A Generalized Maximum Entropy (GME) approach to crisp-input/fuzzy-output regression model

Antonio Calcagnì

Department of Psychology and Cognitive Science

University of Trento (Italy)

SIS 2013 Statistical Conference
"Advanced in Latent Variables. Methods, Models and Applications"
University of Brescia,
June, 19 2013

My Agenda

Goal of the presentation

- The application of **Generalized Maximum Entropy Method of Estimation** (GME) to **crisp-input/fuzzy-output regression model**
- Compare LS and GME approaches when empirical data are corrupted by **multicollinearity** → monte carlo simulation
- Show some **results** due to the features of GME approach in **variable selection** procedure → case study

Introduction

Fuzzy Set Theory (FST) is useful to:

- handle with fuzzy or vague information
- manage a particular source of uncertainty: **fuzziness**

Fuzzy statistics provides several *models*, *methods* and *techniques* for fuzzy data:

- fuzzy regression models
- fuzzy principal component analysis
- fuzzy random variables

Introduction

- GME is an approach for the estimation of parameters from statistical models based on the **information theory**
- It was firstly introduced by Golan in 1996 as **extension of Jaynes's Maximum Entropy approach**
- It estimates statistical **parameters** by re-parametrizing these **as combination of discrete random variables**
- Several works showed the main **advantages** of GME:
  - No distributional errors assumptions are required
  - Robustness for a general class of error distributions
  - Excellent work with small samples and ill-posed design matrices
  - Use of inequality constraints in the parameters estimation procedure

LS based model

[D'Urso, 2003]

$$\hat{\boldsymbol{Y}} = \begin{cases} \boldsymbol{c} = \boldsymbol{X}\boldsymbol{\beta}^c + \boldsymbol{\epsilon} \\ \boldsymbol{l} = (\boldsymbol{X}\boldsymbol{\beta}^c)\beta^l + \mathbf{1}\alpha^l + \boldsymbol{\lambda} \\ \boldsymbol{r} = (\boldsymbol{X}\boldsymbol{\beta}^c)\beta^r + \mathbf{1}\alpha^r + \boldsymbol{\rho} \end{cases}$$

Where:

$\boldsymbol{X}$ is a $n \times (m+1)$ matrix of **crisp independent variables**

$\boldsymbol{Y} = \{\boldsymbol{c}, \boldsymbol{l}, \boldsymbol{r}\}$ is a **LR-fuzzy dependent variable**

$\boldsymbol{c}$, $\boldsymbol{l}$, $\boldsymbol{r}$ are $n \times 1$ vectors of fuzzy number's parameters (centers, left and right spreads)

$\boldsymbol{\beta}^c$, α^l, α^r, β^l and β^r are the **model's parameters**

$\boldsymbol{\epsilon}$, $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are $n \times 1$ vectors of **error terms**

LS based model

Solutions are founded by minimizing the following distance between fuzzy numbers:

$$\Delta(\boldsymbol{\beta}^c, \beta^l, \beta^r, \alpha^l, \alpha^r) = \\ \|\boldsymbol{c} - \boldsymbol{X}\boldsymbol{\beta}^c\|^2\omega_1 + \|\boldsymbol{l} - (\boldsymbol{X}\boldsymbol{\beta}^c)\beta^l - \mathbf{1}\alpha^l\|^2\omega_2 + \\ \|\boldsymbol{r} - (\boldsymbol{X}\boldsymbol{\beta}^c)\beta^r - \mathbf{1}\alpha^r\|^2\omega_3$$

where ω_1, ω_2 and ω_3 are positive weights.

GME based model

The first step is the **re-parametrization of the model**

Some basic examples

$$\beta_j^c = (\boldsymbol{z}_{k\times 1}^c)^T \cdot \boldsymbol{p}_{k\times 1}^c \qquad \forall j = 1 \ldots m$$
$$\epsilon_i = (\boldsymbol{z}_{h\times 1}^\epsilon)^T \cdot \boldsymbol{p}_{h\times 1}^\epsilon \qquad \forall i = 1 \ldots n$$

Where:

- z^c and z^ϵ are symmetric around zero **support vectors** (with $3 \leq k \leq 7$ and $3 \leq h \leq 7$)
- $\boldsymbol{p}_{k,1}^c$ and $\boldsymbol{p}_{k,1}^\epsilon$ are **vector of probabilities**

Note that: the points of $\boldsymbol{z}$ can be chosen through a sensitivity analysis or the three-sigma rule (for the case of $\boldsymbol{z}^\epsilon$).

