A note on EM algorithm for mixture models

Weixin Yao

A Note On EM Algorithm For Mixture Models

Weixin Yao

Department of Statistics, Kansas State University, Manhattan, Kansas 66506, U.S.A.

wxyao@ksu.edu

Abstract

Expectation-maximization (EM) algorithm has been used to maximize the likelihood function or posterior when the model contains unobserved latent variables. One main important application of EM algorithm is to find the maximum likelihood estimator for mixture models. In this article, we propose an EM type algorithm to maximize a class of mixture type objective functions. In addition, we prove the monotone ascending property of the proposed algorithm and discuss some of its applications.

Key words: Adaptive regression; EM algorithm; Edge-preserving smoothers; Mode; Robust regression.

1 Introduction

Expectation-maximization (EM) algorithm (Dempster, Laird, and Rubin, 1977) has been used to maximize the likelihood function or posterior when the model contains unobserved latent variables. The EM algorithm iterates between an *expectation* (E) step and a *maximization* (M) step. In the E step, we compute the expectation of the log likelihood of complete data with respect to laten variables given the current parameter estimates. In the M step, we maximize the expected log likelihood of complete data. Therefore, the EM algorithm transfer the problem of maximizing the original log likelihood to the problem of maximizing the expected log likelihood of complete data, which is usually much easier to deal with.

One of the important applications of the EM algorithm is to find the maximum likelihood estimator for finite mixture models. They are natural models for unobserved population heterogeneity and are generally applicable when one samples from a population which consists of several homogeneous subpopulation. The homogeneous subpopulations will be called components of the population. The random variable X is said to have a *m-component finite mixture* density if

$$f(x;\boldsymbol{\theta}) = \sum_{j=1}^{m} \pi_j f_j(x;\lambda_j), \tag{1.1}$$

where $\boldsymbol{\theta} = (\lambda_1 \ldots, \lambda_m, \pi_1, \ldots, \pi_m)$, π_js are mixing proportions, and $f_j(x;\lambda_j)$ is the jth component density with parameter λ_j. Mixture models have experienced increased interest over last decades. Mixture models can be used for cluster analysis, latent class analysis, discriminant analysis, image analysis, survival analysis, disease mapping, meta analysis, and more. They provide extremely flexible descriptive models for distributions in data analysis and inference. For a general introduction to mixture models, see Lindsay (1995), Böhning (1999), Mclachlan and Peel (2000), and Frühwirth-Schnatter (2006).

Unlike the traditional EM algorithm for mixture model, which focuses on finding the MLE of the model parameters $\boldsymbol{\theta}$ in (1.1), Li, Ray, and Lindsay (2007) proposed the Modal EM (MEM) algorithm to find the mode of the mixture density (1.1) with *known* $\boldsymbol{\theta}$ and successfully apply it to do nonparametric clustering. In this article, we prove that the MEM algorithm can be applied to maximize a general mixture type objective function

$$f(x) = \sum_{k=1}^{K} w_k \left[\log \left\{ \sum_{l=1}^{L} a_{kl} f_{kl}(x) \right\} \right], \tag{1.2}$$

where K, L, w_ks and a_{kl}s are known positive constants, $f_{kl}(x)$s are positive known functions, and x can be scalar or vector. We will call the MEM algorithm in such situation generalized modal EM algorithm (GMEM).

When $K = 1$, the objective function (1.2) is simplified to

$$f(x) = w_1 \log \left\{ \sum_{l=1}^{L} a_{1l} f_{1l}(x) \right\} \propto \sum_{l=1}^{L} a_{1l} f_{1l}(x). \tag{1.3}$$

Therefore, the MEM algorithm (Li, Ray, and Lindsay, 2007) is a special case of the proposed GMEM if we further assume $\sum_{l=1}^{L} a_{1l} = 1$ and $f_{1l}(x)$s are density functions.

