Tensor Regression

CAIT
Skolkowo Institute of Science and Technology (SKOLTECH), Moscow, Russia

March 17, 2022

Problem (Feature extraction for 2-D samples)

- *A set of K data matrices* $\mathbf{X}^{(k)} \in \mathbb{R}^{I_1 \times I_2}, (k = 1, \ldots, K)$ *that belong to C different classes.*
- *Feature extraction for all samples by simultaneous matrix factorizations:*

$$\mathbf{X}^{(k)} \approx \mathbf{A}^{(1)}\, \mathbf{F}^{(k)}\, \mathbf{A}^{(2)\,T}$$

$\mathbf{A}^{(1)} \in \mathbb{R}^{I_1 \times R_1}$ *and* $\mathbf{A}^{(2)} \in \mathbb{R}^{I_2 \times R_2}$, *are two basis matrices*
$\mathbf{F}^{(k)} \in \mathbb{R}^{R_1 \times R_2}$ *are features,*
$R_n \leq I_n$.

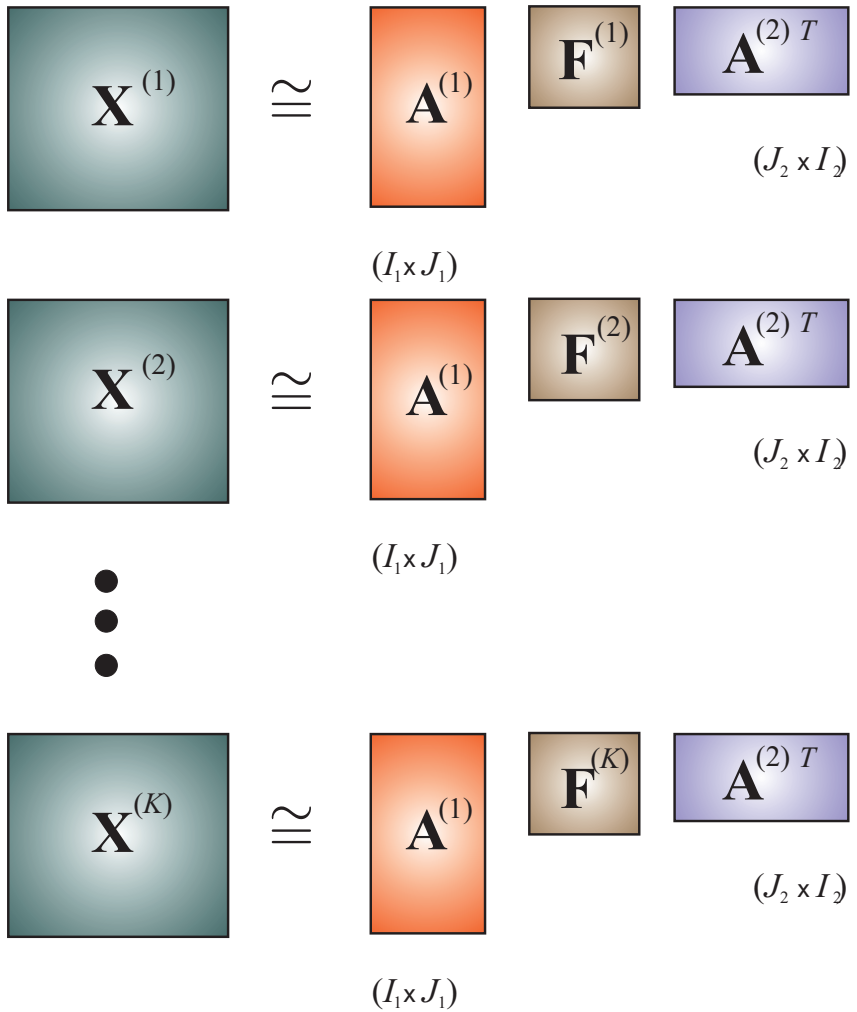


Figure: Simultaneous matrix factorizations

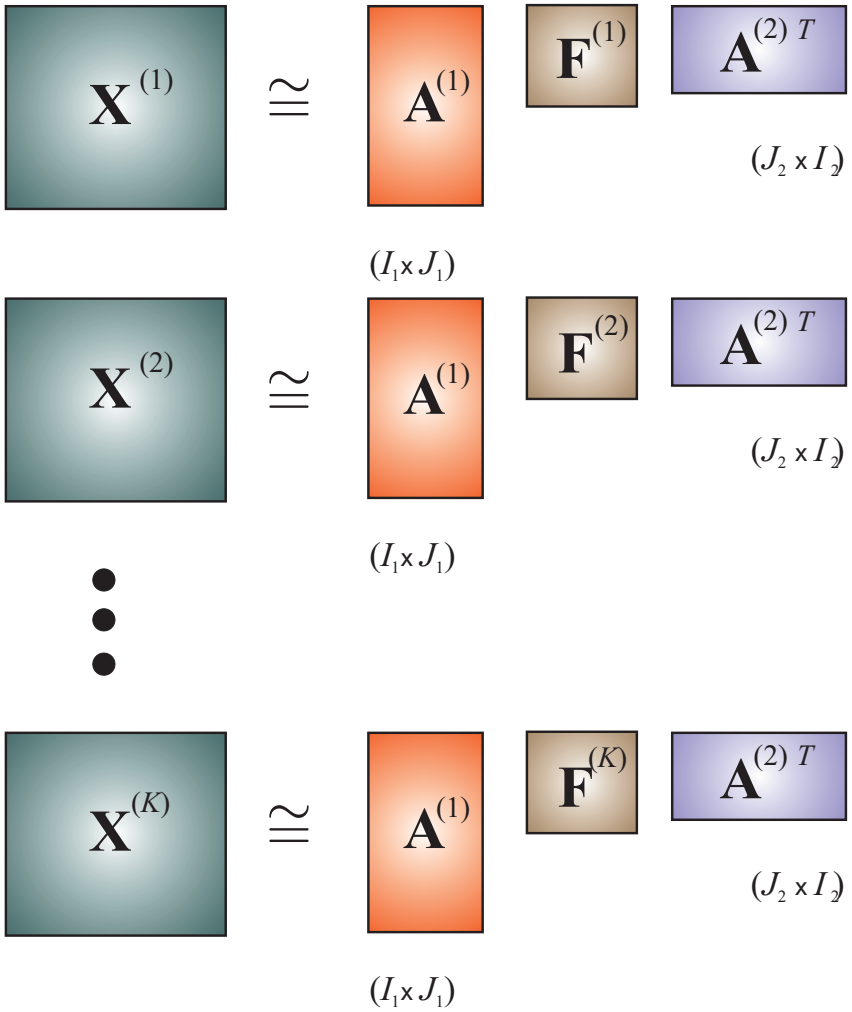


Figure: Simultaneous approximative matrix factorizations of 2-D samples

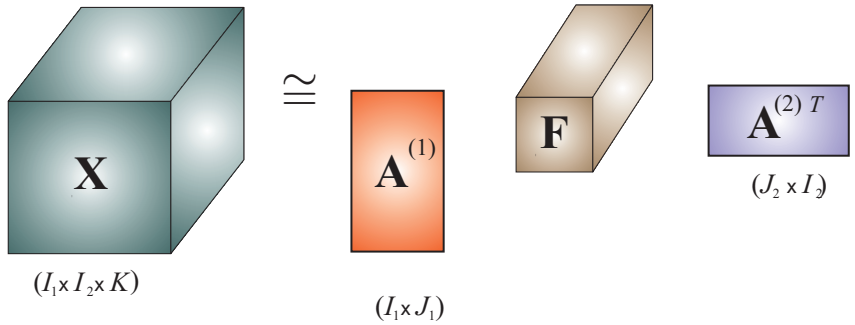


Figure: Equivalent 3D tensor decomposition: TUCKER-2 model.

