Inference rapide dans les modèles GLM à copule avec variables explicatives catégorielles en utilisant une procédure IFM -OSCFE

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan, Tom Rohmer

Inference rapide dans les modèles GLM à copule avec variables explicatives catégorielles en utilisant une procédure IFM -OSCFE

Alexandre Brouste[1], Christophe Dutang[2], Lilit Hovsepyan[1] & Tom Rohmer[3]

[1] *Laboratoire Manceau de Mathématiques, Le Mans Université, F-72000 Le Mans*
[2] *Université Grenoble Alpes, CNRS, Grenoble INP, LJK, F-38000 Grenoble*
[3] *GenPhySE, Université de Toulouse, INRAE, ENVT, F-31326 Castanet Tolosan*

Résumé. Dans les modèles linéaires généralisés multivariés à copule, des approches d'estimation basées sur le maximum de vraisemblance (MLE) joint peuvent être coûteuses en temps de calcul. Des méthodes alternatives (IFM) basées sur l'estimation des modèles marginaux là encore par MLE ont été proposés dans la littérature, pouvant là encore se révéler toujours coûteuses malgré le gain évident par rapport au MLE. Dans ce papier nous proposons une approche basée sur l'estimation des modèles marginaux utilisant un estimateur explicite consistant et asymptotiquement efficace proposé dans un papier récent, lorsque toutes les covariables du modèle sont catégorielles. Ce nouvel estimateur permet un gain réel en temps de calcul, sans perte sur la qualité d'estimation des paramètres du modèle en comparaison avec l'approche IFM classique.

Mots-clés. GLM, copules, IFM

Abstract. In copula multivariate generalized linear models, the approach based on joint maximum likelihood estimator (MLE) may be time consuming. Alternative methods based on inference on the marginals (IFM) which consider MLE marginal estimations was been proposed in the literature. Nevertheless, despite the gain in term of calculation time, these approaches may be again time consuming due to high numbers of explanatory variables or modalities. In this paper, we propose a IFM approach based on an explicit, consistent and asymptotically efficient estimator for the margins which is considered in a recent article when all the explicative variables are categorical. This new estimator allows a real gain in term of computation time comparatively to the classical IFM approach.

Keywords. GLM, copulas, IFM

1 Modèle d'inférence

Considérons un échantillon $\boldsymbol{Y} = (\underline{\boldsymbol{Y}}_1, \ldots, \underline{\boldsymbol{Y}}_n)$ composé de $\mathbb{R}^s$ vecteurs aléatoires indépendants avec pour $i = 1, \ldots, n$, $\underline{\boldsymbol{Y}}_i = (Y_{i,1}, \ldots, Y_{i,s})$. Les distributions marginales des $Y_{i,j}$, $j = 1, \ldots, s$ sont supposées appartenir à la famille exponentielle de paramètre naturelle $\lambda_{1j}, \ldots, \lambda_{nj}$ à valeur dans un espace $\Lambda_j \subset \mathbb{R}$.

En particulier, la vraisemblance $\mathcal{L}_{ij}$ associée à l'expérience statistique engendrée par $Y_{i,j}$, $i \in 1, \ldots, n$ et $j = 1, \ldots, s$ vérifie

$$\log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j \,|\, y_{i,j}) = \frac{\lambda_{ij}(\boldsymbol{\beta}_j) y_{i,j} - b_j\left(\lambda_{ij}(\boldsymbol{\beta}_j)\right)}{a_j(\phi_j)} + c_j(y_{i,j}, \phi_j), \quad y_{i,j} \in \mathbb{Y} \subset \mathbb{R}, \tag{1}$$

et $-\infty$ si $y_{i,j} \notin \mathbb{Y}$, où $a_j : \mathbb{R} \to \mathbb{R}$, $b_j : \Lambda_j \to \mathbb{R}$ et $c_j : \mathbb{Y} \times \mathbb{R} \to \mathbb{R}$ sont des fonctions mesurables (supposées connues) et ϕ_j est le paramètre de dispersion de la distribution, e.g. McCullagh & Nelder (1989, Section 2.2). Les paramètres $\lambda_{1j}, \ldots, \lambda_{nj}$ dépendent des paramètres auxiliaires inconnus $\boldsymbol{\beta}_j \in B_j \subset \mathbb{R}^{p_j}$, à estimer.