We will discuss some applications of the GMEM algorithm. Specifically, we will discuss the applications of the proposed algorithm to adaptive linear regression (Yuan and De Gooijer, 2007), adaptive nonparametric regression (Linton and Xiao, 2007), a class of robust nonparametric regression, and the edge-preserving smoothers for image processing proposed by Chu, et al. (1998). In addition, we will also apply the GMEM algorithm to a special class of Generalized M estimators (GM estimators for short) (Hampel, et al., 1986).

The remainder of this paper is organized as follows. In Section 2, we introduce the proposed EM type algorithm for a general mixture type objective function (GMEM). In Section 3, we introduce some applications of the proposed GMEM algorithm. We give some discussion in Section 4.

2 New GMEM Algorithm

Instead of focusing on estimating the mixture model parameter $\boldsymbol{\theta}$ in (1.1), Li, Ray, and Lindsay (2007) introduced the Modal EM (MEM) algorithm to find the mode of mixture density (1.1), i.e., maximize the mixture density $p(x; \boldsymbol{\theta})$ when $\boldsymbol{\theta}$ is *known*. In this article, we prove that the MEM algorithm can be applied to a more general mixture type objective function (1.2). Specifically, given the initial value $x^{(0)}$, in the $(t+1)^{\text{th}}$ step of proposed GMEM algorithm,

E Step: Let

$$\pi_{kl}^{(t+1)} = \frac{a_{kl} f_{kl}(x^{(t)})}{\sum_{l=1}^{L} a_{kl} f_{kl}(x^{(t)})}, \; k = 1, \ldots, K; \; l = 1, \ldots, L. \tag{2.1}$$

M Step: Update

$$x^{(t+1)} = \arg\max_x Q(x \mid x^{(t)}) \tag{2.2}$$

where

$$Q(x \mid x^{(t)}) = \sum_{k=1}^{K}\sum_{l=1}^{L}\left\{w_k \pi_{kl}^{(t+1)} \log f_{kl}(x)\right\}. \tag{2.3}$$

If $f_{kl}(x)$ is a normal density with mean μ_{kl} and variance σ_{kl}^2, then the above M step has an explicit formula, i.e.,

$$x^{(t+1)} = \frac{\sum_{k=1}^{K}\sum_{l=1}^{L} w_k \pi_{kl}^{(t+1)} \mu_{kl} \sigma_{kl}^{-2}}{\sum_{k=1}^{K}\sum_{l=1}^{L} w_k \pi_{kl}^{(t+1)} \sigma_{kl}^{-2}}.$$

The following Theorem proves that the above GMEM algorithm monotonely increases the objective function (1.2) after each iteration. Its proof is given in the appendix.

Theorem 2.1 *The objective function (1.2) is non-decreasing after each iteration of the above generalized MEM (GMEM) algorithm, i.e., $f(x^{(t+1)}) \geq f(x^{(t)})$, until a fixed point is reached. The GMEM is strictly monotonely increasing at the $(t+1)$th step, i.e., $f(x^{(t+1)}) > f(x^{(t)})$ if one of the following two conditions are satisfied:*

1. *There exists $1 \leq k \leq K$ and $1 \leq l_1 < l_2 \leq L$ such that*

$$\frac{f_{kl_1}(x^{(t+1)})}{f_{kl_1}(x^{(t)})} \neq \frac{f_{kl_2}(x^{(t+1)})}{f_{kl_2}(x^{(t)})}.$$

2. *In the M step of (2.2), $Q(x^{(t+1)} \mid x^{(t)}) > Q(x^{(t)} \mid x^{(t)})$.*

Based on the above theorem, we can see that if $x^{(t+1)}$ is the unique maximizer in (2.2), then the objective function $f(x)$ in (1.2) will increase after the iteration. In addition, it can be seen that the objective function $f(x)$ will also increase if we only increase $Q(x \mid x^{(t)})$ in M step instead of maximizing it.

