Econometrics I

Lecture 11: Maximum Likelihood Estimation

Paul T. Scott
NYU Stern

Fall 2018

Logistics

- PS3 grades, solutions posted
- You should have heard something from me about your projects
- Remaining schedule:
  - 11/15 – Discrete Choice (Chris Conlon guest lecture)
  - 11/22 – No class – Happy Thanksgiving!
  - 11/29 – Workshop with Skand (let us know if there's anything you'd like to review)
  - 12/6 – Last class: group project presentations
  - 12/13 – Group projects due by email
- Group presentations: 12 minutes for individuals (7), 15 minutes for groups (4)
  - 144 minutes total: need to stay on schedule!
  - Pizza? Falafel?

Question

As we all know, if you only have firm fixed effect in a regression, what you analyze is the variation within a firm. Similarly, if you only have time fixed effect in a regression, what you analyze is the variation within a time. My question is how we should interpret results if we have both firm fixed effect and time fixed effect in one regression. Do we look at the variation both within a firm and within a time? This interpretation seems weird to me. So, I am not exactly sure which data variation is used if we have both firm fixed effect and time fixed effect.

Likelihood

- What's a **likelihood**? It's basically the probability of the data conditional on a parameter value θ:

$$Pr\left(\text{observed data}|\theta\right),$$

  but we think of this as a function of θ and telling us something about the plausibility of θ.

- This requires we have a model that says what the probability of the data is.

- $\Rightarrow$ In comparison to GMM estimation, Likelihood-based estimation requires strong assumptions about the data generating process.

Likelihood Function

- Let $f(\cdot|\boldsymbol{\theta})$ represent the probability density of the data conditional on a parameter value $\boldsymbol{\theta}$. If data are independently and identically distributed, the **likelihood function** is

$$L(\boldsymbol{\theta}|\mathbf{y}) = f(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_n|\boldsymbol{\theta}) = \prod_{i=1}^{n} f(\mathbf{y}_i|\boldsymbol{\theta})$$

  where $\mathbf{y}_i$ indicates individual observations (including both dependent and explanatory variables).

- We typically work with **log-likelihood function** because it's computationally simpler:

$$\ln L(\boldsymbol{\theta}|\mathbf{y}) = \sum_{i=1}^{n} \ln f(\mathbf{y}_i|\boldsymbol{\theta}).$$

Maximum Likelihood Estimation

- **Maximum likelihood estimation** entails estimating $\boldsymbol{\theta}$ by maximizing the likelihood function:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\theta} L\left(\boldsymbol{\theta}|\mathbf{y}\right) = \arg\min_{\theta} \ln L\left(\boldsymbol{\theta}|\mathbf{y}\right)$$

- Since the natural log function is strictly increasing, maximizing the likelihood and maximizing log likelihood amount to the same thing.

Likelihood of Normal Errors

- Recall that PDF of normal distribution is

$$f_{\mathcal{N}}(\varepsilon|\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-\varepsilon^2}{2\sigma^2}\right)$$

(for normal ε with zero mean and variance σ^2)

- Thus, log likelihood of an individual observation of ε_i is

$$\ln f_{\mathcal{N}}(\varepsilon_i|\sigma) = -\frac{1}{2}\left(\ln \sigma^2 + \ln 2\pi + \frac{\varepsilon_i^2}{\sigma^2}\right)$$

Likelihood for Linear Regression Model

- For linear model with ε_i mean-zero normal *conditional* on $\mathbf{x}_i$, the likelihood of one observation is

$$L(\boldsymbol{\beta}, \sigma | y_i, \mathbf{x}_i) = f_{\mathcal{N}}\left(y_i - \mathbf{x}_i' \boldsymbol{\beta} | \sigma\right)$$

  noting that this requires the distribution of ε_i to be mean-zero normal *conditional* on $\mathbf{x}_i$.

