# Supervised classification and outliers detection in gene expression data

Laurent Bréhélin and François Major

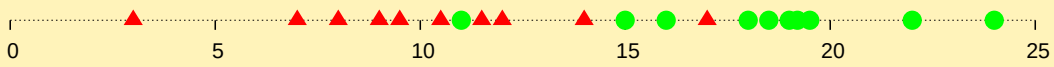
LIRMM, Montpellier, France
LBIT, Montréal, Québec

1. Gene expression data and classification
2. Outliers detection
3. Results

# Gene expression data

| | 1 | 2 | 3 | | p−1 | p |
|---|---|---|---|---|---|---|
| 1 | x11 | x12 | x13 | ··· | | |
| 2 | x21 | x22 | ··· | | | |
| 3 | x31 | ··· | | | | |
| 4 | ··· | | | | | |
| | | | | | | |
| n−1 | | | | | | |
| n | | | | | | |

- Huge number of genes;
- low number of samples;
- high level of noise;
- missing values;
- few discriminant genes.



# Applications

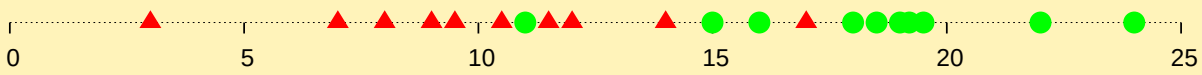
- Cancer diagnosis:
  - annotation (tumor vs. normal);
  - detection (early for better treatment);
  - distinction (cancers with same clinical symptoms);
  - prediction (prognostic).

- Biological interest :
  - what are the genes?
  - what are the classification rules?
  - etc.

# Gene selection

**Why ?**

- eliminate noise ;
- reduce computing time ;
- understand better.




Selection scheme:

1. Gene scoring. e.g., g-score : $s_g = \frac{|m_{g0} - m_{g1}|}{s_{g0} + s_{g1}}$.
2. Selection of the best $k$ genes.

# Learning a classifier - 1

- $G$ a set of selected genes.
- $\mathbf{x} = (x_1, \ldots, x_n)$ a new sample.

**Probabilist approach :**

$$c_{\text{MAP}} = \underset{c \in \{0,1\}}{\text{argmax}}\, P(c|\mathbf{x}_G) \;=\; \underset{c \in \{0,1\}}{\text{argmax}}\, \frac{P(\mathbf{x}_G|c)P(c)}{P(\mathbf{x}_G)} \;=\; \underset{c \in \{0,1\}}{\text{argmax}}\, P(\mathbf{x}_G|c)P(c)$$

- Estimating the $P(c)$ :

$$\widehat{P}(0) = \frac{\#\text{ ex. class 0}}{\#\text{ ex.}}$$

$$\widehat{P}(1) = \frac{\#\text{ ex. class 1}}{\#\text{ ex.}}$$

- The problem is more difficult for $P(\mathbf{x}_G|c)$.

