Diagnostic à partir de biopuces

David Causeur
Laboratoire de Mathématiques Appliquées
Agrocampus Ouest
IRMAR CNRS UMR 6625
http://www.agrocampus-ouest.fr/math/causeur/

Plan du cours

1 Préambule
 Objectifs
 Règle de Bayes
 Classification linéaire

2 Analyse discriminante linéaire de Fisher
 Calcul des scores
 Qualité de discrimination

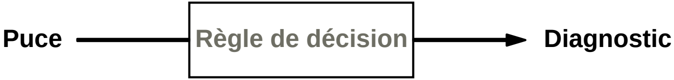
3 Sélection de modèle
 Principes
 Cas de deux groupes
 Cas de plusieurs groupes
 Sélection d'un sous-ensemble de gènes

4 Perspectives

Objectifs

Exemple : comparaison d’états d’un patient, d’un organisme

- Construction d’un outil de diagnostic (prédiction)

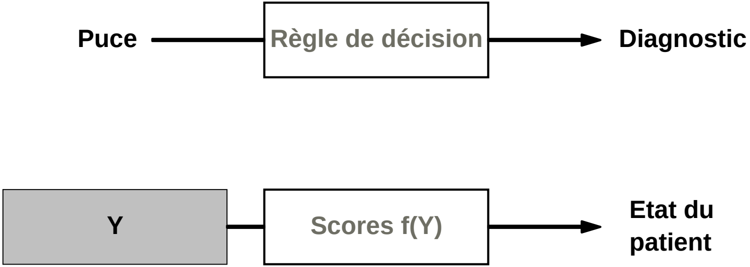


Puce → **Règle de décision** → **Diagnostic**

Objectifs

Exemple : comparaison d’états d’un patient, d’un organisme

- Construction d’un outil de diagnostic (prédiction)



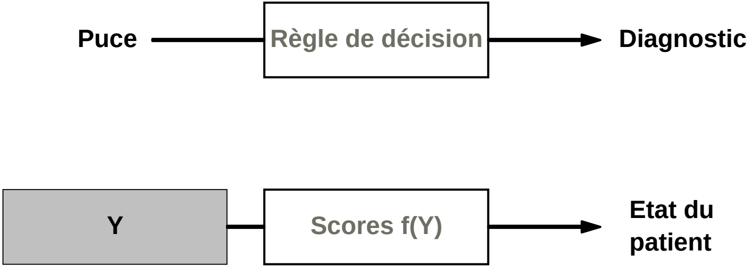
Puce → **Règle de décision** → **Diagnostic**

Y — **Scores f(Y)** → **Etat du patient**

Objectifs

Exemple : comparaison d'états d'un patient, d'un organisme

- Construction d'un outil de diagnostic (prédiction)



Puce → **Règle de décision** → **Diagnostic**

Y — **Scores f(Y)** → **Etat du patient**

- Sélection de gènes (cognition)

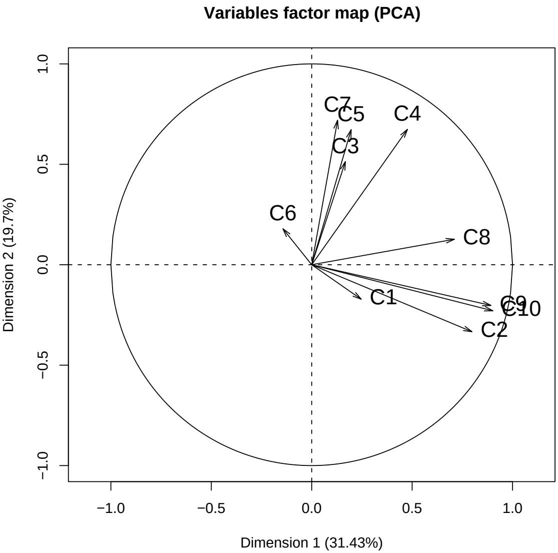
Discrimination à partir d'indices chromosomiques

Indices chromosomiques d'huîtres

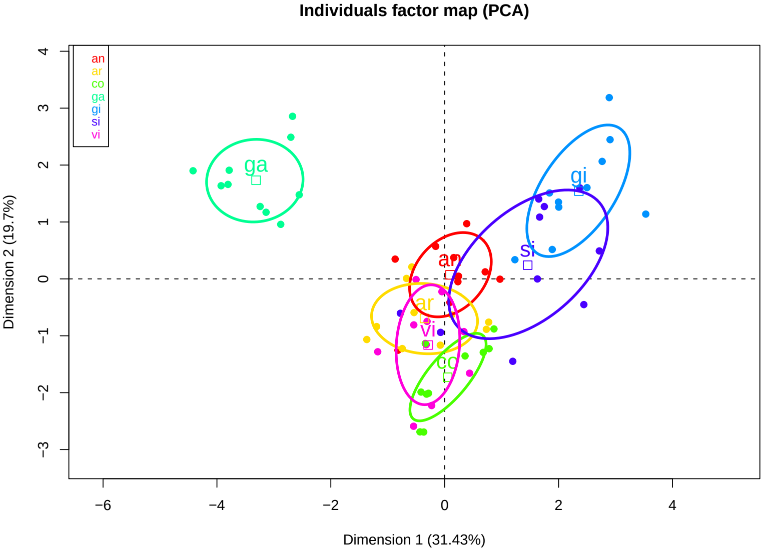
	C1	C2	C3	…	C9	C10	Espèce
1	45.33	49.74	45.43	…	38.31	39.80	vi
2	54.80	50.88	39.22	…	42.97	51.16	si
3	44.32	41.61	50.05	…	45.89	47.97	si
4	46.95	44.99	46.25	…	51.10	46.45	ar
5	40.60	26.36	44.28	…	30.49	25.23	ga
6	46.29	42.40	35.32	…	41.98	46.90	vi
7	39.52	42.27	43.45	…	45.71	47.91	co
⋮	⋮	⋮	⋮		⋮	⋮	

