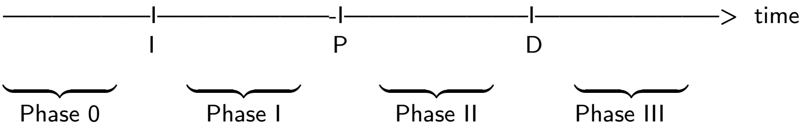
Latent variable model for carcinogenesis.

Sandra Plancade, University of Tromsø, Norway.
Gregory Nuel, Université Paris-Descartes
Eiliv Lund, University of Tromsø.

21 Novembre 2011

Carcinogenesis

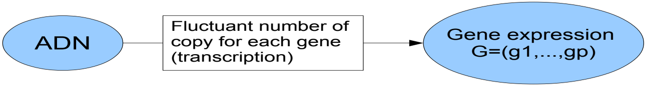


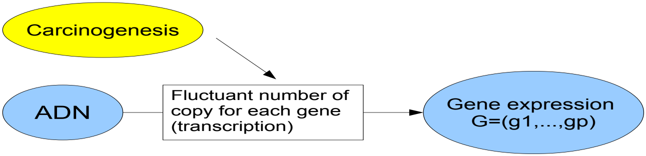
- Cross-sectional study / prospective study
- Exposure / gene expression
- NOWAC : prospective study on gene expression.

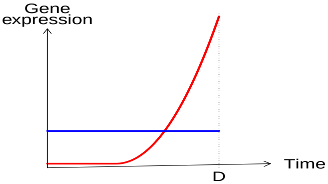
1. Data and biological model
   - Gene expression and carcinogenesis
   - Data
   - Model

2. Linear model with latent variable
   - Statistical model
   - Parameter estimation

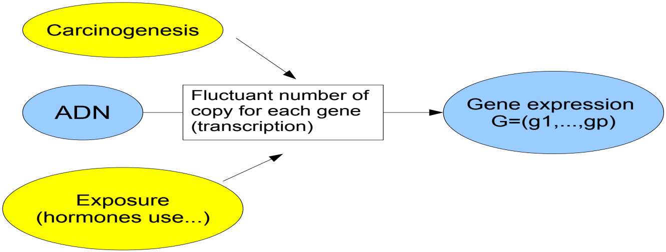
3. Results on simulated data
   - Simulations
   - Performances of the model
   - Comparison with other models

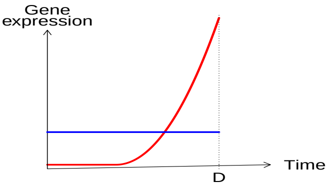

ADN
Fluctuant number of copy for each gene (transcription)
Gene expression G=(g1,...,gp)

