Statistics for high-dimensional data
Overview

Anne–Laure Boulesteix
Institut für Medizinische Informationsverarbeitung, Biometrie und Epidemiologie

February 15th, 2017

Structure

1. **What are high-dimensional data?**

2. High-dimensional data and multiple testing

3. High-dimensional data and prediction models

Objectives of the lectures

- Being able to perform simple statistical analyses with high-dimensional data by yourself (multiple testing, simple prediction models)
- Understanding the main statistical principles to facilitate discussions with statisticians on more complex issues
- Identifying potential "dangers" and problems
- Better interpreting results from the literature from a statistical perspective

High-dimensional

Classical data:

Index	Outcome	Sex	Age	Smoking	BMI
1	bad	F	54	yes	28
2	good	M	50	yes	22
...	...	...	...	...	

High-dimensional data:

Index	Outcome	Sex	Age	Smoking	BMI	V_1	...	...	V_{1000}
1	bad	F	54	yes	28	0.97	...	...	1.33
2	good	M	50	yes	22	1.25	...	...	2.87
...	...	...	...	...		...	...		

High-dimensional data

- **Low dimensional data** (most epidemiological studies):

$$n \gg \text{ number of variables}$$

  Example: $n = 300$ with $p = 15$

- **High-dimensional data** (from high-throughput experiments):

$$n \approx \text{ or even } \ll \text{ number of variables}$$

  Example: $n = 70$ with $p = 2000$, $p = 20000$, $p = 200000$...

Omics data

Omics data = molecular data from high-throughput technologies (e.g. microarrays or NGS)

data type	**number of variables**
metabolomic	a few hundreds
proteomic	a few hundreds/thousands
transcriptomic	a few tens of thousands
CNV	a few hundreds of thousands
SNP	a few hundreds of thousands or millions
...	...

Omics data are by definition high-dimensional.

Issues with high-dimensional data

- high computational cost
- automatization of analysis code is necessary
  e.g. `for (i in 1:1000)` etc
- no easy graphical representation
- **issues with statistical analysis**

Issues with statistical analysis

- Many standard approaches (e.g. least squares regression) are not applicable here.
- Many statistical methods have been designed especially for the analysis of such high-dimensional data.
- **Problem 1:** There is still a lack of guidelines and no commonly accepted standards.
- **Problem 2:** In high-dimensional settings all results of statistical analyses are highly variable (e.g. change a lot if you consider a slightly modified version of the data or a slightly different method)

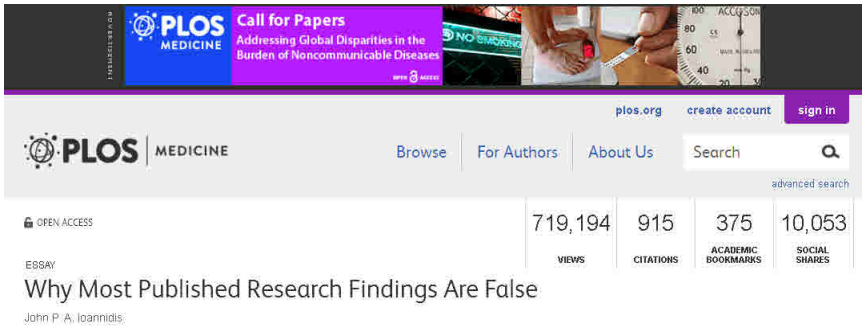
The result

Fishing for significance!

