

Gene Expression Data, Survival Analysis and Power

Master's Project

N. Fournier

Under the supervision of Prof. Stephan Morgenthaler
Chair of Applied Statistics
Swiss Federal Institute of Technology Lausanne (EPFL)

ROeS Seminar Bern
11 september 2007

Table of contents

- 1 Introduction
- 2 Selection
- 3 Survival analysis
- 4 Power
- 5 Conclusion

Introduction

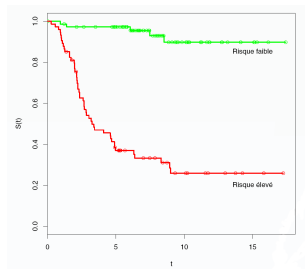
- Genes \longrightarrow Protein production
- Gene *expression* determined by a mesure of RNAm using fluorescence techniques

- Possibility to determine expression of thousands of genes simultaneously

To construct a risk factor using gene expression data.

| | Gene 1 | ... | Gene M |
|-------------|-----------|-----|-----------|
| Patient 1 | $X_{1,1}$ | ... | $X_{M,1}$ |
| Patient 2 | $X_{1,2}$ | ... | $X_{M,2}$ |
| ... | ... | ... | ... |
| Patient N | $X_{1,N}$ | ... | $X_{M,N}$ |

- State : healthy/sick
- Survival time



- 1 Selection of potentially implicated genes
- 2 Construction of the risk factor itself
- 3 Validation

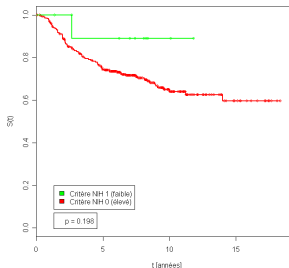
Goal : to predict reappearance of metastasis within 5 years

- 25'000 genes
- First study to construct a selection of 70 genes
- Second study to validate this selection

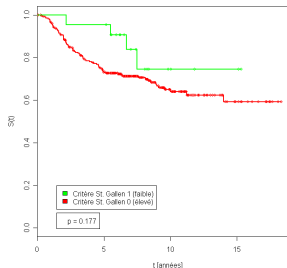
We use the same data as the second study.

Existing risk factors

Based on clinical and/or pathological data.



NIH



St. Gallen

Not significant !

Selection

Potentially implicated genes

N patients, set up into two groups :

- Group 0 : healthy patients (n_0 individuals)
- Group 1 : sick patients (n_1 individuals)

Expression of gene i on patient j denoted by $x_{i,j}$

Mean expression of gene i in group 0 : $\bar{x}_i^{(0)} = \frac{1}{n_0} \sum_{j \in \text{grp } 0} x_{i,j}$

Mean expression of gene i in group 1 : $\bar{x}_i^{(1)} = \frac{1}{n_1} \sum_{k \in \text{grp } 1} x_{i,k}$

Compare $\bar{x}_i^{(0)}$ with $\bar{x}_i^{(1)}$

For every gene i :

$$H_0 : \bar{x}_i^{(0)} = \bar{x}_i^{(1)} \quad \text{vs} \quad H_1 : \bar{x}_i^{(0)} \neq \bar{x}_i^{(1)},$$

using t -statistic for comparison of two groups of unpaired data. Level chosen with the Bonferroni's correction.

Problem : valid only if the distribution is gaussian.

A t -test's modification : the SAM method

Test's statistic :

$$d_i = \frac{\bar{x}_i^{(0)} - \bar{x}_i^{(1)}}{s_i + s_0},$$

where s_i is the standard deviation of repeated expression measurements of gene i , and s_0 a positive constant added to ensure that the variance of d_i is independant of the level of expression.

What is the distribution of d_i under the null hypothesis $H_0 : \bar{x}_i^{(0)} = \bar{x}_i^{(1)}$?

Distribution of d_i under the null hypothesis

Main idea : permutations of patients within the two groups, and \bar{d}_i a mean statistic over the permutations.

| Group 0 | | | Group 1 | | |
|-----------|-----------|-----------|-----------|-----------|-----------|
| Patient 1 | Patient 2 | Patient 3 | Patient 4 | Patient 5 | Patient 6 |

Original (real) configuration : the observed d_i is computed.

| Group 0 | | | Group 1 | | |
|-----------|-----------|-----------|-----------|-----------|-----------|
| Patient 1 | Patient 2 | Patient 5 | Patient 4 | Patient 3 | Patient 6 |

One possible permutation p : $d_i^{(p)}$ is computed.

| Group 0 | | | Group 1 | | |
|-----------|-----------|-----------|-----------|-----------|-----------|
| Patient 6 | Patient 2 | Patient 3 | Patient 4 | Patient 5 | Patient 1 |

Another possible permutation q : $d_i^{(q)}$ is computed.

