Expectation-Maximization

October 18, 2017

1 Estimation using EM

1.1 General Formulation of EM

We consider a general latent variable model. Denote the augmented data (y_i, α_i) and the log-likelihood model is

$$\begin{aligned} l(y_i) &= \log Pr[y_i|\theta] \\ &= \log \sum_k Pr[y_i, \alpha_i{=}\alpha_k|\theta] \end{aligned}$$

To understand the EM we start with a parameter guess θ^τ and we consider the expression of interest, the log-likelihood. For any k the following is true:

$$\forall k, \qquad \sum_i \log Pr[y_i|\theta] = \sum_i \log Pr[y_i, \eta_i{=}\eta_k|\theta] - \log Pr[\eta_i{=}\eta_k|y_i, \theta]$$

We then take the expectation of the previous expression with weights $Pr[\eta_i{=}\eta_k|y_i, \theta^\tau]$

$$\begin{aligned} \sum_i \log Pr[y_i|\theta] &= \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \left(\log Pr[y_i, \eta_i|\theta] - \log Pr[\eta_i|y_i, \theta]\right) \\ &= \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \cdot \log Pr[y_i, \eta_i|\theta] - \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \cdot \log Pr[\eta_i|y_i, \theta] \\ &= Q(\theta|\theta^\tau) + H(\theta|\theta^\tau) \end{aligned}$$

The EM algorithm then consists of 2 steps:

1. E-step: compute the $q_i(k) = Pr[\eta_i{=}\eta_k|y_i, \theta^\tau]$ using the data, the model and the θ^τ guess
2. M-step: choose θ to maximize $Q(\theta|\theta^\tau) = \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \cdot \log Pr[y_i, \eta_i|\theta]$

The proof that the EM algorithm is always increasing compares the likelihood at θ^τ and $\theta^{\tau+1}$. Expand both with $Pr[\eta_i{=}\eta_k|y_i, \theta^\tau]$ to get the difference equal to:

$$\sum_i \log Pr[y_i|\theta^{\tau+1}] - \sum_i \log Pr[y_i|\theta^\tau] = \left(Q(\theta^{\tau+1}|\theta^\tau) - Q(\theta^\tau|\theta^\tau)\right) + \left(H(\theta^{\tau+1}|\theta^\tau) - H(\theta^\tau|\theta^\tau)\right),$$

where we have that $Q(\theta^{\tau+1}|\theta^\tau) - Q(\theta^\tau|\theta^\tau) \geq 0$ since $\theta^{\tau+1}$ is chosen to maximize that quantity $Q(\theta|\theta^\tau)$. A closer look at $H(\theta|\theta^\tau) - H(\theta^\tau|\theta^\tau)$ reveals that it is minus the Kullback–Leibler divergence between $Pr[\eta_i{=}\eta_k|y_i, \theta^\tau]$ and $Pr[\eta_i{=}\eta_k|y_i, \theta]$:

$$\begin{aligned} H(\theta^{\tau+1}|\theta^\tau) - H(\theta^\tau|\theta^\tau) &= \sum_i \sum_k Pr[\eta_i{=}\eta_k|y_i, \theta^\tau] \cdot \log \frac{Pr[\eta_i|y_i, \theta^\tau]}{Pr[\eta_i|y_i, \theta^{\tau+1}]} \\ &= D_{KL}(Pr[\eta_i|y_i, \theta^\tau], Pr[\eta_i|y_i, \theta^{\tau+1}]) \geq 0 \end{aligned}$$

and hence will always be negative. This shows that the likelihood increases at each step.

1.2 Discrete Mixed Multinomial Logit

we are interested in the following likelihood model

$$\begin{aligned} Pr[D_i = j|X_{ij}] &= \sum_k p_k Pr[D_i = j|X_{ij}, k_i = k] \\ &= \sum_k p_k \frac{\exp \beta_k X_{ij}}{\sum_{j'} \exp \beta_k X_{i'}} \end{aligned}$$

Applying the EM we start by computing the posterior probability $q_i(k)$. We call k_i the latent type. Lucky for us it is given by

$$\begin{aligned} q_i(k) &= \frac{Pr[k_i = k, D_i|X_i, \beta^K]}{\sum_{k'} Pr[k_i = k', D_i|X_i, \beta^K]} \\ &= \frac{\hat{p}_i(j; \beta_k) \cdot p_k}{\sum_{k'} \hat{p}_i(j; \beta_{k'}) \cdot p_{k'}} \end{aligned}$$

Then the next step is to maximize the weighted sum of the log likelihoods

$$\begin{aligned} Q(\beta|\beta^{(t)}) &= \sum_i \sum_k q_i(k) \log Pr[D_i = j|X_{ij}, k_i = k] \\ &= \sum_i \sum_k q_i(k) \log \left(\frac{\exp \beta_k X_{ij}}{\sum_{j'} \exp \beta_k X_{i'}} \right) \\ &= \sum_k \sum_i q_i(k) \log \left(\frac{\exp \beta_k X_{ij}}{\sum_{j'} \exp \beta_k X_{i'}} \right) \end{aligned}$$

so we end up with K different multinomial Logit optimizations.

