Modèles hiérarchiques bayésiens 2

Contenu du cours

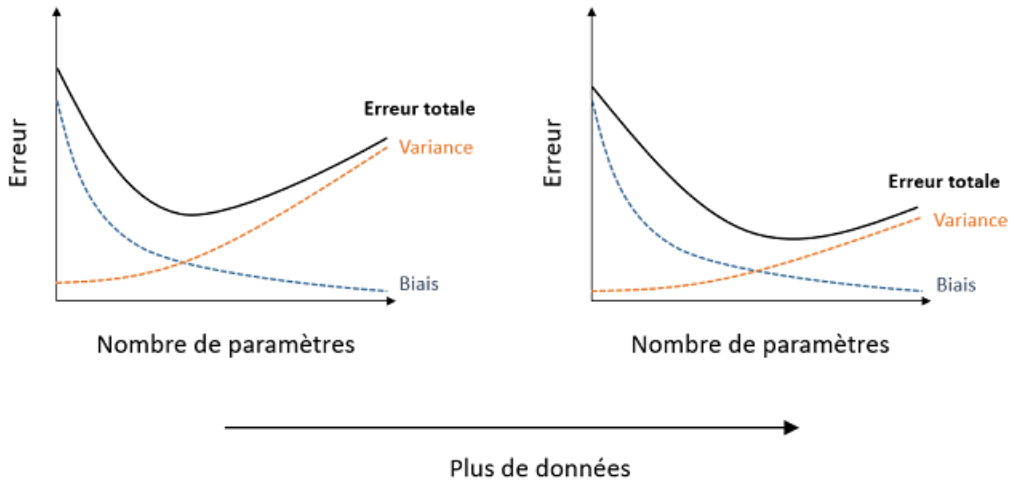
- Révision: Comparaison et sélection de modèles
- Approche bayésienne pour la comparaison de modèles
- Comparaison de modèles avec *loo* et *brms*
- Plus d'exemples de modèles bayésiens en écologie

Comparaison et sélection de modèles

Supposons que nous avons différents modèles statistiques qui visent à expliquer les mêmes données. Les modèles pourraient inclure différents prédicteurs, une distribution différente de la réponse, etc. Comment déterminer quel modèle représente le mieux le phénomène étudié?

La plupart des méthodes de comparaison de modèles cherchent à optimiser la capacité du modèle à **prédire de nouvelles observations** du phénomène. Autrement dit, il ne suffit pas d'évaluer si le modèle s'approche des données utilisées pour son ajustement. Un modèle plus complexe, avec davantage de paramètres ajustables, va toujours être plus près de ces données.

De façon générale, un modèle trop simple comporte une grande erreur systématique (biais ou sous-ajustement), car il omet des effets importants sur la variable réponse; un modèle trop complexe comporte une grande erreur aléatoire (variance ou surajustement), car il tend à représenter des associations "accidentelles" d'un échantillon particulier qui ne se généralisent pas à la population. Le compromis idéal entre ces deux types d'erreur, qui minimise l'erreur totale, dépend de la quantité de données, car un grand échantillon diminue la variance associée à l'estimation de nombreux paramètres dans un modèle complexe.



Avec des grands jeux de données, il est possible de mettre de côté une partie des données (souvent ~20 à 30%) pour créer un ensemble de validation, tandis que le reste des données forment l'ensemble d'entraînement (*training set*). Dans ce cas, chacun des modèles candidats est ajusté à partir des données d'entraînement et la performance prédictive des modèles ajustés est évaluée sur l'ensemble de validation.

Validation croisée

La mise de côté d'une partie des données pour la validation n'est pas pratique si la taille de l'échantillon est modeste. Avec relativement peu de données, chaque point est important pour estimer précisément les paramètres du modèle; aussi, plus l'ensemble de validation est petit, plus il a des chances d'être non-représentatif de la population.

La validation croisée (*cross-validation*) offre une façon d'évaluer la performance prédictive sur des nouvelles observations sans avoir à mettre de côté un ensemble de validation. Cette méthode consiste à diviser aléatoirement les observations en groupes et mesurer la qualité de la prédiction des observations d'un groupe selon un modèle ajusté au reste des observations.

