An Introduction to EM and ECM Algorithms

C. D. Fuh†
Institute of Statistical Science, Academia Sinica
Taipei, Taiwan, ROC

OUTLINE

I. EM Algorithms
  I.1 A Censored Data Example
  I.2 Mixture Models

II. Hidden Markov Models (HMM)
  II.1 Definition
  II.2 Baum-Welch (EM) Algorithms
  II.3 Switch Autoregression

III. ECM Algorithms
  III.1 Normal Mixtures with Equal Correlations
  III.2 Rate of Convergence

VI. Some Possible Further Researches

I.1 Censored data:

Let $X_1, \cdots, X_n$ be $i.i.d.$ from $f(x;\theta) = 1/\theta e^{-x/\theta}$, $x > 0$ (exponential distribution). Now if we have censor time T_0 and exactly r observations $1 < r < n$, such that $X_i < T_0$. W.l.o.g. we may assume $X_1, \cdots, X_r$ and from X_{r+1} to X_n, we have only know that they are $\geq T_0$. Then, $L_y(\theta)$, the likelihood function is

$$\begin{aligned} \prod_{i=1}^{n} g(y_i;\theta) &= (\prod_{i=1}^{r} \frac{1}{\theta} e^{-x_i/\theta})(\prod_{i=r+1}^{n} e^{-T_0/\theta}) \\ &= \frac{1}{\theta^r} e^{-1/\theta(\sum_{i=1}^{r} x_i + (n-r)T_0)} \end{aligned}$$

where $\int_{T_0}^{\infty} \frac{1}{\theta} e^{-x/\theta} dx = e^{-T_0/\theta}$. Therefore,

$$\hat{\theta}_n = \frac{T}{r}, \quad \text{where} \quad T = \sum_{i=1}^{r} x_i + (n-r)T_0.$$

We have $X_1, \cdots, X_r$ and missing $X_{r+1}, \cdots, X_n$. Now we replace X_i by

$$E(X \,|\, X \geq T_0, \theta)(= T_0 + \theta) \quad \text{for} \quad i = r+1, \cdots, n.$$

Since θ is unknown, we replace it by $\hat{\theta}_p$ (the p^{th} step) and choose a "reasonable $\hat{\theta}_0$". Then,

$$L_x(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} e^{-x_i/\theta} = \frac{1}{\theta^n} e^{(-\frac{1}{\theta})[\sum_{i=1}^{r} x_i + (n-r)(T_0 + \hat{\theta}_p)]},$$

we have

$$\hat{\theta}_{p+1} = \frac{1}{n}[\sum_{i=1}^{r} x_i + (n-r)(T_0 + \hat{\theta}_p)]$$

$\hat{\theta}_p$ converges to θ^*, then solving the EQ, we get

$$\theta^* = \frac{1}{r} \sum_{i=1}^{r} X_i + \frac{n-r}{r} T_0.$$

I. EM algorithm

We have two sample spaces $\mathcal{X}$ and $\mathcal{Y}$ and a many-to-one mapping from $\mathcal{X}$ to $\mathcal{Y}$. Instead of observing the "complete data" x in $\mathcal{X}$, we observe the "incomplete data" $y = y(x)$ in $\mathcal{Y}$.

Let the density function of x be $f(x|\theta)$ with parameters $\theta \in \Omega$ and let the

$$g(y|\theta) = \int_{\mathcal{X}(y)} f(x|\theta) dx,$$

where $\mathcal{X}(y) = \{x : y(x) = y\}$. The parameters θ are to be estimated by the method of maximum likelihood, i.e., by maximizing $g(y|\theta)$ over $\theta \in \Omega$.

In many statistical problems, maximization of the complete-data specification $f(x|\theta)$ is simpler than that of the incomplete-data specification $g(x|\theta)$.

A main feature of the EM algorithm is maximization of $f(x|\theta)$ over $\theta \in \Omega$ (M-step). Since x is unobservable, we replace $\log f(x|\theta)$ by its *conditional expectation* given y and the current fit θ_p (E-step).

We now define the EM iteration as follows:
E-step: Determine $Q(\theta|\theta_p) = E\{\log f(x|\theta)|y, \theta_p\}$.
M-step: Choose θ_{p+1} to be any value of $\theta \in \Omega$ which maximizes $Q(\theta|\theta_p)$.

