The EM algorithm II: theory and inference

Julia Wrobel

Overview

Today, we cover:

- EM theory: why does it work?
- Inference for EM estimates

Announcements

- HW3 posted and due 3/4 at 10:00AM

Readings:

- Chapter 4: The EM Algorithm, in Peng
- Givens and Hoeting Chapter 4

EM: review

(1) **E-Step**: Let θ_0 be the current estimate of θ. Define

$$Q(\theta|\theta_0) = E_z \left[\log p(y, z|\theta) | y, \theta_0 \right]$$

(2) **M-Step**: Maximize $Q(\theta|\theta_0)$ with respect to θ to get next value of θ

(3) Iterate between E and M steps until convergence.

Note: E-step expectation taken WRT missing data density,

$$p(z|y, \theta) = \frac{p(y, z|\theta)}{p(y|\theta)}$$

EM Issues

1. Local vs. global max: may be multiple modes, EM may converge to a saddle point

- **Solution**: try multiple starting values

2. Bad initialization can be a problem

- Use information from the context
- Use a crude method to find initial values (such as method of moments, grid search)

EM: intuition

Idea: In order to estimate θ via MLE *using only the observed data*, need to be able to maximize $l(\theta|y) = \log f(y|\theta) = \int_z p(y, z|\theta) dz$

- BUT $l(\theta|y)$ difficult to maximize because of the integral
- INSTEAD: assuming $p(y, z|\theta)$ has some nice form (like EF)

    - If we have estimate of missing data Z, can easily evaluate $p(y, z|\theta)$

To do this, we construct surrogate function (called Q function)

- Q is expected value of log likelihood for $p(y, z|\theta)$ *with respect to conditional distribution of missing given observed data*, $p(z|y, \theta)$, for current estimate of parameters, θ_0
- **M-Step** maximizes this surrogate function

    - Akin to filling in the missing data then taking the MLE for θ

EM: proof of ascent property

First, some definitions

- **Bayes rule**: $P(A|B) = \frac{P(B|A)P(A)}{P(B)}$
- **Kullback-Leibler divergence** (aka "relative entropy"):

$$\int \log \frac{p(x)}{q(x)} p(x) dx \geq 0 \text{ for densities } p(x), q(x)$$

- KL divergence is non-negative
- Attains its minimum 0 when $p(x)$ and $q(x)$ are equal

EM: proof of ascent property

Theorem: at each iteration of the EM algorithm,

$$\log p(y|\theta^{t+1}) \geq \log p(y|\theta^t),$$

and equality holds if and only if $\theta^{t+1} = \theta^t$.

Proof: The definition of θ^{t+1} gives

$$Q(\theta^{t+1}|\theta^t) \geq Q(\theta^t|\theta^t) \implies$$

$$E_z\left[\log p(y, z|\theta^{t+1})|y, \theta^t\right] \geq E_z\left[\log p(y, z|\theta^t)|y, \theta^t\right]$$

Using Bayes rule and the law of conditional probability, this can be expanded to...

EM: proof of ascent property

$$E\left[\log p(z|y, \theta^{t+1})|y, \theta^t\right] + \log p(y|\theta^{t+1}) \geq E\left[\log p(z|y, \theta^t)|y, \theta^t\right] + \log p(y|\theta^t) \quad (1)$$

Also, by non-negativity of KL divergence,

$$\int_z \log \frac{p(z|y, \theta^t)}{p(z|y, \theta^{t+1})} p(z|y, \theta^t) dz = E\left[\log \frac{p(z|y, \theta^t)}{p(z|y, \theta^{t+1})}|y, \theta^t\right] \geq 0 \quad (2)$$

Combining (1) and (2) yields

$$\log p(y|\theta^{t+1}) \geq \log p(y|\theta^t)$$

EM: proof of ascent property

EM: proof of ascent property

Combining (3) and (4), we have

$$\log p(y, z|\theta^{t+1}) = \log p(y, z|\theta^t).$$

The uniqueness of θ leads to $\theta^{t+1} = \theta^t$

EM: Why does it work?

