MAP 574
Méthodes statistiques pour la biologie

Christophe Giraud

CMAP, Ecole Polytechnique

cours introductif à la statistique

Content

Modélisation: la démarche

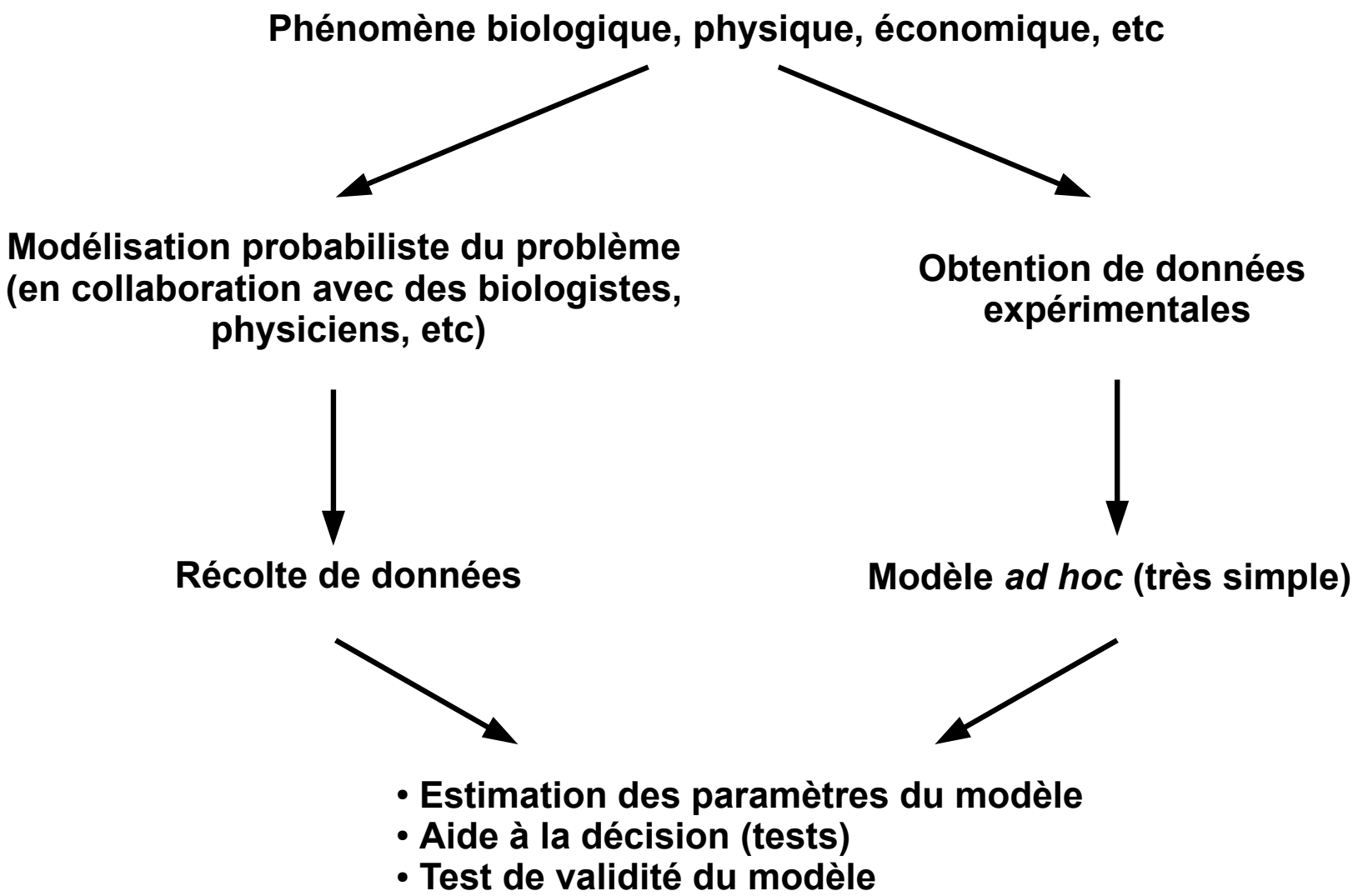
Deux statistiques

Deux approches en statistiques:

- **Statistiques descriptives:** décrire / résumer des données (avec des indices, des graphiques, etc).
  Aucune référence aux probabilités dans cette approche.

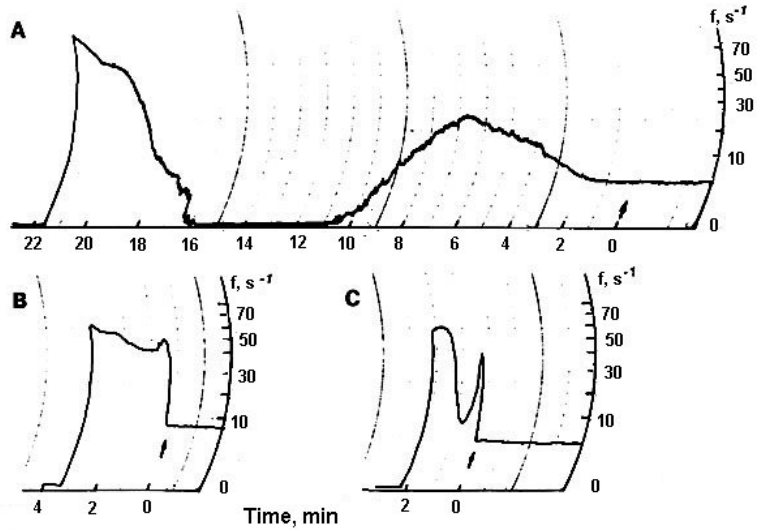
- **Statistiques "probabilistes":** (*model based*) les données sont analysées à travers un modèle probabiliste. Ce modèle est soit construit à partir de connaissances précises du phénomène, soit il est purement *ad hoc*.

Démarche de modélisation "probabiliste"



Exemple

Exemple: réponse neuronale. Deux approches possibles:



- **modèle fin:** pour comprendre le formation du signal. Intérêt pour le modèle (comprendre ses propriétés)
- **modèle simple:** (souvent minimal) $s(t) = f(t) + \varepsilon(t)$ avec f décomposée sur une base d'ondelettes: $f(t) = \sum_{j,k} c_{j,k}\psi_{j,k}(t)$. Intérêt pour la réponse neuronale.

Bon modèle?

Qu'est-ce qu'un bon modèle?

- **Trop simple:** ne parvient pas à décrire (même approximativement) le phénomène étudié.

- **Trop complexe:**
  - l'analyse mathématique est impossible et l'analyse numérique ne donne qu'une vue très partielle du modèle,
  - estimation des paramètres très mauvaise, voir impossible (non indentifiable).

$\longrightarrow$ **Trouver le bon compromis: qualité du modèle / capacité à l'analyser**

Bon modèle?

Un bon modèle n'est pas un modèle exact, mais un modèle qui permet de comprendre des choses!

Problème très important en biologie:
le choix du bon modèle / des bonnes variables

- souvent beaucoup de variables / peu d'observations
- le bon modèle dépend du nombre d'observations!

Statistiques descriptives

Objet: résumer / décrire des données.

Pas de modélisation probabiliste des données!

Données: p variables (Ensoleillement, température, %CO2, %H2O, rendement blé, rendement colza, lattitude, altitude, etc). On a n mesures de ces variables.

$$\text{tableau } n\times p: \quad X = \left[X_i^{(a)}\right]_{\substack{i=1,\dots,n \\ a=1,\dots,p}} = \left[X^{(1)},\dots,X^{(p)}\right] = \begin{bmatrix} X_1^T \\ \vdots \\ X_n^T \end{bmatrix}$$

Description des données

On veut décrire / résumer les données:

- moyenne empirique:

$$\bar{X}^{(a)} = \frac{1}{n} \sum_{i=1}^{n} X_i^{(a)},$$

- variance empirique:

$$\overline{var}\left(X^{(a)}\right) = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i^{(a)} - \bar{X}^{(a)}\right)^2,$$

- etc.

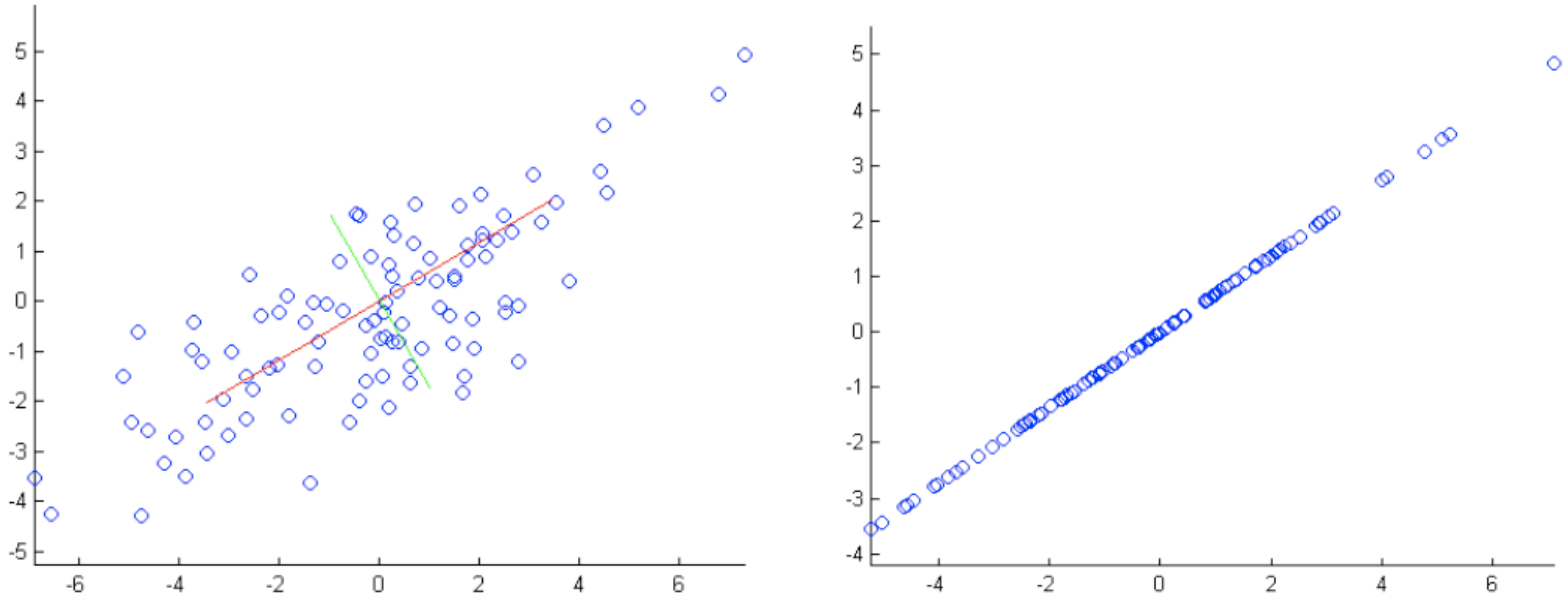
Quelques analyses populaires

- **ACP:** recherche de variables résumant au mieux les données
- **clustering:** recherche de "groupes" de variables "homogènes" dans les données
- et bien d'autres...

1. Analyse en Composantes Principales

ACP (1/13)

But: représenter les obervations $X_i \in \mathbb{R}^p$ dans un espace de plus petite dimension avec le moins de perte d'information possible.



Ex: avec $p = 2$ variables: axes de projections (1er en rouge, 2nd en vert).

