Introduction à l'économétrie bayésienne

Guillaume Horny*

*Banque de France et UCLouvain
guillaume.horny@banque-france.fr

Plan de la présentation

Partie I

Introduction

Qui est Thomas Bayes?

Révérend Thomas Bayes (1763)
An essay towards solving a problem in the doctrine of chances,
Philosophical Transaction of the Royal Society of London, 53, 370-418
republié dans *Biometrika*, 45, 3/4, 293-315, 1958.

Harold Jeffreys

Formulation moderne: Jeffreys (1939), *Theory of Probability*.

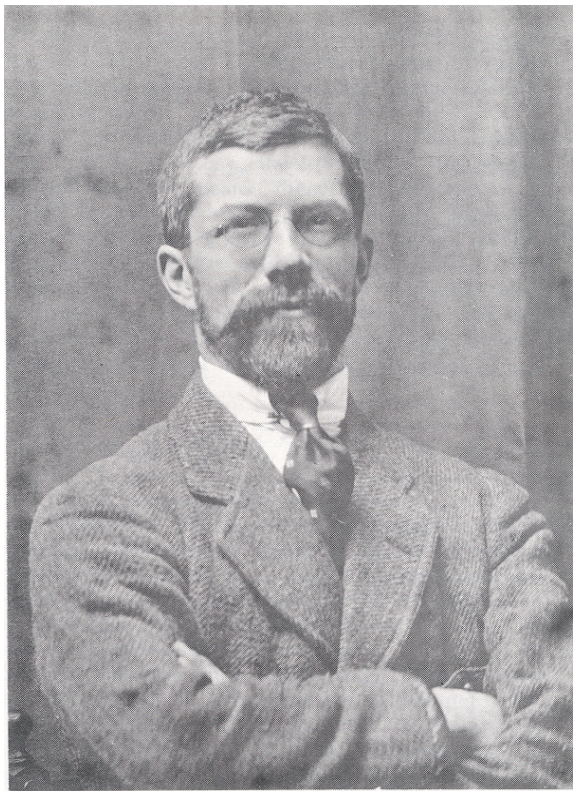
Harold Jeffreys (1891-1989)

Mathématicien, statisticien, géophysicien et astronome. Enseigna au St John's College, fut fait chevalier en 1953 et reçut la médaille d'or la Royal Astronomical Society en 1937.

Fisher et Pearson

50 années suivantes marquées par les méthodes MV et MM

Ronald Fisher (1890-1962)



Karl Pearson (1857-1936)

Le renouveau

années 1990: découverte MCMC et progrès informatiques
$\Rightarrow$ littérature bayésienne foisonnante

> *"Bayesianism has obviously come a long way. It used to be that you could tell a Bayesian by his tendency to hold meetings in isolated parts of Spain and his obsession with coherence, self-interrogation and other manifestations of paranoia. Things have changed..."*

Clifford (1993), JRSS.

Partie II

L'approche bayésienne

Plan de la partie

Q'est-ce qu'une probabilité?

> *"there is no problem about probability: it is simply a non-negative additive set function, whose maximum value is unity"*
>
> Kyburg and Smokler (1980)

La théorie définit le propriétés de $Pr()$, mais comment l'interprèter?
Deux visions concurrentes: objective vs subjective

Interprétation objective (fréquentiste)

La probabilité de l'événement A est la limite de la fréquence empirique de A. Pour n expériences aléatoires et m réalisations de A, on a:

$$Pr(A) = \lim_{n \to \infty} \frac{m}{n}.$$

Limites de l'interprétation objective

Comment s'interprète $Pr()$ lorsqu'une expérience n'est pas répétée?
Quelle est la probabilité :

- que Lehmann Brothers disparaisse?
- qu'une tornade frappe dans le sillage de Ike le comté de Cameron (Texas)?
- que le mât de la grue du chantier d'à côté ne soit pas fiable?

Interprétation subjective

La probabilité de l'événement A mesure la croyance en la proposition que A représente

- A n'est plus nécessairement répliquable
- interprétation proche du langage courant
- permet de conclure:
  "Au vu des données, l'hypothèse A est moins probable que B"

Pour aller plus loin: di Finetti (1974)

En quoi consiste l'approche bayésienne?

Elle consiste à mettre à jour les croyances, suite à l'acquisition d'une nouvelle information, en utilisant la formule de Bayes

L'approche bayésienne est-elle scientifique?

Leamer (1978): *Specification searches*, sous-titré:
"Ad hoc inference with nonexperimental data"
Leamer (1983): "Let's take the con out of econometrics", *AER*

Message

- les données ne parlent jamais sans analyste: l'exercice est inévitablement subjectif!
- autant modéliser explicitement les croyances et les mettre à jour au moyen d'une formule cohérente.

La formule de Bayes (I)

Soient $y \in \mathcal{R}^n$ les observations et $\theta \in \Theta \in \mathcal{R}^q$ les paramètres

Formule de Bayes

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{p(y)}$$

La formule de Bayes (II)

Un modèle est caractérisé par:

- densité jointe des observations (même expression que la vraisemblance)

$$l(y|\theta) : \mathcal{R}^n \times \Theta \rightarrow \mathcal{R}^+$$

- distribution des paramètres

$$\pi(\theta) : \Theta \rightarrow \mathcal{R}^+$$

Distribution a posteriori

$$p(\theta|y) = \frac{l(y|\theta)\pi(\theta)}{\int_\Theta l(y|\theta)\pi(\theta)d\theta} \propto l(y|\theta)\pi(\theta).$$

La vraisemblance

Reformulation de la règle de Bayes

La densité a posteriori est proportionnelle à la vraisemblance multipliée par les densités a priori

- seul le noyau de la vraisemblance est utilisé
- des vraisemblances différentes peuvent conduire à une même loi a posteriori

Exemple

- **Experience A**: 20 épreuves de Bernoulli, où 7 succès sont observés
  Distribution jointe: Bernoulli

$$l_A(\theta; n = 20, s = 7) = C_{20}^{7}\theta^7(1-\theta)^{13}.$$

- **Experience B**: on répète une épreuves de Bernoulli jusqu'à 7 succès. Le dernier succès est observé à la 20ème répétition.
  Distribution jointe: Negbin

$$l_B(\theta; n = 20, s = 7) = C_{19}^{6}\theta^7(1-\theta)^{13}.$$

Le principe de vraisemblance

Le principe du critère d'arrêt

Principe de vraisemblance

des vraisemblances proportionnelles conduisent aux même inférences

$\Leftrightarrow$

Les expériences auraient put conduire à des données radicalement différentes! La règle d'arrêt n'intervient pas dans l'interprétation, l'information est uniquement dans ce qui observé.

Contraste fréquentistes vs bayésiens:

- fréquentiste: distribution d'échantillonnage
- bayésien: seules les données comptent, pas ce qui n'est pas observé

Pour aller plus loin: Berger et Wolpert (1988)

Comment choisir son a priori?

