Estimation de paramètres de modèles d'EDP

—

méthodes et exemples

Samuel Soubeyrand et Lionel Roques
Unité Biostatistique et Processus Spatiaux
INRA PACA

Atelier ModStatSP
5-6 octobre 2013, Avignon

Plan de la présentation

- Approche mécanistico-statistique avec EDP
- Introduction à l'inférence bayésienne et au MCMC
- Estimation d'un paramètre de croissance spatialement hétérogène
- Estimation de la date et du lieu d'introduction d'une espèce invasive

Approche mécanistico-statistique avec EDP

Cadre d'étude :

- Exemples de phénomènes d'intérêt : dynamique d'invasion, dynamique endémique... d'une espèce donnée
- On dispose de données issues d'un suivi spatio-temporel de la population d'intérêt : données bruitées, dégradées (binarisation), agrégées, non-exhaustives, censurées...
- On dispose d'un modèle mécaniste représentant la dynamique (e.g. EDP)
- On cherche à estimer les paramètres mécanistes

Approche mécanistico-statistique avec EDP

Les données (bruitées, dégradées, agrégées, non-exhaustives, censurées...) sont des fonctions aléatoires de l'évolution spatio-temporelle de la densité réelle de la population

Proposition :

- Construction d'un modèle du processus d'observation
  - conditionnel au modèle mécaniste
  - décrivant le lien stochastique entre densité de population et données
- Estimation des paramètres à l'aide d'une méthode statistique (maximum de vraisemblance, moindres carrés, moments, estimation bayésienne, ABC...)

Structure hiérarchique (state-space model) :

- Modèle du processus d'observation
- Modèle mécaniste de la dynamique
- Modèle a priori pour les paramètres (si inférence bayésienne)

Introduction à l'inférence bayésienne et au MCMC

- Ce que l'on cherche à quantifier en bayésien :
  - la loi a posteriori

$$f(\theta \mid Y) = \frac{f(Y \mid \theta)\pi(\theta)}{\int_\Theta f(Y \mid \alpha)\pi(\alpha)d\alpha}$$

  - et ses caractéristiques : moments a posteriori, maximum a posteriori, quantiles a posteriori, intervalles de crédibilité...

- Exemple : Nombre de succès sur n essais indépendants
  - modèle binomial : $Y \mid \theta \sim \text{Binomiale}(n, \theta)$
  - prior beta : $\theta \sim \text{Beta}(a, b)$
  - posterior [1] : $\theta \mid Y \sim \text{Beta}(a + Y, b - n + Y)$

1. Détail du calcul :

$$\begin{aligned} f(\theta \mid Y) &= \frac{C_n^Y \theta^Y (1-\theta)^{n-Y} \frac{\theta^{a-1}(1-\theta)^{b-1}}{B(a,b)}}{\int_0^1 C_n^Y \alpha^Y (1-\alpha)^{n-Y} \frac{\alpha^{a-1}(1-\alpha)^{b-1}}{B(a,b)} d\alpha} \\ &= \frac{\theta^{(a+Y)-1}(1-\theta)^{(b+n-Y)-1}}{B(a+Y, b+n-Y)} \end{aligned}$$

Introduction à l'inférence bayésienne et au MCMC

- Ce que l'on cherche à quantifier en bayésien :
  - la loi a posteriori

$$f(\theta \mid Y) = \frac{f(Y \mid \theta)\pi(\theta)}{\int_\Theta f(Y \mid \alpha)\pi(\alpha)d\alpha}$$

  - et ses caractéristiques : moments a posteriori, maximum a posteriori, quantiles a posteriori, intervalles de crédibilité...

- Mais la loi a posteriori peut être difficilement calculable
  - Modèle avec nombreux paramètres et variables latentes :

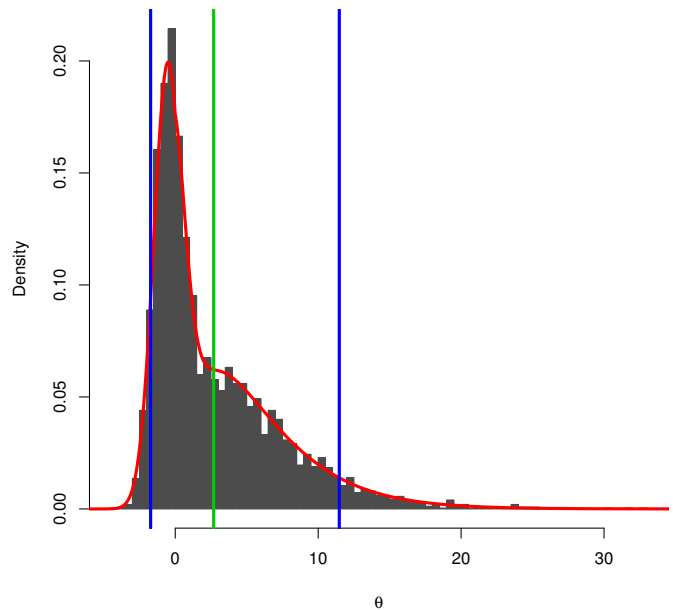
$$f(\theta_1, \ldots, \theta_K \mid Y) = \frac{f(Y \mid \theta_1, \ldots, \theta_K)\pi(\theta_1, \ldots, \theta_K)}{\int_{\Theta_1} \cdots \int_{\Theta_K} f(Y \mid \theta_1, \ldots, \theta_K)\pi(\alpha_1, \ldots, \alpha_K)d\alpha_1 \cdots d\alpha_K}$$

  - Les intégrales multiples (grande dimension) rendent difficile le calcul de la posterior jointe $f(\theta_1, \ldots, \theta_K \mid Y)$, des posteriors marginales $f(\theta_k \mid Y)$, des moments a posteriori $E(\theta_i^q \mid Y)$...