# Learning a classifier - 2

**The naive Bayes approach :**
Gene expression levels are conditionally independent given the class:

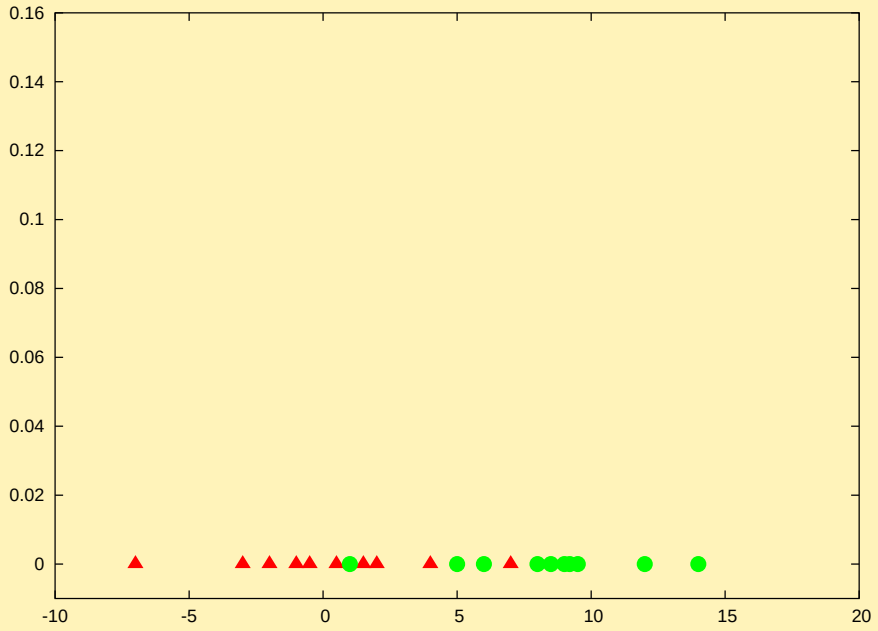
$$P(\mathbf{x}_G|c) = \prod_{g \in G} P(x_g|c).$$

**Normal assumption :**

$$P(x_g|c) \sim \mathcal{N}(x_g; \mu_{gc}, \sigma^2_{gc}).$$

and we have :

- $\widehat{\mu}_{gc} = m_{gc}$
- $\widehat{\sigma^2}_{gc} = s^2_{gc}$



# Learning a classifier - 2

**The naive Bayes approach :**
Gene expression levels are conditionally independent given the class:

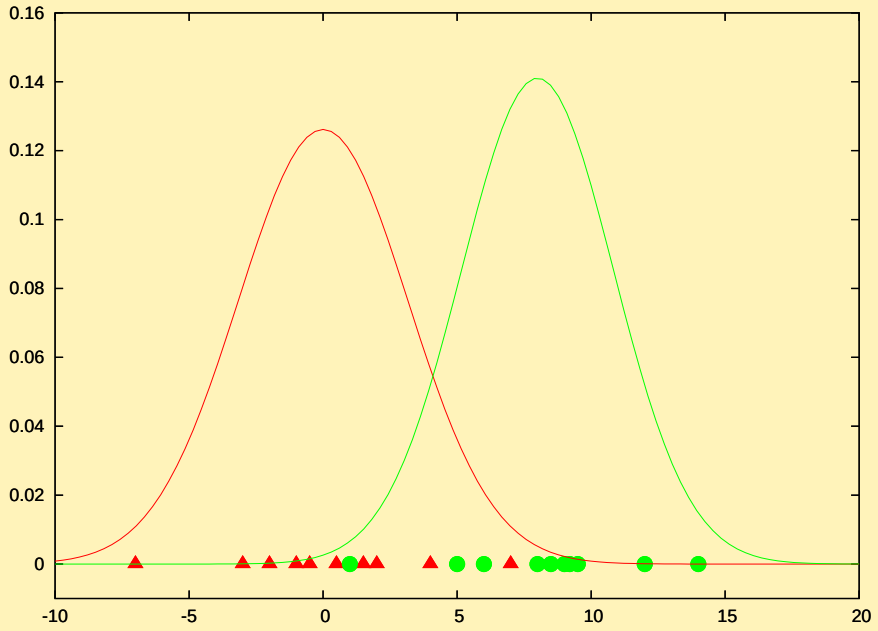
$$P(\mathbf{x}_G|c) = \prod_{g \in G} P(x_g|c).$$

**Normal assumption :**

$$P(x_g|c) \sim \mathcal{N}(x_g; \mu_{gc}, \sigma^2_{gc}).$$

and we have :

- $\widehat{\mu}_{gc} = m_{gc}$
- $\widehat{\sigma^2}_{gc} = s^2_{gc}$



# Evaluating the classifier

Low number of samples $\rightarrow$ cross-validation.

**Leave-one-out procedure:**

**Data** : $X$, the complete set of samples.

**foreach** $x \in X$ **do**
  Learning using $X - x$;
  Classify $x$;

Return the fault coverage;

# Evaluating the classifier

Low number of samples $\rightarrow$ cross-validation.

