Cours 4 : Diagnostic de Convergence

- 1) Principes généraux
- 2) Convergence vers la loi stationnaire
  - Méthodes graphiques
  - Distance à la loi stationnaire
  - Autres méthodes
- 3) Convergence des moyennes
  - Méthodes graphiques (CUSUM, ...)
  - Variance intra/inter chaînes (Gelman et Rubin, 1992)

Principes généraux

On doit en pratique régler deux problèmes

- Comment doit-on règler le nombre d'itérations de chauffage (burn-in) nécessaire pour que $\theta^{(t)}$ soit distribué suivant la loi cible ?
- Quand doit-on arrêter l'algorithme pour que les données générées permettent d'avoir une bonne estimation des paramètres inconnus ?

Types de convergence

- Convergence vers la loi stationnaire
- Convergence des moyennes empiriques

$$\frac{1}{T}\sum_{t=1}^{T} h\left(\theta^{(t)}\right) \to E_f[h(\theta)]$$

  Quelle valeur de T doit-on choisir ? (Convergence importante pour l'estimateur MMSE).

- Indépendance entre les valeurs simulées

Une ou plusieurs chaînes ?

M chaînes indépendantes en parallèle $\left(\theta_m^{(t)}\right)$, $m = 1, ..., M$ ou une seule chaîne ?

- Motivations pour la simulation de chaînes en parallèle
  - Dépendance aux valeurs initiales de la chaîne réduite
  - On obtient différentes estimations des paramètres
  - You've only been where you have been
- mais
  - Convergence gouvernée par la chaîne la plus lente
  - Comparer des chaînes de vitesses de cv différentes
  - Loi initiale basée sur des infos partielles sur la loi cible

Le débat "une seule chaîne" contre "plusieurs chaînes en parallèle" est loin d'être clos!!

Convergence vers la loi cible

- Méthodes graphiques
  - L'idée la plus simple est de représenter la valeurs des éléments de la chaîne $\theta_m^{(t)}$ en fonction de t pour plusieurs chaînes $\Rightarrow$ très utile pour détecter des non-stationarités fortes
  - Évaluation d'une distance entre la loi obtenue à l'itération k et la loi cible (obtenue avec un grand nombre d'itérations)

