

Lecture 29: The EM Algorithm — Motivation, and Gaussian Mixture Distributions

Charles A. Bouman

Purdue University

Monday, November 2, 2026

Outline

- 1 How should the model be chosen? Broad families with parameters
- 2 The catch: the undistorted image is not available
- 3 The plant-height example: labels X_n , heights Y_n
- 4 A heuristic: threshold, classify, average — and why it fails
- 5 Direct ML estimation of θ , and why it is hard
- 6 Missing data, complete data, incomplete data
- 7 Gaussian mixture distributions
- 8 ML estimates when the labels *are* known

Core idea

Two populations, lost labels. Separating the plants by height gives estimates that are **inconsistent**, not merely biased. ML on the incomplete data is the right goal, but $p(y|\theta)$ requires a sum over the missing labels. That is the problem the EM algorithm solves.

Today's Big Ideas

- 1 Model parameters must be estimated from indirect or incomplete measurements, because the undistorted image is not available.
- 2 The plant-height example holds the whole problem: there are two populations of plants, and the record of which plant belongs to which population was lost.
- 3 Separating the plants by a height threshold and averaging each group gives estimates that are inconsistent, so they do not approach the true means no matter how many plants are measured.
- 4 Maximum likelihood on the incomplete data is the right goal, but the likelihood is a sum over the missing labels with no closed form, and removing that obstacle is the purpose of the EM algorithm.
- 5 The Gaussian mixture distribution generalizes the example to any number of classes, and when the labels are known its parameters are estimated by simple counts and averages.

How should the model be chosen?

The previous chapters made a case that model-based image processing is promising. One crucial question that arises is **how the model should be chosen**. Clearly, a poor choice of models will produce bad results.

- One solution is to select a broad family of models that are controlled by a scalar or vector parameter.
- For example, the prior model for an image may be controlled by a parameter which determines properties such as variation or edginess.
- These model parameters can then be estimated along with the unknown image.

If the family of models is sufficiently broad, and the parameter estimation is sufficiently robust, then it should be possible to select a model from the family that accurately describes the data.

The catch

The undistorted image is typically not available

In fact, if it were available, then we would not need to solve the image restoration or reconstruction problem in the first place.

- So the parameters that specify a particular prior model must be estimated from **indirect measurements** of the corrupted, distorted, or transformed image.
- The problem of estimating parameters from indirect or incomplete observations is a recurring theme in model-based image processing.
- Among the tools available to address this problem, the **expectation-maximization (EM)** algorithm is perhaps the most widely used.

What EM does, at its core

At its core, the EM algorithm provides a two-step process for parameter estimation.

- In the first step, one conjectures values for the missing data and uses them to calculate an ML estimate of the unknown parameters (M-step).
- In the second step, the conjectured values are updated using the estimated parameter values (E-step).
- The two-step process is then repeated, with the hope of convergence.

Will it work?

This cyclic process is quite intuitive, but naive implementation of this approach leads to **very bad estimates**. The EM algorithm provides a formal framework for this iterative heuristic, which ensures well-defined properties of convergence and good behavior.

A concrete problem: plants in a garden

You have measured the height of each plant in a garden.

- There are N plants, and you know that some have been regularly fertilized, and the remainder have not been fertilized at all.
- Unfortunately, the **fertilizing records were lost**.
- Your measurements, Y_n , of the plant height for the n^{th} plant could have been modeled as Gaussian with mean μ and variance σ^2 , if they had all been treated equally, but since they have not, the fertilized plants will, on average, be taller.

We can model this unknown treatment of the plants by a random variable

$$X_n = \begin{cases} 1 & \text{if plant has been fertilized ,} \\ 0 & \text{if plant has **not** been fertilized} \end{cases} .$$

The conditional distribution of height given the label

If X_n were known, then the distribution of plant heights for each category could be assumed to be Gaussian, but with different means and perhaps the same variance, σ^2 . This implies that the conditional distribution of Y_n given X_n will be Gaussian:

$$p(y_n|x_n) = \begin{cases} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mu_1)^2\right\} & \text{for } x_n = 1, \\ \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_n - \mu_0)^2\right\} & \text{for } x_n = 0, \end{cases}$$

where μ_0 and μ_1 denote the means of the two different classes.