**Leave-one-out procedure:**

**Data** : $X$, the complete set of samples.

Genes selection;
**foreach** $x \in X$ **do**
    Learning using $X - x$;
    Classify $x$;

Return the fault coverage;

# Non-biased Leave-one-out

**Data** : $X$, the complete set of samples.

**foreach** $x \in X$ **do**

    Gene selection using $X - x$;
    Learning using $X - x$;
    Classify $x$;

Return the fault coverage;

# Non-biased Leave-one-out

**Data** : $X$, the complete set of samples.

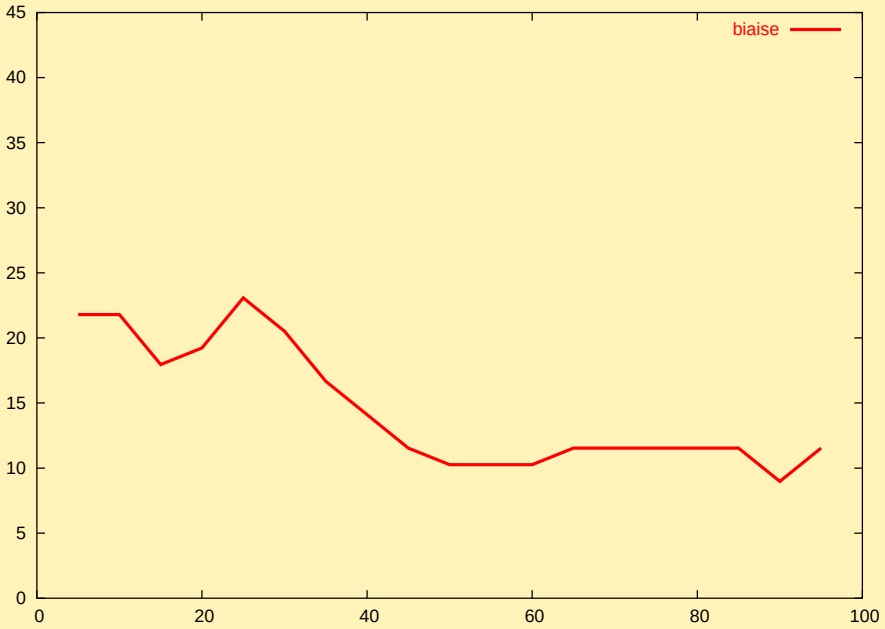
**foreach** $x \in X$ **do**

  Gene selection using $X - x$;

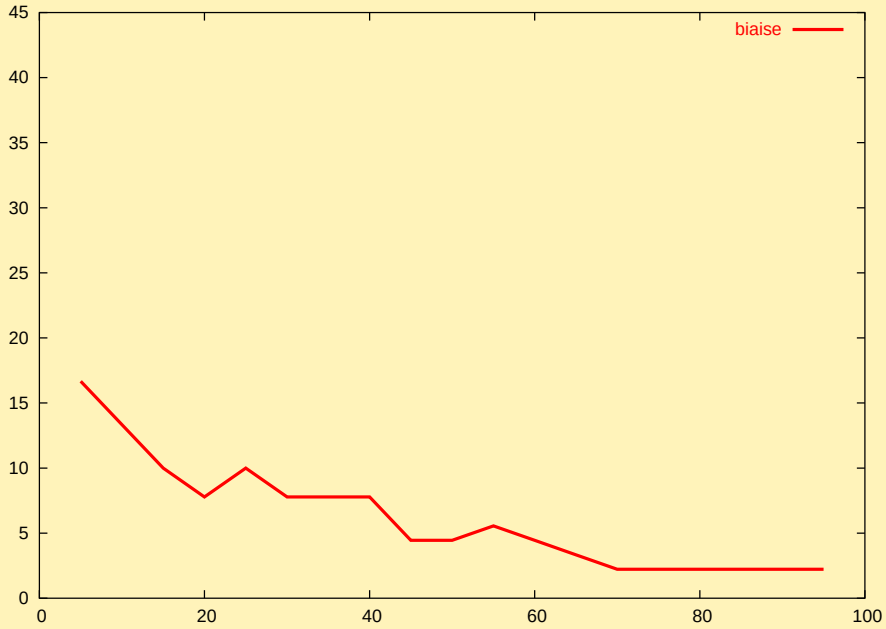
  Learning using $X - x$;