- Assuming the data are i.i.d across observations, the conditional likelihood of all the data is then

$$\begin{aligned} \ln L(\boldsymbol{\beta}, \sigma | \mathbf{y}, \mathbf{X}) = & \textstyle\sum_{i=1}^{n} \ln f_{\mathcal{N}}\left(y_i - \mathbf{x}_i' \boldsymbol{\beta} | \sigma\right) \\ = & -\frac{1}{2} \textstyle\sum_{i=1}^{n} \left(\ln \sigma^2 + \ln 2\pi + \frac{\left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right)^2}{\sigma^2} \right) \end{aligned}$$

MLE for Linear Model I

- Linear model log-likelihood:

$$\ln L(\boldsymbol{\beta}, \sigma | \mathbf{y}, \mathbf{X}) = -\frac{1}{2} \sum_{i=1}^{n} \left(\ln \sigma^2 + \ln 2\pi + \frac{\left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right)^2}{\sigma^2} \right)$$

- Focus on the term that involves β:

$$\frac{-1}{2\sigma^2} \sum_{i=1}^{n} \left(y_i - \mathbf{x}_i' \boldsymbol{\beta}\right)^2$$

NB: maximizing the likelihood with respect to β is equivalent to least squares

MLE for Linear Model II

- MLE estimate of $\boldsymbol{\beta}$ is the same as OLS.
- MLE estimate of σ^2 comes from setting $\frac{d}{d\sigma} \ln L\left(\hat{\boldsymbol{\beta}}, \sigma | \mathbf{y}, \mathbf{X}\right) = 0$:

$$\hat{\sigma}^2_{MLE} = n^{-1} \sum_{i=1}^{n} e_i^2$$

where $e_i = y_i - \mathbf{x}_i' \hat{\boldsymbol{\beta}}$.

- Note that this is a bit different than the estimate of σ^2 we saw before:

$$s^2 = (n-K)^{-1} \sum_{i=1}^{n} e_i^2$$

but the difference will be small in large samples. Recall: s^2 is a unbiased estimate of σ^2, so this means that the ML estimate is biased, and substantially biased in small samples.

Asymptotic Efficiency

- An estimator is **asymptotically efficient** if its asymptotic covariance matrix is not larger than any other consistent estimator (i.e., standard errors are as small as any other estimator).
- It can be shown that (under regularity conditions), MLE is asymptotically efficient.
- Thus, MLE always performs well in large samples.

Estimating Standard Errors I

- The first way to estimate the asymptotic covariance matrix is to take second derivatives of the likelihood function:

$$\mathbf{\Gamma}^{-1} = \left(-\frac{\partial^2 \ln L\left(\hat{\boldsymbol{\theta}}\right)}{\partial \hat{\boldsymbol{\theta}} \partial \hat{\boldsymbol{\theta}}'} \right)^{-1}$$

- A second way is to compute the covariance of the first derivatives:

$$\mathbf{S}^{-1} = \left[\sum_{i=1}^{n} \hat{\mathbf{g}}_i \hat{\mathbf{g}}_i' \right]^{-1}$$

  where

$$\hat{\mathbf{g}}_i = \frac{\partial \ln f\left(\mathbf{x}_i, \hat{\boldsymbol{\theta}}\right)}{\partial \hat{\boldsymbol{\theta}}}.$$

- Either of the above is an asymptotically consistent estimator of $V\left(\hat{\theta}_{MLE}\right)$. The latter is usually easier to compute.

