APPLICATION OF LOGISTIC REGRESSION CLUSTERING IN BANK'S PROPENSITY MODELS

Sergej Sirota – Hana Řezanková

Abstract

Propensity models are very demanding in banks and every improvement of these models could bring a lot of additional business value. The classification is solved by them and the logistic regression is a frequent statistical method for the model estimation. An application of cluster analysis could discover a hidden data structure in the data set, which could possibly improve the usability of the bank's propensity models. A special type of cluster analysis is regression clustering, which stands on the model-based approach. It is considered, that each cluster is presented by regression hyperplane (an application of the linear regression is the most frequent in the literature). The objective is to determine the underlying number of clusters in the data set with the simultaneous application of regression functions to each created cluster. The aim of this contribution is to estimate the optimal number of clusters in logistic regression clustering of a data set for a propensity model by using an information-based technique, such as the Bayesian information criterion and the integrated complete-data likelihood criterion. The lowest absolute value of a certain criterion determines the optimal number of clusters. For the model-based clustering application, the *R* program with the *Mclust* package is used to fulfill the stated goal.

Key words: regression clustering, logistic regression, propensity model, Bayesian information criterion, integrated complete-data likelihood criterion

JEL Code: C25, C38, D12

Introduction

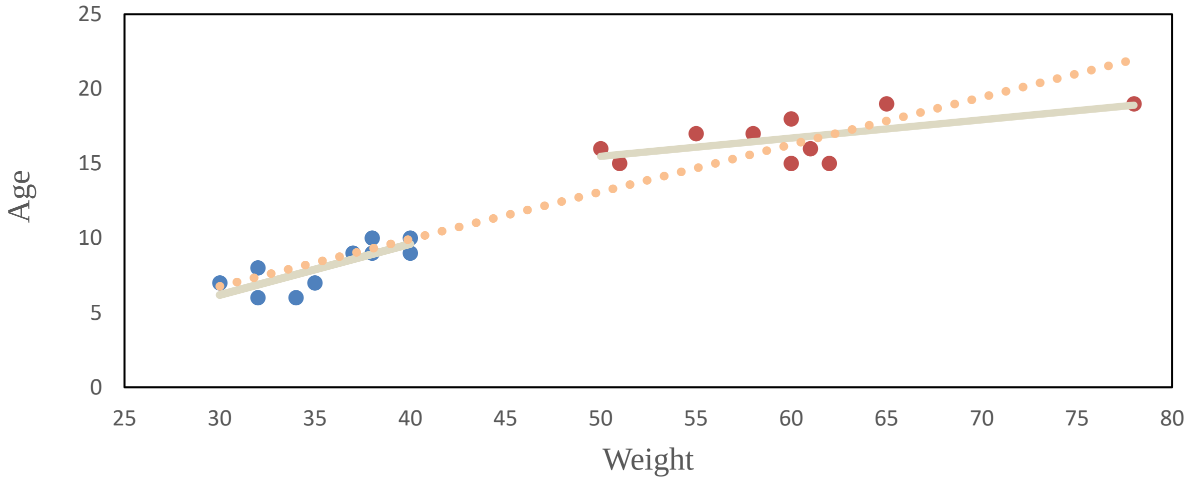
Logistic regression is one of the most used statistical methods in banking, especially for estimating propensity models, which are developed for a client's segmentation. The goal is to estimate single regression hyperplane. As Sirota and Řezanková (2018) mention in, it is possible to improve the usability of propensity models by discovering a hidden data structure in data set by cluster analysis. Combination of regression and clustering is considered e.g. for performing cluster-wise linear regression, where regression functions and objects belonging to clusters are estimated simultaneously (DeSarbo et al., 1988). The selected approach is also appropriate when the responses of the explanatory variable on each observation are not independent of each other (Jayatillake et al., 2011). This approach is also referred to as regression clustering, see e.g. (Lou et al., 1993, Shao and Wu, 2005, Qian and Wu, 2011, Zhang, 2003) for linear regression consideration and (Li et al., 2016) for consideration of logistic regression. Basically, all selected articles are based on the same idea. It is assumed, that the examined population is composed of an unknown but fixed number of sub-populations or clusters (components), which are characterized by class-specific regression hyperplanes (density function) and the objective is to determine the underlying number of clusters in the data set with simultaneous application of regression functions (the number of clusters is greater than 1) to each created cluster.

