18.650: Fundamentals of Statistics October 4 & 6, 2023

Lecture 10 & 11— MLE Algos, EM for mixtures, Method of Moments

Prof. Philippe Rigollet *Scribe: Anya Katsevich*

Recap & Motivation. We saw in Lecture 9 that for a regular statistical model — meaning, a model for which we can find the MLE by solving $\nabla \ell_n(\theta) = 0$ for θ — then $\hat{\theta}^{\mathrm{MLE}}$ is asymptotically normal, and

$$\sqrt{n}\left(\hat{\theta}^{\mathrm{MLE}} - \theta^*\right) \rightsquigarrow \mathcal{N}(0, 1/I(\theta^*)).$$

Here, $I(\theta^*)$ is the Fisher information. Using this asymptotic variance, we can build confidence intervals. For example, the following is an approximate 95% CI:

$$\hat{\theta}^{\mathrm{MLE}} \pm \frac{1.96}{\sqrt{nI(\hat{\theta}^{\mathrm{MLE}})}}.$$

Note that we used $I(\hat{\theta}^{\mathrm{MLE}})$ in place of $I(\theta^*)$, since we don't know θ^*.

For example, for the Bernoulli(p) model, one can compute that the MLE is given by $\hat{p}^{\mathrm{MLE}} = \bar{X}_n$, and the Fisher information $I(p) = 1/p(1-p)$. This implies

$$\sqrt{n}(\bar{X}_n - p^*) \rightsquigarrow \mathcal{N}(0, p^*(1-p^*)).$$

In this example, we did not actually need to appeal to the Fisher information. We could have computed the asymptotic variance directly by the CLT.

But there are other cases in which the MLE is not a sample average, and is not even available in closed form. In such cases, we cannot directly analyze the MLE, so it is very useful to have available the general formula for the asymptotic variance in terms of $I(\theta^*)$.

1 Algorithms to compute the MLE

Recall that the MLE is the solution $\hat{\theta}^{\mathrm{MLE}} = \mathrm{argmax}_\theta\, \ell_n(\theta)$. This is an optimization problem, so we can use one of the many optimization algorithms out there.

1. Gradient ascent. The most common algorithm used these days is gradient descent (or in our case, gradient ascent, since we're maximizing rather than minimizing):

$$\theta_{j+1} = \theta_j + \eta \nabla_\theta \ell_n(\theta_j), \quad j = 0, 1, 2, \ldots.$$

This is an iterative algorithm in which we "climb the hill" of the function ℓ_n (the gradient points in the direction of steepest ascent). Here, η is a step size, typically $\eta \in (0, 1)$. We terminate the algorithm once the magnitude of the gradient gets small enough; or once the change in the function value from one iteration to the next is small enough.

2. Newton-Raphson. Newton's method is a root-finding algorithm, i.e. it solves $f(x) = 0$ for x. Our optimization problem can also be framed this way, since we're looking for a point θ such that $\nabla \ell_n(\theta) = 0$, so $f = \nabla \ell_n$ in our case. We solve the equation by linearizing (Taylor-expanding) $\nabla \ell_n$ around our current estimate θ_j. In 1-d, we get the following linear approximation:

$$\ell_n'(\theta) \approx \ell_n'(\theta_j) + \ell_n''(\theta_j)(\theta - \theta_j).$$

We then set this linear approximation to zero and solve for θ. The result is our updated estimate θ_{j+1}.

$$0 = \ell_n'(\theta_j) + \ell_n''(\theta_j)(\theta - \theta_j) \implies \theta = \theta_j - \frac{\ell_n'(\theta_j)}{\ell_n''(\theta_j)}.$$

To summarize, the Newton-Raphson algorithm takes the form

$$\theta_{j+1} = \theta_j - \frac{\ell_n'(\theta_j)}{\ell_n''(\theta_j)}, \quad j = 0, 1, 2, \ldots$$

Note that this method has no step size parameter. This algorithm works well when it is initialized close to the true maximizer of ℓ_n, but it can be unstable when initialized far away. In higher dimensions, Newton-Raphson takes the form

$$\theta_{j+1} = \theta_j - \nabla^2 \ell_n(\theta_j)^{-1} \nabla \ell_n(\theta_j).$$

This is computationally heavier than gradient ascent, since you have to invert a matrix.

2 Mixture Models and the Expectation Maximization (EM) Algorithm

The EM algorithm is a method of approximately finding the MLE in mixture models.