Note, however, like other general optimization algorithms, the proposed GMEM algo-

rithm is only guaranteed to converge to a local maximum of (1.2). Therefore, it is prudent to run the GMEM algorithm starting from different initial values, if we want to find the global maximum of (1.2).

3 Some Applications of GMEM

In this section, we will discuss some, but not exhaustive, applications of the proposed GMEM algorithm.

3.1 Adaptive linear regression

Suppose $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled from the linear regression model

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon, \tag{3.1}$$

where $\mathbf{x}$ is a p-dimensional vector of covariates independent of the error ϵ with $\mathrm{E}(\epsilon) = 0$. The least squares estimator (LSE) is traditionally used to estimate $\boldsymbol{\beta}$. For normally distributed errors, LSE is exactly the maximum likelihood estimate (MLE). However, LSE will lose some efficiency when the error is not normally distributed. Yuan and De Gooijer (2007) proposed to adaptively estimate the slope parameters by maximizing

$$\sum_{i=1}^{n} \log \left[\frac{1}{n} \sum_{j \neq i} \phi_h \left(y_i - \mathbf{x}_i^T \boldsymbol{\beta} - y_j - \mathbf{x}_j^T \boldsymbol{\beta} \right) \right], \tag{3.2}$$

where $\phi_h(t)$ is a normal kernel with bandwidth h. One might also use some other kernels. However, it is well known that the choice of kernel is not crucial. The Gaussian kernel is used for the simplicity of computation.

Yuan and De Gooijer (2007) used the Newton-Raphson algorithm to maximize (3.2). However, the Newton-Raphson algorithm requires to calculate first and second derivatives

of (3.2). In addition, it cannot guarantee to converge. Moreover the found solution by the Newton-Raphson algorithm can even be a local minima.

Note that the above objective function (3.2) has the mixture form (1.2). Therefore, we can use the proposed GMEM algorithm to maximize (3.2). Assuming that $\mathbf{x}$ doesn't contain the intercept term, then given the the initial value $\boldsymbol{\beta}^{(0)}$, in the $(k+1)^{\text{th}}$ step,
E Step: Let

$$\pi_{ij}^{(k+1)} = \frac{\phi_h\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)} - y_j - \mathbf{x}_j^T\boldsymbol{\beta}^{(k)}\right)}{\sum_{j\neq i}\phi_h\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)} - y_j - \mathbf{x}_j^T\boldsymbol{\beta}^{(k)}\right)},\ 1 \leq i \neq j \leq n\,.$$

M Step: Update

$$\begin{aligned}\boldsymbol{\beta}^{(k+1)} =& \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n}\sum_{j\neq i} \pi_{ij}^{(k+1)} \log\left\{\phi_h\left(y_i - \mathbf{x}_i^T\boldsymbol{\beta} - y_j - \mathbf{x}_j^T\boldsymbol{\beta}\right)\right\}\\ =&(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}\,,\end{aligned}$$

where $\mathbf{X} = (\mathbf{x}_1 - \mathbf{x}_2, \ldots, \mathbf{x}_1 - \mathbf{x}_n, \ldots, \mathbf{x}_n - \mathbf{x}_{n-1})^T, \mathbf{Y} = (y_1 - y_2, \ldots, y_n - y_{n-1})^T$, and $\mathbf{W} = \text{diag}\{\pi_{12}^{(k+1)}, \ldots, \pi_{n,n-1}^{(k+1)}\}$.

The idea of the above GMEM algorithm can be also applied to find adaptive nonlinear regression estimator if the linear regression function in (3.1) is replaced by a parametric nonlinear regression function.

