Ecole chercheur BioBayes 2019

Evaluation des modèles

David Makowski
INRA
david.makowski@inra.fr

1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

1. Introduction

2. Analyse de sensibilité

3. Vérification de la qualité des prédictions a posteriori

4. Validation croisée

5. Facteur de Bayes

6. Vraisemblance pénalisée et déviance

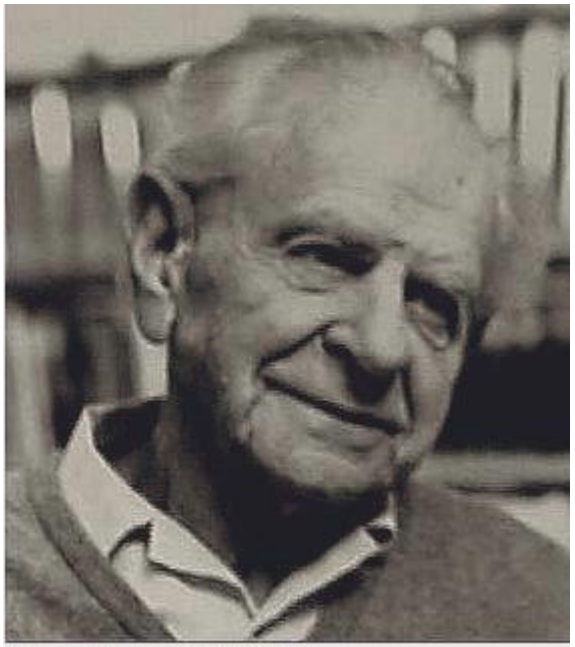
7. Combinaison de modèles

8. Conclusion

Karl Popper (philosophe né en 1902 à Vienne)

- K. Popper a publié plusieurs ouvrages importants, notamment “The logic of scientific discovery” en 1959

- Certaines de ses idées sont utiles pour comprendre les objectifs et limites de l’évaluation des modèles



Sir Karl Popper (1902-1994)

Quelques idées de Popper

- Scientifiques = « problem-solvers »

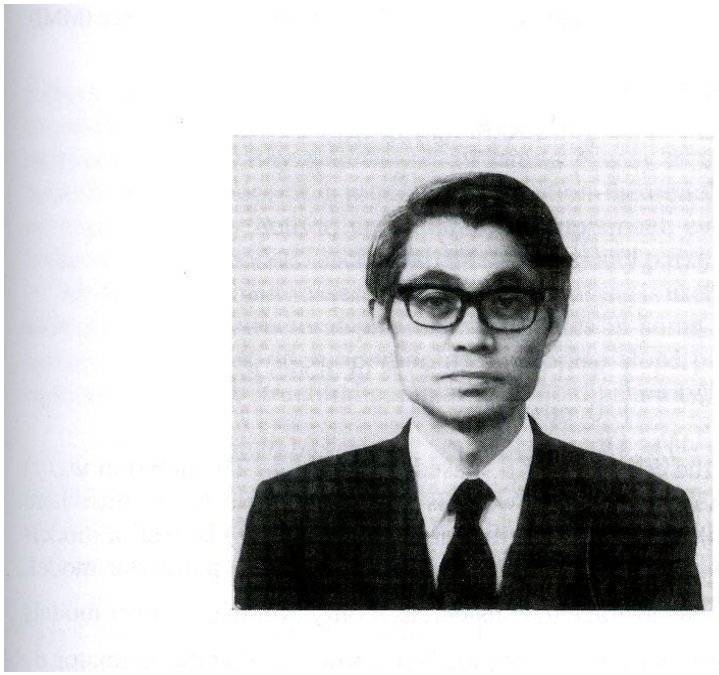
- Nous ne pouvons jamais prouver nos théories scientifiques, nous pouvons simplement les confirmer (provisoirement) ou réfuter (définitivement)

- Les théories non falsifiées restent, les autres sont remplacées par d'autres théories

Hirotugu Akaïke (statisticien né en 1927 au Japon)

- H. Akaïke est célèbre pour avoir proposé un critère d'évaluation pour les modèles : le critère d'Akaïke

- Ce critère est actuellement l'un des plus utilisés pour évaluer les modèles



Le critère d'Akaike (AIC, 1973)

$$AIC = -2 \times \ln(L) + 2 \times p$$

Vraisemblance — Nombre de paramètres

- Estimation de la distance entre deux distributions de probabilité :

  la distribution définie par le modèle et le vrai processus qui génère les données

- Un outil populaire pour le choix des modèles :

  le meilleur modèle est celui avec le plus petit AIC

Georges Box (statisticien né en 1919 en Angleterre)

Il est connu pour avoir défini une procédure pour développer, évaluer et sélectionner des modèles statistiques permettant d'analyser des séries chronologiques.

La procédure de Box-Jenkins pour l'analyse des séries chronologiques (1979)

i. **Définir un modèle**
ii. **Estimer les paramètres**
iii. **Evaluer le modèle**
iv. **Prédire**

"Essentially, all models are wrong, but some are useful"

Quelques idées importantes

- Un modèle **n'est jamais validé**, mais peut être **évalué**

- Il est souvent utile de considérer **plusieurs modèles candidats** pour résoudre un problème donné

- Important d'évaluer les modèles avec des **critères explicites** liés à leurs **utilisations pratiques**

- L'évaluation de modèle est un **processus itératif**

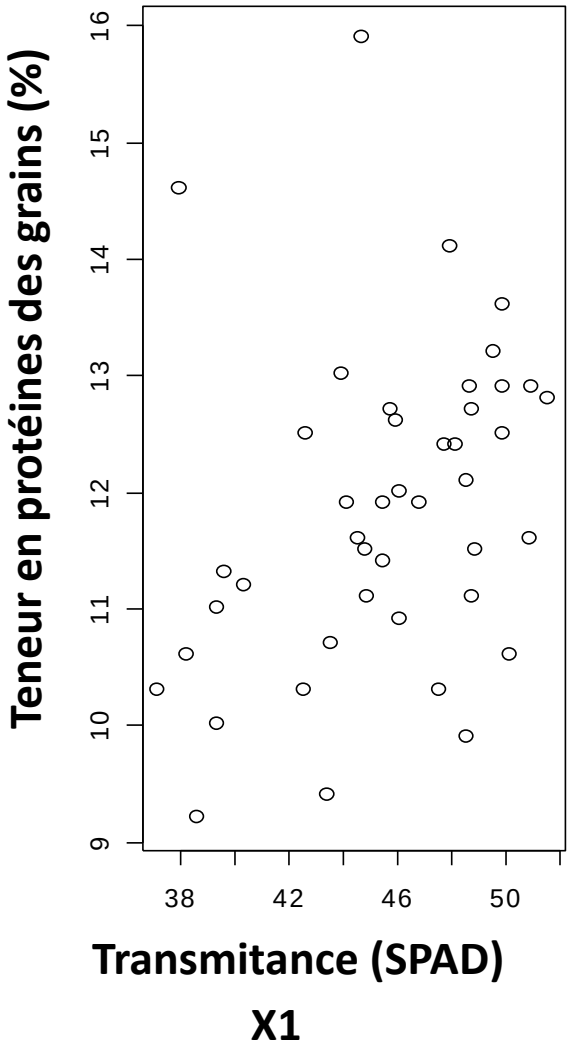
1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

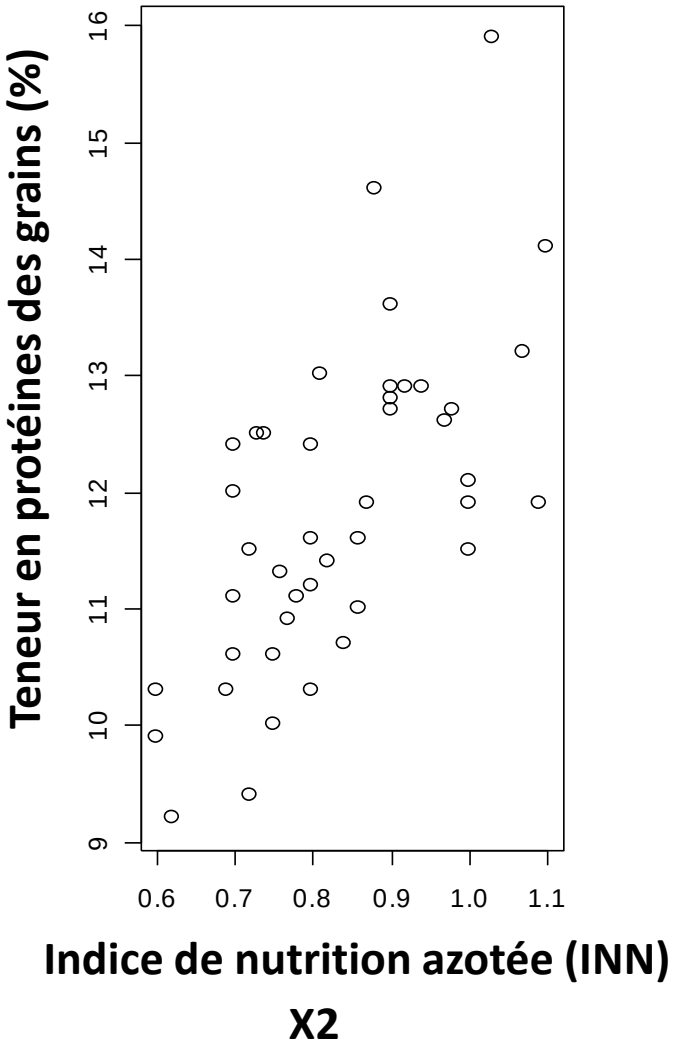
Sensibilité à quoi ?

- Sensibilité à la distribution a priori

- Sensibilité à certaines caractéristiques de la fonction de vraisemblance

- Sensibilité à certaines données





Barbottin et al. (2008)

