Modèles À Espace d'États Non Linéaires/Non Gaussiens et Inférence Bayésienne par Méthode MCMC

Une Application en Évaluation des Stocks Halieutiques[1]

Fabien Campillo
INRIA/IRISA
Campus de Beaulieu
35042 Rennes
FRANCE
Fabien.Campillo@inria.fr

Rivo Rakotozafy
Faculté des Sciences
Université de Fianarantsoa
301 Fianarantsoa
MADAGASCAR
rrakotoz@mail.univ-fianar.mg

Résumé L'évolution de la biomasse d'un stock de poissons peut être modélisée à l'aide d'équations aux différences avec retard (delay difference models). Une formulation de type modèle à espace d'états permet alors une analyse bayésienne pertinente à partir de la donnée des captures annuelles. On peut, par exemple, faire appel à un algorithme de Metropolis-Hastings "composante à composante" (single–component Metropolis–Hastings algorithm).

Mots clefs Inférence bayésienne, Monte Carlo par chaînes de Markov, échantillonneur de Gibbs, algorithme de Metropolis–Hastings "composante à composante", modèles non linéaire à espace d'états, modèles aux différences avec retard.

Abstract Difference equations with delay are widely used to model the evolution of the biomass of a fish stock (delay difference models). Represented as a state–space model they allow, starting from the data of the annual catches, a relevant Bayesian analysis. For this purpose we can use an hybrid MCMC method combining a Metropolis-Hastings algorithm within a Gibbs sampler, namely the single–component Metropolis–Hastings algorithm.

Key words Bayesian inference, Monte Carlo Markov chains, Gibbs sampler, single–component Metropolis–Hastings algorithm, nonlinear state-space models, delay difference model.

Les techniques d'estimation des stocks de pêche visent à évaluer l'impact de différents scénarios d'exploitation et de capture sur l'évolution de l'abondance d'un stock de poissons. Cette abondance s'exprime en poids total de poissons vulnérables à la pêche (biomasse). L'accroissement de cette biomasse intègre l'augmentation du poids moyen de chaque individu ainsi que le recrutement, c'est à dire l'arrivée chaque année de nouvelles générations de poissons. À l'inverse, cette biomasse décroît du fait de la capture (mortalité par pêche) ou d'autres causes (mortalité naturelle : prédation, maladies, etc.).

1. Présenté aux XXXVIèmes Journées de Statistique, 24 au 28 mai 2004, Montpellier.

Pour chaque année $t = 1 \cdots T$ d'une période de T années, B_t désigne la biomasse en début d'année, C_t la biomasse capturée en cours d'année et I_t un indice d'abondance du stock. Cet indice, mesuré chaque année, est la capture par unité d'effort (CPUE). Les séries $C_{1:T}$ et $I_{1:T}$ sont données. Le but est de déterminer, au sein d'une classe de modèles, lequel s'ajuste le mieux aux observations.

À la suite des travaux de Meyer et Millar [4], on adopte une formulation à espace d'états non linéaire d'un modèle aux différences avec retard (delay difference model) de la dynamique de la biomasse et de l'observation de l'indice d'abondance de cette biomasse. Ce modèle a été introduit par Deriso [2] et généralisé par Schnute [5] (voir Hilborn et Walters [3] pour une présentation générale). Dans la représentation à espace d'états, les indices $I_{1:T}$ seront considérées comme les observations, $B_{1:T}$ comme les états (non observés) et $C_{1:T}$ comme les entrées.

Les paramètres du modèle sont ajustés aux données par une procédure de Metropolis–Hastings "composante à composante".

Notre propos n'est pas de valider le modèle aux différences avec retard, mais de tenter d'évaluer la pertinence de l'algorithme de MCMC proposé dans ce cadre.

Un modèle à espace d'états pour l'évaluation de stock

À l'année t la biomasse B_t est :

$$B_t = \sum_{a \geq k} w_a \, N_{a,t} \tag{1}$$

où $N_{a,t}$ est la taille de la population d'âge a à l'année t, w_a est le poids moyen d'un individu à l'âge a, k est l'âge de recrutement (âge auquel on estime que les poissons viennent s'ajouter à la biomasse). Le premier terme $R_t = w_k \, N_{k,t}$ dans la somme (1) est le recrutement de l'année t.

