Estimation par maximum de vraisemblance par paires de champs gaussiens multivariés spatio-temporels. Application à une fonction de covariance entièrement non séparable.

Marc Bourotte [1] & Denis Allard [2]

[1] *INRA BioSP Avignon et bourotte@paca.inra.fr*
[2] *INRA BioSP Avignon et allard@paca.inra.fr*

Résumé. Lors de l'analyse de données spatio-temporelles, le statisticien cherche à modéliser les liens directs et croisés entre le temps, l'espace et les différentes variables dans un but d'estimation, de prédiction ou de simulation. Dans un cadre gaussien, cela revient à proposer des modèles pertinents de covariance qui assurent à toute matrice de covariance issue de ce modèle d'être semi-définie positive.

On construit facilement des modèles valides en utilisant la propriété de séparabilité. Dans ce cas, une matrice de covariance Σ issue de ce processus est simplement le produit de Kronecker d'une matrice de covariance temporelle, d'une matrice de covariance spatiale et une matrice de corrélation. Par rapport au cas général, la séparabilité conduit à réduire le nombre de paramètres et permet de calculer plus rapidement l'inverse et le déterminant de la matrice Σ. Cependant c'est une hypothèse qui peut être trop simpliste pour certains jeux de données comme les données climatiques.

Nous proposons une famille paramétrique de fonctions de covariances croisées entièrement non séparables pour les champs aléatoires multivariés spatio-temporels. Néanmoins, proposer un modèle valide de covariance croisée n'est pas la seule difficulté. En effet, estimer l'ensemble des paramètres de la fonction de covariance croisée est une tâche importante et délicate. L'approche par maximum de vraisemblance classique fonctionne bien mais devient rapidement inutilisable lorsque le nombre d'observations dépasse quelques milliers de données. Dans ce cas, une stratégie consiste à maximiser la vraisemblance composite et notamment la vraisemblance par paires.

Dans ce travail, nous utilisons la vraisemblance par paires pour inférer les paramètres d'une fonction de covariance entièrement non séparable. Nous présenterons les difficultés rencontrées dans la procédure d'estimation et les solutions proposées.

Mots-clés. Champ gaussien aléatoire, fonction de covariance, vraisemblance par paires, statistiques spatio-temporelles, non séparabilité

Abstract. Multivariate space-time data are increasingly recorded in various scientific disciplines. When analyzing these data, one of the (geo)statistician's goal is thus to model

the multivariate space-time dependencies. In a Gaussian framework, this necessitates to propose relevant models for multivariate space-time covariance functions, mathematically described as matrix-valued covariance functions for which nonnegative definiteness must be ensured.

Straightforward nonnegative definite models can be built using separability property. It follows that any covariance matrix Σ is simply the Kronecker product of a purely temporal covariance matrix, a purely spatial covariance matrix and a correlation matrix. Thus compared to general forms of Σ, separability leads to reduced number of parameters and faster computation of the inverse and of the determinant of the matrix. However, it is in many cases an overly simplified assumption for climate data.

Consequently, we have proposed a fully non separable parametric class of cross-covariance functions for multivariate spatio-temporal random fields. Ensuring the nonnegative definiteness is not the only difficulty. Estimating all the parameters of the covariance function is an important and delicate task. Classical maximum likelihood approach works well but becomes impractical when the number of observations is very large, more than a few thousands of data. In this case, a solution consists in maximizing the composite likelihood, in particular the pairwise likelihood.

In this work we use the pairwise marginal Gaussian likelihood with a fully non separable covariance function. We will present the difficulties encountered in estimation procedure on a French dataset of climatic variables and the solutions proposed.

Keywords. Gaussian random field, multivariate spatio-temporal covariance model, composite likelihood, non separability

1 Présentation du modèle

Considérons un champ aléatoire multivarié à p dimensions $\mathbf{Z}(\mathbf{s},t) = \{Z_1(\mathbf{s},t), ..., Z_p(\mathbf{s},t)\}^\top$, où $(\mathbf{s},t) \in \mathbb{R}^{d+1}, d \geq 1$. En supposant le champ stationnaire et gaussien, le processus $\mathbf{Z}(\mathbf{s},t)$ est entièrement caractérisé par sa fonction de matrice de covariance $\mathbf{C}(\mathbf{h},u) = [C_{ij}(\mathbf{h},u)]_{i,j=1}^p$ qui dépend uniquement du lag spatio-temporel $\mathbf{k} = (\mathbf{h},u) \in \mathbb{R}^{d+1}$

$$\text{Cov}\left\{Z_i(\mathbf{s},t), Z_j(\mathbf{s}+\mathbf{h}, t+u)\right\} = C_{ij}(\mathbf{h},u), \qquad i,j = 1,\ldots,p.$$

