A MULTIVARIATE MIXTURE MODEL FOR INCOMES OF THE CZECH HOUSEHOLDS IN 2006-2010

Ivana Malá

Abstract

In the contribution a multivariate model of the probability distribution of nominal annual net incomes of the Czech households (in CZK) in 2006-2010 is presented. A mixture of five dimensional normal distributions is fitted into the (five dimensional) vector of logarithms of net annual equalized incomes in analysed five years. The model divides the population of the households into selected number of "artificial" subgroups in which the distribution of incomes is more homogenous than in the whole population. In the text the models with two to five components are fitted (maximum likelihood estimates are numerically found with the use of EM algorithm) and compared. Information criteria are used to compare quality of fits. Estimates of parameters and their estimated standard errors (bootstrap estimates are performed) are given in the tables and the development of estimates of location and variability in time is shown in figures. The R program is used for all computations.

Key words: multivariate normal distribution, income distribution, finite mixture

JEL Code: C33, C38, D31

Introduction

Incomes and wages are in the centre of interest of experts and practitioners from various fields of economy as well as of general public. The analyses from different points of view enable description of their level, variability and also the differences between groups in population, regions or countries. Its development in time is also very useful characteristics of the development of national economies. A characteristic related to incomes should be also included into all indexes quantifying quality of life.

Modelling of the probability distributions of incomes and wages is the frequent goal of the statistical analyses. There exists a large spectrum of statistical (and econometric) methods that are applicable for this problem. In this text a mixture of multivariate normal distributions is used. It is a flexible model for the modelling of multivariate probability distribution. Fully parametric approach to multivariate mixtures is used in this text and for all unknown

parameters (mixing proportions and component parameters) the maximum likelihood estimates were evaluated with the use of EM algorithm. Bayesian approach to the problem is given for example in (Dellaportas, Papageorgiou, 2006), the inference in such problems is described in Jiahua Chena, Xianming Tanb, 2009. For a detailed problem description we refer to the classical monograph McLachlan, Peel, 2000.

In the presented text incomes of the Czech households in 2006-2010 are of interest. The possible approach is to analyse income distribution every year separately. But we have data from the survey EU-SILC (CZSO, 2013) and it is possible to reflect the fact that households form a rotating and to use multivariate analysis of data. The random vector of five consecutive annual equalized net incomes (in CZK) was analysed. All incomes are nominal, only households with complete information about analysed years were included in the data. The artificial subgroups were fitted and no explanatory variables were used in order to improve fits.

1 Methods

Let **X** is a five component random vector of annual net year equalized incomes (European Union methodology, 2006-2010) and $\mathbf{Y} = \ln \mathbf{X}$. We will suppose that **Y** has a mixture distribution of *K* (five dimensional) multivariate normal distributions $N_5(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, $j=1, ..., K$, $K = 1 - 8$ with unknown mixing probabilities $\boldsymbol{\pi} = (\pi_1, ..., \pi_K)', 0 < \pi_j < 1,\ j = 1, ..., K, \sum_{j=1}^{K} \pi_j = 1,$ (McLachlan, Peel, 2000).

Random vectors $\mathbf{Y}_j,\ j = 1, ..., K$ are distributed $N_5(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, where $\boldsymbol{\mu}_j = (\mu_{j1}, ..., \mu_{j5})'$ is a vector of expected values and $\boldsymbol{\Sigma}_j$ is a regular covariance matrix (type 5x5) in the *j*-the component. Then a vector of unknown parameters $\boldsymbol{\psi}$ in the mixture model consists of $21K - 1$ parameters and

$$\boldsymbol{\psi} = (\boldsymbol{\theta}_j, \pi_1, ..., \pi_{K-1},\ j = 1, ..., K),$$

where $\boldsymbol{\theta}_j = (\mu_{jl}, \sigma_{jl}^2, l = 1, ..., 5, \sigma_{jlt}, 1 \le l < t \le 5)'$.

