STA 112
Modèles hiérarchiques bayésiens : Epidémiologie Spatiale

A. Latouche
aurelien.latouche@cnam.fr

Modèles Hiérarchiques [1]

Fréquent d'avoir des structures hiérarchique dans les données, i.e. que les données ont été récoltées à différents niveaux.

- individus qui sont imbriqués dans une unité géographique ou physique
  - Cas d'une maladie
  - Patients au sein d'hôpitaux

$\Rightarrow$ Modèles hiérarchiques pour saisir la variation dans **chacun** des différents niveaux.

[1] aussi appelé modèle multi-niveaux

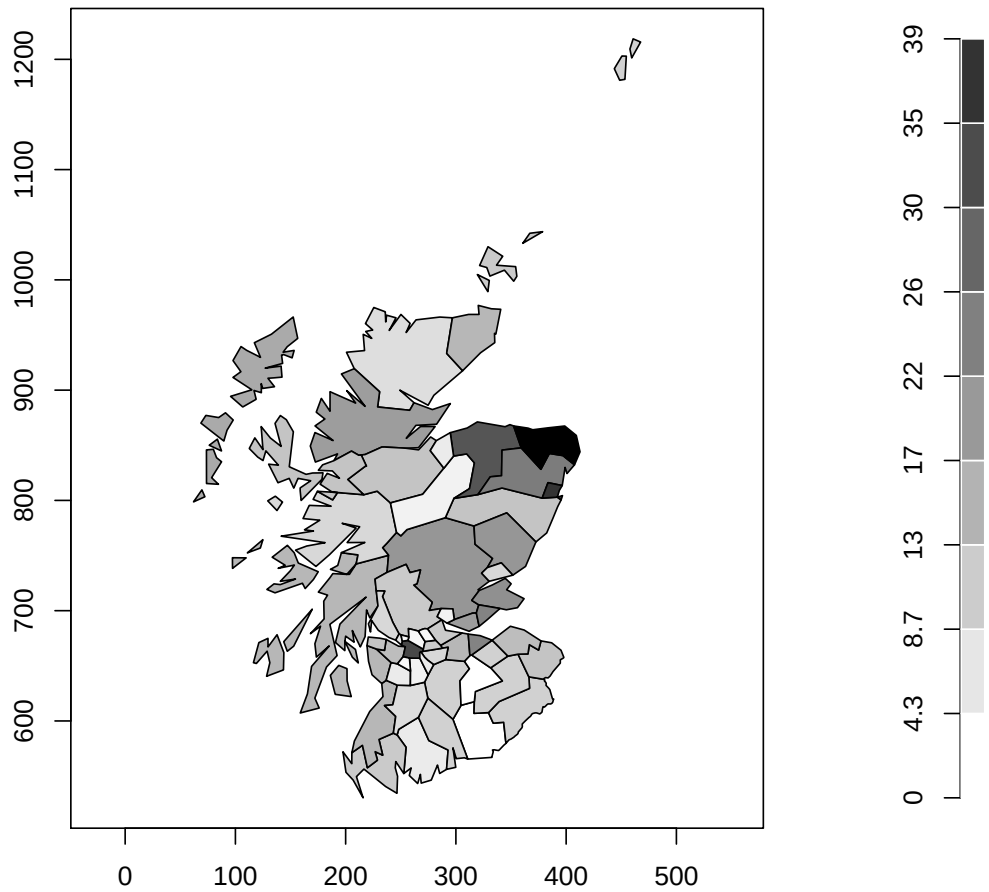
Epidémiologie spatiale

Objectif : étudier les variations spatiales des taux d'*incidence* de maladies

- Peu de cas observés par zone : maladie rare et/ou zones petites)

- Observations corrélées spatialement

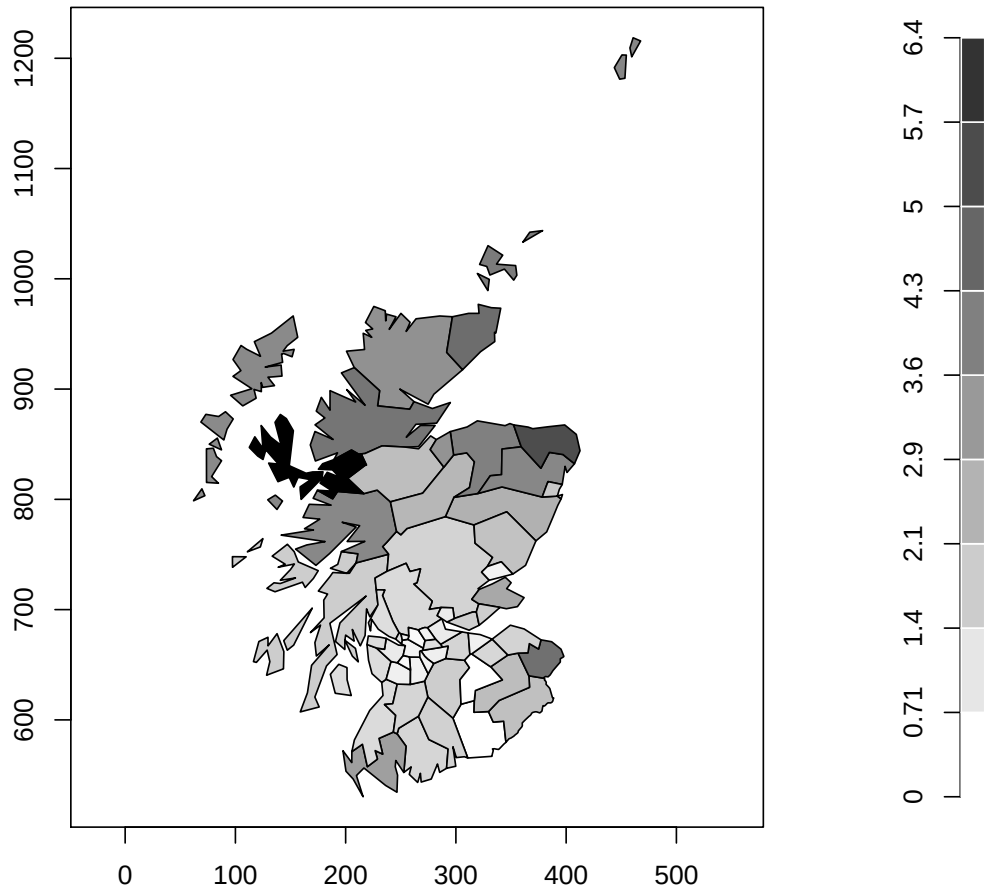
Cas de cancer poumon en Ecosse (1975-1980)



