2413, Machine Learning, Tutorial 10
Universität Bern

Simon Jenni (jenni@inf.unibe.ch)

PCA

PCA is a dimensionality reduction algorithm where each data point is projected to the first k eigenvectors of their covariance matrix. It is important that the data is first normalised, as follows:

- Let $\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$
- Replace each $x^{(i)}$ with $x^{(i)} - \mu$
- Let $\sigma_j^2 = \frac{1}{m}\sum_i (x_j^{(i)})^2$
- Replace each $x_j^{(i)}$ with $x_j^{(i)}/\sigma_j$

If you have prior knowledge that the input data has zero mean, you can skip the first two steps. If you have prior knowledge that the dimension of the data coordinates have the same scale, you can skip the third and fourth step.

1. There was a beer drinking competition at the Octoberfest. There were two rounds, one on Saturday and one on Sunday. The participants needed to drink as many mugs of beer as possible in 30 minutes. The data below shows their results. The first row shows the amount they drank on Saturday, and the second show their result on Sunday.

	$x^{(1)}$	$x^{(2)}$	$x^{(3)}$	$x^{(4)}$
$mugs(Saturday)$	3	2	5	4
$mugs(Sunday)$	1	2	3	4

   - Find the first principal component.

     Solution. Because the data coordinates have the same meaning (number of mugs), we do not need to perform the normalization (3^{rd} and 4^{th}) steps. First we extract the mean,

$$\mu = \frac{1}{m}\sum_{i=1}^{m} x^{(i)} = \frac{1}{4}(\begin{pmatrix}1\\3\end{pmatrix} + \begin{pmatrix}2\\2\end{pmatrix} + \begin{pmatrix}5\\3\end{pmatrix} + \begin{pmatrix}4\\4\end{pmatrix}) = \begin{pmatrix}3.5\\2.5\end{pmatrix}, \quad (1)$$

$$X_c = X - \mu \cdot \mathbf{1}^T = \begin{pmatrix} -0.5 & -1.5 & 1.5 & 0.5 \\ -1.5 & -0.5 & 0.5 & 1.5 \end{pmatrix}, \tag{2}$$

where X_c denotes the centred data. Then we need to compute the covariance matrix, which can be done with the following formula,

$$\Sigma = \frac{1}{m} X_c \cdot X_c^T = \begin{pmatrix} 5/4 & 3/4 \\ 3/4 & 5/4 \end{pmatrix}. \tag{3}$$

To compute the eigenvalues, we need to set the characteristic polynomial to 0.

$$det(\Sigma - \lambda I) = p(\lambda) = (5/4 - \lambda)(5/4 - \lambda) - 9/16 = 0, \tag{4}$$

which leads to the eigenvalues $\lambda_1 = 2$ and $\lambda_2 = 0.5$. The first principal component is simply the eigenvector $\mathbf{v}_1$ corresponding to the largest eigenvalue. This can computed from the following equation,

$$(\Sigma - \lambda_1 I)\mathbf{v}_1 = \begin{pmatrix} -3/4 & 3/4 \\ 3/4 & -3/4 \end{pmatrix} \begin{pmatrix} v_{11} \\ v_{21} \end{pmatrix} = 0, \tag{5}$$

which yields $v_{11} = v_{21}$. Using the convention of $\|\mathbf{v_1}\| = 1$, we get $\mathbf{v}_1 = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})^T$.

- Find the second principal component.

  Solution. The second principal component is the eigenvector corresponding to the second eigenvalue. It can be computed in the same way as above, $\mathbf{v}_2 = (\frac{1}{\sqrt{2}}, -\frac{1}{\sqrt{2}})^T$.

- Model the data as the sum of one principal component and some noise. Remove the noise from the data.