We obtain component densities $f_j(\mathbf{y}; \boldsymbol{\theta}_j)$

$$f_j(\mathbf{y}; \boldsymbol{\theta}_j) = \frac{1}{\sqrt{(2\pi)^5 |\boldsymbol{\Sigma}_j|}} \exp\left\{-(\mathbf{y} - \boldsymbol{\mu}_j)' \boldsymbol{\Sigma}_j^{-1} (\mathbf{y} - \boldsymbol{\mu}_j)\right\},\ \mathbf{y} \in R^5.$$

Mixture density is then given in the form

$$f\ (\mathbf{y};\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{(2\pi)^p \left|\boldsymbol{\Sigma}_j\right|}} \exp\left\{-\left(\mathbf{y}-\boldsymbol{\mu}_j\right)' \boldsymbol{\Sigma}_j^{-1} \left(\mathbf{y}-\boldsymbol{\mu}_j\right)\right\},\ \mathbf{y} \in R^p\ . \quad (1)$$

From (1) we obtain

$$E(\mathbf{Y}) = \sum_{j=1}^{K} \pi_j E\left(\mathbf{Y}_j\right) = \sum_{j=1}^{K} \pi_j \boldsymbol{\mu}_j$$

and

$$\boldsymbol{\Sigma}_{\mathbf{Y}} = \sum_{j=1}^{K} \pi_j \boldsymbol{\Sigma}_j + \sum_{j=1}^{K} \pi_j \left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{\mathbf{Y}}\right)\left(\boldsymbol{\mu}_j - \boldsymbol{\mu}_{\mathbf{Y}}\right)'.$$

Moreover the univariate marginal distributions are univariate mixtures of normal distributions, for t =1, …, 5 we obtain

$$f_t(y_t;\boldsymbol{\psi}) = \sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{2\pi}\sigma_{jt}} \mathrm{e}^{-\frac{(y_t-\mu_{jt})^2}{2\sigma_{jt}^2}},\ y_t \in R.$$

All computations were performed in the program R (R CORE TEAM, 2014), the package MIXTOOLS (Benaglia et al., 2009, Young et al., 2014) was used for multivariate estimation. Multivariate approximation with EM algorithm is highly time consuming and its successful use depends on good (or at least acceptable) initial values for unknown parameters. Moreover numeric problems arising from the shape of logarithmic likelihood (McLachlan, Peel, 2000) appeared especially for higher number of components. The same problems were met when bootstrap was used in order to estimate standard errors of estimates of the parameters.

2. Results

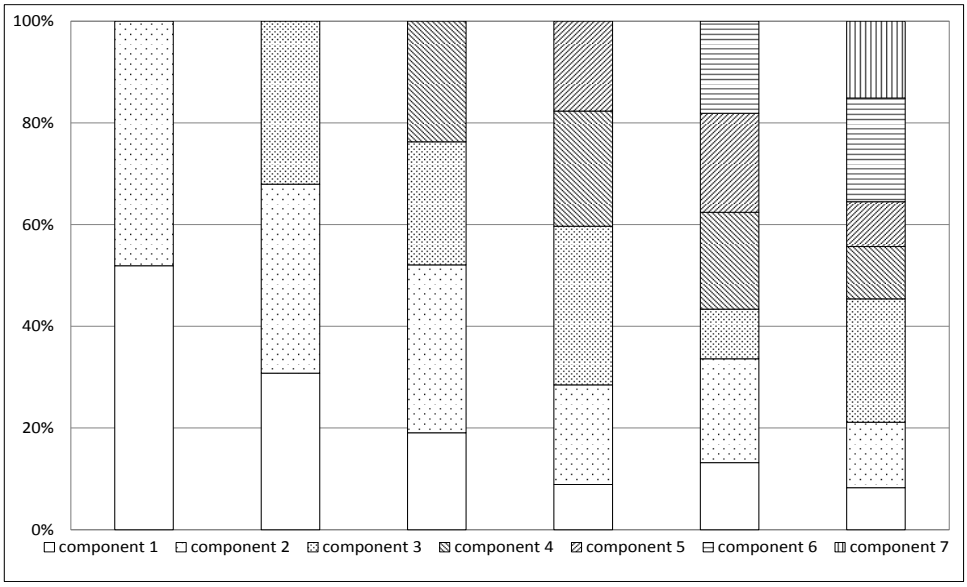
Living Conditions Survey is a part of the EU-SILC (European Union – Statistics on Income and Living Conditions) program that is obligatory for all members of the European Union (CZSO, 2014). The households surveyed in all analysed years were included in the dataset and equalized net annual income was evaluated as a total net annual income of a household divided by number of equalized units according to the European Union methodology (weight 1 to the first adult, other members above 13 0.5 and children under 13 0.3). The equalized incomes reflect possible sharing of spending by members in the household. For this reason

they are supposed to be more informative than frequently used total income or income per capita. Annual exchange rates (CZK/EUR) in analysed years were 28.30, 27.70, 24.90, 26.40 and 25.30. Inflation rate was 2,8 %, 6,3 %, 1,0 % and 1,5 %, overall inflation was 10.6 %.