Introduction à l'inférence bayésienne et au MCMC

- Ce que peuvent les méthodes numériques :
  - générer un échantillon issu de la loi a posteriori
  - sans passer par le calcul d'intégrales multiples

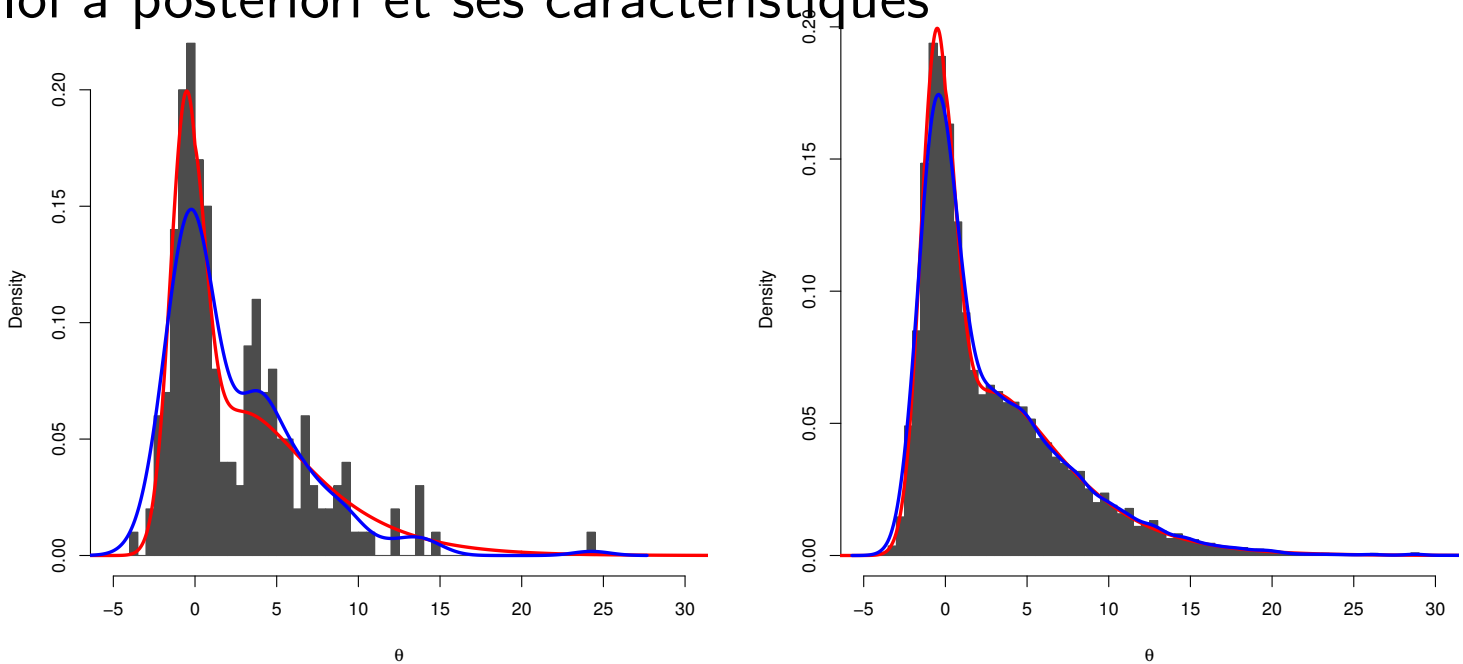
- A quoi sert cet échantillon ?
  - connaître intimement la loi a posteriori
  - estimer la densité a posteriori
  - estimer les moments a posteriori
    ex : $E(\theta \mid Y) \approx \frac{1}{I} \sum_{i=1}^{I} \theta^{(i)}$
  - estimer des intervalles de crédibilité
  - ...



Introduction à l'inférence bayésienne et au MCMC

- Ce que peuvent les méthodes numériques :
  - générer un échantillon issu de la loi a posteriori
  - sans passer par le calcul d'intégrales multiples

- Générer un grand échantillon pour correctement approcher la loi a posteriori et ses caractéristiques

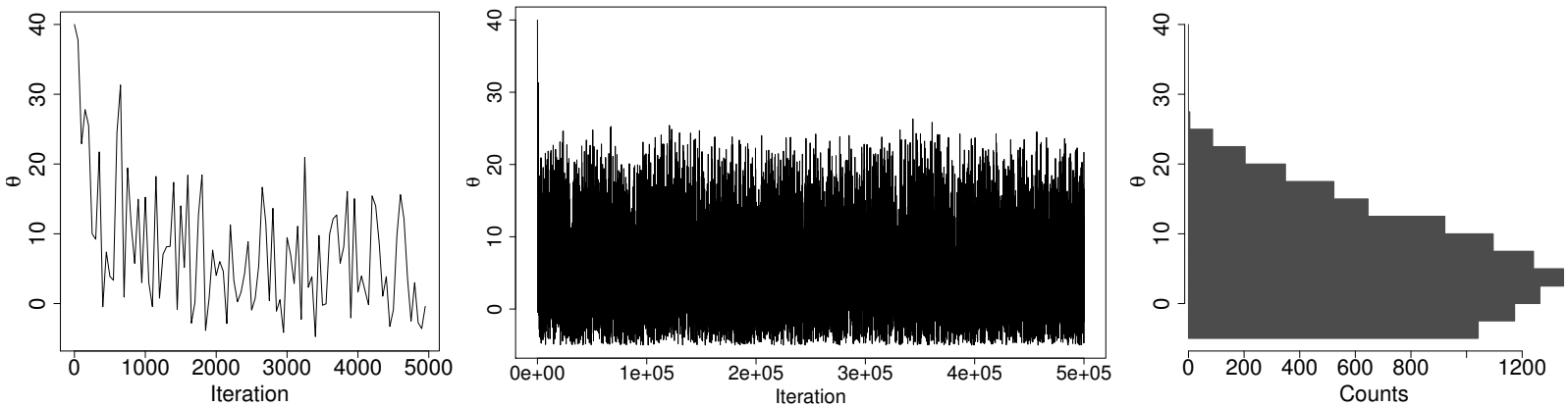


Echantillons de taille 200 (histo gauche) et 10000 (histo droite)

Densités a posteriori vraies (rouge) et estimées (bleu)

MCMC : Présentation

- Méthodes de Monte Carlo par Chaînes de Markov
- Algorithmes séquentiels : une séquence de réalisations dépendantes (i.e. une chaîne) de θ est générée
- Exploration ciblée de l'espace des paramètres (et des variables latentes)
- Qu'est-ce qu'une chaîne ?