$Q(\theta|\theta_0)$ function serves as a lower bound to the observed data density $p(y|\theta)$.

The EM is a **minorization** approach. Instead of directly maximizing. the log-likelihood, which is hard, the algorithm constructs a minorizing function and optimizes that function instead.

A function g *minorizes* f over $\mathcal{X}$ at y if:

1. $g(x) \leq f(x)$ for all $x \in \mathcal{X}$
2. $g(y) = f(y)$

EM: Why does it work?

Because $Q(\theta|\theta_0)$ minorizes $l(\theta|y)$, maximizing it is guaranteed to increase (or at least not decrease) $l(\theta|y)$.

- This is because if θ_n is our current estimate of θ and $Q(\theta|\theta_n)$ minorizes $l(\theta|y)$ at θ_n, then we have

$$l(\theta_{n+1}|y) \geq Q(\theta_{n+1}|\theta_n) \geq Q(\theta_n|\theta_n) = l(\theta_n|y)$$

# Example: minorization in Two-part Gaussian mixture model

Suppose we have data $y_1, \ldots, y_n$ that are sampled independently from a two-part mixture of Normals with density

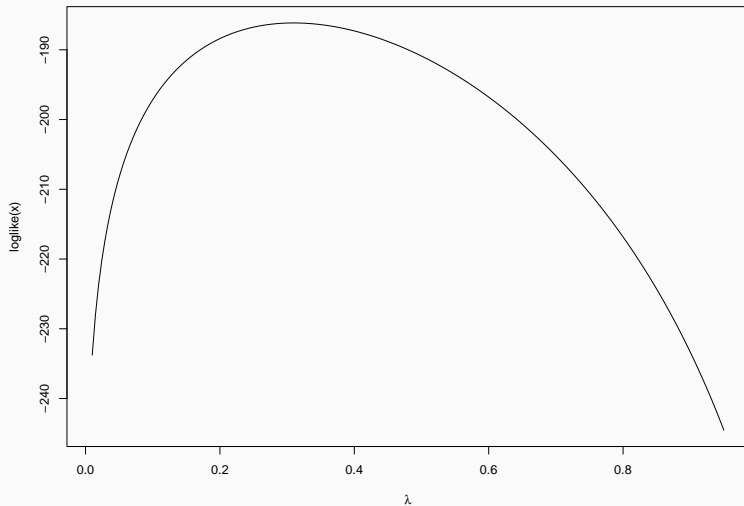$$p(y|\theta) = \lambda \mathcal{N}(y|\mu_1, \sigma_1^2) + (1 - \lambda)\mathcal{N}(y|\mu_2, \sigma_2^2).$$

