Un modèle de mélange pour la segmentation de données spatiales

Allou Samé[1] & Jean-Philippe Tarel[2] & Nadir Ait Saidi[1]

[1] *Université Paris-Est, COSYS, GRETTIA, IFSTTAR, F-77447 Marne-la-Vallée, France*
[2] *Université Paris-Est, COSYS, LEPSIS, IFSTTAR, F-77447 Marne-la-Vallée, France*

Résumé. Cet article décrit une approche basée sur les mélanges de lois, pour la modélisation et la segmentation de données spatiales. La dépendance spatiale des données y est prise en compte par le biais des proportions du mélange, qui sont modélisées par des transformations logistiques de fonctions polynomiales des coordonnées spatiales. Les paramètres du modèle proposé sont estimés par la méthode du maximum de vraisemblance via un algorithme EM spécifique, qui incorpore un algorithme de Newton-Raphson pour l'estimation des coefficients des fonctions logistiques. Les expérimentations, menées sur des images simulées, donnent des résultats encourageants en termes de précision de segmentation.

Mots-clés. Segmentation et modélisation, données spatiales, modèle de mélange, champ logistique latent, algorithme EM

Abstract. A mixture-model-based approach is introduced in this article for spatial data modeling and segmentation. The spatial dependence of the data is taken into account through the proportions of the mixture, which are modeled as logistic transformations of polynomial functions of the spatial coordinates. The parameters of the proposed model are estimated by maximizing the likelihood criterion through a specific EM algorithm which incorporates a Newton-Raphson algorithm dedicated to the estimation of the logistic functions coefficients. The experiments conducted on synthetic images have shown encouraging results in term of accuracy of segmentation.

Keywords. Segmentation and modeling, spatial data, mixture model, hidden logistic random field, EM algorithm

1 Introduction

Dans plusieurs domaines applicatifs, l'extraction de classes à partir des données spatialement référencées, constitue une tâche primordiale. Parmi les approches probabilistes permettant d'effectuer ce type de partitionnement, les modèles à base de champ de Markov caché constituent une référence en la matière (Besag, 1986 ; Ambroise et Govaert, 1998 ; Celeux et al., 2003). La classification à base de modèles de mélange gaussiens (Mclachlan et Peel, 2000), implémentée via l'algorithme Expectation-Maximization (EM) (Dempster

et al., 1977) constitue quant à elle une méthode de référence conçue pour la classification de données multivariées non nécessairement spatiales.

Cet article propose une extension des modèles de mélange permettant de partitionner et modéliser des données spatiales. La dépendance spatiale des données y est prise en compte par le biais des proportions du mélange, qui sont modélisées sous la forme de transformations logistiques de fonctions polynomiales des coordonnées spatiales. Ce modèle constitue l'extension spatiale du modèle de mélange dédié à la segmentation de signaux temporels, proposé par Chamroukhi et al. (2009).

L'article est organisé comme suit : la section 2 décrit le modèle proposé et l'algorithme d'estimation des paramètres. La section 4 évalue l'approche proposée sur des données synthétiques, dans le cadre de la segmentation d'images. Quelques perspectives de ce travail sont données dans la dernière section.

2 Modèle probabiliste à champ logistique latent

2.1 Définition du modèle

Dans la suite de l'article, les données spatiales à partitionner seront notées $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, où chaque observation $\boldsymbol{x}_i \in \mathbb{R}^d$ sera associée à un site $\boldsymbol{s}_i$. Dans le domaine des sciences géographiques par exemple, l'unité statistique $\boldsymbol{x}_i$ représente généralement un vecteur de caractéristiques démographiques, environnementales ou socio-économiques associées à une parcelle, une ville ou un pays, et le site $\boldsymbol{s}_i$ à ses coordonnées géographiques (latitude, longitude). Dans le domaine du traitement d'images, les unités statistiques sont les pixels, caractérisés par des attributs tels que le niveau de gris ou l'intensité des couleurs rouge/vert/bleu. Ces pixels sont repérés par leurs coordonnées dans l'image. Nous supposerons, pour simplifier, que $\boldsymbol{s}_i$ $(i = 1, \ldots, n)$ est défini par le couple de coordonnées spatiales non aléatoires $(u_i, v_i) \in \mathbb{R}^2$. Le modèle que nous proposons ici fait l'hypothèse, comme pour les modèles de mélange gaussiens classiques (Mclachlan et Peel, 2000), que chaque observation est distribuée suivant un mélange de K densités gaussiennes mais dont les proportions sont des fonctions spécifiques de coordonnées spatiales. Ce mélange est défini par :