GME based model

$$\hat{\boldsymbol{Y}} = \begin{cases} \boldsymbol{c}_{n,1} = \boldsymbol{X}_{n,m}(\boldsymbol{Z}^{c}_{m,mk}\boldsymbol{p}^{c}_{mk,1}) + (\boldsymbol{Z}^{\epsilon}_{n,nh}\boldsymbol{p}^{\epsilon}_{nh,1}) \\ \boldsymbol{l}_{n,1} = [\boldsymbol{X}_{n,m}(\boldsymbol{Z}^{c}_{m,mk}\boldsymbol{p}^{c}_{mk,1})]((\boldsymbol{z}^{l}_{k,1})^{T}\boldsymbol{p}^{l}_{k,1}) + \boldsymbol{1}_{n,1}((\boldsymbol{z}^{\alpha^{l}}_{k,1})^{T}\boldsymbol{p}^{\alpha^{l}}_{k,1}) + (\boldsymbol{Z}^{\lambda}_{n,nH}\boldsymbol{p}^{\lambda}_{nh,1}) \\ \boldsymbol{r}_{n,1} = [\boldsymbol{X}_{n,m}(\boldsymbol{Z}^{c}_{m,mk}\boldsymbol{p}^{c}_{mk,1})]((\boldsymbol{z}^{r}_{k,1})^{T}\boldsymbol{p}^{r}_{k,1}) + \boldsymbol{1}_{n,1}((\boldsymbol{z}^{\alpha^{r}}_{k,1})^{T}\boldsymbol{p}^{\alpha^{r}}_{k,1}) + (\boldsymbol{Z}^{\rho}_{n,nh}\boldsymbol{p}^{\rho}_{nh,1}) \end{cases}$$

Where:

$$\boldsymbol{\beta}^{c} = \boldsymbol{Z}^{c}_{m,mk}\boldsymbol{p}^{c}_{mk,1} \equiv (\boldsymbol{I}_{m,m} \otimes \boldsymbol{z}^{c}_{k,1}) \cdot (\boldsymbol{1}_{m,1} \otimes \boldsymbol{p}^{c}_{k,1}) \qquad \beta^{l} = (\boldsymbol{z}^{l}_{k,1})^{T}\boldsymbol{p}^{l}_{k,1}$$

$$\boldsymbol{\epsilon} = \boldsymbol{Z}^{\epsilon}_{n,nh}\boldsymbol{p}^{\epsilon}_{nh,1} \equiv (\boldsymbol{I}_{n,n} \otimes \boldsymbol{z}^{\epsilon}_{h,1}) \cdot (\boldsymbol{1}_{n,1} \otimes \boldsymbol{p}^{\epsilon}_{h,1}) \qquad \alpha^{l} = (\boldsymbol{z}^{\alpha^{l}}_{k,1})^{T}\boldsymbol{p}^{\alpha^{l}}_{k,1}$$

$$\boldsymbol{\lambda} = \boldsymbol{Z}^{\lambda}_{n,nh}\boldsymbol{p}^{\lambda}_{nh,1} \equiv (\boldsymbol{I}_{n,n} \otimes \boldsymbol{z}^{\lambda}_{h,1}) \cdot (\boldsymbol{1}_{n,1} \otimes \boldsymbol{p}^{\lambda}_{h,1}) \qquad \beta^{r} = (\boldsymbol{z}^{r}_{k,1})^{T}\boldsymbol{p}^{r}_{k,1}$$

$$\boldsymbol{\rho} = \boldsymbol{Z}^{\rho}_{n,nh}\boldsymbol{p}^{\rho}_{nh,1} \equiv (\boldsymbol{I}_{n,n} \otimes \boldsymbol{z}^{\rho}_{h,1}) \cdot (\boldsymbol{1}_{n,1} \otimes \boldsymbol{p}^{\rho}_{h,1}) \qquad \alpha^{r} = (\boldsymbol{z}^{\alpha^{r}}_{k,1})^{T}\boldsymbol{p}^{\alpha^{r}}_{k,1}$$