1.3 Gaussian mixture

Let's run through using the EM algorithm in details for the finite Mixture model, considering a normal mixture model. We write the probability model as:

$$L(Y_1, Y_2, Y_3; \theta) = Pr[Y_1{=}y_1, Y_2{=}y_2, Y_3{=}y_3; \theta] = \sum_{k=1}^{K} p_k \prod_{t=1}^{3} \phi(Y_t; \mu_k, \sigma_k)$$

where the parameter space is given by $\theta = \{p_k, \mu_k, \sigma_k\}_{k=1..K}$. The Maximum Likelihood estimator solves:

$$\theta^{MLE} = \arg\max_{\theta} \sum_i \log L(Y_{i1}, Y_{i2}, Y_{i3}; \theta)$$

The EM algorithm is an iterative procedure that climbs the likelihood surface. Given a parameter guess $\theta = \{p_k, \mu_k, \sigma_k\}_{k=1..K}$, the algorithm is composed of two steps:

In the Expectation step we compute:

$$q_i(k) = Pr[\eta_i = \eta_k|Y_{i1}, Y_{i2}, Y_{i3}; \theta^\tau]$$

a natural procedure is to compute this posterior probability using the likelihood model:

$$\begin{aligned} q_i(k) &= \frac{Pr[\eta_i = \eta_k|Y_{i1}, Y_{i2}, Y_{i3}; \theta^\tau]}{\sum_l Pr[\eta_i = \eta_l|Y_{i1}, Y_{i2}, Y_{i3}; \theta^\tau]} \\ &= Pr[\eta_i = \eta_k|Y_{i1}, Y_{i2}, Y_{i3}; \theta^\tau] \\ &= \frac{p_k^\tau \prod_{t=1}^{3} \phi(Y_t; \mu_k^\tau, \sigma_k^\tau)}{\sum_{l=1}^{K} p_l^\tau \prod_{t=1}^{3} \phi(Y_t; \mu_l^\tau, \sigma_l^\tau)} \end{aligned}$$

The second step uses these probabilities in the maximization step:

$$\max_{\theta} \sum_i \sum_k q_i(k) \log Pr(Y_1, Y_2, Y_3, \eta_k; \theta)$$

where in our case

$$Pr(Y_1, Y_2, Y_3, \eta_k; \theta) = p_k \prod_{t=1}^{3} \phi(Y_t; \mu_k, \sigma_k)$$

where we get

$$\max_{\theta} \sum_i \sum_k q_i(k) \left(\log p_k^{\tau+1} - \log \sqrt{\pi} \sigma_k^{\tau+1} - \sum_t \frac{\left(Y_{it} - \mu_k^{\tau+1}\right)^2}{2\left(\sigma_k^{\tau+1}\right)^2} \right)$$

the FOC of p_k gives

$$p_k^{\tau+1} = \frac{\sum q_i(k)}{N}$$

and the FOCs for μ_k gives

$$\sum_i \sum_t q_i(k) \left(Y_{it} - \mu_k\right) = 0$$

or in other in other words:

$$\mu_k^{\tau} = \frac{\sum_i q_i(k) \sum_t Y_{it}}{T \cdot \sum_i q_i(k)}$$

and finally the estimate of the variance is given by

$$\sigma_k^{\tau+1} = \sqrt{\frac{\sum_i \sum_t q_i(k) \left(Y_{it} - \mu_k^{\tau+1}\right)^2}{T \cdot \sum_i q_i(k)}}$$

2 CCP approach

look at Aguirregabiria and Mira (2010)

Remember from the notes we are given :

1. transition probabilities $f(x_{t+1}|x_t, a_t; \theta_x)$
2. payoff function $f(y_t|a_t, x_t; \theta_y)$
3. preferences $u(a_t, y_t, x_t; \theta_u) + \epsilon_t(a_t)$
   (a) with $\mathbb{E}\left[u(a_t, y_t, x_t; \theta_u)|x_t\right] = u_1(a_t, x_t)'\theta_u + u_0(a_t, x_t)$

The agent wants to maximize present discounted utility

$$\mathbb{E}_0 \sum_{t=0}^{T} \beta^t u(a_t, y_t, x_t) + \epsilon_t(a_t)$$

we are given data (a_t, y_t, x_t).