I.2 Mixture Models:

1. The normal mixture model is

$$f(w;\theta) = \sum_{i=1}^{d} \pi_i f_i(w;\theta_i),$$

where $f_i(w;\theta_i) = (2\pi\sigma^2)^{-1/2}\exp\{-1/2(w-\mu_i)^2/\sigma^2\}$.

The unknown parameters is

$$\theta = ((\mu_1,\cdots,\mu_d,\sigma^2)^t,\pi_1,\cdots,\pi_d)^t.$$

Let x be the complete-data vector

$$x = (y^t, z^t)^t.$$

The complete-data log likelihood function for θ is

$$\begin{aligned}
&\log L_c(\theta) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{n} z_{ij}\log\pi_i + \sum_{i=1}^{d}\sum_{j=1}^{n} z_{ij}\log f_i(w_j;\theta_i) \\
&= \sum_{i=1}^{d}\sum_{j=1}^{n} z_{ij}\log\pi_i - \frac{1}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{d}\sum_{j=1}^{n} z_{ij}\{\log\sigma^2 + (w_j-\mu_i)^2/\sigma^2\},
\end{aligned}$$

where z_{ij} is the indicator function of observation w_j from the mixture component i, for $i = 1,\cdots,d$, and $j = 1,\cdots,n$.

Now, the MLE's of μ_i and σ are

$$\frac{\Sigma_{j=1}^{n} z_{ij}w_j}{\Sigma_{j=1}^{n} z_{ij}} \quad \text{and} \quad \frac{\Sigma_{i=1}^{d}\Sigma_{j=1}^{n} z_{ij}(w_j-\mu)^2}{n},$$

respectively, if the z_{ij} were observed. As $\log L_c(\theta)$ is linear in the z_{ij}. In the E-step, we replace z_{ij} by their current conditional expectations

$$z_{ij}^{(k)} = \frac{\pi_i^{(k)} f_i(w_j;\theta_i^{(k)})}{\Sigma_{i=1}^{d} f_i(w_j;\theta_i^{(k)})}, \quad \text{for } i = 1,\cdots,d,$$

where $\pi_i^{(k+1)} = \Sigma_{j=1}^n z_{ij}^{(k)}/n$, for $i = 1, \cdots, d$. This yields

$$\mu_i^{(k+1)} = \frac{\Sigma_{j=1}^n z_{ij}^{(k)} w_j}{\Sigma_{j=1}^n z_{ij}^{(k)}} \quad i = 1, \cdots, d,$$

and

$$\sigma^{(k+1)^2} = \frac{\Sigma_{i=1}^d \Sigma_{j=1}^n z_{ij}^{(k)} \big(w_j - \mu_i^{(k+1)}\big)^2}{n}.$$

2. Multivariate Component Densities

On the M-step at the $(k+1)$th iteration, the updated estimates of the ith component mean μ_i and the common covariance matrix σ are given by

$$\mu_i^{(k+1)} = \frac{\Sigma_{j=1}^n z_{ij}^{(k)} \mathbf{w}_j}{\Sigma_{j=1}^n z_{ij}^{(k)}}, \quad i = 1, \cdots, d,$$

and

$$\Sigma^{(k+1)} = \frac{\Sigma_{i=1}^d \Sigma_{j=1}^n z_{ij}^{(k)} \big(\mathbf{w_j} - \mu_i^{(k+1)}\big)\big(\mathbf{w_j} - \mu_i^{(k+1)}\big)^t}{n}.$$

Theorem 1 *(Render and Walker, 1984)*

Suppose the Fisher information $I\big(\theta\big)$ is positive definite at θ^* and that $\pi_i^* > 0$ for $i = 1, \cdots, d$. Denote $\theta^{(k)}$, the sequence generated by the EM algorithm. Then with probability 1, whenever N is sufficiently large, the unique strongly consistent solution θ^N of the likelihood equations is well-defined and there is a constant λ, $0 \leq \lambda < 1$, for which

$$\big\|\theta^{(k+1)} - \theta^N\big\| \leq \lambda \big\|\theta^{(k)} - \theta^N\big\|, \quad k = 0, 1, 2, \cdots,$$

whenever $\theta^{(0)}$ is sufficiently near θ^N.

II. Hidden Markov Models

II.1a Definition of a HMM

Definition: A HMM is a doubly stochastic process with an underlying stochastic process that is "not" observable, but can only be observed through another set of stochastic processes that produce the sequence of observed symbols.

