

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/371905426>

# TDbasedUFE and TDbasedUFEadv: bioconductor packages to perform tensor decomposition based unsupervised feature extraction

Poster · June 2023

CITATIONS

0

READS

20

1 author:



Y-h. Taguchi

Chuo University

394 PUBLICATIONS 3,032 CITATIONS

SEE PROFILE

Some of the authors of this publication are also working on these related projects:



Tensor decomposition based unsupervised feature extraction [View project](#)



Principal Component Analysis based unsupervised Feature extraction [View project](#)

# TDbasedUFE and TDbasedUFEadv: bioconductor packages to perform tensor decomposition based unsupervised feature extraction

Y-h.Taguchi

Department of Physics, Chuo University, Tokyo 112-8551, Japan

tag@granular.com



## Abstract

We have proposed tensor decomposition (TD) based unsupervised feature extraction (FE) six years ago and applied it to wide range of bioinformatics problem. Although TD based unsupervised FE was generally applied to bioinformatics, it can have capability to select features under the situation of **large p small n** problem. In spite of successful applications, TD based unsupervised FE cannot be popular in the field of bioinformatics. In order to let the researchers who are not familiar with TD to perform TD based unsupervised FE, we developed R packages, TDbasedUFE and TDbasedUFEadv and submit it to Bioconductor, which is a long running R package repository for bioinformatics. In this poster, we introduce **mathematical background** behind TD based unsupervised FE.

## 1. Introduction

In bioinformatics, **large p small n** problem is very usual, since the number of genes (features) is as many as  $10^4$  whereas the number of subjects (conditions) is as small as 10 to  $10^2$ . Thus, it is required to have some method to deal with **large p small n** problem effectively. We have proposed a method, TD based unsupervised FE six years ago and applied it to various problems in bioinformatics. In spite of successful applications to various problems, its popularity is not enough. Then we have released two bioconductor packages by which one can make use of TD based unsupervised FE.

## 2. Feature selection procedure

TD base unsupervised FE follows the following procedures.

### Feature selection

1. Suppose that we have a tensor  $x_{ijk} \in \mathbb{R}^{N \times M \times K}$  that represents the amount of the value of the  $i$ th feature of the  $j$ th and  $k$ th conditions (although we assume here three mode tensor, extension to tensors with higher modes is straightforward).

2. Apply higher order singular value decomposition (HOSVD) to  $x_{ijk}$  and get Tucker decomposition as

$$x_{ijk} = \sum_{\ell_1=1}^N \sum_{\ell_2=1}^M \sum_{\ell_3=1}^K G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \quad (1)$$

where  $G(\ell_1 \ell_2 \ell_3) \in \mathbb{R}^{N \times M \times K}$  is a core tensor that represents the weight of product,  $u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k}$  to  $x_{ijk}$ ,  $u_{\ell_1 i} \in \mathbb{R}^{N \times N}$ ,  $u_{\ell_2 j} \in \mathbb{R}^{M \times M}$ ,  $u_{\ell_3 k} \in \mathbb{R}^{K \times K}$  are singular value matrices and orthogonal matrices.

3. Identify singular value vectors of interest,  $u_{\ell_2 j}$  and  $u_{\ell_3 k}$ , among those attributed to conditions,  $j$  and  $k$  (e.g., distinction between class labels, etc).

4. Select a  $u_{\ell_1 i}$  associated with the largest absolute value of  $G(\ell_1 \ell_2 \ell_3)$  with fixed  $\ell_2$  and  $\ell_3$  selected in the previous step among those attributed to features.

5. Optimize the standard deviation,  $\sigma_{\ell_1}$ , (see below) of the Gaussian distribution that the selected  $u_{\ell_1 i}$  is supposed to obey (the null hypothesis).

6. Attribute  $P$ -values to  $i$ th feature as

$$P_i = P_{\chi^2} \left[ > \left( \frac{u_{\ell_1 i}}{\sigma_{\ell_1}} \right) \right] \quad (2)$$

where  $P_{\chi^2}[> x]$  is the cumulative  $\chi^2$  distribution where the argument is larger than  $x$ .

7.  $P_i$ s are corrected by Benjamini-Hochberg (BH) criterion (multiple comparison correction) and  $i$ s associated with the threshold value (typically, 0.01 or 0.05) are selected.

## 3. Optimization of standard deviation

The standard deviation is overestimated if we include outliers that are supposed to be deviated from the null distribution. To exclude outliers from the estimation of the standard deviation, we perform as follows.