Nous proposons un modèle paramétrique de fonction de covariance entièrement non séparable pour le processus $\mathbf{Z}$. Ce modèle est valide c'est-à-dire que pour tout $n \in \mathbb{N}$, pour tout ensemble de points spatio-temporels $(\mathbf{s}_1,t_1), ..., (\mathbf{s}_n,t_n)$ et pour tout vecteur $\boldsymbol{\lambda} \in \mathbb{R}^{np}$, nous avons $\boldsymbol{\lambda}\boldsymbol{\Sigma}\boldsymbol{\lambda} \geq 0$, où $\boldsymbol{\Sigma}$ est la matrice $np \times np$ composée de blocs $\mathbf{C}(\mathbf{s}_\alpha - \mathbf{s}_\beta, t_\alpha - t_\beta)$, avec $\alpha, \beta = 1, ..., n$.

A l'instar du modèle multivarié spatial défini dans Apanasovich et al. (2012), chaque variable a sa propre régularité ν et son propre paramètre d'échelle r (ou $1/r$ la portée). Le

modèle s'appuie sur la fonction de covariance univariée spatio-temporelle proposée dans Gneiting (2002), et sur les covariances de Matérn que nous noterons

$$M(\mathbf{h}|\nu, r) = \frac{2^{1-\nu}}{\Gamma(\nu)} \left(r\|\mathbf{h}\|\right)^{\nu} \mathcal{K}_{\nu}\left(r\|\mathbf{h}\|\right), \quad \mathbf{h} \in \mathbb{R}^d. \tag{1}$$

La fonction de covariance croisée utilisée dans l'application s'écrit :

$$\begin{aligned} C_{ij}(\mathbf{h}, u) =& \frac{\sigma_i \sigma_j \beta_{ij}}{a|u|^{2\alpha}+1} \frac{\Gamma\left\{(\nu_i+\nu_j)/2\right\}}{\Gamma(\nu_i)^{1/2}\Gamma(\nu_j)^{1/2}} \frac{r_i^{\nu_i} r_j^{\nu_j}}{\left\{(r_i^2+r_j^2)/2\right\}^{(\nu_i+\nu_j)/2}} \\ & \times \; M\left(\frac{\mathbf{h}}{(a|u|^{2\alpha}+1)^{\beta/2}} | \frac{\nu_i+\nu_j}{2}, \sqrt{\frac{r_i^2+r_j^2}{2}}\right), \end{aligned} \tag{2}$$

avec $i, j = 1, ..., p$ et $(\mathbf{h}, u) \in \mathbb{R}^2 \times \mathbb{R}$. Ici, β représente le paramètre de séparabilité. Lorsque $\beta = 0$ le modèle est entièrement séparable.

Le modèle (1) possède $(p^2 + 5p + 6)/2$ paramètres à estimer. Nous avons utilisé la vraisemblance par paires car le nombre de données augmente rapidement avec le nombre d'instants de mesure, de sites et de variables observées. On minimise l'opposé de la log-vraisemblance par paires qui s'écrit, en reprenant les notations précédentes,

$$\sum_{i}^{p}\sum_{j}^{p}\sum_{\alpha}^{n}\sum_{\beta>\alpha}^{n} l_{ij}(\mathbf{s}_\alpha, \mathbf{s}_\beta, t_\alpha, t_\beta; \theta) = \sum_{i}^{p}\sum_{j}^{p}\sum_{\alpha}^{n}\sum_{\beta>\alpha}^{n} -\frac{1}{2}\left\{\log|S| + (z_{i,\alpha}\ z_{j,\beta})S^{-1}(z_{i,\alpha}\ z_{j,\beta})^T\right\} \tag{3}$$

où $S = \begin{pmatrix} C_{ii}^{(Z)}(\mathbf{0}, 0) & C_{ij}^{(Z)}(\mathbf{h}, u) \\ C_{ij}^{(Z)}(\mathbf{h}, u) & C_{jj}^{(Z)}(\mathbf{0}, 0) \end{pmatrix}$. Une pondération peut être utilisée car les paires d'observations trop éloignées spatio-temporellement sont non informatives. Nous avons utilisé une pondération simple (0 ou 1) qui permet ainsi d'alléger le temps de calcul. Les propriétés asymptotiques des estimateurs sont présentées dans Bevilacqua & Gaetan (2014).