2.1 the agent's problem

given the set of assumption we can write the agent's problem as:

$$\bar{V}(x_{it}) = \int \max_{a \in A} \left\{ \mathbb{E}_{y_t} u(a, x_{it}, y_t) + \epsilon_{it}(a) + \beta \sum_{x_{i,t+1}} \bar{V}(x_{i,t+1}) f_x(x_{i,t+1}|a, x_{it}) \right\}$$

$$= \log \sum_j \exp \left\{ \mathbb{E}_{y_t} u(a_j, x_{it}, y_t) + \beta \sum_{x_{i,t+1}} \bar{V}(x_{i,t+1}) f_x(x_{i,t+1}|a, x_{it}) \right\} + \gamma$$

from which we could derive numerically the random policy function function $\alpha_t(j, x; \theta) = Pr[A_t = a_j | X_t = x_{it}]$ which is a function of the full parameter vector θ. Here γ the Euler constant which is equal to 0.5772156649. Defining

$$v_t(j, x_t; \theta) = \mathbb{E}_{y_t} u(a_j, x_t, y_t) + \beta \sum_{x_{i,t+1}} \bar{V}(x_{t+1}) f_x(x_{t+1}|a_j, x_t)$$

we have that

$$\alpha_t(j, x; \theta) = Pr[A_t = a_j | X_t = x_{it}] = \frac{\exp v(j, x_{it}; \theta)}{\sum_{j'} \exp v(j', x_{it}; \theta)}$$

2.2 The likelihood

we want to write the following

$$\ell_i = \log Pr[x_0, a_0, y_0 ..... x_T, a_T, y_T | \theta]$$

in order to maximize it. We can express this in terms of model objects

$$\begin{aligned}
\ell_i &= \log Pr[x_0, a_0, y_0] + \sum_{t=1}^{T} \log Pr[x_t, a_t, y_t | x^{t-1}, a^{t-1}, y^{t-1}; \theta] \\
&= \log Pr[x_0, a_0, y_0] + \sum_{t=1}^{T} \log Pr[x_t, a_t, y_t | x_{t-1}, a_{t-1}, y_{t-1}; \theta] \\
&= \log Pr[x_0] + \log Pr[a_0 | x_0] + \log Pr[y_0 | x_0, a_0] + \\
&\qquad \sum_{t=1}^{T} \log Pr[x_t | x_{t-1}, a_{t-1}; \theta] + \log Pr[a_t | x_t; \theta] + \log Pr[y_t | a_t, x_t; \theta]
\end{aligned}$$

we notice that part of the likelihood is only a function of some subset of the parameters. In particular both θ_x and θ_y can be estimated first using:

$$\theta_x = \arg\max \sum_{i=1}^{n} \sum_{t=1}^{T} \log Pr[X_t = x_{it} | X_{t-1} = x_{it-1}, A_{t-1} = a_{it-1}; \theta_y]$$

$$\theta_y = \arg\max \sum_{i=1}^{n} \sum_{t=1}^{T} \log Pr[Y_t = y_{it} | X_t = x_{it}, A_t = a_{it}]$$

Next we are left with

$$\sum_{t=0}^{T} \log Pr[a_t | x_t; \theta] = \sum_{t=0}^{T} \sum_j \mathbf{1}[a_t = a_j] \cdot \log \alpha_t(j, x_{it}; \theta)$$

where all $(\theta_u, \theta_x, \theta_y)$ enter very non-linearly inside α_t. Remember that

$$\alpha_t(x_{it};\theta) = Pr[A_t = a_j | X_t = x_{it}] = \frac{\exp v(j, x_{it};\theta)}{\sum_{j'} \exp v(j', x_{it};\theta)}.$$