Par exemple, si chaque groupe ne comporte qu'une seule observation (*leave-one-out cross-validation*), nous pouvons évaluer la prédiction de chaque valeur de la réponse y_i pour un modèle ajusté sans l'observation i. Cependant, cette méthode requiert de réajuster le modèle n fois, où n est le nombre d'observations.

Si le nombre d'observations est grand, il peut être plus pratique de diviser les observations en k groupes (*k-fold cross-validation*), par exemple $k = 10$, et d'ajuster chaque modèle à évaluer k fois en laissant une fraction $1/k$ des observations de côté.

Critère d'information d'Akaike

Puisque les méthodes de validation croisée sont coûteuses en terme de calcul, il est utile de pouvoir approximer l'erreur de prédiction qui serait obtenue en validation croisée sans avoir à réajuster le modèle plusieurs fois.

Pour les modèles ajustés par la méthode du maximum de vraisemblance, le critère d'information d'Akaike (AIC) offre une mesure d'ajustement basée sur la théorie de l'information, qui tend à produire le même résultat que la validation croisée *leave-one-out* si la taille de l'échantillon est assez grand. L'AIC est calculé ainsi:

$$AIC = -2 \log L + 2K$$

où L est la fonction de vraisemblance à son maximum et K est le nombre de paramètres estimés par le modèle. Une petite valeur de l'AIC représente un meilleur pouvoir prédictif du modèle. Le premier terme de l'équation représente l'ajustement aux données observées, tandis que le deuxième terme pénalise les modèles plus complexes.

L'AIC est défini à une constante additive près, donc sa valeur absolue ne donne aucune information. C'est plutôt la différence d'AIC entre les modèles candidats qui est interprétable. Cette différence est définie par rapport à la valeur minimale de l'AIC parmi les modèles comparés: $\Delta AIC = AIC - \min AIC$. Le meilleur modèle a un $\Delta AIC = 0$.

L'expression:

$$e^{-\frac{\Delta AIC}{2}}$$

correspond au rapport de la plausibilité (*evidence ratio*) de chaque modèle vs. celui ayant l'AIC minimal. Par exemple, $\Delta AIC = 2$ correspond à un ratio de ~0.37 (~3 fois moins probable), tandis que $\Delta AIC = 10$ correspond à un ratio de ~0.0067 (~150 fois moins probable).

Prédictions multi-modèles

Avec m modèles candidats, on peut se servir des rapports de plausibilité décrits ci-dessus pour définir le poids d'Akaike w pour chaque modèle:

$$w_i = \frac{e^{\frac{-\Delta AIC_i}{2}}}{\sum_{j=1}^{m} e^{\frac{-\Delta AIC_j}{2}}}$$

Le dénominateur normalise chaque rapport par leur somme, de façon à ce que la somme des poids w_i égale 1.

Si plusieurs modèles sont plausibles et ont un poids d'Akaike non-négligeable, alors il est possible de faire la moyenne de leurs prédictions d'une nouvelle observation de la réponse (cette prédiction est notée $\tilde{y}$), en pondérant la prédiction $\tilde{y_j}$ de chaque modèle candidat par son poids w_j.

$$\tilde{y} = \sum_{j=1}^{m} w_j \tilde{y_j}$$

Les prédictions multi-modèles sont souvent plus précises que celles obtenues en considérant seulement le meilleur modèle, car elles tiennent compte de l'incertitude sur la forme du modèle.

Approche bayésienne pour la comparaison de modèles

Densité prédictive

Pour un modèle estimé par maximum de vraisemblance, les prédictions de nouvelles observations sont obtenues en fixant les paramètres du modèles à leur valeur estimée. La vraisemblance de cette nouvelle observation est donc $p(\tilde{y}|\hat{\theta})$, où $\hat{\theta}$ sont les estimés des paramètres au maximum de vraisemblance.

Dans une approche bayésienne, les prédictions de nouvelles observations sont obtenues en faisant la moyenne des prédictions en fonction de la distribution *a posteriori* de la valeur des paramètres. La densité prédictive de $\tilde{y}$ en fonction du modèle ajusté aux observations y, notée $p(\tilde{y}|y)$, est égale à la moyenne de la vraisemblance $p(\tilde{y}|\theta)$ pour la distribution *a posteriori* conjointe des θ:

$p(\tilde{y}|y) = \int p(\tilde{y}|\theta)p(\theta|y)\mathrm{d}\theta$.

