Basic statistics for bioinformatics: differential expression, multiple testing, classification

Ramon Diaz-Uriarte
Dept. Biochemistry, Universidad Autónoma de Madrid
Instituto de Investigaciones Biomédicas “Alberto Sols” (UAM-CSIC)
Madrid, Spain

ramon.diaz@iib.uam.es, rdiaz02@gmail.com

http://ligarto.org/rdiaz

2017-01-20 (Release 1.2: Rev: 89d3a88)

License and copyright

****

Please, **respect the copyright and license**. This material is provided freely. If you use it, I only ask that you use it according to the (very permissive) terms of the license: acknowledging the author, and redistributing copies and derivatives under the same license. If you have any doubts, ask me.

Availability

All of the files are available from `https://github.com/rdiaz02/Stats-bioinfo-intro`.
You will need knitr, beamer, etc, to produce the pdfs.
Again, please respect the copyright and license.

Outline

- Introduction to Omics and statistics issues
- Differential expression and multiple testing
- Classification/prediction (and clustering)

Objectives

- Be aware that from data to biomedical conclusions there are several steps that require statistics. We want to make inferences in a noisy world.
- Be aware of the “big themes”.
- Understand the origin of some of the statistical issues.
- Know when you need to talk to a statistician (almost always).
- Be aware of the kinds of things a statistician is thinking.

Key idea

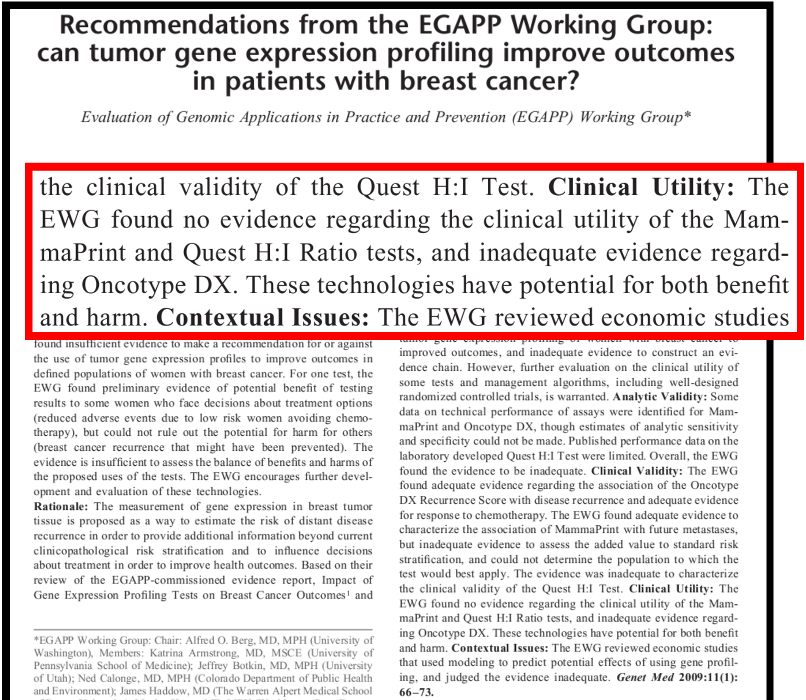
- We measure:
    - “Expression” or “Mutation status” or . . . of genes/probes
    - For a set of samples (subjects, patients, mice, cell lines, etc)
- Many, many genes/probes (tens of thousands, millions)
- Tens to thousands of samples
- And we want to do something with that

What general issues matter to us?

- What are you targeting? What do you want to measure?
  - Abundance of transcripts
  - Copy number changes
  - Relative abundance of polymorphisms
  - . . .
- What type of variable is that?
  - The number in the cell of the spreadsheet
  - Continuous-like vs. count
- Need for normalization
  - E.g., microarrays: GC content

Some common questions

- Are there groups in the genes?
- Are there groups in the subjects?
- Do groups of subjects differ in the expression of some genes?
- Can we find genes that will allow us to differentiate between the groups of patients?
- All of this, in a context of high expectations . . .



Recommendations from the EGAPP Working Group: can tumor gene expression profiling improve outcomes in patients with breast cancer?

*Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group**

the clinical validity of the Quest H:I Test. **Clinical Utility:** The EWG found no evidence regarding the clinical utility of the MammaPrint and Quest H:I Ratio tests, and inadequate evidence regarding Oncotype DX. These technologies have potential for both benefit and harm. **Contextual Issues:** The EWG reviewed economic studies

found insufficient evidence to make a recommendation for or against the use of tumor gene expression profiles to improve outcomes in defined populations of women with breast cancer. For one test, the EWG found preliminary evidence of potential benefit of testing results to some women who face decisions about treatment options (reduced adverse events due to low risk women avoiding chemotherapy), but could not rule out the potential for harm for others (breast cancer recurrence that might have been prevented). The evidence is insufficient to assess the balance of benefits and harms of the proposed uses of the tests. The EWG encourages further development and evaluation of these technologies.

Rationale: The measurement of gene expression in breast tumor tissue is proposed as a way to estimate the risk of distant disease recurrence in order to provide additional information beyond current clinicopathological risk stratification and to influence decisions about treatment in order to improve health outcomes. Based on their review of the EGAPP-commissioned evidence report, Impact of Gene Expression Profiling Tests on Breast Cancer Outcomes[1] and

improved outcomes, and inadequate evidence to construct an evidence chain. However, further evaluation on the clinical utility of some tests and management algorithms, including well-designed randomized controlled trials, is warranted. **Analytic Validity:** Some data on technical performance of assays were identified for MammaPrint and Oncotype DX, though estimates of analytic sensitivity and specificity could not be made. Published performance data on the laboratory developed Quest H:I Test were limited. Overall, the EWG found the evidence to be inadequate. **Clinical Validity:** The EWG found adequate evidence regarding the association of the Oncotype DX Recurrence Score with disease recurrence and adequate evidence for response to chemotherapy. The EWG found adequate evidence to characterize the association of MammaPrint with future metastases, but inadequate evidence to assess the added value to standard risk stratification, and could not determine the population to which the test would best apply. The evidence was inadequate to characterize the clinical validity of the Quest H:I Test. **Clinical Utility:** The EWG found no evidence regarding the clinical utility of the MammaPrint and Quest H:I Ratio tests, and inadequate evidence regarding Oncotype DX. These technologies have potential for both benefit and harm. **Contextual Issues:** The EWG reviewed economic studies that used modeling to predict potential effects of using gene profiling, and judged the evidence inadequate. ***Genet Med*** **2009:11(1): 66–73.**

*EGAPP Working Group: Chair: Alfred O. Berg, MD, MPH (University of Washington), Members: Katrina Armstrong, MD, MSCE (University of Pennsylvania School of Medicine); Jeffrey Botkin, MD, MPH (University of Utah); Ned Calonge, MD, MPH (Colorado Department of Public Health and Environment); James Haddow, MD (The Warren Alpert Medical School

Moral (moraleja, in Spanish)

> *Gene expression technologies show great promise to improve predictions of prognosis and treatment benefit (. . .). The multidimensional nature of these predictors demands that (. . .) that exceptional rigor and discipline be applied in evaluation.*

L. Marchionni et al., *Ann Intern Medl*, 2008

The dangers of “capitalizing on chance”

Statistical context: many genes, few subjects. $p \gg n$.

Differentially expressed genes Risk of too many false positives ⇒ adjustments in the screening of p-values.

Classification/prediction Very easy to obtain algorithms that classify, perfectly, our data, but not new data ⇒ validate algorithms and classifiers

Hypotheses/questions Tempting to make them vague, or ask none and wait until “the data say something” ⇒ define objectives and how we will measure what we are interested in.

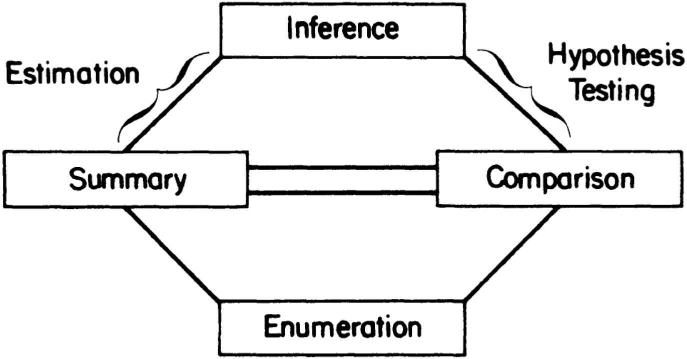


Figure 1. Four Basic Statistical Operations and How They Relate to Estimation. Source: Efron (1982b, fig. 2).

(Efron, 1986, *The American Statistician*, 40: 1–11)

Are there differences in the expression of certain genes between/among groups of subjects?

If we have 2 (or 3, or 4, or ...) kinds of subjects (e.g., breast cancer vs. colon cancer), what genes behave differently?

Differential expression vs. classification

[Classification/prediction]

Can we classify subjects into their true groups if we know the expression of certain genes?

(Are there genes that can allow us to differentiate between groups of subjects?)

vs.

[Differential expression]

What genes show differences between groups of subjects?

Differential expression

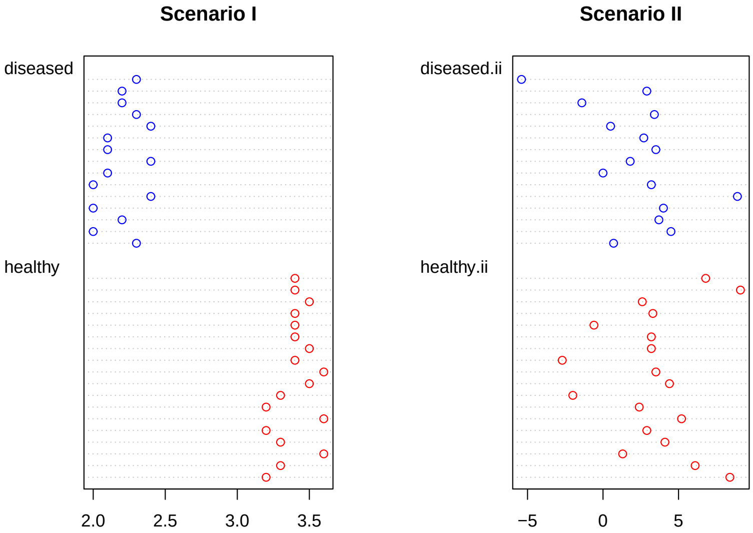
- What does it mean “to show differences”?

