University of St.Gallen

Course and Examination Fact Sheet: Spring Semester 2021

10,382: Econometrics of Big Data

ECTS credits: 4

Overview examination/s

(binding regulations see below)
Decentral - examination paper written at home (individual) (100%)
Examination time: term time

Attached courses

Timetable -- Language -- Lecturer
10,382,1.00 (GSERM) Econometrics of Big Data -- Englisch -- Spindler Martin

Course information

Course prerequisites

The course is a PhD level course. Basic knowledge of parametric statistical models and associated asymptotic theory is expected.

Learning objectives

Students will be introduced to several modern methods, largely coming from statistics and machine learning, which are useful for exploring high dimensional data and for building prediction models in high dimensional settings. Students will learn how to adapt high dimensional methods to the problem of doing valid inference about model parameters and illustrate applications of these proposals for doing inference about economically interesting parameters.

Course content

As in many other fields, economists are increasingly making use of high-dimensional models - models with many unknown parameters that need to be inferred from the data. Such models arise naturally in modern data sets that include rich information for each unit of observation (a type of "big data") and in nonparametric applications where researchers wish to learn, rather than impose, functional forms. High-dimensional models provide a vehicle for modeling and analyzing complex phenomena and for incorporating rich sources of confounding information into economic models.

Our goal in this course is two-fold. First, we wish to provide an overview and introduction to several modern methods, largely coming from statistics and machine learning, which are useful for exploring high-dimensional data and for building prediction models in high-dimensional settings. Second, we will present recent proposals that adapt high-dimensional methods to the problem of doing valid inference about model parameters and illustrate applications of these proposals for doing inference about economically interesting parameters.

Course structure

Lecture 1: Introduction to High-Dimensional Modeling

- Breiman, L. (1996), "Bagging Predictors," Machine Learning 26: 123-140
- Friedman, J., T. Hastie, and R. Tibshirani (2000), "Additive logistic regression: A statistical view of boosting (with discussion)," Annals of Statistics, 28, 337-407
- Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer. [Elements from Chapters 2, 5, 7, 8.7, 10]
- James, G., D. Witten, T. Hastie, and R. Tibshirani (2014), An Introduction to Statistical Learning with Applications in R, Springer. [Elements from Chapters 2, 3, 5, 7, 8.2]
- Li, Q. and J. S. Racine (2007), Nonparametric Econometrics: Theory and Practice, Princeton University Press. [Elements

from Chapters 2, 14]
- Schapire, R. (1990), "The strength of weak learnability," Machine Learning, 5, 197-227

Lecture 2: Introduction to Distributed Computing for Very Large Data Sets

Lecture 3: Tree-based Methods

- Athey, S. and G. Imbens (2015), "Machine Learning Methods for Estimating Heterogeneous Causal Effects," working paper, http://arxiv.org/abs/1504.01132
- Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer. [Chapters 9, 10, 15, 16]
- James, G., D. Witten, T. Hastie, and R. Tibshirani (2014), An Introduction to Statistical Learning with Applications in R, Springer. [Chapter 8]
- Wager, S. and S. Athey (2015), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," working paper, http://arxiv.org/abs/1510.04342
- Wager, S. and G. Walther (2015), "Uniform Convergence of Random Forests via Adaptive Concentration," working paper, http://arxiv.org/abs/1503.06388
- Wager, S., T. Hastie, and B. Efron (2014), "Confidence Intervals for Random Forests: The Jackknife and the Infinitesimal Jackknife," Journal of Machine Learning Research, 15, 1625−1651

Lecture 4: An Overview of High-Dimensional Inference

- Belloni, A. and V. Chernozhukov (2013), "Least Squares After Model Selection in High-dimensional Sparse Models," Bernoulli, 19(2), 521-547.
- Belloni, A., D. Chen, V. Chernohukov, and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica, 80(6), 2369-2430
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "High-Dimensional Methods and Inference on Structural and Treatment Effects," Journal of Economic Perspectives, 28(2), 29-50
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "Inference on Treatment Effects after Selection amongst High-Dimensional Controls," Review of Economic Studies, 81(2), 608-650
- Belloni, A., V. Chernozhukov, and C. Hansen (2015), "Inference in High Dimensional Panel Models with an Application to Gun Control," forthcoming Journal of Business and Economic Statistics
- Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2013), "Program Evaluation with High-Dimensional Data," working paper, http://arxiv.org/abs/1311.2645
- Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," American Economic Review, 105(5), 486-490
- Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," Annual Review of Economics, 7, 649-688

Lecture 5: Penalized Estimation Methods

- Belloni, A. and V. Chernozhukov (2013), "Least Squares After Model Selection in High-dimensional Sparse Models," Bernoulli, 19(2), 521-547.
- Fan, J. and J. Lv (2008), "Sure independence screening for ultrahigh dimensional feature space," Journal of the Royal Statistical Society, Series B, 70(5), 849-911
- Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer. [Chapters 3, 4, 5, 18]
- James, G., D. Witten, T. Hastie, and R. Tibshirani (2014), An Introduction to Statistical Learning with Applications in R, Springer. [Chapter 6]

