CHAPITRE 8

Mélanges de lois

MTH3302 - Méthodes probabilistes et statistiques pour l'I.A.

Jonathan Jalbert – Hiver 2023

Les modèles de mélanges de distributions sont très utiles lorsque la population étudiée est hétérogène et lorsqu'elle peut se partitionner en sous-groupes homogènes. Les mélanges de lois peuvent ainsi servir à la classification, également appelée *clustering*. La classification par les mélanges de lois constitue une méthode d'apprentissage non supervisée. Les mélanges de lois sont également utiles lorsque les lois de probabilités usuelles ne sont pas adéquates pour modéliser la population d'intérêt. Un mélange de ces distributions permet d'enrichir l'éventail des distributions possibles.

Les modèles de mélanges s'écrivent plus simplement lorsqu'une variable non observée, appelée*variable latente* ou *variable manquante*, est ajoutée au modèle. Les variables latentes sont très communes dans les modèles statistiques avancés. Les modèles de mélanges de distributions sont possiblement les modèles à variables latentes les plus simples. En statistique bayésienne, les variables latentes sont traitées de la même façon que les paramètres inconnus à estimer, tandis qu'en statistique classique, un algorithme particulier de maximisation doit être utilisé, l'algorithme EM.

À la fin du chapitre, vous devriez être en mesure de :

— Interpréter de façon probabiliste un mélange de lois.
— Implémenter l'échantillonnage de Gibbs pour générer un échantillon de la loi *a posteriori* des paramètres d'un mélange de lois.
— Implémenter l'algorithme EM pour calculer les estimations du maximum de la vraisemblance d'un mélange de lois.

8.1 Mélange de deux lois normales

La densité du mélange des deux lois normales $\mathcal{N}(\mu_0, \sigma_0^2)$ et $\mathcal{N}(\mu_1, \sigma_1^2)$ s'écrit de la façon suivante :

$$f_{(Y|\boldsymbol{\theta})}(y) = (1-\omega)\,\mathcal{N}(y \mid \mu_0, \sigma_0^2) + \omega\,\mathcal{N}(y \mid \mu_1, \sigma_1^2), \tag{8.1}$$

où $0 \leq \omega \leq 1$ dénote la proportion de la composante $\mathcal{N}(\mu_1, \sigma_1^2)$ dans le mélange et où $\boldsymbol{\theta}$ dénote le vecteur des paramètres $\boldsymbol{\theta} = \begin{bmatrix}\omega & \mu_0 & \sigma_0^2 & \mu_1 & \sigma_1^2\end{bmatrix}^\top$. La proportion de la première composante $\mathcal{N}(\mu_0, \sigma_0^2)$ dans le mélange correspond à $(1-\omega)$ et la proportion de la deuxième composante $\mathcal{N}(\mu_1, \sigma_1^2)$ correspond à ω. Autrement dit, si y est une réalisation de la variable aléatoire Y distribuée selon le mélange de l'équation (8.1), alors la probabilité que y provienne de la première composante $\mathcal{N}(\mu_0, \sigma_0^2)$ est de $(1-\omega)$ et la probabilité qu'elle provienne de la composante $\mathcal{N}(\mu_1, \sigma_1^2)$ est de ω.

Supposons que $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ dénote un vecteur composé de n réalisations indépendantes et identiquement distribuées selon le mélange de deux lois normales exprimé à l'équation (8.1). La fonction de vraisemblance des paramètres s'écrit donc de la façon suivante :

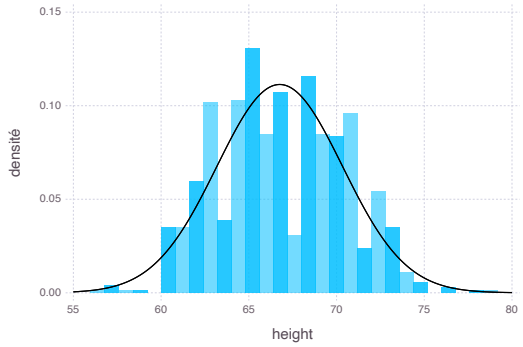
$$\begin{aligned} f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) &= \prod_{i=1}^{n} \left\{(1-\omega)\,\mathcal{N}(y \mid \mu_0, \sigma_0^2) + \omega\,\mathcal{N}(y \mid \mu_1, \sigma_1^2)\right\}, \\ &= \prod_{i=1}^{n} \left[\frac{(1-\omega)}{\sqrt{2\pi\sigma_0^2}} \exp\left\{-\frac{1}{2\sigma_0^2}\left(y_i - \mu_0\right)^2\right\} + \frac{\omega}{\sqrt{2\pi\sigma_1^2}} \exp\left\{-\frac{1}{2\sigma_1^2}\left(y_i - \mu_1\right)^2\right\}\right]. \end{aligned} \tag{8.2}$$

Cette expression ne se factorise pas sous une forme pratique pour l'estimation des paramètres.

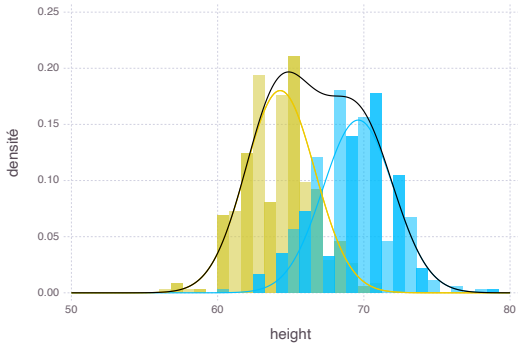
Exemple 1

Soit la taille de 898 adultes mesurés dans les années 1880 par sir Francis Galton pour étudier l'hérédité [a]. La figure 8.1a illustre les tailles mesurées en pouces superposées à la densité de la loi normale ajustée. L'adéquation n'est pas parfaite. L'ensemble hétérogène des tailles peut se décomposer en deux sous-ensembles assez homogènes : la taille des hommes et celle des femmes telle qu'illustré à la figure 8.1b. Ces deux sous-populations peuvent être modélisée par une loi normale et l'ensemble complet des tailles constitue le mélange. Dans ce chapitre, nous tenterons de séparer les tailles en deux sous-ensembles homogènes en n'utilisant que la taille. Nous vérifierons si la classification non supervisée obtenue est cohérente avec le sexe des personnes.

a. Le jeu de données a été transcris par J.A. Hanley et il est disponible en ligne ici.