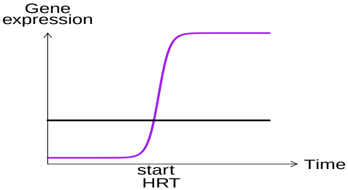




red : gene involved in carcinogenesis
blue : gene non involved





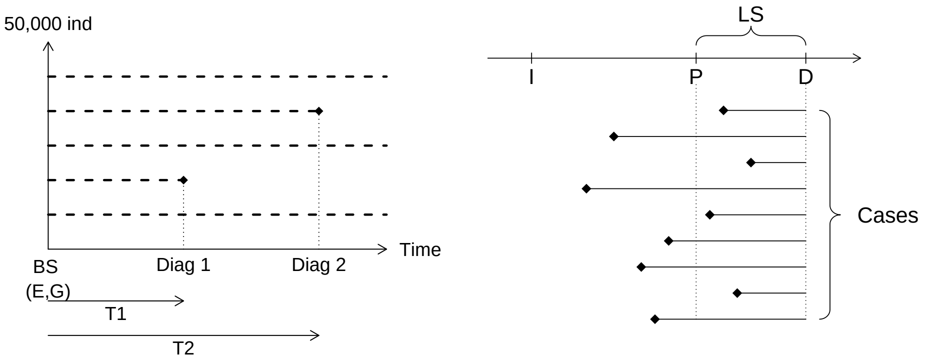


red : gene involved in carcinogenesis
blue : gene non involved

purple : gene linked to HRT
black : gene non-linked to HRT

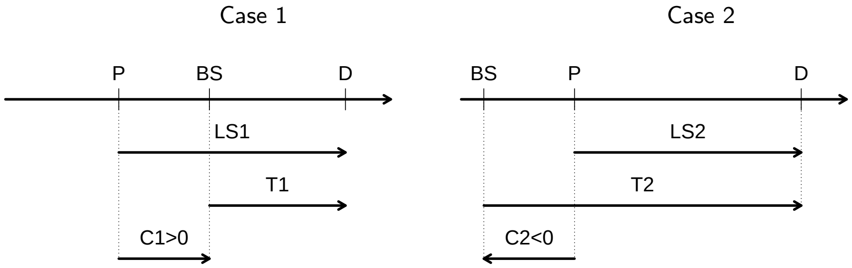
Nowac Data

- Cohort of 50,000 women.



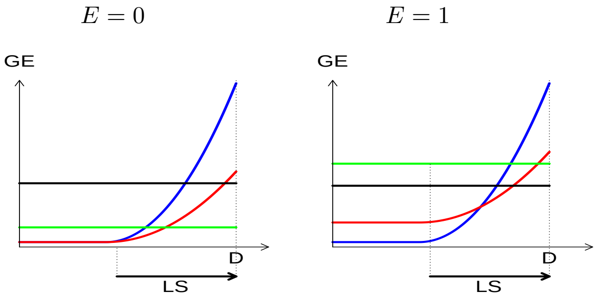
- 700 cases / 5 years of follow-up.
- For each case :
  - Follow-up time T
  - Gene expression G at time of blood sample.
  - Exposure E at time of blood sample.

Last-stage model



- $LS_i = C_i + T_i$
- $\mathbb{P}[LS_i | E_i]$
- $C_i \sim \mathbb{P}[LS_i - T_i | E_i]$

Gene expression model



4 genes

blue : carcinogenesis
red : carcinogenesis and exposure.
green : exposure.
black : invariant gene.

G_i^g gene expression of gene g, and every case i :

$$\mathbb{P}[G_i^g | C_i \mathbb{I}(C_i > 0), E_i]$$

- Notations : for each case $i = 1, \dots, n$.
  - ⋆ T_i follow-up time.
  - ⋆ $C_i = LS_i - T_i$ algebraic distance to start of last stage at time of BS.
  - ⋆ $G_i = (G_i^1, \dots, G_i^p)$ gene expression at time of BS.
  - ⋆ $E_i = (E_{i,1}, \dots, E_{i,d})$ exposure vector.
- $LS_i \sim \Gamma(k, \theta)$ with $\begin{cases} k = 1 + \exp(\langle \kappa, (1, E_i) \rangle), \\ \theta = \exp(\langle \tau, (1, E_i) \rangle). \end{cases}$

  where $\langle \kappa, (1, E_i) \rangle = \kappa_0 + \kappa_1 E_{i,1} + \cdots + \kappa_d E_{i,d}$.
- For each gene g,

$$G_i^g \ = \ \beta_0^g \ + \ \langle \beta_1^g, E_i \rangle + \beta_2^g C_i \mathbb{I}(C_i > 0) \ + \ \varepsilon_i^g$$

  where $\{\varepsilon_i^g\}$ are independent with distribution $\mathcal{N}(0, \sigma_g)$.
  - Gamma distribution : cell reproduction model + flexible model
  - Linear dependence of time.
  - Independence between genes.

◇ Gene g involved in last stage iif $\beta_2^g \neq 0$.

