

# Analysis of high-dimensional data: Opportunities, challenges and goals

Riccardo De Bin<sup>1</sup>

Department of Mathematics - University of Oslo

---

<sup>1</sup>on behalf of the TG9 – High-Dimensional Data of the STRATOS initiative



## Outline of the talk

- Introduction
- Main
- Remarks

## Introduction: TG9 – high-dimensional data

### About TG9 – high-dimensional data: **Aim and scope**

From the web page ([www.stratos-initiative.org/group\\_9](http://www.stratos-initiative.org/group_9)):

*[...] The goal of the 'high-dimensional data' topic group of the STRATOS initiative (TG9) is to **provide guidance** amid the jungle of opportunities and pitfalls inherent **in the analysis of high-dimensional biological and medical data**. [...] **in-depth evaluation and discussion** of various **statistical and computational approaches** aim to reinforce concepts and support specific recommendations for best practices. [...]*

## Introduction: TG9: high-dimensional data

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

Currently 11 members from 7 countries:



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);

## Introduction: TG9: high-dimensional data

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

**From the start**

Lisa McShane



**From the start**

Jörg Rahnenführer



**From this year**

Riccardo De Bin



many slides are taken from Jörg's talk at ISCB42 (Lyon, 2021)

## Introduction: Definition

When a dataset is “high-dimensional”?

- More than 10 variables?
  - ▶ interpretation difficulties;
  - ▶ basic methods (e.g.,  $kNN$ ) start to fail;
- More variables ( $p$ ) than observations ( $n$ )?
  - ▶ popular methods (e.g., OLS) fail;
- Large  $n$ ?
  - ▶ computational (e.g., matrix inversion) issues;
- Large  $n$  and large  $p$ ?
  - ▶ computational and memory issues.

## Introduction: Prediction with high-dimensional data

- Main situation: Many **more variables than samples** ( $p \gg n$ ).
- Prediction models (regression, classification, survival):
  - ▶ **Bias-Variance trade-off**;
  - ▶ “Model fit” vs “**Model complexity**”.
- Solutions for high-throughput data with **variable selection**:
  - ▶ **Filtering**: Select “best” variables before modelling;
  - ▶ **Wrapping**: Select variables “within” modelling algorithm.
- Problems:
  - ▶ **Curse of dimensionality**;
  - ▶ **Overfitting**.

## Main: Overview manuscript

Currently working on a 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<sup>th</sup> century):



## Main: Overview manuscript

1 Introduction	2 Overview	3 Data	4 Methods	5 Results	6 Discussion	7 Conclusion	8 Acknowledgements	9 References	10 Appendix
11	12	13	14	15	16	17	18	19	20
21	22	23	24	25	26	27	28	29	30
31	32	33	34	35	36	37	38	39	40
41	42	43	44	45	46	47	48	49	50
51	52	53	54	55	56	57	58	59	60
61	62	63	64	65	66	67	68	69	70
71	72	73	74	75	76	77	78	79	80
81	82	83	84	85	86	87	88	89	90
91	92	93	94	95	96	97	98	99	100
101	102	103	104	105	106	107	108	109	110
111	112	113	114	115	116	117	118	119	120
121	122	123	124	125	126	127	128	129	130
131	132	133	134	135	136	137	138	139	140
141	142	143	144	145	146	147	148	149	150
151	152	153	154	155	156	157	158	159	160
161	162	163	164	165	166	167	168	169	170
171	172	173	174	175	176	177	178	179	180
181	182	183	184	185	186	187	188	189	190
191	192	193	194	195	196	197	198	199	200
201	202	203	204	205	206	207	208	209	210
211	212	213	214	215	216	217	218	219	220
221	222	223	224	225	226	227	228	229	230
231	232	233	234	235	236	237	238	239	240
241	242	243	244	245	246	247	248	249	250
251	252	253	254	255	256	257	258	259	260
261	262	263	264	265	266	267	268	269	270
271	272	273	274	275	276	277	278	279	280
281	282	283	284	285	286	287	288	289	290
291	292	293	294	295	296	297	298	299	300
301	302	303	304	305	306	307	308	309	310
311	312	313	314	315	316	317	318	319	320
321	322	323	324	325	326	327	328	329	330
331	332	333	334	335	336	337	338	339	340
341	342	343	344	345	346	347	348	349	350
351	352	353	354	355	356	357	358	359	360
361	362	363	364	365	366	367	368	369	370
371	372	373	374	375	376	377	378	379	380
381	382	383	384	385	386	387	388	389	390
391	392	393	394	395	396	397	398	399	400
401	402	403	404	405	406	407	408	409	410
411	412	413	414	415	416	417	418	419	420
421	422	423	424	425	426	427	428	429	430
431	432	433	434	435	436	437	438	439	440
441	442	443	444	445	446	447	448	449	450
451	452	453	454	455	456	457	458	459	460
461	462	463	464	465	466	467	468	469	470
471	472	473	474	475	476	477	478	479	480
481	482	483	484	485	486	487	488	489	490
491	492	493	494	495	496	497	498	499	500
501	502	503	504	505	506	507	508	509	510
511	512	513	514	515	516	517	518	519	520
521	522	523	524	525	526	527	528	529	530
531	532	533	534	535	536	537	538	539	540
541	542	543	544	545	546	547	548	549	550
551	552	553	554	555	556	557	558	559	560
561	562	563	564	565	566	567	568	569	570
571	572	573	574	575	576	577	578	579	580
581	582	583	584	585	586	587	588	589	590
591	592	593	594	595	596	597	598	599	600
601	602	603	604	605	606	607	608	609	610
611	612	613	614	615	616	617	618	619	620
621	622	623	624	625	626	627	628	629	630
631	632	633	634	635	636	637	638	639	640
641	642	643	644	645	646	647	648	649	650
651	652	653	654	655	656	657	658	659	660
661	662	663	664	665	666	667	668	669	670
671	672	673	674	675	676	677	678	679	680
681	682	683	684	685	686	687	688	689	690
691	692	693	694	695	696	697	698	699	700
701	702	703	704	705	706	707	708	709	710
711	712	713	714	715	716	717	718	719	720
721	722	723	724	725	726	727	728	729	730
731	732	733	734	735	736	737	738	739	740
741	742	743	744	745	746	747	748	749	750
751	752	753	754	755	756	757	758	759	760
761	762	763	764	765	766	767	768	769	770
771	772	773	774	775	776	777	778	779	780
781	782	783	784	785	786	787	788	789	790
791	792	793	794	795	796	797	798	799	800
801	802	803	804	805	806	807	808	809	810
811	812	813	814	815	816	817	818	819	820
821	822	823	824	825	826	827	828	829	830
831	832	833	834	835	836	837	838	839	840
841	842	843	844	845	846	847	848	849	850
851	852	853	854	855	856	857	858	859	860
861	862	863	864	865	866	867	868	869	870
871	872	873	874	875	876	877	878	879	880
881	882	883	884	885	886	887	888	889	890
891	892	893	894	895	896	897	898	899	900
901	902	903	904	905	906	907	908	909	910
911	912	913	914	915	916	917	918	919	920
921	922	923	924	925	926	927	928	929	930
931	932	933	934	935	936	937	938	939	940
941	942	943	944	945	946	947	948	949	950
951	952	953	954	955	956	957	958	959	960
961	962	963	964	965	966	967	968	969	970
971	972	973	974	975	976	977	978	979	980
981	982	983	984	985	986	987	988	989	990
991	992	993	994	995	996	997	998	999	1000