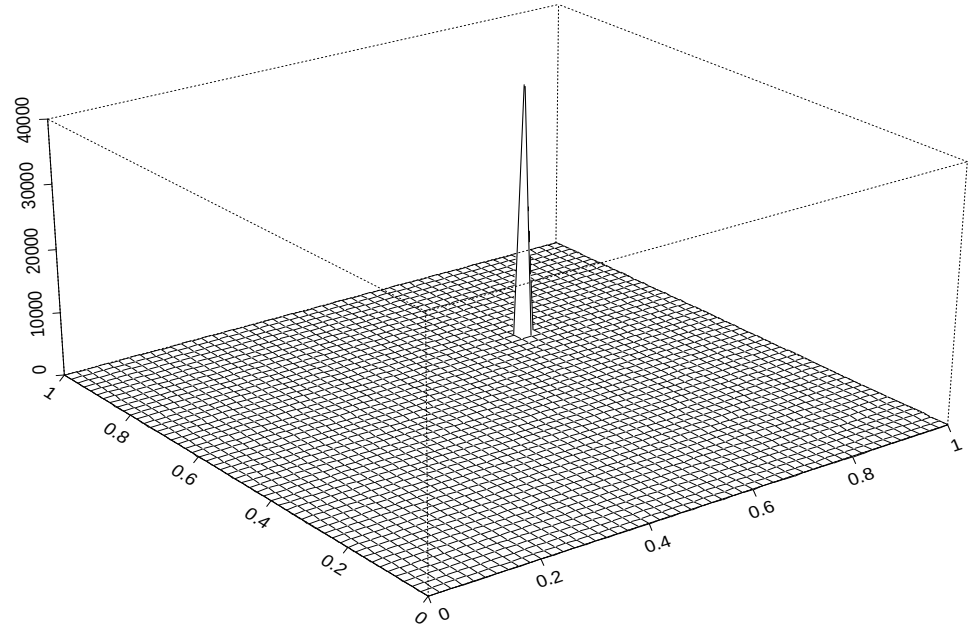
La distribution du chapeau de sorcière

- Un exemple classique

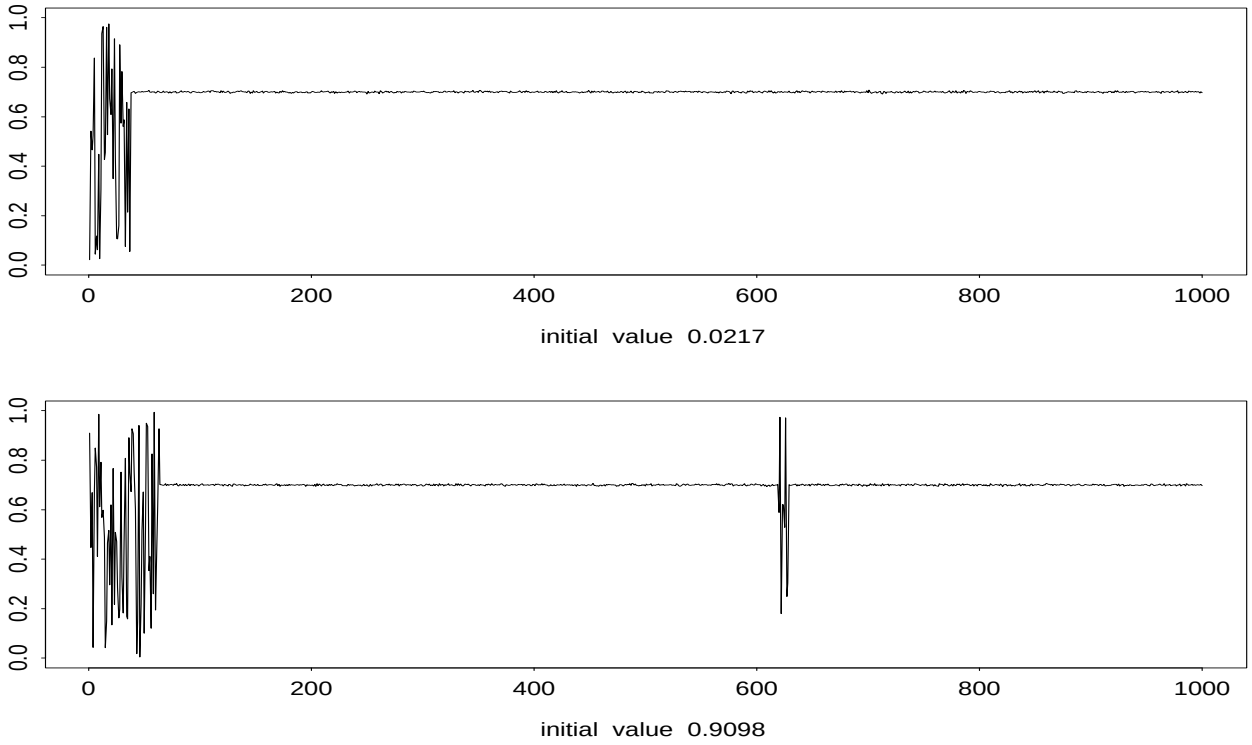
$$\pi(\theta|y) \propto \left\{(1-\delta)\sigma^{-d}e^{-\frac{\|\theta-y\|^2}{2\sigma^2}} + \delta\right\} \mathbb{I}_C(\theta),\ y \in \mathbb{R}^d,\ \theta \in [0,1]^d$$

Un mode très concentré autour de y pour δ et σ “petits".