data(scotland) du package **SpatialEpi**

Cancer du poumon (1975-1980) : SMR

ratio de mortalité standardisé[2]



[2](SMR=# cas observés/ # cas attendus)

Objectifs

Les analyses spatiales en épidémiologie ont plusieurs objectifs

- Représentation des variations spatiales : comparaison entre maladies, évolution temporelle . . . (
- Etudes de corrélations géographiques
- Suggérer des hypothèses concernant l'origine de l'hétérogénéité des risques

Ce sont des études dites *écologiques*

Etudes Ecologiques

Problèmes spécifiques à ce type d'études:

- L'interprétation est dépendante du niveau géographique de l'analyse
- Comment modéliser les variations résiduelles et prendre en compte les corrélations spatiales ?
- Comment prendre en compte les facteurs de confusion et des modifications d'effet potentiellement liées à la structure spatiale?

A quel niveau géographique s'effectue la modélisation ?

- Données ponctuelles: la localisation exacte des cas est observée
- Données groupées : le nombre total de cas dans une zone donnée est enregistré : Données agrégées.
- Les mesures d'exposition peuvent être disponibles soit au niveau individuel, soit (le plus courant) comme caractéristiques du groupe

Données groupées

+ Les plus courantes, données facilement disponibles

+ Correspondent à des contrastes environmentaux pour certains types d'exposition $\rightarrow$ larges intervals de variation

- Le niveau d'aggrégation est souvent administratif et artificiel

- Taille de la population à risque varie de manière substantielle entre les zones . Lien individuel entre exposition et effet sur la santé est perdu

Notations

- Y_i le nombre de cas observés (typiquement faibles) dans la zone i
- Ensemble de zones A_i représentant une partition
- E_i est le nombre attendu calculé en fonction de la distribution d' âge, sexe
- R_i est un risque relatif (RR) spécifique à la zone i . Objet de l'inférence

Modèle non-hiérarchiques spatiaux

- L' estimateur du maximum de vraisemblance du SMR :

$$SMR_i = \frac{Y_i}{E_i}$$

  ne tient pas compte de la structure spatiale, a une variabilité proportionnelle à $1/E_i$

- Modèles Poisson ne sont pas flexibles : souvent le données sont sur-dispersées

- On adopte une formulation hiérarchique où les R_i sont traités comme des variables aléatoires (cadre du modèle mixte)

- Cadre bayésien pour l'inférence sur la distribution conjointe des R_i

Modèle hiérarchiques spatiaux

- 1er Niveau : $Y_i|E_i, R_i \sim Poisson(R_i E_i)$
- Second niveau : La distribution de R_i est spécifiée :
  - De manière paramétrique ou semi-paramétrique
  - En faisant appel à un graphe de contiguïté predéfini: $i \sim j \leftrightarrow i$ voisin de j

Modèle de poisson hiérarchique

$$\begin{array}{lcll} Y_i|R_i & \sim & \text{Poisson}(E_i R_i) & \\ \log(R_i)|\alpha,\beta,X_i & = & \alpha+\beta X_i & (M_0) \end{array}$$

- Le deuxième niveau modélise la relation entre le risque et les variables d'exposition
- M_0 n'introduit pas d'hétérogénéité spatial (ou non spatial)

Hétérogénéité Spatiale et Non-Spatiale

$$
\begin{array}{lcl}
Y_i|R_i & \sim & \text{Poisson}(E_i R_i) \\
\log(R_i)|\alpha, \beta, U_i, V_i & = & \alpha + \beta X_i + U_i + V_i \\
U_i|U_{j\neq i} & \sim & N(\sum_{j\in\delta_i} U_j/n_i, 1/\tau_U^2 n_i) \\
V_i|\tau_V & \sim & N(0, \tau_V^2)
\end{array}
\qquad (BYM)
$$