GI (*C. gigas*), AN (*C. angulata*), SI (*C. sikamea*), VI (*C. virginica*), AR (*C. ariakensis*), GA (*C. gasar*), CO (*S. commercialis*).

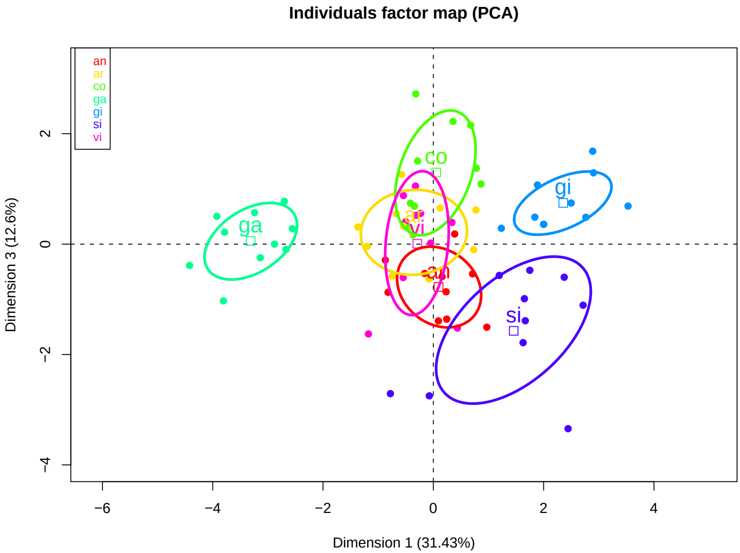
Discrimination à partir d'indices chromosomiques



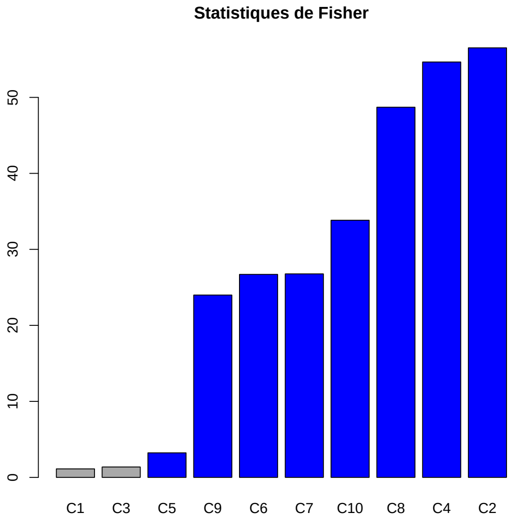
Discrimination à partir d'indices chromosomiques



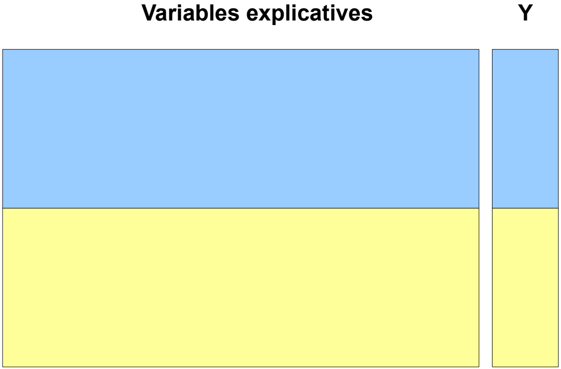
Discrimination à partir d'indices chromosomiques



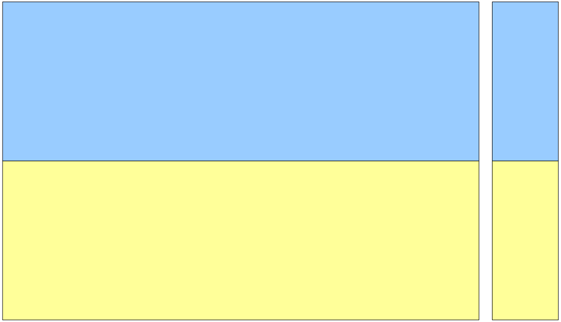
Discrimination à partir d'indices chromosomiques



