Estimation des paramètres d'un mélange de deux gaussiennes

Sujet de travail pratique

Les modèles de mélange de deux distributions gaussiennes apparaissent assez naturellement dans de nombreux problèmes pratiques. Fondamentalement, ils reposent sur deux variables.

1. Une variable d'étiquette L (pour Label en anglais) binaire. On note arbitrairement $L = 1$ et $L = 2$ les deux valeurs possibles. Cette variable décrit un « statut », par exemple : l'état d'un sytème (nominal / défaillant), d'un patient (sain / malade),...
2. Une variable réelle notée X. Pour reprendre les deux exemples ci-dessus : ce peut être une quantité caractéristique du système (pression, température, tension électrique,...) ou bien la concentration d'une certaine molécule dans le sang du patient.

Le modèle semble plus naturellement décrit sous forme hiérarchique : (1) une loi (marginale) pour L et (2) une loi (conditionnelle) pour X, dépendant de la valeur de L. Plus précisément :

1. L suit une loi de Bernoulli de paramètre p :

$$L \sim \mathcal{B}(\ell\,;\,p) \quad \equiv \quad \Pr\left[L = \ell \,|\, \boldsymbol{\theta}\right] = \begin{cases} p & \text{si } \ell = 1 \\ 1-p & \text{si } \ell = 2 \end{cases}$$

2. X sachant L suit une distribution gaussienne dont les paramètres dépendent de la valeur de L

$$\begin{cases} X \,|\, L = 1, \boldsymbol{\theta} & \sim \quad f_{X|L,\boldsymbol{\Theta}}(x \,|\, L = 1, \boldsymbol{\theta}) = \mathcal{N}(x\,;\,\mu_1, \gamma_1) \\ X \,|\, L = 2, \boldsymbol{\theta} & \sim \quad f_{X|L,\boldsymbol{\Theta}}(x \,|\, L = 2, \boldsymbol{\theta}) = \mathcal{N}(x\,;\,\mu_2, \gamma_2) \end{cases}$$

   lorsque $L = 1$, la distribution est paramétrée par (μ_1, γ_1) et lorsque $L = 2$, par (μ_2, γ_2).

Dans ces relations, le vecteur $\boldsymbol{\theta}$ collectionne l'ensemble des cinq paramètres inconnus d'intérêt : $\boldsymbol{\theta} = [\mu_1, \gamma_1, \mu_2, \gamma_2, p]$.

Remarque 1 — ***Notation.*** *Concernant la loi pour L, on peut la mettre sous la forme plus compacte :*

$$\Pr\left[L = \ell \,|\, \boldsymbol{\theta}\right] = p^{\delta(\ell,1)}\,(1-p)^{\delta(\ell,2)} \tag{1}$$

où δ est le symbole de Kronecker : $\delta(u, v) = 1$ si $u = v$ et 0 sinon. Pour ce qui est de la loi pour $X \,|\, L, \boldsymbol{\Theta}$, on peut aussi l'écrire de manière plus compacte :

$$f_{X|L,\boldsymbol{\Theta}}(x \,|\, L = \ell, \boldsymbol{\theta}) = \mathcal{N}(x\,;\,\mu_\ell, \gamma_\ell)\,. \tag{2}$$

Ces deux écritures seront utiles dans la suite, surtout dans la partie 3.

On déduit la distribution jointe pour (X, L) comme produit de ces deux distributions et on obtient la marginale pour X en sommant sur L :

$$\begin{aligned} f_{X|\boldsymbol{\Theta}}(x \,|\, \boldsymbol{\theta}) &= f_{X|L,\boldsymbol{\Theta}}(x \,|\, L=1, \boldsymbol{\theta}) \Pr[L=1 \,|\, \boldsymbol{\theta}] \;+\; f_{X|L,\boldsymbol{\Theta}}(x \,|\, L=2, \boldsymbol{\theta}) \Pr[L=2 \,|\, \boldsymbol{\theta}] \\ &= p\,\mathcal{N}(x\,;\,\mu_1,\gamma_1) \;+\; (1-p)\,\mathcal{N}(x\,;\,\mu_2,\gamma_2) \end{aligned} \tag{3}$$

Il s'agit d'une combinaison convexe des deux densités gaussiennes.