  Classify $x$;

Return the fault coverage;




Breast cancer




SAGE data

# Non-biased Leave-one-out

**Data** : $X$, the complete set of samples.

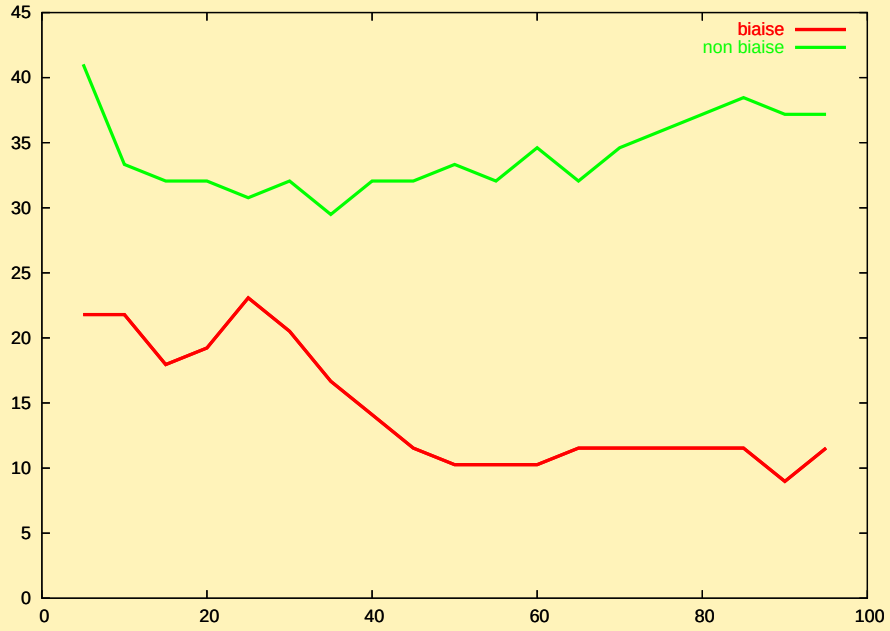
**foreach** $x \in X$ **do**

  Gene selection using $X - x$;

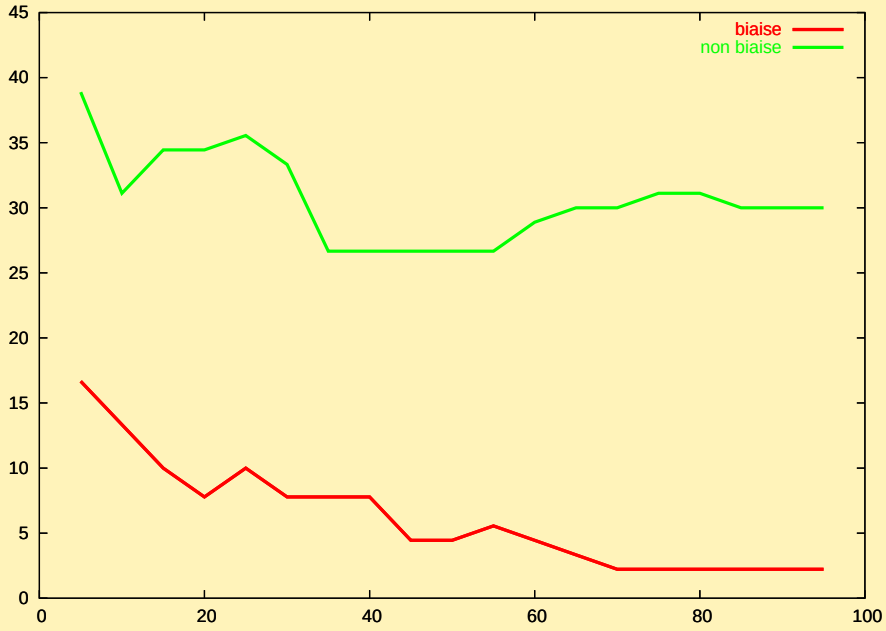
  Learning using $X - x$;

