Segmentation non-supervisée dans les champs de Markov couples gaussiens

Hugo GANGLOFF[1, 2], Jean-Baptiste COURBOT[3], Emmanuel MONFRINI[4], Christophe COLLET[1]

[1]ICube, Université de Strasbourg - CNRS UMR 7357
300 bd Sébastien Brant, 67400 Illkirch-Graffenstaden, France

[2]Groupe Européen de Recherche sur les Prothèses Appliquées à la Chirurgie Vasculaire
4 rue Kirschleger, 67000 Strasbourg, France

[3]IRIMAS EA 7499, Université de Haute-Alsace
2 rue des Frères Lumière, 68100 Mulhouse, France

[4]SAMOVAR - CNRS UMR 5157, Télécom SudParis, Institut Polytechnique de Paris
9 rue Charles Fourier, 91000 Évry, France

hugogangloff@unistra.fr, jean-baptiste.courbot@uha.fr
emmanuel.monfrini@telecom-sudparis.eu, c.collet@unistra.fr

Résumé – Nous présentons un modèle probabiliste à données cachées pour la modélisation de variables fortement corrélées. Une méthode d'estimation non-supervisée des paramètres se distinguant par sa robustesse est proposée. Le nouveau modèle s'illustre face à d'autres méthodes dans les problèmes pratiques où la modélisation par bruit corrélé est pertinente.

Abstract – We present a new probabilistic latent variable model to characterize strong correlations. A robust unsupervised parameter estimation method is proposed. The new model performs well against other classical methods in practical cases where a model including correlated noise is useful.

1 Introduction

L'intérêt des outils probabilistes pour les modélisations mathématiques réside dans la possibilité d'introduire des corrélations complexes entre variables aléatoires. Parmi les modèles à données cachées, à partir des champs de Markov cachés (CMCa) [12], plusieurs propositions de généralisation ont été faites pour enrichir la loi des variables aléatoires considérées. C'est le cas des modèles de Markov couples ou triplets [11]. Des études théoriques de modèles similaires pour divers mélanges de lois de probabilité font l'objet de nombreaux articles [17] [14]. Les Champs Aléatoires Gaussiens Markoviens (CAGM) [15] apportent également des solutions à cette problématique.

Nous présentons dans cet article un nouveau modèle, appelé Champ de Markov Couple Gaussien (CMCoG) qui fait le lien entre les deux théories précédentes pour répondre à des problèmes où les corrélations sont fortes. Ce modèle peut être vu comme une machine de Boltzmann avec de nombreuses connections à l'intérieur d'une même couche [10]. Grâce à l'hypothèse de Markov, nous proposons une procédure d'estimation des paramètres efficaces dans ces modèles complexes.

Dans la suite, la notation $p(X = x) = p(x)$ renvoie à la densité p évaluée en la réalisation x, et les lettres grasses minuscules (resp. majuscules) désignent les vecteurs (resp. matrices).

2 Les champs de Markov couples gaussiens

2.1 Définition du modèle

Soit $\boldsymbol{X} = (X_1, \dots, X_N)$ une variable alétoire discrète à valeurs dans Ω^N, avec $\Omega = (\omega_1, \dots, \omega_L)$. Les réalisations de $\boldsymbol{X}$ sont inobservables. $\boldsymbol{Y} = (Y_1, \dots, Y_N)$ est une variable aléatoire réelle dont les réalisations sont observées. On considère que $(\boldsymbol{X}, \boldsymbol{Y})$ est un processus de Markov défini par rapport à un voisinage $\mathcal{N}$. L'ensemble des sites est $\mathcal{S}$ avec $|\mathcal{S}| = N$. Nous associons à ce champ couple la densité suivante :

$$p(\boldsymbol{x}, \boldsymbol{y}) = \frac{\exp\left(-E(\boldsymbol{x}, \boldsymbol{y})\right)}{Z(\boldsymbol{x}, \boldsymbol{y})}, \tag{1}$$

où l'énergie vaut :

