9.47 Theorem. *Let X have density in an exponential family. Then,*

$$\mathbb{E}(T(X)) = A'(\eta),\ \mathbb{V}(T(X)) = A''(\eta).$$

If $\theta = (\theta_1, \ldots, \theta_k)$ is a vector, then we say that $f(x;\theta)$ has exponential family form if

$$f(x;\theta) = h(x) \exp\left\{ \sum_{j=1}^{k} \eta_j(\theta) T_j(x) - B(\theta) \right\}.$$

Again, $T = (T_1, \ldots, T_k)$ is sufficient. An IID sample of size n also has exponential form with sufficient statistic $(\sum_i T_1(X_i), \ldots, \sum_i T_k(X_i))$.

9.48 Example. Consider the normal family with $\theta = (\mu, \sigma)$. Now,

$$f(x;\theta) = \exp\left\{ \frac{\mu}{\sigma^2} x - \frac{x^2}{2\sigma^2} - \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right) \right\}.$$

This is exponential with

$$\eta_1(\theta) = \frac{\mu}{\sigma^2},\ T_1(x) = x$$

$$\eta_2(\theta) = -\frac{1}{2\sigma^2},\ T_2(x) = x^2$$

$$B(\theta) = \frac{1}{2}\left(\frac{\mu^2}{\sigma^2} + \log(2\pi\sigma^2) \right),\ h(x) = 1.$$

Hence, with n IID samples, $(\sum_i X_i, \sum_i X_i^2)$ is sufficient. ∎

As before we can write an exponential family as

$$f(x;\eta) = h(x) \exp\left\{ T^T(x)\eta - A(\eta) \right\},$$

where $A(\eta) = \log \int h(x) e^{T^T(x)\eta} dx$. It can be shown that

$$\mathbb{E}(T(X)) = \dot{A}(\eta)\ \ \mathbb{V}(T(X)) = \ddot{A}(\eta),$$

where the first expression is the vector of partial derivatives and the second is the matrix of second derivatives.

9.13.4 Computing Maximum Likelihood Estimates

In some cases we can find the MLE $\widehat{\theta}$ analytically. More often, we need to find the MLE by numerical methods. We will briefly discuss two commonly

used methods: (i) Newton-Raphson, and (ii) the EM algorithm. Both are iterative methods that produce a sequence of values $\theta^0, \theta^1, \ldots$ that, under ideal conditions, converge to the MLE $\widehat{\theta}$. In each case, it is helpful to use a good starting value θ^0. Often, the method of moments estimator is a good starting value.

NEWTON-RAPHSON. To motivate Newton-Raphson, let's expand the derivative of the log-likelihood around θ^j:

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta^j) + (\widehat{\theta} - \theta^j)\ell''(\theta^j).$$

Solving for $\widehat{\theta}$ gives

$$\widehat{\theta} \approx \theta^j - \frac{\ell'(\theta^j)}{\ell''(\theta^j)}.$$

This suggests the following iterative scheme:

$$\widehat{\theta}^{j+1} = \theta^j - \frac{\ell'(\theta^j)}{\ell''(\theta^j)}.$$

In the multiparameter case, the mle $\widehat{\theta} = (\widehat{\theta}_1, \ldots, \widehat{\theta}_k)$ is a vector and the method becomes

$$\widehat{\theta}^{j+1} = \theta^j - H^{-1}\ell'(\theta^j)$$

where $\ell'(\theta^j)$ is the vector of first derivatives and H is the matrix of second derivatives of the log-likelihood.

THE EM ALGORITHM. The letters EM stand for Expectation-Maximization. The idea is to iterate between taking an expectation then maximizing. Suppose we have data Y whose density $f(y;\theta)$ leads to a log-likelihood that is hard to maximize. But suppose we can find another random variable Z such that $f(y;\theta) = \int f(y,z;\theta)\,dz$ and such that the likelihood based on $f(y,z;\theta)$ is easy to maximize. In other words, the model of interest is the marginal of a model with a simpler likelihood. In this case, we call Y the observed data and Z the hidden (or latent or missing) data. If we could just "fill in" the missing data, we would have an easy problem. Conceptually, the EM algorithm works by filling in the missing data, maximizing the log-likelihood, and iterating.

