Un modèle pour les nids d'oiseaux

CARVAILLO, CÔME, PRALON

Soutenance de projet de Master 1

3 juin 2022

Introduction

Introduction

Je bite dans vos culs, je bite dans vos bouches ! ! !

Définition et hypothèses

Loi de mélange

Si l'on se donne J densités $f_1(x), \cdots, f_J(x)$, alors toute variable aléatoire X dont la densité f s'exprime, pour tout $x \in \mathbb{R}$, sous la forme

$$f(x) := \sum_{j=1}^{J} \alpha_j f_j(x)$$

où

$$\alpha_j \in \mathbb{R}_+^* \text{ et } \sum_{j=1}^{J} \alpha_j = 1$$

suit une loi de mélange continue.

Définition et hypothèses

Une histoire de variables

Nous introduisons les deux variables aléatoires (V.A.) suivantes :

- la V.A. X, modélisant le volume des nids, de densité f
- la V.A. discrète $Z \in [\![1, J]\!]$, représentant l'espèce d'oiseau

Hypothèse 1

X conditionnellement à $(Z = j)$ suit une loi normale $\mathcal{N}(\mu_j, v_j)$

Hypothèse 2

$$\exists \theta = (\alpha_j, \mu_j, v_j) \in \{\}$$

à voir

Une histoire de densités

- Densité de la loi X conditionnellement à $(Z = j)$:

$$f(x|Z = j) = \gamma_{\mu_j, v_j}(x)$$

- Densité de la loi de X :

$$f_\theta(x) = \sum_{j=1}^{J} \alpha_j \gamma_{\mu_j, v_j}(x)$$

- Probabilité de la loi de Z conditionnellement à $(X = x)$:

$$\mathbb{P}_\theta(Z = j|X = x) = \frac{\gamma_{\mu_j, v_j} \times \alpha_j}{f_\theta(x)}$$

Une approche idéaliste

- Nous observons et le volume et l'espèce d'oiseau
- Log-vraisemblance du modèle :

$$
\begin{aligned}
&\mathcal{L}_\theta(X_1, \cdots, X_n, Z_1, \cdots, Z_n) \\
&= ln\left(\prod_{i=1}^{n} h_\theta(X_i, Z_i)\right) \\
&= \sum_{j=1}^{J} \#A_j ln(\alpha_j) + \sum_{j=1}^{J}\sum_{i\in A_j} ln(\gamma_{\mu_j, v_j}(X_i))
\end{aligned}
$$

où

$$A_j := \{i \in [\![1, n]\!] \text{ tels que } Z_i = j\}$$

Une approche idéaliste

Estimateurs du maximum de vraisemblance (EMV)

$$\widehat{\alpha_j} = \frac{\#A_j}{n}$$

$$\widehat{\mu_j} = \frac{\sum_{i \in A_j} X_i}{\#A_j}$$

$$\widehat{v_j} = \frac{\sum_{i \in A_j} (X_i - \widehat{\mu_j})^2}{\#A_j}$$

Une approche réaliste

- Nous observons seulement le volume du nids
- Log-vraisemblance du modèle :

$$\begin{aligned}&\mathcal{L}_{obs}(\theta, X_1, \cdots, X_n)\\&= ln\left(\prod_{i=1}^{n} f_\theta(X_i)\right)\\&= \sum_{i=1}^{n} ln\left(\sum_{j=1}^{J} \alpha_j \gamma_{\mu_j, v_j}(X_i)\right)\end{aligned}$$

- L'existence d'une expression analytique des EMV n'est pas assurée

Log-vraisemblance conditionnelle

- texte
- Nous définissons ainsi la log-vraisemblance conditionnelle comme :

$$\mathcal{L}_c(\theta, \tilde{\theta}, X_1, \cdots, X_n) = \mathbb{E}_{\tilde{\theta}}[\mathcal{L}_\theta(X_1, \cdots, X_n, Z_1, \cdots, Z_n)|X_1, \cdots, X_n]$$

L'algorithme EM

Pseudo code de l'algorithme EM

Algorithm 1 L'algorithme EM (Dempster et al., 1977).

