SUPPLEMENT TO KYBERNETIKA VOLUME *28* (1992), PAGES 50 - 53

PROBABILISTIC AND FUZZY PANEL MODELLING

SÁNDOR JÓZSEF

In panel data based econometric modelling the main problem is how to formalize the (eventual) homogeneity existing in the data. In the traditional approach there are prior restrictions (suppositions) on the heterogeneity. The purpose of the article is to give a probabilistic and fuzzy based foundation of the problem and to propose an alternative modelling algorithm which is not based on prior restrictions on the heterogeneity.

The main problem in panel data based econometrical modelling is how to formalize the (eventual) heterogeneity existing in the data. In the traditional approach there are prior restrictions (suppositions) on the heterogeneity. These prejudices are not always justified. This is for example the case when we have a huge data basis. In general in this case we do not have any information about the "true" panels but only suppositions and we use the seemingly appropriate method (based on these prejudices) to handle the data base (and estimate the model). These new methods accept that we do not have any information about the real heterogeneity.

On probabilistic basis this problem can be treated as a parameter estimation problem of a mixture of (in general: normal) distributions. Each panel can be represented with the function of its conditioned expected value and with the distribution of its error term. So our probabilistic model is based on such a mixture density function as the following linear equation system (K-panel model):

$$\xi_k = \mathbf{x}^T \mathbf{a}_k + \varepsilon_k \quad k = 1, \ldots, K$$

where the kth equation is fulfilled with probability p_k and $\sum_{k=1}^{K} p_k = 1$. But we can only observe the mixed random variable ξ ($= \xi_k = \mathbf{x}^T \mathbf{a}_k + \varepsilon_k$ with probability p_k) at given points of $\mathbb{R}^M$. Assuming the existence of the density functions f_k of ξ_k $(k = 1, \ldots, K)$ we can say that the panels in this model are represented by the density functions f_k. In case of normally distributed error terms, i.e. when $\varepsilon_k \sim \mathrm{N}(0, \sigma_k^2)$ and the kth density function at point $\mathbf{x}$ is denoted by $f_{k,\mathbf{x}}(y)$, the density function of ξ is

$$f_{\mathbf{x}}(y) = \sum_{k=1}^{K} p_k f_{k,\mathbf{x}}(y).$$

The "information" for the panels in this probabilistic model are in the conditional probabilities defined by

$$p_{k,\mathbf{x}}(y) = P(\xi_k = y | \mathbf{x}, \xi = y) = \frac{p_k f_{k,\mathbf{x}}(y)}{f_{\mathbf{x}}(y)}.$$

Let's suppose that we have a sample y_i at the point $\mathbf{x}_i$ $(i = 1, \ldots, N)$ for the mixed random variable ξ and we know also the conditioned probabilities $p_{k,\mathbf{x}}$ for this sample. Let $f_i = f_{\mathbf{x}_i}$, $f_{ki} = f_{k,\mathbf{x}_i}$, $p_{ki} = p_{k,\mathbf{x}_i}$, $P_k = \operatorname{diag}(p_{k1}, \ldots, p_{kN})$ and

$$\mathbf{y} = \begin{pmatrix} y_1 \\ \vdots \\ y_N \end{pmatrix}; \quad X = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix} = \begin{pmatrix} x_{11} & \ldots & x_{1M} \\ \vdots & \ddots & \vdots \\ x_{N1} & \ldots & x_{NM} \end{pmatrix}.$$

In this model we have $K(M+2)$ parameters to identify: p_k, $a_{k1}, \ldots, a_{kM}$, σ_k^2; $k = 1, \ldots, K$. For identifying the line parameter vectors we suppose "to have enough information", i.e. the existence of the inverse of the product matrix $X^T P_k X$, $k = 1, \ldots, K$. Based on the maximum likelihood equations we can define an iterative process to identify the model parameters [3], since the solution of the ML estimation p_k^*, $\mathbf{a}_k^*$, σ_k^* $(k = 1, \ldots, K)$ has the following properties:

$$p_k^* = \frac{1}{N} \sum_{i=1}^{N} p_{ki}, \quad \mathbf{a}_k^* = \left(X^T P_k X\right)^{-1} X^T P_k \mathbf{y},$$

$$(\sigma_k^*)^2 = \frac{1}{\sum_{i=1}^{N} p_{ki}} (\mathbf{y} - X\mathbf{a}_k)^T P_k (\mathbf{y} - X\mathbf{a}_k).$$

Our proposition for the parameter estimation is the following: first choose an initial K-panel system, then

- compute the actual sample partition for the panel system,
- estimate the model parameters (find a new, "better" panel system),
- and iterate these steps until a specified convergence criterion is satisfied.

This algorithm can be considered as an iterative weighted maximum likelihood method. The weights for each observations are the respective probabilities that the observation is an element of the given panel.

In the fuzzy approach to panel modelling we represent the K-panel model structure by a fuzzy set system $\mathcal{F} = \{ F_1, \ldots, F_K \}$ defined on the sample space $X \times Y \subset \mathbb{R}^M \times \mathbb{R}$. The membership functions give us "soft informations" about the panels and describe, how the observations determine the panels, furthermore how the data can be characterized by a K-panel structure. Similarly we use K linear functions $f_k(\mathbf{x}) = \mathbf{x}^T \mathbf{a}_k$ and by the fuzzy set F_k we describe the property $y \approx \mathbf{x}^T \mathbf{a}_k$, i.e. "the observation y belongs approximately to the kth line (hyperline)".