Note that: $\otimes$ is the Kronecker-product while $\boldsymbol{I}$ is an identity matrix

GME based model

The parameters are estimated by recovering the probabilities vectors associated, **maximizing** the following functional:

$$\mathcal{H}(\boldsymbol{p}^c, \boldsymbol{p}^l, \boldsymbol{p}^r, \boldsymbol{p}^{\alpha^l}, \boldsymbol{p}^{\alpha^r}, \boldsymbol{p}^{\epsilon}, \boldsymbol{p}^{\lambda}, \boldsymbol{p}^{\rho}) =$$
$$-(\boldsymbol{p}^c_{mk,1})^T \log(\boldsymbol{p}^c_{mk,1}) - (\boldsymbol{p}^l_{k,1})^T \log(\boldsymbol{p}^l_{k,1}) - (\boldsymbol{p}^r_{k,1})^T \log(\boldsymbol{p}^r_{k,1})$$
$$-(\boldsymbol{p}^{\alpha^l}_{k,1})^T \log(\boldsymbol{p}^{\alpha^l}_{k,1}) - (\boldsymbol{p}^{\alpha^r}_{k,1})^T \log(\boldsymbol{p}^{\alpha^r}_{k,1}) - (\boldsymbol{p}^{\epsilon}_{nh,1})^T \log(\boldsymbol{p}^{\epsilon}_{nh,1})$$
$$-(\boldsymbol{p}^{\lambda}_{nh,1})^T \log(\boldsymbol{p}^{\lambda}_{nh,1}) - (\boldsymbol{p}^{\rho}_{nh,1})^T \log(\boldsymbol{p}^{\rho}_{nh,1})$$

GME based model

Subject to the following **normalization constraints**:

(i) $(\boldsymbol{I}_{m,m} \otimes \mathbf{1}_{k,1})^T(\boldsymbol{p}^c_{mk,1}) = \mathbf{1}_{m,1}$
(ii) $(\boldsymbol{p}^l_{k,1})^T \cdot \mathbf{1}_{k,1} = 1$
(iii) $(\boldsymbol{p}^r_{k,1})^T \cdot \mathbf{1}_{k,1} = 1$
(iv) $(\boldsymbol{p}^{\alpha^l}_{k,1})^T \cdot \mathbf{1}_{k,1} = 1$
(v) $(\boldsymbol{p}^{\alpha^r}_{k,1})^T \cdot \mathbf{1}_{k,1} = 1$
(vi) $(\boldsymbol{I}_{n,n} \otimes \mathbf{1}_{h,1})^T(\boldsymbol{p}^\epsilon_{nh,1}) = \mathbf{1}_{n,1}$
(vii) $(\boldsymbol{I}_{n,n} \otimes \mathbf{1}_{h,1})^T(\boldsymbol{p}^\lambda_{nh,1}) = \mathbf{1}_{n,1}$
(viii) $(\boldsymbol{I}_{n,n} \otimes \mathbf{1}_{h,1})^T(\boldsymbol{p}^\rho_{nh,1}) = \mathbf{1}_{n,1}$

and the three **consistency constraints** represented by the equations of the regression model for the centers and left/right spreads.

Note that: the problem is solved by **NLP optimization techniques**.

Experimental scenario

In order to compare GME and LS approaches we set up two experiments:

- GC (**general case**) $\rightarrow$ without multicollinearity in the design matrix
- IC (**ill-posed case**) $\rightarrow$ by increasing the multicollinearity in the design matrix

The results were evaluated by considering:

- the *mean values* of the regression coefficients (as average of the Monte Carlo replications) and their *standard deviations*
- the *relative bias*: $\mathrm{RB} = (E(\hat{\theta}) - \theta)/\theta$
- the *RMSE* on the predicted values
- the *relative efficiency*: $\mathrm{RE}(\hat{\theta}_{gme}, \hat{\theta}_{LS}) = \mathrm{MSE}(\hat{\theta}_{LS})/\mathrm{MSE}(\hat{\theta}_{gme})$

