Projet BAYES

Théo Duquesne et Stéphane Courtault

Avril 2025

1 Introduction

Cette étude concerne une population de singes, ainsi que les sensibilités maximales de leurs yeux. Plus précisemment, on mesure pour chaque singe la longueur d'onde qu'il capte le mieux, et on analyse les résultats.

On supposera comme modèle (cf justification plus loin) que cet échantillon de données a été généré par un mélange de deux gaussiennes de même variance. Le but de cette étude statistique va être de déterminer à l'aide de méthodes de MCMC (Markov Chain Monte Carlo) les paramètres de la loi du modèle, ainsi que de pouvoir échantillonner selon cette loi.

2 Partie 1 : Justification du modèle

Pour vérifier la pertinence d'un mélange gaussien à deux composantes, nous avons commencé par effectuer une reconstruction à noyeaux, afin d'effectuer une confirmation visuelle.

2.1 Justification mélange gaussien

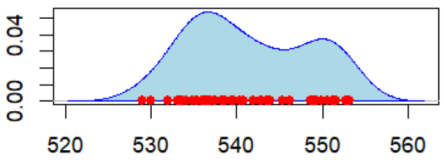


Figure 1: Résultat noyaux gaussiens

Si se baser seulement cette image n'est pas suffisant pour valider l'hypothèse du mélange gaussien, elle semble clairement un argument en sa faveur. De plus, notons que c'est un modèle très simple, et classique dans les répartitions de populations.

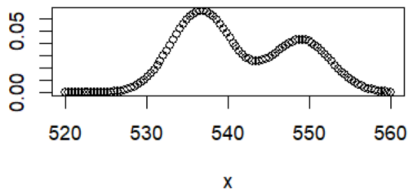
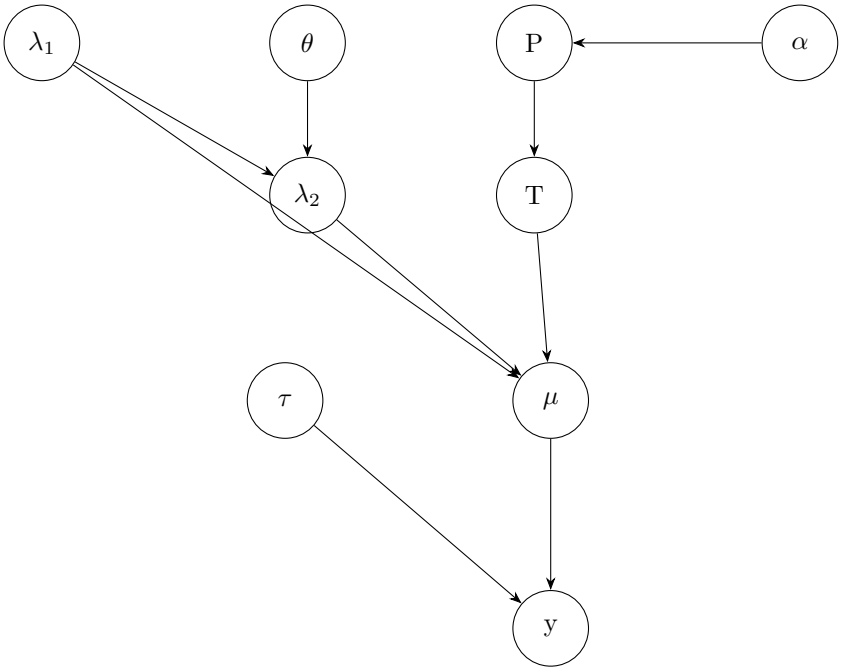


Figure 2: Résultat E.M.

2.2 Modèle

On suppose pour ce modèle l'existance de multiples variables latentes explicatives qui permettent de construire une chaîne de dépendance eplicitée par le DAG ce dessous



Voici un résumé rapide de ce que ces variables représentes :

- λ_1 :
  Représente la moyenne du groupe de plus faible moyenne. On suppose que cette variable suit une loi a priori gaussienne centrée de précision

10^{-6}. Comme cette valeur est très grande devant celles du problème, cela nous permet d'avoir une loi a priori non informative qui passe bien à la conjugaison.

