Blocs de distributions à un facteur pour données binaires.

Mohammed Sedki [1] & Matthieu Marbac [2]

[1] *University Paris-Sud, France, mohammed.sedki@u-psud.fr*
[2] *University of McMaster, Canada, marbacm@math.mcmaster.ca*

Résumé. Nous introduisons une nouvelle famille de distributions à un facteur pour modéliser un vecteur de données binaires. Ce modèle fournit la probabilité de chaque événement, évitant ainsi les approximations numériques généralement nécessaires lorsque des dépendances sont modélisées pour ce type de données. De plus, il permet de décrire chaque variable par deux paramètres continus (donnant la probabilité marginale et la force de dépendance avec les autres variables) et par un paramètre binaire (définissant le signe de la dépendance). Une extension de ce modèle est proposée en supposant que les variables sont reparties en blocs indépendants qui suivent cette nouvelle distribution à un facteur. Les algorithmes d'estimation par maximum de vraisemblance et de choix de modèle (estimation des blocs de variables) sont implémentés dans le package R MvBinary. L'intérêt de cette distribution est illustré par une application sur le jeu de données "USA plants".

Mots-clés. Algorithme EM, copules à un facteur, données binaires, procédure IFM.

Abstract. We introduce a new family of one factor distributions for modeling binary vectors. This model gives the probability of each event, thus avoiding the numerical approximations often made when dependencies are considered for this kind of data. Moreover, it permits to describe each variable with two continuous parameters (given the marginal probability and the strength of dependencies with the other variables) and one binary parameter (given the sign of the dependency). An extension of this model is proposed by assuming that variables are split into independent blocks which follow the new one factor distribution. The algorithms developed for maximum likelihood inference and model selection (estimation of the blocks of variables) are implemented in the R package MvBinary. The benefit of this distribution is illustrated with one application in natural science.

Keywords. Binary data, EM algorithm, IFM procedure, one-factor copulas.

1 Introduction

Puisque les variables binaires sont facilement accessibles mais peu discriminantes, les données binaires sont souvent composées de beaucoup de variables. Bien que ces données

apparaissent dans de nombreux domaines, les statisticiens se retrouvent souvent face à un manque de distributions multivariées pour les modéliser (Genest and Nešlehová, 2007). Plusieurs auteurs se sont intéressés aux propriétés nécessaires pour faciliter l'interprétation et l'estimation d'une distribution. Ainsi, Nikoloulopoulos (2013) liste cinq caractéristiques essentielles que nous résumons comme: grand panel de dépendance (positives et négatives); nombre de paramètres de dépendance; calcul explicite de la vraisemblance; stabilité lors de la marginalisation; pas d'influence d'un paramètre sur l'espace d'autres paramètres.

Les copules à un facteur permettent de modéliser des données en grande dimension tout en conservant un nombre de paramètres raisonnable. L'idée principale est de supposer que les dépendances entre les variables observées sont expliquées par une variable continue latente. Cette approche est utilisée pour modéliser des données continues (Krupskii and Joe, 2015), des valeurs extrêmes (Mazo et al., 2015) ou des variables ordinales (Nikoloulopoulos and Joe, 2013).

Dans ce travail, nous introduisons une nouvelle distribution à un facteur. Pour modéliser des structures de dépendances plus complexes, nous proposons de repartir les variables par blocs indépendants où chaque bloc suit la nouvelle distribution à un facteur. La famille de distribution ainsi obtenue respecte les caractéristiques de Nikoloulopoulos (2013). Cette distribution permet de décrire chaque variable par trois paramètres indiquant: sa probabilité marginale, sa force de dépendance avec les autres variables du bloc et le signe de cette dépendance.

Ce papier s'organise comme suit. La partie 2 introduit la nouvelle famille de distribution. La partie 3 discute de l'estimation des paramètres et de la sélection de modèle. La partie 4 illustre la nouvelle distribution sur un jeu de données "USA plants". La partie 5 conclut ce travail.

2 Modélisation d'un vecteur binaire

2.1 Indépendance entre blocs de variables

L'objectif est de modéliser un vecteur aléatoire de d variables binaires $\boldsymbol{X} = (X_1, \ldots, X_d)$. Pour considérer différents types de dépendance, les variables sont réparties en B blocs indépendants. Le vecteur $\boldsymbol{\omega} = (\omega_1, \ldots, \omega_d)$ détermine cette répartition des variables en blocs puisque $\omega_j = b$ signifie que X_j est affectée au bloc b avec $1 \leq b \leq \text{B}$. Ainsi, $\boldsymbol{\omega}$ définit un modèle qu'il faudra estimer à partir des observations. Les variables du bloc b, notées $\boldsymbol{X}_{\{b\}} = (X_j : \omega_j = b)$, sont mutuellement dépendantes (voir parties 2.2 et 2.3). La function de masse de probabilité (fmp) de la réalisation $\boldsymbol{x} = (x_1, \ldots, x_d)$ est

