Université de Strasbourg
M2 - Statistique
Outils pour la statistique avancée

Ségolen Geffray
geffray@math.unistra.fr
Année 2018/2019

TD 2 : Algorithme EM

Exercice 1.
Estimation dans un modèle de mélange.

1. **Estimation paramétrique d'une densité dans un modèle de mélange**.
   Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées de variable parente X dont la densité est de la forme :

   $$f_X(x) = p_1\lambda_1 e^{-\lambda_1 x} + p_2\lambda_2 e^{-\lambda_2 x} + p_3\lambda_3 e^{-\lambda_3 x}, \quad x > 0$$

   où $\lambda_1, \lambda_2, \lambda_3 > 0$ et $p_1, p_2, p_3 \in]0, 1[$ avec $p_1 + p_2 + p_3 = 1$ sont des paramètres inconnus. Proposer une méthode d'estimation de $\theta = (p_1, p_2, p_3, \lambda_1, \lambda_2, \lambda_3)$.
2. **Estimation paramétrique d'une densité dans un modèle de mélange**.
   Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées de variable parente X dont la densité est de la forme :

   $$f_X(x) = p_1\pi_1^x(1-\pi_1)^{1-x} + p_2\pi_2^x(1-\pi_2)^{1-x} + p_3\pi_3^x(1-\pi_3)^{1-x}, \quad x > 0$$

   où $\pi_1, \pi_2, \pi_3, p_1, p_2, p_3 \in]0, 1[$ avec $p_1 + p_2 + p_3 = 1$ sont des paramètres inconnus. Proposer une méthode d'estimation de $\theta = (p_1, p_2, p_3, \pi_1, \pi_2, \pi_3)$.
3. **Estimation paramétrique d'une densité dans un modèle de mélange**.
   Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées de variable parente X dont la densité est de la forme :

   $$f_X(x) = p + (1-p)g_{a,b}(x), \quad x \in [0, 1]$$

   où $p \in]0, 1[$, $a > 0$ et $b > 0$ sont des paramètres inconnus et où

   $$g_{a,b}(x) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} x^{a-1}(1-x)^{b-1}, \quad x \in [0, 1]$$

   (a) De quel mélange s'agit-il ?
   (b) Ecrire l'algorithme EM permettant de déterminer l'estimateur du maximum de vraisemblance de $\theta = (p, a, b)$.
4. **Estimation paramétrique d'une densité dans un modèle de mélange**.
   Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées de variable parente X dont la densité est de la forme :

   $$f_X(x) = p\delta_{\{0\}}(x) + (1-p)\frac{\lambda^x e^{-\lambda}}{x!}, \quad x \in \mathbb{N}$$

   où $p \in]0, 1[$ et $\lambda > 0$ sont des paramètres inconnus.
   (a) De quel mélange s'agit-il ?

(b) Ecrire l'algorithme EM permettant de déterminer l'estimateur du maximum de vraisemblance de $\theta = (p, \lambda)$.

(c) Construire un intervalle de confiance bilatéral pour λ au niveau de confiance $(1-\alpha)$.

5. **Estimation dans un mélange de modèles de régression linéaire gaussiens standards**.
Soient $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ pour $i = 1, ..., n$ des vecteurs aléatoires indépendants avec $Y_i \in \mathbb{R}$ et $X_i^{(j)} \in \mathbb{R}$ pour $j = 1, ..., p$. On suppose que la loi de Y_i conditionnelle à $\mathbb{X}_i$ est la suivante :

$$f(y_i|\mathbb{X}_i) = \frac{p}{\sqrt{2\pi\sigma_1^2}} \exp\left(-\frac{1}{2\sigma_1^2}(y_i - \mathbb{X}_i.\beta^{[1]})^2\right) + \frac{(1-p)}{\sqrt{2\pi\sigma_2^2}} \exp\left(-\frac{1}{2\sigma_2^2}(y_i - \mathbb{X}_i.\beta^{[2]})^2\right)$$

où $0 < p < 1$, où $\beta^{[1]}, \beta^{[2]} \in \mathbb{R}^{p+1}$ et en notant $\mathbb{X}_i = (1, X_i^{(1)}, ..., X_i^{(p)})$. Proposer une méthode d'estimation de $\theta = (p, \beta^{[1]}, \beta^{[2]}, \sigma_1^2, \sigma_2^2)$.

