Appl. Math. J. Chinese Univ.
2019, 34(4): 393-402

The EM algorithm for ML Estimators under nonlinear inequalities restrictions on the parameters

SHEN Qi-xia[1] MIAO Peng[1] LIANG Yin-shuang[2,*]

Abstract. One of the most powerful algorithms for obtaining maximum likelihood estimates for many incomplete-data problems is the EM algorithm. However, when the parameters satisfy a set of nonlinear restrictions, It is difficult to apply the EM algorithm directly. In this paper, we propose an asymptotic maximum likelihood estimation procedure under a set of nonlinear inequalities restrictions on the parameters, in which the EM algorithm can be used. Essentially this kind of estimation problem is a stochastic optimization problem in the M-step. We make use of methods in stochastic optimization to overcome the difficulty caused by nonlinearity in the given constraints.

§1 Introduction

One of the most powerful algorithms for obtaining maximum likelihood estimates for many incomplete-data problems is the EM algorithm. Since the EM algorithm is computationally simple and numerically stable, it is used for a broad range of applications, such as Analysis of variance component models in normal data, finite mixture models and multivariate normal models with missing data [1, 7], Gaussian copula with missing data[17] , Robust Gaussian process modeling[14]. In the EM algorithm it is usually necessary to find the conditional distribution in the E-step, then the standard maximum likelihood estimation for the complete-data problem is used in the M-step. Wu[18] showed the convergence properties of EM sequence. Many statistician introduced the extensions of the EM algorithm in their papers. Meng and Rubin [10] discussed a kind of generalized EM algorithms which they called the Expectation-Conditional Maximization (ECM) algorithms. They took advantage of the simplicity of complete-data conditional maximization by replacing a complicated M-step of the EM algorithm with several

Received: 2016-08-02. Revised: 2019-06-11.
MR Subject Classification: 35B40, 62H12, 62J12.
Keywords: Linear regression, Maximum likelihood estimation, Nonlinear constraints, Asymptotic properties.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-019-3484-9.
Supported by Teaching reform project of Zhengzhou University of Science and Technology (KFCZ201909), National Foundation for Cultivating Scientific Research Projects of Zhengzhou Institute of Technology (GJJKTPY2018K4), Henan Big Data Double Base of Zhengzhou Institute of Technology (2017410154650302265) and the Key Scientific Research Foundation of Education Bureau of Henan Province (20B110020).
∗ Corresponding author.

computationally simpler CM-steps. Little and Rubin [8] gave a generalization of the ECM algorithm which replaced some of the CM-steps with steps that maximized the constrained actual (incomplete-data) log-likelihood. And they called this algorithm the Expectation-conditional Maximization Either (ECME) algorithm. This algorithm shares with both the EM and the ECM algorithms' stable monotone convergence and basic simplicity of implementation relative to faster converging competitors.

When there are no restrictions on the parameters, each step of the EM algorithm is usually simple and straightforward. But when the parameters must satisfy a set of linear or nonlinear restrictions, the M-step will usually involve complicated procedures to find the solutions, and no closed form may exist, and in this case the constrained optimization routines are needed. The constrained optimization problems have attracted many researchers. For example, Liew[6] considered linear regression with linear constraints. Nagaraj and Fuller [11] studied linear time series subject to nonlinear equality constraints. Eicker [3] studied the asymptotic normality and consistency of the least squares estimators for families of linear regression. Nyquist[12] proposed iteratively reweighted least squares to estimate parameters under a set of linear restrictions and applied the method to generalized linear models. Wang[16] considered the asymptotics of least-squares estimators for constrained nonlinear regression.

Kim and Taylor[5] proposed a modification to the EM algorithm that incorporates linear equalities restrictions on the parameters. Shi, Zheng and Guo[15] studied the incomplete-data problem for the case that the parameters were restricted on a linear subspace. They proposed a restricted EM algorithm to find MLEs under the linear inequalities restrictions in which projection algorithm was used in the M-step. However, when parameters are restricted by some nonlinear inequalities, the incomplete-data approach has important applications in many clinical experiments, agricultural research, public-opinion polls and so on. In this article, we construct the restricted EM algorithm for maximum likelihood estimation under nonlinear inequalities restrictions on the parameters, and investigate the asymptotic behavior of the maximum likelihood (ML) estimators in linear regression problems.

