PCA in Practice

Recap

Given $x_1, \ldots, x_m \in \mathbb{R}^n$ i.i.d random vectors with

$$\underbrace{\mu = \mathbb{E}[x]}_{\text{mean vector}} \quad \text{and} \quad \underbrace{C = \mathbb{E}[(x-\mu)(x-\mu)^T]}_{\text{covariance matrix}}$$

Cov. matrix C is symmetric positive semidefinite.

eigenvalue decomposition

$U^T U = U U^T = I$

$$C = U \Lambda U^T$$

$\lambda_1 \geq \lambda_2 \geq \cdots \lambda_n \geq 0$

Principle component analysis (PCA) analyzes $x_1, \ldots, x_m$ in the eigenbasis:

$$y_j = U^T (x_j - \mu)$$

$\Rightarrow$ New random vectors are mean zero and have diagonal covariance (uncorrelated entries)

mean $\mathbb{E}[y] = 0$ and $\mathbb{E}[y y^T] = \Lambda$ covariance matrix

$\Rightarrow$ Direction u_j maximizes $\mathrm{var}(y_j)$ s.t. $u_j \perp \mathrm{span}\{u_1, \ldots, u_{j-1}\}$

$$u_j = \operatorname{argmax} \mathbb{E}\left[|u_j^T (x - \mu)|^2\right]$$

Computing Principle Components

In practice, the mean and covariance of the random variable $x \in \mathbb{R}^n$ are not usually known. Instead, they can be estimated directly from the data.

$$\tilde{u} = \frac{1}{m} \sum_{j=1}^{m} x_j$$ (Sample mean)

$$\tilde{C} = \frac{1}{m-1} \sum_{j=1}^{m} (x_j - \tilde{u})(x_j - \tilde{u})^T$$ (Sample covariance)

$$= \frac{1}{m-1} BB^T \quad \text{where} \quad B = X - \tilde{u}\underline{1}^T$$

subtract mean from each column

Note that $\tilde{C}$ remains symmetric PSD.

$\Rightarrow$ We can calculate the eigenvalue decomp. of the sample covariance of the data:

$$\tilde{C} = U \Lambda U^T$$

$\Rightarrow$ Maximizes the sample variance along each successive eigendirection s.t. orthogonal constraints

Note that the sample covariance matrix

$$\tilde{C} = BB^T = (X - \tilde{u}1^T)(X - \tilde{u}1^T)^T$$

does not actually need to be formed in practice.

Rather, we compute principle components via the SVD of the mean-shifted data, $B = X - \tilde{u}1^T$.

$$\underbrace{\tilde{C} = U\Lambda U^T}_{\text{Eigenmatrix}} \iff \underbrace{B = U\Sigma V^T}_{\text{eigenmatrix}}$$

where $\Lambda = \Sigma^2$ and $B^T B = V \Lambda V^T$.

Often, one is only interested in the first, or first several principle components of the data, so one only needs to compute the first few singular vectors of B. This can mitigate large computational costs when m, n are large.

$\Rightarrow$ See MATLAB script for demos.