Jianqing Fan

Frederick L. Moore '18 Professor of Finance

Professor of Statistics and Machine Learning

Professor of Operations Research and Financial Engineering

ORF 525: Statistical Foundations of Data Science

Spring Semester, 2023

MW 1:30pm - 2:50pm

Text Book	Details
	Fan, J., Li, R., Zhang, C.-H., and Zou (2020). **Statistical Foundations of Data Science.** (/fan/classes/525/TableOfContent.pdf) CRC Press. **Homepage of the book** (/DataScience/) To order the book **from amazon.com** (https://www.amazon.com/Statistical-Foundation-Monographs-Statistics-Probability/dp/ crid=1K4UN50WQQSQ1&dchild=1&keywords=statistical+foundations+of+data+science&qid=1601335402&sprefix=Statistical+ 1) or **from CRC Press** (https://www.routledge.com/Statistical-Foundations-of-Data-Science/Fan-Li-Zhang-Zou/p/book/978146

General Information

Instructor: Jianqing Fan, Frederick L. Moore'18 Professor of Finance.
Office: 205 Sherred Hall
Phone: 258-7924
E-mail: **jqfan@princeton.edu** (mailto:jqfan@princeton.edu)

Office Hours: Monday 3:00pm--4:00pm, Wednesday 10:30am--11:30am, or by appointments.

Precept: Arranged by the AI as needed

Assistants in Instruction (AIs):

- Bingyan Wang **bingyanw@princeton.edu** (mailto:bingyanw@princeton.edu), 258-8787, Office: 213 Sherred Hall
- Xiaonan Zhu (half) **xz8451@princeton.edu** (mailto:xz8451@princeton.edu), 258-9433, Office: 222 Sherred Hall
  - Office Hours and Locations.
    - Tuesday 1:30pm-2:30pm, Sherrerd Hall 003
    - Thursday 10:00am-11:00am, Sherrerd Hall 107
    - Friday 11:00am-12:00pm, Sherrerd Hall 107
- Financial Econometric Lab, 222 Sherred Hall, 258-9433
- Statistics Lab, 213 Sherred Hall, 258-8787

Text Book

- Fan, J., Li, R., Zhang, C.-H., and Zou, H. (2020). **Statistical Foundations of Data Science.** (/fan/classes/525/chapters1-3.pdf) CRC Press.
- Lectures are primarily based on the lecture notes which is taken from text book.

Reference Books

- James, G., Witten, D., Hastie, T.J., Tibshirani, R. and Friedman, J. (2013). *An Introduction to Statistical Learning with Applications in R* . Springer, New York.
- Hastie, T.J., Tibshirani, R. and Friedman, J. (2009). *The elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed). Springer, New York.
- Buehlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer, New York.
- Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity*. CRC press, New York.
- Wainwright, M. J. (2019). *High-dimensional statistics: A non-asymptotic viewpoint.* Cambridge University Press.

Syllabus

This course gives in depth introduction to statistics and machine learning theory, methods, and algorithms for data science. It covers multiple regression, kernel learning, sparse regression, sure screening, generalized linear models and quasi-likelihood, covariance learning and factor models, principal component analysis, supervised and unsupervised learning, deep learning, and other related topics such as community detection, item ranking, and matrix completion. Applicability and limitations of these methods will be illustrated using mathematical statistics and a variety of modern real world data sets and manipulation of the statistical software R.

Course material will be covered the following topics; some topics will be assigned as reading materials.

1. Rise of Big Data and Dimensionality*
    - Impact of Big Data;
    - Impact of Dimensionality
    - Aims of High-dimensional statistical learning
    - Aims of Big Data
    - **Chapters 1--3** (/fan/classes/525/chapters1-3.pdf)
2. Multiple and Nonparametric Regression
    - Multiple Linear Regression
    - Model Building and Basis Expansions
    - Ridge Regression
    - Reproducing Kernel Regression
    - **Assigned Reading** (/document/1231), **Lecture Notes 1** (/document/1236), **Homework 1** (/document/1241)
3. Penalized Least Squares
    - Best subset and L_0 penalty
    - Folded-concave Penalized Least Squares
    - Lasso and L_1-regularization
    - Numerical Algorithms
    - Regularization parameters
    - Refitted Cross-validation
    - Extensions to Nonparametric Modeling
    - **Lecture Notes 2** (/fan/classes/525/Notes2.pdf), **Homework 2** (/fan/classes/525/Homework2.pdf)
4. Generalized Linear Models and Penalized Likelihood
    - Generalized Linear Models
    - Variable Selection via Penalized Likelihood
    - Numerical Algorithms
    - Statistical Properties
5. Feature Screening
    - Correlation Screening
    - Generalized and Rank Correlation Screeing
    - Nonparametric Screening
    - Sure Screening and False Selection
6. Supervised Learning
    - Model-based Classifiers
    - Kernel Density Classifiers and Naive Bayes
    - Nearest Neighbor Classifiers

- Classification Trees and Ensemble Classifiers
- Support Vector Machine
- Sparsier classifiers
- Sparse Discriminant Analysis
- Sparse Additive Classifiers

