Standard and Convex NMF in Clustering UCI wine and sonar data

ACS AISBITS 2016, October 19-21, Międzyzdroje.

Anna M. Bartkowiak

aba@cs.uni.wroc.pl

Institute of Computer Science, University of Wroclaw PL

Outline

Considered topics

1. Introduction: NMF, Non-negative Matrix Factorization
   Variants: the standard NMF and the convex NMF
   What-for can they serve?
   Can they serve for clustering?
2. Let's see their work for two data sets: 'wine' and 'sonar'
   What kind of data are these?
3. Methods used for analysis: Standard and Convex NMF and Exploratory Visualization
4. Analysis of the wine data
5. Analysis of the sonar data
6. Summary with discussion of the results

1. What is NMF

We consider a size $m \times n$ data matrix $\mathbf{X}$ composed from n column data vectors denoting objects or individuals:

$$\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \text{ with } \mathbf{x}_j \in I\!R^{m \times 1},\ j = 1, \ldots n.$$

Each data vector $\mathbf{x}_i$ has m elements (attributes of the objects). The same data vector $\mathbf{x}_i$ may be also viewed as a data point in the m-dimensional data space All elements of $\mathbf{X}$ are **are nonnegative**: $\mathbf{X} \geq 0$, which may be denoted as $\mathbf{X} \in I\!R_+^{m \times n}$.

The NMF method (short for Non-negative Matrix Factorization) provides for the given matrix $\mathbf{X}_{m \times n}$, and an assumed constant integer k ($k \leq min(m, n)$), a kind of **approximation** by lower rank matrices $\mathbf{A}$ and $\mathbf{H}$:

(1) $$\mathbf{X}_{m \times n} \approx \mathbf{A}_{m \times k} * \mathbf{H}_{k \times n},$$

where the factorizing matrices $\mathbf{A}$ and $\mathbf{H}$ should be also non-negative: $\mathbf{A}_+$ and $\mathbf{H}_+$.

1. What is NMF cnt.

The factor $\mathbf{A}$ of size $m \times k$ is called the *basis*. Its columns constitute basis vectors, and act as representatives (prototypes) of all the column vectors $\mathbf{x}_j,\ j = 1, \ldots, n$ contained in $\mathbf{X}$.
The factor $\mathbf{H}$ of size $k \times n$ is referred to as *encoder* (*coefficient*) matrix, and its columns provide coefficients permitting to approximate (reproduce, reconstruct) from the basis $\mathbf{A}$ the respective columns of $\mathbf{X}$. For example, to reconstruct the column vector $\mathbf{x}_j \in \mathbf{X}_{m \times n}$, one uses the formula:

(2) $$\mathbf{x}_j \approx \hat{\mathbf{x}}_j = A * \mathbf{h}_j, \quad j = 1, \ldots, n.$$

The entire $\mathbf{X}$ matrix is reconstructed as

(3) $$\mathbf{X} \approx [\hat{\mathbf{x}}_1, \ldots, \hat{\mathbf{x}}_n] = \hat{\mathbf{X}} = \mathbf{A} * [\mathbf{h}_1, \ldots, \mathbf{h}_n].$$

The basis $\mathbf{A}$ and the encoders $\mathbf{H}$ are derived by minimizing error of the approximation using the LS method or by minimizing the divergence (e.g. Kullback-Leibler) derived from quasi-likelihood of the distributions of $\mathbf{X}$ and $\hat{\mathbf{X}}$.

1. What is NMF cnt.

The main APPLICATIONS of NMF are:
- matrix decomposition and matrix approximation,
- reduction of dimensionality and data reduction,
- clustering of the data vectors $\mathbf{x}_j \in \mathbf{X}$

NMF has a multitude of applications, the number of them increases steadily

About CLUSTERING: It has been notified that
- the information on the clusters (group assignment) is in the encoders $\mathbf{H}$
- basic vectors are similar to group centroids obtained by k-means and spectral clustering algorithm
- sparseness of $\mathbf{A}$ and $\mathbf{H}$ may lead to better results of classification

Special algorithms were devised – by adding sparseness constraints – to construct sparse NMF.