On suppose les individus de la biomasse uniformément vulnérables à la mortalité naturelle et à la capture. L'évolution annuelle de chaque cohorte (population d'individus de la même classe d'âge) est donc décrite par :

$$N_{a,t} = s_{t-1} \, N_{a-1,t-1} \, , \tag{2}$$

où s_{t-1} est le taux de survie à la mortalité naturelle et à la pêche. L'évolution de w_a est donnée selon la relation classique de croissance du poids selon l'âge : $w_a = w_\infty \, (1 - e^{-\kappa \, (a - a_0)})$ pour $a \geq a_0$ où a_0 est l'âge auquel $w_{a_0} = 0$ et $\kappa > 0$. Cela conduit à :

$$w_a = (1 + \rho) \, w_{a-1} - \rho \, w_{a-2} \tag{3}$$

où $\rho = e^{-\kappa} \in (0, 1)$ est le taux de croissance. Les expressions (2) et (3) dans (1) conduisent à la récurrence :

$$B_t = (1 + \rho) \, s_{t-1} \, B_{t-1} - \rho \, s_{t-1} \, s_{t-2} \, B_{t-2} + R_t - \rho \, s_{t-1} \, w \, R_{t-1} \tag{4}$$

avec $w = w_{k-1}/w_k \in (0,1)$.

On suppose que les causes de mortalité naturelle et de mortalité par capture agissent de façon indépendante sur l'ensemble des poissons, ainsi $s_t = s_t^M\, s_t^F$ où le taux de survie à la mortalité naturelle s_t^M est supposé constant ($s^M = e^{-M}$ avec $M > 0$), et où le taux de survie à la capture est $s_t^F = \frac{B_t - C_t}{B_t}$. Le recrutement est ensuite supposé constant, i.e. $R = R_t$. L'équation (4) devient donc :

$$B_t = (1+\rho)\, e^{-M}\, \tfrac{B_{t-1}-C_{t-1}}{B_{t-1}}\, B_{t-1} - \rho\, e^{-2M}\, \tfrac{B_{t-1}-C_{t-1}}{B_{t-1}}\, \tfrac{B_{t-2}-C_{t-2}}{B_{t-2}}\, B_{t-2} + R\left(1 - \rho\, e^{-M}\, w\, \tfrac{B_{t-1}-C_{t-1}}{B_{t-1}}\right).$$

Avant la première année $t = 1$, les captures sont supposées nulles $C_{t<0} = 0$. La biomasse est supposée être à son équilibre jusqu'à l'année 1, i.e. $B_{t\leq 1} = K$ où K est un paramètre inconnu (la biomasse vierge). La biomasse à l'année $t = 2$ est donc :

$$B_2 = (1+\rho-\rho\, e^{-M})\, e^{-M}\, (B_1 - C_1) + R\left(1 - \rho\, e^{-M}\, w\, \tfrac{B_1-C_1}{B_1}\right).$$

L'équation d'observation non bruitée est de la forme $I_t = q\, B_t$ pour $t = 1 \cdots T$, où I_t est un indice de biomasse relative et q est un coefficient de "capturabilité". Les processus de bruit d'état et d'observation sont multiplicatifs et lognormaux :

$$B_t = F_t(R, K, B_{t-1}, B_{t-2}) \times e^{\sigma_W\, W_t}\,, \tag{5}$$

$$I_t = q\, B_t \times e^{\sigma_V\, V_t}\,, \tag{6}$$

pour $t = 1 \cdots T$, $W_{1:T}$ et $V_{1:T}$ sont des bruits blancs gaussiens $N(0,1)$ et

$$\begin{aligned}
F_1(R, K, B_0, B_{-1}) &= K\,,\\
F_2(R, K, B_1, B_0) &= (1+\rho-\rho\, e^{-M})\, e^{-M}\, (B_1 - C_1) + R\left(1 - \rho\, e^{-M}\, w\, \tfrac{B_1-C_1}{B_1}\right),\\
F_t(R, K, B_{t-1}, B_{t-2}) &= (1+\rho)\, e^{-M}\, \tfrac{B_{t-1}-C_{t-1}}{B_{t-1}}\, B_{t-1} - \rho\, e^{-2M}\, \tfrac{B_{t-1}-C_{t-1}}{B_{t-1}}\, \tfrac{B_{t-2}-C_{t-2}}{B_{t-2}}\, B_{t-2}\\
&\quad + R\left(1 - \rho\, e^{-M}\, w\, \tfrac{B_{t-1}-C_{t-1}}{B_{t-1}}\right), \quad t = 3 \cdots T\,.
\end{aligned}$$