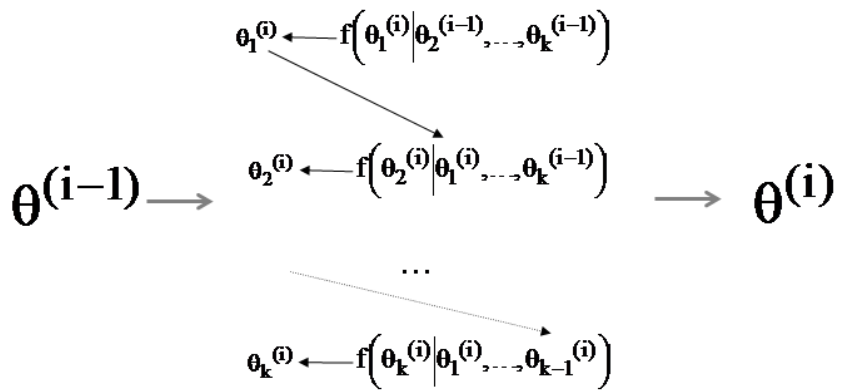
MCMC : Echantillonneur de Gibbs

Algorithme

0. Initialisation : donner des valeurs initiales aux K composantes de θ : $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})$

i. A l'itération $i \in \{1, \ldots, I\}$:

- Générer $\theta_1^{(i)}$ selon la loi $f(\theta_1 \mid Y, \theta_2^{(i-1)}, \ldots, \theta_K^{(i-1)})$
- Générer $\theta_2^{(i)}$ selon la loi $f(\theta_2 \mid Y, \theta_1^{(i)}, \theta_3^{(i-1)}, \ldots, \theta_K^{(i-1)})$
- Générer $\theta_3^{(i)}$ selon la loi $f(\theta_3 \mid Y, \theta_1^{(i)}, \theta_2^{(i)}, \theta_4^{(i-1)}, \ldots, \theta_K^{(i-1)})$
- ...
- Générer $\theta_K^{(i)}$ selon la loi $f(\theta_K \mid Y, \theta_1^{(i)}, \ldots, \theta_{K-1}^{(i)})$



Extrait de Parent et Bernier (2007)

MCMC : Algorithme de Metropolis-Hastings

- Généralisation du MCMC–Gibbs
- Une loi de proposition arbitraire remplace la loi conditionnelle
- La mise à jour des paramètres n'est pas systématique (étape d'acceptation-rejet)
- La loi de proposition intervient dans la proba de mise à jour

MCMC : Algorithme de Metropolis-Hastings (suite)

Algorithme

0. Initialisation : donner des valeurs initiales aux K composantes de θ : $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_K^{(0)})$

i, k. A l'itération $i \in \{1, \ldots, I\}$, pour chaque k :

- Générer θ_k^{cand} selon la loi de proposition $g(\theta_k \mid \theta_k^{(i-1)})$
- Mettre à jour le paramètre $(\theta_k^{(i)} = \theta_k^{cand})$ avec la probabilité [a] :

$$\min\left[1, \frac{Posterior(\theta_k^{cand}) g(\theta_k^{(i-1)} \mid \theta_k^{cand})}{Posterior(\theta_k^{(i-1)}) g(\theta_k^{cand} \mid \theta_k^{(i-1)})}\right]$$

$$= \min\left[1, \frac{(Vraisemblance \times Prior)(\theta_k^{cand}) g(\theta_k^{(i-1)} \mid \theta_k^{cand})}{(Vraisemblance \times Prior)(\theta_k^{(i-1)}) g(\theta_k^{cand} \mid \theta_k^{(i-1)})}\right]$$

- Ne pas mettre à jour $(\theta_k^{(i)} = \theta_k^{(i-1)})$ sinon

---

a. Expression de la probabilité de mise à jour :

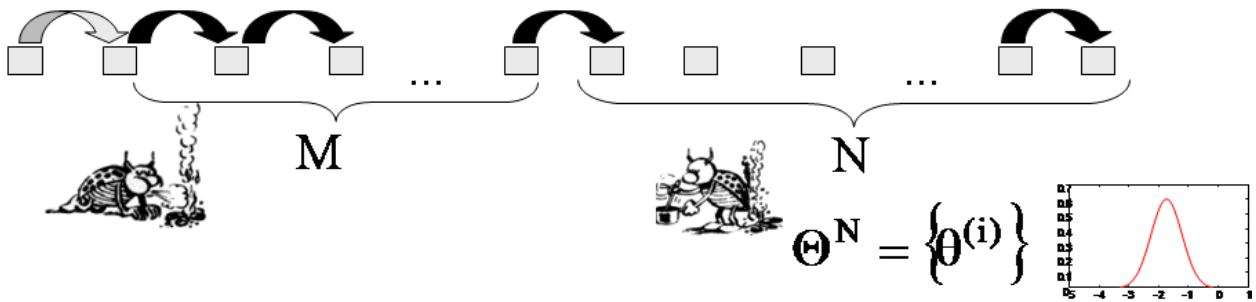
$$\min\left[1, \frac{f(Y \mid \theta_{1:k-1}^{(i)}, \theta_k^{cand}, \theta_{k+1:K}^{(i-1)}) \pi(\theta_{1:k-1}^{(i)}, \theta_k^{cand}, \theta_{k+1:K}^{(i-1)}) g(\theta_k^{(i-1)} \mid \theta_k^{cand})}{f(Y \mid \theta_{1:k-1}^{(i)}, \theta_k^{(i-1)}, \theta_{k+1:K}^{(i-1)}) \pi(\theta_{1:k-1}^{(i)}, \theta_k^{(i-1)}, \theta_{k+1:K}^{(i-1)}) g(\theta_k^{cand} \mid \theta_k^{(i-1)})}\right]$$