We can simulate some data from this model:

```
mu1 = 1; mu2 = 4
s1 = 2; s2 = 1
lambda0 = 0.4
n = 100
set.seed(2017-09-12)
z = rbinom(n, 1, lambda0) ## "Missing" data
x = rnorm(n, mu1 * z + mu2 * (1-z), s1 * z + (1-z) * s2)
```
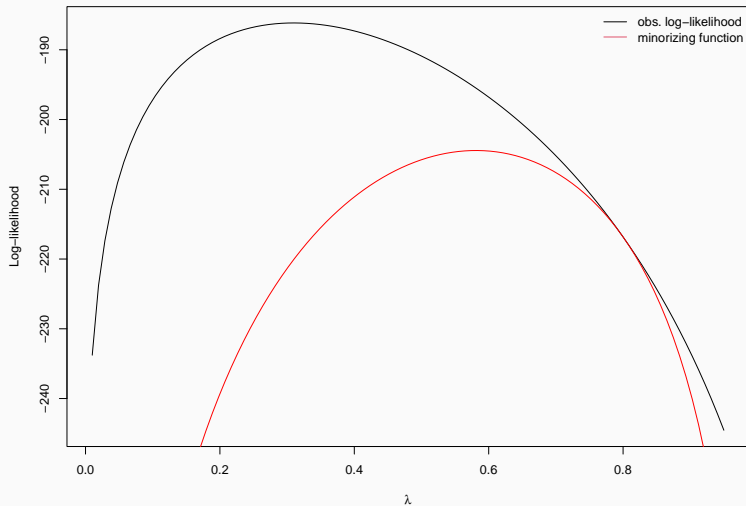
Two-part Gaussian mixture model
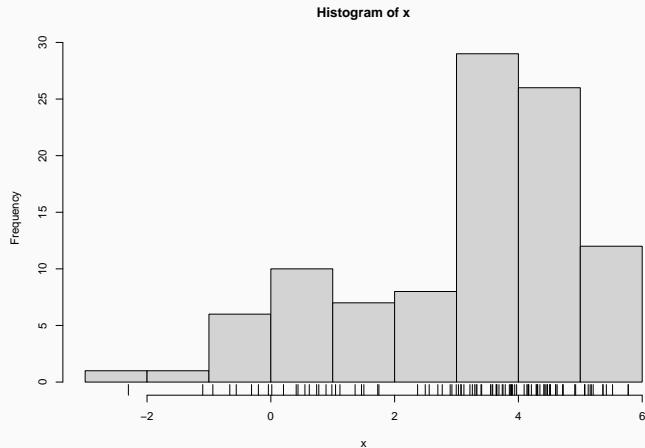
Histogram of x

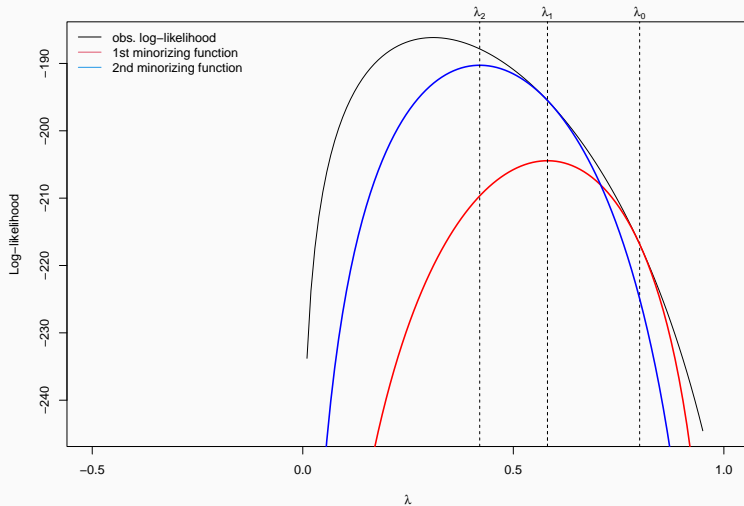# Two-part Gaussian mixture model

Two-part Gaussian mixture model

Two-part Gaussian mixture model

EM Inference

Original EM paper did not discuss how to obtain any measures of uncertainty, such as standard errors.

- *Observed information matrix* would provide inference
  - Like observed data log-likelihood, often difficult to compute because of the missing data
- Louis' method
- Supplemented EM (SEM)
- Bootstrap

EM Inference

$$p(y|\theta) = \frac{p(y, z|\theta)}{p(z|y, \theta)}$$

$$-\log p(y|\theta) = -\log p(y, z|\theta) - [-\log p(z|y, \theta)]$$

$$E\left[-\nabla^2 \log p(y|\theta)\right] = E\left[-\nabla^2 \log p(y, z|\theta)\right] - E\left[-\nabla^2 \log p(z|y, \theta)\right]$$

$$I_Y(\theta) = I_{Y,Z}(\theta) - I_{Z|Y}(\theta)$$

Louis's method

$$I_Y(\theta) = I_{Y,Z}(\theta) - I_{Z|Y}(\theta)$$

- Presumably, $I_{Y,Z}(\theta)$ is reasonable to compute because based on complete data
- What is $I_{Z|Y}(\theta)$?
- $S(y|\theta) = \nabla \log p(y|\theta)$: observed score function
- $S(y,z|\theta) = \nabla \log p(y,z|\theta)$: complete data score function