6. **Estimation dans un mélange de modèles de régression de Poisson**.
Soient $(Y_i, X_i^{(1)}, ..., X_i^{(p)})$ pour $i = 1, ..., n$ des vecteurs aléatoires indépendants avec $Y_i \in \mathbb{R}$ et $X_i^{(j)} \in \mathbb{R}$ pour $j = 1, ..., p$. On suppose que la loi de Y_i conditionnelle à $\mathbb{X}_i$ est la suivante :

$$f(y_i|\mathbb{X}_i) = p\,\frac{e^{-\lambda_i^{[1]}}(\lambda_i^{[1]})^{y_i}}{y_i!} + (1-p)\,\frac{e^{-\lambda_i^{[2]}}(\lambda_i^{[2]})^{y_i}}{y_i!}$$

où $0 < p < 1$, où $\lambda_i^{[1]} = e^{\mathbb{X}_i.\beta^{[1]}}$ et $\lambda_i^{[2]} = e^{\mathbb{X}_i.\beta^{[2]}}$ avec $\beta^{[1]}, \beta^{[2]} \in \mathbb{R}^{p+1}$ et en notant $\mathbb{X}_i = (1, X_i^{(1)}, ..., X_i^{(p)})$. Proposer une méthode d'estimation de $\theta = (p, \beta^{[1]}, \beta^{[2]})$.

7. **Estimation dans un mélange de modèles de régression exponentielle**.
Soient (Y_i, X_i) pour $i = 1, ..., n$ des vecteurs aléatoires indépendants avec $Y_i \in \mathbb{R}$ et $X_i \in \mathbb{R}$. On suppose que la loi de Y_i conditionnelle à X_i est la suivante :

$$f(y_i|X_i) = p\lambda_i^{[1]}e^{-\lambda_i^{[1]}y_i} + (1-p)\lambda_i^{[2]}e^{-\lambda_i^{[2]}y_i}\,, y_i > 0$$

où $0 < p < 1$, où

$$\lambda_i^{[1]} = e^{\beta_0^{[1]} + X_i.\beta_1^{[1]}}$$

et où

$$\lambda_i^{[2]} = e^{\beta_0^{[2]} + X_i.\beta_1^{[2]}}$$

avec $\beta_0^{[1]}, \beta_1^{[1]}, \beta_0^{[2]}, \beta_1^{[2]} \in \mathbb{R}$. Proposer une méthode d'estimation de $\theta = (p, \beta_0^{[1]}, \beta_1^{[1]}, \beta_0^{[2]}, \beta_1^{[2]})$.

Exercice 2.

CLUSTERING.

Considérons un échantillon de données de comptage indépendantes $(X_1, ..., X_n)$. On fait l'hypothèse que cet échantillon est constitué de trois sous-groupes que l'on cherche à reconstituer. Pour cela, on travaille avec l'hypothèse que les données sont issues d'un modèle de mélange de distributions de Poisson.

1. Ecrire le modèle de mélange en question.
2. Détailler un algorithme permettant d'estimer les paramètres de ce modèle par la méthode du maximum de vraisemblance.
3. En déduire un estimateur de la probabilité que X_i appartienne à chacun des trois sous-groupes conditionnellement à la valeur de l'observation de X_i $(i = 1, ..., n)$.

4. Prédire le sous-groupe auquel X_i appartient $(i = 1, ..., n)$.
5. Ecrire un programme qui réalise ceci.

Exercice 3.
Problèmes de données manquantes.