- θ :
  Représente l'écart entre les moyennes des deux groupes. On suppose que cette variable suit une loi "demi-normale centrée" (valeur absolue d'une normale centrée en 0) de précision 10^{-6}, ce qui permet d'avoir une loi a priori non informative tout en forçant une difference positive entre la moyenne du groupe deux et celle du groupe 1, ce qui nous permettra d'éviter le phénimène du "label-switching" lors de l'execution de nos algorithmes de MCMC.

- λ_2 :
  Somme de λ_1 et de θ. Représente la moyenne de la loi normale selon laquelle est tirée le deuxièle groupe.

- alpha :
  Hyperparamètre qui va déterminer la répartition des singes en deux groupes. Une loi de Dirichlet est donc l'outil privilégié De plus, on a pas d'information a priori sur les proportions de singes dans chaque groupe, donc on choisit une loi de Dirichlet (1,1) (ie $\alpha = (1, 1)$)

- P:
  Représente la proportion des singes qui ira dans chaque groupe. Suit la loi a priori non informative de Dirichlet évoquée plus tôt.

- T :
  Variable latente. C'est un vecteur dont les composantes sont tirées a priori selon une loi de Bernouilli de paramètre P, et qui représente les attributions de groupe à chaque singe.

- μ :
  Variable latente. Ce vecteur réunit les informations sur les attributions aux groupes et sur les moyennes.

- τ :
  Va représenter plus tard la précision commune des gaussiennes du mélange gaussien. On suppose a priori que τ suive une loi $\Gamma(0.001, 0.001)$, ce qui permet :

-De faire en sorte d'avoir de petites valeurs pour tau, (la denstité décroit très rapidement depuis 0), donc une loi sans information a priori,
-D'avoir une loi qui se conjugue bien

- y :
  Le vecteur contenant nos données, supposées tirées selon un mélange gaussien. (Chaque y_i est tirée selon une $\mathcal{N}(\lambda_{T_i}, \tau^{-1})$

3 Calculs des lois marginales

Maintenant que nous avons un modèle bien définit, afin de pouvoir réaliser un echantilloneur de Gibbs, nous allons avoir besoin des lois marginales de chacune de ces variables aléatoires sachant toutes les autres

3.1 λ_1

$$\pi(\lambda_1 \mid \_) \propto \mathcal{L}(\mu \mid T, \lambda_1, \lambda_2) \cdot \mathcal{L}(\lambda_2 \mid \theta, \lambda_1) \cdot \pi(\lambda_1)$$

μ est déterministe en fonction de T, λ_1 et λ_2. Ainsi :

$$\mathcal{L}(\mu \mid T, \lambda_1, \lambda_2) \propto \prod_i \mathbb{1}_{\mu_i = \lambda_{T_i}}$$

De plus, λ_2 est déterministe en fonction de λ_1 et θ. Donc :

$$\mathcal{L}(\lambda_2 \mid \theta, \lambda_1) \propto \mathbb{1}_{\lambda_2 = \lambda_1 + \theta}$$

λ_1 suit une loi a priori normale centrée, d'écart-type 10^6. On en conclut que λ_1 suit une loi déterministe en $\lambda_2 - \theta$.

3.2 λ_2

$$\pi(\lambda_2 \mid \_) \propto \mathcal{L}(\mu \mid T, \lambda_1, \lambda_2) \cdot \pi(\lambda_2 \mid \theta, \lambda_1)$$

μ est déterministe en fonction de T, λ_1 et λ_2 :

$$\mathcal{L}(\mu \mid T, \lambda_1, \lambda_2) \propto \prod_i \mathbb{1}_{\mu_i = \lambda_{T_i}}$$

De plus :

$$\pi(\lambda_2 \mid \theta, \lambda_1) \propto \mathbb{1}_{\lambda_2 = \lambda_1 + \theta}$$

On en conclut que :

$$\pi(\lambda_2 \mid \_) \propto \prod_i \mathbb{1}_{\mu_i = \lambda_{T_i}} \cdot \mathbb{1}_{\lambda_2 = \lambda_1 + \theta}$$

3.3 θ

θ est déterministe en fonction de λ_1 et λ_2, donc on peut déjà affirmer que

$$\pi(\theta|\_) \propto \mathbb{1}_{\theta=\lambda_1-\lambda_2}$$

3.4 P

$$\pi(P \mid \_) \propto \mathcal{L}(T \mid P) \cdot \pi(P \mid \alpha)$$

P suit une loi de Dirichlet de paramètre α. T est un vecteur dont les composantes suivent des lois de Bernoulli iid de paramètre P.