Parameter estimation

$$\begin{cases} C_i + T_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i \rangle), \quad \theta = \exp(\langle \tau, E_i \rangle) \\ G_i^g = \langle \beta^g, (1, E_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

- Starting point from an heuristic.
- Iteration.

$$\Theta_j = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$$

Sample $(C_i^{(j),1}, \dots, C_i^{(j),N})$ from $\mathbb{P}_{\Theta^{(j)}}[C_i | G_i, T_i, E_i]$

$(\beta^{(j+1)}, \sigma^{(j+1)})$ MLE from $\mathbb{P}_{\beta,\sigma}[G_i | E_i, C_i^{(j)}]$
$(\kappa^{(j+1)}, \tau^{(j+1)})$ MLE from $\mathbb{P}_{\kappa,\tau}[C_i^{(j)} | E_i, T_i]$

Algorithm SEM

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i \rangle), \quad \theta = \exp(\langle \tau, E_i \rangle) \\ G_i^g = \langle \beta^g, (1, E_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

1) Let $\Theta^{(j)} = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$

2) Simulated expectation.

$$\overbrace{\sum_{i=1}^n \int_{C_i} \log \mathbb{P}_\Theta[G_i, C_i | E_i, T_i] \mathbb{P}_{\Theta^{(j)}}[C_i | G_i, E_i, T_i]}^{\mathbb{E}_{\Theta^{(j)}}[\log \mathbb{P}_\Theta[G_i, C_i|E_i,T_i]]}$$
$$= \quad \sum_{i=1}^n \mathbb{E}_{\Theta^{(j)}}[\log \mathbb{P}_{\beta,\sigma}[G_i|E_i, C_i]] + \sum_{i=1}^n \mathbb{E}_{\Theta^{(j)}}[\log \mathbb{P}_{\kappa,\tau}[C_i|E_i,T_i]]$$

Sample N repetitions of $\{C_i^{(j)}\}_{i=1:n}$ from distribution $\mathbb{P}_{\Theta^{(j)}}[C_i|G_i,T_i,E_i]$.

$$\begin{aligned} \mathbb{P}_{\Theta^{(j)}}[C_i|G_i,T_i,E_i] &= \frac{\mathbb{P}_{\Theta^{(j)}}[G_i|C_i,E_i,T_i] \cdot \mathbb{P}_{\Theta^{(j)}}[C_i|E_i,T_i]}{\mathbb{P}_{\Theta^{(j)}}[G_i|E_i,T_i]} \\ &\propto \Pi_{g=1}^p \underbrace{\mathbb{P}_{\Theta^{(j)}}[G_i^g|C_i,E_i]}_{\mathcal{N}(\langle \beta_j^g, (1,E_i,C_i)\rangle, \sigma_g)} \cdot \underbrace{\mathbb{P}_{\Theta^{(j)}}[C_i|E_i,T_i]}_{\Gamma(k_j,\theta_j) - T_i} \end{aligned}$$

2) Maximization.

$$(\beta_g^{(j+1)}, \sigma_g^{(j+1)}) = \arg\max \sum_{i=1}^{n} \left(\frac{1}{N} \sum_{\ell=1}^{N} \phi(G_i^g - \langle \beta_g, (1, E_i, C_{i,\ell}^{(j)}) \rangle) \right)$$

where ϕ is the standard normal density and

$$(\kappa^{(j+1)}, \tau^{(j+1)}) = \arg\max$$
$$\sum_{i=1}^{n} \left(\frac{1}{N} \sum_{\ell=1}^{N} \psi(C_{i,\ell}^{(j)} + T_i | k = 1 + \exp(\langle \kappa, E_i \rangle),\ \theta = \exp(\langle \tau, E_i \rangle) \right)$$

where ψ is the gamma distribution density.

3) Parameter estimation.

$$\widehat{\Theta} = \sum_{j \geq \textsf{burn-in}} \Theta^{(j)}.$$

Simulations

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i \rangle), \quad \theta = \exp(\langle \tau, E_i \rangle) \\ G_i^g = \langle \beta^g, (1, E_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

- Observed follow-up times $(T_1, \ldots, T_{150})$.

  Observed exposure $(E_1, \ldots, E_{150})$: HRT = 0 or 1.