Let $\{X_t\}$ be a S state Markov chain, generated by some $s \times s$ stochastic matrix $\{a_{ij}\}$ with positive entries. Let $\{Y_t\}$ be a probabilistic function of $\{X_t\}$, viz:

$$P\{Y_t = k | X_t = j, Y_{t-1}, X_{t-1}, \cdots\} = b_{jk},$$

where $\{b_{jk}\}$ is a $s \times r$ matrix with positive entries and row sums 1. A HMC is (X_t, Y_t) with only observed value $\{Y_1, \cdots, Y_T\}$.

1. Noninvertible functional of Markov chains.

2. Generalization of Mixtures.

Coin toss example:

$$
\begin{array}{rcccccccccc}
O = & \mathcal{H} & \mathcal{H} & \mathcal{T} & \mathcal{H} & \mathcal{T} & \mathcal{T} & \mathcal{H} & \mathcal{H} & \cdots & \mathcal{T} \\
 & O_1 & O_2 & O_3 & O_4 & O_5 & O_6 & O_7 & O_8 & \cdots & O_T
\end{array}
$$

II.1b Some Applications

-Speech Recognition (cf. Rabiner and Juang 1993).

-Signal Processes (cf. Elliott et al. 1995)'

-Ion Channels (cf. Ball and Rice 1992).

-Molecular Biology (cf. Krogh et al. 1994).

-Economics (cf. Hamilton 1994).

-Brief Summary (kunsch, Geman and Kehagias 1995).

Basic Issues:

-State Space Estimation.

-Algorithms for the Fitting of These Models.

-Implementation of Likelihood Based Methods.

II.1c Elements of a HMM

i) There are a finite number, say S, of states in the model.

ii) At each clock time, t, a new state is entered based on a transition probability distribution which depends on the previous state (the Markov property).

iii) After each transition is made, an observation output symbol is produced according to a probability distribution which depends on the current state.

We now formally define the following model notation for a discrete observation HMC:
T = length of the observation sequence,
S = number of states in the model,
R = number of observation symbols,
$A = \{a_{ij}\},\ a_{ij} = Pr(X_t = j \,|\, X_{t-1} = i)$, state transition probability distribution,
$B = \{b_{jk}\},\ b_{jk} = Pr(Y_t = k \,|\, X_t = j)$, observation symbol probability distribution in state j,
$a = \{a_i\},\ a_i = Pr(X_0 = i)$, initial state distribution.

Using the model, an observation sequence, $Y = Y_1, Y_2, \cdots, Y_T$ is given. We use the compact notation $\lambda = (a, A, B)$ to represent a HMC. Specification of a HMC involves
1) Choice of the number of states S,
2) The number of discrete symbols R,
3) Specification of the three probability densities a, A and B.

II.1d The Three Problems in HMM's

1) Given the observation sequence Y_t, and the model λ, how to compute $Pr(Y|\lambda)$? (likelihood function)
2) How do we choose a state sequence which is optimal in some meaningful sense. (model selection)
3) How do we adjust the model parameters $\lambda = (a, A, B)$ to maximize $Pr(Y|\lambda)$. (parameter estimation)

Solutions to the three HMC problems:
1)

$$\begin{aligned} & Pr(Y|\lambda) \\ = & \sum_{all\ l} Pr(Y|l,\lambda) Pr(l|\lambda) \\ = & \sum_{i_1,\cdots,i_T} a_{i_1} b_{i_1}(Y_1) a_{i_1 i_2} b_{i_2}(Y_2) \cdots a_{i_{T-1} i_T} b_{i_T}(Y_T). \end{aligned}$$

More efficient procedure: "forward-backward procedure".

2) Viterbi algorithm.

3) There's no known way to solve MLE analytically. An iterative procedure, such as the Baum-Welch method (EM-algorithm) must be used.

II.2 Baum-Welch (EM) algorithm

Let $P(a, A, B) = \Sigma_{x \in S}\, p(x, a, A, B)$ where

$$p(x, a, A, B) = a_{x_0} \prod_{t=1}^{T} a_{x_{t-1} x_t} b_{x_t}(y_t),$$

and $y_t \in \{1, 2, \cdots, r\}$.