Règle d'affectation à un groupe


Variables explicatives
Y

Règle d'affectation à un groupe



Variables explicatives **Y**

Règle d'affectation : $\hat{Y} = \eta(x) = \pm 1$

Problème : estimer η

Règle de Bayes

Risque d'erreur (sachant *x*) :

$$\begin{aligned}\pi(x) &= \underbrace{\mathbb{P}\big[Y=+1 \mid x\big]}_{f(x)} \text{ si } \eta(x) = -1 \\ &= \underbrace{\mathbb{P}\big[Y=-1 \mid x\big]}_{1-f(x)} \text{ si } \eta(x) = +1\end{aligned}$$

Règle de Bayes

Risque d'erreur (sachant *x*) :

$$\begin{aligned}\pi(x) &= \underbrace{\mathbb{P}\big[Y=+1 \mid x\big]}_{f(x)} \text{ si } \eta(x)=-1\\ &= \underbrace{\mathbb{P}\big[Y=-1 \mid x\big]}_{1-f(x)} \text{ si } \eta(x)=+1\end{aligned}$$

Règle optimale (Règle de Bayes) :

$$\eta(x) = \text{signe}\Big[\underbrace{\log\frac{f(x)}{1-f(x)}}_{\text{logodds}(x)}\Big]$$

Estimer η : estimer la loi de Y sachant x

Risques d'erreurs associés à une règle d'affectation

Probabilité de mauvaise affectation

$$\pi = \underbrace{\mathbb{P}\Big[\eta(X) = -1 \mid Y = +1\Big]}_{\text{1-Sensibilité = taux de faux négatifs}} \underbrace{\mathbb{P}(Y = +1)}_{p_1} + \underbrace{\mathbb{P}\Big[\eta(X) = +1 \mid Y = -1\Big]}_{\text{1-Spécifité = taux de faux positifs}} \underbrace{\mathbb{P}(Y = -1)}_{1-p_1}$$

Risques d’erreurs associés à une règle d’affectation

Probabilité de mauvaise affectation

$$\begin{aligned}\pi \quad = \quad & \underbrace{\mathbb{P}\Big[\eta(X) = -1 \mid Y = +1\Big]}_{\text{1-Sensibilité = taux de faux négatifs}} \quad \underbrace{\mathbb{P}(Y = +1)}_{p_1} + \\ & \underbrace{\mathbb{P}\Big[\eta(X) = +1 \mid Y = -1\Big]}_{\text{1-Spécifité = taux de faux positifs}} \quad \underbrace{\mathbb{P}(Y = -1)}_{1-p_1}\end{aligned}$$

Ex. : Diagnostic d’une maladie rare ($p_1 = 1/10.000$) avec Sens.=0.9 et Spéc.=0.9 Un individu choisi au hasard est déclaré positif : quellle probabilité pour qu’il soit malade ?

$$\begin{aligned}\mathbb{P}\Big[Y = +1 \mid \eta(X) = +1\Big] \quad &= \quad \frac{\text{Sensibilité}p_1}{\text{Sensibilité}p_1 + (1 - \text{Spécificité})(1 - p_1)} \\ &=\end{aligned}$$

Risques d'erreurs associés à une règle d'affectation

Probabilité de mauvaise affectation

$$\pi = \underbrace{\mathbb{P}\Big[\eta(X) = -1 \mid Y = +1\Big]}_{\text{1-Sensibilité = taux de faux négatifs}} \underbrace{\mathbb{P}(Y = +1)}_{p_1} + \underbrace{\mathbb{P}\Big[\eta(X) = +1 \mid Y = -1\Big]}_{\text{1-Spécifité = taux de faux positifs}} \underbrace{\mathbb{P}(Y = -1)}_{1-p_1}$$

Ex. : Diagnostic d'une maladie rare ($p_1 = 1/10.000$) avec Sens.=0.9 et Spéc.=0.9 Un individu choisi au hasard est déclaré positif : quellle probabilité pour qu'il soit malade ?

$$\begin{aligned}\mathbb{P}\Big[Y = +1 \mid \eta(X) = +1\Big] &= \frac{\text{Sensibilité}p_1}{\text{Sensibilité}p_1 + (1 - \text{Spécificité})(1 - p_1)} \\ &= 0.0009\end{aligned}$$

Règle linéaire d'affectation

Classification linéaire : logodds(x) $= \beta_0 + x'\beta$

Règle linéaire d'affectation

Classification linéaire : $\text{logodds}(x) = \beta_0 + x'\beta$

Risque empirique : $\hat{R}_\beta = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{1}_{\left[y_i \neq \text{signe}(\beta_0 + x_i'\beta)\right]}}_{\text{Coût d'une erreur}}$

Minimiser $\hat{R}_\beta$ est numériquement impossible !