MCMC : Algorithme de Metropolis-Hastings (suite)

- Connaissance nécessaire de la loi cible (la loi a posteriori) qu'à une constante près

$$f(\theta_1, \ldots, \theta_K \mid Y) \propto f(Y \mid \theta_1, \ldots, \theta_K)\pi(\theta_1, \ldots, \theta_K)$$

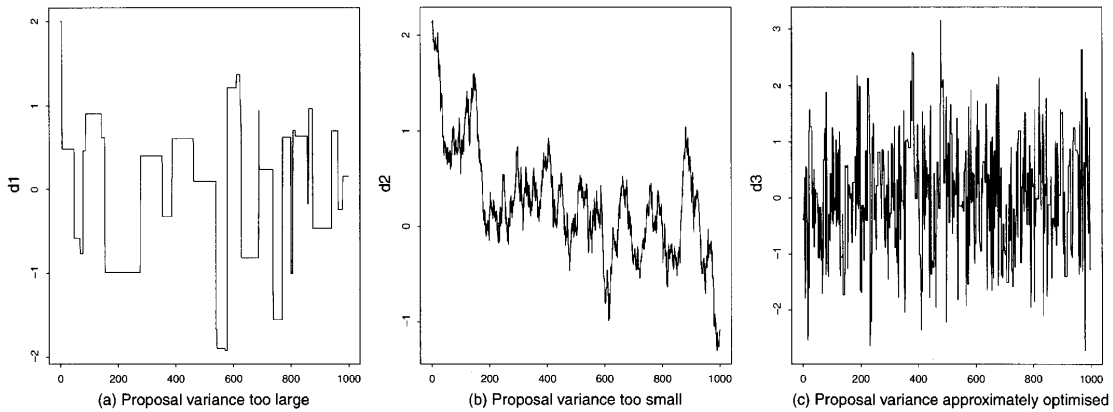
- Mise à jour par bloc
- Algorithmes hybrides combinant Gibbs et M-H
- Comme dans le MCMC–Gibbs, période de chauffe (burn-in) et régime de croisière (stationnaire)



- Contrairement au MCMC–Gibbs, choix des lois de proposition (paramètres de réglages, tuning)
- Question de la convergence de la chaîne vers son régime de croisière (vrai aussi pour le MCMC–Gibbs)

MCMC : Choix des lois de proposition g

- La rapidité de convergence de la chaîne dépend du choix des lois de proposition (formes et valeurs des paramètres)

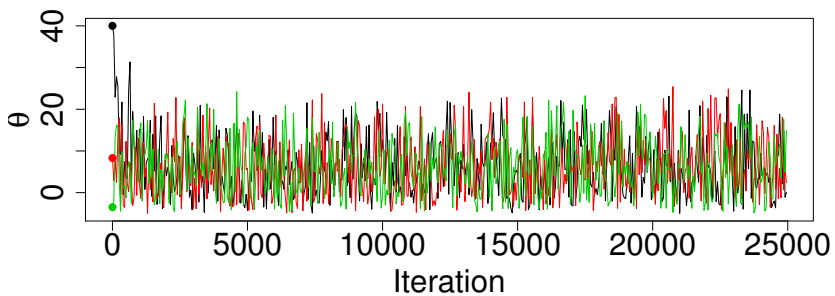


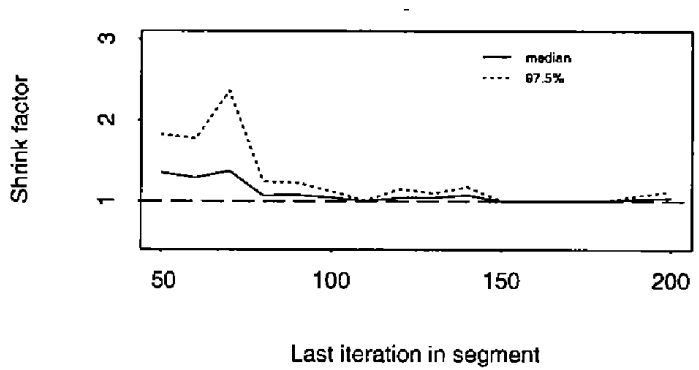
Extrait de Roberts and Rosenthal (2001)

- Tirage indépendant de la valeur courante du paramètre
  ex1 : $g(\theta \mid \theta^{(i-1)})$ est la densité d'une $Normale(\mu, \sigma^2)$
  ex2 : $g(\theta \mid \theta^{(i-1)})$ est la densité d'une $Gamma(\alpha, \beta)$
- Marche aléatoire homogène
  ex1 : $g(\theta \mid \theta^{(i-1)})$ est la densité d'une $Normale(\theta^{(i-1)}, \sigma^2)$
  ex2 : $g(\theta \mid \theta^{(i-1)})$ est la densité d'une $Gamma((\frac{\theta^{(i-1)}}{\sigma})^2, \frac{\sigma^2}{\theta^{(i-1)}})$ tq $E(\theta \mid \theta^{(i-1)}) = \theta^{(i-1)}$ et $Var(\theta \mid \theta^{(i-1)}) = \sigma^2$
- Taux de mise à jour à viser (par essai–erreur) : 25%
- Méthodes adaptatives

