Topic group 9 'High-Dimensional Data": updates and plans

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

- Updates
- Current work
- Plans for the future

Updates: members

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

Currently 11 members from 7 states:

- 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).

Updates: members

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

Pending applications:

- Early career adjunct members:

  🇮🇹 Ilaria Gandin (University of Trieste).

Updates: co-chairs

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

From the beginning
Lisa McShane

From the beginning
Jörg Rahnenführer

From this year
Riccardo De Bin

NEW!

Updates: conferences

We presented our work at workshops/conferences. In 2021:

- Mini-symposium of the STRATOS initiative at ISCB 42 (Lyon, July 22^{nd}, 2021);
- 13^{th} Virtual Conference of the Italian Region of the IBS (Milan, September 20^{th}, 2021);

Current work: Overview manuscript

Currently working on an overview manuscript:

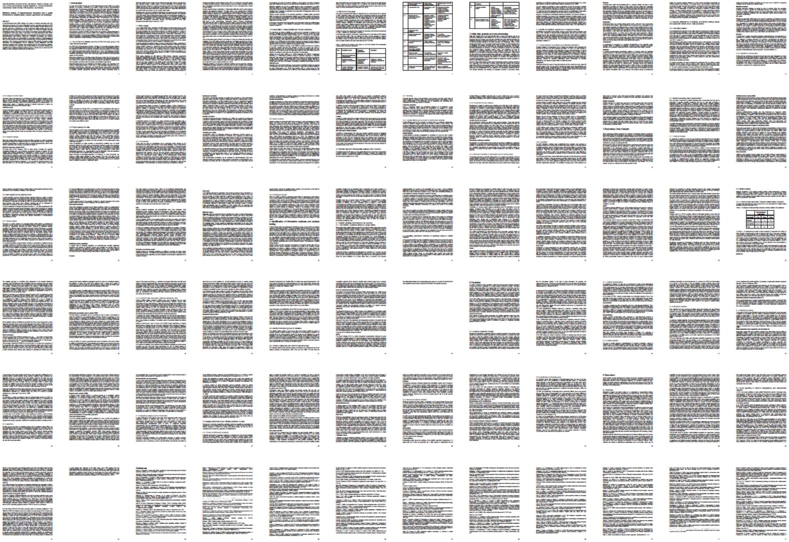
- Title: *Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges*;
- Authors: basically all TG9 members;
- discuss in particular where methods developed for low-dimensional data are inadequate in high-dimensional data (hereafter,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

- Long term project, almost finished:

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

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

Current work: Overview manuscript



Current work: Overview manuscript

Table of contents:

1. Introduction
2. Initial data analysis and preprocessing
3. Exploratory data analysis
4. Identification of informative variables and multiple testing
5. Prediction
6. Discussion

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.

Current work: Overview manuscript

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

Current work: Overview manuscript

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

Current work: Overview manuscript

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

Current work: Overview manuscript

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

Current work: Overview manuscript

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

Plans for the future: simulations

While finishing our overview paper, we have a few projects *in fieri*:

- simulations of high-dimensional data:
  - 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);
  - moreover, how to simulate in the context of correlated mixed data types?
  - can copulas help here? What about more machine-learning-ish techniques (e.g., GAN)?

Plans for the future: other topics

- influence and choice of the tuning parameters:
  - the role and the importance of the tuning parameters for statistical learning techniques used in HDD is often not clear;
  - guidance on how to choose them.
- non-linearities when "modelling" HDD:
  - should be considered at all?
  - if not, what are the arguments against?
  - if yes, which kind of approaches are feasible in HDD?
- influential points in HDD:
  - how do current approaches work?
  - can available knowledge from LDD analysis be transferred into HDD contexts?

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