ST4Health – ~~MLGenom~~ Stat Genom

High-dimensional statistics and machine learning for genomics

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)
Mines ParisTech – Institut Curie – INSERM U1331
PSL Research University & PRAIRIE-PSAI, Paris, France

Nov 3, 2026

http://cazencott.info chloe-agathe.azencott@minesparis.psl.eu @cazencott@lipn.info

Statistics (and machine learning) for genomics

Detecting differences in genomes between patients & healthy controls

Find patterns (eg Methylation patterns)

GWAS / biomarker discoveries / mutations that are associated with a phenotype

Predict chromatin accessibility

Causality?

Inferring evolutionary trees

Identifying disease subtypes

Predicting function of a DNA sequence

High dimensional data

Data matrix: features

samples



wide ("fat data")

number samples << number of features

a few thousand (or less)

gene expression data : ≃ 20 000 features

Single nucleotide polymorphisms ≃ 500 000 – 1 million features

Course programme

Nov 3	Introduction
Nov 5	Gene expression analysis
Nov 17 & Dec 1	Genome-wide association studies & Mendelian randomization
Dec 8 & 15	Dimensionality reduction & Clustering
Jan 5	Exam

- Each topic: part lecture, part lab (Python or R)
- **Prepare the lab before the session**
- **Lecturers: Chloé Azencott**, **Florian Massip** and **Marie Verbanck**
- **Teaching assistant: Katia Antonenko**
- https://moodle.psl.eu/user/index.php?id=39582

Exam (Jan. 5)

60% oral evaluation + 30% written quiz + 10% engagement

- **Written quiz** 15 minutes

Exam (Jan. 5)

60% oral evaluation + 30% written quiz + 10% engagement

- **Written quiz** 15 minutes
- **Oral evaluation** in **4 or 5 groups of 2 or 3 students**
- Present a **publication** chosen among those proposed (see on Moodle):
  - Biological question
  - Limitations of state-of-the-art approaches
  - Methodological contribution
  - Selected results
  - Discussion
- 15 minutes for the presentation + 15 minutes for questions

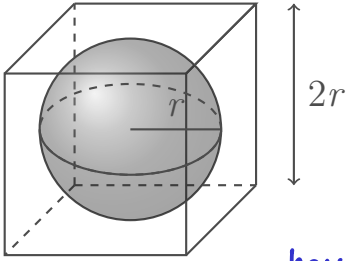
Exam (Jan. 5)

60% oral evaluation + 30% written quiz + 10% engagement

- **Written quiz** 15 minutes
- **Oral evaluation** in **4 or 3 groups of 2 or 3 students**
- Present a **publication** chosen among those proposed (see on Moodle):
  - Biological question
  - Limitations of state-of-the-art approaches
  - Methodological contribution
  - Selected results
  - Discussion
- 15 minutes for the presentation + 15 minutes for questions
- **Form groups and chose publication (full list on Nov. 5) on Moodle by Nov. 17**

I. The curse of dimensionality

Points are far apart in high dim
Nothing is local



In dimension d:

Volume of the cube	Volume of the sphere
$(2r)^d$	$\frac{2r^d \pi^{d/2}}{d\Gamma(d/2)}$

Γ generalizes the factorial: $\Gamma(n) = (n-1)!$ if $n \in \mathbb{N}$

how much of the circle is inside the square?

In 2D = 

area of the square = $4r^2$
area of the circle = πr^2

↳ $\frac{\pi r^2}{4r^2} = \frac{\pi}{4}$

In 3D = volume of the cube = $8r^3$
volume of the sphere = $\frac{4}{3}\pi r^3$

$\frac{4/3 \pi r^3}{8r^3} = \frac{\pi}{6}$

Goes to 0 as d grows

Bellman (1961), **Adaptive control processes: a guided tour,** Princeton University Press