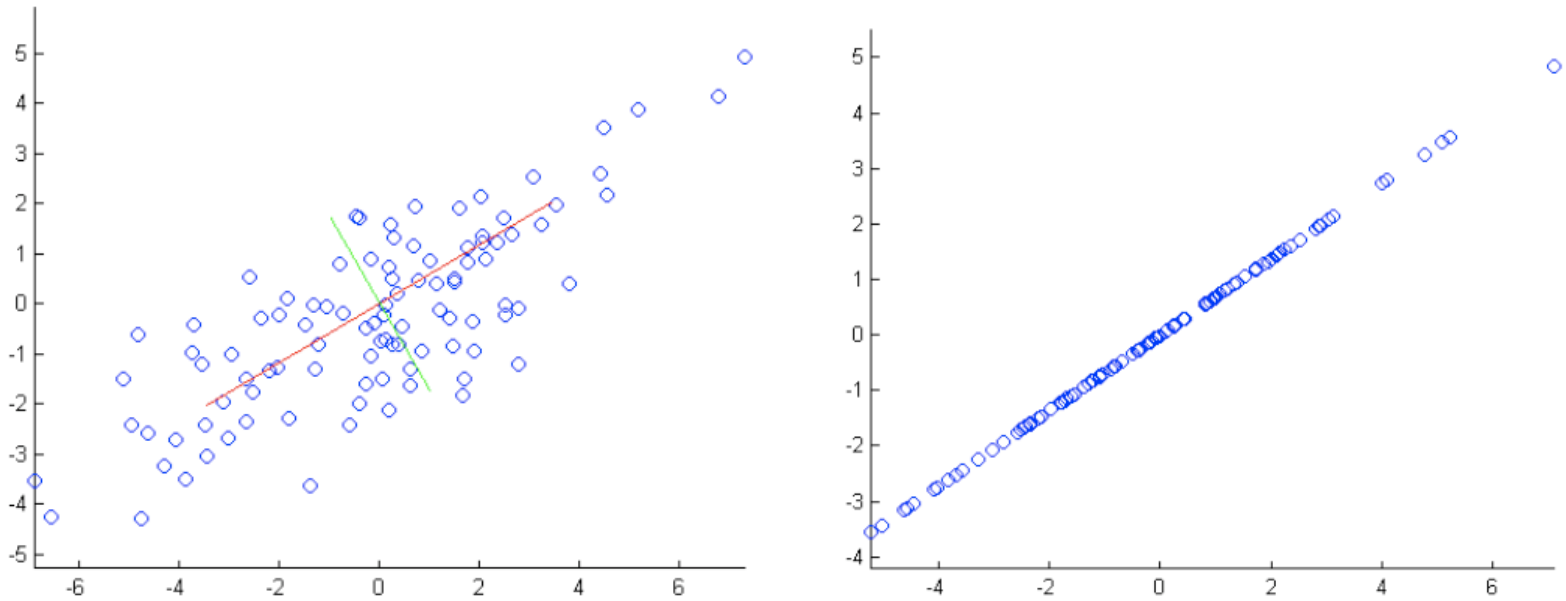
Normalisation: étape préliminaire de normalisation des données:

- centrer: $X^{(a)} \leftarrow X^{(a)} - \bar{X}^{(a)}$
- réduire: $X^{(a)} \leftarrow X^{(a)} / \sqrt{\overline{var}\left(X^{(a)}\right)}$ (recommandé)

ACP (2/13) : premier axe

Recherche d'une composante $C_1 \in \mathbb{R}^n$ expliquant au mieux les données:

$$C_1 = X\beta = \sum_{a=1}^{p} \beta_a X^{(a)} \quad \text{tel que} \begin{cases} \|\beta\| = 1 \\ \\ \overline{var}(X\beta) \text{ maximum} \end{cases}$$



Ex: avec $p = 2$ et $n = 100$: en rouge, le 1er axe $\beta^{(1)}$

Première composante: $C_1 = X\beta^{(1)} \in \mathbb{R}^n$ avec $\beta^{(1)} \in \mathbb{R}^p$ (axe) donné par le problème de maximisation:

$$\begin{aligned} \max_{\|\beta\|=1} \overline{var}(X\beta) &= \max_{\|\beta\|=1} \frac{\|X\beta\|^2}{n-1} \quad \text{(données centrées)} \\ &= \max_{\|\beta\|=1} \beta^T \bar{\Sigma} \beta, \end{aligned}$$

où $\bar{\Sigma} := \frac{1}{n-1} X^T X$ est la matrice $p \times p$ de covariance empirique.

$\Rightarrow \beta^{(1)}$ vecteur propre associé à la plus grande valeur propre λ_1 de $\bar{\Sigma}$.

On a alors $\overline{var}(C_1) = \lambda_1$.

ACP (4/13) : axes suivants

Second axe: $\beta^{(2)} \in \mathbb{R}^p$ solution de

$$\max_{\substack{\beta \perp \beta^{(1)} \\ \|\beta\| = 1}} \overline{var}(X\beta) = \max_{\substack{\beta^T \beta^{(1)} = 0 \\ \|\beta\| = 1}} \beta^T \bar{\Sigma} \beta$$

$\Longrightarrow \beta^{(2)}$ vecteur propre associé à la seconde plus grande valeur propre λ_2 de $\bar{\Sigma}$.

Remarque: on a $\beta^{(2)} \perp \beta^{(1)}$ et aussi $C_2 \perp C_1$.

Axes suivants: idem.

ACP (5/13): résultat

Résultat: après q iterations on obtient q composantes orthogonales $C_1, \ldots, C_q \in \mathbb{R}^n$ (composantes principales) données par $C_k = X\beta^{(k)}$ où $\beta^{(1)}, \ldots, \beta^{(q)} \in \mathbb{R}^p$ (axes principaux) sont les vecteurs propres associés aux q plus grandes valeurs propres de $\bar{\Sigma}$.

- $\beta^{(1)} \perp \ldots \perp \beta^{(q)} \in \mathbb{R}^p$
- $C_1 \perp \ldots \perp C_q \in \mathbb{R}^n$
- $\|\beta^{(k)}\| = 1$ et $\overline{var}(C_k) = \lambda_k$
- $\mathrm{Proj}_{<\beta^{(1)},\ldots,\beta^{(q)}>}(X_i) = \sum_{k=1}^{q} (X_i^T \beta^{(k)})\beta^{(k)} \in \mathbb{R}^p$
- $\mathrm{Proj}_{<C_1,\ldots,C_q>}(X^{(a)}) = \sum_{k=1}^{q} \beta_a^{(k)} C_k \in \mathbb{R}^n$

(car $\mathrm{Proj}_{<C_1,\ldots,C_q>} = \frac{1}{n-1} \sum_k \lambda_k^{-1} C_k C_k^T$)

ACP (6/13): réduction de dimension

Réduction de dimension: les observations X_i sont projetées sur $< \beta^{(1)}, \ldots, \beta^{(q)} >$.

Combien d'axes faut-il choisir?

- Indice de qualité de la représentation: $\frac{\lambda_1 + \ldots + \lambda_q}{\lambda_1 + \ldots + \lambda_p}$ donne le taux de variance expliquée par les q premiers axes.
- Si l'objectif est une représentation visuelle des données: prendre $q = 2$ ou 3.

Interprétation? recherche de groupes de variables, utilisations des composantes principales comme indice agrégé, etc.

ACP (7/13): exemple

Exemple: budget de l'état français sur 24 années.

Les variables: part du budget alloué à différents postes (en pourcentage du budget)

PVP:	Pouvoirs publics	AGR:	Agriculture
CMI:	Commerce et industrie	TRA:	Travail
LOG:	Logement	EDU:	Éducation
ACS:	Action sociale	ANC:	Ancien combattants
DEF:	Défense	DET:	Remboursement dette
DIV:	Divers		

donc $p = 11$

Observations: on a 24 observations pour chaque variable ($n = 24$)

ACP (8/13): exemple

Budgets de l'état de 1872 à 1971 : ACP normée

Donnees brutes

OBS	AN	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
1	1872	18.0	0.5	0.1	6.7	0.5	2.1	2.0	0.0	26.4	41.5	2.1
2	1880	14.1	0.8	0.1	15.3	1.9	3.7	0.5	0.0	29.8	31.3	2.5
3	1890	13.6	0.7	0.7	6.8	0.6	7.1	0.7	0.0	33.8	34.4	1.7
4	1900	14.3	1.7	1.7	6.9	1.2	7.4	0.8	0.0	37.7	26.2	2.2
5	1903	10.3	1.5	0.4	9.3	0.6	8.5	0.9	0.0	38.4	27.2	3.0
6	1906	13.4	1.4	0.5	8.1	0.7	8.6	1.8	0.0	38.5	25.3	1.9
7	1909	13.5	1.1	0.5	9.0	0.6	9.0	3.4	0.0	36.8	23.5	2.6
8	1912	12.9	1.4	0.3	9.4	0.6	9.3	4.3	0.0	41.1	19.4	1.3
9	1920	12.3	0.3	0.1	11.9	2.4	3.7	1.7	1.9	42.4	23.1	0.2
10	1923	7.6	1.2	3.2	5.1	0.6	5.6	1.8	10.0	29.0	35.0	0.9
11	1926	10.5	0.3	0.4	4.5	1.8	6.6	2.1	10.1	19.9	41.6	2.3
12	1929	10.0	0.6	0.6	9.0	1.0	8.1	3.2	11.8	28.0	25.8	2.0
13	1932	10.6	0.8	0.3	8.9	3.0	10.0	6.4	13.4	27.4	19.2	0.0
14	1935	8.8	2.6	1.4	7.8	1.4	12.4	6.2	11.3	29.3	18.5	0.4
15	1938	10.1	1.1	1.2	5.9	1.4	9.5	6.0	5.9	40.7	18.2	0.0
16	1947	15.6	1.6	10.1	11.4	7.6	8.8	4.8	3.4	32.2	4.6	0.0
17	1950	11.2	1.3	16.5	12.4	15.8	8.1	4.9	3.4	20.7	4.2	1.5
18	1953	12.9	1.5	7.0	7.9	12.1	8.1	5.3	3.9	36.1	5.2	0.0
19	1956	10.9	5.3	9.7	7.6	9.6	9.4	8.5	4.6	28.2	6.2	0.0
20	1959	13.1	4.4	7.3	5.7	9.8	12.5	8.0	5.0	26.7	7.5	0.0
21	1962	12.8	4.7	7.5	6.6	6.8	15.7	9.7	5.3	24.5	6.4	0.1
22	1965	12.4	4.3	8.4	9.1	6.0	19.5	10.6	4.7	19.8	3.5	1.8
23	1968	11.4	6.0	9.5	5.9	5.0	21.1	10.7	4.2	20.0	4.4	1.9
24	1971	12.8	2.8	7.1	8.5	4.0	23.8	11.3	3.7	18.8	7.2	0.0

Principal Component Analysis

24 Observations
11 Variables

Simple Statistics

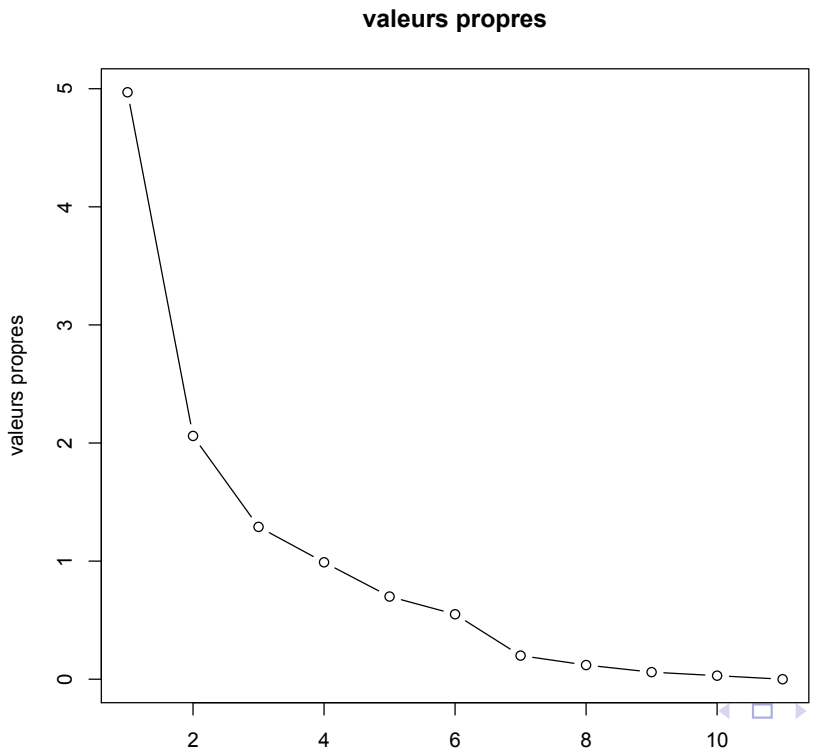
	PVP	AGR	CMI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
Mean	12.21250000	1.995833333	3.941666667	8.320833333	3.958333333	9.941666667	4.816666667	4.275000000	30.25833333	19.14166667	1.183333333
StD	2.19113983	1.645822785	4.489053043	2.467789018	4.181897164	5.223258615	3.408771365	4.154841553	7.30952100	12.19371138	1.025778837

