M A T H E M A T I C A L E C O N O M I C S

No. 10(17) 2014

MODELING INCOME ON THE BASIS OF DISTRIBUTION MIXTURE

Grzegorz Sitek

Abstract. Finite mixtures of probability distributions may be successfully used in the modeling of probability distributions of incomes. These distributions are typically heavy tailed and positively skewed. This article deals with the problem of determining the number of components in mixture modeling. This paper considers the likelihood of ratio-based testing of the null hypothesis of homogeneity in mixture models. The number of components is an important parameter in the applications of finite mixture models.

Keywords: finite mixture, income distribution, maximum likelihood estimate, EM algorithm, number of components.

JEL Classification: C46.

DOI: 10.15611/me.2014.10.06.

1. Introduction

Finite mixtures of probability distributions may be successfully used in the modeling of probability distributions of incomes. These distributions are typically heavy tailed and positively skewed [Kot 1996].

Income distribution, as any other probability distribution, is completely determined by the cumulative distribution $F(x)$ or density functions $f(x)$.

2. Income distribution

The study of income distribution has a long history. The probability modeling of income distribution started with the work of Italian economist Vilfredo Pareto in 1897 and his work, Cours d' economie politique. He described the principle which states that for many events, roughly 80% of the effects come from 20% of the causes. The original observation was in

Grzegorz Sitek
Department of Statistics, University of Economics in Katowice, ul. 1 Maja 50, 40-287 Katowice, Poland.
E-mail: grzegors12@wp.pl

connection with population wealth. Pareto noticed that 80% of Italy's land was owned by 20% of the population. He carried out several surveys on a variety of other countries and found a similar distribution. This is nowadays known as a Pareto law. Since the work of Pareto distribution, a large number of models have been introduced to describe the distribution of incomes.

Lognormal Distribution

Two parameter lognormal distribution is given by the density

$$f(y) = \frac{1}{\sqrt{2\pi}y\sigma} \exp\left\{-\frac{1}{2\sigma^2}(\ln y - \mu)^2\right\} \quad y > 0. \tag{1}$$

If a random variable Y has a lognormal distribution, LN(μ, σ^2) then a variable log(Y) has a normal distribution with expectation μ and variance σ^2. In empirical studies of wage and income distributions they are considered as three-parameter distribution, which in addition μ and σ there is a third parameter τ. The parameter τ is the theoretical minimal value of Y. The central moments of the two-parameter distribution are given by

$$\lambda_r = \exp(r\mu + \frac{1}{2}r^2\sigma^2). \tag{2}$$

The maximum likelihood estimates m and s^2 for the parameter μ and the parameter σ are

$$m = \frac{1}{n}\sum_{i=1}^{n} \ln y_i \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(\ln y_i - m)^2. \tag{3}$$

Gamma Distribution

The gamma (more precisely, the Pearson type III) distribution is certainly among the five most popular distributions in applied statistics when unimodal and positive data are available. Here we shall briefly sketch the basic properties of the gamma distribution and concentrate on aspects more closely related to size and income distributions.

The density of the gamma distribution is

$$f(y) = \frac{\beta^{-\alpha}}{\Gamma(\alpha)} y^{\alpha-1} e^{-y/\beta} \quad \text{for} \quad y > 0,\ \beta,\ \alpha > 0, \tag{4}$$

where α, $\beta > 0$, with α being a shape and β a scale parameter. The likelihood equations for a simple random sample of size n are

$$\begin{cases} \sum_{i=1}^{n} \log X_i - n \log \hat{\beta} - n\Psi(\hat{\alpha}) = 0 \\ \sum_{i=1}^{n} X_i - n\hat{\alpha}\hat{\beta} = 0 \end{cases}. \tag{5}$$

These can be solved iteratively, and indeed procedures for estimation in the gamma distribution are nowadays available in many statistical software packages. The Gini coefficient is given by [McDonald and Jensen 1979]

$$G = \frac{\Gamma(\alpha + 0{,}5)}{\Gamma(\alpha + 1)\sqrt{\pi}}. \tag{6}$$

Pareto Distributions

The classical Pareto distribution is defined in terms of its c.d.f.

