Projet de MAP568

janvier-mars 2024

Josselin Garnier (Ecole polytechnique)

à remettre pour le 22 mars 2024

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 des variantes des équations de prédation de Lotka-Volterra, qui sont utilisées pour décrire la dynamique de systèmes biologiques dans lesquels des prédateurs et leurs proies interagissent. On utilisera des données disponibles en ligne.

On attend un notebook jupyter pour le projet, à remettre le 22 mars 2024 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 (mais pas obligatoire) de faire le projet en binôme.

2 Modèle de Lotka-Volterra

Lotka (1925) et Volterra (1926) ont formulé des équations différentielles ordinaires paramétriques qui caractérisent la dynamique des populations de prédateurs et de proies. Dans le modèle de Lotka-Volterra [2, 4], l'évolution des populations est régie par :

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}y_1(t)}{\mathrm{d}t} & = & y_1(t)\big(\alpha_1 - \alpha_2 y_2(t)\big), \\ \dfrac{\mathrm{d}y_2(t)}{\mathrm{d}t} & = & y_2(t)\big(\alpha_3 y_1(t) - \alpha_4\big), \end{array} \right. \tag{1}$$

où :
- $y_1(t)$, la population des proies en fonction du temps,
- $y_2(t)$, la population des prédateurs en fonction du temps.

Les variables $y_1(t)$ et $y_2(t)$ sont les sorties observées (avec du bruit) et la dynamique est caractérisée par les quatre paramètres suivants :
- α_1, taux de reproduction intrinsèque des proies,
- α_2, taux de mortalité des proies dû aux prédateurs rencontrés,
- α_3, taux de reproduction des prédateurs en fonction des proies rencontrées,

- α_4, taux de mortalité intrinsèque des prédateurs.
Les conditions initiales du système sont données en termes de la date initiale t_0 et des populations initiales de proies y_{10} et de prédateurs y_{20}.

En fait, les biologistes savent que le système de Lotka-Volterra est très -trop- simplifié, et que dans un système plus correct les paramètres $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^4$ devraient dépendre des nombres de proies et de prédateurs. Le système devrait en fait avoir la forme

$$\left\{ \begin{array}{rcl} \dfrac{\mathrm{d}y_1(t)}{\mathrm{d}t} & = & A(y_1(t))y_1(t) - B(y_1(t), y_2(t))y_2(t), \\ \dfrac{\mathrm{d}y_2(t)}{\mathrm{d}t} & = & a_5 B(y_1(t), y_2(t))y_2(t) - a_6 y_2(t). \end{array} \right. \tag{2}$$

Les biologistes s'accordent sur la forme de A :

$$A(y_1) = a_1(1 - a_2 y_1),$$

mais il existe dans la littérature (au moins) deux formes possibles pour la fonction B :

$$B(y_1, y_2) = \frac{a_3 y_1 y_2^{-m}}{1 + a_3 a_4 y_1 y_2^{-m}},$$

avec $m = 0$ ou 1, qui correspondent à deux mécanismes différents d'inetraction prédateur-proie (pour les lecteurs interessés : le modèle avec $m = 0$ est appelé Holling type II et sinon il est appelé Hassell-Varley-Holling [1]). Le but final de ce projet est de déterminer quel modèle est le plus approprié (d'après un jeu de données réelles).

Les paramètres $\boldsymbol{a} = (a_i)_{i=1}^6$, $\boldsymbol{y}_0 = (y_{10}, y_{20})$, t_0, constituent les paramètres d'entrée du modèle (pour $m = 0, 1$ fixé). Lorsque ces paramètres sont fixés, on obtient une trajectoire des variables de sortie $\boldsymbol{y}(t) = (y_1(t), y_2(t))$ par résolution du système (2) avec les conditions initiales $\boldsymbol{y}_0$ à t_0.

