

# Dealing with correlation in regression models: application to microarray data analysis

F. Valet, E. Gravier, Y De Rycke and B. Asselain

Institut Curie, Ecole des Mines de Paris,  
INSERM U900, Paris, FRANCE

# Introduction

Microarray analysis: transcriptomic context → large P, small N

**Patients**

|                      | <b>P<sub>1</sub></b> | <b>P<sub>2</sub></b> | <b>P<sub>3</sub></b> | <b>P<sub>4</sub></b> | ... | <b>P<sub>i</sub></b> | ...    | <b>P<sub>N</sub></b> |
|----------------------|----------------------|----------------------|----------------------|----------------------|-----|----------------------|--------|----------------------|
| <b>V<sub>1</sub></b> | -4.031               | -0.520               | 0.559                | 0.112                | ... | 0.677                | ...    | -0.207               |
| <b>V<sub>2</sub></b> | -1.467               | -0.440               | 1.153                | 0.086                | ... | -0.487               | ...    | -0.423               |
| <b>V<sub>3</sub></b> | 1.376                | 0.735                | -1.344               | -0.367               | ... | 0.433                | ...    | -0.051               |
| <b>V<sub>4</sub></b> | 1.398                | 0.169                | -0.817               | -1.233               | ... | -1.967               | ...    | -0.402               |
| ...                  | ...                  | ...                  | ...                  | ...                  | ... | ...                  | ...    | ...                  |
| <b>V<sub>K</sub></b> | 0.123                | -0.569               | 7.236                | 5.001                | ... | -1.799               | -0.655 | 9.788                |
| ...                  | ...                  | ...                  | ...                  | ...                  | ... | ...                  | ...    | ...                  |
| <b>V<sub>P</sub></b> | 2.297                | -0.001               | 0.378                | -0.810               | ... | -2.151               | ...    | -0.458               |

P ~ 55000

N ~ 100

# Introduction

we usually want to know...

**Patients (  group A,  group B)**

|                      | <b>P<sub>1</sub></b> | <b>P<sub>2</sub></b> | <b>P<sub>3</sub></b> | <b>P<sub>4</sub></b> | ... | <b>P<sub>i</sub></b> | ...    | <b>P<sub>N</sub></b> |
|----------------------|----------------------|----------------------|----------------------|----------------------|-----|----------------------|--------|----------------------|
| <b>V<sub>1</sub></b> | -4.031               | -0.520               | 0.559                | 0.112                | ... | 0.677                | ...    | -0.207               |
| <b>V<sub>2</sub></b> | -1.467               | -0.440               | 1.153                | 0.086                | ... | -0.487               | ...    | -0.423               |
| <b>V<sub>3</sub></b> | 1.376                | 0.735                | -1.344               | -0.367               | ... | 0.433                | ...    | -0.051               |
| <b>V<sub>4</sub></b> | 1.398                | 0.169                | -0.817               | -1.233               | ... | -1.967               | ...    | -0.402               |
| ...                  | ...                  | ...                  | ...                  | ...                  | ... | ...                  | ...    | ...                  |
| <b>V<sub>K</sub></b> | 0.123                | -0.569               | 7.236                | 5.001                | ... | -1.799               | -0.655 | 9.788                |
| ...                  | ...                  | ...                  | ...                  | ...                  | ... | ...                  | ...    | ...                  |
| <b>V<sub>P</sub></b> | 2.297                | -0.001               | 0.378                | -0.810               | ... | -2.151               | ...    | -0.458               |

What are the differentially expressed genes (DEG) between the two groups ?

We usually want to select DEG that can predict P(A)

## Context : predictive signatures

---

- Few preliminary steps
  - data reduction (I.Q.R.)
  - selection of the most significant genes (S.A.M.)
  