Monte Carlo Statistical Methods/October 29, 2001 286



Éléments de la chaîne

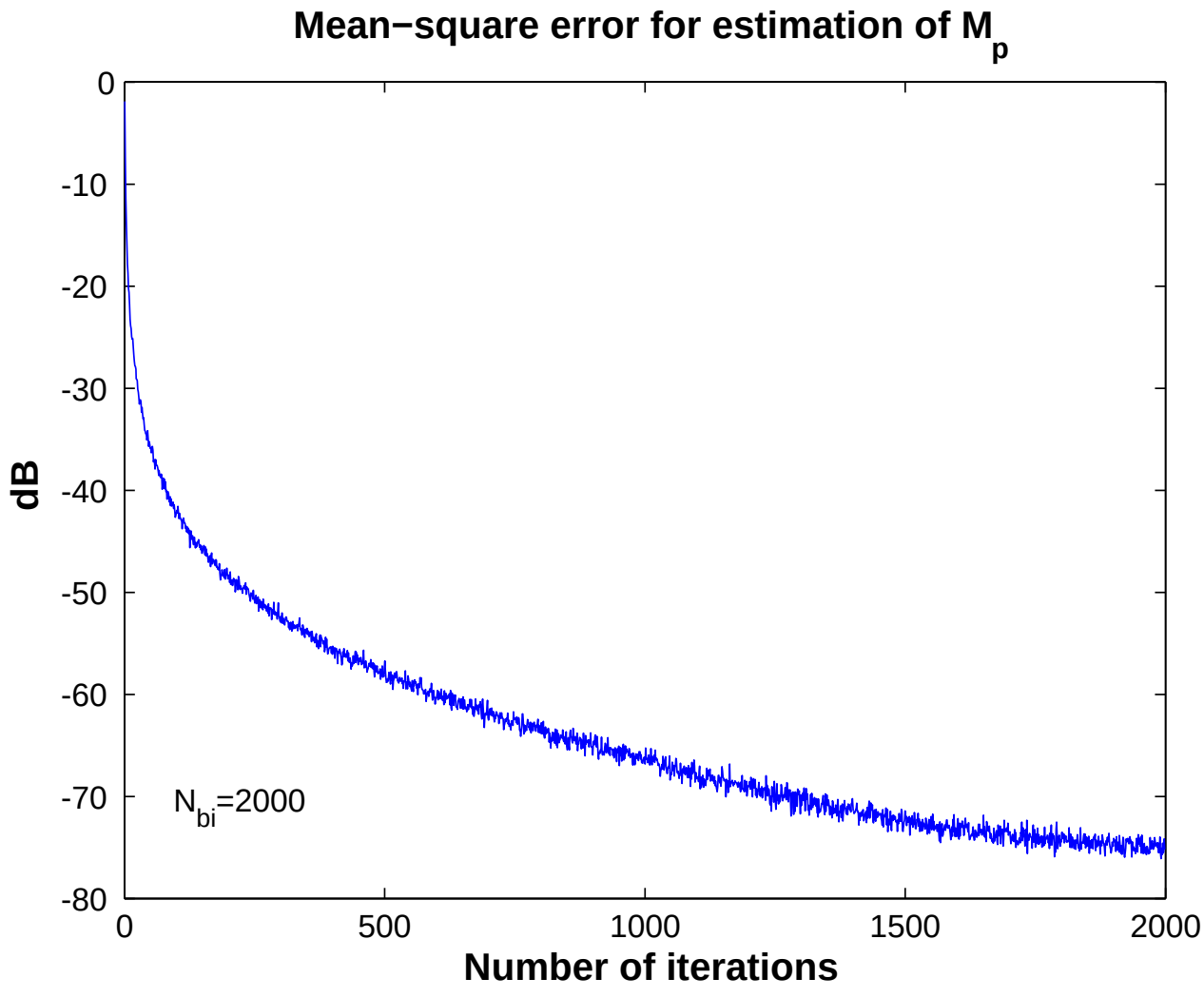


Chain $(\theta_1^{(t)})$ for two initial values, 0.0217 *(top)* **and** 0.9098 *(bottom)*

Distance à la loi cible

- Principes
  - On choisit une distance entre lois de probabilités
  - On fait tourner l'algorithme avec un grand nombre d'itérations $\Rightarrow$ obtention d'une loi de référence
  - On calcule la distance entre la loi estimée à l'instant t et la loi de référence

Exemple


Mean−square error for estimation of M_p
0
-10
-20
-30
-40
-50
-60
-70
-80
dB
N_bi=2000
0
500
1000
1500
2000
Number of iterations

Distance en ligne avec plusieurs chaînes

- Estimation de la distance entre f et $f^{(t)}$ en ligne, où $f^{(t)}$ est la loi marginale de $\theta^{(t)}$ et f est la loi cible.

$$\|f - f^{(t)}\| \simeq -1 + \frac{1}{M(M-1)} \sum_{1 \leq l \neq s \leq M} \frac{K_-\left(\widetilde{\theta}_l^{(0)}, \theta_s^{(t)}\right)}{f(\theta_s^{(t)})},$$

où $\widetilde{\theta}^{(t)}$ est obtenue à l'aide d'un échantillonneur de Gibbs construit à partir des lois conditionnelles $f_k, ..., f_1$ et K_- est le noyau de transition de cette nouvelle chaîne

- Problèmes
  - on doit construire deux échantillonneurs de Gibbs
  - Calcul de la cste de normalisation de K_- peut être coûteux

Contrôle binaire de Raftery et Lewis (1992)

Idée : tester certains quantiles de la loi a posteriori $P[U < u|\text{Données}]$, où U est une fonction du vecteur paramètre inconnu θ (e.g. $U = \theta$ ou $U = |\theta|$ en dimension 1).

- Indicatrices

$$Z_t = \begin{cases} 1 \text{ si } U_t < u, \\ 0 \text{ sinon} \end{cases}$$

- Sous-Chaîne

$$Z_t^{(k)} = Z_{1+(t-1)k}$$

- Quantile

$$q = P[U < u|\text{Données}] \text{ (e.g. } q = 0.025)$$

Rq: u (associé à $q = 0.025$) sera estimé à partir d'une chaîne "pilote". On pourra tester plusieurs valeurs de u et garder le max des burn-in (Brooks, Roberts, 1999)

Détermination de k

- Propriété

  La sous-chaîne $Z_t^{(k)}$ est une chaîne de Markov d'ordre 1 asymptotiquement (k "grand")

- Choix de k

  On prend la valeur de k la plus petite donnant une préférence à un modèle de Markov d'ordre 1 par rapport à un modèle de Markov d'ordre 2 à l'aide d'un test d'hypothèses

$$T = \text{LRT} - 2\log n,$$

Nombre d'itérations de chauffage n_0

- Matrice de transition de $Z_t^{(k)}$: $\begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix}$
- Matrice de transition après l itérations

$$\begin{pmatrix} \pi_0 & \pi_1 \\ \pi_0 & \pi_1 \end{pmatrix} + \frac{\lambda^l}{\alpha+\beta} \begin{pmatrix} \alpha & -\alpha \\ -\beta & \beta \end{pmatrix},$$

avec $\pi_0 = \beta/(\alpha+\beta)$, $\lambda = 1-\alpha-\beta$ et $\pi_1 = 1-\pi_0$.

- Condition $|P[Z_m^{(k)} = i | Z_0^{(k)} = j] - \pi_i| < \epsilon$ (e.g. $\epsilon = 0.0125$):

$$\lambda^m < \frac{(\alpha+\beta)\epsilon}{\max(\alpha,\beta)} \Rightarrow m = m^\star = \frac{\log\left[\frac{(\alpha+\beta)\epsilon}{\max(\alpha,\beta)}\right]}{\log(\lambda)} \Rightarrow n_0 = km^\star$$

Rq: α_u et β_u estimés à l'aide de la chaîne "pilote"

Nombre d'itérations de calcul N

- Estimation du quantile :

$$\overline{Z}_n^{(k)} = \frac{1}{n} \sum_{t=1}^{n} Z_t^{(k)}$$

- Théorème de la limite centrale

$$\frac{\overline{Z}_n^{(k)} - q}{\nu^2/n} \underset{n\to\infty}{\to} \mathcal{N}(0,1), \quad \nu^2 = \frac{\alpha\beta(2-\alpha-\beta)}{(\alpha+\beta)^3}$$

