Advanced Regression: 1a Overview and motivations

Verena Zuber

Epidemiology and Biostatistics, Imperial College London

17th January 2019

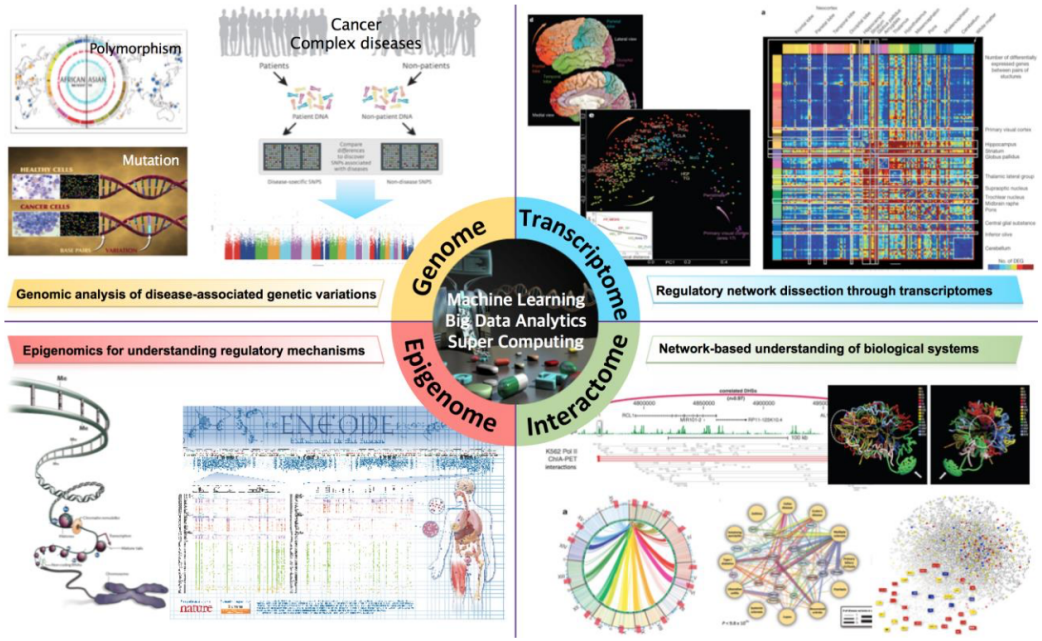
Advanced Regression: Motivation

- Course aims
- High-dimensional data
- Omics data
- Classical approaches and high-dimensional statistics

Advanced Regression: Course details

- Practicals
- Timetable
- Learning outcomes and exam
- Questions?

Advanced Regression: Course aims



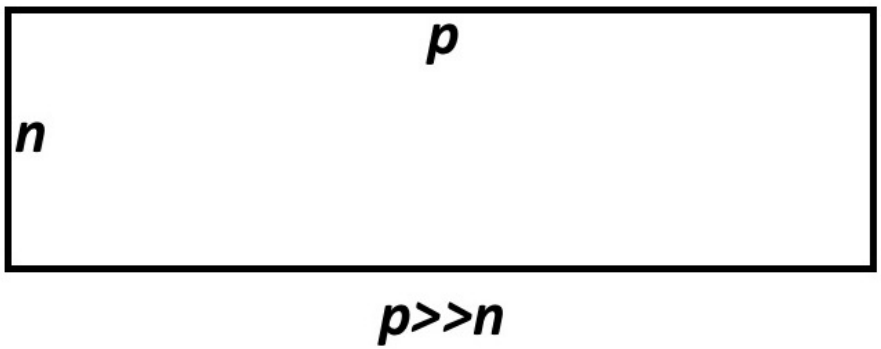
- Learn principles of advanced regression for high-dimensional data analysis.
- Apply these techniques on real-world data problems.