On va développer une approche bayésienne. On considère la loi a priori suivante pour $(\boldsymbol{a}, \boldsymbol{y}_0) \in \mathbb{R}^8$: Les paramètres a_i et y_{0j} sont indépendants, de lois log-normales. Plus exactement, $\ln a_i \sim \mathcal{N}(\ln \lambda_i, \zeta^2)$, $i = 1, \ldots, 6$, $\ln y_{0j} \sim \mathcal{N}(\ln(\lambda_{6+j}), \zeta^2)$, $j = 1, 2$, avec $\lambda_1 = 1$, $\lambda_2 = 0.01$, $\lambda_3 = 0.1$, $\lambda_4 = 0.1$, $\lambda_5 = 1$, $\lambda_6 = 1$, $\lambda_7 = \lambda_8 = 10$, $\zeta = 0.5 \ln(10)$. L'hyperparamètre ζ quantifie la concentration de la loi a priori autour du point central $(\lambda_i)_{i=1}^8$. On suppose aussi $t_0 = 0$.

Question 1 : Programmez la résolution du système (2) (l'unité de temps est le jour).
Remarque : Il y a des routines de résolution d'équations différentielles ordinaires dans python !
Par échantillonnage Monte Carlo de la loi a priori de $(\boldsymbol{a}, \boldsymbol{y}_0)$ (en prenant $\zeta = 0.25 \ln(10)$), tracez quelques trajectoires.

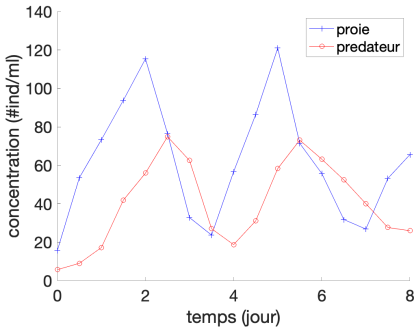

FIGURE 1 – Visualisation des données

3 Calibration

On va considérer des données qui concernent l'évolution des concentrations de protozoaires et arthropodes en culture mesurées toutes les douze heures. On reprend les données présentées dans [3] et utilisées à de nombreuses reprises dans la littérature (plus exactement, on prend les données correspondant à la figure 11a de [3]). On peut les trouver numérisées sur le site `https://robjhyndman.com/tsdldata/`. Elles sont accessibles sur le moodle du cours.

Les tailles des populations sont exprimées en nombres d'individus par millilitre et le temps est exprimé en jours. On note n la longueur de la série temporelle ($n = 17$). On note t_i, $i = 0, \dots, n-1$ les instants où les données sont recueillies. Les données ont la forme d'un tableau $\mathbf{data} = (\text{time}_i, \text{data}_{proie,i}, \text{data}_{preda,i})_{i=0}^{n-1}$.

Question 2 : Dessinez les données collectées (vous devez trouver quelque chose qui ressemble à la figure 1).

3.1 Calibration déterministe

On cherche à ajuster au mieux le modèle (2) avec $m = 0$ au sens des moindres carrés :

$$(\boldsymbol{a}^*, \boldsymbol{y}_0^*) = \text{argmin}_{\boldsymbol{a}, \boldsymbol{y}_0} \mathcal{E}(\boldsymbol{a}, \boldsymbol{y}_0),$$
$$\mathcal{E}(\boldsymbol{a}, \boldsymbol{y}_0) = \sum_{i=0}^{n-1} e_{proie,i}(\boldsymbol{a}, \boldsymbol{y}_0)^2 + e_{preda,i}(\boldsymbol{a}, \boldsymbol{y}_0)^2,$$

où les résidus sont définis par :

$$e_{proie,i}(\boldsymbol{a}, \boldsymbol{y}_0) = \ln \mathcal{M}_{proie}(t_i; \boldsymbol{a}, \boldsymbol{y}_0, t_0) - \ln \text{data}_{proie,i}, \tag{3}$$
$$e_{preda,i}(\boldsymbol{a}, \boldsymbol{y}_0) = \ln \mathcal{M}_{preda}(t_i; \boldsymbol{a}, \boldsymbol{y}_0, t_0) - \ln \text{data}_{preda,i}, \tag{4}$$

