Classification de données mixtes par un modèle de mélange de copules gaussiennes.

Matthieu Marbac [1] & Christophe Biernacki [2] & Vincent Vandewalle [3]

[1] *DGA & Inria Lille, matthieu.marbac-lourdelle@inria.fr*
[2] *Université Lille 1 & CNRS & Inria Lille, Christophe.Biernacki@math.univ-lille1.fr*
[3] *Université Lille 2 EA 2694 & Inria Lille, vincent.vandewalle@univ-lille2.fr*

Résumé. Nous proposons un modèle de mélange de copules gaussiennes pour la classification non supervisée de données mixtes. Les marginales de chaque composante sont des distributions standard, ce qui facilite l'interprétation des classes. Les corrélations intra-classe, quant à elles, sont prises en compte au travers des copules gaussiennes qui estiment un coefficient de corrélation, par couple de variables et par classe, ayant des propriétés de robustesse. De plus, les copules gaussiennes permettent de visualiser les données par classe. Dans un cadre bayésien, l'estimateur du maximum *a posteriori* est obtenu par un échantillonneur de Gibbs permettant d'explorer efficacement l'espace des paramètres. La classification d'un jeu de données médical illustre notre modèle.

Mots-clés. Classification, copules gaussiennes, données mixtes, modèles de mélanges.

Abstract. We propose a mixture model of Gaussian copulas to cluster mixed data sets. The margins of each component are classical distributions facilitating the interpretation of the classes. The intra-class correlations are taken into account by the Gaussian copulas which estimate one coefficient of correlation, per class, for each couple of variables, having robustness properties. Furthermore, Gaussian copulas allow to visualize the individuals per class. In a Bayesian framework, the maximum *a posteriori* estimate is obtained via a Gibbs sampler allowing efficient exploration of the parameter space. The model is illustrated by a medical data set clustering.

Keywords. Clustering, Gaussian copula, mixed data, mixture models.

1 Introduction

Avec le développement de l'informatique, les données à analyser se sont complexifiées. En particulier, elles comportent souvent plusieurs types de variables (données mixtes). Le *clustering* est un outil permettant d'extraire l'information des données car il regroupe les individus en classes caractéristiques. Il peut être conduit par des *méthodes probabilistes* qui modélisent le processus de génération des données. En considérant qu'une classe regroupe les individus issus de la même distribution de probabilité, l'approche classique consiste à utiliser un *modèle de mélange* fini de distributions paramétriques (McLachlan and Peel,

2000), permettant ainsi une interprétation facile des classes. Cependant, il existe peu de distributions multivariées lorsque les données sont mixtes.

Le modèle des *classes latentes*, en faisant l'hypothèse de l'*indépendance* des variables *conditionnellement à la classe*, permet de résoudre ce problème puisqu'il n'utilise que des distributions univariées. Composé de distributions standard pour les marginales de chaque composante, il résume chaque classe par ses paramètres marginaux. Cependant, il souffre de biais sévères lorsque les données sont effectivement corrélées au sein d'une classe.

Le modèle proposé a alors pour but de conserver des distributions standard pour les marginales de chaque composante tout en relâchant l'hypothèse d'indépendance conditionnelle. Comme les *copules* permettent de définir de manière dissociée le modèle de dépendance et la nature des distributions marginales (Hoff, 2007), on propose tout d'abord d'utiliser un modèle de *mélange de copules gaussiennes* car elles permettent, d'une part, de modéliser la dépendance intra-classe par une matrice de corrélation robuste et, d'autre part, de visualiser les données par classe. Ensuite, on choisira des distributions de la famille exponentielle pour les marginales des composantes car leur interprétation est facile.

Cet article est organisé de la manière suivante. La partie 2 introduit le nouveau modèle de dépendance dont les propriétés principales sont présentées dans la partie 3. La partie 4 traite des problématiques d'estimation des paramètres et de choix de modèle, dans un cadre bayésien. Une illustration du modèle, sur un jeu de données médical, est proposée dans la partie 5. La partie 6 conclut ce travail et discute de ses extensions futures.