Règle linéaire d'affectation

Classification linéaire : $\text{logodds}(x) = \beta_0 + x'\beta$

Risque empirique : $\hat{R}_\beta = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\mathbf{1}_{\left[y_i \neq \text{signe}(\beta_0 + x_i'\beta)\right]}}_{\text{Coût d'une erreur}}$

Minimiser $\hat{R}_\beta$ est numériquement impossible !

Différentes méthodes d'estimation

- Discrimination linéaire de Fisher
- Modèle de régression logistique
- SVM, Forêts aléatoires, ...

Règle linéaire d'affectation

Pour un individu avec $X = x$

$$\text{Si } \beta_0 + x'\hat{\beta} \geq 0, \quad \text{alors} \quad \hat{Y} = +1$$
$$\text{Si } x'\hat{\beta} \geq -\beta_0, \quad \text{alors} \quad \hat{Y} = +1$$

Règle linéaire d'affectation

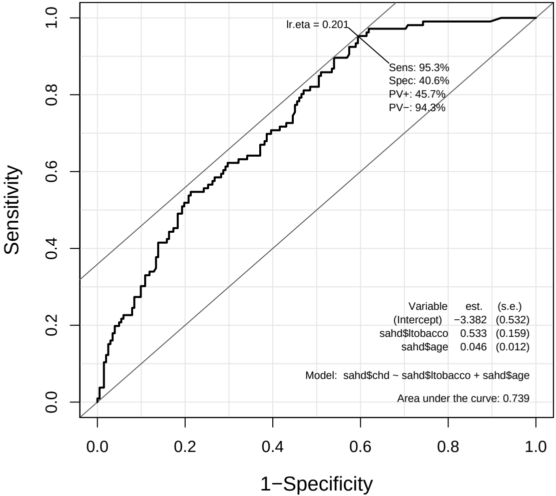
Pour un individu avec $X = x$

$$\begin{aligned} \text{Si } \beta_0 + x'\hat{\beta} \geq 0, &\quad \text{alors} \quad \hat{Y} = +1 \\ \text{Si } x'\hat{\beta} \geq -\beta_0, &\quad \text{alors} \quad \hat{Y} = +1 \end{aligned}$$

Choix de β_0

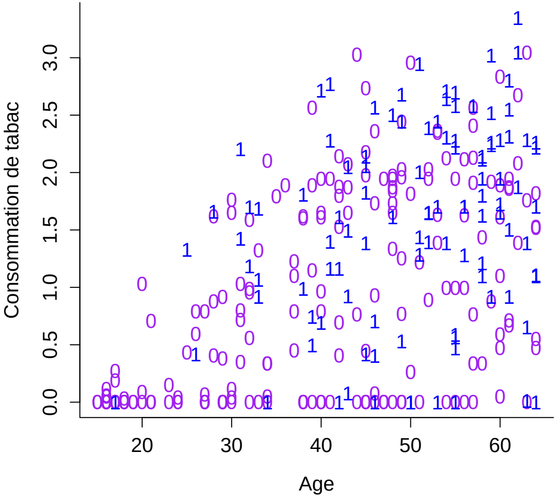
- Estimation par minimisation du coût quadratique
- Estimation par minimisation du risque empirique
- Compromis sensibilité/spécificité (Courbe ROC)

Règle linéaire d’affectation



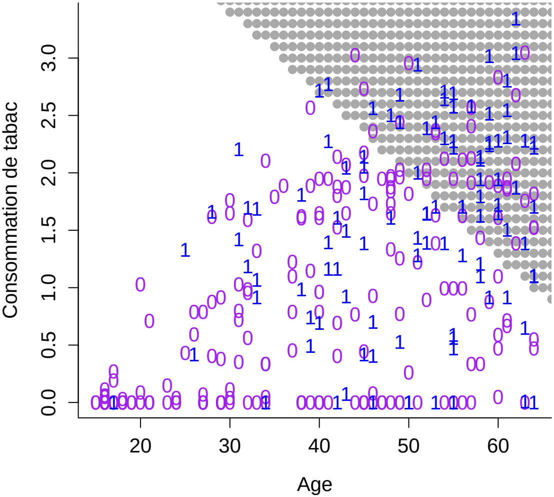
Classifications linéaire et quadratique

Modélisation du risque d'infarctus



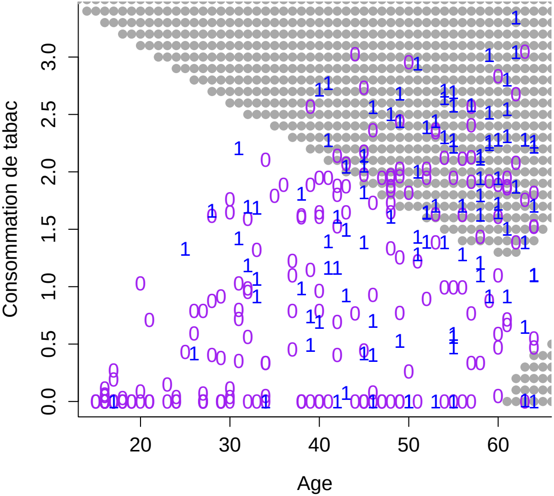
Classifications linéaire et quadratique

Modélisation du risque d'infarctus



Classifications linéaire et quadratique

Modélisation du risque d'infarctus



Plan du cours

1. Préambule
   - Objectifs
   - Règle de Bayes
   - Classification linéaire
2. Analyse discriminante linéaire de Fisher
   - Calcul des scores
   - Qualité de discrimination
3. Sélection de modèle
   - Principes
   - Cas de deux groupes
   - Cas de plusieurs groupes
   - Sélection d’un sous-ensemble de gènes
4. Perspectives