où

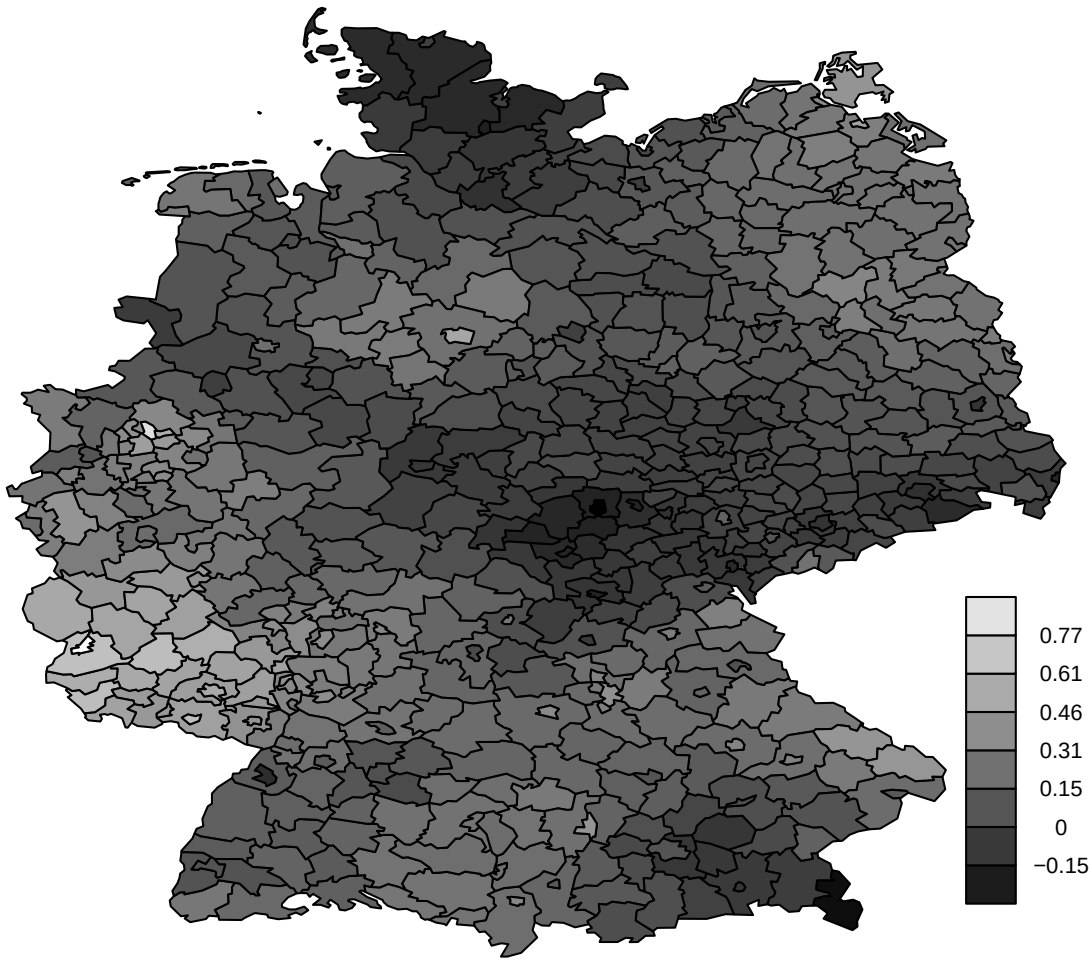
- δ_i voisins de la zone i,
- n_i le nombre de voisin i
- U_i $(i = 1, \ldots, m)$ effet aléatoire spatialisé : Conditional Auto Regressive
- V_i $(i = 1, \ldots, m)$ effet aléatoire non-spatialisé