3.2 Adaptive nonparametric regression

Suppose that $(x_1, y_1), \ldots, (x_n, y_n)$ are an independent and identically distributed random sample from

$$y = m(x) + \epsilon,$$

where $\mathrm{E}(\epsilon \mid X = x) = 0$, $\mathrm{var}(\epsilon \mid X = x) = \sigma^2(x)$, and $m(\cdot)$ is an unknown nonparametric smoothing function to be estimated. Local polynomial regression (Fan and Gijbels, 1996) is to locally approximate $m(x)$ by a polynomial function. That is, for x in a neighborhood of x_0, we approximate

$$m(x) \approx \sum_{l=0}^{p} \frac{m^{(l)}(x_0)}{l!}(x - x_0)^l \equiv \sum_{l=0}^{p} \beta_l (x - x_0)^l,$$

where $\beta_l = m^{(l)}(x_0)/l!$. Then the local polynomial regression estimates local parameter $\boldsymbol{\theta} = (\beta_0, \ldots, \beta_p)$ by minimizing the following weighted least squares

$$\sum_{i=1}^{n} K_h(x_i - x_0) \left\{ y_i - \sum_{l=0}^{p} \beta_l (x_i - x_0)^l \right\}^2, \tag{3.3}$$

where $K_h(t) = h^{-1}K(t/h)$, a rescaled kernel function of $K(t)$ with a bandwidth h. The above least squares based local polynomial regression estimator will lose some efficiency if the error density is not normal.

Linton and Xiao (2007) proposed to adaptively estimate the local parameter $\boldsymbol{\theta}$ by maximizing the estimated local log-likelihood

$$\sum_{i=1}^{n} K_h(x_i - x_0) \log \left\{ \tilde{f}(y_i - \sum_{l=0}^{p} \beta_l (x_i - x_0)^l) \right\}, \tag{3.4}$$

where $\tilde{f}$ is a kernel density estimator of error term ϵ

$$\tilde{f}(\epsilon_i) = \frac{1}{n} \sum_{j \neq i}^{n} \phi_g \left(\epsilon_i - \tilde{\epsilon}_j \right),$$

where $\tilde{\epsilon}_j = y_j - \tilde{m}(x_j)$ is the residual based on some initial estimator $\tilde{m}(\cdot)$ (such as the least squared based local polynomial regression estimator), and $\phi_g(\cdot)$ is the normal kernel with bandwidth g. Linton and Xiao (2007) proved that the above adaptive nonparametric

regression estimator has the asymptotic "oracle" property, i.e., it has the same asymptotic efficiency as the local log-likelihood estimator assuming $f(\cdot)$ were known.

Linton and Xiao (2007) proposed to use Newton-Raphson method or one-step Newton-Raphson method to maximize (3.4). Note that (3.4) has the mixture form (1.2). Therefore, we can apply the proposed GMEM algorithm to maximize (3.4): given the the initial value $\boldsymbol{\theta}^{(0)}$, in the $(k+1)^{\text{th}}$ step,

E Step: Let

$$\pi_{ij}^{(k+1)} = \frac{K_h(x_i - x_0)\phi_g\left(y_i - \sum_{l=0}^{p}\beta_l^{(k)}(x_i - x_0)^l - \tilde{\epsilon}_j\right)}{\sum_{j\neq i} K_h(x_i - x_0)\phi_g\left(y_i - \sum_{l=0}^{p}\beta_l^{(k)}(x_i - x_0)^l - \tilde{\epsilon}_j\right)},\ 1 \le i \neq j \le n\,.$$

M Step: Update

$$\begin{aligned}\boldsymbol{\theta}^{(k+1)} =& \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{n}\sum_{j\neq i} \pi_{ij}^{(k+1)} K_h(x_i - x_0)\log\left\{\phi_g\left(y_i - \sum_{l=0}^{p}\beta_l(x_i - x_0)^l - \tilde{\epsilon}_j\right)\right\}\\ =&(\mathbf{X}^T\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y}\,,\end{aligned}$$