7. Unsupervised Learning
   - Cluster Analysis
   - Variable Selection in Clustering
   - Choice of Number of Clusters
   - Sparse PCA
8. Introduction to Deep Learning
   - CNN and RNN
   - Generative adversary networks
   - Training Algorithms
   - A Glimpse of Theory
9. Covariance Regularization and Graphical Models
   - Sparse Covariance Matrix Estimation
   - Robust Covariance Inputs
   - Sparse Precision Matrix and Graphical Models
   - Latent Gaussian Graphical Models
10. Covariance Learning and Factor Models
    - Principal Component Analysis
    - Factor Models and Structured Covariance Learning
    - Covariance and Precision Learning with Known Factors
    - Augmented Factor Models and Projected PCA
    - Asymptotic Properties
11. Applications of PCA and Factor Models
    - Factor-adjusted Regularized Model Selection
    - Factor-adjusted Robust Multiple Testing
    - Augmented Factor Regression
    - Applications to Statistical Machine Learning

Computation

The software package for this class is **R** (https://www.r-project.org/) or **RStudio** (https://rstudio.com/). See R-labs below. Most of computation in this class can be done through a laptop. Laptops with wireless communication off can be used during the exams, and so are the calculators.

Attendance

Attendance of the class is required and essential. The course materials are mainly from the notes. Many conceptual issues and statistical thinking are only taught in the class. They will appear in the midterm and final exams.

Homework

Problems will be assigned through **Canvas** (https://canvas.princeton.edu/) approximately biweekly and submitted online. No late homework will be accepted. Missed homework will receive a grade of zero. The homework will be graded, and each assignment carries equal weight. You are allowed to work with other students on the homework problems, however, verbatim copying of homework is absolutely *forbidden*. Therefore each student must ultimately produce his or her own homework to be handed in and graded.

Exams

There will be one in-class midterm exam, and a final exam. All exams are required and there will be no make-up exams. Missed exams will receive a grade of zero. All exams are open-book and open-notes. Laptops with wireless off and calculators may be used during the exams.

Schedules and Grading Policy

Assignment	Schedule
Homework (25%)	Various due dates (approx 5 sets)
Midterm Exam (25%)	Wednesday, March 22, 2023 (1:30pm--2:50pm, in class)
Final Exam (50%)	9:00am--12:00pm, Friday, May 5, 2023 (tentative)

R-labs

The following files intend to help you familiar with the use of R-lab commands.

Here are some useful materials too.

- **An Introduction to R** (http://cran.r-project.org/doc/manuals/R-intro.pdf), by W. N. Venables, D. M. Smith and the R Core Team.
- **U-Tube video: An introduction to R** (https://www.youtube.com/playlist?list=PLOU2XLYxmsIK9qQfztXeybpHvru-TrqAP)
- **Labs 1-5: Basic skills** (/fan/classes/504/labs/lab1-5.pdf)and **their associated data set** (/fan/classes/504/labs/boston.housing.dat)(Boston housing data)
  - The following extended skills are not used in the class, but is provided here for your convinience.
    - **Extended Skills: ANOVA** (/fan/classes/504/labs/anova.pdf)and **their associated data set** (/fan/classes/504/labs/labor.suppl.dat)(labor data).
    - **Extended Skills: GLIM** (/fan/classes/504/labs/glim.pdf)and **their associated data set** (/fan/classes/504/labs/burn.dat) (burn data). **Description of the data set** (/fan/classes/504/labs/burn.des.txt)
- **Lab 6: Linear time series analysis** (/fan/classes/504/labs/lab6.pdf)
- **Lab 7: Discrete volatility models** (/fan/classes/504/labs/lab7.pdf)
- **Lab 8: Capital Asset Pricing Model** (/fan/classes/504/labs/lab8.pdf)

Data Sets used in the class

- **Zillow House Price Prediction: training data** (/fan/classes/525/DataSets/train.data.csv)
- **Zillow test data** (/fan/classes/525/DataSets/test.data.csv), **Source and Details** (https://www.kaggle.com/c/zillow-prize-1/data)
- **Monthly Macroeconomics Data** (/fan/classes/525/DataSets/Macro2019-01.csv), **Source and Details** (https://research.stlouisfed.org/econ/mccracken/fred-databases/)
- **Transformed Macroeconomics Data** (/fan/classes/525/DataSets/Macro2019-01-transformed.csv), **Details of transformation and Meanings of variable** (https://s3.amazonaws.com/files.fred.stlouisfed.org/fred-md/Appendix_Tables_Update.pdf)
- **Autism Data** (/fan/classes/525/DataSets/autism.csv), Expressions of top 5 differently expressed genes and other variables
- **Neuroblastoma Data** (/fan/classes/525/DataSets/Neuroblastoma/neuroblastoma.csv), Gene expressions of 246 neuroblastoma patients and indicator whether a patient has a 3-year event free survival
- **mice protein expression data (preprocessed)** (/fan/classes/525/DataSets/mice-protein-expressions.csv)
- **Image Data** (/fan/classes/245/DataSets/pictures.zip): 500 photos with people in the pictures and 500 photos without people in pictures and its associated R-code to preprocess the data **human.r** (/fan/classes/245/DataSets/human.r)
- **Yields of SP500 and more** (http://www.multpl.com/s-p-500-dividend-yield/)
- **Expenditure and Personal Expenditure and Other Macroeconomics Data** (http://research.stlouisfed.org/fred2/series/PCE/downloaddata?cid=110)