- Eg., differences in means: the mean of expression of gene MYC in group 1 is larger than the mean of expression of gene MYC in group 2 (group 1 over-expressed compared to group 2).
- “different things” ⇒ “differ in the mean” **of the populations**
- But it need not refer to the mean.

We want to compare the mean of expression of MYC between 15 diseased subjects and 18 non-diseased (“healthy”) subjects. How?

More formally: can the “true” (mean of) expression of the two groups be the same? (Do the two groups have the same mean of expression?)

We compute the mean of the two groups: 2.2 and 3.4. So what?

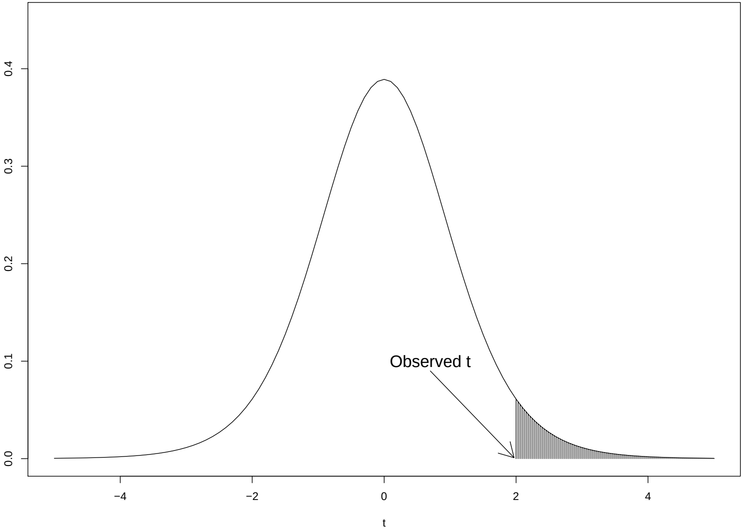

Scenario I
diseased
healthy
2.0
2.5
3.0
3.5
Scenario II
diseased.ii
healthy.ii
−5
0
5

How can we compare means?

- If there were no differences (null hypothesis, H_0), what would we expect?
- If there were no differences, what relationship would there be between labels (diseased, healthy) and values?
- t-test, permutation tests, non-parametric tests, etc. ▸ Appendix: Permutation tests.

What about probability and the strength of evidence?

1. Using differnt approaches (analysis, permutation) we can obtain the distribution of “t” under the null hypothesis. Null hypothesis: in this case that the two true means are equal.: Obtain the distribution of the “t” that one would find if there were, really, no differences.
2. Compute how likely it is to observe our “t” if the null hypothesis were true.
3. p-value: how likely our result would be if the null were true. p-value: a measure of evidence against the null hypothesis.


Observed t
0.0
0.1
0.2
0.3
0.4
−4
−2
0
2
4
t

p-value

- Differential expression: our hypothesis ($\mu_1^{MYC} \neq \mu_2^{MYC}$)
- p-value: how likely our results if the null hypothesis were true
- (So there is a null hypothesis: $H_0 : \mu_1^{MYC} = \mu_2^{MYC}$)

- p-value: measure of evidence against the null hypothesis.
- p-value: **it is NOT the probability of the null hypothesis (nor of the alternative hypothesis).**

- We compute one p-value for one null hypothesis (one per gene). E.g., MYC.

The p-value and the bag of interesting genes

- We can think about a statistical test as . . .
- a procedure to assign a gene to one of two groups
  - “Interesting ones” (differentially expressed)
  - Non “interesting”

Multiple testing

We know how to obtain a p-value to compare two groups. (And there are similar approaches for other comparisons.). We have, e.g., 10000 genes. So 10000 p-values . . .

Stats for bioinfo

Omics et al: the data

p-values

Multiple testing

FDR

Multiple testing, forensics, etc

Design and analysis

Appendix: FDR and permutation tests

Intro classif.

Algorithms

Error estimation

Survival

Added value

Clustering

Appendix: survival analysis

Help!

(Remember) p-values and the bag of interesting genes

If we are studying, e.g., differential expression of genes . . .

- We can think about a statistical test as . . .
- a procedure to assign a gene to one of two groups
  - “Interesting ones” (probably differentially expressed)
  - Non “interesting”
- We apply now the procedure to each of the genes.

Can we just compute a p-value for each gene and select the relevant genes as those with small (say, $p < 0.05$) p-value?

The fish (or the fishing expeditions)

- We go fishing.
- In this sea, there is one specific fish (fish A) with a probability of being caught of 0.05.
- In this sea, there are another 1000 fish like A (but only one is A, of course). These are “i.i.d” fish (independent of A, but with identical behavior to A).
- What is $Pr\{eat\ fish\ A\}$?
- What is $Pr\{eat\ fish\}$?
- (In this case it is simple to see the differences, because the wording makes obvious we are, or not, restricting ourselves to “A”. But what if we say “eat fish A” vs. “have dinner”?).

The fish (II)

- $Pr\{eat\ fish\ A\} = 0.05$.
- $Pr\{eat\ fish\} \simeq 1$.
- The two events (eat A, eat fish) are very different.
- $Eat\ fish = \bigcup(eat\ A, eat\ B, eat\ C, \ldots, eat\ A\ and\ B, \ldots)$.

p-values are like fish

- If we have 30000 genes, and there is no differential expression at all in any . . .
- and we declare as “interesting” those genes with p-value < 0.05 we will make lots of false positives ($\sim$ 1500).
- We need to control this.
- (Note the differences between testing a pre-specified hypothesis about a specific gene, and “anything goes” —any gene with a significant result will do for writing a paper).

The p-value case

(An example modified from Westfall and Young, 1993 “Resampling-based multiple testing”).

- Suppose we have 100 independent genes. Thus, 100 null hypotheses, one for each gene.
- Suppose also that there are no differences in gene expression between the two groups of patients (i.e., the null is true, and we are using the appropriate test so that the p-value is Uniform on [0,1]).
- Thus, the probability that a particular test (say, for gene 3) is declared significant at level 0.05 is exactly 0.05. Good.

p-value case (II)

- However, the probability of declaring at least one of the 100 hypotheses false (i.e., rejecting at least one, or finding at least one result significant) is:

$$Pr(\text{at least one null rejected}) = 1 - Pr(\text{all } p_i > 0.05) = \\ 1 - (1 - 0.05)^{100} = 1 - 0.95^{100} = 0.994$$

- So now, even if the 100 genes are not differentially expressed, there is a probability of 0.994 (yes, that is 99%!!!) of “finding” at least one which we declare as significantly different.
- The more genes, the more serious is the problem.
- In summary, without control for multiple testing, we would end up rejecting the null much more often than we should.

FDR

	# non rejected	# rejected
# same expression (H_0 *true*)	U	V
# different expression (H_0 *false*)	T	S

FDR False Discovery Rate: expected proportion of type I error among the rejected nulls: $(V + S)$. $FDR = E(Q)$ where $Q = V/(V + S)$ if $V + S > 0$ (and $Q = 0$ otherwise).

FWER $P(V \geq 1)$

False positives

- Why are false positives worse than false negatives?

  Even if the false positive rate were zero, we still don't have nearly enough resources to experimentally verify all the claims
  (Cited en X.-L. Meng, *The American Statistician*, 2009)

FDR: interpreting output

- p-value
- FDR
- FDR-adjusted p-values
- Properties of lists! See output.
- E.g.: `http://pomelo2.iib.uam.es/Examples/LeukemiaGolub/results.html`

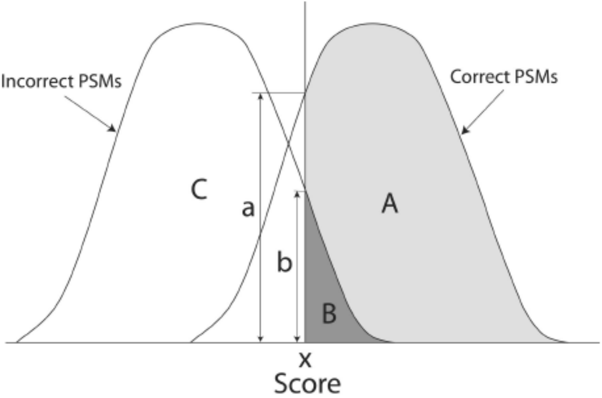
Further details in ▸ Appendix. FDR: the algorithm

pFDR, q-values, PEP

- Storey, Efron, et al.
- $FDR = E\left[\frac{V}{R}|R > 0\right] \mathrm{Pr}(R > 0)$
- Only interested in FDR if there are positive results: $pFDR = E\left[\frac{V}{R}|R > 0\right]$
- pFDR: “posterior Bayesian Type I error”.

- PEP, “posterior error probability” or “local FDR”: the probability that a given feature be incorrect. For instance, “the probability that gene XYZ is NOT a differentially expressed one”.
- A property of a feature. The probability of a given gene being a false positive **in the context of** a collection of genes (p-values).

PEP vs. FDR



FDR = B/(A + B)

PEP = b/(a + b)

(Kall et al., 2008. J. Proteome Research, 7: 40–44)

Forensics

- A crime.
- DNA test: $\frac{1}{100000}$ of match at random.
- Two scenarios:
  - The suspect (suspect because of something else) matches
  - You search in a large database of individuals and find a match
- Beware of the prosecutor's fallacy.

Sally Clark case

- $P(2\ SIDS)$ is rare.
- $P(2\ murders)$ might be even rarer.
- $P(Innocent|Data) \neq P(Data|Innocence)$
- Before someone is sent to jail you probably want: $\frac{P(Guilty|Data)}{P(Innocent|Data)}$ very large
- Beware of the prosecutor's fallacy.

Sample size

- I choose, randomly, 2 men and 3 women from this class and measure their height. Can I say anything about the differences in height between sexes in the Spanish population?

Sample size

- I choose, randomly, 2 men and 3 women from this class and measure their height. Can I say anything about the differences in height between sexes in the Spanish population?
- Significant results vs. repeatable results.
- Each poorly conducted study is a wasted opportunity.
- The argument of money . . . is it an argument?
- See Dobbin and Simon, 2005 and 2007, *Biostatistics*.

What test to use?

- Even in the simplest of cases (comparing two groups) there are many ways to analyze the data.
- Non-parametric vs. parametric statistics.
- Non-parametric and permutation tests.

