Ecole Centrale de Nantes

Projet methodes bayésiennes et modèles hiérarchiques

Seeds: Random effect logistic regression

Auteurs:
Paul Trincklin
Jiacheng Fu
Enzo Mori
Grégory Perret

Encadrant:
Mathieu Ribatet

March 30, 2022

1 Problématique

L'objectif de ce projet est de modéliser la capacité de germination de deux types de graines *(O. aegyptiaco 75 et O. aegyptiaco 73)* pour deux variétés différentes (haricot et concombre). Pour cela on étudie 21 assiettes dans lesquelles ont été planté ces différents types de graines.

On note n_i le nombre total de graines plantées dans la ième assiette, r_i le nombre de graines ayant germées dans la i-ème assiette, et p_i la probabilité de germination d'une graine dans la i-ème assiette. Nous disposons du jeu de données suivant :

O. aegyptiaco											
Bean (75)			Cucumber (75)			Bean (73)			Cucumber(73)		
r	n	r/n	r	n	r/n	r	n	r/n	r	n	r/n
10	39	0.26	5	6	0.83	8	16	0.50	3	12	0.25
23	62	0.37	53	74	0.72	10	30	0.33	22	41	0.54
23	81	0.28	55	72	0.76	8	28	0.29	15	30	0.50
26	51	0.51	32	51	0.63	23	45	0.51	32	51	0.63
17	39	0.41	46	79	0.58	0	4	0.00	3	7	0.43
			10	13	0.77						

On utilise le modèle suivant :

- $r_i \sim Bin(p_i, n_i)$ avec $logit(p_i) = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + b_i$
- $b_i \sim N(0, \tau)$ où $\tau = \frac{1}{\sigma^2}$

$\rightarrow$ x_{1i} *est le type de graine (0 pour O. aegyptiaco 75, 1 pour O. aegyptiaco 73)*
$\rightarrow$ x_{2i} *est la variété (0 pour haricot, 1 pour concombre) dans la i-ème assiette*

2 Modélisation mathématique

Pour chacun de nos paramètres, on a:

$$\left.\begin{aligned} \alpha_0 &\sim N(0, 10^6) \\ \alpha_1 &\sim N(0, 10^6) \\ \alpha_2 &\sim N(0, 10^6) \\ \alpha_{12} &\sim N(0, 10^6) \\ \sigma^2 &\sim InverseGamma(10^{-3}, 10^{-3}) \end{aligned}\right\} \textit{Lois a priori}$$

On calcule les lois à posteriori pour les différents paramètres pour notre algorithme MCMC associé à un échantilloneur de Gibbs *(sans passer à l'échelle log ici, mais dans notre code pour simplifier les calculs)*:

- $$\begin{aligned}\pi(b_i \mid ...) &\propto \pi(b_i \mid \sigma^2) \times \pi(r_i \mid ...)\\ &\propto e^{-\frac{b_i^2}{2\sigma^2}} \times (p_i)^{r_i} \times (1-p_i)^{n_i-r_i}\end{aligned}$$

- $$\begin{aligned}\pi(\alpha_0 \mid ...) &\propto \pi(\alpha_0) \times \prod_{i=1}^{N} \pi(r_i \mid ...)\\ &\propto e^{-\frac{\alpha_0^2}{2\times 10^6}} \times \prod_{i=1}^{N} [(p_i)^{r_i} \times (1-p_i)^{n_i-r_i}]\end{aligned}$$

- $$\begin{aligned}\pi(\alpha_1 \mid ...) &\propto \pi(\alpha_1) \times \prod_{i=1}^{N} \pi(r_i \mid ...)\\ &\propto e^{-\frac{\alpha_1^2}{2\times 10^6}} \times \prod_{i=1}^{N} [(p_i)^{r_i} \times (1-p_i)^{n_i-r_i}]\end{aligned}$$

- $$\begin{aligned}\pi(\alpha_2 \mid ...) &\propto \pi(\alpha_2) \times \prod_{i=1}^{N} \pi(r_i \mid ...)\\ &\propto e^{-\frac{\alpha_2^2}{2\times 10^6}} \times \prod_{i=1}^{N} [(p_i)^{r_i} \times (1-p_i)^{n_i-r_i}]\end{aligned}$$