2 Modèle de mélange de copules gaussiennes

Modèle de mélange Soit le vecteur de d variables mixtes $\boldsymbol{x} = (x^1, \ldots, x^d) \in \mathcal{X}$ dont les d_{C} premiers éléments correspondent aux variables continues notées $\boldsymbol{x}^{\mathrm{C}}$, et dont les $d - d_{\mathrm{C}}$ autres correspondent aux variables discrètes (entières, ordinales et binaires) notées $\boldsymbol{x}^{\mathrm{D}}$. Les données $\boldsymbol{x}$ sont supposées issues d'un mélange de g distributions paramétriques dont la fonction de distribution de probabilité (fdp) est

$$p(\boldsymbol{x}; \boldsymbol{\theta}) = \sum_{k=1}^{g} \pi_k p(\boldsymbol{x}; \boldsymbol{\alpha}_k), \tag{1}$$

où $\boldsymbol{\theta} = (\boldsymbol{\pi}, \boldsymbol{\alpha})$ regroupe l'ensemble des paramètres. Le vecteur $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)$ rassemble les proportions de chaque classe k notées π_k, avec $0 < \pi_k \leq 1$ et $\sum_{k=1}^{g} \pi_k = 1$, tandis que le vecteur $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_g)$ regroupe les paramètres de chaque composante k notés $\boldsymbol{\alpha}_k$.

Dépendances intra-classe et fonctions de répartition Les copules déterminent les dépendances au travers des fonctions de répartition (fdr). Le modèle suppose que la composante k suit une copule gaussienne de matrice de corrélation $\boldsymbol{\Gamma}_k$ dont la fdr s'écrit

$$P(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \Phi_d(\Phi_1^{-1}(u_k^1), \ldots, \Phi_1^{-1}(u_k^d); \mathbf{0}, \boldsymbol{\Gamma}_k), \tag{2}$$

où $u_k^j = P(x^j; \boldsymbol{\beta}_{kj})$ est la valeur de la fdr de la marginale j, pour la composante k, paramétrée par $\boldsymbol{\beta}_{kj}$ et où $\boldsymbol{\alpha}_k = (\boldsymbol{\beta}_k, \boldsymbol{\Gamma}_k)$, avec $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}, \ldots, \boldsymbol{\beta}_{kd})$. Les fonctions $\Phi_d(.; \mathbf{0}, \boldsymbol{\Gamma}_k)$ et $\Phi_1(.)$ correspondent respectivement aux fdr de $\mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_k)$ et de $\mathcal{N}_1(0, 1)$.

Un vecteur latent continu sachant la classe Le *modèle de mélange de copules gaussiennes* peut alors s'exprimer à partir de *deux vecteurs latents* : un vecteur *qualitatif* utilisant un codage condensé $z \in \{1, \ldots, g\}$ indiquant l'appartenance des individus aux classes et un vecteur *gaussien* $\boldsymbol{y} = (y^1, \ldots, y^d) \in \mathbb{R}^d$. En effet, si $\boldsymbol{y}|z \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_z)$ et que $\forall j = 1, \ldots, d$, $x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{zj})$ alors la composante z est une copule gaussienne. On en déduit le modèle génératif suivant en trois étapes :

- échantillonnage de la classe : $z \sim \mathcal{M}(\pi_1, \ldots, \pi_g)$
- échantillonnage de la copule : $\boldsymbol{y}|z \sim \mathcal{N}_d(\mathbf{0}, \boldsymbol{\Gamma}_z)$
- calcul des données observées : $\boldsymbol{x}|z, \boldsymbol{y}$ est déterministe $\forall j\ x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{zj})$.

Fonction de distribution de probabilité Pour chaque composante, nous utilisons les marginales suivantes car elles facilitent l'interprétation du modèle et sont d'estimation simple : gaussiennes (x^j continue), de Poisson (x^j entière), multinomiales (x^j binaires ou ordinales). L'application $P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{zj})$ est alors bijective si $1 \leq j \leq d_{\mathrm{C}}$ et surjective sinon. En notant $\boldsymbol{y}^{\mathrm{C}}$ les d_{C} premiers éléments de $\boldsymbol{y}$ et $\boldsymbol{y}^{\mathrm{D}}$ les $d - d_{\mathrm{C}}$ derniers éléments de $\boldsymbol{y}$, on définit l'espace des antécédents de $\boldsymbol{x}^{\mathrm{D}}$ pour la classe k par $\mathcal{D}_k = \mathcal{D}_k^{d_{\mathrm{C}}+1} \times \ldots \times \mathcal{D}_k^d$ où $\mathcal{D}_k^j = \{y^j : x^j = P^{-1}(\Phi_1(y^j); \boldsymbol{\beta}_{zj})\}$, $1 \leq j \leq d$. Ainsi, la fdp de la composante k s'écrit