Motivation: High-dimensional data

- Number of samples or observations: n
- Number of variables: p

Data types:

- Big data: $p >> n$

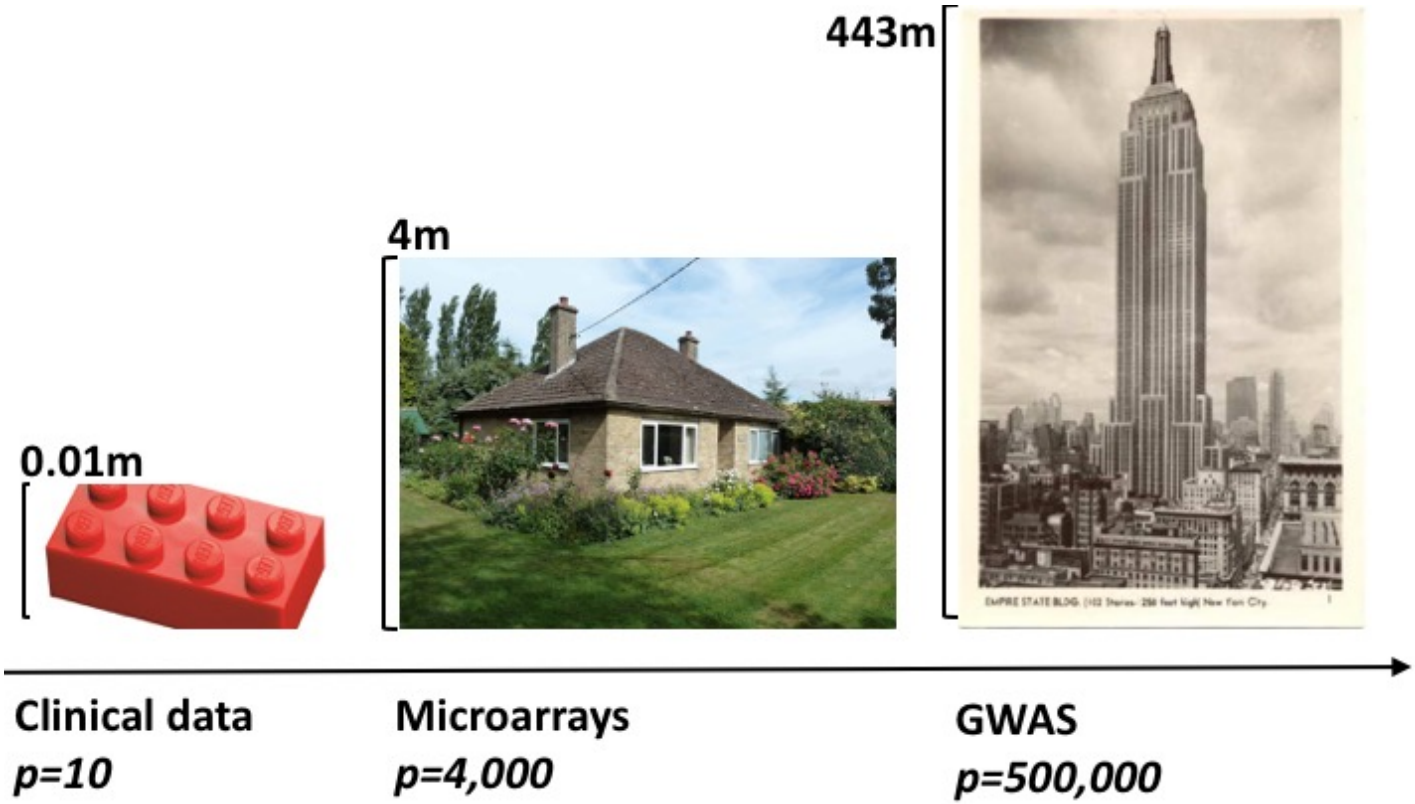


- (Tall data: Summary-level data $p \times 1$)