Les paramètres ρ, M, w sont ici supposés connus. Les paramètres K, R, q, σ_W^2, σ_V^2 sont supposés aléatoires et inconnus. Ces coefficients et les processus de bruit sont supposés mutuellement indépendants.

Ce modèle propose un bon compromis entre les modélisations globales (surplus production models) et les modélisations structurées en âge. Les premières sont réductrices dans la mesure où elles agrègent en un seul paramètre toutes les causes d'évolution de la biomasse, les secondes sont plus sophistiquées mais nécessitent des données de capture ventilées en taille (ainsi qu'un modèle reliant la taille à l'âge), le plus souvent ce type de données est inaccessible.

Méthode MCMC

Il s'agit d'identifier K, R, q, σ_W^2, σ_V^2, $B_{1:T}$ à partir des observations $I_{1:T}$. Notons :

$$X_{1:N} = [K, R, q, \sigma_W^2, \sigma_V^2, B_{1:T}] \text{ avec } N = T + 5\,.$$

Le but est d'échantillonner selon la densité a posteriori :

$$p(X_{1:N}|I_{1:T}) \tag{7}$$

de $X_{1:N}$ sachant que $I_{1:T}$, ce qui ne peut se faire directement, on fait donc appel à une méthode MCMC : on construit une chaîne de Markov ergodique $\{X_{1:N}^{(k)}\}_{k\geq 0}$ dont la densité (de la loi) stationnaire est (7).

Appliquons l'échantillonneur de Gibbs : on simule successivement selon chacune des densités conditionnelles marginales de X_n sachant les autres composantes $\{X_{n'}; n' \neq n\}$ (et $I_{1:T}$) i.e. on simule selon la densité :

$$p(X_n|X_{\setminus n}, I_{1:T}) \quad \text{où } \setminus n = \{n' = 1, \ldots, N\,;\, n' \neq n\} \tag{8}$$

i.e. la densité conditionnelle de X_n sachant $X_{\setminus n}$ et $I_{1:T}$, pour $n = 1 \cdots N$. Ces densités conditionnelles admettent des représentations analytiques mais à nouveau il n'est pas toujours possible d'échantillonner directement selon elles. On fait donc appel à un algorithme de Metropolis–Hastings à chaque itération de l'échantillonneur de Gibbs : chacune des densités conditionnelles (8) est décomposée en produit d'une densité de proposition (une densité selon laquelle il est possible d'échantillonner simplement) et d'une fonction de vraisemblance (que l'on sait calculer explicitement) :

$$p(X_n|X_{\setminus n}, I_{1:T}) \propto \underbrace{\pi_n(X_n|X_{\setminus n}, I_{1:T})}_{\text{densité de proposition}} \times \underbrace{\Psi_n(X_n, X_{\setminus n}, I_{1:T})}_{\text{vraisemblance}}\,. \tag{9}$$

Partant d'une configuration initiale $\mathbf{X}_{1:N}^{(k)}$ on simule la configuration suivante $\mathbf{X}_{1:N}^{(k+1)}$ de la façon suivante : on pose $\mathbf{X}_{1:N}^{(k+1)} \leftarrow \mathbf{X}_{1:N}^{(k)}$, et pour tout indice $n \in \{1, \ldots, N\}$, on simule ensuite un candidat pour la $n$ième composante

$$\mathbf{X}'_n \sim \pi_n(X_n|\mathbf{X}_{\setminus n}^{(k+1)}, I_{1:T})\, dX_n \quad \text{et calcule } \alpha = \frac{\Psi_n(\mathbf{X}'_n, \mathbf{X}_{\setminus n}^{(k+1)}, I_{1:T})}{\Psi_n(\mathbf{X}_n^{(k)}, \mathbf{X}_{\setminus n}^{(k+1)}, I_{1:T})}\,,$$

avec probabilité $\min(1, \alpha)$ (resp. $1 - \min(1, \alpha)$) on accepte (resp. on rejette) la nouvelle configuration, dans le premier cas $\mathbf{X}_n^{(k+1)} \leftarrow \mathbf{X}'_n$. Toutes les composantes sont ainsi mises à jour.

