EuroCIM Day 4 – 2 September – Schedule

hosted by BIPS, Bremen, Germany

Times are in pm CEST

4-4:10 (7am in Seattle) ***Welcome and Intro***

4:10-4:30
Maurice O'Connell (National University of Ireland Galway) "Pathway specific population attributable fractions"

4:30-4:50
Chiara Di Maria (University of Palermo) "Longitudinal mediation analysis with latent variables: a separable effect approach"

4:50-5:20
Lightning talks (4 talks of 4 minutes each + collective questions)
- Juha Karvanen (University of Jyvaskyla): "Causal effect identification from multiple incomplete data sources"
- Tyrel Stokes (McGill University): "Causal simulation experiments: How to intervene on a DAG"
- Caleb Miles (Columbia University): "Optimal tests of the composite null hypothesis arising in mediation analysis"
- Christian Gische (Humboldt University of Berlin): "Graph-based causal inference using parametric structural equation models"
- Chengchun Shi (London School of Economics): "Testing mediation effects using logic of Boolean matrices"

5:20-5:35 Break

5:35-5:55
Maximilian Ilse (University of Amsterdam) "Efficient causal inference from combined observational and interventional data through causal reductions"

5:55-6:15
Rohit Bhattacharya (Williams College) "Differentiable causal discovery under unmeasured confounding"

6:15-6:40
Lightning talks (4 talks of 4 minutes each + collective questions)
- Johannes Huegle (University of Potsdam): "Causal structure learning for heterogeneous data characteristics of real-world scenarios"
- Marco Doretti (University of Perugia): "On mediation analysis for a binary outcome and multiple binary mediators"
- Noam Finkelstein (Johns Hopkins University): "Bounds on non-restrictive latent variable cardinalities in Bayesian networks with discrete observed variables"
- Eleonora Iob (University College London): "Adverse childhood experiences, cortisol, and depressive symptoms in early adulthood: a causal mediation approach."

6:40-6:55 Break

6:55-7:35 (9:55am in Seattle)
Keynote talk: Thomas Richardson (University of Washington): "Single world intervention graphs"

7:35-8:00 Open discussion

Abstracts

Maurice O'Connell; John Ferguson

Pathway specific population attributable fractions

Population attributable fractions (PAF) represent the relative change in disease prevalence expected if an exposure was absent from the population. What percentage of this effect acts through particular pathways may be of interest, e.g. the effect of sedentary lifestyle on stroke may be mediated by blood pressure, BMI and several other mediators. Path-specific PAFs (PS-PAFs) represent the relative change in disease prevalence from an intervention that, conditional on observed covariates, shifts the distribution of the mediator to its expected distribution in a hypothetical population where the risk factor was eliminated. A related (more mechanistic) definition examines disease prevalence expected from an individual-level intervention assigning each individual the mediator they would have received if the risk factor had been eliminated.

Our aim here is not to decompose the total PAF for a risk factor into an additive sum over mediating pathways, but to instead fairly compare disease burden attributable to differing mediating pathways and as a result gain insights into the dominant mechanisms by which the risk factor affects disease on a population level. While PS-PAFs corresponding to differing pathways (mediating the same risk factor-outcome relationship) will each usually be less than the total PAF, they will often sum to more than the total PAF. In this manuscript, we present definitions, identifiability conditions and estimation approaches for PS-PAFs under various study designs. We illustrate results using INTERSTROKE, an international case-control study designed to quantify disease burden attributable to a number of known causal risk factors. A R package will be available.

