Financial Statistics

Markus Pelger

General Information

- **Class Meeting Times:** 5-day block course (4/9, 4/10, 4/13, 4/14, 4/15)
  4 hours of class each day (exact times TBA)
- **Instructor:** Markus Pelger, mpelger@stanford.edu
- **Target audience:** Ph.D. students.

Short Description

This Ph.D. course covers topics in financial statistics with a focus on current research. Topics will include large dimensional factor modeling, high-frequency statistics, missing financial data and machine learning in finance. The course requires a strong background in statistics and mathematics and some knowledge of economics and finance. The evaluation will be based on a final project.

Course Outline

Financial statistics is the intersection of statistical techniques and finance. Financial statistics provides a set of tools that are useful for modeling financial data and testing beliefs about how markets work and prices are formed. Conversely, new techniques in analyzing financial data can lead to empirical facts inconsistent with existing theories, begging for new models or investment strategies.

The goal of this course is to introduce students to the frontiers of financial statistics, financial econometrics and machine learning in finance. The course will focus on the statistical theory of the estimation approaches, but we will also spend time covering a range of significant applications to the estimators. As we often have a large number of observations for financial data, we will develop an asymptotic inferential theory for our estimators.

The course is not intended as an introduction class to financial statistics, but to lead Ph.D. students to the research frontiers of statistically analyzing financial data. It is not possible within a quarter to cover all relevant topics in financial statistics in depth. Hence we will focus on a selected number of topics including:

- High-dimensional factor modeling: Curse of dimensionality, principal component analysis, random matrix theory and spiked covariance models, number of factors, matrix completion

- High-frequency statistics: Limit theorems, non-parametric volatility and jump estimation
- Machine-learning in empirical asset pricing
- Missing Financial data: Latent low rank structures, applications to causal inference

Evaluation

This class is intended to prepare students for research in financial statistics. The final presentation is one of the most important parts of the course. This is an opportunity for PhD students to receive feedback on their research and practise its presentation in an academic setup. Students have the choice between two options:

1. Students can present their own research if it broadly relates to financial statistics (this includes methodological work applied to finance or empirical financial research).
2. Students can select a paper that is at the research frontier. Students should replicate the paper and provide a critical discussion on it.

The presentations are scheduled for Wednesday, 4/15. We have 10 presentations slots and each student has 20 minutes for presentation and 5 minutes for Q&A and feedback. The length of the presentations and their number might be adjusted depending on the enrollment. As this is a block course, students are expected to start preparing for the presentations in advance.

If you decide to replicate an existing paper, you can choose from over 30 different papers that are listed at the end of the syllabus. If you are interested in a paper that is not on the list, please send me an email. If it is appropriate for this class I am happy to include it. If you present your own research, please also send me an email in advance.

- Your contribution will consist of two elements: a presentation and a brief summary report.
- Presentation:
  - Hold a 20 minute presentation on your topic. This presentation should include the intuition of the theory behind the estimation approach and your empirical results. The presentation should be understandable by your classmates who have not read the paper. You should view your presentation as a conference presentation.
  - Send me your presentation slides on the evening of Tuesday, 4/14 by email. They will be part of your evaluation.
- Summary report:
  - Write a brief summary report which summarizes the main insights of your presentation. If you are presenting your own research, you can also attach a draft of the research paper as well. This report is due on Tuesday, 4/14 as well.
  - There is no prescribed page number for the report. The report should include enough details to understand the main ideas of the paper.

Active participation in class and during the student presentations will be rewarded.

Course Material

Unfortunately, there is no Ph.D. level textbook for financial statistics. The course readings are assigned to each topic based on book chapters, papers and notes.

The following textbooks are good introductory master level textbooks:

- Tsay: Analysis of Financial Time Series
- Ruppert and Matteson: Statistics and Data Analysis for Financial Engineering
- Fan: The Elements of Financial Econometrics
- Lai: Statistical Models and Methods for Financial Markets
- Carmona: Statistical Analysis of Financial Data in R
- Campbell, Lo and MacKinlay: The Econometrics of Financial Markets

These are advanced graduate textbooks that cover some of the topics in this course

- Singleton: Empirical Dynamic Asset Pricing
- Aït Sahalia and Hansen: Handbook of Financial Econometrics, vol I and II
- Newey and McFadden: Handbook of Econometrics, Large Sample Estimation and Hypothesis Testing
- Aït Sahalia and Jacod: High-Frequency Financial Econometrics
- Bai, Yao and Zheng: Large Sample Covariance Matrices and High-Dimensional Data Analysis
- Cochrane: Asset Pricing
- Hamilton: Time Series Analysis
- Nagel: Machine Learning in Asset Pricing

Detailed Course Outline and Reading List

The readings are categorized into required readings and recommended additional readings. If you intend to do research in this area I would strongly recommend to read all the material. You can find the required readings on my website https://mpelger.people.stanford.edu:

1. **High-Dimensional Factor Modeling**
   Required readings:
   - Lettau and Pelger (2020): Estimating Latent Asset-Pricing Factors
   - Lettau and Pelger (2020): Factors that Fit the Time Series and Cross-Section of Stock Returns
   - Pelger and Xiong (2020): State-Varying Factor Models of Large Dimensions
   - Pelger and Xiong (2020). Interpretable Proximate Factors for Large Dimensions
   - Bryzgalova, DeMiguel, Li and Pelger (2022): Asset-Pricing Factors with Economic Targets

Recommended readings:

- Bai and Ng (2008): Large Dimensional Factor Analysis
- Bai (2003): Inferential Theory for Factor Models of Large Dimensions
- Cochrane (2000): Asset Pricing: Chapter 12
- Tsay (2002): Analysis of Financial Time Series 2002: Chapter 11
- Ruppert and Matteson (2015): Statistics and Data Analysis for Financial Engineering: Chapter 18
- Fan, Liao and Mincheva (2013): Large covariance estimation by thresholding principal orthogonal complements
- Bai, Yao and Zheng (2015): Large Sample Covariance Matrices and High-Dimensional Data Analysis: Chapter 1-12
- Benaych-Georges and Nadakuditi (2011): Eigenvalues and eigenvectors of low rank perturbation
- Ahn and Horenstein (2013): Eigenvalue Ratio Test for the Number of Factors
- Bai (2003): Inferential Theory for Factor Models of Large Dimensions
- Bai and Ng (2002): Determining the number of factors in approximate factor models
- Onatksi (2010): Determining the Number of Factors from Empirical Distribution of Eigenvalues
- Bai, Yao and Zheng (2015): Large Sample Covariance Matrices and High-Dimensional Data Analysis: Chapter 11
- Pelger and Zou (2022): Inference for Large Panel Data with Many Covariates
- Jagannathan, Skoulakis and Wang (2010): Handbook of Financial Econometrics: The Analysis of the Cross-Section of Security Returns

2. **Missing Data**
Required readings:

- Xiong and Pelger (2020): Large Dimensional Latent Factor Modeling with Missing Observations and Applications to Causal Inference
- Bryzgalova, S, Lettau, M., Lerner, S. and Pelger, M. (2022): Missing Financial Data
- Duan, Pelger and Xiong (2022): Target-PCA: Transfer Learning Large Dimensional Panel Data
- Duan, Pelger and Xiong (2024): Factor Analysis for Large Non-Stationary Panels with Endogenous Missingness and Applications to Causal Inference
- Duan and Pelger (2025): Imputation-Powered Inference for Missing Covariates

Recommended readings:

- Bai and Ng (2020): Matrix Completion, Counterfactuals and Factor Analysis of Missing Data
- Freyberger, Höppner, Neuhierl and Weber (2022): Missing Data in Asset Pricing Panel

- Cahan, Bai, and Ng (2022): Factor-based imputation of missing values and covariances in panel data of large dimensions
- Athey, Bayati, Doudchenko, Imbens and Khosravi (2020): Matrix Completion Methods for Causal Panel Data Models

3. **High-Frequency Statistics**
Required readings:

- Pelger (2019): Large-dimensional factor modeling based on high-frequency observations
- Pelger (2020): Understanding Systematic Risk: A High-Frequency Approach
- Podolskij and Vetter (2009): Understanding limit theorems for semimartingales: a short survey

Recommended readings:

- Aït-Sahalia and Jacod (2012): Analyzing the Spectrum of Asset Returns: Jump and Volatility Components in High Frequency Data
- Lee and Mykland (2008): Jumps in Financial Markets: A New Nonparametric Test and Jump Dynamics
- Barndorff-Nielsen and Shephard (2006): Econometrics of Testing for Jumps in Financial Economics Using Bipower Variation
- Aït-Sahalia and Jacod (2014): Chapters 1-4, 6-8, 10
- Christensen, Oomen, Podolskij (2014): Fact or friction: Jumps at ultra high frequency
- Aït-Sahalia and Jacod (2009): Testing for Jumps in a Discretely Observed Process

4. **Machine-Learning in Finance**
Required readings:

- Pelger (2022): Asset Pricing and Investment with Big Data
- Chen, Pelger and Zhu (2022): Deep-Learning in Asset Pricing
- Bryzgalova, Pelger and Zhu (2019): Forest through the Trees: Building Cross-Sections of Stock Returns
- Guijarro-Ordonez, Pelger and Zanotti (2019): Deep Learning Statistical Arbitrage
- Filipovic, Pelger and Ye (2022): Stripping the Discount Curve - A Robust Machine Learning Approach
- Filipovic, Pelger and Ye (2022): Shrinking the Term Structure
- Kaniel, Lin, Pelger and Van Nieuwerburgh (2021): Machine-Learning the Skill of Mutual Fund Managers
- Gu, Kelly and Xiu (2020): Empirical Asset Pricing via Machine Learning

Recommended readings:

- Kozak, Nagel, and Santosh (2018). Shrinking the Cross-Section.
- Freyberger, Neuhierl and Weber, (2020)

- Da, Nagel and Xiu (2022).The Statistical Limit of Arbitrage.
- Kelly, Malamud and Zhou (2022). The Virtue of Complexity in Return Prediction.
- Kelly. Pruitt, and Su (2019). Characteristics Are Covariances: A Unified Model of Risk and Return.