Entrée(s) : $\tilde{\theta}_0 \in \Theta$, un jeu de données $X_1 \cdots X_n$, $K \in \mathbb{N}$;

1: **pour** k allant de 1 à K **faire**

2: **ETAPE E :** *Calculer la probabilité* $\mathbb{P}_{\tilde{\theta}_{k-1}}(Z = j | X = X_i) = \dfrac{\alpha_j \times \gamma_{\mu_j, j_v}}{\sum_{k=1}^{J} \alpha_k \times \gamma_{\mu_k, v_k}}$, $\forall i \in [\![1, n]\!]$

3: **ETAPE M :** *Calculer* $\tilde{\theta}_k = \underset{\theta=(\alpha_j, \mu_j, v_j)_{j \in [\![1,J]\!]}}{\operatorname{argmax}} \mathbb{P}_{\tilde{\theta}_{k-1}}(Z = j | X = X_i)$;

4: **fin du pour**

5: **retourner** $\tilde{\theta}_K$;

Les étapes de l'algorithme EM

L'étape E (Expectation)

Consiste à déterminer $\mathbb{P}_{\tilde{\theta}}(Z = j|X = X_i)$ à l'aide de la formule suivante :

$$\mathbb{P}_{\tilde{\theta}}(Z = j|X = X_i) = \frac{\alpha_j \times \gamma_{\mu_j, v_j}}{\sum_{k=1}^{J} \alpha_k \times \gamma_{\mu_k, v_k}}$$

Les étapes de l'algorithme EM

L'étape M (Maximization)

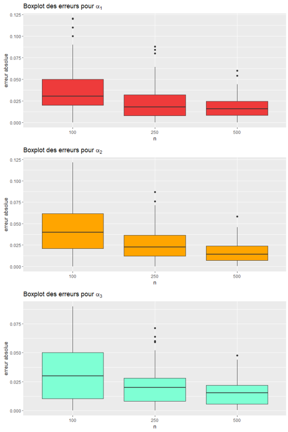
Consiste à déterminer les EMV $(\widehat{\alpha_j}, \widehat{\mu_j}, \widehat{\sigma_j})$ de la log-vraisemblance conditionnelle via les formules suivantes :

$$\widehat{\alpha_j} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)$$

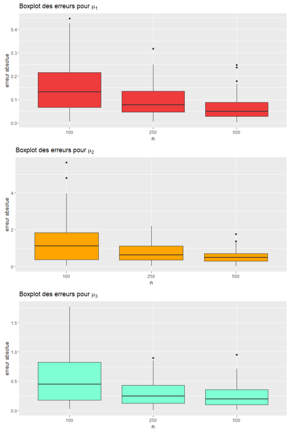
$$\widehat{\mu_j} = \frac{\sum_{i=1}^{n} X_i \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}{\sum_{i=1}^{n} \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}$$

$$\widehat{v_j} = \frac{\sum_{i=1}^{n} (X_i - \widehat{\mu_j})^2 \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}{\sum_{i=1}^{n} \mathbb{P}_{\tilde{\theta}}(Z = j | X = X_i)}$$

Cas des variables à "fortes séparations"

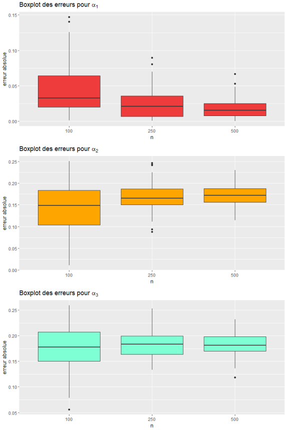


(a) Boxplot des erreurs pour α_1, α_2 et α_3

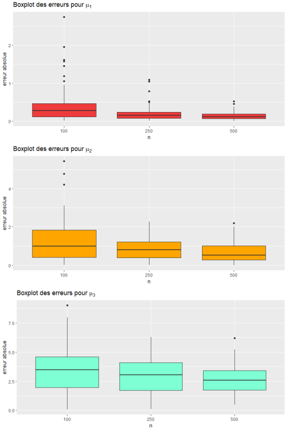


(b) Boxplot des erreurs pour μ_1, μ_2 et μ_3

Cas des variables à "faibles séparations"



