Projet de MAP568

janvier-mars 2023

Josselin Garnier (Ecole polytechnique)

à remettre pour le 17 mars 2023

1 Introduction

Le but de ce projet est d'effectuer une calibration bayésienne d'un modèle d'évolution issu de l'écologie, en utilisant les outils présentés dans le cours. On va considérer les équations de prédation de Lotka-Volterra, qui sont utilisées pour décrire la dynamique de systèmes biologiques dans lesquels un prédateur et sa proie interagissent, et des données disponibles en ligne.

On attend un notebook jupyter pour le projet, à remettre le 17 mars 2023 au plus tard par mail à josselin.garnier@polytechnique.edu. Le notebook devra être envoyé pré-exécuté et ne doit pas utiliser d'import non standard. Il est recommandé de faire le projet en binôme.

2 Modèle de Lotka-Volterra

Lotka (1925) et Volterra (1926) ont formulé des équations différentielles paramétriques qui caractérisent la dynamique des populations de prédateurs et de proies. On va effectuer une inférence bayésienne complète pour résoudre le problème inverse de l'inférence des paramètres à partir de données bruitées.

Dans le modèle de Lotka-Volterra, l'évolution des populations est régie par :

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}y(t)}{\mathrm{d}t} & = & y(t)\,\Big(\alpha - \beta z(t)\Big), \\ \dfrac{\mathrm{d}z(t)}{\mathrm{d}t} & = & z(t)\,\Big(\delta y(t) - \gamma\Big). \end{array} \right. \tag{1}$$

où :
- $y(t)$, l'effectif des proies en fonction du temps,
- $z(t)$, l'effectif des prédateurs en fonction du temps.

Les variables $y(t)$ et $z(t)$ sont les sortie observées (avec bruit de mesure).

Les paramètres suivants caractérisent les interactions entre les deux espèces :
- α, taux de reproduction intrinsèque des proies (constant, indépendant du nombre

Paramètre	Description	Loi
α	taux de reproduction des proies	$\mathcal{LN}(\mu_\alpha = 0.5, \sigma_\alpha^2 = 0.2^2)$
β	taux de mortalité des proies dû aux prédateurs	$\mathcal{LN}(\mu_\beta = 0.05, \sigma_\beta^2 = 0.02^2)$
δ	taux de reproduction des prédateurs en fonction des proies	$\mathcal{LN}(\mu_\delta = 0.05, \sigma_\delta^2 = 0.02^2)$
γ	taux de mortalité des prédateurs	$\mathcal{LN}(\mu_\gamma = 0.5, \sigma_\gamma^2 = 0.2^2)$
y_0	population initiale de proies	$\mathcal{LN}(\lambda_{y_0} = \log(10), \zeta_{y_0}^2 = 1)$
z_0	population initiale de prédateurs	$\mathcal{LN}(\lambda_{z_0} = \log(10), \zeta_{z_0}^2 = 1)$

TABLE 1 – Lois des 6 paramètres d'entrée du modèle.
Les paramètres $\alpha, \beta, \delta, \gamma$ sont exprimés en année^{-1}, le temps est exprimé en année, les tailles de population sont exprimées en milliers.
On a $X \sim \mathcal{LN}(\lambda, \zeta^2)$ ssi $\ln X \sim \mathcal{N}(\lambda, \zeta^2)$. On a alors $\mu = \mathbb{E}[X] = \exp(\lambda + \zeta^2/2)$ et $\sigma^2 = \mathrm{Var}(X) = \exp(2\lambda + \zeta^2)(\exp(\zeta^2) - 1)$.

de prédateurs),
- β, taux de mortalité des proies dû aux prédateurs rencontrés,
- δ, taux de reproduction des prédateurs en fonction des proies rencontrées et mangées,
- γ, taux de mortalité intrinsèque des prédateurs (constant, indépendant du nombre de proies).

Les conditions initiales du système sont données en termes de la date initiale t_0 et des populations initiales de proies y_0 et de prédateurs z_0.

Les 6 paramètres $\alpha, \beta, \delta, \gamma, y_0, z_0$, constituent les paramètres d'entrée incertains du modèle. Lorsque ces 6 paramètres sont fixés, on obtient une trajectoire des variables de sortie $y(t), z(t)$ par résolution du système (1).

3 Propagation d'incertitudes

Les lois des 6 paramètres d'entrée (qu'on regroupe dans le vecteur d'entrée $\boldsymbol{x} \in \mathbb{R}^6$ dans la suite) sont données dans la table 1. On suppose les paramètres indépendants.