Let $(\mathbf{x}_i, y_i)$ $(i = 1, \ldots, N)$ be N observed values (a "sample" at the points $\mathbf{x}_i$) and denote the distance from a value y_i to the kth fuzzy panel

$$d_{ki} = |y_i - \mathbf{x}_i^T \mathbf{a}_k|.$$

The problem can now be formulated as follows:
Find the $0 \leq \mu_{ki} = \mu_{F_k}(\mathbf{x}_i, y_i) \leq 1$ $i = 1, \ldots, N$; $k = 1, \ldots, K$ values of memberships

by minimizing the cost function

$$C = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ki}^{\alpha} d_{ki}^2 \quad \text{subject to the constraint} \quad \sum_{k=1}^{K} \mu_{ki} = 1 \quad \forall i \in \{1, \ldots, N\},$$

where $\alpha \geq 1$ is a fixed parameter. With α the fuzziness of the panels can be determined and for $\alpha = 1$ we would have totally separated nonfuzzy sets.

For the model we will search only "pure fuzzy panels" at the point $y_i \notin \{\mathbf{x}_i^T \mathbf{a}_k : k = 1, \ldots, K\}$ i.e. it will be assumed that all $\mu_{ki} > 0$ if $\prod_{k=1}^{K} d_{ki} > 0$. Let $M_k^{\alpha} = \operatorname{diag}(\mu_{k1}^{\alpha}, \ldots, \mu_{kN}^{\alpha})$ and we assume the existence of the inverse of $X^T M_k^{\alpha} X$ $(k = 1, \ldots, K)$. Using the matrix notation given for the probabilistic case the fuzzy linear panels have the following properties:

$$\mu_{ki_0}^{*} = \begin{cases} 0, & \text{if } k \notin K_0; \\ \in [0, 1] & \text{otherwise} \end{cases} \qquad \text{fulfilling} \qquad \sum_{k=1}^{K} \mu_{ki}^{*} = 1$$

if $d_{ki_0} = 0$ for $k \in K_0 \subset \{1, \ldots, K\}$, $i_0 \in \{1, \ldots, N\}$ fixed, and

$$\mu_{ki}^{*} = \frac{\left(d_{ki}^2\right)^{\frac{-1}{\alpha-1}}}{\sum_{l=1}^{K} \left(d_{li}^2\right)^{\frac{-1}{\alpha-1}}}$$

if $\prod_{k=1}^{K} d_{ki} > 0$ (i fixed), furthermore

$$\mathbf{a}_k^{*} = \left(X^T M_k^{\alpha} X\right)^{-1} X^T M_k^{\alpha} \mathbf{y}.$$

Like the probabilistic case the proposition for the estimation of the fuzzy line parameters is: first choose an initial K-panel system, then
- compute the actual fuzzy partition for the line system,
- estimate the model parameters (find new, "better" lines),
- and iterate these steps until a specified convergence criterion is satisfied.

The basic idea of our methods comes from the study [2]. The construction (estimations, algorithm) of our panels is also similar. In this study the ML estimation of parameters of a mixture density function of K random vector variables was analysed. The authors estimated the vectors of the expected values and the covariance matrices of each (homogeneous) density function under Gaussian assumption. The fuzzy clustering problem was solved like this, using fuzzy covariance matrix and the solution of their problem gives special "ellipsoid-type" fuzzy clusters for every fuzzy part like the homogeneous, normally distributed parts of a mixture density function. Our algorithms are also similar to the algorithms proposed by M. J. Hartley (in a comment to the paper [1]) for the switching regression model based on the so called EM-algorithms (see [1], for example), but in addition we can estimate the probabilities of the panels.

We note that if we do not know the number of panels (K) we can fix different K's and from these K–panel models we can choose one. For example the choice can be based on the cumulative sums

$$\sum_{k=1}^{K} p_k \sigma_k^2 \quad \text{or} \quad C = \sum_{i=1}^{N} \sum_{k=1}^{K} \mu_{ki}^{\alpha} d_{ki}^2.$$

In the fuzzy case the rule of the p_k probabilities is taken by the "number (or percentages) of elements" of the fuzzy panels:

$$n_k \stackrel{\text{def}}{=} \sum_{i=1}^{N} \mu_{ki} \qquad \text{i. e.} \qquad n_k(\%) \stackrel{\text{def}}{=} \frac{1}{N} \sum_{i=1}^{N} \mu_{ki}.$$

The estimation was analysed by Monte–Carlo experiments on specific models. The experiments were carried out with three different models to compare the properties of the two algorithms: model with two parallel lines; model with two crossing linear panels and model with three panels in a triangular form. 50 samples (with 26 or 30 observations) were generated for each panel model. The initial parameters were chosen with special changes (or rotations) of the regression line. The value $\alpha = 2$ was fixed. So all estimations were correct and the procedures converged in 10–20 iteration steps. The probabilistic and the fuzzy panels were always reliably estimated in average, and the ranges of the estimations were small enough.

We note that this method can also be applied for general function fitting problem. If the data come from a functional relation and we calculate some linear panels then using the weights p_{ki} or μ_{ki} for each fixed i, the function values can be approximated by these weighted sums of linear panels (functions).

REFERENCES

[1] A. Dempster, N.M. Laird and D.B. Rubin: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B *39* (1977), 1–38.

[2] D.E. Gustafson and W.C. Kessel: Fuzzy clustering with a fuzzy covariance matrix. In: Advances in Fuzzy Set Theory and Applications (M.M. Gupta, R.K. Ragade and R.R. Yager, eds.), North-Holland, Amsterdam 1979, pp. 605–620.

[3] S. József and L. Mátyás: Probabilistic and fuzzy panel modelling. Working Paper, University of Economics, Budapest 1989.

[4] R.E. Quandt and J.B. Ramsey: Estimating mixtures of normal distributions and switching regressions. J. Amer. Statist. Assoc. *73* (1978), 730–738.

Dr. Sándor József, Research Institute for Agricultural Economics, Zsil u. 3–5, H–1093 Budapest. Hungary.