The E-step gives

$$\begin{aligned} & Q\left(a', A', B' | a, A, B\right) \\ = & \sum_x p(x, a, A, B)(\log a'_{x_0} + \sum_t \log a'_{x_{t-1}x_t} + \sum_t \log b'_{x_t}(y_t)), \end{aligned}$$

For M-step, an elementary computation shows that there is an unique $\bar{a}'$, $\bar{A}'$ and $\bar{B}'$ which maximizes Q as a function of the primed coordinated. we find:

$$\begin{aligned} \bar{a}'_i &= \frac{\Sigma_{x_0=i}\, p(x, a, A, B)}{\Sigma_x\, p(x, a, A, B)}, \\ \bar{a}'_{ij} &= \frac{\Sigma_x\, p(x, a, A, B)\, \Sigma_{x_{t-1}=i, x_t=j}\, 1}{\Sigma_x\, p(x, a, A, B)\, \Sigma_{x_{t-1}=i, t\geq 1}\, 1}, \\ \bar{b}'_j(k) &= \frac{\Sigma_x\, p(x, a, A, B)\, \Sigma_{x_t=j, y_t=k}\, 1}{\Sigma_x\, p(x, a, A, B)\, \Sigma_{x_t=j}\, 1}. \end{aligned}$$

The reestimation formula for a_i is trivially the probability of being in state i at $t = 1$.

The reestimation formula for a_{ij} is the ratio of the expected number of transitions from state i to j, divided by the expected number of transitions out of state i.

The reestimation formula for $b_j(k)$ is the ratio of the expected number of times of being in state j and observing symbol k divided by the expected number of times of being in state j. Note that the summation for $b_j(k)$ is from $t = 1$ to $t = T$.

If we define the initial model as λ and the reestimation model as $\bar{\lambda}$, consisting of the above $\bar{a}_i, \bar{a}_{ij}$ and $\bar{b}_j(k)$, then it can be proven that either:

1. The initial model λ defines a critical point of the likelihood function, in which case $\bar{\lambda} = \lambda$, or

2. Model λ is more likely in the scene that $Pr(Y|\bar{\lambda}) > Pr(Y|\lambda)$, i.e. we have found another model $\bar{\lambda}$, from which the observation sequence is more likely to be produced.

More generally, let X be a totally finite measure space with measure μ. Let

$$P(\lambda) = \int_X p(x, \lambda) d\mu(x),$$

with

$$p(x, \lambda) = \prod_{t=1}^{T} b_{x_t}(\lambda, y_t),$$

and

$$Q(\lambda'|\lambda) = \int_X p(x, \lambda) \log p(x, \lambda') d\mu(x).$$

Lemma 1 *If $Q(\bar{\lambda}|\lambda) \geq Q(\lambda|\lambda)$ then $P(\bar{\lambda}) \geq P(\lambda)$. The inequality is strictly unless $p(x, \lambda) = p(x, \bar{\lambda})$ almost everywhere $d\mu(x)$.*

Proof. $\log t$ is strictly concave for $t > 0$ since $d^2/dt^2(\log t) = -t^{-2} < 0$. Hence

$$\begin{aligned} & \log P(\bar{\lambda})/P(\lambda) \\ = \; & \log \int_x p(x, \bar{\lambda}) d\mu(x)/P(\lambda) \\ = \; & \log \int_x [p(x, \lambda) d\mu(x)/P(\lambda)] \, p(x, \bar{\lambda})/p(x, \lambda) \end{aligned}$$

$$
\begin{aligned}
&\geq \int_x [p(x,\lambda)d\mu(x)/P(\lambda)] \log [p(x,\bar{\lambda})/p(x,\lambda)] \\
&= (P(\lambda))^{-1} [Q(\lambda,\bar{\lambda}) - Q(\lambda,\lambda)] \\
&\geq 0.
\end{aligned}
$$

For the first inequality we have used Jensen's inequality for the probability measure $dv_\lambda(x) = p(x,\lambda)d\mu(x)/P(\lambda)$. This inequality is strict unless $p(x,\bar{\lambda})/p(x,\lambda)$ is constant almost everywhere $dv_\lambda(x)$, hence unless $p(x,\bar{\lambda}) = p(x,\lambda)$ almost everywhere $d\mu(x)$.