Question 1 : Programmez la résolution du système (1) avec $t_0 = 1900$ (l'unité de temps est l'année).

Question 2 : Par échantillonnage Monte Carlo (en utilisant les lois des paramètres d'entrée de la table 1), représentez la loi des populations de proies et prédateurs en 1900, en 1901 et en 1920.

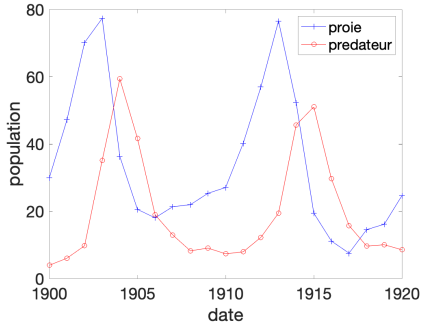

FIGURE 1 – Visualisation des données

4 Calibration

On va considérer des données qui concernent l'évolution des populations de lynx (prédateurs) et de lièvres (proies) entre 1900 et 1920. Ces données ont été obtenues sur la base du nombre de fourrures collectées annuellement par la Compagnie de la Baie d'Hudson. Ces données (et des informations sur ce problème) se trouvent sur : https://jmahaffy.sdsu.edu/courses/f09/math636/lectures/lotka/qualde2.html

On les reproduit ici :
annee = [1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1920]
lievre = [30 47.2 70.2 77.4 36.3 20.6 18.1 21.4 22 25.4 27.1 40.3 57 76.6 52.3 19.5 11.2 7.6 14.6 16.2 24.7]
lynx = [4 6.1 9.8 35.2 59.4 41.7 19 13 8.3 9.1 7.4 8 12.3 19.5 45.7 51.1 29.7 15.8 9.7 10.1 8.6]
Les tailles de population sont exprimées en milliers.

Question 3 : Dessinez les données collectées de 1900 à 1920 (vous devez trouver quelque chose qui ressemble à la figure 1).

4.1 Calibration déterministe

On note t_i, $i = 0, \dots, 20$ les dates où les données sont recueillies ($t_0 = 1900$, $t_{20} = 1920$). On cherche à ajuster au mieux le modèle au sens des moindres carrés :

$$\boldsymbol{x}^* = \text{argmin}_{\boldsymbol{x}} \mathcal{E}(\boldsymbol{x}),$$
$$\mathcal{E}(\boldsymbol{x}) = \sum_{i=0}^{20} e_{proie}(\boldsymbol{x}, t_i)^2 + e_{preda}(\boldsymbol{x}, t_i)^2,$$

où les résidus sont définis par :

$$e_{proie}(\boldsymbol{x}, t_i) = \log \mathcal{M}_{proie}(\boldsymbol{x}, t_i) - \log \text{data}_{proie}(t_i),$$
$$e_{preda}(\boldsymbol{x}, t_i) = \log \mathcal{M}_{preda}(\boldsymbol{x}, t_i) - \log \text{data}_{preda}(t_i),$$

$\mathcal{M}(\boldsymbol{x}, t) = (\mathcal{M}_{proie}(\boldsymbol{x}, t), \mathcal{M}_{preda}(\boldsymbol{x}, t))$ est la prédiction à l'instant t du nombre de proies et de prédateurs par le modèle de Lotka-Volterra avec les paramètres $\boldsymbol{x}$. Le fait de prendre le log sera expliqué dans la section suivante, cela correspond à supposer un modèle d'erreur multiplicative.

Question 4 : Evaluez numériquement $\boldsymbol{x}^*$ dans le domaine de $\mathbb{R}^6$ déterminé par le support essentiel des lois a priori (l'hypercube $\prod_{j=1}^{6}[\exp(\lambda_j - 2\zeta_j), \exp(\lambda_j + 2\zeta_j)]$). Comparez sur une figure les données $(\text{data}_{proie}(t_i), \text{data}_{preda}(t_i))_{i=0}^{20}$ et les prédictions $\mathcal{M}(\boldsymbol{x}^*, t)$.
Remarques : Il y a des routines d'optimisation dans python ! Attention, la fonction $\mathcal{E}$ peut posséder des minima locaux !

4.2 Calibration bayésienne

Le modèle statistique est le suivant :

- On connaît la loi a priori des paramètres $\boldsymbol{x} \in \mathbb{R}^6$ (voir table 1).