$$p(\boldsymbol{x}; \boldsymbol{\alpha}_k) = \frac{\phi_{d_{\mathrm{C}}}(\boldsymbol{y}^{\mathrm{C}}; \mathbf{0}, \boldsymbol{\Gamma}_{k\mathrm{CC}})}{\prod_{j=1}^{d_{\mathrm{C}}} \sigma_{kj}} \int_{\mathcal{D}_k} \phi_{d_{\mathrm{D}}}(\boldsymbol{y}^{\mathrm{D}}; \boldsymbol{\mu}_k^{\mathrm{D}}, \boldsymbol{\Sigma}_k^{\mathrm{D}}) d\boldsymbol{y}^{\mathrm{D}}, \tag{3}$$

où σ_{kj} est l'écart-type de la gaussienne j de la composante k, où $\boldsymbol{\Gamma}_k = \begin{bmatrix} \boldsymbol{\Gamma}_{k\mathrm{CC}} & \boldsymbol{\Gamma}_{k\mathrm{CD}} \\ \boldsymbol{\Gamma}_{k\mathrm{DC}} & \boldsymbol{\Gamma}_{k\mathrm{DD}} \end{bmatrix}$ est décomposée en sous-matrices, par exemple $\boldsymbol{\Gamma}_{k\mathrm{CC}}$ est la sous-matrice composée des d_{C} premières lignes et colonnes de $\boldsymbol{\Gamma}_k$, où $\boldsymbol{\mu}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \boldsymbol{y}_k^{\mathrm{C}}$ et où $\boldsymbol{\Sigma}_k^{\mathrm{D}} = \boldsymbol{\Gamma}_{k\mathrm{DD}} - \boldsymbol{\Gamma}_{k\mathrm{DC}} \boldsymbol{\Gamma}_{k\mathrm{CC}}^{-1} \boldsymbol{\Gamma}_{k\mathrm{CD}}$.

3 Propriétés du modèle

En faveur de composantes interprétables Chaque classe est résumée par sa proportion, par les paramètres de ses distributions marginales et par la matrice des corrélations de la copule quantifiant la dépendance intra-classe entre chaque couple de variables. Ainsi, en notant $\nu_{kj} = \mathrm{card}(\boldsymbol{\beta}_{kj})$ le nombre de paramètres de la marginale j pour la composante k , le modèle nécessite $\nu = (g-1) + g(\frac{d(d-1)}{2}) + \sum_{k=1}^{g} \sum_{j=1}^{d} \nu_{kj}$ paramètres.

Des coefficients de corrélations uniformisés La copule gaussienne de chaque classe fournit un coefficient de corrélation robuste. En effet, dans le cas continu, il est égal

à la borne supérieure des coefficients de corrélation obtenus par toutes transformations monotones de ces variables (Klaassen and Wellner, 1997). De plus, dans le cas de variables discrètes, il est égale au coefficient de corrélation *polychorique* (Olsson, 1979).

Visualisation des données par classe On utilise les paramètres du modèle pour effectuer une représentation des individus par classe. Pour une classe fixée, cela consiste à faire une ACP sur l'espace de la copule gaussienne, ce qui revient à faire la décomposition spectrale de $\mathbf{\Gamma}_k$. Les individus sont représentés par les coordonnées de $\mathbb{E}[\boldsymbol{y}|\boldsymbol{x}, z = k; \boldsymbol{\alpha}_k]$ sur le plan factoriel comme illustré par la figure 1. Les individus issus de la classe k suivent une loi normale centrée dans le plan ACP et sont donc "proches" de l'origine, tandis que ceux issus d'une autre classe, ne sont pas d'espérance nulle et s'en retrouvent "éloignés".