Distribution of d_j under the null hypothesis

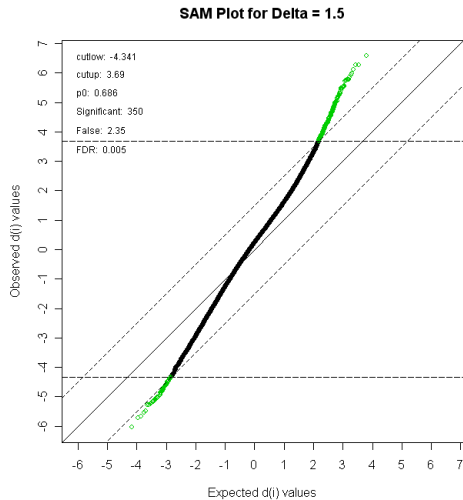
We make B permutations and define :

$$\bar{d}_j = \frac{1}{B} \sum_{b=1}^B d_i^{(b)}$$

If H_0 is true, we should have $d_j \simeq \bar{d}_j$.

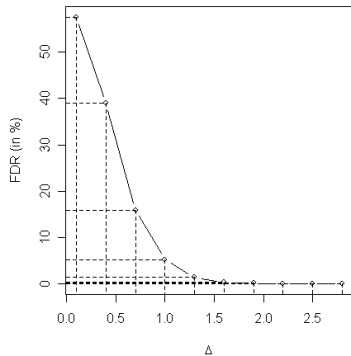
Scatterplot of d_j against \bar{d}_j . Threshold Δ . Differently expressed genes are distant from the diagonal by a distance greater than Δ .

Scatterplot of d_i against \bar{d}_i

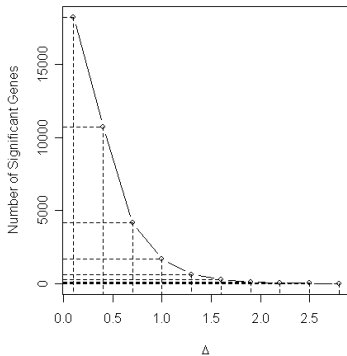


The SAM method and the type I error

Delta vs. FDR



Delta vs. Significant Genes



Strengthen the selection

Use a cross-validation procedure to reinforce the selection.

Divide the population into eight groups randomly, and use only 6 of them to establish the selection. $\binom{8}{2} = 28$ possibilities.

| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
|---|---|---|---|---|---|---|---|
| + | + | + | + | + | + | - | - |
| + | + | + | + | + | - | + | - |
| + | + | + | + | - | + | + | - |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| - | - | + | + | + | + | + | + |

A very strong criteria for a gene's selection.

Survival analysis

We have now a selection of K genes that seem to be expressed differently between healthy and sick patients.

How can the expression of these genes be used to classify the patients into different risk groups ?

Average good profile and correlation

Average good profile : mean expression of the K selected genes over the *healthy patients only* :

$$b = (b_1, \dots, b_K) \quad \text{where} \quad b_i = \frac{1}{n_0} \sum_{j \in \text{grp } 0} x_{i,j}.$$

Correlation between the j^{th} patient's profile and the average good profile :

$$\rho_j = \frac{\sum_{k=1}^K (x_{k,j} - \bar{x}_j)(b_k - \bar{b})}{\left[\sum_{k=1}^K (x_{k,j} - \bar{x}_j)^2 \right]^{1/2} \left[\sum_{k=1}^K (b_k - \bar{b})^2 \right]^{1/2}},$$

where

$$\bar{x}_j = \frac{1}{K} \sum_{k=1}^K x_{k,j} \quad \text{and} \quad \bar{b} = \frac{1}{K} \sum_{k=1}^K b_k.$$

Classification :

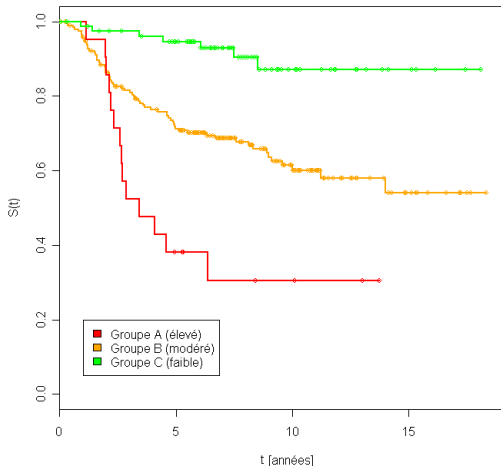
Group A high risk group. Correlation lower than -0.5 ;

Group B moderate risk group. Correlation between -0.5 and 0.5 ;

Group C low risk group. Correlation greater than 0.5 .

