1. **Analiza glavnih komponent (PCA).** Podatke (vrstične vektorje) $\mathbf{x}_1^\mathsf{T}, \mathbf{x}_2^\mathsf{T}, \ldots, \mathbf{x}_n^\mathsf{T}$ predstavimo kot vrstice matrike

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

Komponente vsakega vektorja $\mathbf{x}_i^\mathsf{T}$ si lahko predstavljamo kot (nabor) značilnosti opazovanih objektov. Stolpcem $\mathbf{c}_j$ matrike $X = [\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_d]$ pogosto pravimo *značilni vektorji* oz. *vektorji značilk*.

Cilj naloge je poiskati t.i. *glavne komponente* $\mathbf{y}_1, \ldots, \mathbf{y}_d \in \mathbb{R}^n$, ki so nekorelirane projekcije podatkov $\mathbf{x}_i^\mathsf{T}$ na enotske vektorje $\mathbf{v}_1^\mathsf{T}, \ldots, \mathbf{v}_d^\mathsf{T}$, ki maksimizirajo variance $\mathrm{var}(\mathbf{y}_i)$. Oporne točke so:

- *Centralizacija podatkov:* Od vsakega stolpca X odštejemo njegovo srednjo vrednost
$$\overline{X} := X - [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_d],$$
kjer je $\boldsymbol{\mu}_j = \mu_j[1, \ldots, 1]^\mathsf{T}$, μ_j pa je srednja vrednost komponent značilnega vektorja $\mathbf{c}_j$.
- *Izračun singularnega razcepa* $\overline{X}$: $\overline{X} = USV^\mathsf{T}$, kjer je $U = [\mathbf{u}_1, \ldots, \mathbf{u}_n] \in \mathbb{R}^{n \times n}$, $V = [\mathbf{v}_1, \ldots, \mathbf{v}_d] \in \mathbb{R}^{d \times d}$ ter $S \in \mathbb{R}^{n \times d}$ diagonalna matrika s singularnimi vrednostmi $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_d$ na diagonali.
- *Glavne komponente* X so $\mathbf{y}_1, \ldots, \mathbf{y}_d \in \mathbb{R}^n$, ki jih dobimo kot
$$\mathbf{y}_j = \overline{X}\mathbf{v}_j = \sigma_j \mathbf{u}_j.$$

Odgovori na spodnja vprašanja.

(a) Naj bo $\Sigma = \frac{1}{n-1}\overline{X}^\mathsf{T}\overline{X}$. Utemelji, da za $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$ velja $\mathrm{cov}(X\mathbf{v}, X\mathbf{w}) = \mathbf{v}^\mathsf{T}\Sigma\mathbf{w}$.

(b) Kako izrazimo $\mathrm{var}(\mathbf{y}_j) := \mathrm{cov}(\mathbf{y}_j, \mathbf{y}_j)$ s singularnimi vrednostmi matrike $\overline{X}$?

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

Napiši tri Octave funkcije:

- `[mu, Vk, Uk, Dk]=pca(X, k)`, ki za dano matriko s podatki X in za celo število k, $0 \leq k \leq \min(n, d)$, vrne srednje vrednosti `mu`, matriki `Vk` ter `Uk` s prvimi k levimi oz. desnimi glavnimi smermi in vektor `Dk` prvih k varianc $\mathrm{var}(\mathbf{y}_j)$,
- `Z=proj(X)`, ki za matriko s podatki X vrne projekcijo $\mathbf{x}_i^\mathsf{T} - [\mu_{i1}, \ldots, \mu_{id}]$ na največji dve desni glavni smeri in izriše obe glavni komponenti ter projekciji podatkov na isto sliko,
- `r=threshold(X, p)`, ki za matriko X ter za število $p \in [0, 1]$ vrne najmanjše celo število r, za katerega velja
$$\frac{\mathrm{var}(\mathbf{y}_1) + \cdots + \mathrm{var}(\mathbf{y}_r)}{\mathrm{var}(\mathbf{y}_1) + \cdots + \mathrm{var}(\mathbf{y}_d)} \geq p.$$