where $\mathbf{Y} = (y_1 - \tilde{\epsilon}_2, \ldots, y_1 - \tilde{\epsilon}_n, \ldots, y_n - \tilde{\epsilon}_{n-1})^T$, $\mathbf{W} = \text{diag}\{\pi_{12}^{(k+1)}K_h(x_1 - x_0), \ldots, \pi_{1n}^{(k+1)}K_h(x_1 - x_0), \ldots, \pi_{n,n-1}^{(k+1)}K_h(x_n - x_0)\}$, and $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, with

$$\mathbf{x}_i = \begin{pmatrix} 1 & 1 & \ldots & 1 \\ (x_i - x_0) & (x_i - x_0) & \ldots & (x_i - x_0) \\ \ldots & \ldots & \ldots & \\ (x_i - x_0)^p & (x_i - x_0)^p & \ldots & (x_i - x_0)^p \end{pmatrix}_{(p+1)\times(n-1)}\,.$$

The idea of the above GMEM algorithm can be also applied to find adaptive estimator for some other nonparametric or semiparametric regression models, such as varying coefficient models (Cleveland, Grosse, and Shyu, 1992) and varying-coefficient partially linear model (Zhang, Lee, and Song, 2002; Xia, Zhang, and Tong, 2004; Fan and Huang, 2005).

3.3 Mode detection

Given the observation $(x_1, \ldots, x_n)$ from the population X with density $f(x)$, suppose we want to estimate the mode of $f(x)$. Parzen (1962) and Eddy (1980) proposed to estimate the model of $f(x)$ by maximizing the kernel density estimator of $f(x)$

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} \phi_h(x_i - x). \tag{3.5}$$

Here x can be a scalar or vector. If x is a vector, then we use multivariate normal kernel.

Note that the above kernel density estimator has the mixture form (1.3) ((1.2) with $K = 1$). We can apply GMEM to find the mode of $\hat{f}(x)$: given the initial value $x^{(0)}$, in the $(k+1)^{\text{th}}$ step,

E Step: Let

$$\pi(j \mid x^{(k)}) = \frac{\phi_h(x_j - x^{(k)})}{\sum_{i=1}^{n} \phi_h(x_i - x^{(k)})},\ j = 1, \ldots, n\,.$$

M Step: Update

$$x^{(k+1)} = \arg\max_x \sum_{j=1}^{n} \pi(j \mid x^{(k)}) \log\{\phi_h(x_j - x)\} = \sum_{j=1}^{n} \pi(j \mid x^{(k)}) x_j\,.$$

From the above algorithm, we can see that the estimated mode can be also considered as a weighted average of the observations and the weights depend on the distance between each observation and the mode.

Note that the kernel density estimator (3.5) can be considered as a mixture density with n components. Li, Ray, and Lindsay (2007) has successfully applied the MEM algorithm to do nonparametric clustering by locating the local modes of kernel density (3.5) when starting from each observation, assuming that the observations converged to the same mode are in the same cluster.

3.4 Edge-preserving smoothers for image processing and robust nonparametric regression

Chu, et al. (1998) proposed an edge-preserving smoothers for image processing. Suppose that $(x_1, y_1), \ldots, (x_n, y_n)$ are an independent and identically distributed random sample from

$$y = m(x) + \epsilon,$$

where $\mathrm{E}(\epsilon \mid X = x) = 0$, $\mathrm{var}(\epsilon \mid X = x) = \sigma^2$, and $m(\cdot)$ is an unknown smoothing function except for *some jump discontinuities*. The focus is to estimate $m(\cdot)$ at data points, i.e., $m(x_1), \ldots, m(x_n)$. The traditional nonparametric smoothers have limited usefulness in image processing, because sharp "edges" tend to be blurred. The edge-preserving smoother of $m(x_i)$ proposed by Chu, et al. (1998) is the local maximizer of

$$S(\theta) = \sum_{j=1}^{n} \phi_g(y_j - \theta)\phi_h(x_i - x_j), \tag{3.6}$$

when starting from y_i, where ϕ_h is Gaussian kernel with bandwidth h.