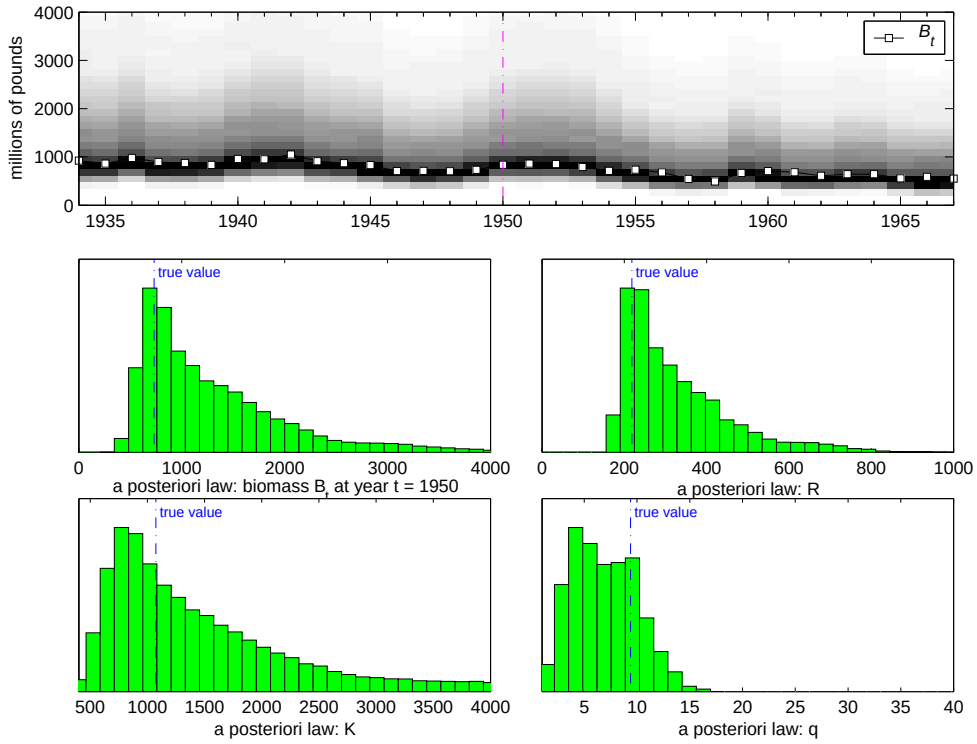

FIG. 1 – *La figure du haut présente l'évolution de la biomasse $B_{1:T}$ ainsi, que pour chaque année t, la représentation en niveaux de gris de la loi a posteriori empirique associée. Les quatre autres figures présentent les lois a posteriori empiriques correspondant respectivement à B_{1950}, R, K et q (les loi a priori des paramètres R, K, q ne sont pas représentée : elles sont uniformes respectivement sur* $[0, 2000]$, $[400, 4000]$,$[1, 100]$*).*

Résultats préliminaires

Nous appliquons cet algorithme sur des données simulées comparables à celles utilisées dans Meyer et Millar [4] (eux mêmes font référence aux données de Pella–Tomlinson sur le thon albacore 1934–1967). Dans ces premiers essais, les variances σ_W^2 et σ_V^2 sont supposées connues.

Les densités a priori des paramètres R, K, q sont uniformes sur $[R_{\min}, R_{\max}] = [0, 2000]$, $[K_{\min}, K_{\max}] = [400, 4000]$, $[q_{\min}, q_{\max}] = [1, 100]$ respectivement. Dans cet exemple "simpliste" (d'une part simulé et d'autre part où les paramètres σ_W^2 et σ_V^2 sont supposés connus) il est en effet possible d'utiliser des densités a priori aussi peu informatives. Le choix des décompositions (9) des densités conditionnelles est présenté en annexe.

