Harvard University, spring 2019, Syllabus for: Economics 2148 - Topics in Econometrics

Advances in causality and foundations of machine learning

instructor	Maximilian Kasy
office	Littauer 121
office hours	send me an email to coordinate
email	teachingmaxkasy@gmail.com
class time	Tue & Thur, 10:30pm-11:45
location	Littauer M17
webpage	https://maxkasy.github.io/home/TopicsInEconometrics2019/

Overview and Objectives

Economics 2148 is one of the second-year econometrics field classes. We will begin the class with a survey of the literature on **identification using instrumental variables**, taking the linear model as a point of departure. The linear model imposes strong restrictions on the heterogeneity of causal effects. Generalizing this model to allow for nonlinear and heterogeneous effects leads to a variety of approaches discussed in the literature, including a re-interpretation of classic estimands as LATE, bounds on objects such as the ATE that are not point identified, conditional moment restrictions, and control function approaches.

The next part of class will cover some of the **theoretical foundations of machine learning**, including regularization and data-driven choice of tuning parameters. We will discuss in some detail the canonical normal means model. In this model, we will motivate shrinkage estimators in different ways, and will prove the famous result that shrinkage estimators can uniformly dominate conventional estimators. We will then move from normal means to function estimation using Gaussian process priors. We will show the equivalence of (empirical) Bayes estimation using such priors to penalized least squares regression with penalties corresponding to so-called reproducing kernel Hilbert space norms. This part of 2148 concludes with some applications of Gaussian process priors to experimental design and to optimal taxation.

After the spring break, we will cover some selected additional **topics in machine learning**. Jann Spiess will give two guest lectures (topics to be determined). We will next discuss (deep) neural nets, including some numerical methods used for training them, such as stochastic gradient descent. We will then consider methods developed for use with text data, in particular topics models. After that, we will review methods for active learning in the context of multi-armed bandit settings. We will review some theoretical results providing performance guarantees (regret bounds) for algorithms used for learning in bandit settings. Lastly, we will talk a bit about data visualization.

At some point early in the semester, I will provide an introduction to R. R is an open source statistical software with a large and growing community, which (I believe) will increasingly supplant other environments such as Stata and Matlab

We will conclude the semester with in-class discussions of research proposals by you that are related to the topics covered. We will do so in a fairly informal manner, based on brief presentations by you and subsequent open discussion.

Requirements and policies

Your grade for Economics 2148 will be determined based the following assignments. You are asked to complete two computer-based problem sets, and to submit summaries of two papers of your choice from the references at the end of this Syllabus. I am happy to make recommendations if you are not sure which ones to pick. You will also have to prepare a research proposal, of 3-10 pages, and to give a brief presentation of your proposal in class. Proposals could be applied or theoretical, but are ideally related to some of the topics covered in class. Please upload both your problem set solutions, summaries, and research proposal via Canvas.

These assignments contribute to your grade as follows.

1. Two **summaries** of about 3 pages length each (10% of grade each).

2. Two **problem set** solutions (10% of grade each).

3. In-class **midterm exam on October 12** (30% of grade).

4. A **research proposal** and presentation. (30% of grade).

Additionally, the slides contain a lot of **"practice problems,"** which you will have to solve in class. The idea is to have you complete most of the proofs, after I pointed you in the right direction. After a few minutes, we will discuss the solutions to these problems. These problems provide good guidance for what you might expect from the midterm exam.

To help me improve the course, I will ask you to give me anonymous feedback at some point, writing what you like about the class and what you think I should change.

I encourage you to come to office hours with any questions. If you need any special accommodations for physical or medical reasons, please see me after class or send me an email.

Outline of the course

Instrumental variables part I – origins and binary treatment

- Origins of instrumental variables: Systems of linear structural equations

- Strong restriction: Constant causal effects.

- Modern perspective: Potential outcomes, allow for heterogeneity of causal effects

- Keep IV estimand, reinterpret it in more general setting:
  Local Average Treatment Effect (LATE)

- Keep object of interest: Average Treatment Effect (ATE)
  Partial identification (Bounds)

Instrumental variables part II – continuous treatment

- Restricting heterogeneity in the structural equation:
  Nonparametric IV (conditional moment equalities)

- Restricting heterogeneity in the first stage:
  Control functions

- Linear IV:
  Continuous version of LATE

Review of decision theory

- Basic definitions
- Optimality criteria
- Relationships between optimality criteria
- Analogies to microeconomics
- Two justifications of the Bayesian approach

Shrinkage in the normal means model

- Setup: the normal means model $\boldsymbol{X} \sim N(\boldsymbol{\theta}, I_k)$ and the canonical estimation problem with loss $\|\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2$.
- The James-Stein (JS) shrinkage estimator.
- Three ways to arrive at the JS estimator (almost):
  1. Reverse regression of θ_i on X_i.
  2. Empirical Bayes: random effects model for θ_i.
  3. Shrinkage factor minimizing Stein's Unbiased Risk Estimate.
- Proof that JS uniformly dominates $\boldsymbol{X}$ as estimator of $\boldsymbol{\theta}$.
- The normal means model as asymptotic approximation.