Ces deux lois sont conjuguées. D'après Wikipédia, la loi a posteriori de P est une Dirichlet de paramètres proportionnels à $(\alpha + \#2, 1 - \alpha + \#1)$, où $\#i$ est le nombre d'éléments attribués au groupe i.

3.5 T

$$\pi(T \mid \_) \propto \mathcal{L}(\mu \mid T, \lambda_1, \lambda_2) \cdot \pi(T \mid P)$$

μ est déterministe en fonction de T, λ_1 et λ_2 :

$$\mathcal{L}(\mu \mid T, \lambda_1, \lambda_2) \propto \prod_i \mathbb{1}_{\mu_i=\lambda_{T_i}}$$

Ainsi, comme T est parfaitement défini par les valeurs prises par μ, il n'est même pas nécessaire de s'intéresser à $\pi(T \mid P)$.

Cependant, si on le fait :

$$\pi(T \mid P) \propto \prod_i (\mathbb{1}_{T_i=1} \cdot (1 - P) + \mathbb{1}_{T_i=2} \cdot P)$$

Donc, on retrouve encore :

$$\pi(T \mid \_) \propto \prod_i \mathbb{1}_{\mu_i=\lambda_{T_i}} \cdot \mathbb{1}_{\lambda_2=\lambda_1+\theta}$$

3.6 μ

$$\pi(\mu \mid \_) \propto \mathcal{L}(y \mid \mu, \tau) \cdot \pi(\mu \mid T, \lambda_1, \lambda_2)$$

Pour tout i, y_i suit une loi $\mathcal{N}(0, 1)$. Donc (en notant ϕ la densité d'une loi normale):

$$\mathcal{L}(y \mid \mu, \tau) \propto \prod_i \phi(y_i, \mu_i, \tfrac{1}{\tau})$$

De plus, μ_i est déterministe sachant T, λ_1, et λ_2. Donc on ne peut tirer les valeurs que parmi une unique configuration de μ.

3.7 τ

$$\pi(\tau \mid \_) \propto \mathcal{L}(y \mid \mu, \tau) \cdot \pi(\tau)$$

Or, $\tau \sim \Gamma(0.001, 0.001)$
$y \mid \mu, \tau \sim \mathcal{N}(\mu, \frac{1}{\tau})$
Donc on a avec les lois conjuguées et on a :

$$\tau \mid \_ \sim \Gamma(\alpha + \frac{n}{2}; \beta + \frac{\sum\limits_i (x_i - \lambda_1)^2}{2})$$

Comme nous venons de le voir, la création de beaucoup de variables latentes posent des problèmes : En effet, certaines de ces variables étant déterministes en fonction des autres (comme μ ou λ_2), les lois marginales sont "bloquées" par ces variables aléatoires "parents" ou "enfants", ce qui empêche artificiellement les différentes variables aléatoires de changer de valeur. Pour la suite de l'étude, nous avons ainsi décidé d'"élaguer" certaines variables aléatoires afin de ne garder que celles qui apportent des informations.
Ci dessous nous avons tracé le nouveau DAG correspondant.

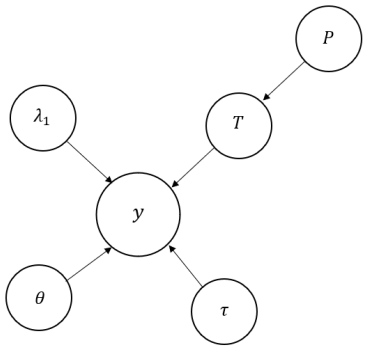


Figure 3: Nouveau DAG

3.8 Loi marginale de λ_1

$$\pi(\lambda_1 \mid \_) \propto \mathcal{L}(y \mid T, \lambda_1, \tau, \theta) \cdot \pi(\lambda_1)$$

Or,

$$\pi(y \mid \_) \propto \prod_i \phi(y_i, \lambda_1 + \mathbb{1}_{T_1=2}\theta, \frac{1}{\tau})$$

$$\pi(y \mid \_) \propto \prod_i \phi(y_i - \mathbb{1}_{T_1=2}\theta, \lambda_1, \frac{1}{\tau})$$

