Lesson 4: High dimensional statistics and multiple hypothesis testing

Zacharie Naulet

zacharie.naulet@math.u-psud.fr

Contents

1 The curse of dimensionality

1.1 The problem

- We want to consider problems where $p \gg n$.
  Typically, in biology the typical setup is $p \approx 10^5$ or $p \approx 10^6$ but $n \approx 10$ or at best $n \approx 20$.

- This is an issue for least squares estimation, because it is impossible to consider models with $|J| \geq n$. here we use the same conventions as in the previous lecture: The complete model includes $\{1, \dots, p\}$ covariates, and we let $J \subseteq \{1, \dots, p\}$ be a subset of those covariates. We say that we use model J when we consider the model using the covariates with indices in J only.

- We could use the strategy we saw in the previous lecture to search for the **best** model within all the models with $|J| < n$.

  - This is in general not computationally not feasible: If we have p covariates and want to try all possible models using k covariates among the p available, this gives $\binom{p}{k}$ models to try. When $k \gg 2$ and $p \gg k$ this is just **huge**.

- We could try the greedy algorithms too, but they also easily fail, especially when the covariates are correlated.
  Example: If $y_i = x_2 + x_3 + \varepsilon_i$ (true model), and we have also the covariate $x_1 = \frac{2}{3}x_2 + \frac{1}{3}x_3$, then the algorithm will select x_1 as a first covariate, and then we will end up with the model $\{1, 2, 3\}$ instead of the ideal model $\{2, 3\}$.

- There are also **statistical limitations**: If there are too many covariates, it is simply impossible to say anything, or at least it is difficult to be confident about what we say.

1.2 The solutions

- **From the statistics viewpoint:** If $p \gg n$ we need to assume something more on the model to be able to do statistics.
- The classical assumption is **sparsity**:
  The true model J_* is small, *i.e.* $|J_*| \ll n$. Typically we need at least to have $|J_*| \ll \frac{n}{\log(p)}$.
- Assumptions on the complexity of the model are **necessary** when $p \gg n$, otherwise we cannot say anything. We also note a subtlety here: though we have to assume $|J_*|$ is small, this do not totally simplifies the problem: we just know that only $|J_*| \ll n$ covariates are useful, but we don't know exactly how much, neither which they are!
- If we have unlimited computational resources: we can do model selection using a penalty depending on $|J|$, the noise (difficult to estimate when $p \gg n$, indeed) and p and n. But as we discussed earlier, this requires to many computations.
- A solution that actually works: the **LASSO** algorithm *(Least Absolute Shrinkage and Selection Operator).*
  - For $\lambda \geq 0$ to be chosen, use as an estimator:

$$\hat{\beta}_\lambda^{lasso} \in \arg\min_{\beta \in \mathbb{R}^p} \Big\{ \|Y - \boldsymbol{X}\beta\|^2 + \lambda \sum_{j=1}^{p} |\beta_j| \Big\}.$$

  - If λ is large enough: actually does select variables (indeed, it sets automatically $\hat{\beta}_{\lambda,j} = 0$ whenever there is not enough signal to estimate β_j and/or $\beta_j = 0$.
  - A good choice for λ (but this does not tell the whole story!) is to take $\lambda \propto \sigma\sqrt{\log(p)}$. Under certain assumption, we can show this choice guarantee that with large probability $\{j \,:\, \hat{\beta}_{\lambda,j} \neq 0\} \approx J_*$.
  - The good-news: it is computationally **fast**: can be implemented to be computed in $O(n^2 p)$ operations (compare with exhaustive search!).

2 Multiple Hypothesis testing

2.1 Issues with multiple testing

- If we do a lot of tests on the same data, we will always find something, even if there is nothing to find (false discovery).

- Example:
  - Assume we have p covariates (in biology: genes), and we want to test the effect of each of these covariates on the response variable Y (it can be for instance testing the effect of each gene on the glycemia).
  - Then, we want to run p tests, corresponding to

$$H_0 : Y = \beta_1 + \varepsilon \qquad H_{1,j} : Y = \beta_1 + x_j\beta_j + \varepsilon,$$

    i.e. we want to test if the gene number j has en effect on Y or not, and that for every genes $\{1, \dots, p\}$.
  - So we do p tests with level α, one per alternative $H_{1,j}$.
  - Now suppose that none of the gene has an effect on Y, that is H_0 is always true. In this case, by definition of α, we will fail to choose H_0 with probability α, and since we did p tests and p is very large, in average we will be mistaken $\alpha \times p$ times, ie if $p \gg 1/\alpha$, we are very likely to be mistaken at least one time. We will indeed do a **false discovery**.
- Note that in biology, it is very common to have $\alpha = 0.05$ and $p \approx 10^5$, meaning that $p\alpha \approx 5000$: If we are not careful we might do as much as thousands false discovery.
- This issue is not always well taken into account by scientists, so please be careful!