### Optimization of SD

1. Set threshold adjusted  $P$ -value,  $P_0$ , and the initial value of standard deviation,  $\sigma_{\ell_1}$ .
2. Compute  $P_i$  and correct  $P_i$  with BH criterion.
3. Exclude  $i$ s with adjusted  $P$ -values less than  $P_0$  as outliers.
4. Compute histogram,  $h_n$ , of  $P_i$  (with arbitrary bins).
5. Compute the standard deviation,  $\sigma_{h_n}$ , of  $h_n$ .
6. Update  $\sigma_{\ell_1}$  such that  $\sigma_{h_n}$  decreases (with arbitrary minimization algorithm).
7. Go back to the step 2 and repeat until we can find  $\sigma_{\ell_1}$  that enables  $\sigma_{h_n}$  to have minimum values.

The purpose of the above procedure is to find a set of  $i$ s whose associated  $P_i$ s fully obey Gaussian, since  $h_n$  should be constant, i.e.,  $\sigma_{h_n} = 0$ , if a set of  $i$ s is that associated  $P_i$ s fully obey Gaussian.

## 4. Integration of multiple profiles

### Shared conditions

Suppose that we have  $K$  multiple profiles  $x_{ijk} \in \mathbb{R}^{N_k \times M \times K}$  that represents the amount of value of  $i_k$ th feature of  $j$ th condition of  $k$ th profile. We compute matrices as  $x_{jj'k} = \sum_{i_k=1}^{N_k} x_{ijk} x_{i'jk} \in \mathbb{R}^{M \times M \times K}$  to which HOSVD is applied and we get

$$x_{jj'k} = \sum_{\ell_1=1}^M \sum_{\ell_2=1}^M \sum_{\ell_3=1}^K G(\ell_1 \ell_2 \ell_3) u_{\ell_1 j} u_{\ell_2 j'} u_{\ell_3 k} \quad (3)$$

After identifying  $u_{\ell_2 j}$  of interest,  $u_{\ell_2 j'}$  is recovered as  $u_{\ell_2 j'k} = \sum_{j=1}^M x_{ijk} u_{\ell_2 j}$ . The remaining procedure till feature selection is similar to the above.

### Shared features

Suppose that we have  $K$  multiple profiles  $x_{ijk} \in \mathbb{R}^{N \times M_k \times K}$  that represents the amount of value of  $i$ th feature of  $j_k$ th condition of  $k$ th profile. We compute matrices as  $x_{i'jk} = \sum_{j_k=1}^{M_k} x_{ijk} x_{i'jk} \in \mathbb{R}^{N \times N \times K}$  to which HOSVD is applied and we get

$$x_{i'jk} = \sum_{\ell_1=1}^N \sum_{\ell_2=1}^N \sum_{\ell_3=1}^K G(\ell_1 \ell_2 \ell_3) u_{\ell_1 i} u_{\ell_2 j} u_{\ell_3 k} \quad (4)$$

Missing singular value vectors attributed to conditions,  $u_{\ell_2 j}$ , can be recovered as  $u_{\ell_2 jk} = \sum_{i=1}^N x_{ijk} u_{\ell_1 i}$ . The remaining procedure till feature selection is similar to the above.

## 5. Integration of two profiles

### Shared conditions

Suppose that we have two profiles  $x_{ij} \in \mathbb{R}^{N \times M}$  and  $x_{kj} \in \mathbb{R}^{K \times M}$  that represents the values of  $i$ th and  $k$ th feature of  $j$ th condition, respectively. Generate tensor  $x_{ijk} \in \mathbb{R}^{N \times M \times K}$  as  $x_{ijk} = x_{ij} x_{kj}$  and apply the feature selection procedure as described above. Only modification is that we need to identify only one singular value vector,  $u_{\ell_2 j}$ , of interest whereas we need to identify two singular value vectors,  $u_{\ell_1 i}$  and  $u_{\ell_3 k}$  that have the largest absolute value of  $G(\ell_1 \ell_2 \ell_3)$  with fixed  $\ell_2$ .

### Reduction of required memory

Since  $N \times M \times K$  can be very large, we often ought to reduce the size of matrices. For this we take partial summation of  $x_{ijk}$  as  $x_{ik} = \sum_{j=1}^M x_{ijk}$  and singular value decomposition (SVD) was applied to  $x_{ik} \in \mathbb{R}^{N \times K}$  as

$$x_{ik} = \sum_{\ell=1}^L \lambda_{\ell} u_{\ell i} u_{\ell k} \quad (5)$$

and missing singular value vectors attributed to conditions,  $j$ s, can be recovered as  $u_{\ell j}^{(i)} = \sum_{i=1}^N x_{ij} u_{\ell i}$ ,  $u_{\ell j}^{(k)} = \sum_{k=1}^K x_{kj} u_{\ell k}$ . After identifying the singular value vector,  $u_{\ell j}^{(i)}$  or  $u_{\ell j}^{(k)}$ , of interest, the corresponding  $u_{\ell i}$  or  $u_{\ell k}$  is used to attribute  $P$ -values to  $i$ th or  $k$ th feature with eq. (2) (For  $k$ ,  $i$  must be replaced with  $k$ ). Features  $i$ s and  $k$ s with adjusted  $P$ -values less than threshold value are selected.