9.49 Example (Mixture of Normals). Sometimes it is reasonable to assume that the distribution of the data is a mixture of two normals. Think of heights of people being a mixture of men and women's heights. Let $\phi(y;\mu,\sigma)$ denote a normal density with mean μ and standard deviation σ. The density of a mixture of two Normals is

$$f(y;\theta) = (1-p)\phi(y;\mu_0,\sigma_0) + p\phi(y;\mu_1,\sigma_1).$$

The idea is that an observation is drawn from the first normal with probability p and the second with probability $1-p$. However, we don't know which Normal it was drawn from. The parameters are $\theta = (\mu_0, \sigma_0, \mu_1, \sigma_1, p)$. The likelihood function is

$$\mathcal{L}(\theta) = \prod_{i=1}^{n} \left[(1-p)\phi(y_i; \mu_0, \sigma_0) + p\phi(y_i; \mu_1, \sigma_1)\right].$$

Maximizing this function over the five parameters is hard. Imaging that we were given extra information telling us which of the two normals every observation came from. These "complete" data are of the form $(Y_1, Z_1), \ldots, (Y_n, Z_n)$, where $Z_i = 0$ represents the first normal and $Z_i = 1$ represents the second. Note that $\mathbb{P}(Z_i = 1) = p$. We shall soon see that the likelihood for the complete data $(Y_1, Z_1), \ldots, (Y_n, Z_n)$ is much simpler than the likelihood for the observed data $Y_1, \ldots, Y_n$. ■

Now we describe the EM algorithm.

The EM Algorithm

(0) Pick a starting value θ^0. Now for $j = 1, 2, \ldots$, repeat steps 1 and 2 below:

(1) (The E-step): Calculate

$$J(\theta|\theta^j) = \mathbb{E}_{\theta^j}\left(\log \frac{f(Y^n, Z^n; \theta)}{f(Y^n, Z^n; \theta^j)} \,\middle|\, Y^n = y^n\right).$$

The expectation is over the missing data Z^n treating θ^i and the observed data Y^n as fixed.

(2) Find θ^{j+1} to maximize $J(\theta|\theta^j)$.

We now show that the EM algorithm always increases the likelihood, that is, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$. Note that

$$\begin{aligned} J(\theta^{j+1}|\theta^j) &= \mathbb{E}_{\theta^j}\left(\log \frac{f(Y^n, Z^n; \theta^{j+1})}{f(Y^n, Z^n; \theta^j)} \,\middle|\, Y^n = y^n\right) \\ &= \log \frac{f(y^n; \theta^{j+1})}{f(y^n; \theta^j)} + \mathbb{E}_{\theta^j}\left(\log \frac{f(Z^n|Y^n; \theta^{j+1})}{f(Z^n|Y^n; \theta^j)} \,\middle|\, Y^n = y^n\right) \end{aligned}$$

and hence

$$\frac{\mathcal{L}(\theta^{j+1})}{\mathcal{L}(\theta^j)} = \log \frac{f(y^n; \theta^{j+1})}{f(y^n; \theta^j)}$$

$$
\begin{aligned}
&= J(\theta^{j+1}|\theta^j) - \mathbb{E}_{\theta^j}\left(\log \frac{f(Z^n|Y^n;\theta^{j+1})}{f(Z^n|Y^n;\theta^j)} \;\middle|\; Y^n = y^n\right) \\
&= J(\theta^{j+1}|\theta^j) + K(f_j, f_{j+1})
\end{aligned}
$$

where $f_j = f(y^n;\theta^j)$ and $f_{j+1} = f(y^n;\theta^{j+1})$ and $K(f,g) = \int f(x)\log(f(x)/g(x))\,dx$ is the Kullback-Leibler distance. Now, θ^{j+1} was chosen to maximize $J(\theta|\theta^j)$. Hence, $J(\theta^{j+1}|\theta^j) \geq J(\theta^j|\theta^j) = 0$. Also, by the properties of Kullback-Leibler divergence, $K(f_j, f_{j+1}) \geq 0$. Hence, $\mathcal{L}(\theta^{j+1}) \geq \mathcal{L}(\theta^j)$ as claimed.

9.50 Example (Continuation of Example 9.49)**.** Consider again the mixture of two normals but, for simplicity assume that $p = 1/2$, $\sigma_1 = \sigma_2 = 1$. The density is

$$f(y;\mu_1,\mu_2) = \frac{1}{2}\phi(y;\mu_0,1) + \frac{1}{2}\phi(y;\mu_1,1).$$

Directly maximizing the likelihood is hard. Introduce latent variables $Z_1,\ldots,Z_n$ where $Z_i = 0$ if Y_i is from $\phi(y;\mu_0,1)$, and $Z_i = 1$ if Y_i is from $\phi(y;\mu_1,1)$, $\mathbb{P}(Z_i = 1) = P(Z_i = 0) = 1/2$, $f(y_i|Z_i = 0) = \phi(y;\mu_0,1)$ and $f(y_i|Z_i = 1) = \phi(y;\mu_1,1)$. So $f(y) = \sum_{z=0}^{1} f(y,z)$ where we have dropped the parameters from the density to avoid notational overload. We can write