Vraisemblance

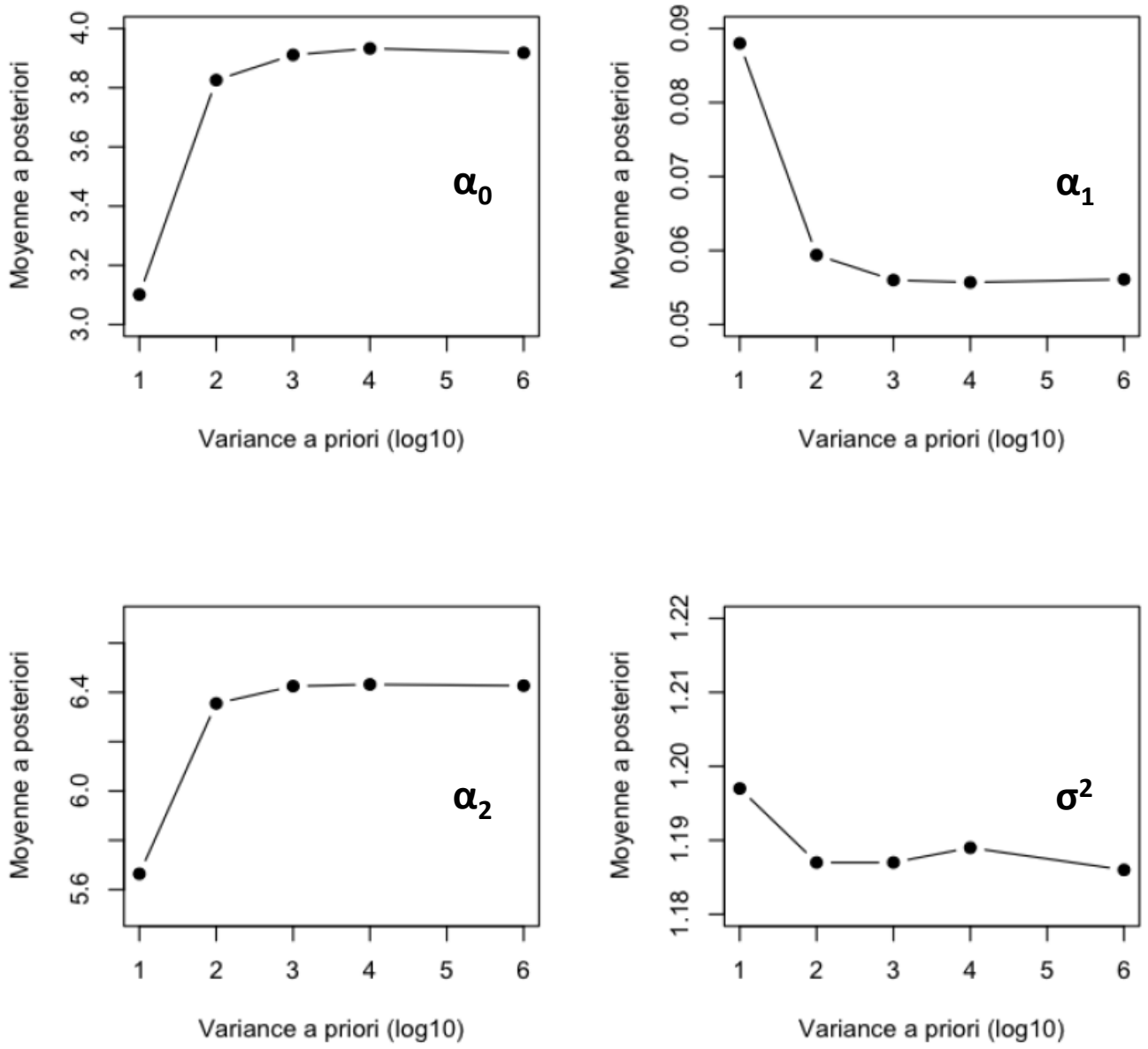
$$Y = a_0 + a_1 X_1 + a_2 X_2 + e$$

$$e \sim N(0, s^2)$$

A priori

$$a_0, a_1, a_2 \sim N(0, 10^4)$$

$$s^2 \sim Unif(0, 1000)$$



Moyennes a posteriori de α_0 (A), α_1 (B), α_2 (C) et σ^2 (D) obtenues pour cinq variances a priori différentes (10, 10^2, 10^3, 10^4, 10^6) à partir d'une chaîne MCMC de 30000 valeurs.

1. Introduction

2. Analyse de sensibilité

3. Vérification de la qualité des prédictions a posteriori

4. Validation croisée

5. Facteur de Bayes

6. Vraisemblance pénalisée et déviance

7. Combinaison de modèles

8. Conclusion

Principe

• Vérifier que le modèle s'ajuste de manière satisfaisante aux données

• Comparer une quantité $Q(y)$ à des valeurs simulées de cette quantité, notées $Q(y_s)$
où *y* et y_s sont respectivement le jeu de données utilisé pour construire le modèle et le jeu de données simulées avec ce modèle.

• On calcule la probabilité que $Q(y_s) > Q(y)$ ou $Q(y_s) < Q(y)$, c'est-à-dire la probabilité que le modèle soit plus extrême que les observations.

• Exemples

$$Q(y) = y_{\min} \qquad Q(y) = y_{\max} \qquad Q(y_i, q) = \frac{[y_i - E(Y_i \mid q)]^2}{\mathrm{var}(Y_i \mid q)}$$

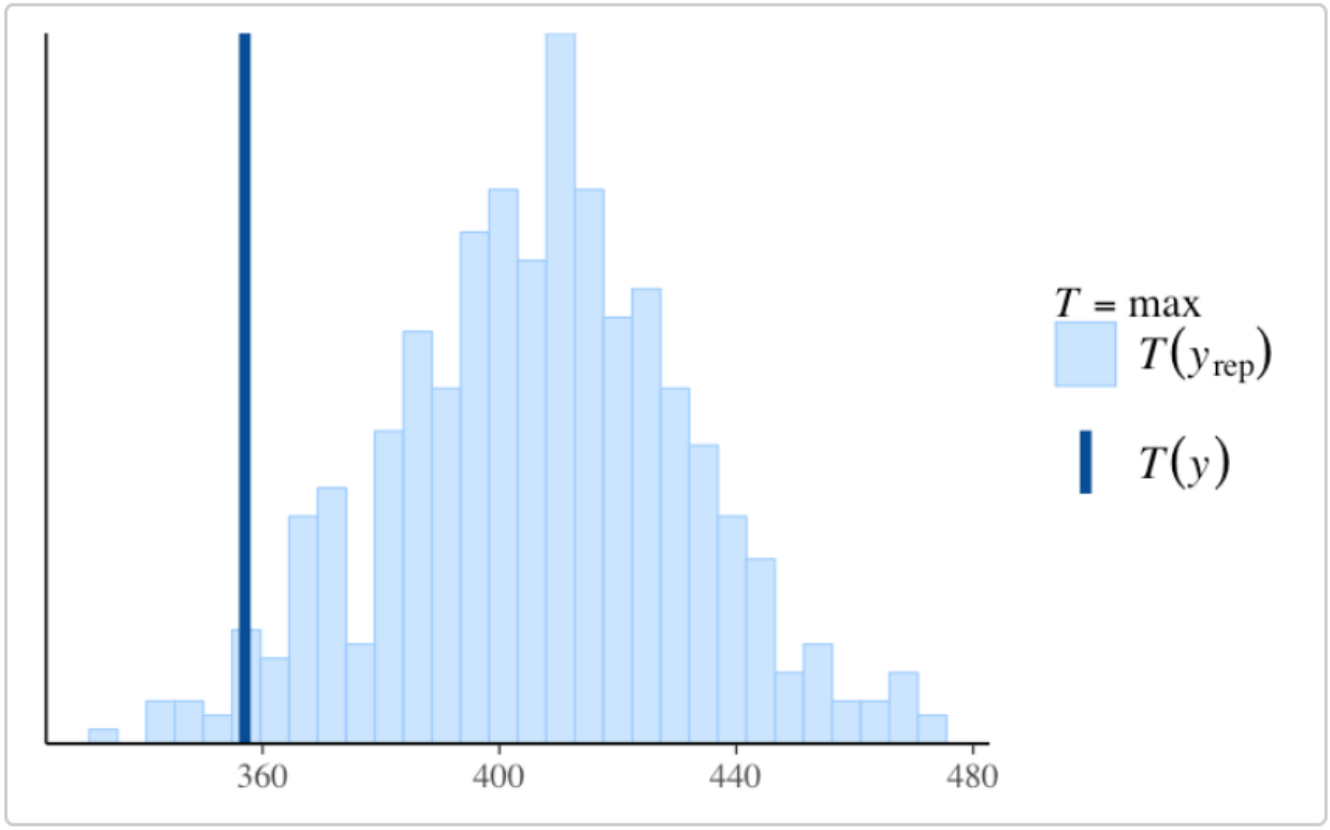
$$P[Q(y_s) < Q(y)] \qquad P[Q(y_s) > Q(y)] \qquad P[Q(y_{si}, q) > Q(y_i, q)]$$

```
> summary(roaches$y)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00    0.00    3.00   25.65   24.00  357.00
```

$$Y \sim Poisson(\lambda)$$

$$\ln(\lambda) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3$$

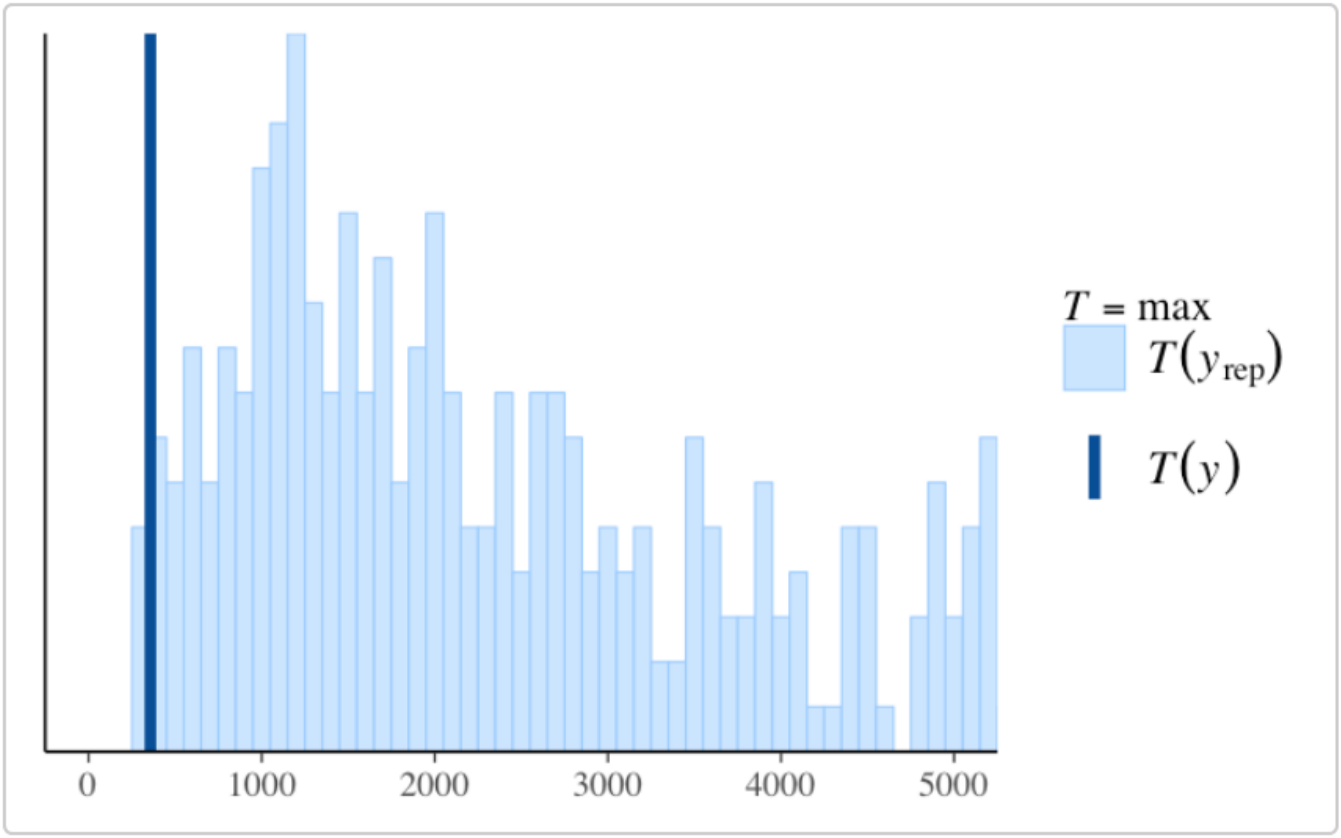
https://cran.r-project.org/web/packages/bayesplot/vignettes/graphical-ppcs.html