On reconnaît ici la vraisemblance d'un échantillon $(x_i)_{i\in[\![1,n]\!]} = (y_i - \mathbb{1}_{T_1=2})_{i\in[\![1,n]\!]}$

$$\pi(\lambda_1) \propto \phi(\lambda_1, 0, 10^6)$$

Les deux se conjuguent et nous donnent :

$$\lambda_1 \mid \_ \propto \mathcal{N}\left(\frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum\limits_i x_i}{\sigma^2}\right); \frac{1}{\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}}\right)$$

avec $\sigma_0^2 = 10^6$, $\sigma^2 = \frac{1}{\tau}$ et $x_i = y_i - \mathbb{1}_{T_1=2}\theta$

theta va suivre une loi demi normale qui est très facile à simuler à partir d'une loi normale (avec une méthode d'acceptation-rejet par exemple, métropolis-hastings étant ici peu adapté (trop lourd en calcul par rapport à d'autres méthodes)

3.9 Loi marginale de θ

$$\pi(\theta \mid \_) \propto \mathcal{L}(y \mid T, \lambda_1, \tau, \theta) \cdot \pi(\theta)$$

Or,

$$\pi(y \mid \_) \propto \prod_{i, T_i=2} \phi(y_i - \lambda_1, \theta, \frac{1}{\tau})$$

(On reconnaît encore la vraisemblance de données générées suivant une loi gaussienne) De plus,

$$\pi(\theta) \propto \mathbb{1}_{\theta\geq 0} \cdot \phi(\theta, 0, 10^6)$$

Pour se simplifier la vie dans les expressions suivantes, on utilisera les notations suivantes :
- $\sigma_0^2 = 10^6$
- $\sigma^2 = \frac{1}{\tau}$
- $\mu_0 = 0$
- $\underset{\sim}{(x_i)}_{i\in[\![1,k]\!]} = (y_j - \lambda_1)_{j\in\{u \text{ tels que } T_u=2\}}$
- $\sigma_f^2 = \frac{1}{\frac{1}{\sigma_0^2}+\frac{k}{\sigma^2}}$
- $\mu_f = (\frac{1}{\frac{1}{\sigma_0^2}+\frac{k}{\sigma^2}}\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum\limits_{i,T_i=2}\tilde{x_i}}{\sigma^2}\right)$

Or, on sait grâce à la formule de conjugaison d'une vraisemblance gaussienne de moyenne inconnue et d'une loi a priori gaussienne que :

$$\prod_{i\in[\![1,k]\!]} \phi(\tilde{x_i}, \theta, \frac{1}{\tau}) \cdot \phi(\theta, 0, 10^6) \propto \phi(\theta, \mu_f, \sigma_f^2)$$

et ainsi,

$$\prod_{i\in[\![1,k]\!]} \phi(\tilde{x_i}, \theta, \frac{1}{\tau}) \cdot \mathbb{1}_{\theta\geq 0} \cdot \phi(\theta, 0, 10^6) \propto \phi(\theta, \mu_f, \sigma_f^2)\mathbb{1}_{\theta\geq 0}.$$

Nous avons codé un Métropolis Hastings afin de rester dans le cadre du cours, mais, la plupart du temps, lors de l'execution de Gibbs nous avons plutôt utilisé la formule obtenue ci dessus afin de ne pas multiplier par 1000 le temps de calcul

3.10 Loi marginale de P

$$P \sim Dirichlet(\alpha + k, \alpha + n - k)$$

3.11 Loi marginale de τ

$$\pi(\tau \mid \_) \propto \mathcal{L}(y \mid \_) \cdot \pi(\tau)$$

Or, $\tau \sim \Gamma(\alpha, \beta)$
Ainsi, en conjugant les lois on retrouve :

$$\tau \mid \_ \propto \Gamma(\alpha + \frac{n}{2}; \beta + \frac{\sum\limits_i x_i}{2})$$

3.12 Loi marginale de T

On voit les T_i comme des variables latentes

$$\pi(T_i \mid \_) \propto \mathcal{L}(y_i \mid \_) \cdot \pi(T_i \mid P)$$

Or, $\pi(T_i \mid P) = P$ si $T_i = 2$ et $1 - P$ sinon

$$\pi(T \mid \_) \propto \phi(y_i, \lambda_1 + \mathbb{1}_{T_1=2}\theta, \frac{1}{\tau}) \cdot (P\mathbb{1}_{T_i=2} + (1-P)\mathbb{1}_{T_i=1})$$