MCMC : Diagnostics de convergence des chaînes

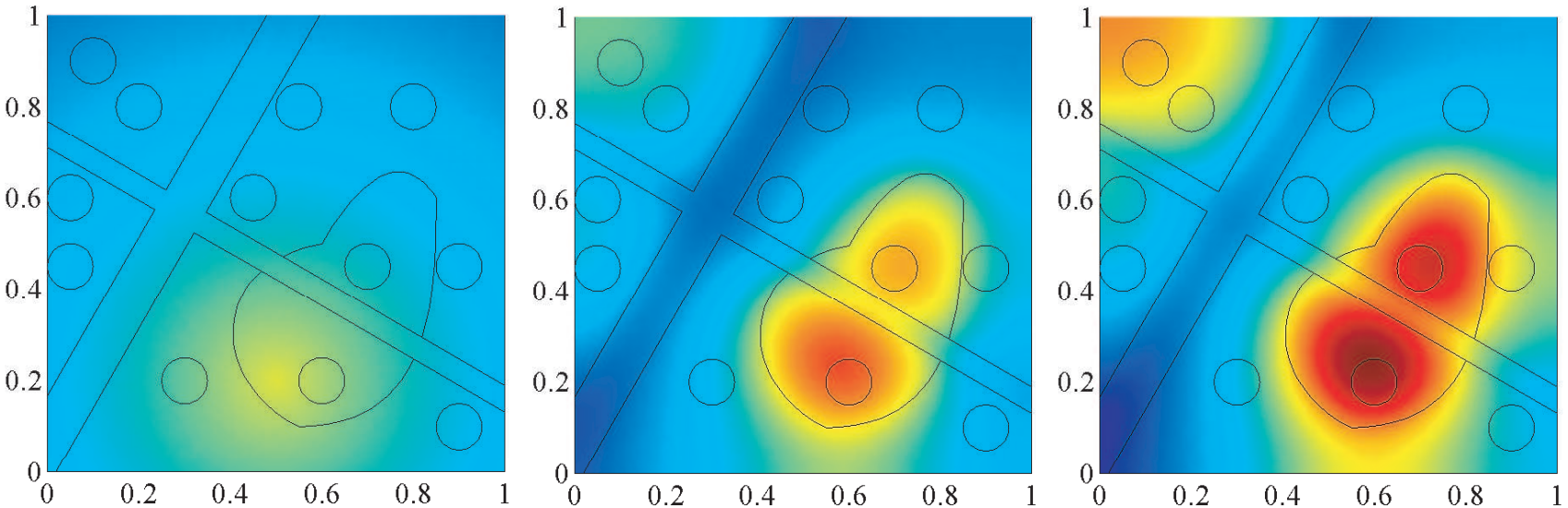
- Quelle convergence ?
  - Convergence vers la stationarité
  - Convergence des moyennes empiriques
    La chaîne a-t-elle exploré toutes les facettes de la distribution cible ?
  - Convergence vers un échantillonnage i.i.d.
- Un exemple de diagnostic : la méthode de Gelman–Rubin qui est basée
  - sur plusieurs chaînes initialisées différemment
  - sur les variances inter-chaîne et intra-chaîne





Estimation d'un paramètre de croissance spatialement hétérogène

$t = 0$ $t = 0.5$ $t = 4$



Mechanistic model

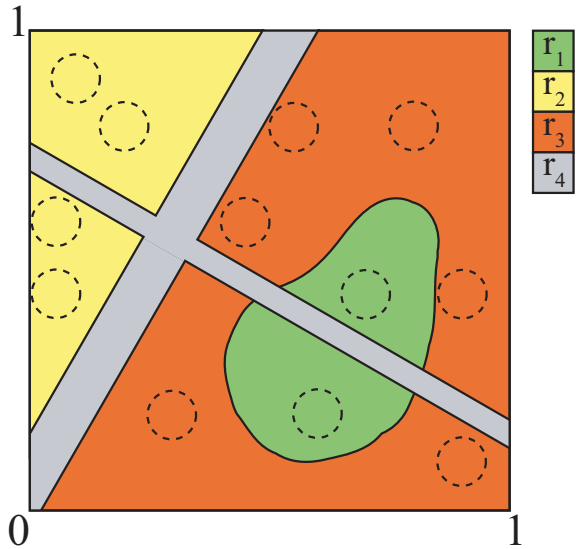
$$\frac{\partial u}{\partial t} = D\,\Delta\, u + u(r(x) - \gamma u),\ \ t > 0,\ \ x \in \Omega \subset \mathbb{R}^2,$$

- $u = u(t, x)$: population density at time t and space location $x \in \Omega$
- $D > 0$ measures the dispersion rate
- $\Delta u = \sum_{i=1}^{d} \frac{\partial^2 u}{\partial x_i^2}$: spatial dispersion operator
- spatially heterogeneous coefficient $r(x)$: intrinsic growth rate of the species (i.e. growth rate in the absence of competition)
- $\gamma > 0$ measures the effect of competition

Given some conditions on the boundary of Ω and an initial condition $u_0(x) = u(0, x)$, this equation is well-posed in the sense that it admits a unique solution $u(t, x)$, for all $t > 0$ and $x \in \Omega$

Mechanistic model

- Partition of the spatial domain Ω into four regions where the intrinsic growth rate $r(x)$ takes constant values : r_1, r_2, r_3, r_4. The dashed circles correspond to the observation regions ω_i



- u satisfies Neumann conditions on the boundary
- At time $t = 0$, the initial population density is $u(0, x) = \exp(-\|x - x_0\|)$ with $x_0 = (0.5, 0.2)$