  Classify $x$;

Return the fault coverage;



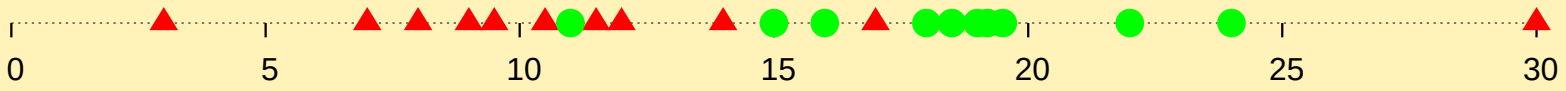

Breast cancer




SAGE data

# Outliers

**Outlier :** A gene expression measurement which differs surprisingly from the other measurements obtained for the same gene on other samples of the same class.




**Outliers bias :**

- the estimates of the model parameters ($\mu_{cg}$ and $\sigma^2_{cg}$) ;
- the gene score.

Ex. g-score : $s_g = \frac{|m_{g0} - m_{g1}|}{s_{g0} + s_{g1}}$

# Origins

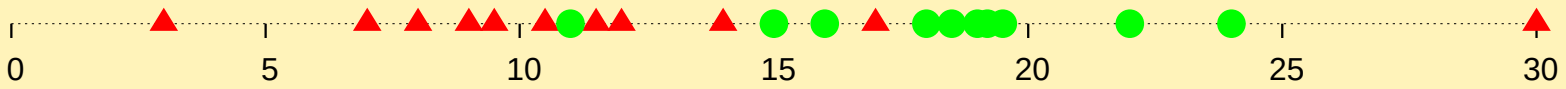
- **Intrinsic factors :** the surprising measurement is actually the true measure. It results from rare but non impossible biological phenomena.

- **Extrinsic factors :** Error measurement :
  - material reasons;
  - human reasons;
  - inherent limits of the measurement method;
  - ...

# Outlier detection - 1

**Principle :**

- assume that the data, with the possible exception of any outlier, form a sample of a given distribution —here the normal distribution;
- use a reasonable test statistical to decide whether or not the suspect measurement is an outlier.

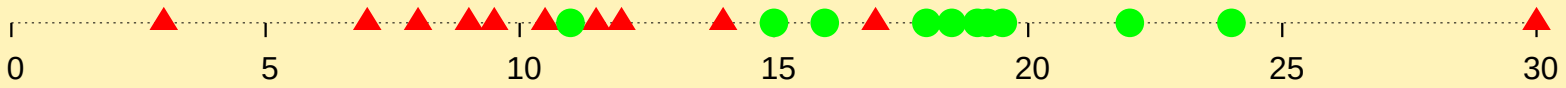



The Thompson statistic :

$$T_{gc} = \frac{|x^*_{gc} - m_{gc}|}{s_{gc}}$$

The greater $T_{gc}$ the more $x^*_{gc}$ is unlikely.

# Outlier detection - 2




**The rule :**

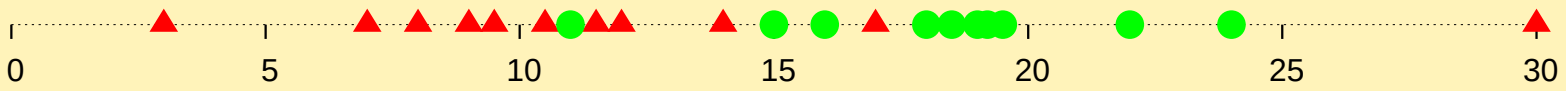
If $T_{gc} \geq \tau_{\alpha c}$ then $x^*_{gc}$ is an outlier.

