Annex : Classification - Description of the method

This method involves two steps : i) the **singular value decomposition** theorem is used to extract the important features from the data; ii) then the extracted features are combined in a **classification** step to form new features, each of which describing a subset of the data

1 Singular value decomposition

Formally, the singular value decomposition (SVD) of an $m \times n$ matrix $\mathbf{A}$ is a factorization of the form $\mathbf{U\Sigma V}^{\mathsf{T}}$, where $\mathbf{U}$ is an $m \times m$ unitary matrix, $\mathbf{\Sigma}$ is a $m \times n$ rectangular diagonal matrix with non-negative real numbers on the diagonal, and $\mathbf{V}$ is an $n \times n$ unitary matrix. The diagonal entries σ_i of $\mathbf{\Sigma}$ are known as the singular values of $\mathbf{A}$. The columns of $\mathbf{U}$ and the columns of $\mathbf{V}$ are called the left-singular vectors and right-singular vectors of $\mathbf{A}$, respectively.

Let $\mathbf{A} = \{a_{i,j}\}_{1 \leq i \leq m, 1 \leq j \leq n}$, and suppose $m \leq n$. Then

$$a_{i,j} = \sum_{k=1}^{m} \sigma_k u_{i,k} v_{j,k} \tag{1}$$

with $\mathbf{U} = \{u_{i,j}\}_{1 \leq i \leq m, 1 \leq j \leq m}$, $\mathbf{V} = \{v_{i,j}\}_{1 \leq i \leq n, 1 \leq j \leq m}$, and singular values

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_m \geq 0$$

The singular vectors are orthonormal, i.e.,

$$\sum_{i=1}^{m} u_{i,k} u_{i,l} = \sum_{j=1}^{n} v_{j,k} v_{j,l} = \begin{cases} 0 & \text{if } k \neq l \\ 1 & \text{if } k = l \end{cases}$$

The first L leading terms ($1 \leq L \leq m$) of the singular value expansion (Eq. (1)) can be used to approximate $\mathbf{A}$ by $\mathbf{A}^{(L)} = \{a_{i,j}^{(L)}\}_{1 \leq i \leq m, 1 \leq j \leq n}$, with

$$a_{i,j}^{(L)} = \sum_{k=1}^{L} \sigma_k u_{i,k} v_{j,k} \tag{2}$$

Moreover, the variance explained by each term is in proportion to the magnitude of the respective singular value, since

$$\sum_{i=1}^{m} \sum_{j=1}^{n} (a_{i,j} - a_{i,j}^{(L)})^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} a_{i,j}^2 - \sum_{k=1}^{L} \sigma_k^2 \tag{3}$$

So an efficient compaction of the data set is thus achieved and, as a result, the meaningful information is separated from the redundancy and noise inherent in the data. The choice of the number L of leading terms is crucial for the next step (classification) since the number of classes will be the same. For this purpose, Eq. (3) will be a decisive assistance in the choice of this number.

2 Classification

In this step, the features extracted by SVD are recombined to form mutually exclusive classes. The goal is to rearrange the leading terms of SVD to form characteristic shape functions. The left and singular vectors are rotated so that the approximation is preserved.

2.1 Rotation of the singular vectors

In the following, we consider that the number of leading terms L is selected, and, in order to symplify the notation, $\mathbf{A}$ means $\mathbf{A^{(L)}}$, $\mathbf{\Sigma}$ is the diagonal matrix of the first L singular values, and $\mathbf{U}$ and $\mathbf{V}$ are respectively the first L left and right singular vectors.

$$\mathbf{A} = \mathbf{U\Sigma V}^\mathsf{T}$$

Let $\mathcal{R}$ be a square $L \times L$ rotation matrix; its transpose is equal to its inverse : $\mathcal{R}\mathcal{R}^\mathsf{T} = \mathsf{I}$ where I is the identity matrix

$$\mathbf{A} = \mathbf{U}\mathcal{R}\mathcal{R}^\mathsf{T}\mathbf{\Sigma}(\mathbf{V})^\mathsf{T}$$

and, as $\mathcal{R}^\mathsf{T}$ and $\mathbf{\Sigma}$ commute :

$$\mathbf{A} = \mathbf{U}\mathcal{R}\mathbf{\Sigma}\mathcal{R}^\mathsf{T}\mathbf{V}^\mathsf{T}$$

$$\mathbf{A} = \mathbf{U}\mathcal{R}\mathbf{\Sigma}(\mathbf{V}\mathcal{R})^\mathsf{T}$$