Type of response variable

- Continuous-like: microarray
- Count: NGS

Type of response variable

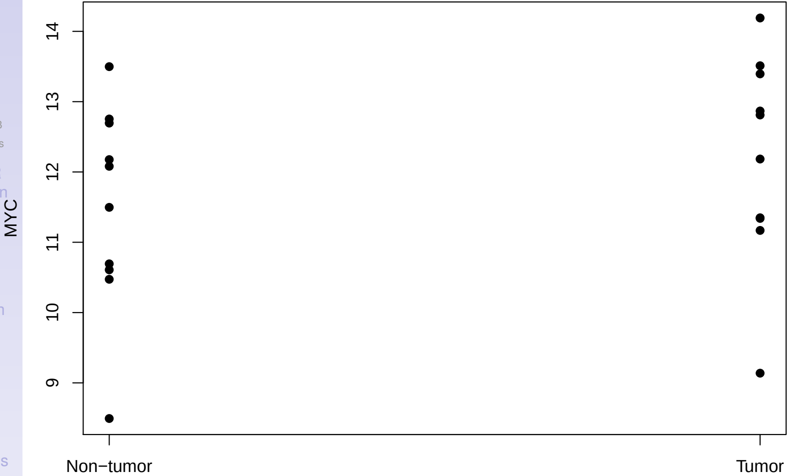
- Continuous-like: microarray
- Count: NGS
- Can we have a unified view of this mess?
- Linear models et al. (▸ Linear models)

Samples have many characteristics

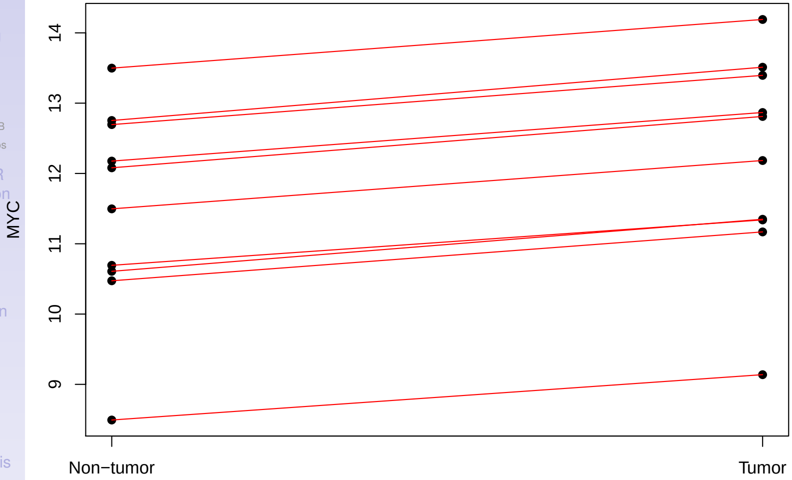
E.g., human subjects

- Age
- Sex
- Hospital, region, date of diagnostic, . . .
- Patients measured multiple times
- Family relationships, same doctors, . . .
- . . .
- Include other variables to increase power (decrease variance) and avoid biases

Paired vs. non-paired



Paired vs. non-paired



Which is the experimental unit?

- 20 mice
- 10 assigned to drug A, 10 assigned to drug B
- Each mouse, in one leg a corticoid ointment, on the other a placebo ointment
- ointment: nested within mouse
- Which is the experimental unit?

- Two types of experimental unit: mouse, leg within mouse.
- To compare drugs: use mice
- To compare ointment: use leg within mouse
- Interaction: we can study it
- split-plot designs, mixed-effects models

Replicates and pseudoreplicates

- 20 arrays, 10 of one kind, 10 of another
- Scenarios
  - 20 subjects total
  - 5 subjects in each group, each subject measured twice
  - 2 subjects in each group, each subject measure 5 times
  - 5 families in each group, some with 1 representative, others with 2, others with 3, . . .

Blocking

- Mice are “blocks”: ointment effect is within-mouse. We keep mouse effects constant. Each mouse is its own control.
- “Block what you can, randomize what you can’t”
- Randomization: a tool to deal with possible systematic sources of variation that we cannot control and avoids biases.

Real life is complicated . . .

A simple t-test or simple-whatever will rarely be the most appropriate approach

When you go to GEO or ArrayExpress or . . . you must keep the above in mind.

Observational vs. experimental studies

- Random assignment of treatments to subjects vs. observational studies
- Carefully use additional covariate: sex, age.
- Prostate cancer: what is the control?
- Controls: qualitative difference between observational and experimental studies. (Randomization principle).
- Inference with observational data a lot more tricky. Complex interpretation.

More covariates

- Even if there is no confounding, including covariates in analysis can increase statistical power.
- Why?
- (Note: we use models to explain this.)

Linear models: introduction

- Key in statistcs.

$$y_i = \sum_{j=1}^{j=p} x_{ij}\beta_j + e_i$$

- Often assume $e_i \sim Normal(0, \sigma^2)$
- Simple regression $y_i = \alpha + \beta x_i + e_i$ is a special case.
- Also multiple regression.
- And ANOVA (analysis of variance).
- Many other models derived from the general linear model. modelo lineal.

Linear models and their derivatives

(Following Faraway, 2008)

- Linear model:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + e$$

- GLM: generalize the *y*. GLMs can analyze binary data, categorical data, survival data, etc.
- Mixed models: generalize the *e*. Data with nested structures, longitudinal, multilevel, etc, that induce correlations in *e*. Mixed models, GEE, weighted least squares, etc.

Linear models and their derivatives

“Roadmap” in fig. 57, Kuhnert and Venables (p. 141).

- Linear models
- Generalized linear models
- Multivariate linear models (MANOVA)
- Mixed effects linear models
- Nonlinear models
- Nonlinear mixed-effects models
- Generalized linear mixed effects models (GLMM)
- Generalized additive models (GAM)

Linear models and the matched-pairs design

- $P53_{subject,condition} = Subject + Condition + e$
- If we remove Subject (do not use that info) . . .
- . . . we move “Subject” to the *e*.

Extending the t-test/linear model approach

- For each gene, use ONLY information for that gene
- A poor job estimating variances

Moderated statistics, limma, Empirical Bayes

- Can use information from all other genes when making inferences about each particular gene, specially in the estimation of variances.
  - Empirical Bayes approach of G. Smyth among the most widely used.
  - Moderated statistics lead to both increases in power and decreases in Type I errors.

Comparing two groups

Example question	Model	H_0	Test	Other
Is the expression of MYC different between Cancer (C) and Non-cancer (N) patients?	$MYC \sim Group$	$\mu_{\text{NC}} = \mu_{\text{C}}$	t-test	
Is the expression of any genes in this array different between Cancer (C) and Non-cancer (N) patients?	$MYC \sim Group$ $P53 \sim Group$ $\ldots$ $\ldots \sim Group$	$\mu_{\text{NC}}^{\text{MYC}} = \mu_{\text{C}}^{\text{MYC}}$ $\mu_{\text{NC}}^{\text{P53}} = \mu_{\text{C}}^{\text{P53}}$ $\ldots$ $\mu_{\text{NC}}^{\ldots} = \mu_{\text{C}}^{\ldots}$	Many t-tests	Empirical Bayes (EB). FDR
Is the expression of MYC different between Cancer (C) and Non-cancer (N) patients when we control for other factors (age, sex, ...)?	$MYC \sim Group + Age + Sex$	$\mu_{\text{NC}} = \mu_{\text{C}}$	t-test	Type or relationship of others (non-linear, etc). Interactions
Is the expression of any genes in this array different between Cancer (C) and Non-cancer (N) patients when we control for other factors?	$MYC \sim Group + Age + Sex$ $P53 \sim Group + Age + Sex$ $\ldots$ $\ldots \sim Group + Age + Sex$	$\mu_{\text{NC}}^{\text{MYC}} = \mu_{\text{C}}^{\text{MYC}}$ $\mu_{\text{NC}}^{\text{P53}} = \mu_{\text{C}}^{\text{P53}}$ $\ldots$ $\mu_{\text{NC}}^{\ldots} = \mu_{\text{C}}^{\ldots}$	Many t-tests	Empirical Bayes. FDR. Type or relationship of others (non-linear, etc). Interactions

Comparing three or more groups

Example question	Model	H_0	Test	Other
Is the expression of MYC different between patients with Colon (C), Prostate (P), Lung (L), and Neck (N) cancer?	$MYC \sim Group$	$\mu_\mathrm{C} = \mu_\mathrm{P} = \mu_\mathrm{L} = \mu_\mathrm{N}$	ANOVA (F test)	Might want to do some pairwise comparisons.
Is the expression of any genes in this array different between patients with Colon (C), Prostate (P), Lung (L), and Neck (N) cancer?	$MYC \sim Group$ $P53 \sim Group$ $\ldots$ $\ldots \sim Group$	$\mu_\mathrm{C}^\mathrm{MYC} = \mu_\mathrm{P}^\mathrm{MYC} = \mu_\mathrm{L}^\mathrm{MYC} = \mu_\mathrm{N}^\mathrm{MYC}$ $\mu_\mathrm{C}^\mathrm{P53} = \mu_\mathrm{P}^\mathrm{P53} = \mu_\mathrm{L}^\mathrm{P53} = \mu_\mathrm{N}^\mathrm{P53}$ $\ldots$ $\mu_\mathrm{C}^{\cdots} = \mu_\mathrm{P}^{\cdots} = \mu_\mathrm{L}^{\cdots} = \mu_\mathrm{N}^{\cdots}$	Many ANOVAs	EB. FDR. Might want to do some pairwise comparisons.
Is the expression of MYC different between patients with Colon (C), Prostate (P), Lung (L), and Neck (N) cancer when we control for other factors?	$MYC \sim Group + Age + Sex$	$\mu_\mathrm{C} = \mu_\mathrm{P} = \mu_\mathrm{L} = \mu_\mathrm{N}$	ANOVA	Might want to do some pairwise comparisons. Type or relationship of others (non-linear, etc). Interactions
Is the expression of any genes in this array different between patients with Colon (C), Prostate (P), Lung (L), and Neck (N) cancer when we control for other factors?	$MYC \sim Group + Age + Sex$ $P53 \sim Group + Age + Sex$ $\ldots$ $\ldots \sim Group + Age + Sex$	$\mu_\mathrm{C}^\mathrm{MYC} = \mu_\mathrm{P}^\mathrm{MYC} = \mu_\mathrm{L}^\mathrm{MYC} = \mu_\mathrm{N}^\mathrm{MYC}$ $\mu_\mathrm{C}^\mathrm{P53} = \mu_\mathrm{P}^\mathrm{P53} = \mu_\mathrm{L}^\mathrm{P53} = \mu_\mathrm{N}^\mathrm{P53}$ $\ldots$ $\mu_\mathrm{C}^{\cdots} = \mu_\mathrm{P}^{\cdots} = \mu_\mathrm{L}^{\cdots} = \mu_\mathrm{N}^{\cdots}$	Many ANOVAs	EB. FDR. Might want to do some pairwise comparisons. Type or relationship of others (non-linear, etc). Interactions

