The EM Algorithm

Econ 674

Purdue University

- Today, we discuss the *EM Algorithm*.

- This algorithm is very useful in nonlinear models, many of which are linear in suitably defined *latent data*.

- It's name comes from two steps: First, an *Expectation* step, where expectations are taken with respect to the latent data, given the observed data and a particular state of the parameter vector. The second step is a a *maximization* step.

- In the following slides we offer an explanation behind why the method works and illustrate its application for the probit and Gaussian mixture models.

EM Algorithm

First, let us define some notation. Let

$$y = g(y^*)$$

be the link between the latent data y^* and observed data y.
Denote the density of y^* as

$$f(y^*|\theta)$$

and let

$$L(\theta; y^*) = \log f(y^*|\theta).$$

[In the context of the probit, for example, $f(y^*|\theta) = \phi(y^*|x\beta, I_n)$.]
Finally, define

$$Q(\theta, \theta_t; y) = E\left[L(\theta; y^*)\right] = E_{y^*|\theta=\theta_t, Y=y}\left[L(\theta; y^*)\right].$$

EM Algorithm

Theorem

Whenever

$$Q(\theta, \theta_t; y) > Q(\theta_t, \theta_t; y)$$

it must be the case that

$$L(\theta; y) > L(\theta_t; y).$$

Let's pause to appreciate what the theorem states. If we define in an iterative fashion, for example,

$$\theta_t = \text{argmax}_\theta Q(\theta, \theta_{t-1}; y),$$

then the sequence of θ_t values obtained in this fashion lead us to higher values of the log likelihood. So, if the expectation and maximization are easily performed, this provides an alternative to traditional MLE.

EM Algorithm

We will sketch a proof of this theorem. First, note that

$$f_{y|y^*}(y|y^*) = I[y = g(y^*)].$$

That is, the distribution of y is degenerate given y^*.
Now, consider:

$$\begin{aligned} p(y, y^*|\theta) &= p(y|y^*, \theta)p(y^*|\theta) \\ &= p(y|y^*)p(y^*|\theta) \\ &= f(y^*|\theta)I[y = g(y^*)] \end{aligned}$$

EM Algorithm

From this joint distribution we seek to obtain $f(y^*|y,\theta)$. We note:

$$f(y^*|y,\theta)f(y|\theta) = p(y,y^*|\theta).$$

Therefore,

$$f(y^*|y,\theta) = \frac{p(y,y^*|\theta)}{f(y|\theta)}$$

or

$$f(y^*|y,\theta) = \frac{f(y^*|\theta)}{f(y|\theta)} I[y = g(y^*)].$$

EM Algorithm

Therefore, the log-likelihood for θ given y^* drawn from $y^*|y,\theta$ is

$$\begin{aligned} L(\theta; y^*|y) &= \log f_{y^*|y,\theta}(y^*|y,\theta) \\ &= \log\left[f(y^*|\theta)/f(y|\theta)\right] \\ &= \log f(y^*|\theta) - \log f(y|\theta) \\ &= L(\theta; y^*) - L(\theta; y) \end{aligned}$$

Note that the second line follows since the sampling is from $y^*|y,\theta$.

EM Algorithm

Now, let

$$H(\theta, \theta_t; y) \equiv Q(\theta, \theta_t; y) - L(\theta; y).$$

It follows that

$$\begin{aligned} H(\theta, \theta_t; y) &= Q(\theta, \theta_t; y) - L(\theta; y) \\ &= E_{y^*|y,\theta=\theta_t}\left[L(\theta; y^*)\right] - L(\theta; y) \\ &= E_{y^*|y,\theta=\theta_t}\left[L(\theta; y^*) - L(\theta; y)\right] \\ &= E_{y^*|y,\theta=\theta_t}\left[L(\theta; y^*|y)\right] \end{aligned}$$

using our notation above. By Jensen's inequality (like our proof for the expected log likelihood inequality), it is clear that $H(\theta, \theta_t; y)$ is maximized at $\theta = \theta_t$.