Max mumber of roaches in an apartment

Model Poisson

https://cran.r-project.org/web/packages/bayesplot/vignettes/graphical-ppcs.html



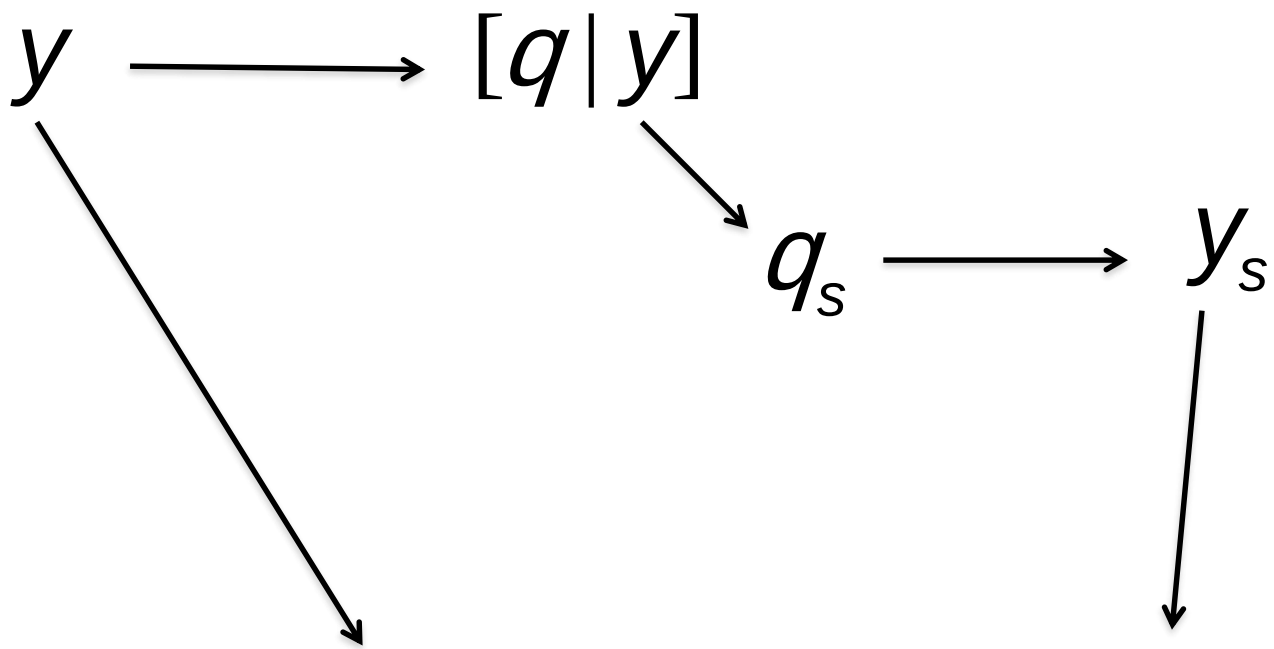
Max mumber of roaches in an apartment

Model Negative binomial

$$[q \mid y] \longrightarrow q_s = \left(a_{0s}, a_{1s}, a_{2s}, s_s\right)'$$

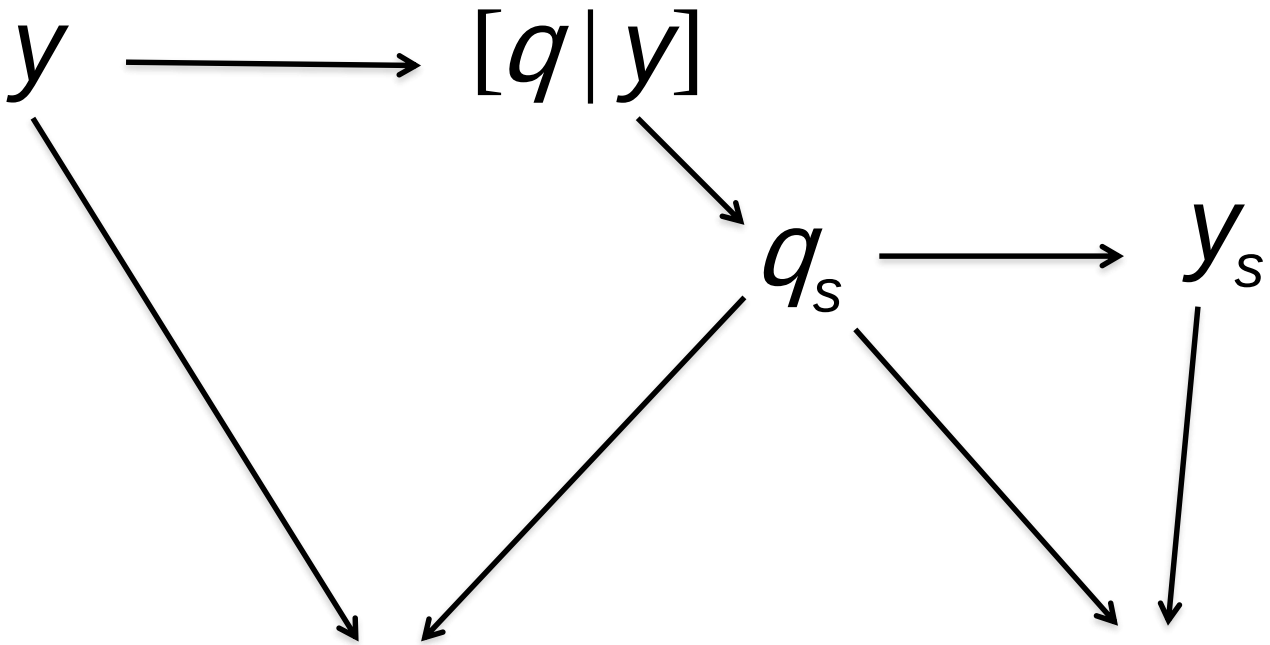
$$y_{si} = a_{0s} + a_{1s}x_{1i} + a_{2s}x_{2i} + e_{si}$$

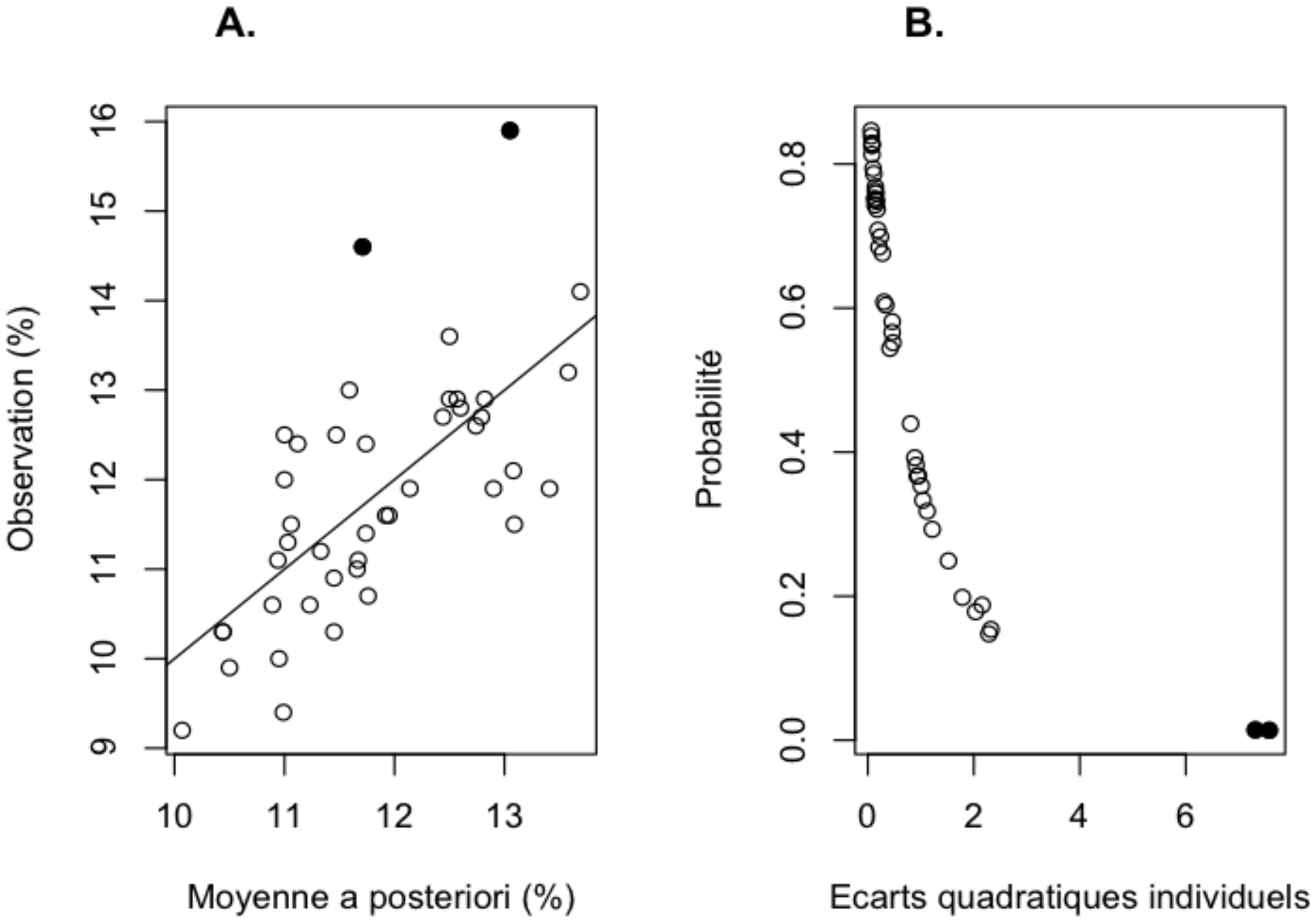
$$e_{si} \sim N(0, s_s^2)$$



$$Q(y) = y_{\max} = \max\{y\}$$

$$Q(y_s) = \max\{y_s\}$$


y
$[q \mid y]$
q_s
y_s
$Q(y_i, q_s) = \frac{[y_i - E(Y_i \mid q_s)]^2}{\mathrm{var}(Y_i \mid q_s)}$
$Q(y_{si}, q_s) = \frac{[y_{si} - E(Y_i \mid q_s)]^2}{\mathrm{var}(Y_i \mid q_s)}$


A.
B.
Observation (%)
Moyenne a posteriori (%)
Probabilité
Ecarts quadratiques individuels
9
10
11
12
13
14
15
16
0.0
0.2
0.4
0.6
0.8
0
2
4
6

Limites

- Pas de seuil de probabilité objectif

- Les mêmes données sont utilisées pour estimer les paramètres et prédire

- Pas de vraies prédictions

- Vision trop optimiste

1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

Principe

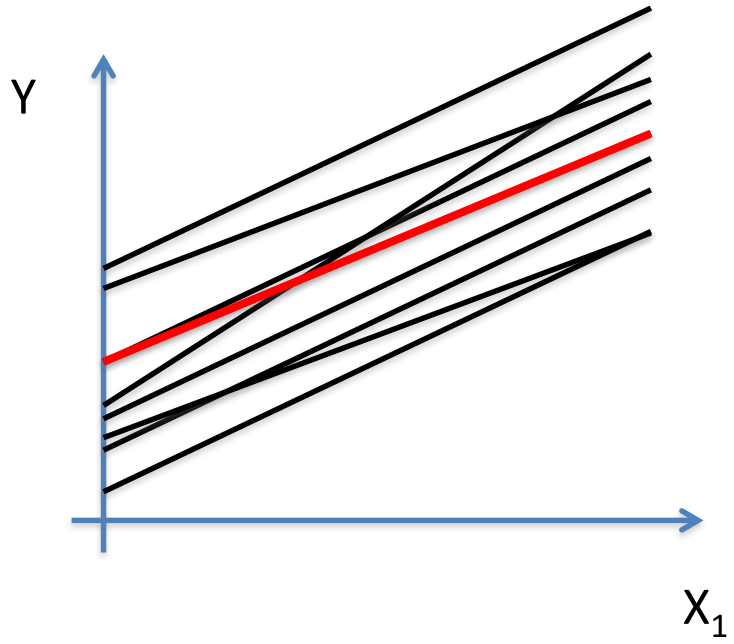
- Méthode souvent utilisée pour évaluer la précision des prédictions d'un modèle