The convergence of Baum-Welch algorithm:

Define $\mathcal{F} : \Lambda \to \Lambda$ by

$$\mathcal{F}(\lambda) = \max_{\lambda' \in \Lambda} Q(\lambda'|\lambda),$$

then $Q(\mathcal{F}(\lambda)|\lambda) \geq Q(\lambda|\lambda)$. By Lemma 1, we have

$$P(\mathcal{F}(\lambda)) \geq P(\lambda).$$

Let $\Lambda = R^s$. For almost all x let $\log p(x,\lambda) = \Sigma_{i=1}^s \log p_i(x,\lambda_i)$ where for each i and almost all x, $\log p_i(x,\lambda_i)$ is strictly concave in λ_i and
$\lim_{|\lambda_i|\to\infty} \log p_i(x,\lambda_i) = -\infty$.
Define $Q_i(\lambda_i'|\lambda)$ by

$$Q_i(\lambda_i'|\lambda) = \int_X p(x,\lambda) \log p_i(x,\lambda_i') d\mu(x).$$

Then for λ fixed, $Q_i(\lambda_i'|\lambda)$ is a strictly concave function of λ_i' which $\to -\infty$ at $\pm\infty$ and hence has a "unique" global maximum $\bar{\lambda}_i$, which is a critical point of $Q_i(\lambda_i'|\lambda)$.

Define $\mathcal{F} : \lambda \to \bar{\lambda} = \{\bar{\lambda}_i\}$.

Theorem 2 *Under the above assumptions, for all* $\lambda \in \Lambda$, $P\big(\mathcal{F}(\lambda)\big) \geq P\big(\lambda\big)$ *with equality if and only if* λ *is a critical point of* P *or equivalently is a fixed point of* $\mathcal{F}$.

Proof.

$$Q\left(\bar{\lambda}|\lambda\right) = \sum_{i=1}^{s} Q_i(\bar{\lambda}_i|\lambda\big) \geq \sum_{i=1}^{s} Q_i(\lambda_i|\lambda\big) = Q\left(\lambda|\lambda\right).$$

So Lemma 1 implies $P\left(\bar{\lambda}\right) \geq P\left(\lambda\right)$.

Since for each i, $Q_i(\lambda_i|\lambda\big)$ is strictly concave in λ_i the inequality $Q\left(\bar{\lambda}|\lambda\right) \geq Q\left(\lambda|\lambda\right)$ (and hence the inequality $P\left(\bar{\lambda}\right) \geq P\left(\lambda\right)$) will be strict unless $\bar{\lambda}_i = \lambda_i$ for each i. This latter is the case iff $\partial Q_i/\partial \lambda_i^{'}|_{\lambda'=\lambda} = 0, i = 1, \cdots, s$; i.e., $\partial P/\partial \lambda_i|_{\lambda} = 0, i = 1, \cdots, s$.

Proposition 1 *If* $y_t,\ t = 1, \cdots, T$ *is a sequence of real numbers and* $c_x = c_{x_1 \cdots x_T}$ *then*

$$P\left(\lambda\right) = \sum_{x} c_x \prod_{t=1}^{T} \exp(-|\lambda_{X_t} - y_t|)$$

assumes its global maximum at a point $\bar{\lambda}$ *where each* $\bar{\lambda}_i$ *is some* y_{t_i}.

Proof. We have $Q\left(\lambda^{'}|\lambda\right) = \Sigma_i Q_i(\lambda_i^{'}|\lambda\big)$ where $Q_i(\lambda_i^{'}|\lambda\big) = -\Sigma\, \gamma_t(i)|\lambda_i^{'} - y_t|$ with $\gamma_t(i) \geq 0$. Then $Q_i(\lambda_i^{'}|\lambda\big) \to -\infty$ as $|\lambda_i| \to \infty$.

$$\begin{aligned} &\frac{\partial}{\partial \lambda_i^{'}} Q_i(\lambda_i^{'}, |\lambda\big) \\ &= -\sum \gamma_i(t) sgn(\lambda_i^{'} - y_t), \qquad \lambda_i^{'} \neq \text{any } y_t, \end{aligned}$$

is monotonic decreasing from ∞ at $\lambda_i^{'} = -\infty$ to $-\infty$ at $\lambda_i^{'} = \infty$, constant in intervals between y_t, an changes sign at some $y_i =$

$\bar{\lambda}_i$. We then have $Q_i(\bar{\lambda}_i|\lambda) = \max_{\lambda_i'} Q_i(\lambda_i'|\lambda) \geq Q_i(\lambda_i|\lambda)$. By Lemma 1 $P(\bar{\lambda}) \geq P(\lambda)$. Now if λ_i is any parameter which appears in $P(\lambda)$ with a non-zero coefficient c_x, then $P(\lambda) \to -\infty$ as $|\lambda_i| \to \infty$ so that $P(\lambda)$ assumes its maximum at some point $\hat{\lambda}$. But $\mathcal{F}(\hat{\lambda}) = \bar{\lambda}$ is a point with each $\bar{\lambda}_i$ a y_{t_i} and $P(\bar{\lambda}) \geq P(\hat{\lambda}) = \max P(\lambda)$.

