Statistical Analysis of High-Dimensional Biomedical Data

Analytical Goals, Common Approaches and Challenges

Axel Benner

German Cancer Research Center (DKFZ), Heidelberg, Germany

July 18, 2019

STRATOS INITIATIVE

Motivation & Relevance

- **Increasing use and availability of health-related metrics**
  - Omics data (e.g., genomics, transcriptomics, proteomics)
  - Electronic health records
- **Big data / high dimensionality**
  - **Big data**
    typically characterized by very large sample size n
  - **High dimensionality**
    number of unknown parameters p is of much larger order than sample size n ($p >> n$)

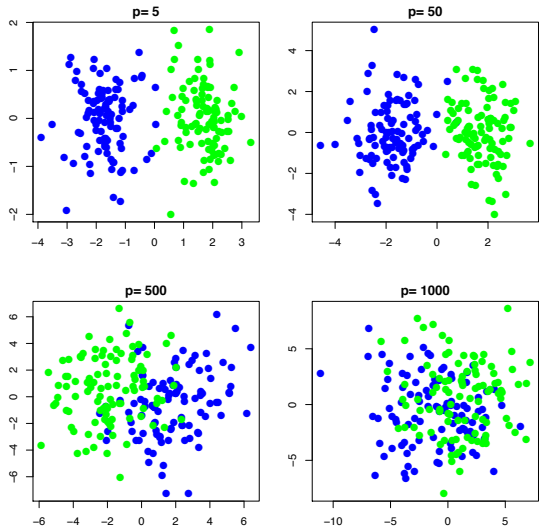
Motivation & Relevance

High dimensionality introduce computational and statistical challenges

- Heterogeneity (e.g., different sources, technologies)
- Noise accumulation (accumulation of estimation errors)

Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014

Example: Noise accumulation



Projections of the observed data (n=100, each class) onto the first two principal components of the p-dimensional feature space.

Motivation & Relevance

High dimensionality introduce computational and statistical challenges

- Heterogeneity (e.g., different sources, technologies)
- Noise accumulation (accumulation of estimation errors)
- Spurious correlation (uncorrelated variables may have high sample correlations in high dimensions)
- Incidental endogeneity (correlations between predictors and residual noise)

These features of high-dimensional data often make traditional statistical methods invalid!

Guidance required on how to deal with these challenges.

Fan J, Han F, Liu H. Challenges of big data analysis. Natl Sci Rev. 2014

14 Members (July 2019)

- Federico Ambrogi (University of Milan, Italy)
- Axel Benner (DKFZ Heidelberg, Germany)
- Harald Binder (Freiburg University, Germany)
- Anne-Laure Boulesteix (LMU Munich, Germany)
- Tomasz Burzykowski (Hasselt University, Belgium)
- Riccardo De Bin (University Oslo, Norway)
- W. Evan Johnson (Boston University, USA)
- Lara Lusa (University of Ljubljana, Slovenia)
- **Lisa McShane (NCI, USA)**
- Stefan Michiels (University Paris-Sud, France)
- Eugenia Migliavacca (Nestle Institute of Health Sciences Lausanne, Switzerland)
- **Jörg Rahnenführer (TU Dortmund, Germany)**
- Sherri Rose (Harvard Medical School, USA)
- Willi Sauerbrei (Freiburg University, Germany)

Subtopics

(All in context of very large number of predictor, descriptor, or outcome variables)

1. Data pre-processing
2. Exploratory data analysis
3. Data reduction
4. Multiple testing
5. Prediction modeling/algorithms
6. Comparative effectiveness and causal inference
7. Design considerations
8. Data simulation methods
9. Resources for publicly available high-dimensional data sets

Current Goals

- Overview paper
  - Statistical analysis of high-dimensional biomedical data: A gentle introduction to analytical goals, common approaches and challenges
- Simulation paper
  - Guidance for planning, conducting and reporting simulation studies for comparing analytic approaches for biomedical data: General concepts with additional considerations for high-dimensional data
- Guidance for analysis processes
  - Examples for data analysis processes for specific types of HDD
  - Recommendations for best practices
  - R-Code with interpretations