(a) Histogramme des tailles avec la loi normale estimée superposée.



(b) Histogramme des tailles en fonction du sexe des personnes (bleu pour les hommes et or pour les femmes).

FIGURE 8.1 – Tailles des personnes étudiées par sir Francis Galton.

Exercice 1

Soit le mélange de deux lois normales suivant :

$$f_{(Y|\boldsymbol{\theta})}(y) = \frac{1}{2}\mathcal{N}(y \mid 0, 1^2) + \frac{1}{2}\mathcal{N}(y \mid 1, 1^2)$$

Quelle est la valeur de la vraisemblance pour $\boldsymbol{y} = (-1/2, 0, 1/2)$.

8.2 Introduction d'une variable manquante

Supposons que nous sachions à quelle composante du mélange la variable aléatoire Y_i provient, alors la densité s'écrirait beaucoup plus facilement. En effet, si on savait que Y_i provient de la première composante, alors sa densité conditionnelle serait simplement $\mathcal{N}(y_i \mid \mu_0, \sigma_0^2)$. Ou bien sa densité conditionnelle serait $\mathcal{N}(y_i \mid \mu_1, \sigma_1^2)$ si elle provenant de la deuxième composante. Posons la variable aléatoire indicatrice suivante

$$Z_i = \begin{cases} 0 & \text{si } Y_i \text{ provient de la composante 1;} \\ 1 & \text{si } Y_i \text{ provient de la composante 2.} \end{cases} \tag{8.3}$$

Alors la densité conditionnelle $f_{(Y_i|Z_i=z_i,\boldsymbol{\theta})}(y_i)$ pourrait se décomposer de la façon suivante :

$$f_{(Y_i|Z_i=0,\boldsymbol{\theta})}(y_i) = \mathcal{N}(y_i \mid \mu_0, \sigma_0^2);$$
$$f_{(Y_i|Z_i=1,\boldsymbol{\theta})}(y_i) = \mathcal{N}(y_i \mid \mu_1, \sigma_1^2).$$

Pour l'ensemble des variables aléatoires, on aurait le vecteur des variables indicatrices suivantes $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$. De cette façon, la densité conditionnelle de $\boldsymbol{Y}$ sachant $\boldsymbol{Z} = \boldsymbol{z}$ s'écrit sous la forme simplifiée suivante :

$$\begin{aligned} f_{(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta})}(\boldsymbol{y}) &= \left\{ \prod_{\{i:z_i=0\}} \mathcal{N}(y_i \mid \mu_0, \sigma_0^2) \right\} \left\{ \prod_{\{i:z_i=1\}} \mathcal{N}(y_i \mid \mu_1, \sigma_1^2) \right\} \\ &= \prod_{i=1}^{n} \left\{ \mathcal{N}(y_i \mid \mu_0, \sigma_0^2) \right\}^{1-z_i} \left\{ \mathcal{N}(y_i \mid \mu_1, \sigma_1^2) \right\}^{z_i} \end{aligned} \tag{8.4}$$

La variable indicatrice Z permet donc de factoriser la vraisemblance conditionnelle en fonction des différentes composantes du mélange. Or cette variable n'est généralement pas observée. Ce genre de variable non observée simplifiant l'écriture du modèle statistique est appelé variable manquante ou variable latente. En statistique bayésienne, l'estimation des paramètres d'un modèle à variables manquantes s'effectue assez directement à l'aide de l'échantillonnage de Gibbs (voir la section 8.3). En statistique classique, l'estimation des paramètres est plus ardue. Il faut recourir à l'algorithme EM (voir la section 8.4) pour estimer les paramètres.

Puisque la variable latente $\boldsymbol{Z}$ n'est pas observée, elle doit être introduite dans la vraisemblance. Pour ce faire, on multiplie la densité conditionnelle $f_{(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta})}(\boldsymbol{y})$ par la densité de $f_{(\boldsymbol{Z}|\boldsymbol{\theta})}(\boldsymbol{z})$ pour obtenir la loi conjointe $f_{\{(\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{\theta}\}}(\boldsymbol{y},\boldsymbol{z})$, aussi appelée vraisemblance augmentée. En raison de l'indépendance supposée des Y_i, on a que

$$f_{(\boldsymbol{Z}|\boldsymbol{\theta})}(\boldsymbol{z}) = \prod_{i=1}^{n} f_{(Z_i|\boldsymbol{\theta})}(z_i).$$

Or, Z_i ne peut prendre que deux valeurs :

$$Z_i = \begin{cases} 0 & \text{si } Y_i \text{ provient de la composante 1;} \\ 1 & \text{si } Y_i \text{ provient de la composante 2.} \end{cases}$$

Alors Z_i est la loi de Bernoulli dont la probabilité de succès correspond au poids de la deuxième composante :

$$f_{(Z_i|\boldsymbol{\theta})}(z_i) = \mathcal{B}ernoulli(z_i|\omega).$$

La vraisemblance augmentée de $\boldsymbol{Z}$ s'écrit donc de la façon suivante :

$$\begin{aligned} f_{\{(\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{\theta}\}}(\boldsymbol{y},\boldsymbol{z}) &= f_{(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta})}(\boldsymbol{y}) \times f_{(\boldsymbol{Z}|\boldsymbol{\theta})}(\boldsymbol{z}); \\ &= \prod_{i=1}^{n} \left\{ \mathcal{N}(y_i \mid \mu_0, \sigma_0^2) \right\}^{1-z_i} \left\{ \mathcal{N}(y_i \mid \mu_1, \sigma_1^2) \right\}^{z_i} \times (1-\omega)^{1-z_i} \omega^{z_i}; \\ &= \prod_{i=1}^{n} \left\{ (1-\omega)\, \mathcal{N}(y_i \mid \mu_0, \sigma_0^2) \right\}^{1-z_i} \left\{ \omega\, \mathcal{N}(y_i \mid \mu_1, \sigma_1^2) \right\}^{z_i}. \end{aligned} \tag{8.5}$$

Exercice 2

Soit le mélange de deux lois normales suivant :

$$f_{(Y|\boldsymbol{\theta})}(y) = \frac{1}{2}\mathcal{N}(y \mid 0, 1^2) + \frac{1}{2}\mathcal{N}(y \mid 1, 1^2)$$

Quelle est la valeur de la vraisemblance pour $\boldsymbol{y} = (-1/2, 0, 1/2)$ sachant que $\boldsymbol{z} = (0, 0, 1)$.

