Int. J. Contemp. Math. Sciences, Vol. 9, 2014, no. 2, 57 - 64
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ijcms.2014.312133

The EM Algorithm for the Finite Mixture of Exponential Distribution Models

WANG Yanling and WANG Jixia

College of Mathematics and Information Science
Henan Normal University, Xinxiang 453007, China

Abstract

In this paper, we first propose a finite mixture of exponential distribution model with parametric functions. By using the local constant fitting, the local maximum likelihood estimations of parametric functions are obtained, and their asymptotic biases and asymptotic variances are discussed. Moreover, we propose the EM algorithm to carry out the estimation procedure and give the ascent property of the EM algorithm.

Keywords: finite mixture model; EM algorithm; local maximum likelihood estimation; local constant fitting

1. Introduction

Mixture models are widely used in social science and econometrics. The work for mixture models have been well studied, for example, see Lindsay [1]. The finite mixture model, which has been applied in various fields, is a useful class of mixture models. A lot of efforts have been made to the finite mixture

The Foundation and Frontier Technology Research Projects of Henan Province (122300410396); the Soft Science Project of Henan Province(122400450180); Natural Science Foundation of Henan Educational Committee(2011B110018).

models, such as Früihuirth-Schnatter [2], Rossi, Allenby and McCulloch [3], Hurn, Justel and Robert [4] and Huang, Li and Wang [5] and so on.

In this paper, we propose a finite mixture of exponential distribution model with parametric functions. We allow that the parametric functions of our models are smooth. In order to estimating unknown parametric functions, we develop the estimation procedure by using the local constant fitting. The local maximum likelihood estimations of parametric functions are obtained via local weighted likelihood method. Local likelihood method(see Tibshirani and Hastie [6]) extends the idea of kernel regression. Furthermore, the asymptotic biases and asymptotic variances of local estimations proposed are discussed. Moreover, we propose the EM algorithm to carry out the estimation procedure. For each estimation procedure of parametric functions, it is desirable to estimate the curves over a set of the grid points. We further show that the EM algorithm has the monotone ascent property in an asymptotic sense.

The rest of this paper is organized as follows. In Section 2, we introduce our model and propose the Local Maximum Likelihood Estimations of the parametric functions. The EM algorithm of the local estimations is proposed in Section 3. In Section 4 we give the property of our EM algorithm.

2. Models and Local Maximum Likelihood Estimation

Now we discuss the local maximum likelihood estimations of the parametric functions $p_m(x)$ and $\lambda_m(x)$, $m = 1, 2, \cdots, M$. The log-likelihood function for the collected data $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ is

$$\sum_{i=1}^{n} \log \left[\sum_{m=1}^{M} p_m(X_i)\lambda_m(X_i) exp\{-\lambda_m(X_i)Y_i\} \right]. \tag{2.2}$$

Note that $p_m(x)$ and $\lambda_m(x)$ are parametric functions. In this paper, we will employ the local constant fitting(see Fan and Gijbels [7]) for model (2.1). That is, for a given point x, we use local constants p_m and λ_m to approximate $p_m(x)$ and $\lambda_m(x)$, respectively. So the local weighted log-likelihood function for data $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ is

$$l_n(p, \lambda; x) = \sum_{i=1}^{n} \log \left[\sum_{m=1}^{M} p_m \lambda_m exp\{-\lambda_m Y_i\} \right] K_h(X_i - x), \tag{2.3}$$

where $p = (p_1, p_2, \cdots, p_M)^T$, $\lambda = (\lambda_1, \lambda_2, \cdots, \lambda_M)^T$, $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ be a nonnegative weighted function and h is a properly selected bandwidth. Let $(\tilde{p}, \tilde{\lambda})$ be the maximizer of the local weighted log-likelihood function (2.3). Then, the local maximum likelihood estimations of $p_m(x)$ and $\lambda_m(x)$ are

$$\tilde{p}_m(x) = \tilde{p}_m, \quad \text{and} \quad \tilde{\lambda}_m(x) = \tilde{\lambda}_m, \tag{2.4}$$

respectively.