$$F(x) = 1 - \left(\frac{x_0}{x}\right)^{\alpha} \qquad x \geq x_0 > 0. \tag{7}$$

The density is

$$f(x) = \alpha x_0^{\alpha} x^{-1-\alpha} \qquad x \geq x_0 > 0, \tag{8}$$

where $\alpha > 0$ is a shape parameter (also measuring the heaviness of the right tail) and x_0 is a scale. The expected value of a random variable following a Pareto distribution is

$$E(X) = \frac{x_0 \alpha}{\alpha - 1} \quad \alpha > 1. \tag{9}$$

The variance of a random variable following a Pareto distribution is

$$D^2(X) = \frac{x_0^2 \alpha}{(\alpha - 2)(\alpha - 1)^2} \quad \alpha > 2. \tag{10}$$

Burr Distributions

The c.d.f.'s of all Burr distributions satisfy the differential equation

$$F'(x) = F(x)[1 - F(x)]g(x), \tag{11}$$

where F is distribution and g is some nonnegative function. The most widely known of the (non-uniform) Burr distributions is the Burr XII distribution, frequently just called the Burr distribution. The density is

$$f(x) = \frac{\alpha\gamma(x/\theta)^{\gamma}}{x[1+(x/\theta)^{\gamma}]^{\alpha+1}} \text{ for } x > 0,\ \alpha > 0,\ \gamma > 0,\ \theta > 0, \tag{12}$$

where α and β are a shape parameter and θ scale parameter.

3. Maximum Likelihood Method

To use the method of maximum likelihood, one first specifies the joint density function for all observations. For an independent and identically distributed sample, this joint density function is

$$L = \prod_{i=1}^{n} f(x_i|\theta). \tag{13}$$

The maximum likelihood estimate (MLE) of θ is the value of θ that maximizes (13): it is the value that makes the observed data the "most probable". Rather than maximizing this product which can be quite tedious, one often uses the fact that the logarithm is an increasing function so it will be equivalent to maximizing the log likelihood. As the sample size increases to infinity, sequences of maximum-likelihood estimators have the following properties [Fisz 1969]: consistency, asymptotic normality, efficiency, it achieves the Cramér-Rao lower bound when the sample size tends to infinity.

4. Fitting income distributions

Fitting distributions to data is a common task in statistics and consists in choosing a probability distribution modeling the random variable, as well as finding parameter estimates for that distribution. The *fitdistr* function estimates distribution parameters by maximizing the likelihood function using the *optim* function. Data on basic salaries are taken from the book "Analiza ekonometryczna kształtowania się płac w Polsce w okresie transformacji". Computations of goodness-of-fit can be judged using Akaike information criterion.

$$\text{AIC} = -2*l(\boldsymbol{\psi}) + 2*k, \tag{14}$$

where $l(\boldsymbol{\psi})$ – log-likelihood function, k – number of parameters.

If different models are compared, the smaller the value of AIC the better the fit. When fitting continuous distributions, two goodness-of-fit statistics: Cramer-von Mises and Kolmogorov-Smirnov are classically considered.

Table 1. Goodness-of-fit statistics

Theoretical distribution	AIC	Statistic Kolmogorov-Smirnov	Statistic Cramer-von Mises
Lognormal	18 812,21	0,10089280	4,06621720
Gamma	19 275,35	0,15255850	10,6059535
Burra B12	18 560.90	0,04670285	0,53944619
Pareto	20 016,83	0,32876380	39,1321924

Source: own calculations.

On the basis of the table above, we conclude that the best quality of the fit to the empirical data is obtained in the case of distribution Burr B12, and the worst for the Pareto distribution.