1. *Grace à la routine* `TrMixture`*, tracez cette densité pour différentes valeurs des paramètres.*

1a. *Un cas particulier :* $\mu_1 = -1$, $\mu_2 = +1$ *et* $p = 1/2$*. On se place dans la situation* $\gamma_1 = \gamma_2 = \gamma$*. Pour quelles valeurs de* γ *distingue-t-on visuellement les deux gaussiennes ? A quel écart-type cela correspond-il ?*

1b. *Un autre cas particulier :* $\mu_1 = \mu_2 = 0$ *et* p *proche de 1. Que se passe-t-il lorsque* γ_1 *est grand et* γ_2 *petit. A quoi peut servir ce modèle ?*

Remarque 2 — *Ce modèle se généralise volontiers au cas vectoriel (*X *devient un vecteur* $\boldsymbol{X}$ *de distribution paramétrée par une moyenne* $\boldsymbol{\mu}$ *et une précision* $\boldsymbol{\Gamma}$*) et au cas d'un nombre de composantes supérieur :*

$$f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x} \,|\, \boldsymbol{\theta}) = \sum_{c=1}^{C} p_c\,\mathcal{N}(\boldsymbol{x}\,;\,\boldsymbol{\mu}_c, \boldsymbol{\Gamma}_c) \quad \text{avec} \quad \sum_{c=1}^{C} p_c = 1\,.$$

Il se généralise également à une infinité de composantes et même à un « continuum » de composantes : on parle de continuous mixture of Gaussians (e.g., location mixture of Gaussians ou encore scale mixture of Gaussians. On obtient ainsi une très large classe de modèles non-gaussiens qui sont relativement aisés à manipuler car conditionnellement gaussiens.

1 Observations

Dans ce travail, on observe N réalisations indépendantes et identiquement distribuées sous la densité $f_{X|\boldsymbol{\Theta}}(x \,|\, \boldsymbol{\theta})$. On les regroupe dans le vecteur $\boldsymbol{x} = [x_1, x_2, \ldots x_N]$.

Chacune de ces quantités est implicitement ou explicitement liée à une variable d'étiquette L_n qui prend une valeur $\ell_n \in \{1, 2\}$. On précise ici que ces étiquettes ne sont pas observées.

Remarque 3 — *Parfois, les données sont indirectement observées et peuvent être déformées ou bruitées. On aurait par exemple :* $y_n = \varphi(x_n) + \varepsilon_n$*, pour* $n = 1, 2, \ldots N$*. Ce ne sera pas le cas ici.*

2. *Utilisez la fonction Matlab* `SimulMixture`*, qui simule des échantillons (des observations) d'un mélange de gaussiennes scalaires avec différentes valeurs des paramètres.*

3. *Des « vraisemblances ».*

3a. *Justifiez que la vraisemblance de $\boldsymbol{\theta}$ attachée aux données $\boldsymbol{x}$ prend la forme :*

$$f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \prod_{n=1}^{N} \left[\, p\,\mathcal{N}(x_n\,;\,\mu_1,\gamma_1) \;+\; (1-p)\,\mathcal{N}(x_n\,;\,\mu_2,\gamma_2)\,\right] \tag{4}$$

ce qui n'est pas très amical à manipuler.