$$E(\boldsymbol{x}, \boldsymbol{y}) = \exp\Big(- \big(\sum_{s \in \mathcal{S}} V_1(x_s) + \sum_{s \in \mathcal{S}} \sum_{s' \in \mathcal{N}_s} V_2(x_s, x_{s'}) + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu_x})^T \boldsymbol{Q} (\boldsymbol{y} - \boldsymbol{\mu_x}) \big) \Big), \tag{2}$$

et $Z(\boldsymbol{x}, \boldsymbol{y})$ est la constante de normalisation inconnue. Pour que $p(\boldsymbol{x}, \boldsymbol{y})$ existe, il faut que $\exp(-E(\boldsymbol{x}, \boldsymbol{y}))$ soit intégrable sur $\Omega^N \times \mathbb{R}^N$ ce qui est le cas dans notre définition de l'Équation 2.

Nous pouvons montrer que $\boldsymbol{Y}$ est un CAGM, conditionnellement aux réalisations de $\boldsymbol{X}$. En effet :

$$p(\boldsymbol{y}|\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^N \det(\boldsymbol{Q}^{-1})}} \exp\big(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu_x})^T \boldsymbol{Q}(\boldsymbol{y}-\boldsymbol{\mu_x})\big). \tag{3}$$

Ce qui correspond à la définition d'un CAGM de moyenne $\boldsymbol{\mu_x}$ et de matrice de covariance $\boldsymbol{\Sigma} = \boldsymbol{Q}^{-1}$. Cette propriété offre de nombreuses possibilités de modélisations. Notons que nous n'incluons pas la non-stationnarité de la variance du champ gaussien. Nous verrons à la Section 2.3 que la simulation du CAGM doit se faire à travers ses *équations conditionnelles* [15] [3] qui varient au cours des itérations. Dans ce contexte, introduire une variance non-stationnaire est hors de portée de cet article [8].

2.2 Paramètres du modèle

Paramètres liant les variables cachées

Le choix d'un modèle de type Multi-Level Logistic (MLL) est fait ici [12]. Le potentiel V_1 (resp. V_2) modélise le biais (resp. la granularité), $\forall \omega_l \in \Omega$:

$$V_1(x_s) = \alpha_{[x_s=\omega_l]} \text{ et } V_2(x_s, x_{s'}) = (-1)^{\mathbb{1}_{[x_s \neq x_{s'}]}}\beta. \tag{4}$$

Paramètres du CAGM

D'une part, les moyennes varient avec les réalisations des X_s, $\forall s \in \mathcal{S} : \mu_s = \mu_{x_s=\omega_l}$. D'autre part, deux autres paramètres sont liés à la matrice de précision $\boldsymbol{Q}$ du CAGM : σ^2 et r. Par définition d'un CAGM, $\boldsymbol{Q}$ est une matrice semi-définie positive. Nous faisons le choix de la caractériser par la fonction de corrélation exponentielle ρ avec un taux de décroissance r et une variance σ^2. On a, $\forall (s,s') \in \mathcal{S}^2$:

$$Cov(Y_s, Y_{s'}) = \Sigma_{s,s'} = \sigma^2 \rho(s,s';r), \tag{5}$$

et $\boldsymbol{Q} = \boldsymbol{\Sigma}^{-1}$. L'hypothèse de champ à bordure périodique est faite, ce qui correspond à considérer les indices des variables modulo N, ce qui est une manière de traiter les indices hors des limites (par exemple, l'indice $N+2$, qui n'existe pas, devient l'indice 2). Notons que, sans cette hypothèse, nous n'avons pas de garantie que les calculs soient faisables en un temps raisonnable. Les notions abordées ici sont décrites en détail dans [15].

Ensemble final des paramètres

Sans perte de généralité, dans la suite, nous prenons $L = 2$ classes. Dans ce cas, le modèle est décrit par 7 paramètres. Soit $\boldsymbol{\theta} \triangleq \{\alpha_0, \alpha_1, \beta, \mu_0, \mu_1, \sigma, r\}$ le vecteur des paramètres. $\boldsymbol{\theta} \in \boldsymbol{\Theta} = \mathbb{R}^5 \times (\mathbb{R}^{+*})^2$. La Section 3 traite de l'estimation de ces paramètres. Notons que les cas $L > 2$ sont théoriquement possibles avec le modèle MLL et en considérant L moyennes pour le CAGM, au prix d'une plus grande complexité algorithmique.

