Department of Statistics, Stanford University

“Maximum Likelihood From Incomplete Data via the EM Algorithm"
(AP Dempster, NM Laird & DB Rubin)

Wenlong Ji & Dileka Gunawardana

March 17, 2024

EM Algorithm (High-Level Idea): Dream World

Consider an *imaginary dream world* perhaps without a replicability crisis.

One may posit that we would have access to the following two settings where for any population of interest we have access to:

1. Large sample of data
2. Underlying parametric model (i.e., that it has a Normal distribution with mean 10 and standard deviation 5)

EM Algorithm (High-Level Idea): E-Step

Do we really need both of these? What if we only had access to the **parametric model**?

1. What would be the best estimate of a missing data point given our model?
   - If we knew that our model $\sim$ Normal$(10, 5)$, what would be the best guess for the missing point?
2. Underlying parametric model

EM Algorithm (High-Level Idea): M-Step

Do we really need both of these? What if we only had access to a **large sample of the data** ?

1. Large sample of data
2. What would be the best estimate for the model parametrization given access to infinite data from our population of interest?
   - If our observed data was $\{5, 10, 15, \ldots\} \sim$ Normal$(x, 5)$ what would be the best guess for x?

EM Algorithm (High-Level Idea): Real World

Of course, we typically don't have access to either of these settings in the **real world**.

Figure 1: The Real World

This is why the **EM algorithm** is such a powerful and essential tool!
By assuming one of the two settings to do inference in the other setting and vice versa, we can eventually get statistically valid estimates of the model underlying our population of interest while *also* filling in missing data.

EM Algorithm (Math): Missing Data

There are numerous data analysis settings where we have to deal with **missing data**. Some **examples** include:

- A factory sensor that malfunctions for a brief period of time
- Census data that isn't able to access particularly rural communities
- Genetic studies where have only access to data on a subset of the genes

We can represent this mathematically by defining a many-one mapping $\mathcal{X} \to \mathcal{Y}$ where $X \in \mathcal{X}$ is our world of **complete data**, and $Y(X) \in \mathcal{Y}$ is the world of **data we observe** (so $y(x)$ is the complete data x with some of the points missing).

We **assume a family of sampling densities** $f(x|\phi)$, so the distribution corresponding to the missing data actually observed integrates over every possible dataset $x \in \mathcal{X}$ that would lead to the observation y:

$$g(y|\phi) := \int_{\mathcal{X}(y)} f(x|\phi) dx$$

EM Algorithm (Math): Algorithm

$$Q(\phi'|\phi) := \mathbb{E}[\log f(x|\phi')|y,\phi] = \int_{\mathcal{X}(y)} \log f(x|\phi') f(x|\phi) dx$$

We can think of Q as weighting the possible points that the "complete data" can be based on how likely they are under our current guess of ϕ.

At each step $p = 1, 2, \dots$ until we converge, the **EM algorithm** iterates between the following two steps:

1. **E-Step**: Compute $Q(\phi|\phi^{(p)})$
2. **M-Step**: Choose $\phi^{(p+1)}$ to be the value of $\phi \in \Omega$ that maximizes $Q(\phi|\phi^{(p)})$

Note that the EM algorithm has simple extensions to the **Bayesian** setting where we replace the MLE with the posterior mode in the M-Step.

Why Does it Work? (Theory)

Idea: Every step of EM increases the value of the likelihood.
Denote a step of EM by the function M: $\phi^{(p+1)} = M(\phi^{(p)})$

$$k(x|y,\phi) := \frac{f(x|\phi)}{g(y|\phi)}$$

$$L(\phi) := \log g(y|\phi)$$

$$H(\phi'|\phi) := \mathbb{E}[\log k(x|y,\phi')|y,\phi]$$

$$L(\phi) = \log f(x|\phi) - \log k(x|y,\phi) \implies Q(\phi'|\phi) = L(\phi') + H(\phi'|\phi)$$

$$\mathbf{L}(\mathbf{M}(\phi)) - \mathbf{L}(\phi) = \{\mathbf{Q}(\mathbf{M}(\phi)|\phi) - \mathbf{Q}(\phi|\phi)\} + \{\mathbf{H}(\phi|\phi) - \mathbf{H}(\mathbf{M}(\phi)|\phi)\} \geq \mathbf{0}$$

The inequality follows by the construction in the M-Step (for the first term involving Q) and Jensen's inequality (for the second term involving H).
Settings for convergence are discussed in detail in the paper.

Important Considerations with use of the EM Algorithm

We particularly note the following assumptions required in the derivations and thus applications of EM:

1. Parametric assumption on the population distribution
2. Missing data comes from population whose distribution is equivalent to that of the observed data
    - Typically, there's a reason why that data is missing that could be a result of that population's distribution being different
    - Among our 3 examples of missing data (Slide 6), the factory sensor one would be a valid application, but the census data example would not necessarily be appropriate

These are interesting areas for **future research** perhaps through sensitivity analysis or extensions of the algorithm.

