Jour.Soc.Sci. & Humn. 1996-2000
pp.59-64

EM-ALGORITHM AND LATENT CLASS MODEL

Dr Junaid Sagheer Siddiqui

ABSTRACT

This article gives a brief account of the latent class model and the EM-algorithm for obtaining' maximum likelihood estimates. The EM routine is presented here in the special setting of latent class model, The relevant score equations also are *derived which could be used to obtain the iterative solution of the maximum likelihood estimates and the missing*

I INTRODUCTION

There are several types of latent variable models depending upon the assumption on latent and manifest variables'. One of them is a latent *class* model when we assume the existence of single latent categorical variable. The level of the latent variable refers to as latent class. The basic assumption, in this case, knowing the exact NUMBER of latent classes, is that the manifest variables in all the latent classes are independent. In this article the aim is to derive equations for the solution of maximum likelihood estimator using the EH-algorithm of Dumpster et al (1977) in the setting of latent lass model.

The EM-algorithm of Dempster, Laird and Rubin (1977) is a -particular iterative procedure which provides a general approach for computing the maximum likelihood estimates when the observed data can be viewed as "incomplete data." The term incomplete *data* refer to a situation where the data may be viewed as coming from two sample spaces, say x and γ. The observed data which is a realisation from x is referred to as incomplete *data* and the variable Y which is a realisation from the sample space γ, is not directly observable and can be observed only indirectly through X. In other words, we view Y = (X, Z) where Z is referred to as *missing data,* i.e. incomplete data is the union of observed and missing data and taking values from the sample space γ.

I.I LATENT CLASS MODEL

Suppose $X^T = (x_1..., X_p)$ denotes a vector of p-manifest variables, maybe discrete 011: continuous or a mixture of the two. Suppose that there is a single categorical latent variable Y, with K. categories (classes) i.e. Y-l...y, ..., K, is sufficient to explain the association among the observed variables. The probability that an independently chosen individual from the ps.pu.lat ion belong to the k-th latent class is:

$$P(Y=y) = \lambda_y \qquad (1)$$

where

$$\sum_{y=1}^{K} \lambda_y = 1. \qquad (2)$$

Under the assumption that the observable random variable x are independent when observed in a. latent class, this assumption is, in latent class terms, known as axiom of local independence. Then the joint density function of x can be written as:

$$F(x; \varnothing) = \sum_{i=1}^{K} \sum_{y=1} \lambda_y \prod^{p} g(x_I \mid \theta_y, y)$$

where $\varnothing^T = (\lambda_1, \ldots\ldots, \lambda_K, \theta_1, \ldots.., \theta_k)$ are the parameters of the latent class model. A latent class model will be completely specified when the assumption on the form of g (.) is made. However there can be several possibilities, but in this article we consider the general case that g (.) has an arbitrary distribution. For further details on latent class models see Lazarsfel and Henry. (1968) Aitkin et al (1981), and Clogg and Goodman (1984) and Everitt (1984).

I.2 THE EM-ALGORITHM

Suppose that the complete data variable Y has a density function f (y| $\varnothing$) and let g (x| $\varnothing$) be the corresponding function of the incomplete data. Then g (.) and f (.) are related by

$$g(x \mid \varnothing) = \zeta \; f(y \mid \varnothing)\, dz \qquad (4)$$

where $\varnothing = (\varnothing_1, \ldots, \varnothing_q)^T$ is a vector of unknown parameters (reference). In many problems the maximisation of the complete data specification is simpler than that of the incomplete data specification. The main feature of the EM-algorithm is the maximisation of f (y | $\varnothing$) over $\varnothing$. Since y is not directly observable, f (y | $\varnothing$) cannot be calculated, so it is replaced by its conditional expectation given x and the current estimate parameters. Let

$$h(y \mid x, \varnothing) = \frac{f(y \mid \varnothing)}{g(y \mid \varnothing)} \qquad (5)$$

be the conditional density of y given x and $\varnothing$. Then the log-likelihood of incomplete data can be written, (in Dempster et al's notations we will not use these notation in other sections),

$$L(\varnothing') = Q(\varnothing' \mid \varnothing) - H(\varnothing' \mid \varnothing) \qquad (6)$$

