Corrélations : explications et limites

Marie-Laure Martin-Magniette

Institut des Sciences des Plantes de Paris-Saclay (IPS2)

Responsable de l'équipe Réseaux Génomiques

MIA-Paris à l'AgroParisTech

Membre de l'équipe Statistique and Genome

Corrélations : explications et limites

Marie-Laure Martin-Magniette

Institut des Sciences des Plantes de Paris-Saclay (IPS2)

Responsable de l'équipe Réseaux Génomiques

MIA-Paris à l'AgroParisTech

Membre de l'équipe Statistique and Genome

Remerciements : Guillem Rigaill pour les simulations et G. Saporta pour son livre

La corrélation de Pearson

La corrélation de Pearson entre X et Y est définie par

$$-1 \leq \frac{Cov(X,Y)}{\sigma(X)\sigma(Y)} \leq 1$$

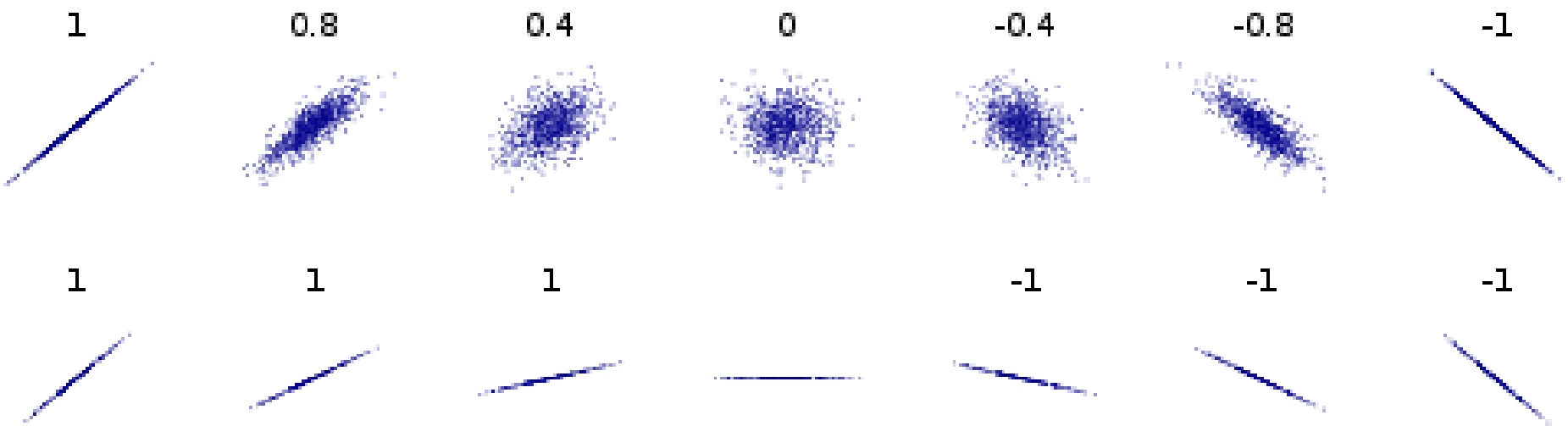
Elle est estimée par

$$\frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{s_x s_y} \text{ où } s_x = \frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2$$

Ce coefficient mesure exclusivement le caractère plus ou moins linéaire du nuage de points

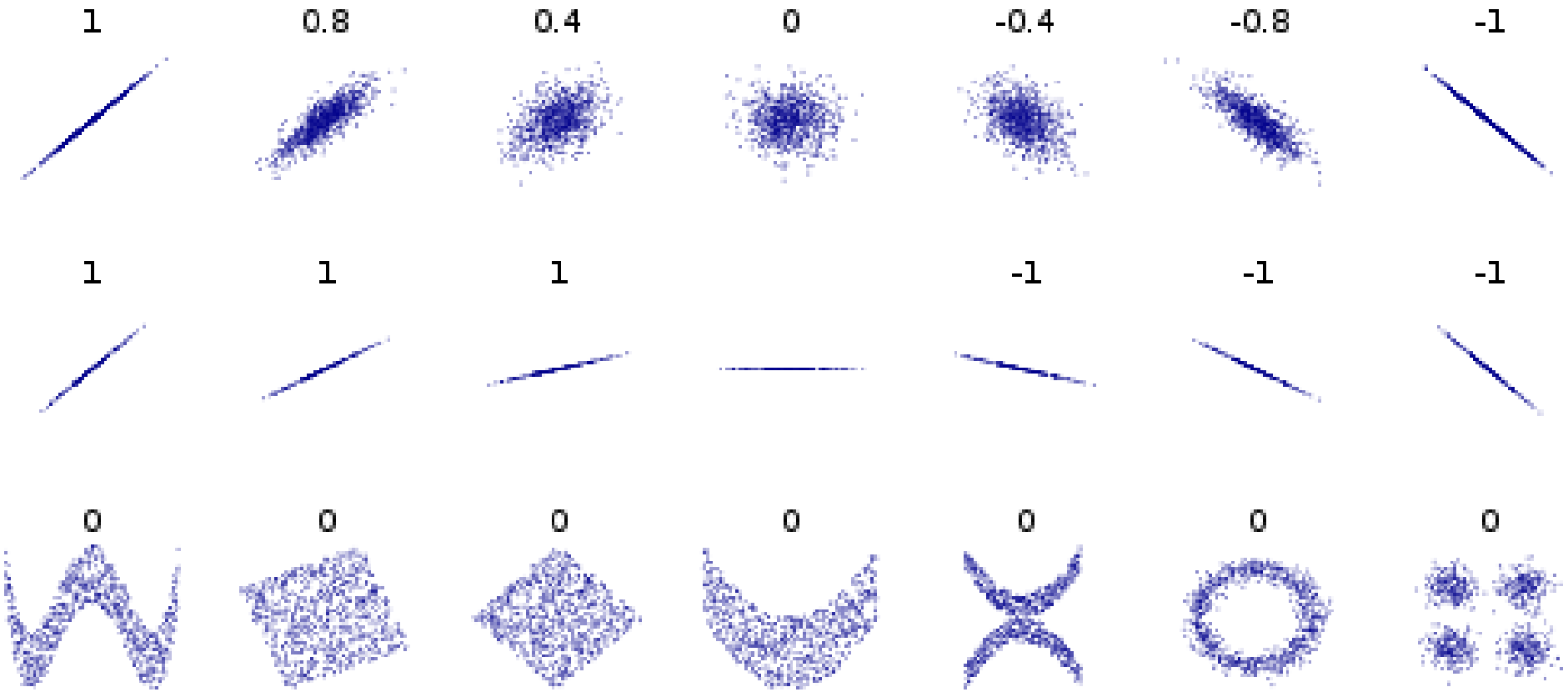
Signification de la corrélation

- C'est une notion de liaison qui contredit leur indépendance.
- La valeur absolue du coefficient mesure la prépondérance de la relation affine sur les variations internes des variables.