$$p(\boldsymbol{x}_i; \boldsymbol{\Phi}) = \sum_{k=1}^{K} \pi_k(u_i, v_i; \boldsymbol{\alpha}) \, \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k), \tag{1}$$

où les $\boldsymbol{\mu}_k$ et les $\boldsymbol{\Sigma}_k$ sont les moyennes et les matrices de covariance des composantes gaussiennes définies dans $\mathbb{R}^d$ et $\boldsymbol{\Phi} = \{\boldsymbol{\alpha}, \boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \ldots, \boldsymbol{\Sigma}_K\}$ désigne l'ensemble des paramètres du modèle. Pour prendre en compte la régularité spatiale des données dans la classification, nous définissons la probabilité $\pi_k(u_i, v_i; \boldsymbol{\alpha})$ par la transformation logistique suivante d'une fonction polynôme des coordonnées spatiales :

$$\pi_k(u_i, v_i; \boldsymbol{\alpha}) = \frac{\exp(\boldsymbol{r}(u_i, v_i)^T \boldsymbol{\alpha}_k)}{\sum_{\ell=1}^{K} \exp(\boldsymbol{r}(u_i, v_i)^T \boldsymbol{\alpha}_\ell)}, \tag{2}$$

où $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, \ldots, \boldsymbol{\alpha}_K^T)^T$ est l'ensemble des coefficients des fonctions logistiques, $\boldsymbol{r}(u_i, v_i) \in \mathbb{R}^q$ est le vecteur des monômes associés à une fonction polynomiale du couple (u_i, v_i) et $\boldsymbol{\alpha}_k \in \mathbb{R}^q$ est le vecteur des coefficients qui lui est associé. Le tableau 1 donne des exemples de vecteurs $\boldsymbol{r}(u_i, v_i)$ pour différents ordres polynomiaux. On peut remarquer que les proportions logistiques qui viennent d'être définies satisfont bien les contraintes $\sum_{k=1}^K \pi_k(u_i, v_i; \boldsymbol{\alpha}) = 1$ et $0 < \pi_k(u_i, v_i; \boldsymbol{\alpha}) < 1$.

TABLE 1 – Vecteurs $\boldsymbol{r}(u_i, v_i)$ associés à différents ordres polynomiaux

polynôme d'ordre 1	$q = 3$	$\boldsymbol{r}(u_i, v_i) = (1, u_i, v_i)^T$
polynôme d'ordre 2	$q = 6$	$\boldsymbol{r}(u_i, v_i) = (1, u_i, v_i, u_i^2, u_i v_i, v_i^2)^T$
polynôme d'ordre 3	$q = 10$	$\boldsymbol{r}(u_i, v_i) = (1, u_i, v_i, u_i^2, u_i v_i, v_i^2, u_i^3, u_i^2 j, u_i v_i^2, v_i^3)^T$

Les densités exploitées dans ce modèle étant gaussiennes, son identifiabilité, c'est-à-dire l'équivalence $p(\boldsymbol{x}_i; \boldsymbol{\Phi}_1) = p(\boldsymbol{x}_i; \boldsymbol{\Phi}_2) \Longleftrightarrow \boldsymbol{\Phi}_1 = \boldsymbol{\Phi}_2$, est assurée à une permutation près des composantes du mélange, dès lors que le dernier paramètre des fonctions logistiques $(\boldsymbol{\alpha}_K)$ est fixé au vecteur nul (Jiang et Tanner, 1999). Sous ces conditions, que nous considérerons dans la suite de cet article, le vecteur $\boldsymbol{\alpha}$ appartient donc à l'espace $\mathbb{R}^{q(K-1)}$.

Soulignons ici que notre modèle peut également être posé comme un modèle génératif à variables latentes, les variables latentes $(z_1, \ldots, z_n)$ étant générées indépendamment suivant la loi multinomiale $\mathcal{M}(1; \pi_1(u_i, v_i), \ldots, \pi_K(u_i, v_i))$ telle que $p(z_i; \boldsymbol{\Phi}) = \pi_{z_i}(u_i, v_i)$, et les observations $(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ étant générées conditionnellement aux $(z_1, \ldots, z_n)$ suivant la densité gaussienne $p(\boldsymbol{x}_i | z_i; \boldsymbol{\Phi}) = \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i})$. Deux situations de données simulées, correspondant à des probabilités logistiques d'ordre polynômial 1 et 2, sont présentées dans la section 4 (figure 1). On constate notamment que les données simulées respectent la contrainte spatiale qui leur est imposée par les probabilités logistiques.

