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

Fast inference in copula GLMs for categorical explanatory variables using Inference for margins coupling with OneStep approach

from a joint work with

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan, and Tom Rohmer

JdS 2024, bordeaux

multivariate GLMs I

Let the sample $\boldsymbol{Y} = (\underline{\boldsymbol{Y}}_1, \ldots, \underline{\boldsymbol{Y}}_n)$ composed of $\mathbb{R}^s$–valued independent random vectors. For $i = 1, \ldots, n$, the vector $\underline{\boldsymbol{Y}}_i = (Y_{i,1}, \ldots, Y_{i,s})$ has marginals $Y_{i,j}$, $j = 1, \ldots, s$ belonging to a family of probability measures of one-parameter exponential type with respective natural parameters $\lambda_{1j}, \ldots, \lambda_{nj}$ which depend on parameters $\boldsymbol{\beta}_j$.

The likelihood $\mathcal{L}_{ij}$ associated to the statistical experiment generated by $Y_{i,j}$ verifies

$$\log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j \mid y_{i,j}) = \frac{\lambda_{ij}(\boldsymbol{\beta}_j) y_{i,j} - b_j\left(\lambda_{ij}(\boldsymbol{\beta}_j)\right)}{a_j(\phi_j)} + c_j(y_{i,j}, \phi_j).$$

multivariate GLMs I

Let the sample $\boldsymbol{Y} = (\underline{\boldsymbol{Y}}_1, \ldots, \underline{\boldsymbol{Y}}_n)$ composed of $\mathbb{R}^s$–valued independent random vectors. For $i = 1, \ldots, n$, the vector $\underline{\boldsymbol{Y}}_i = (Y_{i,1}, \ldots, Y_{i,s})$ has marginals $Y_{i,j}$, $j = 1, \ldots, s$ belonging to a family of probability measures of one-parameter exponential type with respective natural parameters $\lambda_{1j}, \ldots, \lambda_{nj}$ which depend on parameters $\boldsymbol{\beta}_j$.

The likelihood $\mathcal{L}_{ij}$ associated to the statistical experiment generated by $Y_{i,j}$ verifies

$$\log \mathcal{L}_{ij}(\boldsymbol{\beta}_j, \phi_j \mid y_{i,j}) = \frac{\lambda_{ij}(\boldsymbol{\beta}_j) y_{i,j} - b_j\left(\lambda_{ij}(\boldsymbol{\beta}_j)\right)}{a_j(\phi_j)} + c_j(y_{i,j}, \phi_j).$$

The GLMs are defined by assuming the following relation between the expectation $\mathbb{E} Y_{i,j} = b_j'(\lambda_{ij}(\boldsymbol{\beta}_j))$ and the linear predictors η_{ij} through link functions g_j:

$$g_j(\mathbb{E} Y_{i,j}) = \boldsymbol{x}_{ij}^T \boldsymbol{\beta}_j = \eta_{ij}.$$

Here, $\boldsymbol{x}_{ij}$ are vectors constituting by m_j deterministic explanatory variables.

multivariate GLMs II

In this setting, the variables $Y_{i1}, \dots, Y_{is}$ constituting $\underline{\boldsymbol{Y}}_i$ are not assumed independent. We consider a parametric copula for the joint distribution of $(Y_{i1}, \dots, Y_{is})$:

Sklar's Theorem (1959):

Let $\boldsymbol{Y} = (Y_1, \dots, Y_s)$ be a s-dimensional random vector with c.d.f. $\boldsymbol{F}$ and let $F_1, \dots F_s$ be the marginal c.d.f. of $\boldsymbol{Y}$ assuming continuous. Then it exists a unique function $C : [0,1]^s \to [0,1]$ such that:

$$\boldsymbol{F}(\boldsymbol{y}) = C\{F_1(y_1), \dots, F_s(y_s)\}, \qquad \boldsymbol{y} = (y_1, \dots, y_s) \in \mathbb{R}^s.$$

▷ The so called copula C characterize the dependence between the components of $\boldsymbol{Y}$.

Estimation procedure, IFM approach

Let $\boldsymbol{\alpha}_j = (\boldsymbol{\beta}_j, \phi_j)$. The log-likelihood of $\boldsymbol{y} = (\underline{\boldsymbol{y}}_1, \dots, \underline{\boldsymbol{y}}_n)$ can be written as:

$$\log \mathcal{L}(\boldsymbol{\alpha}, \theta | \boldsymbol{y}) = \sum_{i=1}^{n} \log c_\theta(F_1(y_{i,1}|\boldsymbol{\alpha}_1), \dots, F_s(y_{i,s}|\boldsymbol{\alpha}_s)) + \sum_{j=1}^{s}\sum_{i=1}^{n} \log \mathcal{L}_{ij}(\boldsymbol{\alpha}_j | y_{i,j}).$$

Estimation:

- MLE approach: $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\alpha}}_1, \dots, \hat{\boldsymbol{\alpha}}_s, \hat{\theta})$ is solution of

$$\left(\frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\alpha}_1}, \dots, \frac{\partial \log \mathcal{L}}{\partial \boldsymbol{\alpha}_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\boldsymbol{\xi}) = 0.$$

- IFM approach: $\hat{\boldsymbol{\xi}} = (\hat{\boldsymbol{\alpha}}_1, \dots, \hat{\boldsymbol{\alpha}}_s, \hat{\theta})$ is solution of

$$\left(\frac{\partial \log \mathcal{L}_1}{\partial \boldsymbol{\alpha}_1}, \dots, \frac{\partial \log \mathcal{L}_s}{\partial \boldsymbol{\alpha}_s}, \frac{\partial \log \mathcal{L}}{\partial \theta}\right)(\boldsymbol{\xi}) = 0.$$