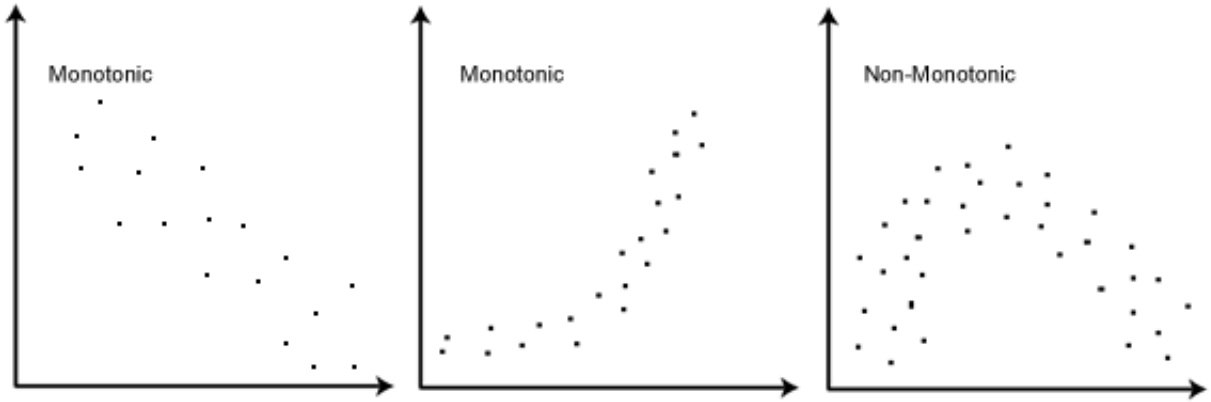
Signification de la corrélation

Si 2 variables sont indépendantes, alors la corrélation est nulle
La réciproque est fausse



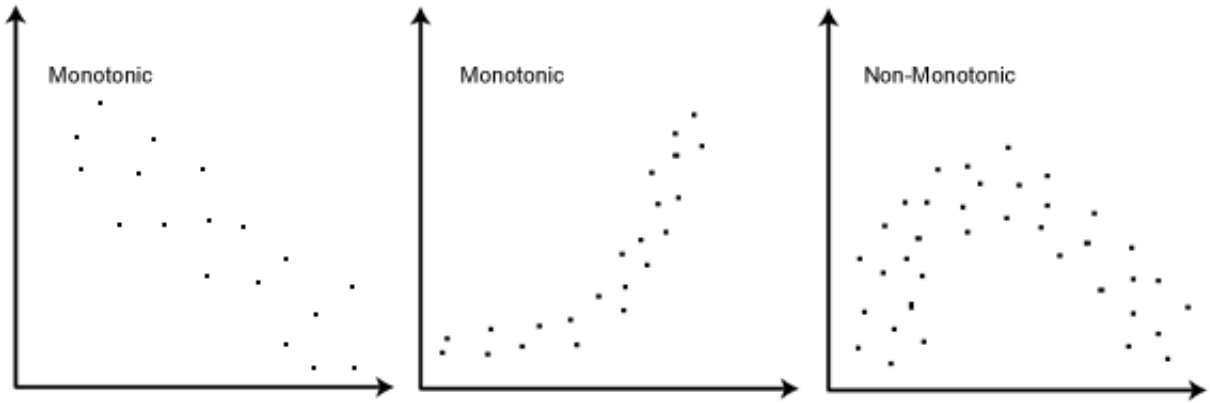
Corrélation de Spearman

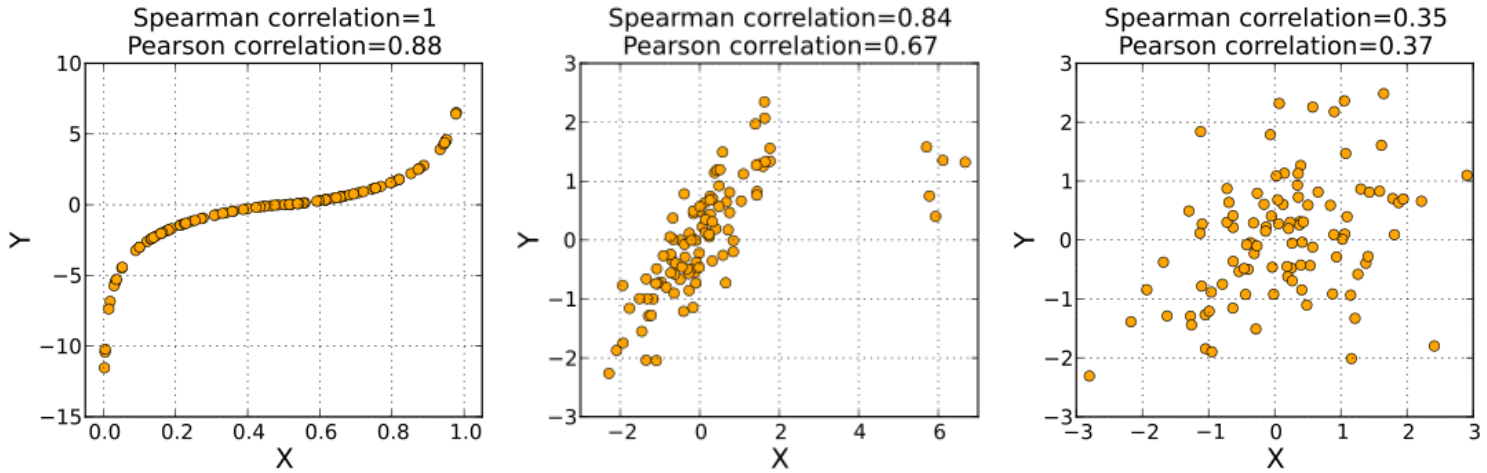
- C'est la corrélation de Pearson calculée sur le rang des valeurs
- Mesure une relation monotone (linéaire ou non)