Louis's method

$$I_Y(\theta) = I_{Y,Z}(\theta) - I_{Z|Y}(\theta)$$

- $S(y|\theta) = \nabla \log p(y|\theta)$: observed score function
- $S(y,z|\theta) = \nabla \log p(y,z|\theta)$: complete data score function

Louis (1982) showed that

$$I_{Z|Y}(\theta) = E\left[S(y,z|\theta)S(y,z|\theta)^T\right] - S(y|\theta)S(y|\theta)^T$$

Where expectation is taken WRT missing data density $p(z|y,\theta)$.

Louis's method

$$I_Y(\theta) = I_{Y,Z}(\theta) - I_{Z|Y}(\theta)$$

- $I_{Z|Y}(\theta) = E[S(y,z|\theta)S(y,z|\theta)^T] - S(y|\theta)S(y|\theta)^T$
  - At MLE $\hat{\theta}$, $S(y|\hat{\theta}) = 0$

$$I_Y(\hat{\theta}) = I_{Y,Z}(\hat{\theta}) - E\left[S(y,z|\theta)S(y,z|\theta)^T\right]$$

Louis's method

$$I_Y(\hat{\theta}) = I_{Y,Z}(\hat{\theta}) - E\left[S(y,z|\theta)S(y,z|\theta)^T\right] \tag{1}$$

$$= -E\left[\nabla^2 \log p(y,z|\theta)|\hat{\theta},y\right] - E\left[S(y,z|\theta)S(y,z|\theta)^T\right] \tag{2}$$

$$= -Q''(\hat{\theta}|\hat{\theta}) - E\left[S(y,z|\theta)S(y,z|\theta)^T\right] \tag{3}$$

Louis's estimator should be evaluated at last iteration of EM algorithm.

Supplemented EM (SEM)

Meng & Rubin, 1991: *Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm.*

Background: EM defines a mapping $\Psi : \theta^{t+1} = \Psi(\theta^t)$

- $\Psi(\theta) = \big(\Psi_1(\theta), \dots, \Psi_p(\theta)\big)$ and $\theta = (\theta_1, \dots, \theta_p)$
- When EM converges, it converges to a fixed point of this mapping, so $\hat{\theta} = \Psi(\hat{\theta})$
- $\Psi'(\theta)$ is the Jacobian matrix where $[\Psi'(\theta)]_{i,j} = \frac{\partial \Psi_i(\theta)}{\partial \theta_j}$

Supplemented EM (SEM)

Dempster et al showed that

$$\Psi'(\hat{\theta})^T = I_{Z|Y}(\hat{\theta})I_{Y,Z}(\hat{\theta})^{-1} \tag{1}$$

The missing information principle says that

$$I_Y(\hat{\theta}) = I_{Y,Z}(\hat{\theta}) - I_{Z|Y}(\hat{\theta})$$
$$= \left[\mathcal{I} - I_{Z|Y}(\hat{\theta})I_{Y,Z}(\hat{\theta})^{-1} \right] I_{Y,Z}(\hat{\theta})$$

Then, substituting (1) and inverting gives

$$\widehat{Var}(\hat{\theta}) = I_Y(\hat{\theta})^{-1} = I_{Y,Z}(\hat{\theta})^{-1} \left[\mathcal{I} - \Psi'(\hat{\theta}^T) \right]^{-1}. \tag{2}$$

Supplemented EM (SEM)

$$\widehat{Var}(\hat{\theta}) = I_Y(\hat{\theta})^{-1} = I_{Y,Z}(\hat{\theta})^{-1} \left[\mathcal{I} - \Psi'(\hat{\theta}^T) \right]^{-1} \qquad (2)$$

- (2) means that the observed-data asymptotic variance can be obtained by inflating the complete-data asymptotic variance by the factor $\left[\mathcal{I} - \Psi'(\hat{\theta}^T) \right]^{-1}$
- Smaller missingness $\implies$ smaller Ψ' $\implies$ less variance inflation and faster convergence.

