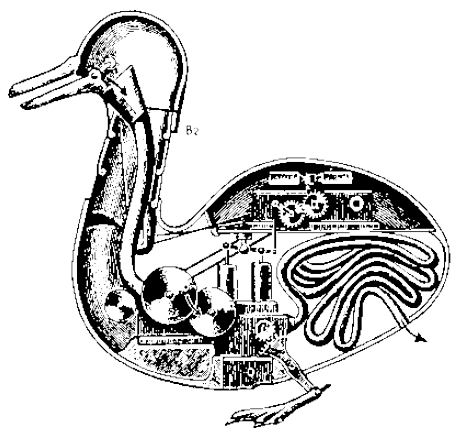
Modèles probabilistes pour l'apprentissage

le hasard au service des algorithmes

olivier.francois@grenoble-inp.fr olivier.francois@imag.fr

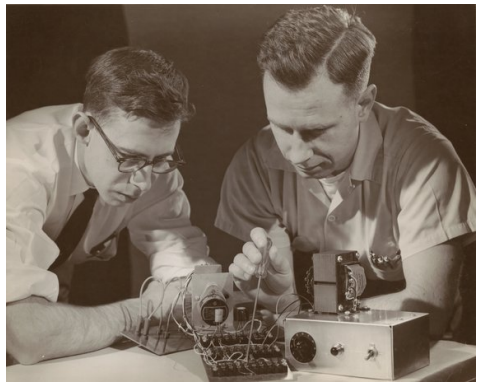
Qu'entend-t-on par apprentissage?

- Méthodes qui permettent à une « machine » d'évoluer grâce à un processus automatique
- Principe : Modéliser la réponse d'un système à partir des variables d'entrées et des données empiriques recueillies pour ce système
- Adaptatif et flexible, prenant en compte l'évolution de la base d'information des comportements enregistrés



Histoire

- 1943 : McCulloch et Pitts, Turing, von Neuman, considérent les fonctions mentales comme des fonctions mathématiques auto-régulées.
  A Logical Calculus of Ideas Immanent in Nervous Activity, 1943, Bull. Math. Biophys. 5:115-133.
- 1957 : Perceptron de Rosenblatt (Mark 1) “The fundamental laws of organization which are common to all information handling systems, machines and men included, may eventually be understood”



Applications

- Reconnaissance des formes, reconnaissance vocale, vision par ordinateur
- Aide au diagnostic, détection de fraude, d'anomalies, de spams, analyse financière, bio-informatique
- Sites web adaptatifs : Recommandation de produits (Amazon, Netflix)
- Moteurs de recherche

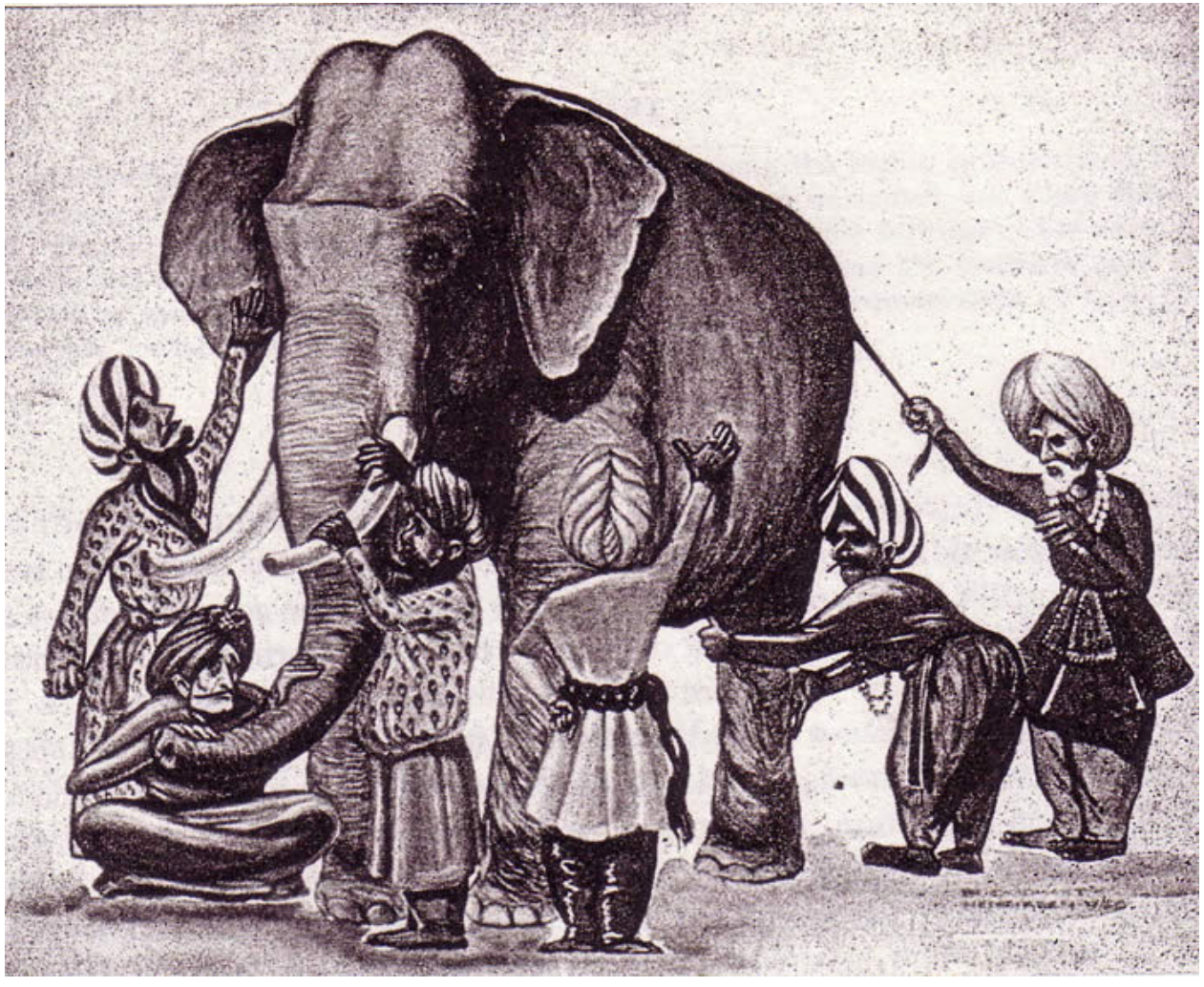
Contenu de cet exposé

- Principes d'analyse bayésienne
- Recherche d'un objet
- Estimation d'une probabilité
- Reconnaître des groupes dans les données
- Prédire vos préférences cinématographiques

Objectifs d'une analyse bayésienne

- Apprentissage : Mettre à jour l'information disponible sur un phénomène à partir d'observations de ce phénomène
- Quantifier l'incertitude sur les paramètres expliquant le phénomène
- Prédire les valeurs de nouvelles données et évaluer l'incertitude de cette prédiction

The blind men and the elephant



Lien avec “la” statistique

- En statistique, l'objectif est généralement de tester une hypothèse et de déterminer la zone de rejet du test
- L'incertitude sur le paramètre à tester n'est pas modélisée. La seule source d'incertitude prise en compte provient de l'échantillonnage.

La formule de Bayes

- Soit A, B des événements de probabilité non-nulle

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\mathrm{P}(B)}{\mathrm{P}(A)} \propto \mathrm{P}(A|B)\mathrm{P}(B)$$

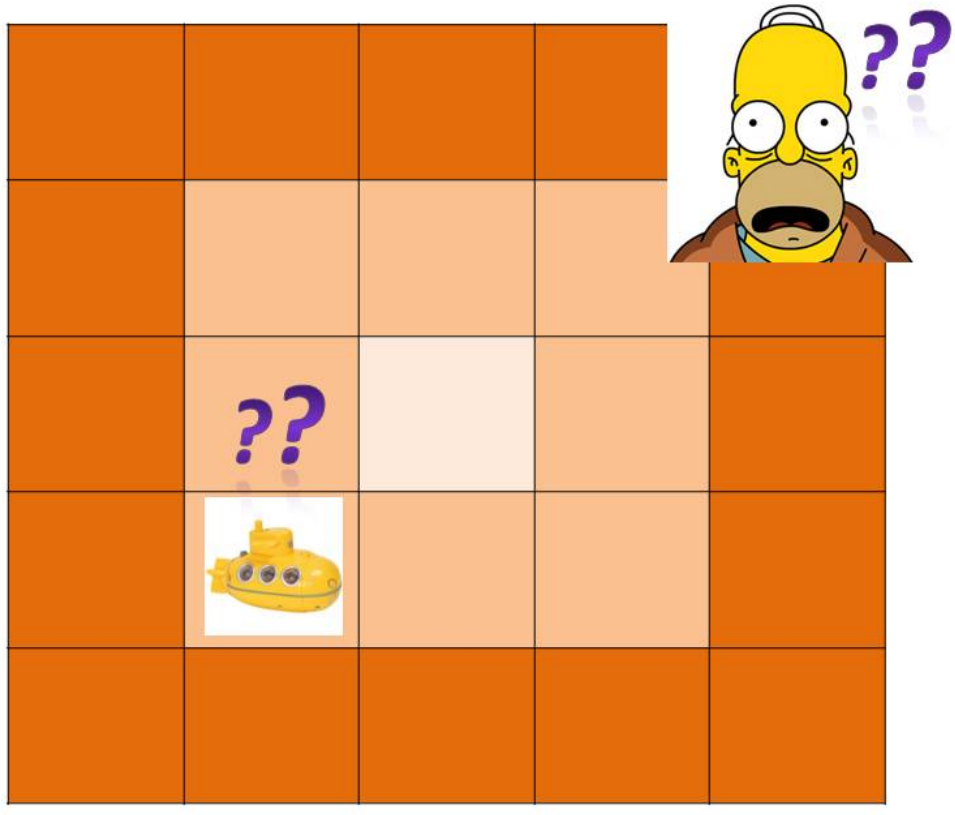
Exemple 1 : Recherche d'un objet

- Un exemple élémentaire où l'observation permet de (re)mettre à jour des connaissances a priori.