Relationship with a numerical variable: regression

Example question	Model	H_0	Test	Other
Does the expression of MYC change (increase or decrease) with (as we increase) cholesterol (numerical variable)?	$MYC \sim cholest$	$\beta = 0$	Regression t-test	(Non-linear?)
Does the expression of any genes in this array change with cholesterol?	$MYC \sim cholest$ $P53 \sim cholest$ $\ldots$ $\ldots \sim cholest$	$\beta^{MYC} = 0$ $\beta^{P53} = 0$ $\ldots$ $\beta^{\cdots} = 0$	Many regressions	Empirical Bayes. FDR
Does the expression of MYC change with cholesterol when we control for other factors?	$MYC \sim cholest + Age + Sex$	$\beta = 0$	Regression	Type or relationship of others (non-linear, etc). Interactions
Does the expression of any genes in this array change with cholesterol when we control for other factors?	$MYC \sim cholest + Age + Sex$ $P53 \sim cholest + Age + Sex$ $\ldots$ $\ldots \sim cholest + Age + Sex$	$\beta^{MYC} = 0$ $\beta^{P53} = 0$ $\ldots$ $\beta^{\cdots} = 0$	Many regressions	Empirical Bayes. FDR. Type or relationship of others (non-linear, etc). Interactions

The logic of a permutation test

- Define the statistic (e.g., differences between means).
- Obtain their distribution under the null hypothesis (H_0).
- Calculate how likely our observed statistic is under the null hypothesis.

- (Permutation tests are very general approaches that can be used for testing a variety of hypothesis —e.g., Dupuy and Simon paper— but their use in real life requires a lot of care.)

The logic of a permutation test

- Key idea: under H_0 labels and values are not related.
- That's it!
- How could we generate a data set that is compatible with H_0?
- And another? . . .

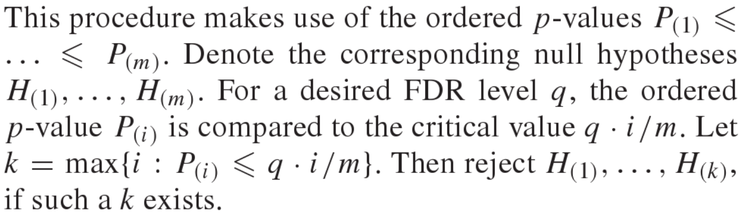
t-test to compare two groups

1. Compute the means
2. Subtract one from the other
3. Compute a quantity related to the variance of the differences of the means (this comes from the variance of each group).
4. Divide the difference in means by the standard deviation of the difference in means.
5. Now we have a standardized difference: the t statistic.

Differences between both procedures?

- With a t-test: if certain assumptions are true, there is a statistic of know distribution under H_0. From here, the p-value is immediate.
- permutation test: we define a statistic. We do not derive analytically its distribution. We obtain it numerically by counting events generated under H_0.
- Permutation tests are not “assumption free”!!!
- permutation tests might be testing a hypothesis we are not interested in (e.g., dispersion vs. mean).
- Some assumptions of parametric tests might be verifiable and/or reasonable. And parametric models give us extra stuff (model checking).
- Numerically: are results similar?

FDR: the algorithm



This procedure makes use of the ordered p-values $P_{(1)} \leqslant \ldots \leqslant P_{(m)}$. Denote the corresponding null hypotheses $H_{(1)}, \ldots, H_{(m)}$. For a desired FDR level q, the ordered p-value $P_{(i)}$ is compared to the critical value $q \cdot i/m$. Let $k = \max\{i : P_{(i)} \leqslant q \cdot i/m\}$. Then reject $H_{(1)}, \ldots, H_{(k)}$, if such a k exists.

(Reiner et al., 2003, Bioinformatics)

(Note: a “step-up procedure”).

This procedure controls FDR. (Does not say “estimates”).

Examples

What will I reject at 0.1?

- 0.1, 0.1, 0.1, 0.1
- all

- 0.1, 0.01, 0.01, 0.01
- all

- 0.2, 0.1, 0.1, 0.1
- none
- Threshold: 0.1
- $p \leq threshold * i/m$?
- $0.1 * 3/4 = 0.075$

- 0.2, 0.075, 0.075, 0.075
- last three

Adjusted p-values

> *The results of a multiple testing procedure can be reported as multiplicity adjusted p-values. As with the regular p-value, each adjusted p-value is compared to the desired significance level, and if smaller, the hypothesis is rejected. Therefore, the way adjusted p-values are used and interpreted remains conveniently familiar, regardless of the adjustment procedure complexity.*

(Reiner et al., 2003, Bioinformatics)

Adjusted p-values: FDR

> *For an FDR controlling procedure, the adjusted p-value of an individual hypothesis is the lowest level of FDR for which the hypothesis is first included in the set of rejected hypotheses. Thus the adjusted p-value of $P_{(j)}$ using the BH procedure, is $P_{(j)}^{BH} = \min_{j \leq i}\{P_{(i)}\frac{m}{i}\}$.*

(Reiner et al., 2003, Bioinformatics)

Examples of FDR-adjusted p-values

- 0.2, 0.08, 0.08, 0.08
- 0.2, 0.1067, 0.1067, 0.1067
- 0.08 * 4/3 = 0.1067; 0.08 * 4/2 = 0.16; . . .

```
p.adjust(c(0.2, 0.08, 0.08, 0.08),
         method = "BH")
```

- 0.2, 0.08, 0.07, 0.07
- 0.2, 0.1067, 0.1067, 0.1067
- 0.08 * 4/3 = 0.1067; 0.07 * 4/2 = 0.14;

```
p.adjust(c(0.2, 0.08, 0.07, 0.07),
          method = "BH")
```

- 0.2, 0.08, 0.05, 0.015
- 0.2, 0.1067, 0.1, 0.06
- 0.05 * 4/2 = 0.1; 0.015 * 4/1 = 0.06

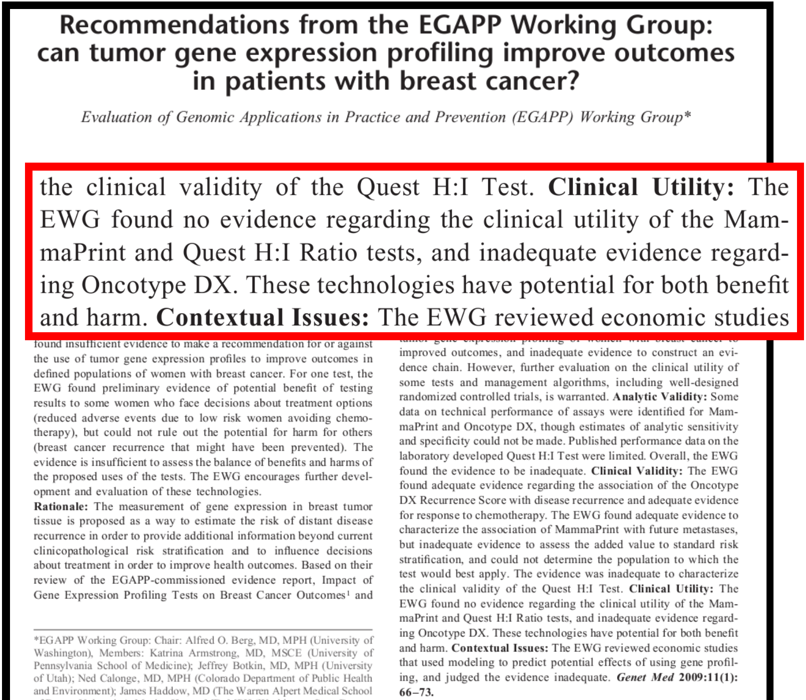
```
p.adjust(c(0.2, 0.08, 0.05, 0.015),
           method = "BH")
```

Differentiate between groups of patients

Classification (or prediction if a continuous variable)

A classical problem in statistics and machine learning.

What do we want? A good classifier. Something that, given a new sample, will assign it to its appropriate group.



Recommendations from the EGAPP Working Group: can tumor gene expression profiling improve outcomes in patients with breast cancer?

*Evaluation of Genomic Applications in Practice and Prevention (EGAPP) Working Group**

the clinical validity of the Quest H:I Test. **Clinical Utility:** The EWG found no evidence regarding the clinical utility of the MammaPrint and Quest H:I Ratio tests, and inadequate evidence regarding Oncotype DX. These technologies have potential for both benefit and harm. **Contextual Issues:** The EWG reviewed economic studies

found insufficient evidence to make a recommendation for or against the use of tumor gene expression profiles to improve outcomes in defined populations of women with breast cancer. For one test, the EWG found preliminary evidence of potential benefit of testing results to some women who face decisions about treatment options (reduced adverse events due to low risk women avoiding chemotherapy), but could not rule out the potential for harm for others (breast cancer recurrence that might have been prevented). The evidence is insufficient to assess the balance of benefits and harms of the proposed uses of the tests. The EWG encourages further development and evaluation of these technologies.