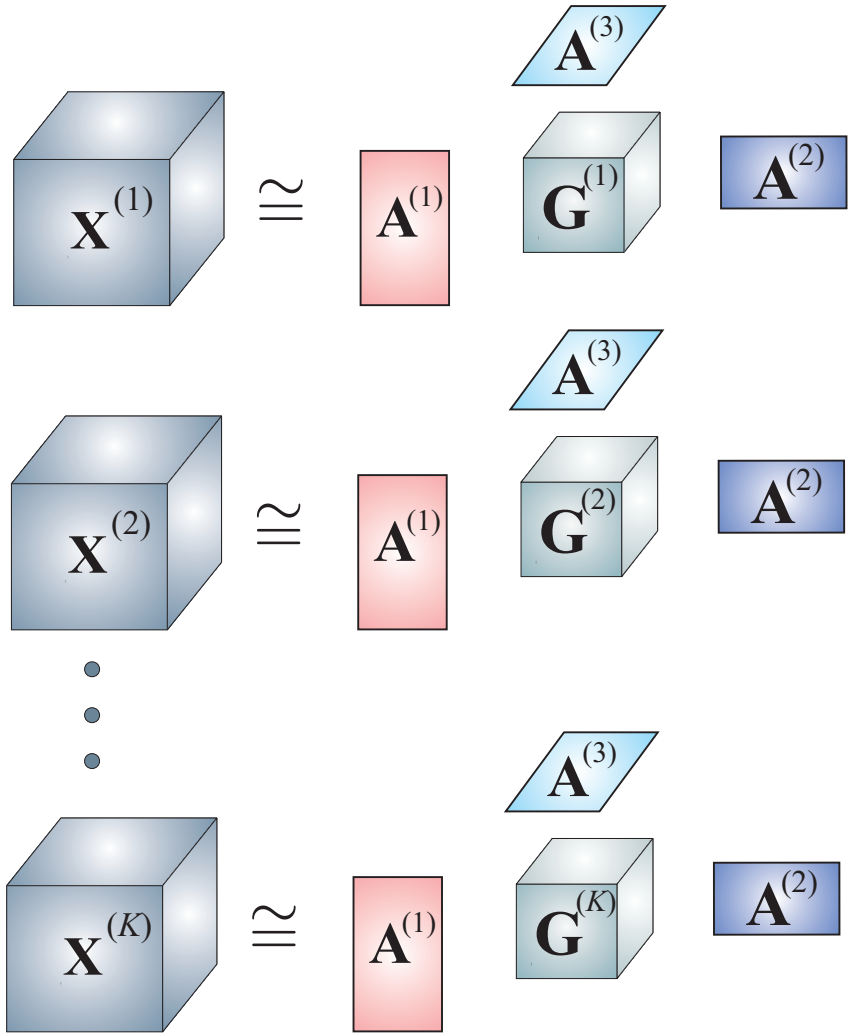
- If $\mathbf{F}^{(k)}$ are diagonal matrices, the model corresponds to CPD
- If $\mathbf{A}^{(n)}$ are orthogonal and $\mathbf{F}^{(k)}$ are dense, the model becomes HOSVD or multi-way PCA De Lathauwer et al. (2001)
- If $\mathbf{A}^{(n)}$ are nonnegative, the model becomes Tri Nonnegative Matrix Factorization of data $\mathbf{X}^{(k)}$.
- if $\mathbf{X}^{(k)}$ are positive-definite covariance or cumulant matrices, model becomes closely related to Joint Diagonalization, ICA.

Problem (N-way feature extraction)

Find N common factors $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R_n}, (n = 1, 2, \ldots, N)$ *from K simultaneous decompositions of K samples* $\mathcal{X}^{(k)} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$

$$\mathcal{X}^{(k)} \approx \mathcal{G}^{(k)} \times_1 \mathbf{A}^{(1)} \ldots \times_N \mathbf{A}^{(N)},$$

where $\mathcal{G}^{(k)} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N}$ *consist of features of* $\mathcal{X}^{(k)}$.