ACP (9/13): exemple

Correlation Matrix

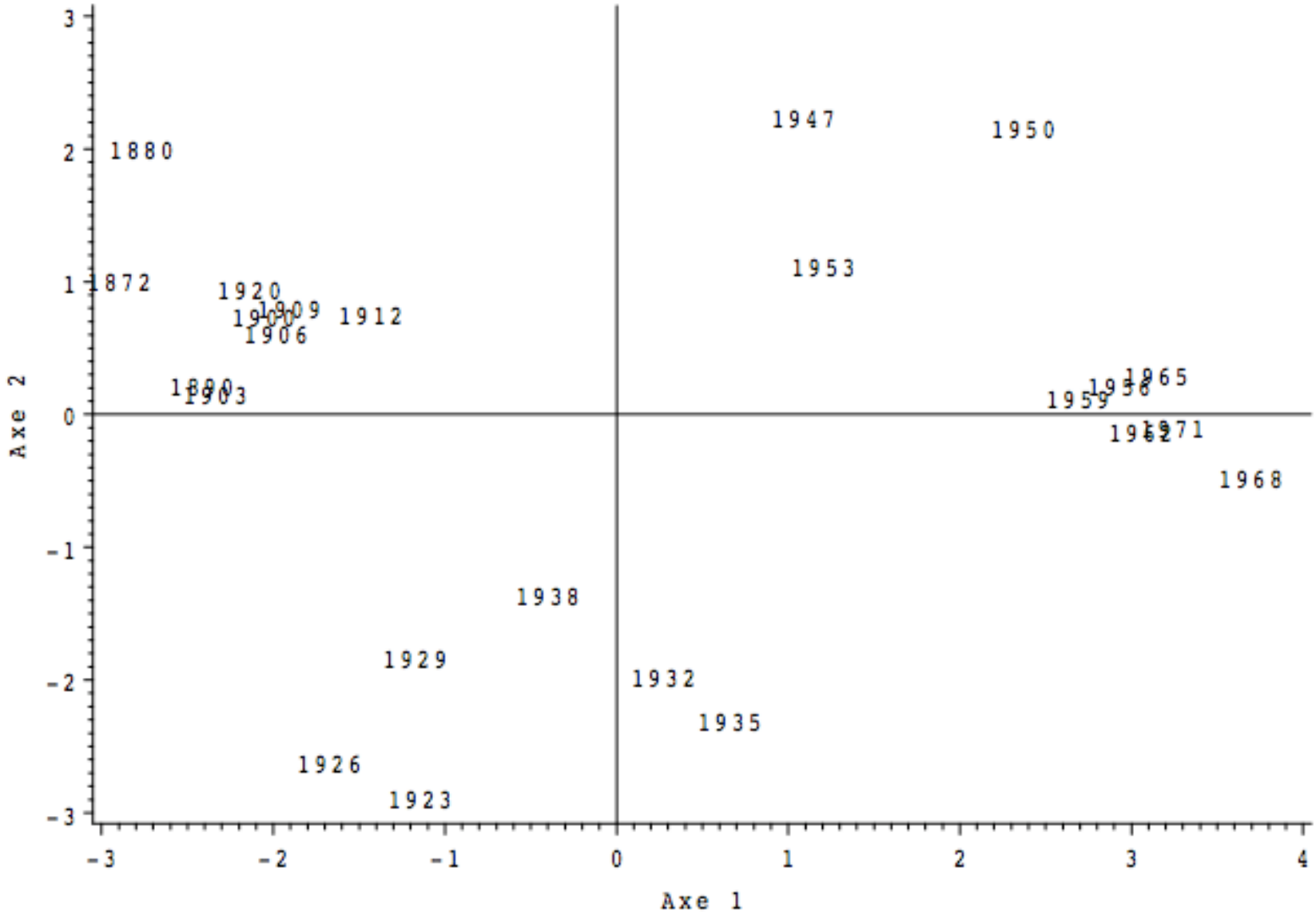
	PHP	AGR	CHI	TRA	LOG	EDU	ACS	ANC	DEF	DET	DIV
PHP	1.0000	-.0846	-.0004	0.2327	0.0356	-.1500	-.1314	-.6869	0.1011	0.0336	0.1493
AGR	-.0846	1.0000	0.6001	-.2758	0.4357	0.7313	0.8057	0.0443	-.4484	-.6949	-.2772
CHI	-.0004	0.6001	1.0000	0.0930	0.8910	0.4671	0.6212	0.0225	-.5363	-.8042	-.3480
TRA	0.2327	-.2758	0.0930	1.0000	0.1661	-.2132	-.2031	-.3132	0.1580	-.1483	0.1144
LOG	0.0356	0.4357	0.8910	0.1661	1.0000	0.2324	0.4878	0.0447	-.3786	-.7580	-.4379
EDU	-.1500	0.7313	0.4671	-.2132	0.2324	1.0000	0.8750	0.1570	-.5241	-.6702	-.2486
ACS	-.1314	0.8057	0.6212	-.2031	0.4878	0.8750	1.0000	0.2882	-.5672	-.8082	-.5296
ANC	-.6869	0.0443	0.0225	-.3132	0.0447	0.1570	0.2882	1.0000	-.4169	-.0494	-.3775
DEF	0.1011	-.4484	-.5363	0.1580	-.3786	-.5241	-.5672	-.4169	1.0000	0.2616	0.0204
DET	0.0336	-.6949	-.8042	-.1483	-.7580	-.6702	-.8082	-.0494	0.2616	1.0000	0.5539
DIV	0.1493	-.2772	-.3480	0.1144	-.4379	-.2486	-.5296	-.3775	0.0204	0.5539	1.0000

	Eigenvalue
AXE1	4.97296
AXE2	2.06064
AXE3	1.29017
AXE4	0.99306
AXE5	0.70835
AXE6	0.55815
AXE7	0.20425
AXE8	0.12520
AXE9	0.06281
AXE10	0.03500
AXE11	0.00000



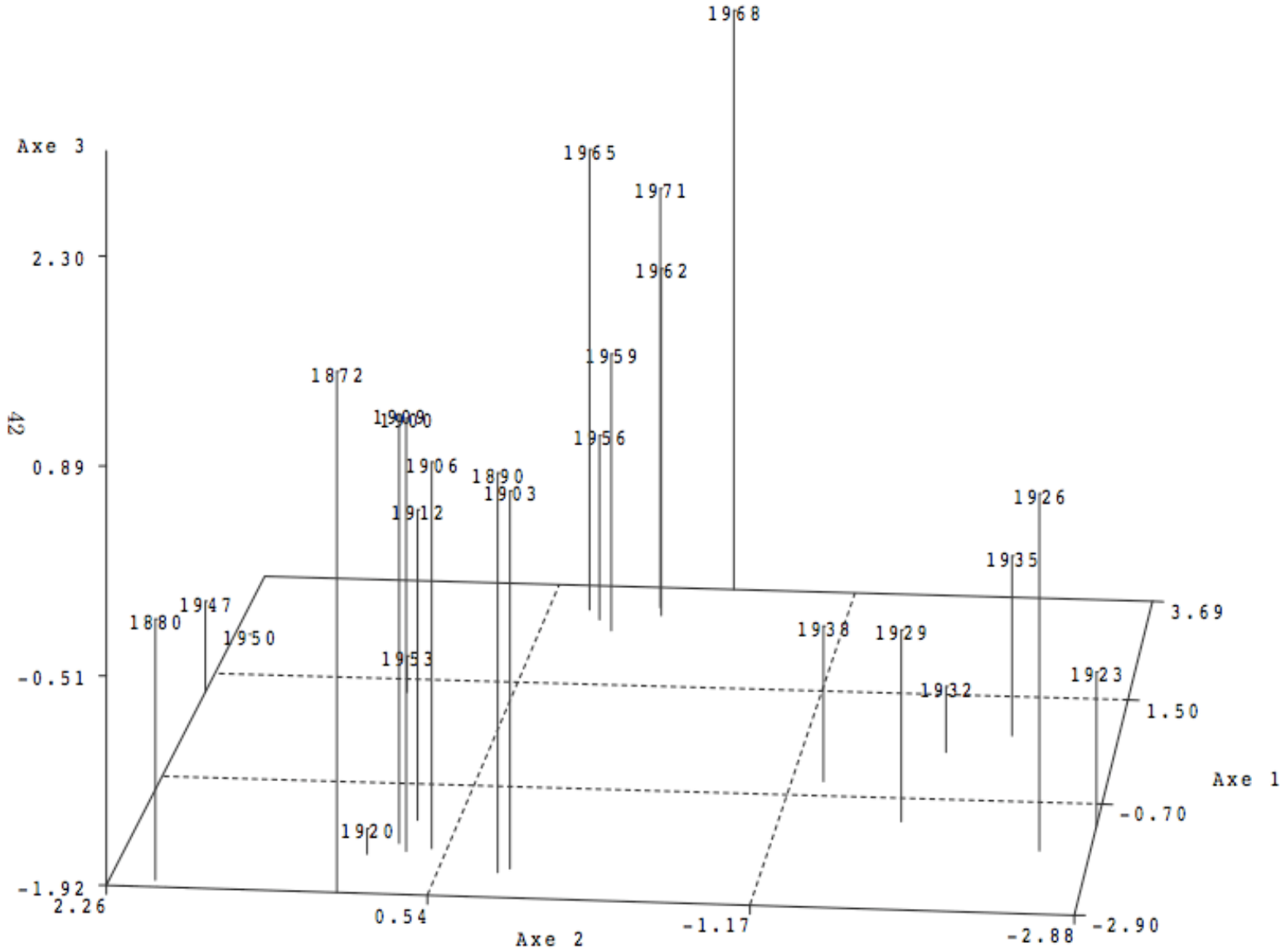
ACP (10/13): exemple

Projection sur les 2 premiers axes



ACP (11/13): exemple

Projection sur les 3 premiers axes



Cercle des corrélations: pour chaque variable a on définit le vecteur $r^{(a)} \in \mathbb{R}^q$ par