Note that (3.6) has the mixture type objective function (1.2) with $K = 1$. Therefore, we can apply GMEM algorithm to maximize (3.6) to estimate $m(x_i)$: given the the initial value $\theta^{(0)} = y_i$, in the $(k+1)^{\text{th}}$ step,

E Step: Let

$$\pi(j \mid \theta^{(k)}) = \frac{\phi_g(y_j - \theta^{(k)})\phi_h(x_i - x_j)}{\sum_{j=1}^{n} \phi_g(y_j - \theta^{(k)})\phi_h(x_i - x_j)},\ j = 1, \ldots, n\,.$$

M Step: Update

$$\theta^{(k+1)} = \arg\max_{\theta} \sum_{j=1}^{n} \pi(j \mid \theta^{(k)}) \log\{\phi_g(y_j - \theta)\} = \sum_{j=1}^{n} \pi(j \mid \theta^{(k)}) y_j\,.$$

Note that the edge-preserving smoother (3.6) has similar form of local M estimator for

any fixed g. Therefore, the above GMEM algorithm can be also applied to produce robust nonparametric regression. However, unlike the traditional local M estimator, Chu, et al. (1998) proved that the conditions $g \to 0$ and $h \to 0$ are required in order to get edge-preserving result.

3.5 Robust generalized M estimator for linear regression

Suppose $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ are sampled from the regression model

$$y = \mathbf{x}^T \boldsymbol{\beta} + \epsilon, \tag{3.7}$$

where $\mathbf{x}$ is a p-dimensional vector of covariates independent of the error ϵ with $\mathrm{E}(\epsilon) = 0$. Traditionally, $\boldsymbol{\beta}$ is estimated by least squares estimate (LSE)

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2. \tag{3.8}$$

However, it is well known that the LSE is very sensitive to outliers. Many robust regression methods have been proposed. One of the commonly used robust regression methods is M-estimator (Huber, 1981; Andrews, 1974; Hampel, 1986), which replaces the square loss in (3.8) by some robust loss function $\rho(\cdot)$, i.e., estimates $\boldsymbol{\beta}$ by minimizing

$$\sum_{i=1}^{n} \rho(y_i - \mathbf{x}_i^T \boldsymbol{\beta}). \tag{3.9}$$

The above robust M estimator regression works well if there are only outliers in y direction. However, it is well known that the M estimator regression does not work well if there are high leverage outliers and in fact has zero breakdown point (Maronna, Martin, and Yohai, 2006).

Generalized M estimators (GM estimators for short) (Hampel, et al., 1986) are an im-

portant class of robust regression estimators which can deal with the high leverage outliers. The GM estimators find $\boldsymbol{\beta}$ by minimizing

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} w(\mathbf{x}_i)\rho(y_i - \mathbf{x}_i^T\boldsymbol{\beta}), \tag{3.10}$$

where $w(\cdot)$ is a weight function used to down-weight the high leverage points. Here we mainly consider the redescending function $\rho'(\cdot)$, since they completely reject gross outliers, while the Huber estimator effectively treats these the same as moderate outliers. In addition, the redescending M-estimators are about 20% more efficient than the Huber estimator for the Cauchy distribution. Based on Chu, et al. (1998), minimizing (3.10) is equivalent to maximizing

$$Q(\boldsymbol{\beta}) = \sum_{i=1}^{n} w(\mathbf{x}_i)K_h(y_i - \mathbf{x}_i^T\boldsymbol{\beta}), \tag{3.11}$$

where K_h is a kernel density. Note that (3.11) has the mixture form (1.2) with $K = 1$. Therefore, we can use GMEM to maximize (3.11): given the the initial value $\boldsymbol{\beta}^{(0)}$, in the $(k+1)^{\text{th}}$ step,