2.3 Équations de simulation du couple

Dans cette section nous décrivons comment simuler des réalisations suivant le nouveau modèle CMCoG.

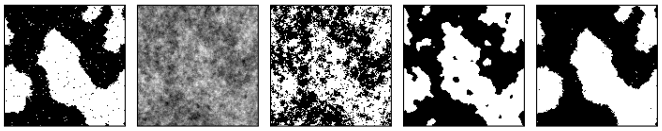


FIGURE 1 – Exemple de simulation et segmentation de CMCoG pour $\Delta\mu = 1.2$ dans le cas supervisé. De gauche à droite : $\boldsymbol{X}$ généré, $\boldsymbol{Y}$ généré, segmentation KMeans, segmentation CMCa et segmentation CMCoG.

Les probabilités de transitions, $p(x_s, y_s | x_{\mathcal{N}_s}, y_{\mathcal{N}_s}), \forall (x_s, y_s) \in \Omega \times \mathbb{R}, \forall s \in \mathcal{S}$, peuvent être écrite sous la forme $p(x_s|x_{\mathcal{N}_s}, y_{\mathcal{N}_s})$ $\times\ p(y_s|x_s, x_{\mathcal{N}_s}, y_{\mathcal{N}_s})$, dans le cas le plus général, en séparant les termes en y_s de ceux en x_s. Ainsi, $\forall s \in \mathcal{S}$, le couple (x_s, y_s) est obtenu en tirant successivement x_s puis y_s avec l'échantillonneur de Gibbs.

Remarquons que $p(y_s|x_s, x_{\mathcal{N}_s^X}, y_{\mathcal{N}_s^Y})$ est une loi gaussienne de moyenne η_s et de variance σ_s^2 données par :

$$\eta_s = \mu_s - \frac{1}{Q_{s,s}} \sum_{s' \in \mathcal{N}_s} (y_{s'} - \mu_{s'}) Q_{s,s'} \text{ et } \sigma_s^2 = \frac{1}{Q_{s,s}}. \tag{6}$$

Ces N équations sont les *équations conditionnelles* d'un CAGM, elles vont être utilisées pour la simulation. La Figure 1 illustre une réalisation d'un CMCoG, sous forme d'image à 2 classes.

2.4 Équations de simulation *a posteriori*

Nous décrivons maintenant comment utiliser notre modèle dans un cadre de classification. Le couple $(\boldsymbol{X}, \boldsymbol{Y})$ est markovien, ce qui permet de conserver la markovianité de la loi *a posteriori*. On a alors, $\forall x_s \in \Omega, \forall s \in \mathcal{S}$, $p(x_s|x_{\mathcal{N}_S}, \boldsymbol{y}) \propto p(x_s, y_s|x_{\mathcal{N}_s}, y_{\mathcal{N}_s})$, et des réalisations peuvent être obtenues via l'échantillonneur de Gibbs [9].

2.5 Segmentation supervisée de données simulées

Pour illustrer la richesse des modélisations de bruit corrélé envisageable avec le modèle CMCoG, nous nous proposons de comparer le modèle CMCoG aux modèles CMCa et KMeans [1]. Nous générons des réalisations de CMCoG, $(\boldsymbol{X}, \boldsymbol{Y})$, sous forme d'image binaire de taille 128×128px, et segmentons les images pour différentes valeurs de μ_0 et μ_1. Nous utilisons le critère *Mode of Posterior Marginals* (MPM) [13] pour la segmentation. Dans cette expérience, les vrais paramètres et la vérité terrain sont connus. Dans la Figure 1, nous présentons les résultats des segmentation pour $\Delta\mu = 1.2$. Dans la Figure 2, nous traçons les taux d'erreur moyens pour les différents niveaux de bruit $\Delta\mu = |\mu_0 - \mu_1|$ et pour chaque modèle. Nous observons des courbes qui rejoignent les attentes théoriques, le nouveau modèle présente des taux d'erreur bien plus faibles que les autres, ce qui illustre sa plus grande généralité. Par exemple, dans le cas de la Figure 1 $\Delta\mu = 1.2$, le modèle CMCoG atteint un taux d'erreur de 4%, contre 38% pour KMeans, et 13% pour CMCa.