To the data models (1) with one (a 5 dimensional normal distribution) to 8 components were fitted. The value of Akaike information criterion *AIC* criterion decreases to 7 components (146 parameters in the model) and then it begins to increase. But it was quite impossible to handle such a big model (with 7 components). In this contribution a model with 3 components (62 unknown parameters) is presented. This model represents a compromise between quality, numerical capability and possibility to show easily results.

In the Figure 1 the estimated mixing proportions are shown for the models with up to 7 components. The models tend to find subgroups with approximately equal proportions. Components in this text are ordered according to the estimated expected value in 2006 (from the component with the lowest expected income to the highest expected income).

Fig. 1: Estimated values of mixing proportions (*K* = 2-7)



Source: own calculations

In the Table 1 all estimated parameters are shown for the model with 3 components. It is necessary to estimate three vectors of expected values (with standard deviations evaluated with the use of bootstrap) and three covariance matrices. Moreover estimated correlation matrices are presented in order to show estimated (linear) dependencies of incomes between years in estimated components.

Estimated vectors $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2$ and $\boldsymbol{\mu}_3$ are also shown in the Figure 3 together with estimated values of these parameters of models with $K = 1 - 4$. Estimated variances increase from the first to the third components (in each element). The opposite is true for estimated covariances, incomes in the first components are stronger (positively) correlated than incomes in other components. There are even small negative correlations in the third component.

Tab. 1: Estimated mixture model with three components. (Standard deviations of estimates are shown in brackets)

component 1: 0.372 (0.016)						component 2: 0.320 (0.009)				
$\hat{\boldsymbol{\mu}}$	11.870	11.924	11.987	12.043	12.076	11.899	11.993	12.077	12.098	12.114
	(0.008)	(0.009)	(0.008)	(0.005)	(0.012)	(0.014)	(0.015)	(0.015)	(0.017)	(0.017)
$\hat{\boldsymbol{\Sigma}}_1$						$\hat{\boldsymbol{\Sigma}}_2$				
	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010
2006	0.094					0.106				
2007	0.092	0.094				0.095	0.109			
2008	0.088	0.090	0.091			0.011	0.015	0.097		
2009	0.043	0.044	0.044	0.105		0.010	0.011	0.084	0.093	
2010	0.003	0.003	0.002	0.048	0.125	0.007	0.007	0.081	0.087	0.092
$\hat{\boldsymbol{\rho}}_1$						$\hat{\boldsymbol{\rho}}_2$				
	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010
2007	0.980	1				0.878	1			
2008	0.954	0.972	1			0.113	0.141	1		
2009	0.435	0.441	0.446	1		0.105	0.108	0.885	1	
2010	0.029	0.025	0.018	0.423	1	0.070	0.074	0.855	0.940	1
component 3: 0.308 (0.013)										
$\hat{\boldsymbol{\mu}}$	11.959	12.032	12.135	12.155	12.131					
	(0.008)	(0.008)	(0.007)	(0.007)	(0.010)					
$\hat{\boldsymbol{\Sigma}}_3$						$\hat{\boldsymbol{\rho}}_3$				
	2006	2007	2008	2009	2010	2006	2007	2008	2009	2010
2006	0.286					1				
2007	0.192	0.258				0.708	1			
2008	0.141	0.162	0.254			0.522	0.631	1		
2009	0.062	0.064	0.106	0.281		0.219	0.405	0.396	1	
2010	-0.007	-0.003	0.021	0.130	0.288	-0.024	-0.011	0.079	0.457	1

Source: own calculations

It is difficult to interpret results from the Table 1, but we obtained three distributions of subgroups of the Czech households ordered by incomes level. Moreover detailed information about components is provided by the model. The mixture model is significally better than one or two components models according to bootstrapped likelihood-ratio test. From maximum likelihood estimates in the Table 1 the maximum likelihood estimates of various characteristics of all components and the mixture can be evaluated.

In the Table 2 the estimated characteristics of location and variability based on moments (expected value, standard deviation) and quantiles (median and quartile deviation) are given for the components and for the mixture. Sample characteristics are shown in the Table 3, if we use the exchange rates given above we obtain mean equivalised incomes (in EUR) 5,708, 6,227, 7,511, 7,370 and 7,791. Good correspondence of estimated values (Table 2) with sample values (Table 3) is visible. The same conclusion can be done from the Figure 2 (for expected values).