Chiara Di Maria; Vanessa Didelez

Longitudinal mediation analysis with latent variables: a separable effect approach

In causal mediation analysis, the main goal is to address research questions about some direct effect of a treatment variable X on a response of interest Y, and indirect effect(s) conveyed by intermediate variables called mediators. It is well established that causal effects may need time to unfold, and this is the reason why longitudinal data are often more suitable to investigate mediational mechanisms. Estimating direct and indirect effects in a longitudinal setting is quite challenging due to the potential for time-varying (post-treatment) confounding, which hinders the identification of natural effects. Recently, so-called separable (in)direct effects have been introduced based on Robins and Richardson's work and were subsequently applied in longitudinal and time-to-event mediation contexts. In this work, we address longitudinal mediation analysis in a framework including latent variables, focusing on two approaches, generalized linear mixed models and latent growth models which are popular in the structural equations literature. We aim to clarify the meaning of structural parameters and the causal interpretation of mediational effects taking a separable effects perspective. We discuss the assumptions for the identifiability of separable effects and address issues with non-parametric (through the g-formula) and parametric identification. In the models considered, separable mediational effects are shown to be time-dependent, and we highlight how they differ between the two approaches. We run a simulation study to show how model misspecification can affect inference and illustrate this with an application to real data.

Juha Karvanen; Santtu Tikka; Antti Hyttinen

Causal effect identification from multiple incomplete data sources

Causal effect identification considers whether an interventional probability distribution can be uniquely determined without parametric assumptions from measured source distributions and structural knowledge on the generating system. While complete graphical criteria and procedures have been presented for many identification problems, there are still challenging but important extensions for which algorithmic solutions do not exist. To tackle these settings, we present a search algorithm directly over the rules of do-calculus. Due to the generality of do-calculus, the search is capable of taking advanced data-generating mechanisms into account along with an arbitrary type of both observational and experimental source distributions. The approach, called Do-search, is provably sound, and it is complete with respect to identifiability problems that have been shown to be completely characterized by do-calculus. When extended with additional rules, the search is capable of handling missing data problems as well. With the versatile search, we are able to approach new problems for which no other algorithmic solutions exist. The R package 'dosearch' provides an interface for a C++ implementation of the search.

Tyrel Stokes; Ian Shrier; Russell Steele

Causal simulation experiments: How to intervene on a DAG

In both theoretical and applied settings, simulation plays an important role in evaluating the properties of causal estimators. In particular, we might simulate to test the performance of competing estimators with respect to changes in underlying assumptions, such as say unmeasured confounding in a sensitivity analysis. In causal inference we have developed robust frameworks for articulating and answering causal questions about data experiments, but there has been less focus on how to apply those insights to simulation experiments.

In this talk we present extensions of our recently accepted work in Statistical Methods in Medical Research. Specifically, we use the challenge of simulating the effects of bias amplification to demonstrate the importance of thinking about simulation experiments as interventions. From the intervention perspective, one must properly control for the confounding factors in order to properly estimate the effect of intervention. For bias amplification we show this requires that the treatment variance be set to a degenerate distribution, which in turn implies constraints over the simulation interventions that can be performed. While this adds some extra burden on the simulation experimenter, it has the added benefit of clearly defining the boundaries of the simulation experiment and more easily generalizing to real data sets where observed quantities are fixed. To demonstrate this, we modify real experimental data in the spirit of a plasmode simulation. This allows us to compare differences in bias amplification of competing OLS estimators after one increases the causal effect of the observed variables (X) on the treatment (A) (i.e modifying the edge from X to A).

Caleb H. Miles; Antoine Chambaz

Optimal tests of the composite null hypothesis arising in mediation analysis

The indirect effect of an exposure on an outcome through an intermediate variable can be identified by a product of regression coefficients under certain causal and regression modeling assumptions. Thus, the null hypothesis of no indirect effect is a composite null hypothesis, as the null holds if either regression coefficient is zero. A consequence is that standard hypothesis tests of mediation are either severely underpowered near the origin (i.e., both coefficients are small with respect to standard errors) or invalid. We propose hypothesis tests that (i) preserve uniform level alpha type 1 error, (ii)

meaningfully improve power when both true underlying effects are small relative to sample size, and (iii) preserve power when at least one is not. One approach uses sparse linear programming to produce an approximately optimal test for a Bayes risk criterion. Another gives a closed-form test that is minimax optimal with respect to local power over the alternative parameter space.