2.2 Segmentation spatiale des données

Le modèle proposé conduit à une segmentation $(\Omega_1, \ldots, \Omega_K)$, où l'ensemble Ω_k, qui est obtenu par maximisation des probabilités logistiques, est défini par :

$$\begin{aligned}
\Omega_k &= \Big\{(u, v) \,:\, \pi_k(u, v; \boldsymbol{\alpha}) = \max_{1 \le \ell \le K} \pi_\ell(u, v; \boldsymbol{\alpha})\Big\} \\
&= \bigcap_{\ell=1}^K \Big\{(u, v) \,:\, \log(\pi_k(u, v; \boldsymbol{\alpha}) / \pi_\ell(u, v; \boldsymbol{\alpha})) \ge 0\Big\} \\
&= \bigcap_{\ell=1}^K \Big\{(u, v) \,:\, (\boldsymbol{\alpha}_k - \boldsymbol{\alpha}_\ell)^T \boldsymbol{r}(u, v) \ge 0\Big\}.
\end{aligned}$$

Par conséquent, dans le cas particulier où les proportions du mélange sont des transformations logistiques de polynômes d'ordre 1 de (u, v), la région Ω_k, qui est l'intersection de parties convexes de $\mathbb{R}^2$ est convexe. Notons que les probabilités a posteriori peuvent aussi être utilisées pour partitionner les données si des contraintes géométriques strictes telles que la convexité ne sont pas imposées aux segments.

3 Algorithme d'estimation des paramètres

L'estimation des paramètres du modèle proposé est effectuée en maximisant la log-vraisemblance

$$\mathcal{L}(\boldsymbol{\Phi}) = \log \prod_{i=1}^{n} p(\boldsymbol{x}_i; \boldsymbol{\Phi}) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k(u_i, v_i; \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \tag{3}$$

via l'algorithme EM (McLachlan et Krishnan, 2008 ; Chamroukhi et al., 2009). Celui-ci consiste à démarrer d'un paramètre initial $\boldsymbol{\Phi}^{(0)}$, puis à calculer de manière itérative le paramètre $\boldsymbol{\Phi}^{(c+1)}$ qui maximise la fonction auxiliaire

$$Q^{(c)}(\boldsymbol{\Phi}) = \sum_{i,k} \tau_{ik}^{(c)} \log(\pi_k(u_i, v_i; \boldsymbol{\alpha}) \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)), \tag{4}$$

avec $\tau_{ik}^{(c)} = \exp(\boldsymbol{r}(u_i, v_i)^T \boldsymbol{\alpha}_k^{(c)}) \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_k^{(c)}, \boldsymbol{\Sigma}_k^{(c)}) / \sum_{\ell=1}^{K} \exp(\boldsymbol{r}(u_i, v_i)^T \boldsymbol{\alpha}_\ell^{(c)}) \mathcal{N}(\boldsymbol{x}_i; \boldsymbol{\mu}_\ell^{(c)}, \boldsymbol{\Sigma}_\ell^{(c)})$, où c désigne l'itération courante. La maximisation de $Q^{(c)}$ par rapport à $\boldsymbol{\mu}_k$ et $\boldsymbol{\Sigma}_k$ est analogue à celle d'un modèle de mélange gaussien classique et la maximisation de $Q^{(c)}$ par rapport au paramètre $\boldsymbol{\alpha}$ se ramène à la résolution du problème de régression logistique $\boldsymbol{\alpha}^{(c+1)} = \arg\max_{\boldsymbol{\alpha}} \sum_{i,k} \tau_{ik}^{(c)} \log \pi_k(u_i, v_i; \boldsymbol{\alpha})$, qui est un problème convexe d'optimisation que nous résolvons par l'algorithme de Newton-Raphson. Celui-ci correspond, dans ce cadre, à l'algorithme IRLS (Iteratively Reweighted Least Squares) (Green, 1984 ; Chamroukhi et al., 2009).