Tab. 2: Estimated characteristics in the mixture model with three components (CZK).

year	comp 1	comp 2	comp 3	mixture	comp 1	comp 2	comp 3	mixture
	expected value				standard deviation			
2006	149 845	155 181	180 349	160 943	86 269	48 705	60 273	69 413
2007	158 214	170 786	191 295	172 422	85 768	53 543	65 039	71 705
2008	168 254	184 677	211 716	186 891	90 550	57 041	67 753	76 321
2009	179 212	188 133	218 737	194 234	102 034	62 598	68 113	82 678
2010	186 923	191 038	214 420	196 704	107 986	69 649	66 389	85 963
	median				quartile deviation			
2006	156 297	142 969	147 179	147 789	57 637	29 769	32 562	38 141
2007	168 173	150 969	161 696	159 094	58 707	31 401	36 371	40 610
2008	186 432	160 760	175 890	172 634	64 649	32 957	37 319	43 501
2009	190 087	170 046	179 625	178 666	69 393	37 462	37 117	46 137
2010	185 664	175 615	182 491	180 910	68 681	42 242	37 501	47 931

Source: own calculations

Tab. 3: Sample characteristics of location and variability (CZK)

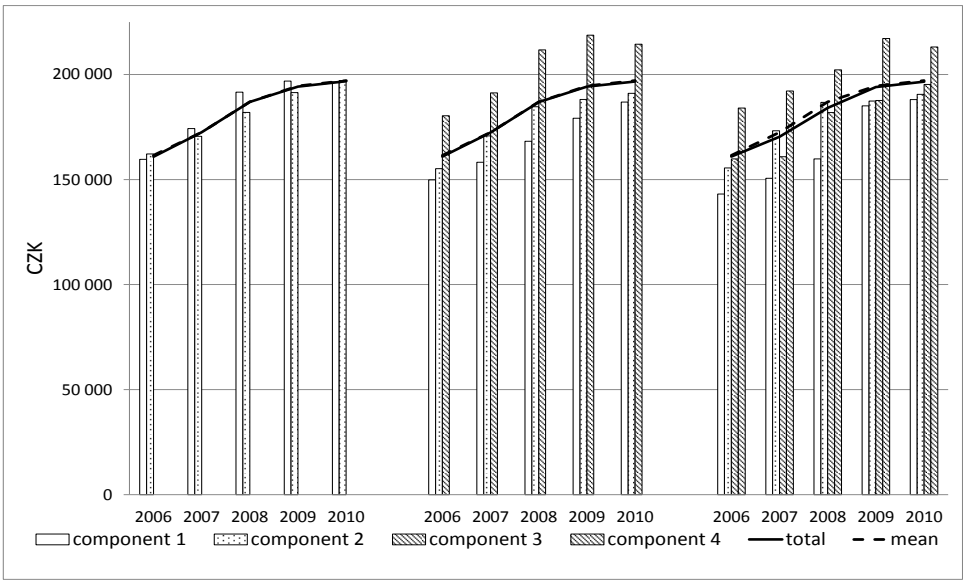
	mean	median	standard deviation	quartile deviation
2006	161 535	143 564	80 665	34 860
2007	172 486	155 724	76 042	38 685
2008	187 028	168 304	82 710	42 200
2009	194 563	173 624	92 643	41 908

2010	197 101	177 551	93 643	43 947

Source: own calculations

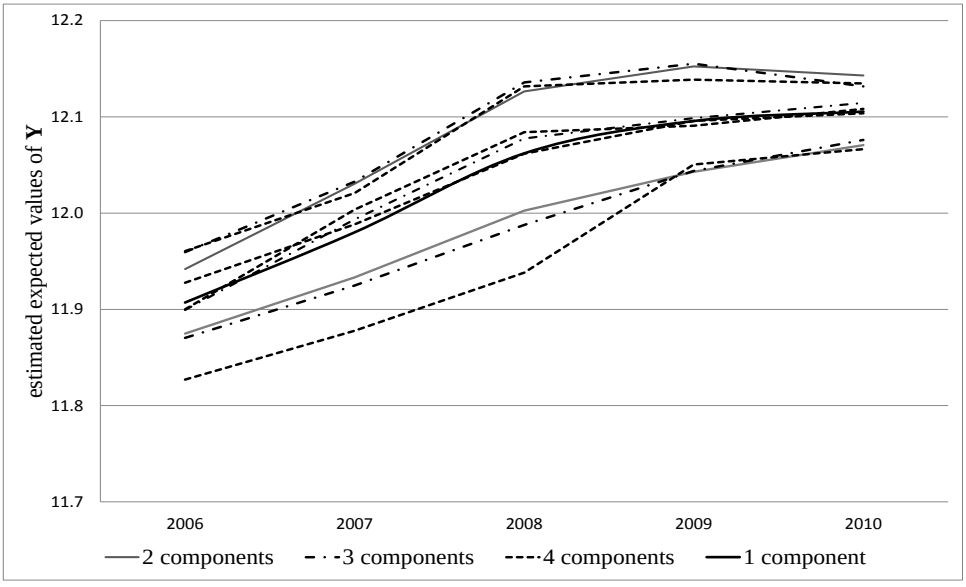
In the Figure 2 estimated expected values of components are shown for the models for $K = 2 – 4$ (values from the Table 2). The lines describes the mean (dashed line) and estimated expected value from the mixture (solid line). These two lines almost coincide.