Noise discovery

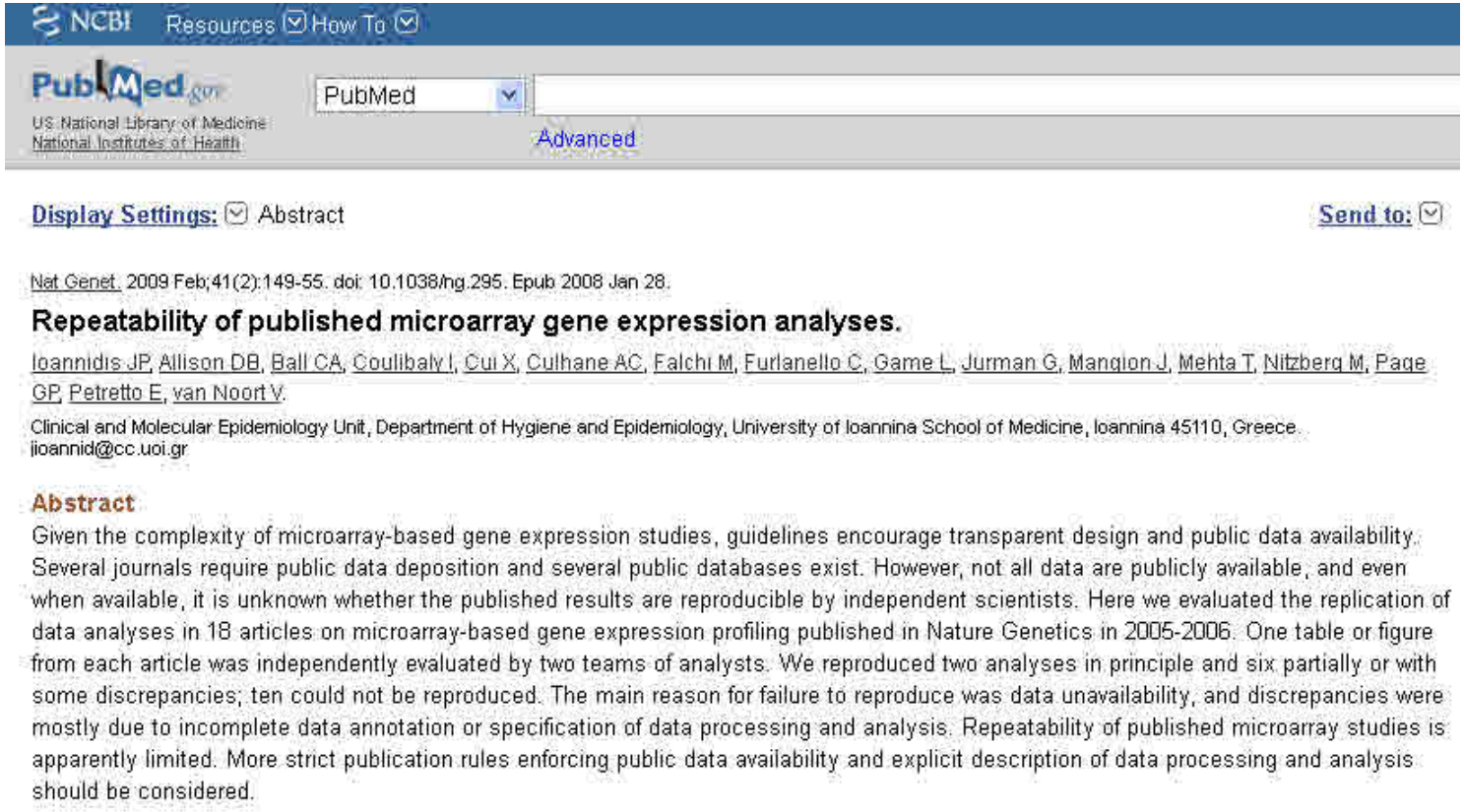
Microarrays and molecular research: noise discovery?

"Give me information on a single gene and 200 patients, half of them dead, please. I bet that I can show that this gene affects survival ($p<0.05$) even if it does not. One can do analyses: counting or ignoring exact follow-up, censoring at different timepoints, excluding specific causes of death, exploiting subgroup analyses, using dozens of different cut-offs to decide what constitutes inappropriate gene expression, and so forth. Without highly specified a priori hypotheses, there are hundreds of ways to analyse the dullest dataset. Thus, no matter what my discovery eventually is, it should not be taken seriously, unless it can be shown that the same exact mode of analysis gets similar results in a different dataset. Validation becomes even more important when datasets become complex and analytical options increase exponentially."