Overview paper

Topics

- Initial data analysis
- Exploratory data analysis
- Multiple testing
- Prediction

Initial Data Analysis

Analytical goals

- Describe distributions of variables and identify inconsistent, suspicious or unexpected values
- Identify missing values and consider strategies to address
- Identify systematic effects due to data collection and adjust if required
- Simplify data and refine/update analysis plan if required

Common approaches

- Graphical displays: Scatterplots, Histograms, Heatmaps, ...
- Descriptive statistics
- Projections: Principal component analysis (PCA)

Exploratory data analysis

Analytical goals

- Identify interesting data characteristics
- Analyze data structure

Common approaches

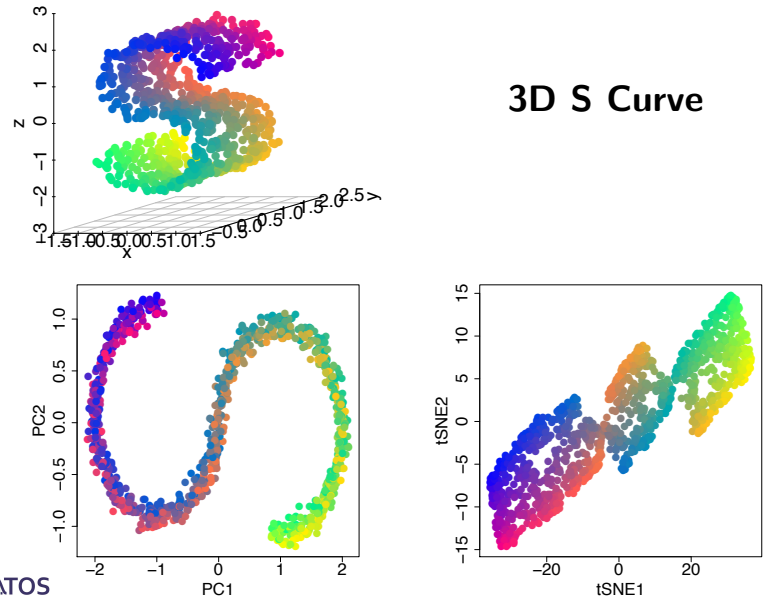
- Projections into fewer dimensions: PCA, t-Distributed Stochastic Neighbor Embedding (t-SNE), ...
- Descriptive statistics (e.g., to identify regions with relatively large data density)
- Cluster analysis
  - Hierarchical clustering, K-means, Partitioning around Medoids (PAM), ...
- Creation of prototypical samples

Exploratory Data Analysis

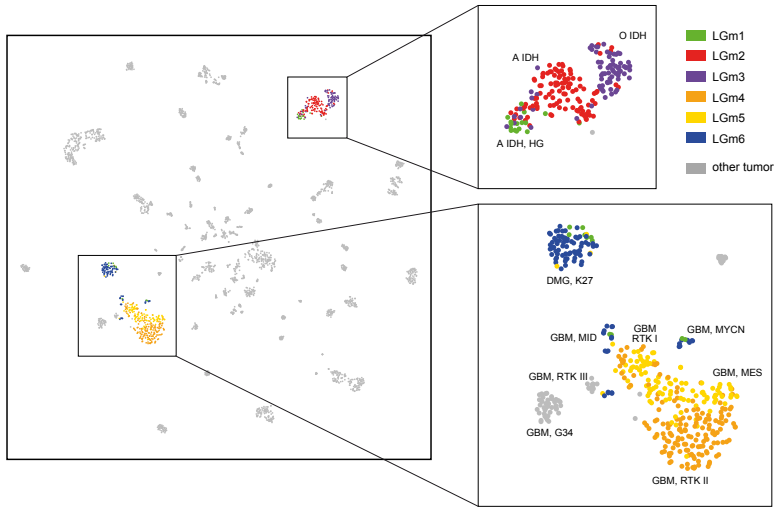
Example: **PCA vs. t-SNE**

- PCA performs a linear mapping of the data to a lower-dimensional space such that the variance of the data is maximized.
- t-SNE (van der Maaten & Hinton, 2008) is a variation of Stochastic Neighbor Embedding (SNE, Hinton & Roweis, 2002) that minimizes the divergence between the distribution of the input data and the distribution in the low-dimensional space.
- The main difference between t-SNE and PCA is that PCA focuses on preserving the distances between widely separated data points whereas t-SNE tries to preserve the distances between nearby high-dimensional data points.
  i.e. t-SNE reduces the dimensionality of data mainly based on local properties of the data