First, we give some results about the EM algorithm, consider the linear model $Y = X\beta + e$, following the notation of Kim and Taylor[5], denote $Y = (Y_{mis}, Y_{obs})$, where Y is the complete-data, Y_{mis} and Y_{obs} represent the missing part and the observed components of Y. The likelihood function of Y can be written as

$$f(Y|\beta) = f(Y_{obs}, Y_{mis}|\beta) = f(Y_{obs}|\beta)f(Y_{mis}|Y_{obs}, \beta), \tag{1}$$

where β is the unknown parameter, $f(Y_{obs}|\beta)$ denotes the observed likelihood and $f(Y_{mis}|Y_{obs}, \beta)$ denotes the conditional likelihood given Y_{obs} and β. Then the log-likelihood is given by

$$l(\beta|Y_{obs}) = l(\beta|Y) - \ln f(Y_{mis}|Y_{obs}, \beta), \tag{2}$$

where $l(\beta|Y_{obs}) = \ln f(Y_{obs}|\beta)$, $l(\beta|Y) = \ln f(Y|\beta)$. Let $\{\beta_n^{(m)}\}$ be an iteration sequence of the EM algorithm and denote

$$Q(\beta|\beta_n^{(m)}) = \int l(\beta|Y)f(Y_{mis}|Y_{obs}, \beta_n^{(m)})dY_{mis},$$

$$H(\beta|\beta_n^{(m)}) = \int \ln f(Y_{mis}|Y_{obs}, \beta)f(Y_{mis}|Y_{obs}, \beta_n^{(m)})dY_{mis}.$$

Then $l(\beta|Y_{obs})=Q(\beta|\beta_n^{(m)})-H(\beta|\beta_n^{(m)})$ and

$$\begin{aligned}&l(\beta_n^{(m+1)}|Y_{obs})-l(\beta_n^{(m)}|Y_{obs})\\ =\ &Q(\beta_n^{(m+1)}|Y_{obs})-Q(\beta_n^{(m)}|\beta_n^{(m)})-[H(\beta_n^{(m+1)}|\beta_n^{(m)})-H(\beta_n^{(m)}|\beta_n^{(m)})].\end{aligned}$$

From $H(\beta_n^{(m+1)}|\beta_n^{(m)})-H(\beta_n^{(m)}|\beta_n^{(m)})\le 0$(Kim and Taylor[5]; Shi, Zheng and Guo [15]), we have

$$l(\beta_n^{(m+1)}|Y_{obs})-l(\beta_n^{(m)}|Y_{obs})\ge Q(\beta_n^{(m+1)}|\beta_n^{(m)})-Q(\beta_n^{(m)}|\beta_n^{(m)}).$$

If $Q(\beta_n^{(m+1)}|\beta_n^{(m)})=\max_\beta Q(\beta|\beta_n^{(m)})$, then $l(\beta_n^{(m+1)}|Y_{obs})\ge l(\beta_n^{(m)}|Y_{obs})$.

The above result implies that the observed likelihood function increases in each step. This is a property of the EM algorithm that if $\beta_n^{(m+1)}$ is chosen to increase $Q(\beta|\beta_n^{(m)})$ with respect to β which under a set of restrictions, this will ensure that the log-likelihood under the set of restrictions on the parameters also increase(Kim and Taylor[5]).

§2 The restricted EM algorithm

The restricted regression we are facing is of the following form:

$$\begin{cases} Y_i &= X_i\beta+e_i \quad i=1,\dots,n,\\ h_j(\beta)\ge 0, & j=1,\dots,r,\\ g_j(\beta)=0, & j=r+1,\dots,q,\end{cases}\tag{3}$$

where $Y_i=(Y_{i1},\dots,Y_{ik})'$, $i=1,\dots,n$ are random samples, and Y_{ij} is the jth variate of Y_i. When Y_i cannot be fully observed, $Y_i's$ missing data vector $Y_{i(mis)}$ and observed data vector $Y_{i(obs)}$ exist (when Y_i is completely observed, $Y_{i(mis)}$ is absent and $Y_{i(obs)}$ is Y_i). Let $Y=(Y_1',\dots,Y_n')$, Y_{mis} and Y_{obs} denote $Y's$ missing data vector $(Y'_{1(mis)},\dots,Y'_{n(mis)})$ and observed data vector $(Y'_{1(obs)},\dots,Y'_{n(obs)})$. X_i is $k\times p$ matrix, and $rank(X_i)=p$, $\beta\in\Omega\subseteq R^p$ is the unknown parameter to be estimated, where Ω is a convex compact subset of R^p, $e_i=(e_{i1},\dots,e_{ik})'$ is error vector and normally distributed with mean zero and known covariance matric $\Sigma_i>0$, $h_j, j=1,\dots,r$; $g_j, j=r+1,\dots,q$ are continuously differentiable functions in Ω.

Now, we consider the linear model (3), the log-likelihood function

$$l(\beta|Y)=-\frac{1}{2}\sum_{i=1}^n(Y_i-X_i\beta)'\Sigma_i^{-1}(Y_i-X_i\beta)+C_0,$$

where C_0 is a constant which does not depend on the unknown parameter β. At first, we compute the conditional expectation, for convenience, substitute $\sum_{i=1}^n$ with $\sum$.

$$\begin{aligned}Q(\beta|\beta_n^{(m)}) &= -\tfrac{1}{2}\int\sum(Y_i-X_i\beta)'\Sigma_i^{-1}(Y_i-X_i\beta)f(Y_{mis}|Y_{obs},\beta_n^{(m)})dY_{mis}+C_0\\ &= -\tfrac{1}{2}\sum E(Y_i'\Sigma_i^{-1}Y_i|Y_{obs},\beta_n^{(m)})+\sum E(\beta'X_i'\Sigma_i^{-1}Y_i|Y_{obs},\beta_n^{(m)})\\ &\quad -\tfrac{1}{2}\sum\beta'X_i'\Sigma_i^{-1}X_i\beta+C_0.\end{aligned}$$