Principes

- Scores
  - Discriminants : scores moyens/classe aussi $\neq$ que possible
  - Synthétiques : combinaisons linéaires des variables
  - Non-redondants : non-covariants entre eux

Principes

- Scores
  - Discriminants : scores moyens/classe aussi $\neq$ que possible
  - Synthétiques : combinaisons linéaires des variables
  - Non-redondants : non-covariants entre eux
- Affectation d'une puce à un groupe
  - Positionnement des groupes : scores moyens par groupe
  - Affectation d'un individu à un groupe : le plus proche . . .

Construction du score le plus discriminant

- $L_1 = \beta_1 \frac{y_1 - \bar{y}_1}{S_1} + \beta_2 \frac{y_2 - \bar{y}_2}{S_2} + \dots$

  On choisit les coefficients β_j tels que

  Fisher (L_1˜Groupe) soit le plus grand possible

Construction du score le plus discriminant

- $L_1 = \beta_1 \frac{y_1 - \bar{y}_1}{S_1} + \beta_2 \frac{y_2 - \bar{y}_2}{S_2} + \dots$

  On choisit les coefficients β_j tels que

  Fisher (L_1˜Groupe) soit le plus grand possible

- Ici, Fisher (L_1˜Groupe) = 142.22

  (Meilleur Fisher : $F(C_2) = 56.52$)

Construction du score le plus discriminant

- $L_1 = \beta_1 \frac{y_1 - \bar{y}_1}{S_1} + \beta_2 \frac{y_2 - \bar{y}_2}{S_2} + \dots$

  On choisit les coefficients β_j tels que

  Fisher (L_1˜Groupe) soit le plus grand possible

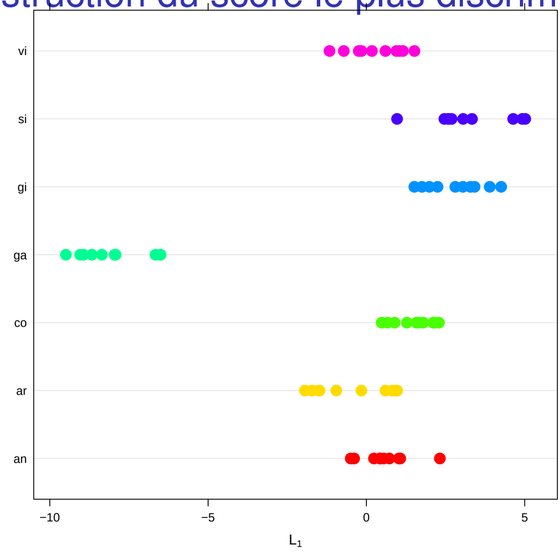
- Ici, Fisher (L_1˜Groupe) = 142.22

  (Meilleur Fisher : $F(C_2) = 56.52$)