Corrélation de Spearman

- C'est la corrélation de Pearson calculée sur le rang des valeurs
- Mesure une relation monotone (linéaire ou non)





Test de corrélation

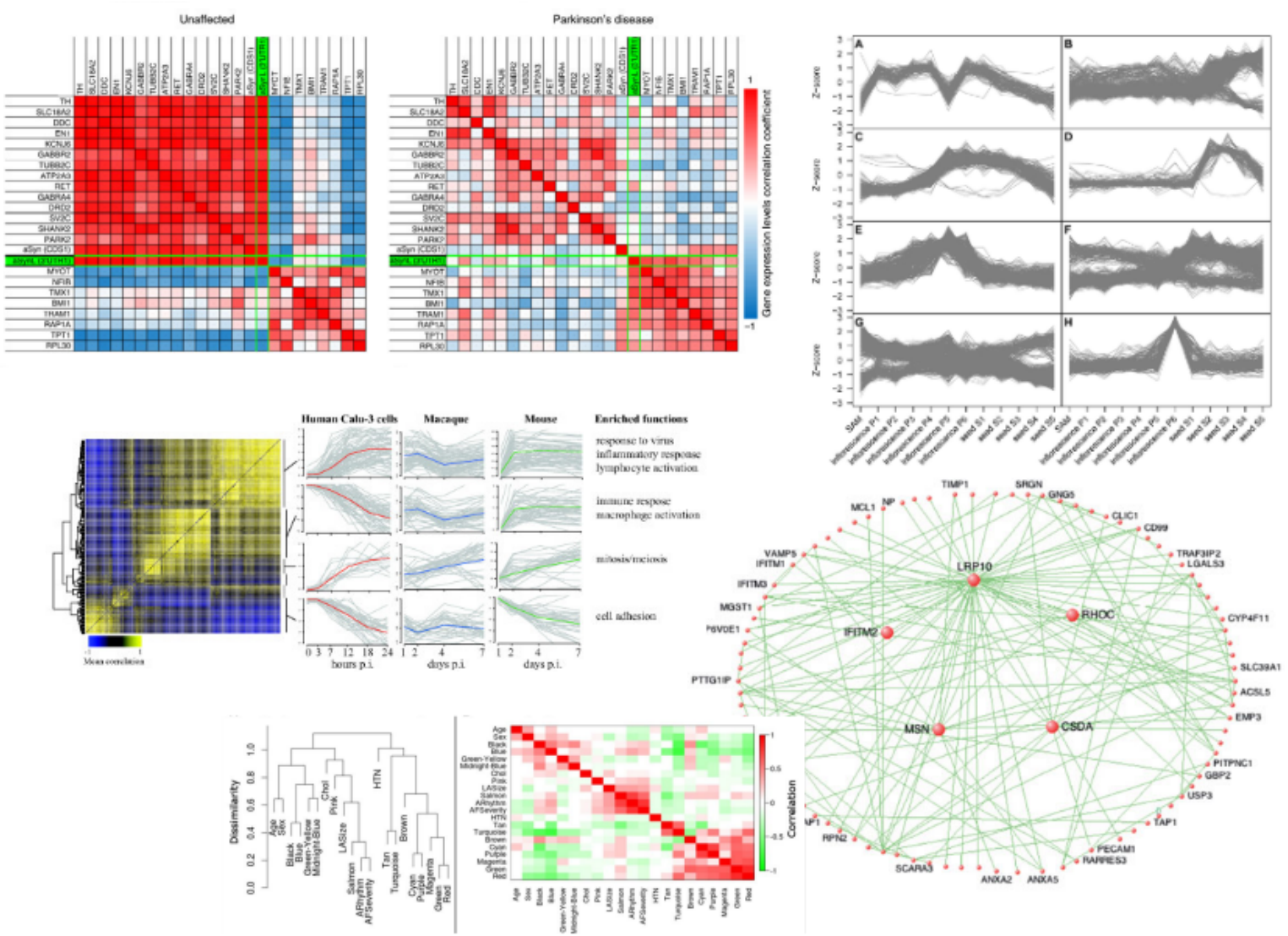
$$H_0 = \{R = 0\} \text{ versus } H_1 = \{R \neq 0\}$$

Si les n observations sont prélevées au hasard dans une population où les 2 variables sont gaussiennes et indépendantes, alors

$$\sqrt{n-2}\frac{R}{\sqrt{1-R^2}} \sim T_{n-2}$$

- Au niveau 5%, on déclare une liaison significative sur un échantillon de 30 observations si $|r| > 0.36$
- Le seuil décroît quand n croît
- Trouver que r diffère significativement de 0 ne garantit nullement que la liaison soit forte
- Si le couple n'est pas gaussien, ces remarques restent vraies si n grand
- Mais accepter H_0 n'entraine pas nécessairement l'indépendance

Co-expression de gènes [1]



[1] Google image search: “Coexpression”

Il y a co-expression quand il y a ...

- une expression simultanée de deux gènes ou plus[2]
- des groupes de gènes co-transcrits [3]
- une similarité d'expression[4] (corrélation, chevauchement topologique, information mutuelle, ...)
- des groupes de gènes avec un pattern d'expression similaire[5] sur un grand nombre d'expériences

[2] https://en.wiktionary.org/wiki/coexpression
[3] http://bioinfow.dep.usal.es/coexpression
[4] http://coxpresdb.jp/overview.shtml
[5] Yeung *et al.* (2001)
[6] Eisen *et al.* (1998)

Il y a co-expression quand il y a ...

- une expression simultanée de deux gènes ou plus[2]
- des groupes de gènes co-transcrits [3]
- une similarité d'expression[4] (corrélation, chevauchement topologique, information mutuelle, ...)
- des groupes de gènes avec un pattern d'expression similaire[5] sur un grand nombre d'expériences

Pourquoi étudier la co-expression ?

---

[2] https://en.wiktionary.org/wiki/coexpression
[3] http://bioinfow.dep.usal.es/coexpression
[4] http://coxpresdb.jp/overview.shtml
[5] Yeung *et al.* (2001)
[6] Eisen *et al.* (1998)

Il y a co-expression quand il y a ...