MLE as GMM

- To maximize the likelihood function we set

$$n^{-1}\sum_{i=1}^{n}\hat{\mathbf{g}}_i = n^{-1}\sum_{i=1}^{n}\frac{\partial \ln f\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}\right)}{\partial\hat{\boldsymbol{\theta}}} = 0.$$

Thus, maximum likelihood is a GMM estimator based on moments

$$E\left[\frac{\partial \ln f\left(\mathbf{x}_i,\hat{\boldsymbol{\theta}}\right)}{\partial\hat{\boldsymbol{\theta}}}\right] = 0.$$

- The GMM estimator for the asymptotic covariance matrix has the form

$$\left(\boldsymbol{\Gamma}\mathbf{S}^{-1}\boldsymbol{\Gamma}\right)^{-1},$$

but in the MLE context it can be shown that $\mathbf{S}$ and $\boldsymbol{\Gamma}$ are asymptotically equivalent, so they effectively cancel and we can use either $\mathbf{S}^{-1}$ or $\boldsymbol{\Gamma}^{-1}$ to estimate the variance.

Conditional Likelihood I

- Our starting point was that likelihoods were about the probability of the data conditional on a parameter value:

$$\ln L\left(\boldsymbol{\theta}|\text{data}\right)=\sum_{i=1}^{n}\ln f\left(\text{data}_i|\boldsymbol{\theta}\right).$$

- The above derivation was about ε_i, or the probability of $y_i|x_i$. But x_i might be a random variable, and it's also part of the data.
- Do we need to consider the randomness in x_i? In econometric models, typically we don't bother to explicitly model the randomness in explanatory variables.

Conditional Likelihood II

- Start with the full log likelihood function

$$\sum_{i=1}^{n} \ln p\left(y_i, \mathbf{x}_i | \boldsymbol{\alpha}\right)$$

- We can decompose this using $Pr\left(y_i, \mathbf{x}_i\right) = Pr\left(y_i | \mathbf{x}_i\right) Pr\left(\mathbf{x}_i\right)$:

$$\sum_{i=1}^{n} \ln f\left(y_i | \mathbf{x}_i, \boldsymbol{\theta}\right) + \sum_{i=1}^{n} \ln g\left(\mathbf{x}_i, \boldsymbol{\delta}\right)$$

  where $\boldsymbol{\theta}$ is the subset of $\boldsymbol{\alpha}$ that dictates the distribution of $y_i | \mathbf{x}_i$ and $\boldsymbol{\delta}$ is the subset of $\boldsymbol{\alpha}$ that dictates the distribution of $\mathbf{x}_i$.
- If we're only interested in $\boldsymbol{\theta}$, then as long as there are no restrictions between $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$, we can just focus on the first component of the likelihood function (i.e., the **conditional likelihood** function)

Endogeneity

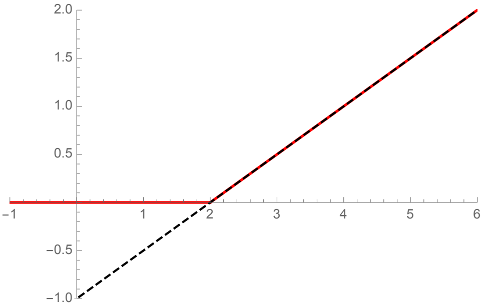
- Note that the likelihood framework does not solve the endogeneity problem.
- The consistency of MLE relies on the model being correctly specified, and when ε_i and $\mathbf{x}_i$ are correlated, the mean of ε_i is generally non-zero conditional on $\mathbf{x}_i$.
- Full information maximum likelihood (FIML) and limited information maximum likelihood (LIML) are the ML analog of IV estimators.

Application: Censored Regression Model I

- Censored data is a common problem
  - Demand for a concert/sporting event with capacity constraints.
  - Meters often only measure outcomes within a bounded range (speedometers, thermometers, etc.)
  - A test is scored on a bounded range (200-800), and we're thinking of the test as marker for ability.

Application 1: Censored Regression Model II

$$
\begin{aligned}
y_i^* &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i & & \textbf{latent variable} \text{ (black dashed)} \\
y_i &= 0 & \text{if } y_i^* \leq 0 & \quad \text{(red line)} \\
y_i &= y_i^* & \text{if } y_i^* > 0 & \quad \text{(red line)}
\end{aligned}
$$



How would we go about estimating this model?

