A Stata package for Cluster Weighted Modeling

Daniele Spinelli [1*] Salvatore Ingrassia [2] Giorgio Vittadini [1]

[1]University of Milan-Bicocca

[2]University of Catania
[*]presenting author (daniele.spinelli@unimib.it)

XVII Italian Stata Conference
Florence - 19 May 2022

Introducing **cwmglm**, a new package that allows users to estimate Cluster Weighted Models (CWM).
This package extends Stata capabilities in estimating finite mixture of regressions.

Cluster Weighted Models (CWM)

Given

- K latent classes
- a response variable Y
- set of covariates X.

A CWM (aka Mixture of Regressions with Random Covariates) assumes that the distribution of (Y, X):

$$p(x, y, \theta) = \sum_{j=1}^{K} \pi_j p(y|x; \beta_j, \phi_j) p(x; \alpha_j) \tag{1}$$

θ: model parameters to be estimated
π_j: the mixing proportion of class j ($\sum_{j=1}^{K} \pi_j = 1, \pi_j > 0$).

Conditional Distribution

- $p(y|x;\beta_j,\phi_j)$: class j-specific conditional part of the model
- the parametric distribution of $Y|X=x$ is modeled as a GLM with regression coefficients β_j and ancillary parameters ϕ_j

Marginal distribution

- $p(x;\alpha_j)$ is the parametric distribution of X in class j given model parameters α_j.
- if $p(x;\alpha_j) = \mathcal{N}(x,\mu_j,\Sigma_j)$ fourteen models are originated (Celeux and Govaert, 1995)
  - μ_j is the means vector
  - Σ_j is the variance-covariance matrix

Marginal distribution of normal covariates

If $p(x;\alpha_j) = \mathcal{N}(x, \mu_j, \Sigma_j)$. The eigenvalue decomposition of variance covariance matrix:

$$\Sigma_j = \lambda_j \boldsymbol{D}_j \boldsymbol{A}_j \boldsymbol{D}_j' \tag{2}$$

Geometrically:

- $\lambda_j = |\Sigma^{\frac{1}{d}}|$ is the cluster volume,
- $\boldsymbol{D}_j$ is the orientation (orthogonal matrix) ,
- $\boldsymbol{A}_j$ is the shape $(|\boldsymbol{A}_j = 1|)$.

Different assumptions can be made on λ, $\boldsymbol{D}$ and $\boldsymbol{A}$

Marginal distribution of normal covariates

Volume	Shape	Orientation	Model	Σ_j	N. parameters
Equal	Spherical		EII	$\lambda \boldsymbol{I}$	1
Variable	Spherical		VII	$\lambda_j \boldsymbol{I}$	K
Equal	Equal	Axis-Aligned	EEI	$\lambda \boldsymbol{A}$	d
Variable	Equal	Axis-Aligned	VEI	$\lambda_j \boldsymbol{A}$	K+d-1
Equal	Variable	Axis-Aligned	EVI	$\lambda \boldsymbol{A}_j$	1+K(d-1)
Variable	Variable	Axis-Aligned	VVI	$\lambda_j \boldsymbol{A}_j$	Kd
Equal	Equal	Equal	EEE	$\lambda \boldsymbol{DAD}'$	d(d+1)/2
Variable	Equal	Equal	VEE	$\lambda_j \boldsymbol{DAD}'$	K+d-1+d(d-1)/2
Equal	Variable	Equal	EVE	$\lambda \boldsymbol{DA}_j\boldsymbol{D}'$	1+K(d-1)+d(d-1)/2
Variable	Variable	Equal	VVE	$\lambda_j \boldsymbol{DA}_j\boldsymbol{D}'$	Kd+d(d-1)/2
Equal	Equal	Variable	EEV	$\lambda \boldsymbol{D}_j\boldsymbol{AD}_j'$	d+Kd(d-1)/2
Variable	Equal	Variable	VEV	$\lambda_j \boldsymbol{D}_j\boldsymbol{AD}_j'$	K+d-1+kD(D-1)/2
Equal	Variable	Variable	EVV	$\lambda \boldsymbol{D}_j\boldsymbol{A}_j\boldsymbol{D}_j'$	1+K(d-1)+Kd(d-1)/2
Variable	Variable	Variable	VVV	$\lambda_j \boldsymbol{D}_j\boldsymbol{A}_j\boldsymbol{D}_j'$	Kd(d+1)/2