(c) Boxplot des erreurs pour α_1, α_2 et α_3



(d) Boxplot des erreurs pour μ_1, μ_2 et μ_3

Un théorème de croissance

Théorème

Soit $(\theta_k)_{k \in [\![1,K]\!]}$ la suite de paramètres construite à l'aide de l'algorithme EM.
La log-vraisemblance $\mathcal{L}_{obs}$ des observations vérifie

$$\mathcal{L}_{obs}(\theta_{k+1}, X_1, \cdots, X_n) \geq \mathcal{L}_{obs}(\theta_k, X_1, \cdots, X_n)$$

Ersatz de preuve

- Nous cherchons donc à montrer que

$$\mathcal{L}_{obs}(\theta_{k+1}, X_1, \cdots, X_n) - \mathcal{L}_{obs}(\theta_k, X_1, \cdots, X_n) \geq 0$$

- Réécriture :

$$\mathcal{L}_c(\theta_{k+1}, \theta_k, X_1, \cdots, X_n) = \mathcal{L}_{obs}(\theta_{k+1}, X_1, \cdots, X_n) + \kappa_{\theta_{k+1}, \theta_k}$$

- Avec

$$\kappa_{\theta_{k+1}, \theta_k} = \sum_{i=1}^{n} \sum_{j=1}^{J} ln(\mathbb{P}_{\theta_{k+1}}(Z = j | X = X_i)) \times \mathbb{P}_{\theta_k}(Z = j | X = X_i)$$

Ersatz de preuve

Ainsi,

$$\mathcal{L}_{obs}(\theta_{k+1}, X_1, \cdots, X_n) - \mathcal{L}_{obs}(\theta_k, X_1, \cdots, X_n)$$

$$= \mathcal{L}_c(\theta_{k+1}, \theta_k, X_1, \cdots, X_n) - \kappa_{\theta_{k+1},\theta_k} - \mathcal{L}_c(\theta_k, \theta_k, X_1, \cdots, X_n) + \kappa_{\theta_k,\theta_k}$$

Ersatz de preuve

- A l'étape M de l'algorithme, la quantité

$$\mathcal{L}_c(\theta, \theta_k, X_1, \cdots, X_n)$$

est maximisée en θ, de maximum θ_{k+1}

- Donc,

$$\mathcal{L}_c(\theta_{k+1}, \theta_k, X_1, \cdots, X_n) - \mathcal{L}_c(\theta_k, \theta_k, X_1, \cdots, X_n) \geq 0$$

Ersatz de preuve

- Il reste donc à prouver que

$$\kappa_{\theta_k,\theta_k}, -\kappa_{\theta_{k+1},\theta_k} \geq 0$$

- On montre que, après quelques fastidieux calculs,

$$\begin{aligned}
&\kappa_{\theta_k,\theta_k}, -\kappa_{\theta_{k+1},\theta_k} \\
&\geq -n \times ln\left(\sum_{i=1}^{n}\sum_{j=1}^{J}\mathbb{P}_{\theta_{k+1}}(Z=j|X=X_i)\times\frac{1}{n}\right) \\
&= -n \times ln(1) \\
&= 0
\end{aligned}$$