### Shared features

Suppose that we have two profiles  $x_{ij} \in \mathbb{R}^{N \times M}$  and  $x_{kj} \in \mathbb{R}^{N \times K}$  that represents the values of  $i$ th feature of  $j$ th and  $k$ th conditions, respectively. Generate tensor  $x_{ijk} \in \mathbb{R}^{N \times M \times K}$  as  $x_{ijk} = x_{ij} x_{ik}$  and apply the feature selection procedure as described above.

### Reduction of required memory

Since  $N \times M \times K$  can be very large, we often ought to reduce the size of matrices. For this we take partial summation of  $x_{ijk}$  as  $x_{jk} = \sum_{i=1}^N x_{ijk}$  and singular value decomposition (SVD) was applied to  $x_{jk} \in \mathbb{R}^{M \times K}$  as

$$x_{jk} = \sum_{\ell=1}^L \lambda_{\ell} u_{\ell j} u_{\ell k} \quad (6)$$

and missing singular value vectors attributed to features,  $i$ s, can be recovered as  $u_{\ell i}^{(j)} = \sum_{j=1}^M x_{ij} u_{\ell j}$ ,  $u_{\ell i}^{(k)} = \sum_{k=1}^K x_{ik} u_{\ell k}$ . After identifying the singular value vector,  $u_{\ell j}$  or  $u_{\ell k}$ , of interest, the corresponding  $u_{\ell i}^{(j)}$  or  $u_{\ell i}^{(k)}$  is used to attribute  $P$ -values to  $i$ th feature with eq. (2). Features  $i$ s with adjusted  $P$ -values less than threshold value are selected, although there are two distinct sets of  $i$ s selected dependent upon whether  $j$  or  $k$  is considered.

## 6. Integration of multiple profiles II

### Shared conditions

Suppose that we have  $K$  multiple profiles  $x_{ijk} \in \mathbb{R}^{N_k \times M \times K}$  that represents the amount of value of  $i_k$ th feature of  $j$ th condition of  $k$ th profile. We applied SVD to them as  $x_{ijk} = \sum_{\ell=1}^L \lambda_{\ell} u_{\ell i_k} u_{\ell jk}$ . HOSVD was applied to  $u_{\ell jk} \in \mathbb{R}^{L \times M \times K}$  as

$$u_{\ell jk} = \sum_{\ell_1=1}^L \sum_{j=1}^M \sum_{k=1}^K G(\ell_1 \ell_2 \ell_3) u_{\ell_1 \ell} u_{\ell_2 j} u_{\ell_3 k}. \quad (7)$$

Missing singular value vectors attributed to features can be recovered as  $u_{\ell_2 i_k} = \sum_{j=1}^M x_{ijk} u_{\ell_2 j}$  after identifying the  $u_{\ell_2 j}$  of interest. The remaining procedure till feature selection is similar to the above.

### Shared features

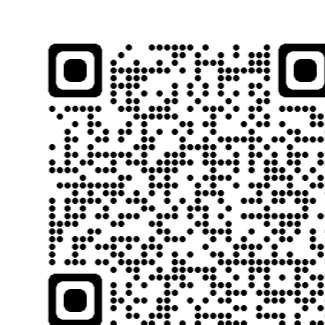
Suppose that we have  $K$  multiple profiles  $x_{ijk} \in \mathbb{R}^{N \times M_k \times K}$  that represents the amount of value of  $i$ th feature of  $j_k$ th condition of  $k$ th profile. We applied SVD to them as  $x_{ijk} = \sum_{\ell=1}^L \lambda_{\ell} u_{\ell i} u_{\ell jk}$ . HOSVD was applied to  $u_{\ell jk} \in \mathbb{R}^{N \times L \times K}$  as

$$u_{\ell jk} = \sum_{\ell_1=1}^L \sum_{i=1}^N \sum_{k=1}^K G(\ell_1 \ell_2 \ell_3) u_{\ell_1 \ell} u_{\ell_2 i} u_{\ell_3 k}. \quad (8)$$

Missing singular value vectors attributed to conditions can be recovered as  $u_{\ell_2 jk} = \sum_{i=1}^N x_{ijk} u_{\ell_2 i}$  then  $u_{\ell_2 j}$  of interest is selected. The remaining procedure till feature selection is similar to the above.

## 7. TDbasedUFE and TDbasedUFEadv

TDbasedUFE



TDbasedUFEadv



We released two bioconductor packages to perform the above analyses easily.

TDbasedUFE implemented "2. Feature selection procedure" and "4. Integration of multiple profiles" whereas TDbasedUFEadv implemented "5. Integration of two profiles" and "6. Integration of multiple profiles II", respectively. Both implemented "3. Optimization of standard deviation" as well.