- $(\tau = (2, 0.5), \kappa = (3, 0.5))$ so that :
  - Shorter last-stage for HRT=1 than HRT=0
  - 42% of positive C.

  Simulate $(LS_1, \ldots, LS_n)$ and compute $C_i = LS_i - T_i$ for each case i.

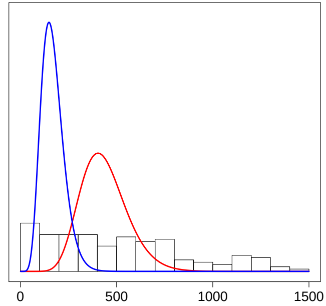
- Simulate $p = 2000$ genes. (β_0^g, β_1^g) sampled from standard gaussian distribution, (σ_g) sampled from chi2 distribution.

  (β_2^g) sampled from $\mathcal{N}(0, 0.01)$ for $g0 = 20$ genes, and 0 for the other genes.

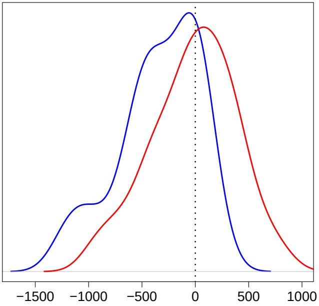
  Simulate G.

Description of the simulated data

Last-stage length distribution

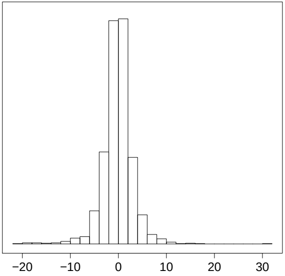


Distribution of the C_i's

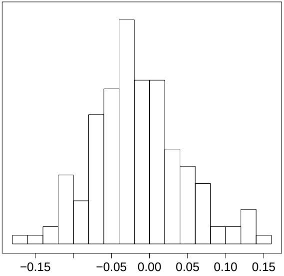


Red : HRT = 0
Blue : HRT = 1

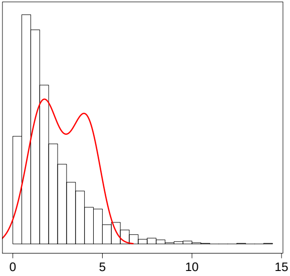
Distribution of all genes for an individual



Distribution of the gene means

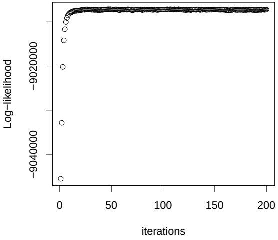


Distribution of the gene sd (red : genes involved in carcinogenesis)

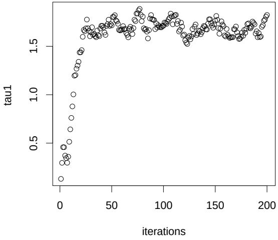


Algorithm convergence

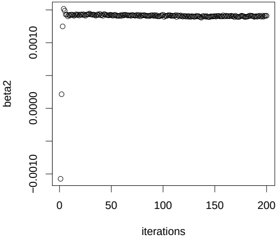
Likelihood



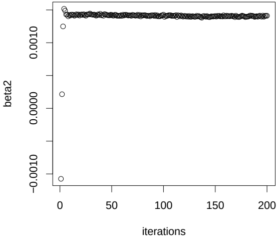
LS distribution parameter (τ_1)



β_2^g (carcinogenesis gene)



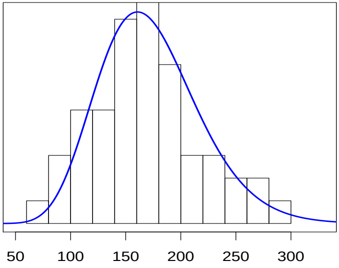
β_2^g (no carcinogenesis gene)



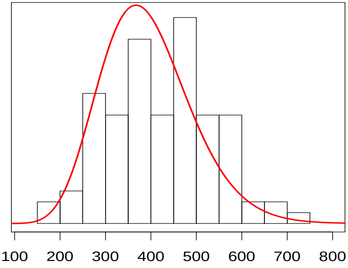
Last stage

- Histogram of the last stage LS.
- Estimated last stage density (Gamma distribution with estimated parameters) : solid line.