2.1 Mixture Models

Mixture models are used when data is heterogeneous; e.g. if we measure the heights of a group of men and women, the distribution of the data will have two modes. Karl Pearson conducted one of the first rigorous studies of a mixture model in the late

1800s: he studied crabs on the beach. He suspected that there were two different kinds of crab, based on their body proportions.

What makes parameter estimation difficult in mixture models is that we don't know which of the groups each data point belongs to; the crabs don't have a label "0" or "1" on their backs.

Formally, suppose we have two pdfs f_0, f_1. Then $f = (1-p)f_0 + pf_1$ is a mixture of the two pdfs, with $p \in (0,1)$. We can construct a random variable with the pdf f as follows: let $Z \sim \text{Ber}(p)$, X_0 have pdf f_0, and X_1 have pdf f_1. First, draw $Z \in \{0,1\}$. This is the latent membership variable which we don't observe. Then, let

$$Y = \begin{cases} X_0, & \text{if } Z = 0, \\ X_1, & \text{if } Z = 1. \end{cases}$$

In other words, draw X_0 if $Z = 0$ and draw X_1 if $Z = 1$. In condensed form:

$$Y = (1-Z)X_0 + ZX_1.$$

It can be shown that the pdf of Y is indeed $(1-p)f_0 + pf_1$. A common kind of mixture is a Gaussian mixture:

$$f_0 = \mathcal{N}(\mu_0, \sigma_0^2), \quad f_1 = \mathcal{N}(\mu_1, \sigma_1^2).$$

There are five unknown parameters: $\theta = (\mu_0, \sigma_0^2, \mu_1, \sigma_1^2, p)$. But for now, we will suppose for simplicity that $p, \sigma_0^2, \sigma_1^2$ are known, with $p = 1/2$ and $\sigma_0^2 = \sigma_1^2 = 1$.

2.2 Intuition for estimating mixture parameters

Given observations $Y_1, \ldots, Y_n$ from the mixture $f = \frac{1}{2}f_0 + \frac{1}{2}f_1$, where $f_0 = \mathcal{N}(\mu_0, 1)$ and $f_1 = \mathcal{N}(\mu_1, 1)$, our goal is to estimate the two-dimensional parameter $\theta = (\mu_0, \mu_1)$. The log likelihood for this model is

$$\ell_n(\mu_0, \mu_1) = \sum_{i=1}^{n} \log \left(\frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i - \mu_0)^2}{2}} + \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{(Y_i - \mu_1)^2}{2}} \right) \tag{1}$$

This is extremely difficult to optimize. How do we do better? To gain intuition, suppose the crabs *did* have numbers on their backs. Intuitively, the best strategy would be to separate the crabs into two groups and compute the Gaussian MLE separately for each group. Formally, the crabs having their number on their back corresponds to observing the latent (hidden) membership variable $Z_i \in \{0,1\}$ for each crab, so that our data is

$$(Y_1, Z_1), \ldots, (Y_n, Z_n).$$

Then the log likelihood would be

$$\begin{aligned}\ell_n(\mu_0,\mu_1) &= \sum_{i=1}^n \log\left(f_0(Y_i)^{1-Z_i} f_1(Y_i)^{Z_i}\right)\\ &= \sum_{i=1}^n \left[(1-Z_i)\log f_0(Y_i) + Z_i \log f_1(Y_i)\right]\\ &= -\frac{1}{2}\sum_{i=1}^n \left[(1-Z_i)(Y_i-\mu_0)^2 + Z_i(Y_i-\mu_1)^2\right] + \text{const.}\\ &= -\frac{1}{2}\sum_{i:Z_i=0}(Y_i-\mu_0)^2 - \frac{1}{2}\sum_{i:Z_i=1}(Y_i-\mu_1)^2 + \text{const.}\end{aligned} \tag{2}$$

using that $f_0 = \mathcal{N}(\mu_0, 1)$ has density $\frac{1}{\sqrt{2\pi}}e^{-(x-\mu_0)^2/2}$ and similarly for f_1. The "const" refers to $n\log(1/\sqrt{2\pi})$ which does not affect the optimization of ℓ_n.

Remark.

> In the first line, we used the compact representation $f_0(Y_i)^{1-Z_i} f_1(Y_i)^{Z_i}$ to express that if $Z_i = 1$ then the density is f_1 and if $Z_i = 0$ then the density is f_0.