Christian Gische; Manuel C. Voelkle

Graph-based causal inference using parametric structural equation models

Graph-based causal models are a flexible tool for causal inference from observational data. In this paper, we develop a comprehensive framework to define, identify, and estimate a broad class of causal quantities in linearly parametrized graph-based models. We link graph-based causal quantities defined via the do-operator to parameters of the model implied distribution of the observed variables using so-called causal effect functions. We use covariance structure models for the statistical modeling of the joint distribution of observed variables. The latter models can be estimated using statistical techniques developed in the traditional literature on structural equation models. We use causal effect functions to construct estimators for causal quantities and show that these estimators are consistent and converge at a rate of N(-1/2) under standard assumptions. Thus, causal quantities can be estimated based on sample sizes that are typically available in the social and behavioral sciences. In case of maximum likelihood estimation, the estimators are asymptotically efficient. We illustrate the proposed method using simulated data based on a prior empirical study. We provide an outlook on our ongoing research where we extend the proposed method to deal with measurement error, nonlinear structural equations, and effect heterogeneity.

Chengchun Shi; Lexin Li

Testing mediation effects using logic of Boolean matrices

Mediation analysis is becoming an increasingly important tool in scientific studies. A central question in high-dimensional mediation analysis is to infer the significance of individual mediators. The main challenge is the sheer number of possible paths that go through all combinations of mediators. Most existing mediation inference solutions either explicitly impose that the mediators are conditionally independent given the exposure, or ignore any potential directed paths among the mediators. In this work, we propose a novel hypothesis testing procedure to evaluate individual mediation effects, while taking into account potential interactions among the mediators. Our proposal thus fills a crucial gap, and greatly extends the scope of existing mediation tests. Our key idea is to construct the test statistic using the logic of Boolean matrices, which enables us to establish the proper limiting distribution under the null hypothesis. We further employ screening, data splitting, and decorrelated estimation to reduce the bias and increase the power of the test. We show our test can control both the size and false discovery rate asymptotically, and the power of the test approaches one, meanwhile allowing the number of mediators to diverge to infinity with the sample size. We demonstrate the efficacy of our method through both simulations and a neuroimaging study of Alzheimer's disease.

Maximilian Ilse; Patrick Forré; Max Welling; Joris M. Mooij

Efficient causal inference from combined observational and interventional data through causal reductions

Unobserved confounding is one of the main challenges when estimating causal effects. We propose a novel causal reduction method that replaces an arbitrary number of possibly high-dimensional latent

confounders with a single latent confounder that lives in the same space as the treatment variable without changing the observational and interventional distributions entailed by the causal model. After the reduction, we parameterize the reduced causal model using a flexible class of transformations, so-called normalizing flows. We propose a learning algorithm to estimate the parameterized reduced model jointly from observational and interventional data. This allows us to estimate the causal effect in a principled way from combined data. We perform a series of experiments on data simulated using nonlinear causal mechanisms and find that we can often substantially reduce the number of interventional samples when adding observational training samples without sacrificing accuracy. Thus, adding observational data may help to more accurately estimate causal effects even in the presence of unobserved confounders.