$$f(z,y) = f(z)f(y|z) = \frac{1}{2}\phi(y;\mu_0,1)^{1-z}\phi(y;\mu_1,1)^z.$$

Hence, the complete likelihood is

$$\prod_{i=1}^{n} \phi(y_i;\mu_0,1)^{1-z_i}\phi(y_i;\mu_1,1)^{z_i}.$$

The complete log-likelihood is then

$$\widetilde{\ell} = -\frac{1}{2}\sum_{i=1}^{n}(1-z_i)(y_i-\mu_0) - \frac{1}{2}\sum_{i=1}^{n} z_i(y_i-\mu_1).$$

And so

$$J(\theta|\theta^j) = -\frac{1}{2}\sum_{i=1}^{n}(1-\mathbb{E}(Z_i|y^n,\theta^j))(y_i-\mu_0) - \frac{1}{2}\sum_{i=1}^{n}\mathbb{E}(Z_i|y^n,\theta^j))(y_i-\mu_1).$$

Since Z_i is binary, $\mathbb{E}(Z_i|y^n,\theta^j) = \mathbb{P}(Z_i = 1|y^n,\theta^j)$ and, by Bayes' theorem,

$$
\begin{aligned}
\mathbb{P}(Z_i = 1|y^n,\theta^i) &= \frac{f(y^n|Z_i = 1;\theta^j)\mathbb{P}(Z_i = 1)}{f(y^n|Z_i = 1;\theta^j)\mathbb{P}(Z_i = 1) + f(y^n|Z_i = 0;\theta^j)\mathbb{P}(Z_i = 0)} \\
&= \frac{\phi(y_i;\mu_1^j,1)\frac{1}{2}}{\phi(y_i;\mu_1^j,1)\frac{1}{2} + \phi(y_i;\mu_0^j,1)\frac{1}{2}} \\
&= \frac{\phi(y_i;\mu_1^j,1)}{\phi(y_i;\mu_1^j,1) + \phi(y_i;\mu_0^j,1)} \\
&= \tau(i).
\end{aligned}
$$

Take the derivative of $J(\theta|\theta^j)$ with respect to μ_1 and μ_2, set them equal to 0 to get

$$\widehat{\mu}_1^{j+1} = \frac{\sum_{i=1}^n \tau_i y_i}{\sum_{i=1}^n \tau_i}$$

and

$$\widehat{\mu}_0^{j+1} = \frac{\sum_{i=1}^n (1-\tau_i) y_i}{\sum_{i=1}^n (1-\tau_i)}.$$

We then recompute τ_i using $\widehat{\mu}_1^{j+1}$ and $\widehat{\mu}_0^{j+1}$ and iterate. ■

9.14 Exercises

1. Let $X_1, \ldots, X_n \sim \text{Gamma}(\alpha, \beta)$. Find the method of moments estimator for α and β.

2. Let $X_1, \ldots, X_n \sim \text{Uniform}(a, b)$ where a and b are unknown parameters and $a < b$.

   (a) Find the method of moments estimators for a and b.

   (b) Find the MLE $\widehat{a}$ and $\widehat{b}$.

   (c) Let $\tau = \int x \, dF(x)$. Find the MLE of τ.

   (d) Let $\widehat{\tau}$ be the MLE of τ. Let $\widetilde{\tau}$ be the nonparametric plug-in estimator of $\tau = \int x \, dF(x)$. Suppose that $a = 1$, $b = 3$, and $n = 10$. Find the MSE of $\widehat{\tau}$ by simulation. Find the MSE of $\widetilde{\tau}$ analytically. Compare.

3. Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$. Let τ be the .95 percentile, i.e. $\mathbb{P}(X < \tau) = .95$.

   (a) Find the MLE of τ.

   (b) Find an expression for an approximate $1 - \alpha$ confidence interval for τ.

   (c) Suppose the data are:

```
3.23 -2.50  1.88 -0.68  4.43  0.17
1.03 -0.07 -0.01  0.76  1.76  3.18
0.33 -0.31  0.30 -0.61  1.52  5.43
1.54  2.28  0.42  2.33 -1.03  4.00
0.39
```

   Find the MLE $\widehat{\tau}$. Find the standard error using the delta method. Find the standard error using the parametric bootstrap.