Problem (Global TUCKER decomposition)

The N common bases of K samples $\mathcal{X}^{(k)} \in \mathbb{R}^{I_1 \times I_2 \times \cdots \times I_N}$ *are factor matrices,* $\mathbf{A}^{(n)} \in \mathbb{R}^{I_n \times R_n}$*, for* $n = 1, \ldots, N$*, in the TUCKER-N decomposition of the concatenation tensor along the mode-(N+1), that is*

$$\mathcal{X} \approx \mathcal{G} \times_1 \mathbf{A}^{(1)} \times_2 \mathbf{A}^{(2)} \cdots \times_N \mathbf{A}^{(N)},$$

$\mathcal{X} = \mathtt{cat}(N+1, \mathcal{X}^{(1)}, \mathcal{X}^{(2)}, \ldots, \mathcal{X}^{(K)})$

$\mathcal{G}$ *represents extracted features for the training samples.*

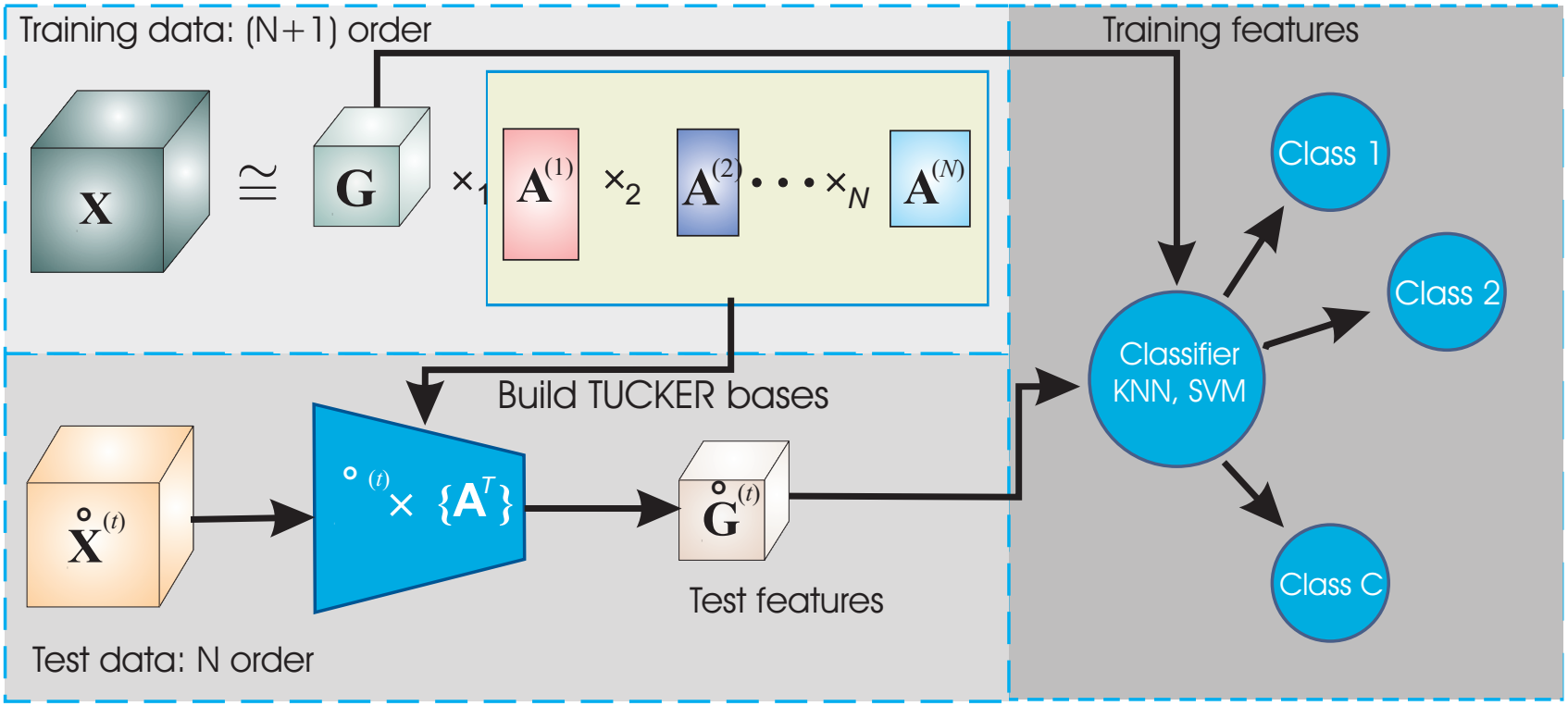


Figure: Conceptual diagram illustrating a classification procedure based on the TUCKER decomposition of the concatenated tensor of all sampling training data. Reduced features are obtained by projecting the data tensor onto the feature subspace spanned by basis factors (bases).