Now, let's group the terms as follows : $\mathbf{U}\mathcal{R}\mathbf{\Sigma} = \mathcal{S}$ and $\mathbf{V}\mathcal{R} = \mathcal{C}$; then

$$\mathbf{A} = \mathcal{S}\mathcal{C}^\mathsf{T}$$

$\mathcal{S} = \{s_{i,k}\}$ is a $m \times L$ matrix whose columns will be called the characteristic shapes, and $\mathcal{C} = \{c_{j,k}\}$ is a $n \times L$ matrix whose columns are the new right vectors, called coefficients of characteristic shapes.

$$a_{i,j} = \sum_{k=1}^{L} s_{i,k} c_{j,k} \tag{4}$$

coefficient $c_{j,k}$ referring to the jth column of matrix $\mathbf{A}$. Note that orthonormality of $\mathcal{C}$ is preserved.

2.2 Criterium for the determination of groups

The goal of the classification is to enhance features, each of which describing a subset of the data. We called these features characteristic shapes. Consider the jth column of matrix $\mathbf{A}$; we call it the jth data profile :

$$\hat{a}_j = \{a_{i,j}\}_{1 \leq i \leq m}$$

The kth characteristic shape is written as :

$$\hat{s}_k = \{s_{i,k}\}_{1 \leq i \leq m} \tag{5}$$

and Eq.(4) as :

$$\hat{a}_j = \sum_{k=1}^{L} c_{j,k} \hat{s}_k$$

The jth data profile is a linear combination of the characteristic shapes. The aim of the optimisation is that such a profile is best described by one feature and only one; suppose it is characteristic shape $\hat{s}_{k_1}$. Then

$$c_{j,k_1} \gg c_{j,k_2} \ \forall k_2 \neq k_1$$

A mathematical statement of this requirement would be

$$\sum_{j=1}^{n} (c_{j,k_1} c_{j,k_2})^2 \text{minimum}$$

So the criterion adopted to determine the rotation matrix is

$$\sum_{\substack{k_1,k_2=1 \\ k_2>k_1}}^{L} \sum_{j=1}^{n} (c_{j,k_1} c_{j,k_2})^2 \text{minimum}$$

2.3 Numerical resolution

Now we have to find the rotation matrix in order to solve the problem of minimization. For that, we use the fact that any rotation matrix of size $L \times L$ can be constructed as a product of at most $L \times (L-1)/2$ Givens rotation matrices. A Givens rotation is represented by a matrix of the form

$$G(i,j,\theta) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix}$$

where $c = \cos(\theta)$ and $s = \sin(\theta)$ appear at the intersections ith and jth rows and columns.

The rotation matrix $\mathcal{R}$ is written as :

$$\mathcal{R} = \prod_{\substack{i,j=1 \\ j>i}}^{L} G(i,j,\theta_{i,j})$$

The minimization focuses on the $L \times (L-1)/2$ $\theta_{i,j}$ angles of Givens rotations. It is achieved by the method of Nelder and Mead (1965), that uses only function values and is robust.

2.4 Achieving classification

Let minimization being achieved; $\mathcal{S}$ and $\mathcal{C}$ are respectively the optimal matrix of characteristic shapes and the optimal matrix of coefficients. We defined L groups corresponding to the L characteristic shapes. At characteristic shape $\hat{s}_k$ correspond the group of profiles G_k

$$G_k = \{\hat{a}_j = \sum_{l=1}^{L} c_{j,l}\hat{s}_l, j \in [\![1;n]\!], \max_{l=1,L} |c_{j,l}| = |c_{j,k}|\}$$

In other words, a profile $\hat{a}_j$ belongs to group k if its coefficient with the greatest absolute value is the coefficient of $\hat{s}_k$.

Once each profile is affected to a group, we define the mean shapes for the groups. Let N_k be the set of indexes of profiles of group G_k

$$N_k = \{j \in [\![1;n]\!], \hat{a}_j \in G_k\}$$

and n_k the number of elements of N_k. The mean coefficient for the group is defined as :

$$\bar{c}_k = 1/n_k \sum_{j \in N_k} c_{j,k} \tag{6}$$

and the mean shape of group k as :

$$\bar{s}_k = \bar{c}_k \hat{s}_k \tag{7}$$