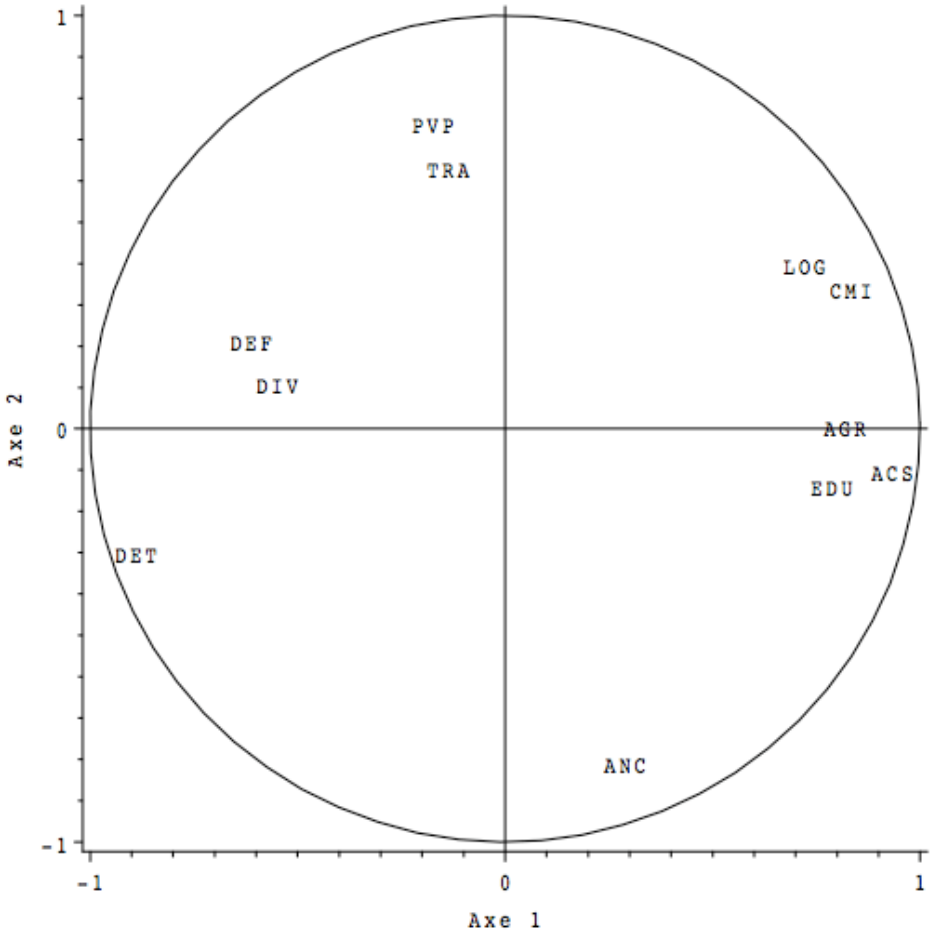
$$r_k^{(a)} = \overline{cor}(X^{(a)}, C_k) = \frac{C_k^T X^{(a)}}{\|X^{(a)}\| \, \|C_k\|}.$$

On a

$$\begin{aligned} \|r^{(a)}\|^2 &= \frac{1}{n-1} \|\mathrm{Proj}_{<C_1,\ldots,C_q>}(X^{(a)})\|^2 \\ &\leq \frac{1}{n-1} \|X^{(a)}\|^2 = \overline{var}(X^{(a)}) = 1. \end{aligned}$$

La norme de $\|r^{(a)}\|$ représente la qualité de la représentation de la variable a par les q premiers axes.

Cercle des corrélations

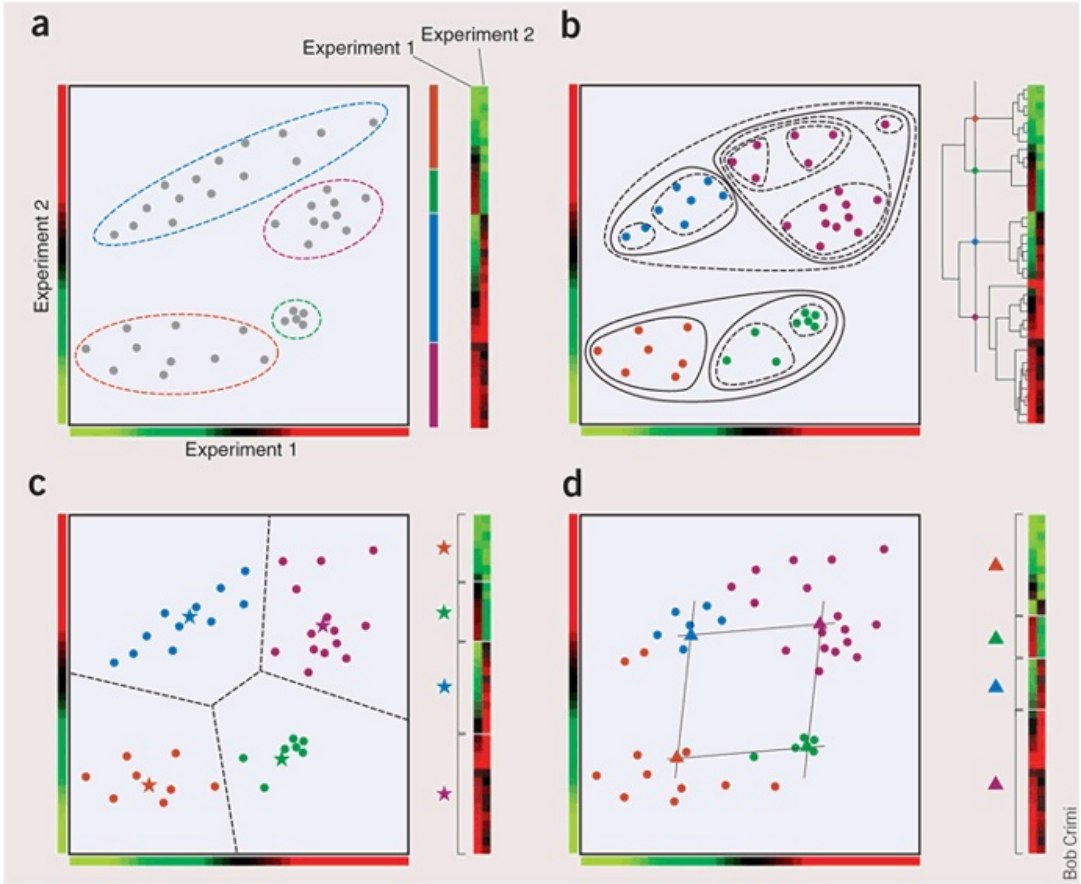


Variables proches du cercle: bien expliquées par les deux premiers axes.

2. Clustering

Objectifs

But: séparer un nuage de points en groupes $\mathcal{C}_1, \ldots, \mathcal{C}_k$. Problème mal posé:



Applications: écologie, phylogénie, post-génomique, imagerie médicale, etc.

k-means: pour k fixé (arbitrairement) l'algorithme vise à minimiser

$$\min_{\mathcal{C}_1,\ldots,\mathcal{C}_k} \left\{ \sum_{l=1}^{k} \sum_{X_i \in \mathcal{C}_l} \left(X_i - \frac{1}{\#\mathcal{C}_l} \sum_{X_j \in \mathcal{C}_l} X_j \right)^2 \right\}$$

Algorithme itératif de type E-M (Expectation-Maximisation)

Iterate:

1. $\bar{C}_l \leftarrow$ centre de gravité $\{X_i : \ X_i \in \mathcal{C}_l\}$
2. $\mathcal{C}_l \leftarrow \left\{ X_i \text{ tel que } \|X_i - \bar{C}_l\| = \text{argmin}_{l'} \|X_i - \bar{C}_{l'}\| \right\}$

Plusieurs algorithmes et interprétations (laplacien discret sur les graphes / "feature space")

Exemple:

1. envoi des données dans un espace de plus grande dimension (où les points sont mieux séparés):

$$\begin{array}{rccl} & \text{data} & \longrightarrow & \text{Feature space} \\ \phi: & \mathbb{R}^p & \longrightarrow & \mathcal{H} = \left\{ f : \mathbb{R}^p \to \mathbb{C} : \int |f(x)| e^{-\|x\|^2/2} dx < \infty \right\} \\ & x & \longmapsto & (t \to e^{i<t,x>}) \end{array}$$

2. faire une ACP sur $\phi(X_1), \ldots, \phi(X_n)$. Cela revient à calculer les vecteurs propres de $\left[e^{-\|X_i - X_j\|^2/2} \right]$.
3. clustering par k-means sur les axes principaux.

Estimation paramétrique

1. Exemple introductif

exemple introductif (1/5)

Problème: on veut estimer la proportion θ de moustiques infectés par le plasmodium (malaria)

Données: échantillon de n moustiques

$$X_i = \begin{cases} 1 & \text{si moustique } i \text{ infecté} \\ 0 & \text{sinon} \end{cases}$$

Estimation de θ: $\hat{\theta}_n = \frac{X_1 + \ldots + X_n}{n}$

Modélisation probabiliste: si $n \ll N =$ taille population totale, on modélise la loi des X_i par: X_i i.i.d. de loi de Bernoulli de paramètre θ.

exemple introductif (2/5)

Propriétés basiques:

- $\mathbb{E}(\hat{\theta}_n) = \frac{1}{n}\sum_i \mathbb{E}(X_i) = \theta$ (estimateur sans biais)
- LGN: $\hat{\theta}_n \longrightarrow \theta$ p.s. (estimateur consistant)

Intervalles de confiance: la proportion $\hat{\theta}_n$ est aléatoire. A quelle distance est-elle de θ?

L'estimateur $\hat{\theta}_n$ a la loi d'une variable binomiale $\mathcal{B}(n, \theta)$ divisée par n. On peut donc calculer ε_α tel que $\mathbb{P}(|\theta - \hat{\theta}_n| \leq \varepsilon_\alpha) = 1 - \alpha$.

$$\Longrightarrow \theta \in \mathcal{I}_\alpha := [\hat{\theta}_n - \varepsilon_\alpha, \hat{\theta}_n + \varepsilon_\alpha] \text{ avec probabilité } 1 - \alpha.$$

$\mathcal{I}_\alpha$ est un intervalle de confiance de niveau α. Il est ALEATOIRE.

exemple introductif (3/5)

Pour n grand (≥ 30) on peut utiliser des approximations pour ε_α:

- approximation gaussienne si θ pas trop proche de 0 ou 1
- approximation poissonnienne si θ proche de 0 ou 1 (typiquement $\theta \leq 5/n$ ou $\theta \geq 1 - 5/n$)

Approximation Gaussienne:

$$\text{TCL: } \sqrt{\frac{n}{\theta(1-\theta)}}(\hat{\theta}_n - \theta) \overset{loi}{\to} Z \sim \mathcal{N}(0,1)$$

Soit z_α tel que $\mathbb{P}(|Z| \leq z_\alpha) = 1 - \alpha$. On a

$$\mathbb{P}\left(\theta \in \left[\hat{\theta}_n - \sqrt{\frac{\theta(1-\theta)}{n}} z_\alpha, \hat{\theta}_n + \sqrt{\frac{\theta(1-\theta)}{n}} z_\alpha\right]\right) \overset{n \text{ grand}}{\approx} 1 - \alpha$$

Problème: $\varepsilon_\alpha = \sqrt{\frac{\theta(1-\theta)}{n}} z_\alpha$ dépend de θ...

Plug-in: l'idée est de remplacer $\sqrt{\frac{\theta(1-\theta)}{n}}z_\alpha$ par $\sqrt{\frac{\hat\theta_n(1-\hat\theta_n)}{n}}z_\alpha$.

Lemme de Slutsky
Si (Z_n) et (Y_n) sont deux suites de variables aléatoires réelles telles que $Z_n \xrightarrow{loi} Z$ et $Y_n \xrightarrow{\mathbb{P}} 1$ alors $Y_n Z_n \xrightarrow{loi} Z$.

permet de jusifier le "plug-in":

$$\sqrt{\frac{n}{\hat\theta_n(1-\hat\theta_n)}}(\hat\theta_n-\theta) = \underbrace{\sqrt{\frac{\theta(1-\theta)}{\hat\theta_n(1-\hat\theta_n)}}}_{\xrightarrow{p.s.} 1 \text{ (LGN)}} \times \underbrace{\sqrt{\frac{n}{\theta(1-\theta)}}(\hat\theta_n-\theta)}_{\xrightarrow{loi} Z \text{ (TCL)}}$$

$$\xrightarrow{loi} \quad Z \sim \mathcal{N}(0,1)$$

Approximation Poisson: lorsque $\theta \approx 1/n$ (évènements rares)

Rappel: Une variable binomiale $\mathcal{B}(n, \lambda/n)$ converge en loi vers la loi de Poisson de paramètre λ.