$HRT = 1$

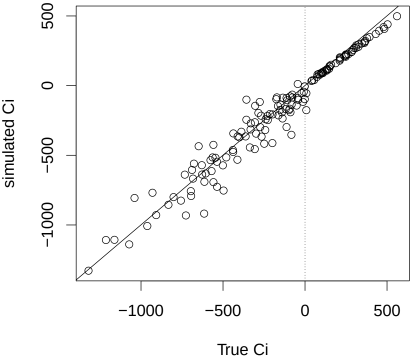


$HRT = 0$

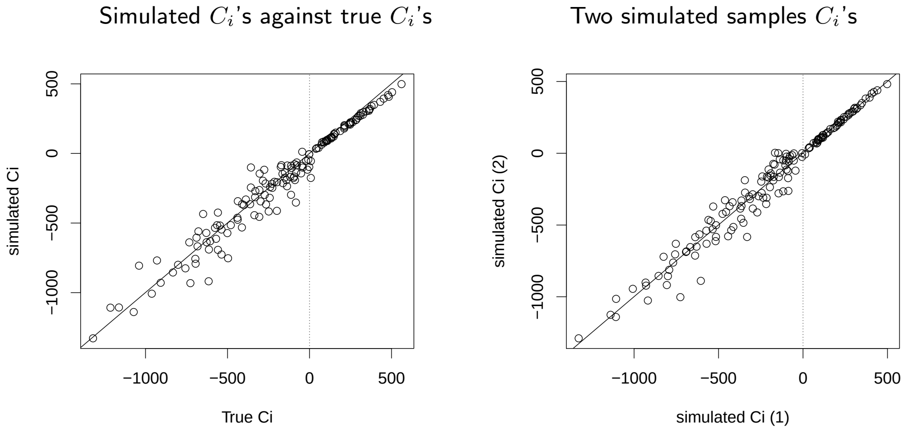


- $\widehat{\Theta}$ estimated parameters.
- C_i simulated from $\mathbb{P}_{\widehat{\Theta}}[C_i|G_i, E_i, T_i]$.

Simulated C_i's against true C_i's



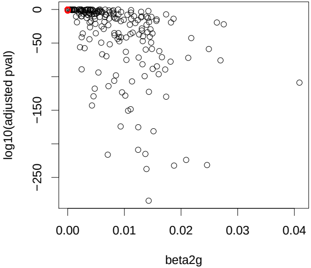
- $\widehat{\Theta}$ estimated parameters.
- C_i simulated from $\mathbb{P}_{\widehat{\Theta}}[C_i|G_i, E_i, T_i]$.



Gene detection : multiple testing

- t-test : $\beta_2^g = 0$.
- Multiple testing procedure : (Benjamini-Hochberg).

log10-adjusted p-value against $|\beta_2^g|$



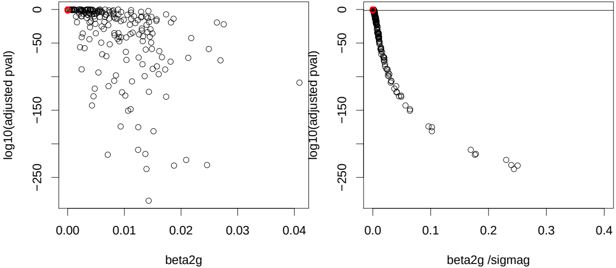
Black : 200 genes involved in carcinogenesis
Red : 1800 genes not involved in carcinogenesis

Gene detection : multiple testing

- t-test : $\beta_2^g = 0$.
- Multiple testing procedure : (Benjamini-Hochberg).