- Condition $P\left[\left|\overline{Z}_n^{(k)} - q\right| < r\right] = s$ (e.g. $s = 0.95$ and $r = 0.0125$):

$$n = n^\star = \frac{(2-\alpha-\beta)\alpha\beta}{(\alpha+\beta)^3} \left\{ \frac{\Phi^{-1}\left(\frac{1}{2}(s+1)\right)}{r} \right\}^2 \Rightarrow N = kn^\star$$

Autres méthodes

- Tests non paramétriques de stationarité
  - Tests standards (Kolmogorov-Smirnov, ...)
  - Lorsque la chaîne est stationnaire, $\theta^{(t_1)}$ et $\theta^{(t_2)}$ ont la même loi pour tout couple (t_1, t_2)

Convergence des moyennes

On cherche les valeurs de T telles que

$$\frac{1}{T}\sum_{t=n_0+1}^{n_0+T} h\left(\theta^{(t)}\right) \simeq E_f[h(\theta)]$$

où n_0 est le nombre d'itérations de chauffage.

- Méthodes graphiques

  Représentations graphiques des sommes cumulées (CUSUM) (Yu et Mykland, 1994)

$$D_T^i = \sum_{t=n_0+1}^{n_0+i} [h(\theta^{(t)}) - S_T], \quad i = 1, ..., T,$$

Mélange de Gaussiennes (Yu, Mykland, 1998)