Background: Truncated Normal

- Suppose v is distributed with standard normal PDF, but only for values above a cutoff a.
- PDF will be
$$\frac{\phi(v)}{1-\Phi(a)}$$
where ϕ is the standard normal PDF and Φ is standard normal CDF.
- Note that we must divide by $1-\Phi(a)$ to make the PDF integrate to 1.

Truncated Normal Moments I

Truncated Normal Properties

Suppose $v \sim \mathcal{N}(0,1)$ has a normal distribution truncated with $v > a$. That is, v takes values in (a, ∞) and has PDF

$$\frac{\phi(v)}{1 - \Phi(a)}.$$

Then,

$$\begin{aligned} E[v] &= \frac{\phi(a)}{1-\Phi(a)} \\ Var[v] &= \left(1 - \frac{\phi(a)}{1-\Phi(a)}\left(\frac{\phi(a)}{1-\Phi(a)} - a\right)\right) \end{aligned}$$

The ratio of a normal density to its CDF, $\frac{\phi(v)}{1-\Phi(a)}$, is known as the **inverse Mills ratio**.

Truncated Normal Moments II

- If original distribution is $v \sim \mathcal{N}\left(\mu, \sigma^2\right)$, truncated for $v > a$, we get similar results:

$$\begin{aligned} E\left[v\right] = & \quad \mu + \sigma \frac{\phi(\alpha)}{1-\Phi(\alpha)} \\ Var\left[v\right] = & \quad \sigma^2 \left(1 - \frac{\phi(\alpha)}{1-\Phi(\alpha)} \left(\frac{\phi(\alpha)}{1-\Phi(\alpha)} - \alpha\right)\right) \end{aligned}$$

where $\alpha = \frac{a-\mu}{\sigma}$.

- If truncation is for $v < a$, then we replace $\frac{\phi(\alpha)}{1-\Phi(\alpha)}$ with $-\frac{\phi(\alpha)}{\Phi(\alpha)}$

Censored Normal

- Suppose $v^* \sim \mathcal{N}\left(\mu, \sigma^2\right)$. Consider

$$v = \begin{cases} v^* & \text{if } v^* > a \\ a & \text{if } v^* \leq a \end{cases}$$

- Note: v will have the normal PDF above the cutoff a, and there will be a point mass at $v = a$.
- $Pr\left(v = a\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$ where Φ is the standard normal CDF.

Censored Normal Mean

- Censored Normal will have mean

$$\begin{aligned} E(v) &= E(v|v=a)\,Pr(v=a) + E(v|v>a)\,Pr(v>a) \\ &= a\Phi + E(v|v>a)(1-\Phi) \\ &= a\Phi + (\mu + \sigma\lambda)(1-\Phi) \end{aligned}$$

where $\lambda = \frac{\phi(\alpha)}{1-\Phi(\alpha)}$, $\Phi = \Phi(\alpha)$, $\alpha = \frac{a-\mu}{\sigma}$

- We can similarly derive the variance from the truncated normal variance

$$Var(v) = \sigma^2 (1-\Phi)\left[(1-\delta) + (\alpha-\lambda)^2 \Phi\right]$$

where $\delta = \lambda^2 - \lambda\alpha$.

Censored Regression

- Let's now return to censored regression framework:

$$\begin{aligned} y_i^* &= \mathbf{x}_i\boldsymbol{\beta} + \varepsilon_i & \\ y_i &= 0 & \text{if } y_i^* \leq 0 \\ y_i &= y_i^* & \text{if } y_i^* > 0 \end{aligned}$$

- What do you expect to happen if we estimate with OLS?
- What if we drop the observations with $y_i = 0$?