En pratique, si un algorithme de Monte-Carlo génère S vecteurs de paramètres $\theta_{(1)}, ..., \theta_{(S)}$ approximant la distribution *a posteriori*, nous calculons $p(\tilde{y}|y)$ par la moyenne des prédictions de chaque vecteur:

$p(\tilde{y}|y) = \frac{1}{S}\sum_{j=1}^{S} p(\tilde{y}|\theta_{(j)})$.

Comme pour la vraisemblance, il est plus facile de travailler avec le logarithme de la densité prédictive.

Validation croisée

Pour déterminer le modèle qui maximise la densité prédictive de nouvelles observations, telle que définie ci-dessus, nous pouvons utiliser la validation croisée. Cependant, puisque l'ajustement des modèles hiérarchiques bayésiens demande parfois considérablement de temps de calcul, il n'est pas pratique dans ces cas de répéter l'estimation du modèle un grand nombre de fois, en laissant de côté une partie des données. Nous utilisons donc le plus souvent des critères qui approximent la performance prédictive d'une validation croisée.

Si l'AIC approxime bien l'erreur de validation croisée pour les modèles ajustés par maximum de vraisemblance avec un nombre d'observations assez grand, ce critère s'applique mal aux modèles bayésiens. D'une part, le maximum de vraisemblance n'est pas directement produit par l'ajustement du modèle bayésien. De plus, il est difficile de définir un nombre de paramètres ajustables K en raison de la structure hiérarchique et des contraintes imposées par les distributions *a priori* des paramètres, qui font que ces paramètres ne varient pas librement.

Critères de sélection pour les modèles hiérarchiques bayésiens

DIC

Le critère d'information de la déviance (DIC), basé sur l'AIC, a été l'un des premiers critères développés pour la comparaison de modèles bayésiens:

$$DIC = -2 \log p(y|\bar{\theta}) + 2p_D$$

où $\bar{\theta}$ est la moyenne de la distribution *a posteriori* de θ et p_D est le nombre effectif de paramètres, qui peut être calculé de plusieurs façons.

Comme l'AIC, le DIC représente bien la performance prédictive relative de modèles sur de nouvelles données, si la taille de l'échantillon est assez grand. Cependant, il ne s'agit pas d'une prédiction bayésienne car elle se base sur un estimé unique de chaque paramètre (sa valeur moyenne) plutôt que sur la distribution *a posteriori* au complet.

WAIC

Le critère de Watanabe-Akaike (WAIC) est semblable au DIC, mais le premier terme est basé sur la densité prédictive conjointe des observations $y_1, ..., y_n$.

$$WAIC = -2 \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{j=1}^{S} p(y_i|\theta_{(j)}) \right) + 2p_W$$

,

où la pénalité p_W est la somme des variances du logarithme de la densité prédictive à chaque point:

$$p_W = \sum_{i=1}^{n} \mathrm{Var}_j \left(\log p(y_i|\theta_{(j)}) \right)$$

Ici, Var_j désigne la variance de l'expression entre parenthèses sur l'ensemble des itérations j.

Le WAIC d'un modèle *brms* peut être calculé avec la fonction `waic`.

PSIS-LOO

Une méthode développée récemment par Vehtari et al. (2017) consiste à estimer la densité prédictive à chaque point qui serait obtenue par validation croisée *leave-one-out*, c'est-à-dire en prédisant y_i à partir du modèle ajusté aux données excluant i, y_{-i}.

$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})\mathrm{d}\theta$.

La méthode PSIS-LOO (PSIS = *Patero smoothed importance sampling*, LOO = *leave-one-out*) vise à estimer cette quantité sans effectuer la validation croisée. En bref, cette approximation est obtenue en faisant la moyenne des $p(y|\theta_{(j)})$, mais avec une pondération particulière des $\theta_{(j)}$ (échantillonnage préférentiel). Cette pondération est ensuite ajustée pour que les poids extrêmes suivent un modèle théorique (distribution de Pareto).

Cette méthode est implémentée dans le package R *loo* et peut être appelée à partir de la fonction `loo` appliquée au résultat d'un modèle dans *brms*.