The aim of this contribution is to estimate the optimal number of clusters in logistic regression clustering of a data set for the propensity model by using the information-based technique. For the analysis performance and calculations, we suppose a modeling data set for a propensity model for consumer loan. We use the statistical program *R* (R Core Team, 2018) for calculations, where the *MClust* package (Scrucca et al., 2016) is applied, which is one of the most popular *R* packages for model-based hierarchical agglomerative clustering with the random-partitioning approach. It stands on finite Gaussian mixture modeling and uses a penalized maximum likelihood approach. It is assumed, that a propensity model estimated by logistic regression can be estimated by a mixture model for computational reasons of the *MClust* package. In the text below, some notations are defined about Gaussian finite mixture modeling and the model-selection based criteria for estimating the number of clusters. The modeling base is also described. This theoretical part is followed by the results of the analysis.

1 Model-based clustering

Model-based methods could be applied in hierarchical or partitioning type clustering. In general, there are two likelihood methods: the random-partitioning approach (the mixture likelihood) and the fixed-partitioning approach (the classification likelihood). Regression clustering is one class of model-based clustering. Fig. 1 shows the basic idea of the regression clustering when the population has been split into two subsets (the objects are displayed as the blue and orange points) based on optimization of the determination coefficient for each regression hyperplane. Regressions on these two subsets (grey lines) provide better prediction than the single regression (on the whole population – the orange line). As an example, we can imagine a population of people with the age between five and nineteen, when the dependency is investigated between the weight (the *x*-axis) and the age (the *y*-axis). More about the model-based clustering could be found in (Banfield and Raftery, 1993).

Fig. 1: The regression clustering principle



Source: own construction

1.1 Gaussian mixture modeling

The *R* package *MClust* use Gaussian finite mixture modeling for model-based clustering and also for classification and density estimation. It is considered the *n* × *p*-dimensional data set $\mathbf{x} = \{x_1, x_2, \ldots, x_i, \ldots, x_n\}$ as a sample of independent and identically distributed observations, which are specified by a probability density functions of finite mixture model with *G* components (clusters):

$$f(x_i; \boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(x_i; \boldsymbol{\theta}_k), \quad (1)$$

where $\boldsymbol{\Psi} = \{\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G\}$ are parameters of the Gaussian mixture model, $\pi_1, \ldots, \pi_{G-1}$ are the weights of probabilities with $\pi_k > 0$ and $\sum_{k=1}^{G} \pi_k = 1$, where *G* means the number of mixture components (which is fixed), $f_k(x_i; \boldsymbol{\theta}_k)$ is the *k*-th component density for observation x_i with selected parameter vector $\boldsymbol{\theta}_k$. The mixture model parameters $\boldsymbol{\Psi}$ are unknown. The maximum likelihood estimator (MLE) for the log-likelihood function, which corresponds to (1), is computationally demanding. The solution for estimating parameters of the mixture model is the application of the expectation-maximization (EM) method, see (Dempster et al., 1977). In the model-based clustering, each component of mixture density is mostly associated with a cluster.

If we considered the Gaussian mixture model, it means that each object and cluster in a data set has Gaussian distribution $f_k(x_i;\boldsymbol{\theta}_k) \sim N(\boldsymbol{\mu}_k, \sum_k)$ and clusters are ellipsoidal, centered at the mean vector $\boldsymbol{\mu}_k$ and has volume, shape and orientation determined by the covariance matrix $\sum_k$:

$$\sum\nolimits_k = \lambda_k \mathbf{D}_k \mathbf{A}_k \mathbf{D}_k^{\mathrm{T}}, \tag{2}$$

where λ_k is a scalar controlling the volume of the ellipsoid, $\mathbf{D}_k$ is an orthogonal matrix determines the orientation of the corresponding ellipsoid and $\mathbf{A}_k$ is a diagonal matrix, which specifies the shapes of the clusters, where the silhouettes of clusters are determined by $\det(\boldsymbol{A}_k) = 1$ (Banfield and Raftery, 1993).