The asymptotic bias, asymptotic variance and asymptotic normality are studied as the following. Let $\theta = (p^T, \lambda^T)^T$ and denote

$$\eta(y|\theta) = \sum_{m=1}^{M} p_m \lambda_m exp\{-\lambda_m y\}, \quad l(\theta, y) = \log \eta(y|\theta).$$

Furthermore, let $\theta(x) = (p^T(x), \lambda^T(x))^T$, and denote

$$I(x) = -E\left[\frac{\partial^2 l(\theta(X), Y)}{\partial\theta\partial\theta^T}|X = x\right]$$

and

$$\Lambda(u|x) = \int_Y \frac{\partial l(\theta(x), y)}{\partial\theta}\eta(y|\theta(u))dy.$$

For $\xi = 1, 2, \cdots, M$, denote $\tilde{\lambda}^*_m = \{\tilde{\lambda}_m - \lambda_m\}$. For $\xi = 1, 2, \cdots, M-1$, denote $\tilde{p}^*_m = \{\tilde{p}_m - p_m\}$. Let $\tilde{\lambda}^* = (\tilde{\lambda}^*_1, \tilde{\lambda}^*_2, \cdots, \tilde{\lambda}^*_M)^T$, $\tilde{p}^* = (\tilde{p}^*_1, \tilde{p}^*_2, \cdots, \tilde{p}^*_{M-1})^T$ and $\tilde{\theta}^* = ((\tilde{p}^*)^T, (\tilde{\lambda}^*)^T)^T$. Furthermore, Let $g(\cdot)$ is the marginal density function of X, $\nu_0(K) = \int K^2(z)dz$ and $\kappa_2(K) = \int z^2 K(z)dz$. Then, The asymptotic bias and asymptotic variance of $\tilde{\theta}^*$ are

$$bias(\tilde{\theta}^*) = I^{-1}(x)\left\{\frac{g'(x)\Lambda'_u(x|x)}{g(x)} + \frac{1}{2}\Lambda''_u(x|x)\right\}\kappa_2(K)h^2$$

and

$$Var(\tilde{\theta}^*) = \frac{\nu_0(K)I^{-1}(x)}{g(x)},$$

respectively.

Under some regularity conditions, $\tilde{\theta}^*$ has the asymptotic normal distribution. The proofs of above results are similar to that of Theorem 2 in Huang, Li and Wang [5]. In this paper, we mainly study the EM algorithm for the finite mixture of exponential distribution model (2.1).

3. The EM Algorithm

The mixture problem is formulated as an incomplete-data problem in the EM framework. The observed data (X_i, Y_i)s are viewed as being incomplete, and the unobserved Bernoulli random variables are introduced as following:

$$Z_{im} = \begin{cases} 1, & \text{if } (X_i, Y_i) \text{is in the } m\text{th group} \\ 0, & \text{otherwise.} \end{cases}$$

Let $Z_i = (Z_{i1}, Z_{i2}, \cdots, Z_{iM})^T$, the associated component identity or label of (X_i, Y_i). Then, $\{(X_i, Y_i, Z_i), i = 1, 2, \cdots, n\}$ are the complete data, and complete log-likelihood function corresponding to (2.2) is

$$\sum_{i=1}^{n}\sum_{m=1}^{M} Z_{im}\left[\log p_m(X_i) + \log \lambda_m(X_i) - \lambda_m(X_i)Y_i\right].$$

For $x \in \{u_1, u_2, \cdots, u_N\}$, The set of grid points at which the unknown functions are to be evaluated. We define a local weighted complete log-likelihood function as

$$\sum_{i=1}^{n}\sum_{m=1}^{M} Z_{im}\left[\log p_m + \log \lambda_m - \lambda_m Y_i\right] K_h(X_i - x).$$

Note that Z_{im}s do not depend on the choice of x. We have $\lambda_m^{(l)}(\cdot)$ and $p_m^{(l)}(\cdot)$ in the lth cycle of the EM algorithm iteration. Then, in the E-step of $(l+1)$th cycle, the expectation of the latent variable Z_{im} is given by

$$r_{im}^{(l+1)} = \frac{p_m^{(l)}(X_i)\lambda_m^{(l)}(X_i)\exp\{-\lambda_m^{(l)}(X_i)Y_i\}}{\sum_{m=1}^{M} p_m^{(l)}(X_i)\lambda_m^{(l)}(X_i)\exp\{-\lambda_m^{(l)}(X_i)Y_i\}}. \tag{3.1}$$