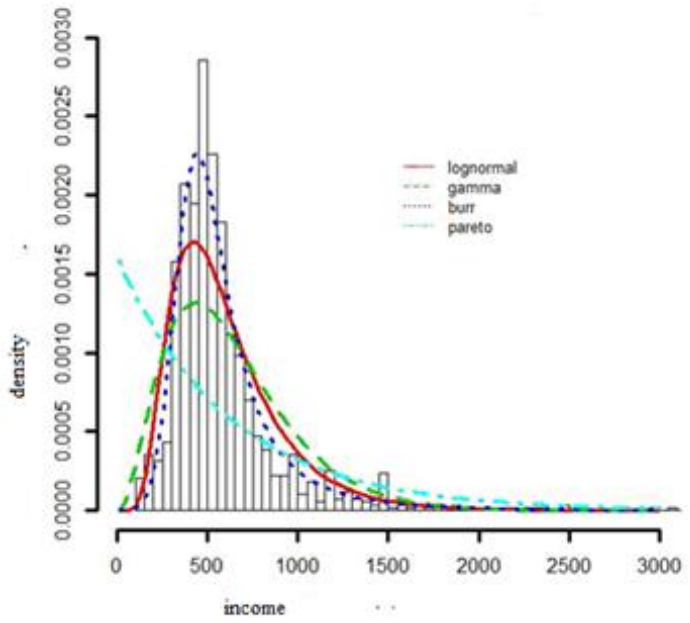

Fig. 1. Histogram and theoretical densities

Source: own elaboration.

5. Finite Mixture Models

Suppose X to be a positive value random variable with continuous distribution. The density function f is given as a weighted average of K component densities $f_j(x|\theta)$ with mixing proportions π_j [Jajuga 1990]:

$$f(x|\theta) = \sum_{j=1}^{K} \pi_j f_j(x|\theta_j), \tag{15}$$

where $\pi_j \geq 0,\ j = 1,2,...,K\ \sum_{j=1}^{K} \pi_j = 1$ and component densities $f_j(x|\theta_j)$ depend on p-dimensional (in general unknown) vector parameters θ_j. For the estimation of unknown parameters (from a random sample x_i, $i = 1,\ldots,n$) the maximum likelihood estimation is usually used in order to obtain the estimate of the parameter. From (13) it follows that the likelihood function is equal to

$$l(x_1, x_2, ..., x_n|\theta, \pi) = \prod_{i=1}^{n} \left[\sum_{j=1}^{K} \pi_j f_j(x|\theta_j) \right]. \tag{16}$$

Now we want to maximize the complete data log likelihood

$$\sum_{i=1}^{n} \ln f(x_i|\theta, \pi) \tag{17}$$

with respect to $\sum_{j=1}^{K} \pi_j = 1$. For this, we actually need to introduce a Lagrange multiplier. This gives us the likelihood function of:

$$L_0 = \sum_{i=1}^{n} \ln f(x_i|\theta, \pi) - \lambda \left[\sum_{j=1}^{K} \pi_j - 1 \right]. \tag{18}$$

First, we have to find the partial first derivatives of L_0 and set them equal to zero.

The task of maximizing the likelihood function can be solved using the EM algorithm. This is a numeric procedure that consists of two steps. The first step is called Expectation (probabilities π_j are estimated) and the second one Maximization, where estimated values from the first step are used in

order to find new approximations of parameters θ. These two steps are repeated until a solution is found. Generally, EM algorithm does not guarantee the absolute maximum of the logarithmic likelihood function but only the local extreme [Titterington, Smith, Makov 1985]. In the model for complete data associated with the model, each random vector $C_i = (X_i; Z_i)$, where $Z_i = (Z_{ij}, j = 1,2,...,K)$ and $Z_{ij} \in \{0,1\}$ is a Bernoulli random variable indicating that individual i comes from component j. Since each individual comes from exactly one component, this implies $\sum_{j=1}^{K} Z_{ij} = 1$ and $P(Z_{ij} = 1) = \pi_j$. The complete-data density for one observation is thus

$$h_\theta(c_i) = h_\theta(x_i, z_i) = \sum_{j=1}^{K} I_{z_{ij}} \pi_j f_j(x_i). \tag{19}$$

Instead of the observed log-likelihood the EM algorithm iteratively maximizes the operator

$$Q(\theta|\theta^{(t)}) = E\left[\log h_\theta(C)|x, \theta^{(t)})\right], \tag{20}$$

where $\theta^{(t)}$ is the current value θ at iteration t, and the expectation is with respect to the distribution $k_\theta(c \mid x)$ of c given x, for the value $\theta^{(t)}$ of the parameter.

E-step: compute $Q(\theta|\theta^{(t)})$.