Looking at Markov samplers through cusum path plots

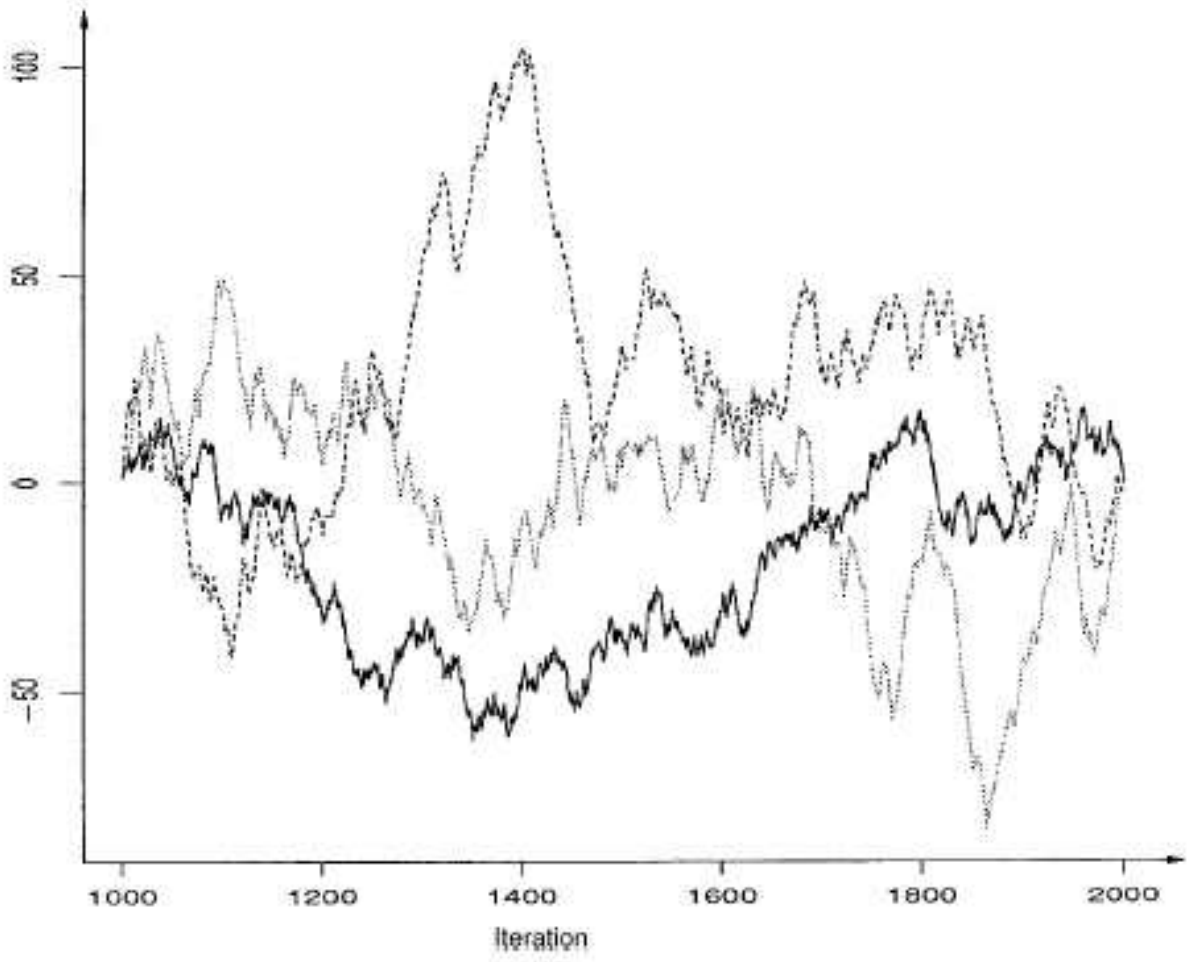
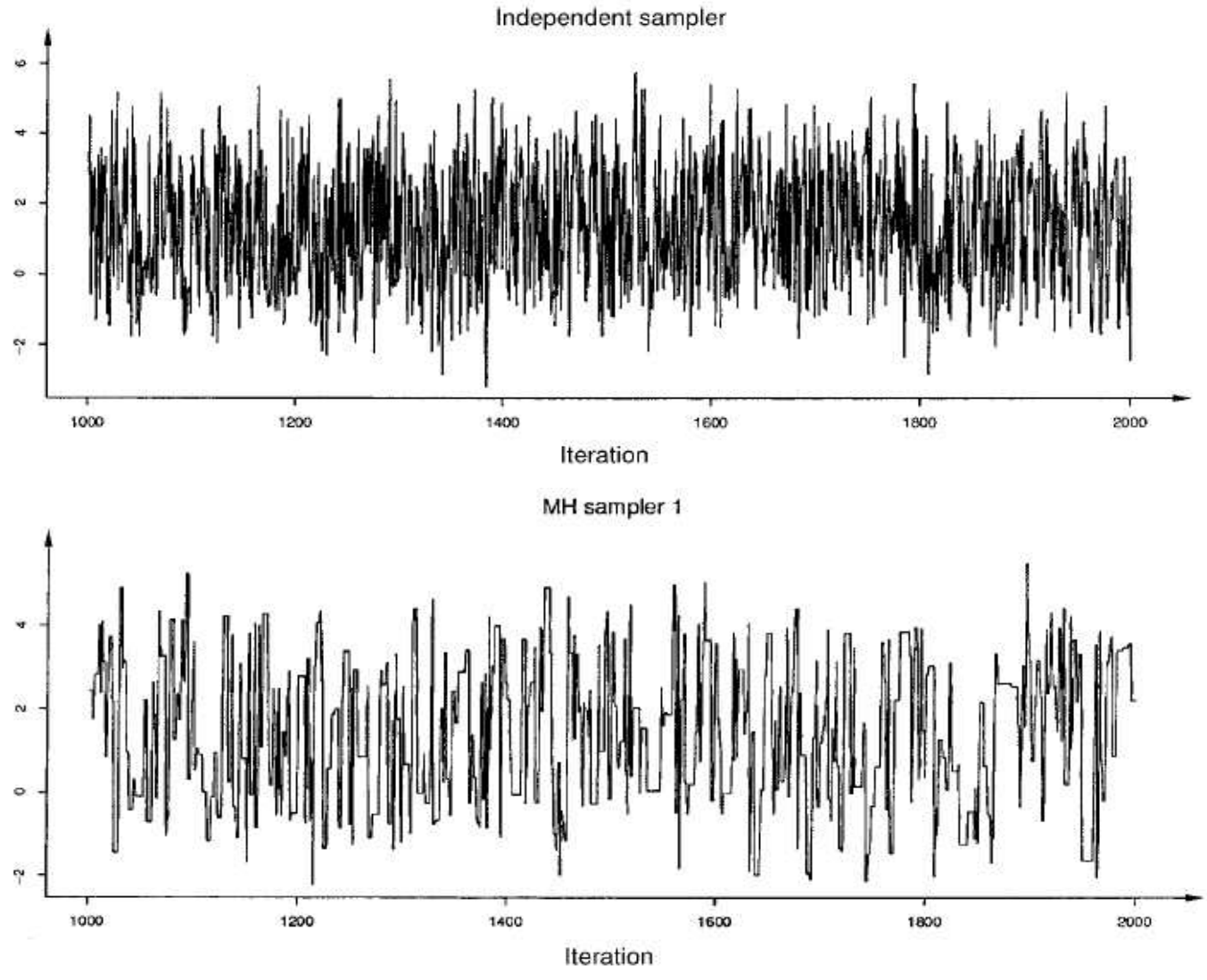


Fig. 2. *Bimodal example. Solid line – independent sampler; dotted line – sampler 1; dashed line – sampler 2*

Mélange de Gaussiennes (Yu, Mykland, 1998)



Estimateurs Multiples

- Moyenne empirique S_T
- Version Rao-Blackwellizée de la Moyenne empirique

$$S_T^C = \frac{1}{T} \sum_{t=n_0+1}^{n_0+T} E[h(\theta^{(t)})|\eta^{(t)}]$$

- Utilisation de l'échantillonnage d'importance

$$S_T^P = \frac{1}{T} \sum_{t=n_0+1}^{n_0+T} \omega_t h(\theta^{(t)}),$$

où $\omega_t = f(\theta^{(t)})/g(\theta^{(t)})$ ($f(.)$ est la loi cible et $g(.)$ est la loi d'importance)

Exemple : Normal-Cauchy

- Loi a posteriori

$$\pi(\theta|x_1,x_2,x_3) \propto e^{-\frac{\theta^2}{2\sigma^2}} \prod_{i=1}^{3} \frac{1}{1+(\theta-x_i)^2}$$