Average good profile : nonparametric analysis

Kaplan-Meier's estimators and log-rank test.



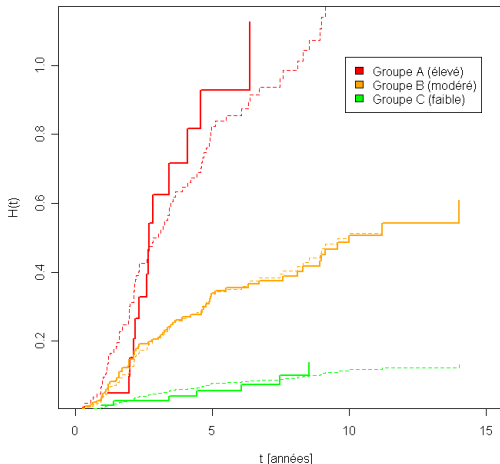
Cox proportionnal hazards model :

$$h_j(t) = h_0(t) \exp(\beta_B \cdot 1_{\{j \in \text{grp B}\}} + \beta_C \cdot 1_{\{j \in \text{grp C}\}}).$$

| Variable | Coef | exp(Coef) | CI 95% | Std err | p-value |
|--------------------------|--------|-----------|---------------|---------|---------|
| Correlation class | | | | | |
| Group B (vs grp A) | -0.892 | 0.41 | [0.23 ; 0.73] | 0.296 | 2.6e-03 |
| Group C (vs grp A) | -2.376 | 0.093 | [0.03 ; 0.23] | 0.465 | 3.2e-07 |

Average good profile : semiparametric analysis

Observed and estimated cumulative hazard curves.



Same data used twice :

- Selection of differently expressed genes
- Survival analysis

Falsify the results : too good !

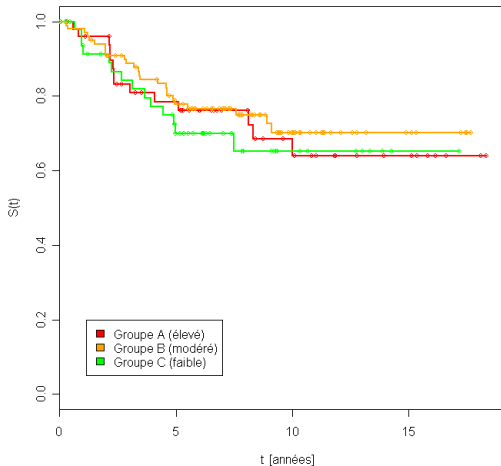
Solutions :

- Another data set
- Cross-validation

- Data are randomly set into two groups
- 1/3 of the data is used to select the genes
- the other part is used for survival analysis

No significant result...

Kaplan-Meier's estimators and log-rank test



Power

What are the effects of a cross-validation procedure on the power of a method such as our one ?

Generate data sets and estimate the power.

Fictive data set

| | Patient | X_1 | ... | X_d | T | Censoring |
|----------------|-----------------|---------------|-------------------------|---------------|---|-------------|
| Group 1 | 1 | $x_{1,1}$ | ... | $x_{1,d}$ | T_1 | C_1 |
| | 2 | $x_{2,1}$ | ... | $x_{2,d}$ | T_2 | C_2 |
| | ⋮ | ⋮ | $\sim \text{Bern}(p_1)$ | ⋮ | $\sim \mathcal{E}(\lambda)$ | ⋮ |
| | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |
| | n_1 | $x_{n_1,1}$ | ... | $x_{n_1,d}$ | T_{n_1} | C_{n_1} |
| Group 2 | $n_1 + 1$ | $x_{n_1+1,1}$ | ... | $x_{n_1+1,d}$ | T_{n_1+1} | C_{n_1+1} |
| | $n_2 + 2$ | $x_{n_1+2,1}$ | ... | $x_{n_1+2,d}$ | T_{n_1+2} | C_{n_1+2} |
| | ⋮ | ⋮ | $\sim \text{Bern}(p_2)$ | ⋮ | $\sim \mathcal{E}((\theta + 1)\lambda)$ | ⋮ |
| | ⋮ | ⋮ | | ⋮ | ⋮ | ⋮ |
| | $n_1 + n_2 = n$ | $x_{n,1}$ | ... | $x_{n,d}$ | T_n | C_n |