K: number of latent classes, d: number of parameters in x

E-Step

The complete data log-likelihood of the CWM:

$$l(\theta)=\sum_{j=1}^{K}\sum_{i=1}^{N}\tau_{ij}ln(\pi_j)+\sum_{j=1}^{K}\sum_{i=1}^{N}\tau_{ij}ln[p(y_i|x_i;\beta_j,\phi_j)]+\sum_{j=1}^{K}\sum_{i=1}^{N}\tau_{ij}ln[p(X_{i,};\alpha_j)] \tag{3}$$

τ_{ij} is the estimated posterior probability for observation i to belong to component j.

During iteration t, and given the current expectation of θ^t:

$$\tau_{ij}^t = \frac{\pi_j^t p(y_i|x_i;\beta_j^t,\phi_j^t)p(x_{i,};\alpha_j^t)}{p(x,y,\theta^t)} \tag{4}$$

M- Step

The complete data log-likelihood of the CWM:

$$l(\theta)=\sum_{j=1}^{K}\sum_{i=1}^{N}\tau_{ij}ln(\pi_j)+\sum_{j=1}^{K}\sum_{i=1}^{N}\tau_{ij}ln[p(y_i|x_i;\beta_j,\phi_j)]+\sum_{j=1}^{K}\sum_{i=1}^{N}\tau_{ij}ln[p(X_{i,};\alpha_j)] \tag{5}$$

In the **M-Step** the three components of the log likelihood are maximized independently.

- the conditional part reduces to a GLM with weighted log-likelihood
- the marginal part reduces to the calculation of weighted means
- for normal covariates the procedures may be more complex (Celeux and Govaert, 1995; Sarkar et al., 2020)

Convergence criterion

To establish convergence the Aitken acceleration is applied, at iteration $t+1$:

$$a^{t+1} = \frac{l^{t+2} - l^{t+1}}{l^{t+1} - l^{t}} \quad (6)$$

Stopping criterion:

$$l_{\infty}^{r+2} - l^{r+1} < \varepsilon \quad (7)$$

where:

$$l_{\infty}^{t+2} = l^{r+1} + \frac{l^{t+2} - l^{t+1}}{1 - a^{t+1}} \quad (8)$$

flexcwm

Mazza et al. (2018) introduced R package **flexcwm**. Main features:

- Binomial, Poisson, Gaussian, t , Gamma, Inv. Gaussian GLMs
- Normal, binomial, Poisson, multinomial covariates
- all the fourteen parsimonious models
- Information criteria based model selection
- OIM standard errors for the GLMs

cwmglm

cwmglm implements in Stata the features of **flexcwm** plus:

- non-parametric bootstrap for inference
- GLM measures of fit: generalized coefficient of determination (Di Mari et al., 2019)

Main

The syntax of **cwmglm** is:
cwmglm *depvar indepvars* [if] [in], **posterior**(stub) [options]

- posterior(*stub*) Required option. Generates a set posterior group probabilities named *stub*1, *stub*2 ...
- k(#) Number of latent classes, the default is 2

Marginalization Options

- xnormal(varlist) variables having normal distributions
- xpoisson(varlist) variables having poisson distributions
- xbinomial(varlist) variables having binomial distributions
- xmultinomial(varlist) variables having multinomial distributions. Factor variable syntax is not allowed. Categories are detected automatically.

If xnormal is specified the user can model the variance-covariance matrix of normal covariates using one of the fourteen parsimonious models of Celeux and Govaert, 1995.