8.3 Estimation bayésienne

Dans le cas du mélange des deux lois normales exprimé à l'équation (8.1), il y a 5 paramètres inconnus à estimer : $\boldsymbol{\theta} = (\omega, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$. La forme fonctionnelle de la loi *a posteriori* des paramètres s'obtient en utilisant le théorème de Bayes :

$$f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}(\boldsymbol{\theta}) \propto f_{(\boldsymbol{Y}|\boldsymbol{\theta})}(\boldsymbol{y}) \times f_{\boldsymbol{\theta}}(\boldsymbol{\theta}).$$

Or, le terme de vraisemblance du mélange de lois exprimée à l'équation (8.2) ne s'exprime pas sous une forme pratique pour l'estimation des paramètres. En effet, dans ce cas-ci, il est impossible de trouver des formes connues pour les lois conditionnelles complètes permettant d'implémenter l'échantillonnage de Gibbs.

Plutôt que d'utiliser la vraisemblance du mélange de loi, la vraisemblance augmentée de $\boldsymbol{Z}$ est utilisée pour simplifier l'écriture du modèle. On cherche donc la loi *a posteriori* de $\boldsymbol{\theta}$ augmentée de $\boldsymbol{Z}$, autrement dit la loi *a posteriori* conjointe de $\boldsymbol{\theta}$ et $\boldsymbol{Z}$: *i.e.* la loi $f_{\{(\boldsymbol{\theta},\boldsymbol{Z})|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\theta}, \boldsymbol{z})$. En statistique bayésienne, une variable latente est traitée comme un autre paramètre à estimer, à l'exception qu'une loi *a priori* n'a pas besoin d'être spécifiée pour la variable latente. La forme fonctionnelle de la loi *a posteriori* augmentée s'exprime sous la forme suivante :

$$\begin{aligned} f_{\{(\boldsymbol{\theta},\boldsymbol{Z})|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\theta}, \boldsymbol{z}) &\propto f_{(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta})}(\boldsymbol{y}) \times f_{(\boldsymbol{Z}|\boldsymbol{\theta})}(\boldsymbol{z}) \times f_{\boldsymbol{\theta}}(\boldsymbol{\theta}); \\ &\propto f_{\{(\boldsymbol{Y},\boldsymbol{Z})|,\boldsymbol{\theta}\}}(\boldsymbol{y}, \boldsymbol{z}) \times f_{\boldsymbol{\theta}}(\boldsymbol{\theta}). \end{aligned}$$

La vraisemblance augmentée, $f_{\{(\boldsymbol{Y},\boldsymbol{Z})|,\boldsymbol{\theta}\}}(\boldsymbol{y}, \boldsymbol{z})$, est celle exprimé à l'équation (8.5).

Peu importe la forme de la loi *a priori* $f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, aucune forme analytique n'existe pour la loi *a posteriori* des paramètres augmentée de $\boldsymbol{Z}$. Il faudra recourir à l'échantillonnage de Gibbs pour obtenir un échantillon aléatoire de la loi *a posteriori*.

Pour implémenter l'échantillonnage de Gibbs, toutes les lois conditionnelles complètes des paramètres sont nécessaire. Puisque les variables indicatrices $\boldsymbol{Z}$ sont elles aussi inconnues, leurs lois conditionnelles complètes sont aussi requises pour les intégrer dans l'échantillonnage de Gibbs.

8.3.1 La loi *a priori* partiellement conjuguée

Supposons que la loi *a priori* se partitionne de la façon suivante :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = f_{\omega}(\omega) \times f_{(\mu_0,\sigma_0^2)}(\mu_0, \sigma^2) \times f_{(\mu_1,\sigma_1^2)}(\mu_1, \sigma_1^2). \tag{8.6}$$

À, l'instar du chpaitre 6, les lois *a priori* suivantes peuvent être utilisées pour les paramètres μ_0, σ_0^2, μ_1 et σ_1^2 des deux lois normales :

$$\begin{aligned} f_{(\mu_0,\sigma_0^2)}(\mu_0, \sigma_0^2) &= \mathcal{N}\left(\mu_0 \mid \nu_0, \sigma_0^2\right) \times \mathcal{I}nv\mathcal{G}amma\left(\sigma_0^2 \mid \alpha_0, \beta_0\right) \\ f_{(\mu_1,\sigma_1^2)}(\mu_1, \sigma_1^2) &= \mathcal{N}\left(\mu_1 \mid \nu_1, \sigma_1^2\right) \times \mathcal{I}nv\mathcal{G}amma\left(\sigma_1^2 \mid \alpha_1, \beta_1\right). \end{aligned}$$

Pour le paramètre ω défini sur l'intervalle $(0,1)$, la loi partiellement conjuguée est la loi bêta :

$$\begin{aligned} f_{\omega}(\omega) &= \mathcal{B}eta\left(\omega \mid \alpha, \beta\right) \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\, \omega^{\alpha-1}\, (1-\omega)^{\beta-1} \qquad \text{pour } 0<\omega<1,\ \alpha>0 \text{ et } \beta>0. \end{aligned}$$

Alors nous montrerons en classe que les lois conditionnelles complètes correspondantes s'expriment sous les formes suivantes :

$$f_{(\omega|-)}(\omega) = \mathcal{B}eta(\omega \mid \alpha + n_1, \beta + n_0)$$

$$f_{(\mu_0|-)}(\mu_0) = \mathcal{N}\left(\mu_0 \left| \frac{n_0\,\bar{y}_0 + \nu_0}{n_0+1}, \frac{\sigma_0^2}{n_0+1}\right.\right)$$

$$f_{(\sigma_0^2|-)}(\sigma_0^2) = \mathcal{I}nv\mathcal{G}amma\left\{\sigma_0^2 \left| \alpha_0 + \frac{n_0+1}{2}, \beta_0 + \frac{(\mu_0-\nu_0)^2}{2} + \frac{1}{2}\sum_{\{i:z_i=0\}}(y_i-\mu_0)^2\right.\right\}$$

$$f_{(\mu_1|-)}(\mu_1) = \mathcal{N}\left(\mu_1 \left| \frac{n_1\,\bar{y}_1 + \nu_1}{n_1+1}, \frac{\sigma_1^2}{n_1+1}\right.\right)$$

$$f_{(\sigma_1^2|-)}(\sigma_1^2) = \mathcal{I}nv\mathcal{G}amma\left\{\sigma_1^2 \left| \alpha_1 + \frac{n_1+1}{2}, \beta_1 + \frac{(\mu_1-\nu_1)^2}{2} + \frac{1}{2}\sum_{\{i:z_i=1\}}(y_i-\mu_1)^2\right.\right\}$$

où

$$n_0 = \text{Card}\{i : z_i = 0\} \quad \text{et} \quad n_1 = \text{Card}\{i : z_i = 1\}$$

et

$$\bar{y}_0 = \frac{1}{n_0}\sum_{\{i:z_i=0\}} y_i \quad \text{et} \quad \bar{y}_1 = \frac{1}{n_1}\sum_{\{i:z_i=1\}} y_i.$$

