# 1 Principal Component Analysis

## 1.1 Motivation

PCA wants to identify a *meaningful* basis to re-express the dataset. PCA assumes that a *meaningful* data representation is one which

1. the features with large variance have meaningful structure and should be preserved

2. the features with small variance are noise and should be discarded

3. correlated features indicate redundancy and should be made uncorrelated

Suppose we have observations $\{x_i\}_{i=1}^N$ where $x_i \in \mathbb{R}^p$ for some random variable $\mathsf{x}$. We want to find linear transformation of $\mathsf{x}$ to obtain $\mathsf{y}$. In particular, let $\mathbf{X} \in \mathbb{R}^{N\times p}$ be stacked observations, we want to find a linear map $\mathbf{P} \in \mathbb{R}^{p\times q}$, where columns of $\mathbf{P}$ are orthonormal basis for feature space, i.e. $row(\mathbf{X})$, to re-express data $\mathbf{X}$ to $\mathbf{Y} \in \mathbb{R}^{N\times q}$.

$$\mathbf{Y} = \mathbf{X}\mathbf{P}$$

$\mathbf{Y}$ has a meaningful representation if $cov(\mathbf{Y})$ is a diagonal matrix (decorrelated), and that successive dimension in $\mathbf{Y}$ are rank-ordered according to variance (preserve, discard noise).

## 1.2 Empirical Covariance Matrix

Note that for a random variable $\mathsf{x}$ with stacked observations $\mathbf{X} \in \mathbb{R}^{N\times p}$, the empirical covariance for $\mathsf{x}_i, \mathsf{x}_j$ is given by

$$\hat{\sigma}^2(\mathsf{x}_i, \mathsf{x}_j) = \frac{1}{N}\sum_i (x_i - \overline{x}_i)(x_j - \overline{x}_j) = \frac{1}{N}\left(\mathbf{X}_i - \overline{\mathbf{X}}_i \mathbf{1}_N\right)^T \left(\mathbf{X}_j - \overline{\mathbf{X}}_j \mathbf{1}_N\right)$$

where $\mathbf{X}_i, \mathbf{X}_j$ are $i$ and $j$-th column of $\mathbf{X}$ and $\overline{\mathbf{X}}_i = \frac{1}{N}\sum_j \mathbf{X}_{ji}$. So then,

$$\hat{cov}(\mathsf{x}) = \left[\hat{\sigma}(\mathsf{x}_i, \mathsf{x}_j)\right]_{i,j=1}^p = \frac{1}{N}\left(\mathbf{X} - \overline{\mathbf{X}}\mathbf{1}_N\right)^T \left(\mathbf{X} - \overline{\mathbf{X}}\mathbf{1}_N\right)$$

where $\overline{\mathbf{X}}$ is column wise feature average of $\mathbf{X}$. For zero mean observation matrix, the empirical covariance matrix is simply $\frac{1}{N}\mathbf{X}^T\mathbf{X}$

## 1.3 Solving PCA using Eigenvector Decomposition

We first write covariance matrix for $\mathbf{Y}$,

$$\hat{cov}(\mathsf{y}) = \frac{1}{N}\mathbf{Y}^T\mathbf{Y} = \frac{1}{N}(\mathbf{X}\mathbf{P})^T(\mathbf{X}\mathbf{P}) = \mathbf{P}^T\left(\frac{1}{N}\mathbf{X}^T\mathbf{X}\right)\mathbf{P} = \mathbf{P}^T\hat{cov}(\mathsf{x})\mathbf{P}$$

We know that $\hat{cov}(\mathsf{x})$ is a symmetric matrix and therefore can be written as $\hat{cov}(\mathsf{x}) = \mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ where $\mathbf{Q} \in \mathbb{R}^{p\times p}$ are eigenvectors of $\hat{cov}(\mathsf{x})$ with corresponding eigenvalues along diagonal entries in $\mathbf{\Lambda}$. Setting projection to be eigenvectors of $\hat{cov}(\mathsf{x})$ diagonalizes $\hat{cov}(\mathsf{y})$,

$$\mathbf{P} \leftarrow \mathbf{Q} \qquad \Rightarrow \qquad \hat{cov}(\mathsf{y}) = \mathbf{P}^T\hat{cov}(\mathsf{x})\mathbf{P} = \mathbf{Q}^T\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T\mathbf{Q} = \mathbf{\Lambda}$$

where $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$. The *principal components* of $\mathbf{X}$ are column vectors of $\mathbf{P}$, i.e. eigenvectors for $\hat{cov}(\mathsf{x})$. $\mathsf{y}$ is decorrelated and $\hat{\sigma}^2(\mathsf{y}_i)$ is the variance of $\mathsf{x}$ along $i$-th principal component.