E Step: Let

$$\pi(j \mid \boldsymbol{\beta}^{(k)}) = \frac{w(\mathbf{x}_j)K_h(y_j - \mathbf{x}_j^T\boldsymbol{\beta}^{(k)})}{\sum_{i=1}^{n} w(\mathbf{x}_i)K_h(y_i - \mathbf{x}_i^T\boldsymbol{\beta}^{(k)})},\ j = 1, \ldots, n\,.$$

M Step: Update

$$\boldsymbol{\beta}^{(k+1)} = \arg\max_{\boldsymbol{\beta}} \sum_{j=1}^{n} \pi(j \mid \boldsymbol{\beta}^{(k)}) \log[K_h(y_j - \mathbf{x}_j^T\boldsymbol{\beta})]\,.$$

If $K_h(\cdot)$ is taken as a Gaussian kernel, as used by Chu, et al. (1998), then M step has an explicit form

$$\boldsymbol{\beta}^{(k+1)} = (\mathbf{X}^T\mathbf{W}^{(k)}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{W}\mathbf{Y},$$

where $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^T, \mathbf{Y} = (y_1, \ldots, y_n)^T$, and $\mathbf{W}^{(k)} = \text{diag}\{\pi(1 \mid \boldsymbol{\beta}^{(k)}), \ldots, \pi(n \mid$

$\boldsymbol{\beta}^{(k)})\}$.

Note that the above GMEM still applies if the weight function $w(\cdot)$ in (3.11) also depends on the response variable y, such as the initial estimated residuals.

4 Discussion

Note that in (1.3) when $K = 1$, the condition that $f_{1l}(x)$ is a positive function can be in fact relaxed to the condition that $f_{1l}(x)$ is bounded below such that $f_{1l}(x) + c_l > 0$ for all x for some $c_l > 0$. Then, maximizing $\sum_{l=1}^{L} a_{1l} f_{1l}(x)$ is equivalent to maximizing $\sum_{l=1}^{L} a_{1l}(f_{1l}(x) + c_l)$.

Note that the proposed algorithm still belongs to the bigger class of EM algorithm. Therefore, all the properties of EM algorithm (McLachlan and Krishnan, 2008) also apply to the GMEM algorithm proposed in this article.

In this article, we just mentioned some statistical applications of equation (1.2) that we are aware of. It requires more research to explore other statistical applications of (1.2) besides the ones mentioned in this article.

Acknowledgements

The author is grateful to the editors and the referee for their insightful comments and suggestions, which greatly improved this article.

Appendix

Proof of Theorem 2.1: Let $Y_k^{(t+1)}$ be a discrete random variable such that

$$P\left(Y_k^{(t+1)} = \frac{f_{kl}(x^{(t+1)})}{f_{kl}(x^{(t)})}\right) = \frac{a_{kl} f_{kl}(x^{(t)})}{\sum_{l=1}^{L} a_{kl} f_{kl}(x^{(t)})} \triangleq \pi_{kl}^{(t+1)}, l = 1, \ldots, m.$$

Then,

$$\begin{aligned} f(x^{(t+1)}) - f(x^{(t)}) &= \sum_{k=1}^{K} w_k \log \left\{ \frac{\sum_{l=1}^{L} a_{kl} f_{kl}(x^{(t+1)})}{\sum_{l=1}^{L} a_{kl} f_{kl}(x^{(t)})} \right\} \\ &= \sum_{k=1}^{K} w_k \log \left\{ \sum_{l=1}^{L} \frac{a_{kl} f_{kl}(x^{(t)})}{\sum_{l=1}^{L} a_{kl} f_{kl}(x^{(t)})} \frac{a_{kl} f_{kl}(x^{(t+1)})}{a_{kl} f_{kl}(x^{(t)})} \right\} \\ &= \sum_{k=1}^{K} w_k \log \left\{ \sum_{l=1}^{L} \pi_{kl}^{(t+1)} \frac{f_{kl}(x^{(t+1)})}{f_{kl}(x^{(t)})} \right\} \\ &= \sum_{k=1}^{K} w_k \log \left\{ \mathrm{E}\left(Y_k^{(t+1)} \right) \right\}. \end{aligned}$$