- Interprétation par les corrélations avec les variables

C1	C2	C3	C4	C5	C6	C7	C8	C9	C10
0.16	0.9	-0.03	0.26	-0.11	-0.3	-0.13	0.63	0.83	0.88

Construction du score le plus discriminant



Constructions des scores suivants

- On choisit les coefficients de L_2 tels que

L_2 et L_1 non-covariants

&

Fisher (L_2˜Groupe) soit le plus grand possible

Constructions des scores suivants

- On choisit les coefficients de L_2 tels que

L_2 et L_1 non-covariants
&
Fisher (L_2˜Groupe) soit le plus grand possible

- Ici, Fisher (L_2˜Groupe) = 76.77

Constructions des scores suivants

- On choisit les coefficients de L_2 tels que

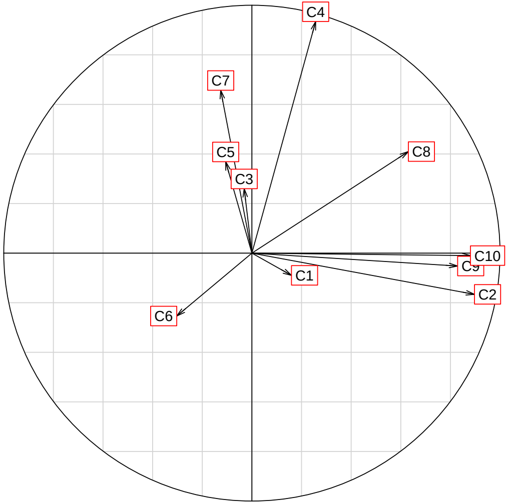
L_2 et L_1 non-covariants

&

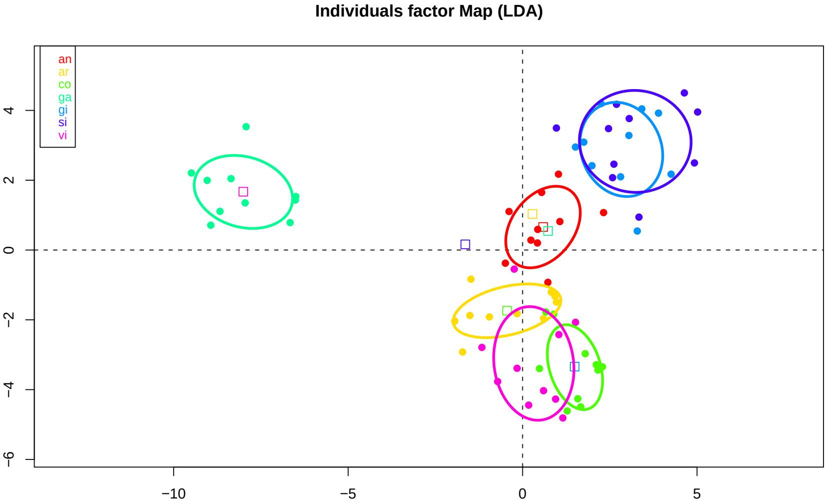
Fisher (L_2˜Groupe) soit le plus grand possible

- Ici, Fisher (L_2˜Groupe) = 76.77
- Interprétation par les corrélations avec les variables

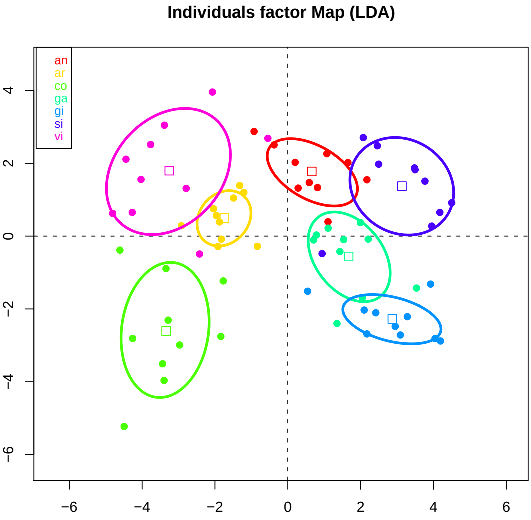
Constructions des scores suivants



Constructions des scores suivants



Constructions des scores suivants



Matrice de confusion

Matrice de confusion

Discriminante : modèle à une dimension

	Observé						
Prédit	an	ar	co	ga	gi	si	vi
an	3	2	1	0	1	0	3
ar	4	6	0	0	0	0	0
co	3	0	5	0	2	0	0
ga	0	0	0	10	0	0	0
gi	0	0	3	0	2	5	0
si	1	0	0	0	4	5	0
vi	3	4	2	0	0	0	1

Matrice de confusion

Matrice de confusion

Discriminante : modèle à deux dimensions

	Observé						
Prédit	an	ar	co	ga	gi	si	vi
an	8	2	0	0	0	0	0
ar	0	10	0	0	0	0	0
co	0	2	7	0	0	0	1
ga	0	0	0	10	0	0	0
gi	0	0	0	0	6	4	0
si	0	0	0	0	5	5	0
vi	0	2	4	0	0	0	4

Matrice de confusion

Matrice de confusion

Discriminante : modèle à trois dimensions

	Observé						
Prédit	an	ar	co	ga	gi	si	vi
an	10	0	0	0	0	0	0
ar	0	10	0	0	0	0	0
co	0	1	9	0	0	0	0
ga	0	0	0	10	0	0	0
gi	0	0	0	0	10	0	0
si	0	0	0	0	1	9	0
vi	1	2	0	0	0	0	7

Plan du cours

1. Préambule
   - Objectifs
   - Règle de Bayes
   - Classification linéaire

2. Analyse discriminante linéaire de Fisher
   - Calcul des scores
   - Qualité de discrimination

3. Sélection de modèle
   - Principes
   - Cas de deux groupes
   - Cas de plusieurs groupes
   - Sélection d'un sous-ensemble de gènes

4. Perspectives

Selection de modèles : objectifs

Réduction de l'information prédictrice

- Objectif cognitif : *Quels gènes sont discriminants ?*
  - Choix de variables
  - Sélection d'un sous-modèle du modèle complet
- Objectif de précision : *Comment prédire au mieux ?*
  - Minimiser l'erreur de prédiction
  - Tenir compte de la redondance de l'information prédictrice

Modèle de régression logistique

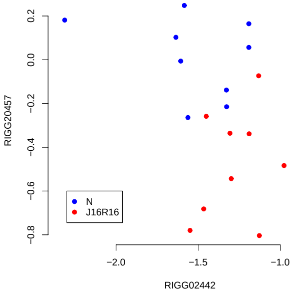
y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

$$\text{logodds}(y) \quad = \quad \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \ldots + \beta_p y_p$$