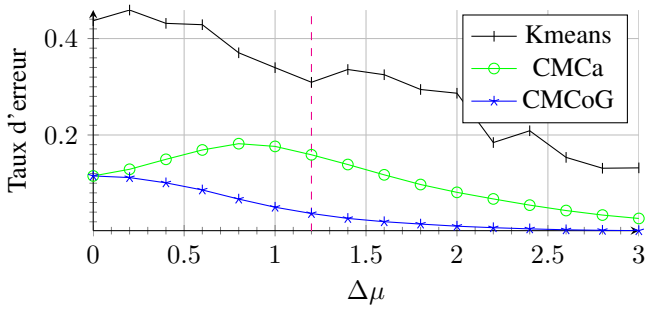


FIGURE 2 – Taux d'erreur des modèles en fonction de $\Delta\mu$ (valeurs moyennées sur 20 simulations). Le cas de la Figure 1 correspond à la ligne verticale en tirets.

3 Estimation des paramètres

3.1 Méthode robuste d'estimation

Nous nous plaçons dans un contexte non-supervisé, les données complètes $(\boldsymbol{X}, \boldsymbol{Y})$ ne sont pas disponibles. Nous proposons une procédure stochastique pour compléter les données puis estimer les paramètres, nommée PSEP.

Lorsque les données complètes sont connues, nous pouvons estimer α_0, α_1 et β avec le Maximum de la Pseudo-Vraisemblance [2] (MPV) et μ_0, μ_1 et σ^2 avec le Maximum de Vraisemblance (MV). r est estimé par ajustement de la fonction de corrélation exponentielle sur le corrélogramme estimé, via la méthode des moindres carrés [6]. L'utilisation d'algorithmes de type Expectation - Maximization, comme Stochastic E-M [4] a été écartée car les estimateurs du MV ne sont pas calculables en un temps raisonnable. La convergence de PSEP est fondée sur la stationnarité des échantillons tirés *a posteriori*, c'est le rôle de la fonction `verifier_convergence()`. L'Algorithme 1 résume les différentes étapes.

3.2 Échantillonneur de Gibbs tempéré

Les problèmes de convergence de l'échantillonneur de Gibbs sont connus dans la littérature. Nous proposons l'utilisation de l'échantillonneur de Gibbs tempéré (Gibbs-T). Son principe est détaillé dans [7].

L'idée est de mieux explorer la distribution *a posteriori* dans l'étape de la ligne 4 de l'Algorithme 1. Gibbs-T introduit un bruit stochastique qui joue un rôle de régularisation dans le problème d'optimisation que constitue l'étape d'estimation des paramètres. Cette approche a déjà été choisie pour une meilleure simulation de divers modèles probabilistes, [5] et [16].

3.3 Estimation des paramètres améliorée

En pratique, Gibbs-T n'est pas utilisé à chacune des itérations pour éviter un temps de calcul trop important lorsque les corrélations estimées couvrent un large voisinage. Nous modifions PSEP de manière à ce qu'il y ait alternance des échantillonneurs Gibbs-T et Gibbs à chaque itération. Nous obtenons un algorithme que nous appelons PSEP alterné (PSEP-A).

Algorithme 1 : Procédure Stochastique d'Estimation des Paramètres (PSEP)

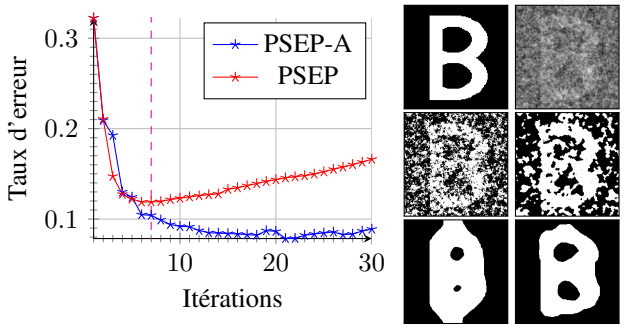
Données : $\boldsymbol{\theta}^0 = \{\alpha_0^0, \alpha_1^0, \beta^0, \mu_0^0, \mu_1^0, (\sigma^2)^0, r^0\}$, les paramètres initiaux et $\boldsymbol{y}$, les observations.
Résultat : $\hat{\boldsymbol{\theta}} = \{\hat{\alpha}_0, \hat{\alpha}_1, \hat{\beta}, \hat{\mu}_0, \hat{\mu}_1, \hat{\sigma}^2, \hat{r}\}$, les paramètres estimés.