EM Algorithm

Thus,

$$H(\theta_t, \theta_t; y) \geq H(\theta, \theta_t; y)$$

which is equivalent to:

$$Q(\theta_t, \theta_t; y) - L(\theta_t; y) \geq Q(\theta, \theta_t; y) - L(\theta; y)$$

or, after rearranging,

$$L(\theta; y) - L(\theta_t; y) \geq Q(\theta, \theta_t; y) - Q(\theta_t, \theta_t; y).$$

This completes the proof. That is, whenever θ is chosen such that $Q(\theta, \theta_t; y) > Q(\theta_t, \theta_t; y)$ it is necessarily the case that $L(\theta; y) > L(\theta_t; y)$. That is, we can iterate to the maximum likelihood estimate.

The EM Algorithm

In practice, the EM algorithm chooses:

$$\theta_{t+1} = \text{argmax}_\theta Q(\theta, \theta_t; y).$$

Thus,

1. With θ_{t+1} defined in this way, it is clear that all updates to new θ values can not decrease the value of the log-likelihood.

2. In practice, the current value θ_t is treated as the "true" parameter vector, and expectations are taken assuming $\theta = \theta_t$. Q, however, remains a function of both θ and θ_t, and setting $\theta_{t+1} = \theta_t$ is not optimal in general.

3. Two examples illustrate use of the EM algorithm.

Probit Example

We illustrate the practical usefulness of the EM algorithm in fitting the probit model:

$$y^* = X\beta + \epsilon, \quad \epsilon | X \overset{iid}{\sim} \mathcal{N}(0, I_n).$$

$$y_i = I(y_i^* > 0).$$

Step 1: E-Step

We need to get $L(\theta; y^*)$. For the probit model, this is easy since:

-

Probit Example

We now need to take the expectation of $L(\beta; y^*)$ over $y^*|y, \beta = \beta_t$. Expanding the quadratic, we get:

$$\begin{aligned} Q(\beta, \beta_t; y) &= -\frac{n}{2}\log(2\pi) - \frac{1}{2}E(y^{*\prime}y^*|\beta = \beta_t, y) \\ &+ \beta' X' E(y^*|\beta = \beta_t, y) - \frac{1}{2}\beta' X' X \beta. \end{aligned}$$

Let

$$\mu(\beta_t, y) \equiv E(y^*|\beta = \beta_t, y).$$

This completes the E-step.

Probit Example

Step 2: M-Step

Using the μ-notation, we can write

$$Q(\beta, \beta_t; y) = c(y^*, \beta_t) + \beta' X' \mu(\beta_t, y) - \frac{1}{2}\beta' X' X \beta.$$

for some c that does not involve β. So

-

Since this is just like least-squares, we obtain:

-

Probit Example

It remains for us to characterize the conditional expectation $E(y^*|\beta = \beta_t, y)$.

Suppose $y = 1$. Then

-

Likewise,

$$E(y^*|\beta = \beta_t, y = 0) = X\beta_t - \frac{\phi(X\beta_t)}{1 - \Phi(X\beta_t)}.$$

So, generally,

$$\mu(\beta_t, y) = X\beta_t + \frac{\phi(X\beta_t)}{\Phi(X\beta_t)[1 - \Phi(X\beta_t)]}\left[y - \Phi(X\beta_t)\right].$$

Probit Example

Putting all the pieces together, application of the EM algorithm to the probit proceeds as follows:

1. Pick a starting value, say β_0.
2. Calculate $\mu(\beta_0, y)$ using the formula on the last slide.
3. Regress $\mu(\beta_0, y)$ on X to obtain β_1.
4. Repeat the process to obtain β_2, β_3, etc.
5. Iterate until the difference in log likelihoods (or β) is negligable.

Mixtures Example

Our second example relates to the use of *Gaussian mixtures*.

Mixture models are rapidly increasing in popularity, for a number of reasons. The most common reasons people use mixtures in practice are:

1. *Added flexibility* - with enough mixture components, you can approximate any well-behaved density with an arbitrary degree of accuracy.

2. The population of interest is known to be comprised of a *discrete set of subgroups*.

Mixtures Example

To fix ideas, we consider the simplest case of a two-component Gaussian mixture:

-

We can define a latent *component indicator* variable z_i as follows:

$$z_i = \begin{cases} 1 & \text{if person i is ``drawn from'' the first component} \\ 0 & \text{if person i is ``drawn from'' the second component} \end{cases}$$

It follows that

$$p(y_i|z_i, \theta) = \phi(y_i; \mu_1, \sigma_1^2)^{z_i} \phi(y_i; \mu_2, \sigma_2^2)^{1-z_i}$$

and we define

$$\Pr(z_i = 1|\theta) = \pi.$$

(So, after integrating out z, we have the same likelihood).

