Modèle spatio-temporel pour des variables mixtes

Carlo Gaetan, Cécile Hardouin

Modèle spatio-temporel pour des variables mixtes

Carlo Gaetan & Cécile Hardouin

SAMOS-MATISSE/Centre d'Economie de la Sorbonne, Université Paris 1
Dipartimento di statistica, Universita Ca'Foscari, Venezia, Italy

Abstract

In several application fields like daily pluviometry data modeling, or motion analysis from image sequences, observations contain two components of different nature. A first part is made with zero values accounting for absence of some phenomenon, and a second part records a continuous (real-valued) measurement. We call such type of observations " mixed-state observations". To deal with data of mixed nature, most of the existing approaches rely on an hierarchical approach. We explore a different approach based on a Markov Chain of Markov fields modelling, the Markovian fields being defined as mixed state auto-models whose local conditional distributions belong to an exponential family and the observations derive from mixed states variables.

Résumé

Nous étudions la modélisation de données spatio-temporelles de nature mixte ; c'est-à-dire que les données sont composées de valeurs discrètes et continues. Dans la plupart des cas, il s'agit de zéros accompagnés de valeurs positives. Ce phénomène est souvent rencontré dans de nombreux domaines. Par exemple en pluviométrie, on mesure la quantité de pluie pendant des périodes, suivies de zéros lorsqu'il ne pleut pas.

Nous proposons ici une approche non hiérarchique pour la modélisation de ce type de données. Nous considérons une chaîne de Markov dans le temps de champs spatiaux de type auto-modèles markoviens specifiés par leurs lois conditionnelles. A titre d'exemple, nous détaillerons un modèle où la composante continue des données mixtes est modélisée par une loi exponentielle, et présentons quelques propriétés de ce modèle (ergodicité, estimation, coopération spatiale). Enfin, nous présentons une application sur des données radars de pluie.

Mots clés: variable à états mixtes, auto-modèle, champ markovien

1 Variable à états mixtes

Les variables aléatoires à états mixtes ont été introduites par Hardouin et Yao (2008). Nous rappelons brièvement leur construction. Pour simplifier, on suppose que l'espace

d'états est $E = \{0\} \cup (0, \infty)$, mais les résultats existent pour des espaces plus généraux. La mesure sur E est définie par

$$m(dx) = \delta_0(dx) + \lambda(dx) \ ,$$

où δ_0 est la mesure de Dirac en zéro et λ la mesure de Lebesgue sur $(0, \infty)$.

Une variable X définie sur E prend la valeur 0 avec probabilité $\gamma \in]0, 1[$, et est strictement positive avec probabilité $1- \gamma$. On suppose que la densité g de X sur $]0, \infty[$ est dans une famille exponentielle de dimension s:

$$g_\xi(x) = H(\xi)L(x) \exp\langle \xi, T(x)\rangle \ , \qquad \xi \in \mathbb{R}^l \ , \quad T(x) \in \mathbb{R}^l$$

Alors la densité f de X sur E est encore dans une famille exponentielle, de dimension $s + 1$, par rapport à $m(dx)$:

$$\begin{aligned} f_\theta(x) &= \gamma\delta(x) + (1 - \gamma)\delta^*(x)g_\xi(x) \\ &= H'(\theta)L'(x) \exp\langle \theta, B(x)\rangle \end{aligned} \tag{1}$$

où on a posé $\delta^*(x) = 1 - 1_{\{0\}}(x)$ et avec $H'(\theta) = \gamma, \quad L'(x) = \exp\{\delta^*(x) \ln L(x)\}$.

Le paramètre naturel et la statistique exhausitive sont obtenus par

$$\theta = \begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} = \begin{pmatrix} \ln \frac{(1-\gamma)H(\xi)}{\gamma} \\ \xi \end{pmatrix} \ , \quad B(x) = \begin{pmatrix} \delta^*(x) \\ \delta^*(x)T(x) \end{pmatrix} \ .$$

Exemple : si g est la densité de la loi exponentielle, on obtient $f_\theta(x) = H'(\theta) \exp\langle \theta, B(x)\rangle$ avec

$$\theta = \begin{pmatrix} \ln \frac{(1-\gamma)\lambda}{\gamma} \\ \lambda \end{pmatrix} \ , \quad B(x) = \begin{pmatrix} \delta^*(x) \\ -x\delta^*(x) \end{pmatrix} \ .$$

2 Chaîne de Markov de champs de Markov

Nous utilisons les résultats sur la généralisation des auto-modèles de Besag d'une part (Cf. Besag (1974)) et Hardouin, Yao (2008)), et la dynamique chaîne de Markov de champs de Markov décrite dans Hardouin, Guyon (2002) d'autre part.