Besag, York, Mollié :
`http://www.ism.ac.jp/editsec/aism/pdf/043_1_0001.pdf`

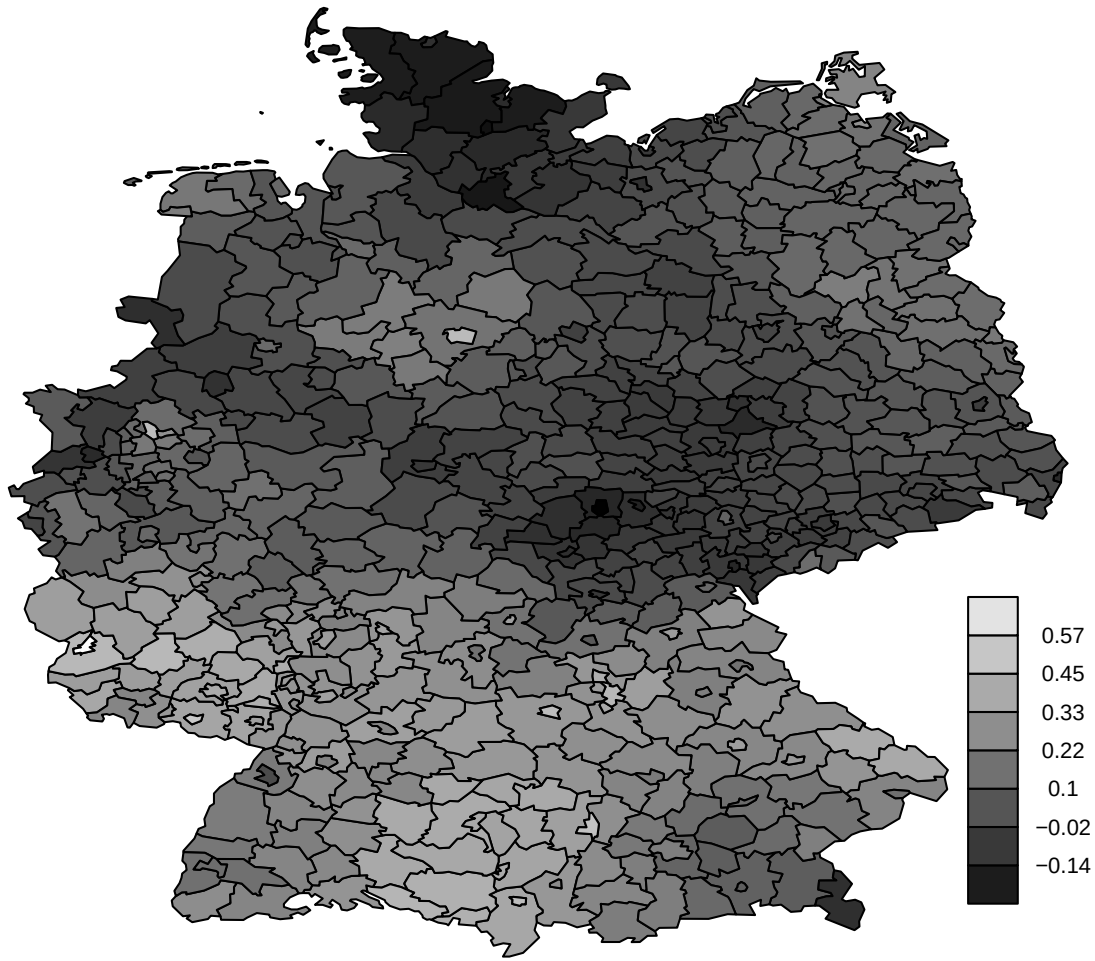
Illustration

- Données : Cancer Oesophage en Allemagne 1986-1990
- Résolution spatiale : 544 regions
- Covariable : Consommation Cigarette
- Pour chaque région i : $Y_i|R_i \sim \text{Poisson}(E_i R_i)$
- $\log(R_i)|\alpha, \beta, U_i, V_i = \alpha + \beta X_i + U_i + V_i$

Estimation de la moyenne *a posteriori* des U_i sans covariable



Estimation de la moyenne *a posteriori* des U_i (avec covariable)



Estimation des modèles hiérarchique bayésiens

‘

- Méthodologie bayésienne
- Exemple
- Méthodes MCMC

Pourquoi utiliser l'inférence bayésienne ?

- Maximum de Vraisemblance : Complexité d'estimation
  - Données non gaussienne
  - Taille d'échantillon grande pour garantir la convergence des estimateurs

Pourquoi utiliser l'inférence bayésienne ?

- Maximum de Vraisemblance : Complexité d'estimation
  - Données non gaussienne
  - Taille d'échantillon grande pour garantir la convergence des estimateurs
- Mise en œuvre pratique
  - Markov Chain Monte Carlo
  - Logiciels : WinBUGS, R2WinBUGS, OpenBUGS, JAGS ...

Un exemple (très) simple

- Soit θ la proportion de lots conformes

$$Y_i \sim \text{Bernoulli}(\theta)$$

- N Lots (indépendants)
- L'estimateur du maximum de vraisemblance (MLE) :

$$\hat{\theta} = \sum_i Y_i / N$$

Cet estimateur maximise quelle quantité ?

Inférence Bayésienne

Un exemple (très) simple

- Soit θ la proportion de lots conformes

$$Y_i \sim \text{Bernoulli}(\theta)$$

- N Lots (indépendants)
- L'estimateur du maximum de vraisemblance (MLE) :

$$\hat{\theta} = \sum_i Y_i / N$$

Cet estimateur maximise quelle quantité ?

$\Rightarrow P(Y|\theta)$

Inférence Bayésienne

Un exemple (très) simple

- Soit θ la proportion de lots conformes

$$Y_i \sim \text{Bernoulli}(\theta)$$

- N Lots (indépendants)
- L'estimateur du maximum de vraisemblance (MLE) :

$$\hat{\theta} = \sum_i Y_i / N$$

Cet estimateur maximise quelle quantité ?

$\Rightarrow P(Y|\theta)$ i.e. la vraisemblance

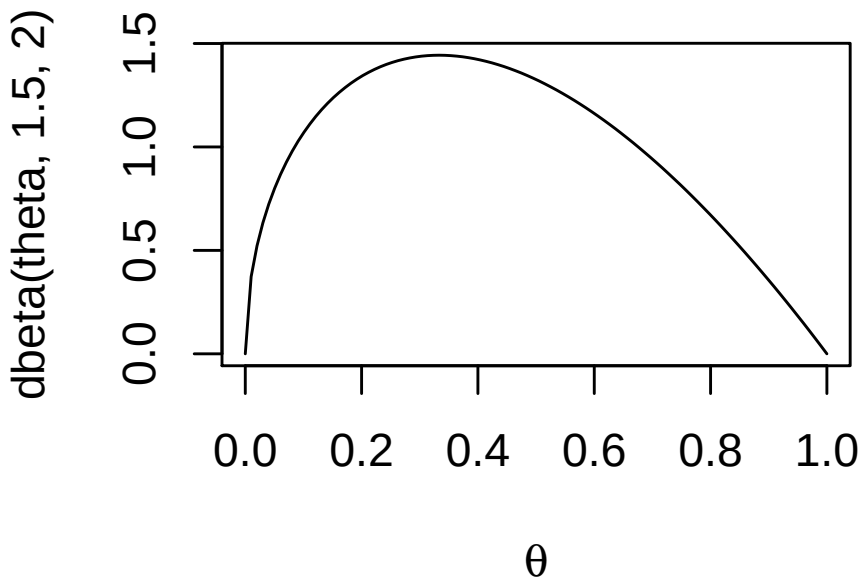
Approche Bayesienne

a priori

- $[\theta]$ *distribution a priori*
- Représente les probabilités que θ peut prendre
- **avant** d'avoir utilisé les données

