Mélange de Bernoulli

Jingzhuo HUI, Gabriel Moran

24/10/2018

Modèle

Considérons un vecteur aléatoire binaire $\boldsymbol{x} \in [0,1]^p$ de p variables x_j suivant chacune une distribution de Bernoulli $\mathcal{B}(\mu_j)$. La distribution du vecteur s'exprime comme:

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{j=1}^{p} \mu_j^{x_j}(1-\mu_j)^{1-x_j},$$

avec $\boldsymbol{x} = (x_1, \cdots, x_p)^T$ et $\boldsymbol{\mu} = (\mu_1, \cdots, \mu_p)^T$.

Soit une distribution mélange à K composantes de Bernoulli

$$p(\boldsymbol{x}|\boldsymbol{\pi}, \boldsymbol{M}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{x}|\boldsymbol{\mu}_k)$$

où les π_k sont les proportions du mélange et les $p(\boldsymbol{x}|\boldsymbol{\mu}_k)$ sont des distributions de Bernoulli multivariées de paramètres $\boldsymbol{\mu}_k = (\mu_{k1}, \cdots, \mu_{kp})^T$, et $M = \{\boldsymbol{\mu}_1, \cdots, \boldsymbol{\mu}_K\}^T$ la matrice des paramètres des densités de classes.

Dans la suite nous considérerons

- un échantillon observé $X = \{\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n\}$ issu de cette distribution mélange,
- des variables latentes $Z = \{z_1, \cdots, z_n\}$ indiquant la composante d'origine de chaque $\boldsymbol{x}_i$.

Exercice 1

Considérons un mélange à 3 composantes de Bernoulli mélangées en proportions égales $\pi_1 = \pi_2 = \pi_3$.

Simulons une matrice M de proportions dont les 3 lignes et les 50 colonnes décrivent 3 vecteurs des proportions d'un mélange de Bernoulli dans un espace de dimension 50.

```
set.seed(3)
K<-3
p<-50
n<-200
pi<-c(1/3,1/3,1/3)
M<-matrix(runif(K*p),K,p)
M[K,]<-1-M[1,]
```

Simulons $Z = \{z_1, \cdots, z_n\}$ pour $n = 200$.

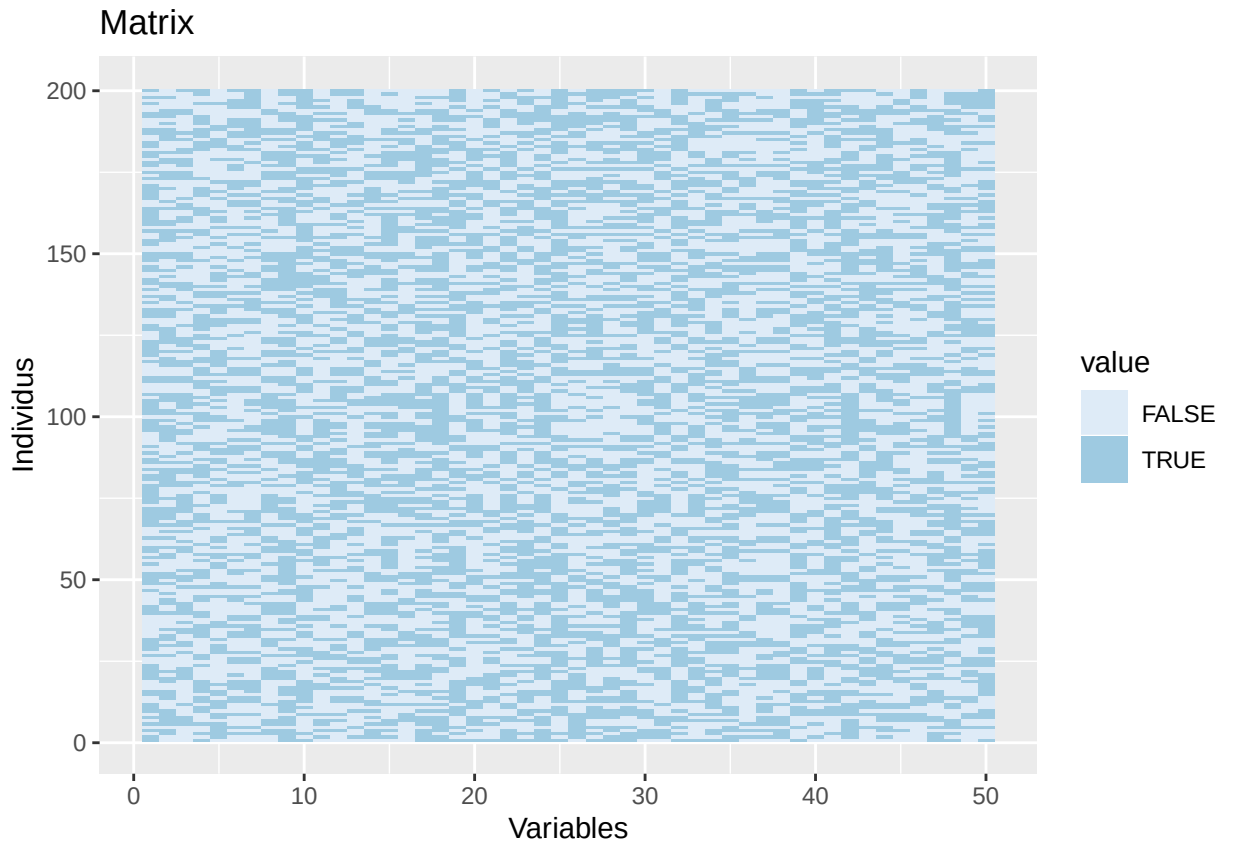
```
nks<-rmultinom(1,200,prob = pi)
Z<-rep(1:length(nks),nks)
```