Modèle de régression logistique

y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

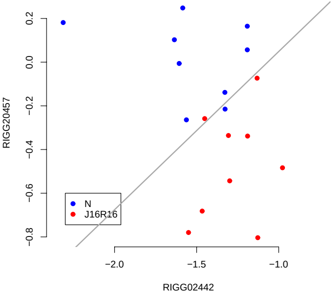
$$\text{logodds}(y) \;=\; \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \ldots + \beta_p y_p > 0$$



Modèle de régression logistique

y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

$$\text{logodds}(y) \quad = \quad \beta_0 + \beta_1 y_1 + \beta_2 y_2 + \ldots + \beta_p y_p > 0$$



Ajustement d'un modèle

x = +1 (si État = N), x = -1 (si État = J16R16)

Déviance, proximité entre observations et modèle :

$$\mathcal{D} = \sum_{i=1}^{n} \log\left\{1 + \exp\left[-x_i\left(\beta_0 + \beta_1 y_i\right)\right]\right\}$$

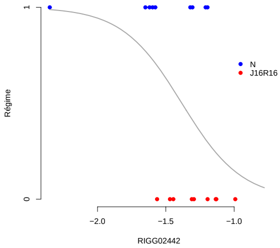
Ajustement d'un modèle

x = +1 (si État = N), x = -1 (si État = J16R16)

Déviance, proximité entre observations et modèle :

$$\mathcal{D} = \sum_{i=1}^{n} \log\left\{1 + \exp\left[-x_i\left(\beta_0 + \beta_1 y_i\right)\right]\right\}$$

Minimisation de la déviance



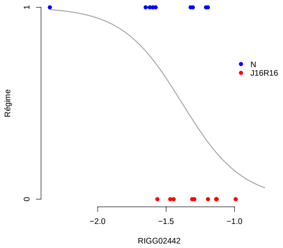
Ajustement d'un modèle

x = +1 (si État = N), x = -1 (si État = J16R16)

Déviance, proximité entre observations et modèle :

$$\mathcal{D} = \sum_{i=1}^{n} \log\left\{1 + \exp\left[-x_i\left(\beta_0 + \beta_1 y_i\right)\right]\right\}$$

Minimisation de la déviance



Coefficients:

	Estimate	Std. Error
(Intercept)	-6.317	3.848
RIGG02442	-4.569	2.795

—

Residual deviance: 20.795, AIC=24.795

Number of Fisher Scoring iterations: 5

Analyse de la déviance

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			17	24.9533	
RIGG02442	1	4.1629	16	20.7904	0.0413

Généralisation à plusieurs gènes

y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

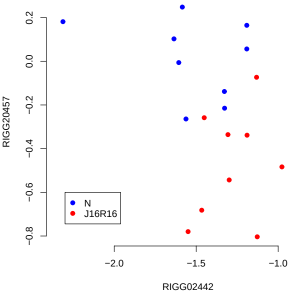
$$\text{logodds}(y) \quad = \quad \beta_0 + \beta_1 y_1 + \beta_2 y_2$$

Généralisation à plusieurs gènes

y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

$$\text{logodds}(y) \;=\; \beta_0 + \beta_1 y_1 + \beta_2 y_2$$

Ajustement

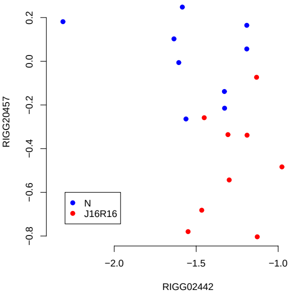


Généralisation à plusieurs gènes

y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

$$\text{logodds}(y) = \beta_0 + \beta_1 y_1 + \beta_2 y_2$$

Ajustement



Coefficients:

	Estimate	Std. Error
(Intercept)	-21.17	17.09
RIGG02442	-19.84	15.64
RIGG20457	27.38	24.74

—

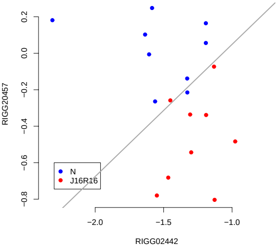
Residual deviance: 5.687, AIC=11.687
Number of Fisher Scoring iterations: 9

Généralisation à plusieurs gènes

y_k expression du $k$ème gène et $y = [y_1, \ldots, y_p]$

$$\text{logodds}(y) \quad = \quad \beta_0 + \beta_1 y_1 + \beta_2 y_2$$

Ajustement



Coefficients:

	Estimate	Std. Error
(Intercept)	-21.17	17.09
RIGG02442	-19.84	15.64
RIGG20457	27.38	24.74

—

Residual deviance: 5.687, AIC=11.687
Number of Fisher Scoring iterations: 9

Analyse de la déviance

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			17	24.9533	
RIGG02442	1	4.1629	16	20.7904	0.0413
RIGG20457	1	15.1038	15	5.6866	0.0001