USS ne répond plus

- Uss Scorpion (SSN-589): un sous-marin nucléaire américain qui a coulé dans l'océan atlantique en 1968
- L'épave a été retrouvée au bout de 5 mois de recherche à l'aide de signaux sonores et d'une méthode de recherche bayésienne
- Principe ?

Recherche d'un objet



Cartes de probabilité

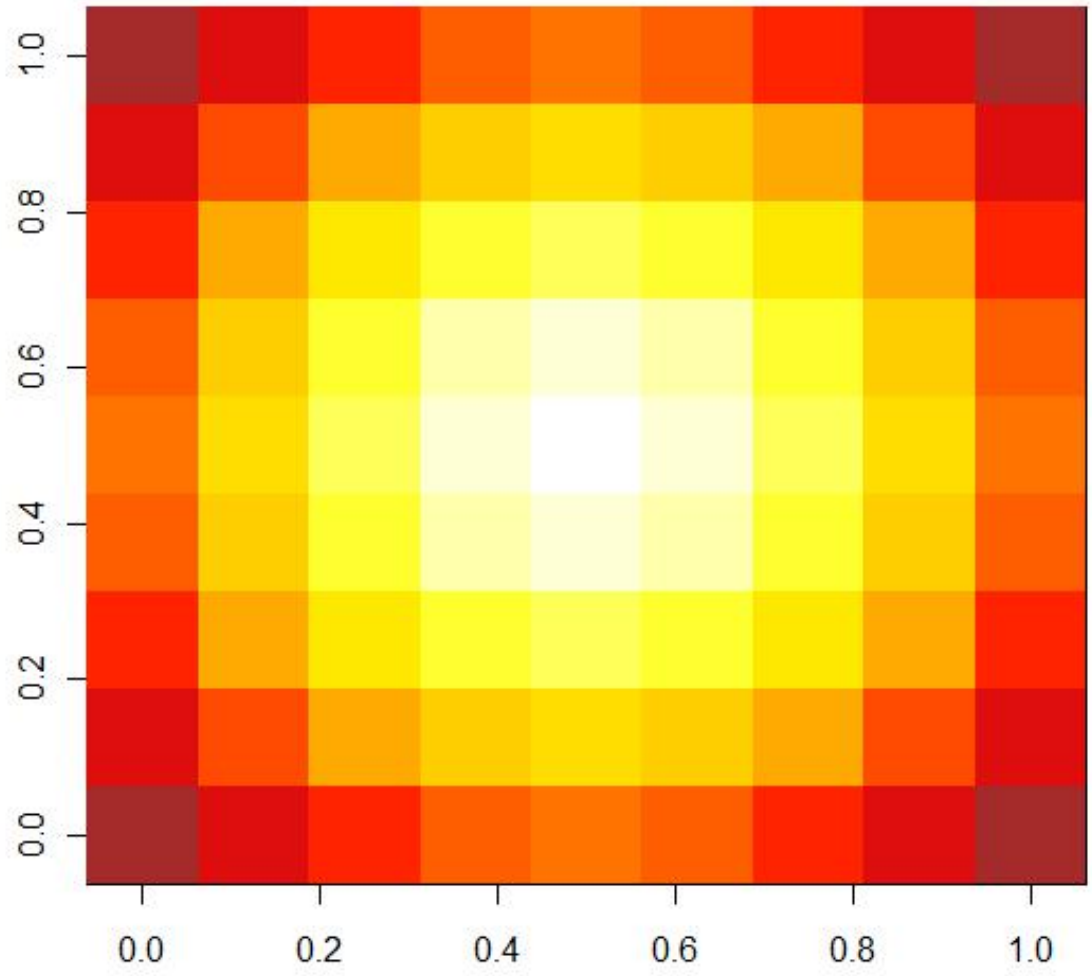
- Carte *a priori* :

$$p_i = \mathrm{P}(\text{Uss est dans la cellule } i)$$

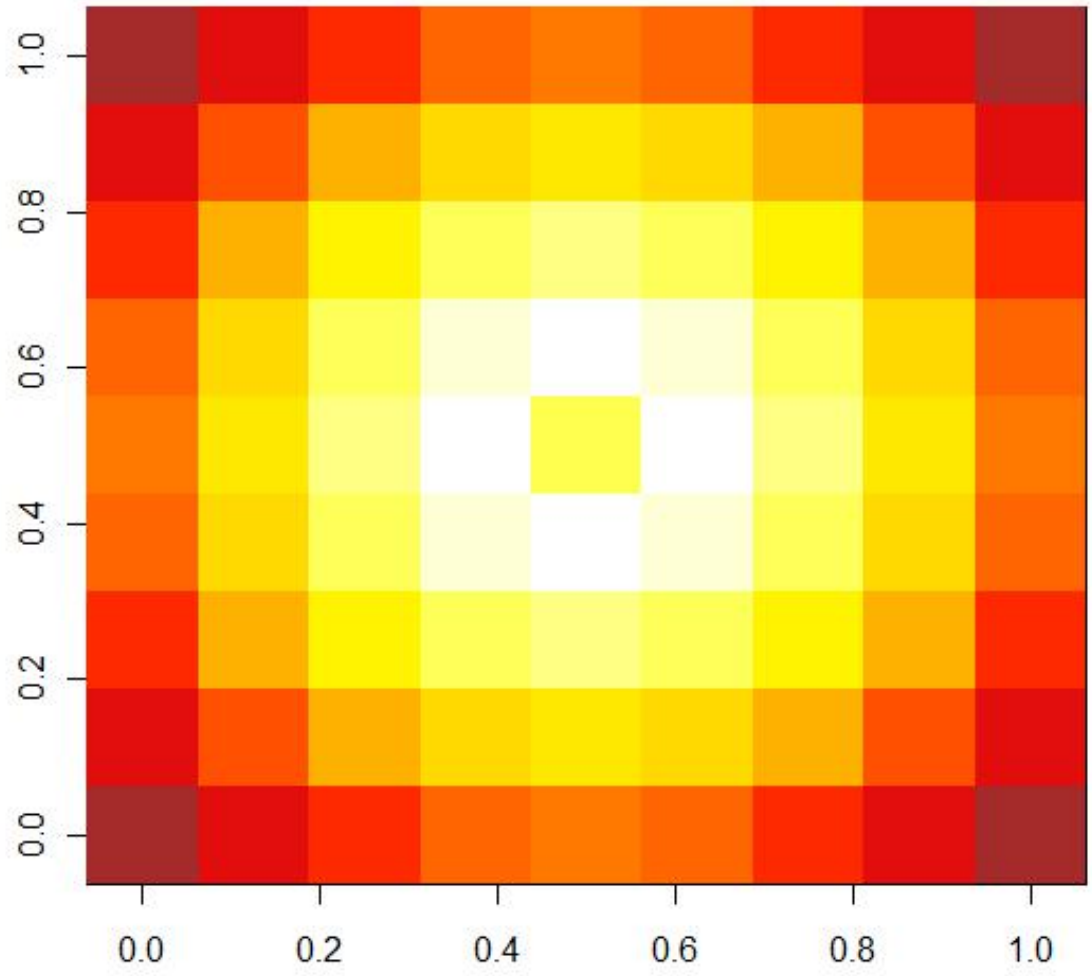
- Probabilité d'erreur de mesure $= 1 - q$
- Supposons que le résultat de la recherche dans la cellule i est infructueux
- Mise à jour des probabilités

$$\mathrm{P}(\text{Uss est dans la cellule } i | \text{Resultat negatif}) = \frac{(1-q)p_i}{1-qp_i}.$$