Removing β-independent terms, let $\overline{Q}(\beta|\ \beta_n^{(m)})=\beta'\sum X_i'\Sigma_i^{-1}X_i\beta-2\beta'\sum X_i'\Sigma_i^{-1}$

$E(Y_i|Y_{obs}, \beta_n^{(m)})$, then the restricted maximum likelihood problem is equivalent to

$$\begin{cases} \min \overline{Q}(\beta|\ \beta_n^{(m)}) = \beta' \sum X_i'\Sigma_i^{-1}X_i\beta - 2\beta' \sum X_i'\Sigma_i^{-1}E(Y_i|Y_{obs}, \beta_n^{(m)}) \\ \qquad h_j(\beta) \geq 0, \quad j = 1, \ldots, r, \\ \qquad g_j(\beta) = 0, j = r+1, \ldots, q. \end{cases} \tag{4}$$

Let $\beta_n^{(m+1)}$ be the optimal solution to problem (4) and β_0 be the true value of β in model:

$$\begin{cases} \quad Y_i \quad = X_i\beta + e_i \quad i = 1, \ldots, n, \\ h_j(\beta) = 0, \quad j = 1, \ldots, r, \\ g_j(\beta) = 0, \quad j = r+1, \ldots, q, \end{cases}$$

β_0 can be estimated by restricted EM algorithm under equality.

Let $\theta = n^{\frac{1}{2}}(\beta - \beta_0)$, which is often used in the statistical literature, for example, in Prakasa[13], in Wang [16] and in Liu and Wang[9]. Then we'll get the ML estimation of β from the ML estimation of θ. Substituting θ into (4), we get

$$\begin{cases} \min \quad \overline{Q}(\theta|\theta_n^{(m)}) = \theta'[n^{-1} \sum X_i'\Sigma_i^{-1}X_i]\theta - 2\theta' n^{-\frac{1}{2}} \sum X_i'\Sigma_i^{-1}[E(Y_i|Y_{obs}, \theta_n^{(m)}) - \\ \qquad X_i\beta_0] + \beta_0' \sum X_i'\Sigma_i^{-1}X_i\beta_0 - 2\beta_0' \sum X_i\Sigma_i^{-1}E(Y_i|Y_{obs}, \theta_n^{(m)}) \\ h_j(n^{-\frac{1}{2}}\theta + \beta_0) \geq 0 \quad j = 1, \ldots, r, \\ g_j(n^{-\frac{1}{2}}\theta + \beta_0) = 0, \quad j = r+1, \ldots, q, \end{cases} \tag{5}$$

where $\theta_n^{(m)} = n^{\frac{1}{2}}(\beta_n^{(m)} - \beta_0)$.

Remove the amount of θ-independent, let

$$\widetilde{Q}_n(\theta|\theta_n^{(m)}) = \theta'[n^{-1} \sum X_i'\Sigma_i^{-1}X_i]\theta - 2\theta' n^{-\frac{1}{2}} \sum X_i'\Sigma_i^{-1}[E(Y_i|Y_{obs}, \theta_n^{(m)}) - X_i\beta_0],$$

and S_n be the objective function and the feasible solution set of problem (5), respectively. Then the problem (5) is equivalent to

$$\begin{cases} \min \quad \widetilde{Q}_n(\theta|\theta_n^{(m)}) = \theta'[n^{-1} \sum X_i'\Sigma_i^{-1}X_i]\theta - 2\theta' n^{-\frac{1}{2}} \sum X_i'\Sigma_i^{-1}[E(Y_i|Y_{obs}, \theta_n^{(m)}) - \\ \qquad X_i\beta_0] \\ h_j(n^{-\frac{1}{2}}\theta + \beta_0) \geq 0 \quad j = 1, \ldots, r, \\ g_j(n^{-\frac{1}{2}}\theta + \beta_0) = 0, \quad j = r+1, \ldots, q, \end{cases} \tag{6}$$

Assume the optimal solution of (6) exists, denote it by $\theta_n^{(m+1)}$, then $\theta_n^{(m+1)} = n^{\frac{1}{2}}(\beta_n^{(m+1)} - \beta_0)$. Due to the appearance of the constraints, especially of the nonlinear inequality constraints, one can not expect to get the actual restricted ML estimation of the parameter of θ as in the unconstrained regression problems. In this paper, we firstly find the asymptotic distribution of $\theta_n^{(m+1)}$. Then, we consider the limit form of problem (6) in the following:

2.1 The asymptotic property of the objective function $\widetilde{Q}_n(\theta|\theta_n^{(m)})$

Let W be a neighborhood of β_0 and such that for β in W it holds that

$$X_i\beta = X_i\beta_0 + (\nabla_\beta(X_i\beta))'(\beta - \beta_0) + r_i(\beta)(||\beta - \beta_0||)^2$$

Where $\nabla_\beta(X_i\beta)$ is the gradient vector of $X_i\beta$ with repect to β at β_0, $||.||$ denotes the Euclidean norm in R^P and $r_i(\beta) = (r_{i1}(\beta), r_{i2}(\beta), ..., r_{ip}(\beta))'$ satisfies

$$\lim_{n\to\infty} n^{-1}\Sigma r_i'(\beta)r_i(\beta) < \infty.$$

Uniformly on W.