Experimental scenario

- $\boldsymbol{X}_{n,m+1} \sim U(1, 20)$
- $n = 25$, $n = 50$ and $m = 3$
- $\boldsymbol{\beta}^c = [-1.5, -0.2, 1.6, 2.3]$, $\beta^l = 1.7$, $\beta^r = 0.9$, $\alpha^l = -4.2$, $\alpha^r = -2.1$
- $\boldsymbol{\epsilon}_{n,1} \sim N(0, 1)$, $\boldsymbol{\lambda}_{n,1} \sim N(0, 1)$ and $\boldsymbol{\rho}_{n,1} \sim N(0, 1)$
- $\boldsymbol{z} = [-100, -50, 0, 50, 100]$ and $\boldsymbol{v} = [-3\sigma, 0, 3\sigma]$
- second variable corrupted with the third variable: $\boldsymbol{x}_m^{new} = \zeta \cdot \boldsymbol{x}_{m+1} + (1 - \zeta) \cdot \boldsymbol{x}_m$
- three level of multicollinearity: $\zeta = 0.9$, $\zeta = 0.93$ and $\zeta = 0.95$
- 1000 simulations

Some Results

LS for n=25

	$\zeta = 0$	$\zeta = 0.9$	$\zeta = 0.93$	$\zeta = 0.95$
β_1^c	-0.200	-0.200	-	-
SE(β_1^c)	0.043	0.043	-	-
β_2^c	1.601	1.596	-	-
SE(β_2^c)	0.042	0.429	-	-
β_3^c	2.301	2.303	-	-
SE(β_3^c)	0.041	0.388	-	-
β^l	1.699	1.700	-	-
SE(β^l)	0.014	0.010	-	-
β^r	0.899	0.900	-	-
SE(β^r)	0.014	0.010	-	-
RMSE($\boldsymbol{c}$)	0.920	0.920	-	-
RMSE($\boldsymbol{l}$)	1.811	1.811	-	-
RMSE($\boldsymbol{r}$)	1.246	1.246	-	-

Some Results

GME for n=25

	$\zeta = 0$	$\zeta = 0.9$	$\zeta = 0.93$	$\zeta = 0.95$
β_1^c	-0.199	-0.199	-0.199	-0.199
SE(β_1^c)	0.044	0.044	0.043	0.043
β_2^c	1.599	1.589	1.949	1.949
SE(β_2^c)	0.041	0.420	0.019	0.019
β_3^c	2.299	2.309	1.949	1.949
SE(β_3^c)	0.041	0.379	0.019	0.019
β^l	1.700	1.700	1.700	1.700
SE(β^l)	0.014	0.010	0.010	0.010
β^r	0.900	0.900	0.900	0.900
SE(β^r)	0.014	0.010	0.010	0.010
RMSE($\boldsymbol{c}$)	0.914	0.914	0.932	0.932
RMSE($\boldsymbol{l}$)	1.799	1.799	1.834	1.834
RMSE($\boldsymbol{r}$)	1.243	1.243	1.263	1.263

Some Results

Relative Bias for n=25

		$\zeta = 0$	$\zeta = 0.9$	$\zeta = 0.93$	$\zeta = 0.95$
	β_1^c	0.002	0.001	-	-
	β_2^c	0.000	-0.002	-	-
LS	β_3^c	0.000	0.001	-	-
	β^l	-0.001	0.000	-	-
	β^r	-0.001	0.000	-	-
	β_1^c	-0.003	-0.003	-0.004	-0.004
	β_2^c	-0.001	-0.007	0.218	0.218
GME	β_3^c	0.000	0.004	-0.152	-0.152
	β^l	0.000	0.000	0.000	0.000
	β^r	0.000	0.000	0.000	0.000

Some Results

Relative Efficiency between LS and GME

		$\zeta = 0$	$\zeta = 0.9$	$\zeta = 0.93$	$\zeta = 0.95$
	$\boldsymbol{c}$	1.0134	1.0132	-	-
n=25	$\boldsymbol{l}$	1.0125	1.0124	-	-
	$\boldsymbol{r}$	1.0056	1.0055	-	-