Censored Regression: Conditional Means

- Assuming ε_i is normal, the formula for the censored normal implies

$$E\left[y|\mathbf{x}\right] = \Phi\left(\frac{\mathbf{x}'\boldsymbol{\beta}}{\sigma}\right)\left(\mathbf{x}'\boldsymbol{\beta} + \sigma\frac{\phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right)}{\Phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right)}\right)$$

  which implies that OLS applies to full data set is biased.
- Using there results from the truncated normal,

$$E\left[y|\mathbf{x}, y > 0\right] = \left(\mathbf{x}'\boldsymbol{\beta} + \sigma\frac{\phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right)}{\Phi\left(\mathbf{x}'\boldsymbol{\beta}/\sigma\right)}\right)$$

  which implies that OLS applies to non-censored data set is biased.

Censored Regression: ML Estimation (Tobit)

- Log likelihood equation:

$$\ln L = \sum_{y_i > 0} -\frac{1}{2}\left[\ln(2\pi) + \ln\sigma^2 + \frac{(y_i - \mathbf{x}_i'\beta)}{\sigma^2}\right] + \sum_{y_i = 0} \ln\left(1 - \Phi\left(\frac{\mathbf{x}_i'\beta}{\sigma}\right)\right)$$

- Maximum likelihood here will give consistent (and asymptotically efficient) estimates of all parameters.
- This is known as a **tobit** regression.
- These mathematical tools are also what's behind the **Heckman selection correction** to deal with *sample selection bias*.

Application 2: Finite Mixture Models

- x - observed variables
- ζ - unobserved variables assumed to have finite support, Z
- θ parameters of interest

- $p(x_i, \zeta_i|\theta)$ - complete data likelihood for ith observation
- $p(x_i|\theta)$ - incomplete data likelihood for ith observation:

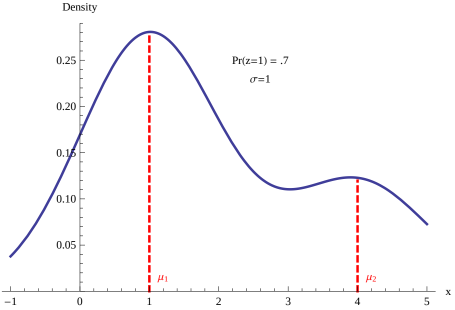
$$p(x_i|\theta) = \sum_{z \in Z} p(x_i, z|\theta)$$

- $q_{iz}(\theta)$ - expectation of incomplete data

$$q_{iz}(\theta) = Pr(\zeta_i = z|x_i, \theta)$$

Example 1: Mixture of Normals

- $\theta = (\mu_1, \mu_2, \sigma, \alpha_1)$
- If $z_i = 1$, then $x_i \sim N(\mu_1, \sigma)$
- If $z_i = 2$, then $x_i \sim N(\mu_2, \sigma)$
- $Pr(z_i = 1) = \alpha_1$



Example 2: collusion (Porter, 1983)

- Rob Porter (1983), "A Study of Cartel Stability: The Joint Executive Committee, 1880-1886"
- $$\begin{aligned} \ln Q_t &= \alpha_0 + \alpha_1 \ln P_t + \alpha_2 D_t + U_{1t} \\ \ln P_t &= \beta_0 + \beta_1 \ln Q_t + \beta_2 S_t + \beta_3 I_t + U_{2t} \end{aligned}$$

  where
  - D_t: demand shifters
  - S_t: supply shifters
  - $I_t \in \{0, 1\}$ indicating whether the cartel was in a price war or not
- In previous notation,
  - $x_t = (Q_t, P_t, D_t, S_t)$
  - $z_t = I_t$
  - $\theta = (\alpha, \beta)$
  - to deal with simultaneity, likelihood function $p(x_i, \zeta_i | \theta)$ is FIML

Complete and incomplete data likelihoods

The *incomplete data log-likelihood function* or *unconditional log-likelihood function* for a mixture model involves a sum within an expectation, which makes it very hard to maximize with standard optimization algorithms:

$$\ln L(x|\theta) = \sum_i \ln \left(\sum_z p(x_i, z|\theta) \right).$$

The EM algorithm is based on the (expected) *complete data log-likelihood function*:

$$Q(x, q|\theta) = \sum_i \sum_z q_{iz} \ln (p(x_i, z|\theta)).$$

Note that Q would simply be the log-likelihood function if ζ were observed.

