Syllabus

High-Dimensional Statistical Methods

Teacher

Quentin Paris

Assistant Professor
National Research University Higher School of Economics
Faculty of Computer Science
School of Data Analysis and Artificial Intelligence
& Laboratory of Stochastic Analysis and its Applications
Moscow, Russia

qparis@hse.ru
http://www.qparis-math.com/teaching

Content of the course

Due to increasingly powerful computing technologies, modern data analysis faces the constant challenge of dealing with very high-dimensional and complex data. Common examples include network data, images, videos, financial data or microarrays in genetics. Many standard statistical methods usually fail in such contexts both on a theoretical and computational level. Popular methods to overcome the so called *curse of dimensionality* (referred to as dimension reduction techniques) usually involve exploiting (potentially) simple and informative structures of complex data. Among such simple structures, sparsity is a characteristic of high-dimensional data that can be efficiently exploited by statistical methodology. A large amount of effort and techniques have been developed by the research community in order to develop statistically and computationally efficient methods allowing to leverage the sparsity of high-dimensional data. The first part of this course (Chapter 1) will provide an introduction to these methods. The second part of the course (Chapters 2 and 3) will provide an introduction to developments and applications, of the questions discussed in Chapter 1, to more complex settings.

chapter 1 - high-dimensional regression

1. Least squares
2. Constrained least squares
3. Penalized least squares
3.1. Bayes Information Criterion (Bic)
3.2. Least Absolute Shrinkage and Selection Operator (Lasso)

chapter 2 - matrix estimation

1. Matrix regression and completion (a.k.a. the Netflix problem)
2. Thresholding estimators
3. Penalized approaches
4. Additional topics
4.1. Covariance matrix estimation with applications
4.2. Probabilistic PCA

chapter 3 - graphical models

1. Directed and non-directed acyclic graphs
2. Gaussian graphical models
3. Estimation of the precision matrix
4. Additional topics

General instructions

The students are encouraged to spend (on their own or in group discussions) at least twice the time of the lectures to read, understand, learn and challenge the material presented in class. In other words, it is expected that each student will spend at least 2 hours (minimum) reviewing 1 hour of lecture material. It is naturally expected that a serious review of the material studied in class should give rise to many questions. The first 10 to 15 minutes of each lecture will be devoted to the discussion and answering of these questions in the form of a regulated group discussion. The regular homework activities are also intended to answer potential questions of the students and provide additional information.

Evaluation process

The final evaluation grade g of the students will be computed based on:

t: a written final test grade,
p: a personal project grade,
h: a home assignment grade,
c: a class participation grade,

according to the formula:

$$g = \frac{3t + 2p + 3h + 2c}{10}.$$

Final written test

The final written test will consist in a 3 hour test that students will perform individually, at the same time and under the supervision of the teacher. During this test, only a single hand-written A4 sheet of paper (front and back) will be allowed for the students to collect lecture material and results. All other documents or electronic devices will be strictly forbidden. The students are expected to work in total silence and to provide manuscript answers in the order of the given questions, on ruled paper with their name and surname.

Personal project

Each student is asked to select one of the following two options:

• `Option 1: Research article` – The student selects a research article of his choice (a list will be provided by the teacher). The personal project will consist in reading (and understanding) the article and present, in 20 minutes, the main matter. Due to the technicality of some research articles, the evaluation of the oral presentation will, mostly (but not only), focus on oral and presentation abilities.

• `Option 2: Implementation` – The student selects a real-life problem and data set of his choice and implements (in R, Python or C) statistical methods seen in class. The results should be presented in 20 min (oral presentation).

Home assignements

Students will be given two home assignments during the course. Each HA can be performed individually or in groups of two people (maximum). The HA should be handed over in time, can be hand-written or typed, with name and surname of participants. Reports are expected to be very well presented.

Classroom intellectual engagement and participation

Students are expected to be present at all lectures, arrive in time and pay close attention to lectures and seminars. Students are also expected to ask questions and try to answer those given in class by the teacher. At the beginning of each lecture, 10 to 15 minutes will be specifically devoted to the discussion and answering of questions concerning the previous lectures (or home assignments) in the form of a regulated group discussion.

Make-up, absence and howework policy

- If the final evaluation grade (described above) does not meet the minimal requirements, a final oral make-up test will be given to the concerned students as a final chance to validate the course. In case of success at the make-up test, the students will only be given the minimal grade authorizing to validate the course.

- Potential absence at the written tests should be justified by exceptional medical reasons. In such a situation, official medical certificates are expected to be handed-out as soon as possible. Only for those students able to justify officially their absence, an additional (final or make-up test) will be prepared. Students unable to justify their absence will unfortunately be given the grade 0 to the related test.

- Howe assignments will be given in advance with a large preparation time. As a result, they are expected to be handed out on time and no additional delay will be allowed. Students unable to hand-out their work in time will be given the grade 0 to the related assignment.