1. What is NMF

about sparseness:
The sparseness do not work for every data.
Increasing sparseness may increase approximation error.

Our main concern in this paper is: **What about the clustering? Can NMF do it?** IT HAPPENS THAT THE OBTAINED CLUSTERS DO NOT CONFORM WITH OUR EXPECTATIONS

Let's see, how NMF is working for some chosen data

Let's look into the multivariate data space and at the clusters therein

2. Two data sets

Both are benchmark data for clustering
Both are downloadable from UCI Data Repository

Wine

- n=168 data vectors, each characterized by m variables (attributes)
- are known to come from 3 vineyards constituting 3 groups (clusters):
- no.s 1–59 from vineyard A, no.s 60–130 from from vineyard B, and no.s 131-168 from vineyard C.

Sonar

- n=208 data vectors, each characterized by 60 variables (attributes)
- data and pattern recognition problem described by Gorman and Sejnowski (1988),
- are known belonging to 2 groups: 'Rock' and 'Mine'
- no.s 1–97 $\in$ group 'Rock'; no.s 98–208 $\in$ group 'Mine'

Our main concern is:
WILL NMF RECOVER THE INDICATED GROUPS AS DISTINCT CLUSTERS?

2. The data sets. Tasks to perform

To find out, what's going on, we will concentrate on three tasks:

1. To find out – for the two data sets – how affine are the respective NMF bases H_{wine} or H_{sonar} to the corresponding K-means centroids?
2. How close are the group assignment obtained from the NMF encoders given in the matrices H_{wine} and H_{sonar} to the respective group assignments obtained by the corresponding K-means results?
3. Is it possible to recognize by the considered techniques the vineyard of the wine data vectors or the rock-or-mine object of the sonar data?

3. Methods used for analysis

FOR COMPUTING NMF AND K-MEANS:

- Ordinary NMF with LS estimation method
- Convex NMF with LS estimation method
- kmeans

For Multivariate visualization of data vectors:

- Kohonen Self-Organizing Maps
- t-distributed stochastic embedding
- Canonical discriminant functions obtained by CDA.

3. Methods used for analysis, cont.

For a given data set, the methods will be carried out in the following order

(a) Exploratory data visualization using Kohonen SOM, t-sne and Canonical discriminant analysis CDA,

(b) Standard and Convex NMF,

(c) Clustering by K-means.

Let us state clearly, that all the used in our analysis methods – except canonical discriminant functions – work in an unsupervised mode . The group colors in the presented figures were added *ex post*, after obtaining the results/projections.

3. Methods used for analysis, cont.

The **Standard NMF** considers the model

(4) $$\mathbf{X}_+ \approx \mathbf{A}_+ * \mathbf{H}_+$$

It represents data vectors as a linear combination of spatially localized components, called parts-base representation.
The factors **A** and **H** may be obtained from the optimization criterion

(5) $$\min_{A,H} \frac{1}{2}||\mathbf{X} - \mathbf{AH}||_{\mathbf{F}}^{\mathbf{2}} \quad \text{s.t.} \quad \mathbf{A}, \mathbf{H} \geq \mathbf{0}.$$

NMF has been successfully applied among others in pattern recognition,image processing, object detection in images, classification and clustering.
However it does not always result in "parts-based" representation, and the classification and clustering may be problematic.
Supervised methods may be helpful in establishing or confirmation of the obtained clusters.

3. Methods used for analysis, cont.

The **Sparse NMF** uses additionally a sparseness constraint added to the optimization criterion.
Hoyer's proposal (n denotes the length of vector **x**):

(6) $$sparseness(\mathbf{x}) = \frac{\sqrt{n} - (\sum |x_i|)/\sqrt{\sum x_i^2}}{\sqrt{n} - 1}$$

Proposal by Kim and Park:

(7) $$\min_{A,H} \frac{1}{2}||\mathbf{X} - \mathbf{AH}||_{\mathbf{F}}^{\mathbf{2}} + \frac{\eta}{\mathbf{2}}||\mathbf{A}||_{\mathbf{F}}^{\mathbf{2}} + \frac{\lambda}{\mathbf{2}}\sum_{\mathbf{j=1}}^{\mathbf{n}}||\mathbf{h}_j||_1^2, \text{ s.t. } A, H \geq 0.$$