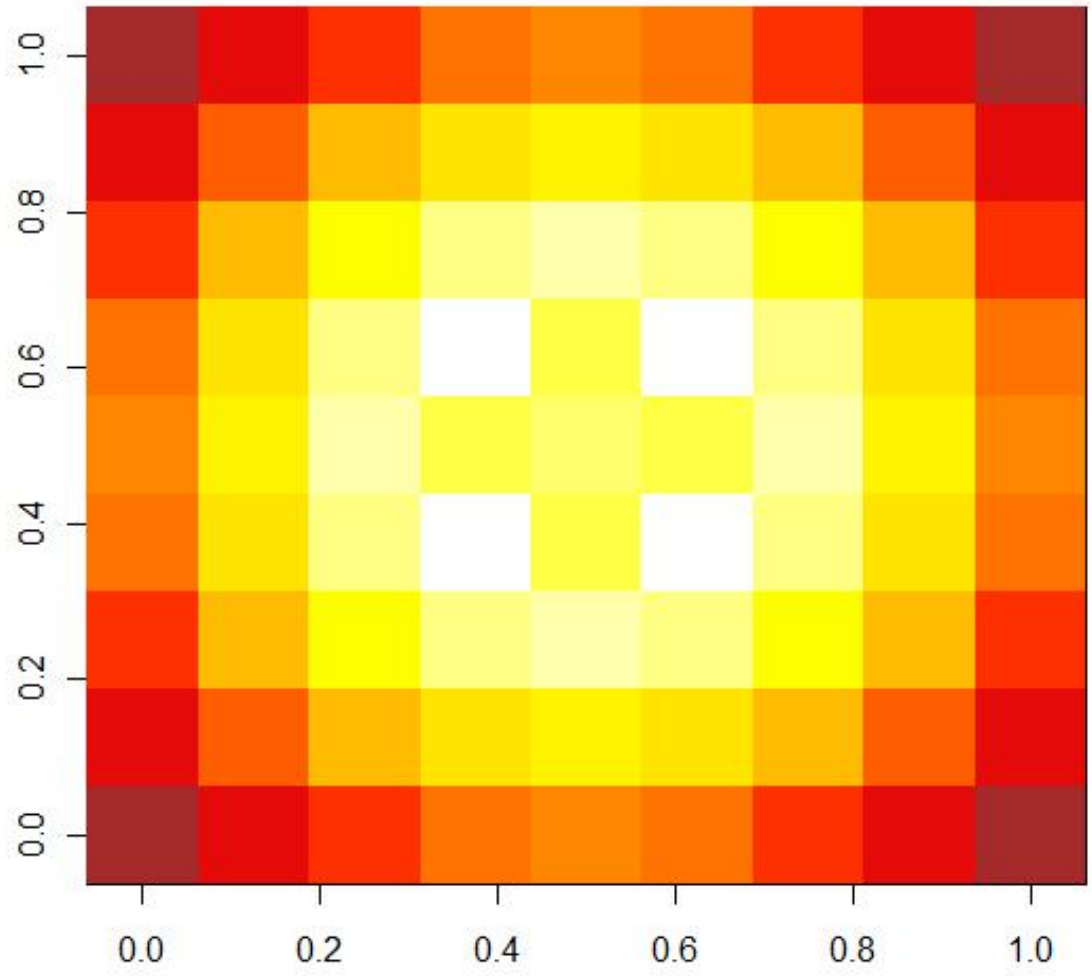
Jour 0



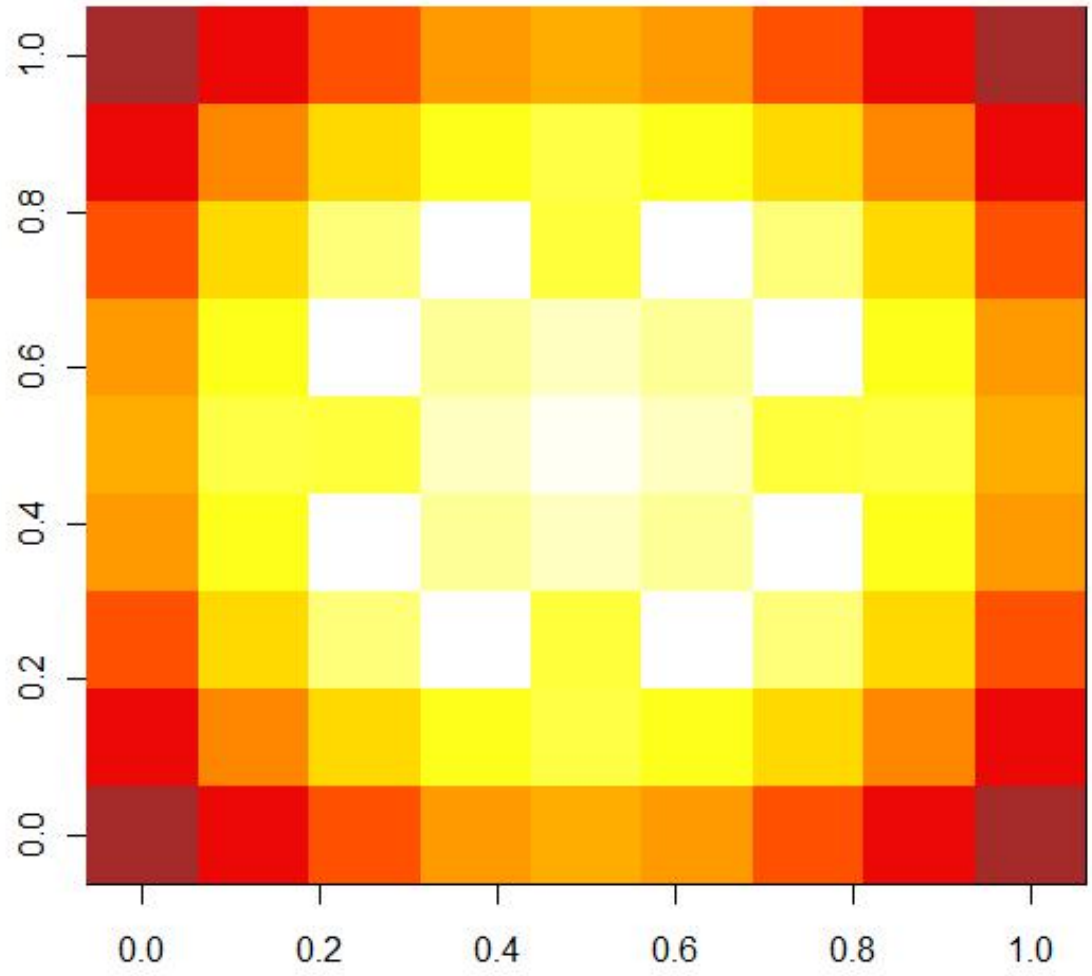
Jour 1



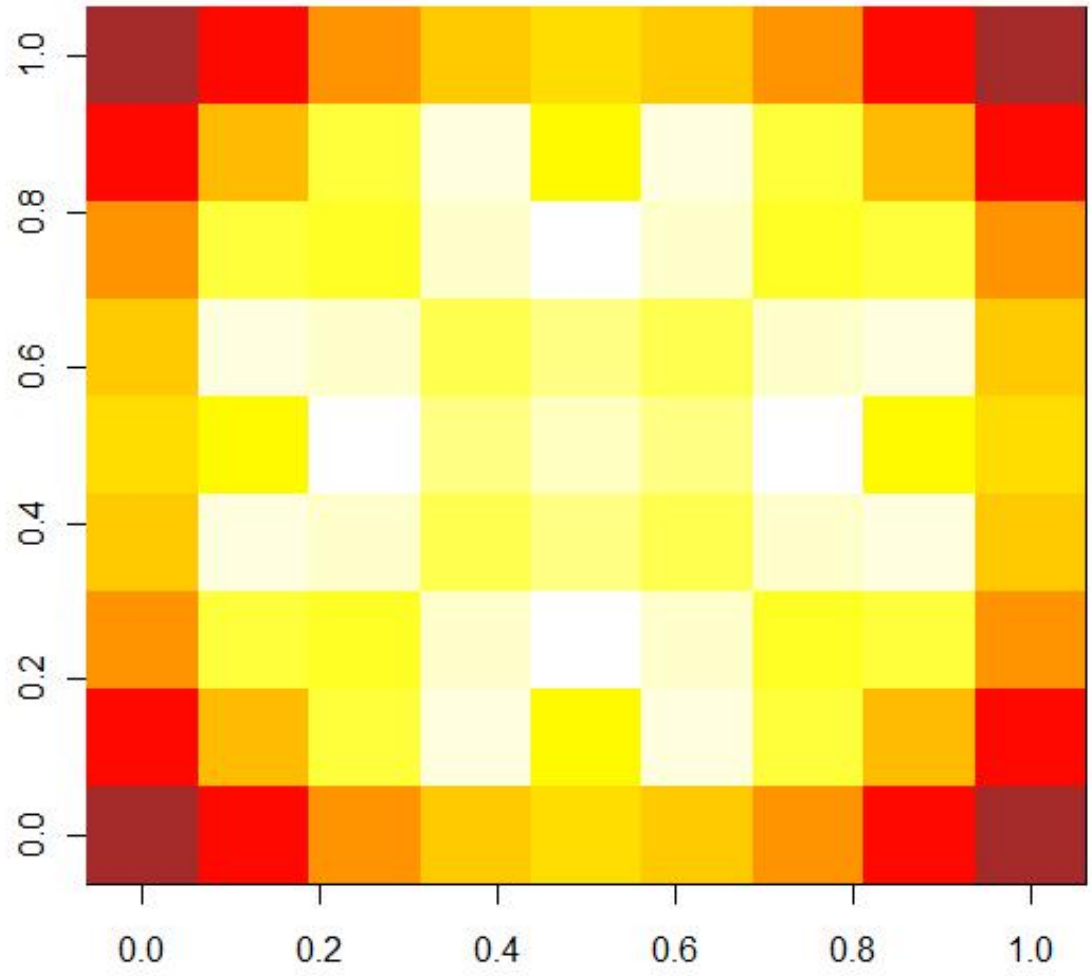
Jour 5



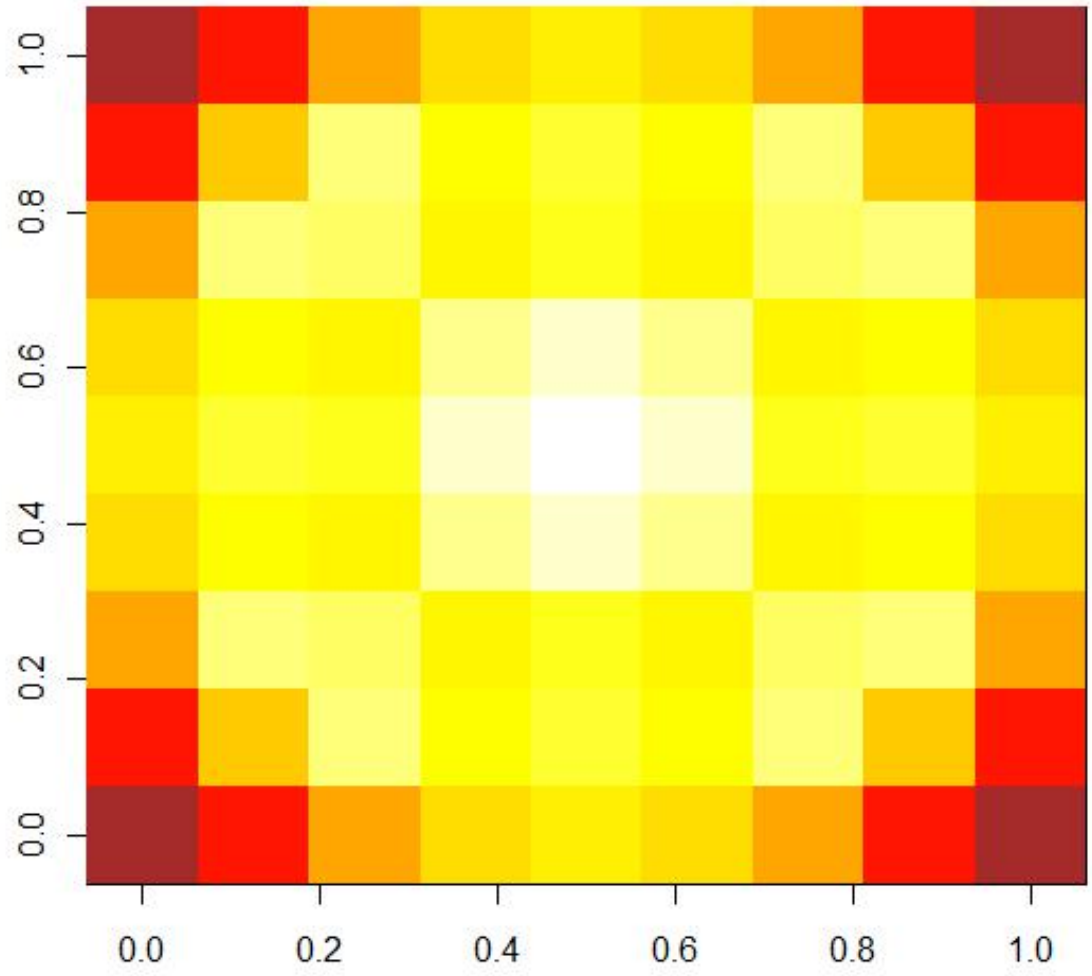
Jour 13



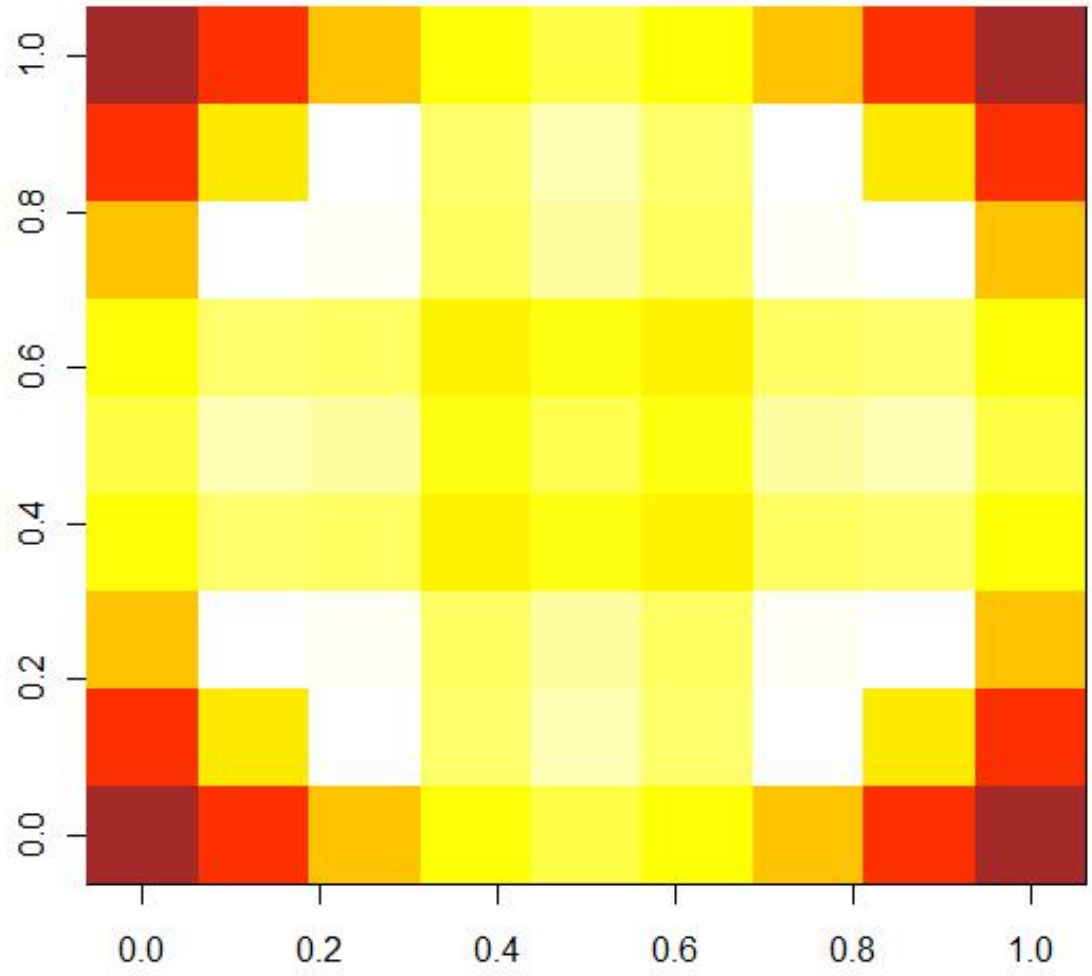
Jour 37



Jour 57



Jour 67



Modèle probabiliste pour l'analyse bayésienne

- Paramètre $\theta = (\theta_1, \ldots, \theta_J)$, $J \geq 1$.
- Données $y = (y_1, \ldots, y_n)$, $n \geq 1$.
- Un modèle est décrit par une loi de probabilité jointe :

$$p(y, \theta) = p(y|\theta)p(\theta)$$

Modèle probabiliste pour l'analyse bayésienne

- $p(\theta)$ est la loi *a priori*. Elle peut être informative ou non-informative.
- $p(y|\theta)$ est la vraisemblance. Elle décrit la manière dont sont générées les données observées dans le modèle.

$$p(y, \theta) = p(y|\theta)p(\theta)$$

Inférence bayésienne