## Main: 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.

## Main: Initial data analysis and preprocessing

### 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

## Main: Initial data analysis and preprocessing

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 re-fine/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

## Main: Initial data analysis and preprocessing

- Important first step in (every) data analysis, can be particularly challenging in HDD settings;
- “data preprocessing” is a term used in biomedical HDD settings, especially in the omics field;
- for HDD, a detailed examination of the distribution of every variable individually is rarely feasible,
  - ▶ instead calculate scores and select interesting cases;
- HDD typically contains uninformative variables that do not vary much across subjects (variability only reflecting noise),
  - ▶ standardization can exaggerate the noise;
- issues with missing values in HDD settings:
  - ▶ multiple imputation approaches typically performs poorly;
  - ▶ complete case analysis may exclude too many observations.

## Main: Exploratory data analysis

### 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

## Main: Exploratory data analysis

- Visual identification of interesting characteristics of HDD requires specialized graphical displays / data reduction;
- difficulties with cluster analysis in HDD:
  - ▶ mixtures of low-dimensional parametric probability distributions cannot be applied at all or perform very poorly;
  - ▶ partitioning algorithms can present computational challenges;
  - ▶ algorithms may not converge to a solution;
  - ▶ hard to identify the right number of cluster (e.g., scree plots suffer from noise accumulating over the variables.

## Main: Identification of informative variables and multiple testing

### 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



## Main: Identification of informative variables and multiple testing

- Frequent goals in HDD settings:
  - ▶ identify variables related to a **single outcome**;
  - ▶ identify variables with a **trajectory over time** affected by experimental factors or exhibiting a prescribed pattern;
  - ▶ identify variables related to **other variables**;
- Specific methods developed for **testing in HDD**:
  - ▶ for hypothesis testing for single variables (e.g. limma, edgeR, DESeq2) – **borrow information** among variables;
  - ▶ for multiple testing correction: control of **false discovery rate**.

## Main: Prediction

## 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

## Main: Prediction

- Numerous **dramatic claims** of performance of prediction models have been made using HDD,
  - ▶ not always withstand **careful validation**;
- Need to identify the optimal level of **model complexity**:
  - ▶ that will yield **interpretable models**;
  - ▶ good **prediction performance** on independent data;
- Balance between **overfitting** ...
  - ▶ too specific to the data at hand;
  - ▶ identifies more or too complex patterns than real ones;
- ... and **underfitting**,
  - ▶ misses important patterns useful for prediction.

## Main: Prediction

- Approaches used to overcome issues related to the  $p \gg n$  situation in the model building process:
  - ▶ variable pre-selection (SIS, ISIS, ...);
  - ▶ dimensionality reduction (PCR, LSR, ...);
  - ▶ constrained optimization (LASSO, ridge, ...);
  - ▶ algorithmic approaches (random forests, neural networks, ...).
- Other challenges particularly interesting in the HDD context:
  - ▶ data integration (data sources, external knowledge, ...);
  - ▶ identification of influential points;
  - ▶ evaluation of prediction models.

## Remarks

Back to TG9. Other projects / topics not considered here:

- simulations in HDD:
  - ▶ 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);
- reporting / transparency / reproducibility.



Visit [https://www.stratos-initiative.org/group\\_9](https://www.stratos-initiative.org/group_9)