- Evaluation des prédictions avec des données qui n'ont pas été utilisées pour développer le modèle
  - ✓ Le jeu de données est découpé en sous-groupes.
  - ✓ Les données de chaque sous-groupe sont prédites en développant le modèle avec les données non incluses dans ce sous-groupe

- Approche pas spécifiquement bayésienne

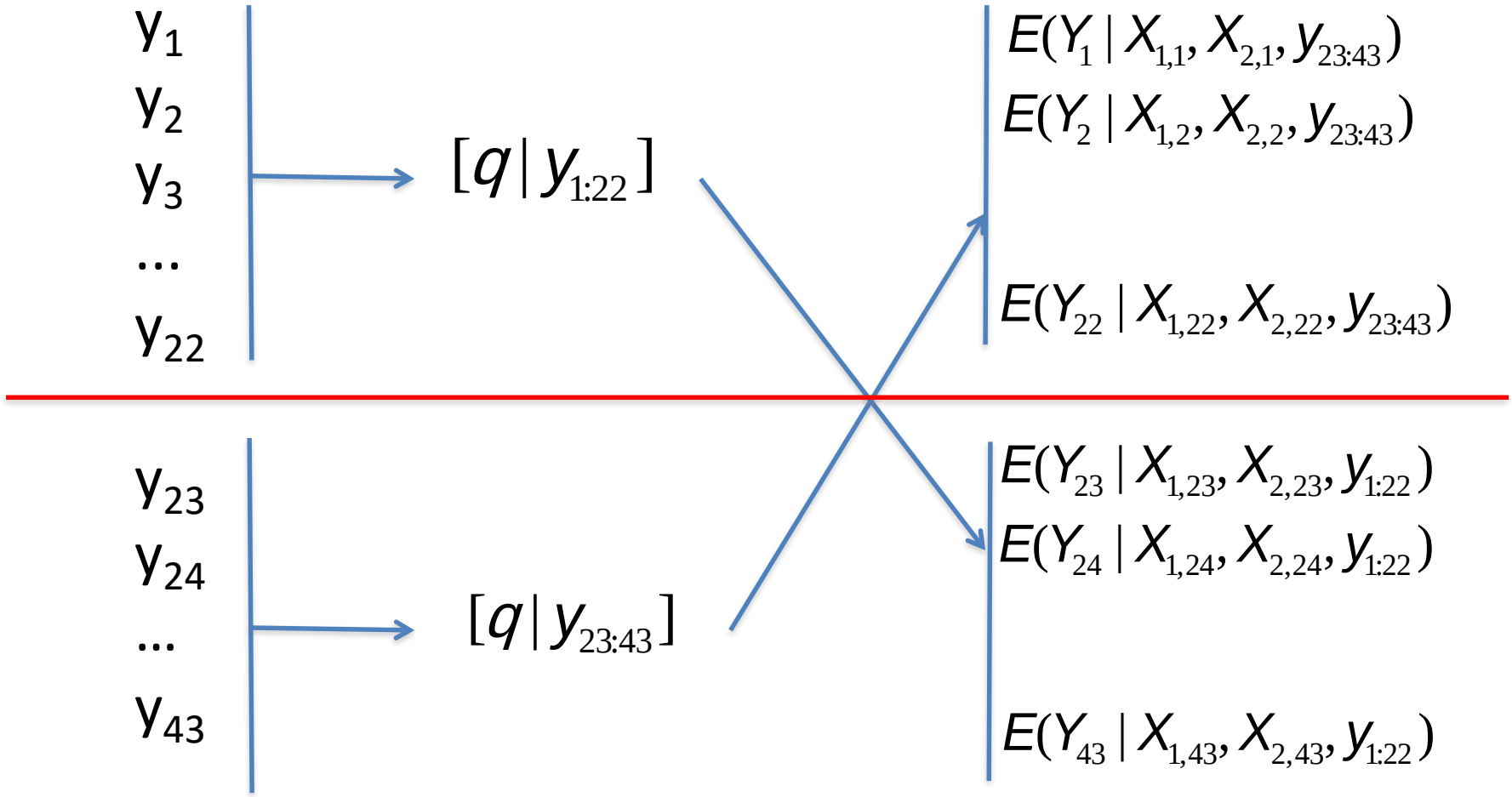
- Etat d'esprit « data science »

Prédire une teneur en protéines

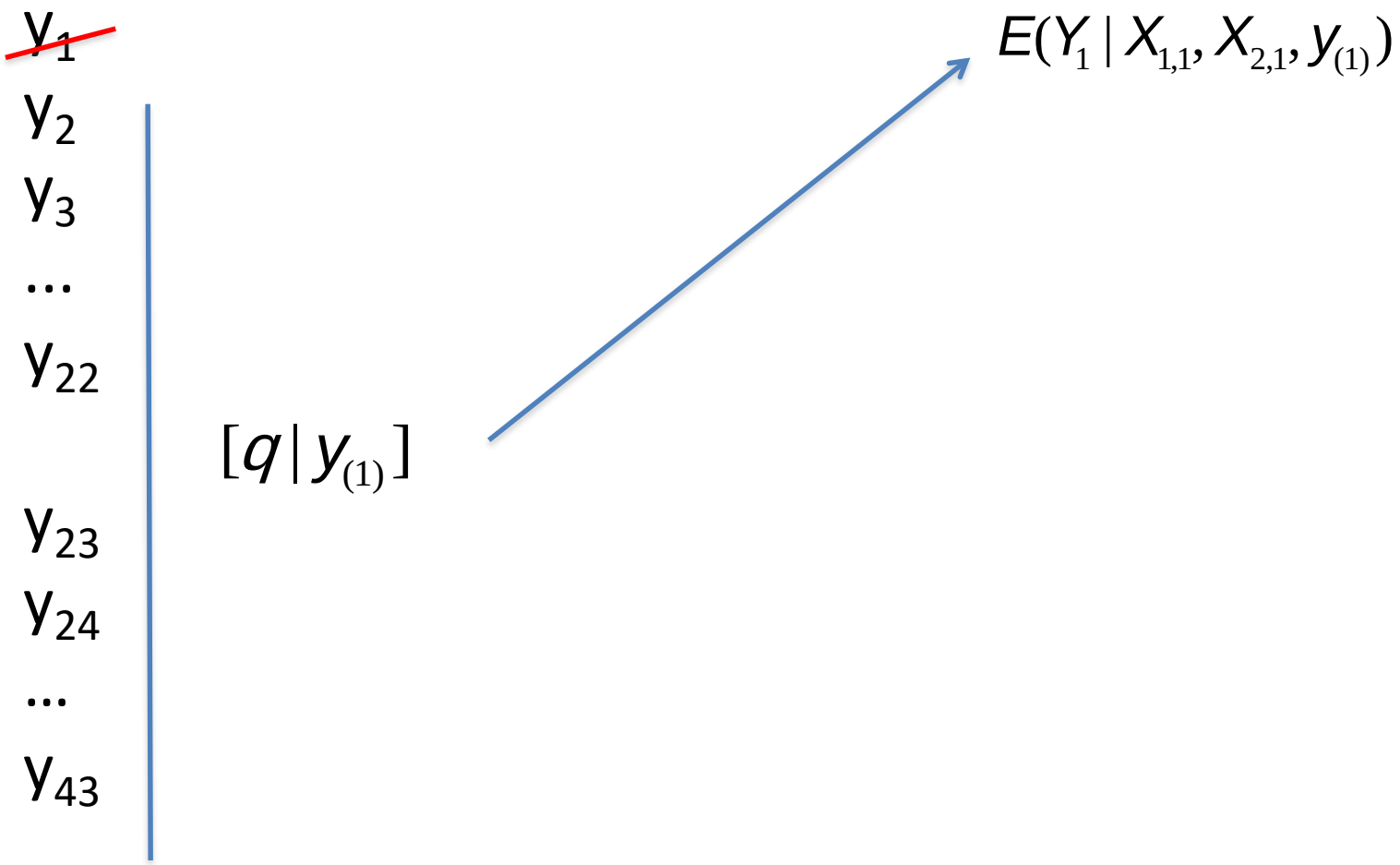


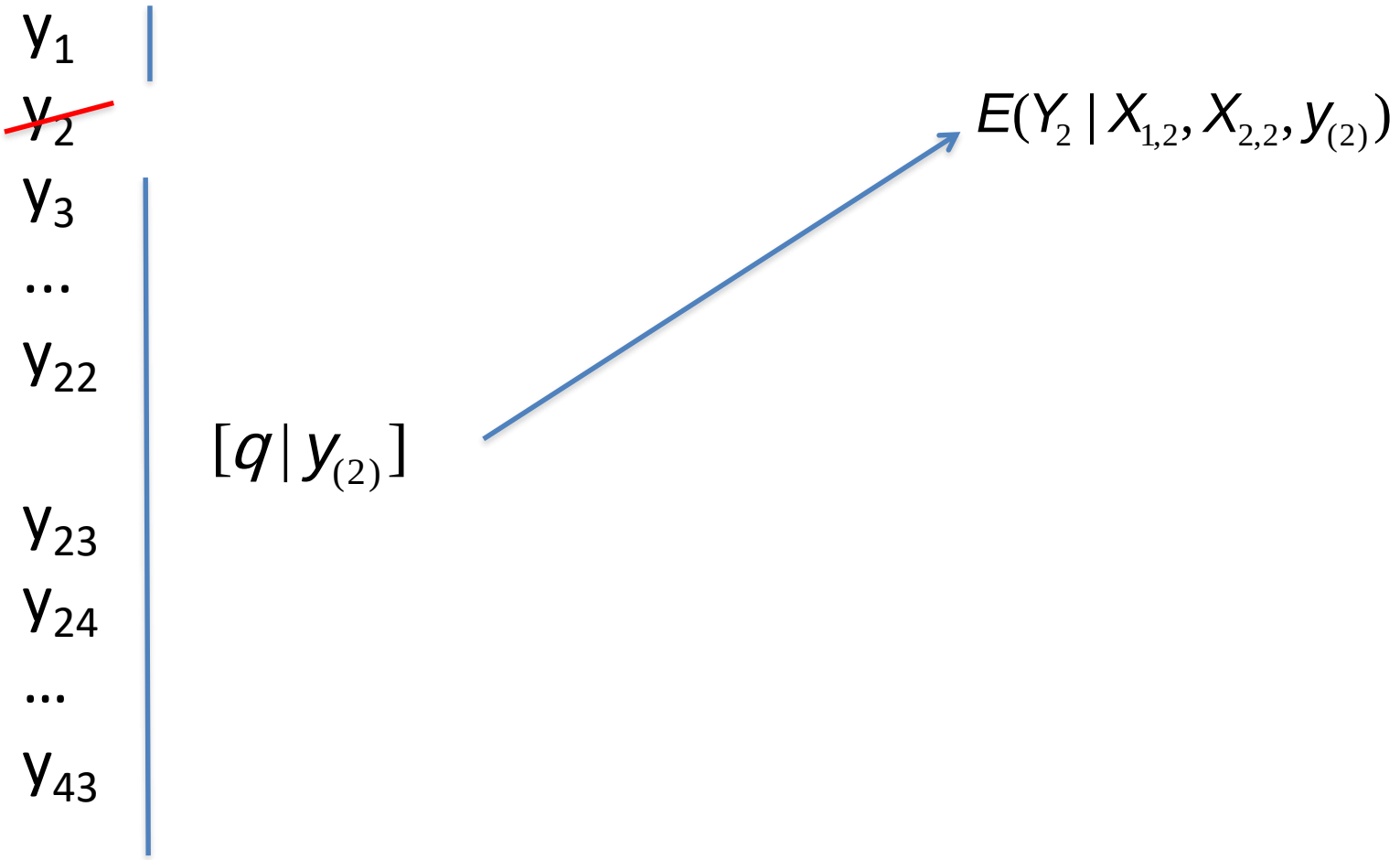
Prédicteur:

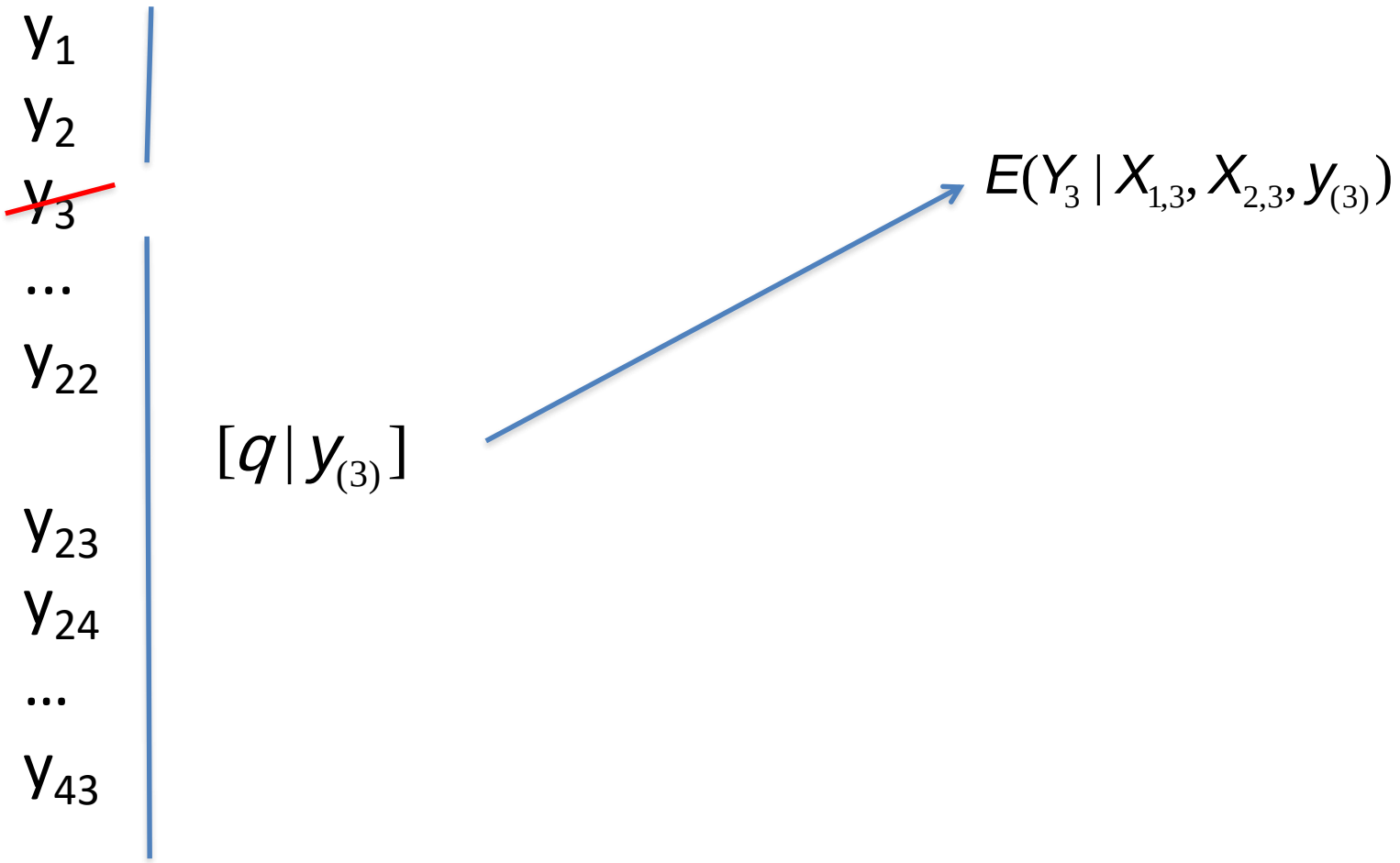
$$P(X_1,X_2) = E\left(Y \middle| X_1, X_2, y_{1:N}\right)$$


y_1
y_2
y_3
...
y_{22}
$[q \mid y_{1:22}]$
y_{23}
y_{24}
...
y_{43}
$[q \mid y_{23:43}]$
$E(Y_1 \mid X_{1,1}, X_{2,1}, y_{23:43})$
$E(Y_2 \mid X_{1,2}, X_{2,2}, y_{23:43})$
$E(Y_{22} \mid X_{1,22}, X_{2,22}, y_{23:43})$
$E(Y_{23} \mid X_{1,23}, X_{2,23}, y_{1:22})$
$E(Y_{24} \mid X_{1,24}, X_{2,24}, y_{1:22})$
$E(Y_{43} \mid X_{1,43}, X_{2,43}, y_{1:22})$

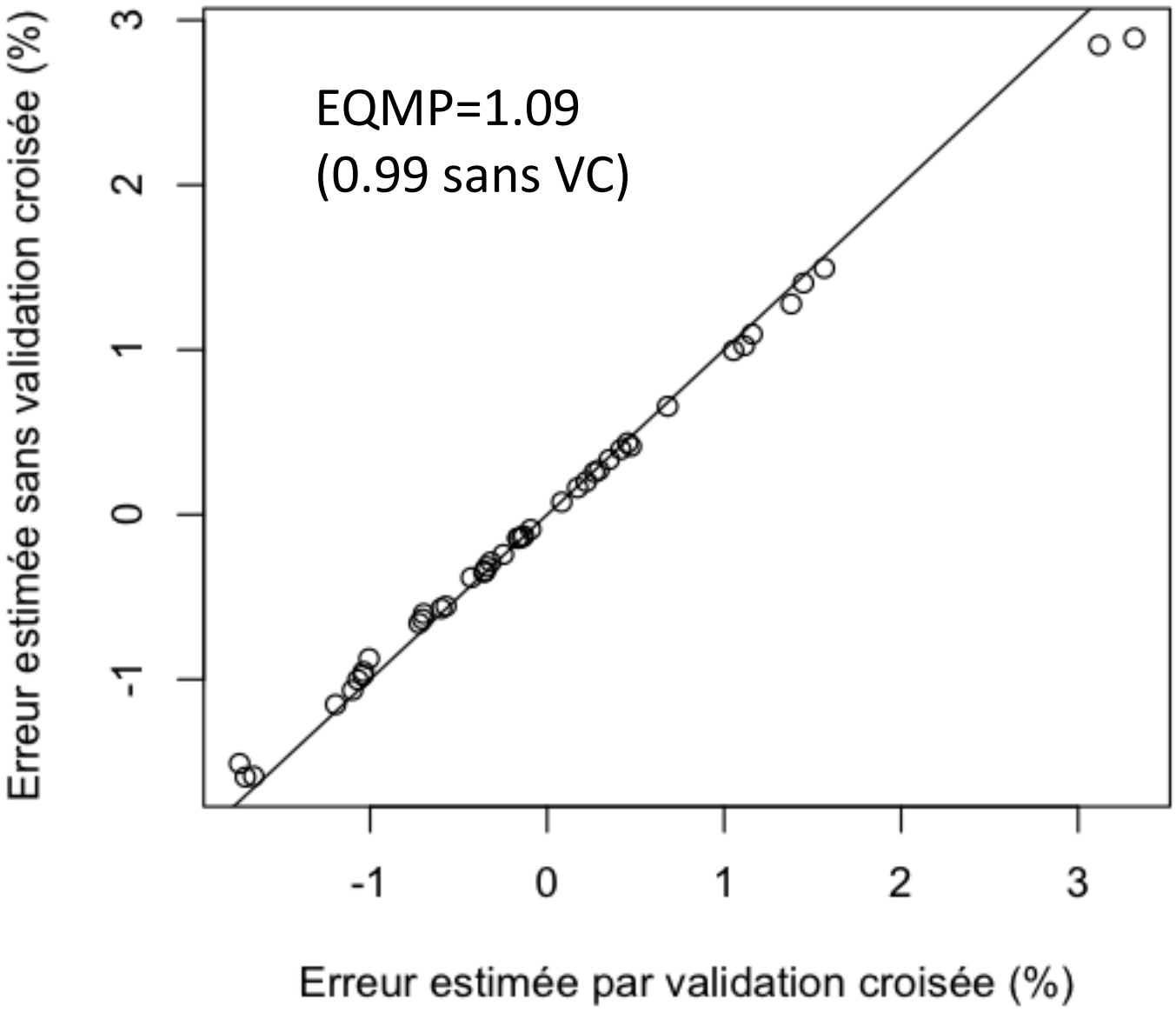
$$E(Y|X_1, X_2, y_{1:i-1}, y_{i+1:N}) = E\left(Y|X_1, X_2, y_{(i)}\right)$$


y_1
y_2
y_3
...
y_{22}
y_{23}
y_{24}
...
y_{43}
$[q \mid y_{(1)}]$
$E(Y_1 \mid X_{1,1}, X_{2,1}, y_{(1)})$


y_1
y_2
y_3
...
y_{22}
y_{23}
y_{24}
...
y_{43}
$[q \mid y_{(2)}]$
$E(Y_2 \mid X_{1,2}, X_{2,2}, y_{(2)})$


y_1
y_2
y_3
...
y_{22}
y_{23}
y_{24}
...
y_{43}
$[q \mid y_{(3)}]$
$E(Y_3 \mid X_{1,3}, X_{2,3}, y_{(3)})$

$$EQMP = \frac{1}{43} \sum_{i=1}^{43} [y_i - E(Y_i \mid X_{1i}, X_{2i}, y_{(i)})]^2$$


EQMP=1.09
(0.99 sans VC)
Erreur estimée sans validation croisée (%)
Erreur estimée par validation croisée (%)
-1
0
1
2
3

1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

Principe

- Permet de comparer deux modèles M_1 et M_2

- Rapport de probabilité des observations pour chacun des modèles

$$B_{12} = \frac{[\mathbf{y}|M_1]}{[\mathbf{y}|M_2]} = \frac{[M_1|\mathbf{y}]}{[M_2|\mathbf{y}]} / \frac{[M_1]}{[M_2]}$$

$$[\mathbf{y}|M_1] = \int_\Theta [\mathbf{y}|\theta_1][\theta_1] d\theta_1$$
$$[\mathbf{y}|M_2] = \int_\Theta [\mathbf{y}|\theta_2][\theta_2] d\theta_2$$

- Plus ce rapport est grand, plus les données sont en faveur du modèle 1

- Difficile à calculer en pratique mais des approximations existent

Approximation par Importance Sampling

$$m_{SI}(\mathbf{y}) = \sum_{t=1}^{T} [\mathbf{y}|\theta_t] \frac{[\theta_t]}{g(\theta_t)} / \sum_{t=1}^{T} \frac{[\theta_t]}{g(\theta)}$$

$$m_{SI}(\mathrm{y}) = \frac{1}{T} \sum_{t=1}^{T} [\mathrm{y}|\theta_t]$$

1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

$$L\left(\hat{\theta}_{ML}\right) = \left[y \mid \hat{\theta}_{ML} \right]$$

$$AIC = -2 \times \ln\left[L\left(\hat{\theta}_{ML}\right) \right] + 2 \times p$$

$$BIC = -2 \times \ln\left[L\left(\hat{\theta}_{ML}\right) \right] + p \times \log(n)$$

$$DIC = E\left\{ -2 \ln\left[L(\theta) \right] \right\} + p_D$$

Ces critères dépendent de la qualité d'ajustement du modèle aux données et de la complexité du modèle