Based on Jensen's inequality, we have

$$\begin{aligned} f(x^{(t+1)}) - f(x^{(t)}) &\geq \sum_{k=1}^{K} w_k \mathrm{E}\left\{ \log(Y^{(k+1)}) \right\} \\ &= \sum_{k=1}^{K} w_k \sum_{l=1}^{L} \pi_{kl}^{(t+1)} \log \frac{f_{kl}(x^{(t+1)})}{f_{kl}(x^{(t)})} \\ &= \sum_{k=1}^{K} \sum_{l=1}^{L} w_k \pi_{kl}^{(t+1)} \log \frac{f_{kl}(x^{(t+1)})}{f_{kl}(x^{(t)})}. \end{aligned}$$

The equality occurs if and only if $f_{kl}(x^{(t+1)})/f_{kl}(x^{(t)})$ are the same for all ls given any $k = 1, \ldots, K$. Based on the property of the M-step of (2.2), we have

$$\sum_{k=1}^{K} \sum_{l=1}^{L} w_k \pi_{kl}^{(t+1)} \log\{ f_{kl}(x^{(t+1)}) \} \geq \sum_{k=1}^{K} \sum_{l=1}^{L} w_k \pi_{kl}^{(t+1)} \log\{ f_{kl}(x^{(t)}) \}.$$

Therefore,

$$f(x^{(t+1)}) - f(x^{(t)}) \geq 0.$$

References

Andrews, D.F. (1974). A robust method for multiple linear regression. *Technometrics*, 16, 523-531.

Böhning, D. (1999). *Computer-Assisted Analysis of Mixtures and Applications*, Boca Raton, FL: Chapman and Hall/CRC.

Chu, C. K., Glad, I., Godtliebsen, F., and Marron, J. S. (1998). Edge-preserving smoothers for image processing (with discussion). *Journal of the American Statistical Association*, 93, 526-556.

Cleveland, W. S., Grosse, E., and Shyu, W. M. (1992). *Local Regression Models*. In Statistical Models in S (ed. J.M. Chambers and T.J. Hastie), pp. 309-376. Pacific Grove: Wadsworth & Brooks.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of Royal Statistical Society, Ser B.*, 39, 1-38.

Eddy, W. F. (1980). Optimum kernel estimators of the mode. *Annals of Statistics*, 8, 870-882.

Fan, J. and Huang, T. (2005). Profile likelihood inferences on semiparametric varying-coefficient partially linear models. *Bernoulli*, 11, 1031-1057.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*, Springer, 2006.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Li, J., Ray, S., and Lindsay, B. G. (2007). A nonparametric statistical approach to clustering via mode identification. *Journal of Machine Learning Research*, 8, 1687-1723.

Lindsay, B. G., (1995). *Mixture Models: Theory, Geometry, and Applications*. NSF-CBMS Regional Conference Series in Probability and Statistics v 5, Hayward, CA: Institure of Mathematical Statistics.

Linton, O. and Xiao, Z. (2007). A nonparametric regression estimator that adapts to error distribution of unknown form. *Econometric Theory*, 23, 371-413.

Maronna, R. A., Martin, R. D., and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. Wiley-Interscience.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. New York: Wiley.

Parzen, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33, 1065-1076.

Xia, Y., Zhang, W., and Tong, H. (2004). Efficient estimation for semivarying-coefficient models. *Biometrika*, 91, 661-681.

Yuan, A. and De Gooijer, J. G. (2007). Semiparametric regression with kernel error model. *Scandinavian Journal of Statistics*, 34, 841-869.

Zhang, W., Lee, S. Y., and Song, X. (2002). Local polynomial fitting in semivarying coefficent models. *Journal of Multivariate Analysis*, 82, 166-188.