3b. *Vérifiez que la « vraisemblance » de $(\boldsymbol{\theta},\boldsymbol{\ell})$ attachée à $\boldsymbol{x}$ prend la forme :*

$$f_{\boldsymbol{X}|\boldsymbol{\Theta},\boldsymbol{L}}(\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\ell}) = \prod_{n=1}^{N} \mathcal{N}(x_n\,;\,\mu_{\ell_n},\gamma_{\ell_n}) \tag{5}$$

ce qui est plus amical. Vérifiez également que le produit se sépare naturellement :

$$f_{\boldsymbol{X}|\boldsymbol{\Theta},\boldsymbol{L}}(\boldsymbol{x}\,|\,\boldsymbol{\theta},\boldsymbol{\ell}) = \prod_{n\in\mathcal{E}_1} \mathcal{N}(x_n\,;\,\mu_1,\gamma_1) \;\; \prod_{n\in\mathcal{E}_2} \mathcal{N}(x_n\,;\,\mu_2,\gamma_2) \tag{6}$$

et préciser les ensembles $\mathcal{E}_1$ et $\mathcal{E}_2$.

2 Cas d'une seule gaussienne

2.1 Données et vraisemblance

Dans cette première partie, on s'intéresse au cas simplifié où on a une seule gaussienne. On a alors $K = 1$, les étiquettes ne prennent qu'une seule valeur $L = 1$, autant dire qu'il n'y a pas d'étiquette… Le vecteur des paramètres est $\boldsymbol{\theta} = [\mu, \gamma]$.

4. *Utilisez la fonction Matlab* `SimulMixture` *pour simuler un jeu de données dans ce cas là. Calculez la moyenne et la précision empiriques. Comparez-les aux vraies valeurs.*

5. *Vérifiez que la vraisemblance de $\boldsymbol{\theta}$ attachée aux données $\boldsymbol{x}$ prend la forme :*

$$f_{\boldsymbol{X}|\boldsymbol{\Theta}}(\boldsymbol{x}\,|\,\boldsymbol{\theta}) = \prod_{n=1}^{N} \mathcal{N}(x_n\,;\,\mu,\gamma)\,.$$

Sachant que la densité gaussienne s'écrit :

$$\mathcal{N}(x\,;\,\mu,\gamma) = (2\pi)^{-1/2}\;\gamma^{1/2}\exp\left[-\gamma\,(x-\mu)^2\,/\,2\right]$$

explicitez cette vraisemblance. On pourra faire intervenir la moyenne empirique et la variance empirique que l'on pourra noter $m(\boldsymbol{x})$ et $v(\boldsymbol{x})$.

2.2 Prior et Posterior

Pour estimer le paramètre $\boldsymbol{\theta}$, on se donne une densité *a priori* $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$. On la choisit de manière à ce qu'elle soit peu informative et qu'elle permette des calculs théoriques et numériques aisés. Un premier choix est celui de la séparabilité :

$$\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \pi_M(\mu)\ \pi_\Gamma(\gamma)$$

qui, en un sens, permet de ne pas se prononcer *a priori* sur un lien potentiel entre μ et γ. Un second argument est celui de la conjugaison : pour μ on adopte une distribution gausienne (paramètres μ_0, γ_0) et pour γ une distribution Gamma (paramètres α_0, β_0) :

$$\pi_M(\mu) = (2\pi)^{-1/2}\,\gamma_0^{1/2}\,\exp\left[-\gamma_0\,(\mu-\mu_0)^2\,/\,2\right] \tag{7}$$

$$\pi_\Gamma(\gamma) = \frac{\beta_0^{\alpha_0}}{\Gamma[\alpha_0]}\,\gamma^{\alpha_0-1}\,\exp\left[-\beta_0\gamma\right]\,\mathbb{1}_{\mathbb{R}_+}(\gamma) \tag{8}$$

Pour être en adéquation avec l'idée de distribution *a priori* peu informative :

- on choisira γ_0 petit : par exemple $\gamma_0 = 10^{-4}$ ce qui correspond à un écart-type de 100 (autour de la valeur moyenne μ_0) ;
- on se placera dans le cas $\alpha_0 = \beta_0$ et on choisira un valeur faible : par exemple $\alpha_0 = \beta_0 = 10^{-4}$ ce qui corespond à une erreur relative de 100 (pour une valeur nominale de 1).