Examples:

The densities

$$\begin{aligned} b_i(\lambda, y) &= ce^{-k|\lambda_i - y|^\epsilon}, \quad \epsilon > 0, k > 0, \\ &\quad and \ -\infty \leq \lambda_i \leq \infty, \\ b_i(\lambda, y) &= \frac{e^{\lambda_i}\lambda_i^y}{y!}, \quad 0 \leq \lambda_i < \infty, \\ b_i(\lambda, y) &= C(N, y)\lambda_i^y(1 - \lambda_i)^{N-y} \quad 0 \leq \lambda_i \leq 1, \end{aligned}$$

satisfies the conditions of Theorem 1.

II.3 Switch Autoregression

A simple scalar-valued fourth-order autoregression around one of two constants μ_1 or μ_2:

$$\begin{aligned}\xi_n - \mu_{x_n} &= \varphi_1(\xi_{n-1} - \mu_{x_{n-1}}) + \varphi_2(\xi_{n-2} - \mu_{x_{n-2}}) \\ &\quad + \varphi_3(\xi_{n-3} - \mu_{x_{n-3}}) + \varphi_4(\xi_{n-4} - \mu_{x_{n-4}}) + \varepsilon_n,\end{aligned}$$

where $\varepsilon_n \sim N(0, \sigma^2)$, x_n is a two-state ergodic Markov chain and $\theta = (\varphi_1, \varphi_2, \varphi_3, \varphi_4, \mu_1, \mu_2, \sigma^2)$ are the unknown parameters.

This model has been studied by Hamilton (1989) in an attempt to analyze the behavior of U.S. real GNP. In this case, the likelihood function for given $X_n = x_n$ for $n \geq 0$ is

$$\begin{aligned}&f(\xi_n | x_n; \theta) \\ &= \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ -[(\xi_n - \mu_{x_n}) - \frac{1}{2\sigma^2} \sum_{k=1}^{4} \varphi_k(\xi_{n-k} - \mu_{x_{n-k}})]^2 \Big\}.\end{aligned}$$

Assume that all the roots of $1 - \Sigma_{k=1}^{4} \varphi_k z^k = 0$ are outside of a unit circle, and that there exists a constant $c > 0$ such that $\sigma^2 > c$. Suppose the identifiability condition holds.

Engel and Hamilton (1990) considered another switch autoregression model in which both the mean vector and the variance-covariance matrix were functions of the state:

$$\xi_n | x_n \sim N(\mu_{x_n}, \Omega_{x_n}), \qquad \text{for } x_n = 1, 2,$$

where $\theta = (\mu_1, \mu_2, \Omega_1, \Omega_2)$ are the unknown parameters. Assume that there exists a constant c such that $0 < c < |\Omega_x|$ for each $x = 1, 2$. Suppose the identifiability condition holds.

In general, let $\xi_1, \cdots, \xi_n$ be a sample from the model

$$\xi_n = \sum_{k=1}^{p-1} a_{x_n}^k \xi_{n-k} + \sigma_{x_n} v_n,$$

where v_n is a normal random variable with zero mean and unit variance, and $a_x = (a_x^1, \cdots, a_x^{p-1}, \sigma_x)$ are the unknown parameters. Assume that all the roots of $1 - \Sigma_{k=1}^{p} a_x^k z^k = 0$ are outside of a unit circle, and that there exists a constant c with $0 < c < \sigma_x^2$ for $x = 1, \cdots, d$.

This formulation also includes a generalization of Engel's ARCH model to allow for occasional discrete shifts in the ARCH parameters.

Hamilton (1990) applied EM algorithm to approximate the MLE.

III. ECM Algorithm

ECM (expectation-conditional maximization) Algorithm

– Replace a "complicated" M-step of the EM algorithm with several computational simpler CM-steps.