Simulons $X|Z$.

```
X <-do.call(rbind,
                  mapply(function(nk,k){
                    matrix(rbernoulli(nk*p,p=M[k,]),
                           nrow = nk,
                           ncol=p,
                           byrow = TRUE)}, nks,1:K))
```

Permutons les lignes de la matrice X à 200 lignes et 50 colonnes et visualisons la matrice ainsi obtenue.

```
permutation <- sample(nrow(X))
X <- X[permutation,]
ggplot(melt(X), aes(x = Var2, y = Var1)) +
  geom_raster(aes(fill=value)) +
  scale_fill_brewer(aesthetics = "fill") +
  labs(x="Variables", y="Individus", title="Matrix")
```



On observe que l'échantillon a bien été mélangé et qu'il est impossible de distinguer graphiquement 3 motifs. On va désormais appliquer l'algorithme des kmeans à X avec 3 classes.

```
kmeans(X,3,nstart = 10)->res.kmeans
```

L'algorithme des kmeans convergeant vers un minimum local, on choisi nstart=10 initialisations différentes pour optimiser la réponse. L'algorithme des kmeans minimise l'inertie intra-classe et maximise l'inertie inter-classes. En assignant l'échantillon à 3 classes au lieu d'une seule, la réduction de variance expliquée par les clusters vaut (en pourcentage):

```
100*res.kmeans$tot.withinss/res.kmeans$totss
```