- une expression simultanée de deux gènes ou plus[2]
- des groupes de gènes co-transcrits [3]
- une similarité d'expression[4] (corrélation, chevauchement topologique, information mutuelle, ...)
- des groupes de gènes avec un pattern d'expression similaire[5] sur un grand nombre d'expériences

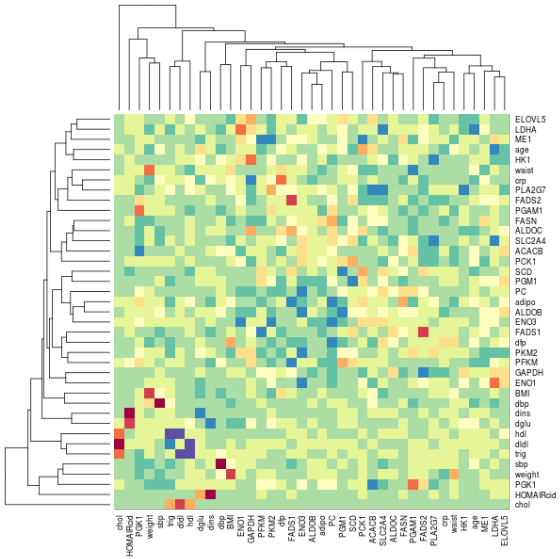
Pourquoi étudier la co-expression ?
car c'est une manière d'identifier des processus biologiques[6]

---

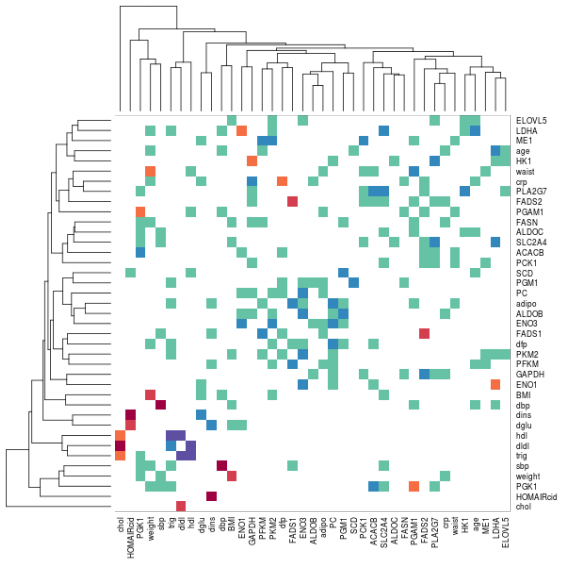
[2] https://en.wiktionary.org/wiki/coexpression
[3] http://bioinfow.dep.usal.es/coexpression
[4] http://coxpresdb.jp/overview.shtml
[5] Yeung *et al.* (2001)
[6] Eisen *et al.* (1998)

Utilisation de la *correlation*: relevance network [7]

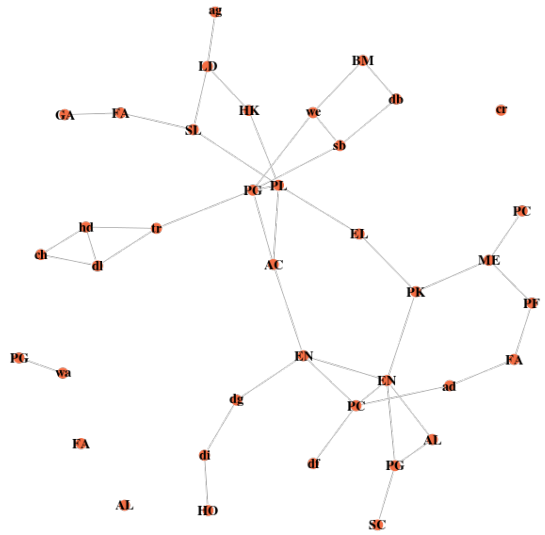
Approche naïve: calculer les corrélations entre les expressions de toutes les paires de gènes, seuiller les plus petites et construire un réseau



Calcul des corrélations



Seuillage



Construction du réseau

[7] Butte and Kohane (1999,2000)

Que signifie une correlation supérieure à ... ?

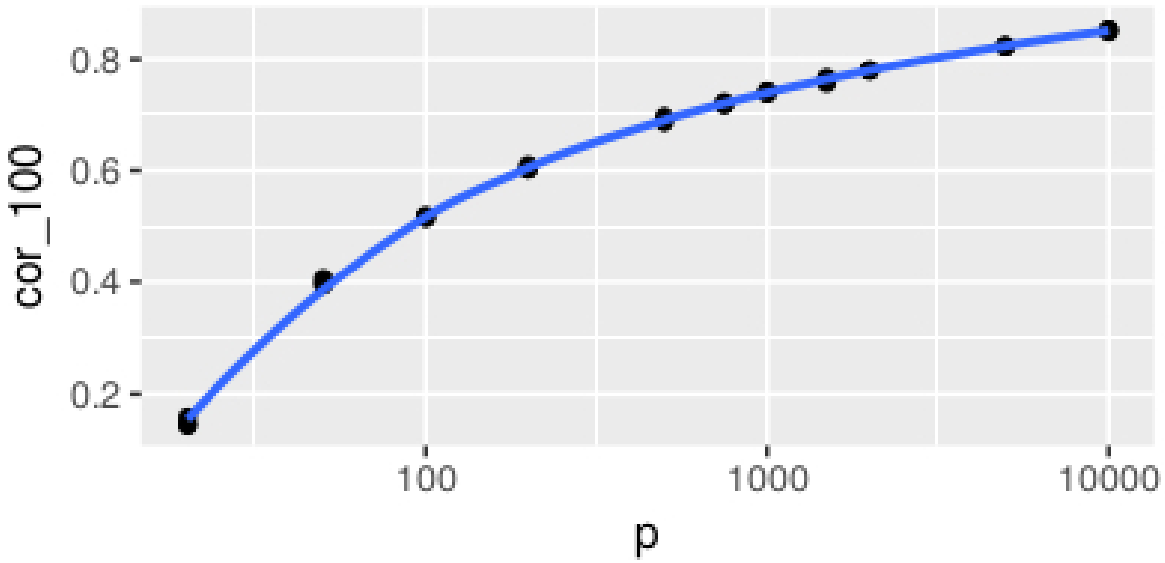
Simulation répétée 100 fois

- 20 expériences avec un nombre croissant de gènes p
- tous les gènes sont indépendants
- calcul de toutes les corrélations entre les génes et enregistrement de la 100 ième valeur la plus forte

Que signifie une correlation supérieure à ... ?