Suppose that the M-Step is replaced by $S > 1$ steps. We let $\theta^{(k+s/S)}$ denote the value of θ on the sth CM-step of the $(k+1)$th iteration, where $\theta^{(k+s/S)}$ is chosen to maximize

$$Q\left(\Phi|\theta^{(k)}\right)$$

subject to the constraint

$$g_s(\theta) = g_s(\theta^{(k+(s-1)/S)}).$$

Here $C = \{g_s(\theta), s = 1, \cdots, S\}$ is a set of S preselected functions.

III.1 Normal Mixtures with Equal Correlations:

The case of two equal component-correlation matrices can be represented by writing the covariance matrices Σ_1 and Σ_2 as

$$\Sigma_1 = \Sigma_0 \quad \text{and} \quad \Sigma_2 = K\,\Sigma_0 K,$$

where Σ_0 is a positive definite symmetric matrix and

$$K = diag(k_1, \cdots, k_p).$$

For the application of the ECM algorithm,

$$\theta = (\theta_1^t, \theta_2^t)^t,$$

where $\theta_1 = (\pi_1, \mu_1, \mu_2, \Sigma_0)$ and $\theta_2 = (k_1, \cdots, k_p)$.

CM-Step 1. Calculate $\theta_1^{(k+1)}$ as the value of θ_1 that maximizes $Q\left(\theta\middle|\theta^{(k)}\right)$ with θ_2 fixed at

$$\theta_2^{(k)} = \left(k_1^{(k)}, \cdots, k_p^{(k)}\right)^t.$$

CM-Step 2. Calculate $\theta_2^{(k+1)}$ as the value of θ_2 that maximizes $Q\left(\theta\middle|\theta^{(k)}\right)$ with θ_1 fixed at $\theta_1^{(k+1)}$.

It follows that $\theta_1^{(k+1)}$ consists of $\pi_1^{(k+1)}$, the elements of $\mu_1^{(k+1)}$ and $\mu_2^{(k+1)}$, and the distinct elements of $\Sigma_0^{(k+1)}$, where

$$\begin{aligned}
\pi_i^{(k+1)} &= \sum_{j=1}^{n} z_{ij}^{(k)}/n \quad (i=1,2), \\
\mu_i^{(k+1)} &= \frac{\sum_{j=1}^{n} z_{ij}^{(k)} \mathbf{w}_j}{\sum_{j=1}^{n} z_{ij}^{(k)}}, \quad (i=1,2), \\
\Sigma_0^{(k+1)} &= \pi_1^{(k+1)} V_1^{(k+1)} + \pi_2^{(k+1)} K^{(k)^{-1}} V_2^{(k+1)} K^{(k)^{-1}},
\end{aligned}$$

where

$$V_i^{(k+1)} = \frac{\sum_{j=1}^{n} z_{ij}^{(k)} \left(\mathbf{w_j} - \mu_i^{(k+1)}\right)\left(\mathbf{w_j} - \mu_i^{(k+1)}\right)^t}{\sum_{j=1}^{n} z_{ij}^{(k)}} \quad (i=1,2),$$

and

$$K^{(k)} = diag\left(k_1^{(k)}, \cdots, k_p^{(k)}\right)^t.$$

Here,

$$z_{1j}^{(k)} = \frac{\pi_1^{(k)} \phi\left(\mathbf{w}_j; \mu_1^{(k)}, \Sigma_0^{(k)}\right)}{\pi_1^{(k)} \phi\left(\mathbf{w}_j; \mu_1^{(k)}, \Sigma_0^{(k)}\right) + \pi_2^{(k)} \phi\left(\mathbf{w}_j; \mu_2^{(k)}, \Sigma_2^{(k)}\right)},$$

where

$$\Sigma_2^{(k)} = K^{(k)} \Sigma_0^{(k)} K^{(k)},$$

and $z_{2j}^{(k)} = 1 - z_{1j}^{(k)}$.

In the CM-Step 2, $k_\alpha^{(k+1)}$ is a solution of the equation

$$k_\alpha = \sum_{i=1}^{p} (\Sigma_0^{(k+1)^{-1}})_{i\alpha} (V_2^{(k+1)})_{i\alpha} / k_i \quad (\alpha = 1, \cdots, p).$$

This equation can be solved iteratively (by Newton-Raphson) to yield $k_\alpha^{(k+1)}$.