Algorithm 1: HOOI for Orthogonal Bases

Input: $\mathcal{X}$: concatenation tensor of all training samples
$I_1 \times I_2 \times \cdots \times I_N \times K$,

Output: N orthogonal factors $\mathbf{U}^{(n)} \in \mathbb{R}^{I_n \times R_n}$ and a core tensor
$\mathcal{G} \in \mathbb{R}^{R_1 \times R_2 \times \cdots \times R_N \times K}$.

begin

  HOSVD or random initialization for all factors $\mathbf{U}^{(n)}$

  repeat

    for $n = 1$ *to* N **do**

      $\mathcal{W}^{(-n)} = \mathcal{X} \times_{-(n,N+1)} \{\mathbf{U}^T\}$

      $[\mathbf{U}^{(n)}, \Sigma^{(n)}, \mathbf{V}^{(n)}] = \text{svds}(\mathbf{W}^{(-n)}_{(n)}, R_n, \text{'LM'})$

    end

  until *a stopping criterion is met*

  $\mathcal{G} = \mathcal{W}^{(-N)} \times_N \mathbf{U}^{(N)T}$

end

Discriminant Analysis

Suppose we have a set of training samples $\mathbf{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$ belonging to C classes. (L)DA maximizes the between-class scatter $\mathbf{S}_b$ while minimizing the within-class scatter $\mathbf{S}_w$.

Problem (Multilinear Discriminant Analysis)

$$\arg \max_{\mathbf{U}^{(1)},\ldots,\mathbf{U}^{(N)}} \frac{\sum_{c=1}^{C} K_c \left\| \bar{\mathcal{G}}^{(c)} - \bar{\bar{\mathcal{G}}} \right\|_F^2}{\sum_{k}^{K} \left\| \mathcal{G}^{(k)} - \bar{\mathcal{G}}^{(c_k)} \right\|_F^2}$$

$\bar{\mathcal{G}}^{(c)} = \left(\sum_{k \in \mathcal{I}_c} \mathcal{G}^{(k)} \right) / K_c$: mean tensor of the c-th class

$\bar{\bar{\mathcal{G}}} = \left(\sum_{k=1}^{K} \mathcal{G}^{(k)} \right) / K$: mean tensor of the whole training features

c_k class to which $\mathcal{X}^{(k)}$ belongs

Within-class scatter matrix

$$\begin{aligned}
\sum_{k=1}^{K} \left\| \mathcal{G}^{(k)} - \bar{\mathcal{G}}^{(c_k)} \right\|_F^2 &= \sum_{k=1}^{K} \left\| \left(\mathcal{X}^{(k)} - \bar{\mathcal{X}}^{(c_k)} \right) \times_{-n} \{\mathbf{U}^T\} \times_n \mathbf{U}^{(n)\,T} \right\|_F^2 \\
&= \sum_{k=1}^{K} \mathrm{tr} \left[\mathbf{U}^{(n)\,T} \langle \tilde{\mathcal{Z}}^{(k)}_{-n}, \tilde{\mathcal{Z}}^{(k)}_{-n} \rangle_{-n} \mathbf{U}^{(n)} \right] \\
&= \mathrm{tr} \left[\mathbf{U}^{(n)\,T} \mathbf{S}_w^{-n} \mathbf{U}^{(n)} \right],
\end{aligned}$$

The within-class scatter matrix $\mathbf{S}_w^{-n}$ is defined as

$$\begin{aligned}
\mathbf{S}_w^{-n} &= \langle \tilde{\mathcal{Z}}^{-n}, \tilde{\mathcal{Z}}^{-n} \rangle_{-n} \\
\tilde{\mathcal{X}}^{(k)} &= \mathcal{X}^{(k)} - \bar{\mathcal{X}}^{(c_k)}, \\
\tilde{\mathcal{Z}}^{-n} &= \tilde{\mathcal{X}} \times_{-(n,N+1)} \{\mathbf{U}^T\},
\end{aligned}$$

