# Homework 4

*Posted: April 10, 2019*
*Due: April 24, 2019*

1. A series of five experiments involves recording three variables and produces the following data matrix:

| | $\mathcal{V}_1$ | $\mathcal{V}_2$ | $\mathcal{V}_3$ |
|---|---|---|---|
| $\mathbf{u}_1$ | 1 | 160 | 168 |
| $\mathbf{u}_2$ | 0 | 150 | 148 |
| $\mathbf{u}_3$ | 1 | 120 | 170 |
| $\mathbf{u}_4$ | 0 | 100 | 120 |
| $\mathbf{u}_5$ | 1 | 200 | 180 |

   Scale the matrix using the **R** function `scale`.
   Using singular value decompositions compute approximations of rank 1 and 2 of the *centered* matrix that corresponds to the data matrix given above.
2. Starting from the approximation of rank 2 of the data matrix defined above construct manually a biplot to represent data. What informations can be extracted from this biplot?
3. Consider the set of points that consists of two "entangled spirals" shown in Figure 1. Extract the coordinates of the points from Figure 1 and compute the `dist` object using the Euclidean metric. Apply spectral clustering to determine if the two sets of points of the two curves can be separated.
4. Let $D \in \mathbb{R}^{m\times n}$ be a centered data matrix and let $D = U\mathsf{diag}(\sigma_1 \ \cdots \ \sigma_r)V'$ be the thin SVD of $D$, where $U \in \mathbb{R}^{m\times r}$, $V \in \mathbb{R}^{n\times r}$, $U$ and $V$ have orthonormal columns. If

$$S = U\mathsf{diag}(\sigma_1 \ \cdots \ \sigma_r) = (\mathbf{s}_1 \ \cdots \ \mathbf{s}_r) = (\sigma_1\mathbf{u}_1 \ \cdots \ \sigma_r\mathbf{u}_r) \in \mathbb{R}^{m\times r}$$

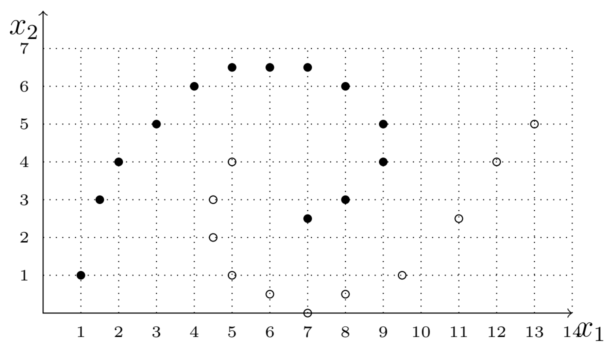


Figure 1: Two entangled curves

is the matrix of scores and $V \in \mathbb{R}^{n\times r}$ is the matrix of loadings, prove that

(a) the variance of a score vector $\mathbf{s}_i$ is $var(\mathbf{s}_i) = \frac{1}{m-1}\sigma_i^2$;
(b) $D = \mathbf{s}_1\mathbf{v}_1' + \cdots + \mathbf{s}_r\mathbf{v}_r'$;
(c) if $D_k = \mathbf{s}_1\mathbf{v}_1' + \cdots + \mathbf{s}_k\mathbf{v}_k'$, where $k \le r$, then

$$\frac{\mathsf{TVAR}(D_k)}{\mathsf{TVAR}(D)} = \frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}.$$

In other words $\frac{\sum_{i=1}^k \sigma_i^2}{\sum_{i=1}^r \sigma_i^2}$ indicates the portion of the total variance of $D$ explained by the first $k$ scores.