EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Wirtschafts- und
Sozialwissenschaftliche
Fakultät

LEHRSTUHL FÜR MARKETING
Prof. Dr. Dominik Papies

Universität Tübingen · LS für Marketing · Nauklerstr. 47 · 72074 Tübingen

Telefon +49 7071 29-76977
Telefax +49 7071 29-5078
dominik.papies@uni-tuebingen.de
www.uni-tuebingen.de/wiwi/marketing

Seminar "New developments in Machine Learning and Causal Inference"

I. Type of seminar

In this seminar, students will work on selected topics that involve modern tools for data analysis, e.g., from the domain of Machine Learning or Causal Inference, or at the intersection of these two.

The topics can be chosen either from the list of suggested topics, or students propose their own topics. In the latter case, the suitability of the topic will be discussed with the supervisors.

In this seminar, students will also acquire relevant tools to be prepared for writing a research-based master thesis. This will be supported by an obligatory workshop on academic research as well as an obligatory workshop on presentation skills, which includes a short presentation of each student's current state of the thesis ("research plan presentation"). On top of that, we expect and encourage active participation and interaction between students.

It is expected that students have **advanced or at least very solid skills in statistical software (preferably R or Python)**, equivalent to, e.g., a successful completion of DS400 Data Science Project Management and/or DS404 Data Science with Python. In addition, we expect that students are willing to **familiarize themselves** with new methods and approaches as well as new tools in R or Python. The respective supervisor will support students in this.

II. Topics and introductory reading material

Topic 1 **How useful are causal discovery methods to business and economics research?**

Research in business and economics is often concerned with the estimation of causal effects. To estimate these causal effects, researchers need to come up with a plausible causal model. This model is typically derived and justified from theory and domain knowledge with appropriate assumptions and can be summarized in a causal graph. However, in recent years, computer scientists have developed methods under the terms “causal discovery” or “causal structure learning” that try to learn (parts of) such a causal model (or causal graph) from data instead of theory/domain knowledge. While there has been a lot of progress on the methodological side, it is still unclear how feasible and useful most of these methods are in real applications in business and economics.

Students therefore should review the literature on causal discovery, before focusing on one method in more detail. They will evaluate this method in simulations and apply it to a real dataset from business and economics to demonstrate its use.

Literature Peters, J., Janzing, D., & Schölkopf, B. (2017). Elements of Causal Inference: Foundations and Learning Algorithms. MIT Press.

Glymour, C., Zhang, K., & Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. Frontiers in Genetics, 10, 524. https://doi.org/10.3389/fgene.2019.00524

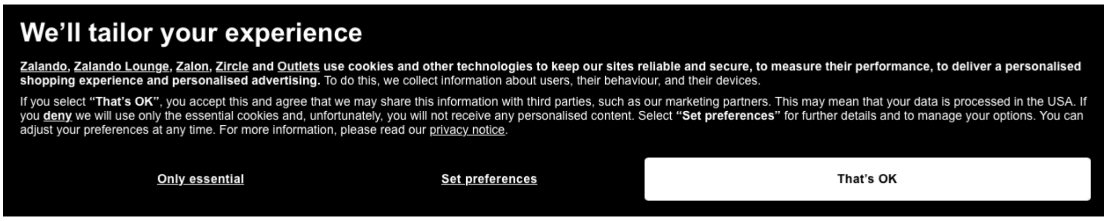
Heinze-Deml, C., Maathuis, M. H., & Meinshausen, N. (2018). Causal Structure Learning. Annual Review of Statistics and Its Application, 5(1), 371–391. https://doi.org/10.1146/annurev-statistics-031017-100630

Burauel, P. (2022). Evaluating Instrument Validity using the Principle of Independent Mechanisms (SSRN Scholarly Paper 3344981). https://doi.org/10.2139/ssrn.3344981

Data Own simulations & applications with openly available data from related literature

Topic 2	**How different disciplines use machine learning for causal inference**
	Answering causal questions is at the core of many scientific disciplines (What is the effect of education on wages? What is the effect of a drug on health outcomes? Etc.). Since interdisciplinary communication is often not very pronounced, these disciplines have tended to develop methodologies almost independently from one another. With the recent popularity of machine learning techniques, researchers from both econometrics and biostatistics have developed methods that try to use these techniques to answer causal questions. Two of the "flagship" methods include "targeted maximum likelihood estimation" (TMLE, originating in biostatistics) and "double/debiased machine learning" (DML, originating in econometrics). The goal of this thesis is to compare these two approaches and to evaluate their relative strengths and weaknesses. Students should synthesize the terminology across the fields as well as develop and communicate the intuitions behind the two methods. The performance of both methods should be assessed on simulated data. Finally, students apply the methods to a classical question in business/economics and compare the respective estimates.
Literature	Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, *21*(1), Article 1. https://doi.org/10.1111/ectj.12097 Laan, M. J. van der, & Rubin, D. (2006). Targeted Maximum Likelihood Learning. *The International Journal of Biostatistics*, *2*(1). https://doi.org/10.2202/1557-4679.1043 Díaz, I. (2020). Machine learning in the estimation of causal effects: Targeted minimum loss-based estimation and double/debiased machine learning. *Biostatistics*, *21*(2), 353–358. https://doi.org/10.1093/biostatistics/kxz042
Data	Own simulations & applications with openly available data from related literature