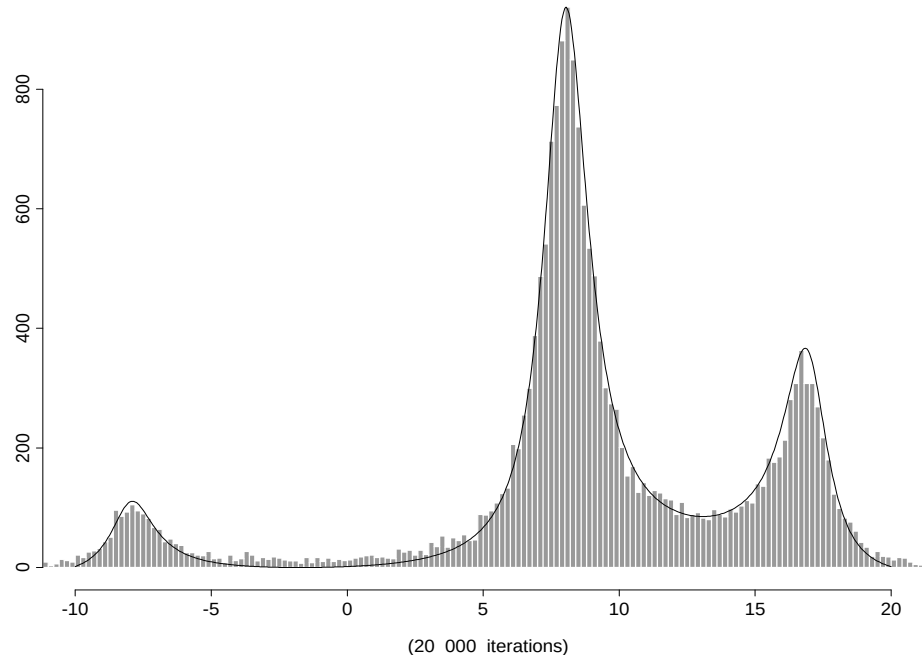
- Complétion

$$\pi(\theta,\eta_1,\eta_2,\eta_3|x_1,x_2,x_3) \propto e^{-\frac{\theta^2}{2\sigma^2}} \prod_{i=1}^{3} e^{-\frac{\eta_i}{2}\left[1+(\theta-x_i)^2\right]}$$

- Lois conditionnelles

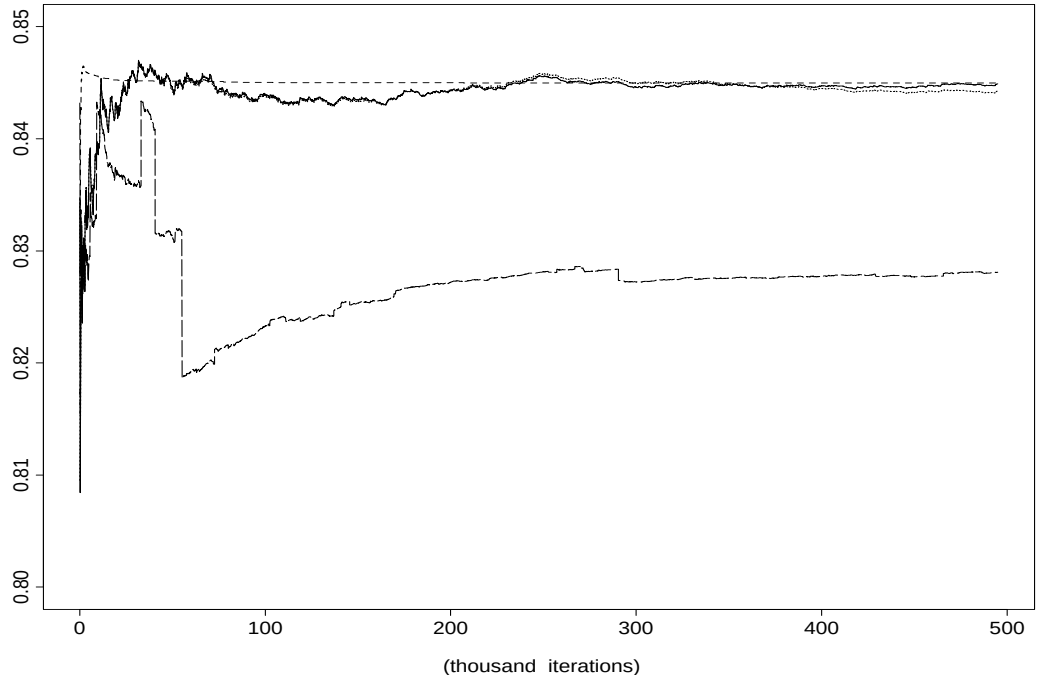
$$\begin{aligned} \eta_i|\theta,x_i &\sim \mathcal{E}\left(\frac{1+(\theta-x_i)^2}{2}\right), \\ \theta|\boldsymbol{x},\boldsymbol{\eta} &\sim \mathcal{N}\left(\frac{\sum \eta_i x_i}{\sum \eta_i+\sigma^{-2}}, \frac{1}{\sum \eta_i+\sigma^{-2}}\right). \end{aligned}$$

Résultats de simulation



Comparison of the normal-Cauchy density and of the histogram $(20,000$ **points)**

Moyenne de $h(\theta) = \exp(-\theta/\sigma)$



Convergence of S_T (full line), S_T^C (dotted line), S_T^R (mixed) and S_T^P (long dashes)

Variances intra et inter-chaînes

- Moyenne de la chaîne m

$$\overline{\psi}_m = \frac{1}{T} \sum_{t=n_0+1}^{n_0+T} \psi_m^{(t)},$$

où $\psi_m^{(t)} = h[\theta_m^{(t)}]$ et $\theta_m^{(t)}$ est l'élément t de la chaîne m.