- $$\begin{aligned}\pi(\alpha_{12} \mid ...) &\propto \pi(\alpha_{12}) \times \prod_{i=1}^{N} \pi(r_i \mid ...)\\ &\propto e^{-\frac{\alpha_{12}^2}{2\times 10^6}} \times \prod_{i=1}^{N} [(p_i)^{r_i} \times (1-p_i)^{n_i-r_i}]\end{aligned}$$

- $$\begin{aligned}\pi(\sigma^2 \mid ...) &\propto \pi(\sigma^2) \times \prod_{i=1}^{N} \pi(b_i \mid \sigma^2)\\ &\propto (\sigma^2)^{(-0.001-1)} e^{-\frac{0.001}{\sigma^2}} \times (\frac{1}{\sigma})^N e^{-\frac{\sum_{i=1}^N b_i^2}{2\sigma^2}}\\ &\propto InverseGamma(0.001 + \frac{N}{2}, 0.001 + \frac{\sum_{i=1}^N b_i^2}{2})\end{aligned}$$

où $p_i = sigmoid(\alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_{12} x_{1i} x_{2i} + bi)$

3 Résultats obtenus et analyse

Vérifions l'allure des chaines de Markov obtenues:

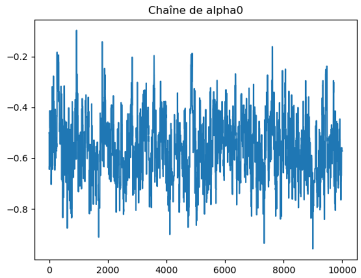

Figure 1: Chaine de Markov de α_0

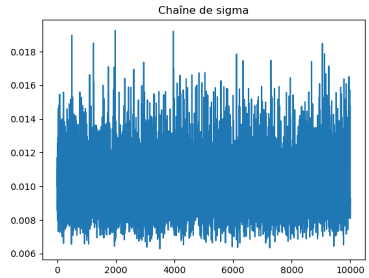

Figure 2: Chaine de Markov de σ

Nous obtenons les résultats suivants:

Résultats					
	α_0	α_1	α_2	α_{12}	σ
Moyenne	-0,558	0,144	1,327	-0,789	0,01
Variance	0,017	0,049	0,032	0,09	$2,7.10^{-6}$

Nous obtenons des résultats similaires à ceux attendus, à part pour α_1 et σ. Interpretons les différents coefficients trouvés afin de voir l'influence de chaque graine sur la probabilité de germer.

En accord avec le modèle choisi, la probabilité de germer est donnée ainsi pour chaque type de graine et chaque variété:

- Graine de type O. aegyptiaco 75, **Haricot** : $p = sigmoid(\alpha_0) = sigmoid(-0,558) = 0,361$
- Graine de type O. aegyptiaco 75, **Concombre** : $p = sigmoid(\alpha_0 + \alpha_2) = sigmoid(-0,558 + 1,327) = sigmoid(0,769) = 0,682$
- Graine de type O. aegyptiaco 73, **Haricot** : $p = sigmoid(\alpha_0+\alpha_1) = sigmoid(-0,558+0,144) = sigmoid(-0,414) = 0,399$
- Graine de type O. aegyptiaco 73, **Concombre** : $p = sigmoid(\alpha_0 + \alpha_1 + \alpha_2 + \alpha_{1,2}) = sigmoid(-0,558 + 0,144 + 1,327 - 0,789) = sigmoid(0,124) = 0,532$

Probabilité de germination		
	seed O. aegyptiaco 75	O. aegyptiaco 75
Haricot	0,361	0,399
Concombre	0,682	0,532

On voit ainsi que pour les **concombres**, il est préférable de choisir des graines de type *O. aegyptiaco 75* puisque la probabilité de germer est supérieure. Au contraire, pour les **haricots**, il vaut mieux planter des graines de type *O. aegyptiaco 73* qui présentent une plus grande probabilité de germination. Pour les deux types de graines, les **concombres** germent avec une plus grande probabilité que les **haricots**.