- Multivariate logistic regression: predictive signature
  - uniqueness ?
  - stability ?
  - many correlations between variables → misleading interpretations

*« It is important to be able to predict in which group patients will be classified, but it is also of prime importance to have an interpretative and comprehensive approach of these predictive signatures »*

**Objective:** to propose an algorithm based on logistic regression that can take into account correlation patterns.

# Method: pruning correlation in multivariate logistic regression

---

## □ We assume that:

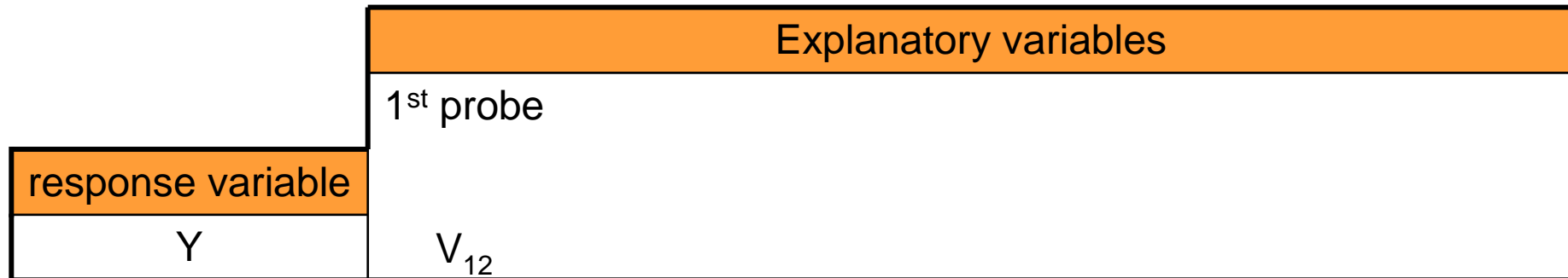
- if two probes show a very high correlation between their parameters, information provided by one of them could be redundant and may affect estimates quality
- two probes are too much correlated if the correlation between their corresponding parameters is greater than a threshold  $\rho_{lim}$

## □ We define a «pruning correlation algorithm» using a forward selection based on AIC



## Method: pruning correlation algorithm

---



- Model with  $V_{12}$  has the smallest AIC value among all univariate logistic models

## Method: pruning correlation algorithm

---

|                   |                       | Explanatory variables |                       |
|-------------------|-----------------------|-----------------------|-----------------------|
|                   | 1 <sup>st</sup> probe |                       | 2 <sup>nd</sup> probe |
| response variable | $V_{12}$              | +                     | $V_{45}$              |
| Y                 |                       |                       |                       |

- Model with  $V_{12}$  has the smallest AIC value among all univariate logistic models
- Model with  $V_{12}$  and  $V_{45}$  has the smallest AIC value among all bivariate logistic models

and  $\left| \rho(\beta_{V_{12}}, \beta_{V_{45}}) \right| \leq \rho_{\text{lim}} \Rightarrow V_{45}$  is kept in the model

## Method: pruning correlation algorithm

---

|                           |   | Explanatory variables |                       |
|---------------------------|---|-----------------------|-----------------------|
|                           |   | 1 <sup>st</sup> probe | 2 <sup>nd</sup> probe |
| response variable         | Y | V <sub>12</sub>       | +                     |
| pruning correlation lists |   | V <sub>45</sub>       |                       |

- Model with V<sub>12</sub> has the smallest AIC value among all univariate logistic models
- Model with V<sub>12</sub> and V<sub>45</sub> has the smallest AIC value among all bivariate logistic models

and  $\left| \rho(\beta_{V_{12}}, \beta_{V_{45}}) \right| > \rho_{\text{lim}} \Rightarrow V_{45}$  is not included in the model

BUT V<sub>45</sub> is stored in a « pruning correlation list of V<sub>12</sub> »

## Method: pruning correlation algorithm

|                           |   | Explanatory variables |                       |
|---------------------------|---|-----------------------|-----------------------|
|                           |   | 1 <sup>st</sup> probe | 2 <sup>nd</sup> probe |
| response variable         | Y   | V <sub>12</sub>       | V <sub>7</sub>        |
| pruning correlation lists | V <sub>45</sub><br>V <sub>81</sub><br>... | +                     |                       |

- Model with V<sub>12</sub> has the smallest AIC value among all univariate logistic models
- Model with V<sub>12</sub> and V<sub>7</sub> has the smallest AIC value among all bivariate logistic models

and  $\left| \rho(\beta_{V_{12}}, \beta_{V_7}) \right| \leq \rho_{\text{lim}} \Rightarrow V_7$  is included in the model