Rationale: The measurement of gene expression in breast tumor tissue is proposed as a way to estimate the risk of distant disease recurrence in order to provide additional information beyond current clinicopathological risk stratification and to influence decisions about treatment in order to improve health outcomes. Based on their review of the EGAPP-commissioned evidence report, Impact of Gene Expression Profiling Tests on Breast Cancer Outcomes[1] and

tumor gene expression profiling of women with breast cancer to improved outcomes, and inadequate evidence to construct an evidence chain. However, further evaluation on the clinical utility of some tests and management algorithms, including well-designed randomized controlled trials, is warranted. **Analytic Validity:** Some data on technical performance of assays were identified for MammaPrint and Oncotype DX, though estimates of analytic sensitivity and specificity could not be made. Published performance data on the laboratory developed Quest H:I Test were limited. Overall, the EWG found the evidence to be inadequate. **Clinical Validity:** The EWG found adequate evidence regarding the association of the Oncotype DX Recurrence Score with disease recurrence and adequate evidence for response to chemotherapy. The EWG found adequate evidence to characterize the association of MammaPrint with future metastases, but inadequate evidence to assess the added value to standard risk stratification, and could not determine the population to which the test would best apply. The evidence was inadequate to characterize the clinical validity of the Quest H:I Test. **Clinical Utility:** The EWG found no evidence regarding the clinical utility of the MammaPrint and Quest H:I Ratio tests, and inadequate evidence regarding Oncotype DX. These technologies have potential for both benefit and harm. **Contextual Issues:** The EWG reviewed economic studies that used modeling to predict potential effects of using gene profiling, and judged the evidence inadequate. ***Genet Med* 2009:11(1): 66–73.**

*EGAPP Working Group: Chair: Alfred O. Berg, MD, MPH (University of Washington), Members: Katrina Armstrong, MD, MSCE (University of Pennsylvania School of Medicine); Jeffrey Botkin, MD, MPH (University of Utah); Ned Calonge, MD, MPH (Colorado Department of Public Health and Environment); James Haddow, MD (The Warren Alpert Medical School

. . . clinical utility

Clinical validity predict risk of recurrence

Clinical utility predict benefit of a treatment over another: added value when making decissions.

- . . . when we already have conventional classifiers/predictors
- **Does the new method/algorithm, based on genomic data, improve our ability to predict a result, compared to what we could predict without those genomic data?**

The dangers of “capitalizing on chance

Statistical context: many genes, few subjects. $p \gg n$.

Differentially expressed genes Risk of too many false positives ⇒ adjustments in the screening of p-values.

Classification/prediction Very easy to obtain algorithms that classify, perfectly, our data, but not new data ⇒ validate algorithms and classifiers

Hypotheses/questions Tempting to make them vague, or ask none and wait until “the data say something” ⇒ define objectives and how we will measure what we are interested in.

Review of methods and good practices.

(***): highly recommended

- Tarca et al., 2013. Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics*, 29: 2892–2899. (***)
- Shi et al. 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28: 827–838. (***)
- Dupuy A, Simon R. 2007. Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *J Natl Cancer Inst.*, 99: 147–157. (***)

Books

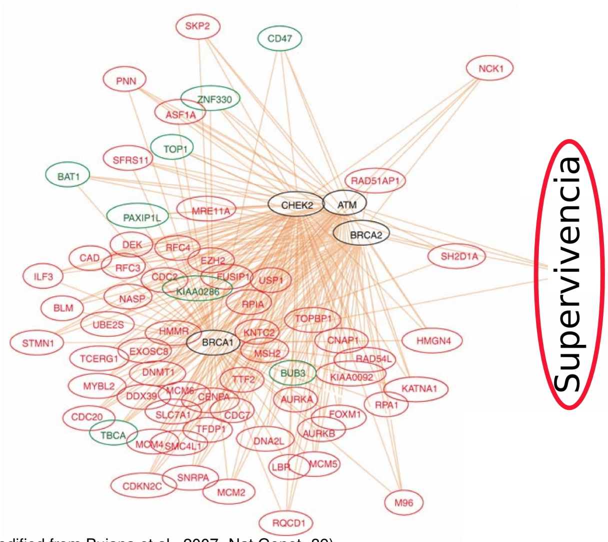
- Kuhn, and Johnson. 2013. Applied Predictive Modeling. *Springer*. (***)
- James et al. 2013. An introduction to statistical learning. *Springer*. (***) (A PDF can be downloaded freely and legally from their web page)

- Lee et al., 2005. *Computational Statistics and Data Analysis*, 48: 869–885.
- Zucknick et al. 2008. *Stat Appl Genet Mol Biol.* 7 (se encuentra en la web, en PubMed central).
- Many other (see outdated list in Diaz-Uriarte, 2005. http://ligarto.org/rdiaz/Papers/chapter-azuaje-dopazo.pdf).
- Slawski et al., *BMC Bioinformatics*, 2008, 9: 439. Description of many methods in a single paper. But it is not a comparison.
- Dudoit, S., J. Fridlyand, and T. P. Speed. 2002. Comparison of discrimination methods for the classification of tumors suing gene expression data. *J Am Stat Assoc* 97(457), 77-87. (You can find the reprint on the web.).

Classification/prediction: key ideas

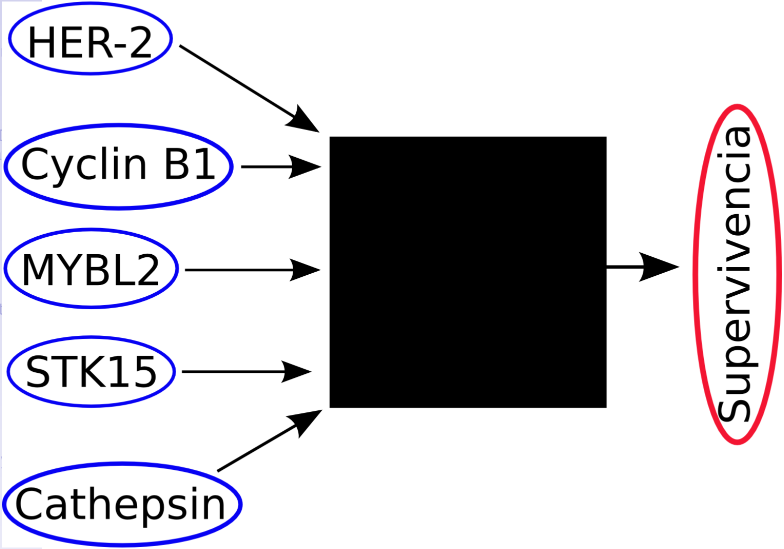
- All we care about is a good classifier.
- We do not care about p-values.
- We will have to choose some genes.
- We will have to, ESPECIALLY, estimate the error of the classifier.

Tell me how it works (dime cómo funciona) . . .



(Modified from Pujana et al., 2007. Nat Genet. 39)

. . . vs. the black box



Prediction: black box

- Rules of the game: that it predicts (classifies) well.
- We are not assessing the “truth” of the model. Only its predictive success.
- Almost all methods eliminate genes with redundant info for classification: this limits interpretability anyway.

Prediction vs. interpretability

- Good classifiers need not be intuitively easy to understand.
- $p \gg n$: many classifiers with similar predictive capacity but different genes.
- Black boxes ameliorate these problems (you do not worry too much about them).
  - Inversions in the signs of coefficients
  - Genes shared between models

- Do not jump between objectives: classify vs. interpret.

Steps in the construction of a classifier with genomic data

- Selection of a classification algorithm.
- Gene selection.
- Classifier construction/training.
- Estimate the error of the classifier.

Some algorithms/models

- Just to say some specific.
- We will mention a few that work well.
- There are lots we say nothing about.
- “Follow the pros”: read reviews, follow recommendations, and understand the methods you use.

Reviews of methods.

- Kuhn, and Johnson. 2013. Applied Predictive Modeling. *Springer*. (***)
- James et al. 2013. An introduction to statistical learning. *Springer*. (***) (A PDF can be downloaded freely and legally from their web page)
- Malley et al., 2011. Statistical learning for biomedical data. *Cambridge University Press*.
- Shi et al. 2010. The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models. *Nature Biotechnology*, 28: 827–838.

A great presentation

http://www-onderzoek.lumc.nl/HumaneGenetica/mgc/2005/presentations/Classification_Wessels.pdf

Nearest mean

- As it says: the closest mean
- (Next two slides from `http://www-onderzoek.lumc.nl/HumaneGenetica/mgc/2005/presentations/Classification_Wessels.pdf`)

Omics et al: the data

p-values

Multiple testing

Design and analysis

Appendix: FDR and permutation tests

Intro classif.

Algorithms

Algorithms

Gene selection

Error estimation

Survival

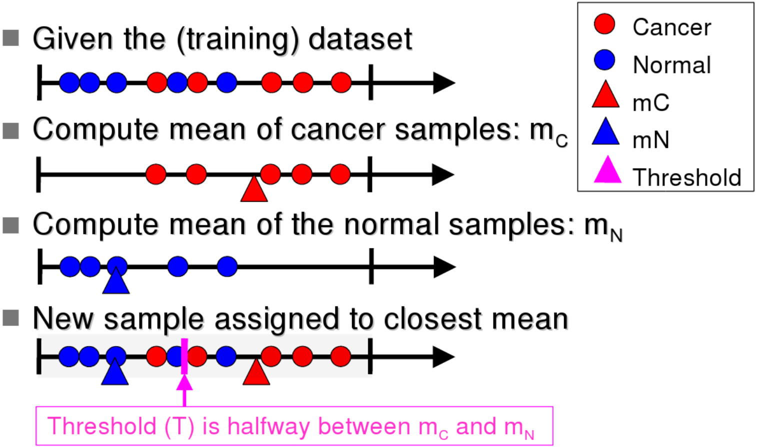
Added value

Clustering

Appendix: survival analysis

Help!

Nearest mean classifier in 1D

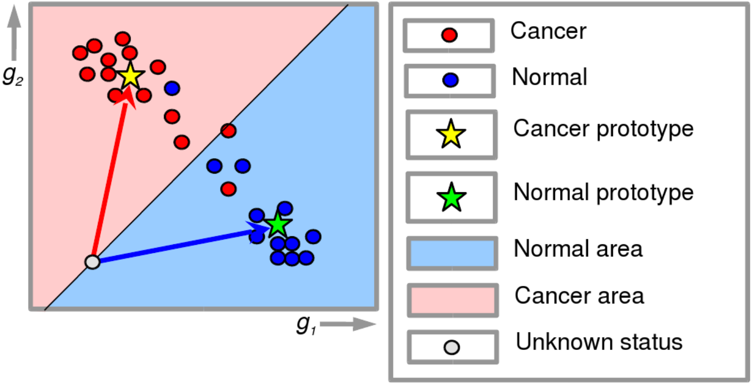


- Given the (training) dataset
- Compute mean of cancer samples: m_C
- Compute mean of the normal samples: m_N
- New sample assigned to closest mean

Threshold (T) is halfway between m_C and m_N

(Taken from Lodewyk Wessels, Classification_Wessels.pdf)

Nearest mean classifier III

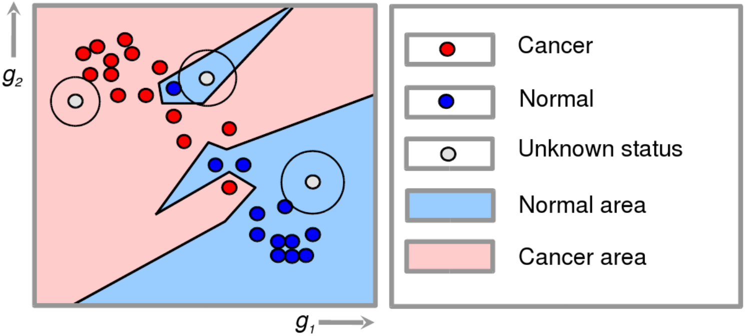


(Taken from Lodewyk Wessels, Classification_Wessels.pdf)

KNN

- K-nearest neighbor.
- Simple non-parametric rule:
- Predicts the sample of a test case as the majority vote among the k nearest neighbors of the test case.
- To decide on “nearest” we often use the Euclidean distance, but other measures of proximity are possible.
- The number of neighbors used (k) is either fixed or chosen by cross-validation.

