Gaussian Mixture Model

Nitay Alon

May 30, 2018

We are testing to see if we can use the given sample size (1120) to perform a likelihood ratio test.
We assume the following mixture of models:

- Men:
$$f_X(x) = p * \mathcal{N}(\theta_{mas}, 1) + (1-p) * \mathcal{N}(\theta_{fem}, 1)$$

- Women:
$$q * \mathcal{N}(\theta_{mas}, 1) + (1-q) * \mathcal{N}(\theta_{fem}, 1)$$

We assume that: $\theta_{fem} = -\theta_{mas}$ and also $p \approx 1, q \approx 0$. Denoting x_i as the i^{th} observation from the men group and y_j as the j^{th} observation from the women group and using EM algorithm for GMM we can estimate the parameters as follows:

$$M_i = P(z_i = M|x_i) = \frac{p * \mathcal{N}(\theta_{mas}, 1)}{p * \mathcal{N}(\theta_{mas}, 1) + (1-p) * \mathcal{N}(\theta_{fem}, \sigma)} = \big(1 + \frac{1-p}{p} exp^{-\frac{(x_i - (\frac{\theta_{mas}+\theta_{fem})}{2}(\theta_{mas}-\theta_{fem})}{2\sigma^2}}\big)^{-1} \quad (1)$$

The same formula can be applied to the women data:

$$M_j = P(z_i = M|y_j) = \frac{q * \mathcal{N}(\theta_{mas}, 1)}{q * \mathcal{N}(\theta_{mas}, 1) + (1-q) * \mathcal{N}(\theta_{fem}, \sigma)} = \big(1 + \frac{1-q}{q} exp^{-\frac{(y_j - (\frac{\theta_{mas}+\theta_{fem})}{2}(\theta_{mas}-\theta_{fem})}{2\sigma^2}}\big)^{-1} \quad (2)$$

And the parameters:

$$A = \frac{1}{\sum_{i=1}^{m} M_i} \sum_{i=1}^{m} M_i x_i \tag{3}$$

$$B = \frac{1}{\sum_{j=1}^{n} M_j} \sum_{j=1}^{n} M_j y_j \tag{4}$$

$$p = \frac{\sum_{i=1}^{m} M_i}{m} \tag{5}$$

$$q = \frac{\sum_{j=1}^{n} M_j}{n} \tag{6}$$

$$\tag{7}$$

Denoting:

$$m_{men} = \bar{x} \tag{8}$$

$$mm_{men} = \bar{x^2} \tag{9}$$

$$m_{women} = \bar{y} \tag{10}$$

$$mm_{women} = \bar{y^2} \tag{11}$$

we can estimate the parameters:

$$\theta_{mas} = \frac{A+B}{p+q} \tag{12}$$

$$\theta_{fem} = \frac{m1 + m2 - A - B}{2 - p - q} \tag{13}$$

$$\sigma^2 = \frac{mm_{mas} + mm_{fem}}{2} - \theta_{fem}^2 + \big(\frac{p+q}{2}\big)(\theta_{fem}^2 - \theta_{mas}^2) \tag{14}$$

1 Log likelihood ratio

Using the EM results we can compute the log likelihood of the mixture model and the log likelihood of the null model

$$H_0 : p = q$$
$$H_0 : p \neq q$$

By computing the delta between the llk we can learn about the significance of the delta (recall that $\lambda \sim exp(\theta)$).

1.1 30/5/2018 - Update

After some not satisfying results, we're going over the EM equations again. This time we begin with a constrained version of the problem:

$$p = \frac{\theta_2 + \mu}{\theta_2 + \theta_1} \tag{15}$$

$$q = \frac{\theta_2 - \mu}{\theta_2 + \theta_1} \tag{16}$$

$$\sigma^2 = 1 + \theta_1\theta_2 \tag{17}$$