In the M-step of the $(l+1)$th cycle, we maximize

$$\sum_{i=1}^{n}\sum_{m=1}^{M} r_{im}^{(l+1)}\left[\log p_m + \log \lambda_m - \lambda_m Y_i\right] K_h(X_i - x), \tag{3.2}$$

for $x = u_j, j = 1, 2, \cdots, N$. The maximization of equation (3.2) is equivalent to maximizing

$$\sum_{i=1}^{n}\sum_{m=1}^{M} r_{im}^{(l+1)}\left[\log p_m\right] K_h(X_i - x) \tag{3.3}$$

and for $m = 1, 2, \cdots, M$

$$\sum_{i=1}^{n}\sum_{m=1}^{M} r_{im}^{(l+1)}\left[\log \lambda_m - \lambda_m Y_i\right] K_h(X_i - x), \tag{3.4}$$

separately. For $x \in \{u_1, u_2, \cdots, u_N\}$, the solution for maximization of equation (3.3) is

$$p_m^{(l+1)}(x) = \frac{\sum_{i=1}^{n} r_{im}^{(l+1)} K_h(X_i - x)}{\sum_{i=1}^{n} K_h(X_i - x)}, \tag{3.5}$$

and the solution for maximization of equation (3.4) is

$$\lambda_m^{(l+1)}(x) = \frac{\sum_{i=1}^{n} r_{im}^{(l+1)} K_h(X_i - x)}{\sum_{i=1}^{n} r_{im}^{(l+1)} Y_i K_h(X_i - x)}. \tag{3.6}$$

Furthermore, we update $p_m^{(l+1)}(X_i)$ and $\lambda_m^{(l+1)}(X_i)$, $i = 1, 2, \cdots, n$ by linearly interpolating $p_m^{(l+1)}(u_j)$ and $\lambda_m^{(l+1)}(u_j)$, $j = 1, 2, \cdots, N$, respectively. We summarize the EM algorithm as the following.

The EM algorithm

Initial Value: Conduct a mixture of polynomial regressions with constant proportions and variance, and obtain the estimations of mean function $\bar{\mu}_m(x)$ and parameters $\bar{p}_m$. Set the initial values $\lambda_m^{(1)}(x) = 1/\bar{\mu}_m(x)$ and $p_m^{(1)}(x) = \bar{p}_m$.

E-step: use equation (3.1) to calculate $r_{im}^{(l+1)}$ for $i = 1, 2, \cdots, n$ and $m = 1, 2, \cdot, M$.

M-step: For $m = 1, 2, \cdot, M$ and $j = 1, 2, \cdot, N$, evaluate $p_m^{(l+1)}(u_j)$ in (3.5) and $\lambda_m^{(l+1)}(u_j)$ in (3.6). Further, we obtain $p_m^{(l+1)}(X_i)$ and $\lambda_m^{(l+1)}(X_i)$ using linear interpolation.

Iteratively update the E-step and the M-step with $l = 2, 3, \cdots$, until the algorithm converges.

It is well known that the bandwidth selection can be tuned to optimize the performance of the estimated parametric functions. At the end of this section, we select the bandwidth of the local estimations for the parametric functions. We select bandwidth h via the Cross-validation method, which is discussed in detail in Fan and Gijbels [8].

4. Property of the EM Algorithm

In this section, we study the EM algorithm we proposed preserves the ascent property. We first give the following Assumptions.

(A1) The sample $\{(X_i, Y_i), i = 1, 2, \cdots, n\}$ is independent and identically distribution from (X, Y), and the support for X, denoted by χ, is a compact subset of R.

(A2) The marginal density function $g(x)$ of X is twice continuously differentiable and positive for all $x \in \chi$.