2.2 Bonferroni correction and FWER

- Say that we want to do N tests $\{H_{0,j} \text{ vs } H_{1,j},\ j = 1, \dots, N\}$, and those tests have corresponding p-values $p_1, \dots, p_N$.
- Usually, we reject $H_{0,j}$ for all j such that $p_j \leq \alpha$.
  But indeed, if we don't take additional cares, we do in average a number of false discoveries which is $\approx \alpha\times$ the number j such that $H_{0,j}$ is true; which can be as large as $\approx \alpha \times N$ (see previous example!).
- **Bonferroni correction**: Reject $H_{0,j}$ for all j such that $p_j \leq \frac{\alpha}{N}$.
  Example: if $\alpha = 0.05$ and $N = 10^5$, then $\frac{\alpha}{N} = 5.10^{-7}$.

  Proposition 1. *Consider the Bonferroni correction. Then,*

$$\underbrace{\text{proba}\Big(\exists j \,\text{such that } H_{0,j} \text{ true and } p_j \leq \frac{\alpha}{N}\Big) \leq \alpha}_{\text{FWER}}.$$

  Here, FWER means ***Family-Wise Error Rate****, which is the probability of having at least one false discovery. The probability is understood under the assumption that all the $H_{0,j}$ are true.*

- As seen in the previous proposition, the Bonferroni correction consists on tweaking the level of each tests so that the probability of having at least one false discovery (under $H_{0,j}$ being true for all j) is no more than α.

2.3 Other notions of errors

- The FWER is quite conservative, we can do other stuff as well.
- Consider N tests for $H_{0,j}$ against $H_{1,j}$, $j = 1, \dots, N$.
- We define the following vocabulary (here it is understood that a discovery is positive, *i.e.* the test is conclusive, if we succeed in rejecting H_0 when H_0 is false).
  - TP: The number of true positives, *i.e.* the number of j for which $H_{0,j}$ is false and for which we rejected $H_{0,j}$
  - TN: The number of true negatives, *i.e.* the number of j for which $H_{0,j}$ is true and for which we accepted $H_{0,j}$.
  - FP: The number of false positives, *i.e.* the number of j for which $H_{0,j}$ is true, but for which we rejected $H_{0,j}$.
  - FN: The number of false negatives, *i.e.* the number of j for which $H_{0,j}$ is false, but for which we accepted $H_{0,j}$.

	H_0 is true	H_1 is true
We choose H_0	True negative	Type II error / False negative
We choose H_1	False positive (proba α)	True positive

Table 1: Summary of possible outcomes for hypothesis testing.

- We note that FWER = proba(FP ≥ 1).
- We can also define the Proportion of False Discoveries FDP:

$$\text{FDP} := \begin{cases} \frac{\text{FP}}{\text{TP}+\text{FP}} & \text{if TP} + \text{FP} \geq 1, \\ 0 & \text{otherwise.} \end{cases}$$

- The **False Discovery Rate** (FDR) is defined as the $\mathbb{E}[\text{FDP}]$, wheere the expectation is understood under the $H_{0,j}$'s.
- We always have FDR $\leq$ FWER.
- So, if we use the Bonferroni correction, we also control the FDR and we have FDR $\leq$ FWER $\leq \alpha$.
- But, we might want to be less conservative that Bonferroni and control only the FDR (it is clear that a small FWER implies small FDR, but the reverse is certainly not true in general). In particular, using the FWER could prevent from doing some discoveries.
- **Benjamini-Yekutieli procedure:**
  - Let $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(N)}$ be the p-values of the tests, but ordered in increasing order.
  - Reject all the $H_{0,j}$ corresponding to the K smallest p-values, where K is given by

$$K := \max\left\{ k \,:\, p_{(k)} \leq \frac{k\alpha}{N \sum_{j=1}^{N} \frac{1}{j}} \right\}.$$

  - This procedures guarantees that FDR $\leq \alpha$.