2 Application

Nous avons modélisé trois variables climatiques journalières (rayonnement solaire, température moyenne, humidité) par un champ gaussien aléatoire. Ces données ont été mesurées dans l'ouest de la France par des stations INRA. Nous les avons standardisées à l'aide d'une moyenne et d'un écart-type calculés pour chaque site, chaque variable et chaque mois afin d'assurer la stationnarité du champ aléatoire.

Un des soucis rencontrés concerne la variable température. En effet, comme on peut l'observer sur la figure 1, le palier du variogramme de la température n'atteint pas la variance (ici égale à 1) comme pour les autres variables. Le même type d'observation peut être faite sur la figure 2 avec la représentation en covariance. Ceci peut être dû à un processus purement temporel et nous avons décidé d'utiliser la modélisation suivante pour $i = 1, 2, 3$:

$$Z_i(\mathbf{s}, t) = X_i(t) + Y_i(\mathbf{s}, t). \tag{4}$$

Ainsi pour $i = j$ et $u = 0$ on peut écrire :

$$C_{ii}^{(Z)}(\mathbf{h}, 0) = C_{ii}^{(X)}(0) + C_{ii}^{(Y)}(\mathbf{h}, 0) \underset{\|\mathbf{h}\| \to \infty}{\longrightarrow} C_{ii}^{(X)}(0) = \left[\sigma_i^{(X)}\right]^2. \tag{5}$$

La covariance du processus spatio-temporel multivarié $\mathbf{Y}$ a été modélisée à l'aide du modèle (1) tandis que celle du processus temporel multivarié $\mathbf{X}$ a été modélisée par :

$$C_{ij}^{(X)}(\mathbf{h}, u) = \sigma_i^{(X)} \sigma_j^{(X)} \beta_{ij}^{(X)} \left(a' |u|^{2\alpha'} + 1\right)^{-1}. \tag{6}$$

Les figures 1 à 4 montrent l'adéquation du modèle de covariance proposé avec les paramètres estimés par rapport aux données. L'estimation a été effectuée à l'aide du maximum de vraisemblance par paires. Nous avons utilisé une méthode profilée à cause du nombre de paramètres (23). Ici, la pondération vaut 0 pour $|u| > 3$jours ou $\|h\| > 300$ kilomètres.

Bibliographie

[1] Apanasovich, Tatiyana V and Genton, Marc G and Sun, Ying (2012), A valid Matérn class of cross-covariance functions for multivariate random fields with any number of components, *Journal of the American Statistical Association*, 497, 180–193.
[2] Gneiting, Tilmann (2002), Nonseparable, stationary covariance functions for space–time data, *Journal of the American Statistical Association*, 458, 590–600.
[3] Bevilacqua, Moreno and Gaetan, Carlo (2014), Comparing composite likelihood methods based on pairs for spatial Gaussian random fields, *Statistics and Computing*, 1-16.