Nous montrerons en classe que la loi conditionnelle complète de Z_i s'exprime sous la forme suivante :

$$f_{(Z_i|-)}(z_i) = \mathcal{B}ernoulli\left\{\frac{\omega\,\mathcal{N}(y_i \mid \mu_1, \sigma_1^2)}{(1-\omega)\,\mathcal{N}(y_i \mid \mu_0, \sigma_0^2) + \omega\,\mathcal{N}(y_i \mid \mu_1, \sigma_1^2)}\right\}$$

L'échantillonnage de Gibbs permettant d'obtenir un échantillon de la loi *a posteriori* augmentée $f_{\{(\boldsymbol{\theta},\boldsymbol{Z})|\boldsymbol{Y}=\boldsymbol{y}\}}(\boldsymbol{\theta})$ est présenté à l'algorithme 1.

8.3.2 Identifiabilité des paramètres

Les mélanges de lois sont invariants par permutation des indices des composantes. Par exemple, le mélange

$$\frac{3}{5}\mathcal{N}(0, 1^2) + \frac{2}{5}\mathcal{N}(1, 2^2)$$

avec $\boldsymbol{\theta} = \left(3/5, 0, 1^2, 1, 2^2\right)$ est identique au mélange

$$\frac{2}{5}\mathcal{N}(1, 2^2) + \frac{3}{5}\mathcal{N}(0, 1^2).$$

avec $\boldsymbol{\theta} = \left(2/5, 1, 2^2, 0, 1^2\right)$. Dans ce cas, on dit que les paramètres du modèle ne sont pas identifiables. Cette caractéristique peut provoquer des difficultés lors de l'échantillonnage de Gibbs. En effet, lorsque l'on échantillonne la loi *a posteriori* à l'aide de l'échantillonnage de Gibbs, il est possible d'échantillonner $\boldsymbol{\theta} = \left(3/5, 0, 1^2, 1, 2^2\right)$ sur une partie de la chaîne et $\boldsymbol{\theta} = \left(2/5, 1, 2^2, 0, 1^2\right)$ sur une autre autre partie.

Un moyen simple pour rendre les paramètres identifiables consiste à imposer une relation d'ordre sur les paramètres. Par exemple, on peut imposer la condition $\mu_0 < \mu_1$. Cette contrainte revient à utiliser la loi *a priori* suivante :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \times \mathbf{1}_{(\mu_0<\mu_1)}(\boldsymbol{\theta}),$$

où

$$\mathbf{1}_{(\mu_0<\mu_1)}(\boldsymbol{\theta}) = \begin{cases} 1 & \text{si } \mu_0 < \mu_1 \\ 0 & \text{sinon.} \end{cases}$$

Par conséquent, lors de l'échantillonnage de Gibbs, on refusera le candidat pour μ_1 s'il est inférieur à l'état actuel de μ_0. Cette solution est souvent utilisée en pratique étant donné sa simplicité. Elle est cependant risquée notamment si la solution se trouve près de la frontière définie par la contrainte.

8.3.3 Permutation des paramètres (OPTIONEL)

Une autre solution au problème d'identifiabilité consiste à effectuer l'échantillonnage de Gibbs sans se soucier du problème et de modifier la chaîne obtenue par après. Dans l'étape

Algorithm 1 Échantillonnage de Gibbs pour un mélange de deux lois normales

Initialiser les paramètres $\boldsymbol{\theta}$ aux valeurs $\boldsymbol{\theta}^{(0)} = \left(\omega^{(0)}, \mu_0^{(0)}, \sigma_0^{2\ (0)}, \mu_1^{(0)}, \sigma_1^{2\ (0)}\right)$.
for t = 1 à N **do**

1. Générer le vecteur des variables latentes.
for i = 1 à n **do**
Générer $z_i^{(t)}$ de la loi conditionnelle complète $f_{(Z_i|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(t-1)})}(z_i)$.
end for

2. Générer $\omega^{(t)}$ de la loi conditionnelle complète

$$f_{\left(\omega\middle|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\mu_0^{(t-1)},\sigma_0^{2\ (t-1)},\mu_1^{(t-1)},\sigma_1^{2\ (t-1)}\right)}(\omega).$$

3. Générer $\mu_0^{(t)}$ de la loi conditionnelle complète

$$f_{\left(\mu_0\middle|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\omega^{(t)},\sigma_0^{2\ (t-1)},\mu_1^{(t-1)},\sigma_1^{2\ (t-1)}\right)}(\mu_0).$$

4. Générer $\sigma_0^{2\ (t)}$ de la loi conditionnelle complète

$$f_{\left(\sigma_0^2\middle|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\omega^{(t)},\mu_0^{(t)},\mu_1^{(t-1)},\sigma_1^{2\ (t-1)}\right)}(\sigma_0^2).$$

5. Générer $\mu_1^{(t)}$ de la loi conditionnelle complète

$$f_{\left(\mu_1\middle|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\omega^{(t)},\mu_0^{(t)},\sigma_0^{2\ (t)},\sigma_1^{2\ (t-1)}\right)}(\mu_1).$$