1. **Données manquantes dans un modèle gaussien i.i.d. univarié**.
   Soient $X_1, ..., X_n$ des variables aléatoires indépendantes et identiquement distribuées de variable parente X de loi $\mathcal{N}(\mu, \sigma^2)$. Les variables $X_1, ..., X_r$ sont observées tandis que les variables $X_{r+1}, ..., X_n$ sont manquantes (avec $1 < r < n$).
   Proposer plusieurs méthodes d'estimation de $\theta = (\mu, \sigma^2)$.
2. **Données manquantes dans un modèle gaussien i.i.d. bivarié**.
   Soient $(X_1, Y_1), ..., (X_n, Y_n)$ v.a.i.i.d. distribuées comme une variable parente (X, Y) de $\mathbb{R}^2$ de loi $\mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$. Les variables $(X_1, Y_1), ..., (X_r, Y_r)$ sont observées tandis que les variables $(X_{r+1}, Y_{r+1}), ..., (X_n, Y_n)$ sont manquantes (avec $1 < r < n$).
   Proposer plusieurs méthodes d'estimation de $\theta = (\sigma_1^2, \sigma_2^2, \rho)$.
3. **Données manquantes dans un modèle gaussien i.i.d. bivarié, autre schéma de données manquantes**.
   Soient $(X_1, Y_1), ..., (X_n, Y_n)$ v.a.i.i.d. distribuées comme une variable parente (X, Y) de $\mathbb{R}^2$ de loi $\mathcal{N}_2\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}\sigma_1^2 & \rho\sigma_1\sigma_2\\ \rho\sigma_1\sigma_2 & \sigma_2^2\end{pmatrix}\right)$. Les variables $(X_{r_1+1}, ..., X_{r_2})$ sont manquantes ainsi que les variables $(Y_{r_2+1}, .., Y_n)$ (avec $1 < r_1 < r_2 < n$).
   Proposer plusieurs méthodes d'estimation de $\theta = (\sigma_1^2, \sigma_2^2, \rho)$.

Exercice 4.
Estimation en génétique.
Le gène codant pour le groupe sanguin d'un individu a trois allèles possibles : A, B et O. Le type O est récessif, les types A et B sont dominants. On observe quatre groupes sanguins (ou phénotypes) : [A] (pour un génotype AA ou AO), [B] (pour un génotype BB ou BO), [AB] (pour un génotype AB) et [O] (pour un génotype OO). On souhaite estimer les probabilités qu'un chromosome porte l'allèle A, B ou O respectivement notées p_A, p_B et $p_O = 1 - p_A - p_B$. On observe pour cela le phénotype de 521 personnes. Les données sont les suivantes :

n_A	n_B	n_{AB}	n_O
208	57	21	235

Déterminer une estimation de p_A, p_B et p_O par la méthode du maximum de vraisemblance en utilisant l'algorithme EM.

Exercice 5.
Estimation en épidémiologie.
D'après les observations rapportées par McKendrick (1926), le choléra affectant les foyers indiens au début du 20ème siècle peut donner lieu à des infections sévères ou à des infections plus légères qui sont alors asymptomatiques de sorte que l'observation des foyers sans cas est non fiable. Les malades asymptomatiques sont néanmoins transmetteurs de la maladie. McKendrick a recensé le nombre de cas par foyer dans une certaine région indienne au cours d'une épidémie donnée. Les données sont les suivantes :

nombre de cas par foyer	0	1	2	3	4	≥ 5
effectif	NA	32	16	6	1	0

En admettant que les comptages suivent une loi de Poisson, proposer une imputation du nombre de foyer sans cas.
Indication : estimer le(s) paramètre(s) du modèle par maximum de vraisemblance (utiliser l'algorithme EM).

Exercice 6.
Données manquantes dans un modèle de régression linéaire.
Soient (Y_i, X_i) des vecteurs i.i.d. de réalisation (y_i, x_i) pour $i = 1, ..., n$. On suppose que l'observation de certains des Y_i est manquante. On cherche à ajuster sur ces données un modèle de régression sur ces données.

1. Ajuster un modèle de régression linéaire gaussien standard.
2. Ajuster un modèle de régression linéaire généralisé de Poisson.

Exercice 7.
Etude démographique.
Une étude démographique est réalisée afin d'étudier le nombre moyen d'enfants par foyer. Une campagne de recensement est effectuée. Pour des raisons techniques, les données suivantes sont recueillies :

nombre d'enfants par foyer	0	1 ou 2	3 ou 4	5	≥ 6
effectif	24	32	16	4	0

1. En admettant que la loi du nombre d'enfants par foyer est une loi de Poisson, proposer une estimation du paramètre λ de cette loi (utiliser l'algorithme EM).
2. Ecrire un programme calculant cette estimation avec le logiciel `R`.
3. Donner la construction d'un intervalle de confiance bilatéral bootstrap studentisé pour λ de niveau de confiance (1-α).
4. Ecrire un programme qui calcule les bornes de cet intervalle de confiance bilatéral bootstrap studentisé.