For finding the limit form of the objective function $\widetilde{Q}_n(\theta|\theta_n^{(m)})$, we have the following theorem.

Theorem 2.1. *Let* $T^{(m)} = \lim_{n\to\infty} n^{-\frac{1}{2}} \sum X_i'\Sigma_i^{-1}[E(Y_i|Y_{obs}, \theta_n^{(m)}) - X_i\beta_0]$, $K = \lim_{n\to\infty} \frac{1}{n} \sum X_i'\Sigma_i^{-1}X_i$, *suppose* $T^{(m)}$ *and* K *exists and* K *is a positive-definite matrix. Then for each* $\theta \in \Omega$, $\widetilde{Q}_n(\theta|\theta_n^{(m)})$ *converges in distribution to* $\theta' K\theta - 2\theta' T^{(m)}$.

Proof. For any β in W we have $\theta = n^{\frac{1}{2}}(\beta - \beta_0)$ and

$$\begin{aligned}\widetilde{Q}_n(\theta|\theta_n^{(m)}) &= \theta'[n^{-1}\sum X_i'\Sigma_i^{-1}X_i]\theta - 2\theta' n^{-\frac{1}{2}} \sum X_i'\Sigma_i^{-1}[E(Y_i|Y_{obs}, \theta_n^{(m)}) - X_i\beta_0] \\ &= \theta'[n^{-1}\sum X_i'\Sigma_i^{-1}X_i]\theta - 2\theta' n^{-\frac{1}{2}} \sum X_i'\Sigma_i^{-1}E(e_i|Y_{obs}, \theta_n^{(m)}).\end{aligned}$$

By Theorem 1 in Wang[16],

$$n^{-\frac{1}{2}} \sum_{i=1}^{n} X_i{}'\Sigma_i^{-1} e_i \to_L N(0, K) \quad (n \to \infty).$$

Hence, for any fixed θ, we have $\widetilde{Q}_n(\theta|\theta_n^{(m)}) \to_L \theta' K\theta - 2\theta' T^{(m)} (n \to \infty)$. □

2.2 The asymptotic property of the feasible solution set S_n

We use the compact of convergence of sets in Kuratowsi's sense, because this kind of convergence of sets will lead to convergence of optimal solutions about the related programming problems. We write $S = (K)\lim S_n$, if

$$limsup S_n \subset S \subset liminf S_n,$$

where $limsup S_n = \{\theta : \exists \ \{\theta_{n_j}\}$ such that $\theta_{n_j} \in S_n$, and $\theta_{n_j} \to \theta\}$, $limsup S_n = \{\theta : \exists \ \{\theta_n\}$ such that $\theta_n \in S_n$, and $\theta_n \to \theta\}$. Then for any $\theta \in S$ there is a sequence $\{\theta_n\}$, $\theta_n \in S_n$ and $\theta_n \to \theta$, and for any sequence $\{\theta_n\}$ with $\theta_n \in S_n$ any accumulation point of $\{\theta_n\}$ must belong to S.

Lemma 2.1. *Suppose that* $h_j, j = 1, \ldots, r$ *and* $g_j, j = r+1, \ldots, q$ *are continuously differentiable functions in* Ω, *then as* $n \to \infty$ *we have*

$$S = \{\theta : \nabla h_j(\beta_0)'\theta \geq 0, \ j = 1, \ldots, r; \ \nabla g_j(\beta_0)'\theta = 0, \ j = r+1, \ldots, q\}.$$

where $\nabla h_j(\beta_0)(j = 1, \ldots, r)$ *and* $\nabla g_j(\beta_0)(j = r+1, \ldots, q)$ *are the gradient vectors of* $g_j(\beta)(j = 1, \ldots, r)$ *and* $h_j(\beta)(j = r+1, \ldots, q)$ *respectively with respect to* β *at* $\beta = \beta_0$.

(For proof one could see Theorem 2 of Wang[16]).