**How can we set $\tau_{\alpha c}$ ?**

- Compare with what is expected in the null hypothesis $H_0$ that there is no spurious observation (i.e. all points belong to the same normal distribution).
- Find $\tau_{\alpha c}$ so that

$$P(T_{gc} > \tau_{\alpha c}|H_0) = \alpha. \text{ (e.g. } \alpha = 10^{-5}.)$$

# Leave-one-out & outliers

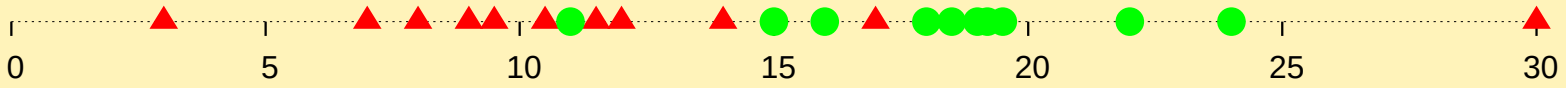



**Data** : $X$ and $\alpha$
**foreach** $x \in X$ **do**

Select a set of genes using $X - x$;
Estimate parameters using $X - x$;
Classify $x$;

Return the fault coverage;

# Leave-one-out & outliers



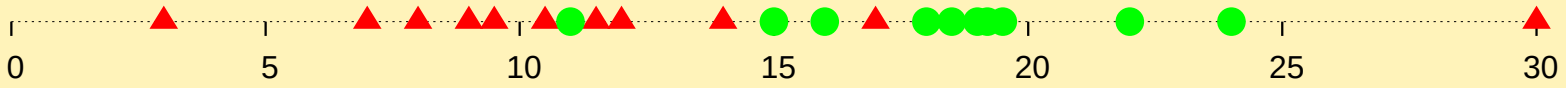

**Data** : $X$ and $\alpha$
**foreach** $x \in X$ **do**

Compute $\tau_{\alpha 0}$ and $\tau_{\alpha 1}$ from $\alpha$;

Select a set of genes using $X - x$;
Estimate parameters using $X - x$;
Classify $x$;

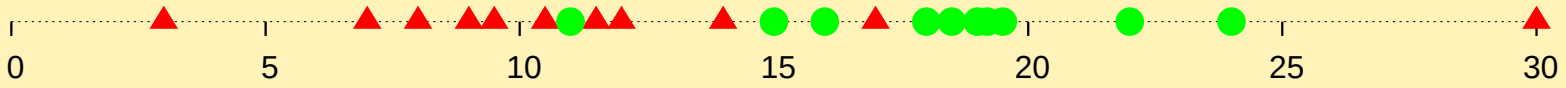
Return the fault coverage;

# Leave-one-out & outliers



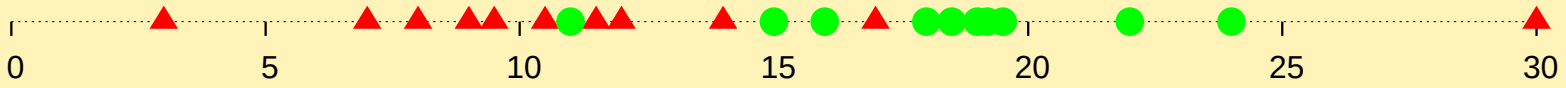

**Data** : $X$ and $\alpha$
**foreach** $x \in X$ **do**
    Compute $\tau_{\alpha 0}$ and $\tau_{\alpha 1}$ from $\alpha$;
    **foreach** *gene* $g$ **do**
        Compute $T_{g0}$ and $T_{g1}$ using $X - x$;
    Select a set of genes using $X - x$;
    Estimate parameters using $X - x$;
    Classify $x$;
Return the fault coverage;