Simulation répétée 100 fois

- 20 expériences avec un nombre croissant de gènes p
- tous les gènes sont indépendants
- calcul de toutes les corrélations entre les génes et enregistrement de la 100 ième valeur la plus forte



Nombre de corrélation supérieure à 0.7

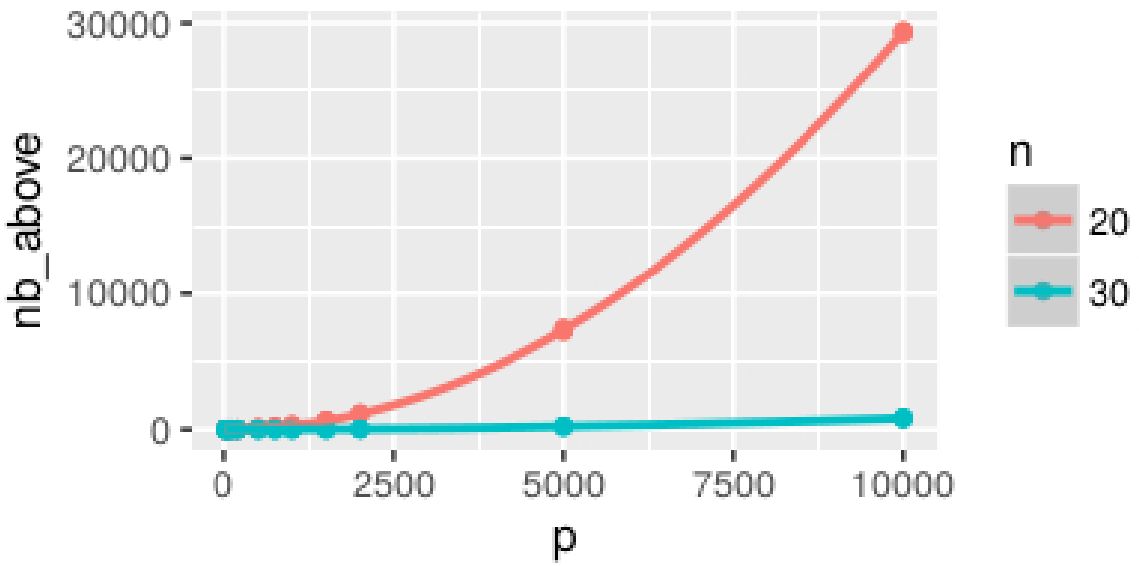
Simulation

- n expériences avec un nombre croissant de gènes p
- tous les gènes sont indépendants

Nombre de corrélation supérieure à 0.7

Simulation

- *n* expériences avec un nombre croissant de gènes *p*
- tous les gènes sont indépendants



Corrélation supérieure à 0.7 dans un contexte particulier

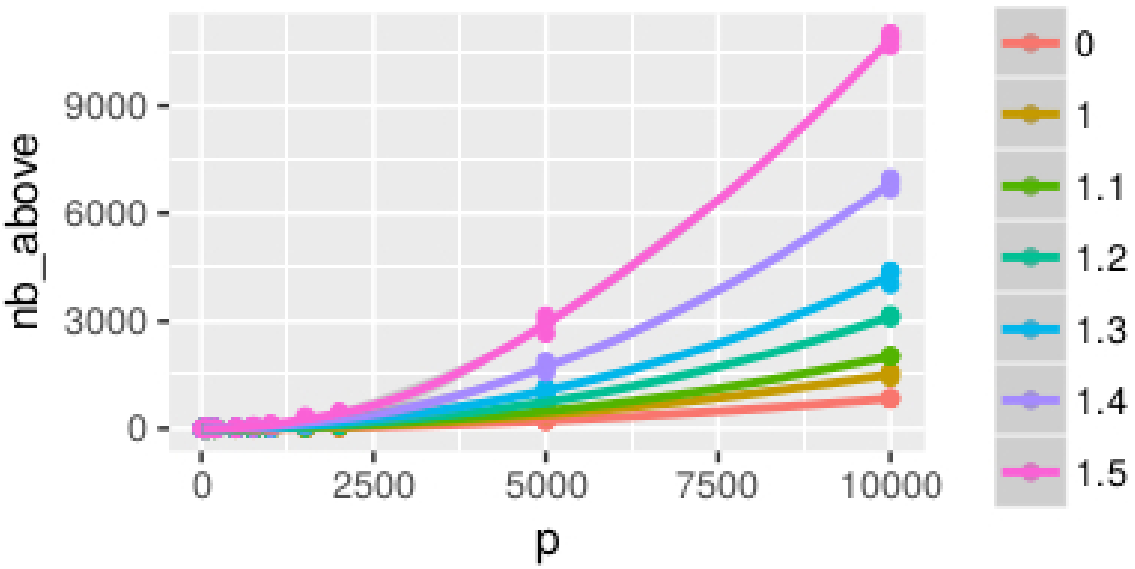
Simulation

- $n = 30$ expériences avec un nombre croissant de gènes p
- les 2 conditions sont décrites chacune avc 15 échantillons
- 20% des gènes on une différence d'expression entre les 2 conditions