6. *Partant des distributions a priori (7) et (8) ainsi que de la vraisemblance obtenue à la question 5, explicitez densité a posteriori $\pi_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$ pour le vecteur des paramètres $\boldsymbol{\theta}$ étant donné les observations $\boldsymbol{x}$, à un coefficient multiplicatif près.*

7. *Question facultative. Calculez numériquement et tracez cette densité ou son co-logarithme sur une grille de valeur de (μ, γ). Déterminez le maximiseur sur la grille et tracez-le aussi. Positionnez également la vraie valeur des paramètres ainsi que les estimées empiriques.*

2.3 Échantillonnage du Posterior

Pour terminer, on s'intéresse à l'échantillonnage de la densité *a posteriori*, c'est-à-dire à la production d'échantillons (de réalisations) de $\boldsymbol{\theta}$ sous $\pi_{\boldsymbol{\Theta}|\boldsymbol{X}}(\boldsymbol{\theta}|\boldsymbol{x})$: il s'agit d'un problème d'*échantillonnage stochastique*. Compte-tenu de la forme obtenue précédemment, il est manifeste qu'il ne s'agit pas d'une distribution usuelle [1].

On montre qu'on obtient des échantillons de cette densité pour $\boldsymbol{\theta}$ en alternant l'échantillonnage de chaque composante du vecteur $\boldsymbol{\theta}$ sous sa densité conditionnelle. Cette structure d'échantillonnage alternée porte de nom de *Gibbs*. Il s'agit d'une structure générale qui permet de découper le problème de l'échantillonnage d'une densité multidimensionnelle en sous problèmes d'échantillonnage de plus faible dimension (ou scalaire). Elle entre dans la classe plus générale des algorithmes de *Monte-Carlo par Chaîne de Markov*. Une forme symbolique de l'algorithme est figurée ci-dessous.

1. En fait, il s'agit d'une distribution gamma-normale, pas si inhabituelle que ça…

- Initialisation (trois possibilités parmi d'autres) :
  - — Valeurs arbitraires : $\mu^{[0]}$ à 0 et $\gamma^{[0]}$ à 1.
  - — Un échantillon de la loi *a priori*.
  - — Les estimées empiriques.
- Boucle pour $k = 1, \ldots, K$
  1. Tirer $\mu^{[k]}$ sous la densité de $(\mu|\boldsymbol{x}, \gamma^{[k-1]})$
  2. Tirer $\gamma^{[k]}$ sous la densité de $(\gamma|\boldsymbol{x}, \mu^{[k]})$

Pour le mettre en œuvre, il faut déterminer les deux densités *a posteriori* conditionnelles pour $(\mu|\boldsymbol{x}, \gamma)$ et pour $(\gamma|\boldsymbol{x}, \mu)$. Nous allons voir qu'il s'agit de distributions usuelles (gaussienne et gamma).

8. *Détermination des conditionnelles a posteriori.*

8a. ***Moyenne.*** *Déduisez la densité a posteriori conditionnelle pour μ, i.e., pour $(\mu|\boldsymbol{x}, \gamma)$. Montrez qu'elle est gaussienne et donnez ses paramètres.*

8b. ***Précision.*** *Déduisez également la densité a posteriori conditionnelle pour γ, i.e., pour $(\gamma|\boldsymbol{x}, \mu)$. Montrez que c'est une densité gamma et identifiez ses paramètres.*

On peut maintenant encoder l'algorithme, le faire fonctionner, observer son comportement et obtenir les paramètres estimés ainsi que d'autres informations sur la (les) distribution(s) *a posteriori*. L'encodage se réalise à l'aide d'une seule boucle (celle de l'algorithme de Gibbs).