GLM options

family(familyname) specifies the distribution of depvar for the GLM (see glm).
Allowed families:

- family(gaussian) *(link identity)*
- family(binomial) *(link logit)*
- family(poisson) *(link log)*.

Initialization

The initialization procedure is controlled by **start(svmethod)**. Possibile options:

- start(kmeans) starting latent class membership is determined by running a kmeans cluster analysis on depvar indepvars. The default.
- start(custom) user-specified starting values. Starting values must be contained in **initial(varlist)** (k variables)

Initialization

The initialization procedure is controlled by **start(svmethod)**.
Possibile options:

- start(randomid) specifies that starting values are computed by randomly assigning observations to initial classes.
- start(randompr) specifies that starting values are computed by randomly assigning initial class probabilities.
- ndraws(#) specifies the number of random draws for selecting the starting values. Applies only to start(randompr) and start(randomid). Default is 10.

Postestimation - prediction

cwmglm allows **predict**. This command creates a new variable *varname* that assigns "hard" group membership according to the *maximum a posteriori probability*. The syntax is:
predict *varname*

Postestimation - prediction

default standard errors for **cwmglm** are based on the OIM of the weighted maximum likelihood problem of the GLM part. These standard errors are an underestimation as the weights are estimated as well. Model parameters can be estimated using **cwmbootstrap**.
The syntax is:
cwmbootstrap, nreps(#)

The students dataset

- 270 students attending the University of Catania. Variables: (height), father's height (heightf), weight (weight) and gender (gender).
- Replicating the model of Mazza et al. 2018 $\rightarrow$ treat gender as an unobservable and evaluate whether the CWM is able to discriminate females and males.
- conditional part, gaussian GLM: the dependent variable is weight while the covariates are height and heightf.
- marginal part, multivariate normal with equal size, equal shape and equal orientation (EEE).

The students dataset

```
. cwmglm w height heightf,  k(2)  posterior(z) xnormal(height heightf) eee
initializing EM...
EM iteration          1:log-likelihood=  -2721.68580
EM iteration         17:log-likelihood=  -2648.16080
Prior Probabilities
```

g1	g2
.5630098	.4369902

Clustering Table

g1	g2
153	117

Information criteria

AIC	BIC
5328.322	5385.896

The students dataset

Deviance measures and coefficient of determination

	g1	g2	Overall
Total	317.4497	285.3551	602.8048
Residual	151.0126	116.9874	268
Explained	77.92581	52.1048	130.0306
Between	88.51127	116.2629	204.7742
R_sq	.2454745	.1825964	.2157093

weight	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
g1						
height	.8983653	.0912865	9.84	0.000	.719447	1.077284
heightf	-.1442846	.0838213	-1.72	0.085	-.3085712	.0200021
_cons	-54.08234	12.12518	-4.46	0.000	-77.84725	-30.31742
g2						
height	.7612449	.1082026	7.04	0.000	.5491717	.9733182
heightf	-.0088664	.0939404	-0.09	0.925	-.1929862	.1752534
_cons	-57.28365	12.3717	-4.63	0.000	-81.53174	-33.03556