En utilisant une fonction dite de lien g_j deux fois continue différentiable et bijective de $b_j'(\Lambda_j)$ à $\mathbb{R}$, les GLM sont définis par une relation liant l'espérance des observations $\mathbf{E}Y_{i,j}$ aux predicteurs linéaires

$$g_j(\mathbf{E}Y_{i,j}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j = \eta_{ij}, \quad \text{pour tout } \boldsymbol{\beta}_j \in B_j,$$

où η_{ij} sont les prédicteurs linéaires, $\boldsymbol{x}_{ij} = (x_{ij}^{(1)}, \ldots, x_{ij}^{(m_j)})$, avec $x_{ij}^{(1)} = 1$ sont des vecteurs constitués par les m_j variables explicatives déterministes. En d'autres termes, les paramètres naturels s'écrivent $\lambda_{ij}(\boldsymbol{\beta}_j) = (b_j')^{-1} \circ g_j^{-1}(\eta_{ij})$.

Dans cette communication, on s'intéresse au cas où pour $j = 1, \ldots, s$, les m_j variables explicatives sont catégorielles avec $d_{\ell,j}$ modalités, $\ell = 1, \ldots, m_j$ et sont encodées en utilisant des variables binaires $x_i^{(\ell),k,j}$ valant 1 si la modalité k de la variable ℓ associée à la variable réponse j est choisie, et 0 sinon; voir Brouste et al. (2019, 2022). Sans perte de généralité, nous supposons que $\boldsymbol{x}_{ij}$ et $\boldsymbol{\beta}_j = (\beta_{1,j}, (\beta_{k,j}^{(\ell)})_{k,\ell})$ sont tels que le modèle soit identifiable voir Brouste et al. (2023). De plus, on s'intéresse (pour simplifier l'écriture) au modèle à effets simples uniquement que l'on peut réécrire

$$\begin{cases} g_1\left(\mathbf{E}Y_{i,1}\right) &= \beta_{1,1} + \displaystyle\sum_{\ell=2}^{m_1+1} \sum_{k=1}^{d_{\ell,1}} x_i^{(\ell),k,1} \beta_{k,1}^{(\ell)}, \\ &\vdots \\ g_s\left(\mathbf{E}Y_{i,s}\right) &= \beta_{1,s} + \displaystyle\sum_{\ell=2}^{m_s+1} \sum_{k=1}^{d_{\ell,s}} x_i^{(\ell),k,s} \beta_{k,s}^{(\ell)}. \end{cases}$$

Pour les modèles marginaux, Brouste et al. (2023) ont proposé un estimateur explicite, consistant et asymptotiquement efficace, alternatif au MLE (dont l'obtention par des méthodes itératives de descente de gradient peut être couteux en temps de calcul lorsqu'on considère un grand nombre de variables explicatives ou de modalités).

Dans ce contexte-ci, les variables réponses $Y_{i1}, \ldots, Y_{is}$ ne sont pas supposées indépendantes. Plus précisément nous supposons que la distribution jointe de $(Y_{i1}, \ldots, Y_{is})$ est caractérisée par (1) et par une copule paramétrique C_θ, où le paramètre θ est à estimer.

En utilisant le théorème de Sklar (1959), la log-vraisemblance de $\boldsymbol{y} = (\underline{\boldsymbol{y}}_1, \ldots, \underline{\boldsymbol{y}}_n)$ peut se réécrire:

$$\begin{aligned}\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\phi}, \theta | \, \boldsymbol{y}) &= \sum_{i=1}^{n} \log c_\theta(F_{i1}(y_{i,1}|\boldsymbol{\beta}_1, \phi_1), \ldots, F_{is}(y_{i,s}|\boldsymbol{\beta}_s, \phi_s)) + \sum_{i=1}^{n}\sum_{j=1}^{s} \log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j | y_{i,j}). \\ &= (a) + (b), \end{aligned} \tag{2}$$

où $\mathcal{L}_{ij}$ correspond à la vraisemblance associée à $y_{i,j}$ et c_θ la densitée de copule donnée par

$$c_\theta\left(u_1, \ldots, u_s\right) = \frac{\partial^s C_\theta\left(u_1, \ldots, u_s\right)}{\partial u_1 \ldots \partial u_s}.$$

Dans ces modèles multivariées, les paramètres du modèle peuvent être estimés par MLE. Néanmoins les méthodes permettant d'accéder au MLE, peuvent là encore se révéler extrêmement coûteuses en temps de calcul. Une approche alternative est d'estimer les paramètres par une méthode d'inférence sur les marges (IFM) voir (Xu 1996, Joe 1997, 2005). Cette méthode consiste à estimer les paramètres marginaux $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_s)$ et de dispersion $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_s)$ de telle sorte à maximiser la somme (b) dans (2), c'est-à dire calculer les MLE des modèles marginaux. Puis, le ou les paramètres de copule θ sont estimés en utilisant les estimations des paramètres marginaux et de dispersion précédents de telle sorte à maximiser la somme (a) dans (2).

