Homework 5

JD Gallego Posada, *University of Amsterdam* 19/10/2016

1 PCA

(a) Provide an expression for $\hat{\mathbf{x}}_n$.

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \quad \rightarrow \quad \hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}} \quad \rightarrow \quad \hat{\mathbf{X}} = [\hat{\mathbf{x}}_1 \cdots \hat{\mathbf{x}}_N]$$

(b) Prove that the average of $\hat{\mathbf{x}}_n$ is the $\mathbf{0}$ vector.

$$\frac{1}{N}\sum_{n=1}^{N}\hat{\mathbf{x}}_n = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n - \frac{1}{N}\sum_{n=1}^{N}\bar{\mathbf{x}} = \bar{\mathbf{x}} - \frac{N}{N}\bar{\mathbf{x}} = \mathbf{0}$$

(c) Provide an expression for $\mathbf{S}$ in terms of $\hat{\mathbf{X}}$.

$$\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}\hat{\mathbf{x}}_n\hat{\mathbf{x}}_n^T = \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T$$

(d) What is the dimensionality of $\mathbf{S}$?

Since $\mathbf{S}$ is the product of a $D \times N$ and a $N \times D$ matrix, it has dimensionality $D \times D$.

(e) What is the expression for the linear projection $\mathbf{L}$ that maps data vectors $\hat{\mathbf{x}}_n$ onto a K-dimensional sub-space, $y_n = \mathbf{L}\hat{\mathbf{x}}_n$, such that it has zero mean and identity covariance. Prove that the average over N of y_n is $\mathbf{0}$. Prove that the covariance of y_n is the identity. What is this operation called?

Note that $\mathbf{S}$ is symmetric and (semi) positive definite. Consider the eigenvector equation $\mathbf{SU} = \mathbf{U\Lambda}$. This equation has solution $\mathbf{U} = [\mathbf{u}_1 \cdots \mathbf{u}_D]$ and $\mathbf{\Lambda} = \text{diag}([\lambda_1 \cdots \lambda_D])$, where $\mathbf{u}_i$ is the i-th orthonormal eigenvector of $\mathbf{S}$, corresponding to the eigenvalue λ_i. Additionally, $\{\mathbf{u}_i\}$ are an orthonormal basis for $\mathbb{R}^D$ and $\mathbf{U}$ is orthogonal. Therefore, $\mathbf{U}^{-1} = \mathbf{U}^T \rightarrow \mathbf{\Lambda} = \mathbf{U}^T\mathbf{SU}$.

$$\mathbf{L} = \mathbf{\Lambda}_K^{-1/2}\mathbf{U}_K^T$$

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n = \frac{1}{N}\sum_{n=1}^{N}\mathbf{L}\hat{\mathbf{x}}_n = \frac{1}{N}\mathbf{L}\sum_{n=1}^{N}\hat{\mathbf{x}}_n = \frac{1}{N}\mathbf{L0} = \mathbf{0}$$

$$\frac{1}{N}\sum_{n=1}^{N}\mathbf{y}_n\mathbf{y}_n^T = \mathbf{\Lambda}_K^{-1/2}\mathbf{U}_K^T\left(\frac{1}{N}\sum_{n=1}^{N}\hat{\mathbf{x}}_n\hat{\mathbf{x}}_n^T\right)\mathbf{U}_K\mathbf{\Lambda}_K^{-1/2} = \mathbf{\Lambda}_K^{-1/2}\mathbf{U}_K^T\mathbf{S}\mathbf{U}_K\mathbf{\Lambda}_K^{-1/2} = \mathbf{\Lambda}_K^{-1/2}\mathbf{\Lambda}_K\mathbf{\Lambda}_K^{-1/2} = \mathbf{I}$$

where $\mathbf{\Lambda}_K$ and $\mathbf{U}_K$ represent the matrices associated with the K chosen eigenvalues and eigenvectors. This operation is called *whitening* or *sphering*.

2 Mixture Models

(a) Write down the likelihood for the data set in terms of $\{x_1, x_2, \ldots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.

$$\mathscr{L}(\{x_n\}|\boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{n=1}^{N} p(x_n|\boldsymbol{\pi}, \boldsymbol{\lambda}) = \prod_{n=1}^{N}\left(\sum_{k=1}^{K}\pi_k\frac{\lambda_k^{x_n}}{x_n!}e^{-\lambda_k}\right)$$

(b) Write down the log-likelihood for the data set in terms of $\{x_1, x_2, \dots, x_N\}$, $\{\pi_k\}$, $\{\lambda_k\}$.

$$\log \mathscr{L}(\{x_n\}|\boldsymbol{\pi}, \boldsymbol{\lambda}) = \sum_{n=1}^{N} \log \left(\sum_{k=1}^{K} \pi_k \frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k} \right)$$

(c) Find the expression for the responsibilities r_{nk}.

$$r_{nk} = p(z_k = 1|x_n) = \frac{p(z_k = 1)p(x_n|z_k = 1)}{\sum_{j=1}^{K} p(z_j = 1)p(x_n|z_j = 1)} = \frac{\pi_k p(x_n|\lambda_k)}{\sum_{j=1}^{K} \pi_j p(x_n|\lambda_j)} = \frac{\pi_k \frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k}}{\sum_{j=1}^{K} \pi_j \frac{\lambda_j^{x_n}}{x_n!} e^{-\lambda_j}}$$

where $\{z_k\}$ represents the set of latent variables.

