STRATOS INITIATIVE

technische universität dortmund

Statistical analysis of high-dimensional biomedical data: Analytical goals, common approaches and challenges

Jörg Rahnenführer

TU Dortmund University, Department of Statistics
Email: rahnenfuehrer@statistik.tu-dortmund.de

STRATOS
Mini-Symposium,
July 22, 2021

Topic group 9: High-dimensional data

Currently 11 members from 7 countries

Chairs together with JR

- Federico Ambrogi (University of Milan, Italy)
- Axel Benner (DKFZ Heidelberg, Germany)
- Harald Binder (Freiburg University, Germany)
- Anne-Laure Boulesteix (LMU Munich, Germany)
- Riccardo De Bin (University Oslo, Norway)
- Lara Lusa (University of Ljubljana, Slovenia)
- Lisa McShane (NCI, USA)
- Stefan Michiels (University Paris-Sud, France)
- Eugenia Migliavacca (Nestle Institute Lausanne, Switzerland)
- Jörg Rahnenführer (TU Dortmund, Germany)
- Willi Sauerbrei (Freiburg University, Germany)

Riccardo De Bin

Lisa McShane

Goals

- Quiz: When is a data set high-dimensional?
- Multiple choices allowed

- A: More than 10 variables
- B: More than 1.000 variables
- C: Data set too big to fit in memory
- D: More than 10.000 samples

Goals

- Quiz: When is a data set high-dimensional?
- Multiple choices allowed

- A: More than 10 variables: Interpretation?
- B: More than 1.000 variables: Omics data!
- C: Data set too big to fit in memory: Big data?
- D: More than 10.000 samples: Not big data, EH data

Prediction with high-dimensional data

- Main situation: Many more variables than samples: p >> n
- Prediction models (regression, classification, survival): Inherent model selection problem

Bias/Variance – „Model fit“ vs. „Model complexity“

1 gene 50.000 genes

- Solutions for high-throughput data with variable selection
  - Filtering: Select “best” variables before modelling
  - Wrapping: Select variables “within” modelling algorithm
- Problem: Curse of dimensionality – Overfitting