- Pour évaluer la vraisemblance, on suppose un modèle statistique prenant en compte une erreur (de mesure et/ou de modèle) multiplicative, de loi log-normale. On a alors

$$\text{data}_{proie}(t_i) = \mathcal{M}_{proi}(\boldsymbol{x}, t_i) \exp\left(\epsilon_{proie,i}\right), \tag{2}$$
$$\text{data}_{preda}(t_i) = \mathcal{M}_{preda}(\boldsymbol{x}, t_i) \exp\left(\epsilon_{preda,i}\right), \tag{3}$$

où $\epsilon_{proie,i} \sim \mathcal{N}(0, \sigma^2_{proie})$, $\epsilon_{preda,i} \sim \mathcal{N}(0, \sigma^2_{preda})$ sont indépendantes en i et entre elles. On obtient ainsi une expression de la vraisemblance $p(\mathbf{data}|\boldsymbol{x}, \boldsymbol{\sigma})$, $\mathbf{data} = (\text{data}_{proie}(t_i), \text{data}_{preda}(t_i))_{i=0}^{20}$, $\boldsymbol{\sigma} = (\sigma_{proie}, \sigma_{preda})$:

$$p(\mathbf{data}|\boldsymbol{x}, \boldsymbol{\sigma}) = \frac{1}{(2\pi)^{21}\sigma_{proie}^{21}\sigma_{preda}^{21}} \exp\Big[-\frac{1}{2}\sum_{i=0}^{20} \frac{e_{proie}(\boldsymbol{x}, t_i)^2}{\sigma^2_{proie}} - \frac{1}{2}\sum_{i=0}^{20} \frac{e_{preda}(\boldsymbol{x}, t_i)^2}{\sigma^2_{preda}} \Big],$$

où e_{proie} et e_{preda} sont définis par (2-3). On remarque que le point $\boldsymbol{x}^*$ obtenu dans la calibration déterministe est le maximum de vraisemblance lorsque $\sigma_{proie} = \sigma_{preda}$ est fixé à une valeur arbitraire.

En fait, on ne connaît pas les valeurs de σ_{proie} et σ_{preda}. On peut alors envisager deux approches :
- Une approche plug-in, dans laquelle on fixe la valeur de $\boldsymbol{\sigma}$ à $\boldsymbol{\sigma}^*$ telle que la vraisemblance $p(\mathbf{data}|\boldsymbol{x}^*, \boldsymbol{\sigma})$ est maximale (c'est ce qu'on va faire dans cette section).
- Une approche full-bayésienne où $\boldsymbol{\sigma}$ suit une loi a priori peu informative (c'est ce qu'on fera dans la section suivante).

On commence par déterminer l'hyper-paramètre $\boldsymbol{\sigma}$ par une méthode du maximum de vraisemblance : on fixe la valeur de $\boldsymbol{\sigma}$ à $\boldsymbol{\sigma}^*$ telle que la vraisemblance $p(\mathbf{data}|\boldsymbol{x}^*, \boldsymbol{\sigma})$ est maximale.

Question 5 : Vérifiez qu'on a $(\sigma^*_{proie})^2 = \frac{1}{21}\sum_{i=0}^{20} e_{proie}(\boldsymbol{x}^*, t_i)^2$ *et* $(\sigma^*_{preda})^2 = \frac{1}{21}\sum_{i=0}^{20} e_{preda}(\boldsymbol{x}^*, t_i)^2$.

Dans l'approche plug-in, la loi a posteriori de $\boldsymbol{x}$ est de la forme

$$p(\boldsymbol{x}|\mathbf{data}, \boldsymbol{\sigma}^*) \approx p(\mathbf{data}|\boldsymbol{x}, \boldsymbol{\sigma}^*) p_{\text{prior}}(\boldsymbol{x})$$

où $\approx$ signifie "à une constante multiplicative près". Elle n'a pas d'expression explicite puisqu'elle implique des appels au modèle de Lotka-Volterra dans la vraisemblance. Par conséquent, nous devons recourir à des algorithmes d'échantillonnage. Ici, nous suggérons d'utiliser un algorithme de Metropolis-Hastings adaptatif.

Une prédiction de la population de proies $M_{proie}(t_{new})$ et prédateurs $M_{preda}(t_{new})$ pour une année t_{new} peut alors être effectuée car, pour toute fonction test ϕ,

$$\mathbb{E}\big[\phi\big(M_{proie}(t_{new})\big)\big] = \int_{\mathbb{R}} \int_{]0,+\infty[^2} \phi\big(\mathcal{M}_{proie}(t_{new}, \boldsymbol{x}) \exp(\sigma^*_{proie}\varepsilon)\big) p(\boldsymbol{x}|\mathbf{data}, \boldsymbol{\sigma}^*) p(\varepsilon) d\boldsymbol{x} d\varepsilon$$

avec $p(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\varepsilon^2}{2})$.