$\varnothing'$ is a value-derived form $\varnothing$

where $Q(\varnothing' \mid \varnothing) = E\{\log f(y \mid \varnothing') \mid x, \varnothing\}$, and

$$H(\varnothing' \mid \varnothing) = E\{\log f(h \mid \varnothing') \mid x, \varnothing\}.$$

Considering the relationship between complete and incomplete data functions, the EH -algorithm suggest to maximize $L(\varnothing' \mid \varnothing)$ over ^ by maximizing $H(\varnothing' \mid \varnothing)$. Since the log-likelihood function of the complete data cannot be determined therefore instead of maximizing this we maximize the expected value of log-likelihood function conditional on observed data and the parameters, i.e.

$$Q(\varnothing' \mid \varnothing) = E\{\log f(y \mid \varnothing') \mid x, \varnothing\}, \qquad (7)$$

This leads to the following iterative procedure involving two steps. Estimation and Maximization (EH), at each cycle. Writing $\varnothing^{[s]}$ for the estimation of the parameters after s-th iteration, the next for the computation $\varnothing^{[s+1]}$ has the general form:

E-step: Compute the expected log-likelihood function using the current estimate

$\varnothing^{[s]}$ i.e. $Q(\varnothing' \mid \varnothing^{[s]})$.

M-step: choose on improved estimate $\varnothing^{[s+1]}$ *which maximise* $Q(\varnothing' \mid \varnothing^{[s]})$.

The E and M steps iterate until the estimates converge and no further improvement is possible in terms of maximising Q. The situation here is that the missing data and the estimates are both functions of each other and to estimate one we need to know the other. If we know the missing data we can estimate the parameters and if we have the parameters we can estimate the missing data, or equivalently the complete data likelihood which is a sufficient statistics based on the observed and missing data. So we start by assigning some initial values to the missing data and then estimate the parameters. Now with the help of the first estimate we can compute a better estimate for the missing values. And then using the improved estimate of the missing value further improved estimate of the parameters can be obtained, and this process is repeated until no further improvement is r We can also start the process other way round i.e. by assigning some initial values to the parameters and then estimate the missing values and repeating the two step as before. For further details see Redner and Walker (1984), Boyles (1983) and Wu (1983).

1. MAXIMUM LIKELIHOOD ESTIMATION USING THE EN-ALGORITHM.

In this section the EM equations suitable for the fitting arbitrary latent class model, described as in (3), are derived. Let $(x_1, \ldots, x_n)$ denote n random observations and suppose that there exist K classes and each observation x_j, j= 1, 2, …, n, is associated with an unobservable variable z_j which is an indicator vector for the true classification. The missing data z_j are K-dimensional indicator vectors:

$$Z_j (z_{1j}, \ldots\ldots, z_{Kj})^T$$

where

$$Z_{kj} = \{ 1 \text{if } x_j \in o_k$$
$$0 \text{ otherwise}$$

and o_k denotes the component population. Assume that the Z_j are. independently and identically distributed with density function $h(z_h \mid \varnothing)$. Let $g_k (z_j \mid \varnothing)$. to k=l, ..., K denote the component density and assume that g_k have identical forms.

If the missing data are known, the only part of the $\varnothing$ we will he concerned with is $\varnothing_k$ $g_k (x \mid z, \varnothing)$ is equivalent to $g (x \mid k, \varnothing_k)$ where k= 1, ..., K to a latent class. From (5) the complete data density can be expressed as:

$$f(y_j \mid \varnothing) = g(x_j \mid Z_j, \varnothing)\, h(Z_j \mid \varnothing)$$
$$= \prod_{k-1}^{k} [g(x_j \mid z_{kj}, \varnothing)\, h(z+k\, j \mid Z_{kj}, \varnothing)]^{z_{kj}} \qquad (8)$$

therefore the complete data log-likelihood, equivalent to Q can be written as:

$$(y \mid \varnothing) = \sum_{j=1}^{n} \sum_{k=1}^{k} Z_{kj} \log g(x_j \mid z_{kj}, \varnothing) + \sum_{j=1}^{n} \sum_{k=1}^{k} Z_{kj} \log g(x_j \mid z_{kj}, \varnothing) \quad (9)$$

At the **E-step** of the EH-algorithm instead .of maximising the usual log-likelihood function, we maximise the expected log-likelihood of the complete data. The complete data involve the missing value z_j, and the E-step i-equire the estimation of the components of z, the observation x and the current estimates of density parameters. The estimated values of z_j are the conditional probabilities that an observation be label led to the respective class.