$$p(\boldsymbol{x}|\boldsymbol{\omega}, \boldsymbol{\theta}) = \prod_{b=1}^{\text{B}} p(\boldsymbol{x}_{\{b\}}|\boldsymbol{\theta}_b), \tag{1}$$

où $\boldsymbol{\theta} = (\boldsymbol{\theta}_b; b = 1, \ldots, \text{B})$ regroupe les paramètres du modèle, où $\boldsymbol{\theta}_b$ regroupe les paramètres du bloc b et où $p(.|\boldsymbol{\theta}_b)$ est la fmp associée aux variables du bloc b.

2.2 Distribution conditionnelle de blocs à un facteur

Dans le bloc b, les dépendances entre variables s'expliquent à travers une variable continue latente U_b qui suit une distribution uniforme sur $[0, 1]$. Plus précisément, les variables du bloc b sont indépendantes conditionnellement à U_b. Ainsi,

$$p(\boldsymbol{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b) = \prod_{j \in \Omega_b} p(x_j|u_b, \boldsymbol{\theta}_j), \tag{2}$$

où $\Omega_b = \{j : \omega_j = b\}$ est l'ensemble des indices des variables du bloc b, $\boldsymbol{\theta}_b = (\boldsymbol{\theta}_j; j \in \boldsymbol{\Omega}_b)$ et où $\boldsymbol{\theta}_j$ correspond aux paramètres de la distribution conditionnelle de X_j. Cette distribution est un produit de Bernoulli dont les paramètres sont définis en fonction de u_b. De sorte que, pour $j \in \Omega_b$,

$$p(x_j|u_b, \boldsymbol{\theta}_j) = p_j^{x_j}(1 - p_j)^{1-x_j} \text{ où } p_j = (1 - \varepsilon_j)\alpha_j + \varepsilon_j \mathbb{1}_{\{u_b < \alpha_j, \delta_j = 1\}} \mathbb{1}_{\{u_b > 1-\alpha_j, \delta_j = 0\}}, \tag{3}$$

où $\boldsymbol{\theta}_j = (\alpha_j, \varepsilon_j, \delta_j)$ correspond aux paramètres de X_j. Le paramètre continu $\alpha_j \in (0, 1)$ indique la probabilité marginale de $X_j = 1$ (*i.e* $\int_0^1 p(X_j = 1|u_b, \boldsymbol{\theta}_j)du_b = \alpha_j$). Le paramètre continu $\varepsilon_j \in (0, 1)$ indique la force de dépendance entre X_j et les autres variables de son bloc. Enfin le paramètre binaire $\delta_j \in \{0, 1\}$ indique la nature de la dépendance, puisque $\delta_j = 1$ si la variable j est corrélée positivement avec U_b et $\delta_j = 0$ sinon. Ainsi, deux variables X_j and $X_{j'}$ affectée au même bloc (*i.e.* $\omega_j = \omega_{j'}$) sont corrélées positivement si $\delta_j = \delta_{j'}$ et négativement si $\delta_j = 1 - \delta_{j'}$. Notons que des contraintes d'identifiabilité doivent être imposées dans certains cas (Marbac and Sedki, 2015).

2.3 Distribution de blocs à un facteur

La paramétrisation définie par (3) permet une interprétation facile des paramètres. Cependant, nous introduisons une seconde paramétrisation de cette distribution pour pouvoir facilement calculer la pmf du bloc b. Conditionnellement à U_{ω_j}, X_j suit une distribution de Bernoulli dont les paramètres sont déterminés par la relation entre u_{ω_j} et le réel $\beta_j = \alpha_j^{\delta_j}(1 - \alpha_j)^{1-\delta_j}$ qui correspond à la probabilité marginale de $X_j = \delta_j$. En effet, pour $u_{\omega_j} \in [0, \beta_j)$, la distribution conditionnelle de $X_j|u_{\omega_j}, \boldsymbol{\theta}_j$ est une distribution de Bernoulli $\mathcal{B}(\lambda_j)$ où $\lambda_j = (1-\varepsilon_j)\alpha_j + \varepsilon_j\delta_j$. De plus, pour $u_{\omega_j} \in [\beta_j, 1]$, $X_j|u_b, \boldsymbol{\theta}_j$ suit une distribution de Bernoulli $\mathcal{B}(\nu_j)$ avec $\nu_j = (1 - \varepsilon_j)\alpha_j + \varepsilon_j(1 - \delta_j)$. Ainsi, (3) s'écrit comme

$$p(x_j|u_b, \boldsymbol{\theta}_j) = \begin{cases} \lambda_j^{x_j}(1 - \lambda_j)^{1-x_j} & \text{si } 0 \leq u_b < \beta_j \\ \nu_j^{x_j}(1 - \nu_j)^{1-x_j} & \text{si } \beta_j \leq u_b < 1 \end{cases}. \tag{4}$$

