

Analysis of High-Dimensional Data: Guidance or (Best) Practice?

Jörg Rahnenführer

Technische Universität Dortmund, Fakultät Statistik

Email: rahnenfuehrer@statistik.tu-dortmund.de



62. Jahrestagung der GMDS
Oldenburg, 18.09.2017

STRATOS



- An efficient way to help researchers *to keep up with recent methodological developments* is to develop guidance documents that are spread to the research community at large.
- The objective of STRATOS is to *provide accessible and accurate guidance* in the design and analysis of observational studies.

Outline

- Analysis of High-Dimensional Data: Guidance or (Best) Practice?
- High-dimensional data: Joy and frustration
- STRATOS TG 9: Goals and structure
- Also statisticians should consider ...
 - ... automated pipelines
 - ... machine learning
 - ... guidance

Analysis of high-dimensional data

- **Situation: Many more variables than samples: $p \gg n$**
- **Prediction models** (regression, classification, survival):
Inherent **model selection** problem

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



1 gene

50.000 genes

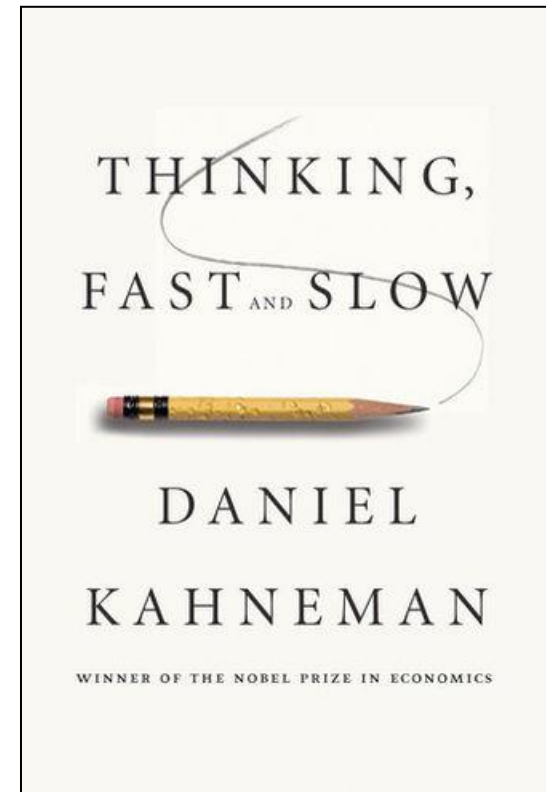
- **Solutions** for high-throughput data with variable selection
 - **Filtering**: Select “best” variables before modelling
 - **Wrapping**: Select variables “within” modelling algorithm
(AIC, BIC, penalized regression/classification, cross-validation)

Analysis of high-dimensional data

- **Joy** of the analysis of high-dimensional data
 - Having so much fun with data
 - Great interdisciplinary research opportunities
- **Frustration** after preparing such a talk
 - Reality check – in practice hard to always do what should be done
- **Pressure to publish – publication bias**
 - Ioannidis, John P. A. (August 1, 2005). "[Why Most Published Research Findings Are False](#)". PLoS Medicine. **2** (8): e124
 - “Proteus phenomenon”: Occurrence of extreme contradictory results in the early studies performed on the same research question

Analysis of high-dimensional data

- **Humans** are definitely not good in avoiding **pitfalls**:
Thinking – fast and slow
(Daniel Kahnemann)
 - Confirmation bias: The tendency to search for, interpret, favor, and recall information in a way that confirms preexisting beliefs or hypotheses
 - Overfitting of numbers and patterns...
 - 18: “62 GMDS Jahrestagung”
 - 09: “Oldenburg”