1 $t \leftarrow 0$
2 **tant que** *pas de convergence* **faire**
3 /* Échantillonage *a posteriori* */
4 $\boldsymbol{x}^{t+1} \sim p(\boldsymbol{X}|\boldsymbol{y}; \boldsymbol{\theta}^t)$
5 /* Estimateurs */
6 Estimateur du MPV pour α_0^{t+1}, α_1^{t+1} et β^{t+1}.
7 Estimateur du MV pour μ_0^{t+1} et μ_1^{t+1}.
8 Estimateur du MV pour $(\sigma^2)^{t+1}$.
9 Estimateur du corrélogramme r^{t+1}.
10 $\boldsymbol{\theta}^{t+1} \leftarrow$
$\{\alpha_0^{t+1}, \alpha_1^{t+1}, \beta^{t+1}, \mu_0^{t+1}, \mu_1^{t+1}, (\sigma^2)^{t+1}, r^{t+1}\}$
11 `verifier_convergence()`
12 $t \leftarrow t + 1$
13 **fin**

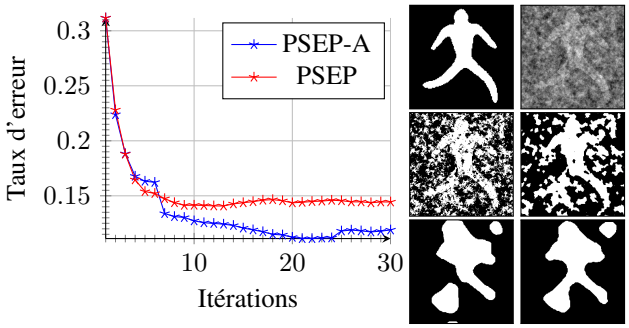
Pour comparer PSEP et PSEP-A, des images réelles binaires sont vues comme des réalisations des variables cachées ($\boldsymbol{x}$) pour les modèles CMCoG et CMCa. Pour chaque image binaire, nous choisissons μ et $\boldsymbol{Q}$ et utilisons l'Équation 6 pour simuler une observation. Cette réalisation correspond aux $\boldsymbol{y}$ des modèles à données cachées. Tous les paramètres sont ensuite considérés comme perdus : nous travaillons en contexte non-supervisé. Cependant la connaissance de la vérité terrain permet de suivre l'évolution des algorithmes d'estimation via le taux d'erreur de reconstruction.

La Figure 3 illustre les effets de l'introduction de Gibbs-T. Dans un premier cas (Figure 3a), PSEP augmente fortement après avoir atteint une valeur d'erreur minimale. C'est typique d'un régime de sur-apprentissage, caractérisé par la partie à droite de la ligne verticale en tirets [10][Section 5.2]. Il n'y a pas de sur-apprentissage pour PSEP-A. Dans un second cas (Figure 3b), PSEP se stabilise mieux, mais l'utilisation de l'échantillonneur tempéré permet une meilleure exploration de la loi *a posteriori* et un taux d'erreur plus favorable pour PSEP-A après 30 itérations. Dans tous les cas rencontrés dans nos expériences, l'utilisation de PSEP-A améliorait les résultats de segmentation.

PSEP-A est également utilisable avec le modèle CMCa, mais les bénéfices sont moins importants que pour CMCoG, nous ne les illustrons donc pas. Nous pouvons expliquer cela en conjecturant que la densité d'un modèle CMCoG présente beaucoup plus d'optima locaux que celle d'un modèle CMCa. Ainsi, dans cet espace de grande dimension, une procédure d'exploration améliorée, telle que PSEP-A, contribue à l'amélioration des résultats plus significativement que dans le modèle CMCa.