It should be noted, that in the model-based approach to clustering, each component of a finite mixture model is related to the cluster.

1.2 Model-selection based criteria

Select how many components (clusters) should be included for Gaussian mixture modeling and also for the model-based clustering is important. To determine the number of components we can use information criteria. There are two criteria included in the *MClust* package – BIC (the Bayesian information criterion) and ICL (the integrated complete-data likelihood criterion). The lowest absolute value of criterion determines the optimal number of components.

BIC is defined as

$$\mathrm{BIC}_{M,G} = 2\ell_{M,G}(\mathbf{x}|\widehat{\boldsymbol{\Psi}}) - \nu\log(n), \tag{3}$$

where $\ell_{M,G}(\mathbf{x}|\widehat{\boldsymbol{\Psi}})$ presents the log-likelihood estimator for the vector of parameters $\boldsymbol{\Psi}$ calculated by EM method for model M with G components, ν is the number of estimated parameters and n is the selection sample. We choose M and G, which maximizes the BIC criterion.

The second one is the ICL criterion in the form

$$\mathrm{ICL}_{M,G} = BIC_{M,G} + 2\sum_{i=1}^{n}\sum_{k=1}^{G} c_{ik}\log(z_{ik}), \tag{4}$$

where z_{ik} stands for the conditional probability that x_i is from the k-th component. The $c_{ik} = 1$ if the i-th object in a data set belongs to cluster k, otherwise $c_{ik} = 0$ (Scrucca et al., 2016).

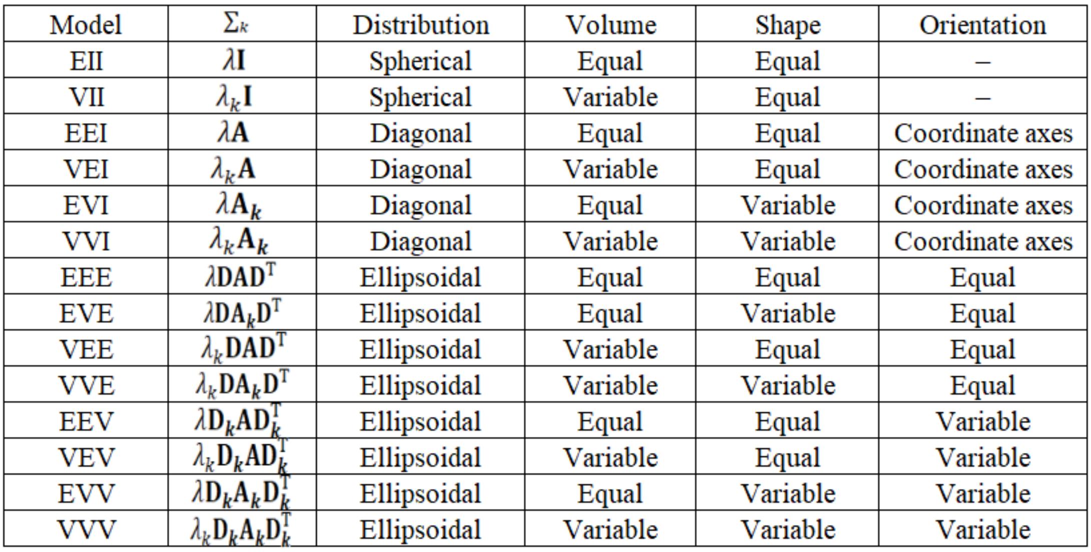
1.3 Modeling base

It is considered the modeling base (where there are 154 113 objects and 2 312 variables), which is used for real developing of the propensity model for consumer loan by logistic regression. Due to calculations reasons, in the first step, we apply a method of the significant variables selection (it is based on a comparison of the correlation between explanatory variables and the target variable, which means, if the client bought (Y = 1) or did not buy (Y = 0) a consumer loan). It chooses the best 100 quantitative variables. In the second step, we apply factor analysis and choose 12 created factors as the final data set (154 113 × 12) for the next calculations – see (Sirota and Řezanková, 2018).