- On utilise la formule de Bayes pour calculer la loi *a posteriori*

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

où $p(y) = \int p(y|\theta)p(\theta)d\theta$ est la loi marginale.

- En général, on cherche à éviter le calcul de la loi marginale et on écrit

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

Exemple 2 : Sondage

- Une source émet des signaux binaires, x_i, et on cherche à estimer la probabilité d'émettre un 1.
- On note cette probabilité θ

$$p(x_i = 1|\theta) = \theta$$

- Puis on observe une suite de signaux émis par la source : $1, 0, 1, \ldots$

Apprentissage : $x = 1, 0, 1, \cdots$

Etape 1:

$$p(\theta|x_1 = 1) \propto p(x_1 = 1|\theta)p(\theta) = \theta \quad (p(\theta) = 1)$$

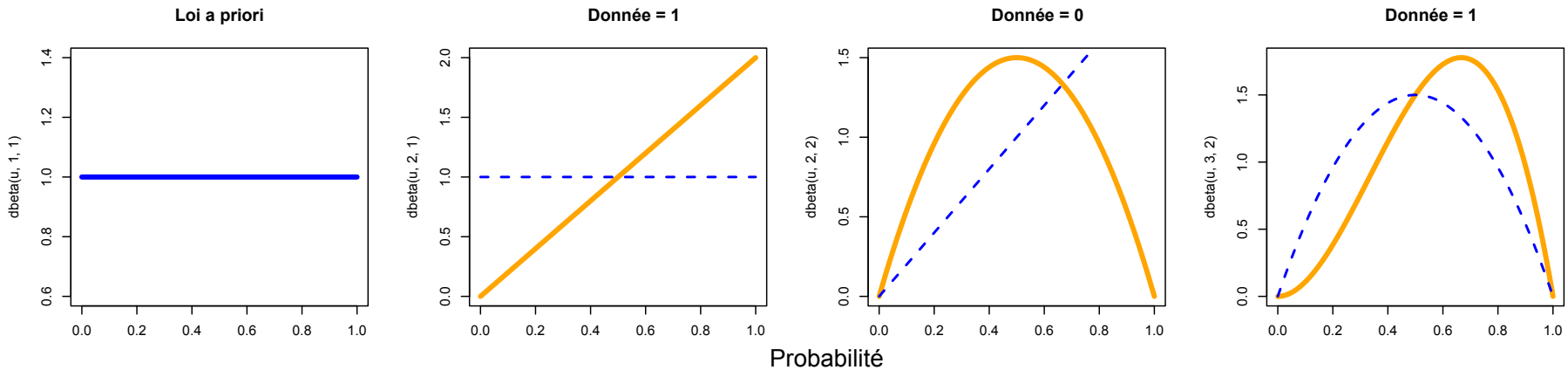
Etape 2:

$$p(\theta|x_2 = 0, x_1 = 1) \propto p(x_2 = 0|\theta)p(\theta|x_1 = 1) = (1 - \theta)\theta$$