Comments

- **GC condition**: GME and LS are overlapped ($\zeta = 0$)
    - slightly differences in terms of RMSEs and RE also for $0.4 \leq \zeta \leq 0.8$
- **IC condition**: GME works better than LS
    - when corruption of $\boldsymbol{X}$ is strong mean values of $\boldsymbol{\beta}$'s are stable and SE($\boldsymbol{\beta}$)'s tend to be low
    - if X_j and X_{j+1} are corrupted variables β_{X_j} and $\beta_{X_{j+1}}$ are equal to $(\beta_{X_j}^{true} + \beta_{X_{j+1}}^{true})/2$
    - RMSEs are stable and small

Purposes

GME approach was applied in order to evaluate its usefulness in:

- **variables selection procedure** through the measurement of the *reduction of uncertainty* with the *normalized entropy index*

By considering $\beta_j = (\boldsymbol{z}_j)^T \boldsymbol{p}_j$ its normalized entropy is:

$$S(\boldsymbol{p}_j) = \frac{-\boldsymbol{p}_j^T \log(\boldsymbol{p}_j)}{\log(k)} \qquad \text{with:} \qquad 0 \leq S(\boldsymbol{p}_j) \leq 1$$

The variable is significant when $S(\boldsymbol{p}_j) < 0.99$

- model's **goodness of fit** evaluation

$$R^2_{pseudo} = 1 - \frac{-\sum_j^m p_j \log(p_j)}{m \cdot \log(k)} \qquad \text{with:} \qquad 0 \leq R^2_{pseudo} \leq 1$$

Data

Dataset used [Coppi et al., 2006]

- Fuzzy dependent variable:
  - $Y_{c,l,r}$=*concentration of carbon monoxide*
- Crisp independent variables:
  - X_1=*temperature*
  - X_2=*relative humidity*
  - X_3=*atmospheric pressure*
  - X_4=*rain*
  - X_5=*radiation*
  - X_6=*wind speed*

Results

GME regression parameters

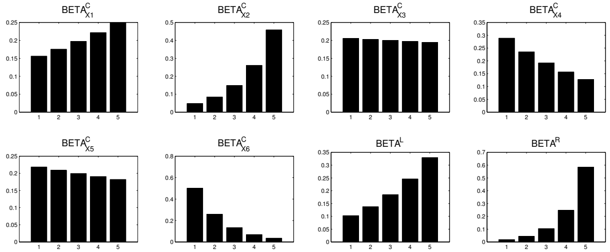
	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$
$\boldsymbol{\beta}^c$	0.248	0.520	-0.252	-0.407	-0.119	-0.650
β^l	0.630					
β^r	0.918					

Normalized Entropy and Goodness of fit indices

	$\boldsymbol{X}_1$	$\boldsymbol{X}_2$	$\boldsymbol{X}_3$	$\boldsymbol{X}_4$	$\boldsymbol{X}_5$	$\boldsymbol{X}_6$
$Sp(\boldsymbol{p}^c)$	0.991	0.836	0.999	0.974	0.998	0.788
$Sp(\boldsymbol{p}^l)$	0.950					
$Sp(\boldsymbol{p}^r)$	0.688					
R^2_{pseudo}	0.605					

Results

Probability distributions for the z vectors



Comments

- Significant variables: X_2 (*relative humidity*), X_4 (*rain*) and X_6 (*wind speed*)
- Their probabilities distribution are far away from uniformity (due to the reduction of uncertainty)
- The global fit of the model is good and it explains about 60% of the overall information stored in the empirical data

Conclusion and Remarks

- GME works with ill-posed problems better than LS
- By using GME researchers can introduce researcher's knowledge in the estimation procedure
- GME provides a useful variables selection procedure without assuming an inferential framework

Conclusion and Remarks

- GME works with ill-posed problems better than LS
- By using GME researchers can introduce researcher's knowledge in the estimation procedure
- GME provides a useful variables selection procedure without assuming an inferential framework

Thank you for attention!

Coppi, R., D'Urso, P., Giordani, P., and Santoro, A. (2006).
Least squares estimation of a linear regression model with lr fuzzy response.
Computational Statistics & Data Analysis, 51(1):267–286.

D'Urso, P. (2003).
Linear regression analysis for fuzzy/crisp input and fuzzy/crisp output data.
Computational Statistics & Data Analysis, 42(1):47–72.