$$\theta \sim \text{Beta}(1.5, 2)$$

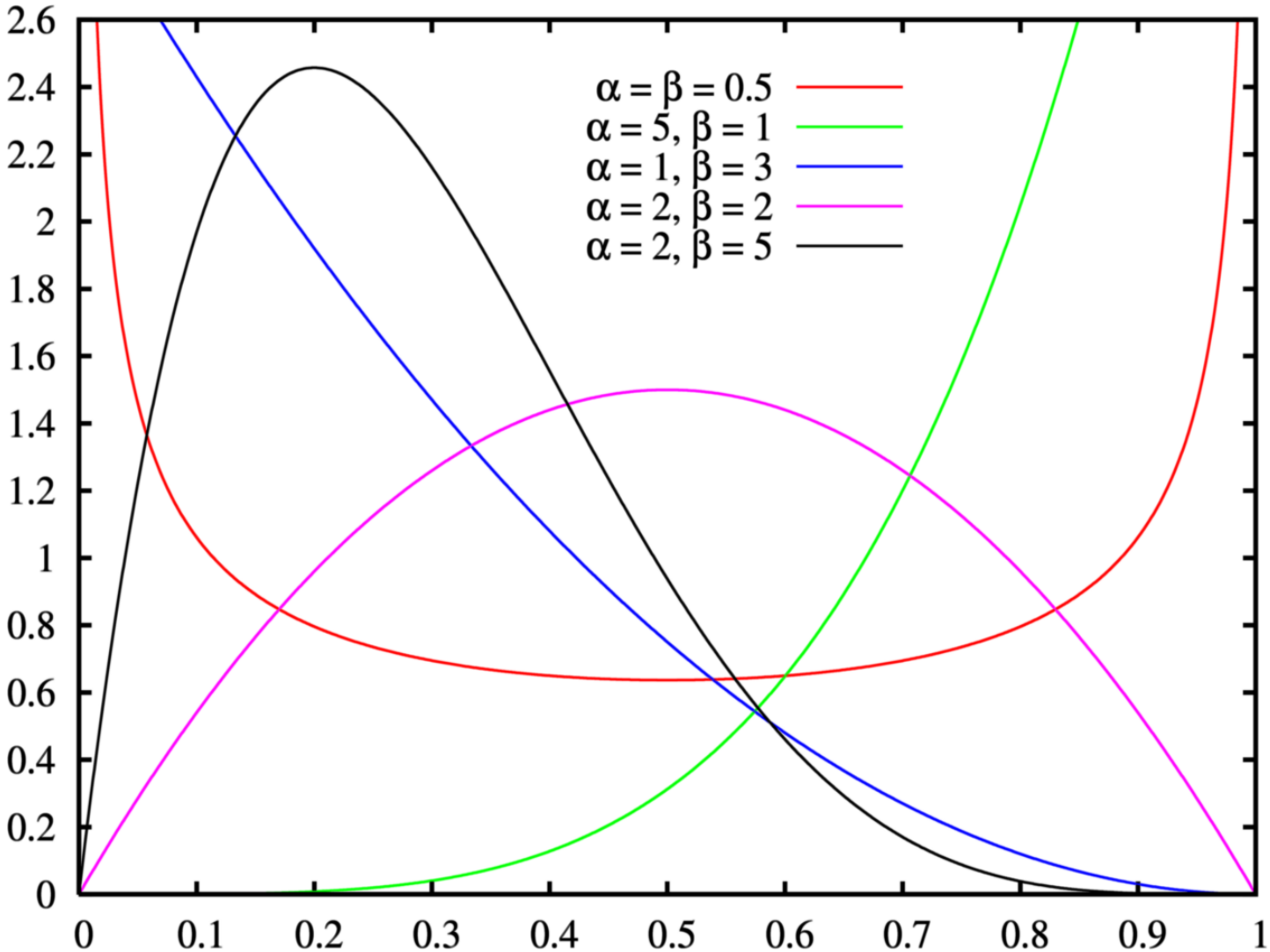
$\theta \sim \text{Beta}(1.5, 2)$



Distribution $Beta(a, b)$

- Souplesse d'utilisation
- Distribution Flexible
- Si $X \sim Beta(a, b)$ alors $E(X) = \frac{a}{a+b}$ et $Var(X) = \frac{ab}{(a+b)^2(a+b+1)}$
- Densité $f(x) = \dfrac{x^{a-1}(1-x)^{b-1}}{\int_0^1 u^{a-1}(1-u)^{b-1}du}$ pour $x \in [0; 1]$

Distribution $Beta(a, b)$



Distribution *a posteriori*

Probabilité conditionnelle

$$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$$

Approche Bayésienne

$$P(\theta|Y) = \frac{P(Y|\theta) \times P(\theta)}{P(Y)}$$

Y données de l'expérience, θ paramètre(s) du modèle

Distribution *a posteriori*

- $[Y|\theta]$ est notre modèle : Loi de Bernoulli
- $[\theta]$ est l' *a priori* : Loi Beta
- $[\theta|Y]$ est la loi *a posteriori*