Model of the observation process

- Y_{ij} : impact of the species towards the environment measured at time τ_j in a subdomain $\omega_i \subset \Omega$
- Impact proportional to the mean time spent by the individuals inside this subdomain :

$$Y_{ij} = \alpha \int_0^{\tau_j} \int_{\omega_i} u(t,x)\, dx\, dt,$$

for some known constant α which measures the mean impact per individual and per unit of time

- $\tilde{Y}_{ij}$: measurement of the impact

$$\tilde{Y}_{ij} \sim_{indep} \text{Poisson}\{Y_{ij}\}.$$

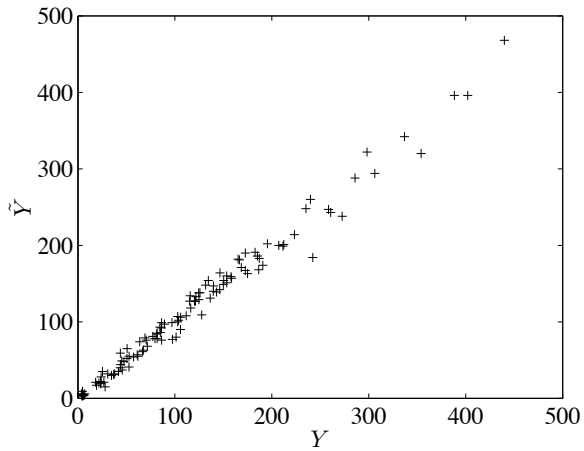
- Likelihood of the model :

$$f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}};\theta) = \prod_{1\leq i\leq I, 1\leq j\leq J} \exp\{-Y_{ij}\} \frac{\{Y_{ij}\}^{\tilde{y}_{ij}}}{\tilde{y}_{ij}!},$$

where $\tilde{\mathbf{y}} = (\tilde{y}_{ij})_{1\leq i\leq I, 1\leq j\leq J}$, and $\theta = (D, \gamma, \mathbf{r})$ governs u

Parameter estimation

- 12 subdomains $\times$ 10 times : 120 observations $\tilde{Y}_{ij}$



- Observation parameter α and initial state $u(0, x)$ are known
- Uniform prior distribution for the parameter vector θ :

$$f(\theta) = \frac{1}{0.99 \times 9.9 \times 20^4} \mathbf{1}(10^{-2} \leq D \leq 1,\ 0.1 < \gamma \leq 10,\ -10 \leq r_1, \ldots, r_4 \leq 10)$$

- Metropolis-Hastings MCMC to estimate the posterior distribution of θ :

$$f(\theta \mid \tilde{\mathbf{y}}) = \frac{f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \theta) f(\theta)}{\int_{\mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}^4} f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \alpha) f(\alpha) d\alpha}$$

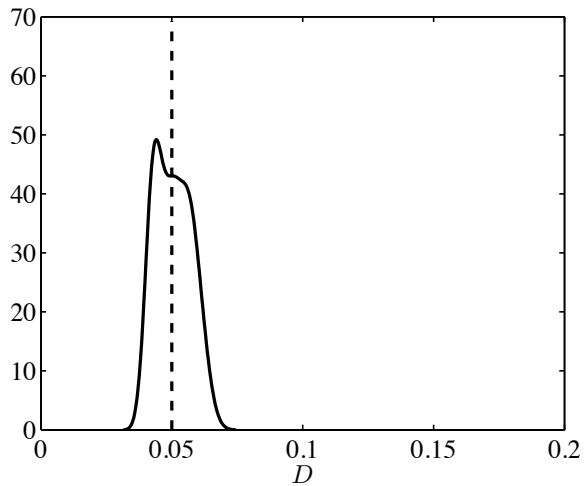
Parameter estimation

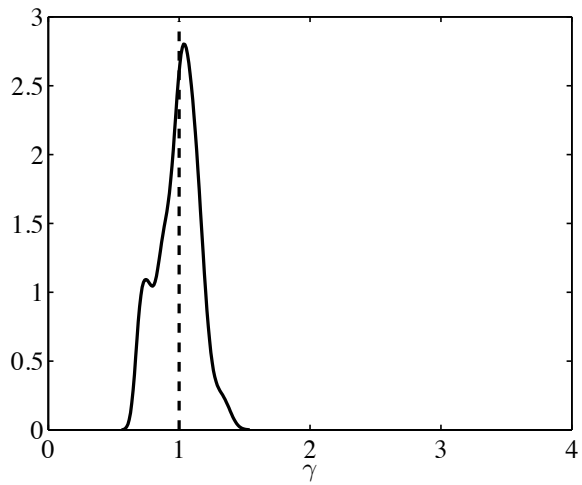
- Algorithm :
  Start at $k = 0$: initialize θ^0.
  while $k \leq N_h$
  - Draw $\hat{\theta}$ from a proposal distribution $Q(\hat{\theta}|\theta^k)$.
  - Draw $\zeta \in (0, 1)$ from a uniform distribution
  - Compute $\delta = \dfrac{f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \hat{\theta}) f(\hat{\theta}) Q(\theta^k|\hat{\theta})}{f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \theta^k) f(\theta^k) Q(\hat{\theta}|\theta^k)}$.
  - **If** $\zeta < \delta$, $\theta^{k+1} = \hat{\theta}$ **else** $\theta^{k+1} = \theta^k$.
  - $k \leftarrow k + 1$