- **Corollary 1:** The smaller the studies conducted in a scientific field, the less likely the research findings are to be true.
- **Corollary 2:** The smaller the effect sizes in a scientific field, the less likely the research findings are to be true.
- **Corollary 3:** The greater the number and the lesser the selection of tested relationships in a scientific field, the less likely the research findings are to be true.
- **Corollary 4:** The greater the flexibility in designs, definitions, outcomes, and analytical modes in a scientific field, the less likely the research findings are to be true.
- **Corollary 5:** The greater the financial and other interests and prejudices in a scientific field, the less likely the research findings are to be true.
- **Corollary 6:** The hotter a scientific field (with more scientific teams involved), the less likely the research findings are to be true.

Lack of reproducibility



NCBI Resources How To

PubMed.gov
US National Library of Medicine
National Institutes of Health

PubMed

Advanced

Display Settings: Abstract

Send to:

Nat Genet. 2009 Feb;41(2):149-55. doi: 10.1038/ng.295. Epub 2008 Jan 28.

Repeatability of published microarray gene expression analyses.

Ioannidis JP, Allison DB, Ball CA, Coulibaly I, Cui X, Culhane AC, Falchi M, Furlanello C, Game L, Jurman G, Mangion J, Mehta T, Nitzberg M, Page GP, Petretto E, van Noort V.

Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, Ioannina 45110, Greece.
jioannid@cc.uoi.gr

Abstract

Given the complexity of microarray-based gene expression studies, guidelines encourage transparent design and public data availability. Several journals require public data deposition and several public databases exist. However, not all data are publicly available, and even when available, it is unknown whether the published results are reproducible by independent scientists. Here we evaluated the replication of data analyses in 18 articles on microarray-based gene expression profiling published in Nature Genetics in 2005-2006. One table or figure from each article was independently evaluated by two teams of analysts. We reproduced two analyses in principle and six partially or with some discrepancies; ten could not be reproduced. The main reason for failure to reproduce was data unavailability, and discrepancies were mostly due to incomplete data annotation or specification of data processing and analysis. Repeatability of published microarray studies is apparently limited. More strict publication rules enforcing public data availability and explicit description of data processing and analysis should be considered.

Specific issues with statistical analysis of $n \ll p$ data

- Multiple testing
- Regression, prediction models
- Other topics:
  - global testing
  - clustering
  - added predictive value
  - reproducibility

Structure

1. What are high-dimensional data?

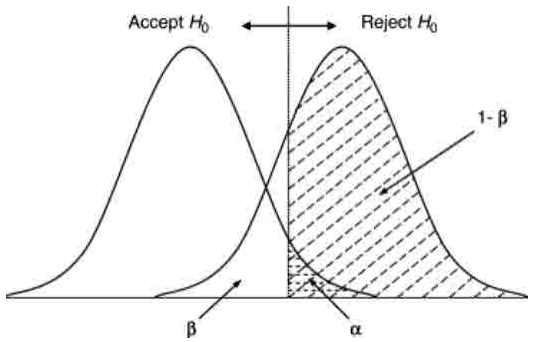
2. **High-dimensional data and multiple testing**

3. High-dimensional data and prediction models

Testing

Univariate analyses:
Test whether the means of $V_1, \ldots, V_{1000}$ are different in groups "bad" and "good".

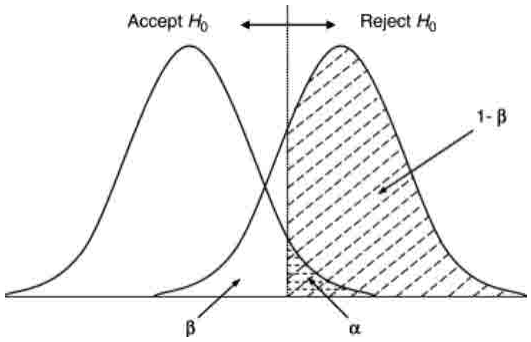
Naive approach: Perform a t-test at the level 0.05 for each variable, i.e. reject the null-hypothesis of equality of the means for all variables having a p-value < 0.05.