One-Step Closed-form IFM (OSCFE-IFM) estimator

- OSCFE-IFM approach:
  - OneStep Closed form estimator for $\boldsymbol{\beta}_j$ (Brouste et al. 2023):

$$\hat{\beta}_j^\star = (Q_j^T Q_j)^{-1} Q_j^T g_j(\bar{Y}_{\cdot,j}), \quad \hat{\boldsymbol{\beta}}_j = \hat{\beta}_j^\star + \mathcal{I}_j(\hat{\beta}_j^\star)^{-1} S_j(\hat{\beta}_j^\star)$$

    where $\hat{\beta}_j^\star$ is a closed-form consistent (but not efficient) mean-based estimator of $\hat{\beta}$, $\mathcal{I}_j$ and S_j the fisher Information and the score function for the jth marginal

  - $\hat{\phi}_j = \arg\max_\phi \log \mathcal{L}_j(\hat{\boldsymbol{\beta}}_j, \phi; y_{1,j}, \dots, y_{n,j})$
  - Finally $\hat{\theta}$ is solution of

$$\frac{\partial \log \mathcal{L}}{\partial \theta}(\hat{\boldsymbol{\alpha}}_1, \dots, \hat{\boldsymbol{\alpha}}_s, \theta) = 0.$$

▷ The OSCFE-IFM $(\hat{\boldsymbol{\alpha}}_1, \dots, \hat{\boldsymbol{\alpha}}_s, \hat{\theta})$ is consistent, asymptotically Gaussian, and asymptotically equivalent to the IFM one!

100 simulations of the gamma-GLM model with single effects only, 2 response variables, 15 parameters to estimate, $n = 10^5$

Spearman's ρ	Copula type	Theo. θ	Mean $\hat{\theta}$		Sd $\hat{\theta}$	
			IFM	**OSCFE-IFM**	**IFM**	**OSCFE-IFM**
0.4	Clayton	0.758	0.758	0.758	0.007	0.007
	Frank	2.610	2.613	2.613	0.021	0.021
	Gumbel	1.382	1.382	1.382	0.004	0.004
	Normal	0.416	0.416	0.416	0.002	0.002
0.8	Clayton	3.188	3.187	3.187	0.018	0.018
	Frank	7.902	7.901	7.902	0.033	0.033
	Gumbel	2.582	2.582	2.582	0.009	0.009
	Normal	0.814	0.813	0.813	0.001	0.001

Computational times

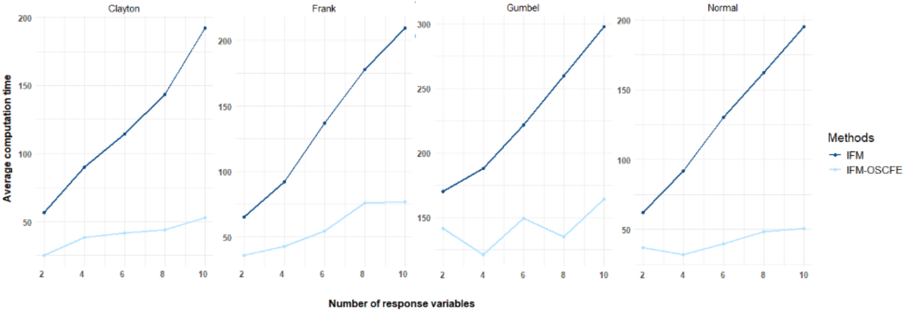

Figure: Copula parameter θ average computation time (sec.) for 4 copula types, $\rho = 0.8$, 100 simulations, 2 explanatory variables with 20 modalities and $n = 10^5$ observations for $s = 2$ to 10 response variables.

Conclusion

In multivariate GLMs,

- MLE is efficient but Fisher-Scoring procedure are totally time consuming
- IFM is a consistent estimator, but again, dealing with categorical explanatory variables with high number of modalities, the marginal estimations (by MLE) can remain time-consuming (Brouste et al. 2023)

Conclusion

In multivariate GLMs,

- MLE is efficient but Fisher-Scoring procedure are totally time consuming
- IFM is a consistent estimator, but again, dealing with categorical explanatory variables with high number of modalities, the marginal estimations (by MLE) can remain time-consuming (Brouste et al. 2023)

▷ IFM-OSCFE is consistent, whose the marginal estimations have closed-form and are asymptotically efficient. On the simulated data, the IFM-OSCFE solution is similar to the IFM but the calculations times are much lower.

▷ The improvement of this new estimator could be to propose a second joint-correction step to obtain a fast and asymptotically efficient estimator of joint parameters

Some biblio..

Alexandre Brouste, Christophe Dutang & Tom Rohmer
Closed form Maximum Likelihood Estimation for Generalized Linear Models in the case of categorical explanatory variables: application to insurance loss modelling
Computational Statistics, 2020

Alexandre Brouste, Christophe Dutang & Tom Rohmer
A closed-form alternative estimator for GLM with categorical explanatory variables
Communications in Statistics, Simulation and computation, 2022

Alexandre Brouste, Christophe Dutang & Tom Rohmer
glmtools: Tools to fit generalized linear models with explicit expressions
disponible à la demande

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan & Tom Rohmer
One-step closed-form estimator generalized linear model with categorigal explanatory variables
Statistics and Computing, 2023

Alexandre Brouste, Christophe Dutang, Lilit Hovsepyan & Tom Rohmer
Fast inference in copula models with categorical explanatory variables using one-step procedures
Article in progress, 2023