$$[\theta|Y] \approx [Y|\theta] \times [\theta]$$

Ce qu'on peut écrire

a posteriori $\approx$ vraisemblance $\times$ *a priori*

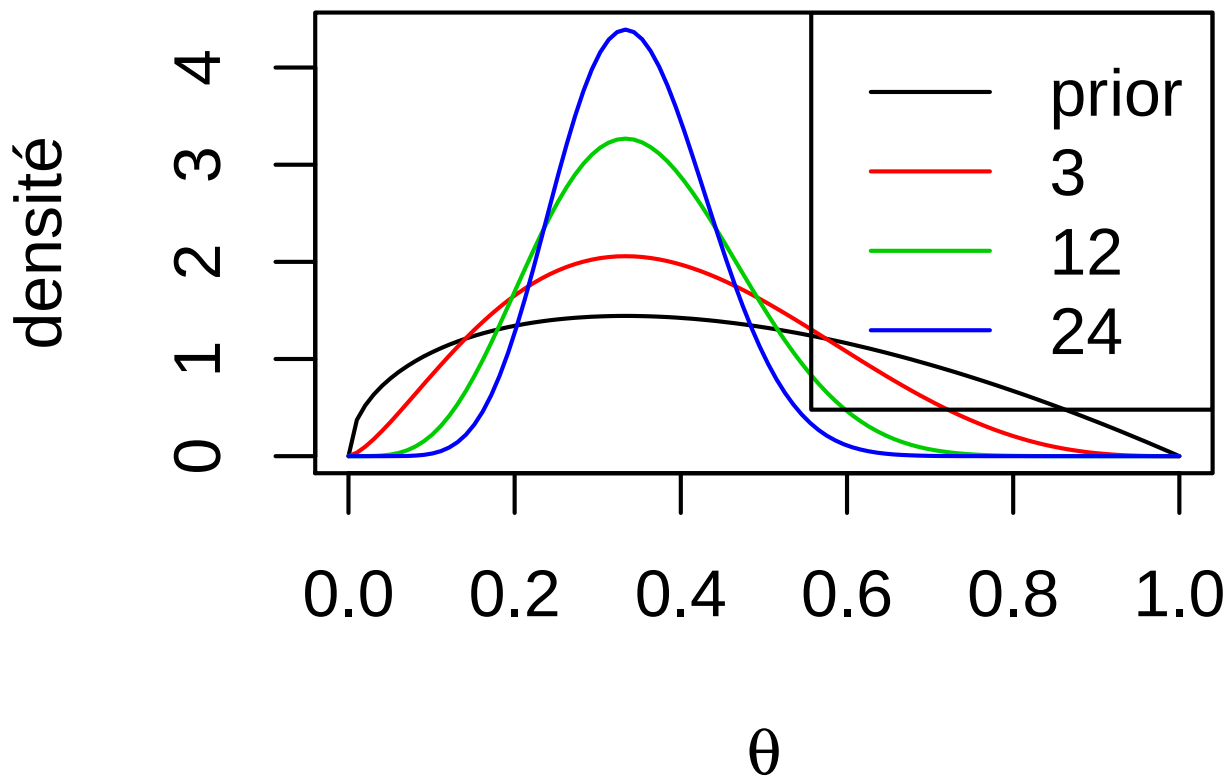
Exemple

- Sans donnée, l' *a posteriori* est l' *a priori*
- Avec des données, l' *a posteriori* est différent de l' *a priori*
- Si la loi a posteriori est différente de la loi a priori alors les données ont apporté une réelle information
- Pour le modèle Beta-Binomial on montre que l' *a posteriori*

$$[\theta|Y] = \text{Beta}(1.5 + \sum_i Y_i, 2 + N - \sum_i Y_i)$$

Densité *a posteriori*

- Supposons que 1/3 des lots échantillonnés soit conforme
- Considérons une taille d'échantillon $N=$ 3, 12 ou 24



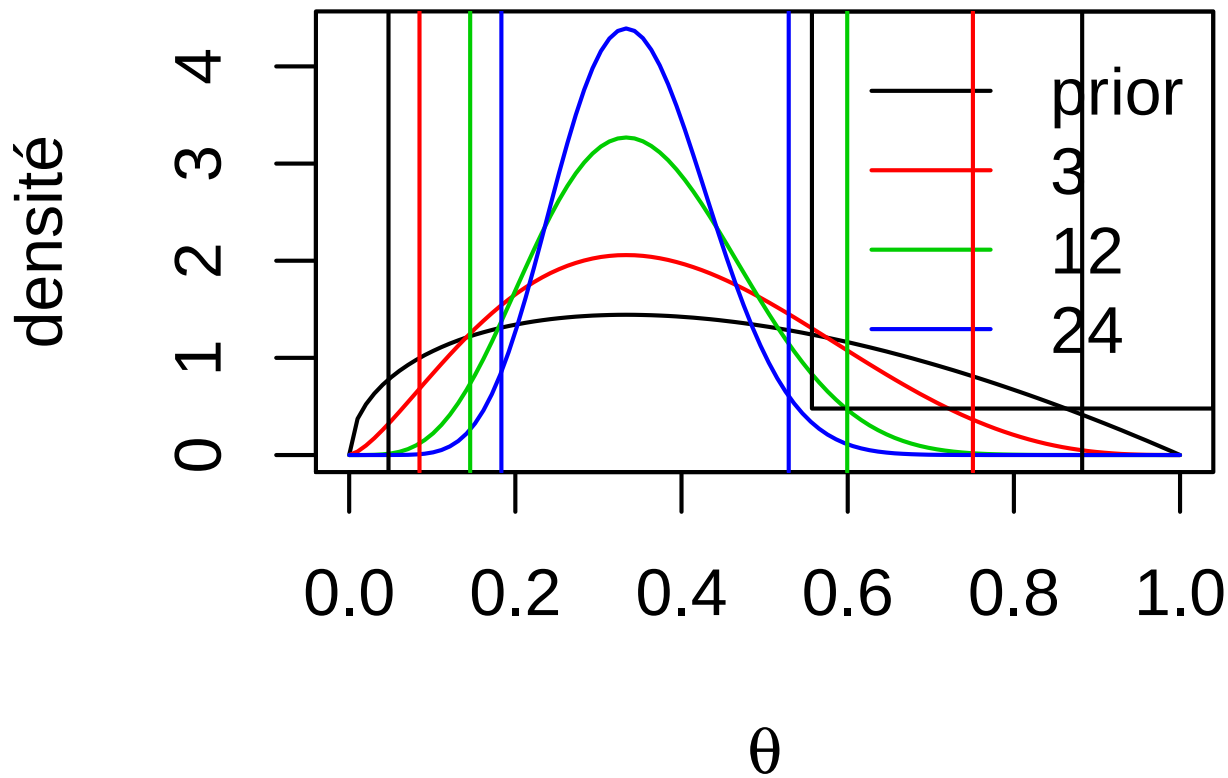
Comment calculer l' *a posteriori*?