III.2 Rate of Convergence

EM and ECM implicityl defines a map $\theta \rightarrow M(\theta)$ such that $\theta^{(k+1)} = M(\theta^{(k)})$. If θ^* is a limit of $\{\theta^{(k)}, k \geq 0\}$ and $M(\theta)$ is differentiable in a nhood of θ^*,

$$\theta^{(k+1)} - \theta^* \approx (\theta^{(k)} - \theta^*) D\,M(\theta^*),$$

where

$$D\,M(\theta) = (\frac{\partial M_j(\theta)}{\partial \theta_i})$$

is the $p \times p$ first derivative matrix for $M(\theta) = (M_1(\theta), \cdots, M_p(\theta))$.

Theorem 3 *Suppose all the outputs of an ECM, $\theta^{(k+s/S)}, k \geq 1, s = 1, \cdots, S$, satisfy the Lagrange multiplier equations for constrained maximization, and $\theta^{(k+s/S)} \rightarrow \theta^*$ as $k \rightarrow \infty$ for $s = 1, \cdots, S$. Then the rate of convergence of ECM is given by*

$$D\,M^{\,ECM}(\theta^*) = D\,M^{\,EM}(\theta^*) + (I_p - D\,M^{\,EM}(\theta^*)) \prod_{s=1}^{S} P_s.$$

Here $D\,M^{\,EM}(\theta^*)$ is the rate of convergence of EM given by

$$D\,M^{\,EM}(\theta^*) = I_{mis}(\theta^*) I_{com}^{-1}(\theta^*),$$

where

$$I_{mis} = \int -\frac{\partial^2 \log f(Y_{mis}|Y_{obs}, \theta)}{\partial \theta^2} f(Y_{mis}|Y_{obs}, \theta) d\,Y_{mis}$$

$$I_{com} = \int -\frac{\partial^2 \log f(Y|\theta)}{\partial \theta^2} f(Y_{mis}|Y_{obs}, \theta) d\,Y_{mis}.$$

$$P_s = \bigtriangledown_s (\bigtriangledown_s^t I_{com}^{-1}(\theta^*) \bigtriangledown_s)^{-1} \bigtriangledown_s^t I_{com}^{-1}(\theta^*), \quad s = 1, \cdots, S,$$

with $\bigtriangledown_s = \bigtriangledown g_s(\theta^*)$ *and* $\Pi_{s=1}^{S} P_s = P_1 \cdots P_S$.

Define

$$r = \lim_{k \to \infty} \frac{\|\theta^{(k+1)} - \theta^*\|}{\|\theta^{(k)} - \theta^*\|}.$$

It is known that

$$r = \lambda_{max} = \text{ the largest eigenvalue of } D\,M\,(\theta^*).$$

Note that a large value of r implies slow convergence. Define $s = 1 - r$ as the *global speed of convergence*. And

$$s \text{ is the smallest eigenvalue of } \mathcal{S} = I_p - D\,M\,(\theta^*).$$

Remarks: No missing data, $D\,M^{\,EM}(\theta^*) = 0$, and thus

$$D\,M^{\,CM}(\theta^*) = \prod_{s=1}^{S} P_s,$$

where $M^{\,CM}(\theta)$ is the mapping determined by the CM algorithm. It follows that

$$I_p - D\,M^{\,ECM}(\theta^*) = (I_p - D\,M^{\,EM}(\theta^*))(I_p - D\,M^{\,CM}(\theta^*)),$$

and hence,

$$\text{Speed of ECM} = \text{ Speed of EM} \times \text{ Speed of CM}.$$

IV. Some possible further researches:

1) The comparison of Fisher information (observed Fisher information) between Mixture Distributions and Hidden Markov Models. (I guess HMM gives larger FI).

2) The performance of EM (ECM, AECM) in a general state HMM and hidden Markov random fields.

2.1. Convergence and Rate of convergence.

2.2. The LIL for the maximum likelihood estimator of λ in HMM.

3) The performance of bootstrap (include EM algorithm in the Monte-Carlo simulation of the bootstrap resample) method in HMM.

4) Model selection methods: AIC, BIC, Stochastic complexity to estimate the number of hidden states in HMM. MCMC (Gibbs sampler) to estimate the distribution of the number of hidden states in HMM.

5) Interaction between MCMC and GEM algorithm.