Comme les réalisation de u_b ne sont pas observées, on modélise la distribution des variables observées $\boldsymbol{X}_{\{b\}}$. Ainsi, la pmf de $\boldsymbol{x}_{\{b\}}$ est définie par

$$p(\boldsymbol{x}_{\{b\}}|\boldsymbol{\theta}_b) = \int_0^1 p(\boldsymbol{x}_{\{b\}}|u_b, \boldsymbol{\theta}_b)du_b. \tag{5}$$

Pour montrer que (5) a une forme explicite, on introduit la permutation σ_b de Ω_b telle que pour $1 \leq j < j' \leq d_b$ on a $\beta_{(b,j)} \leq \beta_{(b,j')}$, où $\beta_{(b,j)} := \beta_{\sigma_b(j)}$ et où $d_b = \text{card}(\Omega_b)$ est le nombre de variables du bloc b. Ainsi, la pmf de $\boldsymbol{x}_{\{b\}}$,

$$p(\boldsymbol{x}_{\{b\}}|\boldsymbol{\theta}_b) = \sum_{j=0}^{d_b} (\beta_{(b,j+1)} - \beta_{(b,j)}) f_b(j; \boldsymbol{\theta}_b), \tag{6}$$

où $\beta_{(b,0)} = 0$, $\beta_{(b,d_b+1)} = 1$ et où la fonction $f_b(.)$ est définie par

$$f_b(j_0; \boldsymbol{\theta}_b) = \prod_{j=1}^{j_0} \nu_{(b,j)}^{x_{(b,j)}} (1 - \nu_{(b,j)})^{1-x_{(b,j)}} \prod_{j=j_0+1}^{d_b} \lambda_{(b,j)}^{x_{(b,j)}} (1 - \lambda_{(b,j)})^{1-x_{(b,j)}}, \tag{7}$$

où $x_{(b,j)} := x_{\sigma_b(j)}$ correspond à la j-ème variable du bloc b (au sens de la permutation σ_b) , $\lambda_{(b,j)} := \lambda_{\sigma_b(j)}$, $\nu_{(b,j)} := \nu_{\sigma_b(j)}$ et où $\prod_{j=j_0+1}^{j_0} = 1$.

3 Inference et sélection de modèle

À partir d'un échantillon *iid*, on souhaite estimer les paramètres du modèle en supposant $\boldsymbol{\omega}$ connu. Comme la distribution de chaque bloc est une copule à un facteur, l'inférence est effectuée par une procédure IFM dont Joe (2005) a montré la consistance. Cette estimation se fait en deux étapes. La première consiste à maximiser la vraisemblance sur les paramètres marginaux (ici α_j). La seconde consiste à maximiser cette fonction sur les paramètres de dépendances (ici ε_j et δ_j) en laissant les paramètres marginaux fixés à leur valeur obtenue lors de l'étape précédente. Si cette première étape est explicite, la seconde étape s'effectue par un algorithme EM qui utilise les variables latentes u_b pour maximiser la vraisemblance.

Le choix de modèle $\boldsymbol{\omega}$ se fait par le critère BIC avec l'estimateur issu de la procédure IFM (Gao and Song, 2010). Comme le nombre de modèles en compétition est trop grand pour effectuer une approche exhaustive, la sélection de modèle se fait en deux étapes. D'abord, une procédure de réduction du nombre de modèles candidats est appliquée afin d'obtenir seulement d modèles candidats. Ensuite, le modèle maximisant le critère BIC parmi ces d modèles est sélectionnés.

Les détails concernant l'estimation des paramètres et la sélection de modèle (notamment la consistance de la procédure de sélection de modèle et l'utilisation du critère BIC avec l'estimateur issu de la procédure IFM) sont présentés dans Marbac and Sedki (2015).

4 Modélisation de la présence de plantes en Amérique

Le jeu de données est issu de la base "USA plants" du 29 juillet 2015 et décrit 35583 plantes en indiquant leur présence dans $d = 69$ états (USA, Canada, Porto Rico, Iles vierges, Groenland et St Pierre and Miquelon). En modélisant la distribution des données, on cherche à caractériser la phytosociologique d'espèces présentes dans chaque état. De plus, on peut espérer mettre en évidence des dépendances géographiques entre les variables. Les données sont disponibles dans le package R MvBinary où la méthode proposée est implémentée.