***V<sub>7</sub> is the first probe that matches the two constraints (AIC and correlation)***

## Method: pruning correlation algorithm → final model

|                           |                  | Explanatory variables |   |                       |   |                       |   |                       |   |                       |  |
|---------------------------|------------------|-----------------------|---|-----------------------|---|-----------------------|---|-----------------------|---|-----------------------|--|
|                           |                  | 1 <sup>st</sup> probe |   | 2 <sup>nd</sup> probe |   | 3 <sup>rd</sup> probe |   | 4 <sup>th</sup> probe |   | 5 <sup>th</sup> probe |  |
| response variable         | Y                | V <sub>12</sub>       | + | V <sub>7</sub>        | + | V <sub>780</sub>      | + | V <sub>1</sub>        | + | V <sub>300</sub>      |  |
| pruning correlation lists | V <sub>45</sub>  |                       |   | V <sub>396</sub>      |   | V <sub>314</sub>      |   | V <sub>159</sub>      |   | -                     |  |
|                           | V <sub>812</sub> |                       |   | V <sub>387</sub>      |   | V <sub>76</sub>       |   | V <sub>143</sub>      |   |                       |  |
|                           | V <sub>696</sub> |                       |   | V <sub>43</sub>       |   | V <sub>823</sub>      |   |                       |   |                       |  |
|                           | V <sub>712</sub> |                       |   | V <sub>25</sub>       |   | V <sub>25</sub>       |   |                       |   |                       |  |
|                           | V <sub>471</sub> |                       |   | V <sub>668</sub>      |   |                       |   |                       |   |                       |  |
|                           | V <sub>261</sub> |                       |   |                       |   |                       |   |                       |   |                       |  |
|                           | V <sub>348</sub> |                       |   |                       |   |                       |   |                       |   |                       |  |

For this example, five probes were selected in the model

 genes that would have brought information to the model

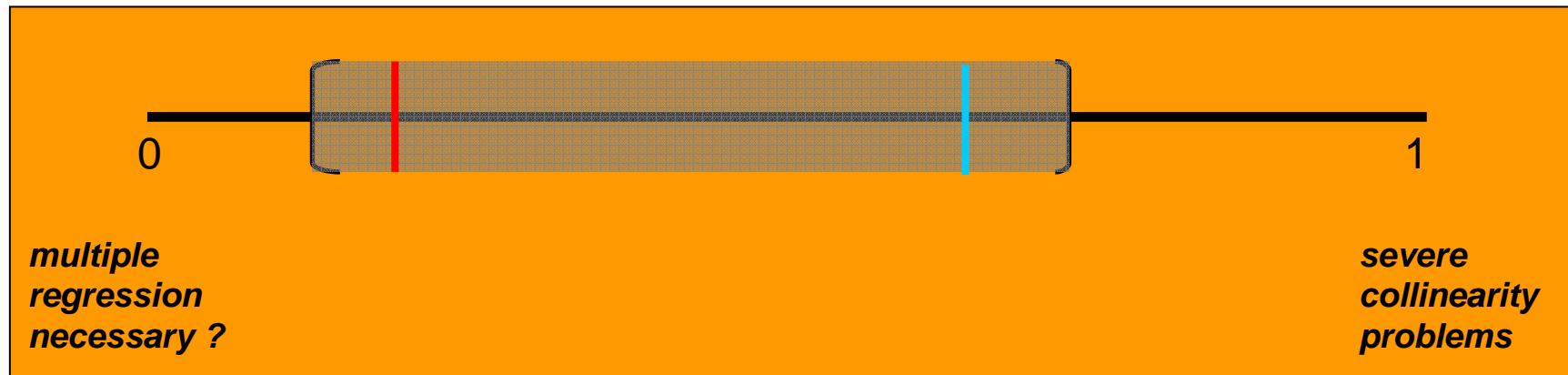
Note: we may observe correlations between one parameter and two or more parameters already in the model (example of probe V<sub>25</sub>)

## Method: correlation threshold(s)

---

- MLE  $\hat{\beta}$  have a large-sample normal distribution with covariance matrix equal to the inverse of the information matrix

Range of thresholds that can be taken for  $\rho_{lim}$



- Estimation of  $\rho_{lim}$  can be done using several approaches
  - tuning parameter
  - maximum variance inflation factor
  - simplest restrictive approach → simulation of multivariate logistic models with normal independent variables