4 Expérimentation sur des images synthétiques

Cette section présente quelques résultats obtenus par l'algorithme proposé, sur des images simulées. Rappelons que dans le cas des images, les coordonnées spatiales (u_i, v_i) appartiennent à une grille régulière bidimensionnelle $\{1, \ldots, U\} \times \{1, \ldots, V\}$, où U et V sont les dimensions de l'image. Nous supposons que les observations $\boldsymbol{x}_i$ sont des vecteurs de $\mathbb{R}^3$ indiquant l'intensité des couleurs rouge/vert/bleu. Les trois algorithmes comparés sont :

- EM : l'algorithme EM classique appliqué au modèle de mélange gaussien qui ne gère pas nécessairement l'aspect spatial des données (Mclachlan et Krishnan, 2008).
- NEM : l'algorithme « Neighborhood EM » qui maximise la log-vraisemblance associée au modèle de mélange gaussien classique pénalisé par un terme prenant en compte la contiguïté spatiale des données (Ambroise et Govaert, 1998). Cette méthode s'appuie implicitement sur le modèle de champ de Markov caché de Potts.
- EM-CL : l'algorithme EM proposé dont le modèle associé repose sur un champ logistique latent.

Nous considérons ici deux situations d'images formées de $K = 3$ segments, dont les fonctions logistiques sont associés à des polynômes d'ordre 1 et 2. Les images de type 1

sont de taille 100×100 ($n = 10000$) et les images de type 2 de taille 140×140 ($n = 19600$). Les paramètres utilisés pour générer les images sont listés dans le tableau 2.

TABLE 2 – Paramètres de simulation pour les situations 1 et 2

	Situation 1	Situation 2
$\boldsymbol{\alpha}_1$	$(16.27 \ -0.47 \ \ 0.23)^T$	$(-55.00 \ \ 1.56 \ \ 0.19 \ -0.0117 \ \ 0.0002 \ -0.0014)^T$
$\boldsymbol{\alpha}_2$	$(-14.03 \ -0.17 \ \ 0.53)^T$	$(-91.00 \ \ 0.32 \ \ 2.42 \ -0.0027 \ \ 0.0008 \ -0.0169)^T$
$\boldsymbol{\alpha}_3$	$(0 \ \ 0 \ \ 0)^T$	$(0 \ \ 0 \ \ 0 \ \ 0 \ \ 0 \ \ 0)^T$
$\boldsymbol{\mu}_1 \ \boldsymbol{\Sigma}_1$	$(255 \ \ 0 \ \ 0)^T \quad \sigma^2 \mathbf{I}$	$(255 \ \ 0 \ \ 0)^T \quad \sigma^2 \mathbf{I}$
$\boldsymbol{\mu}_2 \ \boldsymbol{\Sigma}_2$	$(0 \ \ 255 \ \ 0)^T \quad \sigma^2 \mathbf{I}$	$(0 \ \ 255 \ \ 0)^T \quad \sigma^2 \mathbf{I}$
$\boldsymbol{\mu}_3 \ \boldsymbol{\Sigma}_3$	$(0 \ \ 0 \ \ 255)^T \quad \sigma^2 \mathbf{I}$	$(0 \ \ 0 \ \ 255)^T \quad \sigma^2 \mathbf{I}$

La figure 1 illustre les probabilités logistiques utilisées et les données simulées pour chacune des situations, pour un écart-type de bruit $\sigma = 100$.

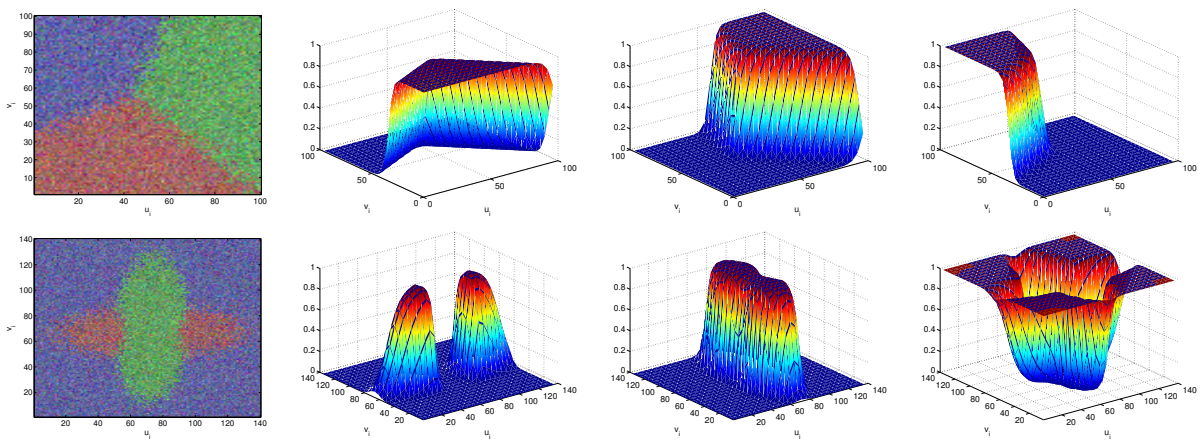

FIGURE 1 – Exemple de données simulées pour la situation 1 (haut) et la situation 2 (bas), et probabilités logistiques correspondantes