Topic 3 Large Scale Analysis of Nudging Practices in Online Cookie Consent Management



We'll tailor your experience

Zalando, Zalando Lounge, Zalon, Zircle and Outlets use cookies and other technologies to keep our sites reliable and secure, to measure their performance, to deliver a personalised shopping experience and personalised advertising. To do this, we collect information about users, their behaviour, and their devices.

If you select **"That's OK"**, you accept this and agree that we may share this information with third parties, such as our marketing partners. This may mean that your data is processed in the USA. If you **deny** we will use only the essential cookies and, unfortunately, you will not receive any personalised content. Select **"Set preferences"** for further details and to manage your options. You can adjust your preferences at any time. For more information, please read our privacy notice.

Cookie consent management systems for several websites could be criticized for not letting the user make a free choice about which cookies to accept. This nudging practice of websites is often manifested in a highlighted "OK" button compared to opting out in smaller underemphasized font or hidden under a second layer of settings. The goal of this thesis is to automatically infer such nudging practices from website screenshots. Students will investigate the prospects of exploiting deep learning methods to extract information relevant to cookie consent management from website screenshot images. Furthermore, students will employ this automated approach to conduct a large-scale analysis of hundreds of popular websites and infer insights about patterns in nudging behavior of online cookie management systems. *Do e-commerce websites employ more aggressive nudging policy compared to news websites? Are websites of businesses headquartered in the EU less likely to nudge users into accepting the use of all cookies?*

Literature

T Gogar, O Hubacek, J Sedivy (2016). Deep Neural Networks for Web Page Information Extraction

A Kumar, K Morabia, W Wang, K Chang, A Schwing (2022). CoVA: Context-aware Visual Attention for Webpage Information Extraction.

Data Dataset to be developed in the project.

Topic 4 **Sensitivity analysis in causal inference**

Causal inference from observational data always relies on untestable assumptions. In many applications, these assumptions might be questionable. Therefore, one important concern is how robust causal estimates are to violations of crucial assumptions. Sensitivity analysis is a tool to determine how strong a violation would need to be to significantly change the research results. Various frameworks for sensitivity analysis are available, but rarely used. Cinelli & Hazlett (2020) propose a new tool for sensitivity analysis that works under weaker assumptions, is easy to implement, and delivers intuitively interpretable results.

In this thesis, students should study how researchers can use sensitivity analysis when estimating causal effects in business and economics. They compare recent methodological developments to traditional ways of doing sensitivity analysis. They review and discuss the methodology, before applying it to simulated as well as real data from business and economics. Advanced students can also engage with a recent extension of the framework to causal estimation with machine learning (Chernozhukov et al., 2022).

Literature Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *82*(1), 39–67. https://doi.org/10.1111/rssb.12348

Chernozhukov, V., Cinelli, C., Newey, W. K., Shamar, A., & Syrgkanis, V. (2022). Long Story Short: Omitted Variable Bias in Causal Machine Learning (Working paper)

Data Own simulations & applications with openly available data from related literature

Topic 5 **Bad controls in Marketing**

A common piece of advice in applied econometric analysis is to control for potential confounders, i.e., variables that may affect the focal regressor as well as the outcome. This approach may lead researchers to include a wide range of covariates. At the same time, controlling for so-called “bad control variables” may bias the results, i.e., it may be better not to control for some variables. This problem, however, is not well understood in applied econometric research and in marketing. It is therefore the goal of this project to summarize the current state of the literature on the topic of bad control variables, examine its relevance for marketing, and assess the severity of the problem through simulations.

Literature Cinelli, C., Forney, A., & Pearl, J. (2022). A Crash Course in Good and Bad Controls. *Sociological Methods & Research*, https://doi.org/10.1177/00491241221099552

Klarmann, M., & Feurer, S. (2018). Control Variables in Marketing Research. *Marketing ZFP*, *40*(2), 26–40. https://doi.org/10.15358/0344-1369-2018-2-26

Data Own simulations & applications with openly available data from related literature

Topic 6 **Understanding Consumer Aesthetic Preferences for Pre-owned Clothing Items**

Recently, the online market for pre-owned clothing items has been growing rapidly.