SEM Algorithm

SEM consists of three steps:

1. The evaluation of $I_{Y,Z}(\hat{\theta})$
2. The evaluation of $\Psi'(\hat{\theta})$
3. The evaluation of $\widehat{Var}(\hat{\theta})$

Evaluation of $I_{Y,Z}(\hat{\theta})$

- For exponential family, $I_{Y,Z}(\hat{\theta}) = -E\left[\nabla^2 \log p(y,z|\theta)|\hat{\theta}, y\right]$ should be easy to obtain.
- This is the second derivative of the Q function evaluated at $\hat{\theta}$

SEM Algorithm

Estimation of $\Psi'(\hat{\theta})$

1. Run EM algorithm to convergence to obtain MLE $\hat{\theta}$.
2. Pick a new starting point, θ^0. θ^0 should be some small distance from $\hat{\theta}$ but not equal to $\hat{\theta}$ in any component.

SEM Algorithm

Estimation of $\Psi'(\hat{\theta})$

1. Run EM algorithm to convergence to obtain MLE $\hat{\theta}$.
2. Pick a new starting point, θ^0.
3. Repeat the following until r_{ij}^k is stable:

- Calculate $\theta^k = \Psi(\theta^{k-1})$ using one step of EM
- For each $i = 1, \ldots, p$:
  - Let $\theta^k(i) = (\hat{\theta}_1, \ldots, \hat{\theta}_{i-1}, \hat{\theta}_i^k, \hat{\theta}_{i+1}, \ldots, \hat{\theta}_p)$., i.e., replace i^{th} element of $\hat{\theta}$ with the ith element of θ^k.
  - Perform one step of EM on $\theta^k(i)$ to obtain $\Psi[\theta^k(i)]$
  - Obtain $r_{ij}^k = \left\{ \Psi[\theta^k(i)] - \hat{\theta} \right\} / \left\{ \theta_i^k - \hat{\theta}_i \right\}$ for $j = 1, \ldots, p$

SEM Algorithm

- The MLE $\hat{\theta}$ should be obtained at a very low tolerance (e.g. $\epsilon = 10^{-12}$)

The final r_{ij} is taken to be the first value of r_{ij}^k satisfying $|r_{ij}^k - r_{ij}^{k-1}| \leq \epsilon$, where k can be different for different (i, j).

Bootstrapping

Goal is to obtain estimate of covariance matrix for EM parameters. To do a simple nonparametric bootstrap given an *iid* sample of observed data $y_1, \ldots, y_n$, do the following:

1. Calculate $\hat{\theta}_{EM}$
2. Sample data $y_1, \ldots, y_n$ with replacement, and for each sample y_b^*, calculate a bootstrap estimate $\hat{\theta}_b^*$
3. Repeat step 2 B times to obtain $\theta_1^*, \ldots, \theta_B^*$ bootstrap parameter estimates.
4. Sample covariance matrix of θ^* can be used as covariance of $\hat{\theta}_{EM}$.

Comparing EM inference approaches

- Louis's Method
  - Requires calculation of the conditional expectation of the square of the complete-data score function, which is specific to each problem
- SEM
  - Obtains covariance matrix by using only the code for computing the complete-data covariance matrix, the code for EM itself, and code for standard matrix operations.
- Bootstrapping
  - Conceptually simple
  - May be prohibitively slow if your EM algorithm is slow to converge

Speeding up EM

Sometimes convergence of EM can be very slow. Some methods to help with this:

- Louis's Acceleration
- SQUAREM

References

- Louis's method original paper
  - Finding observed information using the EM algorithm, JRSSB 1982
- SEM original paper (Meng & Rubin)
  - Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm, JASA 1991

Exercise

Start to implement SEM method for the two-part GMM.