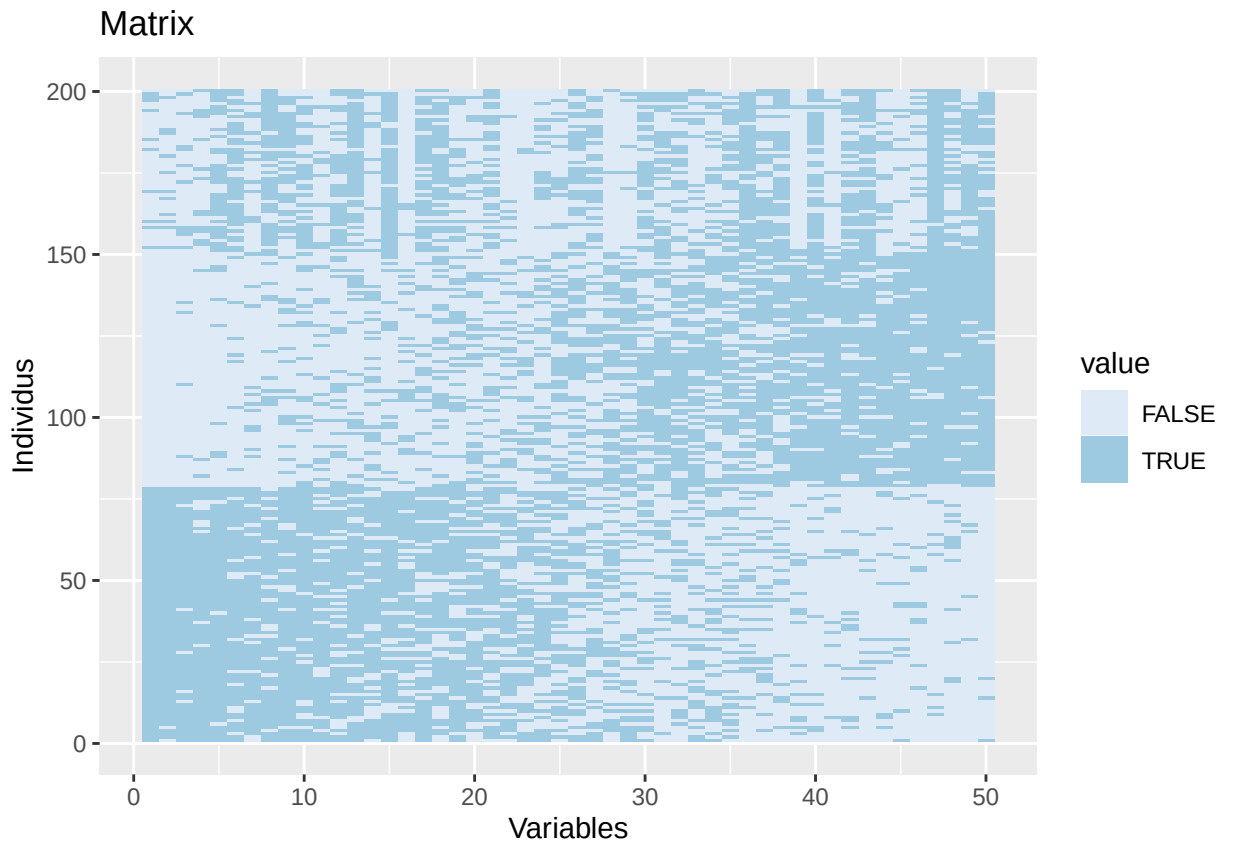
```
## [1] 67.92565
```

Visualisons la matrice classée :

```
tidyData<-melt(X[order(res.kmeans$cluster),order(M[1,])])

ggplot(tidyData, aes(x = Var2, y = Var1)) +
  geom_raster(aes(fill=value)) +
  scale_fill_brewer(aesthetics = "fill") +
```

```
labs(x="Variables", y="Individus", title="Matrix")
```



On observe bien les 3 motifs distincts.

Exercice 2

Q1. Calculons la log-vraisemblance complète $\ln P(\mathbf{X}, \mathbf{Z}|\boldsymbol{\theta} = \{\boldsymbol{\pi}, \boldsymbol{M}\})$.

Calculons d'abord la vraisemblance complète. On a :

$P(\mathbf{X}, \mathbf{Z}|\pi, \mathbf{M}) = P(\mathbf{X}|\mathbf{Z}, \pi, \mathbf{M})P(\mathbf{Z}|\pi, \mathbf{M})$

Or, les observations étant indépendantes, en notant $z_{i,k} = \mathbf{1}_{Z_i=k}$: $P(\mathbf{X}|\mathbf{Z}, \pi, \mathbf{M}) = \prod_{i=1}^{n} P(\mathbf{x_i}|\mathbf{z_i}, \pi, \mathbf{M}) = \prod_{i=1}^{n}\prod_{k=1}^{K} P(\mathbf{x_i}|\mu_\mathbf{k})^{z_{i,k}} = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\prod_{j=1}^{p}\mu_{k,j}^{x_{i,j}}(1-\mu_{k,j})^{1-x_{i,j}}\right)^{z_{i,k}}$

Par ailleurs, comme $\pi_k = P(z_i = k)$,on a: $P(\mathbf{Z}|\pi) = \prod_{i=1}^{n} P(\mathbf{z_i}|\pi) = \prod_{i=1}^{n}\prod_{k=1}^{K}\pi_k^{z_{i,k}}$

On obtient donc pour la vraisemblance complète :

$P(\mathbf{X}, \mathbf{Z}|\pi, \mathbf{M}) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left(\pi_k\prod_{j=1}^{p}\mu_{k,j}^{x_{i,j}}(1-\mu_{k,j})^{1-x_{i,j}}\right)^{z_{i,k}}$

En prenant le logarithme, cela donne :

$\ln P(\mathbf{X}, \mathbf{Z}|\pi, \mathbf{M}) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{i,k}\left(\ln\pi_k + \sum_{j=1}^{p} x_{i,j}\ln\mu_{k,j} + (1-x_{i,j})\ln(1-\mu_{k,j})\right)$

Q2. Calculons $t_{ik}^q = \mathbb{E}[z_{ik}]$ par rapport à la loi $p_{\theta^q}(\mathbf{Z}|\mathbf{X})$. On a, en utilisant le théorème de Bayes et le fait que $P(\mathbf{x_i}|z_{i,k}) = P(\mathbf{x_i}|\mu_k)$:

$t_{ik}^q = \mathbb{E}[z_{i,k}] = P(z_{i,k}|\mathbf{x_i}, \pi, \mathbf{M}) = \frac{P(z_{i,k})P(\mathbf{x_i}|\mu_\mathbf{k})}{P(\mathbf{x_i})}$.