The key simplification in (2) compared to (1) is that in place of a logarithm of the sum, we now have a sum of logarithms. We can now easily optimize (2) with respect to μ_0, μ_1, using either the formula in the third or fourth line. Taking the gradient with respect to μ_0, μ_1 and setting it to zero, we get

$$\begin{aligned}\hat{\mu}_0 &= \frac{\sum_{i=1}^n (1-Z_i)Y_i}{\sum_{i=1}^n (1-Z_i)} = \frac{\sum_{i:Z_i=0} Y_i}{\#\{i : Z_i = 0\}},\\ \hat{\mu}_1 &= \frac{\sum_{i=1}^n Z_i Y_i}{\sum_{i=1}^n Z_i} = \frac{\sum_{i:Z_i=1} Y_i}{\#\{i : Z_i = 1\}}.\end{aligned} \tag{3}$$

This is precisely what we would get by separating the Y_i into two groups based on their level, and then computing the Gaussian MLE separately for each group.

2.3 The EM algorithm

In reality, we do not have access to the labels Z_i. So we simply estimate them based on the Y_i, and replace Z_i by $\hat{Z}_i$ in the third line of (2), the log likelihood in the case of observed Z_i's. To estimate Z_i based on Y_i, we need to know how close Y_i is to μ_0 and to μ_1, since this is what determines how likely it is that Y_i came from $f_0 = \mathcal{N}(\mu_0, 1)$ or from $f_1 = \mathcal{N}(\mu_1, 1)$. However, we don't know μ_0, μ_1! So instead we use an estimate of μ_0, μ_1 to compute $\hat{Z}_i$. We then use $\hat{Z}_i$ to improve our estimate of μ_0, μ_1.

Estimating Z_i using estimate of μ_0, μ_1. (E step) To compute an estimate $\hat{Z}_i$ of

Z_i, we compute the expectation of Z_i given Y_i, which is equivalent to the probability that $Z_i = 1$ given Y_i. Specifically, we take

$$\begin{aligned}
\hat{Z}_i &= \mathbb{E}[Z_i \mid Y_i] = \mathbb{P}(Z_i = 1 \mid Y_i) \\
&= \frac{f(Y_i|Z_i = 1)\mathbb{P}(Z_i = 1)}{f(Y_i|Z_i = 1)\mathbb{P}(Z_i = 1) + f(Y_i|Z_i = 0)\mathbb{P}(Z_i = 0)} \\
&= \frac{f(Y_i|Z_i = 1)}{f(Y_i|Z_i = 1) + f(Y_i|Z_i = 0)} \\
&= \frac{e^{-(Y_i-\mu_1)^2/2}}{e^{-(Y_i-\mu_1)^2/2} + e^{-(Y_i-\mu_0)^2/2}}.
\end{aligned} \tag{4}$$

We used Bayes Rule to get the second line, and the fact that $\mathbb{P}(Z_i = 1) = \mathbb{P}(Z_i = 0) = 1/2$ to get the third line. To get the fourth line, we used that $f(Y_i \mid Z_i = 1) = f_1(Y_i)$ and $f(Y_i \mid Z_i = 0) = f_0(Y_i)$. In the fourth line, we use our current estimate of μ_1 and μ_0. Note that $\hat{Z}_i$ is guaranteed to be a number between 0 and 1. If Y_i is very close to μ_1, then $\hat{Z}_i \approx 1/(1 + e^{-(\mu_1-\mu_0)^2/2})$, so the farther apart μ_1 and μ_0, the closer $\hat{Z}_i$ is to 1 (i.e. we are very sure Y_i came from f_1 if $Y_i \approx \mu_1$ and μ_0, μ_1 are very far apart).

Estimating μ_0, μ_1 using the $\hat{Z}_i$. (M step) We now plug in $\hat{Z}_i$ in the third line of (2) to get an approximate log likelihood:

$$\begin{aligned}
\hat{\ell}_n(\mu_0, \mu_1) &= \sum_{i=1}^n (1 - \hat{Z}_i) \log f_0(Y_i) + \sum_{i=1}^n \hat{Z}_i \log f_1(Y_i) \\
&= \sum_{i=1}^n (1 - \hat{Z}_i) \log \left[\frac{1}{\sqrt{2\pi}} e^{-(Y_i-\mu_0)^2/2}\right] + \sum_{i=1}^n (1 - \hat{Z}_i) \log \left[\frac{1}{\sqrt{2\pi}} e^{-(Y_i-\mu_1)^2/2}\right] \\
&= -\frac{1}{2}\sum_{i=1}^n (1 - \hat{Z}_i)(Y_i - \mu_0)^2 - \frac{1}{2}\sum_{i=1}^n \hat{Z}_i (Y_i - \mu_1)^2 + \text{const}
\end{aligned}$$

The sum involving μ_0 is not interacting with the sum involving μ_1. We maximize by setting the gradient to zero (exercise!) to get

$$\hat{\mu}_0 = \frac{\sum_{i=1}^n (1 - \hat{Z}_i) Y_i}{\sum_{i=1}^n (1 - \hat{Z}_i)}, \qquad \hat{\mu}_1 = \frac{\sum_{i=1}^n \hat{Z}_i Y_i}{\sum_{i=1}^n \hat{Z}_i}.$$

Compare this formula to (3), which we got in the case where we knew the Z_i's. It's exactly the same, expect that we have replaced Z_i by $\hat{Z}_i$.