Lecture 6: Moderate p Asymptotics

Lecture 7: Examples

- Belloni, A., D. Chen, V. Chernohukov, and C. Hansen (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica, 80(6), 2369-2430
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "High-Dimensional Methods and Inference on Structural and Treatment Effects," Journal of Economic Perspectives, 28(2), 29-50
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "Inference on Treatment Effects after Selection amongst High-Dimensional Controls," Review of Economic Studies, 81(2), 608-650
- Belloni, A., V. Chernozhukov, and C. Hansen (2015), "Inference in High Dimensional Panel Models with an Application to

Gun Control," forthcoming Journal of Business and Economic Statistics
- Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2013), "Program Evaluation with High-Dimensional Data," working paper, http://arxiv.org/abs/1311.2645
- Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Post-Selection and Post-Regularization Inference in Linear Models with Many Controls and Instruments," American Economic Review, 105(5), 486-490
- Chernozhukov, V., C. Hansen, and M. Spindler (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," Annual Review of Economics, 7, 649-688
- Gentzkow, M., J. Shapiro, and M. Taddy (2015), "Measuring Polarization in High-Dimensional Data: Method and Application to Congressional Speech," working paper, http://www.brown.edu/Research/Shapiro/
- Hansen, C. and D. Kozbur (2014), "Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE," Journal of Econometrics, 182(2), 290-308
- Kleinberg, J., J. Ludwig, S. Mullainathan, and Z. Obermeyer (2015), "Prediction Policy Problems," American Economic Review: Papers and Proceedings, 105(5), 491-495

Lecture 8: Inference: Computation

Lecture 9: Introduction to Unsupervised Learning

- Blei, D., A. Ng, and M. Jordan (2003), Lafferty, J., ed. "Latent Dirichlet allocation," Journal of Machine Learning Research, 3 (4-5), 993-1022
- Hastie, T., R. Tibshirani, and J. Friedman (2009), The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer. [Chapter 14]
- James, G., D. Witten, T. Hastie, and R. Tibshirani (2014), An Introduction to Statistical Learning with Applications in R, Springer. [Chapter 10]
- Li, Q. and J. S. Racine (2007), Nonparametric Econometrics: Theory and Practice, Princeton University Press. [Chapter 1]
- Stock J. H and Watson M. W (2002), "Forecasting using principal components from a large number of predictors," Journal of the American Statistical Association, 97, 1167-1179

Lecture 10: Very Large p Asymptotics

- Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2012): "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," Econometrica, 80, 2369-2429. (ArXiv, 2010)
- Belloni, A., and V. Chernozhukov (2011): "`1-penalized quantile regression in high-dimensional sparse models," Annals of Statistics, 39(1), 82-130. (ArXiv, 2009)
- Belloni, A., and V. Chernozhukov (2013): "Least Squares After Model Selection in High-dimensional Sparse Models," Bernoulli, 19(2), 521-547. (ArXiv, 2009)
- Belloni, A., V. Chernozhukov, and C. Hansen (2010) "Inference for High-Dimensional Sparse Econometric Models," Advances in Economics and Econometrics. 10th World Congress of Econometric Society, Shanghai, 2010. (ArXiv, 2011).
- Belloni, A., V. Chernozhukov, and C. Hansen (2014), "Inference on Treatment Effects after Selection amongst High-Dimensional Controls," Review of Economic Studies, 81(2), 608-650
- Belloni, A., V. Chernozhukov, K. Kato (2013): "Uniform Post Selection Inference for LAD Regression Models," arXiv:1304.0282. (ArXiv, 2013)
- Belloni, A., V. Chernozhukov, L. Wang (2011a): "Square-Root-LASSO: Pivotal Recovery of Sparse Signals via Conic Programming," Biometrika, 98(4), 791-806. (ArXiv, 2010).
- Belloni, A., V. Chernozhukov, L. Wang (2011b): "Square-Root-LASSO: Pivotal Recovery of Nonparametric Regression Functions via Conic Programming," (ArXiv, 2011)
- Belloni, A., V. Chernozhukov, Y. Wei (2013): "Honest Confidence Regions for Logistic Regression with a Large Number of Controls," arXiv preprint arXiv:1304.3969 (ArXiv, 2013)
- Bickel, P., Y. Ritov and A. Tsybakov, "Simultaneous analysis of Lasso and Dantzig selector", Annals of Statistics, 2009.
- Candes E. and T. Tao, "The Dantzig selector: statistical estimation when p is much larger than n," Annals of Statistics, 2007.
- Donald S. and W. Newey, "Series estimation of semilinear models," Journal of Multivariate Analysis, 1994.
- Tibshirani, R, "Regression shrinkage and selection via the Lasso," J. Roy. Statist. Soc. Ser. B, 1996.
- Frank, I. E., J. H. Friedman (1993): "A Statistical View of Some Chemometrics Regression Tools," Technometrics, 35(2), 109-135.
- Gautier, E., A. Tsybakov (2011): "High-dimensional Instrumental Variables Rergession and Confidence Sets," arXiv:1105.2454v2
- Hahn, J. (1998): "On the role of the propensity score in efficient semiparametric estimation of average treatment effects," Econometrica, pp. 315-331.
- Heckman, J., R. LaLonde, J. Smith (1999): "The economics and econometrics of active labor market programs," Handbook of labor economics, 3, 1865-2097.

