Assignment 5
Machine Learning 1, Fall 2016

Dana Kianfar
University of Amsterdam

January 6, 2018

1 PCA

(a)

$$\bar{\mathbf{x}} = \frac{1}{N}\sum_{n}^{N} x_n$$

$$\hat{\mathbf{x}}_n = \mathbf{x}_n - \bar{\mathbf{x}}$$

(b)

$$\begin{aligned}\frac{1}{N}\sum_{n}^{N}\hat{\mathbf{x}}_n &= \frac{1}{N}\sum_{n}^{N}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\\ &= \frac{1}{N}\sum_{n}^{N}\mathbf{x}_n - \frac{1}{N}\left(N\bar{\mathbf{x}}\right)\\ &= \bar{\mathbf{x}} - \bar{\mathbf{x}} = 0\end{aligned}$$

(c)

$$\begin{aligned}\mathbf{S} = \mathbf{var}(\hat{\mathbf{X}}) &= \frac{1}{N}\sum_{n}^{N}\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)\left(\mathbf{x}_n - \bar{\mathbf{x}}\right)^T\\ &= \frac{1}{N}\sum_{n}^{N}\left(\hat{\mathbf{x}}_n\right)\left(\hat{\mathbf{x}}_n\right)^T\\ &= \frac{1}{N}\hat{\mathbf{X}}\hat{\mathbf{X}}^T\end{aligned}$$

(d)

$$\mathbf{S} \in \mathbb{R}^{D\times D}$$

(e) We reproduce a K-dimensional representation of our matrix $\mathbf{X}$ with zero mean and unit covariance. In the process, we remove linear correlations in our data vectors x_n. This process is called whitening/sphering. Therefore we have: $\mathbf{Z} = \mathbf{U}^T\hat{\mathbf{X}}$.

The projection of $\hat{\mathbf{X}}$ on its eigenvectors $\mathbf{U}_K$ cause the de-correlation as all eigenvectors are orthogonal to each other. Therefore, the covariance matrix of Z is a diagonal matrix Λ_K with eigenvalues λ_k on its diagonals for $\forall k = 1, \cdots, K$.

$$\mathbf{L} = \Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T$$

$$\mathbf{y}_n = \Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_n$$

For the mean of $\mathbf{y}_n$ we have the following.

$$\frac{1}{N}\sum_n^N \mathbf{y}_n = \frac{1}{N}\sum_n^N \mathbf{L}\hat{x}_n = N\mathbf{L}\left(\cancelto{0}{\frac{1}{N}\sum_n^N \hat{x}_n}\right) = 0$$

For the variance of $\mathbf{y}_n$ we have the following. We know that the transpose of the symmetric matrix $\Lambda_K^{\frac{-1}{2}}$ is the same matrix.

$$\begin{aligned}
\mathbf{var}(\mathbf{y}_n) &= \frac{1}{N}\sum_n^N \left(\mathbf{y}_n - \bar{\mathbf{y}}\right)\left(\mathbf{y}_n - \bar{\mathbf{y}}\right)^T \\
&= \frac{1}{N}\sum_n^N \left(\Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_n - \cancelto{0}{\frac{1}{N}\sum_i^N \Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_i}\right)\left(\Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_n - \cancelto{0}{\frac{1}{N}\sum_i^N \Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_i}\right)^T \\
&= \frac{1}{N}\sum_n^N \left(\Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_n\right)\left(\Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_n\right)^T \\
&= \frac{1}{N}\sum_n^N \left(\Lambda_K^{\frac{-1}{2}} \mathbf{U}_K^T \hat{x}_n \cdot \hat{x}_n^T \mathbf{U}_K \Lambda_K^{\frac{-1}{2}}\right) \\
&= \frac{1}{N}\left(\Lambda_K^{\frac{-1}{2}} \cdot \mathbf{U}_K^T \cdot \hat{\mathbf{X}} \cdot \hat{\mathbf{X}}^T \cdot \mathbf{U}_K \cdot \Lambda_K^{\frac{-1}{2}}\right) \\
&= \Lambda_K^{\frac{-1}{2}} \cdot \mathbf{U}_K^T \cdot \mathbf{S} \cdot \mathbf{U}_K \cdot \Lambda_K^{\frac{-1}{2}}
\end{aligned}$$