Corrélation supérieure à 0.7 dans un contexte particulier

Simulation

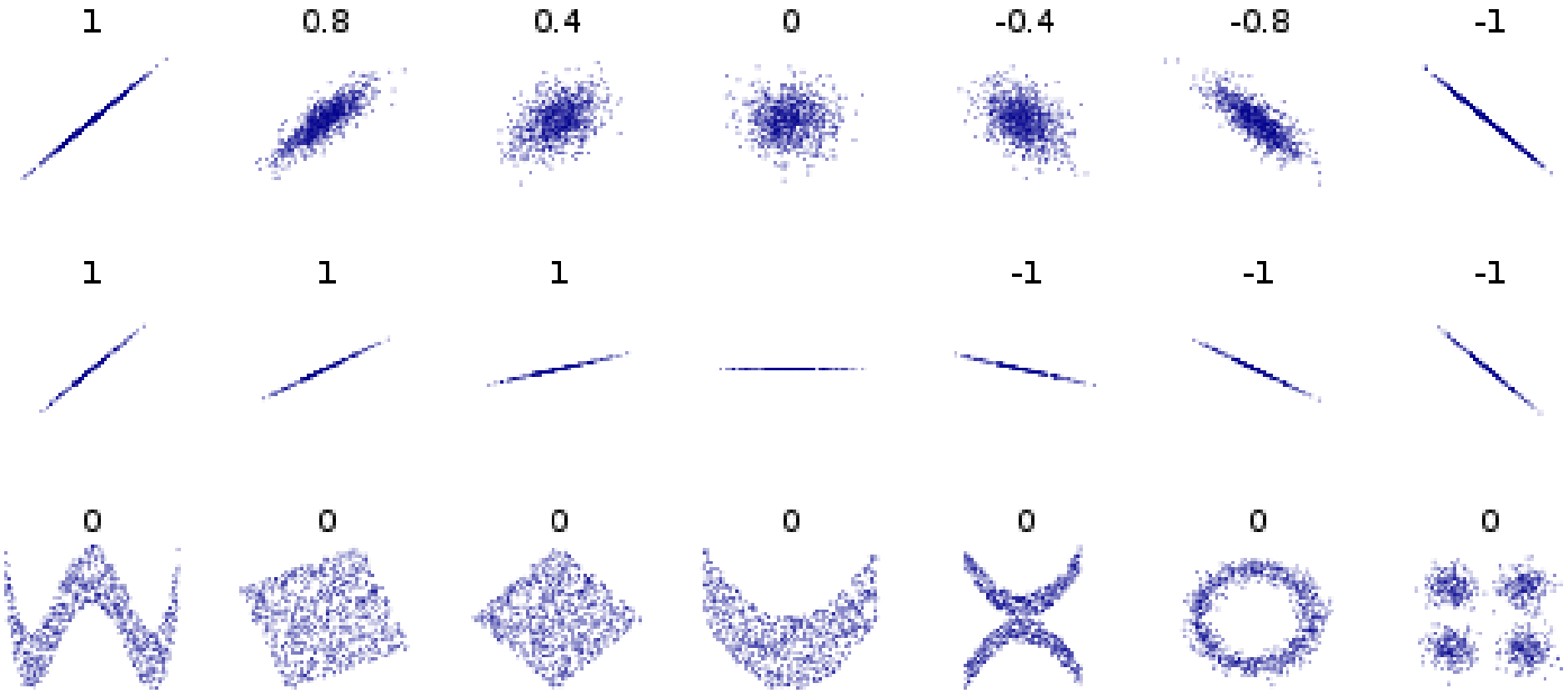
- $n = 30$ expériences avec un nombre croissant de gènes p
- les 2 conditions sont décrites chacune avc 15 échantillons
- 20% des gènes on une différence d'expression entre les 2 conditions



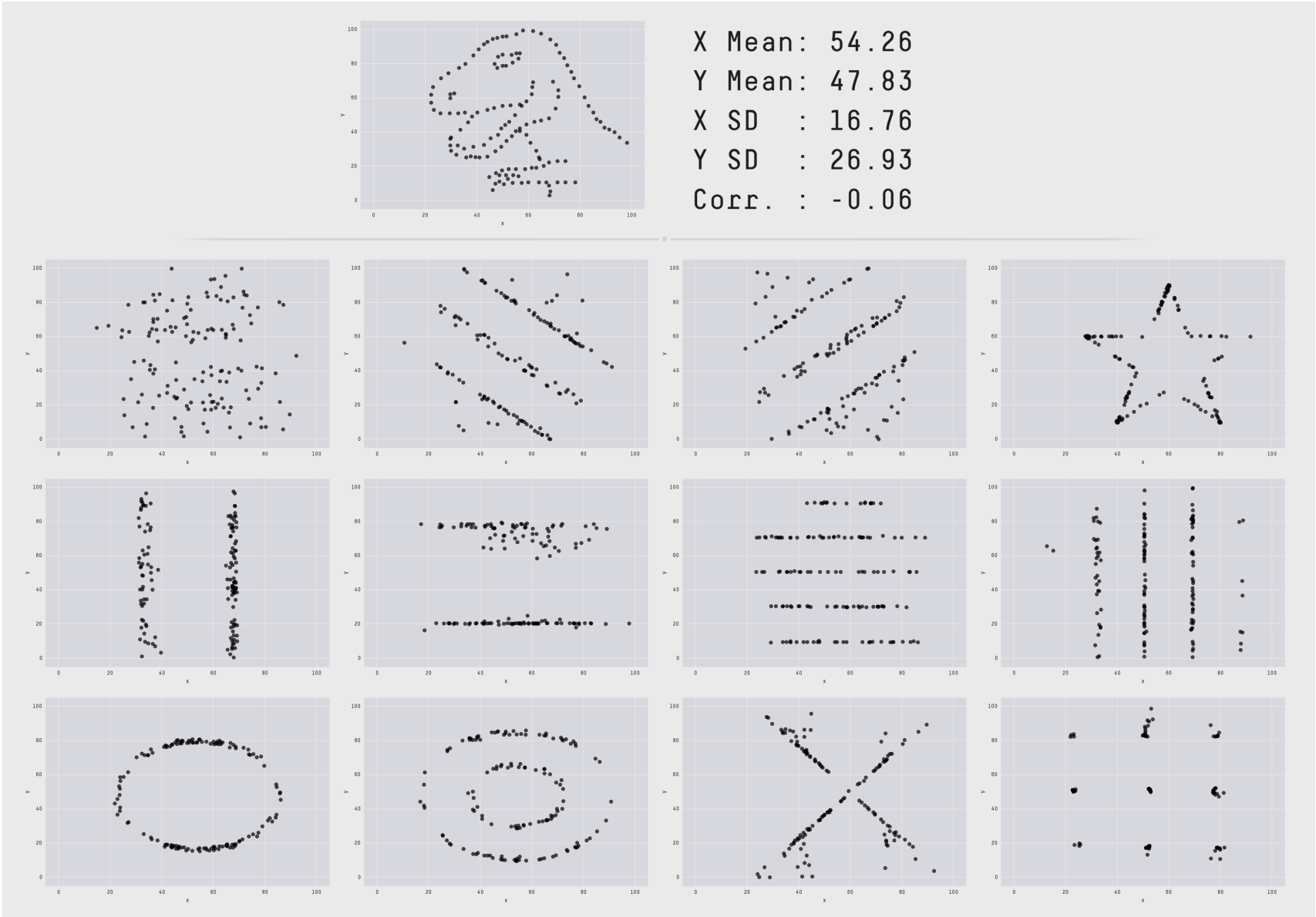
Que signifie une correlation supérieure à ... ?