Examples of the EM Algorithm: Missing Data

Here we discuss a simple example of using EM algorithm with missing data.

- Consider a five-category multinomial population, where the probability is $(\frac{1}{2}, \frac{1}{4}\pi, \frac{1}{4}(1-\pi), \frac{1}{4}(1-\pi), \frac{1}{4}\pi)$, and the complete data is $\mathbf{x} = (x_1, x_2, \cdots, x_5)$
- Due to the missing data issue, we only observe $\mathbf{y} = (y_1, y_2, \cdots, y_4)$, where $y_1 = x_1 + x_2, y_2 = x_3, y_3 = x_4, y_4 = x_5$, and we hope to estimate π.
- The complete log-likelihood is

$$f(x|\pi) \propto x_1 \log(\frac{1}{2}) + (x_2 + x_5) \log(\frac{1}{4}\pi) + (x_3 + x_4) \log(\frac{1}{4}(1-\pi))$$

- The EM steps produce an iterative estimate $\pi^{(p)}$:
  - Given the current estimate $\pi^{(p)}$, the estimates of x_1, x_2 is $x_1^{(p)} = \frac{1/2}{1/2+\pi^{(p)}/4} y_1$ and $x_2^{(p)} = \frac{\pi^{(p)}/4}{1/2+\pi^{(p)}/4} y_1$
  - Given the current estimate $\hat{x}_1^{(p)}, \hat{x}_2^{(p)}$, the maximum likelihood estimate is $\pi^{(p+1)} = \frac{x_2^{(p)}+x_5}{x_2^{(p)}+x_3+x_4+x_5}$.

Examples of the EM Algorithm: Missing Data

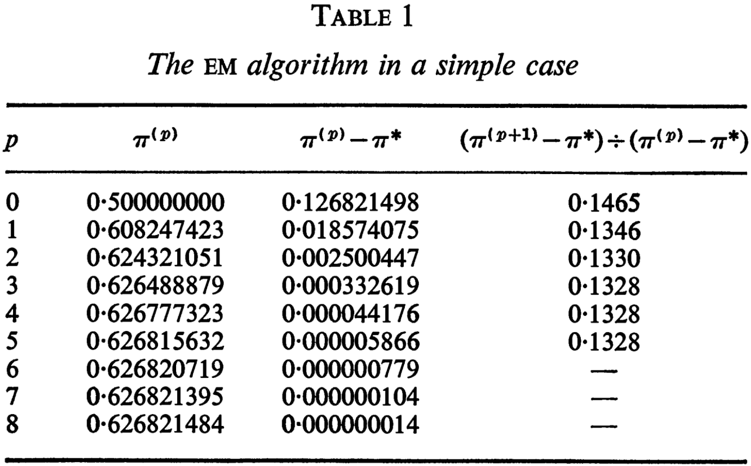

TABLE 1

The EM *algorithm in a simple case*

p	$\pi^{(p)}$	$\pi^{(p)} - \pi^*$	$(\pi^{(p+1)} - \pi^*) \div (\pi^{(p)} - \pi^*)$
0	0·500000000	0·126821498	0·1465
1	0·608247423	0·018574075	0·1346
2	0·624321051	0·002500447	0·1330
3	0·626488879	0·000332619	0·1328
4	0·626777323	0·000044176	0·1328
5	0·626815632	0·000005866	0·1328
6	0·626820719	0·000000779	—
7	0·626821395	0·000000104	—
8	0·626821484	0·000000014	—

Examples of the EM Algorithm: Finite Mixtures

- Suppose we have observations $\mathbf{y} = (y_1, \cdots, y_n)$, and there exists a finite set of R states, such that each y_i is associated with a unique state (unobserved), denoted as $\mathbf{z} = (z_1, \cdots, z_n)$. Specifically, $z_i = k$ indicates y_i belongs to the k-th state.
- $z_i \overset{\text{iid}}{\sim} v(\cdot|\phi)$, $y_i \overset{\text{iid}}{\sim} u(\cdot|z_i, \phi)$
- The complete data likelihood is

$$\log f(\mathbf{y}, \mathbf{z}|\phi) = \sum_{i=1}^{n} (\log v(z_i|\phi) + \log u(y_i|z_i, \phi))$$

The EM steps are:

- **E step:** Estimate the hidden states z_i for given the current parameters $\phi^{(p)}$, i.e., compute $\mathbb{P}(z_i = k|\mathbf{y}, \phi^{(p)})$
- **M step:** Complete-data maximization with estimated states z_i, i.e., solve $\phi^{(p+1)} = \text{argmax}_{\phi} \mathbb{E}_{\mathbf{z}|\mathbf{y},\phi^{(p)}} \log f(\mathbf{y}, \mathbf{z}|\phi)$

Examples of the EM Algorithm: Finite Mixtures

A classical example is the Gaussian Mixture Model.