Papers for Final Project (Choose one paper)

1. Large-dimensional factor modeling

    - Fan, J., Liao, Y. and Wang, W. (2016). Projected Principal Component Analysis in Factor Models. Annals of Statistics
    - Bai, J. and Ng, S. (2019). Principal Components and Regularized Estimation of Factor Models. Journal of Econometrics
    - Fan, J., Liao, Y. and Micheva, M. (2013). Large Covariance Estimation by Thresholding Principal Orthogonal Complements. Journal of Royal Statistical Society
    - Fan, J., Xue, L. and Yao, J. (2017). Sufficient forecasting using factor models. Journal of Econometrics
    - Fan and Liao (2020). Learning Latent Factors From Diversified Projections and Its Applications to Over Estimated and Weak Factors. JASA
    - Kelly, Malamud and Pedersen (2022). Principal Portfolios. Journal of Finance

2. Testing of asset-pricing models

    - Kelly, B. and S. Pruitt (2015). The three-pass regression filter: A new approach to forecasting using many predictors. Journal of Econometrics
    - Fan, J., Liao, Y. and Yao, J. (2015). Power Enhancement in High Dimensional Cross-Sectional Tests. Econometrica
    - Giglio, S. W. and Xiu, D. (2019). Asset Pricing with Omitted Factors. Journal of Political Economy
    - Giglio, Liao and Xiu (2020): Thousands of Alpha Tests. Review of Financial Studies
    - Giglio, Xiu and Zhang (2022): Test Assets and Weak Factors. Working paper
    - Anatolyev and Mikusheva (2020): Factor models with many assets: Strong factors, weak factors, and the two-pass procedure. Journal of Econometrics
    - Manresa, Penaranda, Sentana (2023): Empirical evaluation of overspecified asset pricing models. Journal of Financial Economics
    - Da. R., S. Nagel and D. Xiu (2023): The Statistical Limit of Arbitrage. Working paper
    - Kozak, S. and S. Nagel (2023): When do Cross-Sectional Asset Pricing Factors Span the Stochastic Discount Factor? Working paper
    - Lettau (2024). 3D-PCA: Factor Models with Restrictions. Working paper

3. High-frequency statistics

- Aït-Sahalia, Y. and Xiu, D. (2017). Using Principal Component Analysis to Estimate a High Dimensional Factor Model with High-Frequency Data. Journal of Econometrics
- Li, J., Tauchen, G., Lin, H. and Todorov, V. (2018) Rank Tests at Jump Events. Journal of Business and Economic Statistics.
- Andersen, T. G., Fusari, N. and Todorov, A. (2017). Short-Term Market Risks Implied by Weekly Options. Journal of Finance
- Da and Xiu (2021). When Moving-Average Models Meet High-Frequency Data. Econometrica

4. Covariance matrix estimation and risk modeling

    - Fan, J., Han, F., Liu, H., and Vickers, B. (2016). Robust Inference of Risks of Large Portfolios. Journal of Econometrics
    - Fan, J., Wang, W, and Zhong, Y. (2017). Robust Covariance Estimation for Approximate Factor Models. Journal of Econometrics
    - Ledoit, O. and Wolf, M. (2004). A Well-Conditioned Estimator for Large-Dimensional Covariance Matrices. Journal of Multivariate Analysis
    - Ao, Li and Zheng (2018). Approaching Mean-Variance Efficiency for Large Portfolios. Review of Financial Studies
    - Fan, Fan, Han and Lv (2020). Asymptotic Theory of Eigenvectors for Random Matrices With Diverging Spikes. JASA