(d) Find the expression for λ_k *that maximizes the log-likelihood.*

$$\begin{aligned}
\frac{\partial \log \mathscr{L}}{\partial \lambda_k} &= \sum_{n=1}^{N} \frac{1}{\sum_{j=1}^{K} \pi_j \frac{\lambda_j^{x_n}}{x_n!} e^{-\lambda_j}} \frac{\pi_k}{x_n!} \left[x_n \lambda_k^{x_n - 1} e^{-\lambda_k} - \lambda_k^{x_n} e^{-\lambda_k} \right] \\
&= \sum_{n=1}^{N} \frac{\pi_k \frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k}}{\sum_{j=1}^{K} \pi_j \frac{\lambda_j^{x_n}}{x_n!} e^{-\lambda_j}} \left[\frac{x_n}{\lambda_k} - 1 \right] \\
&= \sum_{n=1}^{N} r_{nk} \left[\frac{x_n}{\lambda_k} - 1 \right] = 0
\end{aligned}$$

$$\frac{1}{\lambda_k} \sum_{n=1}^{N} r_{nk} x_n - \sum_{n=1}^{N} r_{nk} = 0 \quad \rightarrow \quad \lambda_k^* = \frac{1}{N_k} \sum_{n=1}^{N} r_{nk} x_n$$

where $N_k = \sum_{n=1}^{N} r_{nk}$.

(e) Find the expression for π_k *that maximizes the log-likelihood.*

Define $\tilde{\mathscr{L}} = \log \mathscr{L} + \theta(\sum_{k=1}^{K} \pi_k - 1)$, where θ is the Lagrange multiplier associated to the constraint on the vector $\boldsymbol{\pi}$.

$$\begin{aligned}
\frac{\partial \tilde{\mathscr{L}}}{\partial \pi_k} &= \theta + \sum_{n=1}^{N} \frac{\frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k}}{\sum_{j=1}^{K} \pi_j \frac{\lambda_j^{x_n}}{x_n!} e^{-\lambda_j}} \\
&= \theta + \frac{1}{\pi_k} \sum_{n=1}^{N} r_{nk} = 0 \quad \rightarrow \quad \pi_k^* = \frac{N_k}{-\theta^*}
\end{aligned}$$

$$\sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} + \theta \sum_{k=1}^{K} \pi_k = 0 \quad \rightarrow \quad -\theta^* = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} = N \quad \rightarrow \quad \pi_k^* = \frac{N_k}{N}$$

where $N_k = \sum_{n=1}^{N} r_{nk}$.

(f) Write down the log-joint distribution $\log p(\mathbf{x}_1, \dots, \mathbf{x}_N, \{\pi_k\}, \{\lambda_k\}|a, b, \alpha, K)$.

Let $\boldsymbol{\alpha} = (\alpha/K, \dots, \alpha/K)$. Recall that:

- $\log \mathscr{G}(\lambda_k|a, b) = a \log b - \log \Gamma(a) + (a - 1) \log \lambda_k - b\lambda_k$
- $\frac{\partial \log \mathscr{G}(\lambda_k|a,b)}{\partial \lambda_k} = \frac{a-1}{\lambda_k} - b$
- $\log \mathscr{D}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \log C(\boldsymbol{\alpha}) + \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k$
- $\frac{\partial \log \mathscr{D}(\boldsymbol{\pi}|\boldsymbol{\alpha})}{\partial \pi_k} = \frac{\alpha_k - 1}{\pi_k}$

Applying the product rule $p(x, y) = p(x|y)p(y)$, we have:

$$\mathscr{L}(\{x_n\}, \boldsymbol{\pi}, \boldsymbol{\lambda}|a, b, \alpha, K) = \left(\prod_{n=1}^{N} p(x_n|\boldsymbol{\pi}, \boldsymbol{\lambda})\right)\left(\prod_{k=1}^{K} p(\lambda_k|a, b)\right) p(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

$$\log \mathscr{L} = \sum_{n=1}^{N} \log p(x_n|\boldsymbol{\pi}, \boldsymbol{\lambda}) + \sum_{k=1}^{K} \log \mathscr{G}(\lambda_k|a, b) + \log \mathscr{D}(\boldsymbol{\pi}|\boldsymbol{\alpha})$$

$$\log \mathscr{L} = \left(\sum_{n=1}^{N} \log \sum_{k=1}^{K} \pi_k \frac{\lambda_k^{x_n}}{x_n!} e^{-\lambda_k}\right) + \left(\sum_{k=1}^{K} \log \mathscr{G}(\lambda_k|a, b)\right) + \log C(\boldsymbol{\alpha}) + \sum_{k=1}^{K} (\alpha_k - 1) \log \pi_k$$