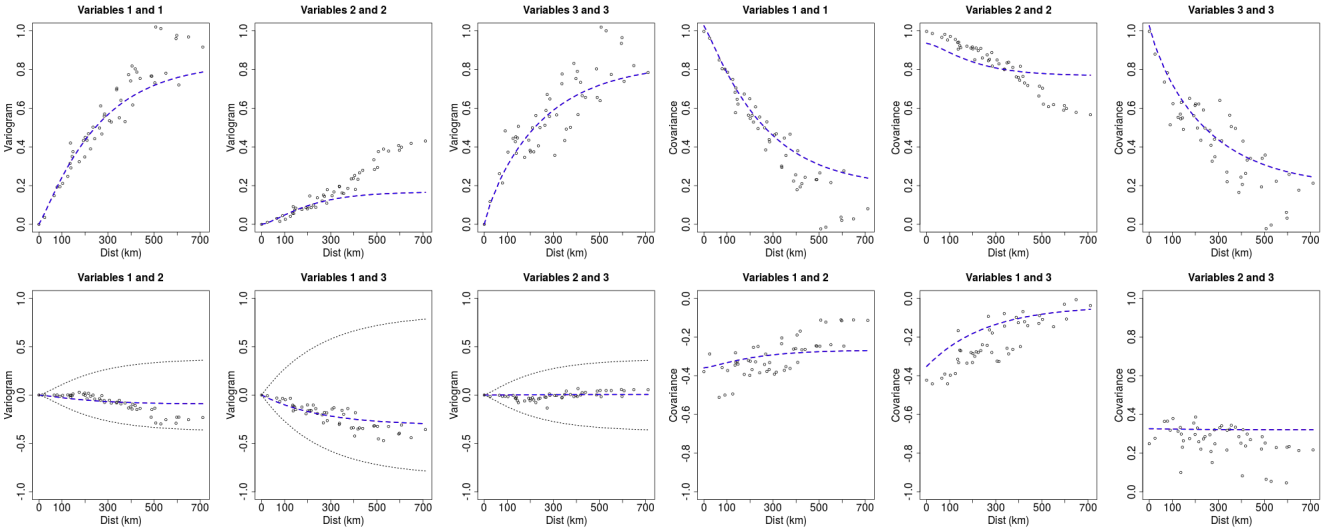

Figure 1: *Variogrammes directs et croisés pour $u = 0$. Variable 1 rayonnement solaire, variable 2 température, variable 3 humidité.*

Figure 2: *Covariances directes et croisées pour $u = 0$. Variable 1 rayonnement solaire, variable 2 température, variable 3 humidité.*

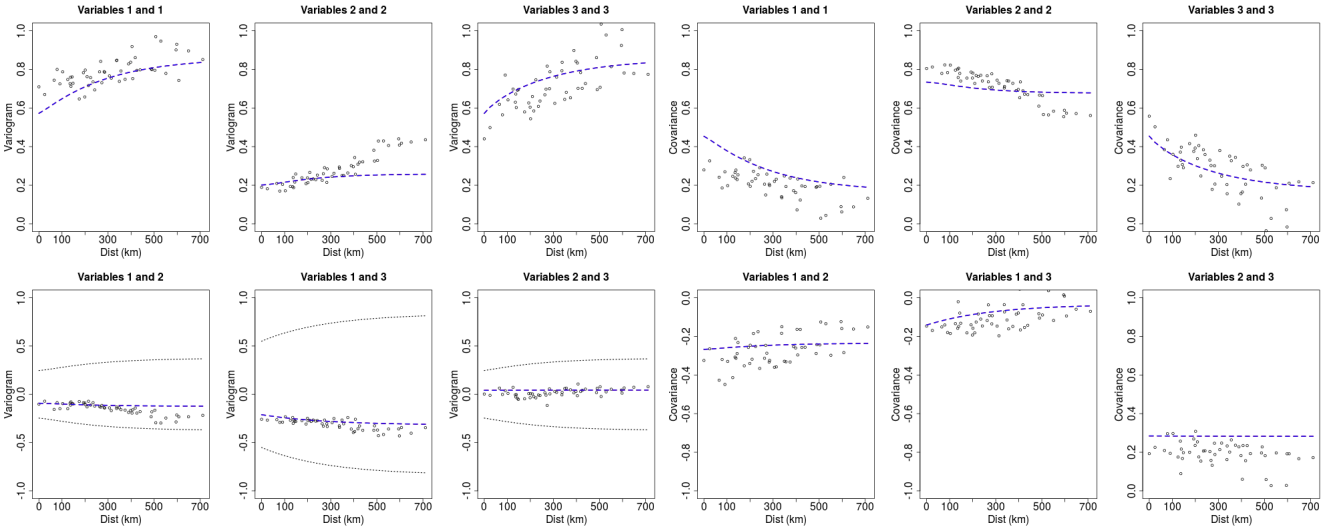

Figure 3: *Variogrammes directs et croisés pour $u = 1$. Variable 1 rayonnement solaire, variable 2 température, variable 3 humidité.*

Figure 4: *Covariances directes et croisées pour $u = 1$. Variable 1 rayonnement solaire, variable 2 température, variable 3 humidité.*