log10-adjusted p-value against $|\beta_2^g|$

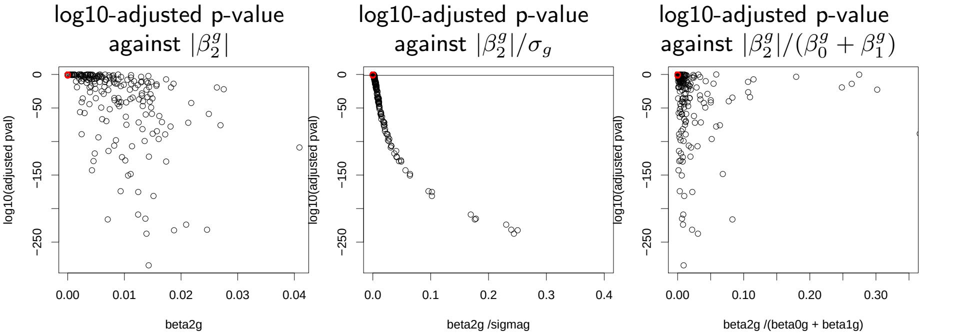
log10-adjusted p-value against $|\beta_2^g|/\sigma_g$



Black : 200 genes involved in carcinogenesis
Red : 1800 genes not involved in carcinogenesis

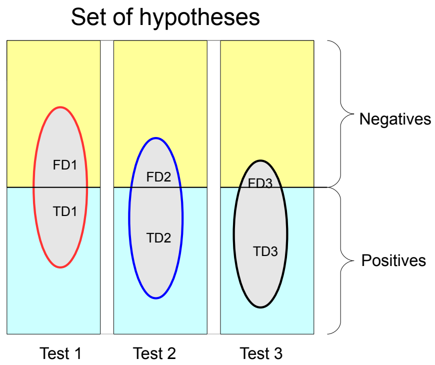
Gene detection : multiple testing

- t-test : $\beta_2^g = 0$.
- Multiple testing procedure : (Benjamini-Hochberg).



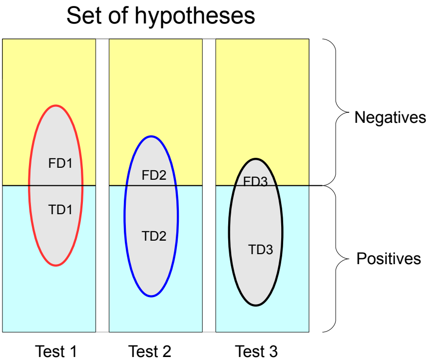
Black : 200 genes involved in carcinogenesis
Red : 1800 genes not involved in carcinogenesis

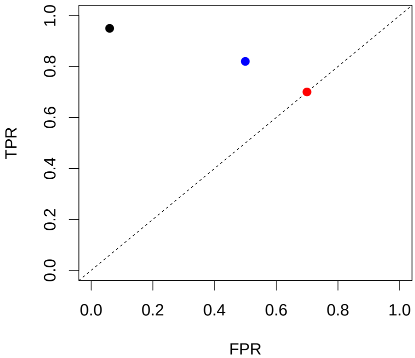
ROC curve



- FD : False Discovery.
- TP : True Discovery.
- FPR : False Positive Rate (proportion of false dicoveries out of the negatives).
- TPR : False Positive Rate (proportion of true dicoveries out of the positives).

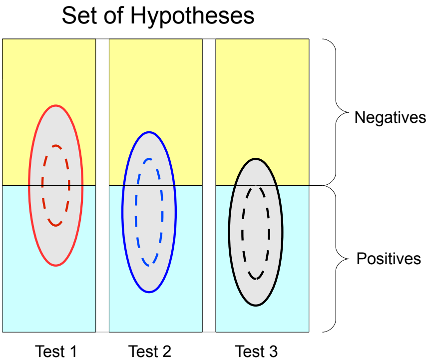
ROC curve

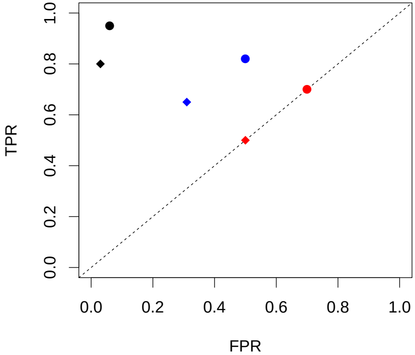




- FD : False Discovery.
- TP : True Discovery.
- FPR : False Positive Rate (proportion of false dicoveries out of the negatives).
- TPR : False Positive Rate (proportion of true dicoveries out of the positives).