Mixtures Example

Note

$$p(z_i|y_i,\theta) \propto \pi^{z_i}(1-\pi)^{1-z_i}\left[\phi_{1i}^{z_i}\phi_{2i}^{1-z_i}\right]$$

where $\phi_{ji} \equiv \phi(y_i;\mu_j,\sigma_j^2),\ j=1,2$, so that

$$\Pr(z_i=1|y_i,\theta) \propto \pi\phi_{1i}$$

and

$$\Pr(z_i=0|y_i,\theta) \propto (1-\pi)\phi_{2i}$$

Mixtures Example

Scaling these quantities up to make the conditional density proper, we obtain:

-

and

$$\Pr(z_i = 0|y_i, \theta) = \frac{(1-\pi)\phi_{2i}}{\pi\phi_{1i} + (1-\pi)\phi_{2i}}$$

Hence,

$$E(z_i|y_i, \theta = \theta_t) \equiv \tau_i(\theta_t, y_i) = \frac{\pi^{(t)}\phi_{1i}^{(t)}}{\pi^{(t)}\phi_{1i}^{(t)} + (1-\pi^{(t)})\phi_{2i}^{(t)}}.$$

Mixtures Example

The (log) joint density of observed and latent data is:

$$\log p(y, z|\theta) = \sum_i z_i \left[\log \phi_{1i} + \log \pi\right] + \sum_i (1 - z_i) \left[\log \phi_{2i} + \log(1 - \pi)\right].$$

Therefore,

$$Q(\theta, \theta_t; y) = \sum_i \tau_i(\theta_t, y_i) \left[\log \phi_{1i} + \log \pi\right] + \left[1 - \tau_i(\theta_y, y_i)\right] \left[\log \phi_{2i} + \log(1 - \pi)\right].$$

This *concludes the E-step*. As for the M-step, consider the FOC for π:

$$\sum_i \left[(1/\pi_{t+1})\tau_i(\theta_t, y_i) - \left[1 - \tau_i(\theta_t, y_i)\right] \frac{1}{1 - \pi_{t+1}}\right] = 0.$$

This yields, after some algebra:

$$\pi_{t+1} = \frac{1}{n} \sum_{i=1}^{n} \tau_i(\theta_t, y_i).$$

Mixtures Example

Next, consider μ_1. (A result for μ_2 will follow analogously). The relevant term in Q is:

$$\sum_i \tau_i \left[-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log[\sigma_1^2] - \frac{1}{2\sigma_1^2}(y_i - \mu_1)^2 \right].$$

Differentiating with respect to μ_1 gives the FOC:

$$\sum_i \tau_i(\theta_t, y_i)(y_i - \mu_{1,t+1}) = 0$$

yielding

Mixtures Example

Finally, consider σ_1^2. (A result for σ_2^2 will follow analogously). The FOC from Q is:

$$-\frac{1}{2}\frac{1}{\sigma_{1,t+1}^2}\sum_i \tau_i + \frac{1}{2\sigma_{1,t+1}^4}\sum_i \tau_i(y_i - \mu_{1,t+1})^2 = 0.$$

Yielding

$$\sigma_{1,t+1}^2 = \frac{\sum_i \tau_i(\theta_t, y_i)(y_i - \mu_{1,t+1})^2}{\sum_i \tau_i(\theta_t, y_i)}.$$

Mixtures Example

If *covariates* are included in the model so that, for example,

$$\phi_{1i} = \phi(y_i; x_i\beta_1, \sigma_1^2)$$

then

1. τ_i is defined in the same way, making the above replacements for ϕ_{1i} and ϕ_{2i}

2. $$\beta_{1,t+1} = (X'TX)^{-1}X'Ty,$$

   where

   $$T = diag\{\tau_i(\theta_t, y_i)\}.$$

3. $$\sigma^2_{1,t+1} = \frac{\sum_i \tau_i(y_i - x_i\beta_{1,t+1})^2}{\sum_i \tau_i}.$$

Mixtures Example

- Though we have illustrated things here for the case of two components, this generalizes easily to the arbitrary case with k components.
- Essentially, the results we obtained for each component are simply repeated for the additional mixture components.
- This also generalizes in a straightforward way to the case of *multivariate* data in which case σ_j is replaced by Σ_j.