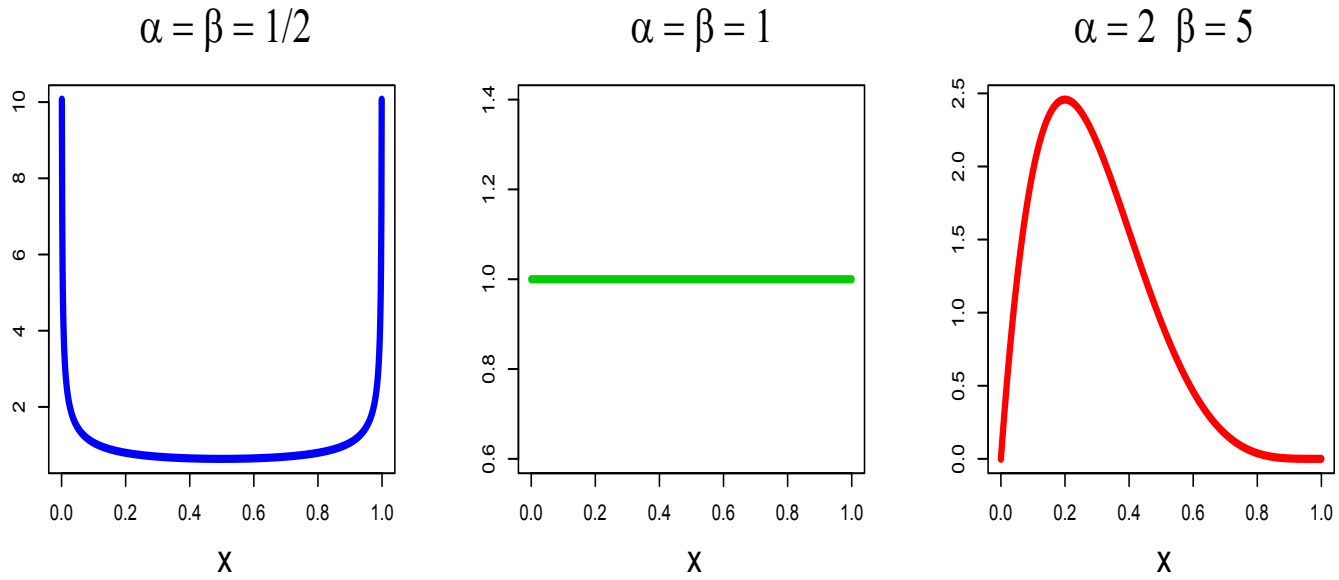
Etape 3:

$$p(\theta|x_3 = 1, x_2 = 0, x_1 = 1) \propto p(x_3 = 1|\theta)(1 - \theta)\theta = (1 - \theta)\theta^2$$

Apprentissage



Loi Beta$(\theta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}$



- Espérance et mode de la loi Beta

$$\mathrm{E}[\theta] = \frac{\alpha}{\alpha+\beta} \quad \mathrm{Mode}(\theta) = \frac{\alpha-1}{\alpha+\beta-2}$$

Modèle Beta-binomial

- La loi *a priori* est uniforme $\theta \sim$ beta(1,1).
- Données: On observe $y = \sum x_i = 9$ fois la valeur 1 dans un échantillon de $n = 20$ répétitions (fréquence = .45)
- Vraisemblance

$$p(y|\theta) = \text{binom}(n, \theta)(y) \propto \theta^y (1-\theta)^{n-y}$$

- Loi *a posteriori*

$$p(\theta|y) = \text{beta}(y+1, n+1-y)(\theta)$$

Remarques

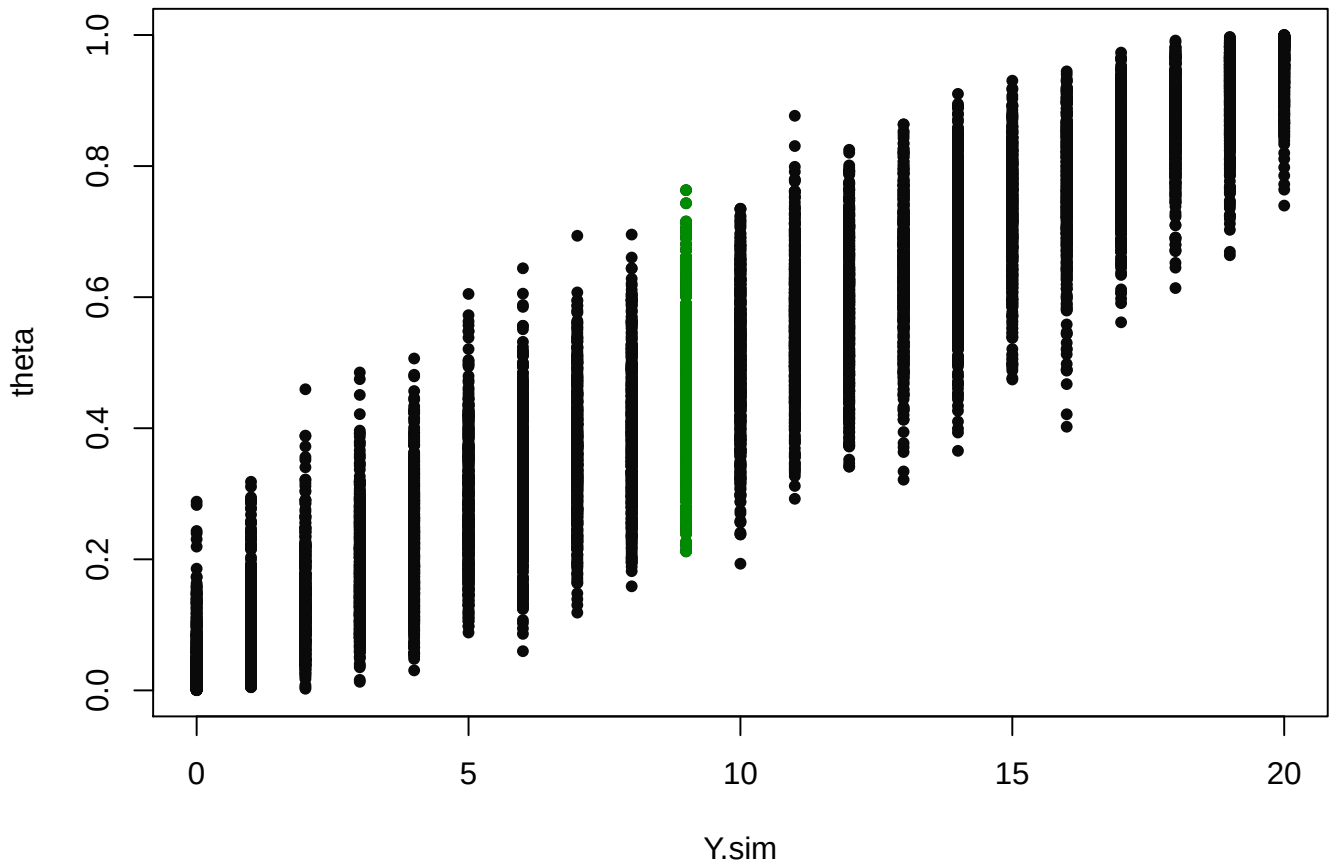
- Estimateur bayésien

$$\mathrm{E}[\theta|y] = \frac{y+1}{n+2} \approx \frac{y}{n}, \quad \text{lorsque } n \to \infty$$

- Intervalle de crédibilité, I, tel que $\Pr(\theta \in I|y) = .95$

$$I = (0.25, 0.65)$$

Loi jointe



Simuler la loi a posteriori

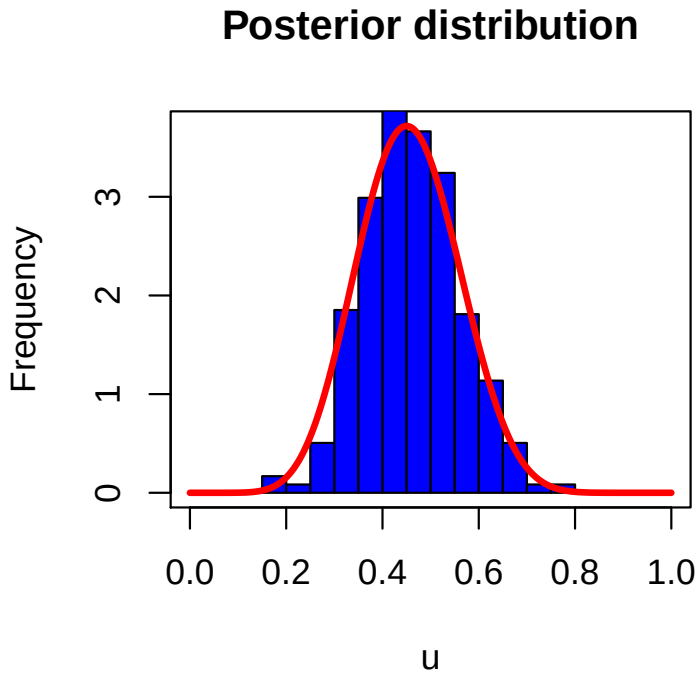
- Algorithme de rejet :