1-Nearest neighbor classifier



(Taken from Lodewyk Wessels, Classification_Wessels.pdf)

DLDA

- Diagonal Linear Discriminant Analysis.
- A form of discriminant analysis (optimal when class densities have the same diagonal variance-covariance matrix).
- Simple linear rule: a sample is assigned to the class k which minimizes $\Sigma_{j=1}^{p}(x_j - \bar{x}_{kj})^2/\hat{\sigma}_j^2$, where p is the number of variables, x_j is the value on variable (gene) j of the test sample, $\bar{x}_{kj}$ is the sample mean of class k and variable (gene) j, and $\hat{\sigma}_j^2$ is the (pooled) estimate of the variance of gene j.
- Unrealistic assumption, but works very well (and often better than other forms of discriminant analysis that require estimation of many more parameters).
- Also called “NaÃ¯ve Bayes.”

Random Forest

- An ensemble of classification trees.
- Each tree is grown using a bootstrap sample of the data set, and at each node only a random subset of the original variables is examined.
- Interactions are implicitly considered.
- Provides ranking of variable importance.

Logistic regression

- We model the (logit of the) probability of belonging to a class as a linear combination of features. Extension of linear models to binary data.
- As well as DLDA, this is a "classic" of statistics.

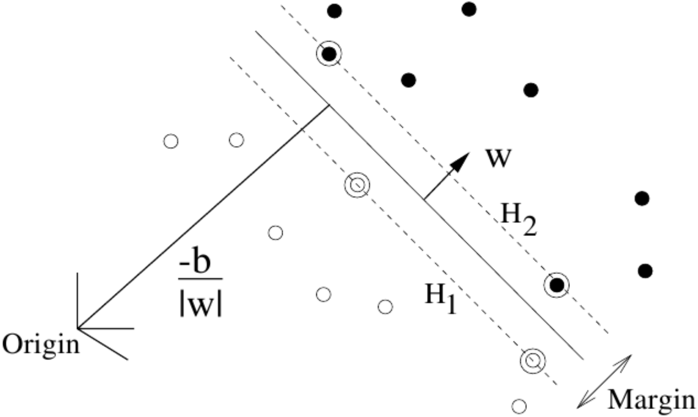
Regularization, ridge, the lasso, . . .

- As it says. What and why?

SVM

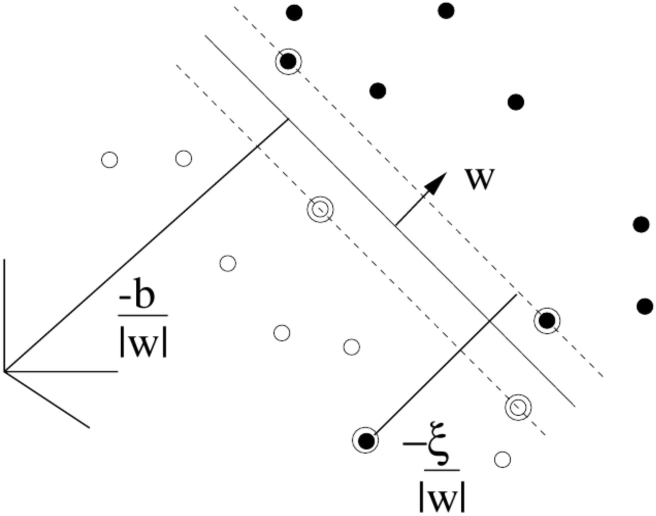
- Support vector machines.
- Obtain the best separating hyperplane between classes; hyperplane is located so that it has maximal margin (i.e., so that there is maximal distance between the hyperplane and the nearest point of any of the classes).
- When the data not separable, there is no separating hyperplane; in this case, we still try to maximize the margin but allow some classification errors subject to the constraint that the total error (distance from the hyperplane in the “wrong side”) is less than a constant.

SVM: separable case



(Taken from Burgues, 1998, *Data Mining and Knowledge Discovery* 2, 121-167)

SVM: non-separable case



(Taken from Burgues, 1998, *Data Mining and Knowledge Discovery* 2, 121-167)

Wisdom of crowds

- Aggregating predictions from different methods; mixtures of experts.
- We saw a similar idea with random forests.
- See Tarca et al. paper.

And boosting?

- . . .

A beautiful and insightful exposition of major themes

- Efron and Hastie, 2016, “Computer age statistical inference”

Gene selection

- Filter approaches: select before training the classifier.
  - Univariate
  - Multivariate
- Wrapper approaches. Within the classifier. A few infamous examples: stepwise regression et al. (There are non-infamous examples too).

Estimating the classifier's error (or "validating the classifier")

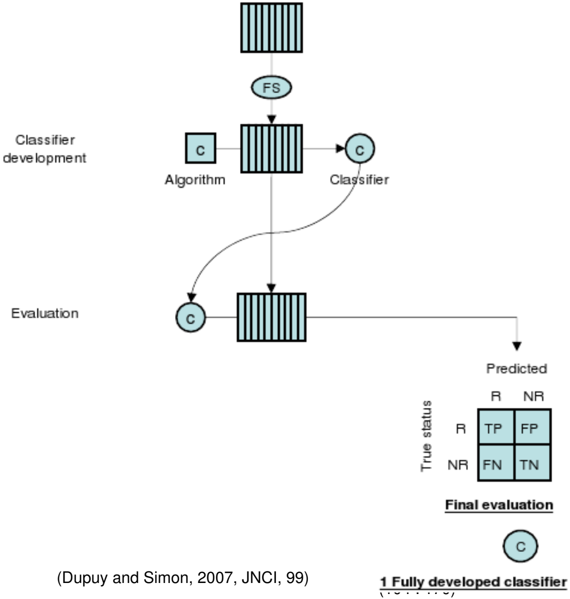
A sample with 50 healthy subjects and 50 diseased ones. We build a classifier with those 100 samples, and on those 100 we make a mistake of 10%.

Estimating the classifier's error (or "validating the classifier")

A sample with 50 healthy subjects and 50 diseased ones. We build a classifier with those 100 samples, and on those 100 we make a mistake of 10%.

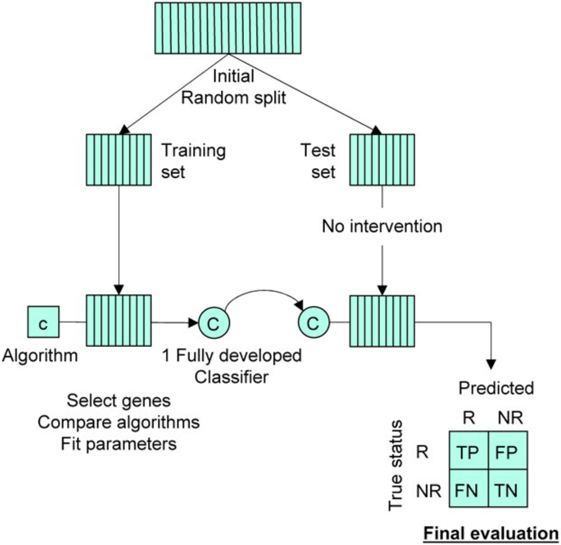
Can we use that 10% as a reasonable estimate of the error we would make with new samples?

Resubstitution



(Dupuy and Simon, 2007, JNCI, 99)

“Split-sample”, “holdout validation”, “data splitting”



Initial Random split

Training set

Test set

No intervention

c

Algorithm

C

C

1 Fully developed Classifier

Select genes
Compare algorithms
Fit parameters

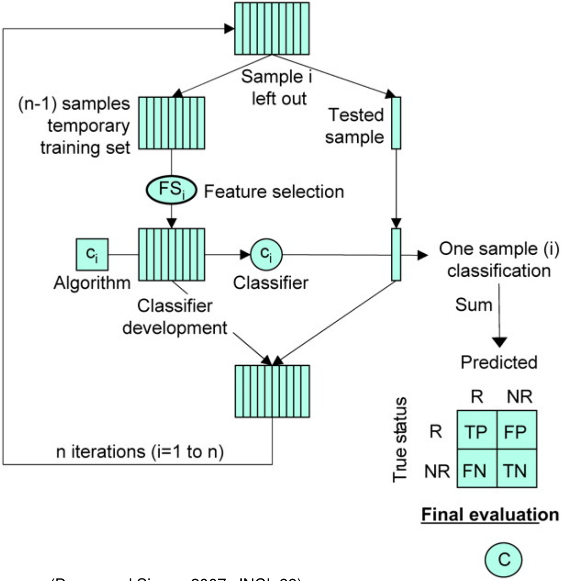
Predicted

True status		R	NR
	R	TP	FP
	NR	FN	TN

Final evaluation

(Dupuy and Simon, 2007, JNCI, 99)

“Cross-validation”

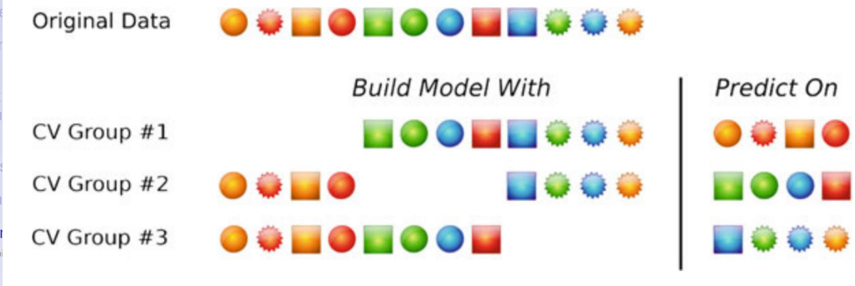


	R	NR
R	TP	FP
NR	FN	TN

(Dupuy and Simon, 2007, JNCI, 99)

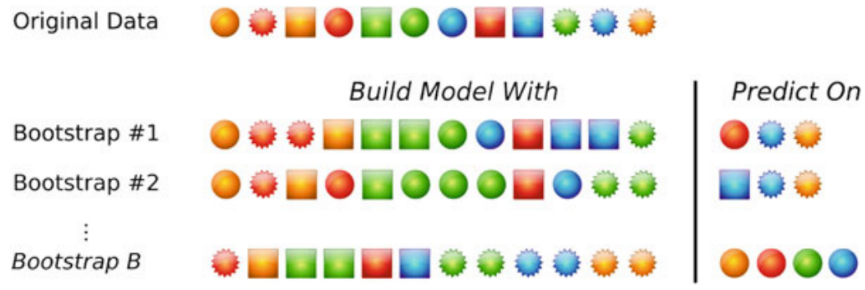
- Suppose 100 subects, 50 healthy, 50 diseased.
- Select at random 10 (“testing set”).
- Usae the other 90 to build the classifier (“training set”).
- Evaluate the classifier with the first 10.
- Repeat the process another 9 times (until all subjects have been used exactly once in the “testing set”).
- We have 10 estimates of error, we compute the mean, and we now have an estimate of the error we would make with a new sample.