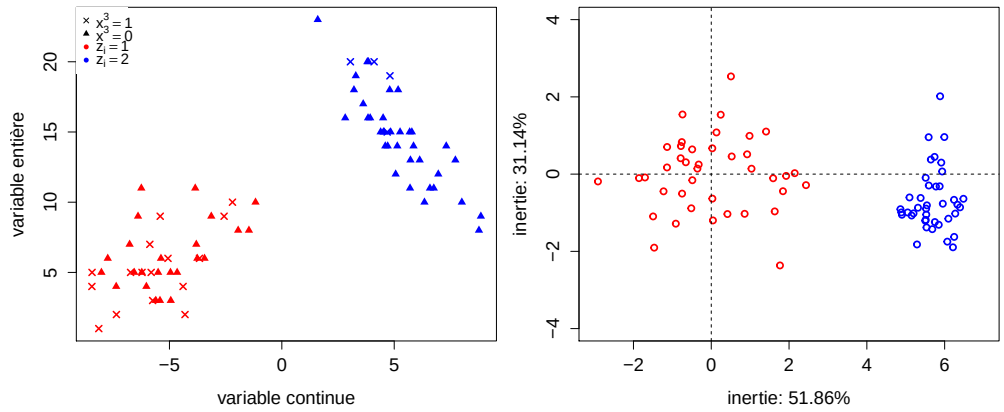

FIGURE 1 – Exemple d'un mélange de copule gaussienne à deux composantes avec une variable continue (abscisse), une entière (ordonnée) et une binaire (symbole), et de sa visualisation des individus, avec les paramètres la classe 1, sur le premier plan ACP.

4 Estimation des paramètres et choix de modèle

Estimation bayésienne des paramètres À partir d'un échantillon $\mathbf{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$ de n individus *i.i.d.*, on introduit le vecteur latent qualitatif $\mathbf{z} = (z_1, \ldots, z_n)$ et les vecteurs latents gaussiens $\mathbf{y} = (\boldsymbol{y}_i; i = 1, \ldots, n)$. L'estimateur du maximum *a posteriori* (MAP) est obtenu par l'échantillonneur de Gibbs, ayant $p(\boldsymbol{\theta}|\mathbf{x})$ comme loi stationnaire. Une de ses itérations se déroule selon les quatre étapes suivantes :

$$\mathbf{z}|\mathbf{x}, \mathbf{y}, \boldsymbol{\theta} \tag{4}$$

$$\boldsymbol{\pi}|\mathbf{x}, \mathbf{y}, \mathbf{z} \tag{5}$$

$$\forall (k, j) \quad (\boldsymbol{\beta}_{kj}, \mathbf{y}^j)|\mathbf{x}, \mathbf{y}^{\setminus j}, \mathbf{z}, \boldsymbol{\theta}^{\setminus kj} \tag{6}$$

$$\forall k \quad \mathbf{\Gamma}_k|\mathbf{x}, \mathbf{y}, \mathbf{z}, \tag{7}$$

où $\mathbf{y}^{\setminus j}$ est l'échantillon gaussien $\mathbf{y}$ privé de la variable j et où $\boldsymbol{\theta}^{\setminus kj}$ le vecteur $\boldsymbol{\theta}$ privé des paramètres de la marginale j de la composante k. Les équations (4) et (5) sont classiques. L'équation (6) se fait par un algorithme de Métropolis-Hastings comme proposé par Pitt et al. (2006). Cependant, les distributions étant de la famille exponentielle, on utilise les *priors* non informatifs de Jeffreys pour définir $p(\boldsymbol{\beta}_{kj}|\mathbf{x}^j, \mathbf{z})$ comme distribution *proposale*, car plus la dépendance intra-classe sera faible, plus elle sera proche de la distribution *a posteriori*. Enfin, l'équation (7), comme proposé par Hoff (2007), consiste à simuler une matrice de covariance par une distribution de Wishart que l'on normalise ensuite.

Choix de modèle par un critère bayésien Ici, le problème du choix de modèle est confiné au choix du nombre de classes. L'échantillonneur de Gibbs permet d'obtenir l'estimateur du maximum *a posteriori* pour un nombre fixé de classes. Ainsi, on peut appliquer un critère BIC pour sélectionner le nombre de classes.

5 Application

Les données On souhaite classifier un jeu de données (Czerniak and Zarzycki, 2003) décrivant 120 patients par cinq variables binaires (nausées fréquentes *Nau*, douleur lombaires *Lom*, envies pressantes d'uriner *Pre*, douleurs en urinant *Uri*, brûlure de l'urètre *Bru*) et une variable continue (température du patient *Tem*).