$\boldsymbol{\mathcal{M}}(t; \boldsymbol{a}, \boldsymbol{y}_0, t_0) = (\mathcal{M}_{proie}(t; \boldsymbol{a}, \boldsymbol{y}_0, t_0), \mathcal{M}_{preda}(t; \boldsymbol{a}, \boldsymbol{y}_0, t_0))$ est la prédiction à l'instant t des nombres de proies et de prédateurs par le modèle (2) avec les paramètres $\boldsymbol{a}$ et les données initiales $\boldsymbol{y}_0$ à t_0. Le fait de prendre le log sera expliqué dans la section suivante, cela correspond à supposer une erreur de type bruit multiplicatif.

Question 3 : Evaluez numériquement $(\boldsymbol{a}^*, \boldsymbol{y}_0^*)$ *dans le domaine de* $\mathbb{R}^8$ *déterminé par le support essentiel de la loi a priori (l'hypercube* $\prod_{j=1}^{8}[\lambda_j \exp(-2\zeta), \lambda_j \exp(2\zeta)]$ *avec* $\zeta = 0.5\ln(10)$*). Comparez sur une figure les données* $(\text{data}_{proie,i}, \text{data}_{preda,i})_{i=0}^{n-1}$ *et les prédictions* $(\boldsymbol{\mathcal{M}}(t; \boldsymbol{a}^*, \boldsymbol{y}_0^*, t_0))_{t\in[t_0,t_{n-1}]}$.
Remarques : Il y a des routines d'optimisation dans python ! Attention, la fonction $\mathcal{E}$ *possède des minima locaux !*

3.2 Calibration bayésienne

Le modèle statistique est le suivant :

- On connaît la loi a priori des paramètres $\boldsymbol{x} = (\boldsymbol{a}, \boldsymbol{y}_0) \in \mathbb{R}^8$.

- Pour évaluer la vraisemblance, on suppose un modèle statistique prenant en compte une erreur de mesure multiplicative, de loi log-normale. On considère alors que les observations ont la forme

$$\text{data}_{proie,i} = \mathcal{M}_{proie}(t_i; \boldsymbol{a}, \boldsymbol{y}_0, t_0) \exp\left(\epsilon_{proie,i}\right), \tag{5}$$
$$\text{data}_{preda,i} = \mathcal{M}_{preda}(t_i; \boldsymbol{a}, \boldsymbol{y}_0, t_0) \exp\left(\epsilon_{preda,i}\right), \tag{6}$$

pour $i = 0, \ldots, n-1$, où $\epsilon_{proie,i} \sim \mathcal{N}(0, \sigma_{proie}^2)$, $\epsilon_{preda,i} \sim \mathcal{N}(0, \sigma_{preda}^2)$ sont indépendantes en i et entre elles. On obtient ainsi une expression de la vraisemblance $p(\mathbf{data}|\boldsymbol{x}, \boldsymbol{\sigma})$, $\mathbf{data} = (\text{data}_{proie,i}, \text{data}_{preda,i})_{i=0}^{n-1}$, $\boldsymbol{\sigma} = (\sigma_{proie}, \sigma_{preda})$:

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

où $e_{proie,i}$ et $e_{preda,i}$ sont définis par (3-4). 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 des hyper-paramètres 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, avec $\boldsymbol{x}^* = (\boldsymbol{a}^*, \boldsymbol{y}_0^*)$ déterminé dans la question 3 (c'est ce qu'on va faire dans cette section).
- Une approche full-bayésienne où $\boldsymbol{\sigma}$ suit une loi a priori peu informative (vous pouvez vous lancer si vous vous sentez d'attaque !).

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 4 : Vérifiez qu'on a $(\sigma^*_{proie})^2 = \frac{1}{n}\sum_{i=0}^{n-1} e_{proie,i}(\boldsymbol{x}^*)^2$ *et* $(\sigma^*_{preda})^2 = \frac{1}{n}\sum_{i=0}^{n-1} e_{preda,i}(\boldsymbol{x}^*)^2$.