Example: PCA vs. t-SNE



Example: PCA vs. t-SNE



Genome-wide DNA methylation profiles of n = 2,801 brain tumor samples (Capper et al., Nature 2018)

Multiple Testing

Analytical goals

- Identify informative variables for an outcome
- Identify informative groups of variables

Multiple Testing

Statistical testing of thousands of hypotheses

- requires alternative procedures to control the false discovery rates and to improve the power of the tests.

Many different scenarios

- Find variables with different distributions between pre-specified classes of subjects or with association with outcome
- Enriched variables classes in a list of selected variables

Common approaches

- Control for false discoveries (e.g. FDR, empirical Bayes)
- Global testing versus one-at-a-time testing
- Enrichment tests (e.g., gene set enrichment analysis)

Prediction

Analytical goals

- Construct prediction models
- Assess performance and validate prediction models

Prediction

Goal: **Construct prediction models**

Common approaches

- Dimension reduction
- Statistical modelling
  - Ridge regression, lasso and their modifications
  - Boosting
  - Support vector machines
  - Trees and Random forests
  - Neural networks and deep learning

Prediction

Problem: **Standard methods break down**

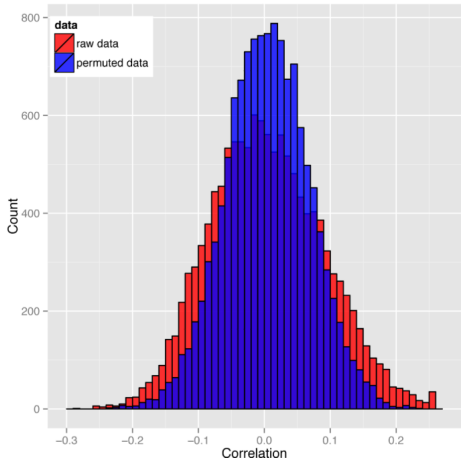
- For $n \ll p$ cannot fit standard regression model
- Redundancy in variables:
  huge correlation as problem for stable variable selection

  Example: **Incidental endogeneity**

  Gene expression example (Fan et al. 2014):

  (Ten-fold cross validated) L1-penalized least squares regression (37 genes are selected) - refit ordinary least squares regression on the selected model to calculate residuals.

Example: Incidental endogeneity



Red: Empirical distribution of the correlations between the predictors and the residuals
Blue: "null distribution" of the spurious correlations by randomly permuting the orders of rows in the design matrix.

Prediction

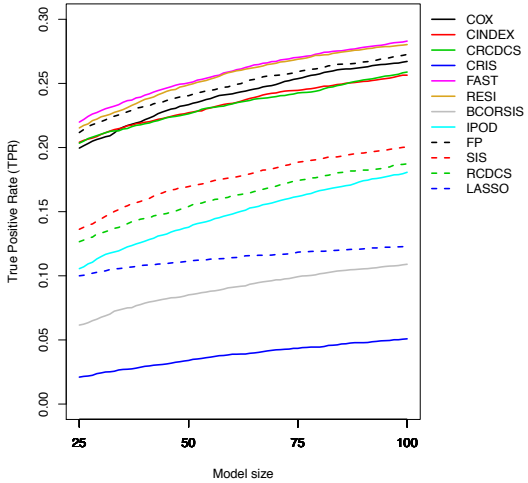
Penalized regression is inefficient when the dimensionality **p** of the covariates is ultrahigh (Fan & Lv, JRSS-B 2008)

Dimension reduction

- Apply screening methods that are based on correlation learning to reduce dimensionality from ultrahigh to a moderate scale
- This enhances finite sample performance of subsequent regression methods (Fan & Lv, JRSS-B 2008)

Speed is essential: Fast distance correlation screening using martingale residuals, which is computationally efficient and easy to implement (Edelmann et al. BiomJ, 2019)

Example: Incidental endogeneity



True positive rate vs. model size considering linear and nonlinear effects. Plasmode simulation of 100,000 CpG sites, sample size n=300 and 50% censoring (500 data sets).