We use the eigen-decomposition of covariance matrix S $\quad \mathbf{U}_K^T\mathbf{S} = \mathbf{U}_K^T\Lambda$

$$= \Lambda_K^{\frac{-1}{2}} \cdot \Lambda_K \cdot \Lambda_K^{\frac{-1}{2}} = \mathbb{I}_{\mathbb{K}}$$

2 Mixture Models

(a)

$$\begin{aligned}
p(\mathbf{X}|\pi_k, \lambda_k) &= \prod_n^N \pi_k p(x_n|\lambda_k) \\
&= \prod_n^N \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}
\end{aligned}$$

$$\begin{aligned}
p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\lambda}) &= \prod_n^N \sum_k^K \pi_k p(x_n|\lambda_k) \\
&= \prod_n^N \sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}
\end{aligned}$$

(b)

$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\lambda}) &= \log \prod_n^N \sum_k^K \pi_k p(x_n|\lambda_k)\\ &= \sum_n^N \log\left(\sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}\right)\end{aligned}$$

(c)

$$\begin{aligned}r_{nk} &= \frac{p(\lambda_k)p(x_n|\lambda_k)}{\sum_k p(\lambda_k)p(x_n|\lambda_k)}\\ &= \frac{\pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}}{\sum_j^K \pi_j \frac{\lambda_j^{x_n} e^{-\lambda_j}}{x_n!}}\end{aligned}$$

(d)

$$\begin{aligned}\log p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\lambda}) &= \sum_n^N \log\left(\sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}\right)\\ \frac{\partial \log p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\lambda})}{\partial \lambda_k} &= \sum_n^N \frac{\pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}}{\sum_j^K \pi_j \frac{\lambda_j^{x_n} e^{-\lambda_j}}{x_n!}} \cdot (\frac{x_n}{\lambda_k} - 1)\\ &= \sum_n^N r_{nk} \cdot (\frac{x_n}{\lambda_k} - 1)\\ \frac{\partial \log p(\mathbf{X}|\boldsymbol{\pi},\boldsymbol{\lambda})}{\partial \lambda_k} &= 0\\ \sum_n^N r_{nk} \cdot \frac{x_n}{\lambda_k} &= \sum_n^N r_{nk}\\ \sum_n^N r_{nk} \cdot x_n &= \sum_n^N r_{nk} \cdot \lambda_k\\ \sum_n^N r_{nk} &= N_k\\ \lambda_k^* &= \frac{1}{N_k} \sum_n^N r_{nk} \cdot x_n\end{aligned}$$

(e)

$$g(\boldsymbol{\pi}) = \sum_j \pi_j - 1$$

$$H = \log p(\mathbf{X}|\boldsymbol{\pi}, \boldsymbol{\lambda}) + \mu g(\boldsymbol{\pi})$$

$$= \sum_n^N \log \left(\sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!} \right) + \mu(\sum_j \pi_j - 1)$$

$$\frac{\partial H}{\partial \pi_k} = \sum_n \frac{\frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}}{\sum_j^K \pi_j \frac{\lambda_j^{x_n} e^{-\lambda_j}}{x_n!}} + \mu$$

$$= \sum_n \frac{r_{nk}}{\pi_k} + \mu$$

$$\frac{\partial H}{\partial \pi_k} = 0$$

$$\sum_n \frac{r_{nk}}{\pi_k} + \mu = 0$$

$$-\sum_n r_{nk} = \mu \pi_k$$

we sum over k

$$\mu = -N$$

$$\pi_k^* = -\frac{\sum_n r_{nk}}{\mu} = \frac{N_k}{N}$$

(f)

$$\alpha_k = \frac{\alpha}{K}, \boldsymbol{\alpha} = \{\alpha_k\}_{k=1}^K$$

$$p(\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\lambda}|a, b, \boldsymbol{\alpha}) = \prod_n^N \sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!} \cdot C(\boldsymbol{\alpha}) \prod_k \pi_k^{\alpha_k - 1} \cdot \prod_k \frac{b^a}{\Gamma(a)} \lambda_k^{a-1} e^{-b\lambda_k}$$

$$\log p(\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\lambda}|a, b, \boldsymbol{\alpha}) = \sum_n^N \log \sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!} + \log C(\boldsymbol{\alpha}) + \sum_k ((\alpha_k - 1)\pi_k)$$