Stratos: Topic Groups

TG 1: Missing data

TG 2: Selection of variables and functional forms in multivariable analysis



TG 3: Initial data analysis

TG 4: Measurement error and misclassification

TG 5: Study design

TG 6: Evaluating diagnostic tests and prediction models

TG 7: Causal inference

TG 8: Survival analysis

TG 9: **High-dimensional data**

Stratos: Topic Groups

TG 1: Missing data

TG 2: Selection of variables and functional forms in multivariable analysis



TG 3: Initial data analysis

TG 4: Measurement error and misclassification

TG 5: Study design

TG 6: Evaluating diagnostic tests and prediction models

TG 7: Causal inference

TG 8: Survival analysis

TG 9: High-dimensional data

Motivation for TG 9

- **Increasing use and availability of health-related metrics**
 - Omics data (genomics, transcriptomics, proteomics, ...)
 - Electronic health records
- **Big data / high dimensionality**
 - **Big data** typically refers to very large sample size n
 - **High-dim**: number of unknown parameters p is of much larger order than sample size n ($p \gg n$)
- **Unique 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

TG 9

- Started in 2016
- Co-Chairs
 - Lisa McShane NCI, USA
 - Jörg Rahnenführer TU Dortmund, Germany
- Talks today at gmDs
 - Jörg Rahnenführer: Analysis of High-Dimensional Data: Guidance or (Best) Practice?
 - Harald Binder: Advances in dimension reduction, manifold learning, and generative models
 - Axel Benner: Simulating high-dimensional molecular data

TG 9: Members

- 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)



TG 9: Subtopics

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

Subtopic 1: Data Preprocessing

- **Omics data: Removal of systematic biases**
 - Intensity effect, batch effect, dye effect, block effect, ...
- **Challenges**
 - Keeping up with new technologies that generate new data types
 - Methods specific to technology or data generating mechanism (microarrays, NGS, mass spectrometry)
- **Typical Tasks**
 - Normalization/calibration, identification of outliers/errors
- **Nice guidance example: Spike-in benchmark data**
 - Affycomp: A benchmark for Affymetrix GeneChip expression measures (Irizarry et al., Biostatistics 2003)
 - Truth known, allows to identify statistical features of the data

Subtopic 2: Exploratory analysis

- **Descriptive statistics**
 - Initial data analysis
 - Univariate, bivariate measures
 - Identify regions with relatively large data density
- **Data visualization**
 - Heatmaps, projection methods
- **Clustering approaches**
 - Biclustering
 - Classical (k-means, ...) and high-dim (subspace clus., DBSCAN)
- **Integrative analyses of different data types**
 - e.g., proteomic, transcriptomic, and genomic data measured on the same subject

Subtopic 3: Data reduction

- **Dimension reduction and variable selection**
 - Central role for analyzing high dimensional data, in terms of statistical accuracy
- **Goals**
 - Visualization of samples or variables
 - Building / finding prototypical samples
 - Building new variables, e.g. meta-genes, for use in subsequent statistical modeling or machine learning approaches
 - Variable selection
- **See talk of Harald Binder**

Subtopic 4: Multiple testing

- **Statistical testing of thousands of hypotheses**
 - Requires alternative procedures to control false discovery rates and to improve 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
- **Statistical approaches**
 - Control of false positives (e.g., FDR, empirical Bayes)
 - Global testing versus one-at-a-time testing
 - Enrichment tests (e.g., gene set enrichment analysis)
 - Variable selection

Subtopic 5: Prediction modelling/algorithms

- **Model choice: parametric versus non-/semi-parametric**
- **Model building**
 - Penalized regression (ridge, lasso, elastic net, SCAD, MCP)
- **Machine learning methods**
 - Trees, support vector machines, multilayer neural networks
 - Random forests, super learners (bagging, bundling), boosting
- **Evaluation of prediction models**
 - Calibration versus prediction accuracy
 - Performance metrics (e.g., MSE, AUC, Brier score)
 - Risk of overfitting (consider stability, validation)
 - **Improper evaluation (e.g., resubstitution) drastically overestimates model performance (and is still extremely common)**

Subtopic 6: Comparative Effectiveness and Causal Inference

- **High-dimensional data driven challenges**
 - Non-randomized observational data
 - Missingness
 - Sparsity
 - Unmeasured confounding
 - Positivity violations
 - Distributed networks
 - Multiple treatments
- **Estimation**
 - Use of machine learning
 - Propensity score methods