La distribution a priori résume des croyances sur θ

- il n'y a pas de règle générale de détermination des a priori
- intérêt à tester plusieurs a priori
- éviter les a priori ayant une probabilité nulle pour des ensembles de Θ
- préférence pour des a priori vagues
- on peut paramétrer les a priori avec des statistiques descriptives

A priori conjugués

A priori conjugué

Un a priori conjugué permet d'obtenir une loi a posteriori de la même famille que lui même

Intérêt:

- facilité analytique
- les a priori peuvent s'interpréter comme des a posteriori issus de travaux antérieurs
- en pratique, la vraisemblance dicte la forme de l'a priori conjugué
- les paramètres peuvent s'interpréter comme des données supplémentaires

Exemple

Epreuves de Bernouilli

$$l(\theta; n, s) \propto \theta^s (1-\theta)^{n-s}$$

Tout a priori proportionnel à $\theta^{a-1}(1-\theta)^{b-1}$ conduit à une loi a posteriori de même famille:

$$p(\theta|y) \propto \theta^{a+s-1}(1-\theta)^{b+n-(s+1)}$$

$\Rightarrow$ la distribution a priori conjuguée est une loi beta

Correspondances

Vraisemblance	A priori et a posteriori
binomial	beta
normale	normale
Poisson	gamma
uniforme	Pareto
Pareto	gamma

A priori diffus (impropres)

A priori impropre

La "distribution de probabilité" de θ est impropre si son intégrale sur Θ ne converge pas

$$\pi(\theta) \propto 1, -\infty < \theta < \infty.$$

Intérêt:

- ne déforment pas la vraisemblance
- loi limite traduisant des croyances très vagues
- des a priori impropres peuvent amener à une a posteriori propre
- souvent utilisés dans les phases exploratoires

A priori de Jeffrey (I)

Des propositions équivalentes devraient avoir les même probabilités: la loi a posteriori devrait être invariante à une reparamétrisation du modèle

A priori de Jeffrey

La loi a posteriori est invariante pour un a priori proportionnel à la racine carrée de la matrice d'information

A priori de Jeffrey (II)

Illustration:

- Modèle 1 formulé en terme de θ
  Loi a posteriori en suivant la règle de Jeffrey: $I(\theta) I_\theta^{1/2}$
- Modèle 1 formulé en terme de $\gamma = h(\theta)$
  On a $I_\gamma = I_\theta \left(\partial\theta/\partial\gamma\right)^2$
  Loi a posteriori en suivant la règle de Jeffrey:
  $I(h(\theta)) I_\gamma^{1/2} = I(\theta) I_\gamma^{1/2} |\partial\gamma/\partial\theta| = I(\theta) I_\theta^{1/2}$

A priori de Jeffrey (III)

Limites:

- application parfois complexes lorsque θ n'est pas scalaire
- amène souvent a des a priori impropres, cad informatifs pour des valeurs peu plausibles de θ

$\Rightarrow$ priori de Jeffrey est un "truc", par défaut on utilise plutôt des a priori impropres

A priori hiérarchiques

Soit $\theta = (\theta_1, \theta_2)$. Si θ_1 et θ_2 ont un rôle similaire, ils peuvent être tirés dans une même distribution $h(\theta|\lambda)$, et λ est tiré dans une distribution a priori de dimension généralement plus petite que θ.

Que reporter?

La loi a posteriori représente les croyances en θ sachant les croyances a priori et celles contenues dans la vraisemblance
On reporte généralement:

- les densité marginales a posteriori
- les moments de la loi a posteriori
- l'intervalle de plus forte densité a posteriori

La dominance de la vraisemblance

Dominance de la vraisemblance

Pour n grand, la loi a posteriori converge vers la vraisemblance.

Intuition:

- $\ln p(\theta|y) \propto \ln l(y|\theta) + \ln \pi(\theta)$
- une nouvelle observation y_{n+1} incrémente la log-vraisemblance de $\ln l(y_{n+1}|\theta)$, tandis que $\pi(\theta)$ reste constant
- pour n grand, le terme dominant est $\ln l(y|\theta)$

Condition: l'argument ne tient pas si $\pi(\theta) = 0$ pour des valeurs de θ où $l(y|\theta)$ concentre sa masse

Résultats asymptotiques

Pour n grand:

1. l'espérance a posteriori est proche de l'estimateur MV
2. la variance est approximativement donnée par l'inverse de l'information de Fisher
3. convergence et efficacité asymptotique ne dépendent pas des lois a priori si $\pi(\theta) > 0$ où $I(y|\theta)$ concentre sa masse
4. la loi a posteriori tend vers une loi normale

Référence: Gouriéroux et Monfort (1991)

Partie III

Méthodes de simulation

Pourquoi simuler?

Calculs ou simulations?

L'estimateur bayésien n'admet généralement pas de solution analytique
$\longrightarrow$ recours à des tirages dans la loi a posteriori

Plan de la partie

Echantillonnage dans la loi a posteriori

Intuition:

- soit la loi a posteriori $p(\theta_1, \theta_2|y)$
- on effectue `nrep` tirages dans $p(\theta_1, \theta_2|y)$ et récupère la matrice:

$$\begin{matrix} \theta_{1,1} & \theta_{2,1} \\ \theta_{1,2} & \theta_{2,2} \\ \vdots & \vdots \\ \theta_{1,nrep} & \theta_{2,nrep} \end{matrix}$$

- les colonnes contiennent des tirages dans les lois marginales de θ_1 et θ_2

Problème: Comment échantillonner dans $p(\theta|y)$?

Différentes méthodes de simulations

Principales méthodes d'échantillonnage:

- **tirages directs** si la loi a posteriori est "standard" (d'où l'utilisation d'a priori conjugués)
- **tirages successifs** si on ne peut pas échantillonner dans $f(x, y)$ mais facilement dans $f(x)$ et $f(y|x)$ (exemples: lois multivariés)
- **transformations de tirages de lois disponibles**: algorithmes d'Acceptation-Rejet, échantillonnage d'importance (voir Robert, 1996)
- **Méthodes de Monte Carlo par Chaînes de Markov** On construit un processus stochastique convergeant vers une loi stationnaire qui est la loi a posteriori.
  - il ne s'agit pas de tirages directs
  - les tirages sont identiquement distribués mais ne sont pas indépendants

Chaînes de Markov homogènes

On note (X_t) une suite de variables aléatoires à valeur dans $\mathcal{X}$ et $\mathcal{A}$ un sous-ensemble de $\mathcal{X}$.

Chaîne de Markov

La suite (X_t) forme une chaîne de Markov si:

$$Pr(X_{t+1} \in \mathcal{A} | x_0, x_1, \ldots, x_t) = Pr(X_{t+1} \in \mathcal{A} | x_t), \forall t \in \mathbb{N}$$

Homogénéité

Une chaîne est homogène si les probabilités de transition ne dépendent pas de t:

$$Pr(X_{m+1} \in \mathcal{A} | x_m) = Pr(X_{n+1} \in \mathcal{A} | x_n), \forall (m, n) \in \mathbb{N}^2$$

Noyeau de transition (I)

Cas général

Noyeau de transition

Un noyau de transition $K(x, A)$ est une fonction de $x \in \mathcal{X}$ et $A \in \mathcal{B}(\mathcal{X})$ telle que:

1. pour tout $x \in \mathcal{X}$, $K(x, .)$ est une mesure de probabilité
2. pour tout $A \in \mathcal{B}(\mathcal{X})$, $K(., A)$ est mesurable

Pour $\mathcal{X}$ continu, le noyau est la densité:

$$Pr(X \in A|y) = \int_{\mathcal{A}} K(x, y)dy.$$

Noyeau de transition (II)

Cas discret

Le noyau de transition d'une chaîne homogène (X_t) est la fonction:

$$K(x, y) = Pr(X_{t+1} = y | X_t = x), x, y \in \mathcal{X}.$$

Le noyau est une matrice de transition

Exemple:

		X_{t+1}	
		1	2
X_t	1	1 - α	α
	2	β	1 - β

Distribution d'état

La distribution de X_{t+1} peut s'écrire:

$$Pr(X_{t+1} = j) = \sum_{i=1}^{M} Pr(X_t = i) Pr(X_{t+1} = j | X_t = i),$$

où M est le nombre d'états et $j = 1, 2, \ldots, M$.
On note p_{t+1} la distribution de X_{t+1}:

$$p_{t+1}(j) = \sum_{i=1}^{M} p_t(i) K(i, j).$$

Sous forme matricielle:

$$p'_{t+1} = p'_t K.$$

Distribution stationnaire

Distribution stationnaire

La distribution p est stationnaire pour K si on a $p = pK$.

