TP: Estimation de processus stochastiques
Master SSD (2019-2020)

Olivier Zahm*

Dans ce TP, on apprendra à estimer les fonctions de covariance de processus gaussiens en calculant leurs variogrammes. Dans un deuxième temps, on réutilisera cette estimation afin de générer des données manquantes, comme illustré en Figure 1.

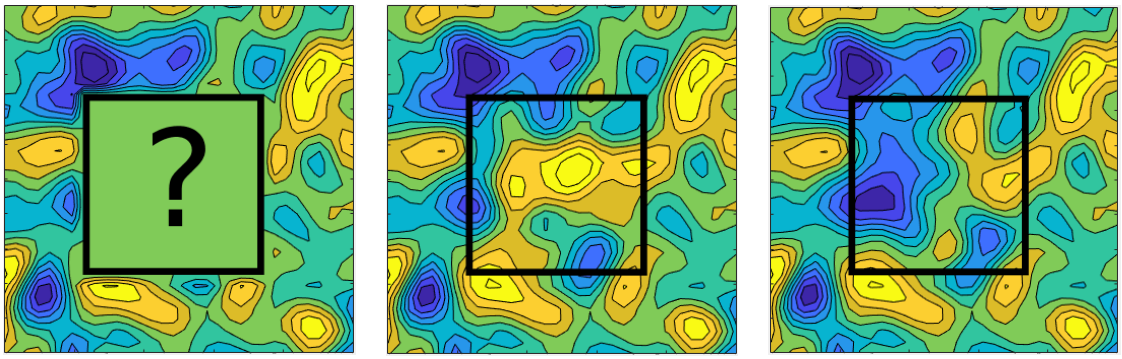

Figure 1: Le but final du TP est de régénérer des données manquantes.

Consignes: Vous travaillerez en binôme en utilisant le langage de programmation de votre choix (R, Python, Matlab...). Les consignes sont assez ouvertes : n'hésitez pas à prendre des initiatives ! Les compte rendus de TP sont à envoyer au plus tard **jeudi 16 janvier 23h59** à l'adresse `olivier.zahm@inria.fr`. N'envoyez qu'un seul document au format pdf **avec le code en annexe**.

*olivier.zahm@inria.fr

1 Variogramme

Soit X_s un processus spatial sur $S = [0,1]^2$. On suppose que X_s est stationnaire du second ordre et de moyenne nulle. Son variogramme $\gamma : \mathbb{R}^2 \to \mathbb{R}$ est une fonction définie par

$$\gamma(h) = \frac{1}{2}\,\mathbb{E}[(X_{s+h} - X_s)^2],$$

pour tout $h \in \mathbb{R}^2$. On rappelle que, par stationnarité de X, cette quantité est indépendante de s. On rappelle aussi que le variogramme est relié à la fonction de covariance $c(h) = \mathrm{Cov}(X_{s+h}, X_s) = \mathbb{E}[X_{s+h}X_s]$ par la relation suivante :

$$\gamma(h) = \frac{1}{2}\left(\underbrace{\mathbb{E}[(X_{s+h})^2]}_{=c(0)} - 2\underbrace{\mathbb{E}[X_{s+h}X_s]}_{c(h)} + \underbrace{\mathbb{E}[(X_s)^2]}_{=c(0)}\right) = c(0) - c(h).$$

On suppose de plus que X_s est **isotrope**, c'est à dire que c (et donc h) n'est fonction que de la norme de h. Avec un abus de notation, on écrira

$$\gamma(h) = \gamma(\|h\|) = \gamma(r),$$

avec $\gamma : \mathbb{R} \to \mathbb{R}$ une fonction d'une seule variable et $r = \|h\|$. Intuitivement, l'hypothèse d'isotropie veut dire que les propriétés de X_s sont les mêmes après rotation de l'espace (cf TP précédent).

1.1 Estimation du variogramme

On observe une réalisation de X_s sur une grille régulière de points $s \in \{s_1, \ldots, s_{n^2}\} \subset S$ de $n = 32$ points par dimension :

$$\begin{array}{ccccc}
 & | & | & | & | & \\
- & s_{(n+1)n+1} & s_{(n+1)n+2} & \cdots & s_{n^2} & - \\
- & \vdots & \vdots & \vdots & \vdots & - \\
- & s_{n+1} & s_{n+2} & \cdots & s_{2n} & - \\
- & s_1 & s_2 & \cdots & s_n & - \\
 & | & | & | & | &
\end{array}$$

À partir de cette seule réalisation, on peut estimer le variogramme par

$$\gamma(r) \approx \frac{1}{2N_\epsilon(r)} \sum_{(s_i,s_j)\in S_\epsilon(r)} \left(X_{s_i} - X_{s_j}\right)^2,$$

avec

$$S_\epsilon(r) = \Big\{(s_i, s_j) : r - \varepsilon \leq \|s_i - s_j\| \leq r + \varepsilon\Big\}.$$

et $N_\epsilon(r) = \#S_\epsilon(r)$ (remarque: on fait ici une hypothèse d'ergodicité de X_s qui permet d'estimer des espérances par des moyennes spatiales). Noter que $\varepsilon > 0$ est un paramètre à choisir le plus petit possible, mais assez gros pour avoir sélectionné suffisamment d'échantillons $N_\epsilon(r)$...