Rohit Bhattacharya; Tushar Nagarajan; Daniel Malinsky; Ilya Shpitser

Differentiable causal discovery under unmeasured confounding

The data drawn from biological, economic, and social systems are often confounded due to the presence of unmeasured variables. In such scenarios, statistical and causal models of a directed acyclic graph (DAG) over the observed variables do not faithfully capture the underlying causal process. Prior work on causal discovery in the presence of latent confounders has focused on search procedures for selecting acyclic directed mixed graphs (ADMGs), specifically ancestral ADMGs, that encode ordinary conditional independence constraints among the observed variables of the system. However, confounded systems also exhibit more general equality restrictions that cannot be represented via these graphs, placing a limit on the kinds of causal processes that can be learned using ancestral ADMGs. In this work, we derive differentiable algebraic constraints that fully characterize the space of ancestral ADMGs, as well as more general classes of ADMGs, arid ADMGs and bow-free ADMGs, that capture all equality restrictions on the observed variables. We use these constraints to cast causal discovery as a continuous optimization problem and design differentiable procedures to find the best fitting ADMG when the data come from a confounded linear system of equations with correlated errors. We demonstrate the efficacy of our method through simulations and application to a protein expression dataset. We end with a short discussion on open problems and challenges.

Johannes Huegle; Christopher Hagedorn

Causal structure learning for heterogeneous data characteristics of real-world scenarios

Causal Structure Learning (CSL) from observational data has received widespread attention in practice as the knowledge of underlying causal relationships is the basis for decision support within many real-world scenarios. For example, in discrete manufacturing, the knowledge about causal relationships is the key for a root cause analysis of failures within the complex production processes. While constraint-based methods for CSL are popular, they require variables of the same type, continuous or discrete, or have strong statistical assumptions on the underlying Functional Causal Model (FCM). In contrast, most real-world scenarios incorporate diverse heterogeneous data characteristics that include non-linear or mixed discrete-continuous causal relationships where the true statistical properties are mostly unknown in advance.

In our work, we present an information-theoretic perspective on the examination of probabilistic conditional independence (CI) within heterogeneous data to allow for weaker assumptions on the underlying FCM. In particular, we propose a non-parametric CI-test based on a kNN-based conditional mutual information estimation that captures non-linear and mixed discrete-continuous causal relationships. On this information-theoretic basis, we demonstrate that incorporation into

constraint-based methods such as the popular PC-Algorithm enables an accurate examination of causal structures in the presence of heterogeneous data within synthetic, and real-world scenarios.

In this talk, we present challenges of CSL within real-world scenarios, e.g., from industrial manufacturing, and demonstrate how an information perspective improves the accuracy and interpretability of CSL in practice. Furthermore, improved accuracy is reflected within various synthetic scenarios on mixed discrete-continuous and non-linear data, too.

Marco Doretti; Elena Stanghellini; Martina Raggi; Paolo Berta

On mediation analysis for a binary outcome and multiple binary mediators

Given a series of univariate logistic regressions, recent contributions exist on the relationship between the marginal effect and the conditional ones (Stanghellini and Doretti, 2019). In a mere associational context, these results have been used to disentangle the total effect of a treatment X on a response variable Y into a direct one and a number of indirect ones, each of them attributable to binary mediators along the pathway between X and Y. With reference to a single mediator, a bridge between the associational context and the causal framework has also been made (Doretti et al. 2021), thereby deriving a parametric decomposition of the total effect into natural causal effects.

In this work, we aim at extending this parallel to a setting with multiple binary mediators. Building from the results obtained within the path analysis framework (Raggi et al., 2021), we investigate the parametric identification of natural effects in the counterfactual framework. This includes some interesting situations like the identification of the natural direct effect in the absence of the cross-world independence assumption, that is, when there is a mediator-outcome confounder affected by the exposure.

Noam Finkelstein; Elie Wolfe; Ilya Shpitser

Bounds on non-restrictive latent variable cardinalities in Bayesian networks with discrete observed variables

We resolve the long-standing question of whether latent variables can be assumed to have finite state spaces in arbitrary latent variable Bayesian networks when observed variables are discrete without restricting the marginal model of the network, and provide upper bounds on these non-restrictive cardinalities. We further demonstrate that it is always possible to assume, without loss of generality, that the observed variables are deterministic functions of the latent variables, and provide upper bounds on the cardinalities of latent variables required to make this stronger assumption. Such deterministic relationships make up a so-called "functional model" of the latent variable Bayesian network.