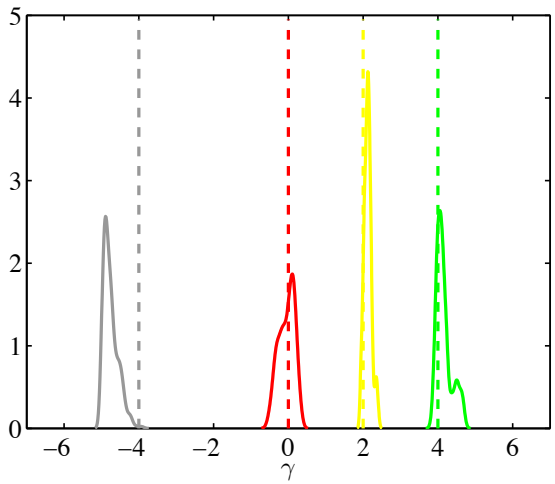
  endwhile
- The likelihood $f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \theta)$ is computed using a numerical solution of the reaction-diffusion equation
- This solution was obtained with Comsol Multiphysics® time-dependent solver, which is based on a second order finite element method (FEM)

Results

Marginal posterior distributions :

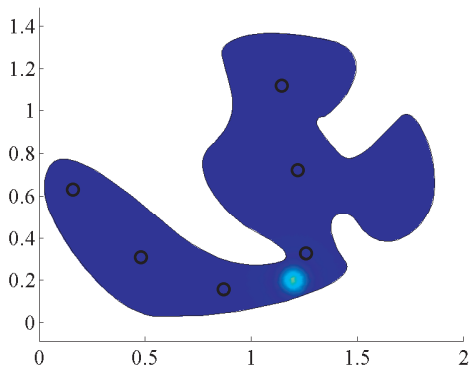


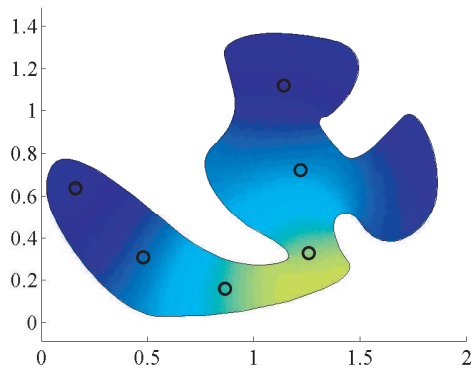


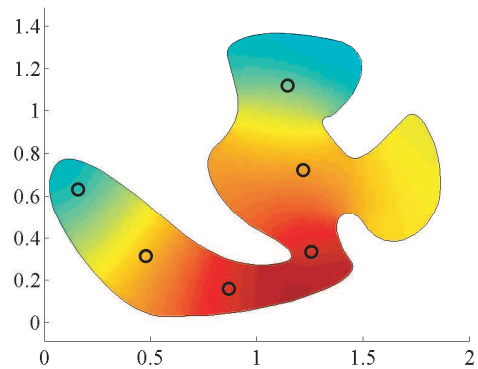


Estimation de la date et du lieu d'introduction d'une espèce invasive

$t = -2$ $t = 0$ $t = 0.9$







Mechanistic model

$$\begin{cases} \dfrac{\partial u}{\partial t} = D\,\Delta u + u(r - \gamma u),\ t > -t_0,\ x \in \Omega, \\ \dfrac{\partial u}{\partial \nu}(t,x) = 0,\ t > -t_0,\ x \in \partial\Omega, u(-t_0,x) = u_0(x),\ x \in \Omega, \\ u_0(x) = \exp\left(-20\,\|x - x_0\|\right) \end{cases}$$

where t_0 is the introduction time and $x_0 \in \Omega$ is the location of introduction

Model of the observation process

- Y_{ij} : impact of the species towards the environment measured at time τ_j in a subdomain $\omega_i \subset \Omega$
- Impact proportional to the number of individuals inside the subdomain at the observation time :

$$Y_{ij} = \alpha \int_{\omega_i} u(\tau_j, x)\, dx,$$

for some known constant α which measures the mean impact per unit of population density

- $\tilde{Y}_{ij}$: measurement of the impact

$$\tilde{Y}_{ij} \sim_{indep} \text{Poisson}\{Y_{ij}\}.$$

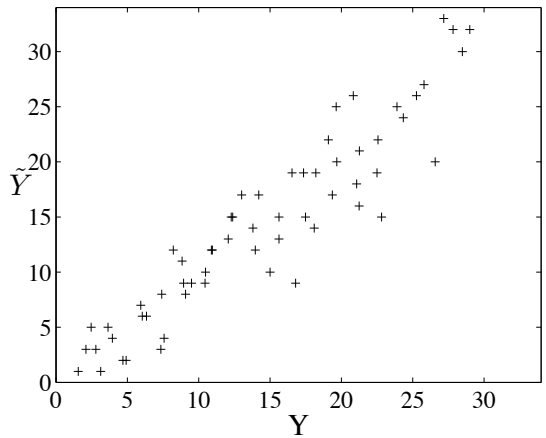
- Likelihood of the model :

$$f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \theta) = \prod_{1 \le i \le I, 1 \le j \le J} \exp\{-Y_{ij}\} \frac{\{Y_{ij}\}^{\tilde{y}_{ij}}}{\tilde{y}_{ij}!},$$

where $\tilde{\mathbf{y}} = (\tilde{y}_{ij})_{1 \le i \le I, 1 \le j \le J}$, and $\theta = (t_0, x_0, D, \gamma, r)$ governs u

Parameter estimation

- 6 subdomains $\times$ 10 times between $t = 0$ and $t = 0.9$ (while introduction arises at $t = -2$) : 60 observations $\tilde{Y}_{ij}$



- Observation parameter α is known
- Uniform prior distribution for the parameter vector θ :