With Theorem 2.1 and Lemma 2.1, we can formulate a limit form of problem (6):

$$\begin{cases} \min & \ddot{Q}(\theta|T^{(m)}) = \theta' K\theta - 2\theta' T^{(m)} \\ & \theta \in S, \end{cases} \tag{7}$$

Let $\theta^{(m+1)}$ be the optimal solution of problem (7). Although for the objective function we have $\widetilde{Q}_n(\theta|\theta_n^{(m)}) \to_L \ddot{Q}(\theta|T^{(m)})(n \to \infty)$ for any fixed θ, it has not been shown that $\widetilde{Q}_n(\theta_n^{(m+1)}|\theta_n^{(m)}) \to_L \ddot{Q}(\theta^{(m+1)}|T^{(m)})(n \to \infty)$. When θ is varying over the connected set $D = \Omega \bigcap \{\theta : \ \|\theta\| \leq M, M > 0, \theta \subseteq R^p\}$, $\{\widetilde{Q}_n(\theta|\theta_n^{(m)}), \theta \in D\}$ and $\{\ddot{Q}(\theta|T^{(m)}), \theta \in D\}$ can be viewed as stochastic processes. We will study the convergence in distribution of the sequence of these stochastic processes in the next section. Now we propose the restricted EM algorithm:

Let $\beta^{(0)}$ be the starting point, and $\theta^{(m)} = n^{1/2}(\beta^{(m)} - \beta_0)$, $m = 0, 1, \cdots$.

E-step: Compute $\ddot{Q}(\theta|T^{(m)}) = \theta' K\theta - 2\theta' T^{(m)}$ from $\theta^{(m)}$.

M-step: Let $\beta^{(m+1)} = n^{-1/2}\theta^{(m+1)} + \beta_0$, where $\theta^{(m+1)}$ is the optimal solution of problem (7) by using Hildreth-D'Esopo algorithm (Hildreth[4], D'Esopo[2]).

§3 Convergence of the restricted EM algorithm

In this section, we discuss some convergence properties of the proposed algorithm. Denote the sequence obtained from the restricted EM algorithm by $\{\theta^{(m+1)}\}$, we'll prove $\theta_n^{(m+1)} \to_L \theta^{(m+1)} (n \to \infty)$, where $\theta_n^{(m+1)}$ and $\theta^{(m+1)}$ are the optimal solutions of (6) and (7) respectively. First, we give two lemmas (for proof one could see Wang[16]):

Lemma 3.1. *Under the assumptions made in Theorem 2.1, $\theta_n^{(m+1)}$ is bounded in probability.*

Lemma 3.2. *Let D be the ball in R^k with center $\theta = 0$ and radius $d > 0$. Suppose the assumptions in Theorem 2.1 hold true. Then the stochastic processes $\{\widetilde{Q}_n(\theta|\theta_n^{(m)}), \theta \in D\}$ converge in distribution to $\{\ddot{Q}(\theta|T^{(m)}), \theta \in D\}$.*

Then we get the following restricted optimal problems:

$$\begin{cases} \min \ \widetilde{Q}_n(\theta|\theta_n^{(m)}) \\ \theta \in S_n \bigcap D, \end{cases} \tag{8}$$

and

$$\begin{cases} \min \ \ddot{Q}(\theta|T^{(m)}) \\ \theta \in S \bigcap D. \end{cases} \tag{9}$$

Let $\theta_n^{(m+1)(D)}$ and $\theta^{(m+1)(D)}$ be the optimal solutions of problem (8) and (9), respectively. We'll prove $\widetilde{Q}_n(\theta_n^{(m+1)}|\theta_n^{(m)}) \to_L \ddot{Q}(\theta^{(m+1)}|T^{(m)})$ $(n \to \infty)$ and $\theta_n^{(m+1)} \to_L \theta^{(m+1)}(n \to \infty)$ in the following theorems.

Theorem 3.1. *Let $\theta_n^{(m+1)}$ and $\theta^{(m+1)}$ be the optimal solutions of (6) and (7) respectively, then $\widetilde{Q}_n(\theta_n^{(m+1)}|\theta_n^{(m)}) \to_L \ddot{Q}(\theta^{(m+1)}|T^{(m)})(n \to \infty)$.*

Proof. Note that the sample function of the stochastic processes $\{\widetilde{Q}_n(\theta|\theta_n^{(m)}),\ \theta \in D\}$ and $\{\ddot{Q}(\theta|\theta_n^{(m)}),\ \theta \in D\}$ are continuous functions on D. Let $C(D)$ be the space of all continuous functions over D whose metric is defined by

$$d(h_1, h_2) = \sup_{\theta \in D} |h_1(\theta) - h_2(\theta)|, \quad h_1, h_2 \in C(D).$$

Define mappings $H_n(.)$ and $H(.)$ on C(D) such that

$$H_n(f_n) = \min_{\theta \in S_n \bigcap D} f_n(\theta) = f_n(\theta_n^{(m+1)(D_f)}), \tag{10}$$

$$H(f) = \min_{\theta \in S \bigcap D} f(\theta) = f(\theta^{(m+1)(D_f)}), \tag{11}$$

for $f_n, f \in C(D)$, where S is the same as in (7), and $\theta_n^{(m+1)(D_f)}$ and $\theta^{(m+1)(D_f)}$ are optimal solution of $\min_{\theta \in S_n \bigcap D} f_n(\theta)$ and $\min_{\theta \in S \bigcap D} f(\theta)$. First we are going to show that

$$\lim_{n \to \infty} H_n(f_n) = H(f), \tag{12}$$

for any $f_n,\ f$ in $C(D)$ with $f_n \to f$ and f is such that (11) has a unique optimal solution.