2 Results

We use the *MClust* package in the *R* program to determine the optimal number of clusters in the final data set by application of the model-based clustering. There are implemented 14 Gaussian finite mixture models with different geometric characteristics, where Tab. 1 shows parameters of the within-group covariance matrix $\sum_k$ for these models – see (Scrucca et al., 2016).

Tab.1: Parameters of covariance matrix $\sum_k$

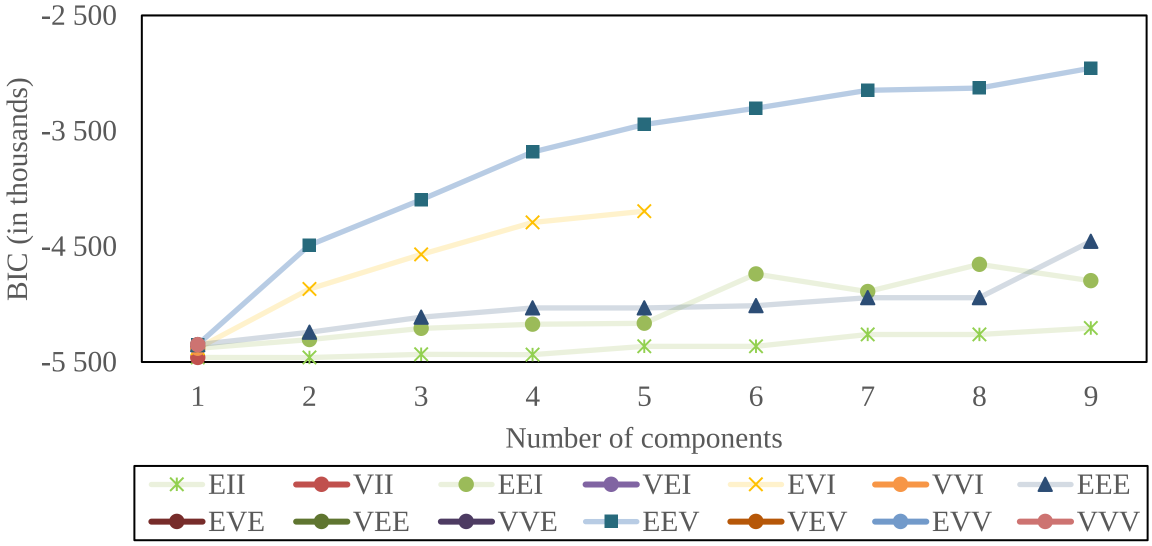


Model	Σ_k	Distribution	Volume	Shape	Orientation
EII	$\lambda\mathbf{I}$	Spherical	Equal	Equal	–
VII	$\lambda_k\mathbf{I}$	Spherical	Variable	Equal	–
EEI	$\lambda\mathbf{A}$	Diagonal	Equal	Equal	Coordinate axes
VEI	$\lambda_k\mathbf{A}$	Diagonal	Variable	Equal	Coordinate axes
EVI	$\lambda\mathbf{A}_k$	Diagonal	Equal	Variable	Coordinate axes
VVI	$\lambda_k\mathbf{A}_k$	Diagonal	Variable	Variable	Coordinate axes
EEE	$\lambda\mathbf{DAD}^T$	Ellipsoidal	Equal	Equal	Equal
EVE	$\lambda\mathbf{DA}_k\mathbf{D}^T$	Ellipsoidal	Equal	Variable	Equal
VEE	$\lambda_k\mathbf{DAD}^T$	Ellipsoidal	Variable	Equal	Equal
VVE	$\lambda_k\mathbf{DA}_k\mathbf{D}^T$	Ellipsoidal	Variable	Variable	Equal
EEV	$\lambda\mathbf{D}_k\mathbf{AD}_k^T$	Ellipsoidal	Equal	Equal	Variable
VEV	$\lambda_k\mathbf{D}_k\mathbf{AD}_k^T$	Ellipsoidal	Variable	Equal	Variable
EVV	$\lambda\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Ellipsoidal	Equal	Variable	Variable
VVV	$\lambda_k\mathbf{D}_k\mathbf{A}_k\mathbf{D}_k^T$	Ellipsoidal	Variable	Variable	Variable