  Solution. Removing the noise of the data means to project it to its principal subspace. It consists of three steps: extract the mean from the data, project it to the first few (in our case one) principal components, and add back the mean. We already computed X_c, now we need to project it to the first principal component,

$$X_p = \mathbf{v}_1^T X_c = (-\frac{2}{\sqrt{2}}, -\frac{2}{\sqrt{2}}, \frac{2}{\sqrt{2}}, \frac{2}{\sqrt{2}}). \tag{6}$$

  X_p can be thought of as the compressed data. We can get the denoised data X_d by the following formula,

$$X_d = \mathbf{v}_1 \mathbf{v}_1^T X_c + \mu \cdot \mathbf{1}^T = \begin{pmatrix} 2.5 & 2.5 & 4.5 & 4.5 \\ 1.5 & 1.5 & 3.5 & 3.5 \end{pmatrix}. \tag{7}$$

NOTE In equation (2) and (7) we used the vector $\mathbf{1}^T$. This is simply a notation for the vector $\mathbf{1}^T = (1, 1, ...1)$, where the length of the vector is the number of data points m.

2. There is a nice theorem in numerical linear algebra, that says the following. Suppose M is an $m \times n$ matrix whose entries are real numbers. Then there exists a factorisation of the form

$$M = USV^T, \tag{8}$$

where U and V are orthogonal matrices, and S is diagonal, and $s_i \geq 0$ for every diagonal element of S. The s_i diagonal elements are called singular values, and the factorisation itself is called singular value decomposition (SVD). This decomposition is unique (in most cases). How would you compute the principal components using SVD?

Solution. The principal components are the eigenvectors of the covariance matrix of the data. The covariance matrix is expressed in the following form,

$$\Sigma = \frac{1}{m} X_c X_c^T. \tag{9}$$

Factorising $\frac{1}{\sqrt{m}} X_c$ with SVD, we get $\frac{1}{\sqrt{m}} X_c = USV^T$. We can write

$$\Sigma = \frac{1}{m} X_c X_c^T = USV^T VSU^T = US^2U^T. \tag{10}$$

It is easy to see, that the columns of U (denoted by u_i) are eigenvectors of Σ and the eigenvalues are the singular values squared $\lambda_i = s_i^2$.

$$\Sigma u_i = US^2U^T u_i = US^2 e_i = U(s_i^2 e_i) = s_i^2 U e_i = s_i^2 u_i, \tag{11}$$

where $e_i = (0, 0, \ldots, 1, \ldots, 0)^T$, the i^{th} element of e_i is 1, all the others are 0. There is an interesting thing to notice here. The factorization of $\Sigma = US^2U^T$ is a valid SVD factorization too. You can compute the SVD of $\Sigma = U_\Sigma S_\Sigma V_\Sigma^T$. Because the SVD is unique (apart from multiplying u_i with ± 1), $U_\Sigma = V_\Sigma = U$ and $S_\Sigma = S^2$. Computing the eigenvectors this way usually works faster in practice.

Factor Analysis

3. Let us assume that the random variable x given z is defined by the following model

$$\epsilon \sim \mathcal{N}(0, \Psi)$$
$$x = \mu + \Lambda R z + \epsilon,$$

where R is a known rotation matrix, $R^\top R = I$, and $z \in \mathbb{R}^n$ and $\epsilon \in \mathbb{R}^m$, with $n < m$, are independent random variables, and $\mathcal{N}(0, \Psi)$ denotes the zero-mean Gaussian distribution with covariance Ψ. Let $\Lambda \in \mathbb{R}^{m \times n}$, $\mu \in \mathbb{R}^m$ and $\Psi \in \mathbb{R}^{m \times m}$ be (unknown) model parameters. Let us assume that the expectations $\mathrm{E}_{z\sim Q}[z]$ and $\mathrm{E}_{z\sim Q}[zz^\top]$ are given, where Q is some probability density function of the vector z.