Observe the convergence of f_n to f means that

$$\max_{\theta\in D}|f_n(\theta)-f(\theta)|\to 0,(n\to\infty), \tag{13}$$

this implies that for any $\theta_n\to\theta$

$$f_n(\theta_n)\to f(\theta),(n\to\infty). \tag{14}$$

Thus to show (12), it suffices to show that $\theta_n^{(m+1)(D_f)}\to\theta^{(m+1)(D_f)}(n\to\infty)$. We first show the following: if $\theta_n^{(m+1)(D_f)}$, $n=1,2,\ldots$ are optimal solutions of problem (10) and $\tilde{\theta}^{(m+1)}$ is an accumulation point of $\{\theta_n^{(m+1)(D_f)}\}$, then $\tilde{\theta}^{(m+1)}$ must be an optimal solution of problem (11). Suppose it is not true, then there is a point $\bar{\theta}^{(m+1)}\in S\bigcap D$ such that $f(\bar{\theta}^{(m+1)})<f(\tilde{\theta}^{(m+1)})$. Without loss of generality, we assume $\bar{\theta}^{(m+1)}$ is a interior point of D(since f is continuous). On the other hand, by the definition of Kuratowski's convergence of sets there is a sequence $\theta_n^{(m+1)}$ such that $\theta_n^{(m+1)}\in S_n$ and $\theta_n^{(m+1)}\to\bar{\theta}^{(m+1)}$, and then $\theta_n^{(m+1)}\in D$, when n is large enough. As $\tilde{\theta}^{(m+1)}$ is a accumulation point of $\{\theta_n^{(m+1)(D_f)}\}$, by (14) we obtain

$$f(\bar{\theta}^{(m+1)})=\lim_{n\to\infty}f(\theta_n^{(m+1)})\ge\lim_{n\to\infty}f_n(\theta_n^{(m+1)(D_f)})=f(\tilde{\theta}^{(m+1)}).$$

This result is contradicted with the assumption $f(\bar{\theta}^{(m+1)})<f(\tilde{\theta}^{(m+1)})$. Hence $\tilde{\theta}^{(m+1)}$ must be an optimal solution of problem (11). Since D is compact and $S_n,\ S$ are closed, $\{\theta_n^{(m+1)(D_f)}\}$ must have accumulation point. Moreover, by the assumption on f the only possible accumulation point is $\theta^{(m+1)(D_f)}$. Thus we get (12).

Next, $\theta^{(m+1)}$ is the optimal solution of problem(7), thus

$$0\ge(\theta^{(m+1)})'K\theta^{(m+1)}-2(\theta^{(m+1)})'T^{(m)}.$$

Then for any $\varepsilon>0$, there exists a constant M such that $\|\theta^{(m+1)}\|\le M$ with a probability larger than $1-\varepsilon$. When $\|\theta_n^{(m+1)}\|\le M$ and $\|\theta^{(m+1)}\|\le M$ note that $\widetilde{Q}_n(\theta_n^{(m+1)}|\theta_n^{(m)})=H_n(Q_n)$, $\ddot{Q}(\theta^{(m+1)}|T^{(m)})=H(Q)$. Therefore

$$P(\widetilde{Q}_n(\theta_n^{(m+1)}|\theta_n^{(m)})\ne H_n(Q_n))<\varepsilon,\quad P(\ddot{Q}(\theta^{(m+1)}|T^{(m)})\ne H(Q))<\varepsilon.$$

By the arbitrariness of ε and $H_n(Q_n)\to_L H(Q)(n\to\infty)$, we get $\widetilde{Q}_n(\theta_n^{(m+1)}|\theta_n^{(m)})\to_L\ddot{Q}(\theta^{(m+1)}|T^{(m)})(n\to\infty)$. □

Also by $\theta_n^{(m+1)(D_f)}\to_L\theta^{(m+1)(D_f)}(n\to\infty)$ and the arbitrariness of f, we know that $\theta_n^{(m+1)(D)}\to_L\theta^{(m+1)(D)}(n\to\infty)$.

Theorem 3.2. *Let $\theta_n^{(m+1)(D)}$ and $\theta^{(m+1)(D)}$ be the optimal solution of problem (8) and (9). If $\theta_n^{(m+1)(D)}\to_L\theta^{(m+1)(D)}(n\to\infty)$, where $D=\Omega\bigcap\{\theta:\ \|\theta\|\le M,\ \theta\subseteq R^p\}$, then for any $M>0$, $\theta_n^{(m+1)}\to_L\theta^{(m+1)}(n\to\infty)$.*

Proof. Observe that the optimal solution $\theta^{(m+1)}$ of problem (7) satisfies

$$0\ge(\theta^{(m+1)})'K\theta^{(m+1)}-2(\theta^{(m+1)})'T^{(m)},$$

because $\theta=0$ is a feasible solution of (7) and K is positive definite matrix, there must be a constant M such that $\|\theta^{(m+1)}\|\le M$.

