Analysis of high-dimensional data: Opportunities, challenges and goals

Riccardo De Bin[1]

Department of Mathematics - University of Oslo

[1]on behalf of the TG9 – High-Dimensional Data of the STRATOS initiative

Outline of the talk

- Introduction
- Main
- Remarks

Introduction: TG9 – high-dimensional data

About TG9 – high-dimensional data: **Aim and scope**

From the web page (`www.stratos-initiative.org/group_9`):

[...] The goal of the 'high-dimensional data' topic group of the STRATOS initiative (TG9) is to provide guidance amid the jungle of opportunities and pitfalls inherent in the analysis of high-dimensional biological and medical data. [...] in-depth evaluation and discussion of various statistical and computational approaches aim to reinforce concepts and support specific recommendations for best practices. [...]

Introduction: TG9: high-dimensional data

About TG9 – high-dimensional data: **Who are we?**

Currently 11 members from 7 countries:

- Federico Ambrogi (University of Milano);
- Axel Benner (DKFZ Heidelberg);
- Harald Binder (Freiburg University);
- Anne-Laure Boulesteix (LMU Munich);
- Riccardo De Bin (University of Oslo);
- Lara Lusa (University of Primorska);
- Lisa McShane (National Cancer Institute Washington);
- Stefan Michiels (Institute Gustave Roussy)
- Eugenia Migliavacca (Nestlé Institute Lausanne)
- Jörg Rahnenführer (TU Dortmund);
- Willi Sauerbrei (Freiburg University);

Introduction: TG9: high-dimensional data

About TG9 – high-dimensional data: **Who are the chairs?**

From the start	**From the start**	**From this year**
Lisa McShane	Jörg Rahnenführer	Riccardo De Bin



many slides are taken from Jörg's talk at ISCB42 (Lyon, 2021)

Introduction: Definition

When a dataset is “high-dimensional”?

- More than 10 variables?
  - interpretation difficulties;
  - basic methods (e.g., kNN) start to fail;
- More variables (p) than observations (n)?
  - popular methods (e.g., OLS) fail;
- Large n?
  - computational (e.g., matrix inversion) issues;
- Large n and large p?
  - computational and memory issues.

Introduction: Prediction with high-dimensional data

- Main situation: Many more variables than samples ($p >> n$).
- Prediction models (regression, classification, survival):
  - Bias-Variance trade-off;
  - "Model fit" vs "Model complexity".
- Solutions for high-throughput data with variable selection:
  - Filtering: Select "best" variables before modelling;
  - Wrapping: Select variables "within" modelling algorithm.
- Problems:
  - Curse of dimensionality;
  - Overfitting.

Main: Overview manuscript

Currently working on a manuscript:

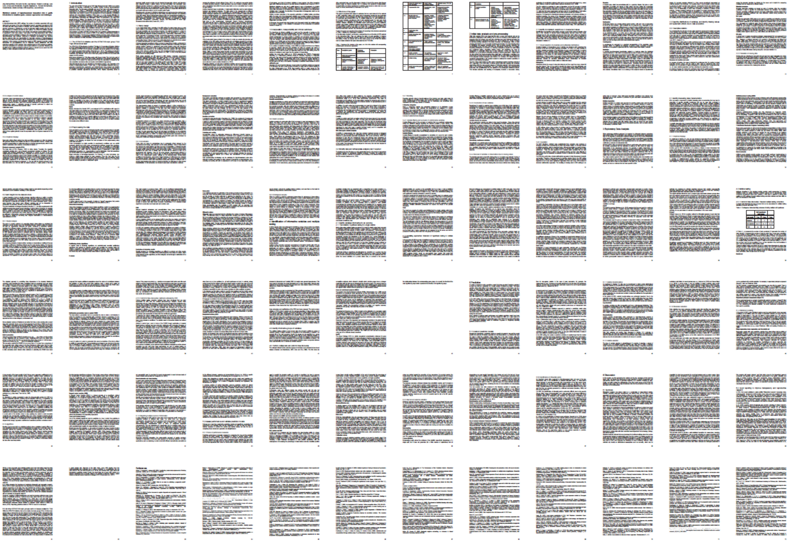
- Title: *Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges*;
- Authors: basically all TG9 members;
- discuss in particular where methods developed for low-dimensional data are inadequate in high-dimensional data (hereafter, HDD) settings.

- Long term project, almost finished:

  "Trees that are slow to grow bear the best fruit."

  (Molière, French playwright, 17^{th} century):

Main: Overview manuscript



Main: Overview manuscript

Table of contents:

Table 1 of the manuscript:

- Overview of the structure of the paper, as a list of the sections with corresponding analytical goals, common approaches, and examples.