High-throughput technologies

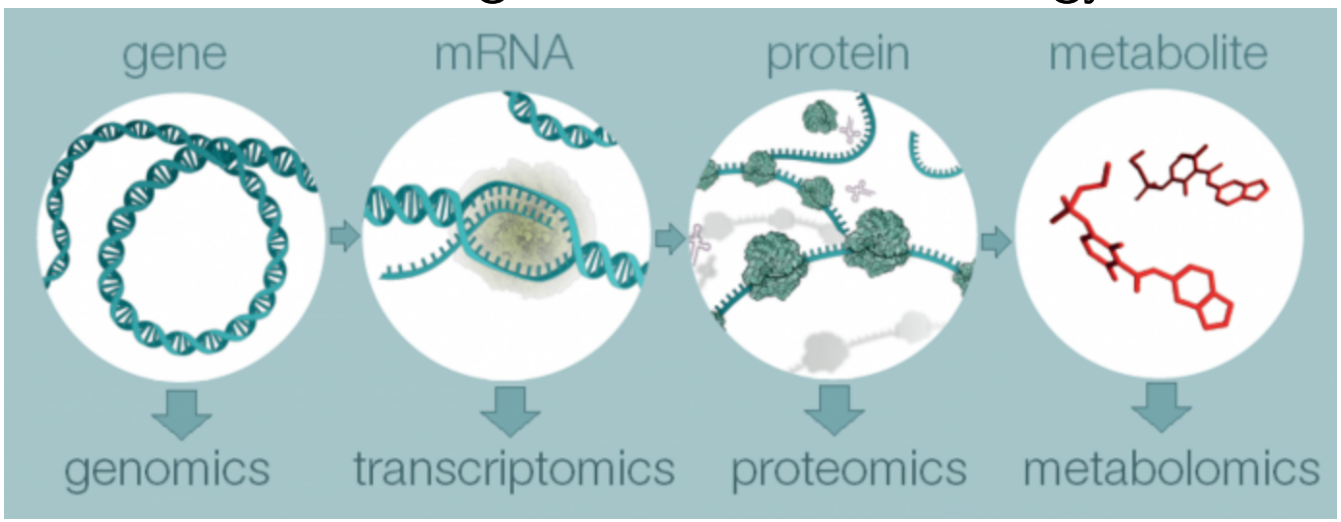
- Automated processes of classical cell biology techniques.
- Allow to capture a large number of molecular entities:
  - Transcriptomics: $21,000$ genes
  - Genomics: $500k$ genetic variants on customised arrays, $30m$ genetic variants in UK biobank
- Technologies:
  - Microarrays
  - Next-generation sequencing

High-throughput data



Omics data

Central dogma of molecular biology:



Omics data

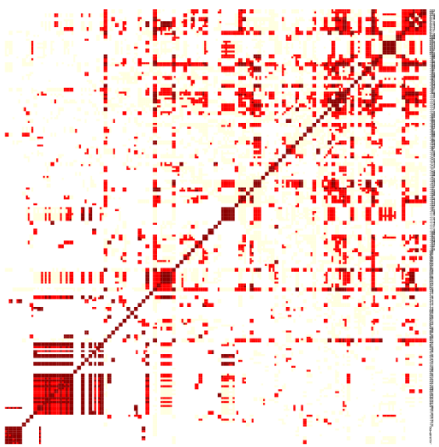
- **Genomics**: The study of genetic variants or single nucleotide polymorphisms (SNPs), i.e. a variation in a single nucleotide that occurs at a specific position in the genome, where there is a certain variation within a population.
- **Transcriptomics**: Set of all RNA molecules, including gene-expression but also micro RNA or small non-coding RNA ('junk RNA').

Omics data

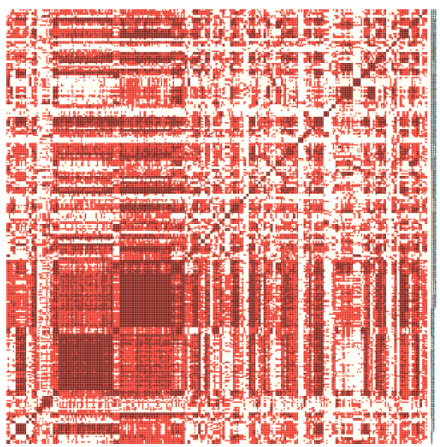
- **Epigenomics**: Study of heritable changes in gene expression that do not involve changes to the underlying DNA sequence; prominent examples are transcription factors, methylation, and histone modifications.
- **Proteomics**: Study of large molecules, or macromolecules, consisting of one or more long chains of amino acid residues.
- **Metabolomics**: Study of small molecules, substrates and products of metabolism, which are influenced by both genetic and environmental factors.