Lorsque la proportion θ qu'on cherche à estimer est de l'ordre de $1/n$ on peut approximer la loi de $\hat{\theta}_n$ par la loi de W_n où nW_n suit une loi de Poisson de paramètre $n\theta$. Ainsi:

$$\mathbb{P}\left(\theta \geq \hat{\theta}_n + \frac{k_\alpha}{n}\right) \approx e^{-n\theta} \sum_{k \leq n\theta - k_\alpha} \frac{(n\theta)^k}{k!}.$$

Objectifs de l'estimation en statistiques

On dispose:

- de données
- d'un modèle probabiliste pour ces données
  Exemples:
  - $X_1, \ldots, X_n$ i.i.d. de loi $\mathcal{B}(\theta)$
  - $Y_i = \theta_1 X_i^{(1)} + \ldots + \theta_p X_i^{(p)} + \varepsilon_i$ pour $i = 1, \ldots, n$. Dans ce modèle, Y est la variable d'intérêt, les $X^{(a)}$ sont des variables explicatives et ε est le bruit.
  - $dS_t/S_t = \theta_1\, dt + \theta_2\, dW_t$ (brownien géométrique)

But: estimer $\theta \in \mathbb{R}^p$.

- Comment construire un estimateur $\hat{\theta}$?
- Comment obtenir un intervalle de confiance?
- Comment quantifier la qualité d'un estimateur?

2. Quelques méthodes d'estimation

Modèle statistique: les données $X_1, \ldots, X_n$ (à valeurs dans $\mathbb{R}$ ou $\mathbb{R}^d$) sont supposées i.i.d. de loi appartenant à une famille (paramétrique) de lois $(\mathbb{P}_\theta)_{\theta \in \Theta}$ avec $\Theta \subset \mathbb{R}^p$.

Estimateur: un estimateur est une application

$$\begin{array}{rccll} \hat{\theta}: & \mathbb{R}^d \times \ldots \times \mathbb{R}^d & \longrightarrow & \mathbb{R}^p & \text{mesurable} \\ & (x_1, \ldots, x_n) & \longmapsto & \hat{\theta}(x_1, \ldots, x_n) & \end{array}$$

Méthode des moments

Si pour une fonction f donnée (typiquement $f(x) = x^\alpha$) on connaît $\phi(\theta) = \mathbb{E}_\theta(f(X_1))$, on définit l'estimateur

$$\hat{\theta}_n(X_1, \dots, X_n) := \phi^{-1}\left(\frac{1}{n}\sum_{i=1}^n f(X_i)\right).$$

Pour allèger, on écrit simplement $\hat{\theta}_n$ pour $\hat{\theta}_n(X_1, \dots, X_n)$.

Si ϕ continue, on a $\hat{\theta}_n \overset{ps}{\longrightarrow} \theta$ (Loi des grands nombres).

Ex: $d\mathbb{P}_\theta(x) = \theta e^{-\theta x}\, dx$ sur $\mathbb{R}^+$ (loi exponentielle). On a par exemple $\mathbb{E}_\theta(X_1^2) = \theta^{-2}$, ce qui incite à définir

$$\hat{\theta}_n = \left(\frac{1}{n}\sum_{i=1}^n X_i^2\right)^{-1/2}.$$

Maximum de vraisemblance (1/3)

Cadre: $X_1, \ldots, X_n$ i.i.d. de loi $\mathbb{P}_{\theta_0}$ appartenant à la famille de lois

$$\{d\mathbb{P}_\theta(x) = p(\theta, x)\, d\mu(x) \ : \ \theta \in \mathbb{R}^p\}$$

où μ est une mesure de référence (en général la mesure de Lebesgue dx ou la mesure de comptage sur $\mathbb{N}$).

Ex:

- loi exponentielle: $p(\theta, x) = \theta e^{-\theta x}$ et $d\mu(x) = \mathbf{1}_{\mathbb{R}_+}(x)\, dx$
- loi de Poisson: $p(\theta, x) = e^{-\theta} \frac{\theta^x}{x!}$ et $d\mu(x) = \sum_{k \in \mathbb{N}} \delta_k(x)$

Maximum de vraisemblance (2/3)

Définition:

- vraisemblance: $V(\theta, X_1, \ldots, X_n) := p(\theta, X_1) \ldots p(\theta, X_n)$
- estimateur du maximum de vraisemblance (s'il existe): $\hat{\theta}_n$ qui maximise $\theta \longrightarrow V(\theta, X_1, \ldots, X_n)$

Ex: loi exponentielle. On a

$$\log V(\theta, X_1, \ldots, X_n) = \sum_{i=1}^{n} (-\theta X_i + \log \theta)$$

et $\frac{\partial}{\partial \theta} \log V(\theta, X_1, \ldots, X_n) = -\sum_i X_i + n/\theta$, d'où

$$\hat{\theta}_n = \frac{n}{\sum_{i=1}^{n} X_i}$$

Maximum de vraisemblance (3/3)

Le principe du maximum de vraisemblance s'étend au cas où les observations ne sont pas i.i.d.

Exemple: supposons que $X_1, \dots, X_n$ sont les n premières valeurs d'une chaîne de Markov à valeurs dans $\{1, \dots, d\}$ et de noyau de transition Q_0 inconnu. La loi du vecteur $(X_1, \dots, X_n)$ appartient à la famille de lois $(\mathbb{P}_Q)_{Q \in [0,1]^{d \times d}}$ où

$$d\mathbb{P}_Q(x_1, \dots, x_n) = Q(x_0, x_1) \dots Q(x_{n-1}, x_n)\, d\mu(x_1) \dots d\mu(x_n).$$

On associe alors à tout noyau Q la vraisemblance

$$V(Q, X_1, \dots, X_n) = Q(x_0, X_1) \dots Q(X_{n-1}, X_n).$$

L'estimateur $\hat{Q}$ du maximum de vraisemblance est dans ce cas le noyau $\hat{Q}$ maximisant $Q \longrightarrow V(Q, X_1, \dots, X_n)$.

Normalité asymptotique

Cadre: $\theta \in \mathbb{R}$ (généralisable à $\theta \in \mathbb{R}^p$)

Information de Fisher: $I_n(\theta) := \mathbb{E}_\theta \left[\left(\frac{\partial \log V(\theta, X_1, \ldots, X_n)}{\partial \theta} \right)^2 \right]$

Lorsque les variables $X_1, \ldots, X_n$ sont i.i.d. on a $I_n(\theta) = n I_1(\theta)$.

Normalité asymptotique: si le modèle est suffisamment régulier (cf MAP 433) l'estimateur $\hat{\theta}_n$ du max de vraisemblance vérifie

1. $\sqrt{I_n(\theta)}\,(\hat{\theta}_n - \theta) \xrightarrow{loi} \mathcal{N}(0,1)$
2. $\sqrt{I_n(\hat{\theta}_n)}\,(\hat{\theta}_n - \theta) \xrightarrow{loi} \mathcal{N}(0,1)$

Optimalité: pour tout estimateur $\tilde{\theta}_n$ sans biais de θ on a

$$\mathrm{var}(\tilde{\theta}_n) \geq 1/I_n(\theta) \quad \text{(borne de Cramer Rao)}.$$

Construction d'un intervalle de confiance (asymptotique)

Dans les cas simples, l'estimateur est souvent de la forme $\hat{\theta}_n = \psi(S_n)$ où S_n vérifie une forme de théorème central limite: $\sqrt{n}\,(S_n - s) \xrightarrow{loi} Z$. Si ψ est différentiable, on a alors

$$\sqrt{n}\,(\hat{\theta}_n - \psi(s)) \xrightarrow{loi} \psi'(Z).$$

La loi de $\psi'(Z)$ étant connue, on peut construire à partir de là un intervalle de confiance asymptotique.

Remarque: pour l'estimateur du max de vraisemblance on a (lorsque le modèle est régulier)

$$\sqrt{n}\,(\hat{\theta}_n - \theta) \xrightarrow{loi} \mathcal{N}(0, 1/I_1(\theta))$$

Construction d'un intervalle de confiance: exemple

Exemple: estimateur des moments:

$$\hat{\theta}_n = \underbrace{\phi^{-1}}_{\psi}\Big(\underbrace{\frac{1}{n}\sum_{i=1}^{n} f(X_i)}_{S_n}\Big)$$

avec

$$\sqrt{n}(S_n - \mathbb{E}_\theta(f(X_1))) \xrightarrow{loi} Z \sim \mathcal{N}(0, \mathbf{var}_\theta(f(X_1)).$$

Si ϕ' existe et $\phi'(\theta) \neq 0$ on a:

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{loi} \psi'(Z) \sim \mathcal{N}\left(0, \frac{\mathbf{var}_\theta(f(X_1))}{\phi'(\theta)^2}\right).$$

Dans les cas plus complexes, les intervalles de confiances sont en général obtenus par simulation Monte Carlo.

3. Estimation bayésienne

Risque d'un estimateur

Problème: comment quantifier la performance d'un estimateur $\hat{\theta}$?

Fonction de perte: $L : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}^+$

Ex:

- $L(\theta, \theta') = \|\theta - \theta'\|^2$ (perte quadratique ou L^2)
- $L(\theta, \theta') = \mathcal{K}(\theta'|\theta) = \mathbb{E}_\theta \left(\log \frac{V(\theta, X_1, \ldots, X_n)}{V(\theta', X_1, \ldots, X_n)} \right)$ (perte Kullback, naturellement associée au max de vraisemblance)

Risque associé à L: $\mathbb{R}_\theta(\hat{\theta}) = \mathbb{E}_\theta \left[L\left(\theta, \hat{\theta}(X_1, \ldots, X_n)\right) \right]$

Ex:

- risque L^2: $\mathbb{R}_\theta(\hat{\theta}) = \mathbb{E}_\theta \left[\|\theta - \hat{\theta}(X_1, \ldots, X_n)\|^2 \right]$
- risque Kullback: $\mathbb{R}_\theta(\hat{\theta}) = \mathbb{E}_\theta \left[\mathcal{K}\left(\hat{\theta}(X_1, \ldots, X_n)|\theta\right) \right]$

Deux points de vue:

1. **Fréquentiste:** il existe un vrai θ inconnu
2. **Bayésien:** il existe une distribution connue sur θ: la loi *a priori* π.

Les objectifs du statisticien vont différer selon le point de vue adopté. Par exemple, le fréquentiste va vouloir minimiser le risque minimax

$$R^*_{\text{minimax}}(\hat{\theta}) = \sup_{\theta} R_\theta(\hat{\theta}),$$

alors que le bayésien va vouloir minimiser le risque bayésien

$$R^\pi(\hat{\theta}) = \int_\theta R_\theta(\hat{\theta})\, d\pi(\theta).$$

Théorème:

Supposons que $L(\theta, \theta') = \|\theta - \theta'\|^2$ et que $\int_\theta \theta^2 \, d\pi(\theta) < \infty$. Définissons

$$\hat{\theta}^\pi := \int_\theta \theta \, d\pi(\theta | X_1, \ldots, X_n)$$

où $\pi(\theta | X_1, \ldots, X_n)$ est la loi *a posteriori* définie par:

$$d\pi(\theta | X_1, \ldots, X_n) = \frac{V(\theta, X_1, \ldots, X_n)}{\int_\alpha V(\alpha, X_1, \ldots, X_n) \, d\pi(\alpha)} \, d\pi(\theta).$$

L'estimateur $\hat{\theta}^\pi$ est appelé estimateur bayésien, il vérifie

$$R^\pi(\hat{\theta}^\pi) = \min_{\hat{\theta}} R^\pi(\hat{\theta}).$$

Avantages comparés des approches fréquentiste versus bayésienne

Bayésien:

\+ - estimateur optimal donné par une formule close
- possibilité d'intégrer une connaissance *a priori*

\− - difficulté de choisir le prior π
- calcul de l'estimateur difficile en grande dimension (intégration par MCMC)

Fréquentiste:

\+ - ne nécessite aucune connaissance *a priori*

\− - pas de formule donnant un estimateur minimax
- calcul de l'estimateur du max de vraisemblance difficile en grande dimension

Introduction aux tests

content

1. Exemple introductif

Exemple introductif

Pour avoir une certification "bio" une coopérative agricole doit garantir pour chaque produit que le taux d'OGM est inférieur à 1%.

