

# The EM algorithm I: introduction

---

Julia Wrobel

# Overview

Today, we cover:

- The EM Algorithm: intro and applications
- Review of some MLE theory

Announcements

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

Readings:

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

# Last lectures

- General optimization problems
  - Steepest descent
  - Newton's method
  - Fisher scoring
  - Quasi-Newton
- GLMs
  - iteratively reweighted least squares

# Expectation–maximization (EM) algorithm

- An iterative algorithm for **maximizing likelihood** when the model contains unobserved latent variables
- The algorithm iterates between **E-step** (expectation) and **M-step** (maximization)
- **E-step**: fill in the missing/latent values
- **M-step**: obtain parameters maximizing the expected log-likelihood from the E step

# EM algorithm

Widely used algorithm!! Some common uses include:

- Gaussian mixture models
- Hidden Markov models
- Missing data imputation
- Latent variable models (i.e. factor analysis, latent growth curves)
- Censored or truncated data

# EM algorithm

## Pros

- Guarantees monotone improvement of the likelihood function
- Handles missing data

## Cons

- Convergence is to a local, not necessarily global, solution
  - Can be heavily dependent on initial values
- Convergence can be slow, especially for high-dimensional problems (lots of parameters)

# EM: notation

- $Y$ : observed data vector
- $Z$ : vector of data that are missing
- $\theta$ : vector of parameters we want to estimate
- $p(y, z|\theta)$ : complete data density
- $p(y|\theta) = \int_z p(y, z|\theta) dz$ : observed data density
  - $l(\theta|y) = \log f(y|\theta)$ : observed data likelihood
- $p(z|y, \theta)$ : conditional density of missing data given observed data

# EM: intuition

**Idea:** In order to estimate  $\theta$  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)

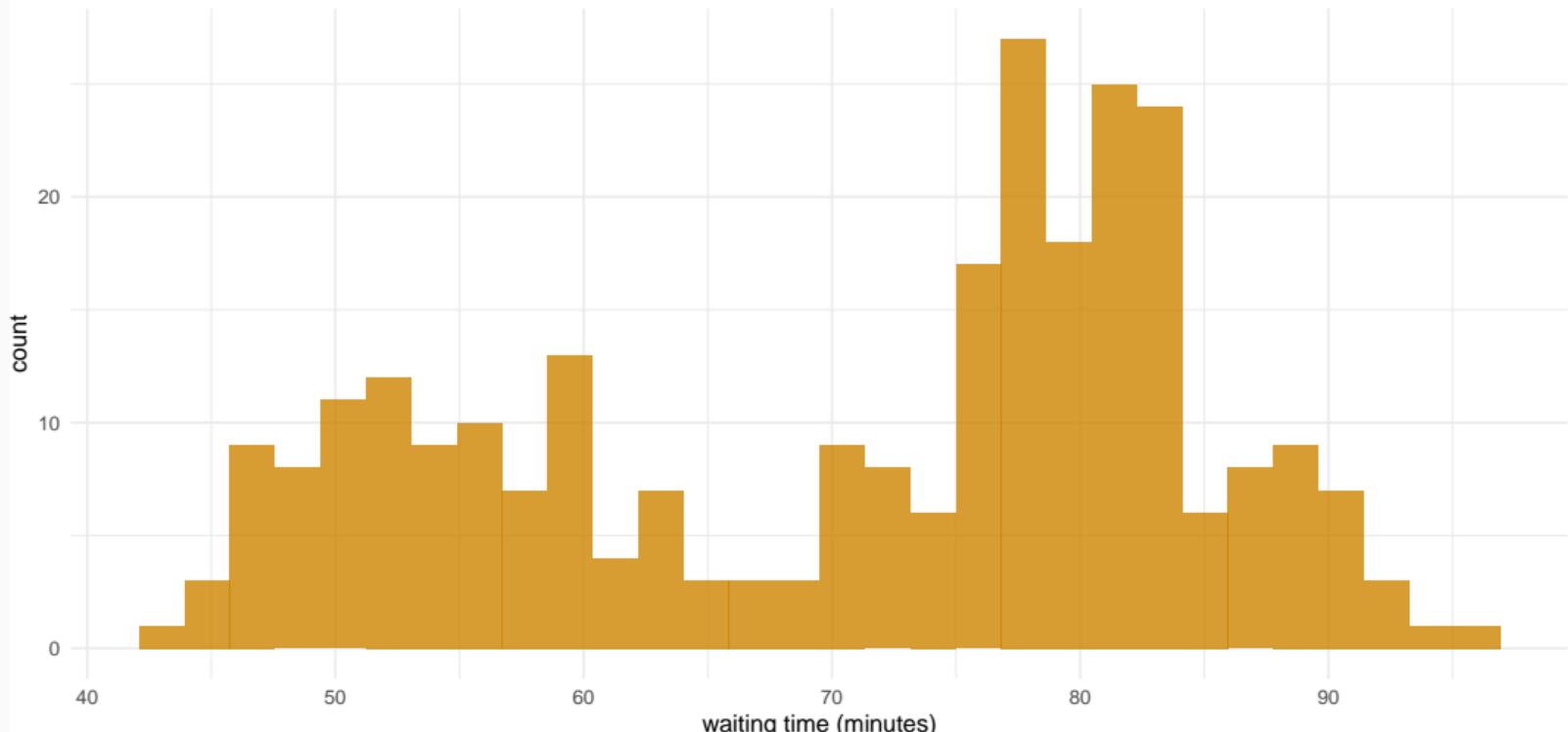
- $Q$  is expected value of log likelihood for  $p(y, z|\theta)$  *with respect to conditional distribution of missing given observed data*,  $p(z|y, \theta)$ , for current estimate of parameters,  $\theta_0$
- **M-Step** maximizes this surrogate function
  - Akin to filling in the missing data then taking the MLE for  $\theta$