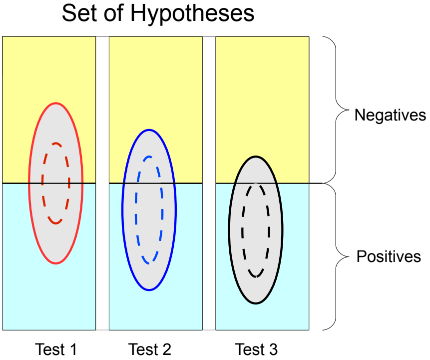
ROC curve

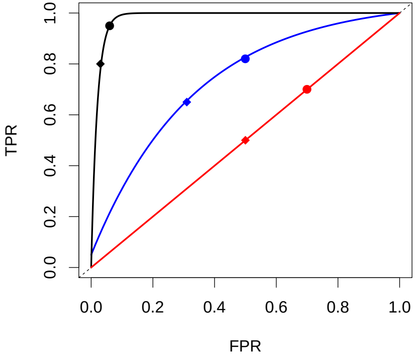




- FD : False Discovery.
- TP : True Discovery.
- FPR : False Positive Rate (proportion of false dicoveries out of the negatives).
- TPR : False Positive Rate (proportion of true dicoveries out of the positives).

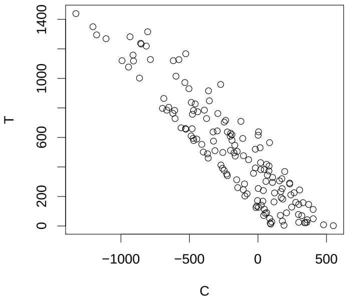
ROC curve





- FD : False Discovery.
- TP : True Discovery.
- FPR : False Positive Rate (proportion of false dicoveries out of the negatives).
- TPR : False Positive Rate (proportion of true dicoveries out of the positives).

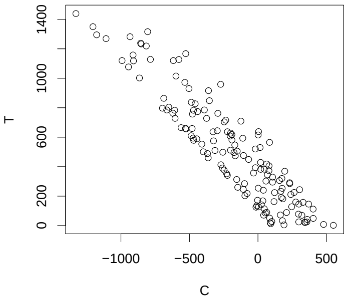
Correlation between C and T

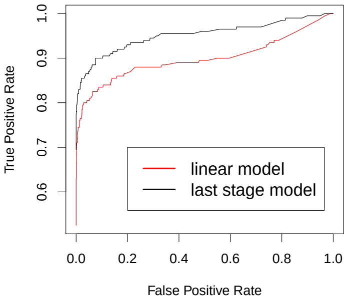


Simple linear model : for every gene g, $G_i^g = \langle \beta^g, (1, E_i, T_i) \rangle + \varepsilon_i^g$.

- Lower sensitivity.
- No estimation about the last stage.

Correlation between C and T





Simple linear model : for every gene g, $G_i^g = \langle \beta^g, (1, E_i, T_i) \rangle + \varepsilon_i^g$.

- Lower sensitivity.
- No estimation about the last stage.

Classical approach on prospective studies : Cox model

- ΔG_i : difference in gene expression between a case and a control.
- Survival model : hazard rate of T_i given (E_i, G_i)

$$\lambda(t|E_i, G_i) = \lambda_0(t)\exp(\langle\alpha, (E_i, G_i)\rangle).$$

◇ Principle : if $\alpha_g \neq 0$, gene g is involved in carcinogenesis.