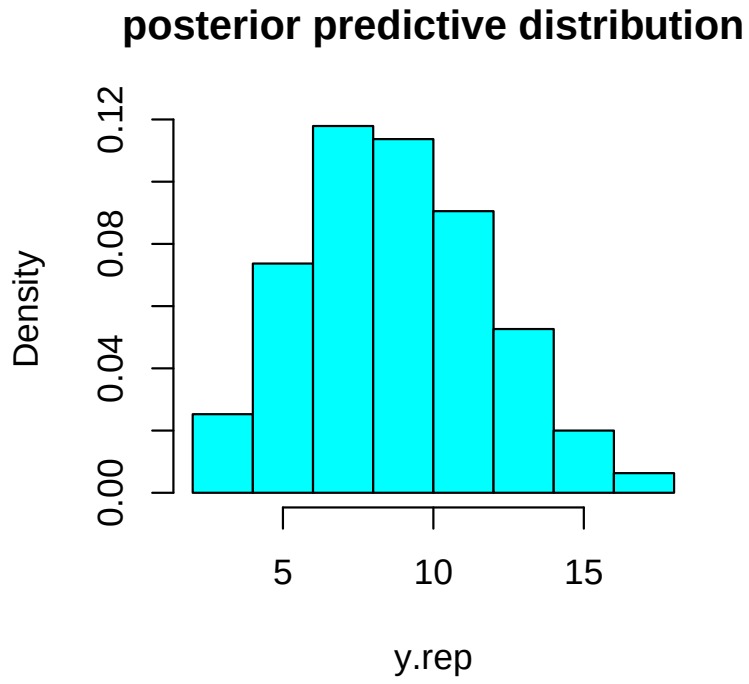
```
Repeat
theta <- unif(0,1)
y.s <- binom(n,theta)
Until (y.s == y)
return(theta)
```

- De manière générale, cet algorithme produit des simulations selon la loi *a posteriori* $p(\theta|y)$

$$p_y(\theta) = \sum_{s=1}^{\infty}(1-p(y))^{s-1}p(y,\theta) = p(\theta|y).$$

Résultat de la simulation





Exemple 3 : Classification non-supervisée

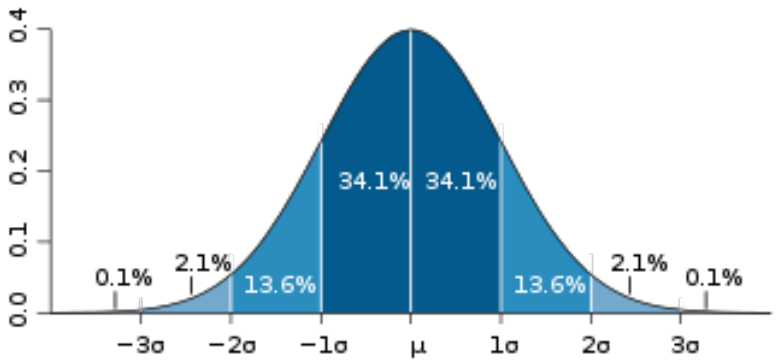
- Une observation, $y \in \mathbb{R}$, est supposée provenir d'un mélange de K sources.
- Combinaison convexe de densités. Soit $(p_k)_{k=1,\dots,K}$, tels que $\sum_{k=1}^{K} p_k = 1$, et $\theta = (\theta_k)_{k=1,\dots,K}$.
- On appelle mélange le modèle suivant

$$p(y|\theta) = \sum_{k=1}^{K} p_k p(y|\theta_k)$$

Lois gaussiennes

- Mélange de gaussiennes. Les composantes du mélanges sont des lois normales de paramètres $\theta_k = (m_k, \sigma_k^2)$

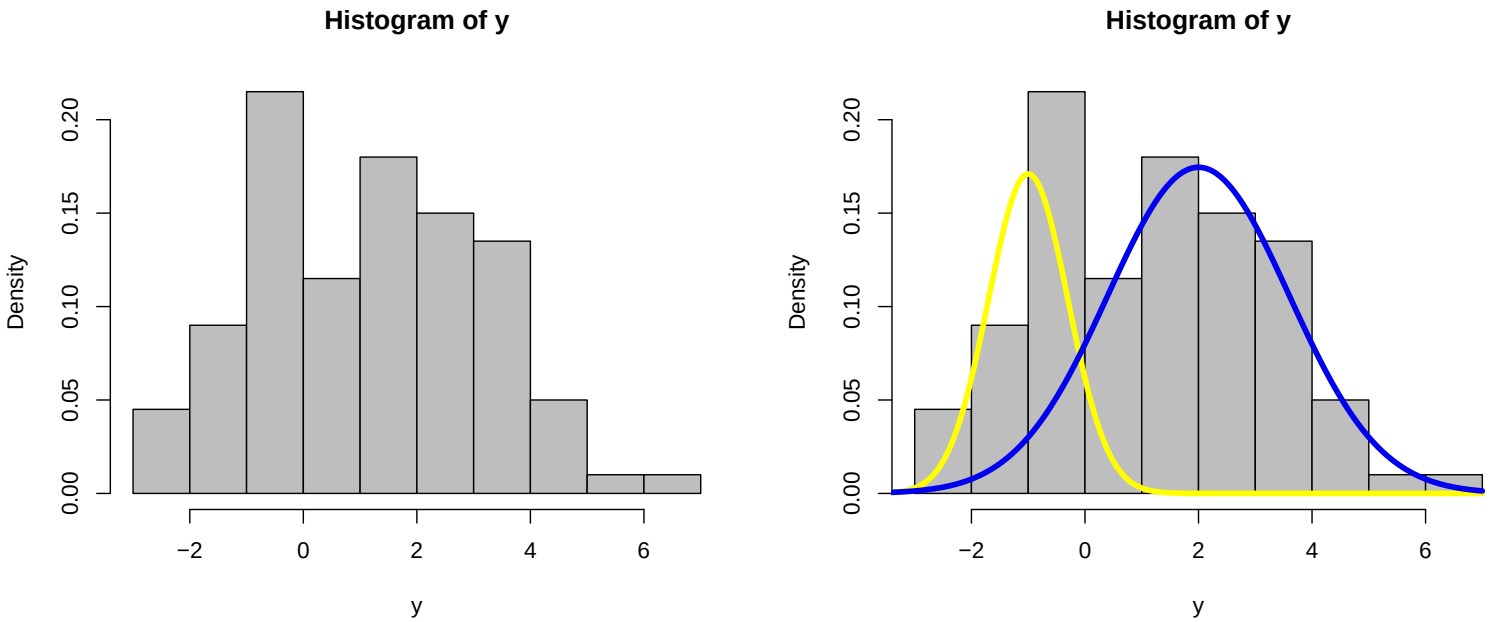
$$p(y|\theta_k) = N(m_k, \sigma_k^2)(y)$$



Questions principales

- Combien de groupes peut on distinguer dans les données ?
- Quelles sont les moyennes et les variances des groupes ?
- Pour un individu donné, quelle est la probabilité pour qu'il provienne du groupe k ?

Exemple de mélange ($K = 2$)



$$p_{\text{jaune}} = 30\% \quad p_{\text{bleu}} = 70\%$$

Modèle de mélange

- Pour un échantillon : $y = (y_1, \ldots, y_n)$
- Vraisemblance

$$p(y|\theta) = \prod_{i=1}^{n} \left(\sum_{k=1}^{K} p_k p(y_i|\theta_k) \right)$$

- Mélange gaussien: Pour tout $i = 1, \ldots, n$,

$$p(y_i|\theta_k) = N(m_k, \sigma_k^2)(y_i)\,.$$

Les mélanges vus comme modèles hiérarchiques

- Introduisons une variable latente (non observée) représentant une étiquette de classe $z_i \in \{1, \ldots, K\}$ pour tout y_i
- Pour tout $i = 1, \ldots, n$, considérons

$$p(y_i|\theta, z_i) = p(y_i|\theta_{z_i}) = N(m_{z_i}, \sigma^2_{z_i})(y_i)$$

- Avec cette représentation, nous retrouvons le mélange

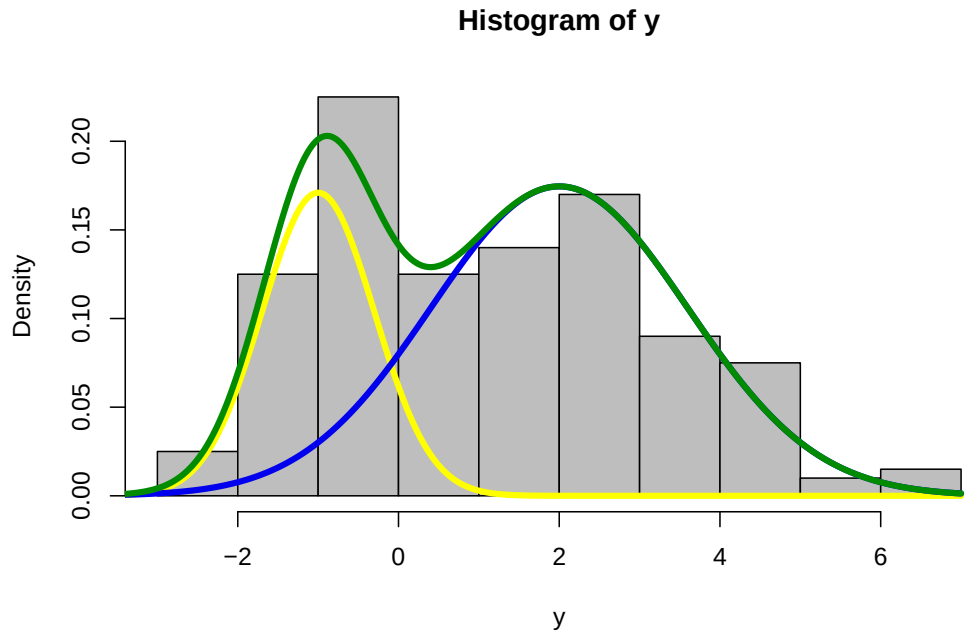
$$p(y_i|\theta) = \sum_{k=1}^{K} p(y_i|\theta_k) p(z_i = k)$$