$$+ \sum_k (a \log b - \log \Gamma(a) + (a-1)\lambda_k - b\lambda_k)$$

(g)

$$\frac{\partial \log p(\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\lambda}|a, b, \boldsymbol{\alpha})}{\partial \lambda_k} = \sum_n r_{nk} \cdot (\frac{x_n}{\lambda_k} - 1) + \frac{a-1}{\lambda_k} - b = 0$$

$$\sum_n r_{nk} \cdot \frac{x_n}{\lambda_k} + \frac{a-1}{\lambda_k} = \sum_n r_{nk} + b$$

$$\sum_n r_{nk} \cdot x_n + a - 1 = \lambda_k (N_k + b)$$

$$\lambda_k^* = \frac{\sum_n r_{nk} \cdot x_n + a - 1}{N_k + b}$$

(h)

$$g(\boldsymbol{\pi}) = \sum_j \pi_j - 1$$

$$H = \log p(\mathbf{X}, \boldsymbol{\pi}, \boldsymbol{\lambda}|a, b, \boldsymbol{\alpha}) + \mu g(\boldsymbol{\pi})$$

$$= \sum_n^N \log \sum_k^K \pi_k \frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!} + \log C(\boldsymbol{\alpha}) + \sum_k (\log(\alpha_k - 1)\pi_k)$$

$$+ \sum_k (a \log b - \log \Gamma(a) + (a-1)\lambda_k - b\lambda_k) + \mu(\sum_j \pi_j - 1)$$

$$\frac{\partial H}{\partial \pi_k} = \sum_n \frac{\frac{\lambda_k^{x_n} e^{-\lambda_k}}{x_n!}}{\sum_j^K \pi_j \frac{\lambda_j^{x_n} e^{-\lambda_j}}{x_n!}} + \frac{\alpha_k - 1}{\pi_k} + \mu$$

$$= \sum_n \frac{r_{nk}}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} + \mu$$

$$\frac{\partial H}{\partial \pi_k} = 0$$

$$\sum_n \frac{r_{nk}}{\pi_k} + \frac{\alpha_k - 1}{\pi_k} + \mu = 0$$

we multiply by π_k

$$N_k + \alpha_k - 1 + \mu\pi_k = 0$$

$$N_k + \alpha_k - 1 = -\mu\pi_k$$

we sum over k

$$-(N + \alpha - K) = \mu$$

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

(i) (a) Initialize variables for all classes in K and compute log likelihood.

$$\textbf{init} \;\; \pi_k^0, \lambda_k^0, N_k^0, \tau = 0 \;\; \textbf{compute} \;\; \log p(\mathbf{X}|\boldsymbol{\pi}^0, \boldsymbol{\lambda}^0)$$

(b) Repeat c, d until convergence

(c) E-step: compute responsibilities r_{nk} for all data points $\forall n = 1, \cdots, N$ and classes $\forall k = 1, \cdots, K$.

$$r_{nk}^\tau = \frac{p(\lambda_k^\tau)p(x_n|\lambda_k^\tau)}{\sum_k p(\lambda_k^\tau)p(x_n|\lambda_k^\tau)} = \frac{\pi_k^\tau \frac{(\lambda_k^{x_n})^\tau e^{-\lambda_k^\tau}}{x_n!}}{\sum_j^K \pi_j^\tau \frac{(\lambda_j^{x_n})^\tau e^{-\lambda_j^\tau}}{x_n!}}$$

(d) M-step: Use r_{nk}^τ to compute maximum likelihood estimators for all $\forall k = 1, \cdots, K$ and compute log-likelihood. Terminate if the change in log-likelihood is below an arbitrarily small threshold ϵ. A similar procedure applies for MAP estimates.

$$N_k^\tau = \sum_n r_{nk}^\tau$$

$$\lambda_k^\tau = \frac{1}{N_k^\tau} \sum_n^N r_{nk}^\tau \cdot x_n$$

$$\pi_k^\tau = \frac{N_k^\tau}{N}$$

$$\tau = \tau + 1$$

$$\text{Convergence criterion: } |\log p(\mathbf{X}|\boldsymbol{\pi}^\tau, \boldsymbol{\lambda}^\tau) - \log p(\mathbf{X}|\boldsymbol{\pi}^{\tau-1}, \boldsymbol{\lambda}^{\tau-1})| \leq \epsilon$$