As these items are sold by non-professional sellers in a wide variety of conditions, the aesthetic features that are relevant for these items might be different from those that are relevant for new items sold by professional sellers. Therefore, with this project we would like to define novel aesthetic features especially relevant for pre-owned clothing items.

Furthermore, we would like to understand how these features are related to user preferences. An example of such a feature could be the presence of clothing tags, which is a feature that is not relevant for new clothing items. Another example could be the amount of clutter in the background while taking a picture of the item, which is a feature that is not relevant for professional sellers. Another example could be the presence of a person or other objects such as a hanger in the picture. Towards these goals, we would employ advanced deep learning models to extract these features from images at scale and then use them to predict consumer preferences. The consumer preferences will be collected through a user study.

Data Web scraping of public pre-owned clothing marketplaces like vinted.de

Topic 7 **Controllable Image Manipulation**

Recent advances in generative image models have enabled the synthesis of high-quality images. Controllable Attribute Manipulation refers to the ability to modify specific aspects of an image in a fine-grained manner, without changing other aspects of the image. For example, it may involve changing the color of a specific object in an image without changing the color of the entire image. While attribute manipulation methods in image editing have advanced significantly, they still come with certain limitations.

1. Limited Attribute Control: Most attribute manipulation methods focus on a specific set of attributes, such as changing the hairstyle or age of a person in an image. They may not support a wide range of attributes simultaneously, and controlling multiple attributes at once can be challenging.
2. Overfitting: Some attribute manipulation models are highly specialized and may overfit to the specific dataset they were trained on. This can lead to unrealistic or inconsistent results when applied to images outside the training distribution.
3. Lack of Fine-grained Control: Achieving fine-grained control over attributes can be challenging. Users might want to make subtle changes, but the model's adjustments can be too pronounced or not precisely aligned with the user's intentions.
4. Semantic Understanding: Many attribute manipulation methods lack a deep understanding of image semantics. For instance, changing the color of an object might not consider its real-world plausibility, leading to unnatural results.
5. Limited Resolution and Detail: Some methods struggle to preserve fine details when modifying attributes. This can result in loss of image quality or artifacts in the edited regions.
6. Dependency on Training Data: Attribute manipulation models are highly dependent on the quality and diversity of the training data. If the training data lacks certain attributes or has biases, the model may not perform well for those attributes.
7. Generalization to Complex Scenes: Attribute manipulation is often easier on simpler scenes and objects. Handling complex scenes with multiple objects and interactions can lead to errors and unrealistic results.
8. Limited Domain Transfer: Models trained on one domain (e.g., human faces) may not generalize well to other domains (e.g., fashion items). Domain-specific models are often needed for accurate attribute manipulation.

In this project, we will survey the recent advances in controllable attribute manipulation and discuss the challenges and limitations of existing methods. We will also explore the future directions for research in this area.

Data Several publicly available datasets

- **III. Dates**

October 11, 2023	Online Application via ILIAS
October 12, 2023	9:00 a.m s.t. – 1:00 p.m. - VG 002, Wilhelmstr. 19 Kick-off and topic assignment Workshop „Academic Writing"
November 10, 2023	9:00 a.m. s.t. – 1:00 p.m. - ÜR 02 Alte Physik Workshop "Presentation Skills"
November 27, 2023	All day - SR 236 NA Research plan presentation
December 21, 2023	Term paper is due by noon (12 p.m. s.t.) (You can drop your term paper in the letterbox outside the faculty (addressed to Chair of Marketing - Nauklerstr. 47) or send it by post (postmark date is relevant).) Containing 2 versions of the term paper with a filing clip (https://de.wikipedia.org/wiki/Heftstreifen) Submit the electronic version (pdf) of the term paper incl. analysis scripts as file upload in ILIAS.
January 12, 2024	All day (dates will be coordinated individually) Feedback Session
January 25, 2024	8:00 p.m. Upload Presentation in ILIAS
January 26, 2024 (tentative, subject to change)	All day Seminar - SR 236 NA

IV. Course credits

Students can obtain course credit (9 ECTS). To obtain course credit students must meet the following criteria:

- Students participate in all meetings listed above
- Students submit their 12-page thesis on time
- Students present their thesis during the seminar
- Students actively participate during the seminar

Approx. 50% of the final grade will be the thesis, and 50% of the final grade will be the presentation and the participation in the seminar.

Please note:
Topics are subject to change - Students are invited to propose their own topics that fit under the general theme of the seminar.

Tübingen, October 2023