## Application

---

- ❑ Evaluation of pathological complete response(pCR) in Breast cancer
- ❑ transcriptomic aim: to detect genes associated to pCR and try to understand biological and independent pathways involved in pCR
- ❑ transcriptomic data → 54675 probes, example based on 153 patients

$$\begin{array}{l} \text{response} \\ \text{variable} \end{array} Y = \begin{cases} 1 \text{ if pCR} \\ 0 \text{ otherwise} \end{cases} \quad \begin{array}{l} \text{explanatory} \\ \text{variables} \end{array} \left\{ \begin{array}{l} V_1 \\ V_2 \\ \dots \\ V_{54675} \end{array} \right.$$

- ❑ Preliminary filtering and supervised analyses
  - I.Q.R. → from 54675 to 27337 probes
  - S.A.M. with a FDR <0.001 → from 27337 to 1092 probes

## Application: usual multiple regression

9 genes were selected in the « classical » logistic regression...

| Probes           | Univariate Logistic |      |        | Multivariate Logistic |      |       |
|------------------|---------------------|------|--------|-----------------------|------|-------|
|                  | $\hat{\beta}$       | se   | pv     | $\hat{\beta}$         | se   | pv    |
| V1 → PFDN5       | -3.72               | 0.74 | <0.001 | -2.18                 | 2.08 | 0.293 |
| V2 → SLC30A1     | -1.91               | 0.51 | <0.001 | -7.92                 | 3.11 | 0.011 |
| V3 → CCND1       | -0.64               | 0.16 | <0.001 | -1.90                 | 0.75 | 0.011 |
| V4 → TOR1B       | -1.80               | 0.58 | 0.002  | -1.74                 | 1.43 | 0.225 |
| V5 → FOXO4       | -2.98               | 0.99 | 0.003  | -14.99                | 6.11 | 0.014 |
| V6 → C15orf24    | -2.22               | 0.63 | <0.001 | -8.08                 | 3.64 | 0.027 |
| V7 → UBXD3       | -0.83               | 0.41 | 0.041  | 2.71                  | 1.48 | 0.066 |
| V8 → '216173_at' | -0.70               | 0.30 | 0.019  | -2.89                 | 1.84 | 0.115 |
| V9 → AAAS        | -1.88               | 0.49 | <0.001 | -3.73                 | 3.05 | 0.222 |

## Application: usual multiple regression

Collinearity may *partly* explain non significant p-values...

| Probes           | Univariate Logistic |      |        | Multivariate Logistic |      |       |
|------------------|---------------------|------|--------|-----------------------|------|-------|
|                  | $\hat{\beta}$       | se   | pv     | $\hat{\beta}$         | se   | pv    |
| V1 → PFDN5       | -3.72               | 0.74 | <0.001 | -2.18                 | 2.08 | 0.293 |
| V2 → SLC30A1     | -1.91               | 0.51 | <0.001 | -7.92                 | 3.11 | 0.011 |
| V3 → CCND1       | -0.64               | 0.16 | <0.001 | -1.90                 | 0.75 | 0.011 |
| V4 → TOR1B       | -1.80               | 0.58 | 0.002  | -1.74                 | 1.43 | 0.225 |
| V5 → FOXO4       | -2.98               | 0.99 | 0.003  | -14.99                | 6.11 | 0.014 |
| V6 → C15orf24    | -2.22               | 0.63 | <0.001 | -8.08                 | 3.64 | 0.027 |
| V7 → UBXD3       | -0.83               | 0.41 | 0.041  | 2.71                  | 1.48 | 0.066 |
| V8 → '216173_at' | -0.70               | 0.30 | 0.019  | -2.89                 | 1.84 | 0.115 |
| V9 → AAAS        | -1.88               | 0.49 | <0.001 | -3.73                 | 3.05 | 0.222 |

...because selection was not based on p-values

## Application: usual multiple regression

More surprisingly, V7 estimate has different signs...