At the **M-step**, the maximum 1likelihood estimates for the parameters can be obtained by solving the set of score equations obtained from the first order derivatives, with respect to the parameters, of the log-likelihood function. In this case the probability distribution of the missing data, equivalent to (1)., is

$$h(z_k = 1 \mid \varnothing) = \lambda_k \text{ for } k = 1, \ldots, K \qquad (10)$$

with constraint $\sum_{k=1}^{k} \lambda_k = 1$. Suppose a sinfflo parameter is required for one variable in a given class, then $\theta_k = (\theta_{1k}, \ldots, \theta_{pk})^p$ denote density parameter in the k-th class, therefore the complete data parameter vector can be written as:

$$\varnothing = (\lambda_1, \ldots, \lambda_k, \varnothing_1, \ldots, \varnothing_k)$$

Using this notation the conditional density of the incomplete x data given the missing data z can be expressed as

$$G(x_j \mid z_{kj}, \varnothing) = g(x_j \mid z_{kj}, \theta_k) \qquad (11)$$

The log-likelihood function under the constrained (2) can be written as:

$$(y \mid \varnothing) = \sum_{j=1}^{n} \sum_{k=1}^{k} Z_{kj} \log g(x_j \mid z_{kj}, \theta_k) + \sum_{j=1}^{n} \sum_{k=1}^{k} Z_{kj} \log \lambda_k) + \mu \sum_{k=1}^{k} \lambda_k - 1) \quad (12)$$

Partially differentiating equation (12) and a little algebraic manipulation, the score equation for the EM iteration can be obtained as below:

$$\lambda_k = \frac{\sum_{j=1}^{n} h(z_{kj} \mid x_j, \theta_k)}{n} \qquad k = 1, 2, \ldots, k \quad (13)$$

and

$$\sum_{j=1}^{n} \frac{h(z_{kj} \mid x_j, \varnothing)}{f(y \mid \varnothing)} \frac{\delta}{\delta \theta_{jk}} g(x_j \mid z_{kj}, \theta_k) = 0 \quad (14)$$

where

$$h(z_{kj} \mid x_j, \varnothing) = \frac{\lambda_k g(x_j \mid z_{kj}, \theta_k)}{\sum \lambda_k g(x_j \mid z_{kj}, \theta_k)} \quad (15)$$

is the posterior probability that an individual is classified to the k-th sub population, give the observation and the parameters. An individual will be classified as a member of a latent class for which (15) has maximum value.

If the distributional form of g (.) is completely specified, the equations (13), (14) and (15) may be used to fit a latent class model for known value of K. To get suitable model, models with several different values of K (e.g. K, K+1, ...) are suggested to fit and the model which explain a large degree of association among the manifest variable can be considered as appropriate model see Aitkin et al (1981) and Everitt (1884).

REFERENCES

Aitkin, Andersen & Hinde "Statistical modelling of data on teaching styles". *J. Roy. Statistic.* Soc., A, 44, 419-61. 1981.

Boyles "On the convergence of the EM algorithm". *J. Roy. Statistic, Soc.*, B, 45, 47-50.1983.

Clogg & Goodman "Latent structure analysis of set of multidimensional contingency table". *J. Amer. Statistic. Assoc.* , 79, 762-71. 1984.

Dempster, Laird & Rubin "Maximum likelihood from incomplete data via EM algorithm". *J. Roy. Statistic.* Soc., B, 39, 1-38. 1977.

Everitt — *An introduction to latent variable models.* London. Chapman and Hall.1984.

Lazaersfedl &Henry — *Latent Structure analysis.* N. Y. Houghton & Mifflin Co.1968.

Redner & Walker — "Mixture density, Maximum likelihood and EM algorithm". *Siam Review,* 26, 195-234. 1984.

Wu — "Jackknife, bootstrap and other resembling methods in regression". *Ann. Statistic.* A, 14, 1261-1343. 1983.