Cross-validation (3-fold, here)



(Kuhn and Johnson, 2013. *Applied predictive modeling*, Springer

Bootstrap



(Kuhn and Johnson, 2013. *Applied predictive modeling*, Springer

Independent validation

Independent validation, with other samples, by other groups: necessary

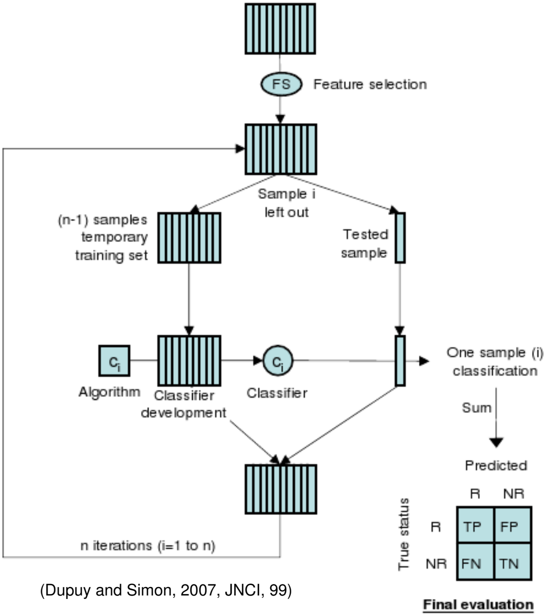
Beware of “selection bias”

What if we have done gene selection?

- Select the 100 genes with smallest p-value.
- Build the classifier.

The validation process has to include the gene selection procedure. We must do the gene selection in each training set.

NEVER DO THIS!



	R	NR
R	TP	FP
NR	FN	TN

(Dupuy and Simon, 2007, JNCI, 99)

CV and others

- There are related techniques, such as bootstrap, etc.
- To leave apart a single testing set is a bad idea.
- Cross-validation: can have high variance.
- Best approaches (?):
  - A variant of CV (repeated splits, 50 times 10)
  - Bootstrap (632+)

Classification, number of genes, etc

And how do we choose the number of genes?

Look, for example, at `http://tnasas.iib.uam.es`

We might not want to use many genes

- Number of features (genes) vs. number of samples
- “Metafeatures” (metagenes):
  - Biological relevance
  - Statistical noise reduction (averages)
- Adding not-very-relevant features generally decreases test set performance.

Validating what?

…what is it we are validating? how do we measure predictive ability?

Specificity and sensitivity

	Predicted	
True	Diseased	Healthy
Diseased	True Positive (TP)	False Negative (FN)
Healthy	False Positive (FP)	True Negative (TN)

- Sensitivity $= \frac{TP}{TP+FN}$
- Specificity $= \frac{TN}{TN+FP}$

Error rates or predictive values?

- (From van Belle, 2002, p. 96).
- Prevalence of colorrectal = 0.003
- Hemoccult test: sensitivity: 50%; specificity: 97%.
- I am positive!!!

Error rates or predictive values?

- (From van Belle, 2002, p. 96).
- Prevalence of colorrectal = 0.003
- Hemoccult test: sensitivity: 50%; specificity: 97%.
- I am positive!!!
- The probability of having colorrectal cancer is only 5%.
- Eh?

- I want to know: $P\left[Diseased|positive\right]$
- Prevalence = $P(D)$
- $P(D|p) = \frac{P(D \cap p)}{P(p)}$
- $P(D \cap p) = P(p|D)P(D)$ (Bayes rule)
- $P(p) = P(p|D)P(D) + P(p|D^c)P(D^c)$
- $P(p|D) = TP/(TP + FN) = Sensitivity$
- $P(p|D^c) = 1 - Specificity = FP/(FP + TN)$
- $P\left[Diseased|positive\right] =$
  $= \frac{P(D) * Sensitivity}{P(D) * Sensitivity + P(D^c)(1 - Specificity)}$
- $P(D^c) = 1 - P(D)$

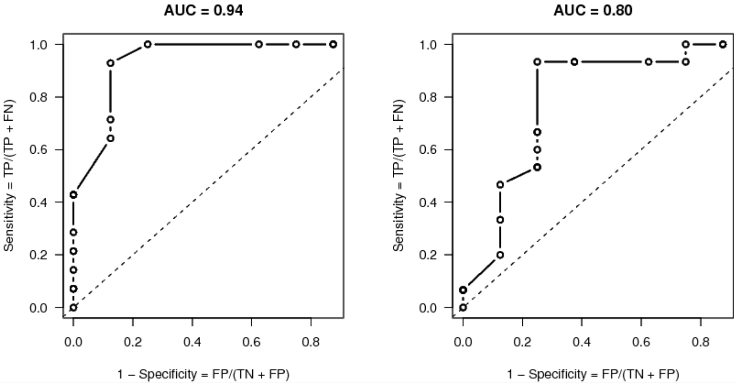
- $\frac{0.003 * 0.5}{0.003 * 0.5 + 0.997 * 0.03} = 0.048$

- $P\left[Diseased|positive\right] =$ Positive predictive value

- $P\left[Healthy|Negative\ test\right] =$ Negative predictive value $= = \frac{(1-P(D))Specificity}{(1-P(D))Specificity+P(D)(1-Sensitivity)}$

- Beware where prevalence estimate is coming from!!!!!

ROC curves


AUC = 0.94
AUC = 0.80
Sensitivity = TP/(TP + FN)
1 – Specificity = FP/(TN + FP)
0.0
0.2
0.4
0.6
0.8
1.0

Predictive ability?

- p-values, hazard-ratios, regression slopes, etc, are measures of association, not of predictive ability.
- Measuring predictive ability: how similar are predicted and observed?

Proportion correctly classified

- Probably NOT what you want.
- Easily “game-able”.(From posts by Frank Harell)
  - You can manipulate the proportion classified correctly in a number of silly ways. The easiest way to see this is if the prevalence of Y=1 is 0.98 you will be 0.98 accurate by ignoring all the data and predicting everyone to have Y=1.
- And we have not even considered asymmetric costs of mistakes.

Measuring predictive ability

- **Brier score** related to $\Sigma_i (Y_i - q_i)^2$, where Y_i is the real status (e.g., class A vs. class B —if A = 1, if B = 0) and q_i is the predicted probability of being of class A.
- **Concordance index (C-index)** Probability that, for all pairs of subjects where one is of one kind and the other of another kind, the patient with larger predicted probability of being of class A is really A. Related to ROC curves (later).
- **Area under ROC curve** As it says: area under ROC curves

Beware: Brier score, C-index, ROC: using “out of bag” predictions!!

Lots of data are survival data

- Time until I have to change the light bulb of my living room.
- Time until death.
- Time until . . .
- (no, not everything qualifies).

Introduction to survival models

- Time until “failure” (death, relapse, change of state, etc)
- Often censored:
  - We observe $\min(T, c)$ where T is life duration, time to death, and c the “censoring time”.
- Distributions such as exponential, weibull, etc
- We should NOT discretize nor use linear regression. **Use methods that are appropriate for the type of response**

Buzz words to remember

- **Cox model**: like a linear model but we model hazard rates ($h(t)$, “instantaneous rate of death at t given that you are alive at t.”)
- **Parametric survival models**: model the distribution of time to death.
- **Log-rank test**: a way of comparing survival curves. `http://signs2.iib.uam.es/Examples/CommentedExample/results.html`

Cross-validation, estimating predictive ability, etc

- All we have seen before applies.
- Estimating predictive ability is more complicated.
- Different criteria and different weights at different times (early vs. late events, for example).

Tools

- Not many.
- Beware of possible issues in the evaluation of predictive performance.
- `http://signs2.iib.uam.es`

Clinical covariates

- We often have clinical covariates
- How are we to include that info?
- Frequently predictors based on other indices
- Does gene expression improve prediction?
- Key question: Does using gene expression add anything? Is it worth it?
- **Does the new method/algorithm, based on genomic data, improve our ability to predict a result, compared to what we could predict without those genomic data?**
- Worth it . . . for what? Clinical utility vs. basic knowledge.

Why would predictions not improve with expression data?

- Expression data can be just noise.
- Expression data are redundant given the clinical covars. (which are often cheaper and faster to measure).
- (BEWARE: no implication about causality. This is irrelevant in this predictive scenario.)

Reasons for caution

- Truntzer et al. 2008.*BMC Bioinformatics*, 9: 434. Survival data.
- “ability of the model to predict outcome with new datasets is overestimated” with expression data.
- No optimism with clinical covars.
- They are two very different kinds of variables.

Simple solutions

- “Put everything in the same bag, and apply the usual methods”
- But “Clinical covariates come first”
- “Same bag” approach can affect negatively to clinical covars if they are correlated with gene expression.
- Coefficients of clinical covars must be estimated without penalization. And a need if we want to compare with models that only have clinical covars (see Binder and Schumacher).
- “Litmus test”: if genes do not provide anything, final model should be as good as if it only had clinical covars (Boulesteix et al., 2008).