Fig. 2: Estimated expected values of components for the models with $K = 2 – 4$, mean and total expected value of the estimated mixture



Source: own calculations

Fig. 3: Estimated values of mixing components ($K = 1 - 4$)



Source: own calculations

In the Figure 3 estimated vectors $\hat{\boldsymbol{\mu}}_j$ are given for the models with one to four components. We can see that with increasing number of components the line for the lower income component tends to be smaller and for the component of highest incomes tends to increase.

Conclusions

The multivariate model of equivalised incomes of the Czech households was proposed for five consecutive years 2006-2010. A mixture of five dimensional normal random vectors was fitted into logarithms of data for number of components 1 to 8.

This model enable to describe not only multivariate mixture distribution, univariate marginal mixture distribution, but also dependence of incomes in years (in the covariance matrix). But in comparison with separate models for analysed years we have only one vector of mixing probabilities for all years. From this multivariate model the predictions of incomes in subgroups could be constructed as well as estimates of the overall set of the Czech households.

Acknowledgement

The research was supported by the project IP 400040 from the Faculty of Informatics and Statistics of University of Economics, Prague.

References

Benaglia, T, Chauveau, D., Hunter, D. R. & Young, D. (2009). Mixtools: An R Package for Analyzing Finite Mixture Models. *Journal of Statistical Software*, 32 (6), 1-29.

Bílková, D. (2012). Recent Development of the Wage and Income Distribution in the Czech Republic. *Prague Economic Papers*, 21, 233–250.

Bílková, D. (2012). Development of wage Distribution of the Czech Republic in Recent Years by Highest Education Attaintment and Forecasts for 2011 and 2012. In: *International Days of Statistics and Economics at VŠE, Prague.* Prague, 13.09.2012 – 15.09.2012. Prague : VŠE, 2012, 162–182.

Czech Statistical Office. (2014). URL: http://www.czso.cz/eng/

Dellaportas, P., Ioulia Papageorgiou, I. (2006). Multivariate mixtures of normals with unknown number of components. *Statistics and Computing*, 16 (1), 57-68

Jiahua Chena, Xianming Tanb. (2009). Inference for multivariate normal mixtures. *Journal of Multivariate Analysis*, 100 (7), 1367–1383

Marek, L. (2013). Some Aspects of Average Wage Evolution in the Czech Republic. In: *International Days of Statistics and Economics.* [online] Prague, 19.09.2013 – 21.09.2013. Slaný: Melandrium, 2013, 947–958.
URL: http://msed.vse.cz/files/2013/208-Marek-Lubos-paper.pdf

Marek, L. & Vrabec, M. (2013). Model wage distribution – mixture density functions. *International Journal of Economics and Statistics*, 1, 113-121.

McLachlan, G. J. & Peel, D. (2000). *Finite Mixture Models.* Wiley series in Probability and Mathematical Statistics: Applied Probability and Statistics Section, New York.

R CORE TEAM. (2014). R: a language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. URL: http://www.r-project.org/.

Young, D., Benaglia, T., Chauveau, D., Hunter, D., Elmore, R., Hettmansperger, T., Hoben T. & Fengjuan Xuan (2014). Mixtools: Tools for analyzing finite mixture models. R package version 1.0.1. URL: http://cran.r-project.org/web/packages/mixtools/index.htm

Contact

Ivana Malá

University of Economics, Prague
W. Churchill Sq. 4
130 67 Prague 3
Czech Republic
malai@vse.cz