A STATA IMPLEMENTATION OF CLUSTER WEIGHTED MODELS: THE CWMGLM PACKAGE

Daniele Spinelli [1], Salvatore Ingrassia [2] and Giorgio Vittadini [1]

[1] Department of Statistics and Quantitative Methods, University of Milano Bicocca (e-mail: daniele.spinelli@unimib.it, giorgio.vittadini@unimib.it)

[2] Department of Economics and Business ,University of Catania, (e-mail: s.ingrassia@unict.it)

ABSTRACT: The Cluster-Weighted Model (CWM) is a member of the family of the Mixtures of Regression Models and it is referred as Mixture of Regression with Random Covariates. Currently, the only procedure for estimating these models is R package **flexcwm**. The aim of this article is to introduce a new software component, the Stata package **cwmglm** which estimates CWMs based on the most common generalized linear models. Our software also extends to Stata users the possibility of estimating parsimonious models of Gaussian distributions with alternative specifications of the variance matrix. **cwmglm** also calculates the the generalized coefficients of determination and bootstrap standard errors that are not currently available in **flexcwm**. We illustrate the use of **cwmglm** with real data on Covid-19 admissions.

KEYWORDS: cluster weighted models, clustering, parsimonious models, Stata.

1 Introduction

The *Cluster-Weighted Model* (CWM) is a member of the family of the Mixtures of Regression Models and it is also referred as Mixture of Regression with Random Covariates. The model has been first proposed under Gaussian assumptions (Gershenfeld *et al.*, 1999). Assuming random covariates relaxes the assumption of assignment independence by allowing the component distribution of the covariates to affect the assignment of the observations to the mixture components (Mazza *et al.*, 2018). A CWM models parametrically the joint density $p(x,y)$ of response variable Y and covariates X using the conditional density $p(y|x)$ and the marginal density $p(x)$. In Ingrassia *et al.* (2012) the CWM has been formulated in the statistical framework under Gaussian assumptions and Ingrassia *et al.* (2015) introduced a broad family of CWMs modeling discrete responses in which the conditional densities are assumed to belong to the exponential family and the covariates are of mixed-type. For such models, Di Mari *et al.* (2019) and Ingrassia & Punzo (2020) introduced local and overall coefficients of determination based on the decomposition of the deviance.

From the software point of view, Mazza *et al.* (2018) underlined the scarcity of packages aimed at estimating CWMs, the same authors developed **flexcwm** for R. To our knowledge, no other software is currently available. The aim of this article is to address such gap by introducing **cwmglm**, a Stata package focused on CMWs. Our software component is based on the framework of Ingrassia *et al.* (2012) and Ingrassia *et al.* (2015) and estimates mixtures of generalized linear models (GLMs) with random covariates. The supported families are Gaussian, Poisson and binomial. The supported marginalizations for the covariates are multivariate Gaussian, multinomial, binomial, and Poisson. The variance matrix of multivariate Gaussian covariates is parametrized according to Celeux & Govaert (1995). This feature is introduced in Stata for the first time with **cwmglm**. Other than extending the possibility of estimating CMWs to Stata users, **cwmglm** introduces new internal validity measures based on the generalized coefficient of determination and bootstrap-based inference, these features are not available in **flexcwm**.

2 Cluster Weighted Models

Assume a sample $(x_1,y_1),\dots,(x_n,y_n)$ concerning a response variable Y and a set of covariates X. Assume that the sample comes from a heterogeneous population formed by K latent classes. The CWM models the density of (Y,X) as outlined by Equation 1.

$$p(x,y,\theta) = \sum_{j=1}^{K} \pi_j p(y|x;\zeta_j)q(x;\psi_j) \tag{1}$$

In Equation 1, π_j is the mixing proportion of latent class j, $p(y|x;\zeta_j)$ is the class j-specific conditional density of the response variable and $q(x;\phi_j)$ is the marginal density of X in class j. Densities are characterized by parameters ζ_j and ϕ_j to be estimated. In our framework, the conditional density belongs to the exponential family and it is modeled as a GLM, while the marginal density $q(x;\psi_j)$ is modeled according to the Gaussian, Bernoulli, multinomial and Poisson distributions. Parameters are obtained by maximizing the log-likelihood corresponding to the density of Equation 1 using the expectation-maximization (EM) algorithm. Assuming $p(y|x;\zeta_j)=1$ in Equation 1 leads to a mixture of distributions, while $q(x;\psi_j) = 1$ leads to a finite mixture of regressions (FMR).

