Seminar on
High Dimensional Probability Theory:
A Data Science Perspective

summer term 2023

Dr. Tobias Kley and Prof. Dr. Axel Munk

Key information

Time:	14/04/2023 – 14/07/2023, on Fridays, 10.15-11.45
Format:	in person, room 5.101 (IMS)
Possible Modules:	B.Mat.3441: Seminar on applied and mathematical stochastics
	M.Mat.4841: Seminar on applied and mathematical stochastics
	B.Mat.3444: Seminar on mathematical statistics
	M.Mat.4844: Seminar on mathematical statistics
	B.Mat.3447: Seminar on statistical foundations of data science
	M.Mat.4847: Seminar on statistical foundations of data science
Instructors:	Dr. Tobias Kley and Prof. Dr. Axel Munk
Intended Audience:	**Bachelor and beginning Master students**
Language:	English

Prerequisites

Participants must have successfully attended:

- Analysis I (B.Mat.0011),
- Analysis II (B.Mat.0021),
- Analytische Geometrie und Lineare Algebra I (B.Mat.0012),
- Maß- und Wahrscheinlichkeitstheorie (B.Mat.1400).

Description

In many contemporary applications (such as high throughput genetics, financial risk assessment or language processing), the data to be analyzed not only has a large number of observations, but also a large number of variables (i. e., the observed data vectors have a high dimension). In such situation, which are ubiquitous nowadays, traditional statistical methods designed for low dimensional methods (such as maximum likelihood) fail.

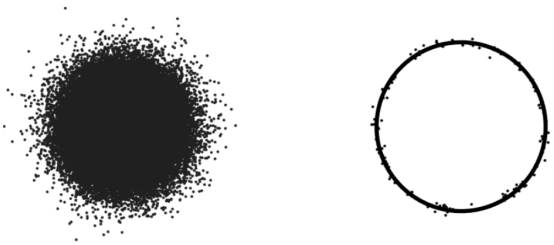

Figure 3.6 A Gaussian point cloud in two dimensions (left) and its intuitive visualization in high dimensions (right). In high dimensions, the standard normal distribution is very close to the uniform distribution on the sphere of radius $\sqrt{n}$.

Source: Vershynin (2018), p. 53.

Instead, modern data analysis tools build on (implicit or explicit) dimension reduction techniques which are indispensable for the analysis of high dimensional and complex data. Fundamental to this are tools and recent results from high dimensional probability theory, the topic of this seminar.

However, they also are of interest by itself as they often provide surprising insight into problems of high dimensional geometry and analysis.

Our focus will be on carefully understanding the basic mathematical principles rather than obtaining results in most generality. Topics include: concentration inequalities for random variables, vectors and matrices (such as Hoeffding's, Chernoff's, Khintchine's, Bernstein's and Hanson-Wright inequality), sub-Gaussian and sub-exponential distributions, isotropic distributions, matrix norms, covering and packing numbers, and isoperimetric inequalities. Applications to be discussed include: semidefinite programming, the kernel trick, random graphs, such as in the stochastic block model, covariance estimation, spectral clustering, Johnson-Lindenstrauss lemma and compressive sensing, community detection in sparse networks, and matrix completion.

This seminar complements the lecture Foundations of Statistical Data Science II (B.Mat.3147) by Prof. A. Munk as well as the lecture Stochastik (B.Mat.2410) by Dr. T. Kley. However, because the material discussed in the seminar is quite elementary, attending the lectures might be helpful but is not necessary for understanding.

Application and admission

To provide participants with the material to be presented at an early stage, we ask you to preregister for this seminar. To this end, please email Tobias Kley (tobias.kley@uni-goettingen.de) and indicate your interest to give a seminar talk. Please include information about relevant courses you have taken in your email. In particular, recall the prerequisites mention above. Deadline for preregistration is 20 March 2023 (12pm/noon).

A preparatory virtual meeting, during which topics will be assigned to participating students, is scheduled for 24 March 2023 (2pm–3pm). Notably, the seminar is limited to 14 participants. Should preregistrations exceed 14, then participants will be chosen based on the information provided in their preregistration email.

Recommended literature

Main Reference

Topics for presentations will be assigned along the lines of

- R. Vershynin (2018): High-Dimensional Probability An Introduction with Applications in Data Science, Cambridge University Press.
  A final draft of the book is available for free download from the author's homepage:
  https://www.math.uci.edu/~rvershyn/papers/HDP-book/HDP-book.pdf

Background material from probability and statistics

- P. Billingsley (1999): Probability and Measure. Wiley.
- A. Klenke (2013): Wahrscheinlichkeitstheorie. Springer.
- A. Munk (2022/23): Statistical Foundations of Data Science. Lecture Notes IMS.