Subtopic 7: Design Considerations

- **Sampling observational units**
 - Random vs. outcome-dependent (e.g., case-control, case-cohort)
 - Many omics studies use case control sampling but ignore it in the analysis
- **Variable ascertainment**
 - Proportion of "complete cases" decreases with increasing number of variables
 - Potential biases
 - Imputation methods
- **Sample size planning**
 - False discovery control under target true positive rate
 - Estimation or testing for model parameters or for performance

Subtopic 8: Data Simulation Methods

- **Issues specific to high-dimensional data**
 - Underlying (biological) mechanism not well understood
 - Difficult to simulate realistic correlation structure and suitable multivariate distributions
- **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)
- **See talk of Axel Benner**
 - Plasmode (from plasm=form, and mode=measure) is a real (i.e., from actual biological specimens) data set for which some aspect of the truth is known (Mehta et al., Physiological Genomics, 2006)

Subtopic 9: Resources for ... Data Sets

- **Subtopic 9: Resources for Publicly Available High-dimensional Data Sets**
- **GEO** (Gene Expression Omnibus)
 - <http://www.ncbi.nlm.nih.gov/geo>
- **GDC** (Genomic Data Commons), including **TCGA** (The Cancer Genome Atlas)
 - <https://gdc.cancer.gov/>
- **Array Express**
 - <https://www.ebi.ac.uk/arrayexpress/>
- Many more ...

Definition of bioinformatics

- **Bioinformatics** is both an umbrella term for the body of **biological studies that use computer programming** as part of their methodology, as well as a reference to specific **analysis "pipelines"** that are repeatedly used, particularly in the field of genomics.

Source: Wikipedia

(Automated) pipelines

- **Pros**
 - **Reproducibility**
 - No intentional/unintentional „overfitting“
 - Standardized procedures enable comparison across different studies
- **Cons**
 - Quality checks not easy to automate – e.g. unexpected batch effects
 - Often specific adjustments to specific data properties (deficits) necessary

Machine learning: DBSCAN

- **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
 - e.g., clustering mass spectra (Schork, master thesis TU Dortmund, 2017, and follow-up research)

Definition of deep learning

- “Deep learning is part of a broader family of machine learning methods based on learning representations of data”.

Source: Wikipedia

- Idea

- “An observation (e.g., an image) can be represented in many ways such as a vector of intensity values per pixel, or in a more abstract way as a set of edges, regions of particular shape, etc.”
- “One of the promises of deep learning is replacing handcrafted features with efficient algorithms for unsupervised or semi-supervised feature learning and hierarchical feature extraction.”