The students dataset

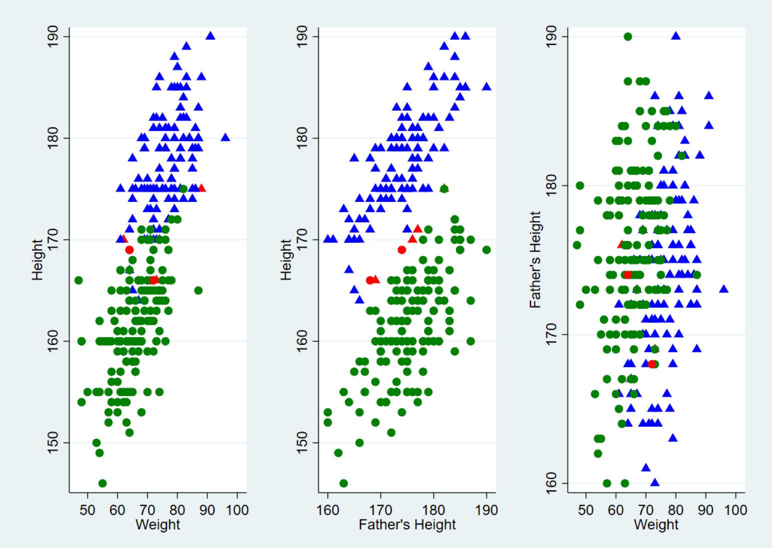
```
. matlist e(mu)

             |    height    heightf
-------------+----------------------
          g1 |  161.7553   175.6054
          g2 |  177.5373   174.1353
. matlist e(sigma)

             | g1                    | g2
             |    height    heightf  |    height    heightf
-------------+-----------------------+----------------------
      height |   27.8441   22.04352  |   27.8441   22.04352
     heightf |  22.04352   34.69652  |  22.04352   34.69652

.
```

The students dataset

```
. predict group
. tab group gender
```

group	Gender F	M	Total
1	149	4	153
2	2	115	117
Total	151	119	270

The students dataset



The students dataset

For females (g1) the estimated mean vector and variance-covariance matrix:

$$\mu_F = (161.75, 177.53)\ ,\ \Sigma = \begin{pmatrix} 27.84 & 22.04 \\ 22.04 & 34.69 \end{pmatrix}$$

```
. sum height  heightf if gender=="F"

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      height |        151    161.6887    5.286695        146        175
     heightf |        151    175.6093    5.816039        160        190

. corr height heightf if gender=="F", cov
(obs=151)

             |   height  heightf
-------------+------------------
      height |  27.9491
     heightf |  20.5709  33.8263
```

The students dataset

For males (g2) the estimated mean vector and variance-covariance matrix:

$$\mu_M = (177.53, 174.13) \; , \; \Sigma = \begin{pmatrix} 27.84 & 22.04 \\ 22.04 & 34.69 \end{pmatrix}$$

```
. sum height  heightf if gender=="M"

    Variable |        Obs        Mean    Std. dev.       Min        Max
-------------+---------------------------------------------------------
      height |        119    177.4874    5.255934        164        190
     heightf |        119    174.1429      6.0328        160        190

. corr height heightf if gender=="M", cov
(obs=119)

             |   height  heightf
-------------+------------------
      height |  27.6248
     heightf |  24.2942  36.3947
```

The students dataset

```
. cwmbootstrap, nreps(100)
Starting  replications
. . . . . . . . . . .  10
. . . . . . . . . . .  20
. . . . . . . . . . .  30
. . . . . . . . . . .  40
. . . . . . . . . . .  50
. . . . . . . . . . .  60
. . . . . . . . . . .  70
. . . . . . . . . . .  80
. . . . . . . . . . .  90
. . . . . . . . . . .  100

.
. matlist r(b)
```

	g1			g2		
	height	heightf	_cons	height	heightf	_cons
mean	.7773018	-.0016771	-61.43893	.8877987	-.1458781	-52.04102
sd	.2118676	.1742196	16.35427	.1328318	.0984212	14.06947
95% CI lcl	.3620414	-.3431475	-93.49329	.6274485	-.3387838	-79.61719
95% CI ucl	1.192562	.3397934	-29.38457	1.148149	.0470275	-24.46485
z	3.66881	-.0096262	-3.756752	6.683633	-1.482181	-3.69886

```
.
. matlist r(mu)
```

	height	heightf
mean	169.5946	174.8445
sd	7.936044	.869638
95% CI lcl	154.0399	173.14
95% CI ucl	185.1492	176.549
z	21.37017	201.0543