Li and Ngom use

(8) $$\min_{A,H} \frac{1}{2}||\mathbf{X} - \mathbf{AH}||_{\mathbf{F}}^{\mathbf{2}} + \lambda\sum_{\mathbf{j=1}}^{\mathbf{n}}||\mathbf{h}_j||_1, \text{ s.t. } H \geq 0, \ ||\mathbf{a}_r||_2^2 = 1 \forall_r.$$

3. Methods used for analysis, cont.

The **Convex NMF** modifies expression for the basis **A**:

(9) $$\mathbf{X}_{\pm} \approx \mathbf{A}_{\pm} * \mathbf{H}_{+}, \text{ where } \mathbf{A}_{\pm} = \mathbf{X}_{\pm} * \mathbf{W}_{+}.$$

Thus the basis $\mathbf{A}_{\pm}$ is supposed to be constructed as

(10) $$\mathbf{A}_{\pm} = \mathbf{X}_{\pm} * \mathbf{W}_{+}.$$

In the Convex NMF model both the data matrix **X** and the factor **A** are allowed to have mixed sign elements.
Moreover, the basis **A** is supposed to be constructed as linear combination of the column vectors from **X**.
This means that the basis is spanned by the data vectors $\mathbf{x}_j$ from **X**, and as such it belongs to the same space. The linear combinations spanning the basis **A** are designated by the matrix $\mathbf{W}_{+}$ of size $n \times k$.

3. Methods used for analysis, cont.

EXPLORATORY DATA VISUALIZATION

The principles of **Kohonen's self-organizing maps (SOM)s**. Free *Matlab SOM Toolbox* implemented by Vesanto et al.
The t-sne is based on the preservation of the similarities $\mathcal{S}$ among objects (d.v.'s) in a high-dimensional space and translation them to lower dimensional (e.g. 2D space) with preserving the stated in higher dimensions similarities $\mathcal{S}$.
The similarities are evaluated as by considering stochastic neighborhoods in both spaces. Algorithm is rather complicated.
Software: t-sne Matlab function implemented by van der Maaten
The **canonical discriminant analysis** is known also as *Linear discriminant analysis*, eventually with further explanation 'using Fisher's between group and within group scatter ratio'. For calculations own Matlab function was used.

The WINE data set (UCI)

