1. **Principal component analysis (PCA).** Assume that we represent given data (row vectors) $\mathbf{x}_1^\top, \mathbf{x}_2^\top, \ldots, \mathbf{x}_n^\top$ as rows of a matrix

$$X = \begin{bmatrix} \mathbf{x}_1^\top \\ \mathbf{x}_2^\top \\ \vdots \\ \mathbf{x}_n^\top \end{bmatrix} \in \mathbb{R}^{n\times d}.$$

We view components of vectors $\mathbf{x}_i^\top$ as various features of observed objects. Columns $\mathbf{c}_j$ of the matrix $X = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_d]$ are often called *feature vectors*.

The objective of this task is to find so-called *principal components* $\mathbf{y}_1, \ldots, \mathbf{y}_d \in \mathbb{R}^n$ which are uncorrelated projections of data $\mathbf{x}_i^\top$ onto unit vectors $\mathbf{v}_1^\top, \ldots, \mathbf{v}_d^\top$, such that the variances $\mathrm{var}(\mathbf{y}_i)$ are maximized. Some anchor points:

- *Centralization of data:* Subtract the mean value from each column of X to obtain
$$\overline{X} := X - [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_d]$$
where $\boldsymbol{\mu}_j = \mu_j[1, \ldots, 1]^\top$ and μ_j is the average value of components of the feature vector $\mathbf{c}_j$.
- *Evaluation of the singular value decomposition* of $\overline{X}$: $\overline{X} = USV^\top$ where $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n\times n}$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_d] \in \mathbb{R}^{d\times d}$, and $S \in \mathbb{R}^{n\times d}$ is a diagonal matrix with singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ on the diagonal.
- *Principal components* of X are $\mathbf{y}_1, \ldots, \mathbf{y}_d \in \mathbb{R}^n$ obtained as
$$\mathbf{y}_j = \overline{X}\mathbf{v}_j = \sigma_j \mathbf{u}_j.$$

Answer questions below.

(a) Let $\Sigma = \frac{1}{n-1}\overline{X}^\top\overline{X}$. Show that for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ we have $\mathrm{cov}(X\mathbf{v}, X\mathbf{w}) = \mathbf{v}^\top\Sigma\mathbf{w}$.

(b) How can $\mathrm{var}(\mathbf{y}_j) := \mathrm{cov}(\mathbf{y}_j, \mathbf{y}_j)$ be expressed with singular values of $\overline{X}$?

(c) Evaluate $\mathrm{cov}(\mathbf{y}_j, \mathbf{y}_k)$ za $j \neq k$.

Write these three Octave functions:

- `[mu, Vk, Uk, Dk]=pca(X, k)` which for a given data matrix X and an integer k, $0 \leq k \leq \min(n, d)$, returns averages `mu`, matrices `Vk` and `Uk` containing first k left/right principal directions, and a vector `Dk` with first k variances $\mathrm{var}(\mathbf{y}_j)$,
- `Z=proj(X)` which for a given data matrix X returns the projection of $\mathbf{x}_i^\top - [\mu_{i1}, \ldots, \mu_{id}]$ onto largest two principal directions and draws a picture of both principal directions and projections of data,

- `r=threshold(X, p)` which for a data matrix X and a number $p \in [0,1]$ returns the smallest integer r, such that

$$\frac{\mathrm{var}(\mathbf{y}_1) + \cdots + \mathrm{var}(\mathbf{y}_r)}{\mathrm{var}(\mathbf{y}_1) + \cdots + \mathrm{var}(\mathbf{y}_d)} \geq p$$

  holds.