Between-class scatter matrix

$$\begin{aligned}
\sum_{c=1}^{C} K_c \left\| \bar{\mathcal{G}}^{(c)} - \bar{\bar{\mathcal{G}}} \right\|_F^2 &= \sum_{c=1}^{C} K_c \left\| \left(\bar{\mathcal{X}}^{(c)} - \bar{\bar{\mathcal{X}}} \right) \times_{-n} \{\mathbf{U}^T\} \times_n \mathbf{U}^{(n)\,T} \right\|_F^2 \\
&= \sum_{c=1}^{C} \operatorname{tr}\left[\mathbf{U}^{(n)\,T} \langle \check{\mathcal{Z}}^{(c)}_{-n}, \check{\mathcal{Z}}^{(c)}_{-n} \rangle_{-n} \mathbf{U}^{(n)} \right] \\
&= \operatorname{tr}\left[\mathbf{U}^{(n)\,T} \mathbf{S}_b^{-n} \mathbf{U}^{(n)} \right],
\end{aligned}$$

The between-class scatter matrix $\mathbf{S}_b^{-n}$ is defined as

$$\begin{aligned}
\mathbf{S}_b^{-n} &= \langle \check{\mathcal{Z}}^{-n}, \check{\mathcal{Z}}^{-n} \rangle_{-n} \\
\check{\mathcal{X}}^{(c)} &= \sqrt{K_c} \left(\bar{\mathcal{X}}^{(c)} - \bar{\bar{\mathcal{X}}} \right) \\
\check{\mathcal{Z}}^{-n} &= \check{\mathcal{X}} \times_{-(n,N+1)} \{\mathbf{U}^T\}
\end{aligned}$$

Problem (Discriminant Bases)

$$\max_{\mathbf{U}^{(n)}} \frac{tr\left[\mathbf{U}^{(n)\,T}\,\mathbf{S}_b^{-n}\,\mathbf{U}^{(n)}\right]}{tr\left[\mathbf{U}^{(n)\,T}\,\mathbf{S}_w^{-n}\,\mathbf{U}^{(n)}\right]} = \max_{\mathbf{U}^{(n)}} \frac{tr\left[\mathbf{S}_b\right]}{tr\left[\mathbf{S}_w\right]}$$

find R_n leading left eigenvectors of matrices

- $\mathbf{S}_w^{-1}\mathbf{S}_b$
- $\mathbf{S}_b - \varphi\,\mathbf{S}_w$

$$\arg\max_{\mathbf{U}^{(n)}} \frac{\mathrm{tr}\left[\mathbf{S}_b\right]}{\mathrm{tr}\left[\mathbf{S}_w\right]} = \arg\max_{\mathbf{U}^{(n)}} \frac{\mathrm{tr}\left[\mathbf{S}_b + \mathbf{S}_w\right]}{\mathrm{tr}\left[\mathbf{S}_w\right]}$$

$$\mathrm{tr}\left[\mathbf{S}_t\right] = \mathrm{tr}\left[\mathbf{S}_b + \mathbf{S}_w\right] = \sum_{k=1}^{K} \left\| \mathcal{G}^{(k)} - \bar{\bar{\mathcal{G}}} \right\|_F^2$$

Problem (Discriminant Bases)

$$\max_{\mathbf{U}^{(n)}} \frac{tr\left[\mathbf{U}^{(n)\,T}\,\mathbf{S}_b^{-n}\,\mathbf{U}^{(n)}\right]}{tr\left[\mathbf{U}^{(n)\,T}\,\mathbf{S}_w^{-n}\,\mathbf{U}^{(n)}\right]} = \max_{\mathbf{U}^{(n)}} \frac{tr\left[\mathbf{S}_b\right]}{tr\left[\mathbf{S}_w\right]}$$

find R_n leading left eigenvectors of matrices

- $\mathbf{S}_w^{-1}\mathbf{S}_b$
- $\mathbf{S}_b - \varphi\,\mathbf{S}_w$

$$\arg\max_{\mathbf{U}^{(n)}} \frac{\mathrm{tr}\left[\mathbf{S}_b\right]}{\mathrm{tr}\left[\mathbf{S}_w\right]} = \arg\max_{\mathbf{U}^{(n)}} \frac{\mathrm{tr}\left[\mathbf{S}_b + \mathbf{S}_w\right]}{\mathrm{tr}\left[\mathbf{S}_w\right]}$$

$$\mathrm{tr}\left[\mathbf{S}_t\right] = \mathrm{tr}\left[\mathbf{S}_b + \mathbf{S}_w\right] = \sum_{k=1}^{K} \left\|\mathcal{G}^{(k)} - \bar{\bar{\mathcal{G}}}\right\|_F^2$$

