ECON E724 Joon Y. Park
Economics Department Spring 2025

Econometrics of Big Data and Machine Learning

The objective of this course is to prepare the Ph.D. students for their dissertation research in the area of *econometrics* of big data and machine learning, both theoretical and empirical. The course is offered for the students who have some serious interests in econometric theory and methodology, not for the students who only wish to learn basic tools on machine learning and large dimensional data analysis. More specifically, the course is intended to provide the students with econometric meanings and perspectives of the techniques and tools that are used widely in machine learning and large dimensional data analysis. In particular, the course is not developed for the students who are mainly interested in learning basic machine learning programs to use them for simple practical applications.

The course consists of three parts. Part I introduces the basic theory of factor models estimated by the principal component analysis (PCA) and large dimensional regressions. Part II provides the basic theory of functional data analysis, which have recently become increasingly more popular to analyze cross-sectional heterogeneities, with applications on various topics across all fields in economics. Part III presents some selected topics in machine learning that are relevant to empirical researches in economics and related fields, such as support vector machine (SVM), tree-based approach including random forest, and various NNs such as LSTM and other RNNs as well as vanilla NNs. The lectures will be based on presentation files, lecture notes or research papers.

The students will be given biweekly homework, one in-class midterm examination, and an empirical project. For the required empirical project, each student will choose any methodology discussed in class, use it to analyze an interesting empirical problem, and present her/his work in class. The course grade will be based on homework, midterm examination, and the empirical project and its presentation.

To fully appreciate the lectures and start meaningful research after taking the course, students should have a good understanding of the materials covered in the standard econometrics sequence offered for the first year economics Ph.D. students. In addition, the basic knowledge in advanced-level linear algebra and elementary functional analysis on Hilbert space of functions will be needed. Though not essential, the materials covered in the second year Ph.D. econometrics courses will also be very helpful to get the best out of this course. Finally, students are expected to have knowledge on computing and programming good enough to use common machine learning tool packages in MATLAB or Python, and to write their own programs in one of these languages. This will be necessary for them to do the required empirical project.

Lectures will be given on Tuesdays and Thursdays at 8:00-10:30am in Wylie Hall Rm 329 until the week of February 23, and student presentations will be scheduled in the weeks

of April 2 and April 9. This will give students four weeks, including spring break, to do the required empirical projects and prepare their presentations. Students who wish to start their research on the areas covered in this course are expected to further develop their own research ideas during the rest of the semester. I will have office hours in my office (Rm 215, Wylie Hall) on Tuesdays at 2:00-4:00pm. Students may meet me outside my office hours by making an appointment in advance. My email address is joon@iu.edu.

Reference Books and Monographs

Functional Data Analysis

Linear Processes in Function Spaces, Lecture Notes in Statistics 149, by D. Bosq, Spinger, 2000.

Functional Data Analysis, by J.O. Ramsay and B.W. Silverman, Springer, 2005.

Applied Functional Data Analysis: Methods and Case Studies, by J.O. Ramsay and B.W. Silverman, Springer, 2007.

Machine Learning

The Elements of Statistical Learning, 2nd edition, by T. Hastie, R. Tibshirani and J. Friedman, Springer, 2009.

An Introduction to Statistical Learning, 2nd edition, by G. James, D. Witten, T. Hastie and R. Tibshirani, Spinger, 2021.

Reference Papers

Here I only list the papers that I will discuss in my lectures. The list below is tentative and will be updated later.

1. *Large Dimensional Regressions and Factor Models*

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the LASSO," *Journal of the Royal Statistical Society: Series B*, 58, 267-288.

Zou, H. (2006). "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418-1429.

Bai, J. and S. Ng (2002). "Determining the Number of Factors in Approximate Factor Model," *Econometria*, 70, 191-221.

Ahn, S.C. and A.R. Horenstein (2013). "Eigenvalue Ratio Test for the Number of Factors," *Econometrica*, 81, 1203-1227.

Zou, H, T. Hastie and R. Tibshirani (2006). "Sparse Principal Component Analysis," *Journal of Computational and Graphical Statistics*, 15, 265-286.

2. *Functional Data Analysis*

Chang, Y., C.S. Kim, and J.Y. Park (2015). "Nonstationarity in Time Series of State Densities," *Journal of Econometrics*, 192, 152-167.

Chang, Y., R.K. Kaufmann, C.S. Kim, J.I. Miller, J.Y. Park and S. Park (2020). "Time Series Analysis of Global Temperature Distributions: Identifying and Estimating Persistent Features in Temperature Anomalies," *Journal of Econometrics*, 214, 274-294.

Chang, Y., B. Hu and J.Y. Park (2018). "Econometric Analysis of Functional Dynamics in the Presence of Persistence," Working Paper, Indiana University.

Park, J.Y. and J. Qian (2012). "Functional Regression of Continuous State Distributions," *Journal of Econometrics*, 167, 397-412.

Chang, Y., M. Choi and J.Y. Park (2023). "A Factor Model for Functional Time Series," Working Paper, Indiana University.

Chang, Y., J.I. Miller and J.Y. Park (2021). "How Does Economic Activity Interact with Climate? What We Learn from Global Temperature Anomaly Distributions," Working Paper, Indiana University.

Chang, Y., S. Durlauf, S. Lee and J.Y. Park (2023). "A Trajectories-Based Approach to Measuring Intergenerational Mobility," NBER Working Paper 31020.

Chang, Y., J.Y. Park and D. Pyun (2022). "From Functional Autoregressions to Vector Autoregressions," Working Paper, Indiana University.

3. *Machine Learning*

Mullainathan, S. and J. Spiess (2017). "Machine Learning: An Applied Econometric Approach," *Journal of Economic Perspectives*, 31, 87-106.

Athey, S. and G.W. Imbens (2019). "Machine Learning Methods that Economists Should Know About," *Annual Review of Economics*, 11, 685-725.

Wager, S. and S. Athey (2018). "Estimation and Inference of Heterogeneous Treatment Effects Using Random Forests," *Journal of the American Statistical Association*, 113, 1228-1242.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey and J. Robins (2018). "Double/Debiased Machine Learning for Treatment and Structural Parameters," *Econometrics Journal*, 21, C1-C68.

Farrell, M.H., T. Liang, and S. Misra (2021). "Deep Neural Networks for Estimation and Inference," *Econometrica*, 89, 181-213.

Chernozhukov, V., W. Newey and R. Singh (2022). "Automatic Debiased Machine Learning of Causal and Structural Effects," *Econometrica*, 90, 967-1027.

Chang, Y., J.Y. Park and G. Yan (2022). "Machine Learning in Econometric Models: Using SVM to Estimate and Predict Binary Choice Models," Working Paper, Indiana University.

Yan, G. (2022). "A Kernelization-Based Approach to Nonparametric Binary Choice Models," Working Paper, Indiana University.