Le tableau 3 donne le taux d'erreur de segmentation (moyenné, pour chaque situation et chaque niveau de bruit, sur 20 images simulées) obtenu avec les trois algorithmes comparés, pour les trois niveaux de bruit considérés. Ce taux d'erreur est calculé par rapport à la segmentation obtenue en appliquant la règle décrite dans la section 2.2 avec les vrais paramètres donnés dans le tableau 2. On peut remarquer que l'algorithme EM-CL donne de meilleurs résultats que les autres approches. Ces bonnes performances, qui peuvent être attribuées au fait que les données aient été simulées suivant notre modèle, nous rassurent néanmoins quant aux qualités escomptées par notre algorithme. On remarque aussi que l'algorithme EM classique, qui ne prend pas en compte l'aspect spatial des données, donne de moins bons résultats en particulier lorsque l'écart-type du bruit est élevé.

TABLE 3 – Taux d'erreur de segmentation (pourcentage de mal classés moyenné sur 20 images simulées) obtenus avec les 4 algorithmes, pour 3 niveaux de bruit ; les écart-types du taux d'erreur sont marqués entre parenthèses

	Situation 1			Situation 2		
	$\sigma = 100$	$\sigma = 200$	$\sigma = 300$	$\sigma = 100$	$\sigma = 200$	$\sigma = 300$
EM	0.76(0.31)	31.67(0.53)	43.10(0.80)	11.37(0.14)	23.42(0.27)	40.88(6.89)
NEM	1.34(0.17)	1.57(0.30)	2.11(0.37)	1.83(0.15	3.05(0.47)	5.95(4.57)
EM-CL	0.49(0.11)	0.58(0.12)	0.68(0.20)	0.44(0.08)	0.67(0.11)	1.02(0.19)

5 Conclusion et perspectives

Cet article a proposé un modèle de mélange pour la segmentation de données spatiales. Celui-ci prend en compte l'aspect spatial des données via les proportions du mélange qui sont des fonctions logistiques des coordonnées spatiales. Les expérimentations menées sur des images simulées suivant le modèle proposé ont montré de bons résultats en termes de segmentation. Les perspectives de ce travail concernent d'une part la poursuite des expérimentations sur des données réelles, et, d'autre part, la segmentation de données spatiales où les segments ont des formes plus complexes. Dans ce cas, notre approche pourrait nécessiter d'augmenter artificiellement le nombre de segments afin de mieux coller aux données. Le choix du nombre de classes constitue ainsi une perspective importante de ce travail.

Bibliographie

[1] Ambroise, C. et Govaert, G. (1998), *Convergence of an EM-type algorithm for spatial clustering*, Pattern Recognition Letters, 19, 919–927.

[2] Besag, J. (1986), *On the statistical analysis of dirty pictures*, JRSS B, 48(3), 259–302.

[3] Celeux, G., Forbes, F. et Peyrard, N. (2003), *Procedures using mean field-like approximations for Markov model-based image segmentation*, Pattern Rec., 36(1), 131–144.

[4] Chamroukhi, F., Samé, A., Govaert, G. et Aknin, P. (2009), *Time series modeling by a regression approach based on a latent process*, Neural Networks, 22, 593–602.

[5] Dempster, A. P., Laird, N. M. et Rubin, D. B. (1977), *Maximum likelihood from incomplete data via the EM algorithm*, JRSS B, 39, 1–38.

[6] Green, P. (1984), *Iteratively reweighted least squares for maximum likelihood estimation, and some robust and resistant alternatives*, Journal of the Royal Statistical Society, Series B, 46(2), 149–192.

[7] Jiang, W. et Tanner, M. A. (1999), *On the identifiability of mixtures-of-experts*, Neural Networks, 12(9), 1253–1258.

[8] McLachlan, G. et Krishnan, T. (2008), *The EM algorithm and extensions*, Wiley, NY.

[9] McLachlan, G. et Peel, D. (2000), *Finite mixture models*, Wiley, NY.