9. *Quelques affichages et quelques réflexions.*

9a. *Affichez les deux séries de paramètres $\mu^{[k]}$ et $\gamma^{[k]}$ pour $k = 1, \ldots, K$, à la fois sous forme de signaux et sous forme d'histogrammes. Il s'agit d'histogrammes de quelle densité ?*

9b. *Représentez également les échantillons en tant que couples $(\mu^{[k]}, \gamma^{[k]})$ dans le plan. Il s'agit de nuage de points sous quelle densité ? On pourra éventuellement superposer la représentation de la densité a posteriori de la question 7.*

10. *Quelques calculs numériques finauds.*

10a. *A partir des $\mu^{[k]}$ et $\gamma^{[k]}$ pour $k = 1, 2, \ldots$ calculez une approximation de l'estimateur « moyenne a posteriori ». Calculez également une approximation des écarts-types a posteriori pour μ et γ. Pour terminer, calculez la corrélation a posteriori du couple (μ, γ).*

10b. *Toujours à partir des $\mu^{[k]}$ et $\gamma^{[k]}$, calculez la probabilité que μ d'une part et γ d'autre part, soient dans un intervalle centré sur leur moyenne et de demi-largeur égale à leur écart-type (intervalle de crédibilité).*

3 Cas de deux gaussiennes

Dans cette partie, on revient au problème central concernant le mélange de deux gaussiennes. Le paramètre d'intérêt est à nouveau $\boldsymbol{\theta} = [\mu_1, \gamma_1, \mu_2, \gamma_2, p]$. Son estimation suit le même schéma que précédemment : on introduit les données et les vraisemblances, puis on précise le choix du prior et on déduit le posterior, finalement on traite la question de l'échantillonnage.

3.1 Données et vraisemblance

Il y a essentiellement deux manières d'aborder la question : l'une rend explicite les variables d'étiquette (bien qu'elles ne soient pas observées) et l'autre les laisse implicites. Pour ce qui est de l'identification des paramètres du modèle, les deux approches sont équivalentes mais du point de vue de la mise en œuvre les choses sont différentes.

La version explicite peut sembler plus compliquée puisque elle va nécessiter l'estimation de ces étiquettes. En fait, elle est plus simple car elle s'appuie sur des modèles conditionnellement gaussiens, ce que ne fait pas la forme implicite. Dit autrement, la version explicite s'appuie sur la vraisemblance (5) alors que l'implicite repose sur la vraisemblance (4).

3.2 Prior et Posterior

Concernant la densité *a priori* $\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$, on la choisit ici encore peu informative. Le premier choix est aussi celui de la séparabilité :

$$\pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta}) = \pi_{M_1}(\mu_1)\, \pi_{\Gamma_1}(\gamma_1)\, \pi_{M_2}(\mu_2)\, \pi_{\Gamma_2}(\gamma_2)\, \pi_P(p)\,.$$

Pour ce qui est de la forme de la distribution, on garde la même que dans la partie précédente :

— gaussienne de paramètres (μ_0, γ_0) pour μ_1 et μ_2

— gamma de paramètres (α_0, β_0) pour γ_1 et γ_2.

Concernant le paramètre p, on se donne une répartition uniforme sur l'intervalle $[0, 1]$:

$$\pi_P(p) = \mathbb{1}(p)\,,$$

et on note qu'il s'agit d'un cas particulier de la distribution Bêta.

On en déduit alors la distribution *a posteriori* (étendue aux étiquettes) :

$$\pi_{\boldsymbol{\Theta},\boldsymbol{L}|\boldsymbol{X}}(\boldsymbol{\theta}, \boldsymbol{\ell}|\boldsymbol{x}) \propto f_{\boldsymbol{X}|\boldsymbol{\Theta},\boldsymbol{L}}(x \,|\, \boldsymbol{\theta}, \boldsymbol{\ell})\, \Pr\left[\boldsymbol{L} = \boldsymbol{\ell} \,|\, \boldsymbol{\theta}\right] \pi_{\boldsymbol{\Theta}}(\boldsymbol{\theta})$$

à un coefficient de proportionalité près.