EM Algorithm overview

- The EM algorithm starts with some initial guess for $\theta^{(0)}$
- In the E-step, we calculate expectations of the q's conditional on the parameter values:

$$q_{iz}^{(m)} = Pr\left(\zeta_i = z|\theta^{(m-1)}\right).$$

- In the M-step, we maximize the value of the complete data likelihood function:

$$\theta^{(m)} = \max_{\theta} Q\left(x, q^{(m)}|\theta\right).$$

- The EM Algorithm iteratively applies E and M steps until $\theta(m)$ converges.

EM Algorithm overview

- The E and M steps are often easy computationally (in contrast to maximization of incomplete data likelihood function).
- Each EM iteration increases $\ln L(x|\theta)$.
- Thus, iterating on the E and M steps will monotonically increase $\ln L\left(x|\theta^{(m)}\right)$, and $\theta^{(m)}$ will typically converge to a local maximum of $\ln L(x|\theta)$.
- $\Rightarrow$ EM Algorithm transforms a hard optimization problem into a series of easy optimization problems

Monotonicity

Monotonicity

$$\ln L\left(x|\theta^{(m)}\right) \geq \ln L\left(x|\theta^{(m-1)}\right)$$

Monotonicity

Monotonicity

$$\ln L\left(x|\theta^{(m)}\right) \geq \ln L\left(x|\theta^{(m-1)}\right)$$

$$\begin{aligned}
\ln L\left(x|\theta^{(m)}\right) &= \sum_i \ln\left(\sum_z p\left(x_i|\zeta_i,\theta^{(m)}\right) p\left(\zeta_i|\theta^{(m)}\right)\right) \\
&= \sum_i \ln\left(\sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(n-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right)
\end{aligned}$$

where the inequality follows from Jensen's inequality

Monotonicity

$$
\begin{aligned}
\ln L\left(x|\theta^{(m)}\right) &= \sum_i \ln\left(\sum_z p\left(x_i|\zeta_i,\theta^{(m)}\right) p\left(\zeta_i\theta^{(m)}\right)\right) \\
&= \sum_i \ln\left(\sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(n-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m-1)}\right)p\left(\zeta_i\theta^{(m-1)}\right)}{p\left(\zeta_i = z|x,\theta^{(m-1)}\right)}\right)
\end{aligned}
$$

where the second inequality follows because $\theta^{(m)}$ is selected to maximize

$$
\sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(p\left(x_i|\zeta_i,\theta\right) p\left(\zeta_i|\theta\right)\right)
$$

Monotonicity

$$
\begin{aligned}
\ln L\left(x|\theta^{(m)}\right) &= \sum_i \ln\left(\sum_z p\left(x_i|\zeta_i,\theta^{(m)}\right) p\left(\zeta_i|\theta^{(m)}\right)\right) \\
&= \sum_i \ln\left(\sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i=z|x,\theta^{(n-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m)}\right)p\left(\zeta_i|\theta^{(m)}\right)}{p\left(\zeta_i=z|x,\theta^{(m-1)}\right)}\right) \\
&\geq \sum_i \sum_z p\left(\zeta_i = z|x,\theta^{(m-1)}\right) \ln\left(\frac{p\left(x_i|\zeta_i,\theta^{(m-1)}\right)p\left(\zeta_i|\theta^{(m-1)}\right)}{p\left(\zeta_i=z|x,\theta^{(m-1)}\right)}\right) \\
&= \mathcal{L}\left(x|\theta^{(m-1)}\right)
\end{aligned}
$$

