Supervised Component-based Generalized Linear Regression with finite mixture models of responses

Julien GIBAUD[1], Xavier BRY[1] and Catherine TROTTIER[1,2]

[1]Institut Montpelliérain Alexander Grothendieck, CNRS, Univ. Montpellier, France.
[2]Univ. Paul-Valéry Montpellier 3, F34000, Montpellier, France.

Contact : julien.gibaud@umontpellier.fr, xavier.bry@umontpellier.fr and catherine.trottier@univ-montp3.fr.

Abstract : Originally, the Supervised Component-based Generalized Linear Regression (SCGLR) methodology, proposed by Bry *et al.* (2013), was designed to find explanatory components in a large set of possibly highly redundant covariates. This methodology optimizes a trade-off criterion between the model's Goodness-of-Fit (GoF) and some Structural Relevance (SR) (Bry *et al.* (2015)) of directions with respect to the explanatory variables. This methodology allows both to find strong explanatory directions and to produce regularized predictors compatible with the high-dimensional framework. However, SCGLR assumes that all the responses are explained by the same explanatory dimensions. To overcome this limitation, we propose to extend this methodology to mixture models of the outcomes, enabling it to identify clusters of responses dependent on common explanatory components. Our work is based on the modeling approach of Dunstan *et al.* (2013) who propose using Finite Mixture Models (FMM) to analyze data communities, which they call Species Archetype Model (SAM). The idea is to assume that the responses (species) can be clustered into a small number of groups. In the spirit of Gibaud *et al.* (2020), we propose to cluster outcomes in groups predicted by the same supervised components built by SCGLR. In an ecological framework for instance, communities of species should be modeled by components characteristic of each community. The flexibility of our method allows to mix several probability distribution functions from the exponential families for the outcomes, dealing thus with presence-absence, count or biomass data. In order to estimate the model parameters, we implement an algorithm which alternatively evaluates the mixture parameters with the Expectation-Maximization (EM) algorithm and then finds the supervised components with the Projected Iterated Normed Gradient (PING) algorithm. For the sake of clarity, a few simple simulation studies illustrating the interest of our model will be presented.

Keywords : SCGLR, response mixture, EM algorithm, clustering.