Multiple testing

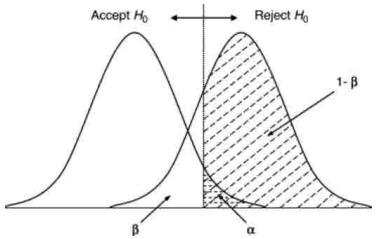
Naive approach: Perform a t-test at the level 0.05 for each variable, i.e. reject the null-hypothesis of equality of the means for all variables having a p-value < 0.05.

Problem: A p-value of 0.05 means that the probability to observe this value of the test statistic or a more extreme value is 0.05.



0.05 = 5% is low, but not zero!

Multiple testing



- 0.05 = 5% is low, but not zero!

- If we consider 1000 tests and assume that all null-hypotheses are true (i.e. should not be rejected), the null-hypothesis will (wrongly) be rejected for 5%× 1000=50 variables!

→ The naive approach yields 50 *false positives* in this case and gives the impression that there are "interesting variables" in the data although it is not the case.

Multiple testing procedures

- The naive approach yields 50 *false positives* in this case and gives the impression that there are "interesting variables" in the data although it is not the case.
- Multiple testing procedures aim to avoid this.
- The idea is to set the significance threshold smaller than 0.05 or - equivalently - to apply an upward correction to the p-values.
- This reduces the amount of false positives.

Structure

1. What are high-dimensional data?

2. High-dimensional data and multiple testing

3. **High-dimensional data and prediction models**

Prediction models

Let us consider a continuous outcome, e.g. IgE level.

If we have the predictors "sex" and "age", we can fit the regression model

$$IgE = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot age$$

to the data at hand ($n = 100$), i.e. estimate the regression coefficients β_0 (intercept), β_1 (sex) and β_2 (age).

Prediction models

In practice such regression analyses are performed for two different purposes:

- testing significance of predictors in the model to assess their importance
- building a prediction model that could be ideally be applied to make (useful) predictions for future patients.

Prediction models

The regression coefficients $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)^\top$ are estimated by minimization of the sum of squares as

$$\hat{\boldsymbol{\beta}} \;=\; (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y},$$

where $\boldsymbol{Y}$ is the vector containing the IgE levels and $\boldsymbol{X}$ is the three-column data matrix containing a column of ones, a column with sex (coded as 0/1) and a column with age.

Dimension problem

- Now suppose we want to fit a regression model for the outcome IgE with sex, age and $V_1, \ldots, V_{1000}$ as predictors.

- **Problem:** We have to estimate as many as **1003 regression coefficients** based on only $n = 100$ patients!

- This is an ill-posed problem: there are infinitely many possible combinations of regression coefficients that would perfectly fit the data.

- One does not know which one to choose, and these solutions are not expected to fit well future independent data.

Dimension problem

Mathematical explanation: in the formula

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y},$$

the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is not invertible if $\boldsymbol{X}$ has 1003 columns and 100 rows, so $\hat{\boldsymbol{\beta}}$ is not uniquely defined.

Dimension problem

- This problem is not limited to least squares linear regression.
- Similar mechanisms hold for logistic regression models of the form

$$\log \frac{P(Y=1|age, sex, \dots)}{P(Y=0|age, sex, \dots)} = \beta_0 + \beta_1 \cdot sex + \beta_2 \cdot age + \dots,$$

  and for Cox regression models of the form

$$\lambda(t|age, sex, \dots) = \lambda_0(t) \exp(\beta_1 \cdot sex + \beta_2 \cdot age + \dots).$$

Prediction models

- Since standard regression methods do not work, alternative methods have to be used to cope with the high dimension of the predictor space.
- A related issue is **overfitting** of the training data.
- The evaluation of **prediction accuracy** is a highly complex topic in the context of high-dimensional data analysis.