Cas discret:
$p = pK \Rightarrow (I - K')p' = 0$. Comme K a des éléments positifs ou nuls et ses lignes somment à un, elle admet (au moins) une valeur propre de 1. Le vecteur propre associée à la valeur propre unitaire, une fois normalisé à 1, est une distribution stationnaire.

Exemple de calculs de distributions stationnaires

- **Exemple A**:

  Pour $K = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$, on a $p' = \begin{pmatrix} \frac{\beta}{\alpha+\beta} \\ \frac{\alpha}{\alpha+\beta} \end{pmatrix}$

- **Exemple B**:

  Pour $K = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 1/2 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, les valeurs propres sont (1,1,0).

  Vecteurs propres $p'_1 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$; $p'_2 = \begin{pmatrix} 0.71 \\ 0.71 \\ 0 \end{pmatrix}$; $p'_3 = \begin{pmatrix} 0.71 \\ -0.71 \\ 0 \end{pmatrix}$

  Il y a deux distributions stationnaires, obtenues en normalisant p'_1 et p'_2 pour que chacun somme à 1.

Irréductibilité

On note $p_{ij}(m) = Pr(X_{t+m} = j | X_t = i)$.
Les états i et j communiquent s'il existe un $m > 0$ tel que $p_{ij}(m) > 0$ et $p_{ji}(m) > 0$. Ils forment une **classe communicante**.

Irréductibilité

Une chaîne est irréductible ssi elle admet une classe communicante

Existence et unicité de la loi stationnaire ($\mathcal{X}$ discret)

Théorème

Un chaîne de Markov irréductible et finie admet une unique distribution stationnaire

Intuition:

1. une des valeurs propre est positive et supérieure (en valeur absolue) aux autres valeurs propres
2. un vecteur propre positif lui correspond
3. cette valeur propre est une racine simple de l'équation caractéristique de K

Extension à $\mathcal{X}$ dénombrable ou continu

Besoin d'hypothèses additionnelles!

Temps d'atteinte

Le temps d'atteinte de l'état i est $T_i = \inf\{n \geq 1; X_n = i\}$

Récurrence

Une état est récurrent si $Pr(T_i < \infty) = 1$

Récurrence positive

Un état est positif récurrent si $\mathrm{E}(T_i) < \infty$

Récurrence au sens de Harris

L'état A est récurrent au sens de Harris si $Pr(\eta_A = \infty | X_0 = x) = 1$, où η_A est le nombre de passage en A et $x \in \mathcal{X}$

Existence et unicité de la loi stationnaire

$\mathcal{X}$ dénombrable

Un chaîne de Markov irréductible et récurrente positive admet une unique distribution stationnaire

$\mathcal{X}$ continu

Un chaîne de Markov irréductible et récurrente au sens de Harris admet une unique distribution stationnaire

Preuves: cf Robert (1996)

Convergence

Apériodicité

Un état i est apériodique si $p_{ii} > 0$ pour tout n suffisamment grand

Convergence pour $\mathcal{X}$ dénombrable

Si une chaîne est irréductible, apériodique et de distribution stationnaire p, alors:

$$Pr(X_n = j) \overset{n \to \infty}{\to} p_j, \forall j,$$

indépendamment de p_0.

Preuve: cf Robert (1996)

Ergodicité

Ergodicité: $\mathcal{X}$ dénombrable

Une chaîne récurrente positive, quelle que soit sa distribution initiale p_0 et de distribution stationnaire p est telle que:

$$Pr\left(\frac{1}{n}\sum_{k=0}^{n-1} f(X_k) \overset{n\to\infty}{\to} \mathrm{E}(f)\right) = 1,$$

où f est bornée.

Des résultats similaires existent pour $\mathcal{X}$ continu
Preuves: cf Robert (1996)

Résumé

On cherche un noyau de transition (donc une chaîne de Markov) tel qu'il:

1. admette une unique distribution stationnaire
2. converge vers la distribution stationnaire
3. le théorème ergodique s'applique

Comment construire un tel noyau?
$\Rightarrow$ Echantillonnage de Gibbs et méthode de Metropolis-Hastings

Echantillonnage de Gibbs

Intuition:
Soient $x = (x_1, x_2)$ et $y = (y_1, y_2)$.
On peut construire le noyau:

$$K(x, y) = p_{Y_1|Y_2}(y_1|x_2)p_{Y_2|Y_1}(y_2|y_1)$$

Et:

$$\begin{aligned}\int K(x, y)p(x)dx &= \int p_{Y_1|Y_2}(y_1|x_2)p_{Y_2|Y_1}(y_2|y_1)p_{Y_1,Y_2}(x_1, x_2)dx_1dx_2 \\ &= p_{Y_2|Y_1}(y_2|y_1)\int p_{Y_1,Y_2}(y_1, x_2)dx_2 \\ &= p_{Y_1,Y_2}(y_1, y_2)\end{aligned}$$

Le noyau ainsi défini admet p comme distribution stationnaire

Algorithme d'échantillonnage de Gibbs (I)

Algorithme

1. tirer y_1 dans la loi de Y_1 sachant $Y_2 = x_2$
2. tirer y_2 dans la loi de Y_2 sachant $Y_1 = y_1$
3. répéter les étapes 1 et 2

Exemple

échantillonnage dans une loi normale bivariée (I)