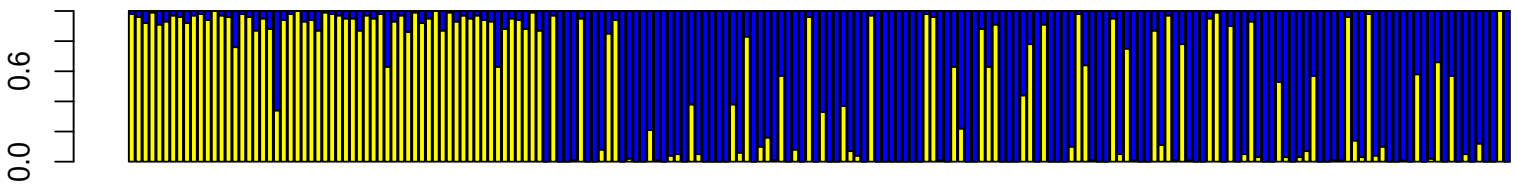
Classification bayésienne

- Classer consiste à estimer l'étiquette de classe z_i, pour tout i.
- On atteint cet objectif en calculant la loi marginale a posteriori $p(z_i|y)$,
- à partir de la loi jointe a posteriori

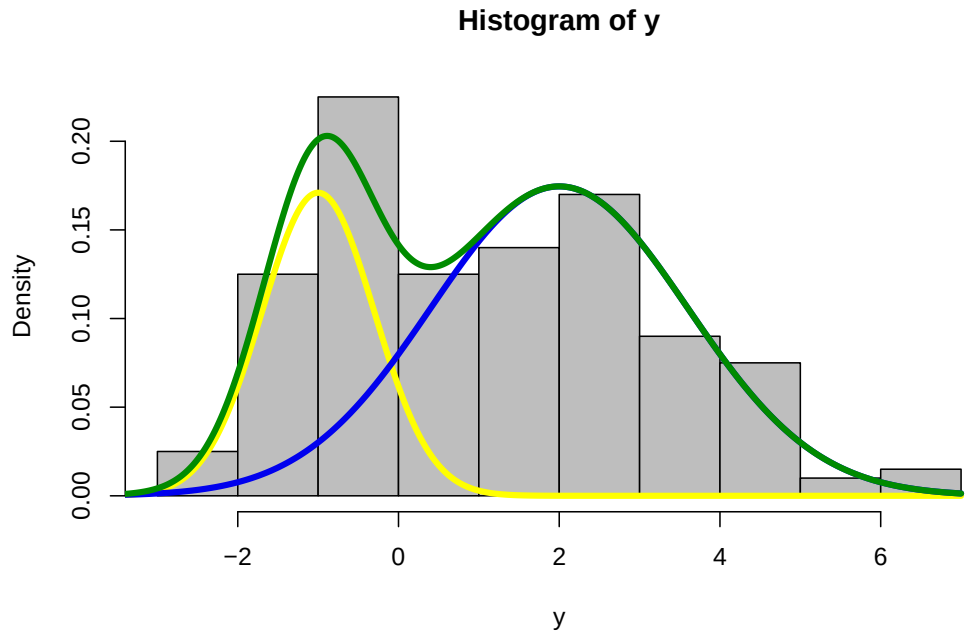
$$p(\theta, z|y) \propto p(y|\theta, z)p(\theta)p(z)$$

Lois marginales *a posteriori* des étiquettes de classe $p(z_i|y)$





Lois marginales *a posteriori* des étiquettes de classe $p(z_i|y)$



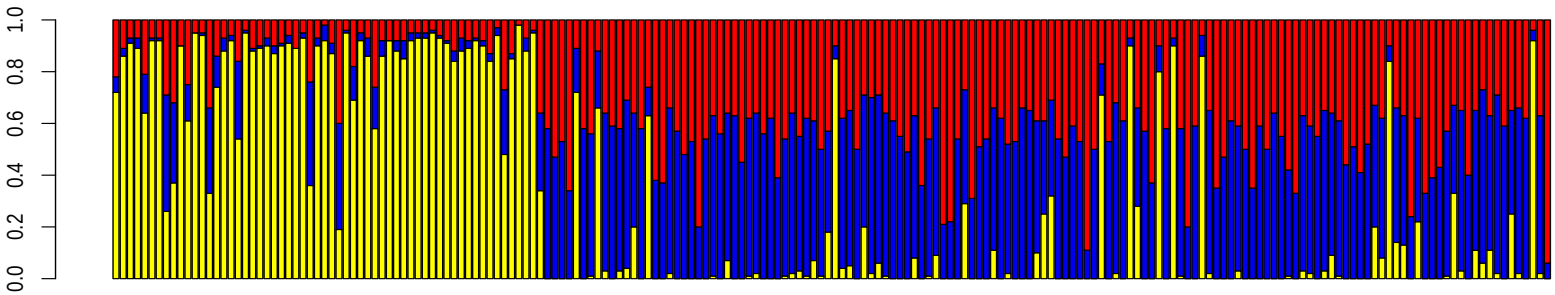


Illustration numérique : Iris de Fisher

Exemple 4 : Systèmes de recommandation

- Un ensemble de n utilisateurs peut accéder à L ressources (par exemple, des films) et indiquer ses appréciations au système
- Au fur et à mesure qu'un utilisateur entre ses appréciations, le système lui suggère de nouvelles ressources, susceptibles de lui plaire.
- Filtrage collaboratif
- Concours Netflix

Matrices de données

- n individus, L films, chacun noté par une partie des n individus seulement
- $y_{i\ell}$ = note de préférence entre 0 et 10, ou manquante.

	Mov1	Mov2	Mov3
ind1	10	8	-
ind2	2	-	4
...			

Factorisation de matrice

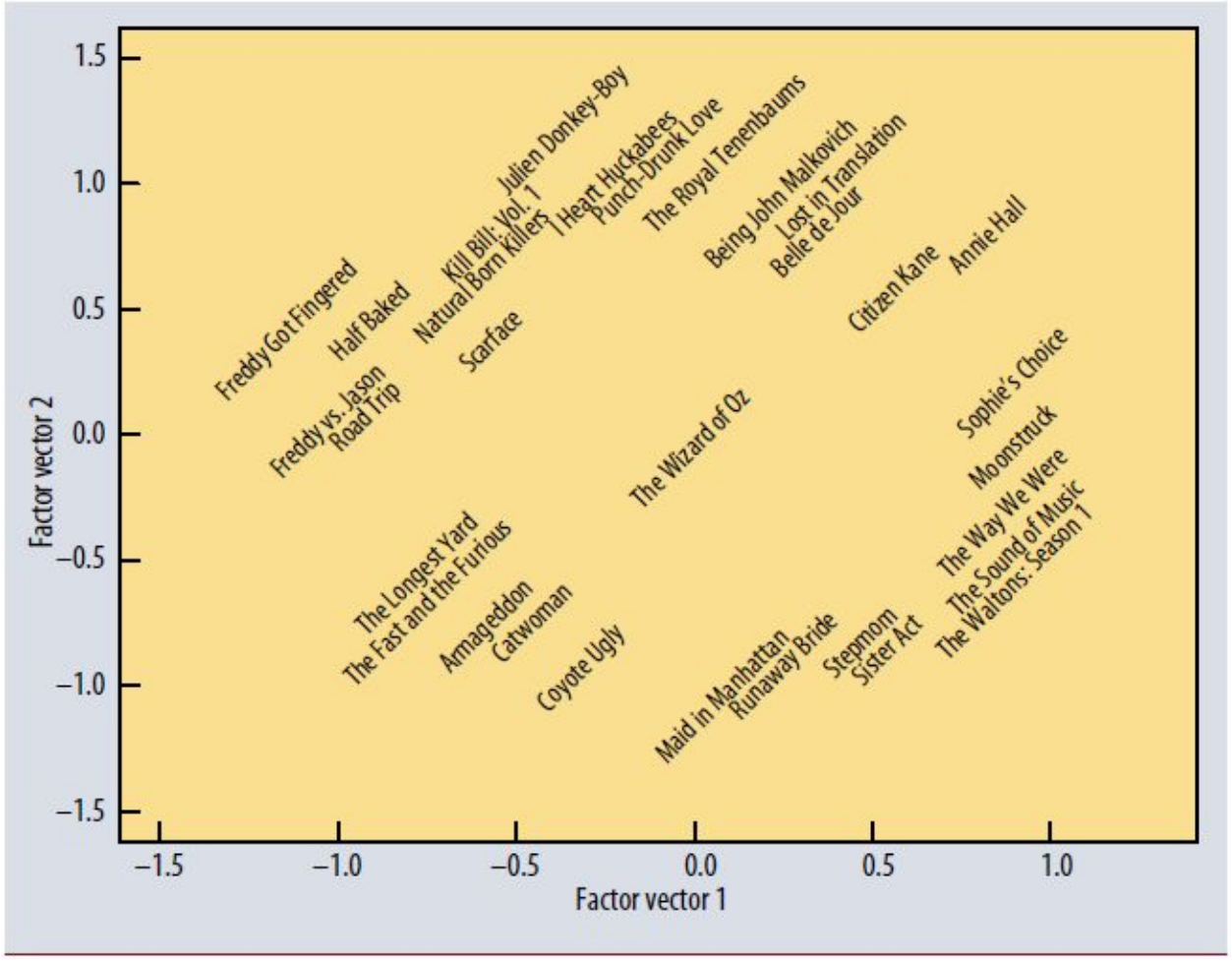
- Modèle :

$$y_{i\ell} = \mu_\ell + \sum_{k=1}^{K} u_{ik} v_{k\ell} + \epsilon_{i\ell}$$

- $\epsilon_{i\ell}$ est un bruit de variance σ^2.
- Les vecteurs V_ℓ sont indépendants et de loi $N(0, \mathrm{Id}_K)$ (orthogonaux).
- Ces vecteurs représentent K axes factoriels principaux.
- En termes matriciels

$$Y = \mu + U^T V + \epsilon$$

Facteurs latents



Koren et al. 2009

Vraisemblance

- Soit $C = UU^T + \sigma^2 \mathrm{Id}$

$$\ln p(y|U, \sigma^2, \mu) = -\frac{n}{2}\left(L \ln(2\pi) + \ln|C| + \mathrm{tr}(C^{-1}S)\right)$$

- où S est la matrice de covariance empirique

$$S = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mu)(y_i - \mu)^T$$

Maximum de vraisemblance : SVD (Tipping et Bishop 1999)

- Projection sur les axes factoriels:

$$U_{\text{max}} = U_K(\Lambda_K - \sigma^2 \text{Id})^{1/2} R$$

- U_K est la matrice $(L \times K)$ formée des K premiers vecteurs propres de la matrice de covariance empirique S
- Λ_K est une matrice diagonale $(K \times K)$ contenant les K plus grandes valeurs propres de S
- R est une matrice de rotation arbitraire.