Modèle logistique multinomial

$$\log\left[\frac{\mathbb{P}(\text{État} = \text{J16})}{\mathbb{P}(\text{État} = \text{N})}\right] = \beta_0^{(1)} + \beta_1^{(1)} y$$

$$\log\left[\frac{\mathbb{P}(\text{État} = \text{J16R5})}{\mathbb{P}(\text{État} = \text{N})}\right] = \beta_0^{(2)} + \beta_1^{(2)} y$$

$$\log\left[\frac{\mathbb{P}(\text{État} = \text{J16R16})}{\mathbb{P}(\text{État} = \text{N})}\right] = \beta_0^{(3)} + \beta_1^{(3)} y$$

$$\log\left[\frac{\mathbb{P}(\text{État} = \text{J48})}{\mathbb{P}(\text{État} = \text{N})}\right] = \beta_0^{(4)} + \beta_1^{(4)} y$$

$$\log\left[\frac{\mathbb{P}(\text{État} = \text{J48R24})}{\mathbb{P}(\text{État} = \text{N})}\right] = \beta_0^{(5)} + \beta_1^{(5)} y$$

Ajustement

Coefficients:			Std. Errors:	
	(Intercept)	RIGG02442	(Intercept)	RIGG02442
J16	3.67	2.77	3.15	2.22
J16R5	25.03	41.52	10.10	24.12
J16R16	5.10	3.67	3.14	2.25
J48	4.78	70.38	22.91	38.21
J48R24	16.36	15.12	5.51	5.28

Residual Deviance: 62.22

AIC: 82.22

Analyse de déviance

	Df	Deviance	Resid. Df	Resid. Dev	P(>|Chi|)
NULL			240	174.99171	
RIGG02442	5	112.7688	235	62.22296	0.000

Sélection selon le critère d'Akaike

Sélection ascendante

- 1ère étape : on choisit le gène G_1 tel que

  AIC (Regime~G_1) soit le plus faible

Gène	RIGG02442	RIGG02858	RIGG13051
AIC	82.22	93.96	96.82

Sélection selon le critère d'Akaike

Sélection ascendante

- 1ère étape : on choisit le gène G_1 tel que

$$\text{AIC} \;(\; \text{Regime}\tilde{}G_1 \;)\; \text{soit le plus faible}$$

Gène	RIGG02442	RIGG02858	RIGG13051
AIC	82.22	93.96	96.82

- 2ème étape : on choisit le gène G_2 tel que

$$\text{AIC} \;(\; \text{Regime}\tilde{}G_1 + G_2 \;)\; \text{soit le plus faible}$$

Gène	RIGG16958	RIGG08709	RIGG03143
AIC	30.16	31.82	32.57

Sélection selon le critère d'Akaike

Sélection ascendante

- 1ère étape : on choisit le gène G_1 tel que

  AIC (Regime˜G_1) soit le plus faible

Gène	RIGG02442	RIGG02858	RIGG13051
AIC	82.22	93.96	96.82

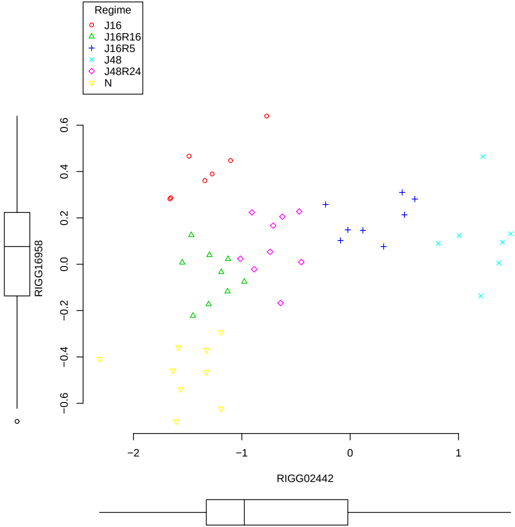
- 2ème étape : on choisit le gène G_2 tel que

  AIC (Regime˜$G_1 + G_2$) soit le plus faible

Gène	RIGG16958	RIGG08709	RIGG03143
AIC	30.16	31.82	32.57

- 3ème étape : on choisit le gène G_3 tel que ...

Meilleur modèle (2 gènes)



Plan du cours

1 Préambule
Objectifs
Règle de Bayes
Classification linéaire

2 Analyse discriminante linéaire de Fisher
Calcul des scores
Qualité de discrimination

3 Sélection de modèle
Principes
Cas de deux groupes
Cas de plusieurs groupes
Sélection d'un sous-ensemble de gènes

4 Perspectives

Perspectives

Stratégie de discrimination

- Exploration : analyse discriminante de Fisher
- Règle de diagnostic : modèle logistique

Perspectives

Stratégie de discrimination

- Exploration : analyse discriminante de Fisher
- Règle de diagnostic : modèle logistique

Attention

- Sélection ascendante : plusieurs modèles équivalents
- Validation externe nécessaire