Omics data: Characteristics

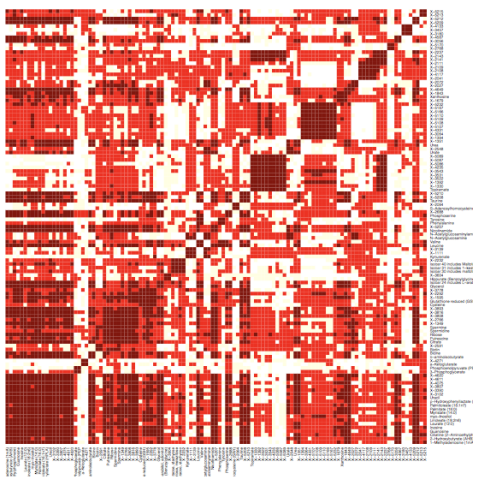
- Big data: $p >> n$
- Complex correlation structure:



Genomics



Transcriptomics



Metabolomics

Why consider omics data?

1. Biomarker discovery: **To predict** a subjects disease risk, or a patients survival or disease progression (Personalised medicine).
   Example: Genetic mutations in the *BRCA1* and *BRCA2* gene regions predict risk of breast cancer.
2. Disease associations: **To understand** molecular mechanisms and to explain the molecular basis of disease.
   Example: Oncogene *MYC* upregulates the expression of a number of factors involved in cell growth and proliferation in prostate cancer.

Other high-dimensional data types

- Health data records, e-records
- Health and fitness apps, location tracker
- Imaging data, e.g. functional and structural fMRI studies
- Credit scoring based on credit files and personal data
- Recommender systems based on user ratings

Modern data science is build on advanced regression models!

Classical statistical approaches

- Provide unbiased estimates of relevant parameters.
- Are designed for datasets with more observations than variables ($n > p$).
- Examples:
  - Linear regression
  - Least squares estimate

$$\hat{\beta}_{LS} = \underbrace{(x^t x)^{(-1)}}_{p \times p} \underbrace{x^t}_{p \times n} \underbrace{y}_{n \times 1}$$

  - Maximum likelihood (Lecture 1b)

Why do classical statistical approaches fail?

- Linear regression requires an inversion of $(x^t x)$, a $p \times p$ matrix.
- Numerical instabilities when $p \approx n$ and not invertible when $p > n$.
- Complex correlation structures can impact the inversion as well.
- Bias-variance trade-off: Classical approaches are unbiased, but have a large variance.

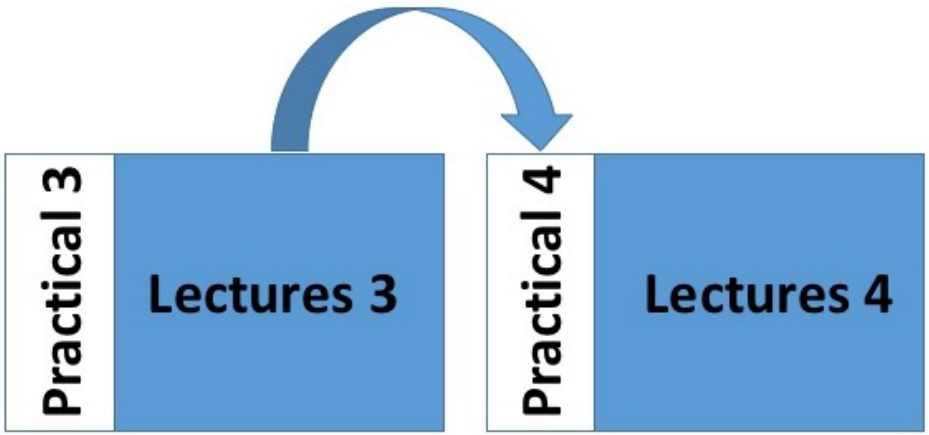
New data structures require new methods!

