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 **Machine learning and instrumental variables**

When trying to estimate causal effects in business and economics applications, researchers often face the problem of unobserved variables, which can lead to biased estimates. In traditional statistics, one important tool to tackle this problem is the instrumental variables approach. In recent years, an emerging literature has developed new tools to answer causal questions by applying methods from the field of machine learning. Some of these new frameworks can also be applied within the context of instrumental variables, with the goal of providing more flexible and robust estimates of the causal effects.
By studying one of these new methods in more detail, students should explore the benefits and drawbacks of using machine learning in an instrumental variables setting. The thesis will evaluate the method(s) in both simulations and applications to classical problems from business and economics.

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

Chen, J., Huang, C.-H., & Tien, J.-J. (2021). Debiased/Double Machine Learning for Instrumental Variable Quantile Regressions. *Econometrics*, *9*(2), Article 2.
https://doi.org/10.3390/econometrics9020015

(Hartford, J., Lewis, G., Leyton-Brown, K., & Taddy, M. (2017). Deep IV: A Flexible Approach for Counterfactual Prediction. *Proceedings of the 34th International Conference on Machine Learning*, 1414–1423.
https://proceedings.mlr.press/v70/hartford17a.html)

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 **Analysis of purchasing behavior and shopping baskets**

The personalization of marketing tools has long been a topic in marketing research and is becoming increasingly important also in practice. Personalized targeting is not only intended to make the purchasing process easier for customers and thus increase long-term customer loyalty, but is also an important factor for possible profit optimization from a firm perspective. Basis for personalized targeting is a detailed analysis of customers and their shopping behavior.
The aim of this thesis is therefore to generate valuable insights about the purchasing behavior of customers using the freely accessible transaction data set of "Instacart". This can include various aspects: from a segmentation of households, to product affinity analyses in shopping baskets, to a forecast of which products will end up in a customer's next purchase. All these topics can be tackled with deep learning and machine learning methods.

Literature

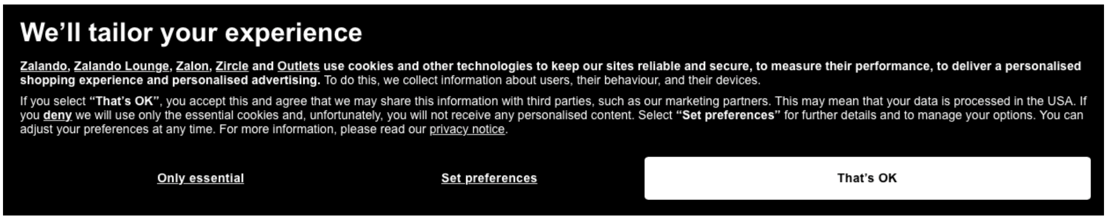
Coussement, K., Harrigan, P., & Benoit, D. F. (2015). Improving direct mail targeting through customer response modeling. *Expert Systems With Applications*, 42(22), 8403-8412. https://doi.org/10.1016/j.eswa.2015.06.054

Gabel, S., Guhl, D., & Klapper, Daniel (2019). P2V-Map: Mapping Market Structures for Large Retail Assortments. *Journal of Marketing Research*, 56(4), 557-580. https://doi.org/10.1177/0022243719833631

Reutterer, T., & Dan, D. (2021). Cluster Analysis in Marketing Research. In Handbook of Market Research (pp. 221-249). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-57413-4_11

Data Own simulations & https://www.kaggle.com/c/instacart-market-basket-analysis

Topic 4 **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 5 **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

- **III. Dates**

March 26, 2023	Online Application closes – please see our website for further information.
April 3, 2023	09:00 a.m. – 1:00 p.m. - SR 236 NA Kick-off and topic assignment Workshop „Academic Writing”
April 27, 2023	3:00 p.m. – 7:00 p.m. - SR 236 NA Workshop “Presentation Skills”
May 17, 2023	All day - SR 236 NA Research plan presentation
June 12, 2023	Term paper is due by noon (12pm) 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.
June 30, 2023	All day (dates will be coordinated individually) Feedback Session
July 13, 2023	8:00 pm Upload Presentation in ILIAS
July 14, 2023 (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

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, February 2023