- Imbens, G. W. (2004): "Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review," The Review of Economics and Statistics, 86(1), 4-29.
- Leeb, H., and B. M. Potscher (2008): "Can one estimate the unconditional distribution of post-model-selection estimators?," Econometric Theory, 24(2), 338-376.
- Robinson, P. M. (1988): "Root-N-consistent semiparametric regression," Econometrica, 56(4), 931-954.
- Rudelson, M., R. Vershynin (2008): "On sparse reconstruction from Foruier and Gaussian Measurements", Comm Pure Appl Math, 61, 1024-1045.
- Jing, B.-Y., Q.-M. Shao, Q. Wang (2003): "Self-normalized Cramer-type large deviations for independent random variables," Ann. Probab., 31(4), 2167-2215.

Course literature

Course notes and a list of readings provided at the beginning of the course.

Additional course information

Only for PhD students of the University of St.Gallen

PEF students may register via regular bidding for the courses offered together by PEF and Global School in Empirical Research Methods (GSERM). Enrolment in a course is binding: students have to attend the course and take the exam. The credits will be shown on the scorecard.

All other PhD students should register for the courses offered by Global School in Empirical Research Methods (GSERM), both via bidding and via GSERM for:

- courses for the curriculum and
- optional courses with an examination. These will be listed on the scorecard under optional work (only possible if all required elective courses have already been completed).

Please register only via GSERM for:

- optional courses without an examination and
- optional courses if not all required elective courses have been completed (not shown on the scorecard).

In the case of the President's Board having to implement new directives due to the SARS-CoV-2 pandemic in SpS2021, the course information listed above will be changed as follows:

- This course may take place in a digital version.

There will be no change in the examination.

Examination information

Examination sub part/s

1. Examination sub part (1/1)

Examination time and form
Decentral - examination paper written at home (individual) (100%)
Examination time: term time

Remark
take-home final exam

Examination-aid rule
Term papers

Term papers must be written without anyone else's help and in accordance with the known quotation standards, and they must contain a declaration of authorship which is a published template in StudentWeb.

The documentation of sources (quotations, bibliography) has to be done throughout and consistently in accordance with the chosen citation standard such as APA or MLA.

For papers in law, the legal standard is recommended (by way of example, cf. FORSTMOSER, P., OGOREK R. et SCHINDLER B., Juristisches Arbeiten: Eine Anleitung für Studierende, newest edition respectively, or according to the recommendations of the Law School).

The indications of the sources of information taken over verbatim or in paraphrase (quotations) must be integrated into texts in accordance with the precepts of the applicable quotation standard, while informative and bibliographical notes must be added as footnotes (recommendations and standards can be found, for example, in METZGER, C., Lern- und Arbeitsstrategien, newest edition respectively.

For any work written at the HSG, the indication of the page numbers is mandatory independent of the chosen citation standard. Where there are no page numbers in sources, precise references must be provided in a different way: titles of chapters or sections, section numbers, acts, scenes, verses, etc.

Supplementary aids
--

Examination languages
Question language: English
Answer language: English

Examination content

Content of the lectures

Examination relevant literature

To be discussed in class

Please note

Please note that only this fact sheet and the examination schedule published at the time of bidding are is binding and takes precedence over other information, such as information on StudyNet (Canvas), on lecturers' websites and information in lectures etc.

Any references and links to third-party content within the fact sheet are only of a supplementary, informative nature and lie outside the area of responsibility of the University of St.Gallen.

Documents and materials are only relevant for central examinations if they are available by the end of the lecture period (CW21) at the latest. In the case of centrally organised mid-term examinations, the documents and materials up to CW 12 are relevant for testing.

Binding nature of the fact sheets:

- Course information as well as examination date (organised centrally/decentrally) and form of examination: from bidding start in CW 04 (Thursday, 28 January 2021);
- Examination information (regulations on aids, examination contents, examination literature) for decentralised examinations: in CW 12 (Monday, 22 March 2021);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised mid-term examinations: in CW 12 (Monday, 22 March 2021);
- Examination information (regulations on aids, examination contents, examination literature) for centrally organised examinations: two weeks before the end of the registration period in CW 14 (Thursday, 8 April 2021).