables. In conditional independence test, it is not straightforward how to obtain the null distribution for continuous variables. The current state of the art and open problems regarding to the approximation of the null distribution of the conditional independence tests are also discussed.

Keywords: Independence, conditional independence, test, positive definite kernel, reproducing kernel, nonparametric

Causal influence of gamma oscillations on sensorimotor rhythms during motor imagery

Moritz Grosse-Wentrup (MPI für biologische Kybernetik - Tübingen, DE)

Evidence is presented that during motor imagery distributed gamma oscillations exert a causal influence on sensorimotor-rhythms. The significance of this observation for research on brain-computer interfaces based on motor imagery is discussed.

Keywords: Gamma oscillations, causal inference, brain-computer interfaces

The bumpy road of the search for a (good) cause

Isabelle Guyon (ClopiNet - Berkeley, US)

The notion of causality is present in our everyday life and is pervasive in science and engineering. It allows us to infer what affects our health, the economy, climate changes, etc. and devise proper actions to obtain beneficial changes or avoid undesired outcomes. Yet there is no consensus on how to approach causal modeling and not even a proper mathematical definition of causality encompassing all the notions implied in science, history, philosophy, law, psychology, history, religion, and engineering. We adopt an engineering point of view in which causality is evidenced by the interventions of an external agent on a self contained system. The agent can learn the underlying causal structure of the system by observing it and devising proper experiments. Classical experimental design has had its successes; we review some of them: Vitamin C and scurvy; Hygene and infectious diseases; Planned experiment in agriculture; Smoking and Lung cancer; Randomized controlled trials to test the efficiency and toxicity of new

drugs. Machine learning may have something to add to improve experimental design by involving the learning machine in the data collection through an "active learning" process, and making better use of observational data. Indeed, observational data is often available in abundance at low cost while experiments are usually expensive and may be unethical or impossible to perform. Preliminary results in genetic epidemiology, system biology, economy, and social sciences are encouraging. To evaluate new algorithms, we started a program of data exchange and benchmarks called the "Causality Workbench". We have already organized two challenges and are creating a virtual laboratory to allow researchers to test experimental design strategies on artificial systems.

Keywords: Causality, experiments, policy, time series

Joint work of: Guyon, Isabelle; Janzing, Dominik; Schölkopf, Bernhard

Full Paper:
http://clopinet.com/causality

Causal inference and identifying confounder with second order statistics

Stefan Harmeling (MPI für biologische Kybernetik - Tübingen, DE)

Causal inference and confounder identification in linear additive models has been shown to be possible if the additive noise is non-Gaussian (see LiNGAM from Shimizu et al 2006). This assumption allows to apply well-developed algorithms from independent component analysis (ICA) to the problem. However, as Cardoso (see "The three easy routes to ICA", 2001) pointed out, non-Gaussianity is only one possibility to depart from the common assumption that data is (i) Gaussian, (ii) independent, (iii) identically distributed. ICA can also be performed if the noise is non-flat, dropping (ii), or non-stationary, dropping (iii).

My talk demonstrates how to infer causal directions and possible confounder in the case of non-stationary and/or non-flat data solely by considering second-order statistics, i.e. covariance matrices. We introduce the Bloch disk (being a non-complex special case of the Bloch sphere from quantum mechanics) as a representation which gives insight into the problem and also suggests algorithms we are going to pursue for identifying confounder.

Joint work of: Harmeling, Stefan; Janzing, Dominik; Schölkopf, Bernhard

Causal discovery inside the d-separation equivalence class

Patrik Hoyer (University of Helsinki, FI)

The discovery of causal relationships from non-experimental data is considered. Such 'causal discovery' relies on well-specified assumptions to allow one to infer causation from correlations observed in the joint distribution over the measured variables. In particular, a well-known result is that given causal Markov, causal faithfulness, a restriction to directed acyclic graphs, and an absence of confounding hidden variables, the joint distribution identifies the d-separation-equivalence class to which the true model belongs, and without further assumptions it is impossible to distinguish members within the equivalence class. Recently, a number of parametric and semi-parametric models have been proposed to be able to fully identify the underlying model. I will discuss some of these models, in particular focusing on linear and non-linear models with independent additive residuals. In addition to describing these models and methods, I hope to raise a discussion on when one is justified in making these stronger assumptions, so as to be able to distinguish models which are equivalent in terms of the independences they represent.