5. Machine-Learning in Finance

    - Gu, Kelly and Xiu (2020): Autoencoder Asset Pricing Models. Journal of Econometrics
    - Bianchi, Buchner and Tamoni (2020): Bond Risk Premia with Machine Learning. Review of Financial Studies
    - Kelly, B., S. Pruitt, and Y. Su (2019). Characteristics Are Covariances: A Unified Model of Risk and Return. Journal of Financial Economics
    - Feng, G., Giglio, S. W. and Xiu, D. (2019). Taming the Factor Zoo. Journal of Finance
    - Freyberger, J., Neuhierl, A. and Weber, M, (2020) Dissecting Characteristics Nonparametrically. Review of Financial Studies
    - Kozak, S., Nagel, S. and Santosh, S. (2018). Shrinking the Cross-Section. Journal of Financial Economics
    - Kim, Korajcyk and Neuhierl (2020): Arbitrage Portfolios. Review of Financial Studies
    - Chinco, Clark-Joseph and Mao Ye (2019): Sparse Signals in the Cross-Section of Returns. Journal of Finance
    - Fan, Ke, Liao and Neuhierl 2021 - Structural Deep Learning in Conditional Asset Pricing. Working paper

- Harvey, Liu, and Zhu 2016- ... and the Cross-Section of Expected Returns. Review of Financial Studies
- Patton and Weller (2021). Risk Price Variation - The Missing Half of Empirical Asset Pricing. Review of Financial Studies
- Da, Nagel and Xiu (2022).The Statistical Limit of Arbitrage. Working paper
- Kelly, Malamud and Zhou (2022). The Virtue of Complexity in Return Prediction. Journal of Finance
- Kelly, Didisheim, Ke and Malamud (2023). Complexity in Factor Pricing Models. Working paper
- Kelly, Jensen, Malamud and Pedersen (2023). Machine Learning and the Implementable Efficient Frontier. Working paper
- Brandt, Goyal, Santa-Clara, and Stroud (2005). A simulation approach to dynamic portfolio choice with an application to learning about return predictability. Review of Financial Studies
- Gabaix, Koijen, Richmond and Yogo (2024). Asset Embeddings. Working paper
- Nagel (2024). Real-time Discovery of Return-Based Anomalies. Working paper
- Neuhierl, Jagannathan and Liao (2024). Robust Stock Index Return Predictions Using Deep Learning. Working paper
- Bell, Kakhbod, Lettau and Nazemi (2024). Glass Box Machine Learning and Corporate Bond Returns. Working paper
- Shen and Xiu (2024). Deep Autoencoders for Nonlinear Factor Models: Theory and Applications. Working paper
- Kelly, Kuznetsov, Malamud and Xu (2024). Artfifical Intelligence Asset Pricing Models. Working paper
- Shen and Xiu (2024). Can Machines Learn Weak Signals?. Working paper

6. Missing Data

- Bai and Ng (2020): Matrix Completion, Counterfactuals and Factor Analysis of Missing Data. JASA
- Jin, Miao and Su (2020): On factor models with random missing EM estimation, inference, and cross validation. Journal of Econometrics
- Chen, Fan, Ma and Yan (2019): Inference and uncertainty quantification for noisy matrix completion. PNAS
- Athey, Bayati, Doudchenko, Imbens and Khosravi (2020): Matrix Completion Methods for Causal Panel Data Models. Working paper
- Freyberger, Höppner, Neuhierl and Weber (2022): Missing Data in Asset Pricing Panel. Working paper
- Cahan, Bai, and Ng (2022): Factor-based imputation of missing values and covariances in panel data of large dimensions. Journal of Econometrics
- Yan, Chen, and Fan (2024). Inference for Heteroskedastic PCA with Missing Data. Annals of Statistics

7. Text Data

    - Kelly, Manela and Moreira (2020). Text Selection. Journal of Business and Economic Statistics
    - Ke, Kelly and Xiu (2020): Predicting Returns with Text Data. Working paper
    - Manela and Moreira (2017): News implied volatility and disaster concerns. Journal of Financial Economics
    - Bybee, Kelly, Su (2023). Narrative Asset Pricing - Interpretable Systematic Risk Factors from News Text. Review of Financial Studies
    - Aleti and Bollerslev (2022). News and Asset Pricing: A High-Frequency Anatomy of the SDF. Working paper
    - Bybee (2023). The Ghost in the Machine: Generating Beliefs with Large Language Models. Working paper
    - Bybee, Kelly, Manela and Xiu (2023). Business News and Business Cycles. Journal of Finance
    - Sarkar (2024). Economic Representations. Working paper

Preliminary Timetable

Date	Day	Topic
4/09/26	Thu	High-Dimensional Statistical Factor Modeling
4/10/26	Fri	High-Frequency Statistics
4/13/26	Mon	Machine-Learning in Finance
4/14/26	Tue	Missing Data
4/15/26	Wed	Student presentation

Learning Outcomes

1. Students should be able to derive the statistical properties for the most important estimators in financial statistics. Based on these techniques students should be in the position to develop their own new estimators and derive their properties.
2. Students should be able to weigh the advantages and disadvantages of different estimation approaches and choose the appropriate technique for an application.
3. Students should become familiar with working with financial data.
4. Students should be able to learn new estimation methods on their own by reading the relevant literature and to present their results in oral and written form.