Question 6 : Générez un échantillon de la loi a posteriori des paramètres $\boldsymbol{x}$ par un algorithme de Metropolis-Hastings adaptatif. Tracez des histogrammes des lois a posteriori des paramètres $\alpha, \beta, \delta, \gamma$. Tracez des histogrammes des prédictions des populations de proies et prédateurs en 1921 et en 1930.

4.3 Calibration full-bayésienne

Dans cette section on va appliquer une approche full-bayésienne. On reprend la situation de la section précédente et on suppose une loi a priori sur $\boldsymbol{\sigma}$ de la forme

$\sigma_{proie}, \sigma_{preda} \sim \mathcal{LN}(\lambda = -1, \zeta^2 = 1^2)$, avec $\sigma_{proie}, \sigma_{preda}$ indépendantes entre elles. Cette loi a priori est peu informative : elle est naturellement concentrée sur les réels positifs, et elle consiste à supposer qu'une valeur de σ_{proie} ou σ_{preda} inférieure à $\exp(\lambda - 2\zeta) \simeq 0.05$ ou supérieure à $\exp(\lambda + 2\zeta) \simeq 3$ est peu vraisemblable.

Dans l'approche full-bayésienne, la loi a posteriori de $\boldsymbol{x}$ est de la forme

$$\begin{aligned} p(\boldsymbol{x}|\mathbf{data}) &= \int_{]0,+\infty[^2} p(\boldsymbol{x}, \boldsymbol{\sigma}|\mathbf{data}) d\boldsymbol{\sigma}, \\ p(\boldsymbol{x}, \boldsymbol{\sigma}|\mathbf{data}) &\approx p(\mathbf{data}|\boldsymbol{x}, \boldsymbol{\sigma}) p_{\text{prior}}(\boldsymbol{x}) p_{\text{prior}}(\boldsymbol{\sigma}), \end{aligned}$$

où $\approx$ signifie "à une constante multiplicative près".

La loi a posterori n'a pas d'expression explicite. Par conséquent, nous devons recourir à des algorithmes d'échantillonnage. Ici, nous suggérons d'utiliser un algorithme de Metropolis-Hastings adaptatif pour échantillonner $(\boldsymbol{x}, \boldsymbol{\sigma})$.

Une prédiction de la population de proies $M_{proie}(t_{new})$ et prédateurs $M_{preda}(t_{new})$ pour une année t_{new} peut alors être effectuée car

$$\mathbb{E}\big[\phi\big(M_{proie}(t_{new})\big)\big] = \int_{\mathbb{R}} \int_{]0,+\infty[^4} \phi\big(\mathcal{M}_{proie}(t_{new}, \boldsymbol{x}) \exp(\sigma_{proie}\varepsilon)\big) p(\boldsymbol{x}, \boldsymbol{\sigma}|\mathbf{data}) p(\varepsilon) d\boldsymbol{x} d\boldsymbol{\sigma} d\varepsilon$$

avec $p(\varepsilon) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{\varepsilon^2}{2})$.

Question 7 : reprenez la question 6 dans le cas full-bayésien.

Question 8 (si vous avez le temps) : refaites l'inférence avec un modèle d'erreur additive, pour lequel (2-3) est remplacé par :

$$\begin{aligned} \text{data}_{proie}(t_i) =& \mathcal{M}_{proi}(\boldsymbol{x}, t_i) + \epsilon_{proie,i}, \\ \text{data}_{preda}(t_i) =& \mathcal{M}_{pred}(\boldsymbol{x}, t_i) + \epsilon_{preda,i}, \end{aligned}$$

où $\epsilon_{proie,i} \sim \mathcal{N}(0, \sigma^2_{proie})$, $\epsilon_{preda,i} \sim \mathcal{N}(0, \sigma^2_{preda})$ sont indépendantes en i et entre elles. Comparez les résultats obtenus avec ceux du modèle d'erreur multiplicative.

Références

[1] Volterra, V. (1926). Fluctuations in the abundance of a species considered mathematically. Nature, 118(2972), 558-560.

[2] Lotka, A. J. (1925). Principles of physical biology. Baltimore : Waverly.