Proceedings of the 4th International Conference on Statistics: Theory and Applications (ICSTA'22)
Prague, Czech Republic – July 28- 30, 2022
Paper No. 126
DOI: 10.11159/icsta22.126

Flexible Semiparametric Kernel Estimation with Bayesian Local Bandwidths and Diagnostics for Multivariate Count Data

Sobom M. SOME[1], Célestin C. KOKONENDJI[2]
[1] Laboratoire Sciences et Techniques/Université Thomas SANKARA,
12 BP 417 Ouagadougou 12, Ouagadougou, Burkina Faso
sobom.some@uts.bf
[2]Laboratoire de Mathématiques de Besançon UMR 6623 CNRS-UFC/Université Bourgogne Franche-Comté,
16 route de Gray, 25030 Besançon cedex, France.
celestin.kokonendji@univ-fcomte.fr

Extended Abstract

While parametric models have fulfilled a prominent role in terms of modelling multivariate data, nonparametric kernel smoothings [1] are recently envisaged for count data. Multivariate count data appear in a wide range of fields like environments (e.g., different kinds of plantation), marketing (e.g., purchases of different products) or epidemiology (e.g., different types of a disease).

In this paper, we elaborate two flexible semiparametric approaches governed by the multivariate Poisson with nonnegative cross correlations (from the common covariance μ_0 such that μ_0 =0 or μ_0 >0) for estimating multivariate probability mass functions. In the jungle of multivariate count distributions, we used the so-called generalized dispersion index [2] to compare several distributions between them. Our semiparametric method is then developed through expectation-maximisation scheme [3] and maximum likelihood method to estimate the parameters of the correlated and uncorrelated parametric Poisson departures. Also, the multiple binomial kernel with local Bayesian bandwidths [4] is used for the nonparametric part. Practical diagnostic criteria like the empirical integrated squared errors ISE_0 and the logarithm of the weight functions [5] are here opted to select the correct approach according to the data analysis. For the latter criterion, we precisely use the percentage of data points of this logarithmic function in pointwise confidence bands of ±1.96 with a view to choose between pure nonparametric approach (<5%), semiparametric ([5%, 95%]), and full parametric models (>95%). Notice that the adoption of a pure nonparametric model suggests the inconvenience of the parametric part considered in this model.

Numerical illustrations on six multivariate count datasets, including sales [6] data, have been performed for several scenarios of nonnegative cross correlations and generalized dispersion indexes. Many choices of conclusions are given according to the diagnostics models and data structure. Our approach is flexible enough and allow to consider either the current semiparametric model, or the started parametric, or a new one with another parametric departure, or simply the pure nonparametric model. Within the framework of improving the semiparametric technique with another *d*-variate count parametric start and therefore the relative dispersion index [2], a challenging problem to tackle would be suitable methods of estimations in this combined multivariate count model: parametric and nonparametric.

References

[1] Y. Esstafa, C.C. Kokonendji and S.M. Somé, “Asymptotic properties of the normalized discrete associated-kernel estimator for probability mass function,” in arXiv, arXiv:2202.10078, 2022.
[2] C.C. Kokonendji and P. Puig, “Fisher dispersion index for multivariate count distributions: a review and a new proposal,” in *J Multiv Anal*, 2018, vol. 165 pp. 180-193.
[3] D. Karlis, “An EM algorithm for multivariate Poisson distribution and related models,” in *J Appl Statist*, 2003, vol. 30 pp. 63-77.
[4] N. Belaid, S. Adjabi, C.C. Kokonendji and N. Zougab, “Bayesian local bandwidth selector in multivariate associated kernel estimator for joint probability mass functions,” in *J Statist Comput Simul*, 2016, vol. 86, pp. 3667-3681.
[5] C.C. Kokonendji and S.M. Somé, “Bayesian bandwidths in semiparametric modelling for nonnegative orthant data with diagnostics,” in *Stats*, 2021, vol. 4 pp. 162-183.
[6] Corporacion Favorita. Grocery Sales Data. Available: https://www.kaggle.com/c/favorita-grocery-salesforecasting/data, 2018 (accessed on 12 November 2021).