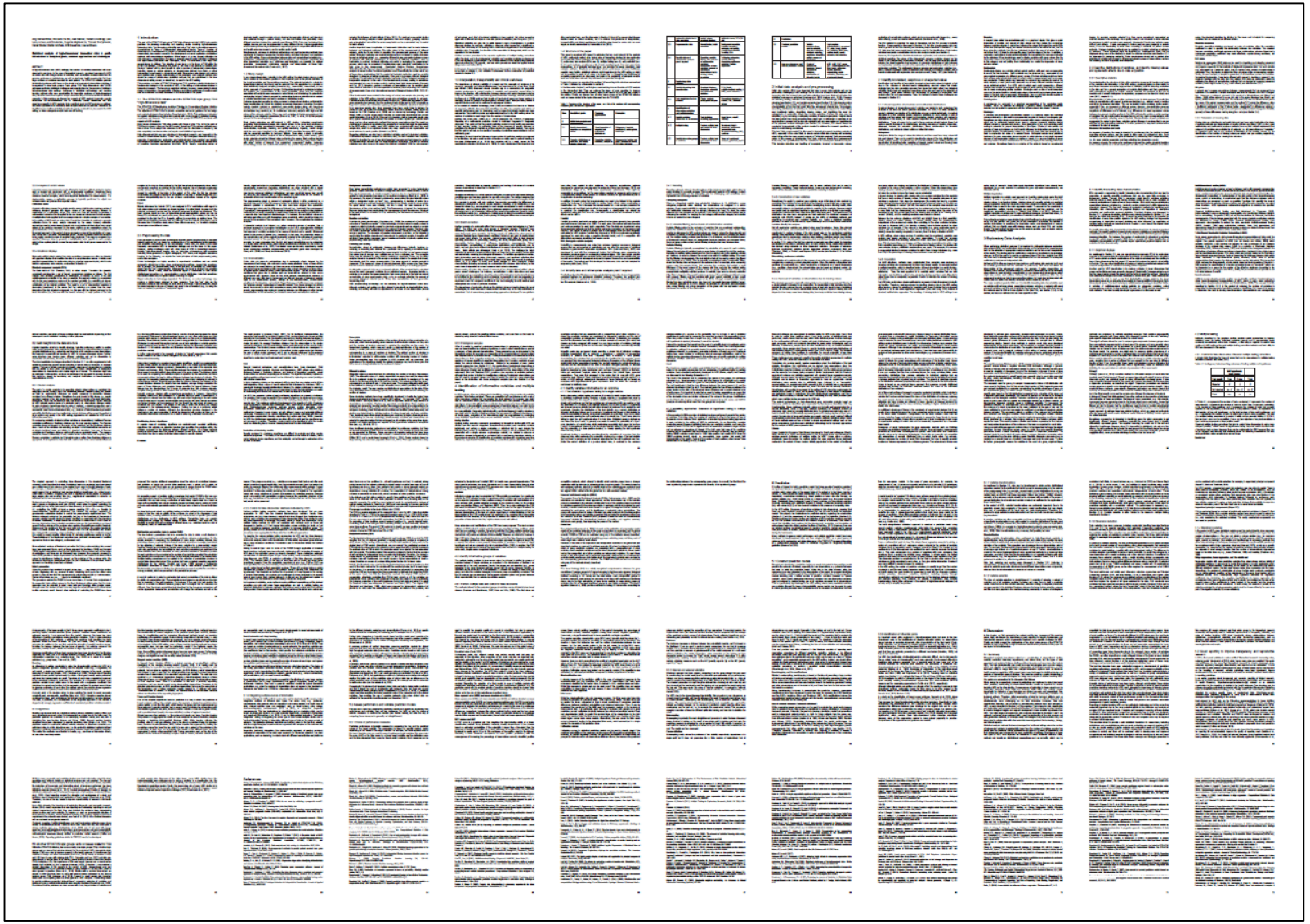
Manuscript

- 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 LDD (low-dimensional data) are inadequate for HDD (high-dimensional data) settings

- Long term project, almost finished:

  Molière (French actor and poet, 17th century):
  “Trees that are slow to grow bear the best fruit.”

Manuscript



Manuscript

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

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	Histograms, boxplots, scatterplots, correlograms, heatmaps
2.2	Describe distributions of variables, and identify missing values and systematic effects due to data acquisition	Descriptive statistics, tabulation, analysis of control values, graphical displays	Measures for location and scale, bivariate measures, RLE plots, MA plots, calibration curve, PCA, Bi-plot

2. Initial data analysis and preprocessing

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

2. Initial data analysis and preprocessing

- IDA important first step in (every) data analysis and can be particularly challenging in HDD settings