Causal inference using the algorithmic Markov condition

Dominik Janzing (MPI für biologische Kybernetik - Tübingen, DE)

Inferring the causal structure that links n observables is usually based upon detecting statistical dependences and choosing simple graphs making the joint distribution Markovian. Here we argue why causal inference is also possible when only single observations are present.

We state that similarities between two objects x and y indicate a causal link whenever their algorithmic mutual information is sufficiently high. It is defined as the number of bits that can be saved when optimally compressing the pair (x,y) jointly compared to compressing them independently. To infer causal graphs among n objects, we replace the notion of conditional *stochastic* independence in the causal Markov condition with the one of conditional *algorithmic* mutual information and describe the corresponding causal inference rules.

In contrast to causal inference methods that rely on statistical dependences, our theory implies rules for distinguishing between the causal hypotheses $X \rightarrow Y$ and $Y \rightarrow X$ for two random variables X,Y. This is because the causal graphs relating the individual observations of the statistical ensemble induce different sets of algorithmic independences for the two cases. Therefore, our theory provides a foundation for a new type of methods for inferring causal structure from statistical data.

Joint work of: Janzing, Dominik; Schölkopf, Bernhard

Full Paper:
http://aps.arxiv.org/abs/0804.3678

Chain graphs for causal inference

Steffen Lauritzen (University of Oxford, GB)

Most causal theory based on graphical models takes a DAG or possibly a directed graph with cycles as base. This lecture describes chain graphs which are graphs with directed and undirected links and their Markov and causal interpretation. Based on the idea that conditional potentials are stable under intervention, an intervention calculus completely analogous to that for DAGs is described. For undirected graphs, which are special cases of chain graphs, this leads to equality between intervention and observation conditioning.

Causal structure learning and inductive inference based on Kolmogorov complexity

Jan Lemeire (Vrije Universiteit Brussel, BE)

These are the results of an analysis of current causal structure learning with the general principles for inductive inference based on Kolmogorov complexity. The principle is that patterns in data do not appear by accident, but reveal some of the structure of the system under study. In causal terms: patterns point to mechanisms. First I will to put forward a new condition, namely the Independence of Conditionals (IC). Causal structure learning through Bayesian networks comes to the mapping of the Conditional Probability Distributions (CPDs) of a factorization to the mechanisms of the underlying system. Causal mechanisms are modular components which can be changed independently. I will argue that, besides local minimality, the fundamental condition for drawing such causal conclusions is the algorithmic independence of the CPDs, which we called the IC condition. Violation of IC gives indications about non-global minimality of the model or the presence of meta-mechanisms. When IC holds for a factorization, the factorization gives you the top-ranked causal hypothesis. It is the hypothesis that should be considered first. But no guarantee of its correctness can be given, since the factorization only describes the behavior of the system under study and the system could be more complex then its behavior suggests. I will argue that IC is more fundamental than faithfulness. First, because the IC condition provides inference rules beyond independence-based learning algorithms. The condition applies for any decomposition, like for example additive noise models in which the CPDs are further decomposed. Secondly, some violations of faithfulness, such as deterministic relations, do not violate IC and should not hinder causal structure learning. Nonetheless, I believe that faithfulness describes an important property of a model. Generally speaking, faithfulness says that the qualitative properties of a system are described by the qualitative part of the model. And this was Pearl's clear intention when creating the Bayesian network approach.

Keywords: Causal structure learning, Kolmogorov complexity, faithfulness

Iterative causal discovery with observational and experimental data

Philippe Leray (Université de Nantes, FR)

Learning causal Bayesian network structure with observational (discrete) data generally does not allow to find a full causally oriented structure. Experimental data is thus needed to discover the full causal graph. We present in [1] a general framework that aims to propose which experiment could be the most interesting to perform, taking into account various costs or that some experiments are not feasible.

We then try to soften some usual assumptions. What happens in real datasets, far away from the large sample limit, when observational data is not sufficient to obtain the right representant of the Markov equivalence class [2]? What happens outside the causal sufficiency framework? In [3, 4, 5, 6], we propose a unifying framework, proposing experiments in order to transform an initial maximum ancestral graph, obtained with observational data, into a semi-Markovian causal model in which we could perform both probabilistic and causal inference.