Let us assume that we are given m independent and identically distributed samples $x^{(1)}, \ldots, x^{(m)}$. Each sample $x^{(i)}$ will have an associated vector $z^{(i)}$, with its own probability density function Q_i. Compute Λ by maximizing the likelihood

$$\sum_{i=1}^{m} \mathrm{E}_{z^{(i)}\sim Q_i}\left[\log p(x^{(i)}|z^{(i)}; \mu, \Lambda, \Psi)\right] =$$
$$\sum_{i=1}^{m} \mathrm{E}_{z^{(i)}\sim Q_i}\left[-\frac{1}{2}\log|\Psi| - \frac{n}{2}\log(2\pi) - \frac{1}{2}\left(x^{(i)} - \mu - \Lambda R z^{(i)}\right)^\top \Psi^{-1}\left(x^{(i)} - \mu - \Lambda R z^{(i)}\right)\right]$$

Hint: Make use of the following *trace* formulas: $\nabla_A \mathrm{tr}(A^\top BAC) = BAC + B^\top AC^\top$ and $\mathrm{tr}(AB) = \mathrm{tr}(BA)$.

Solution:

We first compute the gradient of the likelihood function:

$$\nabla_\Lambda \sum_{i=1}^{m} -E_{z\sim Q}[\frac{1}{2}(x^{(i)} - \mu - \Lambda R z^{(i)})^T \Psi^{-1}(x^{(i)} - \mu - \Lambda R z^{(i)})] \tag{12}$$

$$= \sum_{i=1}^{m} \nabla_\Lambda E_{z\sim Q}[-\mathrm{tr}(\frac{1}{2} z^{(i)^T} R^T \Lambda^T \Psi^{-1} \Lambda R z^{(i)}) + \mathrm{tr}(z^{(i)^T} R^T \Lambda^T \Psi^{-1}(x^{(i)} - \mu))] \tag{13}$$

$$= \sum_{i=1}^{m} \nabla_\Lambda E_{z\sim Q}[-\mathrm{tr}(\frac{1}{2} \Lambda^T \Psi^{-1} \Lambda R z^{(i)} z^{(i)^T} R^T) + \mathrm{tr}(\Lambda^T \Psi^{-1}(x^{(i)} - \mu) z^{(i)^T} R^T)] \tag{14}$$

$$= \sum_{i=1}^{m} E_{z\sim Q}[-\Psi^{-1} \Lambda R z^{(i)} z^{(i)^T} R^T + \Psi^{-1}(x^{(i)} - \mu) z^{(i)^T} R^T] \tag{15}$$

Where

- In (12) we discard terms not containing Λ .

- In (13) we expanded and used $a = \text{tr}(a)$ for $a \in \mathbb{R}$.
- In (14) we used $\text{tr}(AB) = \text{tr}(BA)$.
- In (15) we used $\nabla_A \text{tr}(A^\top BAC) = BAC + B^\top AC^\top$ where $A = \Lambda$, $B = \Psi^{-1}$ and $C = Rz^{(i)}{z^{(i)}}^T R^T$. We also made use of $\nabla_A \text{tr}(A^\top B) = B$ and the fact that Ψ^{-1} and $Rz^{(i)}{z^{(i)}}^T R^T$ are symmetric.

We can then set (15) equal to 0 and solve for Λ:

$$\sum_{i=1}^{m} \Lambda E_{z\sim Q}[Rz^{(i)}{z^{(i)}}^T R^T] = \sum_{i=1}^{m} (x^{(i)} - \mu) E_{z\sim Q}[{z^{(i)}}^T R^T]$$

$$\Lambda R \sum_{i=1}^{m} E_{z\sim Q}[z^{(i)}{z^{(i)}}^T] = \sum_{i=1}^{m} (x^{(i)} - \mu) E_{z\sim Q}[z^{(i)}]^T$$

$$\Lambda = \left[(\sum_{i=1}^{m} (x^{(i)} - \mu) E_{z\sim Q}[z^{(i)}]^T)(\sum_{i=1}^{m} E_{z\sim Q}[z^{(i)}{z^{(i)}}^T])^{-1} \right] R^\top$$