- "Data preprocessing" is a term used in biomedical HDD settings, especially in the omics field, and includes IDA and screening steps
- For HDD, it is rarely feasible to conduct a detailed examination of the distribution of every variable individually – instead calculate scores for each variable and select interesting cases
- HDD typically contains uninformative variables that do not vary much across subjects, with variability reflecting only noise – standardization can exaggerate the noise
- Regarding missing values, in HDD settings, complete case analysis may require exclusion of too many observations, and multiple imputation performing regression in HDD is typically not feasible

3. Exploratory data analysis

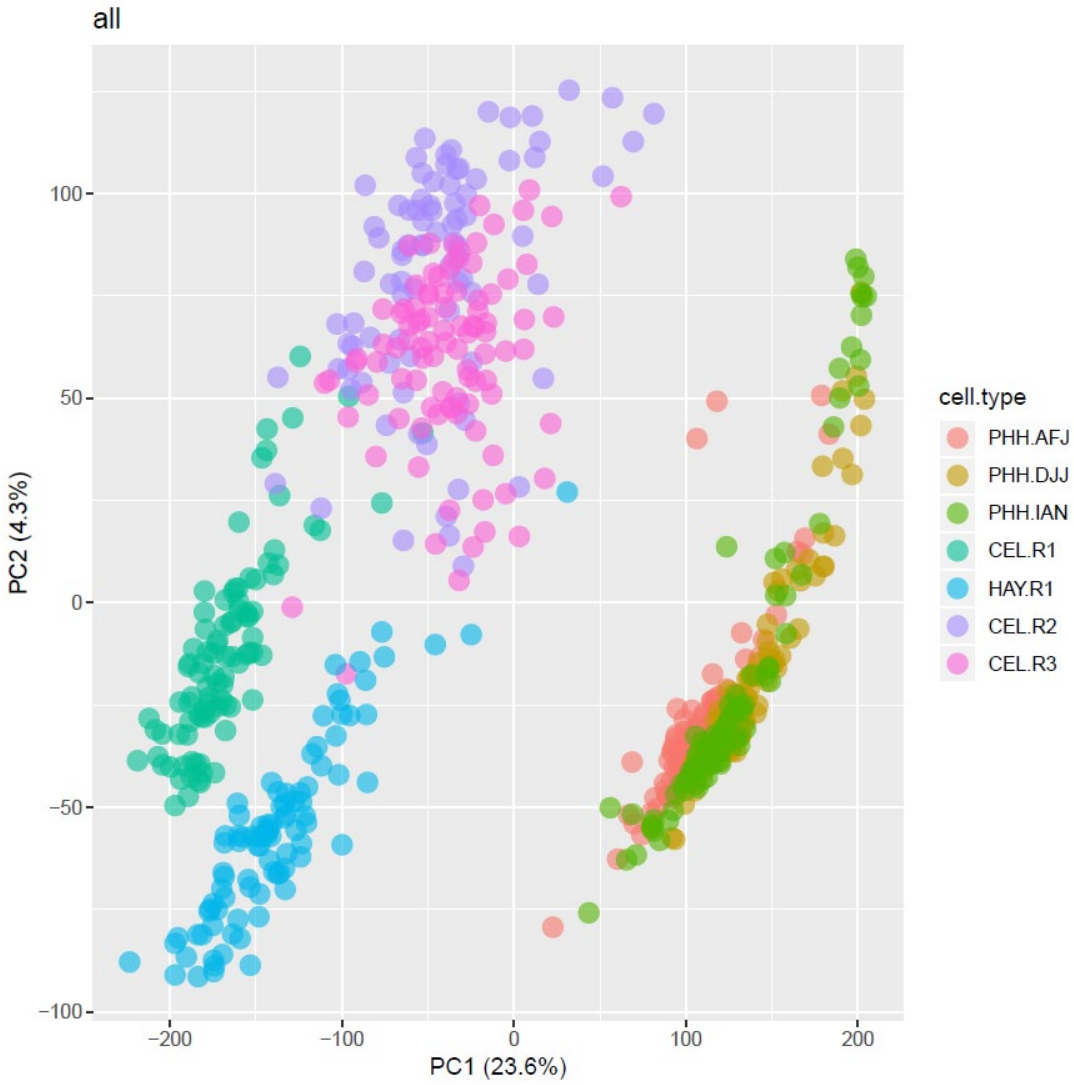
Sec.	Analytical goals	Common approaches	Examples
3.1	Identify interesting data characteristics	Graphical displays, descriptive univariate and multivariate statistics	PCA, Bi-plot, multidimensional scaling, t-SNE, neural networks
3.2	Gain insight into the data structure	Cluster analysis, prototypical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

