Copula Mixture Regression Models for Multivariate Response Data

Claire Cui [1], Orla A. Murphy [1], and Paul D. McNicholas [2]

[1] Department of Mathematics and Statistics, Dalhousie University, Canada

[2] Department of Mathematics and Statistics, McMaster University, Canada

Email: claire.cui@dal.ca

Clustering, an unsupervised method in the realm of data analysis, serves as a powerful tool for uncovering hidden patterns and structures within complex datasets. In recent years, the use of mixtures of multiple linear regression models in clustering has gained popularity due to its ability to account for underlying heterogeneity in the data and provide a more representative interpretation of covariate effects. However, there is a paucity of these models for multivariate response cases, particularly when dealing with dependent responses. One approach that has been applied in the case of multivariate response data is copula regression models. Copulas can be seen as representing the dependence structure of a random vector and are joint distribution functions with uniform margins. In copula regression, a copula function is employed to induce dependence between different response variables through the random error term in the regression model. The concept behind using copula regression models is that they allow us to capture complex dependencies among variables while still maintaining flexibility and interpretability. By incorporating copulas into our analysis, we can better understand how different covariates affect each component of the multivariate response.

In this work, we propose a copula-based finite mixture of regression (CMixR) model for clustering and interpreting covariate effects in heterogeneous multivariate response data and present an ECM algorithm for estimation. The model performance is tested using a simulation study and through data analysis on the morphological properties of purple rock crabs. The results obtained from this model are shown to give excellent clustering performance, as evidenced by the high adjusted Rand index (ARI) values.