# Leave-one-out & outliers




**Data** : $X$ and $\alpha$
**foreach** $x \in X$ **do**
    Compute $\tau_{\alpha 0}$ and $\tau_{\alpha 1}$ from $\alpha$;
    **foreach** *gene* $g$ **do**
        Compute $T_{g0}$ and $T_{g1}$ using $X - x$;
        **if** $T_{g0} > \tau_{\alpha 0}$ **then**
            remove $x^*_{g0}$;
    Select a set of genes using $X - x$;
    Estimate parameters using $X - x$;
    Classify $x$;
Return the fault coverage;

# Leave-one-out & outliers



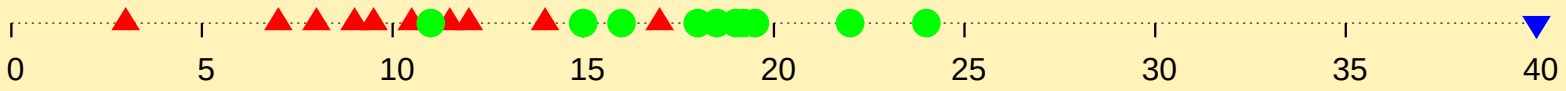

**Data** : $X$ and $\alpha$
**foreach** $x \in X$ **do**
  Compute $\tau_{\alpha 0}$ and $\tau_{\alpha 1}$ from $\alpha$;
  **foreach** *gene* $g$ **do**
    Compute $T_{g0}$ and $T_{g1}$ using $X - x$;
    **if** $T_{g0} > \tau_{\alpha 0}$ **then**
      remove $x^*_{g0}$;
    **if** $T_{g1} > \tau_{\alpha 1}$ **then**
      remove $x^*_{g1}$;
  Select a set of genes using $X - x$;
  Estimate parameters using $X - x$;
  Classify $x$;
Return the fault coverage;