Comme nous verrons dans l'exemple plus loin, la méthode PSIS-LOO produit son propre diagnostic. L'ajustement des poids pour chaque valeur y_i est basé sur un paramètre de la distribution de Pareto

k et lorsque $k > 0.7$, l'approximation de $p(y_i|y_{-i})$ est potentiellement instable. Si ce problème se produit pour quelques observations, il est possible de réajuster le modèle en excluant ces observations seulement afin de calculer directement $p(y_i|y_{-i})$.

Le résultat de cette méthode est l'estimé du logarithme de la densité prédictive $elpd_{loo}$, autrement dit la somme de $\log p(y_i|y_{-i})$. Un critère d'information (LOOIC) semblable au DIC et WAIC peut être obtenu en multipliant $elpd_{loo}$ par -2.

Comparaison des méthodes

La méthode PSIS-LOO est un peu plus précise que le WAIC, surtout pour les petits échantillons, mais le WAIC est généralement plus rapide à calculer.

Puisqu'elles sont basés sur la densité prédictive bayésienne plutôt que sur un seul estimé moyen de chaque paramètre, ces deux méthodes (WAIC et PSIS-LOO) sont actuellement préférées au DIC. Cependant, les deux supposent que les observations individuelles y_i soient indépendantes les unes des autres, conditionnellement à la valeur des paramètres. Typiquement, cette supposition n'est pas respectée si le modèle inclut directement une corrélation entre différentes valeurs de la réponse (ex.: corrélation temporelle ou spatiale).

Prédictions multi-modèles

Dans la section précédente, nous avons vu qu'une prédiction multi-modèles pour une nouvelle observation $\tilde{y}$ est calculée par la moyenne pondérée des prédictions des différents modèles.

$$\tilde{y} = \sum_{j=1}^{m} w_j \tilde{y_j}$$

Comme pour l'AIC, nous pouvons définir des poids selon les différences d'IC entre deux modèles et cela pour différents critères bayésiens (ex.: WAIC, LOOIC).

Cependant, il n'est pas toujours optimal de combiner les modèles proportionnellement à leur plausibilité. Par exemple, les deux meilleurs modèles peuvent produire des prédictions redondantes, tandis que le troisième et quatrième meilleur modèle peuvent aider à corriger certaines prédictions moins bonnes du meilleur modèle.

La **superposition de modèles** (*model stacking*) consiste à chercher les poids w_j qui minimisent l'erreur de prédiction multi-modèles donnée par la moyenne pondérée (Yao et al. 2018). Ce calcul peut être fait directement à partir des résultats de la méthode PSIS-LOO, comme nous verrons dans l'exemple de la prochaine section.

Comparaison de modèles avec *loo* et *brms*

Le jeu de données rikz.csv contient des données sur la richesse de la microfaune benthique (*Richness*) pour 45 sites répartis sur 5 plages (*Beach*) aux Pays-Bas, en fonction de la position verticale du site (*NAP*) et d'un indice d'exposition mesuré au niveau de la plage (*Exposure*).

```
rikz <- read.csv("../donnees/rikz.csv")
rikz$Exposure <- as.factor(rikz$Exposure)
head(rikz)
```

```
##   Sample Richness Exposure    NAP Beach
## 1      1       11       10  0.045     1
## 2      2       10       10 -1.036     1
## 3      3       13       10 -1.336     1
```

```
## 4      4       11       10  0.616     1
## 5      5       10       10 -0.684     1
## 6      6        8        8  1.190     2
```

La semaine dernière, nous avions ajusté avec *brms* un modèle de régression de Poisson pour la richesse spécifique en fonction de *NAP* et *Exposure*, avec un effet aléatoire de la plage sur l'ordonnée à l'origine.

```
library(brms)

rikz_prior <- c(set_prior("normal(0, 1)", class = "b"),
                set_prior("normal(2, 1)", class = "Intercept"),
                set_prior("normal(0, 0.5)", class = "sd"))

mod1 <- brm(Richness ~ NAP + Exposure + (1 | Beach), data = rikz,
            family = poisson, prior = rikz_prior,
            control = list(adapt_delta = 0.99))
```