Problem (Regularization)

$$\varphi = \max_{\mathbf{U}^{(n)}} \frac{tr\left[\mathbf{S}_t\right]}{tr\left[\alpha\,\mathbf{S}_w + (1-\alpha)\,\mathbf{I}\right]}, \tag{1}$$

where $0 \leq \alpha \leq 1$.
The basis factors $\mathbf{U}^{(n)}$ *are* R_n *leading eigenvectors of matrices*

- $(\alpha\,\mathbf{S}_w + (1-\alpha)\,\mathbf{I})^{-1}\mathbf{S}_b$
- $\mathbf{S}_b - \varphi\,(\alpha\,\mathbf{S}_w + (1-\alpha)\,\mathbf{I})$

- For $\alpha = 1$, the optimization problem (1) simplifies MDA.
- For $\alpha = 0$, the optimization problem (1) simplifies the TUCKER decomposition of the training samples after centering.

Algorithm 2: HODA Algorithm for Feature Extraction

input : $\mathcal{X}$: Concanated tensor of K training samples $I_1 \times I_2 \times \cdots \times I_N \times K$

output: $\mathbf{U}^{(n)}$: N orthogonal basis factors $I_n \times R_n$ $(n = 1, 2, \ldots, N)$

begin

Initialize $\mathbf{U}^{(n)}$

Calculate $\tilde{\mathcal{X}}$, and $\check{\mathcal{X}}$

repeat

for $n = 1$ *to* N **do**

$$\tilde{\mathcal{Z}} = \tilde{\mathcal{X}} \times_{-(n,N+1)} \{\mathbf{U}^T\}, \quad \mathbf{S}_w^{-n} = \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{Z}} \rangle_{-n}$$

$$\check{\mathcal{Z}} = \check{\mathcal{X}} \times_{-(n,N+1)} \{\mathbf{U}^T\}, \quad \mathbf{S}_b^{-n} = \langle \check{\mathcal{Z}}, \check{\mathcal{Z}} \rangle_{-n}$$

$$\varphi = \frac{\texttt{trace}\left(\mathbf{U}^{(n)\,T}\, \mathbf{S}_b^{-n}\, \mathbf{U}^{(n)}\right)}{\texttt{trace}\left(\mathbf{U}^{(n)\,T}\, \mathbf{S}_w^{-n}\, \mathbf{U}^{(n)}\right)}$$

$$[\mathbf{U}^{(n)}, \Lambda] = \texttt{eigs}(\mathbf{S}_b^{-n} - \varphi\, \mathbf{S}_w^{-n}, R_n, \text{'LM'})$$

end

until *a criterion is met*

end

Algorithm 2: HODA Algorithm for Feature Extraction

input : $\mathcal{X}$: Concanated tensor of K training samples $I_1 \times I_2 \times \cdots \times I_N \times K$

output: $\mathbf{U}^{(n)}$: N orthogonal basis factors $I_n \times R_n$ $(n = 1, 2, \ldots, N)$

begin

  Initialize $\mathbf{U}^{(n)}$

  Calculate $\tilde{\mathcal{X}}$, and $\check{\mathcal{X}}$

  repeat

    for $n = 1$ *to* N **do**

      $\tilde{\mathcal{Z}} = \tilde{\mathcal{X}} \times_{-(n,N+1)} \{\mathbf{U}^T\}, \quad \mathbf{S}_w^{-n} = \langle \tilde{\mathcal{Z}}, \tilde{\mathcal{Z}} \rangle_{-n}$