Methods covered in this course

- Penalised regression (Ridge, lasso, and elastic net)
- Shrinkage estimation
- False discovery rates
- Building a prediction rule and cross-validation
- Random effects and hierarchical models
- Non-linear methods (lowess, spline, GAM)
- Non-parametric methods (decision trees and random forests)

Advanced Regression: Practicals

- Structure: First lectures, then practicals with one week delay



- Time to digest the lectures and read up, formulate question.
  - Practicals are not assessed.
- But they are an essential part of the course and will help you for the exam.
  - The practicals will be in R.

- Please bring your laptops with R installed!

Why use R?

- R is a language and environment for statistical computing and graphics.
- R is free and published under the GNU licence.
- 13,663 available add-on packages.
- Please download and be prepared to run analysis before the practicals.
- `https://cran.r-project.org/`

Useful links to R

- There are many documentations on R freely available: `https://cran.r-project.org/doc/contrib`
- R Reference Card 2.0 `https://cran.r-project.org/doc/contrib/Baggott-refcard-v2.pdf`
- Search engine: `rseek.org`
- R blog: `https://www.r-bloggers.com/`
- User groups:
  - LondonR `www.londonr.org`
  - Rladies London `https://www.meetup.com/rladies-london/`

What is markdown?

- When coding in R it is important to document and comment the code.
- Markdown is an R package that compiles R code into documents (html, word and many more).
- Package `https://cran.r-project.org/web/packages/rmarkdown`
- Project page `https://rmarkdown.rstudio.com/`

Make your code accessible and reproducible.

Markdown can help you with that.

Week 16: 17th January

11:00-11:50	Lecture 1a	Overview and motivation	Verena Zuber
1:30-2:20	Lecture 1b	Statistical learning - likelihood approach	David Muller
2:40-3.30	Lecture 1c	Simple generalised linear models	Verena Zuber

Week 17: 24th January

9:00-10:40	Practical 1	Using R to analyse data with linear models	Tutors
11:00-11:50	Lecture 2a	Introduction to variable ranking and selection	Verena Zuber
1:30-2:20	Lecture 2b	Variable selection with correlated predictors	Verena Zuber
2:40-3.30	Lecture 2c	Multiple testing	Verena Zuber

Week 18: 31th January

9:00-10:40	Practical 2	High-dimensional regression in R	Tutors
11:00-11:50	Lecture 3a	Principles of high-dimensional data analysis	Verena Zuber
1:30-2:20	Lecture 3b	Penalised regression models (lasso, elastic net)	Verena Zuber
2:40-3.30	Lecture 3c	Prediction accuracy and cross-validation	Verena Zuber

Week 19: 7th February

9:00-10:40	Practical 3	Cross-validation and variable selection in R	Tutors
11:00-11:50	Lecture 4a	Random effects and hierarchical models	Verena Zuber
1:30-2:20	Lecture 4b	Non-linear and non-parametric models - part I (lowess, spline, GAM)	Deborah Schneider-Luftman
2:40-3.30	Lecture 4c	Non-linear and non-parametric models - part II (decision trees and random forests)	Deborah Schneider-Luftman

Week 20: 14th February

9:00-10:40	Practical 4	Random effects models in R	Tutors
11:00-11:50		Questions and answers session	Instructors
1:30-3:10	Practical 5	Non-linear and non-parametric models in R	Tutors

Advanced Regression: Learning outcomes

- Perform advanced statistical analyses, employing penalised likelihood or non-pararametric regression models.
- Discuss the theoretical foundations and limitations of the most widely used advanced regression approaches.
- Identify the challenges of high-dimensional data analysis.
- Identify suitable analysis strategies to address the problems arising from 'small n, large p' data sets.
- Use complex regression models in R, understand which methods are suitable for which data, know the pitfalls of high-dimensional data analysis, and interpret the results.

Advanced Regression: Exam

- This module will be assessed by a 1.5 hour written exam.
- The exam is taking place on Friday, 3rd May 2019.
- Practical sessions offer regular opportunity for receiving formative feedback from the tutors.
- A Q & A session will be scheduled on the last day of the module (14th February).

Advanced Regression: Questions?

Questions?

Please get in touch:

v.zuber@imperial.ac.uk