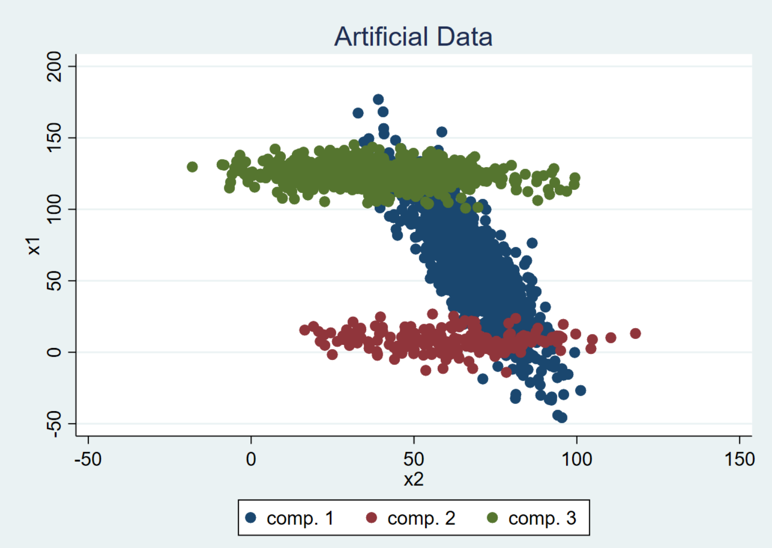
The multinorm dataset

Simulated data from the multivariate normal distribution

- $\mu_1 = (59, 68)$, $\Sigma_1 = \begin{pmatrix} 1351 & -358 \\ -358 & 136 \end{pmatrix}$, $N_1 = 1000(p_1 = 0.5208)$
- $\mu_2 = (8, 61)$, $\Sigma_2 = \begin{pmatrix} 47 & -12 \\ -12 & 378 \end{pmatrix}$, $N_2 = 200(p_2 = .1041)$
- $\mu_3 = (124, 40)$, $\Sigma_3 = \begin{pmatrix} 7407 & 1033 \\ 1033 & 728 \end{pmatrix}$, $N_3 = 720(p_3 = .375)$

cwmglm is used for estimate Gaussian mixture models with different numbers of components and different formulation of the variance covariance matrix (Celeux and Govaert, 1995). Model selection is based on BIC and AIC.

The multinorm dataset



The multinorm dataset

```
. local models vev evv vvv eei vei evi vvi eii vii eee vee eve vve eev
. local bestbic=10e20
. local bestaic=10e20
. cap matrix drop res
.
. foreach model of local models {
  2.         forval i=2/5 {
  3.                 cap drop _tau*
  4.                 qui cwmglm, xnorm(x1 x2) k(`i´) posterior(_tau) `model´
  5.                                 if (e(converged)==1) {
  6.                                         matrix ic=(e(ic),`i´, e(ll))
  7.                                                 matrix rownames ic= "`model´"
  8.                                                 matrix res = nullmat(res) \ ic
  9.                 local current_BIC=e(ic)[1,2]
 10.                 if (`current_BIC´<`bestbic´) {
 11.                         local bestbic=`current_BIC´
 12.                         local bestk_BIC=`i´
 13.                         local bestmodel_BIC `model´
 14.                         }
 15.                 local current_AIC=e(ic)[1,1]
 16.                 if (`current_AIC´<`bestaic´) {
 17.                         local bestaic=`current_AIC´
 18.                         local bestk_AIC=`i´
 19.                         local bestmodel_AIC `model´
 20.                         }
 21.                 }
 22.                 else di in red ///
>               "model `model´ with `i´ mixture component did not converge"
 23.         }
 24. }
model evv with 5 mixture component did not converge
model vvv with 4 mixture component did not converge
model eve with 4 mixture component did not converge
model eev with 5 mixture component did not converge
. di as result "best model according to BIC: k=`bestk_BIC´ type `bestmodel_BIC´"
best model according to BIC: k=3 type vvv
. di as result "best model according to AIC: k=`bestk_AIC´ type `bestmodel_AIC´"
best model according to AIC: k=3 type vvv
```

The multinorm dataset

The AIC- and BIC- minimizing model have $k = 3$ and VVV correlation matrix.