6. Générer $\sigma_1^{2\ (t)}$ de la loi conditionnelle complète

$$f_{\left(\sigma_1^2\middle|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{Z}=\boldsymbol{z},\omega^{(t)},\mu_0^{(t)},\sigma_0^{2\ (t)},\mu_1^{(t)}\right)}(\sigma_1^2).$$

end for

de post-traitement de la chaîne, on souhaite identifier les itérations où la permutation des paramètres s'est effectuée. Plusieurs méthodes existent pour identifier ces itérations. Nous utiliserons ici la méthode basée sur l'entropie relative. Soit $\left(\boldsymbol{\theta}^{(t)} : t = 1, \ldots, N\right)$ un échantillon de la loi *a posteriori* de $\boldsymbol{\theta}$ obtenue par l'échantillonnage de Gibbs. L'idée consiste à comparer chacune des itérations t au mode de la loi *a posteriori*. Une estimation du mode de la loi *a posteriori* $\text{Mo}(\theta \mid \boldsymbol{Y} = \boldsymbol{y})$ peut être obtenue à partir de la chaîne générée par l'échantillonnage de Gibbs :

$$\widehat{\text{Mo}}(\theta \mid \boldsymbol{Y} = \boldsymbol{y}) = \boldsymbol{\theta}^{(t^\star)};$$

où

$$t^* = \arg \max_{t=1,\ldots,M} f_{(\boldsymbol{\theta}|\boldsymbol{Y}=\boldsymbol{y})}\left(\boldsymbol{\theta}^{(t)}\right).$$

À chacune des itérations t de la chaîne, on choisit la permutation τ qui minimise l'entropie relative

$$h(\tau) = \sum_{i=1}^{n}\sum_{j=1}^{2} \mathbb{P}\left\{Z_i = j-1 \mid \widehat{\text{Mo}}(\theta \mid \boldsymbol{Y} = \boldsymbol{y})\right\} \times \log\left[\frac{\mathbb{P}\left\{Z_i = j-1 \mid \widehat{\text{Mo}}(\theta \mid \boldsymbol{Y} = \boldsymbol{y})\right\}}{\mathbb{P}\left\{Z_i = j-1 \mid \tau\left(\boldsymbol{\theta}^{(t)}\right)\right\}}\right].$$

Dans le cas où il n'y a que deux composantes, il n'y a que deux permutations possibles : la permutation identité et la permutation

$$\begin{aligned}
\omega &\to (1-\omega)\\
\mu_0 &\to \mu_1\\
\mu_1 &\to \mu_0\\
\sigma_0^2 &\to \sigma_1^2\\
\sigma_1^2 &\to \sigma_0^2.
\end{aligned}$$

Après la permutation de chaque itération, les estimations Monte-Carlo peuvent être obtenues de la même manière, c'est-à-dire :

$$\mathbb{E}\left\{h(\boldsymbol{\theta})\right\} \approx \frac{1}{N}\sum_{t=1}^{N} h\left(\boldsymbol{\theta}^{(t)}\right)$$

8.3.4 L'utilité d'utiliser des lois *a priori* informatives

Dans le cas des mélanges de lois, une loi *a priori* informative rend plus robuste la procédure numérique d'échantillonnage de la loi *a posteriori*. Les lois impropres sont problématiques dans le cas des mélanges de lois. Soit le mélange de deux lois normales suivant :

$$f_{(Y|\boldsymbol{\theta})}(y) = (1-\omega)\,\mathcal{N}(\mu_0, \sigma_0^2) + \omega\,\mathcal{N}(\mu_1, \sigma_1^2),$$

Considérons la loi non-informative impropre suivante :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto (1-\omega)^{-1}\omega^{-1} \times \frac{1}{\sigma_0^2} \times \frac{1}{\sigma_1^2} \quad \text{pour} \quad 0 \leq \omega \leq 1,\ \sigma_0^2 > 0 \text{ et } \sigma_1^2 > 0.$$

L'exercice suivant vous permettra de calculer les lois conditionnelles complètes correspondantes.

Exercice 3

Montrez que les lois conditionnelles complètes des paramètres s'expriment sous les formes suivantes :

$$\begin{aligned}
f_{(\omega|-)}(\omega) &= \mathcal{B}eta\left(\omega \mid n_1, n_0\right); \\
f_{(\mu_0|-)}(\mu_0) &= \mathcal{N}\left(\mu_0 \mid \bar{y}_0, \frac{\sigma_0^2}{n_0}\right); \\
f_{(\sigma_0^2|-)}(\mu_0) &= \mathcal{I}nverse\mathcal{G}amma\left\{\sigma_0^2 \,\middle|\, \frac{n_0}{2}, \frac{1}{2}\sum_{\{i:z_i=0\}}(y_i-\mu_0)^2\right\}; \\
f_{(\mu_1|-)}(\mu_1) &= \mathcal{N}\left(\mu_0 \mid \bar{y}_1, \frac{\sigma_1^2}{n_1}\right); \\
f_{(\sigma_1^2|-)}(\mu_0) &= \mathcal{I}nverse\mathcal{G}amma\left\{\sigma_1^2 \,\middle|\, \frac{n_1}{2}, \frac{1}{2}\sum_{\{i:z_i=1\}}(y_i-\mu_1)^2\right\}.
\end{aligned}$$

Si aucune observation n'est attribuée à une composante lors d'une itération de l'échantillonnage de Gibbs, alors certaines lois conditionnelles complètes ne sont plus définies et l'algorithme déraillera. Par exemple si $n_0 = 0$, alors les lois conditionnelles complètes de ω, μ_0 et σ_0^2 ne sont plus valides. Ce problème devient de plus en plus important à mesure que le nombre de composantes augmente, par exemple pour un mélange de 5 lois normales.

8.4 Estimation par maximum de la vraisemblance

La vraisemblance du mélange de lois exprimé à l'équation (8.2) ne s'exprime pas sous une forme analytique. L'algorithme EM permet de maximiser la vraisemblance lorsque des variables latentes sont ajoutées au modèles statistique. L'algorithme se décompose en deux étapes, une étape d'estimation de l'espérance (*Expectation*) et une étape de maximisation (*Maximisation*), d'où l'acronyme *EM* pour *Expectation-Maximization*. Il s'agit d'un algorithme itératif permettant de trouver les estimateurs du maximum de la vraisemblance à l'aide de la fonction de vraisemblance augmentée $f_{\{(\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{\theta}\}}(\boldsymbol{y})$.