Soit S un lattice régulier, et $X = \{X(t), \ t \in \mathbb{N}^*\}$ une chaîne de Markov sur $E^S = (\{0\} \cup (0, \infty))^S$. Pour simplifier, nous considérons une chaîne d'ordre 1. Chaque $X(t) = \{X_i(t), \ i \in S\}$ est, conditionnellement à $X(t - 1)$, un champ de Markov sur E^S ; plus précisément, à instant fixé, le champ markovien $X(t)$ est défini comme un auto-modèle à états mixtes (Cf. Hardouin, Yao (2008)). Ce champ est spécifié par ses lois conditionnelles, que nous prenons dans la famille de lois mixtes (1), et est défini sous les hypothèses suivantes.

Notons d'une part $X_i(t) = y_i$ et $X_i(t-1) = x_i$ pour dissocier passé et présent, d'autre part $y^i = \{y_j, \ j \notin A\}$.

[B1] $X = \{X(t),\ t \in \mathbb{N}^*\}$ est une chaîne homogène d'ordre 1 sur E^S, de transition

$$P(x, y) = Z^{-1}(x)\ \exp Q(y \mid x) \tag{2}$$

où $Z(x) = \int_{E^S} \exp Q(y \mid x)\ m^{\otimes S}(dy) < \infty$.

[B2] Pour chaque i, conditionnellement à $(X^i(t) = y^i\ ,\ X(t-1) = x)$, la loi de $X_i(t)$ est

$$\ln \mu_i(y_i \mid y^i, x) = \langle \theta_i(y^i, x), B_i(y_i) \rangle + C_i(y_i) + D_i(y^i, x) \tag{3}$$

où $\theta_i(y^i, x) \in \mathbb{R}^d,\ B_i(y_i) \in \mathbb{R}^d$ avec $B_i(0) = C_i(0) = 0$ pour tout $i \in S$.

[B3] *Pairwise only dependence* : conditionnellement au passé, l'énergie $Q(y \mid x)$ ne s'écrit que sur des potentiels de singletons et de paires.

$$Q(y \mid x) = \sum_{i \in S} G_i(y_i \mid x) + \sum_{\{i,j\}} G_{ij}(y_i, y_j \mid x)\ .$$

avec, presque sûrement en x, $G_i(0 \mid x) = G_{ij}(0, y_j \mid x) = G_{ij}(y_i, 0 \mid x) = 0$ pour tout $i \in S$, $j \in S$.

[B4] Pour tout $i \in S$, Span $\{B_i(y_i),\ y_i \in E\} = \mathbb{R}^d$.

Proposition 2.1 *Supposons que la chaîne X satisfait la condition [B1], et que l'énergie conditionnelle $Q(y \mid x)$ satisfait [B2], [B3], et [B4]. Alors, conditionnellement à $X(t-1) = x$, il existe une famille de vecteurs de dimension d $\{\alpha_i(x),\ i \in S\}$ et une famille de matrices $d \times d$ $\{\beta_{ij}(x),\ i, j \in S,\ i \neq j\}$ verifiant $\beta_{ij}(x) = \beta_{ij}(x)^T$ telles que*

$$\theta_i(y^i, x) = \alpha_i(x) + \sum_{j : \{i,j\}} \beta_{ij}(x) B_j(y_j)$$

De plus, les potentiels s'écrivent

$$\begin{aligned} G_i(y_i \mid x) &= \langle \alpha_i(x), B_i(y_i) \rangle + C_i(y_i) \\ G_{ij}(y_i, y_j \mid x) &= B_i(y_i)^T \beta_{ij}(x) B_j(y_j) \end{aligned}$$

Ce résultat donne la forme nécessaire des paramètres naturels $\theta_i(y^i, x)$. Nous allons voir que la spécification plus détaillée du modèle va induire des contraintes sur les paramètres des vecteurs α_i et matrices β_{ij}. De plus, quel que soit le modèle, nous devons vérifier qu'il est toujours admissible, c'est-à-dire $\int_{E^S} \exp Q(y \mid x)\ m^{\otimes S}(dy) < \infty$.

