Robustness Methods for Modeling Count Data with General Dependence Structures

Marta Nai Ruscone[1], Dimitris Karlis[2]

[1] DIMA - University of Genoa; marta.nairuscone@unige.it
[2] AUEB - Athens; karlis@aueb.gr

Count data

Bivariate Poisson models are appropriate for modeling paired count data. However, the bivariate Poisson model does not allow for a negative dependence structure. Therefore, it is necessary to consider alternatives, which can produce both positive and negative dependence. Several models that can incorporate different structures and marginal properties are available, see, for example, Kocherlakota and Kocherlakota (1992) and Karlis and Ntzoufras (2003). See also the work in Nikoloulopoulos (2013) for defining models with copulas. While several extensions and models have been proposed, up to our knowledge, issues of robustness have been overlooked.

Motivation

Following Grunert da Fonseca and Fieller (2006), robustness that one should consider:

- contamination from outlier observations or, better, from observations that are unexpected under a certain model.
- model deviation, i.e., a researcher would like to fit the model with such a method that, even if the model is not correct, the method would protect from deriving inconsistent results.

Copulas

A **bivariate copula** $C : I^2$ with $I^2 = [0,1] \times [0,1]$ and $I = [0,1]$ is the **cumulative distribution function** (cdf) of a random variable (U, V), with uniform marginal random variable in $[0,1]$:

$$C(u, v; \theta) = P(U \leq u, V \leq v; \theta), \quad 0 \leq u \leq 1 \quad 0 \leq v \leq 1$$

where θ is a parameter measuring the **dependence** between U and V.

The **Sklar's theorem** (Nelsen, 2007) explains the use of the copula in the characterization of a joint distribution.

Bivariate count models based on copulas

For count data, a common starting point is to use the Poisson distribution for the marginals:

$$F(x) = \sum_{j=0}^{x} \frac{\exp(-\lambda_1)\lambda_1^x}{x!}$$

and

$$G(y) = \sum_{j=0}^{y} \frac{\exp(-\lambda_2)\lambda_2^y}{y!}$$

where $\lambda_1, \lambda_2 > 0$.
Take, for instance, Frank copulas:

$$C(u, v) = -\frac{1}{\theta}\log\{1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1}\}$$

Then, the marginal cumulative distribution functions are:

$$P(x, y) = C(F(x), G(y)) - C(F(x-1), G(y)) - C(F(x), G(y-1)) + C(F(x-1), G(y-1))$$

Minimum distance estimation

- Model robustness and efficiency can be achieved almost at the same time, i.e, by appropriately defining distances that in some sense downweight some observations (Lindsay, 1994).
- Minimum distance (MD) estimators can be interpreted (and they are) weighted likelihood estimators. The weights are determined by some kind of distance between observed and expected frequencies. For example, such an estimator can be based on Minimum Hellinger (MH) distance of the form

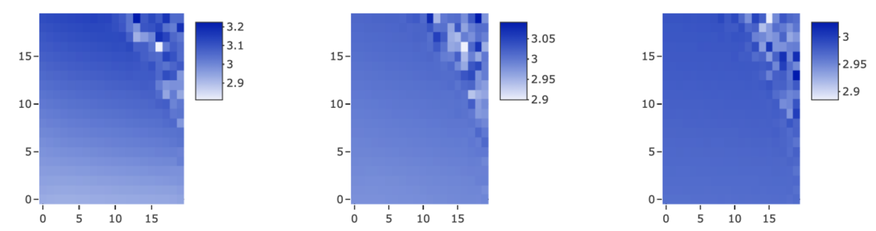
$$\sum_x \left(d(x)^{1/2} - m_\beta(x)^{1/2}\right)^2$$

where $d(x)$ is the observed relative frequency (or some other simple estimate of the probability at x) and $m_\beta(x)$ is the assumed model with parameters of interest β.
- We extend the approach to bivariate count data. x implies a pair of observations; parameters β are marginal distribution plus the copula parameter(s).