Prior Probabilities

g1	g2	g3
.3781614	.5182283	.1036103

Clustering Table

g1	g2	g3
779	966	175

Information criteria

AIC	BIC
34505.37	34599.89

. matlist e(mu)

	x1	x2
g1	124.2815	39.82697
g2	60.27864	67.52907
g3	8.044332	60.91576

. matlist e(sigma)

	g1		g2		g3	
	x1	x2	x1	x2	x1	x2
x1	49.85068	-23.94966	1299.013	-341.5892	50.51839	-8.000548
x2	-23.94966	379.8781	-341.5892	131.7972	-8.000548	399.377

The multinorm dataset

Estimated parameters

- Estimated as **g1**: $\mu_1 = (124.28, 39.82)$, $\Sigma_1 = \begin{pmatrix} 49.85 & -23.94 \\ -23.94 & 379.878 \end{pmatrix}$, $N_1 = 779(p_3 = .378)$
- Estimated as **g2**: $\mu_2 = (60.28, 67.53)$, $\Sigma_2 = \begin{pmatrix} 1299.013 & -341.59 \\ -341.59 & 131.79 \end{pmatrix}$, $N_2 = 966(p_1 = 0.518)$
- Estimated as **g3**: $\mu_3 = (8.04, 60.91)$, $\Sigma_3 = \begin{pmatrix} 50.52 & -8.00 \\ -8.00 & 399.37 \end{pmatrix}$, $N_3 = 175(p_2 = .1036)$

Data generating

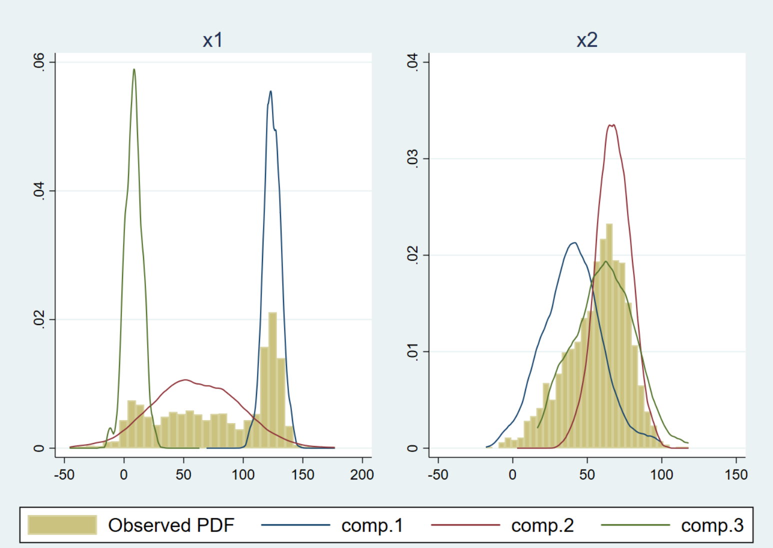
- $\mu_1 = (59, 68)$, $\Sigma_1 = \begin{pmatrix} 1351 & -358 \\ -358 & 136 \end{pmatrix}$, $N_1 = 1000(p_1 = 0.5208)$; **g2** is very similar
- $\mu_2 = (8, 61)$, $\Sigma_2 = \begin{pmatrix} 47 & -12 \\ -12 & 378 \end{pmatrix}$, $N_2 = 200(p_2 = .1041)$; **g3** is very similar
- $\mu_3 = (124, 40)$, $\Sigma_3 = \begin{pmatrix} 7407 & 1033 \\ 1033 & 728 \end{pmatrix}$, $N_3 = 720(p_3 = .375)$; **g1** is very similar

The multinorm dataset

```
. predict map
. tab map group
```

map	group 1	2	3	Total
1	78	0	$\boxed{701}$	779
2	$\boxed{907}$	40	19	966
3	15	$\boxed{160}$	0	175
Total	1,000	200	720	1,920

The multinorm dataset



The multinorm dataset