3 Dynamique auto-exponentielle mixte

Reprenant les hypothèses ci-dessus, nous précisons [B2]. Pour chaque i, et conditionnellement à $(X^i(t) = y^i \ , \ X(t-1) = x)$, la distribution de $X_i(t)$ est une loi exponentielle mixte, c'est-à-dire

$$\ln \mu_i(y_i \mid y^i, x) = \langle \theta_i(y^i, x), B_i(y_i) \rangle + D_i(y^i, x) \tag{4}$$

où

$$\theta_i(y^i, x) = \begin{pmatrix} \ln \frac{(1-\gamma_i(y^i,x))\lambda_i(y^i,x)}{\gamma_i(y^i,x)} \\ \lambda_i(y^i, x) \end{pmatrix} \ , \quad B(y) = \begin{pmatrix} \delta^*(y) \\ -y\delta^*(y) \end{pmatrix} \ .$$

On vérifie $B_i(0) = 0$ pour tout $i \in S$.

Alors, conditionnellement à $X(t-1) = x$, il existe des familles de vecteurs $\alpha_i(x) = (a_i(x), b_i(x))^t$, et de matrices $\beta_{ij}(x) = \begin{pmatrix} c_{ij}(x) & d_{ij}(x) \\ f_{ij}(x) & e_{ij}(x) \end{pmatrix}$ avec $c_{ij}(\bullet) = c_{ji}(\bullet)$, $e_{ij}(\bullet) = e_{ji}(\bullet)$ et $f_{ij}(\bullet) = d_{ji}(\bullet)$ telle que l'énergie s'écrit de la manière suivante:

$$\begin{aligned} Q(y \mid x) = & \textstyle\sum_{i \in S} \{a_i(x)\delta^*(y_i) - b_i(x)y_i\} \\ & + \textstyle\sum_{(i,j):\langle i,j \rangle} \{c_{ij}(x)\delta^*(y_i)\delta^*(y_j) - d_{ij}(x)y_j\delta^*(y_i) - f_{ij}(x)y_i\delta^*(y_j) + e_{ij}(x)y_iy_j\} \end{aligned}$$

Nous voyons ici que l'énergie conditionnelle est liée à deux types de potentiels, les potentiels d'interaction instantanée, et les potentiels d'interaction dans le temps.; on a un double graphe $G = \{\mathcal{G}, \mathcal{G}^-\}$, où $\mathcal{G}$ est le graphe symétrique instantané sur S, tandis que $\mathcal{G}^-$ décrit les dépendances temporelles et est orienté. De plus, les notations y^i et x se réduisent en fait aux voisins instantanés de $X_i(t)$ et à ses voisins à l'instant précédent.

On a aussi la forme des paramètres naturels :

$$\begin{aligned} \theta_{1,i}(y^i, x) &= a_i(x) + \sum_{j:\{i,j\}} \{c_{ij}(x)\delta^*(y_j) - d_{ij}(x)y_j\delta^*(y_j)\} \\ \theta_{2,i}(y^i, x) &= b_i(x) + \sum_{j:\{i,j\}} \{f_{ij}(x)\delta^*(y_j) - e_{ij}(x)y_j\delta^*(y_j)\} \end{aligned}$$

avec la correspondance : $\lambda_i(y^i, x) = \theta_{2,i}(y^i, x)$, $\qquad \gamma_i(y^i, x) = \frac{\theta_{2,i}(y^i,x)}{\theta_{2,i}(y^i,x) + e^{\theta_{1,i}(y^i,x)}}$

Ce modèle est bien défini pourvu que pour tout $i, \ x, \ y, \ \lambda_i(y^i, x) > 0, \ \gamma_i(y^i, x) \in]0, 1[$ d'une part, et que l'énergie est intégrable d'autre part.

Ceci est vérifié sous les conditions suivantes :

(A) (i) Pour tout $i \in S$, pour tout $A \subset S$ et $x, \ b_i(x) + \sum_{j \in A} f_{ij}(x) > 0$.

(ii) Pour tout $\{i, j\} \in S$, et x, $e_{ij}(x) \leq 0$.