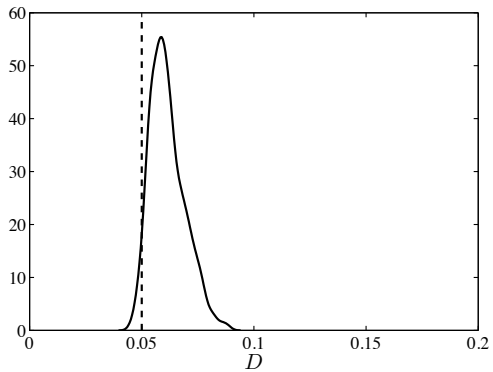
$$f(\theta) = \frac{1}{0.99 \times 9.9 \times 20 \times 10 \times |\Omega|} \mathbf{1}(10^{-2} < D < 1,\ 0.1 < \gamma < 10) \times \mathbf{1}(-10 < r < 10,\ 0 < t < 10,\ x_0 \in \Omega)$$

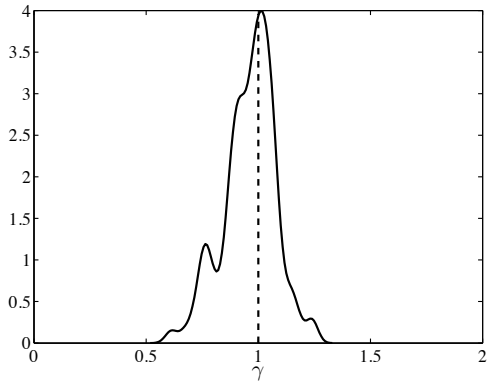
- Metropolis-Hastings MCMC to estimate the posterior distribution of θ :

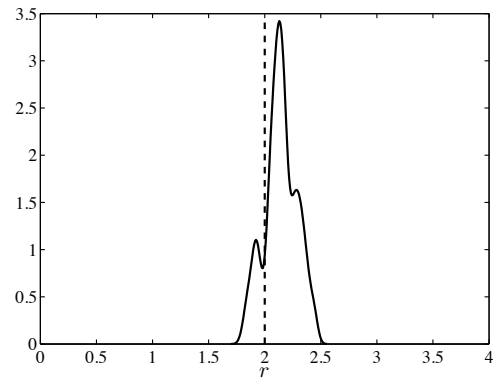
$$f(\theta \mid \tilde{\mathbf{y}}) = \frac{f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \theta) f(\theta)}{\int_{\mathbb{R}_+^* \times \mathbb{R}_+^* \times \mathbb{R}^4} f_{\tilde{\mathbf{Y}}}(\tilde{\mathbf{y}}; \alpha) f(\alpha) d\alpha}$$

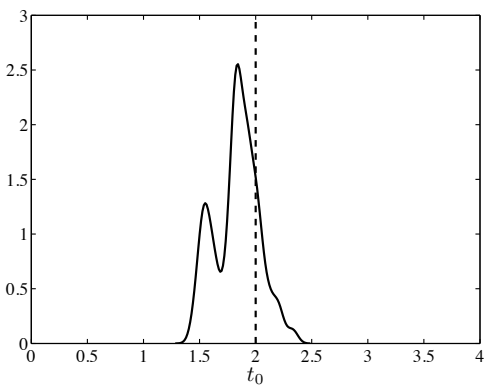
Results

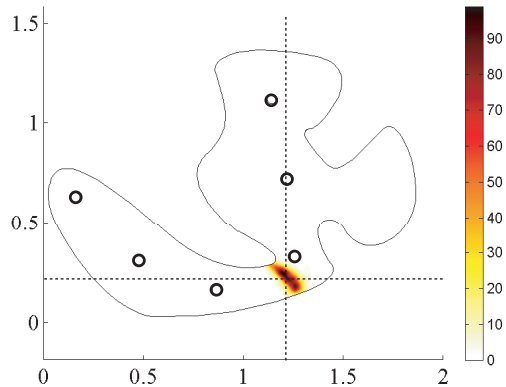
Marginal posterior distributions :











Références

MCMC :

- Casella G. & George E. I. (1992). Explaining the Gibbs sampler. The American Statistician 46 : 167–174.
- Chib S. & Greenberg E. (1995). Understanding the Metropolis-Hastings algorithm. The American Statistician 49 : 327–335.
- Parent E. & Bernier J. (2007). Le raisonnement bayésien – Modélisation et inférence. Springer, Paris.
- Robert C. P. & Casella G. (1999). Monte Carlo Statistical Methods. Springer, New York.
- Roberts G. O. & Rosenthal J. S. (2001). Optimal Scaling for Various Metropolis-Hastings Algorithms. Statistical Science 16 : 351–367.

MCMC and EDP :

- Roques L., Soubeyrand S. and Rousselet J. (2011). A statistical-reaction-diffusion approach for analyzing expansion processes. Journal of Theoretical Biology 274 : 43-51.
- Soubeyrand S. and Roques L. (2013). Problèmes inverses et estimations de paramètres. PDF file. In : Roques L. (Author). Modèles de Réaction-Diffusion pour l'Ecologie Spatiale. Editions QUAE, Versailles. ISBN : 9782759220298.
- Soubeyrand S. and Roques L. (in press). Parameter estimation for reaction-diffusion models of biological invasions. Population Ecology.

Packages R pour MCMC

MCMC :

- adaptMCMC : Implementation of a generic adaptive Monte Carlo Markov Chain sampler
- mcmc : Markov Chain Monte Carlo
- MCMCpack : Markov chain Monte Carlo (MCMC) Package

Interfaçage :

- BRugs : R interface to the OpenBUGS MCMC software
- runjags : Interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS

Diagnostics :

- bmk : MCMC diagnostics package
- boa : Bayesian Output Analysis Program (BOA) for MCMC
- coda : Output analysis and diagnostics for MCMC
- ggmcmc : Graphical tools for analyzing Markov Chain Monte Carlo simulations from Bayesian inference
- mcmcplots : Create Plots from MCMC Output
- scapeMCMC : MCMC Diagnostic Plots