Main: Initial data analysis and preprocessing

2 Initial data analysis and preprocessing:

Sec.	Analytical goals	Common approaches	Examples
2.1	Identify inconsistent, suspicious or unexpected values	Visual inspection of univariate and multivariate distributions	Scatterplots, histograms, boxplots, heatmaps, correlograms, RLE plots, MA plots
2.2	Describe distributions of variables, identify missing values and systematic effects due to data acquisition	Descriptive statistics, tabulation, analysis of batch controls, graphical displays, distribution of summary measures	Measures for location and scale, bivariate measures, calibration curve, PCA, Bi-plot

Main: Initial data analysis and preprocessing

Sec.	Analytical goals	Common approaches	Examples
2.3	Preprocess the data	Normalization, batch correction	Background correction, baseline correction, centering, scaling, quantile normalization, ComBat, SVA
2.4	Simplify data and refine/update analysis plan if required	Recoding, variable filtering, construction of new variables, removal of variables or observations, imputation	Collapsing categories, variance filtering, discretizing continuous variables, multiple imputation

Main: Initial data analysis and preprocessing

- Important first step in (every) data analysis, can be particularly challenging in HDD settings;
- "data preprocessing" is a term used in biomedical HDD settings, especially in the omics field;
- for HDD, a detailed examination of the distribution of every variable individually is rarely feasible,
  - instead calculate scores and select interesting cases;
- HDD typically contains uninformative variables that do not vary much across subjects (variability only reflecting noise),
  - standardization can exaggerate the noise;
- issues with missing values in HDD settings:
  - multiple imputation approaches typically performs poorly;
  - complete case analysis may exclude too many observations.

Main: Exploratory data analysis

3 Exploratory data analysis:

Sec.	Analytical goals	Common approaches	Examples
3.1	Identify interesting data characteristics	Graphical displays, descriptive univariate and multivariate statistics	PCA, Bi-plot, multidimensional scaling, t-SNE, neural networks
3.2	Analyze data structure	Cluster analysis, prototypical samples	Hierarchical clustering, k-means, PAM

Main: Exploratory data analysis

- Visual identification of interesting characteristics of HDD requires specialized graphical displays / data reduction;
- difficulties with cluster analysis in HDD:
  - mixtures of low-dimensional parametric probability distributions cannot be applied at all or perform very poorly;
  - partitioning algorithms can present computational challenges;
  - algorithms may not converge to a solution;
  - hard to identify the right number of cluster (e.g., scree plots suffer from noise accumulating over the variables.

Main: Identification of informative variables and multiple testing

4 Identification of informative variables and multiple testing:

Sec.	Analytical goals	Common approaches	Examples
4.1	Identify informative variables for an outcome	Test statistics and modelling	t-test, c2-test, limma, DESeq, edgeR
4.2	Multiple testing	Perform multiple tests, control for false discoveries	Holm-Bonferroni, BH, q-value
4.3	Identify informative groups of variables	Perform multiple tests, control for false discoveries	Gene set enrichment analysis, global test, topGO, Holm-Bonferroni, BH

Main: Identification of informative variables and multiple testing

- Frequent goals in HDD settings:
  - identify variables related to a single outcome;
  - identify variables with a trajectory over time affected by experimental factors or exhibiting a prescribed pattern;
  - identify variables related to other variables;
- Specific methods developed for testing in HDD:
  - for hypothesis testing for single variables (e.g. limma, edgeR, DESeq2) – borrow information among variables;
  - for multiple testing correction: control of false discovery rate.

Main: Prediction

5 Prediction:

Sec.	Analytical goals	Common approaches	Examples
5.1	Construct prediction models	Variable transformations, variable selection, dimension reduction, statistical modelling, algorithms	Log-transform, supervised PC, ridge, lasso, elastic net, boosting, SVM, trees, random forest, neural networks, deep learning
5.2	Assess performance and validate prediction models	Choice of performance measures, internal and external validation	MSE, MAE, ROC curves, AUC, calibration curves, Brier score, deviance, cross-validation, subsampling, Bootstrap, use of external datasets

Main: Prediction

- Numerous dramatic claims of performance of prediction models have been made using HDD,
  - not always withstand careful validation;
- Need to identify the optimal level of model complexity:
  - that will yield interpretable models;
  - good prediction performance on independent data;
- Balance between overfitting . . .
  - too specific to the data at hand;
  - identifies more or too complex patterns than real ones;
- . . . and underfitting,
  - misses important patterns useful for prediction.

Main: Prediction

- Approaches used to overcome issues related to the $p >> n$ situation in the model building process:
  - variable pre-selection (SIS, ISIS, ...);
  - dimensionality reduction (PCR, LSR, ...);
  - constrained optimization (LASSO, ridge, ...);
  - algorithmic approaches (random forests, neural networks, ...).
- Other challenges particularly interesting in the HDD context:
  - data integration (data sources, external knowledge, ...);
  - identification of influential points;
  - evaluation of prediction models.

Remarks

Back to TG9. Other projects / topics not considered here:

- simulations in HDD:
  - difficult to simulate realistic correlation structure and suitable multivariable distributions;
  - some characteristics of HDD are not uniquely defined;
  - use of plasmode data (real data suitably manipulated);
- reporting / transparency / reproducibility.

Visit https://www.stratos-initiative.org/group_9