Puis, par la formule des probabilités totales :

$t_{i,k}^q = \frac{\pi_k P(\mathbf{x_i}|\mu_\mathbf{k})}{\sum_{m=1}^K \pi_m P(\mathbf{x_i}|\mu_\mathbf{m})} = \frac{\pi_k \prod_{j=1}^p \mu_{k,j}^{x_{i,j}} (1-\mu_{k,j})^{1-x_{i,j}}}{\sum_{m=1}^K \pi_m \prod_{j=1}^p \mu_{m,j}^{x_{i,j}} (1-\mu_{m,j})^{1-x_{i,j}}}$

Q.3 On en déduit $Q(\theta^q|\theta)$, l'espérance de cette log-vraisemblance par rapport à la loi $p_{\theta^q}(\mathbf{Z}|\mathbf{X})$:

$Q(\theta^q|\theta) = \mathbb{E}[\ln P(\mathbf{X}, \mathbf{Z}|\pi, \mathbf{M})] = \sum_{i=1}^n \sum_{k=1}^K t_{i,k}^q \left(\ln \pi_k + \sum_{j=1}^p x_{i,j} \ln \mu_{k,j} + (1 - x_{i,j}) \ln(1 - \mu_{k,j}) \right)$

Q.4 Pour déterminer $\theta^{q+1} = argmax_\theta Q(\theta^q|\theta)$, on peut commencer par maximiser l'argument par rapport à $\boldsymbol{\mu}_k$ en annulant sa dérivée :

$\frac{\partial}{\partial \mu_{k,j}} Q(\theta^q|\theta) = \sum_{i=1}^n t_{i,k}^q (\frac{x_{i,j}}{\mu_{k,j}} - \frac{1-x_{i,j}}{1-\mu_{k,j}}) = \sum_{i=1}^n t_{i,k}^q \frac{x_{i,j}-\mu_{k,j}}{\mu_{k,j}(1-\mu_{k,j})} = 0 \Leftrightarrow \mu_{k,j} = \frac{1}{n_k} \sum_{i=1}^n x_{i,j} t_{i,k}^q$

où $n_k = \sum_{i=1}^n t_{i,k}^q$ correspond au nombre de points affectés au cluster k. Ainsi, pour $k \in \{1, ...K\}$, le maximum de $Q(\theta^q|\theta)$ par rapport à $\boldsymbol{\mu}_k$ est :

$$\boldsymbol{\mu}_k = \frac{1}{n_k} \sum_{i=1}^n \boldsymbol{x}_i t_{i,k}^q$$

Pour maximiser la fonction $Q(\theta^q|\theta)$ par rapport à π, sous la contrainte $\sum_{k=1}^K \pi_k = 1$, on peut introduire le multiplicateur de Lagrange. Le problème d'optimisation devient alors de maximiser la fonction $\Lambda(\theta, \lambda) = Q(\theta^q|\theta) + \lambda(\sum_{k=1}^K \pi_k - 1)$. En dérivant succéssivement par rapport à π_k puis λ, on obtient : $\frac{\partial}{\partial \pi_k} \Lambda(\theta, \lambda) = \frac{1}{\pi_k} \sum_{i=1}^n t_{i,k}^q + \lambda = 0 \Leftrightarrow \pi_k = -\frac{n_k}{\lambda}$,

$\frac{\partial}{\partial \lambda} \Lambda(\theta, \lambda) = \sum_{k=1}^K \pi_k - 1 = 0 \Leftrightarrow \sum_{k=1}^K \pi_k = 1.$

En combinant ces deux résultats, on obtient $\lambda = -\sum_{k=1}^K n_k = -n$ et donc :

$$\pi_k = -\frac{n_k}{\lambda} = \frac{n_k}{n}$$

Q.5 On commence l'algorithme EM en initialisant θ. Ensuite, l'étape E de l'algorithme EM consiste à calculer les $t_{i,k}^q = \mathbb{E}[z_{ik}]$ pour mettre à jour la distribution des variables latentes z_i. Puis l'étape M consiste à affecter à θ le maximum de $Q(\theta^q|\theta) = \mathbb{E}[\ln P(\mathbf{X}, \mathbf{Z}|\pi, \mathbf{M})]$ par rapport à θ. Le θ estimé est obtenu lors de la convergence.

Q.6

Exercice 3