The labels, and the parameters of the model

It might also be reasonable to assume that the class labels of the plants, X_n , are i.i.d. Bernoulli random variables with

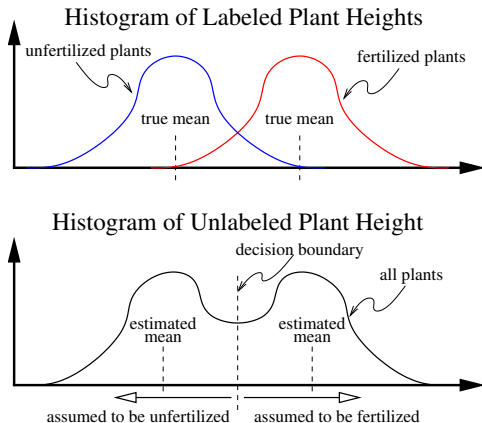
$$p(x_n) = \begin{cases} \pi_1 & \text{for } x_n = 1, \\ \pi_0 = 1 - \pi_1 & \text{for } x_n = 0, \end{cases}$$

where π_i parameterizes the Bernoulli distribution.

Doubly stochastic process

A **doubly stochastic process** is a process with this hierarchical structure, in which the distribution of Y_n depends on the value of the random variable X_n . The parameters of this doubly stochastic process, (X_n, Y_n) , are then given by $\theta = [\pi_0, \mu_0, \pi_1, \mu_1]$, where $\pi_0 + \pi_1 = 1$.

Two populations make two modes



The distributions for fertilized and unfertilized plants are different, with the fertilized plants having a larger mean height. However, if the labels of the plants are unknown, then the distribution appears to have two modes, as shown in the second plot. Notice that naive estimates of the mean are systematically offset from the true values.

Why we care, and what we might try

- Estimating θ is an important practical problem: we may want to measure the effect of fertilization on plant growth, so we would like to know how much μ_0 and μ_1 differ.
- To estimate the mean of each mode, it seems that we would need to know X_n , the label of each plant.
- However, casual inspection of the distribution suggests that one might be able to estimate the unknown means, μ_0 and μ_1 , by looking at the combined distribution of the two populations.

First idea

One possibility for estimating μ_0 and μ_1 is to first estimate the labels X_n .

The threshold heuristic

Apply a threshold at the valley between the two modes, and assign a class of “fertilized” ($X_n = 1$) or “unfertilized” ($X_n = 0$) to each plant:

$$\hat{X}_n = \begin{cases} 1, & Y_n \geq \text{threshold} \\ 0, & Y_n < \text{threshold} \end{cases}.$$

The result of this classification can be compactly represented by the two sets $S_0 = \{n : \hat{X}_n = 0\}$ and $S_1 = \{n : \hat{X}_n = 1\}$. Using this notation, we can estimate the two unknown means as

$$\begin{aligned}\hat{\mu}_0 &= \frac{1}{|S_0|} \sum_{n \in S_0} Y_n, \\ \hat{\mu}_1 &= \frac{1}{|S_1|} \sum_{n \in S_1} Y_n,\end{aligned}$$

where $|S_0|$ and $|S_1|$ denote the number of plants that have been classified as unfertilized and fertilized, respectively.

A very serious flaw

While this is an intuitively appealing approach, it has a very serious flaw.

- Since we have separated the two groups by their height, it is inevitable that we will measure a larger value for $\hat{\mu}_1$ than we should (and also a smaller value of $\hat{\mu}_0$ than we should).
- Even when the two means are quite different, the resulting estimates of $\hat{\mu}_0$ and $\hat{\mu}_1$ will be systematically shifted **no matter how many plants, N , are measured**.