      $\check{\mathcal{Z}} = \check{\mathcal{X}} \times_{-(n,N+1)} \{\mathbf{U}^T\}, \quad \mathbf{S}_b^{-n} = \langle \check{\mathcal{Z}}, \check{\mathcal{Z}} \rangle_{-n}$

      $[\mathbf{U}^{(n)}, \Lambda] = \texttt{eigs}(\mathbf{S}_b^{-n}, \mathbf{S}_w^{-n}, R_n, \text{'LM'})$

    end

  until *a criterion is met*

end

Feature Selection

- Ranking the features in a descending order of their Fisher scores

$$\varphi(i) = \frac{\sum_{c=1}^{C} K_c \, (\bar{g}_i^{(c)} - \bar{\bar{g}}_i)^2}{\sum_{k=1}^{K} (g_i^{(k)} - \bar{g}_i^{(c_k)})^2}, \qquad (i = 1, 2, \ldots, L), \qquad (2)$$

  where $\boldsymbol{g}^{(k)} = \text{vec}\left(\mathcal{G}^{(k)}\right)$.
- Choose significant features

Graph Regularizer I

- Denote by $\tilde{\mathcal{X}}_k = \mathcal{X}_k - \bar{\bar{\mathcal{X}}}$ the centered samples, $\tilde{\mathbf{X}} = [\ldots, \tilde{\mathcal{X}}_k, \ldots]$
- The Total scatter matrix

$$\begin{aligned} \mathbf{S}_t &= \mathbf{S}_w + \mathbf{S}_b \\ &= \tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \end{aligned}$$

- The between-scatter matrix

$$\mathbf{S}_b = \tilde{\mathbf{X}}^T \mathbf{W} \tilde{\mathbf{X}}$$

where $\mathbf{W} = \text{diag}(\frac{1}{K_1}\mathbf{1}_{K_1 \times K_1}, \ldots \frac{1}{K_C}\mathbf{1}_{K_C \times K_C})$

Graph Regularizer I

- Graph Regularizer is given in form of

$$\frac{1}{2}\sum_{i\neq j} w_{i,j}\,\|\mathcal{G}_i - \mathcal{G}_j\|_F^2 \tag{3}$$

where $\mathbf{W} = [w_{i,j}]$ is an adjacency matrix, defines the similarity between i-th and j-th samples.

- Similarity can be calculated as cosine similarity

$$w_{i,j} = \frac{\boldsymbol{x}_i^T \boldsymbol{x}_j}{\|\boldsymbol{x}_i\|\|\boldsymbol{x}_j\|} \tag{4}$$

or

$$w_{i,j} = \begin{cases} \exp(-\frac{d_{i,j}^2}{2\sigma^2}), & \text{if (i, j) in the same class} \\ 0 \end{cases} \tag{5}$$

Graph Regularizer II

where $d^2_{i,j}$ measures Euclidean distance between two samples, σ controls the sensitivity of the graph weights to the distances

- Laplacian matrix $\mathbf{L} = \mathbf{D} - \mathbf{W}$ where $\mathbf{L}$ is Laplacian matrix of the sample similarity based graph and $\mathbf{D}$ is the degree matrix. The degree matrix $\mathbf{D}$ is a diagonal matrix with its i-th diagonal entry computed as $\sum_{j=1} w_{ij}$

Local Discriminant Analysis

Denote by k_w and k_b the number of within and between nearest neighbours

$$\mathbf{W}_w(i,j) = \begin{cases} \frac{1}{k_w} & j \in I_{w,k_w}(i) \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

where $I_{w,k_w}(i)$ denotes index set of top-k_w samples closest to this to the i-th sample and in the same class with this sample.

$$\mathbf{W}_b(i,j) = \begin{cases} \frac{1}{k_w+k_b} - \frac{1}{k_w} & j \in I_{b,k_b}(i) \\ 0, & \text{otherwise} \end{cases} \tag{7}$$

where $I_{b,k_b}(i)$ denotes index set of top-k_b samples closest to this to the i-th sample but not in the same class with this sample.

References I

De Lathauwer, L., de Moor, B., and Vandewalle, J. (2001). A multilinear singular value decomposition. *SIAM Journal of Matrix Analysis and Applications*, 21:1253–1278.