Source: own construction according to Scrucca et al. (2016)

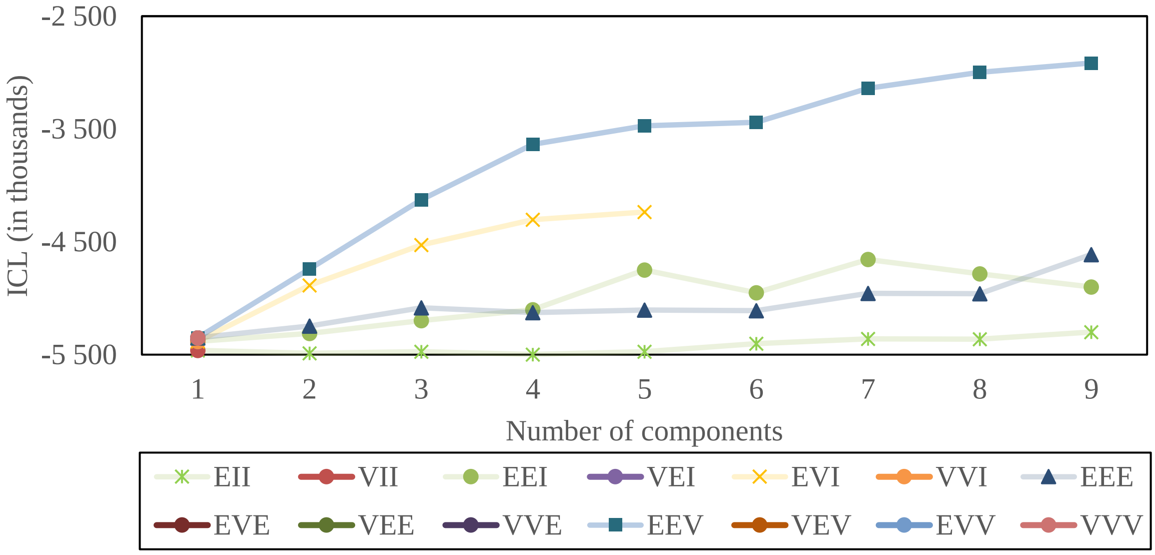
For each model, BIC and ICL criteria are calculated. The lowest value (in absolute) of the criterion determines the optimal number of clusters (the *mclustBIC* and *mclustICL* function). Figs. 2 and 3 show criteria for each model. The best one is the *EEV* model with consideration of 9 clusters in the data set, where clusters are ellipsoidal, centered at the mean vector $\boldsymbol{\mu}_k$ and have the equal volume and shape and variable orientation determined by the covariance matrix Σ_k. We could say, that examined clients (objects in the data set) are from the same population with the same behavior within clusters, that differs from other clusters. Behavior means if the client bought or did not buy the product. Nine models gain values of BIC and ICL criteria only for one component (*VII*, *VEI*, *VVI*, *EVE*, *VEE*, *VVE*, *VEV*, *EVV*, *VVV*).