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

Algorithme

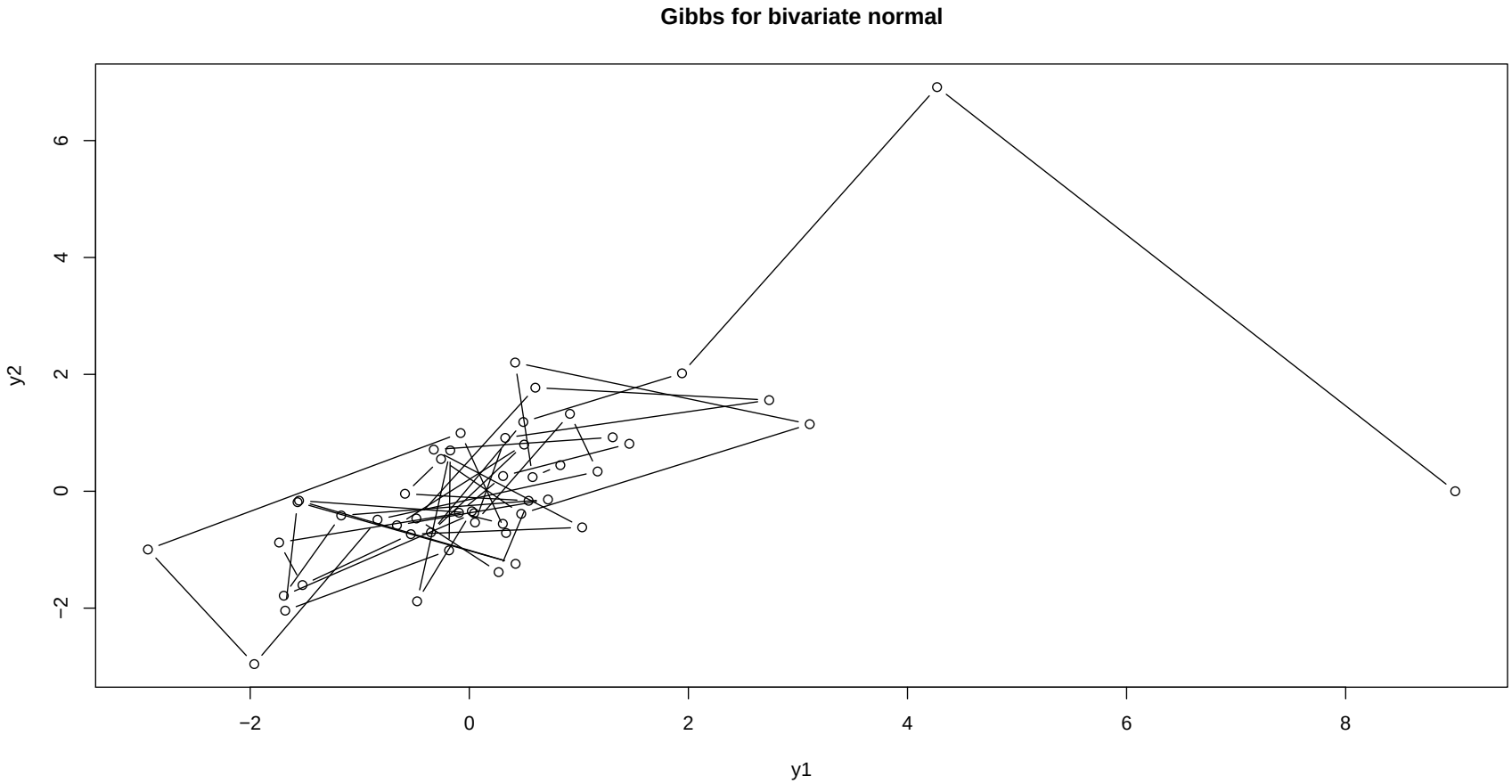
On choisit une valeur initiale $y_2 = y_2^{(0)}$ puis:

1. tirer Y_1 dans $\mathcal{N}(\rho y_2, 1 - \rho^2)$
2. tirer Y_2 dans $\mathcal{N}(\rho y_1, 1 - \rho^2)$
3. répéter les étapes 2 et 3

Exemple

échantillonnage dans une loi normale bivariée (II)

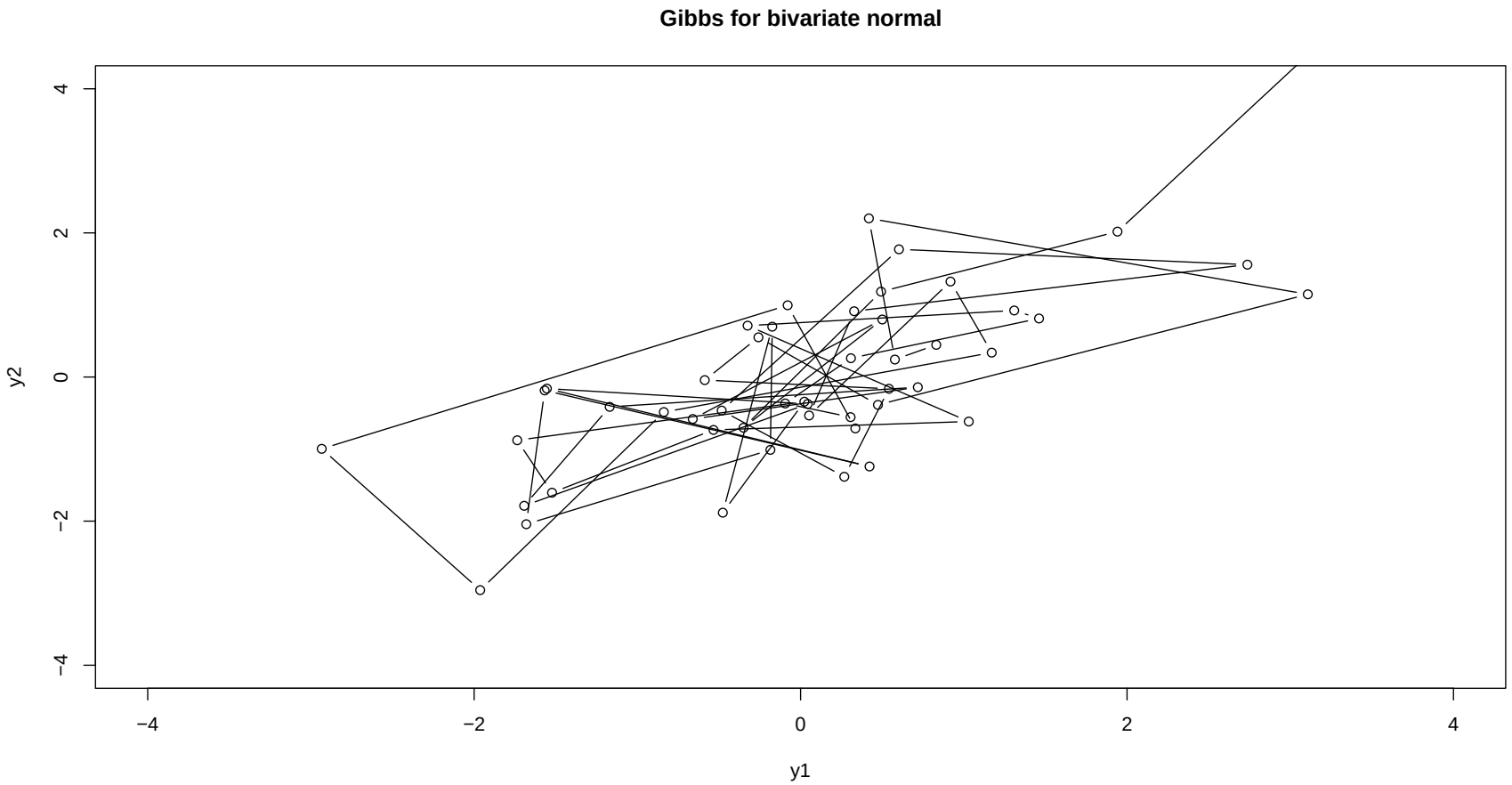
40 tirages



Exemple

échantillonnage dans une loi normale bivariée (II)