In Equation 1, assuming multivariate Gaussian covariates implies that $q(x;\psi_j) = \phi(\boldsymbol{x};\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$ where $\boldsymbol{\mu}_j$ is the mean vector and $\boldsymbol{\Sigma}_j$ is the variance matrix for latent class j (to be estimated). The eigenvalue decomposition of the variance matrix $\boldsymbol{\Sigma}_j = \lambda_j \boldsymbol{D}_j \boldsymbol{A}_j \boldsymbol{D}'_j$ (Celeux & Govaert, 1995) can be used to model cluster volume, shape and orientation. Combining constraints on λ_j (class volume), $\boldsymbol{D}_j$ (orientation) and $\boldsymbol{A}_j$ (shape) define fourteen parsimonious models. Specifically, clusters may be constrained to have equal or variable volume, spherical, equal or variable shape and axis-aligned, equal or variable orientation. For example, possible specifications may be based on the assumption that clusters have equal volume, equal shape, equal orientation (EEE) or that cluster are characterized by variable volume, equal shape and variable orientation (VEV).

3 The cwmglm package

The **cwmglm** module is available in the Statistical Software Components (SSC) archive, can be installed by using the Stata command *ssc install cwmglm* and fits CWMs as mixtures of the most common GLMs with random covariates. To our knowledge, the features of **cwmglm** are completely new for Stata users as CWMs are not estimable with the current availability of Stata commands. Indeed, **gsem** and **fmm** are only capable to estimate FMR and mixtures of distributions, which are nested in CMWs and estimable using **cwmglm**. In **cwmglm**, the parametrization of the class j-specific variance matrix of multivariate Gaussian covariates is based on Celeux & Govaert (1995). Such models are available in R packages such as **mclust** (Fraley & Raftery, 2007) and **clustvarsel** (Scrucca & Raftery, 2018) but not in Stata. Estimation of models with variable orientation and equal shape is based on Browne & McNicholas (2014).

R users can estimate CWMs using **flexcwm**; our package is related to it by extending its capability to Stata. Further, **cwmglm** provides some new procedures based on novel deviance-based measures of model fit (Di Mari *et al.*, 2019) and bootstrap standard errors.

Moreover, besides controlling the number of EM iterations, **cwmglm** users can control the number of iterations of the maximization procedures occurring during each EM iteration. This option is useful when, during a single maximization step, models requiring iterative estimation such as GLMs fail to converge. Such feature is not available in **flexcwm**.

4 Empirical example

The dataset includes a random sample of 1000 hospital admissions during the first Covid-19 wave (Feb 2020 - May 2020) in the hospital of Brescia, Italy. The response variable is the length of stay in days. The covariates are the day of admission, the patient's demographic characteristics, procedures and comorbidities. The empirical strategy is concerned in estimating CWMs for different numbers of mixture components and compare their fit.

References

BROWNE, R., & MCNICHOLAS, P. 2014. Estimating Common Principal Components in High Dimensions. *Advances in Data Analysis and Classification*, **8**(2), 217–226.

CELEUX, G., & GOVAERT, G. 1995. Gaussian Parsimonious Clustering Models. *Pattern recognition*, **28**(5), 781–793.

DI MARI, R., INGRASSIA, S., & PUNZO, A. 2019. A Generalized Coefficient of Determination for Mixtures of Regressions. *Pages 27–35 of: Conference of the International Federation of Classification Societies*. Springer-Verlag.

FRALEY, C., & RAFTERY, A. 2007. Model-Based Methods of Classification: Using the mclust Software in Chemometrics. *Journal of Statistical Software*, **18**, 1–13.

GERSHENFELD, N., SCHÖNER, B., & METOIS, E. 1999. Cluster-Weighted Modelling for Time-Series Analysis. *Nature*, **397**, 329–332.

INGRASSIA, S., & PUNZO, A. 2020. Cluster Validation for Mixtures of Regressions Via the Total Sum of Squares Decomposition. *Journal of Classification*, **37**(2), 526–547.

INGRASSIA, S., MINOTTI, S.C., & VITTADINI, G. 2012. Local Statistical Modeling via the Cluster-Weighted Approach with Elliptical Distributions. *Journal of Classification*, **29**(3), 363–401.

INGRASSIA, S., PUNZO, A., VITTADINI, G., & MINOTTI, S. 2015. The Generalized Linear Mixed Cluster-Weighted Model. *Journal of Classification*, **32**(1), 85–113.

MAZZA, A., PUNZO, A., & INGRASSIA, S. 2018. flexCWM: a Flexible Framework for Cluster-Weighted Models. *Journal of Statistical Software*, **86**(2), 1–30.

SCRUCCA, L., & RAFTERY, A. 2018. clustvarsel: A Package Implementing Variable Selection for Gaussian Model-Based Clustering in R. *Journal of Statistical Software*, **84**.