11. *En se basant sur les cinq distributions a priori, sur la loi des étiquettes (1) et la densité gaussienne conditionnelle (2) ainsi que sur la vraisemblance (5) ou sa forme séparée (6), explicitez complètement cette distibution.*

11a. *Repérez précisément les facteurs qui dépendent de chacune des variables μ_1, γ_1, μ_2, γ_2 et p. Idem pour chacun des ℓ_n.*

11b. *Analysez les deux formes : celle fondée sur la vraisemblance (5) et celle fondée sur la forme séparée (6).*

3.3 Échantillonnage du Posterior

On va obtenir des échantillons de cette distribution pour $(\boldsymbol{\Theta}, \boldsymbol{L})$ en utilisant un algorithme de Gibbs, généralisant celui de la partie précédente. Il doit échantillonner successivement les cinq quantités : μ_1, γ_1, μ_2, γ_2 et p ainsi que chacune des étiquettes l_n sous sa conditionnelle *a posteriori*.

12. *Pour les paramètres des gaussiennes (μ_1, γ_1) et (μ_2, γ_2)*

12a. ***Moyennes.*** *Vérifiez qu'on obtient la même densité a posteriori conditionnelle que précédemment pour μ_1 et μ_2, en ne considérant que les données associées à l'étiquette 1 pour μ_1 et à l'étiquette 2 pour μ_2.*

12b. ***Précisions.*** *Similairement, vérifiez qu'on obtient la même densité que précédemment pour γ_1 et γ_2, en ne considérant que les données associées aux étiquettes 1 et 2 respectivement.*

13. *Pour le paramètre de Bernoulli p, montrez qu'on obtient une densité Bêta de paramètres*

$$a = \nu_1(\boldsymbol{\ell}) \quad \text{et} \quad b = \nu_2(\boldsymbol{\ell})$$

le nombre d'étiquette à 1 et à 2 respectivement (ce sont les cardinaux respectifs de $\mathcal{E}_1$ et $\mathcal{E}_2$, voir la vraisemblance (6)).

14. *Pour les étiquettes L_n.*

14a. *Montrez que la distribution pour l'ensemble des étiquettes $\boldsymbol{L}$ est séparable. On peut donc les échantillonner séparément en parallèle.*

14b. *Montrez que la loi de l'étiquette L_n ne dépend que de la donnée x_n (et de p).*

14c. *Montrez que les deux probabilités s'écrivent :*

$$\Pr\left[L_n = 1 \,|\, \star\right] = \frac{p\,\rho_{1,n}}{p\,\rho_{1,n} + (1-p)\,\rho_{2,n}} \quad \text{et} \quad \Pr\left[L_n = 2 \,|\, \star\right] = \frac{(1-p)\,\rho_{2,n}}{p\,\rho_{1,n} + (1-p)\,\rho_{2,n}}$$

où les quantités $\rho_{1,n}$ et $\rho_{2,n}$ sont définies par

$$\rho_{1,n} = \gamma_1^{1/2}\,\exp\left[-\gamma_1(\mu_1 - x_n)^2\,/\,2\right] \quad \text{et} \quad \rho_{2,n} = \gamma_2^{1/2}\,\exp\left[-\gamma_2(\mu_2 - x_n)^2\,/\,2\right]$$

ce qui détermine le paramètre de Bernoulli a posteriori pour les L_n.

Comme précédemment, on peut alors encoder l'algorithme, le faire fonctionner, observer son comportement et obtenir les paramètres estimés ainsi que diverses informations sur les distributions *a posteriori*. Ici encore, l'encodage se réalise à l'aide d'une seule boucle.

15. *Affichez les cinq séries de paramètres $\mu_1^{[k]}$, $\gamma_1^{[k]}$, $\mu_2^{[k]}$, $\gamma_2^{[k]}$ et $p^{[k]}$, sous forme de signaux et d'histogrammes. Déterminez les paramètres estimés au sens de la moyenne a posteriori ainsi que les écarts-types a posteriori. On pourra également s'intéresser aux intervalles de crédibilité comme dans la question 10b.*

16. *Observez également l'évolution de certaines étiquettes. Déterminez les étiquettes estimées pour chaque donnée en précisant l'estimateur choisi.*