High-dimensional statistics

topics

- High-dimensional (generalized) linear models:
  Lasso and modifications
- Group Lasso for group sparsity
- Additive models and many smooth univariate functions
  Non-convex loss functions and ℓ_1-norm regularization

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

- Statistical inference and uncertainty quantification
  De-biasing and de-sparsified Lasso
  Stability selection
  Multiple testing

Additional text:

`https://stat.ethz.ch/~buhlmann/teaching/highdim-stats-HS2021.html`

- Hidden confounding and deconfounding methods

Additional text (see later)

- Undirected graphical modeling

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

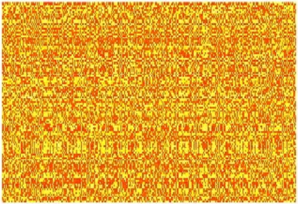
An example: Riboflavin production with Bacillus Subtilis

(in collaboration with DSM (Switzerland))

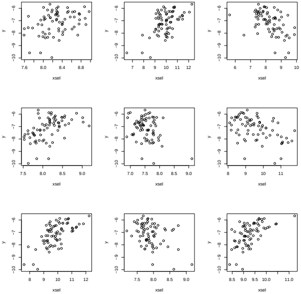
goal: improve riboflavin production rate of Bacillus Subtilis using clever genetic engineering

response variables $Y \in \mathbb{R}$: riboflavin (log-) production rate
covariates $X \in \mathbb{R}^p$: expressions from $p = 4088$ genes
sample size $n = 115$, $p \gg n$

gene expression data



Y versus 9 “reasonable” genes



High-dimensional data

general framework:

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

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, ...