Tâche 1 Pour chaque jeu de donnée qui vous est fourmis (fichiers `data_1.mat`, `data_2.mat` et `data_3.mat`) et pour chaque réalisation `X_1`$, \ldots,$ `X_10` qu'il contient, tracer le variogramme pour $0 \leq r \leq 0.5$. Tracer ensuite la moyenne de ces 10 variogrammes $\bar{\gamma}(r) = \frac{1}{10}\sum_{i=1}^{10} \gamma^i(r)$ pour obtenir une meilleure estimation de $\gamma(r)$. Comment choisir ε ?

Tâche 2 Les données `data_1.mat`, `data_2.mat` et `data_3.mat` ont été obtenues à partir des noyaux suivants:

- Gaussien :
$$c(r) = \sigma^2 \exp\left(-\frac{r^2}{2a^2}\right)$$
- Sphérique + delta :
$$c(r) = \sigma_0\delta(r) + \begin{cases} \frac{2\sigma^2}{\pi}\left(\arccos\left(\frac{r}{a}\right) - \frac{r}{a}\sqrt{1-\frac{r^2}{a^2}}\right) & \text{si } r \leq a \\ 0 & \text{sinon} \end{cases}$$
avec $\delta(r) = 1$ si $r = 0$ et $\delta(r) = 0$ sinon.
- Matérn-3/2 :
$$c(r) = \sigma^2\left(1 + \sqrt{3}\frac{r}{a}\right)\exp\left(-\sqrt{3}\frac{r}{a}\right)$$

Retrouver qui est quoi, et identifier au mieux les paramètres a, σ, σ_0.

2 Régénérer une image (facultatif)

Chaque fichier `data_1.mat`, `data_2.mat` et `data_3.mat` contient en plus une réalisation incomplète `X_incomplet` dont les entrées (i, j) tel que $8 \leq i, j \leq 24$ sont manquantes (elles ont été mises à 0). Dans la partie précédente du TP, on aura déjà identifié la fonction de covariance du processus : grâce à cela, on pourra régénérer cette partie manquante.

Indications : on aura besoin de tirer un vecteur gaussien X conditionné à ce que la partie observée soit égale aux observations. Pour cela, on devra découper le vecteur X en deux parties
$$X = \begin{pmatrix} X_1 \\ X_2, \end{pmatrix}$$

avec X_1 un vecteur contenant les valeurs manquantes (le point d'interrogation sur la Figure 1) et X_2 un vecteur contenant les valeurs connues (le reste observé de l'image). On découpera la matrice de covariance $\Sigma_{i,j} = c(s_i, s_j)$ de façon similaire :

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Enfin, on rappelle que le vecteur X_1 conditionné à l'évènement $X_2 = x_{\text{données observées}}$ est un vecteur gaussien de loi

$$X_1 | X_2 = x_{\text{données observées}} \sim \mathcal{N}(m, \bar{\Sigma})$$

avec

$$m = \Sigma_{12}\Sigma_{22}^{-1} x_{\text{données observées}}$$

et

$$\bar{\Sigma} = \Sigma_{11} - \Sigma_{21}\Sigma_{22}^{-1}\Sigma_{21}.$$

Après avoir réalisé un tirage de $X_1 | X_2 = x_{\text{données observées}}$, il suffira de compléter le vecteur $X = (X_1, X_2)$ et de le plotter comme en Figure 1.

3 Accéder aux données

Chaque fichier `data_1.mat`, `data_2.mat` et `data_3.mat` contient

- deux vecteurs $x \in \mathbb{R}^n$ et $y \in \mathbb{R}^n$ qui contiennent les coordonnées des $n = 32$ points par dimensions,
- dix réalisation du processus `X_1`, ..., `X_10` évalués sur la grille $x \times y$ (les `X_i` sont des matrices de taille $n \times n$)
- une réalisation incomplète `X_incomplet` : les entrée manquantes de `X_incomplet` sont tous les couples (i, j) tel que $8 \leq i, j \leq 24$.

Avec Python. Il suffit de faire :

```
import scipy.io
data = scipy.io.loadmat('data_1.mat')
x = data['x']
y = data['y']
X_1 = data['X_1']
X_2 = data['X_2']
...
X_10 = data['X_10']
X_incomplet = data['X_incomplet']
```

Avec R. On commence par installer le package R.matlab avec la commande `install.packages("R.matlab")`. Puis :

```
library(R.matlab)
data <- readMat('data_1.mat')
x = data$x
y = data$y
X_1 = data$X.1
X_2 = data$X.2
...
X_10 = data$X.10
X_incomplet = data$X.incomplet
```