Exemple Spécifions encore le modèle. Par exemple, nous choisissons $\alpha_i(x) = (a_i(x), b_i(x))^t$ avec

$$\begin{array}{rl} a_i(x) = & a_i + \sum_{l\in\partial i^-} \alpha^1_{li} x_l + \sum_{l\in\partial i^-} \varepsilon^1_{li} \delta^*(x_l) \\ b_i(x) = & b_i + \sum_{l\in\partial i^-} \alpha^2_{li} x_l + \sum_{l\in\partial i^-} \varepsilon^2_{li} \delta^*(x_l) \end{array}$$

et $\beta_{ij}(x) = \beta_{ij} = \begin{pmatrix} c_{ij} & d_{ij} \\ f_{ij} & e_{ij} \end{pmatrix}$ with $d_{ij} = f_{ji}$.

Dans ce cadre, on obtient l'écriture finale de l'énergie :

$$\begin{aligned} Q(y \mid x) = & \sum_{i\in S} \left(a_i + \sum_{l\in\partial i^-} \alpha^1_{li} x_l + \sum_{l\in\partial i^-} \varepsilon^1_{li} \delta^*(x_l)\right) \delta^*(y_i) \\ & - \sum_{i\in S} \left(b_i + \sum_{l\in\partial i^-} \alpha^2_{li} x_l + \sum_{l\in\partial i^-} \varepsilon^2_{li} \delta^*(x_l)\right) y_i \\ & + \sum_{(i,j):\langle i,j\rangle} \left\{c_{ij}\delta^*(y_i)\delta^*(y_j) - d_{ij} y_j \delta^*(y_i) - f_{ij} y_i \delta^*(y_j) + e_{ij} y_i y_j\right\}. \end{aligned}$$

3.1 Propriétés

Ergodicité. On utilise un critère de Lyapounov pour obtenir l'ergodicité sous des conditions non restrcitives.

Estimation. Sous des hypothèses classiques (invariance par translation), les estimateurs de pseudo-vraisemblance conditionnelle sont consistants.

Modèle coopératif tronqué. La corrélation spatiale n'est pas explicite ici. Un modèle est défini comme coopératif si l'espérance conditionnelle en un site augmente avec les voisins de ce site. Le calcul de cette espérance montre que, comme pour le cas connu de l'auto-modèle de Besag à lois conditionnelles exponentielles, la condition (A) est incompatible avec une coopération (instantanée). Un remède est alors de considérer des lois conditionnelles (mixtes) tronquées, ce qui libère toute contrainte sur les paramètres. On peut alors considérer des modèles coopératifs aussi bien que compétitifs.

3.2 Application

Nous présentons une application sur données réelles. Nous avons retenu des données radars de pluviométrie. Nous avons choisi pour cette étude le modèle défini dans l'exemple ci-dessus, en rajoutant des hypothèses d'invariance par translation (dans l'espace et dans le temps), un modèle aux 4 plus proches voisins pour le graphe $\mathcal{G}$ et aux 5 voisins pour le graphe $\mathcal{G}^-$ (le site lui-même et ses 4 plus proches voisins).

L'estimation des paramètres du modèle est faite par la méthode de pseudo-vraisemblance. Sur la base de ces estimations, nous avons resimulé les données.

Bibliographie

[1] B. C. Arnold, E. Castillo et J. M. Sarabia (1999). *Conditional Specification of Statistical Models*. Springer-Verlag, New York

[2] D. J.Allcroft, C. A. Glasbey (2003), A latent Gaussian Markov random-field model for spatiotemporal rainfall disaggregation, *JRSS C* 52 (4), 487–498.
[3] J. Besag (1974), Spatial interactions and the statistical analysis of lattice systems *JRSS B* 148, 1-36
[4] X. Guyon (1995). *Random Fields on a Network: Modeling, Statistics, and Applications.* Springer-Verlag, New York
[5] X. Guyon, C. Hardouin (2002), Markov chain Markov field dynamics: models and statistics. *Statistics*, Vol. 13, pp. 339-363.
[6] C. Hardouin, JF. Yao (2008). Spatial modelling for mixed state observations. *Electronic Journal of Statistics*, Vol. 2, 213-233.
[7] C. Hardouin, JF. Yao (2008). Multi-parameter auto-models with applications to cooperative systems and analysis of mixed state data. *Biometrika* 95, 335 - 349.