3. Exploratory data analysis

- Visual identification of interesting characteristics of HDD typically require specialized graphical displays and/or data reduction.
- In cluster analysis, mixtures of low-dimensional parametric probability distributions such as multivariate normal mixtures, cannot be applied at all or perform very poorly in the HDD setting.
- For HDD, the computer runtime of partitioning algorithms can present a challenge. Distributional assumptions are often difficult to verify and algorithms may not converge to a solution.
- Noise accumulating over the variables in HDD may lead to scree plots that fail to reveal a strong indication for the number of clusters.

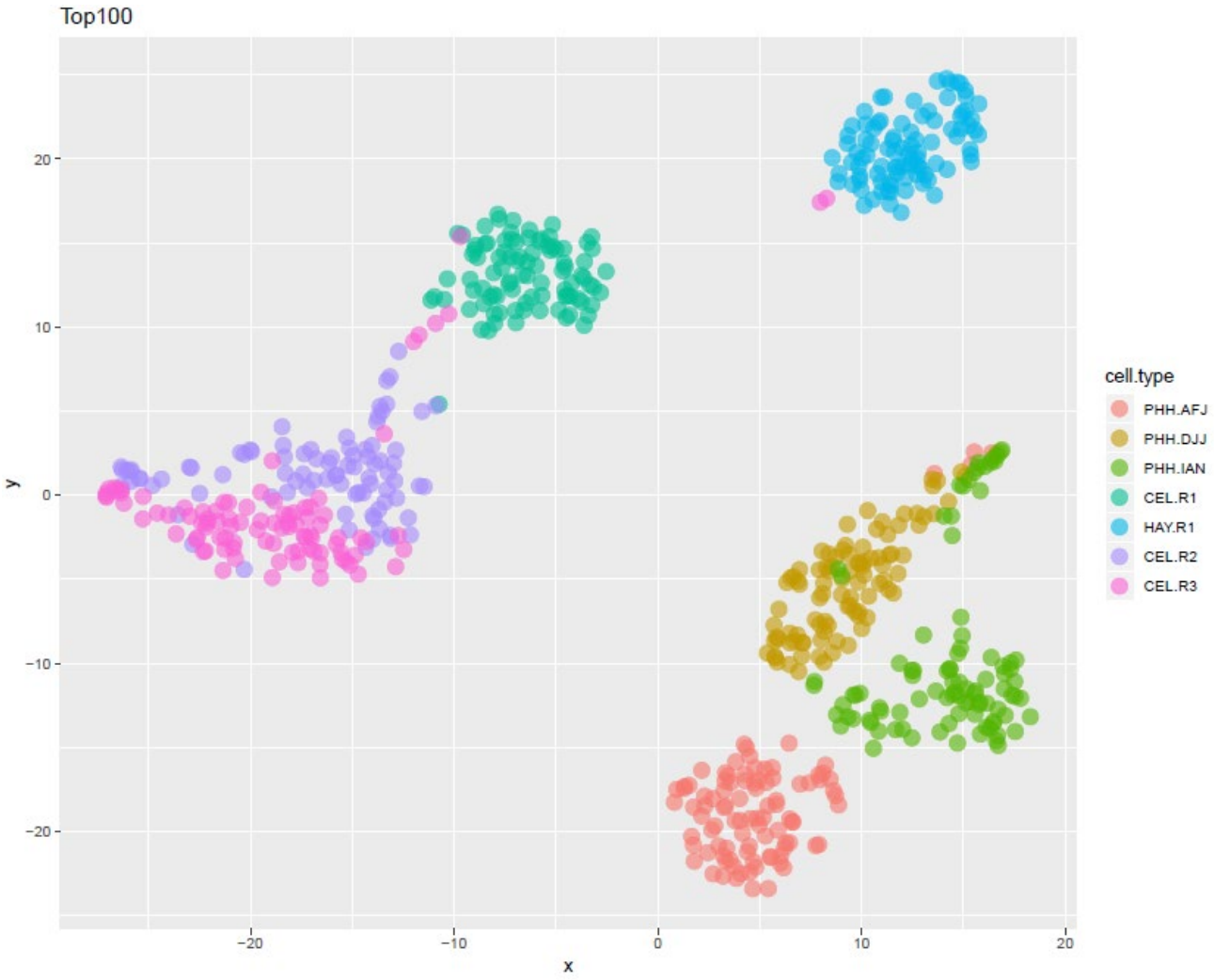
3. Exploratory data analysis

- PCA plot for single-cell transcriptomics
- Color represents experiment
  - 96 values per experiment
- Left
  - 4 experiments with hepatocyte-like cells differentiated from human pluripotent stem cells
- Right
  - 3 experiments with cells from primary human hepatocytes



3. Exploratory data analysis

- t-SNE plot for single-cell experiments
- Color represents experiment
  - 96 values per experiment
- Left
  - 2 replicates separated
- Right
  - 3 replicates clearly separated



3. Exploratory data analysis

- DBSCAN
  - finds clusters of arbitrary shape, is robust to noise, and scales well to large databases (Ester, Kriegel, Sander, Xu, KDD 1996: A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise)
- 2014 SIGKDD Test of Time Award
  - recognizes outstanding papers from past KDD Conferences beyond last decade with important impact on the data mining research community

http://www.kdd.org/News/view/2014-sigkdd-test-of-time-award

- Popular algorithm in computer science and data mining
  - but not much applied in statistics community, although successful/competitive in many applications
  - for example applied to clustering mass spectra (own research)

4. Identification of informative variables and multiple testing

Sec.	Analytical goals	Common approaches	Examples
4.1	Identify variables informative for an outcome	Test statistics, modelling approaches	t-test, limma, edgeR, DESeq2
4.2	Multiple testing	Perform multiple tests and control for false discoveries	Bonferroni, Holm's procedure, multivariate permutation tests, Benjamini-Hochberg (BH), q-values
4.3	Identify informative groups of variables	Perform multiple tests and control for false discoveries	Gene set enrichment analysis, global test, topGO

4. Identification of informative variables and multiple testing

- Frequent goals in HDD settings
  1) Identification of variables that are associated with a single outcome or phenotype variable
  2) Identification of variables with a trajectory over time affected by experimental factors or exhibiting a prescribed pattern
  3) Identification of candidate variables that are associated with a prespecified set of other variables

- Specific methods also developed for multiple testing with HDD
  - For hypothesis testing for single variables (e.g. limma, edgeR, DESeq2): Take advantage of HDD: Information sharing between variables!
  - For multiple testing correction: control of false discovery rate (FDR)

5. Prediction

Sec.	Analytical goals	Common approaches	Examples
5.1	Construct prediction models	Variable transformations, variable selection, dimension reduction, statistical modelling, algorithms, integrating multiple sources of information	Log-transform, standardization, 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, identification of influential points	MSE, MAE, ROC curves, AUC, misclassification rate, Brier score, calibration curves, deviance, subsampling, cross-validation, Bootstrap, use of external datasets

5. Prediction

- Numerous dramatic claims of performance of prediction models have been made using HDD, although unfortunately, these claims do not always withstand careful validation!
- Identify optimal level of model complexity that will yield interpretable models with good prediction performance on independent data
  - If the number of variables *p* is larger than the number of objects *n*, then basic regression models cannot be fitted
  - In HDD settings, filter methods that preselect variables are very popular
  - Regularization approaches (Lasso, ridge regression) to overcome dimensionality challenge
  - Knowledge-based data enrichment, typically based on external knowledge, e.g. by integrating data on biological pathways and networks
- Machine learning methods have been successfully used, but many (in particular deep learning) require estimation of a large number of parameters and thus need a large number of objects as input!

Thank you

- Thank you very much for your attention !

Biology

Computer Science

Statistics