M-step: set $\theta^{(t+1)} = \arg\max_{\theta \in \psi} Q(\theta|\theta^{(t)})$.

E-step: Calculate the "posterior" probabilities (conditional on the data and $\theta^{(t)}$) of component inclusion,

$$p_{ij}^{(t)} \overset{def}{=} P_{\theta^{(t)}}(Z_{ij} = 1|x_i) = \frac{\pi_j^{(t)} f_j^{(t)}(x_i)}{\sum_{j'=1}^{K} \pi_{j'}^{(t)} f_{j'}^{(t)}(x_i)} \tag{21}$$

for $i = 1,2,...,n$ and $j = 1,2,...,K$.

Numerically, it can be dangerous to implement Equation 21 exactly as written due to the possibility of the indeterminate form 0/0 in cases where x_i is so far from any of the components that all $f_{j'}^{(t)}(x_i)$ values result in a numerical underflow to zero. Thus, many of the routines in mixtools [Benaglia et al. 2009] actually use the equivalent expression

$$p_{ij}^{(t)} = \left[1 + \sum_{j' \neq j}^{K} \frac{\pi_{j'}^{(t)} f_{j'}^{(t)}(x_i)}{\pi_j^{(t)} f_j^{(t)}(x_i)}\right]^{-1}. \tag{22}$$

M-step for π

$$\pi_j^{(t+1)} = \frac{1}{n} \sum_{i=1}^{n} p_{ij}^{(t)} \quad \text{for} \quad j = 1, 2, ..., K. \tag{23}$$

6. Modeling the income distributions using a mixture of gamma densities

The parameters of the mixture distribution is estimated using function gammmixEM. The function implements the algorithm in mixtools. Set the following parameter estimates

$$\pi_1 = 0,779 \quad \pi_2 = 0,221, \; \alpha_1 = 12,823, \; \alpha_2 = 1.487, \; \beta_1 = 0,025, \; \beta_2 = 0,001.$$

The AIC criterion values lead to the conclusion that the quality of fitting in the case of mixtures (AIC = 18606) is better than in the case of a single distribution (AIC = 19275).

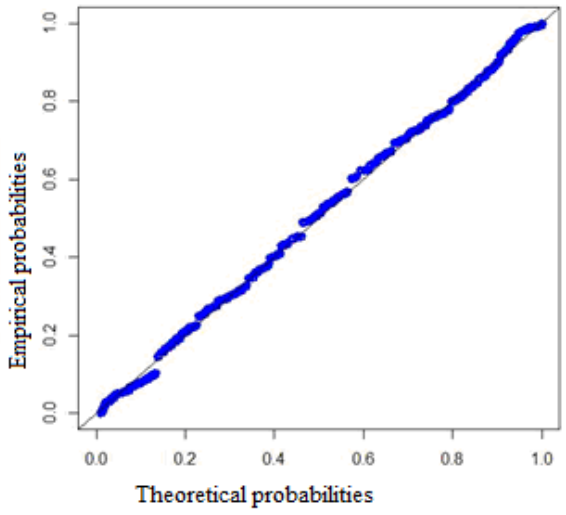

Fig. 2. Theoretical probabilities. Mixture of gamma

Source: own elaboration.

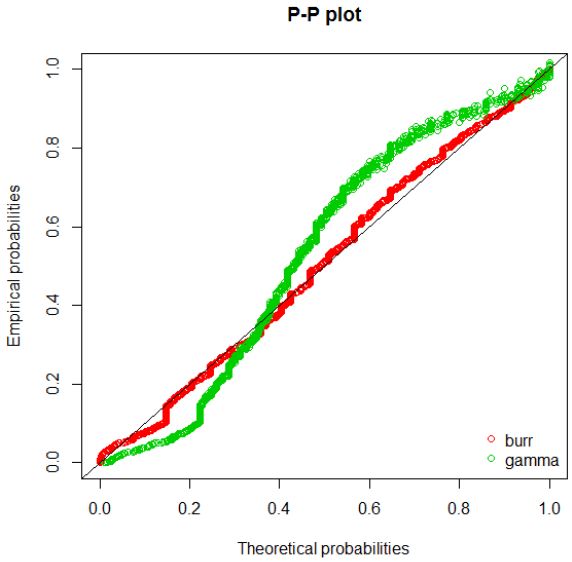

Fig. 3. Theoretical probabilities. Gamma and Burr distribution

Source: own elaboration.