- Partial likelihood in a case-control study :

$$\mathcal{L}(\alpha) = \prod_{i=1}^{n} \frac{1}{1 + \exp(-\langle\alpha, (\Delta E_i, \Delta G_i)\rangle)}$$

◇ Follow-up time not considered.

- Penalized Cox model leads to bias in the estimation.
- Gene-by-gene model : for every gene g,

$$\lambda(t|E_i, G_i^g) = \lambda_0(t)\exp(\langle\alpha, (E_i, G_i^g)\rangle)$$

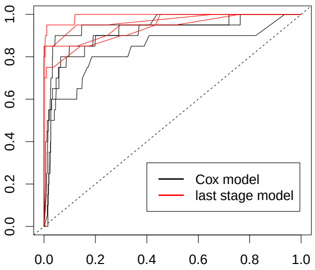
.

Results with Cox on our simulated data.

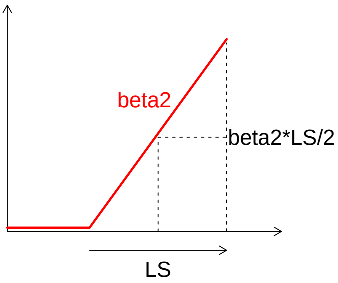
- ΔG_i^g case-control difference : no offset for "ideal" data.
- Simulations with $\beta_g^0 = 0$ for all genes g

Results with Cox on our simulated data.

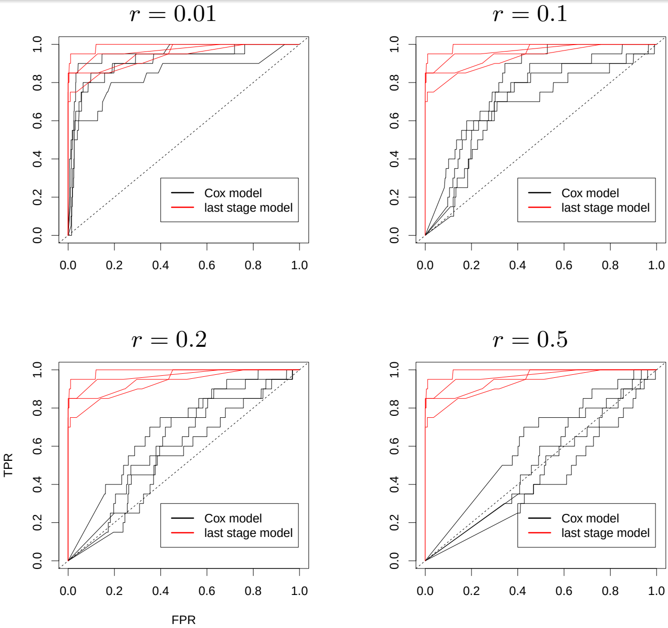
- ΔG_i^g case-control difference : no offset for "ideal" data.
- Simulations with $\beta_g^0 = 0$ for all genes g



- Bias between cases and controls : $\beta_g^0 \neq 0$
- $(\beta_1^0, \ldots, \beta_p^0)$ sampled from $\mathcal{N}(0, \sigma_0)$.
- Comparison between offset (β_0^g) and time effect (β_2^g).



$$r = \frac{\sigma_0}{\sigma_2 * \textbf{mean}(LS)/2}$$


$r = 0.01$
$r = 0.1$
$r = 0.2$
$r = 0.5$
Cox model
last stage model
TPR
FPR
0.0
0.2
0.4
0.6
0.8
1.0

Conclusion and perspectives

- Goal : detect genes involved in the carcinogenesis last-stage.
- Conceptual model based on biological carcinogenesis modeling.
- On simple simulated data, the standard method fail to detect the specific genes.
  - More sophisticated Cox ?
  - Evolve our model ?
- Require further developments to be applied on data.
  - Epidemiology : choice of exposures, type of cancers...
  - Statistics : dependence between genes (a priori knowledge / statistical inference)
- Validation on non parametric simulations.