The above steps are summarized in Algorithm 1.

Algorithm 1 EM Algorithm (simplified setup)

```
Initialize $\mu_0^{(0)}, \mu_1^{(0)}$
for $j = 0, 1, 2, 3, \ldots$ do
    $$\hat{Z}_i^{(j+1)} \leftarrow \frac{e^{-(Y_i-\mu_1^{(j)})^2/2}}{e^{-(Y_i-\mu_1^{(j)})^2/2} + e^{-(Y_i-\mu_0^{(j)})^2/2}}, \quad i = 1, \ldots, n.$$
                                                                    ▷ E step
    $$\mu_0^{(j+1)} \leftarrow \frac{\sum_{i=1}^n (1-\hat{Z}_i^{(j+1)}) Y_i}{\sum_{i=1}^n (1-\hat{Z}_i^{(j+1)})}, \qquad \mu_1^{(j+1)} \leftarrow \frac{\sum_{i=1}^n \hat{Z}_i^{(j+1)} Y_i}{\sum_{i=1}^n \hat{Z}_i^{(j+1)}}.$$
                                                                    ▷ M step
end for
```

Next we consider a more general case, in which the mixture is given by

$$f(y) = (1-p) f_{\theta_0}(y) + p g_{\theta_1}(y).$$

Here, f_θ and g_θ do not even need to belong to the same family; e.g. f_θ could be a normal distribution and g_θ an exponential distribution. The parameters θ_0, θ_1 could be multi-dimensional, e.g. $\theta_0 = (\mu_0, \sigma_0^2)$ in the case of a normal distribution. Note

Algorithm 2 EM Algorithm (general case)

```
Initialize $\theta_0^{(0)}, \theta_1^{(0)}, p^{(0)}$
for $j = 0, 1, 2, 3, \ldots$ do
    $$\hat{Z}_i^{(j+1)} \leftarrow \frac{g_{\theta_1^{(j)}}(Y_i) p^{(j)}}{g_{\theta_1^{(j)}}(Y_i) p^{(j)} + f_{\theta_0^{(j)}}(Y_i)(1-p^{(j)})}$$
    $$p^{(j+1)} \leftarrow \frac{1}{n} \sum_{i=1}^n \hat{Z}_i^{(j+1)}.$$
                                                                    ▷ E step
    $$\theta_0^{(j+1)} \leftarrow \text{argmax}_{\theta_0} \sum_{i=1}^n (1-\hat{Z}_i^{(j+1)}) \log f_{\theta_0}(Y_i)$$
    $$\theta_1^{(j+1)} \leftarrow \text{argmax}_{\theta_1} \sum_{i=1}^n \hat{Z}_i^{(j+1)} \log g_{\theta_1}(Y_i)$$
                                                                    ▷ M step
end for
```

that the M step is the same as setting

$$(\theta_0^{(j+1)}, \theta_1^{(j+1)}) \leftarrow \text{argmax}_{\theta_0,\theta_1} \sum_{i=1}^{n} \left[(1 - \hat{Z}_i^{(j+1)}) \log f_{\theta_0}(Y_i) + \hat{Z}_i^{(j+1)} \log g_{\theta_1}(Y_i) \right].$$

We can break up the sum into two sums, one involving only θ_0 and the other involving only θ_1. We can then maximize the two sums separately.

Remark.

> Sometimes we cannot take the logarithm of the likelihood, e.g. if the likelihood takes the value zero, as for the uniform distribution. We can them replace the M step by the following maximization of the likelihood instead of the log likelihood:

$$\theta_0^{(j+1)} \leftarrow \text{argmax}_{\theta_0} \prod_{i=1}^{n} f_{\theta_0}(Y_i)^{1-\hat{Z}_i^{(j+1)}}$$
$$\theta_1^{(j+1)} \leftarrow \text{argmax}_{\theta_1} \prod_{i=1}^{n} g_{\theta_1}(Y_i)^{\hat{Z}_i^{(j+1)}}$$

3 Method of Moments

The method of moments is an alternative way to estimate parameters.