- Moyenne des moyennes $\overline{\psi} = \frac{1}{M} \sum_{m=1}^{M} \overline{\psi}_m$
- Variance inter-chaînes

$$B_T = \frac{n}{M-1} \sum_{m=1}^{M} \left(\overline{\psi}_m - \overline{\psi}\right)^2$$

- Variance intra-chaînes

$$W_T = \frac{1}{M(T-1)} \sum_{m=1}^{M} \sum_{t=1}^{T} \left(\psi_m^{(t)} - \overline{\psi}_m\right)^2$$

Potential Scale Reduction Factor

- Estimateur de la variance a posteriori de $\psi = h(\theta)$

$$\widehat{\sigma}_T^2 = \frac{T-1}{T} W_T + \left(\frac{M+1}{M}\right) \frac{B_T}{T}$$

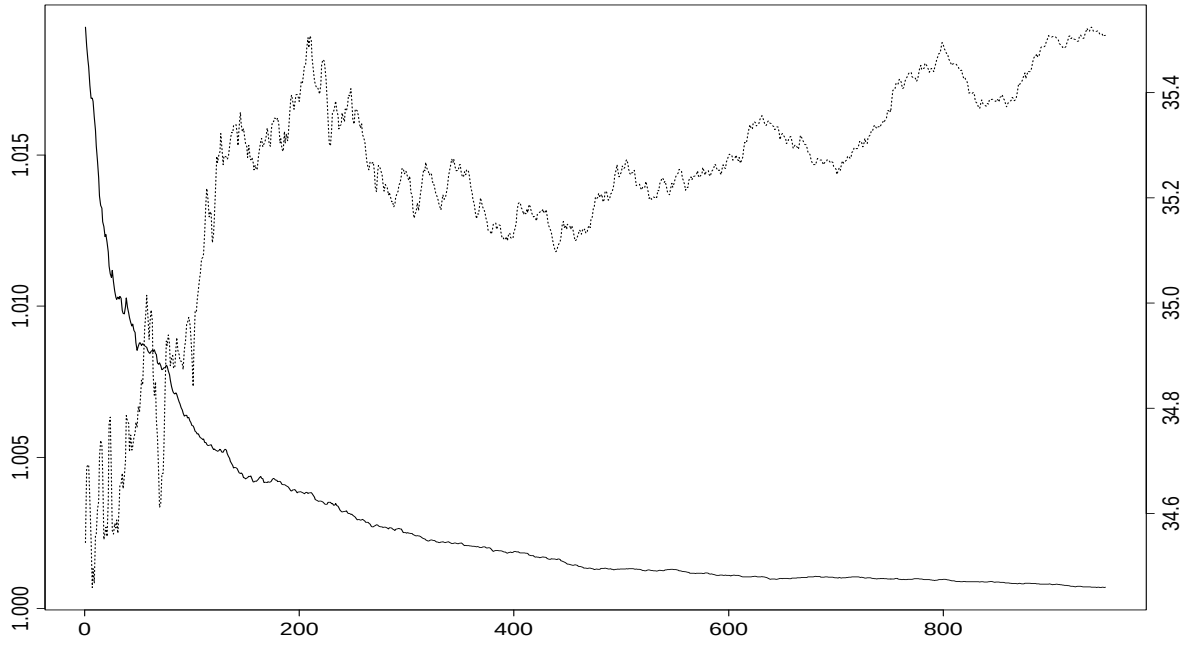
- Potential Scale Reduction Factor

$$R_T = \frac{\widehat{\sigma}_T^2}{W_T} = \frac{T-1}{T} + \frac{1}{T}\left(\frac{M+1}{M}\right) \frac{B_T}{W_T}$$

On compare alors R_T à 1 (une condition de convergence préconisée par Gelman et Rubin est $R_T < 1.2$).

Remarque : la loi de R_T est connue lorsqu'on utilise des approximations normales pour W_T et B_T.

Résultats de simulation



Evolutions of R_T (solid lines and scale on the left) and of W_T (dotted lines and scale on the right)

Commentaires

- Simplicité de cette méthode $\Rightarrow$ Succès
- Nécessite de simuler plusieurs chaînes en parallèle
- Méthode basée sur des approximations normales
- Généralisation au cas multidimensionnel (Brooks and Gelman, 1998)

$$R_T = \frac{T-1}{T} + \left(\frac{M+1}{M}\right)\lambda_1,$$

où λ_1 est la plus grande valeur propre de la matrice symétrique définie positive $W_T^{-1}B_T/T$.