Préambule

	Female Body Mass (g)	Total mass of nest (g)	Cup diameter parallel to long axis (mm)	Cup diameter perpendicular to long axis (mm)	Nest diameter parallel to long axis (mm)	Nest diameter perpendicular to long axis (mm)	Upper wall thickness (mm)	Base Thickness (mm)	Cup depth (mm)	Nest Height (mm)	Volume (cm^3)
Fringillidae											
European Goldfinch (*Carduelis Carduelis*) [10]	16.4	8.3 ± 2.4	62.8 ± 12.1	54.8 ± 7.4	91.4 ± 9.3	77.8 ± 7.9	12.8 ± 3.3	15.7 ± 4.3	26.0 ± 5.5	41.6 ± 7.4	38.0 ± 9.1
Common Linnet (*Linaria cannabina*) [11]	18.0	18.9 ± 5.4	74.7 ± 6.3	59.9 ± 8.6	107.9 ± 8.8	95.1 ± 10.2	16.9 ± 4.9	24.5 ± 8.9	30.6 ± 9.8	55.1 ± 9.2	60.9 ± 20.8
Common Chaffinch (*Fringilla coelebs*) [11]	21.5	14.5 ± 2.9	63.3 ± 8.1	50.8 ± 8.0	98.7 ± 10.9	90.3 ± 9.8	18.5 ± 3.6	23.6 ± 7.6	34.3 ± 7.8	58.0 ± 7.3	58.3 ± 15.0
European Greenfinch (*Chloris chloris*) [5]	25.9	22.4 ± 6.2	75.6 ± 7.8	53.9 ± 11.8	128.6 ± 13.7	99.7 ± 16.2	24.9 ± 7.9	29.4 ± 6.0	35.4 ± 5.7	64.9 ± 9.4	74.5 ± 12.2
Eurasian Bullfinch (*Pyrrhula pyrrhula*) [17]	27.3	12.1 ± 4.6	80.8 ± 12.1	66.4 ± 8.1	129.7 ± 23.4	117.5 ± 19.6	24.8 ± 10.9	24.2 ± 10.7	22.6 ± 4.5	46.8 ± 11.3	45.0 ± 3.8
Hawfinch (*Coccothraustes coccothraustes*) [4]	52.9	27.4 ± 7.3	102.2 ± 17.9	78.8 ± 25.2	153.4 ± 19.1	131.3 ± 27.1	25.4 ± 5.9	23.3 ± 4.9	31.4 ± 10.9	54.7 ± 11.5	71.6 ± 12.9

Figure – Caractéristiques des nids

Hypothèses, outils et démarche

Hypothèses

- la distribution du volume des nids est gaussienne
- le nombre d'espèce J est connu

Outils

- fonction *simulation*
- fonction *algo_EM*

Démarche

- Génération de l'échantillon
- Représentation graphique de la densité de l'échantillon
- Détermination des paramètres initiaux
- Execution de l'algorithme EM

Première exploration des données

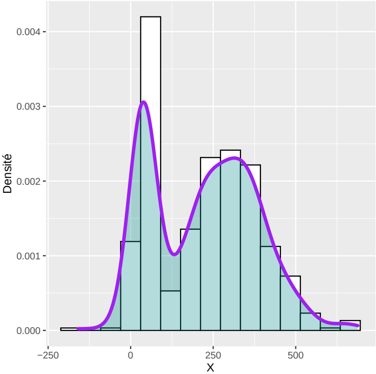


Figure – Densité du mélange

Heuristique graphique

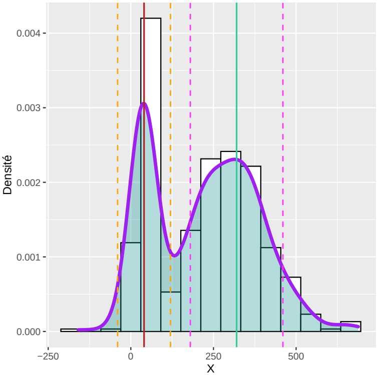


Figure – Détermination des valeurs initiales

Heuristique graphique

Paramètres initiaux

- $\mu 1_{init} = 40$ et $\mu 2_{init} = 320$
- $\sigma 1_{init} = 80$ et $\sigma 2_{init} = 140$
- $\alpha 1_{init} = 0.5$ et $\alpha 2_{init} = 0.5$

Résultats

```
          bird_names     alpha        mu      sigma
1 European Goldfinch 0.2910832  37.76285   9.512478
2         Ring Ouzel 0.7089168 302.51936 125.951894
```

Valeurs théoriques

```
         bird_names2 proportion_alpha mean_volume sd_volume
1 European Goldfinch        0.2878713        38.0       9.1
12        Ring Ouzel        0.7121287       298.6     125.1
```

Détermination automatique

Comme dans le rapport ou fonction de Nicolas ? ?

Conclusion

Conclusion

J'encule vos grosses marraines bien profond ! ! !