Identifiability of Two-Component Normal Mixture Model

Eric Slud STAT 700, 9/12/2022

The two-component normal mixture model introduced in class has (*iid* observations from) the density

$$f(x) \;=\; \alpha\,\frac{1}{\sqrt{2\pi\sigma_1^2}}\,e^{-(x-\mu_1)^2/(2\sigma^2)} \;+\; (1-\alpha)\,\frac{1}{\sqrt{2\pi\sigma_2^2}}\,e^{-(x-\mu_2)^2/(2\sigma^2)}$$

where the 5-dimensional parameter $\theta = (\alpha, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2)$ lies in the parameter space

$$\Theta \;=\; \{\underline{t} \in (0,1)\times\mathbb{R}^2\times\mathbb{R}^2_+ : \quad \sigma_1^2 < \sigma_2^2 \quad \text{or} \quad \sigma_1^2 = \sigma_2^2,\ \mu_1 < \mu_2\}$$

To establish identifiability, we need to find a series of mathematical operations on $f(x)$ in terms of which the components of θ can be defined uniquely. The first is to re-express $f(x)$ as its Moment Generating Function

$$m(s) \;=\; \int e^{xs}\,f(x)\,dx \;=\; \alpha\,e^{\mu_1 s+\sigma_1^2 s^2/2} \;+\; (1-\alpha)\,e^{\mu_2 s+\sigma_2^2 s^2/2}$$

Then the logarithm of this function can be expressed as the sum of three terms

$$\log m(s) \;=\; \log(1-\alpha) \;+\; \mu_2\,s \;+\; \frac{1}{2}\,\sigma_2^2\,s^2 \;+\; \log\,(1 \;+\; \frac{\alpha}{1-\alpha}\,e^{(\mu_1-\mu_2)s\,+\,(\sigma_1^2-\sigma_2^2)s^2/2})$$

As $s\to\infty$, the sum of the first three terms is a quadratic function of s also going to ∞, while the fourth term converges for large s to 0 because of the condition that either $\sigma_1^2 < \sigma_2^2$ holds or both $\sigma_1^2 = \sigma_2^2$ and $\mu_1 < \mu_2$. From these properties of $\log m(s)$, we read off $\log(1-\alpha)$, μ_2, $\sigma^2/2$ as the three coefficients as the leading quadratic. With those parameter components known, we also find from

$$\int x\,f(x)\,dx \;=\; \alpha\,\mu_1 \;+\; (1-\alpha)\,\mu_2$$

that μ_1 also is uniquely determined. Finally, with $(\alpha,\ \mu_1, \mu_2,\ \sigma_2^2)$ known, we determine σ_1^2 uniquely from

$$m(1) \;=\; \alpha\,e^{\mu_1\,+\,\sigma_1^2/2} \;+\; (1-\alpha)\,e^{\mu_2\,+\,\sigma_2^2/2}$$

This establishes identifiability. We will see a little later in the course that the parameters could be estimated from real datasets of *iid* observations from either the Maximum Likelihood method or (more easily) from the EM Algorithm.