$$\lambda > 0, \theta \geq 0, 0 \leq p_1, p_2, \leq 1.$$

- 1 Generate a data set.
- 2 For each covariable $i = 1, \dots, d$ adjust the Cox's model

$$h_j(t) = h_{0,i}(t) \exp(\beta_i x_{i,j}).$$

- 3 Weight for the covariable X_i : z-value of the test for β_i .
Score for the patient j : $S_j = \langle x_j, z \rangle$.
- 4 Classification of the patients into two risk groups.
- 5 Comparison between Kaplan-Meier's curves with the log-rank test.
- 6 Significant difference ?

- 1 Generate a data set.
- 2 For each covariable $i = 1, \dots, d$ adjust the Cox's model

$$h_j(t) = h_{0,i}(t) \exp(\beta_i x_{i,j})$$

using a first half of the data.

- 3 Weight for the covariable X_i : z-value of the test for β_i .
Score for the patient j : $S_j = \langle x_j, z \rangle$.
- 4 Classification of **the second half of the patients** into two risk groups.
- 5 Comparison between Kaplan-Meier's curves with the log-rank test.
- 6 Significant difference ?

Estimation of the power

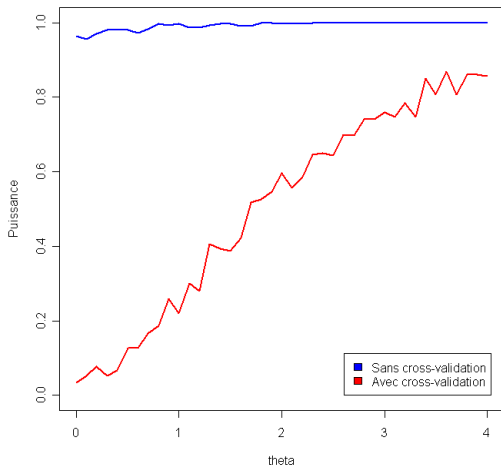
Repeat the method many times.

An estimation of the power will be

$$\frac{\# \text{ of tests that have detected the effect } \theta}{\text{total } \# \text{ of tests}}.$$

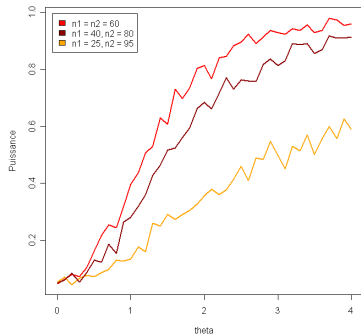
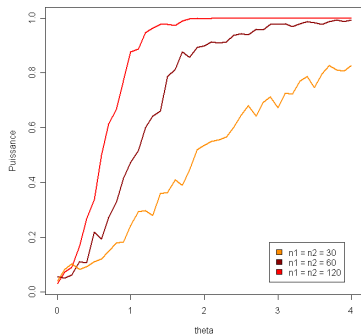
With and without cross-validation.

Estimations of the power for different values of θ .



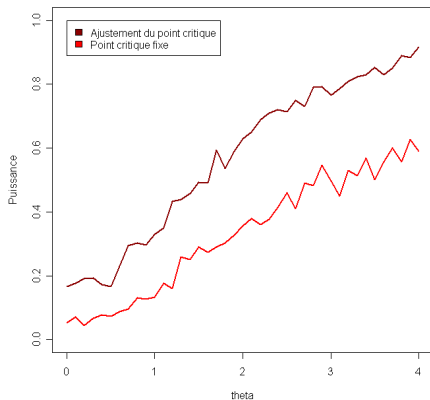
Effect of the number of patients

Power with cross-validation for different values of n_1 and n_2 .



Adjustement of the critical point

The critical point which determine the two risk groups is chosen in a way to optimise the difference between the two resulting survival curves.



Conclusion

A genetic test for breast cancer

A Dutch study based on the same data set has brought a genetic test for breast cancer.

This test was the first genetic test to be approved by the *Food and Drugs Administration* of the USA.

This test seems to be more efficient for low risk patients (95%) than for high risk ones (25-30%).

- Laura van't Veer et al., *Gene expression profiling predicts clinical outcome of breast cancer*, Nature, vol 415, 2002, pp. 530-535.
- Marc van de Vijver et al., *A gene expression signature as a predictor of survival in breast cancer*, New England Journal of Medicine, vol 347, n° 25, 2002, pp. 1999-2009.
- www.agendia.com