# Gene selection & outliers

**Use the outlier detection in the gene selection procedure.**

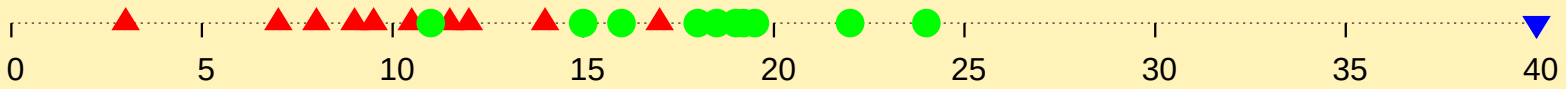
# Gene selection & outliers

**Use the outlier detection in the gene selection procedure.**



# Gene selection & outliers

**Use the outlier detection in the gene selection procedure.**

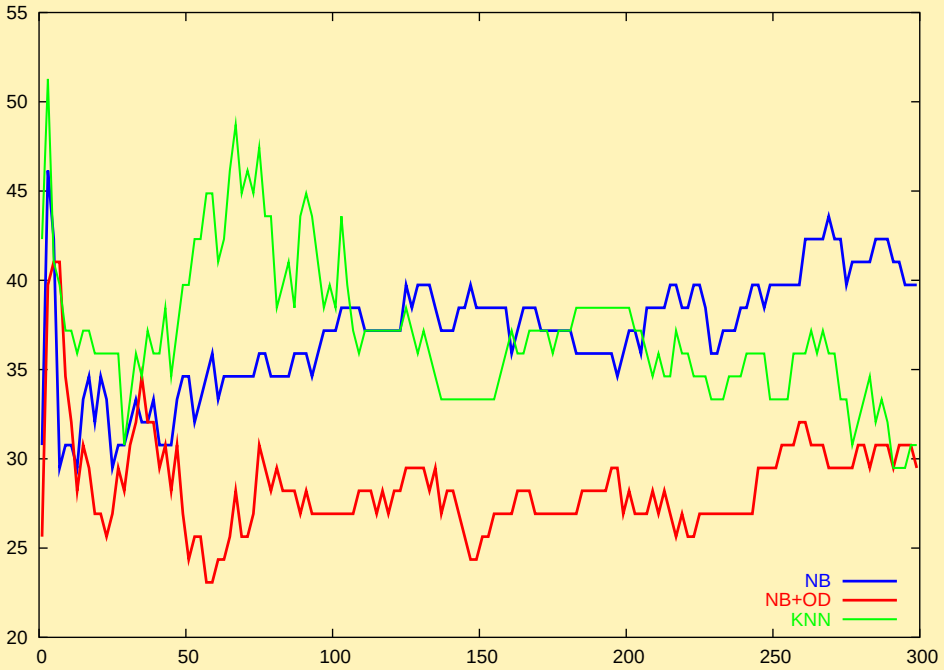


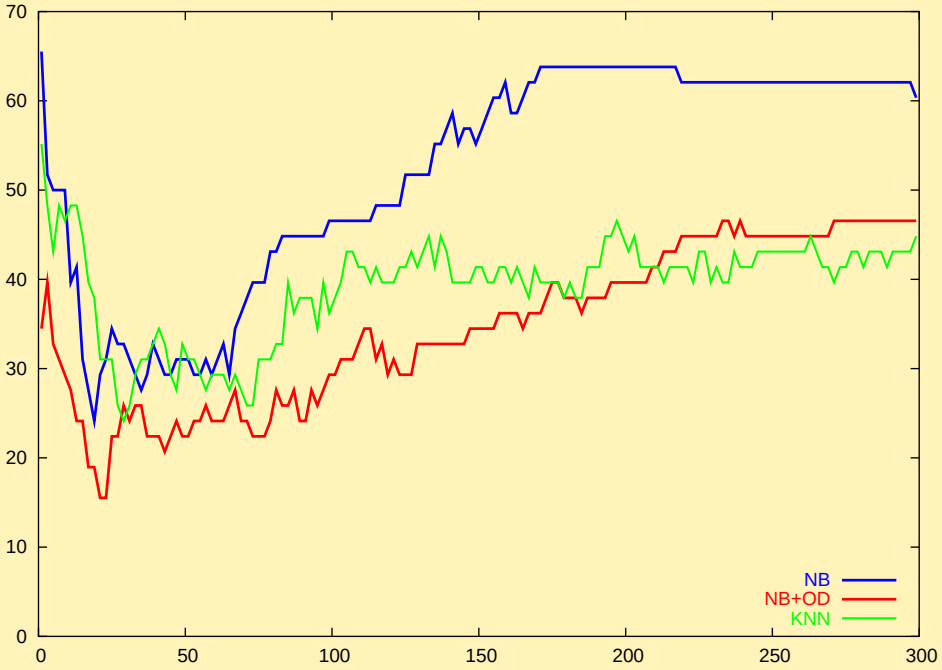

**Data** : $X$, $\alpha'$ and $x$

Compute the $\tau_{\alpha' 0}$ and $\tau_{\alpha' 1}$;
**foreach** *gene* $g$ **do**
    Compute $T'_{g0}$ **and** $T'_{g1}$;
    **if** $T'_{g0} > \tau_{\alpha' 0}$ *and* $T'_{g1} > \tau_{\alpha' 1}$ **then**
        Reject gene $g$;
    **else**
        Compute the score of $g$;
Return the best genes;

# Experiments

- $\alpha' = 10^{-2}$ ;
- $\alpha = 10^{-2},\ 10^{-5},\ 10^{-10},\ 10^{-15},\ 10^{-20}$.




**Breast cancer :**

- 78 samples : 44 vs. 34.
- $\sim$ 24000 genes.

**Lymphome :**

- 58 samples : 32 vs. 26.
- $\sim$ 7000 genes.

# Conclusions

- Outlier detection can improve the performance of the naive Bayes classifier.
- Naive Bayes classifier + outlier detection:
  - simple approach;
  - low computing time;
  - can achieve better results than more sophisticated methods.

Several questions:

- interest of outlier detection combined with other approaches: KNNs, SVMs, weighted voting approach, ...
- comparison with more robust estimates (e.g. median vs. mean);
- outlier origins: intrinsic or extrinsic factors?