Estimation of Mixture of Normals I

- $\theta = (\mu_1, \mu_2, \sigma, \alpha_1)$
- If $z_i = 1$, then $x_i \sim N(\mu_1, \sigma)$
- If $z_i = 2$, then $x_i \sim N(\mu_2, \sigma)$
- $Pr(z_i = 1) = \alpha_1$

In the E step, we just apply Bayes's Theorem to find q's

$$q_{i1}^{(m)} = Pr\left(z_i = 1 | x_i, \theta^{(m)}\right) = \frac{\alpha_1^{(m)} f\left(x_i | \mu_1^{(m)}, \sigma^{(m)}\right)}{\alpha_1^{(m)} f\left(x_i | \mu_1^{(m)}, \sigma^{(m)}\right) + \left(1 - \alpha_1^{(m)}\right) f\left(x_i | \mu_2^{(m)}, \sigma^{(m)}\right)}$$

where $f(x|\mu, \sigma)$ is the density at x of the normal distribution with mean μ and standard deviation σ^2.

Estimation of Mixture of Normals II

- In the M step, maximizing the complete data likelihood function amounts to taking weighted means:

$$\mu_z^{(m)} = \sum_i q_{iz}^{(m)} x_i$$

$$\sigma^{(m)} = \sqrt{\frac{\sum_z \sum_i q_{iz}^{(m)} \left(x_i - \mu_z\right)^2}{\sum_z \sum_i q_{iz}^{(m)}}}$$

$$\alpha_z^{(m)} = N^{-1} \sum_i q_{iz}^{(m)}$$

Estimation of example 1: mixture of normals

- Note: in a mixture model with covariates that enter linearly, the M step involves weighted OLS instead of a weighted mean
- Bottom line: E and M step are both easy computationally, so iterating on them goes quickly.
- In general, the EM algorithm can stop at local maxima, so some care is needed to ensure a global optimum is attained (e.g., multiple starting points).

Model Selection: Likelihood Ratio

- When comparing nested models, the **likelihood ratio test** is simple and powerful
- Let $\boldsymbol{\theta}$ be a vector of parameters to be estimated
  - $\hat{\boldsymbol{\theta}}_U$ is the ML estimate for the full model
  - $\hat{\boldsymbol{\theta}}_R$ is the ML estimate for a restricted model (e.g., with a couple elements fixed to zero)
- **Likelihood ratio**:

$$\lambda = \frac{L\left(\hat{\boldsymbol{\theta}}_R|\text{data}\right)}{L\left(\hat{\boldsymbol{\theta}}_U|\text{data}\right)},$$

which will always be less than one.

Model Selection: Likelihood Ratio Test

- Null hypothesis H_0: the restricted model is correct.
- Given regularity conditions and H_0, then asymptotically asymptotic distribution of

$$-2 \ln \lambda \sim \mathcal{X}_R^2,$$

  where $\mathcal{X}_R^2$ is chi-squared distribution with degrees of freedom equal to number of restrictions.
- Note similarly to testing restrictions in linear models, but no need for linearity and computationally simpler than F test.

Model Selection: Information Criteria

- Just as R^2 always increases as we add parameters, so does the likelihood.
- When comparing models with different numbers of parameters, we should penalize more complex models. Intuitively, evaluating models based on likelihood without a penalty will lead to over fitting the data.
- Two popular criteria for selecting models that reward parsimony:

$$\begin{aligned} \textbf{Akaike information criterion} &= -2\ln L(\theta|\mathbf{y}) + 2K \\ \textbf{Bayes information criterion} &= -2\ln L(\theta|\mathbf{y}) + K\ln n \end{aligned}$$

- To compare two or more models using the AIC (BIC), compute each model's AIC (BIC) score, and select the model with the lowest score (highest penalized likelihood).
- Note: these can be used to compare non-nested models as well as nested models.