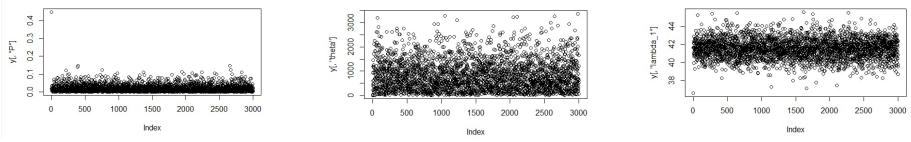
4 Simulation

Nous allons simuler nos variables à l'aide de l'algorithme de Gibbs

5 Résultats

On obtient systématiquement un résultat faux. Alors que la probabilité d'appartenance au groupe 2 aurait dû tendre vers 0.4, elle tendait systematiquement vers 0.5 .
Que ce soit une cause où une consequence, nous avons obtenu dans les meilleurs cas un λ autour de 40 et un θ autour de 5 ou 6, loin des valeurs de 36 et 12 respectivement que nous étions sensés trouver.

Lorsque l'on regarde les chaînes en détail, on remarque que θ reste souvent proche de 0, et parfois prend des valeurs extrêmement grandes.



(a) P (b) θ (c) λ_1

Figure 4: Échantillonnage sans tentatives d'interférence de notre part

Nous avons compris que cela arrive lorsque la variable latente T (qui attribue leurs labels aux points) est un vecteur uniforme en 1.
À ce moment, la vraisemblance n'a donc plus aucun impact sur la loi a posteriori de θ, et celle-ci se retrouve alors être la loi a priori, avec une très forte variance

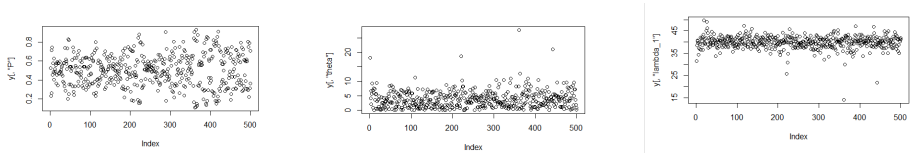
Notre meilleure hypothèse explicative est la suivante :

Nous sommes coincés dans un cercle vicieux car :
-τ est très petit
-les vraisemblances des y sachant leur label n'ont pas beaucoup d'influence sur les attributions des labels (car la vraisemblance est normale avec une faible precision)
- seul P compte dans l'attribution des labels, donc ne fait rien pour augmenter la précision τ (on ne s'attend pas vraiment à ce qu'un vecteur de bernouilli donne les bons labels par hasard, donc la variance doit pouvoir expliquer des résultats dispersés)
-on est dans une boucle de retroaction positive où P tend très rapidement vers 0 ou 1. En effet, P est mis à jour en se basant sur les proportions de y_i dans chaque groupe, et les attributions des labels des y_i ne dépendent pour ainsi dire que de P
-on reste coincés avec un vecteur T constant.
-comme tous les y_i sont dans le même groupe, la précision τ reste très faible puisqu'il faut pouvoir expliquer que toutes ces valeurs soient du même groupe.

Pour régler ce problème nous avons essayé plusieurs choses:

Nous avons forcé quelques labels extrémaux afin d'empêcher de faire entièrement disparaître une catégorie. Cela a consequemment augmenté la stabilité, mais, pour que le résultat soit un minimum concluant, il faut en fixer une quantité très importante, et donc interférer beaucoup avec l'algorithme.
Nous avons essayé de jouer sur les paramètres initiaux, en particulier de tau et de alpha, sans grand succès.

Nous avons essayé de faire du thinning pour retirer les valeurs aberrantes. Cela permis de retirer la majeure partie des valeurs aberrantes, mais les moyennes des paramètres demeurent néanmoins fausses.
Enfin, nous avons également essayé de forcer τ à prendre des valeurs plus grandes. Cependant, cela n'a pas non plus été concluant car, si les valeurs de tau sont trop faibles, le problème reste le même, et sinon, le modèle perd vraiment toute sa pertinence



(a) P (b) θ (c) λ_1

Figure 5: Étude de la chaîne de Markov avec les rectifications évoquées ci-dessus