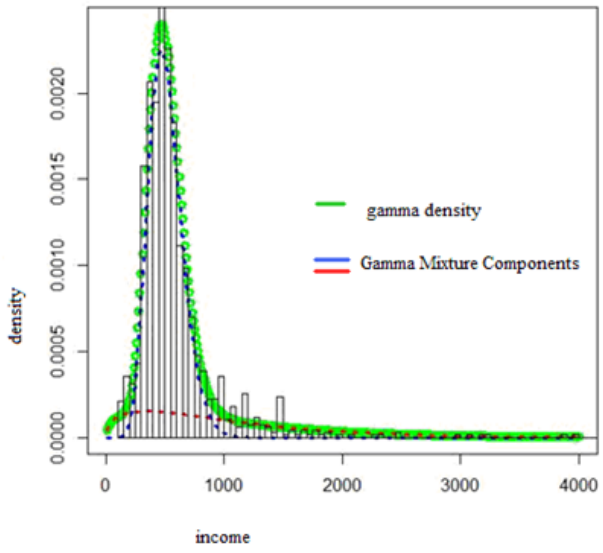

Fig. 4. Gamma Mixture Components

Source: own elaboration.

7. Test for homogeneity in gamma mixture models

We consider two-parameter gamma density [Wong, Li 2012]:

$$f(x,\alpha,\beta)=\frac{\beta^{\alpha}}{\Gamma(\alpha)}x^{\alpha-1}e^{-\beta y} \text{ for } x>0,\ B>0,\ \alpha>1,$$

where α and β are shapes and scale parameters, respectively. Given a set of independent and identically distributed data, we are interested in testing the homogeneity hypothesis H_0 against the alternative hypothesis of a two-component gamma mixture model H_1 where

$$H_0: \quad f(x)=f(x;\alpha,\beta),$$

$$H_1: \ f(x)=\pi\, f_1(x;\alpha_1,\beta_1)+(1-\pi)\,f_2(x;\alpha_2,\beta_2)$$

and $0<\pi<1$ is a mixing proportion. For a parametric hypothesis testing problem, it is customary to use the ordinary Likelihood Ratio Test (LRT) based on the statistic which is defined as

$$LR_n=2\{L(\hat{\pi},\hat{\alpha}_1,\hat{\beta}_1,\hat{\alpha}_2,\hat{\beta}_2)-L(0.5,\hat{\alpha}\,\hat{\beta},\hat{\alpha},\ \hat{\beta})\},$$

where

$$L(\pi,\alpha_1,\beta_1,\alpha_2,\beta_2)=\sum_{i=1}^{n}\log\{\pi\ f(x;\alpha_1,\beta_1)+(1-\pi)\ f(x;\alpha_2,\beta_2) \quad (24)$$

is the log-likelihood function and $\hat{\theta}$ is the MLE of parameter θ. It is well known that the consistency of the MLE, obtained by maximizing (24) directly is not guaranteed. This motivates a penalized procedure coined by Chen and Chen [Chen, Chen 2001], based on the modified log-likelihood function

$$L^{p}(\pi,\alpha_1,\beta_1,\alpha_2,\beta_2)=L(\pi,\alpha_1,\beta_1,\alpha_2,\beta_2)+c\log\{4\pi(1-\pi)\}, \quad (25)$$

where *c* is a positive constant corresponding to the level of modification. An alternative penalty function $c\log(1-|1-2\pi)|$ was suggested by Li [2009].