Remarque. *Le traitement des variables latentes est plus difficile dans le cadre statistique classique puisque les paramètres ne sont pas considérés comme des variables aléatoires.*

Soit des valeurs initiales des paramètres $\boldsymbol{\theta}^{(0)}$. L'étape E consiste à calculer l'espérance conditionnelle suivante en fonction des variables latentes $\boldsymbol{z}$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) = \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})} \left\{ \ln f_{\{(\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{\theta}\}}(\boldsymbol{y}) \right\}.$$

L'étape M consiste à maximiser cette fonction Q pour trouver une estimation améliorer des paramètres $\boldsymbol{\theta}$:

$$\boldsymbol{\theta}^{(1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}).$$

On répète cette procédure jusqu'à ce que l'estimation améliorée est à toute fin pratique identique à l'estimation précédente. La convergence de cet algorithme vers les estimations du maximum de la vraisemblance a été établi en 1977 par Arthur Dempster, Nan Laird et Donald Rubin.

8.4.1 Algorithme EM pour un mélange de deux lois normales

Dans le cas du mélange de deux lois normales exprimée à l'équation 8.1, le vecteur des paramètres est $\boldsymbol{\theta} = (\omega, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$. Étant donnée la fonction de vraisemblance augmentée exprimée à l'équation (8.5), la fonction $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)})$ de l'algorithme EM correspond à

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(0)}) =& \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})} \left\{ -\frac{n_0}{2}\ln(2\pi\sigma_0^2) - \sum_{\{i:z_i=0\}} \frac{(y_i-\mu_0)^2}{2\sigma_0^2} \right. \\ & \left. -\frac{n_1}{2}\ln(2\pi\sigma_1^2) - \sum_{\{i:z_i=1\}} \frac{(y_i-\mu_1)^2}{2\sigma_1^2} + n_0\ln(1-\omega) + n_1\ln\omega \right\} \end{aligned}$$

L'estimation améliorée de μ_1 se trouve en maximisant Q par rapport à μ_1. Dérivons d'abord la fonction Q par μ_1 :

$$\begin{aligned} \frac{\partial Q}{\partial \mu_1} &= \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})} \left\{ \frac{1}{\sigma_1^2} \sum_{\{i:z_i=1\}} (y_i - \mu_1) \right\} \\ &= \frac{1}{\sigma_1^2} \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})} \left(\sum_{\{i:z_i=1\}} y_i - n_1\mu_1 \right) \end{aligned}$$

La valeur qui annule la dérivée est la suivante correspond à l'estimation améliorée :

$$\hat{\mu}_1^{(1)} = \frac{\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})} \left(\sum_{\{i:z_i=1\}} y_i \right)}{\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}(n_1)}$$

Cette dernière expression correspond à l'estimation améliorée de μ_1. Trouvons une façon de réécrire ce ratio d'espérances afin de trouver une expression analytique pour l'estimation améliorée de μ_1 :

$$\begin{aligned}\mu_1^{(1)} &= \frac{\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}\left(\sum_{i=1}^n z_i \times y_i\right)}{\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}\left(\sum_{i=1}^n z_i\right)} \\ &= \frac{\sum_{i=1}^n y_i \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}(z_i)}{\sum_{i=1}^n \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}(z_i)}.\end{aligned}$$

Or, la loi conditionnelle complète de la variable latente Z_i est la loi de Bernoulli suivante :

$$f_{(Z_i|Y_i=y_i,\boldsymbol{\theta})}(z_i) = \mathcal{B}ernoulli\left\{z_i \,\middle|\, \frac{\omega\,\mathcal{N}(y_i \mid \mu_1, \sigma_1^2)}{(1-\omega)\,\mathcal{N}(y_i \mid \mu_0, \sigma_0^2) + \omega\,\mathcal{N}(y_i \mid \mu_1, \sigma_1^2)}\right\}.$$

On a donc que

$$\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}(Z_i) = p_i$$

où

$$p_i = \frac{\omega^{(0)}\mathcal{N}\left\{y_i \mid \mu_1^{(0)}, \sigma_1^{2\ (0)}\right\}}{(1-\omega^{(0)})\mathcal{N}\left\{y_i \mid \mu_0^{(0)}, \sigma_0^{2\ (0)}\right\} + \omega^{(0)}\mathcal{N}\left\{y_i \mid \mu_1^{(0)}, \sigma_1^{2\ (0)}\right\}}.$$

Par conséquent, l'estimation améliorée de μ_1 s'exprime sous la forme suivante :

$$\hat{\mu}_1^{(1)} = \frac{\sum_{i=1}^n p_i^{(0)} y_i}{\sum_{i=1}^n p_i^{(0)}}.$$

Pour trouver l'estimation améliorée de σ_1^2, on dérive la fonction Q par rapport à σ_1^2 :

$$\frac{\partial Q}{\partial \sigma_1^2} = \mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}\left\{\frac{-n_1}{2}\frac{1}{\sigma_1^2} + \frac{1}{2(\sigma_1^2)^2}\sum_{\{i:z_i=1\}}(y_i-\mu_1)^2\right\}.$$

On cherche la valeur de σ_1^2 qui annule la dérivée :

$$\begin{aligned}\hat{\sigma}_1^{2\,(1)} &= \frac{\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}\left\{\sum_{\{i:z_i=1\}}(y_i-\mu_1)^2\right\}}{\mathbb{E}_{(\boldsymbol{Z}|\boldsymbol{Y}=\boldsymbol{y},\boldsymbol{\theta}=\boldsymbol{\theta}^{(0)})}(n_1)}; \\ &= \frac{\sum_{i=1}^n p_i^{(0)}(y_i-\mu_1)^2}{\sum_{i=1}^n p_i^{(0)}}.\end{aligned}$$