Definition 3.1: Moment

The jth moment of X is $\alpha_j = \mathbb{E}[X^j]$.

Using the plug-in method, we can approximate α_j by

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j. \tag{5}$$

Now, suppose the pdf of X is f_θ. Then each moment α_j is some function of θ:

$$\alpha_j(\theta) = \mathbb{E}_\theta[X^j] = \int x^j f_\theta(x) dx. \tag{6}$$

Example.

> Consider $f_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$, i.e. $\theta = (\mu, \sigma^2)$ and f_θ is a normal distribution. Then the functions $\alpha_1(\theta)$ and $\alpha_2(\theta)$ are given as follows:
>
> $$\alpha_1(\mu, \sigma^2) = \mu, \qquad \alpha_2(\mu, \sigma^2) = \mu^2 + \sigma^2, \tag{7}$$
>
> since $\alpha_1 = \mathbb{E}[X] = \mu$ and $\alpha_2 = \mathbb{E}[X^2] = \sigma^2 + \mu^2$.

The main insight of the method of moments is that we can solve for θ in the following equation:

$$\alpha_j(\theta) = \hat{\alpha}_j$$

If θ actually consists of several parameters, e.g. $\theta = (\mu, \sigma^2)$, then a single equation won't be enough to determine θ uniquely. So we compute as many moments as there are unknown parameters, and solve a system of equations.

Definition 3.2: Method of Moments (MoM)

Suppose we are given samples $X_i \overset{\text{i.i.d.}}{\sim} f_\theta$, $i = 1, 2, \ldots, n$, where $\theta = (\theta_1, \ldots, \theta_k)$ is a k-dimensional vector of all the unknown parameters. Compute

$$\hat{\alpha}_j = \frac{1}{n} \sum_{i=1}^{n} X_i^j, \quad j = 1, \ldots, k,$$

i.e. the plug-in estimators for the first k moments. Then the method of moments estimator $\hat{\theta}$ is the solution θ to the following system of equations:

$$\begin{cases} \alpha_1(\theta) = \hat{\alpha}_1, \\ \alpha_2(\theta) = \hat{\alpha}_2, \\ \qquad \cdots \\ \alpha_k(\theta) = \hat{\alpha}_k. \end{cases} \tag{8}$$

Example.

> Returning to the normal example, let $f_{\mu,\sigma^2} = \mathcal{N}(\mu, \sigma^2)$. Recall that $\alpha_1(\mu, \sigma^2) = \mu$ and $\alpha_2(\mu, \sigma^2) = \mu^2 + \sigma^2$, i.e. $\theta = (\mu, \sigma^2)$ and f_θ is a normal distribution. Using the functions $\alpha_1(\theta)$, $\alpha_2(\theta)$ from (7), we see that we need to solve the following system of two equations for μ and σ^2:
>
> $$\begin{cases} \mu & = \hat{\alpha}_1, \\ \mu^2 + \sigma^2 & = \hat{\alpha}_2, \end{cases} \tag{9}$$
>
> The solution is $\mu = \hat{\alpha}_1$ and $\sigma^2 = \hat{\alpha}_2 - \hat{\alpha}_1^2$. Recalling from (5) how the $\hat{\alpha}_k$ are constructed, we see that the method of moments estimator is
>
> $$\hat{\mu} = \hat{\alpha}_1 = \frac{1}{n} \sum_{i=1}^{n} X_i,$$
>
> $$\hat{\sigma}^2 = \hat{\alpha}_2 - \hat{\alpha}_1^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n} \sum_{i=1}^{n} X_i \right)^2.$$

For the normal distribution, the system of equations (9) was solvable in closed form. But more generally, we can't always hope to explicitly solve the system of equations (8). When this isn't possible, we can use an iterative numerical method such as Newton-Raphson instead.

Like the MLE, the method of moments estimator is consistent and asymptotically normal.

Theorem 3.3: Asymptotic properties of the MoM estimator

Under appropriate regularity conditions, the method of moments estimator $\hat{\theta}_n$ satisfies the following properties:

1. Consistency: $\hat{\theta}_n \xrightarrow{\mathbb{P}} \theta$ as $n \to \infty$
2. Asymptotic normality: $\sqrt{n}(\hat{\theta}_n - \theta) \rightsquigarrow \mathcal{N}(0, \Sigma)$.

See Theorem 9.6 in AoS for the asymptotic variance Σ.