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 système (2) dans la vraisemblance. Par conséquent, nous devons recourir à des algorithmes d'échantillonnage. Ici, nous suggérons d'utiliser un algorithme de type Metropolis-Hastings adaptatif.

Question 5 : 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* $\boldsymbol{a}$.

3.3 Calibration avec erreur de modèle

On a considéré dans la section 3.2 un modèle d'erreur de type bruit multiplicatif. Cela correspond à supposer que la vraie dynamique suit le système (2) pour un certain paramètre $\boldsymbol{a}$ inconnu et qu'une erreur d'observation (de type erreur multiplicative) est commise à chaque instant d'observation. On pourrait aussi considérer qu'il n'y a pas d'erreur d'observation, mais que la vraie dynamique ne suit pas le système (2) et qu'une erreur de modèle s'ajoute à chaque pas de temps. Ainsi, pour chaque pas de temps t_i, l'observation est de la forme (5-6) mais avec une prédiction calculée avec une condition initiale en $t = t_{i-1}$ égale à l'observation $\mathbf{data}_{i-1}$:

$$\text{data}_{proie,i} = \mathcal{M}_{proie}(t_i; \boldsymbol{a}, \mathbf{data}_{i-1}, t_{i-1}) \exp\left(\epsilon_{proie,i}\right), \tag{7}$$

$$\text{data}_{preda,i} = \mathcal{M}_{preda}(t_i; \boldsymbol{a}, \mathbf{data}_{i-1}, t_{i-1}) \exp\left(\epsilon_{preda,i}\right), \tag{8}$$

pour $i = 1, \ldots, n-1$. On suppose que les erreurs de modèle $\epsilon_{proie,i} \sim \mathcal{N}(0, \sigma^2_{proie})$, $\epsilon_{preda,i} \sim \mathcal{N}(0, \sigma^2_{preda})$ sont indépendantes en i et entre elles.

On peut alors calibrer le système en considérant et minimisant les erreurs

$$e_{proie,i}(\boldsymbol{a}) = \ln \mathcal{M}_{proie}(t_i; \boldsymbol{a}, \mathbf{data}_{i-1}, t_{i-1}) - \ln \mathrm{data}_{proie,i},$$
$$e_{preda,i}(\boldsymbol{a}) = \ln \mathcal{M}_{preda}(t_i; \boldsymbol{a}, \mathbf{data}_{i-1}, t_{i-1}) - \ln \mathrm{data}_{preda,i},$$

pour $i = 1, \ldots, n-1$. Notez que les données initiales à t_0 sont connues (car il n'y a pas de bruit de mesure), donc elles n'entrent pas dans le vecteur de paramètres à calibrer.

Question 6 : Reprenez la calibration bayésienne de la section 3.2 avec la structure d'erreur de modèle décrite ci-dessus.

3.4 Détermination du meilleur modèle

L'idée est de reprendre l'étude de la section 3.2 ou 3.3 en ajoutant le paramètre m à valeurs dans $\{0, 1\}$ dans le vecteur $\boldsymbol{x}$.

Question 7 : Reprenez la calibration bayésienne avec la structure d'erreur d'observation multiplicative (ou avec la structure d'erreur de modèle) en ajoutant le paramètre m à valeurs dans $\{0, 1\}$ dans le vecteur $\boldsymbol{x}$ (on pourra prendre comme loi a priori pour m la loi de Bernoulli de paramètre $1/2$). Estimez la loi a posteriori de m (c'est toujours une loi de Bernoulli !).

Références

[1] Jost, C. et Arditi, R. (2001). From pattern to process : identifying predator-prey models from time-series data. Popul. Ecol. 43. 229-243 .

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

[3] Veilleux, B.G. (1976). The analysis of a predatory interaction between Didinium and Paramecium. Master's thesis. University of Alberta, Edmondton.

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