(g) Find the expression for λ_k that maximizes the log-joint.

$$\frac{\partial \log \mathscr{L}}{\partial \lambda_k} = \frac{1}{\lambda_k} \sum_{n=1}^{N} r_{nk} x_n - \sum_{n=1}^{N} r_{nk} + \frac{a-1}{\lambda_k} - b = 0$$

$$\frac{1}{\lambda_k}\left(\sum_{n=1}^{N} r_{nk} x_n + a - 1\right) = N_k + b \quad \rightarrow \quad \lambda_k^* = \frac{\sum_{n=1}^{N} r_{nk} x_n + a - 1}{N_k + b}$$

where $N_k = \sum_{n=1}^{N} r_{nk}$.

(h) Find the expression for π_k that maximizes the log-joint.

Define $\tilde{\mathscr{L}} = \log \mathscr{L} + \theta(\sum_{k=1}^{K} \pi_k - 1)$, where θ is the Lagrange multiplier associated to the constraint on the vector $\boldsymbol{\pi}$.

$$\frac{\partial \tilde{\mathscr{L}}}{\partial \pi_k} = \theta + \frac{1}{\pi_k} \sum_{n=1}^{N} r_{nk} + \frac{\frac{\alpha}{K} - 1}{\pi_k} = 0$$

$$\Rightarrow \pi_k \theta + \sum_{n=1}^{N} r_{nk} + \frac{\alpha}{K} - 1 = 0 \quad \rightarrow \quad \pi_k^* = \frac{N_k + \frac{\alpha}{K} - 1}{-\theta^*}$$

$$\sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} + \theta \sum_{k=1}^{K} \pi_k + \sum_{k=1}^{K} \left(\frac{\alpha}{K} - 1\right) = 0 \quad \rightarrow \quad -\theta^* = \sum_{k=1}^{K} \sum_{n=1}^{N} r_{nk} + \alpha - K = N + \alpha - K$$

$$\pi_k^* = \frac{N_k + \frac{\alpha}{K} - 1}{N + \alpha - K}$$

where $N_k = \sum_{n=1}^{N} r_{nk}$.

(i) Write down an iterative algorithm using the above update equations; include initialization and convergence check steps.

Algorithm 1: EM Algorithm for Mixture of Poisson Distributions - Bayesian Approach

Initialize $\pi_k^{(0)}$ (s.t $\sum_k \pi_k^{(0)} = 1$), $\lambda_k^{(0)}$, $N_k^{(0)}$ and compute $\log \mathscr{L}(\{x_n\}, \boldsymbol{\pi}^{(0)}, \boldsymbol{\lambda}^{(0)}|a, b, \alpha, K)$;

E-step: Put $r_{nk}^{(\tau)} := \frac{\pi_k^{(\tau)} \frac{\lambda_k^{(\tau)x_n}}{x_n!} e^{-\lambda_k^{(\tau)}}}{\sum_{j=1}^{K} \pi_j^{(\tau)} \frac{\lambda_j^{(\tau)x_n}}{x_n!} e^{-\lambda_j^{(\tau)}}}$;

M-step: Put $\lambda_k^{(\tau+1)} := \frac{\sum_{n=1}^{N} r_{nk}^{(\tau)} x_n + a - 1}{N_k^{(\tau)} + b}$, $\pi_k^{(\tau+1)} := \frac{N_k^{(\tau)} + \frac{\alpha}{K} - 1}{N + \alpha - K}$ and $N_k^{(\tau+1)} := \sum_{n=1}^{N} r_{nk}^{(\tau)}$;

Compute $\log \mathscr{L}(\{x_n\}, \boldsymbol{\pi}^{(\tau+1)}, \boldsymbol{\lambda}^{(\tau+1)}|a, b, \alpha, K)$ and check for convergence (e.g. absolute change in $\log \mathscr{L}$ less than threshold). If not, **repeat** EM-steps for $\tau := \tau + 1$;

return $\boldsymbol{\pi}^{(\tau+1)}$, $\boldsymbol{\lambda}^{(\tau+1)}$, $r_{nk}^{(\tau+1)}$

This algorithm converges to a local minimum of the log-likelihood. Use estimated $\hat{r}_{nk}$ as a soft assignment of individual n to class k.

NB: For a Maximum Likelihood approach, set $\alpha = K$, $a = 1$, $b = 0$.