Limites de la corrélation

- Elle est très simple calculer ... moins interpréter
- En génomique où le nombre d'entités considérées est grand, la corrélation devient réellement ininterprétable

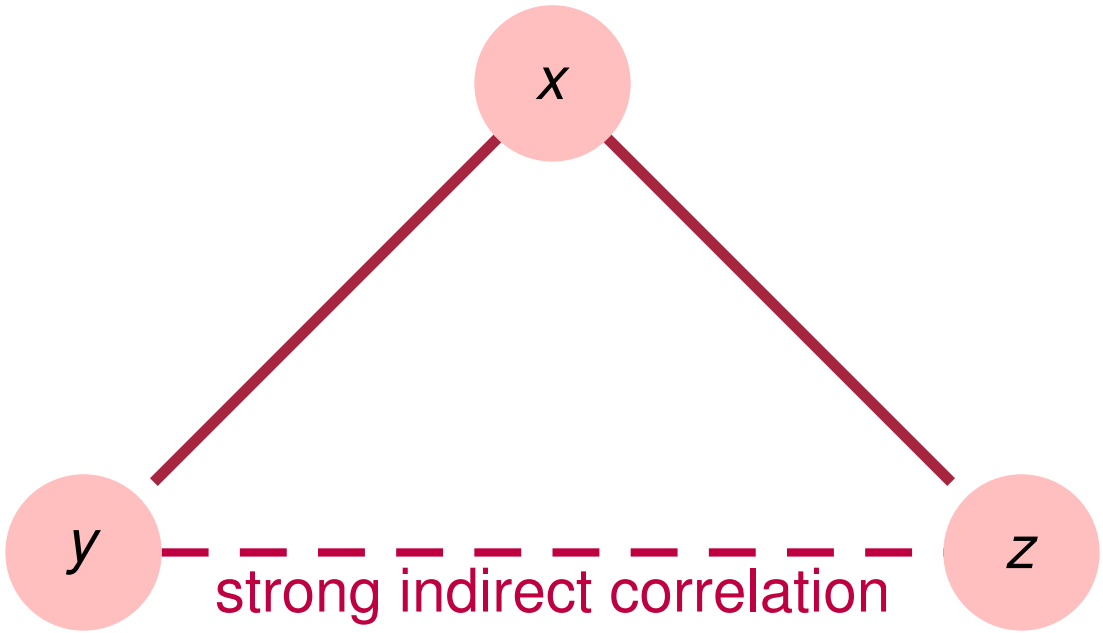


Exemples de corrélations de Pearson



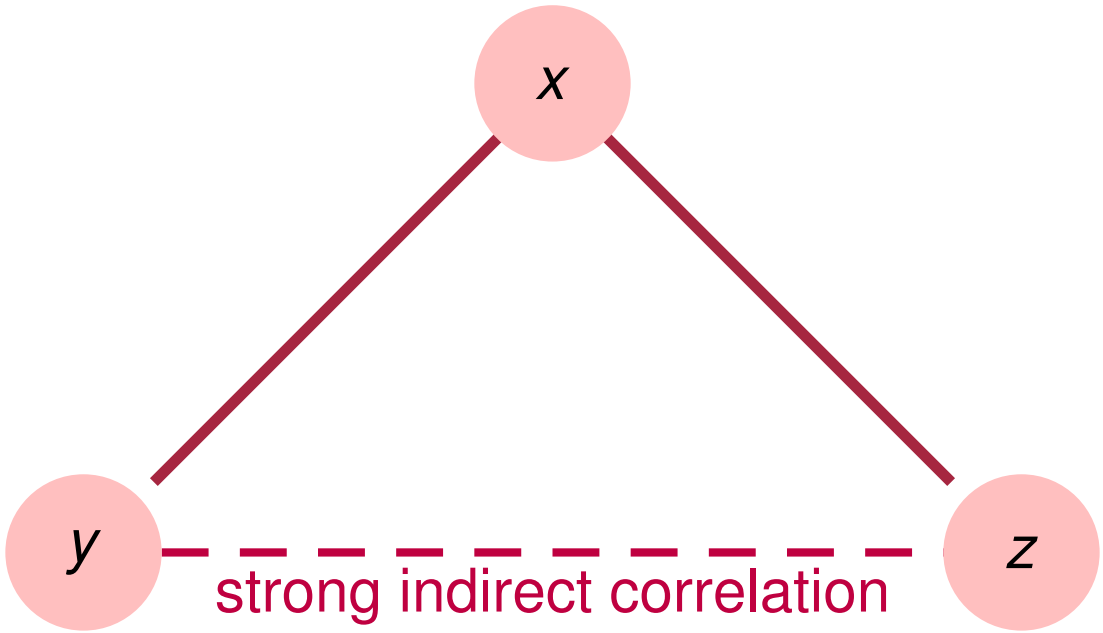
Source: https://www.autodeskresearch.com/publications/samestats

La corrélation partielle




```
set.seed(2807); x <- runif(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y); [1] 0.9988261
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z); [1] 0.998751
cor(y,z); [1] 0.9971105
```

La corrélation partielle




```
set.seed(2807); x <- runif(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y); [1] 0.9988261
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z); [1] 0.998751
cor(y,z); [1] 0.9971105
# Partial correlation
cor(lm(y~x)$residuals,lm(z~x)$residuals) [1] -0.1933699
```

La corrélation partielle

Dans le cas gaussien, la loi du coefficient de corrélation partielle entre 2 variables indépendantes est

$$\sqrt{n-d-2}\frac{R}{\sqrt{1-R^2}} \sim T_{n-d-2}$$

où d est le nombre de variables fixées.

La corrélation partielle

Dans le cas gaussien, la loi du coefficient de corrélation partielle entre 2 variables indépendantes est

$$\sqrt{n-d-2}\frac{R}{\sqrt{1-R^2}} \sim T_{n-d-2}$$

où d est le nombre de variables fixées.

Que signifie une correlation partielle supérieure à ... ?

La corrélation multiple

Soient une variable numérique y et un ensemble de p variables également numériques $x^1, x^2, ...x^p$
Le coefficient de corrélation multiple est alors la valeur maximale prise par le coefficient de corrélation linéaire entre y et une combinaison linéaire des x^j

Cadre gaussien

Soit $(X_i)_{i=1,\dots,n}$ des variables gaussiennes i.i.d.