Can we do something about it? Given the assumption that $\tilde{u}(a_t, y_t, x_t)'\theta_u + \epsilon_t(a_t)$ we can show that

$$v_t(j, x_{it};\theta) = \tilde{v}_{1t}(j, x_t;\theta)'\theta_u + \tilde{v}_{0t}(j, x_t;\theta)$$

where we will then look for an estimate of $\tilde{v}$ to plugin. Let's start with T and assume that $v_{T+1} = 0$, then

$$\begin{aligned} v_T &= \mathbb{E}_{y_t}\left[u(a_j, x_t, y_t)|x_t\right] \\ &= \tilde{u}_1(a_t, x_t)'\theta_u + \tilde{u}_0(a_t, x_t) \end{aligned}$$

and so we get that

$$\begin{aligned} \tilde{v}_{1T}(j, x_t;\theta) &= \tilde{u}_1(a_t, x_t) \\ \tilde{v}_{0T}(j, x_t;\theta) &= \tilde{u}_0(a_t, x_t) \end{aligned}$$

Now assuming the additive separability is true at t let's show it is the case at $t-1$

$$\begin{aligned} v_t(j, x_t;\theta) &= \mathbb{E}_{y_t} u(a_j, x_t, y_t) + \beta \sum_{x_{i,t+1}} \bar{V}(x_{t+1}) f_x(x_{t+1}|a_j, x_t) \\ &= \tilde{u}_1(a_t, x_t)'\theta_u + \tilde{u}_0(a_t, x_t) + \beta \sum_{x_{i,t+1}} \bar{V}(x_{t+1}) f_x(x_{t+1}|a_j, x_t) \\ &= \tilde{u}_1(a_t, x_t)'\theta_u + \tilde{u}_0(a_t, x_t) + \beta \sum_{x_{i,t+1}} \sum_j \alpha_{t+1}(j, x_{t+1};\theta) v_{t+1}(j, x_{t+1};\theta) f_x(x_{t+1}|a_j, x_t) \\ &= \left[\tilde{u}_1(a_t, x_t) + \beta \sum_{x_{i,t+1}} \sum_j \alpha_{t+1}(j, x_{t+1};\theta) \tilde{v}_{1t+1}(j, x_{t+1};\theta) f_x(x_{t+1}|a_j, x_t)\right]' \theta_u + \\ &\quad \left[\tilde{u}_0(a_t, x_t) + \beta \sum_{x_{i,t+1}} \sum_j \alpha_{t+1}(j, x_{t+1};\theta) \tilde{v}_{0t+1}(j, x_{t+1};\theta) f_x(x_{t+1}|a_j, x_t)\right] \end{aligned}$$

and so we get a recursive expression for $\tilde{v}_{0t+1}$ and $\tilde{v}_{1t+1}$. We have that:

$$\begin{aligned} \tilde{v}_{1t}(j, x_t;\theta) &= \tilde{u}_1(a_t, x_t) + \beta \sum_{x_{i,t+1}} \sum_j \alpha_{t+1}(j, x_{t+1};\theta) \tilde{v}_{1t+1}(j, x_{t+1};\theta) f_x(x_{t+1}|a_j, x_t) \\ \tilde{v}_{0t}(j, x_t;\theta) &= \tilde{u}_0(a_t, x_t) + \beta \sum_{x_{i,t+1}} \sum_j \alpha_{t+1}(j, x_{t+1};\theta) \tilde{v}_{0t+1}(j, x_{t+1};\theta) f_x(x_{t+1}|a_j, x_t) \end{aligned}$$

However we still have the problem that we need $\alpha_{t+1}(j, x_{t+1};\theta)$ to construct such quantities. However $Pr[A_t = a_j | X_t = x_{it}]$ in this case is directly observed in the data. It is pretty much just a frequency. So equipped with an $\alpha_{t+1}(j, x_{t+1};\theta)$ we can construct $\tilde{v}_{1t}(j, x_t;\theta), \tilde{v}_{0t}(j, x_t;\theta)$. We are then left with the following likelihood:

$$\sum_{i=1}^{n}\sum_{t=0}^{T} \log Pr[A_t = a_{it} | X_t = x_{it}] = \sum_{i=1}^{n}\sum_{t=0}^{T} \log \frac{\exp\left[\tilde{v}_{1t}(j_{it}, x_t;\theta)'\theta_u + \tilde{v}_{0t}(j_{it}, x_t;\theta)\right]}{\sum_{j'} \exp\left[\tilde{v}_{1t}(j', x_t;\theta)'\theta_u + \tilde{v}_{0t}(j', x_t;\theta)\right]}$$

given that $\tilde{v}_{1t}(j, x_t;\theta)$ and $\tilde{v}_{0t}(j, x_t;\theta)$ are known, this is a multinomial logit to find θ_u.

References

Aguirregabiria, V., and P. Mira (2010): "Dynamic discrete choice structural models: A survey," *J. Econom.*, 156(1), 38–67.