- **Prélèvement:** dans $n = 10$ champs pour chaque produit.
- **Modèle:** $X_i := \log(\%\text{OGM du champ } i) \overset{i.i.d.}{\sim} \mathcal{N}(\theta, 1) = \mathbb{P}_\theta$

Test: si $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i > \underbrace{t}_{\substack{\text{seuil à} \\ \text{choisir}}}$ le produit est refusé.

Quel seuil t prendre?

Point de vue de la coopérative (1/2)

Pour la coopérative le taux d'OGM est inférieur à 1% sauf preuve du contraire.

Elle va tester l'hypothèse **H0** : $\theta \leq 0$ (c'est à dire %OGM$\leq$ 1%) contre l'alternative **H1** : $\theta > 0$.

Si **H0** est vraie, elle veut que la probabilité que le test rejette **H0** soit faible:

$$\boxed{\mathbb{P}_{\mathbf{H0}}\left[\text{rejeter } \mathbf{H0}\right] \leq \alpha \quad (\text{pex } \alpha = 5\%).}$$

Point de vue de la coopérative (2/2)

On a:

$$\begin{aligned}\mathbb{P}_{\mathbf{H0}}\left[\text{rejeter } \mathbf{H0}\right] &\leq \sup_{\theta \leq 0} \mathbb{P}_\theta(\bar{X}_n > t) = \mathbb{P}_0(\bar{X}_n > t) \\ &\leq \mathbb{P}\left[\mathcal{N}\left(0, 1/10\right) > t\right] = \mathbb{P}(\mathcal{N}(0,1) > \sqrt{10}\, t).\end{aligned}$$

Vu que $\mathbb{P}(\mathcal{N}(0,1) > \sqrt{10} \times 0.52) = 5\%$, le test sera

Test (de la coopérative): rejet du produit si $\bar{X}_n > 0.52$

Point de vue association "anti-OGM" (1/2)

Une association "anti-OGM" veut s'assurer qu'il n'y a pas plus de 1% d'OGM.

Elle s'interroge: "si le taux d'OGM est 50% supérieur à la norme (c'est-à-dire %OGM= 1.5%), quelle est la probabilité que le test le détecte?

$$\begin{aligned}\mathbb{P}(\text{détecter}) &= \mathbb{P}_{\log(1.5)}(\bar{X}_n > 0.52) \\ &= \mathbb{P}(\mathcal{N}(0,1) > \sqrt{10}(0.52 - \log(1.5))) = 29\%\end{aligned}$$

SCANDALE!!!

Point de vue association "anti-OGM" (2/2)

Pour l'association le taux d'OGM est supérieur à 1% sauf preuve du contraire: elle va tester

$$\textbf{H0} : \theta > 0 \quad \text{contre} \quad \textbf{H1} : \theta \leq 0.$$

Elle veut que $\mathbb{P}_{\textbf{H0}}[\text{rejeter } \textbf{H0}] \leq \alpha = 5\%$. En raisonnant comme précédemment, elle obtient le test:

Test (de l'association): rejet du produit si $\bar{X}_n > -0.52$

Définition d'un test

Définition: un "test de niveau α" c'est

1. le choix d'une hypothèse **H0** et d'une alternative **H1**,
2. le choix d'un α (petit) appelé "niveau du test",
3. le choix d'une règle de décision $T(X_1, \ldots, X_n) \in \{0, 1\}$ telle que:
   - on accepte **H0** si $T(X_1, \ldots, X_n) = 0$ et on la rejette sinon,
   - $\mathbb{P}_{\mathbf{H0}}[\text{rejeter } \mathbf{H0}] = \mathbb{P}_{\mathbf{H0}}[T(X_1, \ldots, X_n) = 1] \leq \alpha$.

Quelle est l'information fournie par un test?

Le test n'est véritablement informatif que si **H0** est rejetée. En effet:

- **Si H0 est rejetée:** la probabilité d'erreur du test est $\mathbb{P}_{\mathbf{H0}}[\text{rejeter } \mathbf{H0}] \leq \alpha$. (erreur de type I)
- **Si H0 est acceptée:** la probabilité d'erreur du test est $\mathbb{P}_{\mathbf{H1}}[\text{accepter } \mathbf{H0}]$ qui est non controlée. (erreur de type II)

2. Test du rapport de vraisemblance

Test du rapport de vraisemblance

Cadre:

- n observations $X_1, \ldots, X_n$ de loi $d\mathbb{P}_\theta(x_1, \ldots, x_n) = V(\theta, x_1, \ldots, x_n) d\mu(x_1, \ldots, x_n)$
- **H0** :$\theta \in \Theta_0$ et **H1** :$\theta \in \Theta_1$

Rapport de vraisemblance:

$$r(X_1, \ldots, X_n) = \frac{\sup_{\theta \in \Theta_1} V(\theta, X_1, \ldots, X_n)}{\sup_{\theta \in \Theta_0} V(\theta, X_1, \ldots, X_n)}$$

Test du rapport de vraisemblance:

$T^*(X_1, \ldots, X_n) = \mathbf{1}_{\{r(X_1, \ldots, X_n) > t_\alpha\}}$ avec t_α tel que

$$\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\underbrace{r(X_1, \ldots, X_n) > t_\alpha}_{\text{rejette } \mathbf{H0}}) \leq \alpha.$$

T^* un bon test?

Puissance: mesure le pouvoir de discrimination d'un test T

$$\text{Puissance}(\theta) = \mathbb{P}_\theta(T(X_1, \ldots, X_n) = 1) \quad \text{pour } \theta \in \Theta_1$$

Théorème (Neymann-Pearson)

Cadre: Supposons que $\Theta_0 = \{\theta_0\}$ et $\Theta_1 = \{\theta_1\}$. Soit T^* un test de rapport des vraisemblances tel que $\mathbb{P}_{\theta_0}(T^* = 1) = \alpha$.

Optimalité: *Il n'existe pas* de test de niveau α plus puissant que T^*.

(se généralise à un cadre plus large)

3. Test de Wald

Le problème statistique: exemple

Exemple: on veut savoir si la pollution atmosphérique influe sur la croissance du blé

Données: n relevés de taille du blé, %H2O %CO2, indice de pollution ($\%poll.$), etc.

Modèle:
taille$=\theta_0 + \theta_{H20} * \%H2O + \theta_{CO2} * \%CO2 + \theta_{poll.} * \%poll. + \ldots + \varepsilon$

Test: **H0** : $\theta_{poll.} = 0$ contre **H1** : $\theta_{poll.} \neq 0$

Objectif: tester si θ appartient à une sous-variété de $\mathbb{R}^p$.

Plus précisément, pour $g : \mathbb{R}^p \to \mathbb{R}^d$ submersion de classe $\mathcal{C}^1$, on veut tester

$$\textbf{H0} : g(\theta) = 0 \text{ contre } \textbf{H1} : g(\theta) \neq 0$$

Cadre: on suppose que le modèle est suffisamment régulier (voir MAP 433) et que la matrice d'information de Fisher $I(\theta)$ est continue et inversible.

On a en particulier: $\sqrt{n}\left(\hat{\theta}_n - \theta\right) \overset{loi}{\to} \mathcal{N}\left(0, I(\theta)^{-1}\right)$.

Test de Wald

Théorème:
si le modèle est suffisamment régulier et si $\hat{\theta}_n$ est l'estimateur du max de vraisemblance on a sous **H0** lorsque $n \to \infty$

$$W_n^2 := n g(\hat{\theta}_n)^T \left[Dg(\hat{\theta}_n)\, I(\hat{\theta}_n)^{-1} Dg(\hat{\theta}_n)^T \right]^{-1} g(\hat{\theta}_n) \overset{loi}{\to} \chi^2(d)$$

où $Dg(\theta)$ est la matrice de la différentielle de g en θ et $\chi^2(d)$ est la loi de $\varepsilon_1^2 + \ldots + \varepsilon_d^2$ avec $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

Test de Wald: notons $q_d(\alpha)$ tel que $\mathbb{P}(\chi^2(d) \geq q_d(\alpha)) = \alpha$.
Alors le test définit par $T_n = \mathbf{1}_{\{W_n^2 \geq q_d(\alpha)\}}$ est un test de niveau asymptotique α.

4. Tests de comparaison

Le problème statistique

Exemple: on veut comparer l'efficacité d'un nouvel antiviral au traitement standard sur la grippe.

Observations:

- n prélèvements (taux de virus) sur des sujets soignés avec l'antiviral standard:
  $X_1, \dots, X_n$ i.i.d. On note $F_X(t) = \mathbb{P}(X_i \leq t)$
- m prélèvements (taux de virus) sur des sujets soignés avec le nouvel antiviral:
  $Y_1, \dots, Y_m$ i.i.d. On note $F_Y(t) = \mathbb{P}(Y_i \leq t)$

Test: on veut tester

$$\textbf{H0} : F_X = F_Y \text{ contre } \textbf{H1} : F_X(t) < F_Y(t) \ \forall\, t \in \mathbb{R}.$$

Approche paramétrique

Une approche simple consiste à modéliser la loi de X et Y:
$X_i \overset{i.i.d.}{\sim} \mathbb{P}_{\theta_X}$ et $Y_i \overset{i.i.d.}{\sim} \mathbb{P}_{\theta_Y}$ avec $\mathbb{P}_{\theta_X}, \mathbb{P}_{\theta_Y} \in \{\mathbb{P}_\theta : \ \theta \in \Theta\}$.

Le test **H0** : $F_X = F_Y$ contre **H1** : $F_X < F_Y$ se traduit alors souvent en **H0** : $\theta_X = \theta_Y$ contre **H1** : $\theta_X < \theta_Y$ (ou $\theta_X > \theta_Y$).
Il est alors possible d'effectuer un test du rapport des vraisemblances (par exemple).

Défaut: la phase de modélisation de la loi peut être délicate et conduire à des conclusions erronées.

Approches non-paramétriques

Les approches non-paramétriques ne nécessitent pas la modélisation de la loi des X et Y. En voici quelques exemples.

- **Test de Kolmogorov-Smirnov:** basé sur la fonction de répartition empirique. Très général mais peu performant en pratique.
- **Test de Wilcoxon:** suppose $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ indépendants.
- **Test du signe:** suppose $n = m$ et X de loi diffuse $(\forall x,\ \mathbb{P}(X = x) = 0)$
- **Test du rang et du signe:** suppose $n = m$, les (X_i) indépendants des (Y_i) et X de loi diffuse $(\forall x,\ \mathbb{P}(X = x) = 0)$.

Test de Wilcoxon

On suppose $(X_1, \ldots, X_n)$ et $(Y_1, \ldots, Y_m)$ indépendants.

Rang: on considère ensembles les $n + m$ valeurs $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ et on les ordonne par ordre croissant. On note $R(i)$ le rang de X_i dans cette suite de valeurs ordonnées.

Ex: $(X_1, X_2, X_3) = (5, 2, 3)$ et $(Y_1, Y_2) = (7, 4)$.
Ici $R(1) = 4$, $R(2) = 1$ et $R(3) = 2$.

Test: test de **H0** : $F_X = F_Y$ contre **H1** : $F_X < F_Y$:
rejet de **H0** si $W_n = \sum_{i=1}^{n} R(i) > t_\alpha$ avec t_α dans des tables.

Propriétés: sous **H0**

- la loi de R et W_n ne dépend pas de F_X ni F_Y
- $\frac{W_n - n(n+m+1)/2}{\sqrt{nm(n+m+1)/12}} \xrightarrow{loi} \mathcal{N}(0,1)$

Test du signe

suppose $n = m$ et X de loi diffuse ($\forall x,\ \mathbb{P}(X = x) = 0$)

Test: test de **H0** : $F_X = F_Y$ contre **H1** : $F_X < F_Y$:
rejet de **H0** si $S_n = \sum_{i=1}^{n} \mathbf{1}_{\{X_i > Y_i\}} > t_\alpha$ avec t_α dans des tables.