In a lot of real applications, the number of possible experiments is limited. We then study in [7] how to take into account some additional knowledge such as domain ontology in order to best choose what experiment to perform in order to discover the more interesting causal relationship.

See also: [1] S. Meganck, P. Leray, and B. Manderick, “Learning causal bayesian networks from observations and experiments: A decision theoritic approach,” in Proceedings of the Third International Conference, MDAI 2006, vol. 3885 of Lecture Notes in Artificial Intelligence, (Tarragona, Spain), pp. 58–69, Springer, 2006.

[2] S. Meganck, P. Leray, and B. Manderick,“Uncado: Unsure causal discovery,” in Proceedings of 4mes journes francophones de rseaux baysiens JFRB 2008, (Lyon, France), pp. 94–104, 2008.

[3] S. Meganck, S. Maes, P. Leray, and B. Manderick, “Learning semi-markovian causal models using experiments,” in The third European Workshop on Probabilistic Graphical Models PGM’06, (Prague, Czech Republic), pp. 195–206, 2006.

[4] S. Meganck, P. Leray, and B. Manderick, “Causal graphical models with latent variables: Learning and inference,” in Ninth European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty ECSQARU 2007, pp. 5–16, 2007.

[5] S. Maes, S. Meganck, and P. Leray, “An integral approach to causal inference with latent variables,” in Causality and Probability in the Sciences (F. Russo and J. Williamson, eds.), pp. 17–41, Texts In Philosophy series, London College Publications, 2007.

[6] P. Leray, S. Meganck, S. Maes, and B. Manderick, “Causal graphical models with latent variables: learning and inference,” in Innovations in Bayesian Networks: Theory and Applications (D. E. Holmes and L. Jain, eds.), (Germany, 33 pages), Springer, 2008.

[7] M. Ben Messaoud, P. Leray, and N. Ben Amor, "Integrating ontological knowledge for iterative causal discovery and vizualisation," in Proceedings of the 10th European Conference on Symbolic and Quantitative Approaches to Reasoning with Uncertainty (ECSQARU 2009), (Verona, Italy), pp. 168–179, 2009.

Additive noise models for causal inference

Joris Mooij (MPI für biologische Kybernetik - Tübingen, DE)

We consider a special class of causal models with continuous random variables where noise acts in an additive way. Each variable is assumed to be a function of its parent variables, plus noise, where all the noise terms are jointly independent. For the special case of only two variables, this model class is asymmetric in the sense that an additive noise model $X \rightarrow Y$ cannot be written as an additive noise model $Y \rightarrow X$ in the generic case. This asymmetry is exploited for causal inference by postulating that if the observational data is well described by an additive noise model, the DAG structure of the additive noise model likely coincides with the causal structure. We discuss how additive noise models can be learnt from data using regression by dependence minimization. We present preliminary but promising empirical results on real-world data sets consisting of several pairs of variables, showing that the additive noise assumption can indeed be useful for causal inference. We propose an efficient algorithm for learning additive noise models from data in the multivariate case. We also present some first results on the case where two observed variables have a hidden common cause.

Keywords: Causal inference, additive noise models, regression, causality

Joint work of: Mooij, Joris; Janzing, Dominik; Peters, Jonas; Schölkopf, Bernhard; Hoyer, Patrick

Current challenges in statistical analysis of complex psychiatric disease

Bertram Mueller-Myhsok (MPI für Psychiatrie - München, DE)

In this presentation I will give, based on examples from research at the Max-Planck-Institute of Psychiatry with data kindly provided by many colleagues there, an overview of current aspects to be unraveled in complex psychiatric disease. These include the combination of data across many, often disjunct, sources, the joint consideration of many polymorphisms in genetics and, finally, the impact of gene x environment interactions.

Keywords: Statistical genetics; psychiatric genetics; multiomics

Detecting the direction of causal time series

Jonas Peters (MPI für biologische Kybernetik - Tübingen, DE)

We propose a method that detects the true direction of time series, by fitting an autoregressive moving average model to the data. Whenever the noise is independent of the previous samples for one ordering of the observations, but dependent for the opposite ordering, we infer the former direction to be the true one. We prove that our method works in the population case as long as the noise of the process is not normally distributed (for the latter case, the direction is not identifiable). A new and important implication of our result is that it confirms a fundamental conjecture in causal reasoning — if after regression the noise is independent of signal for one direction and dependent for the other, then the former represents the true causal direction — in the case of time series.