(A3) there exists a function $M(y)$ with $E[M(y)] < \infty$, such that for all y, and all θ in a neighborhood of $\theta(x)$, we have $\left|\frac{\partial^3 l(\theta,y)}{\partial\theta_j \partial\theta_k \partial\theta_l}\right| < M(y)$.

(A4) The parametric function $\theta(x)$ has continuous second derivatives. Furthermore, for $m = 1, 2, \cdots, M$, $\lambda_m > 0$ and $p_m > 0$ hold for all $x \in \chi$.

(A5) The kernel function $K(\cdot)$ has a bounded support and satisfies that $\int K(z)dz = 1$, $\int zK(z)dz = 0$, $\int z^2K(z)dz < \infty$, $\int K^2(z)dz < \infty$ and $\int |K^3(z)|dz < \infty$.

Let $\theta^{(l)} = (p^{(l)}(\cdot), \lambda^{(l)}(\cdot))$ be the estimated functions in the lth cycle of the EM algorithm proposed. The local weighted log-likelihood function (2.3) is

rewritten as

$$l_n(\theta) = \sum_{i=1}^{n} l(\theta, Y_i) K_h(X_i - x). \tag{4.1}$$

Then, we have the following theorem.

Theorem 4.1 Assume that conditions (A1)-(A5) hold. For any given point x, suppose that $\theta^{(l)}(\cdot)$ has a continuous first derivative, and $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Then, we have

$$\liminf_{n\to\infty} \frac{1}{n} \left[l_n(\theta^{(l+1)}(x)) - l_n(\theta^{(l)}(x)) \right] \geq 0 \tag{4.2}$$

in probability.

Proof Assume that the unobserved data $\{\xi_i, i = 1, 2, \cdots, n\}$ is a random sample from population ξ. Then, the complete data $\{(X_i, Y_i, \xi_i), i = 1, 2, \cdots, n\}$ can be viewed as a sample from (X, Y, ξ). Let $h(y, m|\theta(x))$ be the joint distribution of (Y, ξ) given $X = x$, and $g(x)$ be the marginal density of X. Conditioning on $X = x$, Y follows a distribution $\eta(y|\theta(x))$. Then, the local weighted log-likelihood function (2.3) can be rewritten as

$$l_n(\theta) = \sum_{i=1}^{n} \log[\eta(Y_i|\theta)] K_h(X_i - x). \tag{4.3}$$

The conditional probability of $\xi = m$ given y and θ is

$$f(m|y, \theta) = h(y, m|\theta)/\eta(y|\theta) = p_m \lambda_m \exp\{-\lambda_m y\} / \sum_{m=1}^{M} p_m \lambda_m \exp\{-\lambda_m y\}. \tag{4.4}$$

For given $\theta^{(l)}(X_i), i = 1, 2, \cdots, n$, it is clear that $\int f(m|Y_i, \theta^{(l)}(X_i))dm = 1$. Then, we have

$$\begin{aligned} l_n(\theta) &= \sum_{i=1}^{n} \log[\eta(Y_i|\theta)] \left[\int f(m|Y_i, \theta^{(l)}(X_i))dm \right] K_h(X_i - x) \\ &= \sum_{i=1}^{n} \left\{ \int \log[\eta(Y_i|\theta)][f(m|Y_i, \theta^{(l)}(X_i))]dm \right\} K_h(X_i - x). \end{aligned} \tag{4.5}$$

By equation (4.4), we have

$$\log[\eta(Y_i|\theta)] = \log[h(Y_i, m|\theta)] - \log[f(m|Y_i, \theta)]. \tag{4.6}$$

Thus, we have

$$l_n(\theta) = \sum_{i=1}^{n} \log[\eta(Y_i|\theta)] \left[\int f(m|Y_i, \theta^{(l)}(X_i))dm \right] K_h(X_i - x)$$

$$= \sum_{i=1}^{n} \left\{ \int \log[h(Y_i, m|\theta)][f(m|Y_i, \theta^{(l)}(X_i))]dm \right\} K_h(X_i - x)$$

$$- \sum_{i=1}^{n} \left\{ \int \log[f(m|Y_i, \theta)][f(m|Y_i, \theta^{(l)}(X_i))]dm \right\} K_h(X_i - x), \quad (4.7)$$

where $\theta^{(l)}(X_i)$ is the estimation of $\theta(X_i)$ at the lth iteration. Taking expectation leads to calculating equation (3.1). In the M-step of the EM algorithm, we update $\theta^{(l+1)}(x)$ such that