| Probes           | Univariate Logistic |      |        | Multivariate Logistic |      |       |
|------------------|---------------------|------|--------|-----------------------|------|-------|
|                  | $\hat{\beta}$       | se   | pv     | $\hat{\beta}$         | se   | pv    |
| V1 → PFDN5       | -3.72               | 0.74 | <0.001 | -2.18                 | 2.08 | 0.293 |
| V2 → SLC30A1     | -1.91               | 0.51 | <0.001 | -7.92                 | 3.11 | 0.011 |
| V3 → CCND1       | -0.64               | 0.16 | <0.001 | -1.90                 | 0.75 | 0.011 |
| V4 → TOR1B       | -1.80               | 0.58 | 0.002  | -1.74                 | 1.43 | 0.225 |
| V5 → FOXO4       | -2.98               | 0.99 | 0.003  | -14.99                | 6.11 | 0.014 |
| V6 → C15orf24    | -2.22               | 0.63 | <0.001 | -8.08                 | 3.64 | 0.027 |
| V7               | -0.83               | 0.41 | 0.041  | 2.71                  | 1.48 | 0.066 |
| V8 → '216173_at' | -0.70               | 0.30 | 0.019  | -2.89                 | 1.84 | 0.115 |
| V9 → AAAS        | -1.88               | 0.49 | <0.001 | -3.73                 | 3.05 | 0.222 |

...this is clearly due to collinearity !

## Application: pruning correlation algorithm

□  $\rho_{\text{lim}} \approx 0.159$


| Probes      | Univariate Logistic |      |        | Multivariate Logistic |      |       |
|-------------|---------------------|------|--------|-----------------------|------|-------|
|             | $\hat{\beta}$       | se   | pv     | $\hat{\beta}$         | se   | pv    |
| V1 → PFDN5  | -3.72               | 0.74 | <0.001 | -2.23                 | 0.97 | 0.022 |
| V2 → NBR1   | -2.65               | 0.90 | <0.001 | -1.86                 | 0.79 | 0.018 |
| V3 → WISP2  | -0.71               | 0.18 | <0.001 | -0.47                 | 0.22 | 0.035 |
| V4 → ULK4   | -7.77               | 2.66 | 0.003  | -7.02                 | 3.80 | 0.065 |
| V5 → HNRPH2 | -3.31               | 0.95 | <0.001 | -2.32                 | 1.32 | 0.078 |
| V6 → MUC1   | -0.63               | 0.17 | <0.001 | -0.47                 | 0.25 | 0.056 |

- No « dramatic differences » (values and / or signs)
- Maximum observed correlation values between parameters = 0.15
- All estimates are negative: when any probe expression level increases, the probability of pCR decreases

## Results: pruning correlations lists from the final model

| 1 <sup>st</sup> Probe  | 2 <sup>nd</sup> Probe  | 3 <sup>rd</sup> Probe                          | 4 <sup>th</sup> Probe                       | 5 <sup>th</sup> Probe    | 6 <sup>th</sup> Probe |
|--|--|--|---|--------------------------|-----------------------|
| <b>PFDN5</b>   | <b>NBR1</b>  | <b>WISP2</b>                                   | <b>ULK4</b>                                 | <b>HNRPH2</b>            | <b>MUC1</b>           |
| SLC30A1<br>PPM1H<br>ARNT<br>FGD6<br>LETMD1<br>EIF4B<br>KIAA0701<br>PDGFD<br>DCN<br>ECM2<br>SMOC2<br>LIMA1<br>C9orf58 | VASN<br>PTGER3<br>PPM1H<br>LETMD1<br>EIF4B<br>'227769_at'<br>SLITRK6<br>PDGFD<br>DCN | LETMD1<br>EIF4B<br>KIAA0701<br>SLITRK6<br>TAF9 | C9orf58<br>COPZ1<br>STK39<br>BLVRA<br>SLC1A | TSPAN3<br>CYBRD1<br>FNTB | -                     |

 « *independant* » predictive variables → leader genes

 « *redundant* » predictive variables

From the biologist point of view:

- find group of genes that contributes « equally » to the prediction
- find genes or group of genes that contribute « independently »

## Conclusion

---

- ❑ The pruning correlation method is quite simple and easy to compute.
- ❑ The correlation structure between parameters is taken into account to deal with multicollinearity
  - biological rationale → find independent pathways
  - statistical rationale → limit multicollinearity
- ❑ Parameter interpretation is more realistic with smaller variances: this method seems to be more suitable for parameter interpretation
- ❑ We can easily extend this method to Cox models

## Discussion : what's next ?

---

- ❑ What is the best way to use information provided by the signature and correlations lists ?
  - correlation lists → to metagenes signatures
  - Gene Ontology → investigate whether correlations lists contain known pathways
  
- ❑ Comparison with others methods
  - Ridge
  - Lasso