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

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
  - Members
  - Overview paper
- Current work
  - Sample size calculation in HDD
  - Influence and choice of the tuning parameters
  - Use of plasmode data for simulations in HDD

Updates: members

TG9 current members:

Federico Ambrogi (Italy)
Axel Benner (Germany)
Harald Binder (Germany)
A.-L. Boulesteix (Germany)
Riccardo De Bin (Norway)
Kevin Dobbin (USA)
Ilaria Gardin (Italy)
Roman Hornung (Germany)
Lara Lusa (Slovenia)
Lisa McShane (USA)
Stefan Michiels (France)
E. Migliavacca (Switzerland)
Jörg Rahnenführer (Germany)
Willi Sauerbrei (Germany)

- 14 members;
- 3 new members (in bold);
- 3 co-chairs (in italics).

Updates: Overview paper

In the last meeting, we used Molière's words

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

to describe the state of our 11-author, ∼40,000-word, ∼60-page manuscript . . .

The "fruit" is finally ripe and the paper has been conditionally accepted in BMC Medicine.

- Title: *Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges*;
- this review provides a solid statistical foundation for researchers, including statisticians and non-statisticians, who are new to research with HDD or simply want to better evaluate and understand the results of HDD analyses.

Updates: Overview paper

Table of contents:

Each section contains analytical goals, common approaches, and examples, related to the specific stage of the HDD analysis.

Updates: Overview paper

Sec.	Analytical goals	Common approaches	Examples
2 Initial data analysis and preprocessing:			
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
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
...			
...	...	...	...
5 Prediction:			
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

Current work: introduction

We now decided to work in parallel. Currently on three projects,

- sample size calculation in HDD:
  - Co-chairs: Federico Ambrogi, Lisa McShane;
  - Participants: Harald Binder, Kevin Dobbin, Stefan Michiels, Eugenia Migliavacca, Willi Sauerbrei, Martin Treppner;
- influence and choice of the tuning parameters:
  - Co-chairs: Riccardo De Bin, Lara Lusa, Stefan Michiels;
  - Participants: Roman Hornung, TBA
- use of plasmode data for simulations in HDD:
  - Co-chairs: Axel Benner, Jörg Rahnenführer;
  - Participants: TBA

Current work: sample size calculation in HDD

- A research protocol should specify the study design, including planned sample size
  - depends on the primary endpoint, analysis goal, ...
- In HDD settings, traditional sample size calculations break down due to:
  - the large number of hypotheses tested;
  - complex modeling or analysis strategies typically employed;
- Several approaches for sample size calculation tailored to certain HDD settings have been proposed in the literature:
  - utility and uptake has not been systematically evaluated;
  - it is unclear what kind of sample size justification is used;
  - if any justification is used at all.

Current work: sample size calculation in HDD

Currently screening the literature. Two reviews underway:

- methodological review,
  - identify statistical methods papers describing HDD samples size methods;
  - starting from specific "key words" for "study design criterion" and "Study goal or method" / "Data type";
  - open to augment the list during the applied literature search;
  - record number of literature citations for each identified method.
- applied review.

Current work: sample size calculation in HDD

- applied review,
  - identify which methods (if any) are actually being used in applied/biomedical papers.
  - papers need to deal with HDD (data of specific types);
  - focus on the 15 journals with the highest impact factor in selected fields (i.e. oncology) in a defined time interval;
  - extended to top 5 (impact factor) for general medicine journals, and possibly biology/biomedical ones . . .
  - compute the percentage of studies (within HDD), where a sample size calculation was performed;
  - record the method and/or the justification used;
    - it involves searching for sample size justification in the text;
  - scope might be further restricted if the number of publications identified will result too large.

Current work: sample size calculation in HDD

Goals:

- describe the methodologies most used in applied research, distinguishing between:
  - Class discovery;
  - Class prediction;
  - Class comparison;
- provide examples when software is available;
- provide recommendations for applied researchers;
- have a starting point for future methodologic work.

Current work: influence and choice of the tuning parameters

- Many approaches, especially in HDD, strongly rely on tuning parameters,
  - choosing the best value of the tuning parameter is often more important than choosing the method;
  - often obtained data-driven.
- Unfortunately, in many cases:
  - there is no understanding on the role of the tuning parameter;
  - default values do not work in broad generality;
  - especially when derived in low dimensional contexts.
- Other issues:
  - when data-driven, tuning parameters must be computed on a dedicated subset of the data (validation set $\neq$ test set);
  - complex methods with many tuning parameters are very hard to handle.

Current work: influence and choice of the tuning parameters

Goals:

- describe the role and the importance of the tuning parameters;
- describe typical procedures used to find them;
- provide examples, using real high-dimensional data
  - chronic obstructive pulmonary disease dataset;
- discuss common issues related to the tuning parameters;
- provide recommendations for their choice in practice.

Current work: use of plasmode data for simulations in HDD

Plasmode data:

- A plasmode dataset is based on a real dataset, but some aspect of the data-generating process is known
  - often resampling from real datasets and using parametric model to generate outcome;

Goal:

- Evaluate potential of using plasmode datasets in high dimensional simulation studies.

Current work: use of plasmode data for simulations in HDD

Ongonig research projects (preparations for STRATOS project):

- comparison of plasmode approaches,
  - A. Benner, N. Schreck (DFKZ, Heidelberg), A. Slynko (University of Waterloo), M. Saadati (Statistical Consulting);
- comparison of methods for quantifying dataset similarity.
  - J. Rahnenführer, M. Stolte, A. Bommert (TU, Dortmund);
- comparison of parametric and plasmode approaches for simulation studies in LDD,
  - all researchers mentioned above.

STRATOS TG9 project (starting in summer 2023)

- comparison of parametric and plasmode approaches for simulation studies in HDD,
- how far must the parametric assumptions deviate from the true model for the plasmode approach to be superior?

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