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \int \log[h(Y_i, m|\theta^{(l+1)}(x))][f(m|Y_i, \theta^{(l)}(X_i))]dm \right\} K_h(X_i - x)$$

$$\geq \frac{1}{n} \sum_{i=1}^{n} \left\{ \int \log[h(Y_i, m|\theta^{(l)}(x))][f(m|Y_i, \theta^{(l)}(X_i))]dm \right\} K_h(X_i - x).$$

It suffices to show that

$$\frac{1}{n} \sum_{i=1}^{n} \left\{ \int \log \left[\frac{f(m|Y_i, \theta^{(l+1)}(x))}{f(m|Y_i, \theta^{(l)}(x))} \right] f(m|Y_i, \theta^{(l)}(X_i))dm \right\} K_h(X_i - x) \leq 0 \quad (4.8)$$

in probability. Let

$$L_{n1} = \frac{1}{n} \sum_{i=1}^{n} \phi(Y_i, X_i) K_h(X_i - x),$$

where

$$\phi(Y_i, X_i) = \int \log \left[\frac{f(m|Y_i, \theta^{(l+1)}(x))}{f(m|Y_i, \theta^{(l)}(x))} \right] f(m|Y_i, \theta^{(l)}(X_i))dm.$$

By using assumptions (A1)-(A4), we have $f(m|Y, \theta^{(l)}(x)) > a > 0$ for some small value a, and $E\{[\phi(Y, X)]^2\} < \infty$. Then, by assumption (A5) and theorem A by Mack and Silverman [9], we have

$$\sup_J |L_{n1} - g(x)E[\phi(Y, X)]| = o_p(1),$$

where J is a compact interval on which the density of X is bounded below from 0. The proof follows that

$$E[\phi(Y, x)] = E \left\{ \int \log \left[\frac{f(\xi|Y, \theta^{(l+1)}(x))}{f(\xi|Y, \theta^{(l)}(x))} \right] f(m|Y, \theta^{(l)}(x))dm \right\}$$

$$\leq E \left\{ \log \left[\int \left[\frac{f(\xi|Y, \theta^{(l+1)}(x))}{f(\xi|Y, \theta^{(l)}(x))} \right] f(m|Y, \theta^{(l)}(x))dm \right] \right\}.$$

The proof of theorem 4.1 is completed.

References

[1] Lindsay, B. Mixture Models: Theory, Geometry and Applications[M]. Hayward, CA: Institute of Mathematical Statistics, 1995.

[2] Frühwirth-Schnatter, S. Markov Chain Monte Carto estimation of classical and dynamic switching and mixture models[J]. Journal of the American Statistical Association, 2001, 96: 194-209.

[3] Rossi, P., Allenby, G. and McCulloch, R. Bayesian Statistics and Marketing[J]. Chichester: Wiley, 2005.

[4] Hurn, M., Justel, A. and Robert, C. Estimating mixtures of regressions[J]. Journal of Computational and Graphical Statistics, 2003, 12:55-79.

[5] Huang, M., Li, R. and Wang, S. Nonparametric mixture of regression models[J]. Journal of the American Statistical Association, 2013, 108:929-941.

[6] Tibshirani, R. and Hastie, T. Local likelihood estiamtion[J]. Journal of the American Statistical Association, 1987, 82: 559-567.

[7] Fan,J and Gijbels,I. Local Polynomial Modelling and Its applications[M]. London: Chapman and Hall,1996.

[8] Fan, J. and Gijbels,I. Variable bandwidth and local linear regression[J]. Annals of Statistics, 1992, 20: 2008-2036.

[9] Mack, Y., Silverman, B. Weak and strong uniform consistency of kernel regression estimates[J]. Probability Theory and Related Fieids, 1982, 61: 405-415.

Received: December 1, 2013