Projections sur les axes principaux

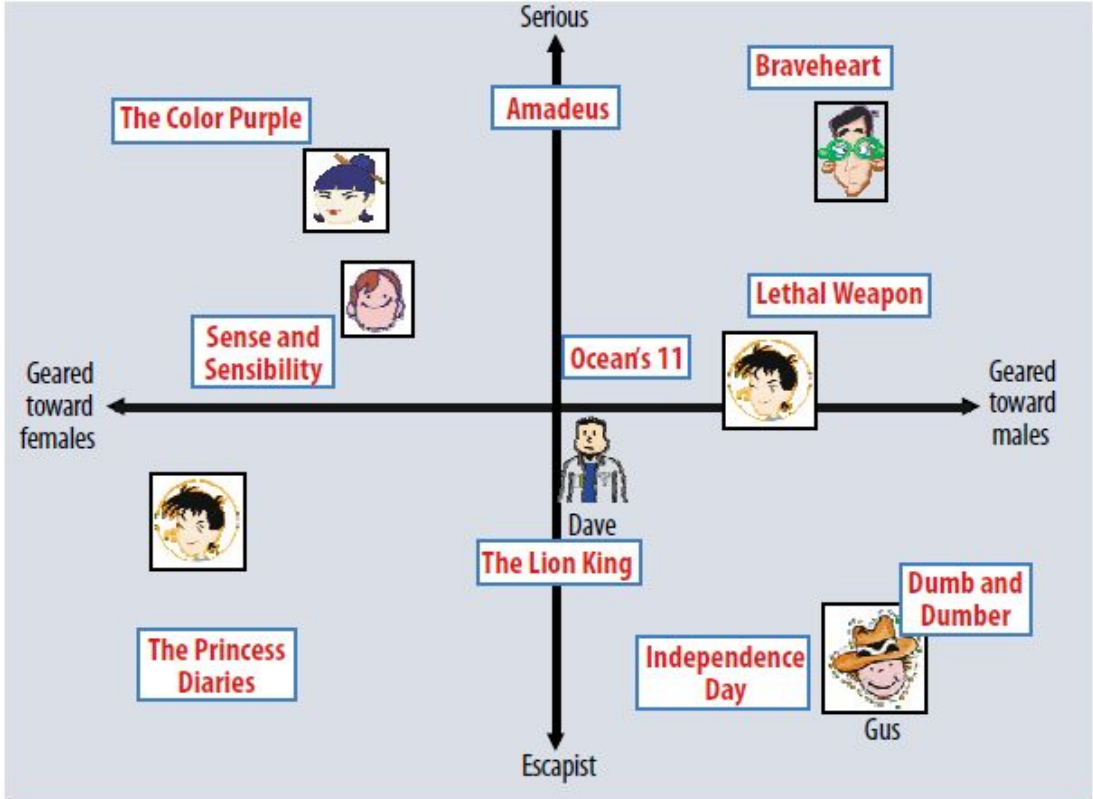
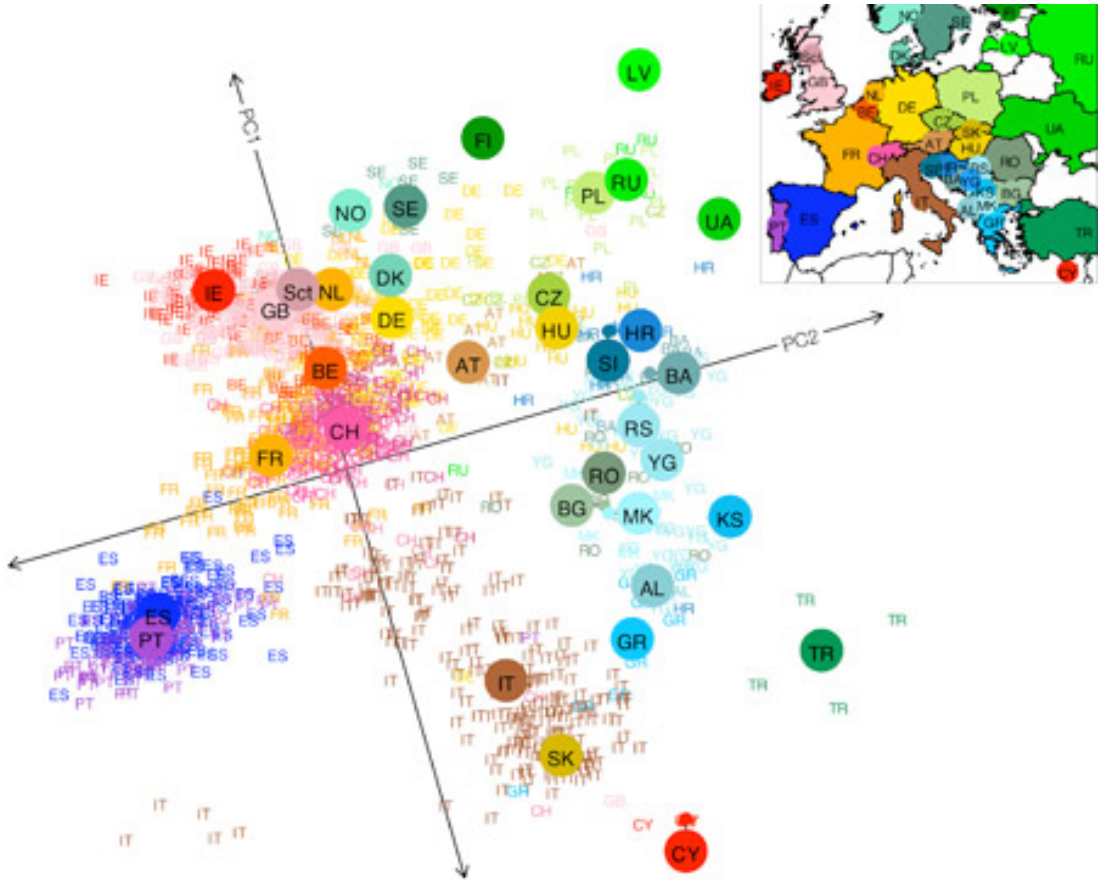


Figure 2. A simplified illustration of the latent factor approach, which characterizes both users and movies using two axes—male versus female and serious versus escapist.

Koren et al. 2009

Aparté: Application en génétique humaine



$y_{i\ell}$ = variant génétique (Novembre et al. 2008)

Prédire les préférences des utilisateurs d'un site web

- Pour la recommandation : la loi prédictive peut être approchée par

$$y_{i\ell} \sim \mu_\ell + N(0, U_{\max} U_{\max}^T) + \epsilon_{i\ell}$$

- problème : L'analyse spectrale ne gère pas les (nombreuses) données manquantes dans le tableau de données de départ.

Loi prédictive

- La loi prédictive *a posteriori* est définie par

$$p(y_{\text{rep}}|y) = \int p(y_{\text{rep}}|\theta)p(\theta|y)d\theta$$

- Elle est obtenue concrètement de la manière suivante
  - simuler θ selon $p(\theta|y)$
  - simuler y_{rep} selon $p(y_{\text{rep}}|\theta)$

Un algorithme stochastique pour la prédiction

- On simule des réalisations *a posteriori* des matrices U et V (U gaussien)
- En alternant la simulation des lois conditionnelles $U|V, y$ et $V|U, y$
- Elles sont calculables facilement car elles correspondent à des régressions linéaires classiques
- Loi prédictive :

$$y \sim \mu + U_{\text{post}}^T V_{\text{post}} + \epsilon$$

- Gestion des données manquantes!

THM

- Utilisation des probabilités pour l'analyse de données
- Approche par modèle, sans approximation asymptotique
- Quantification de l'incertitude
- Données manquantes, données latentes
- Approches prédictives

Pour aller plus loin

- Gelman A et al. (2004) Bayesian Data Analysis, CRC Press.
- Bishop C (2007) Pattern Recognition and Machine Learning, Springer.
- Hastie T et al. (2009) Elements of Statistical Learning, Springer
  `http://www.stanford.edu/ hastie/`

Derniers mots

It was six men of Indostan
To learning much inclined
Who went to see the Elephant
(Though all of them were blind)
That each by observation
Might satisfy his mind

à J.G. Saxe

And so these men of Indostan
Disputed loud and long,
Each in his own opinion
Exceeding stiff and strong,
Though each was partly in the right,
And all were in the wrong!

J.G. Saxe (1816 – 1887)

Merci pour votre attention!