Let $\theta_n^{(m+1)}$ and $\theta^{(m+1)}$ be the optimal solution of problem (6) and (7), then by Lemma 3.1 if $\|\theta_n^{(m+1)}\|\le M$, we have $\theta_n^{(m+1)}=\theta_n^{(m+1)(D)}$, thus

$$P(\theta_n^{(m+1)}\ne\theta_n^{(m+1)(D)})<\varepsilon.$$

Similarly

$$P(\theta^{(m+1)} \neq \theta^{(m+1)(D)}) < \varepsilon.$$

Therefore, for any $\varepsilon > 0$ and any open set G in R^p, we have

$$\begin{aligned} \liminf P(\theta_n^{(m+1)} \in G) &> \liminf P(\theta_n^{(m+1)(D)} \in G) - \varepsilon \geq P(\theta_n^{(m+1)(D)} \in G) - \varepsilon \\ &\geq P(\theta^{(m+1)} \in G) - 2\varepsilon, \end{aligned}$$

where the second inequality holds because of $\theta_n^{(m+1)(D)} \to \theta^{(m+1)(D)}(n \to \infty)$, then by the arbitrariness of ε, we have

$$\liminf P(\theta_n^{(m+1)} \in G) > P(\theta^{(m+1)} \in G).$$

This is equivalent to $\theta_n^{(m+1)} \to_L \theta^{(m+1)}(n \to \infty)$. □

From Theorem 3.2 we have $\theta_n^{(m+1)} \to_L \theta^{(m+1)}(n \to \infty)$. Because $\theta^{(m+1)}$ is a constant for each fixed m, then $\theta_n^{(m+1)}$ converges to $\theta^{(m+1)}$ in probability, that is $\beta_n^{(m+1)}$ converges to $\beta^{(m+1)}$ in probability for each fixed m. At last, By the theorem of Large number and using the restricted EM algorithm given in this paper, we obtain the desired solution for β.

§4 Numerical simulation

In this section we'll give an example to illustrate the theory developed in earlier sections, for computationally convenient, we just consider 2-dimensional normal distribution situation, the restricted problem is of the following form:

$$\begin{cases} Y_i &= X_i\beta + e_i \quad i = 1, \ldots, n, \\ h_1(\beta) &= \beta_1^2 + 2\beta_2 \geq 0, \\ g_2(\beta) &= \beta_1\beta_2 - \beta_2 + 2 = 0. \end{cases} \tag{15}$$

We take n draws from the linear model (15) with mean $X\beta$ and covariance matrix $\Sigma > \mathbf{0}$, where $X = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$, $\beta = (4, -2/3)'$, $\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, $\sigma_1 = 4,\ \sigma_2 = 4$, ρ denotes the correlation coefficient, $Y_i = (Y_{i1}, Y_{i2})'$, $i = 1, \ldots, n$ are random samples which come from the same normal distribution. We'll compare the restricted EM estimators with the true value $\beta = (4, -2/3)'$.

Let $Y_{i1}(i = 1, \cdots, k)$ and $Y_{i2}(i = n-k+1, \cdots, n)$ are missing data, $Y_{i1}(i = k+1, \cdots, n)$ and $Y_{i2}(i = 1, \cdots, n-k)$ are observed data, $\mu^{(m)} = X\beta^{(m)} = (\mu_1^{(m)}, \mu_2^{(m)})'$, where $\beta^{(m)}(m = 0, 1, \cdots)$ is the restricted EM iteration sequence. From the theory developed in earlier sections we have

$$E(Y_{i1}|Y_{i2}, \beta^{(m)}) = \mu_1^{(m)} + \tfrac{\sigma_2\rho}{\sigma_1}(Y_{i2} - \mu_2^{(m)}), i = 1, \cdots, k,$$
$$E(Y_{i2}|Y_{i1}, \beta^{(m)}) = \mu_2^{(m)} + \tfrac{\sigma_1\rho}{\sigma_2}(Y_{i1} - \mu_1^{(m)}), i = n-k+1, \cdots, n.$$

By Lemma 2.1, when n is large enough, the optimization problem is approximately equivalent to the following form:

$$\begin{cases} \ddot{Q}(\theta|T^{(m)}) = \theta' K\theta - 2\theta' T^{(m)} \\ S = \{\theta \in R^2 : 2\theta_1 + \theta_2 \geq 0; 2\theta_1 - 3\theta_2 = 0\}. \end{cases} \tag{16}$$

where $K = X'\Sigma^{-1}X$, $T^{(m)} = n^{-\frac{1}{2}} \sum X'\Sigma^{-1}[E(Y_i|Y_{obs}, \theta^{(m)}) - X\beta_0]$, $\theta = n^{1/2}(\beta - \beta_0)$, β_0 is true value for β in model:

$$\begin{cases} Y_i &= X_i\beta + e_i \quad i = 1, \dots, n, \\ h_1(\beta) &= \beta_1^2 + 2\beta_2 = 0, \\ g_2(\beta) &= \beta_1\beta_2 - \beta_2 + 2 = 0. \end{cases}$$

Then we consider the optimization problem (16), the following figure 1-2 are the simulations for $\beta = (\beta_1, \beta_2)'$ when $n = 100, 1000; k = 10$ and $\rho = 0.9$.