Nous considérons maintenant une autre version du modèle où l'effet du *NAP* varie aussi aléatoirement d'une plage à l'autre.

```
mod2 <- brm(Richness ~ NAP + Exposure + (1 + NAP | Beach), data = rikz,
            family = poisson, prior = rikz_prior,
            control = list(adapt_delta = 0.99))
```

Voici les effets fixes et l'écart-type des effets aléatoires estimés pour les deux modèles. Dans le modèle 2, l'incertitude sur `b_NAP` a augmenté et l'effet aléatoire de la plage sur ce coefficient a un écart-type de 0.34 avec un intervalle de crédibilité de 0.06 à 0.70, comparable à l'effet aléatoire de la plage sur l'ordonnée à l'origine.

```
posterior_summary(mod1, pars = "b|sd")
```

```
##                       Estimate  Est.Error        Q2.5      Q97.5
## b_Intercept          2.3927957 0.26083997  1.80470400  2.8737225
## b_NAP               -0.5035968 0.07314332 -0.64624467 -0.3649724
## b_Exposure10        -0.4822582 0.29030926 -1.02739390  0.1622888
## b_Exposure11        -1.1790398 0.31523957 -1.77229362 -0.4762385
## sd_Beach__Intercept  0.2353781 0.13486006  0.02076051  0.5506889
```

```
posterior_summary(mod2, pars = "b|sd")
```

```
##                       Estimate Est.Error        Q2.5      Q97.5
## b_Intercept          2.3822660 0.3232131  1.67806912  2.9760507
## b_NAP               -0.5820099 0.1540185 -0.90619973 -0.2875833
## b_Exposure10        -0.3875815 0.3665545 -1.07115726  0.4168681
## b_Exposure11        -1.1633889 0.3764942 -1.86588161 -0.3330708
## sd_Beach__Intercept  0.3040575 0.1546702  0.04794410  0.6736681
## sd_Beach__NAP        0.3445586 0.1562193  0.07847283  0.6917369
```

Calcul du LOOIC

La fonction `loo` de *brms* compare différents modèles en fonction du critère estimé avec PSIS-LOO (LOOIC, égal à -2 fois la densité prédictive estimée pour la validtion croisée).

```
loo1 <- loo(mod1, mod2, compare = TRUE)
```

```
## Warning: Found 2 observations with a pareto_k > 0.7 in model 'mod2'. It is
## recommended to set 'reloo = TRUE' in order to calculate the ELPD without
```

```
## the assumption that these observations are negligible. This will refit
## the model 2 times to compute the ELPDs for the problematic observations
## directly.
```

```
loo1
```

```
##              LOOIC    SE
## mod1        211.67 19.28
## mod2        204.27 14.58
## mod1 - mod2   7.40  6.84
```

Le résultat indique que le modèle 1 a un LOOIC plus élevé de 6.42 comparé au modèle 2, ce qui semble une grande différence, sauf que l'erreur-type de cette différence (2e colonne) est de 5.92. Donc il n'est peut-être pas certain que le modèle 2 soit le meilleur.

De plus, R nous avertit que pour 3 observations (1 du modèle 1, 2 du modèle 2), l'estimé PSIS-LOO est instable avec un $k > 0.7$ dans la distribution de Pareto. Cet avertissement signifie que pour ces observations, les poids utilisés pour l'approximation de la densité prédictive de validation croisée ont trop de valeurs extrêmes pour estimer leur variance. Tel que suggéré par le message, nous ré-évaluons le LOOIC avec l'argument `reloo = TRUE`, qui va réajuster le modèle en omettant chacune des observations problématiques, pour calculer la densité prédictive de validation croisée directement.

```
loo_corr <- loo(mod1, mod2, compare = TRUE, reloo = TRUE)
```

```
## No problematic observations found. Returning the original 'loo' object.
```

```
## 2 problematic observation(s) found.
## The model will be refit 2 times.
```

```
##
## Fitting model 1 out of 2 (leaving out observation 10)
```

```
## Start sampling
```

```
##
## Fitting model 2 out of 2 (leaving out observation 22)
```

```
## Start sampling
```

```
loo_corr
```

```
##              LOOIC    SE
## mod1        211.67 19.28
## mod2        206.01 15.18
## mod1 - mod2   5.66  6.40
```

Ici, la valeur des LOOIC a un peu changé par rapport au cas précédent. À titre de comparaison, le WAIC produit une différence plus grande entre les deux modèles.

```
waic(mod1, mod2, compare = TRUE)
```

```
##               WAIC    SE
## mod1        211.21 19.38
## mod2        201.36 13.86
## mod1 - mod2   9.85  7.66
```

Comparaison avec le GLMM

Lorsque nous avions ajusté ces modèles avec des GLMM au cours 5, l'AIC était plus faible pour le modèle 1, avec un effet aléatoire sur l'ordonnée à l'origine seulement. Pourquoi la méthode bayésienne donne-t-elle un

résultat différent?