Worse than bias

This is much worse than simply being biased. These estimates are **inconsistent** because the estimated values do not converge to the true parameters as $N \rightarrow \infty$.

Direct ML estimation of θ

Another approach is to attempt to directly estimate the value of the parameter vector $\theta = [\pi_0, \mu_0, \pi_1, \mu_1]$ from the data Y using the ML estimate. Formally,

$$\hat{\theta} = \arg \max_{\theta \in \Omega} \log p(y|\theta) .$$

This seems to be a much more promising approach, since it is known that the ML estimate is not only **consistent** but **asymptotically efficient** (Section 2.3).

However, there is still a problem

The distribution $p(y|\theta)$ is not explicitly available for this problem. In fact, it requires the evaluation of a sum that makes direct ML estimation difficult:

$$p(y|\theta) = \sum_x p(y|x, \theta)p(x|\theta) .$$

So we will see that it is no longer possible to calculate simple closed-form expressions for the ML estimate of θ .

The purpose of the EM algorithm

The purpose of the EM algorithm is to provide a systematic methodology for estimating parameters, such as μ_k , when there is missing data.

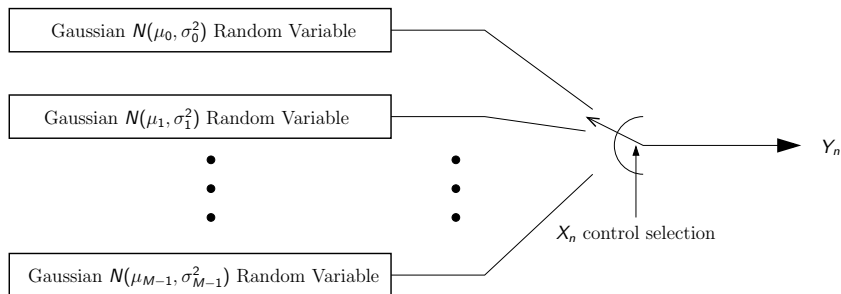
Vocabulary for this problem

- The **missing data** is the set of unknown labels, X .
- The **complete data** is the combination of (X, Y) .
- The **incomplete data** is Y alone.

The EM algorithm provides a method of computing an ML estimate of the parameter vector, θ , from the incomplete data, Y . The derivation is quite ingenious, but in the end, the resulting algorithms are very intuitive. In fact, EM is more than a simple algorithm: it is a way of thinking about incomplete observations of data and the associated optimization algorithms required for ML parameter estimation.

Gaussian mixture distributions

Generalize the previous example: the random variables $\{Y_n\}_{n=1}^N$ are assumed to be conditionally Gaussian and independent given the class labels $\{X_n\}_{n=1}^N$. To be a bit more general, each label takes on M possible values, so that $X_n \in \{0, \dots, M-1\}$.



Each observation can come from one of M possible Gaussian distributions, and the specific Gaussian distribution is selected using a switch controlled by the class label X_n .

The conditional and marginal distributions

For the m^{th} Gaussian distribution, the mean and variance are denoted by μ_m and σ_m^2 , and the distribution by $N(\mu_m, \sigma_m^2)$. If the control variable takes on the value $m = x_n$, then Y_n is conditionally Gaussian with mean μ_{x_n} and variance $\sigma_{x_n}^2$:

$$p(y_n|x_n) = \frac{1}{\sqrt{2\pi\sigma_{x_n}^2}} \exp \left\{ -\frac{1}{2\sigma_{x_n}^2} (y_n - \mu_{x_n})^2 \right\} .$$

From this, we can calculate the marginal distribution of y_n :

$$\begin{aligned} p(y_n) &= \sum_{m=0}^{M-1} p(y_n|m) \pi_m \\ &= \sum_{m=0}^{M-1} \frac{\pi_m}{\sqrt{2\pi\sigma_m^2}} \exp \left\{ -\frac{1}{2\sigma_m^2} (y_n - \mu_m)^2 \right\} , \end{aligned}$$

where $\pi_m = P\{X_n = m\}$.