Résultats Au vu du critère BIC, le modèle le mieux adapté possède deux composantes. La classe 1, majoritaire ($\pi_1 = 0,56$), regroupe les individus ayant une température normale, une absence de nausée (non $= 0,98$) et peu de douleurs lombaires (non $= 0,59$). La classe 2, minoritaire ($\pi_2 = 0,44$), regroupe les individus ayant davantage de problèmes pour ces trois variables. Si les trois autres variables semblent très peu discriminantes au vu de leurs paramètres marginaux (valeurs très proches pour les deux classes), elles interviennent dans la règle de classification à travers les corrélations intra-classe (classe 1 : Nau-Uri $= -0,34$, Lum-Uri $= -0,65$; classe 2 : Nau-Uri $= 0,71$, Lum-Uri $= 0,42$). Les deux variables les plus discriminantes sont la présence de nausées et la température. Ceci est illustré par la figure 2, où l'on constate que les fdp de la variable température se chevauchent très peu. De plus, l'axe 2 de l'ACP construit avec les paramètres de la classe 1, qui est très discriminant, est fortement corrélé avec la variable température.

6 Conclusion

Nous avons, à travers l'application précédente, montré l'intérêt du modèle de mélange de copules gaussiennes pour la classification de données mixtes. Celui-ci permet de classifier des données de différentes natures mais ordonnées (nécessité d'avoir une fdr). L'extension au cas des données qualitatives non binaires est à l'étude. Le modèle actuel donnera

lieu à un package R prochainement. Enfin, remarquons que le modèle peut aussi être utilisé en classification semi-supervisée et supervisée.

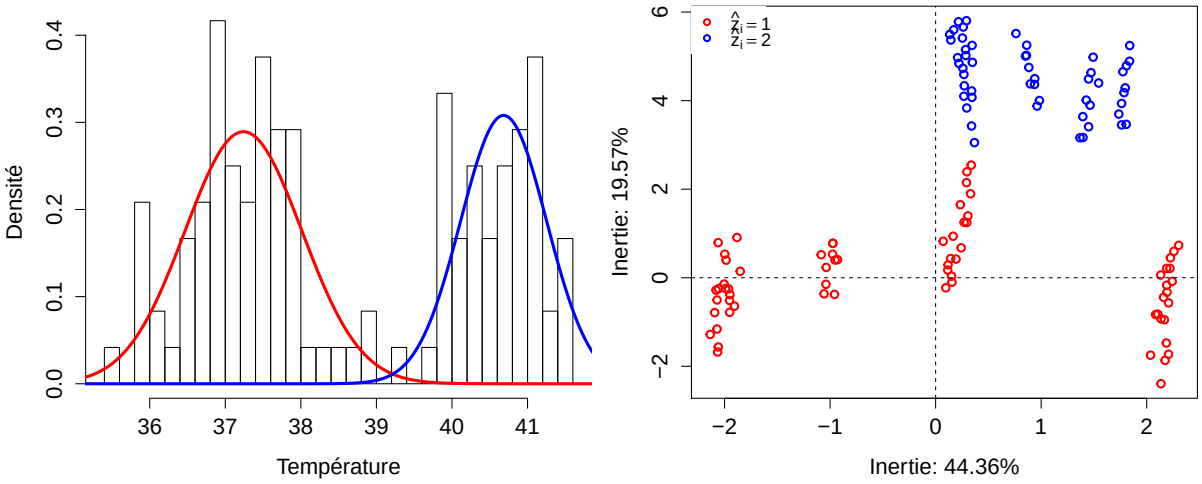

FIGURE 2 – Histogramme et fdp des composantes avec l'estimateur du MAP pour la variable température et nuage de points sur le 1er plan ACP pour la classe 1.

Références

J. Czerniak and H. Zarzycki. Application of rough sets in the presumptive diagnosis of urinary system diseases. *Artifical Inteligence and Security in Computing Systems, ACS'2002 9th International Conference Proceedings*, pages 41–51, 2003.

P.D. Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, pages 265–283, 2007.

C.A.J. Klaassen and J.A. Wellner. Efficient estimation in the bivariate normal copula model : normal margins are least favourable. *Bernoulli*, 3(1) :55–77, 1997.

G.J. McLachlan and D. Peel. *Finite mixutre models*. Wiley Series in Probability and Statistics : Applied Probability and Statistics, Wiley-Interscience, New York, 2000.

U. Olsson. Maximum likelihood estimation of the polychoric correlation coefficient. *Psychometrika*, 44(4) :443–460, 1979.

M. Pitt, D. Chan, and R. Kohn. Efficient Bayesian inference for Gaussian copula regression models. *Biometrika*, 93(3) :537–554, 2006.