Bien que non efficace, on peut montrer que l'estimation résultante des paramètres (joints) possède de bonne propriétés asymptotiques (consistance forte, distribution asymptotique Gaussienne). Néanmoins, en suivant Brouste et al. (2023), les estimations IFM pourront rester très coûteuses en temps de calcul de par les estimations successives par MLE des modèles marginaux.

Dans cette communication, nous proposons une approche IFM-OSCFE, dans laquelle les estimations marginales seront remplacées par les approches one-step explicites (OSCFE), consistantes et asymptotiquement efficaces proposées dans Brouste et al. (2023). Les propriétés asymptotiques concernant l'estimation jointe des paramètres peuvent être facilement démontrées en suivant (Xu 1996).

2 Simulations

Dans la table 1, nous avons considéré un modèle bivarié ($s = 2$) avec deux variables explicatives catégorielles ($m_1 = m_2 = 3$) contenant 2 et 3 modalités ($d_{11} = d_{21} = 2$, $d_{12} = d_{22} = 3$) et $n = 10^5$ observations. Les copules considérées sont Clayton, Frank, Gumbel et Normal, voir Nelsen (2007). Les distributions marginales étaient des distributions gammas avec fonction de lien inverse. Les rhos de Spearman associés aux différentes copules étaient 0.4 ou 0.8, entraînant un paramètre de copule différent selon le type de copule. 100 runs de chaque scénario ont été réalisés. Pour l'ensemble des cas considérés, nous avons obtenu des estimations du paramètre de copule ainsi que des variances aussi proches du vrai paramètre, que ce soit en utilisant un MLE ou en utilisant un OSCFE sur les marginales dans l'approche IFM.

Spearman's ρ	Copula type	Theo. θ	(**IFM**) Mean θ		(**IFM**) Sd θ ($\times 10^3$)	
			MLE	**OSCFE**	**MLE**	**OSCFE**
0.4	Clayton	0.758	0.758	0.758	7.431	7.431
	Frank	2.610	2.613	2.613	20.86	20.83
	Gumbel	1.382	1.382	1.382	3.821	3.823
	Normal	0.416	0.416	0.416	2.242	2.248
0.8	Clayton	3.188	3.186	3.187	18.03	18.03
	Frank	7.902	7.900	7.902	32.98	33.01
	Gumbel	2.582	2.580	2.582	9.249	9.243
	Normal	0.814	0.813	0.813	1.094	1.091

Table 1: Valeur moyenne et écart-types des estimations du paramètre de copule θ en utilisant une approche IFM utilisant une estimation des paramètres marginaux par MLE et par OSCFE, pour $n = 10^5$ observations et $B = 100$ simulations pour le modèle Gamma-GLM bivarié avec liens inverses.

Pour 5 paramètres à estimer par marge, les approches IFM-MLE et IFM-OSCFE étaient comparables en temps de calculs mais pour ce petit nombre de paramètres, les deux approches sont déjà environ 75 fois plus rapides que l'approche MLE (estimation jointe sur 13 paramètres au total). Reprenant les simulations réalisées dans Brouste et al. (2023), à partir de 20 modalités pour chacune des deux variables explicatives, l'estimation des paramètres marginaux était d'environ 95 fois plus rapides en utilisant l'approche OSCFE que l'approche MLE; on peut donc sans trop de risque en conclure que le IFM-OSCFE sera nettement plus rapide que le IFM-MLE dans le cadre d'un modèle GLM multivarié à copule, ouvrant des perspectives prometteuses en terme d'application rapide et sélection de modèle efficace.

References

Brouste, A., Dutang, C., Hovsepyan, L. & Rohmer, T. (2023), 'One-step closed-form estimator for generalized linear model with categorical explanatory variables', *Statistics and Computing* **33**(6), 138.

Brouste, A., Dutang, C. & Rohmer, T. (2019), 'Closed-form maximum likelihood estimator for generalized linear models in the case of categorical explanatory variables: application to insurance loss modeling', *Computational Statistics* .

Brouste, A., Dutang, C. & Rohmer, T. (2022), 'A closed-form alternative estimator for glm with categorical explanatory variables', *Communications in Statistics-Simulation and Computation* pp. 1–17.

Joe, H. (1997), *Multivariate Models and Multivariate Dependence Concepts*, CRC press.

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

McCullagh, P. & Nelder, J. A. (1989), *Generalized linear models*, Vol. 37, CRC press.

Nelsen, R. B. (2007), *An introduction to copulas*, Springer Science & Business Media.

Sklar, M. (1959), 'Fonctions de repartition an dimensions et leurs marges', *Publ. inst. statist. univ. Paris* **8**, 229–231.

Xu, J. J. (1996), Statistical modelling and inference for multivariate and longitudinal discrete response data, PhD thesis, University of British Columbia.