Formule de Bayes

$$\begin{aligned} pr(\theta|Y) &= pr(Y|\theta)pr(\theta)/pr(Y) \\ &= pr(Y|\theta)pr(\theta)/\int pr(Y|\theta)pr(\theta)d\theta \end{aligned}$$

- Expression analytique de la loi aposteriori du modèle beta-binomial (quelques calculs)

- Utilisé pour calculer des Intervalles de crédibilité à 95%


densité
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0
θ
prior
3
12
24

Lois *a priori* conjuguées pour certaines familles

$[Y\|\theta]$	$[\theta]$	$[\theta\|Y]$
Normale	Normale	
Poisson	Gamma	
Binomiale	Gamma	
Normale	Gamma	

Conjugaison : la distribution de l' a priori et de l' a posteriori sont dans la même famille

Interpretations

Vraisemblance

- θ est un nombre fixe
- L'intervalle de confiance I est aléatoire
- L' intervalle de confiance construit contient θ 95% du temps
- pr($\theta \in I$) =0 ou 1

Bayes

- θ est une variable aléatoire
- Choisir un intervalle I de telle sorte qu'il contienne 95% de la masse de la distribution *a posteriori* $[\theta|Y]$
- pr($\theta \in I$) = 0.95

Choix de l' *a priori*

On distingue

- *a priori* informatif (souvent conjugué)
  - Voir tableau de conjugaison
  - Même support : Uniforme sur $[0, 1]$
- *a priori* non informatif :
  - $Y = \beta X + \varepsilon$ et $\varepsilon \sim N(0, \sigma^2)$
  - Sur β : $N(0, 10^6)$
  - Sur σ : $Gamma(10^-3, 10^-3)$
  - Loi uniforme impropre $U[-\infty, +\infty]$

a priori non informatif

- *a priori* de Jeffrey $[\theta] \sim I(\theta)^{1/2}$
- $I(\theta)$ est l'information de Fisher

$$I(\theta) = -E\left[\frac{d^2 \log f(Y|\theta)}{d\theta^2}\right]$$

Comparaison de modèles

Pour 2 modèles M_1 et M_2 et des données $\mathbf{Y}$

Bayes Factor

$$BF_{21} = \frac{P(M_2|Y)}{P(M_1|Y)} \Big/ \frac{P(M_2)}{P(M_1)}$$

$$BF_{21} = \frac{[Y|M_2]}{[Y|M_1]} = \frac{\int [Y|\theta M_2][\theta|M_2] d\theta}{\int [Y|\theta M_1][\theta|M_1] d\theta}$$

Se généralise à K modèles

Table "pratique" de Jeffreys pour la décision en faveur de M_2 par rapport à M_1

- en termes de BF_{21}
- ou de probas a posteriori $P_{21} = \frac{[D|M_2]}{[D|M_1]+[D|M_2]}$

$$1 < BF_{21} < 3 \quad \Leftrightarrow \quad 0.5 < P_{21} < 0.75 \quad \Rightarrow \textit{faible}$$

$$3 < BF_{21} < 12 \quad \Leftrightarrow \quad 0.75 < P_{21} < 0.92 \quad \Rightarrow \textit{positif}$$

$$12 < BF_{21} < 150 \quad \Leftrightarrow \quad 0.92 < P_{21} < 0.99 \quad \Rightarrow \textit{fort}$$

$$BF_{21} > 150 \quad \Leftrightarrow \quad P_{21} > 0.99 \quad \Rightarrow \textit{decisif}$$

où D sont les Données, implémenté dans `MCMCpack`

Remarques sur le Bayes Factor

1. ce n'est pas une maximisation mais une intégration sur l'espace des paramètres
2. pas nécessairement de modèles emboités
3. interprétation par le rapport des "cotes"
4. même écriture que le rapport des vraisemblances dans le cas d'hypothèses simples
5. dans le cas général, BF s'exprime en rapport d'intégration
6. BF n'est pas défini dans le cas de lois a priori impropres
7. BF peut être sensible aux choix des lois a priori

WinBUGS

Bayesian **I**nference **U**sing **G**ibbs **S**ampling

`http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml`

- Installation
- Mise à jours, Clé
- Intuitif pour la spécification des modèles , mais pas pour la gestion des résultats de simulations

$\Rightarrow$ Le meilleur des 2 mondes R2WinBUGS

R2winBUGS

- La commande R `sink` permet de spécifier le modèle pour WinBUGS
- Task view Bayes
  `http://cran.r-project.org/web/views/Bayesian.html`
- Gibbs-Sampling ?

Méthodes MCMC

En très court

MCMC est une méthode de Monte Carlo pour simuler la distribution *a posteriori* $[\theta|Y]$

- Quand un modèle a été postulé pour θ
- Quand la vraisemblance $[Y|\theta]$ est connue
- Quand on dispose de données Y !