- D'abord, l'utilisation de distributions *a priori* contraint les valeurs des paramètres de façon à ce qu'un modèle plus complexe présente un surajustement moindre.
- Ensuite, l'AIC et les critères bayésiens sont basés sur des prédictions différentes. Supposons que deux modèles diffèrent par un paramètre θ. L'AIC compare les prédictions lorsque ce paramètre est omis, ce qui implique par exemple $\theta = 0$, avec les prédictions à la valeur estimée du maximum de vraisemblance $\hat{\theta}$. En contrepartie, les prédictions bayésiennes du modèle incluant θ sont une moyenne réalisée à partir de la distribution *a posteriori* de θ, qui va inclure des valeurs proches de 0 si ce cas a une probabilité *a posteriori* non-négligeable.

Pour ces deux raisons, les prédictions du maximum de vraisemblance et de l'approche bayésienne diffèrent sauf dans des cas très particuliers, ex.: un modèle linéaire où la distribution *a priori* a peu d'importance et la distribution *a posteriori* de tous les paramètres est symétrique.

Superposition des modèles

Le résultat de `loo` contient un élément pour chacun des modèles comparés. Chacun de ces éléments contient une matrice `pointwise` qui présente notamment la valeur estimée du log de la densité prédictive $\log p(y_i|y_{-i})$ pour chaque point i (`elpd_loo`) et l'erreur-type de cet estimé (`mcse_elpd_loo`).

```
head(loo1$mod1$pointwise)
```

```
##       elpd_loo mcse_elpd_loo      p_loo    looic
## [1,] -3.076247   0.010779330 0.23796768 6.152494
## [2,] -2.632016   0.011134332 0.17941677 5.264032
## [3,] -2.559708   0.010424598 0.12829771 5.119415
## [4,] -4.644181   0.021457459 0.70355712 9.288362
## [5,] -2.207510   0.003593317 0.02734295 4.415020
## [6,] -2.222511   0.005394127 0.06324080 4.445023
```

Si nous voulions faire combiner les prédictions de ces deux modèles, la fonction `stacking_weights` du package *loo* permet de déterminer les poids pour la superposition optimale des deux modèles. Cette fonction requière une matrice avec une colonne par modèle, correspondant à la colonne `elpd_loo` de la matrice `pointwise` mentionnée ci-dessus.

```
library(loo)
stacking_weights(cbind(loo1$mod1$pointwise[,1], loo1$mod2$pointwise[,1]))
```

```
## Method: stacking
## ------
##        weight
## model1 0.036
## model2 0.964
```

```
stacking_weights(cbind(loo_corr$mod1$pointwise[,1], loo_corr$mod2$pointwise[,1]))
```

```
## Method: stacking
## ------
##        weight
## model1 0.142
## model2 0.858
```

Pour l'estimé du PSIS-LOO avec correction des valeurs problématiques, nous voyons que presque tout le poids est donné au modèle 2, donc il suffit probablement de faire des prédictions avec le modèle plus complexe. Tel que mentionné plus haut, puisque les prédictions sont basées sur la distribution *a posteriori* entière de

`sd_Beach__NAP`, cela inclut des cas où l'écart-type de cet effet aléatoire s'approche de 0 et on s'approche donc du modèle 1.

Références

Vehtari, A., Gelman, A. et Gabry, J. (2017) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing* 27(5), 1413–1432. doi:10.1007/s11222-016-9696-4.

Yao, Y., Vehtari, A., Simpson, D. et Gelman, A. (2018) Using stacking to average Bayesian predictive distributions. Bayesian Analysis 13(3), 917–1007. doi:10.1214/17-BA1091.