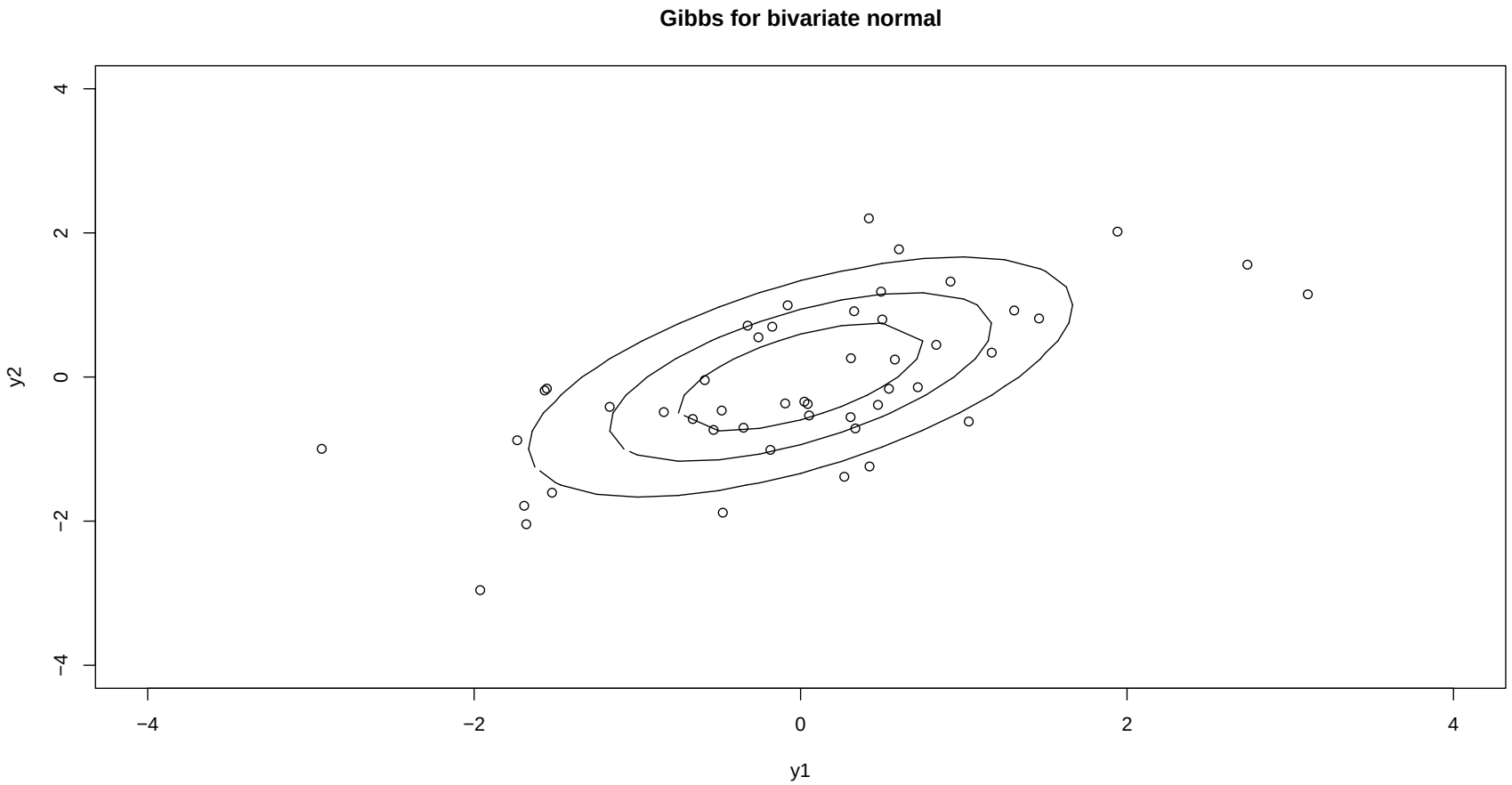
40 tirages



Exemple

échantillonnage dans une loi normale bivariée (II)

40 tirages



Algorithme d'échantillonnage de Gibbs (II)

Algorithme général

Soit $\theta = (\theta_1, \theta_2, \ldots, \theta_q)$. L'algorithme est initialisé à l'itération (k) en $\theta^{(k)} = (\theta_1^{(k)}, \theta_2^{(k)}, \ldots, \theta_q^{(k)})$ et se déroule comme suit:

1. tirer $\theta_1^{(k+1)}$ dans $p(\theta_1 | \theta_2^{(k)}, \theta_3^{(k)}, \ldots, \theta_q^{(k)}, Y)$
2. tirer $\theta_2^{(k+1)}$ dans $p(\theta_2 | \theta_1^{(k+1)}, \theta_3^{(k)}, \ldots, \theta_q^{(k)}, Y)$

$\vdots$

3. tirer $\theta_q^{(k+1)}$ dans $p(\theta_q | \theta_1^{(k+1)}, \theta_2^{(k+1)}, \ldots, \theta_{q-1}^{(k+1)}, Y)$
4. répéter les étapes de 1 à 3

Liens avec l'algorithme de Métropolis-Hastings

Echantillonnage de Gibbs:

- cas particulier de la méthode de Métropolis-Hastings
- nécessite d'appeler d'autres méthodes de simulation si on ne peut pas tirer directement dans les lois marginales conditionnelles

Algorithme de Metropolis (I)

$p(y)$: distribution objectif
$q(y|x)$: distribution de passage symétrique ($q(y|x) = q(x|y)$).

Algorithme de Metropolis

On choisit une valeur initiale $y^{(0)}$. L'itération (k) se déroule comme suit:

1. tirer y_c dans $q(.|y^{(k)})$
2. évaluer:

$$r = \frac{p(y_c)q(y^{(k)}|y_c)}{p(y^{(k)})q(y_c|y^{(k)})}$$

Algorithme de Metropolis (II)

Algorithme de Metropolis (suite)

3

$$r \begin{cases} \geq 1, & y^{(k+1)} = y_c \\ < 1, & \begin{cases} y^{(k+1)} = y_c & \text{avec probabilité } r \\ y^{(k+1)} = y^{(k)} & \text{avec probabilité } 1 - r \end{cases} \end{cases}$$

4 répéter les étapes de 1 à 3

Algorithme de Métropolis-Hastings (I)

La probabilité que y_c soit accepté est:

$$\min\left(\frac{p(y_c)q(y^{(k)}|y_c)}{p(y^{(k)})q(y_c|y^{(k)})}, 1\right).$$

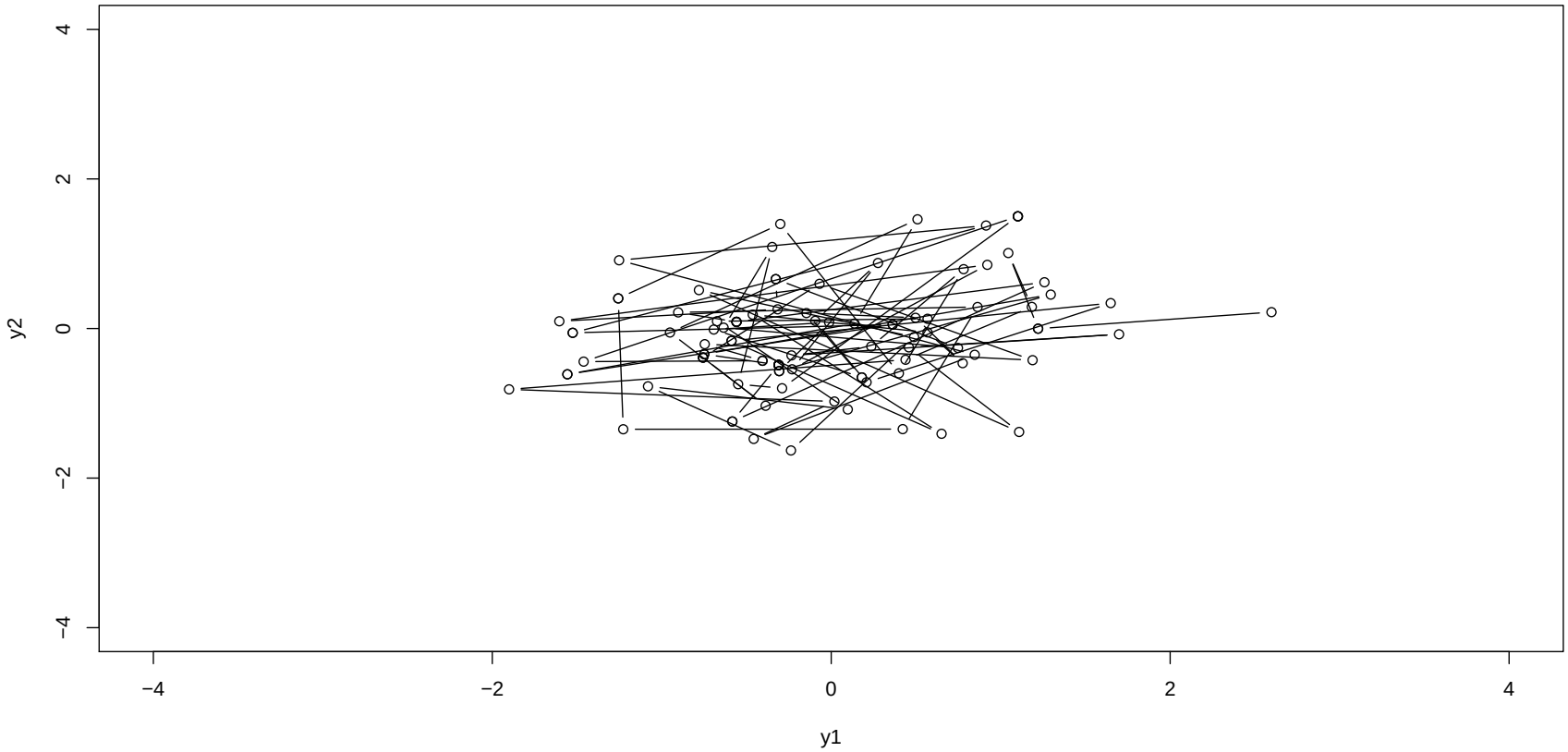
Intuition:

- si la probabilité a posteriori est plus importante en y_c que $y^{(k)}$, on accepte y_c
- sinon, on accepte néanmoins parfois y_c pour échapper à un extremum local

Exemple

échantillonnage dans une loi normale bivariée

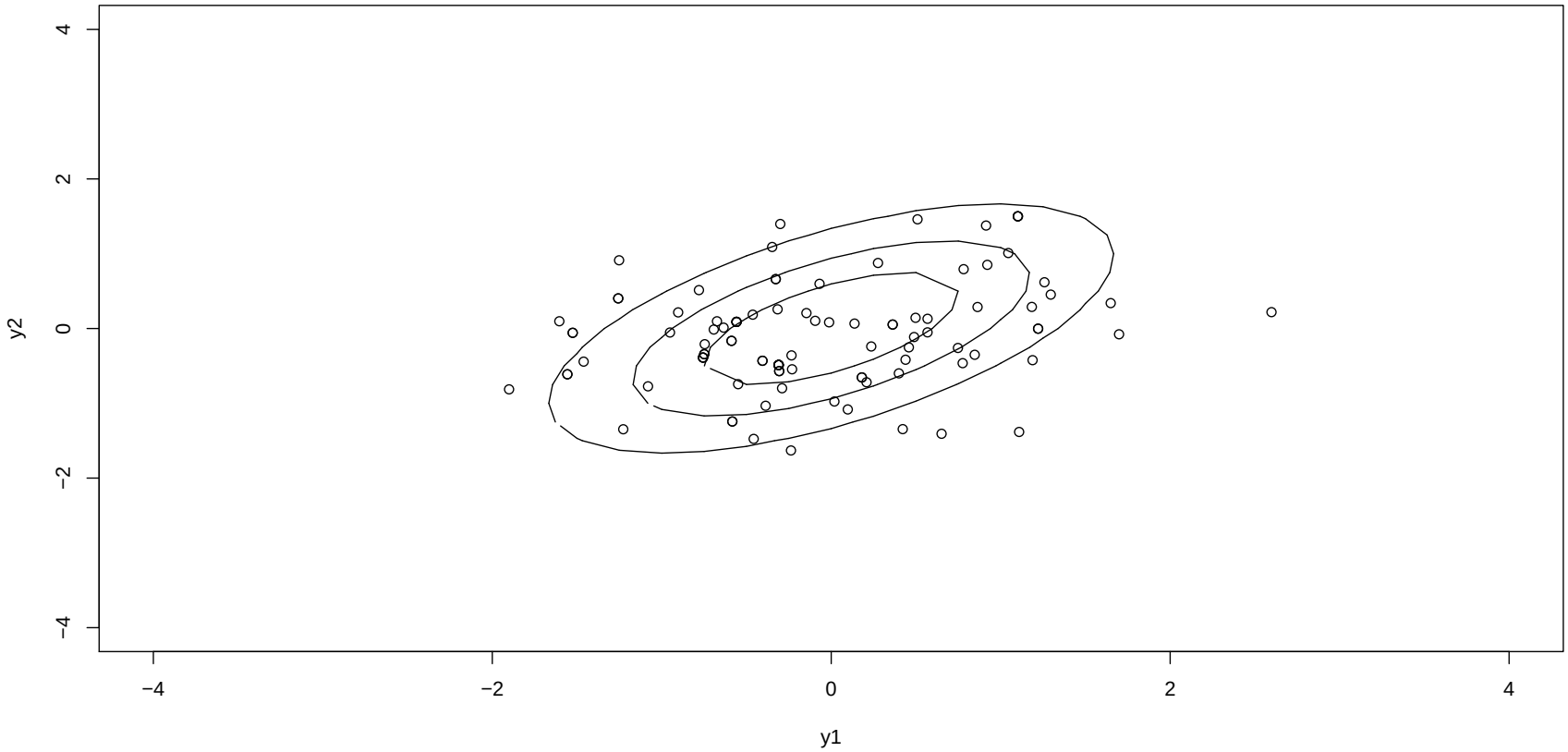
40 tirages



Exemple

échantillonnage dans une loi normale bivariée

40 tirages



C'est encore loin la convergence?

statistique de Gelman et Rubin

Problème: les méthodes MCMC produisent des échantillons et non pas un critère optimisé

Intuition: simuler plusieurs chaînes et comparer les variances intra- et inter-chaînes. Si la convergence est acquise, elles doivent être proches.

Statistique de Gelman et Rubin (I)

Variances intra- et inter-chaînes

Soit ω_{ij} le i-ème $(i = 1, \ldots, n)$ élément de la chaîne j $(j = 1, \ldots, m)$.
Variance inter-chaînes:

$$B = \frac{n}{m-1} \sum_{j=1}^{m} (\overline{\omega_j} - \overline{\omega})^2,$$

où $\overline{\omega_j}$ est la moyenne des n réalisations de la chaîne j et $\overline{\omega}$ la moyenne des mn réalisations de toutes les chaînes. Variance intra-chaînes:

$$W = \frac{1}{m(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (\omega_{ij} - \overline{\omega_j})^2,$$

Statistique de Gelman et Rubin (II)

B et W sont des estimateurs convergents de la variance de ω:

$$\widehat{\sigma}_{\omega}^{2} = (1 - 1/n)W + (1/N)B.$$

D'où la statistique de diagnostique:

$$R = \sqrt{\frac{\widehat{\sigma}_{\omega}^{2}}{W}} \overset{n \to \infty}{\to} 1.$$

Remarques sur la convergence

Généralement:

- utiliser plusieurs chaînes aux valeurs initiales surdispersées
- surveiller l'autocorrélation des éléments de chaque chaîne
- surveiller la forme des lois marginales a posteriori (multimodalité)