(a) Cas 1 - Taux d'erreur KMeans : 32%, CMCa : 24.9%, CMCoG (PSEP) : 16.6%, CMCoG (PSEP-A) : 8.8%.



(b) Cas 2 - Taux d'erreur KMeans : 31%, CMCa : 23.9%, CMCoG (PSEP) : 14.4%, CMCoG (PSEP-A) : 11.8%.

FIGURE 3 – Taux d'erreur des algorithmes dans la reconstruction *a posteriori* non-supervisée en fonction des itérations de l'algorithme d'estimation des paramètres pour le modèle CMCoG. Les images sont, de gauche à droite et de haut en bas : vérité terrain (inconnue), observations, segmentation KMeans, et les échantillons *a posteriori* en sortie de l'estimation des paramètres (après 30 itérations) pour CMCa (PSEP), pour CMCoG (PSEP) et pour CMCoG (PSEP-A).

4 Conclusions

Le modèle CMCoG présenté dans cet article s'avère très utile pour la modélisation de données très corrélées. L'étape d'estimation des paramètres, connue pour être un problème d'optimisation difficile dans ces modèles probabilistes, a été améliorée en intégrant l'échantillonneur de Gibbs tempéré. Notre travail est amené à être étendu avec les liens que ces modèles entretiennent avec les machines de Boltzmann, bien étudiées dans le domaine de l'apprentissage profond.

Références

[1] D. ARTHUR et al. "K-means++ : The advantages of careful seeding". In : *Proceedings of the 18th ACM-SIAM symposium on Discrete algorithms*. 2007, p. 1027–1035.

[2] J. BESAG. "Statistical analysis of non-lattice data". In : *The statistician* (1975), p. 179–195.

[3] D. A. BROWN et al. "Sampling strategies for fast updating of Gaussian Markov random fields". In : *The American Statistician* (2019), p. 1–32.

[4] G. CELEUX et al. "A stochastic approximation type EM algorithm for the mixture problem". In : *Stochastics : An International Journal of Probability and Stochastic Processes* 41.1-2 (1992), p. 119–134.

[5] K. CHO et al. "Improved learning of Gaussian-Bernoulli restricted Boltzmann machines". In : *International conference on artificial neural networks*. Springer. 2011, p. 10–17.

[6] N. CRESSIE. *Statistics for spatial data*. Wiley Online Library, 1992.

[7] G. DESJARDINS et al. "Tempered Markov chain Monte Carlo for training of restricted Boltzmann machines". In : *Proceedings of the 13th international conference on artificial intelligence and statistics*. 2010, p. 145–152.

[8] G.-A. FUGLSTAD et al. "Exploring a new class of non-stationary spatial Gaussian random fields with varying local anisotropy". In : *Statistica Sinica* (2015), p. 115–133.

[9] S. GEMAN et al. "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images". In : *Readings in computer vision*. Elsevier, 1987, p. 564–584.

[10] I. GOODFELLOW et al. *Deep learning*. MIT press, 2016.

[11] I. GORYNIN et al. "Assessing the segmentation performance of pairwise and triplet Markov models". In : *Signal Processing* 145 (2018), p. 183–192.

[12] S. LI. *Markov random field modeling in image analysis*. Springer Science & Business Media, 2009.

[13] J. MARROQUIN et al. "Probabilistic solution of ill-posed problems in computational vision". In : *Journal of the american statistical association* 82.397 (1987), p. 76–89.

[14] Y. PARK et al. "Learning the network structure of heterogeneous data via pairwise exponential Markov random fields". In : *Proceedings of machine learning research* 54 (2017), p. 1302.

[15] H. RUE et al. *Gaussian Markov random fields : theory and applications*. CRC press, 2005.

[16] Ruslan R SALAKHUTDINOV. "Learning in Markov random fields using tempered transitions". In : *Advances in neural information processing systems*. 2009, p. 1598–1606.

[17] Wesley TANSEY et al. "Vector-space Markov random fields via exponential families". In : *International Conference on Machine Learning*. 2015, p. 684–692.