A corollary to these results is that the causal compatibility problem -- the problem of determining whether a distribution over observed variables is in the marginal model of a latent variable Bayesian network -- can be expressed as a polynomial satisfiability problem when observed variables are discrete, such that the distribution is in the model if and only if a polynomial system has a solution. The cardinality bound approach and the functional model approach each suggests a different polynomial system.

Because polynomial satisfiability is exponentially slow, it is important to simplify the system as much as possible. We provide graphical conditions under which each of the two approaches yields the simpler system, and a strategy for choosing functional models that yield simple systems when more

than one is available. Finally, we explore several additional strategies for further simplifying the system.

Eleonora Iob; Jessie R. Baldwin; Robert Plomin; Andrew Steptoe

Adverse childhood experiences, cortisol, and depressive symptoms in early adulthood: a causal mediation approach

Dysregulated hypothalamic-pituitary-adrenal (HPA)-axis function might underlie the relationship between adverse childhood experiences (ACEs) and depression. However, limited research has examined the possible mediating role of the HPA-axis among young people using longitudinal data. Moreover, it remains unclear whether genetic influences could contribute to these associations.

Participants were 290 children from the Twins Early Development Study. ACEs were assessed from age 3 to 11 years. We calculated a cumulative risk score and also assessed different ACEs dimensions (Abuse, Bullying, Separation/Divorce, and Dysfunctional Parenting). HPA-axis activity was indexed by daytime salivary cortisol measured at age 11. Depressive symptoms were ascertained using the Short Mood and Feelings Questionnaire at age 21. Genetic liability to altered cortisol levels and elevated depressive symptoms was measured using a twin-based method. We performed causal mediation analysis with mixed-effects regression models.

The results showed that ACEs cumulative exposure (b=-0.20,p=0.03) and bullying (b=-0.61,p=0.01) were associated with lower cortisol levels at age 11. Among participants exposed to multiple ACEs, lower cortisol at age 11 was related to higher depressive symptoms at age 21 (b=-0.56,p=0.05). Lower cortisol levels mediated around 10-20% of the total associations of ACEs cumulative exposure and bullying with higher depressive symptoms. These mediation effects were smaller when genetic confounding was accounted for (ACEs cumulative exposure:b=0.16[0.02,0.34]; bullying:b=0.18[0.01,0.43]).

In conclusion, ACEs were linked to elevated depressive symptoms in early adulthood partly through lower cortisol levels in early adolescence. ACEs and cortisol remained as risk factors for depression after genetic factors had been taken into account.

Thomas Richardson

Single World Intervention Graphs: A simple framework for unifying graphs and potential outcomes with applications to mediation analysis

Causal models based on potential outcomes, also known as counterfactuals, were introduced by Neyman (1923) and later popularized by Rubin. Causal Directed Acyclic Graphs (DAGs) are another approach, originally introduced by Wright (1921), but subsequently significantly generalized and extended by Spirtes and Pearl among others.

In this talk I will first present a simple approach to unifying these two approaches via Single-World Intervention Graphs (SWIGs). The SWIG encodes the counterfactual independences associated with a specific hypothetical intervention on a set of treatment variables. The nodes on the SWIG are the corresponding counterfactual random variables. This represents a counterfactual model originally introduced by Robins (1986) using event trees.

This synthesis permits a simplification of Pearl's do-calculus that clarifies and separates the underlying concepts. In turn this leads to a simple counterfactual formulation of a complete identification algorithm for causal effects in models with hidden variables.

By expanding the graph, SWIGs may also be used to describe a novel interventionist approach to mediation analysis whereby treatment is decomposed into multiple separable components. This provides a means of discussing direct effects without reference to cross-world (nested) counterfactuals or interventions on the mediator. The theory preserves the dictum ``no causation without manipulation'' and makes questions of mediation empirically testable in future randomized controlled trials.

This is joint work with James M. Robins (Harvard) and Ilya Shpitser (Johns Hopkins).