Propriétés: sous **H0**

- S_n suit une loi Binomiale $\mathcal{B}(n, 1/2)$ sous **H0** donc on prend t_α tel que

$$\sum_{k=t_\alpha}^{n} 2^{-n} C_n^k \leq \alpha.$$

- $\frac{S_n - n/2}{\sqrt{n/4}} \xrightarrow{loi} \mathcal{N}(0,1)$

Test du rang et du signe

suppose $n = m$, les (X_i) indépendants des (Y_i) et X de loi diffuse $(\forall x,\ \mathbb{P}(X = x) = 0)$.

Test: test de **H0** : $F_X = F_Y$ contre **H1** : $F_X < F_Y$:

- $R(i)$:= rang de la valeur $|X_i - Y_i|$ dans la suite de valeurs $|X_1 - Y_1|, \ldots, |X_n - Y_n|$ ordonnées par ordre croissant.
- rejet de **H0** si $W_n^+ = \sum_{i=1}^n R(i)\mathbf{1}_{\{X_i > Y_i\}} > t_\alpha$ avec t_α donné dans des tables.

Propriétés: sous **H0**

- $W_n^+ = \sum_{r=1}^n r\varepsilon_r$ où $\varepsilon_r \overset{i.i.d.}{\sim} \mathcal{B}(1/2)$
- $\frac{W_n^+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \xrightarrow{loi} \mathcal{N}(0,1)$

Le fléau de la dimension

content

Données de grandes dimensions

Données actuelles: le développement des (bio)technologies permet de mesurer beaucoup de paramètres en même temps, mais la taille de l'échantillon reste limité par les coûts expérimentaux:

- le nombre p de paramètres mesurés (= dimension des observations) devient très grand ($\approx$ 100 à 10000)
- la taille n de l'échantillon reste faible ($\approx$ 10 à 100)

Exemples:

- Génomique: pour chaque échantillon une puce à ADN mesure le niveau d'expression de $\approx$ 4000 gènes. Le nombre d'échantillons reste limité à quelques dizaines.
- Spectrometrie: étude sur un petit nombre de patients diabétiques. Pour chaque patient de nombreuses lignes spectrales sont relevées.
- Imagerie médicale: moins de mesures que de pixel (!)

Changement de point de vue

Cadre statistique classique:

- nombre p de paramètres fixé
- on étudie le comportement asymptotique des estimateurs lorsque $n \longrightarrow \infty$

Données actuelles:

- inflation du nombre p de paramètres
- taille d'échantillon réduite: $n \approx p$ voir $n \ll p$

$\Longrightarrow$ penser différemment les statistiques!
(penser $n \longrightarrow \infty$ ne convient plus)

Le fléau de la dimension: exemple 1

On mesure p paramètres (à valeurs dans [0,1]) sur un échantillon de n individus. On a donc n observations $X_1, \ldots, X_n \in [0,1]^p$.

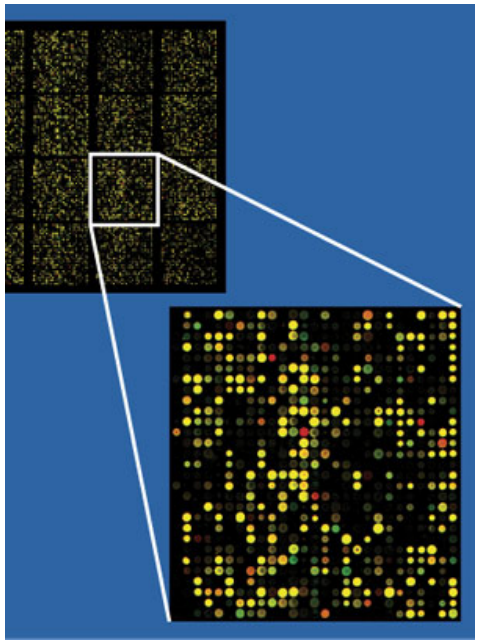
On souhaite évaluer la loi des X avec un histogramme. Faisons une grille 10 par 10: en moyenne on a $n/10^p$ observations par cases.

- Pour $p = 1$ une centaine d'observations permet d'avoir un histogramme raisonnable.
- Pour $p = 100$ il faut une masse énorme d'observations ($n \geq 10^{100}$) pour avoir en moyenne plus d'une observation par case.

$\Longrightarrow$ il est impossible d'inférer la distribution des X lorsque $p = 100$, sauf si les X prennent leurs valeurs dans une variété de $[0,1]^p$ (inconnue) de petite dimension.

Le fléau de la dimension: exemple 2

Puces ADN:

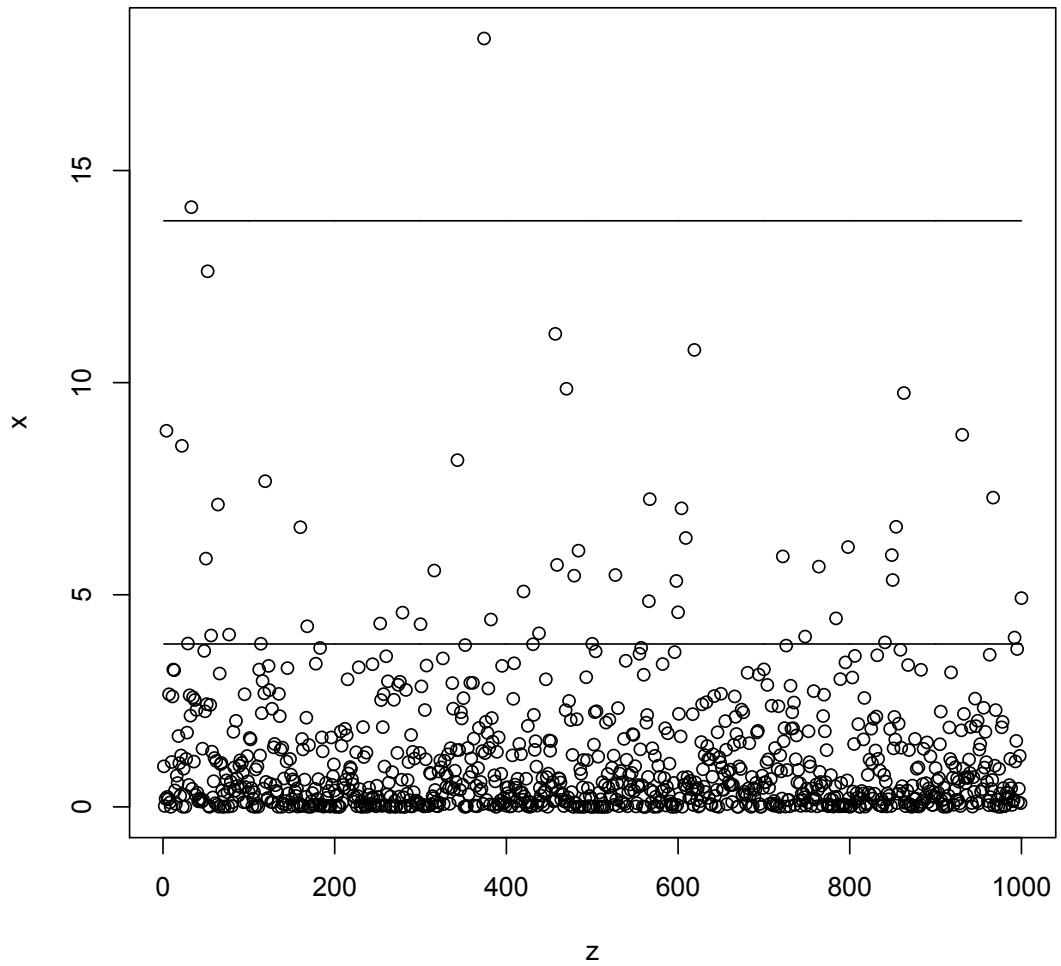


- **Modèle:** intensité du spot $X_i = \theta_i^2 + \sigma^2 \epsilon_i^2$ avec $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ et σ^2 supposé connu.

- **Déviation gaussienne à 5%:** on a $\mathbb{P}\left[(\mathcal{N}(0,1))^2 > 3.84\right] \approx 5\%$

Les valeurs X_i supérieures à $3.84\,\sigma^2$ sont-elles significatives?

Le fléau de la dimension: exemple 2



Avec $n = 1000$, $\sigma^2 = 1$ et $\theta_i = 0 \ \forall i$ (donc $X_i = \varepsilon_i^2$).
Niveaux représentés: 3.84 et $2 \log n$.

Le fléau de la dimension: exemple 2

$$\mathbb{P}\left(\max_{i=1,\dots,n} \varepsilon_i^2 > t_n\right) \quad \overset{t_n \sim \alpha \log n}{\longrightarrow} \quad \begin{cases} 0 & \text{si } \alpha \geq 2 \\ 1 & \text{si } \alpha < 2 \end{cases}$$

- Si on veut tester simultanément $\theta_i = 0$ contre $\theta_i \neq 0$ pour tout $i = 1, \dots, n$ il faut corriger le niveau 3.84
- Si on veut estimer les θ_i les valeurs de X_i en dessous de $2\sigma^2 \log n$ sont peu significatives:

  estimateur seuillé: $\hat{\theta}_i = \sqrt{X_i}\ \mathbf{1}_{\{X_i > 2\sigma^2 \log n\}}$.

1. Tests multiples

Tests multiples

Tests multiples: On effectue simultanément p tests:

$$\begin{array}{lllll} \text{Test } T_1 & \mathcal{H}_0^{(1)} : \theta \in \Theta_0^{(1)} & \text{contre} & \mathcal{H}_1^{(1)} : \theta \in \Theta_1^{(1)} \\ \dots & & & \dots \\ \text{Test } T_p & \mathcal{H}_0^{(p)} : \theta \in \Theta_0^{(p)} & \text{contre} & \mathcal{H}_1^{(p)} : \theta \in \Theta_1^{(p)} \end{array}$$

Niveau: si chaque test T_i est de niveau α, de nombreux tests vont rejeter $\mathcal{H}_0$ à tort: Si $m_0 := \left\{ i : \mathcal{H}_0^{(i)} \text{ vraie} \right\}$,
en moyenne $\alpha \times \text{Card}(m_0)$ tests rejettent $\mathcal{H}_0$ à tort.

Exemple: pour $\alpha = 5\%$, $p = 4000$ et $\text{Card}(m_0) = 3900$ on aura environ 200 fausses découvertes!

FWER: (Family-Wise Error Rate)
Un point de vue est de contrôler la probabilité de rejeter à tort (au moins) un des $\mathcal{H}_0^{(i)}$ (probabilité de faire une erreur ou plus):

$$\mathbb{P}\left(\cup_{i\in m_0}\{T_i = 1\}\right) \leq \alpha.$$

Bonferroni: De façon générale on peut prendre $\beta = \alpha/p$ car

$$\mathbb{P}\left(\cup_{i\in m_0}\{T_i = 1\}\right) \leq \sum_{i\in m_0} \mathbb{P}(T_i = 1) = |m_0|\beta \leq p\beta \leq \alpha.$$

Sime: Si les test T_i sont indépendants et de niveau β, on a

$$1 - \prod_{i\in m_0} \mathbb{P}(T_i = 1) = 1 - (1-\beta)^{|m_0|} \leq \alpha$$

ce qui conduit à prendre $\beta = 1 - (1-\alpha)^{1/p} > \alpha/p$.

Motivation: les tests multiples avec contrôle du FEWR sont très conservatifs. Benjamini et Hocheberg (1995) proposent de contrôler à la place le "taux de fausses découvertes".

FDR: (False Discovery Rate)

	$T=0$	$T=1$
$\mathcal{H}_0$	VN	FP
$\mathcal{H}_1$	FN	VP

Le taux de fausses découvertes est

$$FDR = \mathbb{E}\left[\frac{FP}{FP+VP}\,\mathbf{1}_{\{FP+VP>0\}}\right]$$

FDR moins conservatif que FEWR?

$$FDR = \mathbb{E}\left[\frac{FP}{FP + VP}\ \mathbf{1}_{\{FP+VP>0\}}\right]$$

Remarque:

- Si $VP = 0$ alors $FDR = \mathbb{P}(\exists\ FP)$
- Si $VP > 0$ alors $FDR < \mathbb{P}(\exists\ FP)$

donc conrôler le FDR est moins conservatif que contrôler le FEWR.