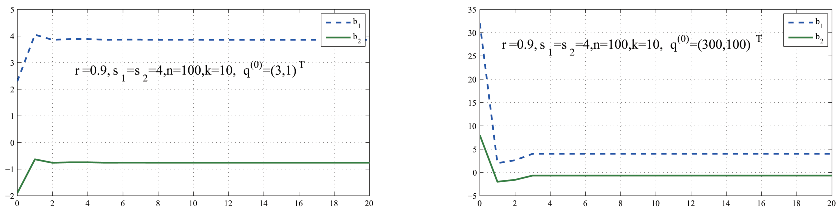

Figure 1: The simulation results for $\beta, n = 100$.

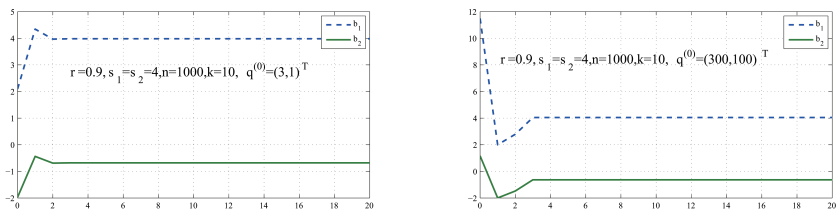

Figure 2: The simulation results for $\beta, n = 1000$.

From the simulation results, we found that the parameters converge to the same true value $\beta = (4, -2/3)'$ when given different initial values, and the effect of convergence is better when n is larger. This is exactly consistent with the theory of the algorithm.

§5 Discussion

In this paper, we propose a restricted EM algorithm for parameters under nonlinear inequality restrictions, by using the asymptotic properties of the maximum likelihood estimators of parameter, the convergence of the algorithm is proved and numerical simulation is given. Next, we are trying to generalize the proposed algorithm to other more general models.

References

[1] A P Dempster, N Laird, D B Rubin. *Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion)*, Journal of the American Statistical Association B, 1977, 39(1), 1-38.

[2] D A D'Esopo. *A convex programming procedure*, Naval Research Logistics Quarterly, 1959, 6(1): 33-42.

[3] F Eicker. *Asymptotic normality and consistency of the least squares estimators for families of linear regressions*, Annals of Mathematical Statistics, 1963, 34(2): 447-456.

[4] C G Hildreth. *A quadratic programming procedure*, Naval Research Logistics Quarterly, 1957, 4(1): 79-85.

[5] D K Kim, J M G Taylor. *The restricted EM Algorithm for maximum likelihood estimation under linear restrictions on the parameters*, Journal of the American Statistical Association, 1995, 90(430): 708-716.

[6] C K Liew. *Inequality constrained least-squares estimation*, Journal of the American Statistical Association, 1976, 71(355): 746-751.

[7] R J A Little, D Rubin. *Statistical Analysis with Missing Data*, Wiley, New York, 1987.

[8] R J A Little, D Rubin. *The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence*, Biometrika, 1994, 81(4): 633-648.

[9] X S Liu, J D Wang. *Testing for increasing convex order in several populations*, Annals of Institute Statistical Mathematics, 2003, 55(1): 121-136.

[10] X L Meng, D B Rubin. *Maximum likelihood estimation via the ECM algorithm: a general framework*, Biometrika, 1993, 80(2): 267-278.

[11] N Nagaraj, W Fuller. *Estimation of the parameters of linear time series models subject to nonlinear regressions*, Annals of Statistics, 1991, 19(3): 1143-1154.

[12] H Nyquist. *Restricted estimation of generalized linear models*, Applied statistics, 1991, 40(1): 133-141.

[13] R Prakasa. *Asymptotic theory of statistical inference*. Wiley, New york, 1987.

[14] R Ranjana, B Huanga, A Fatehiab. *Robust Gaussian process modeling using EM algorithm*, Journal of Process Control, 2016, 42(6): 125-136.

[15] N Z Shi, S R Zheng, J H Guo. *The restricted EM algorithm under inequality restrictions on the parameters*, Journal of Multivariate Analysis, 2005, 92(1): 53-76.

[16] J D Wang. *The asymptotics of least-squares estimators for constrained nonlinear regression*, Annals of Statistics, 1996, 24(3): 1316-1326.

[17] D Wei , X K S Peter. *EM algorithm in Gaussian copula with missing data*, Computational Statistics and Data Analysis, 2016, 101(12): 1-11.

[18] C F J Wu. *on the convergence properties of the EM algorithm*, Annals of Statistics, 1969, 11(1): 95-103.

[1] Department of Basic Course, Zhengzhou University of Science and Technology, Zhengzhou 450064, China.

[2] Information Engineering School of Zhengzhou Institute of Technology, Zhengzhou 450044, China.

Email: 49586700@qq.com