Fig. 2: Values of the BIC criterion according to the numbers of components (clusters)



Source: own construction in the *R*

Fig. 3: Values of the ICL criterion according to the numbers of components (clusters)

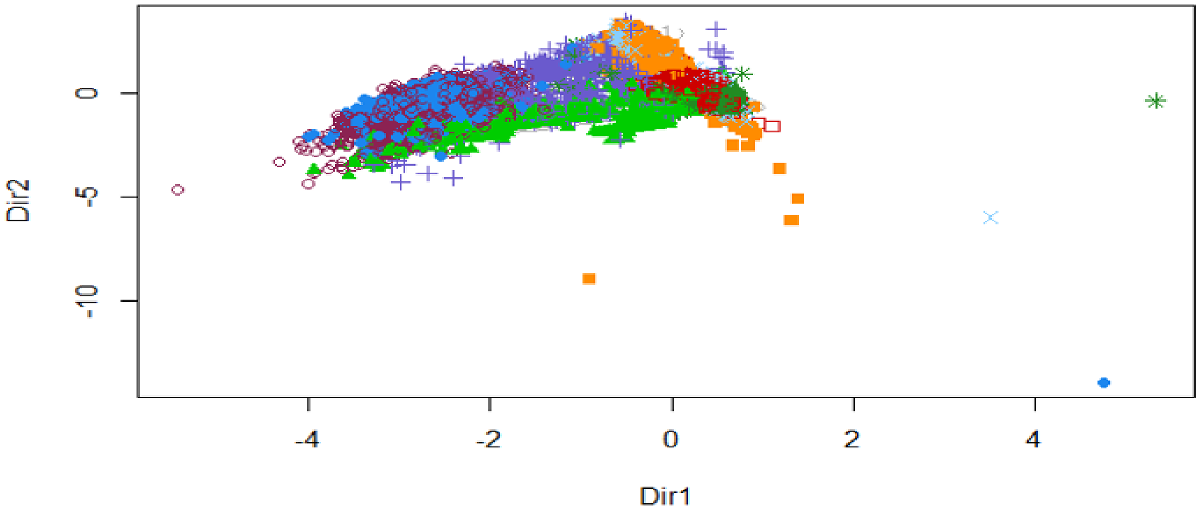


Source: own construction in the *R*

For evaluating a clustering solution, the adjusted Rand index (ARI) is implemented (the *adjustedRandIndex* function). Higher values (which are from 0 to 1) of the index means a higher agreement between partitions of objects into clusters (Hubert and Arabie, 1985). The quality of partitioning is not so high, ARI = 0.183.

The *MclustDR* function transforms the data set from multi-dimensional to two-dimensional subspace (*Dir1* and *Dir2* variables). Fig. 4 shows how objects are partitioned into nine clusters. We can see from the graph that green, orange and purple clusters could be considered as well separated and others 6 clusters are not, which is reflected by the low value of ARI.

Fig. 4: The clustering structure of objects



Source: own construction in the *R*

Tab. 2 shows the sizes of individual clusters and Tab. 3 shows their proportional distribution according to the target variable. Clusters 1 and 2 are the largest (22.66 % and 43.84 %), where there are mostly clients, who did not buy the consumer loan (class = 0). On the other hand, most clients who bought the consumer loan (class = 1) are in clusters 3 and 7 (18.01% and 27.97 %), which means that if we wanted to reach clients most likely to purchase the consumer loan, they would be from these clusters.

Tab.2: The sizes of individual clusters within the classes

class/cluster	1	2	3	4	5	6	7	8	9
0	30 502	66 039	2 016	4 548	1 548	6 300	2 279	2 812	8 682
1	4 419	1 519	5 293	1 684	702	3 917	8 219	1 403	2 231
Total	34 921	67 558	7 309	6 232	2 250	10 217	10 498	4 215	10 913

Source: own construction

Tab.3: The proportions of individual clusters within the classes

class/cluster	1	2	3	4	5	6	7	8	9
0	24.46%	52.95%	1.62%	3.65%	1.24%	5.05%	1.83%	2.25%	6.96%
1	15.04%	5.17%	18.01%	5.73%	2.39%	13.33%	27.97%	4.77%	7.59%
Total	22.66%	43.84%	4.74%	4.04%	1.46%	6.63%	6.81%	2.74%	7.08%