# Canonical examples

- Two-part Gaussian mixture model
  - Data  $Y_1, \dots, Y_n$  come from a mixture of two Gaussian distributions
  - Soft clustering/unsupervised learning technique
  - **Example:** A new blood biomarker shows promise as an early Alzheimer's detection biomarker. Values of the biomarker in a sample of patients have a bimodal distribution: healthy subjects, those with Alzeimers
  - **Example:** a clinical trial is evaluating response to a new cancer drug. There are three subpopulations: non-responders, partial responders, complete responders
- Censored exponential data

# Example: Old Faithful waiting times

Time between Old Faithful eruptions in Yellowstone National Park



# EM: steps

- (1) **E-Step:** Let  $\theta_0$  be the current estimate of  $\theta$ . Define

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

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

How to monitor convergence in EM?

- Each iteration is designed to increase the **observed data log likelihood**,  $p(y|\theta)$ .
  - Check if falls below a certain threshold, then stop
    - $p(y|\theta^{k+1}) - p(y|\theta^k) < \epsilon$
  - In practice, can be very sensitive to starting values
    - Can fail due to numerical difficulties if starting values are far from the truth

However,  $p(y|\theta)$  **cannot always be computed!**

- Another option:  $(\theta^{t+1} - \theta^t)^T(\theta^{t+1} - \theta^t) < \epsilon$
- Another option:  $|Q(\theta^{t+1}|\theta^t) - Q(\theta^t|\theta^t)| < \epsilon$

# Two-part Gaussian mixture model

- $Y_1, \dots, Y_n$  are sampled independently from a mixture of two Normal distributions with density

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

- $\theta = (\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda)$
- $Z_1, \dots, Z_n$ : labels identifying which observation came from which population
  - $Z_i = 1$  if  $Y_i$  from  $\mathcal{N}(y|\mu_1, \sigma_1^2)$ ;  $Z_i = 0$  otherwise

$$z_i \sim Bernoulli(\lambda)$$

## Two-part Gaussian mixture model

Joint density of observed and missing data (i.e. complete data density) is then

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

**Exercise:** show that integrating out the missing data gives the observed data density

# Two-part Gaussian mixture model

## Two-part Gaussian mixture model

Then, complete-data log likelihood is

$$\begin{aligned}\log p(y, z|\theta) &= \sum_i^n [z_i \log (\lambda \mathcal{N}_1) + (1 - z_i) \log ((1 - \lambda) \mathcal{N}_2)] \\ &= \sum_i [z_i \log(\lambda) + z_i \log \mathcal{N}_1 + (1 - z_i) \log(1 - \lambda) + (1 - z_i) \log \mathcal{N}_2]\end{aligned}$$

## Two-part Gaussian mixture model

Missing data density is

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

$$= Bernoulli \left( \frac{\lambda \mathcal{N}(y|\mu_1, \sigma_1^2)}{\lambda \mathcal{N}(y|\mu_1, \sigma_1^2) + (1 - \lambda) \mathcal{N}(y|\mu_2, \sigma_2^2)} \right)$$

This allows us to define  $E[z_i|y_i, \theta] := \pi_i$  which will be used in find  $Q(\theta|\theta_0)$  in the E-step

# Two-part Gaussian mixture model

Next, **E-Step!** Construct  $Q()$  function

$$\begin{aligned} Q(\theta|\theta_0) &= E_z [\log p(y, z|\theta)|y, \theta_0] \\ &= E \left( \sum_i^n [z_i \log (\lambda \mathcal{N}_1) + (1 - z_i) \log ((1 - \lambda) \mathcal{N}_2)] \right) \\ &= \sum_i^n [E(z_i) \log (\lambda \mathcal{N}_1) + E(1 - z_i) \log ((1 - \lambda) \mathcal{N}_2)] \\ &= \sum_i^n [\pi_i \log (\lambda \mathcal{N}_1) + (1 - \pi_i) \log ((1 - \lambda) \mathcal{N}_2)] \end{aligned}$$

Need current estimates of  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda$  - Also, compute  $E[z_i|y_i, \theta] := \pi_i$