Ils ont différentes origines et différentes significations

Akaïke Information Criterion

- L'AIC est dérivé de la distance Kullback-Leibler
- Il mesure la perte d'information qui résulte de l'utilisation d'un modèle pour estimer le processus à l'origine des données
- Son calcul est basé sur l'estimation des paramètres par maximum de vraisemblance

$$AIC = -2 \times \ln\left[L\left(\hat{\theta}_{ML}\right)\right] + 2 \times p$$

Bayesian Information Criterion

(Schwarz's criterion)

- Le BIC est dérivé d'une approximation du facteur de Bayes
- Son calcul est basé sur une estimation des paramètres par maximum de vraisemblance

$$BIC = -2 \times \ln\left(L(\hat{q}_{ML})\right) + p \times \ln(n)$$

$$\ln(B_{21}) \approx \ln\left(\frac{L_1(\hat{q}_{ML.1})}{L_2(\hat{q}_{ML.2})}\right) + \frac{(p_2 - p_1)}{2}\ln(n)$$

$$\approx \ln\left(L_1(\hat{q}_{ML.1})\right) - \frac{p_1}{2}\ln(n) - \left[\ln\left(L_2(\hat{q}_{ML.2})\right) - \frac{p_2}{2}\ln(n)\right]$$

$$\approx 0.5BIC_2 - 0.5BIC_1$$

Deviance Information Criterion

- Calculé à partir des simulations MCMC
- Pas besoin de l'estimateur du maximum de vraisemblance
- Calcul intégré à WinBUGS

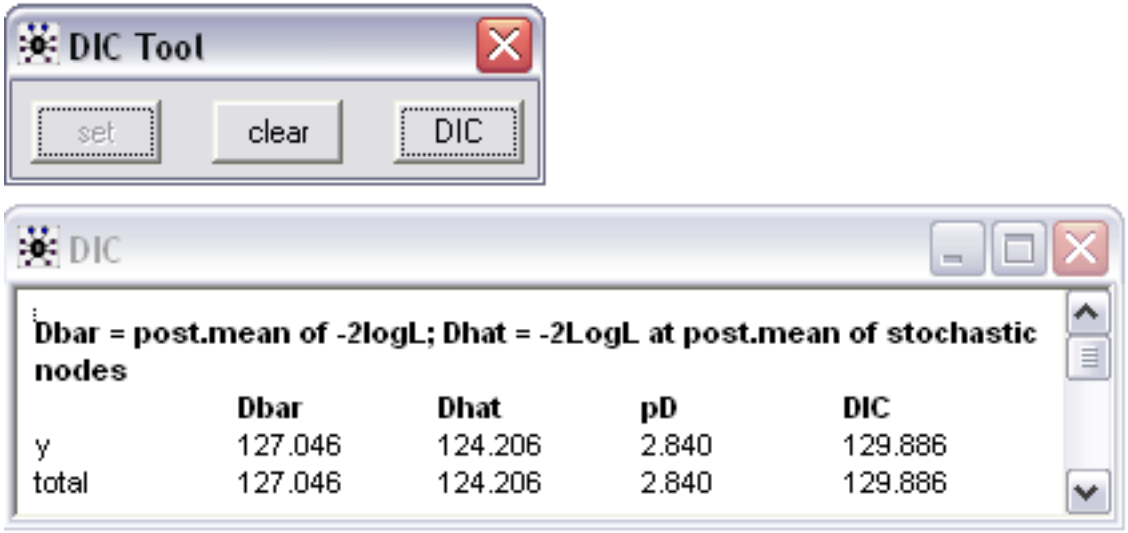
$$Dev(\theta) = -2\ln[L(\theta)]$$

$$DIC = E_\theta[Dev(\theta) \mid y] + \{E_\theta[Dev(\theta) \mid y] - Dev[E_\theta(\theta \mid y)]\}$$

$$DIC = E_\theta[Dev(\theta) \mid y] + p_D$$

Spiegelhalter et al. (2002)

Calcul du DIC avec WinBUGS



Exemple

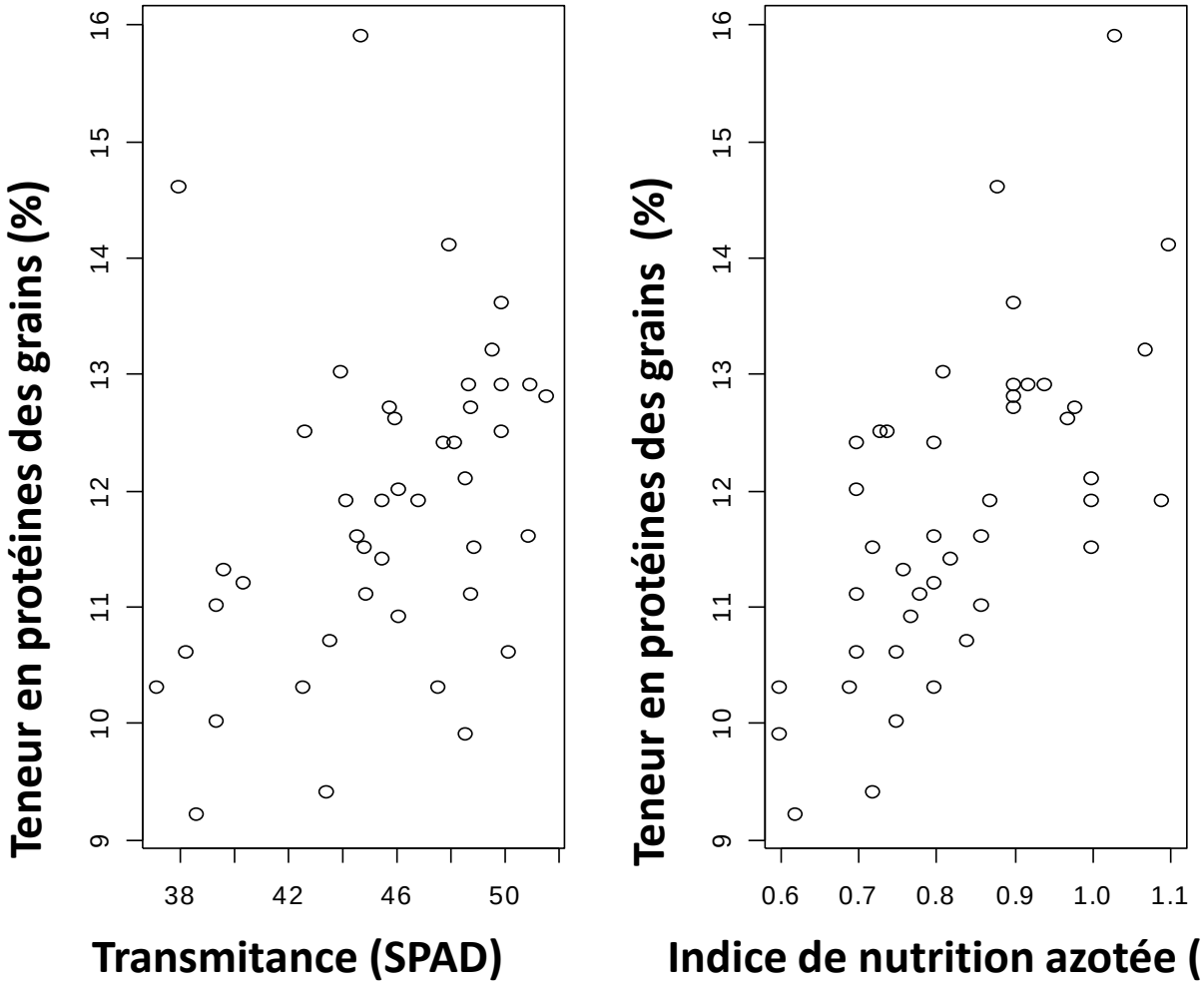
Evaluation de trois modèles

$$M_1 : Y = a_0 + a_1 X_1 + e$$ seulement SPAD

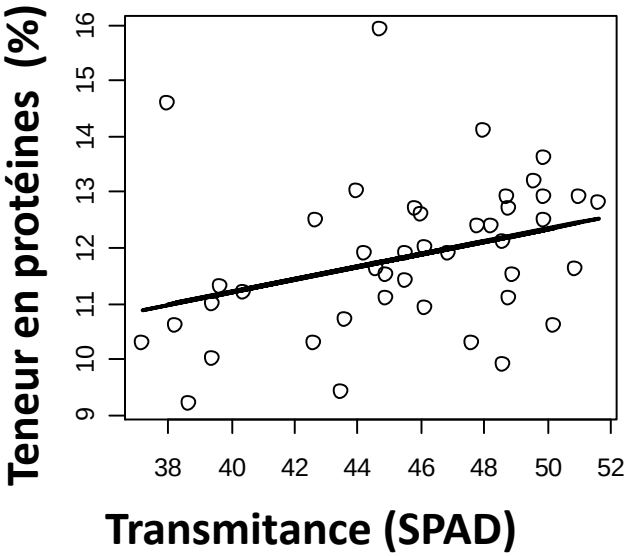
$$M_2 : Y = a_0 + a_2 X_2 + e$$ seulement INN

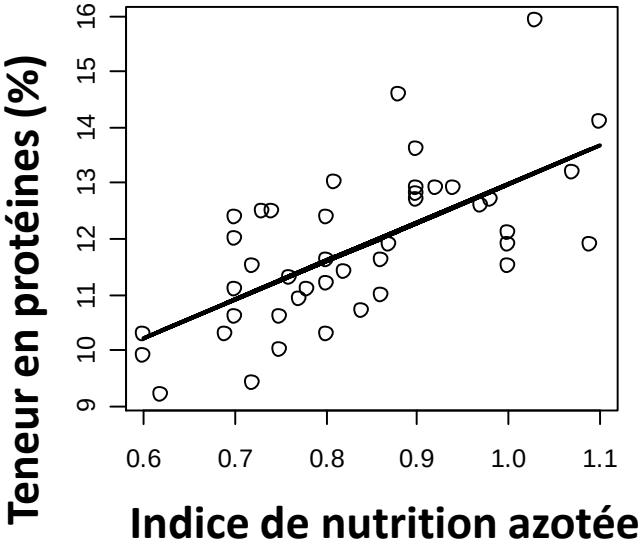
$$M_3 : Y = a_0 + a_1 X_1 + a_2 X_2 + e$$ SPAD et INN

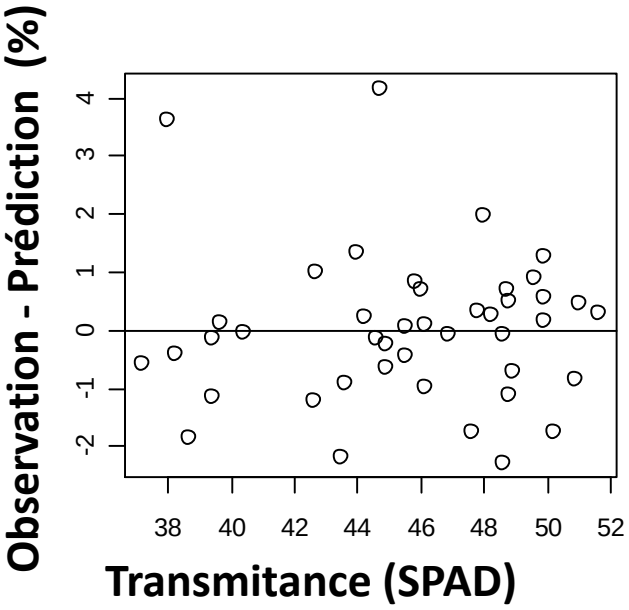
Teneur en protéines des grains de blé vs. Transmitance et INN pour 43 sites-années en France

