5.1 Solution

Antonio Moretti

April 2017

1. Consider the Multinomial PDF:

$$f(x,m,k) = \frac{m!}{x_1! \cdots x_n!} \mu_1^{x_1} \cdots \mu_n^{x_n} \tag{1}$$

$$= \frac{\Gamma(\sum_i x_i + 1)}{\prod_i \Gamma(x_i + 1)} \prod_{i=1}^{n} \mu_i^{x_i} \tag{2}$$

Let $\mu_j^{x_n} = \prod_{l=1}^{M} \mu_j(l)^{x_n(l)}$. The E-step is written as follows.

$$\tau_{n,j} = p(z_n = j | x_n, \theta) \tag{3}$$

$$= \frac{\pi_j p(x_n|\mu_j)}{\sum_{i=1}^{K} \pi_i p(x_n|\mu_i)} \tag{4}$$

$$= \frac{\pi_j \mu_j^{x_n}}{\sum_{i=1}^{K} \pi_i \mu_i^{x_n}} \tag{5}$$

In the M-step we seek to maximize the parameters holding $\tau_{n,j}$ fixed:

$$\theta = \arg\max_{\theta} \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{n,j} \times log\left(\frac{p(x_n, z = j|\theta)}{\tau_{n,j}} \right) \tag{6}$$

Note that $p(x_n, z = j|\theta) = p(x_n|z = j, \theta) p(z_n = j|\theta) = \pi_j \mu_j^{x_n}$ by the probability chain rule.

$$\arg\max_{\mu_1, \cdots, \mu_k, \pi_1, \cdots, \pi_k} \sum_{n=1}^{N} \sum_{j=1}^{K} \tau_{n,j} \Big(log(\mu_j^{x_n}) + log(\pi_j) \Big) \tag{7}$$

$$\text{subject to} \quad \sum_{l=1}^{M} \mu_j(l) = 1 \qquad \forall j \in \{1, \cdots, K\} \tag{8}$$

$$\sum_{i=1}^{K} \pi_i = 1 \tag{9}$$

Form the Lagrangian:

$$\mathcal{L}(\mu, \pi, \alpha, \beta) = \sum_{n=1}^{N}\sum_{j=1}^{K} \tau_{n,j}\Big(log(\mu_j^{x_n}) + log(\pi_j)\Big) - \alpha\Big(\sum_{j=1}^{K}\pi_j - 1\Big) - \sum_{j=1}^{K}\beta_j\Big(\sum_{l=1}^{M}\mu_j(l) - 1\Big) \tag{10}$$

Differentiate with respect to the parameters of interest starting with π_j:

$$\frac{\partial \mathcal{L}}{\partial \pi_j} = \sum_{n=1}^{N}\frac{\tau_{n,j}}{\pi_j} - \alpha = 0 \tag{11}$$

$$\pi_j = \frac{\sum\limits_{n=1}^{N}\tau_{n,j}}{\alpha} \tag{12}$$

Plug into the primal constraint:

$$= \frac{\sum\limits_{n=1}^{N}\tau_{n,j}}{\sum\limits_{i=1}^{K}\sum\limits_{n=1}^{N}\tau_{n,i}} \tag{13}$$

Not surprisingly we find the MLE of the mixing components $\hat{\pi}_j$ is the sample average of $\tau_{n,j}$

$$\pi_j = \frac{1}{N}\sum_{n=1}^{N}\tau_{n,j} \tag{14}$$

Repeating this process for $\mu_j(l)$:

$$\frac{\partial \mathcal{L}}{\partial \mu_j(l)} = \sum_{n=1}^{N}\tau_{n,j}x_n(l)\frac{1}{\mu_j(l)} - \beta_j = 0 \tag{15}$$

$$\mu_j(l) = \frac{1}{\beta_j}\sum_{n=1}^{N}\tau_{n,j}x_n(l) \tag{16}$$

Plug into the constraint on μ:

$$\mu_j(l) = \sum_{n=1}^{N}\frac{\tau_{n,j}}{\sum\limits_{n'=1}^{N}\tau_{n',j}}x_n(l) \tag{17}$$

2. The marginal distribution of $x_n(\Delta)$ for each word Δ is Binomial. By the Poisson limit theorem as $l_n \to \infty$ and $x_n/l_n \to 0$:

$$\binom{n}{k}p^k(1-p)^{n-k} \approx e^{-\lambda}\frac{\lambda^k}{k!} \tag{18}$$

3. PCA assumes that the data is generated by a multivariate Gaussian distribution parametrized by its first two moments μ and $\mathbf{\Sigma}$. The spectral decomposition gives a projection or linear transformation $\mathbf{\Lambda}$ that maximizes the variance of a linear combination of the original variables.

$$\mathbf{\Sigma} = \mathbf{\Lambda} \times \mathbf{D} \times \mathbf{\Lambda}^T \tag{19}$$

Rather than seeking vectors that are linear combinatons of the original variables, we are asked to find variables that can best approximate the document vectors: $\hat{x}_n = \sum_{\delta=1}^{d} u_n(\delta) v_\delta$ where both $u_n(\delta)$ and v_δ are hidden. This bears more similarity to factor analysis than principal component analysis. Also note that for the Poisson distribution whose first and second centered moment are λ, a projection that maximizes the variance of a linear combination of variables cannot be obtained via a spectral decomposition of the covariance matrix and the EM algorithm would also be needed.

4. We are given that the marginal distribution $\hat{x}_n(\delta) \sim Bin(l_n, \frac{\hat{x}_n}{l_n})$ which we shall approximate with a Poisson distribution.

$$\hat{x}_n = \sum_{\delta=1}^{d} u_n(\delta) v_\delta \tag{20}$$

Denote the Poisson parameter $\lambda = \sum\limits_{\delta=1}^{d} u_n(\delta) v_\delta$

$$P(x_n(\delta)|u_n, v) = \frac{\lambda^{x_n(\delta)} e^{-\lambda}}{x_n(\delta)!} \tag{21}$$

Writing the expected complete log likelihood and taking sample averages as maximizers of the sufficient statistics yields the following.

E step:

$$\tau_{n,i} = \frac{\pi_i \sum\limits_{\delta=1}^{d} \left(\frac{u_n(\delta) v_\delta}{l_n} \right)}{\sum\limits_j \pi_j \sum\limits_{\delta=1}^{d} \frac{u_n(\delta) v_\delta}{l_n}} \tag{22}$$

M step:

$$\pi_i = \frac{\sum\limits_{n=1}^{N} \tau_{n,i}}{N} \tag{23}$$

$$u_n(\delta) = \frac{\sum\limits_{n=1}^{N} v_\delta \tau_{n,i}}{\sum\limits_{\delta'=1}^{d} \sum\limits_{n=1}^{N} v_{\delta'} \tau_{n,i}} \tag{24}$$