Prediction

Goal: **Assess performance and validate prediction models**

Problem: Improper evaluation (e.g., resubstitution) drastically overestimates model performance
(and is still extremely common)

Common approaches

- Calibration and prediction accuracy
- Choice of performance measures (e.g., MSE, AUC, Brier score)

Risk of overfitting
⇒ Stability of model selection

Simulation paper

Goal: **Guidance for planning, conducting and reporting simulation studies for comparing analytic approaches for biomedical data**

For high-dimensional data heavier reliance on simulated data necessary

- Data often generated to address complex research questions, and analytical methods may be correspondingly tailored
- Wide range of specialized data and analysis approaches, thus often not sufficient number of data sets available

TG9 cooperation with Simulation Panel obvious!

Simulation paper

Issues specific to high-dimensional data (HDD)

- Underlying (biological) mechanism not well understood
- Difficult to simulate realistic correlation structure and suitable multivariate distributions

Common Approaches

- Simulations based on assumed distributions (e.g. normal, Poisson, negative binomial)
- Simulation using extracted parameters from pilot data
- Simulation using real data (e.g., plasmode data)

More about this:
Victor Kipnis: Issues in modern biomedical simulation studies

Example: "Real data" simulation of HDD

Useful approach for realistic high-dimensional data generation

- Plasmode data:

  Real data (e.g., omics data from actual biological specimens) which are manipulated such that the parameters of interest are known with certainty.

  - Name from plasm=form, and mode=measure
  - References:
    Cattell, R. B. (1966). Handbook of Multivariate Experimental Psychology. Rand McNally, Chicago.
    Mehta et al., Physiological Genomics 2006;28(1):24-32

Example: "Real data" simulation of HDD

Advantages of plasmode data

- Distributions/correlations are taken directly from real data
- Appropriate permutation, resampling, or modification of real data offers flexibility to generate data with desired features
- Can combine with outcome models to generate dependent variables associated with realistic HDD as independent variables

"Real data" simulation of HDD

Example: Generate data for evaluation of multiple testing methods

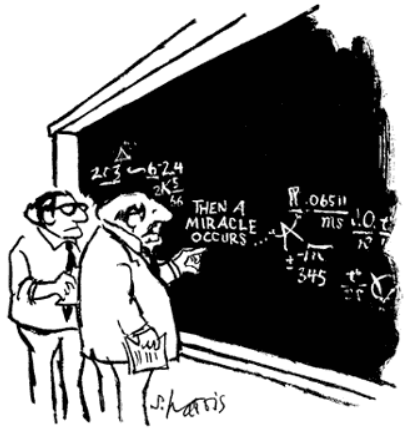
- Permute subject/specimen IDs to generate a null distribution
  - Global null allows assessment of "weak control" of false positives for a multiple testing procedure
- Add back defined effects on specific individual variables
  - Allows assessment of both "power" for true positives and "strong control" of false positives for a multiple testing procedure

Outlook

- TG 9: Large topic with links to Bioinformatics and Molecular Epidemiology
- But: high-dimensional challenges also in non-omics settings
- Overlap with many other topic groups, but always with high-dimensional flavor
- Cooperation with other Topic Groups is essential

Outlook

Guidance is essential!



"I think you should be more explicit here in step two."

(Source: Sidney Harris)

Thank You

Federico Ambrogi
Axel Benner
Harald Binder
Anne-Laure Boulesteix
Tomasz Burzykowski
Riccardo De Bin
Evan Johnson
Lara Lusa
Lisa McShane
Stefan Michiels
Eugenia Migliavacca
Jörg Rahnenführer
Sherri Rose
Willi Sauerbrei