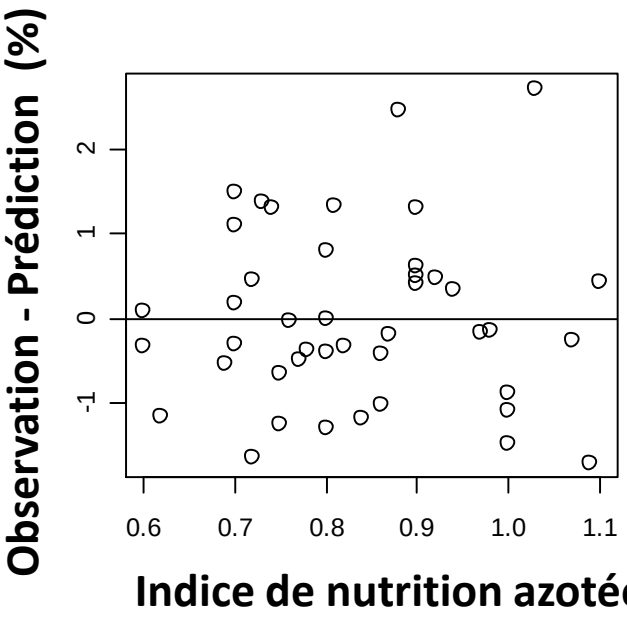


Barbottin et al. (2008)


Teneur en protéines (%)
9
10
11
12
13
14
15
16
38
40
42
44
46
48
50
52
Transmitance (SPAD)


Teneur en protéines (%)
9
10
11
12
13
14
15
16
0.6
0.7
0.8
0.9
1.0
1.1
Indice de nutrition azotée (INN)


Observation - Prédiction (%)
-2
-1
0
1
2
3
4
38
40
42
44
46
48
50
52
Transmitance (SPAD)


Observation - Prédiction (%)
-1
0
1
2
0.6
0.7
0.8
0.9
1.0
1.1
Indice de nutrition azotée (INN)

AIC, BIC et DIC des trois modèles

DIC calculé avec 20000 itérations de WinBUGS

AIC, BIC et R^2 calculés avec des valeurs de paramètres estimées avec glm de R

Modèle	*AIC*	*BIC*	*DIC*	R^2
M_1	149.22	152.74	149.41	0.11
M_2	129.74	133.26	129.93	0.43
M_3	129.82	135.10	130.11	0.46

Watanabe AIC (WAIC)

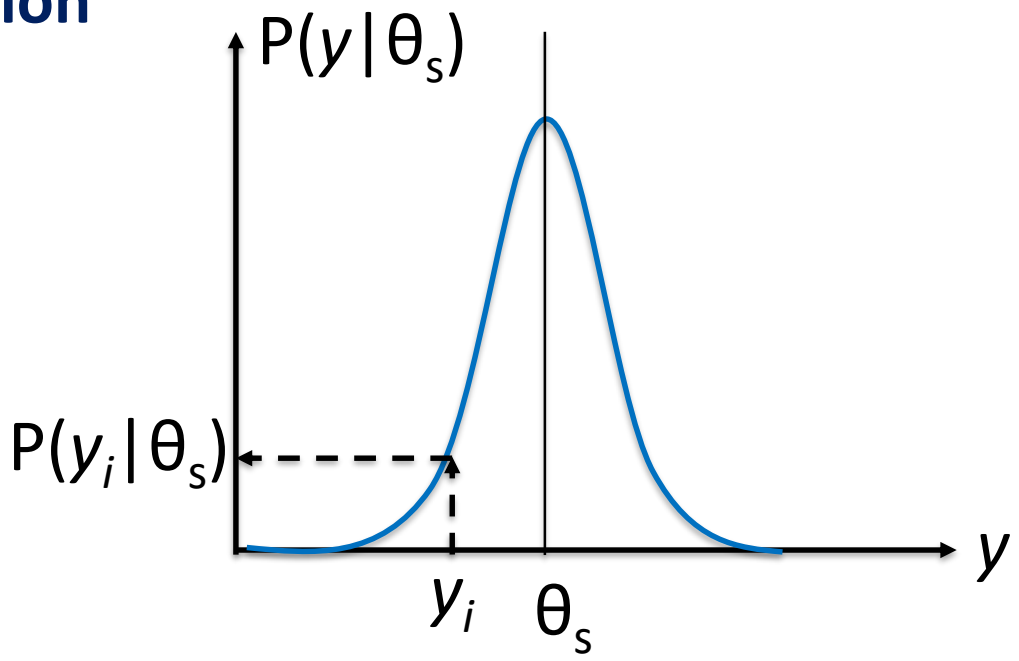
Sumio Watanabe

Tokyo Institute of Technology

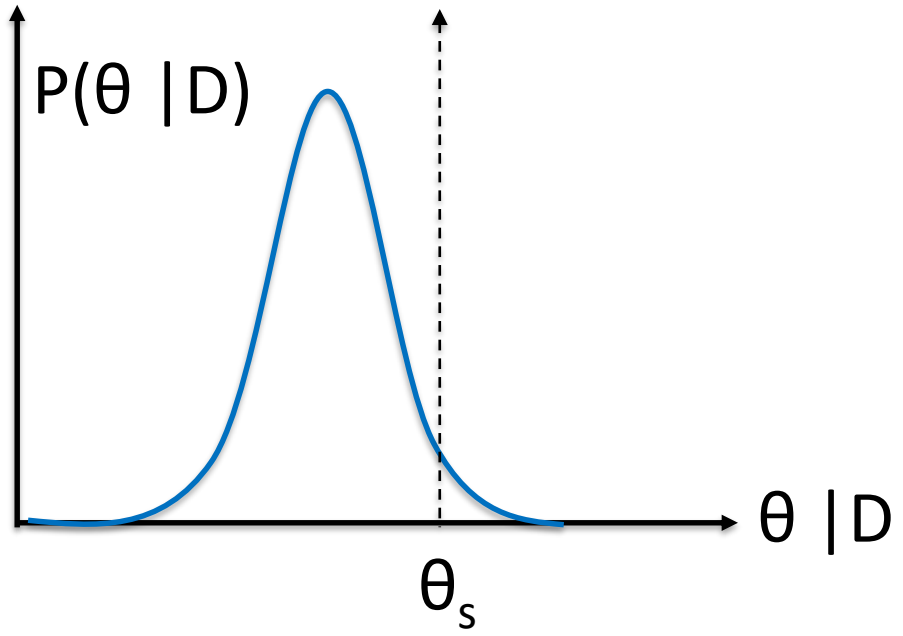
$$
\begin{aligned}
\text{lppd} &= \text{log pointwise predictive density} \\
&= \log \prod_{i=1}^{n} p_{\text{post}}(y_i) = \sum_{i=1}^{n} \log \int p(y_i|\theta) p_{\text{post}}(\theta) d\theta
\end{aligned}
$$

$$
\begin{aligned}
\widehat{\text{lpd}} &= \text{computed log pointwise predictive density} \\
&= \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right).
\end{aligned}
$$

Predictive distribution



Posterior distribution



$$\begin{aligned}\text{lppd} &= \text{log pointwise predictive density} \\ &= \log \prod_{i=1}^{n} p_{\text{post}}(y_i) = \sum_{i=1}^{n} \log \int p(y_i|\theta) p_{\text{post}}(\theta) d\theta\end{aligned}$$

$$\begin{aligned}\widehat{\text{lpd}} &= \text{computed log pointwise predictive density} \\ &= \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right).\end{aligned}$$

Trop optimiste

$y_1, \ldots, y_n$ utilisées deux fois (estimation & prédiction)

Méthode de Watanabe (2010) :

$$\widehat{\mathrm{elpd}}_{\mathrm{waic}} = \widehat{\mathrm{lpd}} - \widehat{p}_{\mathrm{waic}}$$

$$p_{\mathrm{waic}} = \sum_{i=1}^{n} \mathrm{var}_{\mathrm{post}} \left(\log p(y_i|\theta)\right)$$

$$\widehat{p}_{\mathrm{waic}} = \sum_{i=1}^{n} V_{s=1}^{S} \left(\log p(y_i|\theta^s)\right)$$

$$V_{s=1}^{S} a_s = \frac{1}{S-1} \sum_{s=1}^{S} (a_s - \bar{a})^2$$

$$\text{WAIC} = -2\,\widehat{\text{elpd}}_{\text{waic}}.$$

Egale à la validation croisée (asymptotiquement)

Peut être facilement calculé avec le R package loo

School	Estimate (y_j)	Standard Error (σ_j)
A	28	15
B	8	10
C	-3	16
D	7	11
E	-1	9
F	1	11
G	18	10
H	12	18

$$
\begin{aligned}
y_j &\sim \text{Normal}(\theta_j, \sigma_j), \quad j = 1, \ldots, 8 \\
\theta_j &\sim \text{Normal}(\mu, \tau), \quad j = 1, \ldots, 8 \\
p(\mu, \tau) &\propto 1,
\end{aligned}
$$

```r
library(rstan)

schools_data <- list(
  J = 8,
  y = c(28,  8, -3,  7, -1,  1, 18, 12),
  sigma = c(15, 10, 16, 11,  9, 11, 10, 18)
)

fit1 <- stan(
  file = "school.stan",  # Stan program
  data = schools_data,    # named list of data
  chains = 4,             # number of Markov chains
  warmup = 1000,          # number of warmup iterations per chain
  iter = 2000,            # total number of iterations per chain
  cores = 2,              # number of cores (could use one per chain)
  refresh = 0             # no progress shown
  )

print(fit1, pars=c("theta", "mu", "tau", "lp__"), probs=c(.1,.5,.9))

library(loo)
log_lik_1<-extract_log_lik(fit1)
loo_1<-loo(log_lik_1)
loo_1
waic_1<-waic(log_lik_1)
waic_1
```

```
Inference for Stan model: school.
4 chains, each with iter=2000; warmup=1000; thin=1;
post-warmup draws per chain=1000, total post-warmup draws=4000.

          mean se_mean   sd    10%    50%    90% n_eff Rhat
theta[1]  11.36    0.16 8.35   2.29  10.20  22.15  2651    1
theta[2]   7.93    0.09 6.37   0.05   7.92  15.61  4619    1
theta[3]   6.09    0.13 7.71  -3.50   6.74  14.83  3378    1
theta[4]   7.66    0.10 6.58  -0.34   7.68  15.66  4164    1
theta[5]   5.26    0.11 6.39  -3.10   5.74  12.82  3458    1
theta[6]   6.22    0.10 6.72  -2.44   6.68  14.19  4248    1
theta[7]  10.73    0.12 6.86   2.73  10.13  19.68  3282    1
theta[8]   8.53    0.14 7.81  -0.41   8.23  17.91  3109    1
mu         7.88    0.11 5.18   1.41   7.78  14.32  2074    1
tau        6.55    0.15 5.58   0.87   5.20  13.75  1397    1
lp__     -39.54    0.07 2.60 -43.01 -39.35 -36.38  1221    1
```

```
> waic_1<-waic(log_lik_1)
> waic_1

Computed from 4000 by 8 log-likelihood matrix

          Estimate  SE
elpd_waic    -31.0 1.0
p_waic         1.4 0.3
waic          61.9 1.9
```