Robustness under outliers contamination and model misspecification

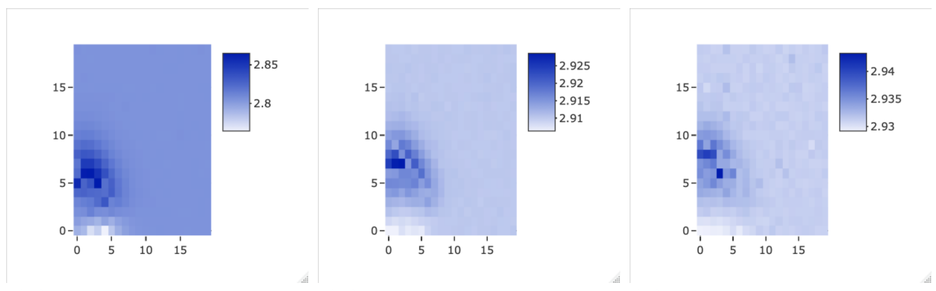
Particular focus is on the robustness of copula related parameters that measure the association exhibited by paired count data.

- Contamination from outliers observations in a bivariate Poisson with a Frank copulas (with $n = 100, 500, 1000$ observations) located in a different regions of the copula support.



(a) n=100 observations (b) n=500 observations (c) n=1000 observations

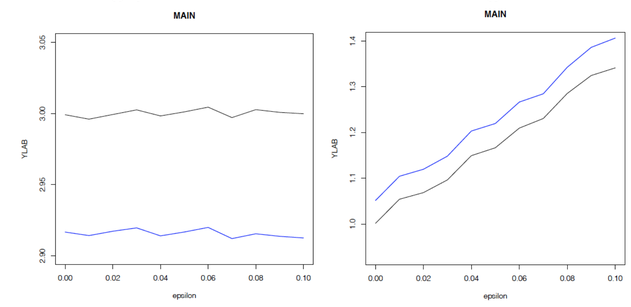
Figure: ML estimator with contamination from outliers observation in bivariate Poisson with Frank copula with parameter $\theta = 1$ and for the marginal Poisson model with means $\lambda_1 = \lambda_2 = 3$. The plots show the value of λ when contaminating the data with one observation located at x, y



(a) n=100 observations (b) n=500 observations (c) n=1000 observations

Figure: MHD estimator with contamination from outliers observation in bivariate Poisson with Frank copula with parameter $\theta = 1$ and for the marginal Poisson model with means $\lambda_1 = \lambda_2 = 3$. The plots show the value of λ when contaminating the data with one observation located at x, y

- ϵ-contaminated bivariate Poisson distribution with Frank copula containing varying proportions ϵ of contaminating points located at different regions at the copula support.



(a) $Poisson(\lambda = 1)$ (b) $Frank(\theta = 1)$

Figure: Bivariate Poisson with Frank copula. The copula parameter is $\theta = 1$ and the marginal distributions are Poisson with $\lambda = 1$ with n=500 observations ϵ-contaminated with a Gumbel copula with parameter $\gamma = 2$. Black = ML; Blue = MHD

Future Work

- Move up to higher dimensions

Main References

Grunert da Fonseca, V., Fieller, N. R. J. (2006). Distortion in statistical inference: the distinction between data contamination and model deviation. Metrika, 63, 169–190 (2006)

Karlis, D., Ntzoufras, I. (2003). Analysis of sports data using bivariate Poisson models. J. R. Stat.Soc. 52(3), 381–393.

Kocherlakota, S., Kocherlakota, K. (1992). Bivariate Discrete Distributions. First edition. CRC Press.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance andrelated methods. Ann. Stat. 22(2), 1081–1114

Nelsen, R.B. (2007). An introduction to copulas. Second edition. Springer, New York.

Nikoloulopoulos, A. K. (2013). Copula-based models for multivariate discrete response data. In P. Jaworski, F. Durante and W. K. Härdle (Eds.), Copulae in Mathematical and Quantitative Finance, pp. 231 – 249, Springer, Heidelberg.