

High-dimensional statistics

topics

- ▶ High-dimensional (generalized) linear models:
Lasso and modifications → Book
- ▶ Group Lasso for group sparsity → Book
- ▶ Additive models and many smooth univariate functions
→ Book
- ▶ De-biasing and de-sparsified Lasso → Additional Material
- ▶ Stability selection → Book
- ▶ Multiple sample splitting → Book
- ▶ Hidden confounding and deconfounding methods
- ▶ Undirected graphical modeling → Book

(Text-) Book:

Bühlmann, P. and van de Geer, S. (2011). Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer.

Additional Material:

<https://stat.ethz.ch/~buhlmann/teaching/desparsifiedLasso.pdf>

James, Witten, Hastie and Tibshirani (2023). An Introduction to Statistical Learning with Applications in R/Python, 2nd ed. Springer

See also: <https://stat.ethz.ch/~buhlmann/teaching/highdim-stats-HS2023.html>

[//stat.ethz.ch/~buhlmann/teaching/highdim-stats-HS2023.html](https://stat.ethz.ch/~buhlmann/teaching/highdim-stats-HS2023.html)

An example: Riboflavin production with Bacillus Subtilis

(in collaboration with DSM (Switzerland))

goal: improve riboflavin production rate of Bacillus Subtilis
using clever genetic engineering

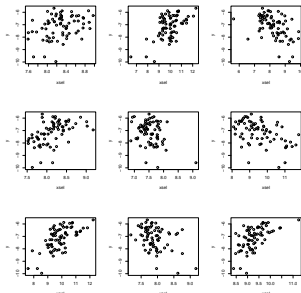
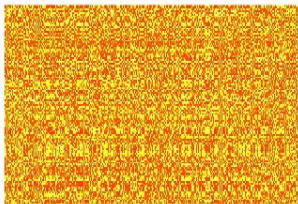
response variables $Y \in \mathbb{R}$: riboflavin (log-) production rate

covariates $X \in \mathbb{R}^p$: expressions from $p = 4088$ genes

sample size $n = 115$, $p \gg n$

Y versus 9 “reasonable” genes

gene expression data



High-dimensional data

general framework:

$$Z_1, \dots, Z_n \text{ i.i.d.}, \quad \dim(Z_i) \gg n$$

for example:

$Z_i = (X_i, Y_i)$, $X_i \in \mathbb{R}^p$, $Y_i \in \mathbb{R}$: regression with $p \gg n$

$Z_i = (X_i, Y_i)$, $X_i \in \mathbb{R}^p$, $Y_i \in \{0, 1\}$: classification with $p \gg n$

$Z_i = (X_i)$, $X_i \in \mathbb{R}^p$: graphical modeling with $p \gg n$

numerous applications:

biology, imaging, economy, environmental sciences, ...