size = 178 x 13

n = 178 = 59 + 71 + 47 instances

k= 3 groups

Results of analysis shown in 6 figures below

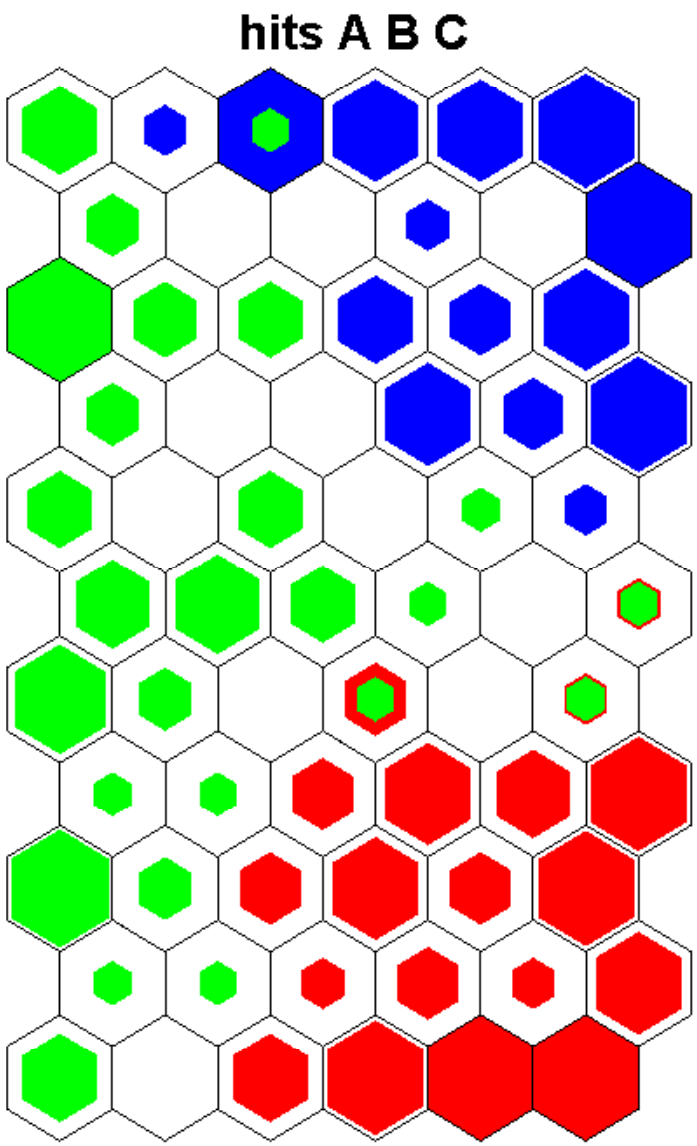
4. Analysis of the 'wine' data

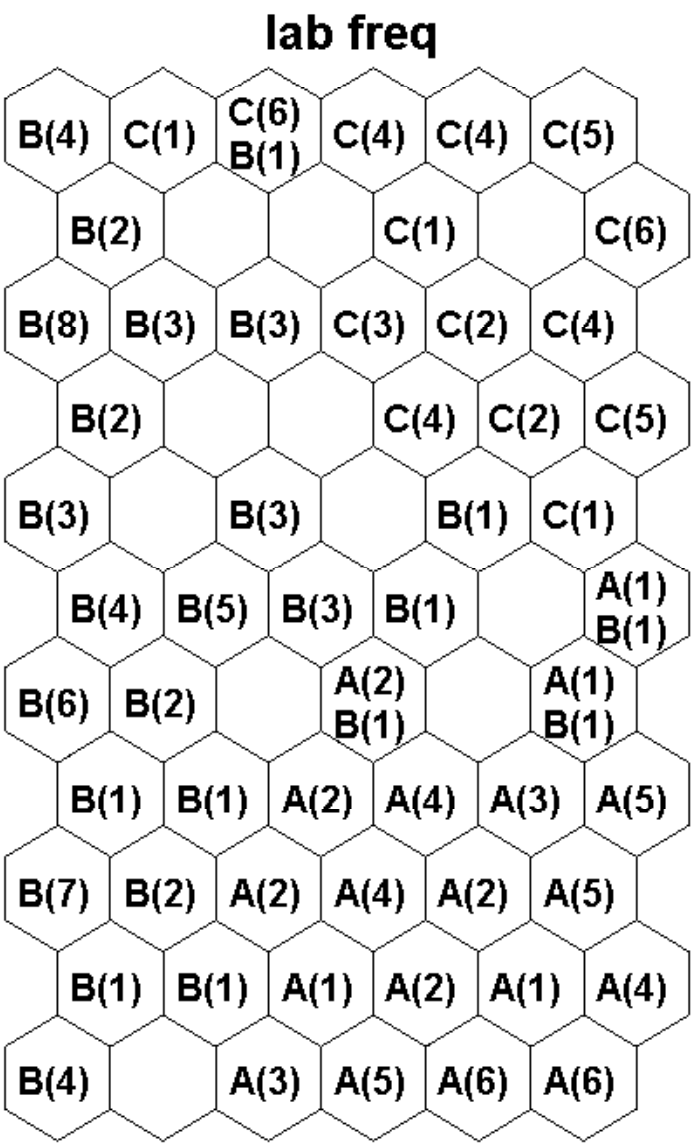
We proceed here along the points (a), (b) and (c) announced in slide 10 above.