We test our approach on two types of data: simulated data sets conforming to our modeling assumptions, and real world EEG time series. Our method makes a decision for a significant fraction of both data sets, and these decisions are mostly correct.

Keywords: Causality, additive noise models, arrow of time

Joint work of: Peters, Jonas; Janzing, Dominik; Gretton, Arthur; Schölkopf, Bernhard

Full Paper:
http://www.cs.mcgill.ca/∼icml2009/papers/503.pdf

Using causal Bayesian networks and information theory for the construction of minimal bio-inspired agent models

Daniel Polani (University of Hertfordshire, GB)

The central importance of (Shannon) information as a resource for the adaptation of living organisms or agents has been increasingly established in the last years. The information perspective for the characterization of an agent's operation is highly attractive since it provides a universal currency and language for any kind of "information processing" taking place both within the agent and in the dynamics of its interaction with the environment, and allows one to adopt an extreme bottom-up view. Furthermore, it is "coordinate-free" in the sense that it allows to formulate principles and "balance sheets" without having to refer to a particular information processing mechanism

The use of Causal Bayesian Networks (CBNs) has been established as a successful technique to create informational models of agents and their perception-action loop. This technique allows the tracking of information flows through the composite agent-environment system, the generalization of Ashby's Law of Requisite Variety, or the application of generalized Infomax principles. In particular,

the latter provide a path for the generation of structured information processing architectures with no assumptions beyond the agent being "embodied" in some structured environment. Phenomena such as active sensing emerge from the principle as a natural side effect.

The transparency of modeling the perception-action loop using the CBN formalism allows one to identify additional phenomena and quantities of interest. Specifically, in the present talk, I will introduce and discuss "empowerment" which is essentially the amount of potential information that an agent could inject into the environment via its actuators and recapture via its sensors. In the simplest of cases, this reduces to an agent-external channel capacity, but in general one requires CBNs to formulate empowerment.

In a situation where an agent has no prior preferences, its empowerment turns out to provide a utility which draws it to "interesting" states in the system. Since empowerment only depends on the embodiment of the agent, it can assign sensible preferences to states even in absence of any other prespecified drives (quantities of this kind we term "universal utility"). Understanding properties of possible universal utilities is particularly relevant for the success of adaptive systems, as the latter frequently have to be able to cope with novel situations that have not been previously encountered and for which the systems' innate drives are not appropriate or suitable drives may not exist yet at all.

I will show how, in a number of scenarios of varied quality and characteristics, the behavior resulting from empowerment optimization is strikingly close to our intuitive expectations, sometimes achieved in a surprisingly nontrivial way. In the discussion, I will suggest possible reasons for this and discuss lines for future research.

Keywords: Causal modeling, agents, information theory, empowerment, intrinsic motivation

Joint work of: Klyubin, Alexander; Polani, Daniel; Nehaniv, Chrystopher

Full Paper:
http://dx.doi.org/10.1371/journal.pone.0004018

See also: Klyubin, A. S., Polani, D., and Nehaniv, C. L., (2008). Keep Your Options Open: An Information-Based Driving Principle for Sensorimotor Systems PLoS ONE, 3(12):e4018.

Independence preserving graphs and their relationships

Kayvan Sadeghi (University of Oxford, GB)

In this talk we introduce three known classes of independence preserving graphs with three types of edge, called MC, summary and maximal ancestral graphs, that are closed under marginalization and conditioning and contain all DAG independence models. We also derive algorithms to generate these graphs from given DAG or from each other. We treat these algorithms as functions and find their properties.

We finally find the relationships between these functions and between different types of independence preserving graph.

Keywords: DAG, marginalization and conditioning, Markov equivalence, maximal ancestral graph, MC graph, separation criterion, summary graph.

Towards causal discovery with dormant independences

Ilya Shpitser (Harvard School of Public Health, US)

The edifice of graphical models research is built on understanding and taking advantage of conditional independence constraints. However, it was known for a long time that graphical models with latent variables can induce additional constraints on the data, sometimes termed Verma constraints. In this talk, I give a review of some recent work on a causal interpretation of Verma constraints as conditional independences of post-intervention distributions which are identifiable. This work also gives a graphical condition for such independences analogous to d-separation, shows that such independences can always be obtained by a simple operation on the joint distribution termed truncation or pseudo-intervention, and finally gives a graphical representation of independences in post-truncation distributions.