Comme l'illustre la Figure 1, la correspondance entre les données simulées et les lois a posteriori empiriques obtenues par l'échantillonneur de Gibbs est relativement satisfaisante. Toutefois, un des principaux inconvénients de cette approche est le nombre im-

portant d'itérations nécessaires à sa convergence. Il serait donc souhaitable de se tourner vers des méthodes plus sophistiquées comme les procédures de Monte Carlo en interaction (voir Del Moral and Doucet [1]). De plus, il est aussi nécessaire de comparer ces résultats avec des approches non bayésiennes.

Remerciements M. Rakotozafy a bénéficié d'une bourse de l'Ambassade de France à Madagascar. Ce travail a été partiellement réalisé lors d'un de ses séjours à l'Irisa.

Références

[1] P. Del Moral and A. Doucet. On a class of genealogical and interacting metropolis models. In J.-M. Morel, F. Takens, and B. Teissier, editors, *Séminaire de Probabilités XXXVII*, LNM 1832, pages 415 – 446. Springer–Verlag, 2003.

[2] R. B. Deriso. Harvesting strategies and parameter estimation for an age-structured model. *Canadian Journal of Fisheries and Aquatic Sciences*, 37:268–282, 1980.

[3] R. Hilborn and C. Walters. *Quantitative Fisheries Stock Assessment: Choice, Dynamics and Uncertainty.* Chapman and Hall, 1992.

[4] R. Meyer and R.B. Millar. Bayesian stock assessment using a state-space implementation of the delay difference model. *Canadian Journal of Fisheries and Aquatic Sciences*, 56:37–52, 1999.

[5] J. Schnute. A general theory for the analysis of catch and effort data. *Canadian Journal of Fisheries and Aquatic Sciences*, 42:414–429, 1985.

Décompositions densités de proposition/vraisemblance. Notons simplement $F_t = F_t(K, R, B_{t-1}, B_{t-2})$. Pour $t \geq 2$, F_t est linéaire en R, i.e. $F_t = F'_t + F''_t R$. Pour chacun des paramètres et de composantes de B_t on donne ici le choix de la densité de proposition (i.e. l'échantillonneur) et la vraisemblance Ψ associée :

$$K \leftarrow B_1 \, e^{\sigma_W \, \texttt{randn}}, \quad \Psi(K) = K$$

$$R \leftarrow \frac{F'_{t_0}}{B_{t_0}} \, e^{\sigma_W \, \texttt{randn}} - \frac{F'_{t_0}}{F''_{t_0}} \qquad \text{où } t_0 \text{ est choisi au hasard dans } \{2, \ldots, T\}$$

$$\Psi(R) = \mathbf{1}_{[R_{\min}, R_{\max}]}(R) \left[R - \frac{F''_{t_0}}{F'_{t_0}}\right] \prod_{\substack{2 \leq t \leq T \\ t \neq t_0}} \exp\left(-\frac{1}{2\sigma_W^2}\left[\log \frac{F'_t + R\,F''_t}{B_t}\right]^2\right)$$

$$q \leftarrow \exp\left(\frac{1}{T}\sum_{t=1}^{T} \log\left(\frac{I_t}{B_t}\right) + \frac{\sigma_V}{\sqrt{T}} \, \texttt{randn}\right), \quad \Psi(q) \propto \mathbf{1}_{[q_{\min}, q_{\max}]}(q)$$

$$B_t \leftarrow \exp\left(\frac{\sigma_V^2 \, \log(F_t) + \sigma_W^2 \log(\frac{I_t}{q})}{\sigma_W^2 + \sigma_V^2} + \frac{\sigma_W \, \sigma_V}{\sqrt{\sigma_W^2 + \sigma_V^2}} \, \texttt{randn}\right)$$

$$\Psi(B_t) = \exp\left(-\frac{1}{2\sigma_W^2}\left\{[\log \frac{B_{t+1}}{F_{t+1}}]^2 + [\log \frac{B_{t+2}}{F_{t+2}}]^2\right\}\right)$$

où `randn` est un générateur pseudo–aléatoire $N(0, 1)$.