Analyse Spectrale (Geweke, 1992)

- Densité spectrale de puissance de $h(\theta^{(t)})$

$$S_h(\omega) = \frac{1}{2\pi} \sum_{t=-\infty}^{\infty} \text{cov}\left[(h(\theta^{(0)}), h(\theta^{(t)})\right] e^{in\omega}$$

  - $S_h(0)$ estimée à l'aide de $\{\theta^{(1)}, ..., \theta^{(T_A)}\}$: σ_A^2
  - $S_h(0)$ estimée à l'aide de $\{\theta^{(T-T_B+1)}, ..., \theta^{(T)}\}$: σ_B^2

- Estimation des moyennes empiriques à l'aide des deux échantillons

$$\delta_A = \frac{1}{T_A} \sum_{t_0+1}^{t_0+T_A} h\left(\theta^{(t)}\right) \qquad \delta_B = \frac{1}{T_B} \sum_{T-T_B+1}^{T} h\left(\theta^{(t)}\right)$$

Test de Normalité

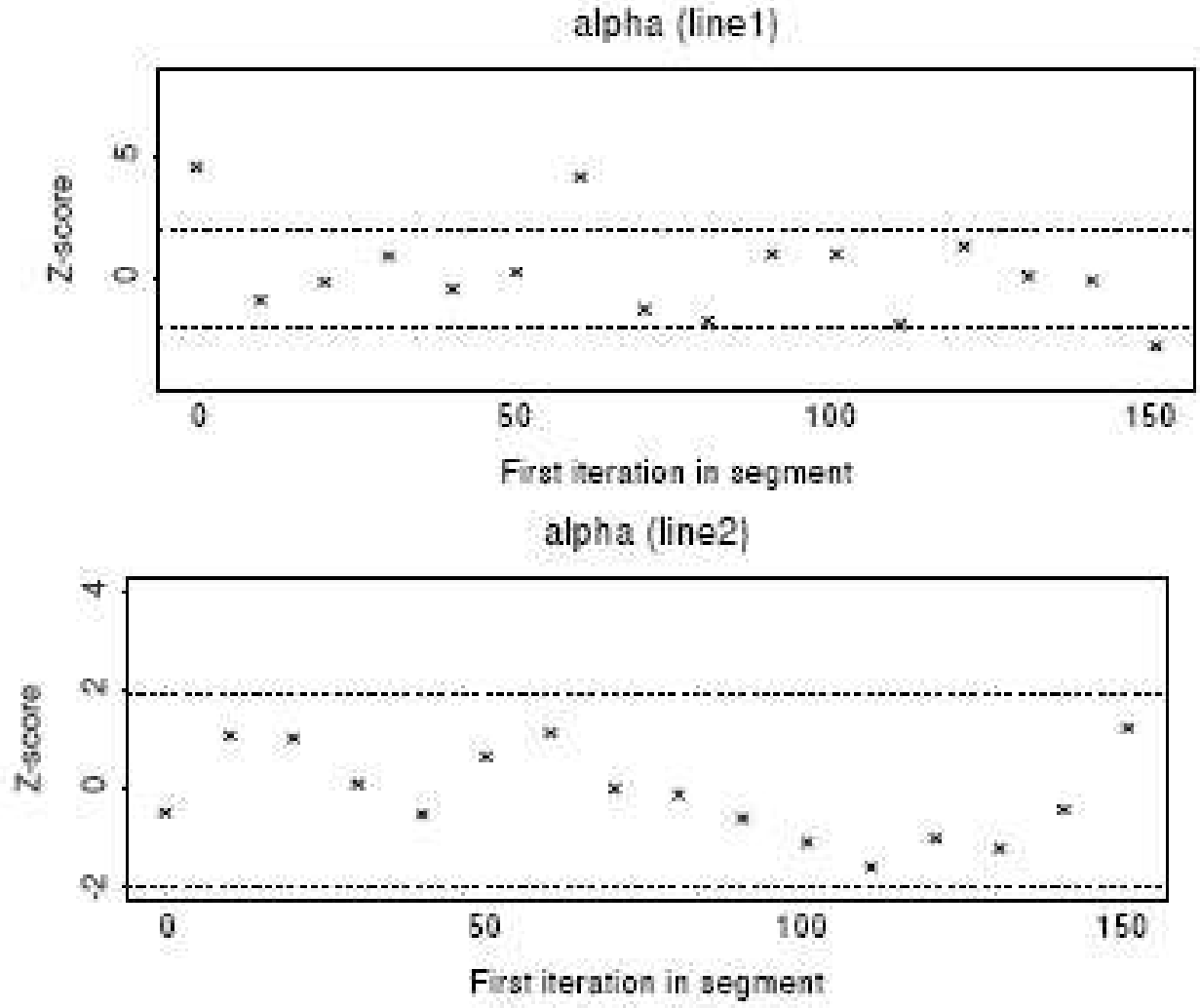
- Z-score

$$\frac{\sqrt{T}\left(\delta_A - \delta_B\right)}{\sqrt{\frac{\sigma_A^2}{\tau_A} + \frac{\sigma_B^2}{\tau_B}}} \underset{T\to\infty}{\to} \mathcal{N}(0,1)$$

  si on est dans la zone de stationarité et si $\tau_A = T_A/T$ et $\tau_B = T_B/T$ sont fixes (e.g. $\tau_A = 0.1$ et $\tau_B = 0.5$).

- Représentation graphique
  - Intervalle de confiance de la loi normale : [-1.96,1.96]
  - Z-scores sur les intervalles
    $[1, ..., T]$,$[n, ..., T]$,$[2n, ..., T]$,...,$[T - 50, ..., T]$

Résultats de simulation



Remarques

- Méthodes de contrôle par
  - Renouvellement (Robert, 1996)
    Théorème limite sur des sommes partielles construites à partir de temps de renouvellement
  - Régénération (Mykland, Tierney et yu, 1995)
- Convergence Diagnosis and Output Analysis (CODA)