- $z_i \overset{\text{iid}}{\sim} \text{Categorical}(\pi_1, \cdots, \pi_R)$, $y_i | z_i = k \overset{\text{iid}}{\sim} \text{Normal}(\mu_k, \sigma_k^2)$

For E-step, the posterior distribution for states $\mathbf{z}$ is

$$\omega_{k,i} = \mathbb{P}(z_i = k | \mathbf{y}, \phi) = \frac{\mathbb{P}(z_i = k|\phi)\mathbb{P}(y_i|z_i = k, \phi)}{\sum_{j=1}^{R} \mathbb{P}(z_i = j|\phi)\mathbb{P}(y_i|z_i = j, \phi)} = \frac{\pi_k \mathcal{N}(y_i|\mu_k, \sigma_k^2)}{\sum_{j=1}^{R} \pi_j \mathcal{N}(y_i|\mu_j, \sigma_j^2)}$$

For M-step, the expected likelihood is expressed as

$$\mathbb{E}_{\mathbf{z}|\mathbf{y},\phi}[\log f(\mathbf{y}, \mathbf{z}|\phi)] = \sum_{i=1}^{n} \sum_{k=1}^{R} \mathbb{P}(z_i = k|\mathbf{y}, \phi) \log \mathbb{P}(y_i, z_i = k|\phi)$$

$$= \sum_{i=1}^{n} \sum_{k=1}^{R} \omega_{k,i} \left[\log \pi_k - \frac{1}{2} \log(\sigma_k^2) - \frac{1}{2\sigma_k^2}(y_i - \mu_k)^2 \right] + C$$

Examples of the EM Algorithm: Finite Mixtures

The optimizer of M-step is then given by:

$$\hat{\pi}_k = \frac{1}{n}\sum_{i=1}^{n} \omega_{k,i}, \quad \hat{\mu}_k = \frac{\sum_{i=1}^{n} \omega_{k,i} y_i}{\sum_{i=1}^{n} \omega_{k,i}}, \quad \hat{\sigma}_k^2 = \frac{\sum_{i=1}^{n} \omega_{k,i}(y_i - \hat{\mu}_k)^2}{\sum_{i=1}^{n} \omega_{k,i}}$$

Intuition:

- the E steps computes responsibilities $\omega_{k,i}$ of each point i to each state k.
- the M steps update the state probability, mean, and variance weighted by the computed responsibilities.
- This is often viewed as the soft assignment version of K-means algorithm, where we assign each point to the closest state, and update the state mean by the sample mean in each state.

Examples of the EM Algorithm: Finite Mixtures

Real-world application: Image Segmentation. The image is represented by a collection of pixels $y_i \in \mathbb{R}^3$, and in the segmentation problem, we simply use the Gaussian mixture model to classify them into two clusters.

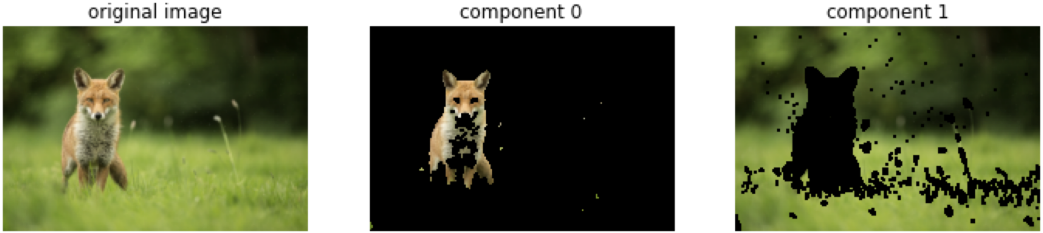

Figure 2: An example of image segmentation using EM algorithm on Gaussian Mixture Models. Taken from Scott Linderman's Stats 305C material.