Simple solutions (II)

- If there are discrete groups (sex, tumor marker) do separate analysis
- But we often have small sample sizes.
- Does not answer the original question directly: do gene expression data improve anything?

Simple solutions (III)

- Two classifiers: only with clinical covs. and only with gene expression data.
- We can compare models (though not obvious: they are not nested).
- Does not answer the basic question.

Not so simple solutions

- Do not penalize or remove clinical covariates.
- Adjust for those.
- Then add “omics” data.
- Assess if omics data adds something to previous model with only clinical covariates.

Conclusions (?)

- Many reasonable methods with similar solutions. Includes methods that are rather straightforward (DLDA, KNN).
- Instability and multiplicity of solutions: are they a problem?
- Which is the best number of genes is difficult to tell.
- Why are we doing this? Biological interpretation/understanding or for diagnostic test development?

Are there groups?

- Can we find groups of genes that behave in a similar way, but different from other genes?
- Likewise for subjects?

“Class discovery”, clustering.

Only makes sense if . . .

we do not know, before hand, that there are different groups of genes/subjects.

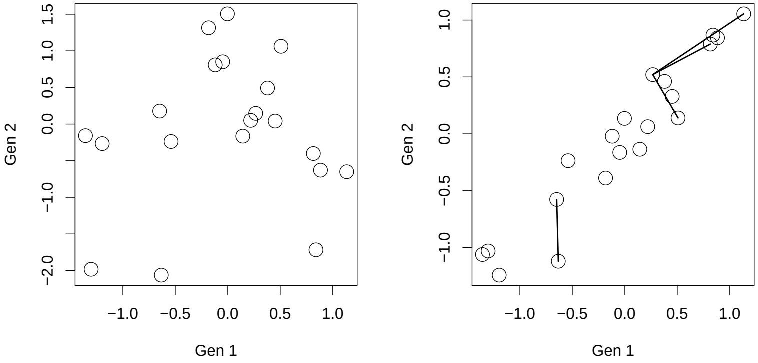
Two needed pieces

What does it mean to “behave similarly” and measuring similarity.

Describing how we will group based on those similarities.

First piece: similiarity (or “dis-similarity”)

- Distances (e.g., Euclidean distance).
- Correlations.


Gen 2
1.5
1.0
0.5
0.0
−1.0
−2.0
−1.0
−0.5
0.0
0.5
1.0
Gen 1
Gen 2
1.0
0.5
0.0
−0.5
−1.0
−1.0
−0.5
0.0
0.5
1.0
Gen 1

We end with a matrix of similarities between all pairs of subjects or genes.

Now what?

	s1	s2	s3	s4
s1	-	2	7	3
s2	-	-	8	4
s3	-	-	-	9
s4	-	-	-	-

???

Second piece: clustering algorithms

- Hierarchical:
  - Divisive
  - Agglomerative (UPGMA again!!!)
- No hierarchical (need to specify number of clusters).

Problems ...

- What measure of similarity should we use?
- What is the appropriate clustering algorithm?
- Should we use all genes when we cluster subjects?

Precautions

- Clustering is class discovery: it is an exploratory tool, not a confirmatory one.
- Clustering ALWAYS returns clusters, whether or not there is any real structure.
- If a cluster is “relevant” and “stable” is a different question.
- Clustering is not the right tool if we know about groups before hand.

t-test and cluster

What do you think about the idea of doing clustering and then a t-test?

An interesting idea: searching for transcription factors

- Cluster genes
- Search in up-stream regions for the most frequent l-mers
- (Details and references in Cristianini and Hahn 2006 and Harmer et al. 2000.)

And there is biclustering

- Cluster according to both dimensions, at the same time.

Appendix: Survival analysis with genomic data

- Many methods suggested
- Few comparisons that settle the issue
- How do we compare performance? How do we assess model quality?

- We observe $\min(T, c)$ where T is life duration, time to death, and c the “censoring time”. $F(t)$: cumulative distr. function of X.
- $F(t) = P(T \leq t) = \int_0^t f(x)dx$
- Survival function, $S(t)$: probability of being alive at time t.
- $S(t) = 1 - F(t) = \int_t^\infty f(x)dx$.

- “hazard function”, $h(t)$: instantaneous rate of death
- $h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$
- $h(t)\Delta t$ = probability of dying in the interval $[t, t + \Delta t]$, given that the subject is alive at time t
- $f(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$
- $h(t) = \frac{f(t)}{S(t)}$

- Cumulative hazard: $H(t) = \int_0^t h(u)du = -\log(S(t))$
- $S(t) = 1 - F(t) = P(T \geq t) = \exp(-H(t)) = exp(-\int_0^t h(u)du)$
- Median survival time:
  - time beyond which 50% of the subjects of the cohort are expected to be alive.
  - The first t where $\hat{S}(t) \leq 0.5$
- Mean survival time: $\int_0^\infty tf(t)dt$. But censored data create problems. Several approaches, e.g., “Efron’s tail correction”.

Kaplan-Meier estimator

- n_i: number of cases at risk just before time t_i (i.e., those that are part of the study and are still alive and not censored at t_i).
- d_i: the ones who die in the interval i.
- Kaplan-Meier estimator: $\hat{S}(t) = \Pi \frac{n_i - d_i}{n_i}$

```
T = 9, 13, 13+, 18, 23, 28+, 31, 34, 45+
```

```
 time n.risk n.event survival
    9      9       1    0.889
   13      8       1    0.778
   18      6       1    0.648
   23      5       1    0.519
   31      3       1    0.346
   34      2       1    0.173
```

- $S(0) = 1$
- $S(9) = \hat{S}(0) * (9 - 1)/9 = 0.889$
- $S(13) = \hat{S}(9) * (8 - 1)/8 = 0.778$ (Note that there are 8 at risk and the one that dies has survival 13).
- $S(13+) = \hat{S}(13+) * (7 - 0)/7 = 0.778$ (Note that there is no death event here).

Log-rank test

- One of the several ways to compare two or more survival curves.
- Related to categorical data analysis by strata: like a Mantel-Haenszel test where each stratum is each period.
- H_0: survivals are equal.
- Compute a “pooled sample estimator” of number of events and number at risk (d_i y n_i from before).
- Compute differences between observed and predicted (ej., $(d_{i1}/n_{i1}) - (d_i/n_i)$).
- Compute variance of those expected values.
- Sum over all periods, weighted by i (log-rank: weight = 1).
- Compare with appropriate distribution (Z, Chi if MH^2).

Cox Model

- $h(t) = h_0(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q)$
- $h_0(t)$: baseline hazard function. Common, independent of x.
- β_i: the effect of the given covariate (e.g.., gene x_i) on $h(t)$.
- Hazard ratio is constant: $\frac{h(t|\mathbf{x}_1)}{h(t|\mathbf{x}_2)} = \frac{exp(\beta^T\mathbf{x}_1)}{exp(\beta^T\mathbf{x}_2)} = \frac{exp(\Sigma\beta_i x_{1i})}{exp(\Sigma\beta_i x_{2i})} = exp(\Sigma\beta_i(x_{1i} - x_{2i}))$

- Estimating β is what matters. The h_0 is ignored ("Partial likelihood").
- Likelihood depends only on the ranking of the times to death ("non-parametric").
- To obtain predictions of time to death the h_0 is estimated with another procedure.
- "Linear scores" ("prognostic index"): $\hat{\beta}^T\mathbf{X}$
- $\log(h(t)) = log(h_0(t)) + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_q x_q$

Other models

- General: model time to death (or a transformation of time to death, such as log(*time*)) as a function of covariates (and that function could be, e.g., and exponential).
- Regression models using Weibull, exponential, etc.
- Less used than Cox (need to chose scale parameters).
- No proportional hazards assumption, and some times more flexible.
- Not very much used with gene expression data. (But see Schmid y Hothorn, 2008, *BMC Bioinformatics*, 9: 269).

How to assess predictive abilities?

- Censored: simple correlation observed-predicted will not work.
- Continuous data: we cannot discretize.

Not great ideas

- Log-rank test between groups formed from predictions? Why is it a bad idea?

Not great ideas

- Log-rank test between groups formed from predictions? Why is it a bad idea?
- Mentioned in Dupuy and Simon: categorization, chi-square not valid.

Not great ideas

- Log-rank test between groups formed from predictions? Why is it a bad idea?
- Mentioned in Dupuy and Simon: categorization, chi-square not valid.
- Survival model on the linear score, and assess slope (and its p-value). Why bad idea?

Not great ideas

- Log-rank test between groups formed from predictions? Why is it a bad idea?
- Mentioned in Dupuy and Simon: categorization, chi-square not valid.

- Survival model on the linear score, and assess slope (and its p-value). Why bad idea?

- Bovelstad et al., van Wieringen et al., Haibe-Kains et al., use questionable approaches.
- (Yes, SignS implements a few bad ideas . . .).

Predicted and observed: better ideas

(Suggestion: read quickly)

R^2 extended to survival data : $R^2 = 1 - exp(-\frac{2}{n}(l(\hat{\beta}) - l(0)))$ (similar to deviance)

Brier score related to $\Sigma_i(Y_i(t) - q_i(t))^2$, where $Y_i(t) = 1$ if subject i is alive at t, $Y_i(t) = 0$ o.w., and $q_i(t)$ is the probability of surviving until t of subject i (and this we obtain from Cox model). Brier score integrates over all t.

Concordance index (C-index) Probability that, for a pair of patients chosen at random, the one with higher risk really dies earlier. Related to ROC curves.

ROC curves The cutoff is the risk score (or the linear predictor). The event is “dead”. For each time *t* we can compute the sensitivity and specificity as we move the cutoff. Then, compute area under the curve. Finally, integrate that AUC for all *t*.

Again: Brier score, C-index, ROC: using “out of bag” predictions!!

- There are yet other measures. Those are the most widely used.
- Which is best? What if method A is best with C-index and method B with Brier?
- A recurrent theme: **survival analysis is complicated**. Thus, **shortcuts are unlikely to work** (only with good weather and if you know the area).

Tools

- R and BioConductor: several packages.
- Many, many, many (way toooooo many?) web-based tools. Some cited on first set of slides.

Statistical autopsies

To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of.

Sir Ronald Aylmer Fisher, Indian Statistical Congress, 1938

. . . the alternative

We want to foster the team concept, not the image of a statistical policeman arriving at the scene of a crime. Let's nip those false positives in the bud, not in the galleys.

R. G. Easterling, *The American Statistician*, 2010