(Book equation 12.2)

The full sequence, and the parameter vector

If we further assume that the samples, Y_n , are i.i.d., then the distribution of the entire sequence, $\{Y_n\}_{n=1}^N$, can be written as

$$p(y) = \prod_{n=1}^N \sum_{m=0}^{M-1} \frac{\pi_m}{\sqrt{2\pi\sigma_m^2}} \exp \left\{ -\frac{1}{2\sigma_m^2} (y_n - \mu_m)^2 \right\} .$$

A **Gaussian mixture distribution** is a distribution with the form of the marginal above. Its parameter vector is

$$\theta = [\pi_0, \mu_0, \sigma_0^2, \dots, \pi_{M-1}, \mu_{M-1}, \sigma_{M-1}^2] .$$

Why mixtures are useful

For sufficiently large M , Gaussian mixtures can approximate almost any distribution. Intuitively, any density function can be approximated by a weighted sum of Gaussian density functions.

ML estimates when the labels *are* known

How can we estimate θ from observations of Y_n ? For now, make the simplifying assumption that we know both the values Y_n and their associated labels X_n . First compute N_m , the number of labels taking on class m :

$$N_m = \sum_{n=1}^N \delta(X_n - m) ,$$

where $\delta(k)$ is one when $k = 0$ and zero otherwise. Then the ML estimate of π_m is the fraction of labels taking on the class m ; the ML estimate of the mean averages only the values of Y_n with class label $X_n = m$; and the ML estimate of the variance sums only over terms with class label $X_n = m$:

$$\hat{\pi}_m = \frac{N_m}{N} , \quad \hat{\mu}_m = \frac{1}{N_m} \sum_{n=1}^N Y_n \delta(X_n - m) ,$$

$$\hat{\sigma}_m^2 = \frac{1}{N_m} \sum_{n=1}^N (Y_n - \hat{\mu}_m)^2 \delta(X_n - m) .$$

Where this leaves us

- Of course, for our problem the class labels, X_n , are not known, so the problem remains of how we are going to calculate the ML estimate of θ from only the incomplete data Y .
- The EM algorithm provides a solution in the form of an algorithm for computing the ML parameter estimate that **does not require the missing data**.
- In essence, this is done by replacing the unknown labels, X_n , with their **expected values**.

However, in order to fully understand this, we must first introduce the underlying theory of the method.

Summary

- Model parameters must be estimated from indirect or incomplete measurements, because the undistorted image is not available.
- Plant heights: labels $X_n \in \{0, 1\}$ Bernoulli with π_1 ; Y_n given X_n Gaussian with means μ_0, μ_1 and shared σ^2 ; $\theta = [\pi_0, \mu_0, \pi_1, \mu_1]$.
- Thresholding and averaging within S_0 and S_1 gives estimates that are systematically shifted for every N — inconsistent, not just biased.
- ML on the incomplete data, $\hat{\theta} = \arg \max_{\theta \in \Omega} \log p(y|\theta)$, is consistent and asymptotically efficient, but $p(y|\theta) = \sum_x p(y|x, \theta)p(x|\theta)$ has no simple closed form. The missing data is X , the complete data is (X, Y) , and the incomplete data is Y .
- A Gaussian mixture selects among M classes with a switch, so $\theta = [\pi_0, \mu_0, \sigma_0^2, \dots, \pi_{M-1}, \mu_{M-1}, \sigma_{M-1}^2]$. With known labels, ML gives $\hat{\pi}_m = N_m/N$, $\hat{\mu}_m$, and $\hat{\sigma}_m^2$ — the estimates EM will compute without the labels.

Reading: Sections 12.1–12.2. Next lecture: Sections 12.3–12.4, theoretical foundation of the EM algorithm.