Utilisation de l'échantillon simulé

- on peut utiliser les tirages de θ_2 pour calculer différentes statistiques...
- ou on peut utiliser l'information véhiculée par les tirages de θ_1 sur la distribution de θ_2: **Rao-Blackwellisation**

Rao-Blackwellisation

On a: $\mathrm{V}(Y) \geq \mathrm{V}(\mathrm{E}(Y|X))$
Or:

$$\begin{aligned} p_2(y) &= \int p(\theta_1, y) d\theta_1 \\ &= \mathrm{E}\left[p_{2|1}(y|\theta_1)\right]. \end{aligned}$$

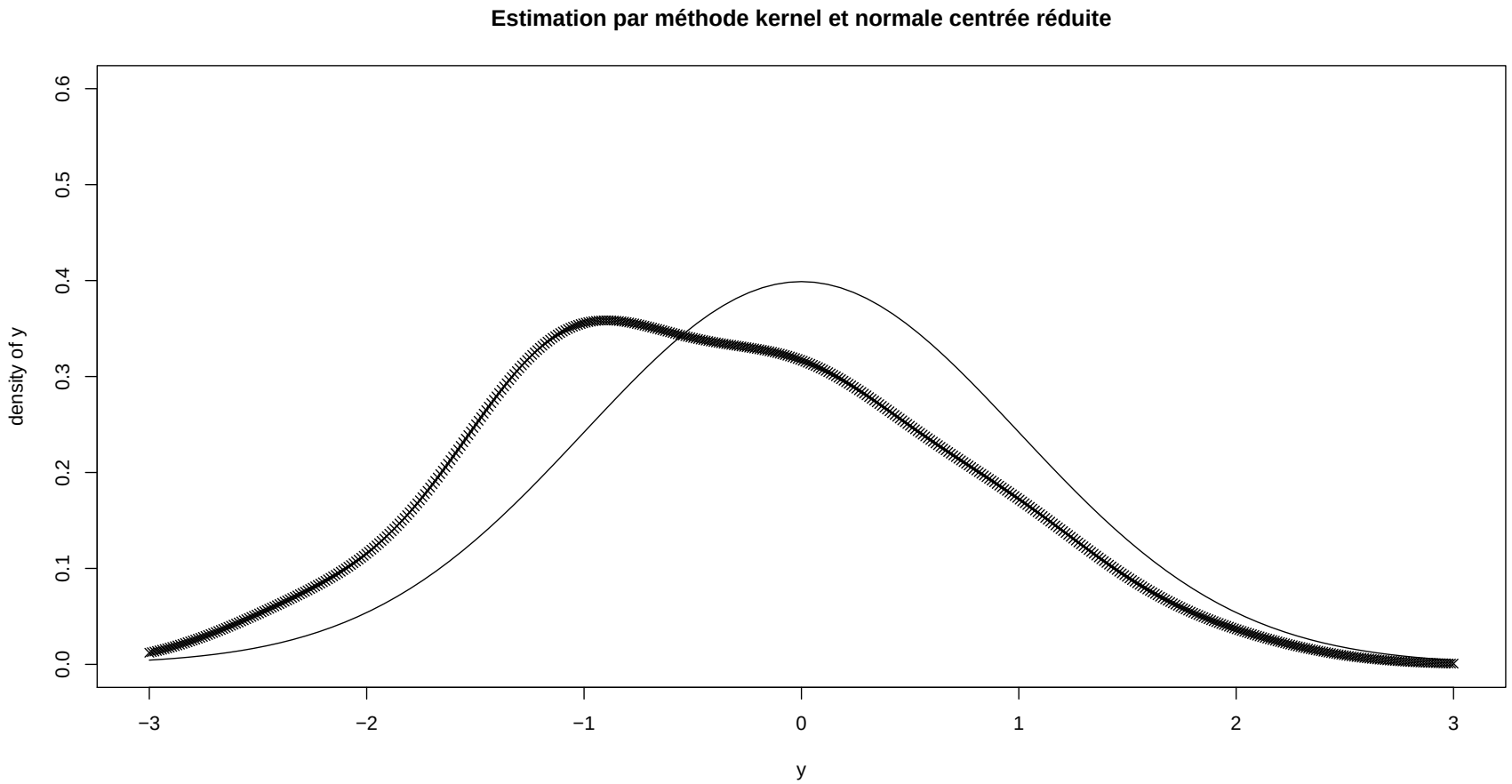
D'où:

$$\widehat{p}_2(y) = \frac{1}{n} \sum_{i=1}^{n} p_{2|1}(y|\theta_{1i}).$$

Exemple

Rao-Blackwellisation d'un échantillon d'une loi normale bivariée

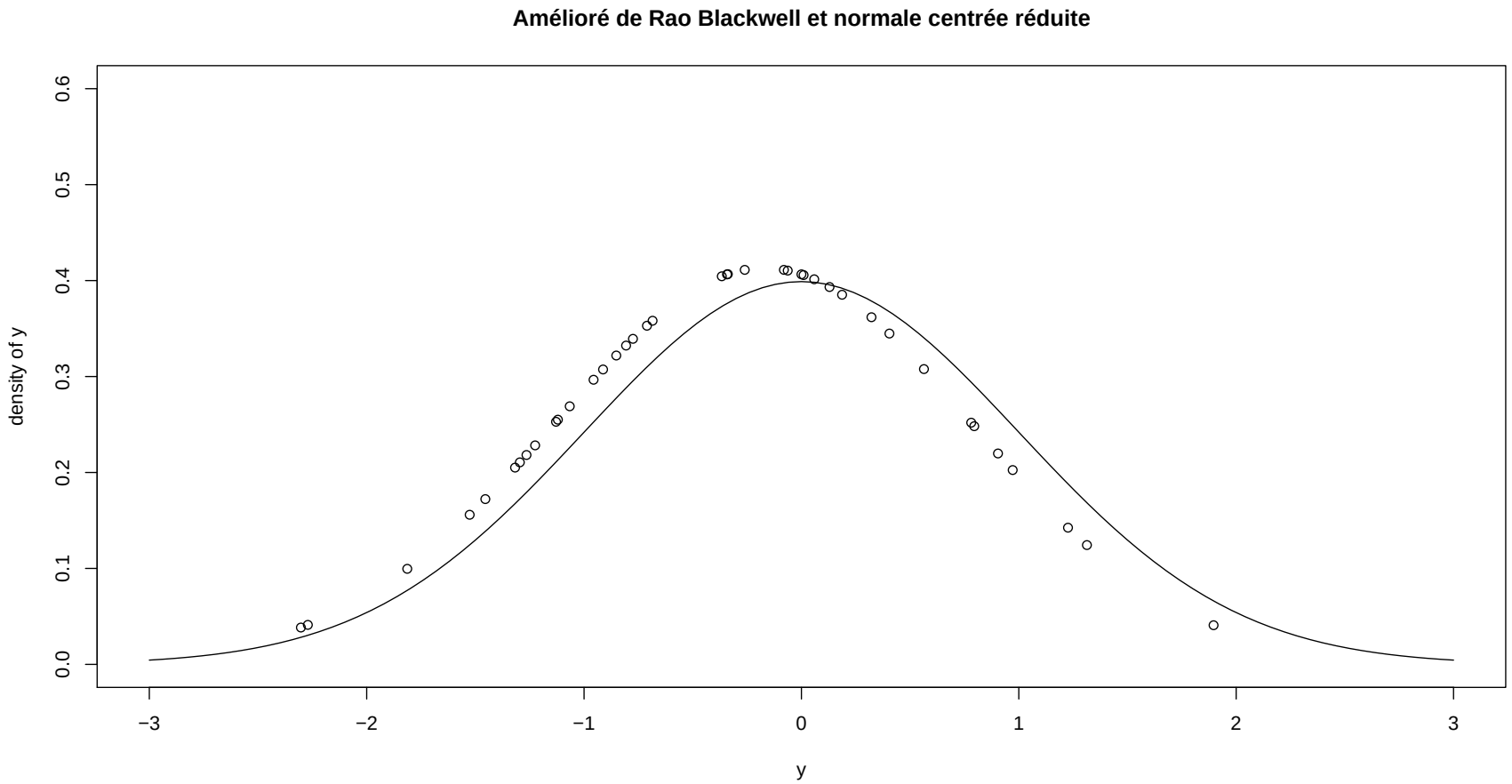
40 tirages



Exemple

Rao-Blackwellisation d'un échantillon d'une loi normale bivariée

40 tirages



Partie IV

Comparaison de modèles

Le mystère de la foi

Comment éprouver les croyances?