Exercice 4

Montrez que les estimations améliorées des paramètres restant s'expriment sous les formes suivantes :

$$\hat{\mu}_0^{(1)} = \frac{\sum_{i=1}^n (1 - p_i^{(0)})\, y_i}{\sum_{i=1}^n (1 - p_i^{(0)})}$$

$$\hat{\sigma}_0^{2\ (1)} = \frac{\sum_{i=1}^n (1 - p_i^{(0)})\ \left(y_i - \mu_0^{(1)}\right)^2}{\sum_{i=1}^n (1 - p_i^{(0)})}$$

$$\hat{\omega}^{(1)} = \frac{1}{n} \sum_{i=1}^n p_i^{(0)}$$

8.5 Mélange de $k > 2$ lois normales

Soit la variable aléatoire Y distribuée selon un mélange fini de $k > 2$ composantes normales exprimé par la densité suivante :

$$g_{(Y|\boldsymbol{\theta})}(y) = \omega_1\, \mathcal{N}(y \mid \mu_1, \sigma_1^2) + \omega_2\, \mathcal{N}(y \mid \mu_2, \sigma_2^2) + \ldots + \omega_k\, \mathcal{N}(y \mid \mu_k, \sigma_k^2),$$

où

$$\omega_j \geq 0 \text{ pour } 1 \leq j \leq k \text{ et } \sum_{j=1}^k \omega_j = 1$$

et où $\boldsymbol{\theta} = \{(\omega_j, \mu_j, \sigma_j^2) : 1 \leq j \leq k\}$. À l'instar du mélange à deux composantes, la vraisemblance d'un échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ s'écrirait plus facilement si on savait à quelle composante du mélange correspondent chacune des observations Y_i. Posons la variable catégorielle Z_i de la façon suivante :

$$Z_i = \begin{cases} 1 & \text{si } Y_i \text{ provient de la composante 1}\,; \\ 2 & \text{si } Y_i \text{ provient de la composante 2}\,; \\ \vdots & \\ k & \text{si } Y_i \text{ provient de la composante k}\,; \end{cases}$$

La variable aléatoire Z_i est donc distribuée selon la loi catégorielle suivante :

$$Z_i \sim \mathcal{C}at(\boldsymbol{\omega}),$$

où $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_k)$. On a alors que

$$\mathbb{P}(Z_i = j) = \omega_j \text{ pour } 1 \leq j \leq k.$$

Un rappel de la loi catégorielle est présenté à l'annexe 8.A.

Remarque. *Lorsqu'il n'y a que deux catégories, l'écriture du modèle est plus simple en utilisant les valeurs* $\{0, 1\}$ *pour la variable indicatrice* Z*. Lorsqu'il y a plus de deux catégories, l'écriture du modèle est plus simple en utilisant les valeurs* $\{1, 2, \ldots, k\}$.

Conditionnellement à $Z_i = j$, la densité de Y_i s'écrit très facilement :

$$f_{(Y_i|Z_i=j,\boldsymbol{\theta})}(y_i) = \mathcal{N}(y_i \mid \mu_j, \sigma_j^2).$$

Conditionnellement à $\boldsymbol{Z} = \boldsymbol{z}$, la densité de $\boldsymbol{Y}$ s'écrit aussi très facilement :

$$f_{(\boldsymbol{Y}|\boldsymbol{Z}=\boldsymbol{z},\boldsymbol{\theta})}(\boldsymbol{y}) = \left\{ \prod_{\{i:z_i=1\}} \mathcal{N}(y_i \mid \mu_1, \sigma_0^2) \right\} \times \ldots \times \left\{ \prod_{\{i:z_i=k\}} \mathcal{N}(y_i \mid \mu_k, \sigma_k^2) \right\}.$$

Puisque l'on observe pas les variables Z_i, on doit plutôt travailler avec la vraisemblance augmentée :

$$f_{\{(\boldsymbol{Y},\boldsymbol{Z})|\boldsymbol{\theta}\}}(\boldsymbol{y}) = \left\{ \prod_{\{i:z_i=1\}} \mathcal{N}(y_i \mid \mu_1, \sigma_0^2) \right\} \times \ldots \times \left\{ \prod_{\{i:z_i=k\}} \mathcal{N}(y_i \mid \mu_k, \sigma_k^2) \right\} \times \omega_1^{n_1} \ldots \omega_k^{n_k},$$

où

$$n_j = \text{Card}\{i : z_i = j\}.$$

L'estimation des paramètres avec l'échantillonnage de Gibbs ou l'algorithme EM se fait de façon similaire au cas du mélange à deux composantes. La seule différence notable est que la loi *a priori* pour le vecteur des paramètres $\boldsymbol{\omega}$ est différente. En effet, on a maintenant plus que deux composantes. Comme loi *a priori* non informative pour le vecteur des paramètres $\boldsymbol{\omega}$, la loi de Dirichlet de paramètre $\boldsymbol{\alpha} = \mathbf{1}_k$ est utilisée :

$$f_{\boldsymbol{\omega}}(\boldsymbol{\omega}) = Dirichlet(\boldsymbol{\omega}|\mathbf{1}_k) = 1.$$

L'annexe 8.B présente la loi de Dirichlet.

Exercice 5

Montrez que la loi conditionnelle complète de Z_i s'exprime de la façon suivante :

$$f_{(Z_i|Y_i=y_i,\boldsymbol{\theta})}(y_i) \propto \left\{\omega_1 \, \mathcal{N}(y_i \mid \mu_1, \sigma_0^2)\right\}^{\mathbf{1}_{\{1\}}(z_i)} \times \ldots \times \left\{\omega_k \, \mathcal{N}(y_i \mid \mu_k, \sigma_k^2)\right\}^{\mathbf{1}_{\{k\}}(z_i)}.$$

Donc,

$$f_{(Z_i|Y_i=y_i,\boldsymbol{\theta})}(y_i) = \mathcal{C}at(z_i \mid \boldsymbol{p}),$$

où

$$p_j = \frac{\omega_j \, \mathcal{N}(y_i \mid \mu_j, \sigma_j^2)}{\sum_{\ell=1}^{k} \omega_\ell \, \mathcal{N}(y_i \mid \mu_\ell, \sigma_\ell^2)}$$

Exercice 6

Montrez que la loi conditionnelle complète de $\boldsymbol{\omega}$ s'exprime de la façon suivante :

$$f_{(\boldsymbol{\omega}|-)}(\boldsymbol{\omega}) = Dirichlet(\boldsymbol{\omega} \mid \boldsymbol{n} + 1),$$

où $\boldsymbol{n} = (n_1, n_2, \ldots, n_k)$.

8.6 Exercices

1. Le panier A contient 5 boules rouges et 5 boules vertes. Le panier B contient 2 boules rouges et 8 boules bleues. On choisit un panier au hasard avec la probabilité de 1/3 de choisir le panier A et la probabilité 2/3 de choisir le panier B. Du panier choisi, on effectue 5 tirages avec remise. On s'intéresse à N, le nombre de boules rouges tirées parmi les 5 tirages. Quelle est la fonction de masse de N ?