Supposons les tests T_i de la forme $T_i = \mathbf{1}_{S_i > s_{i,\alpha}}$ où $S_i = S_i(X_1, \ldots, X_n)$. On note

$$F_i(t) = \sup_{\theta \in \Theta_0^{(i)}} \mathbb{P}_\theta(S_i \geq t)$$

p-value: la p-value d'une valeur observée S_i^{obs} correspond à la probabilité d'observer sous $\mathcal{H}_0^{(i)}$ une valeur aussi élevée que S_i^{obs}:

$$p_i := F_i(S_i^{obs})$$

Remarque: Si $\Theta_0^{(i)} = \{\theta_0^{(i)}\}$ la variable $F_i(S_i^{obs})$ suit sous $\mathcal{H}_0^{(i)}$ une loi uniforme sur $[0, 1]$. Dans le cas général, $F_i(S_i^{obs})$ est sous $\mathcal{H}_0^{(i)}$ stochastiquement supérieure à une variable uniforme sur $[0, 1]$.

Procedure de Benjamini et Hocheberg:

1. On ordonne les p-value par ordre croissant

$$p(1) \leq p(2) \leq \ldots \leq p(p)$$

2. Rejet de toutes les hypothèses $\mathcal{H}_0^{(i)}$ correspondant aux p-values $p(1), \ldots, p(k)$ où

$$k = \operatorname{argmax} \{j : p(j) \leq \alpha j / n\}$$

Théorème: (BH95) Si les tests T_i sont indépendants, alors cette procédure permet de contrôler le FDR au niveau α

Remarque: encore vrai si les T_i sont positivement corrélés (Benjamini-Yekutelli '99). Sinon il y a un terme correctif (voir S.Dudoit).

2. Sélection de variables

Modèle de régression linéaire

Modèle:

$$\underbrace{Y_i}_{\text{observations}} = \sum_{a=1}^{p} \underbrace{\theta_a}_{\text{inconnu}} \underbrace{X_i^{(a)}}_{\substack{\text{covariables} \\ \text{(connues)}}} + \underbrace{\sigma\varepsilon_i}_{\text{bruit}} \text{ pour } i = 1, \ldots, n$$

Peut s'écrire $Y = X\theta + \sigma\varepsilon$. On supposera $\varepsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$.

Exemples:

- Modèle pour expliquer le rendement du blé en fonction de diverses covariables (ensoleillement, % CO_2, etc).
- Modèle de débruitage: $Y_i = f(t_i) + \sigma\varepsilon_i$, pour $i = 1, \ldots, n$. Si $\{\phi_a : a \in \mathcal{A}\}$ est un dictionnaire de fonctions (Fourier, ondelettes, Haar, etc) on a

$$Y_i = \sum_{a \in \mathcal{A}} \theta_a \underbrace{\phi_a(t_i)}_{=X_i(a)} + \sigma\varepsilon_i \,.$$

Estimateur des moindres carrés

Moindres carrés: (OLS) On suppose $\text{rang}(X) = p \leq n$.

L'estimateur des moindres carrés $\hat{\theta}_{ols}$ minimise $\|Y - X\hat{\theta}\|^2$. Notant $\mathcal{S} = \text{vect}\{X^{(1)}, \ldots, X^{(p)}\}$, on a $X\hat{\theta}_{ols} = \text{Pr}_{\mathcal{S}}\, Y$ et

$$\hat{\theta}_{ols} = (X^T X)^{-1} X^T Y.$$

Risque:

$$R_\theta(\hat{\theta}_{ols}) := \mathbb{E}\left[\|X\hat{\theta}_{ols} - X\theta\|_n^2\right] = \frac{p}{n}\,\sigma^2.$$

Bon estimateur? Non si p de l'ordre de n mais peu de covariables sont réellement influentes.

Collection d'estimateurs: pour $m \subset \{1, \dots, p\}$,

$$X\hat{\theta}_m = \underset{\mathcal{S}_m}{\Pr} Y \quad \text{où } \mathcal{S}_m = \text{vect}\{X^{(a)} : a \in m\}.$$

Risque:

$$R_\theta(\hat{\theta}_m) = \mathbb{E}\left[\|X\hat{\theta}_m - X\theta\|_n^2\right] = \underbrace{\|X\theta - \underset{\mathcal{S}_m}{\Pr}(X\theta)\|_n^2}_{\text{biais}} + \underbrace{\frac{\dim(\mathcal{S}_m)}{n}\sigma^2}_{\text{variance}}.$$

Idéal: prendre m^* (inconnu!) tel que

$$R_\theta(\hat{\theta}_{m^*}) = \inf_m R_\theta(\hat{\theta}_m).$$

Idée 1: sélection (1/2)

Idée 1: prendre $\hat{m}$ minimisant $\hat{R}_m$ où $\hat{R}_m$ est un estimateur de $R_\theta(\hat{\theta}_m)$.

Akaike '69: $\hat{R}_m = \|Y - X\hat{\theta}_m\|_n^2 + \frac{2|m|}{n}\sigma^2 - \sigma^2$ vérifie $\mathbb{E}(\hat{R}_m) = R_\theta(\hat{\theta}_m)$ d'où le critère:

$$\underline{\text{Critère AIC:}} \quad \hat{m} = \text{argmin}_m \Big\{ \|Y - X\hat{\theta}_m\|_n^2 + \underbrace{\frac{2|m|}{n}\sigma^2}_{\text{pénalité}} \Big\}.$$

Problème: mauvaises performances dans ce cadre dû à une trop grand variance de $\hat{R}_m - \hat{R}_{m^*}$ lorsque $|m|$ est grand.

Idée 1: sélection (2/2)

Il faut modifier le critère AIC pour tenir compte de la variance des $\hat{R}_m$. Pour $K > 0$:

$$\underline{\text{Critère:}} \quad \hat{m} = \text{argmin}_m \left\{ \|Y - X\hat{\theta}_m\|_n^2 + K \left(1 + \sqrt{2 \log p}\right)^2 \frac{|m|}{n} \sigma^2 \right\}.$$

Théorème: (Birgé & Massart '00) Il existe $c_K > 1$ telle que pour tout $\theta \in \mathbb{R}^p$

$$R_\theta(\hat{\theta}_{\hat{m}}) \leq c_K \min_{m \neq \varnothing} R_\theta(\hat{\theta}_m) \times \underbrace{\log p}_{\text{inévitable}}.$$

De plus ce "type d'inégalité" n'est pas vraie si $K < 1$.

Idée 2: prendre $\hat{\theta} = \sum_m w_m \hat{\theta}_m$ avec $\sum_m w_m = 1$.

Théorème: (Leung & Barron '06)

Pour $w_m \propto e^{-\hat{R}_m/(4\sigma^2)} \pi_m$ avec

- $\hat{R}_m = \|Y - X\hat{\theta}_m\|^2 + 2|m|\sigma^2$,
- $\pi_m = \left[(p+1)C_p^{|m|}\right]^{-1}$,

on a:

$$R_\theta(\hat{\theta}) \leq \min_m \left[R_\theta(\hat{\theta}_m) + 4\sigma^2 \log\left(C_p^{|m|}(p+1) \right) \right].$$

Problème algorithmique

$$\text{card}(\mathcal{P}\{1,\dots,p\}) = 2^p \quad \longrightarrow \quad \text{temps de calcul explose}$$

Idée: convexifier le problème en remarquant que $\hat{\theta}_{\hat{m}}$ avec $\hat{m} = \text{argmin}_m \left\{ \|Y - X\hat{\theta}_m\|^2 + \lambda |m| \right\}$ vérifie:

$$\hat{\theta}_{\hat{m}} = \text{argmin}_{\hat{\theta}\in\mathbb{R}^p} \left\{ \|Y - X\hat{\theta}\|^2 + \lambda |\hat{\theta}|_0 \right\}$$

Lasso: estimateur défini par:

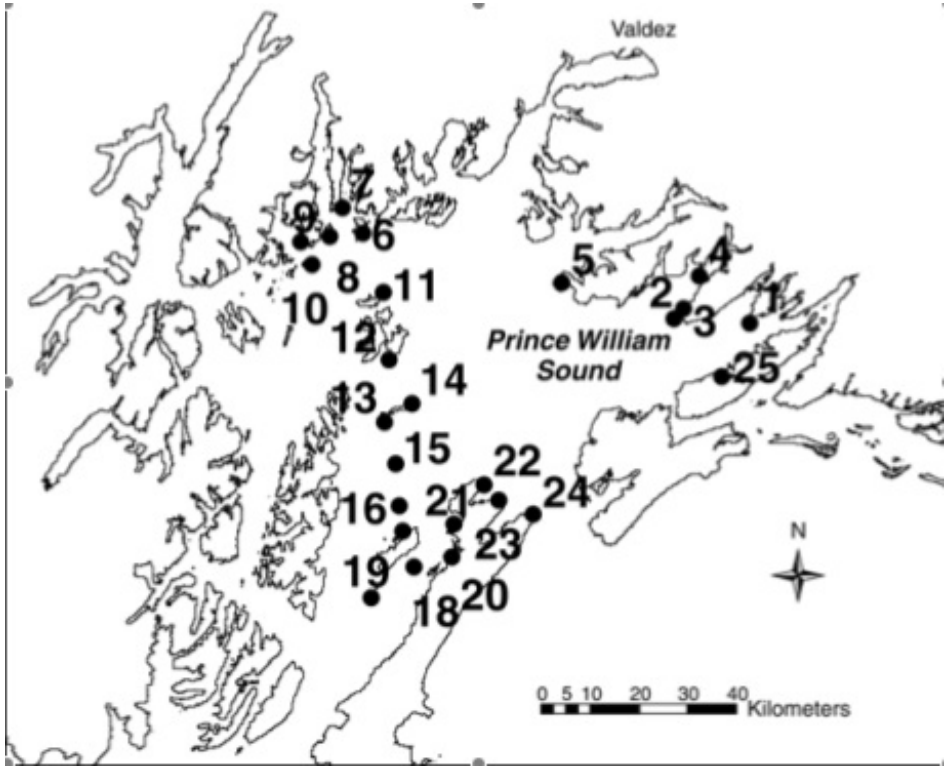
$$\hat{\theta}_{lasso} = \text{argmin}_{\hat{\theta}\in\mathbb{R}^p} \left\{ \|Y - X\hat{\theta}\|^2 + \lambda |\hat{\theta}|_1 \right\}$$

- sélectionne des variables
- algorithme de calcul efficace: LARS '04
- bonnes propriétés si les $X^{(a)}$ ne sont pas trop corrélés.

Sujet de mémoire

Le problème:

Suivi de la population de phoques en baie Prince William (Alaska) suite à la marée noire Exxon-Valdez.



Relevés sur 25 sites pendant 10 ans après Exxon-Valdez.

Objectif principal: estimer la tendance globale et pour chaque site.

Les données:

- Comptage (visuel) aérien sur 12 sites de 1990 à 2002.
- Période d'observation: pendant la mue (Aout-septembre) entre BM-2 et BM+2
- 7 à 10 comptages par an
- certaines données sont manquantes (brouillard)

Covariables:

j = Year-1989,
k = flight number
i = site number
z_{ijk} = number of seal at site i, year j, flight k
x_{1ijk} = Date with August 1 as day 1,
x_{2ijk} = Time-of-day from midnight (in hours),
x_{3ijk} = Time-relative-to-low-tide (in hours).

Références introductives

- **Statistique exploratoire multidimensionelle.** Cours en ligne de Philippe Besse.
  http://www.math.univ-toulouse.fr/∼besse/pub/enseignement.html

- **Statistique exploratoire multidimensionnelle.** L. Lebart, M. Piron, A. Morineau. Edition Dunod.

- **All of statistics.** Larry Wasserman. Edition Springer.

- **Introduction aux méthodes statistiques.** Marc Hoffmann. Poly MAP 433.