Equivalence

- Corrélation partielle

$$j \longleftrightarrow j' \text{ (gènes } j \text{ et } j' \text{ sont liés)} \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k \neq j,j'}\right) \neq 0$$

Cadre gaussien

Soit $(X_i)_{i=1,\dots,n}$ des variables gaussiennes i.i.d.

Equivalence

- Corrélation partielle

$$j \longleftrightarrow j' \text{ (gènes } j \text{ et } j' \text{ sont liés)} \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

- Corrélation multiple

$$X^j = \sum_{j'\neq j} \beta_{jj'} X^{j'} + \varepsilon \qquad \beta_{jj'} \neq 0 \Leftrightarrow j \longleftrightarrow j' \text{ (gènes } j \text{ and } j' \text{ sont liés)}$$

Cela revient faire de l'estimation d'un support d'une régression

Cadre gaussien

Soit $(X_i)_{i=1,\dots,n}$ des variables gaussiennes i.i.d.

Equivalence

- Corrélation partielle

$$j \longleftrightarrow j' \text{ (gènes } j \text{ et } j' \text{ sont liés)} \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

- Corrélation multiple

$$X^j = \sum_{j'\neq j} \beta_{jj'} X^{j'} + \varepsilon \qquad \beta_{jj'} \neq 0 \Leftrightarrow j \longleftrightarrow j' \text{ (gènes } j \text{ and } j' \text{ sont liés)}$$

Cela revient faire de l'estimation d'un support d'une régression

Que faire avec la corrélation simple ?

Cadre gaussien

Soit $(X_i)_{i=1,\dots,n}$ des variables gaussiennes i.i.d.

Equivalence

- Corrélation partielle

$$j \longleftrightarrow j' \text{ (gènes } j \text{ et } j' \text{ sont liés)} \Leftrightarrow \mathbb{C}\text{or}\left(X^j, X^{j'} | (X^k)_{k\neq j,j'}\right) \neq 0$$

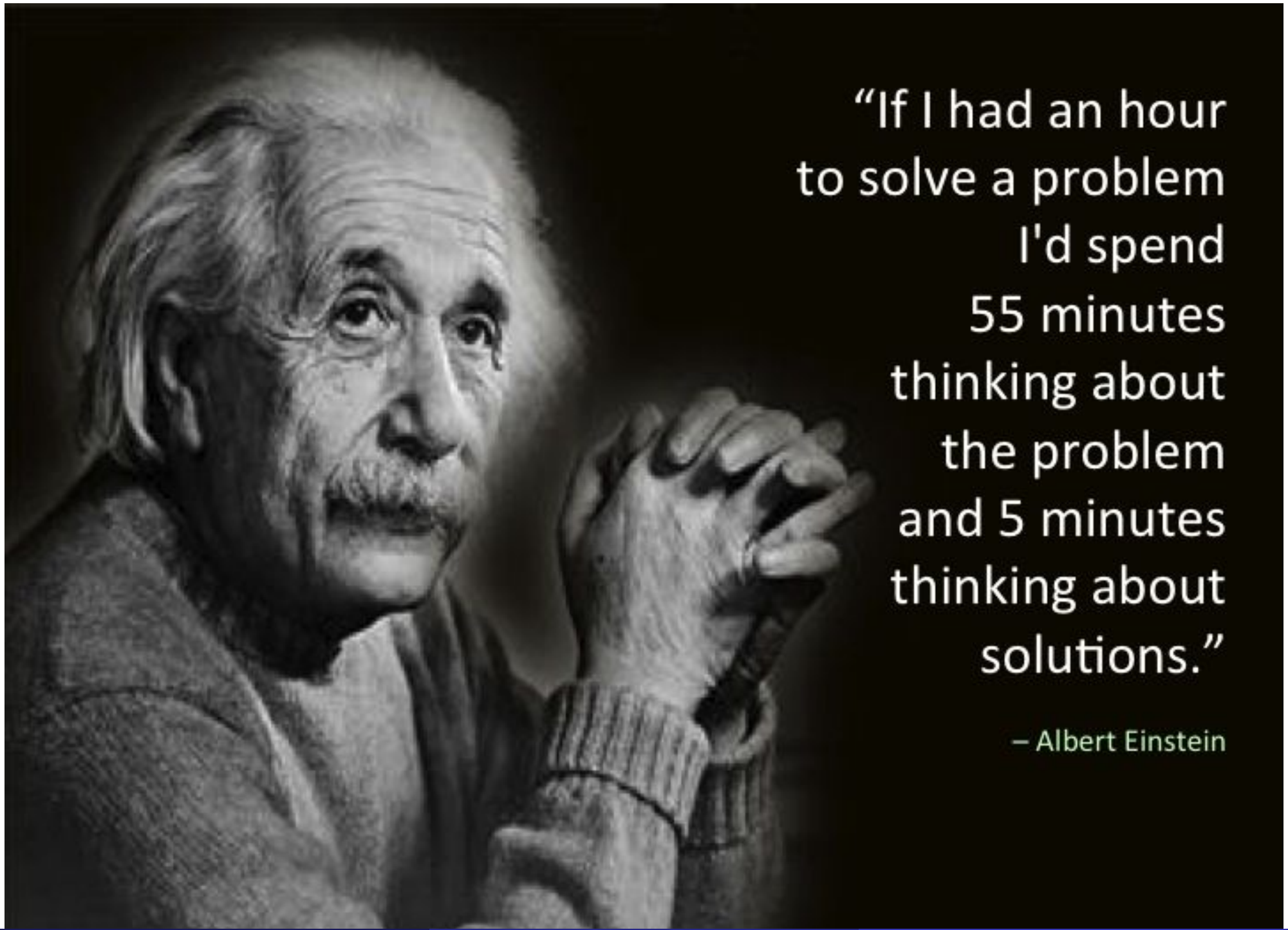
- Corrélation multiple

$$X^j = \sum_{j'\neq j} \beta_{jj'} X^{j'} + \varepsilon \qquad \beta_{jj'} \neq 0 \Leftrightarrow j \longleftrightarrow j' \text{ (gènes } j \text{ and } j' \text{ sont liés)}$$

Cela revient faire de l'estimation d'un support d'une régression

Que faire avec la corrélation simple ?

.... l'abandonner


“If I had an hour
to solve a problem
I'd spend
55 minutes
thinking about
the problem
and 5 minutes
thinking about
solutions.”
– Albert Einstein