Le modèle estimé est composé de 10 blocs de variables dépendantes. Toutes les dépendances estimées sont positives (*i.e.* $j = 1, \ldots, d$, $\hat{\delta}_j = 1$) et chaque bloc est composé de variables fortement dépendantes (*i.e.* grande valeurs de $\hat{\varepsilon}_j$). Par conséquent, la connaissance d'une variable d'un bloc apporte une grande information sur les autres variables du bloc. Par example, le bloc le plus dépendant est composé des huit états suivants: Ile du Prince Edward, Nouvelle Ecosse, New Brunswick, New Hampshire, Vermont, Maine, Québec et Ontario. Une plante est présente en Ontario avec une probabilité $\hat{\alpha}_{Ontario} = 0.14$ mais si on sait qu'elle est présente au Québec alors sa probabilité d'être aussi présente en Ontario est de 0.83. De plus, le bloc ayant le moins de dépendance est composé des trois états: Hawaii, Porto Rico et Iles vierges. Cette faible dépendance entre les variables s'explique par l'éloignement géographique entre ces trois états. Enfin, les paramètres α_j caractérisent les états par leur diversité. On observe que les régions froides obtiennent des faible valeurs de $\hat{\alpha}_j$ tandis que les états de la "sun-belt" obtiennent les plus grandes valeurs de ce paramètre. Notons que la répartition des variables par bloc présente une cohérence géographique comme le montre le Figure 1.

5 Discussion

Nous avons proposé une famille de distributions pour données binaires qui possède les cinq propriétés définissant une "bonne" distribution au sens de Nikoloulopoulos (2013). Cette famille de distributions s'interprète facilement puisque chaque variable est décrite par deux paramètres continus et un paramètre binaire. Comme le nombre de paramètres du modèle est une fonction linéaire du nombre de variables observées, on peut utiliser cette distribution pour modéliser des données binaires de dimension assez grande.

Plusieurs extensions de ce travail sont envisagées. Des versions parcimonieuses du modèles peuvent être obtenues en imposant des contraintes d'égalité entre les paramètres d'un même bloc. De plus des dépendances plus complexes pourrait être modélisées en considérant plus qu'un facteur pour la distribution des blocs. Cependant, l'estimation des paramètres et le calcul de la vraisemblance deviendrait alors plus complexe. En effet, la fmp du block b serait alors définie comme une somme de $(d_b + 1)^2$ termes alors qu'elle

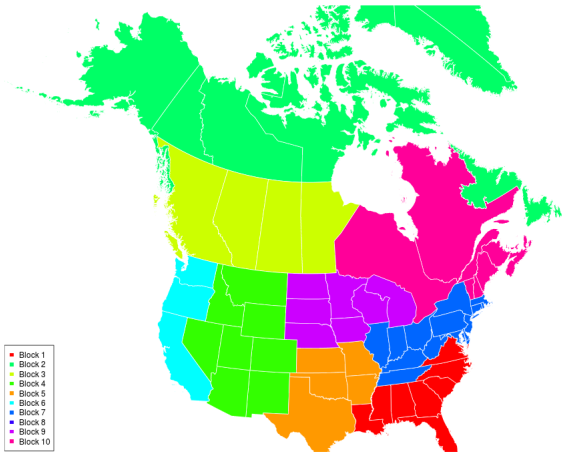

Figure 1: Carte indiquant (par la couleur) le bloc de chaque état.

n'est actuellement qu'une somme de $d_b + 1$ termes.

References

Gao, X. and Song, P. X.-K. (2010). Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492):1531–1540.

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bulletin*, 37(02):475–515.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.

Krupskii, P. and Joe, H. (2015). Structured factor copula models: theory, inference and computation. *J. Multivariate Anal.*, 138:53–73.

Marbac, M. and Sedki, M. (2015). A Family of Blockwise One-Factor Distributions for Modelling High-Dimensional Binary Data. *ArXiv e-prints - 1511.01343*.

Mazo, G., Girard, S., and Forbes, F. (2015). A flexible and tractable class of one-factor copulas. *Statistics and Computing*, pages 1–15.

Nikoloulopoulos, A. K. (2013). *Copula-based models for multivariate discrete response data*. Copulae in Mathematical and Quantitative Finance, Lecture Notes in Statistics, Springer-Verlag Berlin Heidelberg.

Nikoloulopoulos, A. K. and Joe, H. (2013). Factor copula models for item response data. *Psychometrika*, 80(1):126–150.