MCMC produit un échantillon de la distribution a posteriori

Méthode de Monte Carlo

- **Simulation**
- Méthodes "MC classiques" : on suppose les simulations i.i.d

Soit X_1, X_2, ... des réalisations simulées i.i.d d'une distribution On veut estimer

$$\mu = E\{g(X_i)\}.$$

Méthode de Monte Carlo

La loi des grands nombres nous fournit :

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$$

converge en probabilité vers μ.
Le Théorème de la limite centrale nous fournit

$$\sqrt{n}(\hat{\mu}_n - \mu) \Rightarrow N(0, \sigma^2)$$

où

$$\sigma^2 = var\{g(X_i)\}$$

qu'on peut estimer par

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} (g(X_i) - \hat{\mu}_n)^2$$

On peut ensuite en déduire des IC pour μ

$$\hat{\mu}_n \pm 1.96\frac{\hat{\sigma}_n}{\sqrt{n}}$$

Chaînes de markov

- Une *Chaine de Markov* est une suite de v.a. dépendantes $X_1, X_2, \ldots, X_n$ qui a la propriété suivante: La distribution du futur conditionnellement au passé ne dépend que de l'état présent
- Ce qu'on exprime $[X_{n+1}|X_1, \ldots, X_n]$ ne dépend que de X_n.

Une chaine de markov est dite stationnaire si la loi $[X_{n+1}|X_n]$ ne dépend pas de n

MCMC : Monte Carlo par Chaînes de Markov

WinBUGS construit cet algorithme, une fois le modèle spécifié

MCMC principe

On a k paramètres $\theta = (\theta_1, \ldots, \theta_k)$ et on veut simuler un échantillon de θ selon la loi multidimensionnelle a posteriori $[\theta|Y]$: la loi d'intérêt, difficile à simuler

- une itération t de l'algorithme va fournir une réalisation θ^t de la loi $[\theta|x]$ de dimension k, $\theta^t = (\theta_1^t, \ldots, \theta_k^t)$
- l'ensemble des réalisations θ^t est une **chaîne de Markov** (le temps est ici l'itération) (rem : échantillon avec dépendance markovienne).
- Si on obtient beaucoup d'itérations, alors on pourra avoir une bonne estimation de la loi a posteriori (ainsi que des caractéristiques des distributions a posteriori marginales de chaque paramètre comme la moyenne, variance, quantiles a posteriori ...), ceci d'après le théorème ergodique.

Théoréme ergodique : TCL pour des données non i.i.d.

MCMC : echantilloneur de Gibbs

Y les données, θ vecteur de paramètre de dim k

1. Choix des valeurs initiales des paramètres $\theta^t = (\theta_1^{(0)}, \ldots, \theta_k^{(0)})$
2. Tirer $\theta_1^{(1)}$ de $[\theta_1 | \theta_2^{(0)}, \theta_3^{(0)}, \ldots \theta_k^{(0)}, Y]$
3. Tirer $\theta_2^{(1)}$ de $[\theta_2 | \theta_1^{(1)}, \theta_3^{(0)}, \ldots \theta_k^{(0)}, Y]$
4. ...
5. Tirer $\theta_k^{(1)}$ de $[\theta_k | \theta_1^{(1)}, \theta_2^{(1)}, \theta_3^{(0)}, \ldots \theta_{k-1}^{(1)}, Y]$
6. On itère

Problèmes et caractéristiques

- Méthodes itératives (peut être long...)
- Savoir simuler des lois
- Choix du nombre d'itérations : ce nombre doit être assez grand pour l'oubli des points de départ (*burn in* ou temps de chauffe) et la bonne couverture de la loi a posteriori.
- En pratique, on se fixe une valeur M itérations (temps de chauffe pour atteindre la loi stationnaire) et T itérations (T très grand) afin d'obtenir T réalisations de la chaîne de Markov sous la loi a posteriori recherchée
- L'algorithme comporte donc (M+T) itérations dont les M premières ne sont pas utilisées dans les inférences.
- Diagnostics de convergence (pb encore actuel) ... M et T assez grands ?

Diagnostics de convergence

La détection de non-convergence est essentielle car l'ensemble de la théorie est fondé sur le fait que l'échantillon provenant de l'algorithme provient de la loi stationnaire de la chaîne de Markov $[\theta|Y]$

1. Choix du *Burn in* ? (Régime stationnaire)

2. Combien d'itération post chauffe ? Ne pas être économe !

Conseils

Facilement accessibles sous WinBUGS

- Le graphe des réalisations de chaque paramètre en fonction des itérations peut permettre de détecter un problème de lenteur de mélange (*slow mixing*) et dans ce cas, il est indispensable de faire un plus grand nombre d'itérations.
- La vérification de la stabilité (par rapport aux itérations) de quelques statistiques comme la moyenne, la variance et surtout, les quantiles d'ordre α (pour α petit ou grand) est importante car ces statistiques correspondent, en général, aux résumés utilisés pour caractérisées la loi marginale a posteriori de chaque paramètre.

Conseils

- La comparaison des lois a priori et des lois marginales a posteriori permet de voir si les données ont apporté une information.
- Vérification pour chaque paramètre que la "MC error" est inférieure à 5% de l'écart-type a posteriori du paramètre
- Etude de la sensibilité aux lois a priori

Le *Rhat*

Gelman et Rubin :

- Ce diagnostic nécessite de lancer plusieurs fois l'algorithme avec des valeurs initiales différentes et donc d'obtenir plusieurs réalisations des chaînes de Markov.
- Il vérifie que chaque algorithme amène bien à un échantillon provenant de la même loi en comparant les variances intra-chaîne et inter-chaîne
- (Var totale/Var within proche de 1).

Valeurs acceptables de 1 à 1.1