Typologie des croyances

$$\text{Croyances} \begin{cases} \text{dogmatiques} & \text{(le modèle est linéaire, erreurs iid, } \ldots\text{)} \\ \text{non-dogmatiques} & \text{(les a priori)} \end{cases}$$

Révision des croyances

- dogmatiques : étudiant les données, le phénomène, spécifications alternatives
- non-dogmatiques: avec le théorème de Bayes

Plan de la partie

Distributions des prévisions

Vérifier un modèle revient à comparer des prévisions avec les données
Deux types de distributions des prévisions:

- distribution a priori des prévisions
- distribution a posteriori des prévisions

La distribution a priori des prévisions

$$p(y) = \int l(y|\theta)\pi(\theta)d\theta.$$

$p(y)$ est le dénominateur de la formule de Bayes

Algorithme

1. tirer θ dans $\pi(\theta)$
2. tirer y dans $l(y|\theta)$
3. répéter les étapes 1 et 2

Utilisation de la distribution a priori des prévisions (I)

Soient Y^{obs} le vecteur des n observations et $T()$ une fonction scalaire

Algorithme

1. tirer Y dans la distribution a priori des prévisions
2. évaluer $T(Y, \theta) - T(Y^{\text{obs}}, \theta)$

Si la distribution empirique des réalisations de $T(Y, \theta) - T(Y^{\text{obs}}, \theta)$ attribue une faible probabilité à 0, les données ne sont pas en adéquation avec les croyances que $T()$ permet de tester.

Utilisation de la distribution a priori des prévisions (II)

Cas des a priori impropres

Un a priori impropre peut conduire à une distribution a priori des prévisions impropre

Exemple: $y \sim \mathcal{N}(\mu, 1)$ et $\pi(\mu) \propto 1$.

$$p(y) \propto \int_{-\infty}^{\infty} \exp\left[(-1/2)(y-\mu)^2\right] d\mu = \sqrt{2\pi}.$$

Solution:

1. utiliser une partie de l'échantillon pour construire un a priori propre
2. effectuer des prévisions
3. comparer avec le reste de l'échantillon

La distribution a posteriori des prévisions

Soient y^{obs} les observations et $\widetilde{y}$ un nouvel échantillon

$$p(\widetilde{y}|y^{\text{obs}}) = \int p(\widetilde{y}|y^{\text{obs}}, \theta) p(\theta|y^{\text{obs}}) d\theta.$$

Algorithme

1. tirer θ dans la loi a posteriori
2. tirer $\widetilde{y}$ dans $p(\widetilde{y}|y^{\text{obs}}, \theta)$
3. répéter les étapes 1 et 2

Utilisation de la distribution a posteriori des prévisions

Algorithme

1. tirer $\widetilde{Y}$ dans la distribution a posteriori des prévisions
2. évaluer $T(\widetilde{Y}, \theta) - T(Y^{\text{obs}}, \theta)$

Si la distribution empirique des réalisations de $T(Y, \theta) - T(Y^{\text{obs}}, \theta)$ attribue une faible probabilité à 0, les données ne sont pas en adéquation avec les croyances que $T()$ permet de tester.

Approche bayésienne du choix de modèle

Comment choisir entre plusieurs modèles?

Pour chaque modèle:

1. calculer la probabilité que les données en soient issues
2. calculer les probabilités a priori (si possibles égales pour chaque modèle)
3. déduire des étapes 1-2 la probabilité a posteriori
4. comparer les probabilités a posteriori

Rapport de probabilités

On considère M_j modèles ($j = 1, \ldots, J$), chacun doté des probabilités a priori π_j.
Probabilité a posteriori de chaque modèle:

$$p(M_j|y) \propto \frac{p(y|M_j)\pi_j}{p(y)}.$$

Rapport de probabilités des modèles i et j:

$$\frac{p(M_i|y)}{p(M_j|y)} = \frac{p(y|M_i)}{p(y|M_j)} \frac{p(M_i)}{p(M_j)},$$

où $\frac{p(y|M_i)}{p(y|M_j)}$ est le **facteur de Bayes**

Approximation du facteur de Bayes

Calculer le facteur de Bayes revient à évaluer la **densité des prévisions a priori**:

$$p(y|M_j) = \int p(y|\theta_j, M_j)p(\theta_j)d\theta_j.$$

une approximation a été développée pour éviter cela: **Bayesian Information Criteria** (BIC, Schwarz, 1978, *Ann Stat*)

Le BIC

BIC

$$\frac{p(y|M_i)}{p(y|M_j)} \approx \frac{l_i(\widehat{\theta}_i; y)}{l_j(\widehat{\theta}_j; y)} n^{(q_j - q_i)/2},$$

où q_i (respectivement q_j) est le nombre de paramètre du modèle i (resp. j).

principe de parcimonie: le second terme pénalise le modèle avec le plus de paramètres

Le DIC

DIC (Spiegelhalter *et al.*, 2002, *JRSS*):

- extension du BIC
- autorise des a priori impropres

DIC

$$\text{DIC} = \text{E}_{\theta|y}\left[-2\ln l(y|\theta)\right] + 2\ln p\left(y|\text{E}(\theta|y)\right).$$

p_D = "espérance a posteriori de la deviance"
− "deviance des espérances a posteriori".

Le modèle avec le plus petit DIC est préféré

Partie V

Références

Références transversales

Jackman: *Bayesian modeling in the social Sciences: an introduction to Markov Chain Monte Carlo*, notes de cours.
Lancaster (2004): *An introduction to modern bayesian econometrics*, Blackwell Publishing.
Robert (1996): *Méthodes de Monte Carlo par chaînes de Markov*, Economica.
Robert (2006): *Theory of probability revisited: A reassessment of a Bayesian classic*, notes de cours.

Bibliographie

Berger et Wolpert (1988): *The Likelihood Principle*, IMS.
Clifford (1993): Discussion on the meeting on the Gibs sampler and other Markov chain Monte Carlo methods, *Journal of the Royal Statistical Society*, Series B, 55, 53-102.
Kyburg et Smokler (1980): *Studies in subjective probability*, Huntington.
di Finetti (1974): *Theory of probability*, Wiley.
Gourieroux et Monfort (1991): *Statistique et modèles économétriques*, Economica.
Schwarz (1978): Estimating the dimension of a model, *The Annals of Statistics*, 6, 461-464.
Spiegelhalter, Best, Carlin, van der Linde (2002): Bayesian measures of a model complexity and fit, *Journal of the Royal Statistical Society*, Series B, 64, 583-639.

Merci de votre attention!

guillaume.horny@banque-france.fr