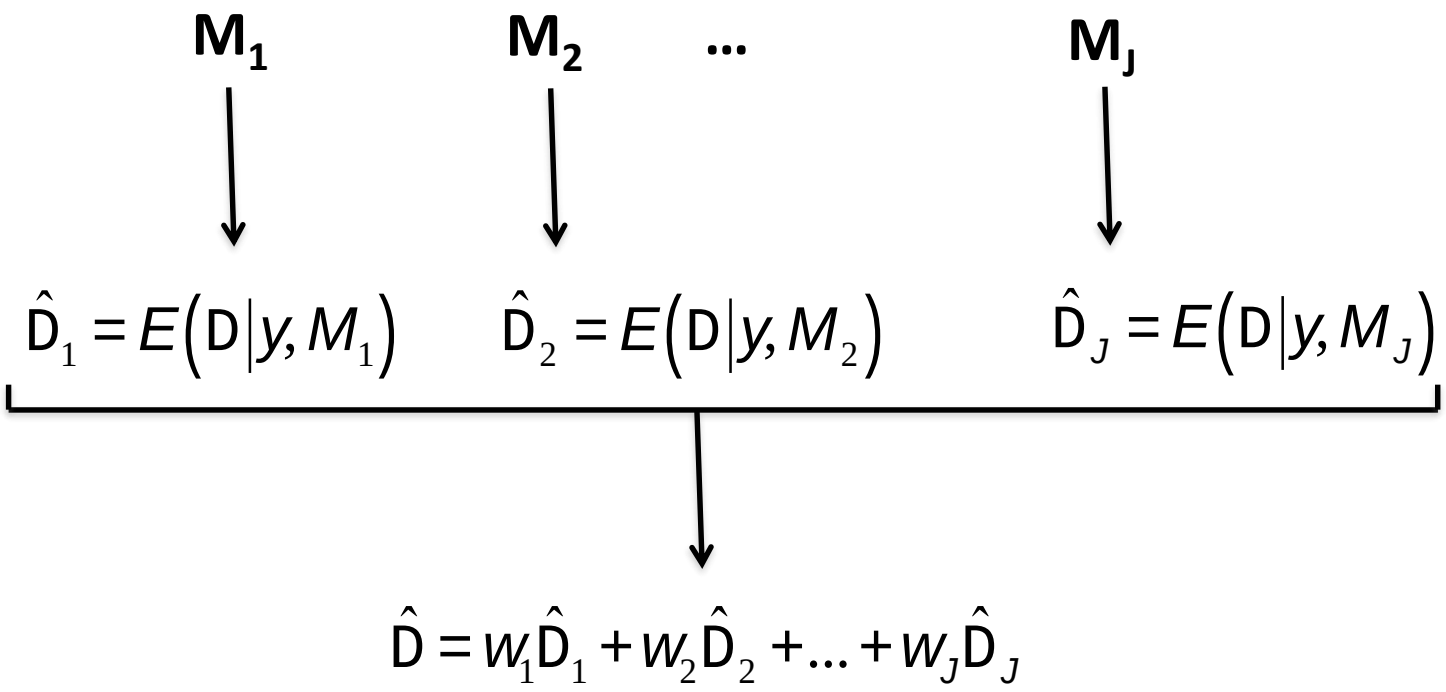
```
> loo_1

Computed from 4000 by 8 log-likelihood matrix

         Estimate  SE
elpd_loo    -31.2 0.9
p_loo         1.6 0.3
looic        62.4 1.9
------
```

1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

Les limites de la sélection

- Parfois, plusieurs modèles ont des performances très proches
- La sélection d'un modèle conduit à ignorer l'incertitude dans la procédure de sélection
- La sélection peut conduire à sous-estimer l'incertitude sur les valeurs des paramètres

Bayesian Model Averaging (BMA)



Bayesian Model Averaging (BMA)

$$E\left(\Delta \middle| y\right) = \sum_{i=1}^{J} w_i \hat{\Delta}_i$$

$$\hat{\Delta}_i = E\left(\Delta \middle| y, M_i\right)$$

$$w_i = [M_i \mid y]$$

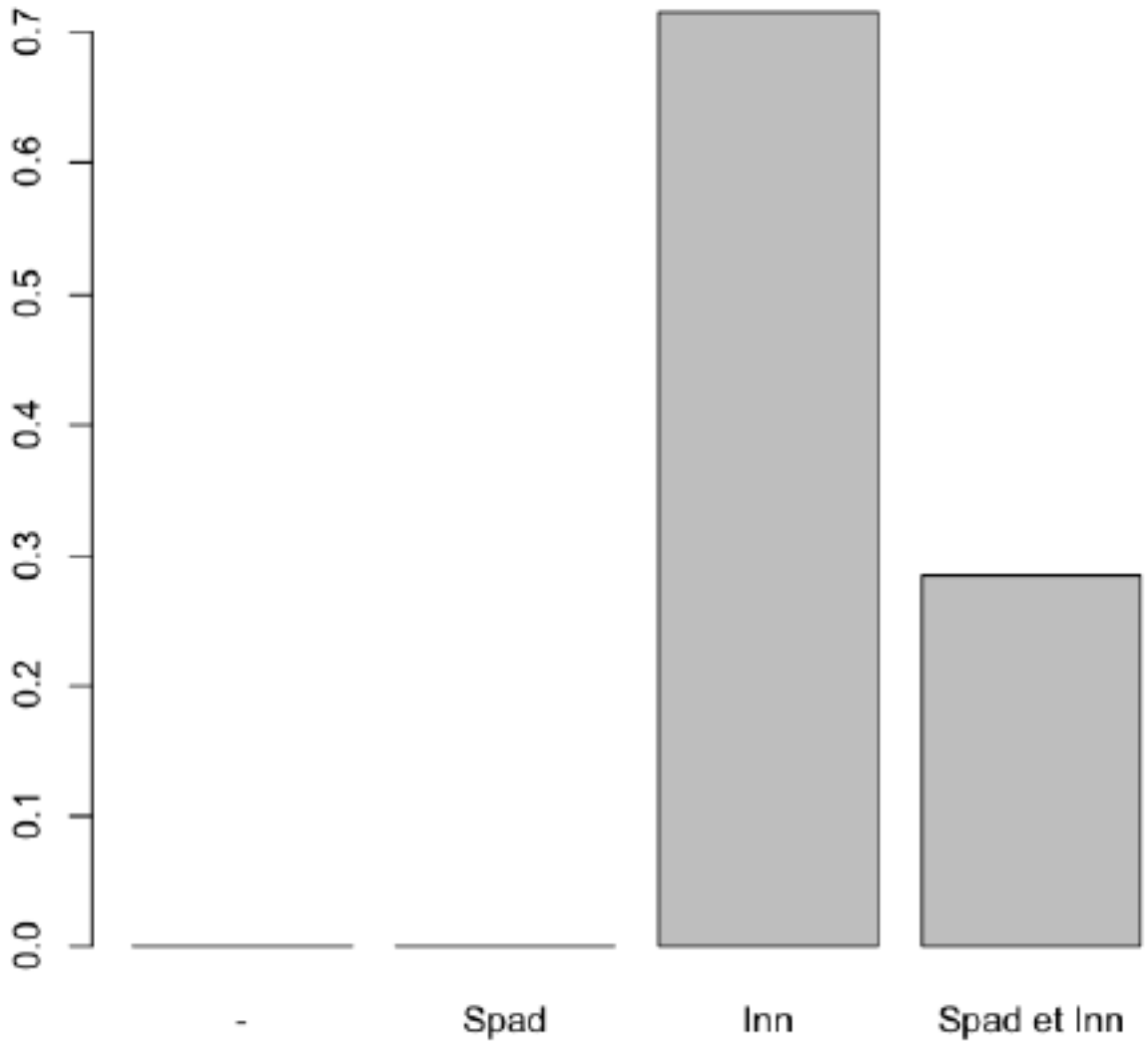
Approximation avec le BIC

$$[M_i \mid y] = \frac{[y \mid M_i][M_i]}{\sum_{k=1}^{J} [y \mid M_k][M_k]}$$

$$\hat{P}\left(M_i \mid y\right) = \frac{\exp\left\{-0.5 \times BIC_i\right\}}{\sum_{k=1}^{J} \exp\left\{-0.5 \times BIC_k\right\}}$$

$$BIC_i = -2 \times L\left(\hat{\theta}_{ML.i}\right) + p_i \times \log\left(n\right)$$

$$\hat{E}\left(\theta \mid y\right) = \sum_{i=1}^{J} \hat{\theta}_{ML.i} \times \hat{P}\left(M_i \mid y\right)$$


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
-
Spad
Inn
Spad et Inn

Utilisation de variables indicatrices

- Définition de variables aléatoires binaires indiquant si les variables explicatives sont incluses ou non dans le modèle
- Estimation des distributions a posteriori des variables binaires avec les données
- La probabilité de chaque variable binaire indique si la variable explicative correspondante a de bonne chance d'être incluse ou non

$$Y = \alpha_0 + \pi_1 \times \alpha_1 \times X_1 + \pi_2 \times \alpha_2 \times X_2 + \epsilon$$

0 ou 1 0 ou 1

A priori

$$p_1 \sim Bern(1/2)$$

$$p_2 \sim Bern(1/2)$$

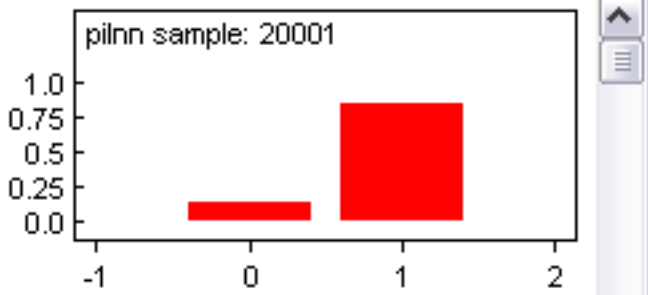
Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
pilnn	0.8525	0.3546	0.004002	0.0	1.0	1.0	10000	20001

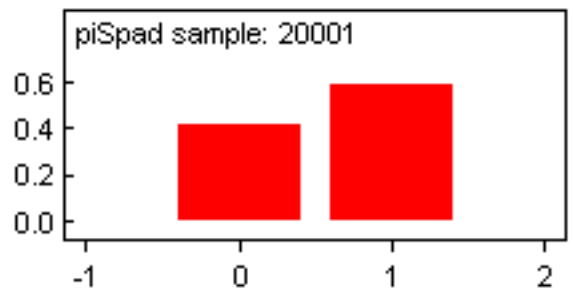
Node statistics

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
piSpad	0.5825	0.4931	0.003945	0.0	1.0	1.0	10000	20001

Kernel density



Kernel density



- Pas toujours possible de tester tous les modèles possibles

  Ex : 20 variables explicatives → 1 048 576 modèles linéaires

- On doit alors … sélectionner un sous ensemble de modèles

1. Introduction
2. Analyse de sensibilité
3. Vérification de la qualité des prédictions a posteriori
4. Validation croisée
5. Facteur de Bayes
6. Vraisemblance pénalisée et déviance
7. Combinaison de modèles
8. Conclusion

Conclusion

- L’analyse de sensibilité permet d’étudier la robustesse des résultats à certaines hypothèses du modèle
- L’analyse prédictive a posteriori permet d’évaluer la plausibilité du modèle
- Différents critères peuvent être utilisés pour sélectionner un modèle parmi plusieurs modèles candidats. Il est recommandé d’utiliser plusieurs critères pour évaluer et sélectionner des modèles.
- Il peut parfois être intéressant de combiner plusieurs modèles, plutôt que d’en utiliser un seul