II. Feature selection in high dimension

Lots of problems in genomics can be framed as feature sel

→ find relevant genomic features for the phenomenon of interest

- 1. Filter methods → scoring each feature individually.
- 2. Embedded methods

II.1. Filter methods: hypothesis testing

Statistical testing for feature selection : independance testing

feature X label Y

null hypothesis H_0 : X and Y are independant

alternative hypothesis : X and Y are correlated

test statistic : Q chosen so that we know its distribution under H_0

If $\underbrace{\mathbb{P}_{H_0}(Q > q)}_{\text{p-value}} \leq \alpha$ ← significance level

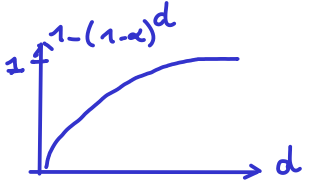
q ↑ observed value of Q on the data

then we reject H_0

Multiple testing = as many as features (= d)

for $j = 1..d$: $\mathbb{P}_{H_0}(Q > q_j) \leq \alpha$ (p-value)

- Single test : probability of rejecting H_0 when it's actually true ? $= \alpha$
  (FD) false discovery / positive
  ≡ selecting the feature when we shouldn't
- d independant tests
  $\mathbb{P}(\text{at least 1 FD}) = 1 - \underbrace{\mathbb{P}(\text{no FD})}_{(1-\alpha)^d}$



The probability of making at least 1 FD grows with d towards 1.
→ need for "Multiple Hypotheses Testing Correction"

Control of the Family-Wise Error Rate

$\mathbb{P}$ (at least 1 FD)

Bonferroni correction

significance level $\alpha \longrightarrow \frac{\alpha}{d}$ equivalently: multiply all p-values by d

- Assumes independance between tests
- Very conservative

Control of the False Discovery Rate

$$\mathbb{E}\left[\frac{\text{number of FD}}{\text{number of discoveries}}\right]$$

← significant test = selected feature

Benjamini – Hochberg correction

$$p_1 \leq p_2 \leq \cdots \leq p_j \cdots \leq p_d$$

$$p_j \leq \frac{j}{d}\,\alpha$$

stop at the largest such j

- Less stringent than Bonferroni
- Also assumes independance between the tests → this is necessary to guarantee that the FDR is controlled.

Correlation between features

– Genomic features will be correlated

→ Tests are not really independant

Benjamini – Yuketeli for non-independent tests

$$p_j \leq \frac{j}{d} \alpha$$

replace with $\frac{j}{d \cdot c(d)}$

Harmonic number

$$c(d) = \sum_{k=1}^{d} \frac{1}{k}$$

If $d = 15\,000$ then $c(d) = 10$

$d = 2 \cdot 10^6$ $c(d) = 15$

II.2. Embedded methods

- Consider all the features together
- Building a predictive model and selecting at the same time

Ex) Lasso ℓ_1-regularized linear or logistic regression

$X \in \mathbb{R}^{n \times d}$, $\vec{y} \in \mathbb{R}^n$

$$\arg\min_{\vec{\beta} \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} L\left(\langle \vec{\beta}, \vec{x}_i \rangle, y^i\right)}_{\text{empirical risk}} + \lambda \underbrace{\sum_{j=1}^{d} |\beta_j|}_{\text{regularizer}}$$

↳ sparse model

i.e. some of the β_j are equal to 0

the features with non-0 coefficients are selected

Possible to get p-values (H_0: $\beta_j = 0$) but not yet mainstream.

Group Lasso

$$\arg\min_{\vec{\beta} \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=1}^{n} L(\langle \vec{\beta}, \vec{x}_i \rangle, y^i)}_{\text{empirical risk}} + \lambda \underbrace{\sum_{j=1}^{d} |\beta_j|}_{\text{regularizer}}$$

predefined groups of features that should be either selected all together or not at all

G groups

$$\sum_{g=1}^{G} \sqrt{p_g} \sum_{j \in \text{group } g} \beta_j^2$$

$\sqrt{p_g}$: size of group g

→ sparsity at the group level

Other variants

* graph regularizers → selected features are connected on the provided network
* tree regularizes

III. Other challenges

- Confounding
  e.g. Identifying ≠ between batches rather than between conditions

- Underrepresented populations