# Two-part Gaussian mixture model

**M-Step!** Maximize  $Q$  to get current estimates of  $\mu_1, \mu_2, \sigma_1^2, \sigma_2^2, \lambda$ .

$$\hat{\mu}_1 = \frac{\sum_i \pi_i y_i}{\sum_i \pi_i} \quad (1)$$

$$\hat{\mu}_2 = \frac{\sum_i (1 - \pi_i) y_i}{\sum_i (1 - \pi_i)} \quad (2)$$

$$\hat{\sigma}_1^2 = \frac{\sum_i \pi_i (y_i - \mu_1)^2}{\sum_i \pi_i} \quad (3)$$

$$\hat{\sigma}_2^2 = \frac{\sum_i (1 - \pi_i) (y_i - \mu_2)^2}{\sum_i (1 - \pi_i)} \quad (4)$$

$$\hat{\lambda} = \frac{1}{n} \sum_i \pi_i \quad (5) \quad 19$$

## Two-part Gaussian mixture model

**Class exercise:** finish implementing this algorithm in R by doing first lab problem. Starter code is provided in the file EM\_GMM.R

# Canonical examples

- Two-part Gaussian mixture model
- Censored exponential data
  - Survival analysis, survival times exponentially distributed
  - Substantial right censoring
    - For censored individuals, true survival time is unknown

# Censored exponential data

Suppose we have survival times  $t_1, \dots, t_n \sim \text{Exponential}(\lambda)$ .

- Do not observe all survival times because some are censored at times  $c_1, \dots, c_n$ .
- Actually observe  $y_1, \dots, y_n$ , where  $y_i = \min(t_i, c_i)$ 
  - Also have an indicator  $\delta_i$  where  $\delta_i = 1$  if  $t_i \leq c_i$ 
    - i.e.  $\delta_i = 1$  if not censored and  $\delta_i = 0$  if censored

# Censored exponential data

Suppose we have survival times  $t_1, \dots, t_n \sim \text{Exponential}(\lambda)$ .

- Do not observe all survival times because some are censored at times  $c_1, \dots, c_n$ .
- Actually observe  $y_1, \dots, y_n$ , where  $y_i = \min(t_i, c_i)$ 
  - Also have an indicator  $\delta_i$  where  $\delta_i = 1$  if  $t_i \leq c_i$ 
    - i.e.  $\delta_i = 1$  if not censored and  $\delta_i = 0$  if censored
- What is  $p(y, z|\theta)$ , the complete data density? - What is  $z$ ?

# Censored exponential data

# EM algorithm

## Pros

- Guarantees monotone improvement of the likelihood function
- Handles missing data

## Cons

- Convergence is to a local, not necessarily global, solution
  - Can be heavily dependent on initial values
- Convergence can be slow, especially for high-dimensional problems (lots of parameters)

# Asymptotic properties of MLEs

If it converges to the global maximum, EM finds the **MLE** of your likelihood function. This means that theory about MLEs holds for EM parameter estimates. Specifically:

- **Consistency:** Let the sequence of MLEs of  $\theta_0$  be denoted by  $\hat{\theta}_n$ . For any fixed  $\epsilon > 0$ , as  $n \rightarrow \infty$

$$P(|\hat{\theta}_n - \theta_0| > \epsilon) \rightarrow 0$$

- Ensures estimate converges in probability to the true value
- **Asymptotic efficiency:**  $\hat{\theta}$  achieves minimum variance among all asymptotically unbiased estimators

# Asymptotic properties of MLEs

If it converges to the global maximum, EM finds the **MLE** of your likelihood function. This means that theory about MLEs holds for EM parameter estimates. Specifically:

- **Asymptotic Normality:** Let the sequence of MLEs of  $\theta_0$  be denoted by  $\hat{\theta}_n$ .

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0, \sigma^2)$$

- A properly centered and scaled sequence is distributed normally with 0 mean and variance  $\sigma^2$  as  $n \rightarrow \infty$

# Invariance Property of MLEs

Allows us to find the MLE of transformations of an MLE

- If  $\hat{\theta}$  is the MLE of  $\theta$ , then for any function  $\tau(\hat{\theta})$  is the MLE of  $\tau(\theta)$ !

# Invariance Property of MLEs

Suppose  $Y_1, Y_2, \dots, Y_n$  is a sample of independent Normal  $N(\mu, \sigma^2)$  random variables with  $E(Y_i) = \mu$ .

- Sample mean  $\hat{\mu} = \bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$  is the MLE of  $\mu$

What is the MLE of  $1/\mu$ ? Using invariance property of MLEs,

- $1/\hat{\mu} = 1/\bar{Y}$  is the MLE of  $1/\mu$

# Final thoughts

- *Ascent property of EM* is what guarantees stability via monotonically increasing likelihood
- Example of a minorization approach
  - Instead of maximizing the log-likelihood directly, which is difficult to evaluate, the algorithm constructs a minorizing function and optimizes that function instead

# Resources

- good notes
- exercises in EM