The next slide illustrates results from a generated data experiment.

We generate $n = 10,000$ observations from a Lognormal(1,.1) distribution.

We then plot the true density function against a two-component Gaussian mixture approximation.

The mixture is fit via the EM algorithm.

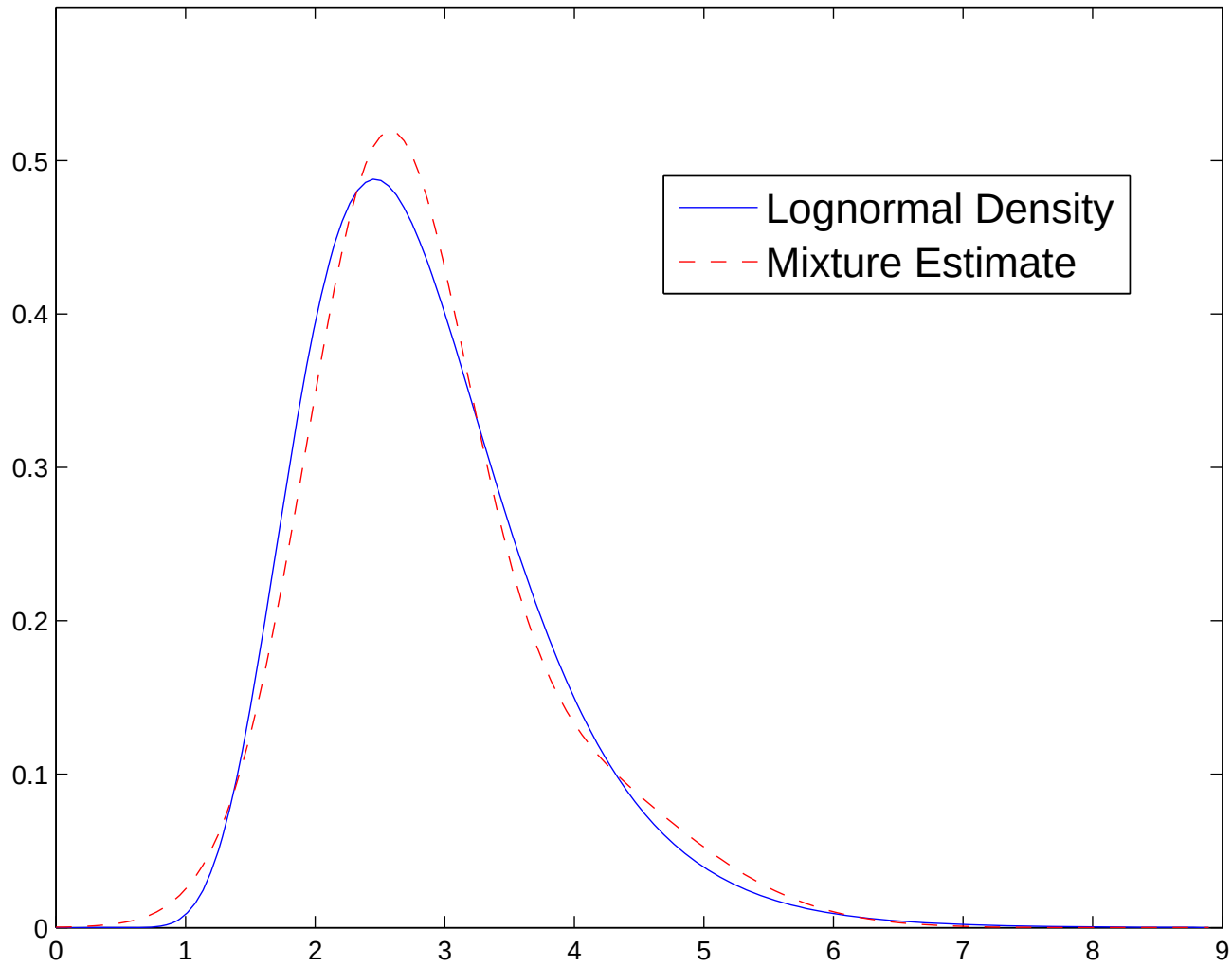

Figure: 2 Component Mixture: $.661\phi(x; 2.47, .357) + .339\phi(x; 3.60, 1.00)$

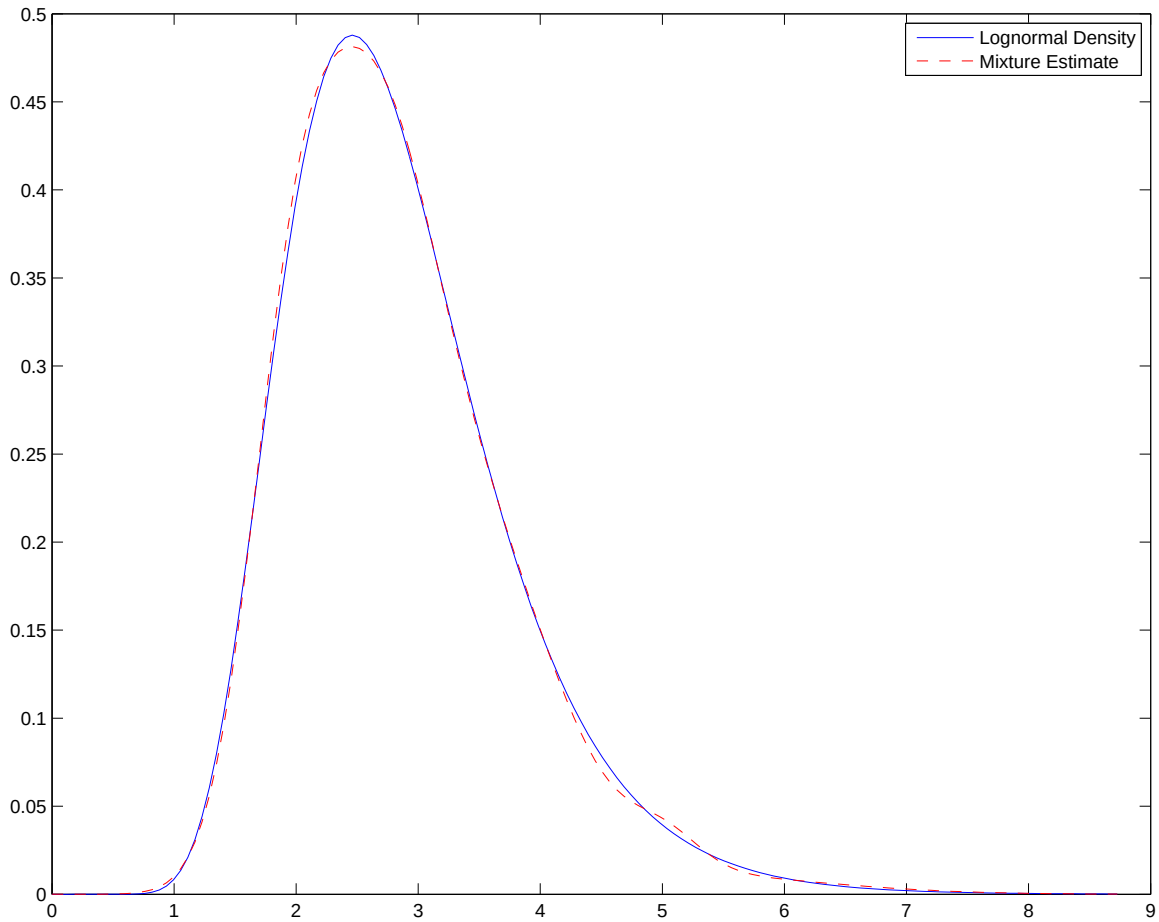

Figure: 5 Component Mixture: $.217\phi(x; 1.95, .138) + .406\phi(x; 2.60, .238) + .328\phi(x; 3.43, .453) + .014\phi(x; 5.01, .086) + .035\phi(5.02, 1.44)$