Source: own construction

Conclusion

Logistic regression is the most applicable statistic method in banking. On the other hand, there is cluster analysis, which is not so much used in practice although it discovers a hidden data structure in a data set. This contribution follows the paper (Sirota and Řezanková, 2018) and applies the model-based clustering for estimating the optimal number of clusters in the data set, which is used for the real development of the propensity model for consumer loan. To fulfill the aim, we used the *R* package *MClust*, which is one of the most popular *R* packages for model-based hierarchical agglomerative clustering with the random-partitioning approach. It stands on finite Gaussian mixture modeling and uses a penalized maximum likelihood approach, where parameters are estimated by EM method. For this reason, we considered, that logistic regression could be expressed as a mixture model. First, we mentioned some notations about Gaussian mixture modeling and BIC and ICL criteria for determining the optimal number of components (clusters). We also described the final data set (154 113 × 12), where we used factors from factor analysis as the input. The final data set could be considered as extensive compared to data sets, which are used in other articles dealing with model-based clustering. The theoretical part follows the results of the analysis. For determining the optimal number of clusters by selected criteria, the *MClust* implemented 14 Gaussian finite mixture models, which have different geometric characteristics. The lowest absolute value of BIC and ICL criteria (which indicates the optimal number of clusters) had the *EEV* model containing nine components (clusters). Most of the clients who did not buy the product were in clusters 1 and 2. These clusters were also the largest. On the contrary, most clients who bought the product were in clusters 3 and 7. Unfortunately, the quality of partitioning objects to created clusters measured by the adjusted Rand index was not so high (ARI = 0.183). For further research, it would be appropriate to implement (in the *R* program) fixed-partitioning approach (the classification likelihood) in the model-based clustering and apply the other suitable criteria.

Acknowledgment

This work was supported by the University of Economics, Prague under Grant IGA F4/44/2018.

References

Banfield, J. D., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, 803-821.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B (Methodological), 39*(1), 1-22.

DeSarbo, W. S., & Cron, W. L. (1988). A maximum likelihood methodology for clusterwise linear regression. *Journal of classification*, *5*(2), 249-282.

Hubert, L., & Arabie, P. (1985). Comparing partitions. *Journal of classification, 2*(1), 193-218.

Jayatillake, R. V., Sooriyarachchi, M. R., & Senarathna, D. L. P. (2011). Adjusting for a cluster effect in the logistic regression model: an illustration of theory and its application. *Journal of the National Science Foundation of Sri Lanka*, *39*(3).

Li, J., Weng, J., Shao, C., & Guo, H. (2016). Cluster-based logistic regression model for holiday travel mode choice. *Procedia Engineering*, 137, 729-737.

Lou, S., Jiang, J., & Keng, K. (1993). Clustering objects generated by linear regression models. *Journal of the American Statistical Association*, *88*(424), 1356-1362.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 289.

Shao, Q., & Wu, Y. (2005). A consistent procedure for determining the number of clusters in regression clustering. *Journal of Statistical Planning and Inference*, *135*(2), 461-476.

Sirota, S. & Řezanková, H. (2018). Factor and cluster analysis in context of bank's propensity score matching. In Löster T., Pavelka T. (Eds.), *12th International Days of Statistics and Economics*, Prague: Melandrium, pp. 1625-1634.

Qian, G., & Wu, Y. (2011). Estimation and selection in regression clustering. *European Journal of Pure and Applied Mathematics*, 4(4), 455-466.

Zhang, B. (2003, November). Regression clustering. In *Third IEEE International Conference on Data Mining* (pp. 451-458). IEEE.

Contact

Ing. Sergej Sirota
University of Economics, Prague, Department of Statistics and Probability
Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic
xsirs00@vse.cz

prof. Ing. Hana Řezanková, CSc.
University of Economics, Prague, Department of Statistics and Probability
Sq. W. Churchill 1938/4, 130 67 Prague 3, Czech Republic
hana.rezankova@vse.cz