Denote by $\hat{\theta}^{p}$ the penalized MLE of θ obtained by maximizing () given a suitable value of *c*. The MLRT statistic is

$$LR_n^{p}=2\{L(\hat{\pi}^{p},\hat{\alpha}_1^{p},\hat{\beta}_1^{p},\hat{\alpha}_2^{p},\hat{\beta}_2^{p})-L(0.5,\hat{\alpha}\,\hat{\beta},\hat{\alpha},\ \hat{\beta})\}. \quad (26)$$

Under H_0, the asymptotic distribution of LR_n^p degenerates to zero with a weight $0 < p < 1$ and has χ_2^2 distribution with a weight 1–*p*,

$$LR_n^p \sim p + (1-p)\chi_2^2, \tag{27}$$

where p is the probability that the matrix $n^{-\frac{1}{2}}\sum_{i=1}^{n} U_i$ is negative definite. U_i is defined as

$$U_i = Z_i - V_i Z_i, \tag{28}$$

where $V_i = \left(n\sum_{j=1}^{n} Y_j^T \right)\left(\sum_{i=1}^{n} Y_j Y_j^T \right)^{-1} Y_i$, and Y_i is a random vector given by

$$Y_i = \begin{Bmatrix} -\Gamma^{(1)}(\alpha_0) + \log\beta_0 + \log X_i \\ \alpha_0\beta_0^{-1} - X_i \end{Bmatrix}. \tag{29}$$

$\Gamma^{(k)}(\alpha) = d^k \ln\Gamma(\alpha)/d^k\alpha$, and Z_i is a symmetric random matrix whose elements

$$\begin{aligned} Z_{i[1,1]} &= -\Gamma^{(2)}(\alpha_0) + \{-\Gamma^{(1)}(\alpha_0) + \log\beta_0 + \log X_i\}^2 \\ Z_{i[1,2]} &= \beta_0^{-1} + \{-\Gamma^{(1)}(\alpha_0) + \beta_0 + \log X_i\}(\alpha_0\beta_0^{-1} - X_i) \\ Z_{i[2,2]} &= \alpha_0\beta_0^{-2} + (\alpha_0\beta_0^{-1} - X_i)^2. \end{aligned} \tag{30}$$

The limiting distribution in (27) is known as chi-bar-square distributions. From the definition of U_i in (28), we observe its dependence on random vector Y_i and random matrix Z_i given by (29) and (30), respectively, which are related to the parameter (α_0, β_0) under H_0. In addition the estimates of *p* may also depend on *n* as the random matrix concerned involves a summation of *n* random matrices.

8. Conclusions

In the paper the use of the mixtures of gamma distributions is proposed as a suitable model for the incomes. The concept of mixture distributions is very applicable to income data, as these values form usually a very non-homogenous set. The AIC criterion values lead to the conclusion that the quality of fitting

in the case of mixtures gamma distributions is better than in the case of a single gamma distribution.

We investigate the modified likelihood test for homogeneity in two-component gamma mixture models. The limiting distribution of the test statistic is the parameter-dependent chi-bar-square distributions given by a degeneration to zero with weight p and a chi-square distributions with two degrees of freedom with weight $1–p$.

References

Benaglia T., Chauveau D., Hunter D.R., Young D.S. (2009). *An R Package for Analyzing Finite Mixture Models*. Journal of Statistical Software.

Chen J., Chen H. (2001). *The Likelihood ratio test for homogeneity in finite mixture models*. The Canadian Journal of Statistics.

Chen J., Li P. (2009). *Hypothesis test for normal mixture models: the EM Approach*. The Annals of Statistics.

Fisz M. (1969). *Rachunek prawdopodobieństwa i statystyka matematyczna*. PWN. Warszawa.

Jajuga K. (1990). *Statystyczna teoria rozpoznawania obrazów*. PWN. Warszawa.

Kot S.M. (Ed.) (1999). *Analiza ekonometryczna kształtowania się płac w Polsce w okresie transformacji*. PWN. Warszawa.

Li, P., Chen, J., Marriott, P. (2009) *Non-finite Fisher information and homegenity: an EM approach*, Biometrica 96, 411-426

Mc Donald J.B., Jensen B. (1979). *An analysis of some properties of alternative measures of income inequality based on the gamma distribution function*. Journal of the American Statistical Association.

Titterington D.M., Smith A.F., Makov U.E. (1985). *Statistical analysis of finite mixture distributions*. Wiley.

Wong T.S.T., Li W.K. (2012). *Test for homogeneity in gamma mixture models using likelihood ratio*. Research Report. The University of Hong Kong.