Gaussian process priors, reproducing kernel Hilbert spaces, and Splines

- 6 equivalent representations of the posterior mean in the normal-normal model.
- Gaussian process priors for regression functions.
- Reproducing Kernel Hilbert Spaces and splines.

Applications of Gaussian process priors from my own work

- Optimal treatment assignment in experiments.
  - Setting: Treatment assignment given baseline covariates
  - General decision theory result:
    Non-random rules dominate random rules
  - Prior for expectation of potential outcomes given covariates
  - Expression for MSE of estimator for ATE
    to minimize by treatment assignment
- Optimal insurance and taxation.
  - Review: Envelope theorem.
  - Economic setting: Co-insurance rate for health insurance
  - Statistical setting: prior for behavioral average response function
  - Expression for posterior expected social welfare
    to maximize by choice of co-insurance rate

Deep neural nets

- What are neural nets?
- Network design:
  Activation functions, network architecture, output layers.
- Calculating gradients for optimization:
  Backpropagation, stochastic gradient descent.
- Regularization using early stopping.

Text analysis

- Representing text as data.
- Text regression.
- Generative language models.
- Hierarchical Dirichlet processes.

Bandit problems

- Setup: The multi-armed bandit problem.
  Adaptive experiment with exploration / exploitation trade-off.
- Two popular approximate algorithms:
  1. Thompson sampling
  2. Upper Confidence Bound algorithm
- Characterizing regret.
- Characterizing an exact solution: Gittins Index.
- Extension to settings with covariates (contextual bandits).

References

Instrumental variables – binary treatment

Angrist, J., Imbens, G., and Rubin, D. (1996). Identification of causal effects using instrumental variables. *Journal of the American Statistical Association*, 91(434):444–455.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Verlag, chapter 2 and 7.

Instrumental variables – continuous treatment

Newey, W. K. and Powell, J. L. (2003). Instrumental Variable Estimation of Nonparametric Models. *Econometrica*, 71(5):1565–1578.

Horowitz, J. L. (2011). Applied Nonparametric Instrumental Variables Estimation. *Econometrica*, 79(2):347–394.

Hahn, J. and Ridder, G. (2011). Conditional moment restrictions and triangular simultaneous equations. *The Review of Economics and Statistics*, 93(2):683–689

Imbens, G. W. and Newey, W. K. (2009). Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity. *Econometrica*, 77:1481–1512.

Kasy, M. (2011). Identification in triangular systems using control functions. *Econometric Theory*, 27(03):663–671.

Angrist, J. D., Graddy, K., and Imbens, G. W. (2000). The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. *The Review of Economic Studies*, 67(3):499–527.

Review of decision theory

Robert, C. (2007). *The Bayesian choice: from decision-theoretic foundations to computational implementation.* Springer Verlag, chapter 2.

Shrinkage in the normal means model

Wasserman, L. (2006). *All of nonparametric statistics.* Springer Science & Business Media, chapter 7.

Stigler, S. M. (1990). The 1988 Neyman memorial lecture: a Galtonian perspective on shrinkage estimators. *Statistical Science*, pages 147–155.

Morris, C. N. (1983). Parametric empirical Bayes inference: Theory and applications. *Journal of the American Statistical Association*, 78(381):pp. 47–55.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge University Press, chapter 7.

Hansen, B. E. (2016). Efficient shrinkage in parametric models. *Journal of Econometrics*, 190(1):115–132.

Abadie, A. and Kasy, M. (2018). Choosing among regularized estimators in empirical economics - the risk of machine learning. *Review of Economics and Statistics*, forthcoming.

Gaussian process priors, reproducing kernel Hilbert spaces, and Splines

Williams, C. and Rasmussen, C. (2006). *Gaussian processes for machine learning.* MIT Press, chapters 2 and 7.

Wahba, G. (1990). *Spline models for observational data*, volume 59. Society for Industrial Mathematics, chapter 1.

Applications of Gaussian process priors from my own work

Kasy, M. (2016). Why experimenters might not always want to randomize, and what they could do instead. *Political Analysis*, 24(3):324–338.

Kasy, M. (2018). Optimal taxation and insurance using machine learning. *Journal of Public Economics.*

Deep neural nets

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep learning.* MIT Press.

Text analysis

Gentzkow, M., Kelly, B. T., and Taddy, M. (2019). Text as data. *Journal of Economic Literature.*

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581.

Bandit problems

Bubeck, S. and Cesa-Bianchi, N. (2012). Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122.

Russo, D. J., Roy, B. V., Kazerouni, A., Osband, I., and Wen, Z. (2018). A Tutorial on Thompson Sampling. *Foundations and Trends® in Machine Learning*, 11(1):1–96.

Weber, R. et al. (1992). On the Gittins index for multiarmed bandits. *The Annals of Applied Probability*, 2(4):1024–1033.