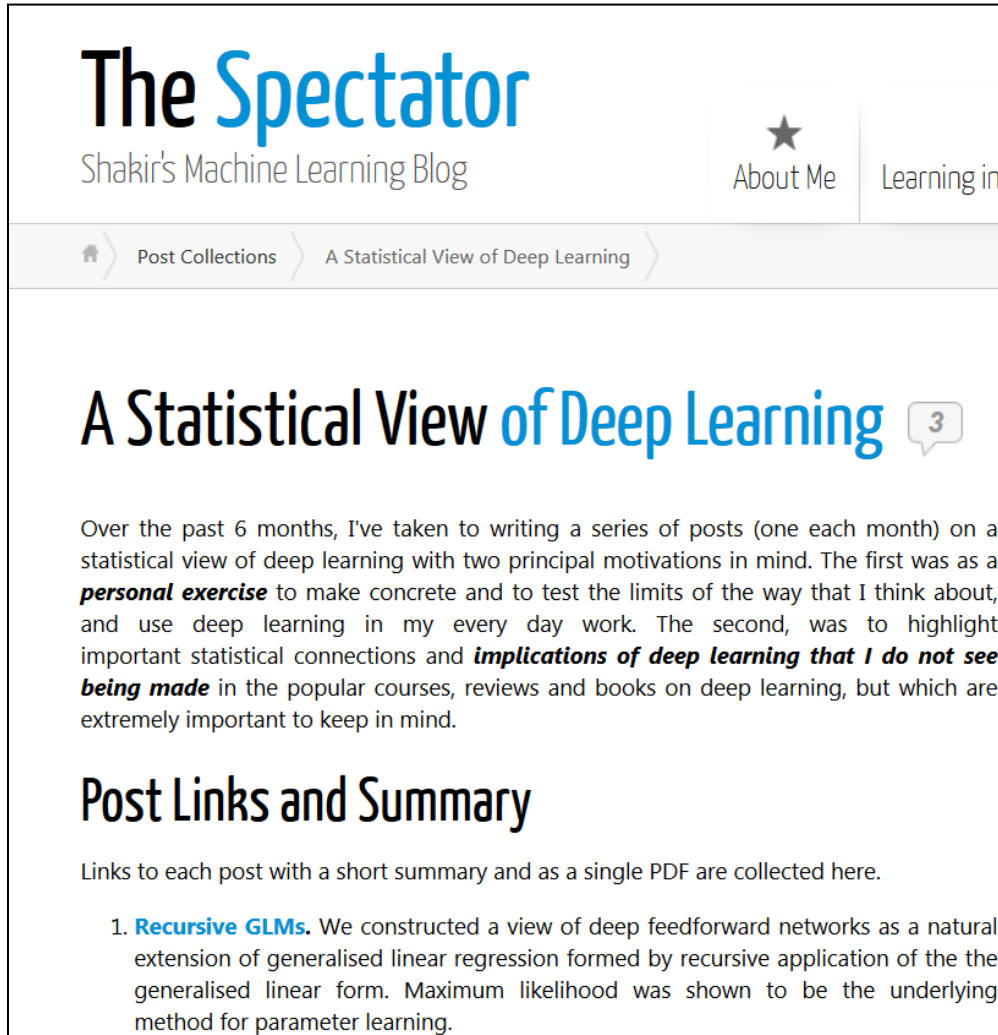
- Types

- “Deep neural networks, convolutional deep neural networks, deep belief networks and recurrent neural networks”

Definition of deep learning

- **Rebranding of neural networks**
 - “Some of the representations are ... loosely based on ... communication patterns in a nervous system, such as neural coding which attempts to define a relationship between various stimuli and associated neuronal responses in the brain.”
- **Competitive results**
 - “in computer vision, automatic speech recognition, natural language processing, audio recognition, and bioinformatics”
- Similar hype than with neural networks in the 90s
- Extremely successful especially in vision with $n \gg p$, but **overfitting for moderate n**

A statistical view of deep learning



The Spectator
Shakir's Machine Learning Blog

About Me Learning in E

Post Collections > A Statistical View of Deep Learning >

A Statistical View of Deep Learning 3

Over the past 6 months, I've taken to writing a series of posts (one each month) on a statistical view of deep learning with two principal motivations in mind. The first was as a **personal exercise** to make concrete and to test the limits of the way that I think about, and use deep learning in my every day work. The second, was to highlight important statistical connections and **implications of deep learning that I do not see being made** in the popular courses, reviews and books on deep learning, but which are extremely important to keep in mind.

Post Links and Summary

Links to each post with a short summary and as a single PDF are collected here.

1. **Recursive GLMs.** We constructed a view of deep feedforward networks as a natural extension of generalised linear regression formed by recursive application of the the generalised linear form. Maximum likelihood was shown to be the underlying method for parameter learning.

<http://blog.shakirm.com/ml-series/a-statistical-view-of-deep-learning/>

A statistical view of deep learning

- **Deep feedforward networks**
 - Natural extension of generalized linear regression
 - Recursive application of the generalized linear form, with maximum-likelihood for parameter learning
- **Recurrent networks**
 - State-space models or dynamical systems
 - Recurrent networks assume that hidden states are deterministic, state-space models have stochastic hidden states
 - Maximum-Likelihood reasoning, innovative new models
- **Various forms of statistical regularization implemented**

Thank you

- Thank you very much for your attention !

Biology



Computer Science



Statistics