2. Soit la variable aléatoire Y distribuée selon le mélange de deux lois normales défini par la densité suivante :

$$f(y) = (1-\omega)\mathcal{N}(y \mid 0, \sigma_0^2) + \omega\mathcal{N}(y \mid 0, \sigma_1^2).$$

   (a) Quelle est l'espérance de Y ?

   (b) Quelle est la variance de Y ?

   (c) Quelle est la probabilité que Y soit inférieur à 1 ? Vous pouvez laisser votre réponse sous la forme d'intégrales.

3. Un cas particulier important du mélange de deux lois normales est celui où les deux variances sont inconnues mais égales. La densité s'exprime donc comme suit :

$$f_{(Y|\boldsymbol{\theta})}(y) = (1-\omega)\mathcal{N}(y \mid \mu_0, \sigma^2) + \omega\mathcal{N}(y \mid \mu_1, \sigma^2)$$

   avec $0 < \omega < 1$, $\mu_0 \in \mathbb{R}$, $\mu_1 \in \mathbb{R}$ et $\sigma^2 > 0$.

   (a) Si vous utilisez la loi *a priori* impropre suivante :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \frac{1}{\sigma^2}\omega^{-1}(1-\omega)^{-1},$$

   quelles sont les lois conditionnelles complètes des paramètres requises pour l'échantillonnage de Gibbs ?

   (b) Avec une estimation initiale des paramètres $\boldsymbol{\theta}^{(0)}$, quelle est l'estimation améliorée de σ^2 en utilisant l'algorithme EM ?

4. Soit le mélange de loi

$$Y \sim (1-\omega)\,\mathcal{P}oisson(\lambda_0) + \omega\,\mathcal{P}oisson(\lambda_1)$$

avec $0 \leq \omega \leq 1$, $\lambda_0 > 0$ et $\lambda_1 > 0$. Supposons que l'on obtienne l'échantillon aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ où les observations sont mutuellement indépendantes et identiquement distribuées selon le mélange précédent. Soit la variable latente

$$Z_i = \begin{cases} 0 & \text{si l'observation } Y_i \text{ a été générée par la première composante} \\ 1 & \text{si l'observation } Y_i \text{ a été générée par la seconde composante.} \end{cases}$$

(a) Quelle est la loi conditionnelle de Y_i sachant les paramètres $\boldsymbol{\theta}$ et la variable manquante $Z_i = z_i$? Autrement dit, trouvez la loi de $f_{(Y_i|Z_i=z_i\boldsymbol{\theta})}(\boldsymbol{y})$ où $\boldsymbol{\theta} = (\omega, \lambda_0, \lambda_1)^\top$.

(b) Quelle est la probabilité que l'observation Y_i ait été générée par la deuxième composante ? Autrement dit, que vaut $\mathbb{P}(Z_i = 1 \mid Y_i = y_i)$?

(c) Soit la loi *a priori* suivante :

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \mathcal{B}eta(\omega \mid \alpha, \beta) \times \mathcal{G}amma(\lambda_0 \mid \alpha_0, \beta_0) \times \mathcal{G}amma(\lambda_1 \mid \alpha_1, \beta_1).$$

Quelles sont les lois conditionnelles complètes de paramètres ? Autrement dit, calculez $f_{(\omega|-)}(\omega)$, $f_{(\lambda_0|-)}(\lambda_1)$ et $f_{(\lambda_1|-)}(\lambda_2)$.

(d) Écrivez l'échantillonnage de Gibbs permettant d'obtenir un échantillon aléatoire de la loi *a posteriori* des paramètres.

5. Reprenez le mélange des deux lois de Poisson du numéro précédent. Si on utilisait l'algorithme EM pour estimer les paramètres, quelle serait l'estimation améliorée du vecteur de paramètres initiaux $\boldsymbol{\theta}^{(0)}$?

Annexe

8.A Loi catégorielle

La loi catégorielle est une loi de probabilité discrète qui généralise la loi de Bernoulli à plus de deux résultats possibles. Soit la variable aléatoire Y pouvant prendre des valeurs dans l'ensemble des k éléments suivants $\{1, 2, \ldots, k\}$. Posons

$$\alpha_j = \mathbb{P}(Y = j) \text{ pour } 1 \leq j \leq k$$

et $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$. Alors la fonction de masse de Y peut s'écrire de la façon suivante :

$$p_{(Y|\boldsymbol{\alpha})}(y) = \alpha_1^{\mathbf{1}_{\{1\}}(y)} \times \ldots \times \alpha_k^{\mathbf{1}_{\{k\}}(y)},$$

où $\mathbf{1}_I(y)$ dénote la fonction indicatrice suivante :

$$\mathbf{1}_I(y) = \begin{cases} 1 & \text{si } y \in I; \\ 0 & \text{si } y \notin I. \end{cases}$$

On dénote par

$$Y \sim \mathcal{C}at(\boldsymbol{\alpha})$$

le fait que la variable aléatoire Y soit distribuée selon la loi catégorielle avec le vecteur de paramètres $\boldsymbol{\alpha}$.

8.B Loi de Dirichlet

La loi de Dirichlet est une densité de probabilité qui généralise la loi bêta à plusieurs dimensions. Soit le vecteur aléatoire $\boldsymbol{Y} = (Y_1, \ldots, Y_k)$ où

$$0 \leq Y_j \leq 1 \text{ pour tout } 1 \leq j \leq k,$$

et

$$\sum_{j=1}^{k} Y_j = 1.$$

On dit que le vecteur aléatoire $\boldsymbol{Y}$ est distribuée selon la loi de Dirichlet avec le vecteur de paramètres $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)$ si il possède la densité suivante :

$$f_{(\boldsymbol{Y}|\boldsymbol{\alpha})}(\boldsymbol{y}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{k} y_j^{\alpha_j - 1},$$

où la constante de normalisation $B(\boldsymbol{\alpha})$ est donnée par :

$$B(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{k} \Gamma(\alpha_j)}{\Gamma\left(\sum_{j=1}^{k} \alpha_j\right)}$$

pour $\alpha_j > 0$ pour tout $1 \leq j \leq k$.

On peut montrer que la loi marginale d'une composante Y_j du vecteur est la suivante :

$$f_{(Y_j|\boldsymbol{\alpha})}(y) = \mathcal{B}eta(y \mid \alpha_j, \alpha_0 - \alpha_j),$$

où $\alpha_0 = \sum_{j=1}^{k} \alpha_j$.