Keywords: Verma constraints, causal discovery, dormant independence

The recent history of graphical causal modeling

Peter Spirtes (Carnegie Mellon University - Pittsburgh, US)

This talk traced the history of recent developments in the graphical modeling of causal inference. Several things were needed for theories of graphical causal estimation to appear and to flower: well defined mathematical objects to represent causal relations; well defined connections between aspects of these objects and sample data; and a way to compute those connections. A sequence of studies beginning with Dempster's work on the factorization of probability distributions and culminating with Kiiveri and Speed's study of linear structural equation models provided the first, in the form of directed acyclic graphs, and the second, in the form of the local Markov condition. Pearl and his students, and independently, Stefan Lauritzen and his collaborators, provided the third, in the form of the global Markov condition, or d-separation in Pearl's formulation, and the assumption of its converse, which came to be known as stability or faithfulness. These developments set the stage for automated search algorithms for graphical causal models on large numbers of variables, including the PC algorithm

of Spirtes and Glymour. Besides the development of estimation or search algorithms, and proofs of their properties, a theory of search for causal explanations required a theory of interventions that would both justify the causal interpretation of directed graphical models and also provide a coherent normative theory of inference using causal premises. That effort can be traced back to Neyman, and Strotz and Wold, and then adaptations of these concepts to graphical causal models in the work of Spirtes, Glymour, & Scheines.

Keywords: Causal modeling, graphs, Bayesian networks

Justifying additive-noise-model based causal discovery via algorithmic information theory

Bastian Steudel (MPI für Mathematik in den Naturwissenschaften, DE)

A recent method for causal discovery is in many cases able to infer whether X causes Y or Y causes X for just two observed variables X and Y. It is based on the observation that there exist (non-Gaussian) joint distributions $P(X, Y)$ for which Y may be written as a function of X up to an additive noise term that is independent of X and no such model exists from Y to X. Whenever this is the case, one prefers the causal model $X \to Y$.

Here we justify this method by showing that the causal hypothesis $Y \to X$ is unlikely because it requires a specific tuning between $P(Y)$ and $P(X|Y)$ to generate a distribution that admits an additive noise model from X to Y. To quantify the amount of tuning required we derive lower bounds on the *algorithmic* information shared by $P(Y)$ and $P(X|Y)$. This way, our justification is consistent with recent approaches for using algorithmic information theory for causal reasoning.

We extend this principle to the case where $P(X, Y)$ *almost* admits an additive noise model. Our results show that additive-noise-model based causal inference gets more reliable if the complexity of $P(Y)$ is high.

Keywords: Algorithmic information theory, additive noise models

Joint work of: Janzing, Dominik; Steudel, Bastian

Full Paper:
http://arxiv.org/abs/0910.1691

Bayesian learning of causal Bayesian networks

Jin Tian (Iowa State University, US)

Bayesian networks are being widely used for probabilistic inference and causal modeling. For example, in causal discovery, we are interested in the causal relations among variables, represented by the edges in the network structure. In a Bayesian approach, we would compute the posterior probability of an edge given data by model averaging over all possible networks. However, the number of possible network structures is super-exponential in the number of variables, and therefore it is impractical to sum over all possible structures unless for tiny domains. In this talk I will present an algorithm that can compute the exact posterior probabilities of structural features in Bayesian networks in exponential time and demonstrate the applicability of the algorithm in Bayesian networks of a moderate size.

Keywords: Bayesian learning, Bayesian networks

Nonlinear directed acyclic structure learning with and without additive noise models

Robert E. Tillman (Carnegie Mellon University - Pittsburgh, US)

The recently proposed additive noise model has advantages over previous structure learning algorithms, when attempting to recover some true data generating mechanism, since it (i) does not assume linearity or Gaussianity and (ii) can recover a unique DAG rather than an equivalence class. However, its original extension to the multivariate case required enumerating all possible DAGs, and for some special distributions, e.g. linear Gaussian, the model is invertible and thus cannot be used for structure learning. We present a new approach which combines a PC style search using recent advances in kernel measures of conditional dependence with local searches for additive noise models in substructures of the equivalence class. This results in a more computationally efficient approach that is useful for arbitrary distributions even when additive noise models are invertible. Experiments with synthetic and real data show that this method is more accurate than previous methods when data are nonlinear and/or non-Gaussian.