Examples of the EM Algorithm: Finite Mixtures

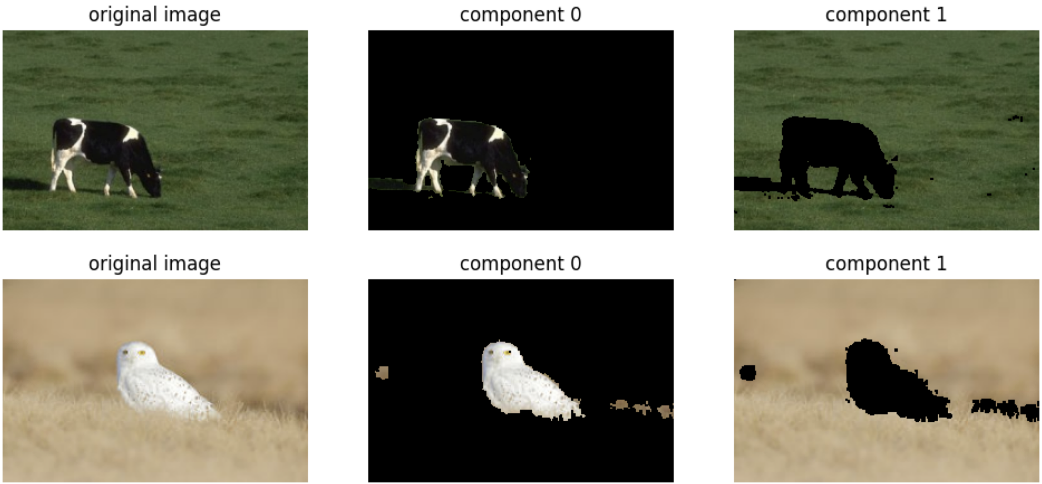

Figure 3: More examples

Variants of EM Algorithm

EM algorithm produces convenient solutions for many simple problems, but: [Gupta et al., 2011]

- EM only finds stationary points of the likelihood function
  - Use EM in conjunction with a global optimizer to explore the space more efficiently. [Ali et al., 2005]
- the computations required may not be computationally tractable
  - Generalized EM: Only ensure the likelihood is increasing in M-step, use gradient ascent or Newton's method.
  - MCMC to approximate the E step.

Variants of EM Algorithm

- the convergence may be too slow
  - Aitken's acceleration [Meilijson, 1989]: Write $\phi^{(p+1)} = \phi^{(p)} + S(\hat{\phi} - \phi^{(p)})$, use Taylor expansion to find the optimal S, the analogy of Newton's method for EM.
- the maximum likelihood estimate may not be the desired output
  - Stochastic EM [Celeux, 1985]: draw a random sample in E step to produce a posterior distribution for ϕ.
  - Data Augmentation [Tanner and Wong, 1987]: also randomize M step, draw $\phi \sim \mathbb{P}(\phi|x) \propto \mathbb{P}(x|\phi)\mathbb{P}(\phi)$, the original M step corresponds to find the posterior mode if we assume a non-informative prior.

Historical Notes

EM was formalized as an approach to solving arbitrary maximum likelihood problems and named EM in [Dempster et al., 1977], but the idea was used earlier.

- In 1958 Hartley presented the main ideas of EM, rooted in the special case of counting data. [Hartley, 1958]
- Baum and Welch developed an algorithm for fitting hidden Markov models (HMMs) that is often called the Baum–Welch algorithm, which is equivalent to applying the EM algorithm. [Baum et al., 1970, Welch, 2003]
- The earliest reference to literature on an EM-type of algorithm is [Newcomb, 1886], who considers estimation of parameters of a mixture of two univariate normals

"I felt like the old minstrel who has been singing his song for 18 years and now finds, with considerable satisfaction, that his folklore is the theme of an overpowering symphony" —— Hartley

References I

Ali, M. M., Khompatraporn, C., and Zabinsky, Z. B. (2005).
A numerical evaluation of several stochastic algorithms on selected continuous global optimization test problems.
Journal of global optimization, 31:635–672.

Baum, L. E., Petrie, T., Soules, G., and Weiss, N. (1970).
A maximization technique occurring in the statistical analysis of probabilistic functions of markov chains.
The annals of mathematical statistics, 41(1):164–171.

Celeux, G. (1985).
The sem algorithm: a probabilistic teacher algorithm derived from the em algorithm for the mixture problem.
Computational statistics quarterly, 2:73–82.

References II

Dempster, A., Laird, N., and Rubin, D. (1977).
Maximum likelihood from incomplete data via the sems algorithm.
Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

Gupta, M. R., Chen, Y., et al. (2011).
Theory and use of the em algorithm.
Foundations and Trends® in Signal Processing, 4(3):223–296.

Hartley, H. O. (1958).
Maximum likelihood estimation from incomplete data.
Biometrics, 14(2):174–194.

References III

Meilijson, I. (1989).
A fast improvement to the em algorithm on its own terms.
Journal of the Royal Statistical Society Series B: Statistical Methodology, 51(1):127–138.

Newcomb, S. (1886).
A generalized theory of the combination of observations so as to obtain the best result.
American journal of Mathematics, pages 343–366.

Tanner, M. A. and Wong, W. H. (1987).
The calculation of posterior distributions by data augmentation.
Journal of the American statistical Association, 82(398):528–540.

References IV

Welch, L. R. (2003).
Hidden markov models and the baum-welch algorithm.
IEEE Information Theory Society Newsletter, 53(4):10–13.