`(a) Exploratory data visualization`

We visualize the data using Kohonens' self-organizing maps, t-distributed stochastic neighbors embedding, and canonical discriminants variates. For the exploratory data analysis the data matrix $\mathbf{X}^T$ was standardized to mean equal zero and std=1 for all 13 variables.
The obtained graphical visualizations are shown in fig:1 and fig:2

4. Analysis of the ’wine’ data (a) [fig:1]





4. Analysis of the ’wine’ data (a) fig:1

Fig. 1. caption
Visualization of the wine data using self-organizing maps. Left: Som-hits into the hexagons of the map – by data vectors belonging to 3 subgroups of the wine data. The number of hits into each hexagon (its frequency) is shown by the magnitude of the painted (in red, green or blue) area inside each hexagon. Right: The same, as at left, however now the frequency of hits is printed directly as integer number.

4. Analysis of the 'wine' data (a) fig:2

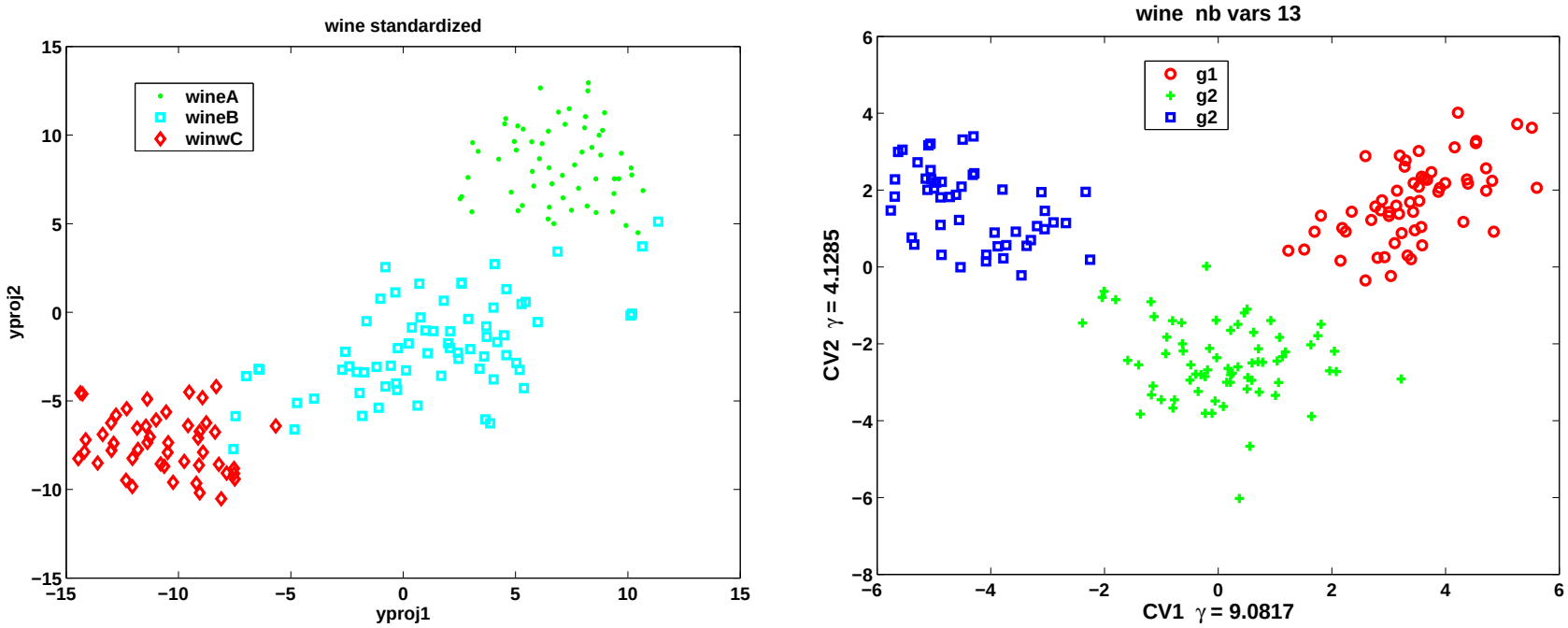


Fig. 2. Caption
Wine. Left: Visualization using **tsne**. Points from the three groups appear practically separated. Right: Visualization using canonical discriminant functions. The 3 groups appear completely separated

4. Analysis of the ’wine’ data (b) fig:3

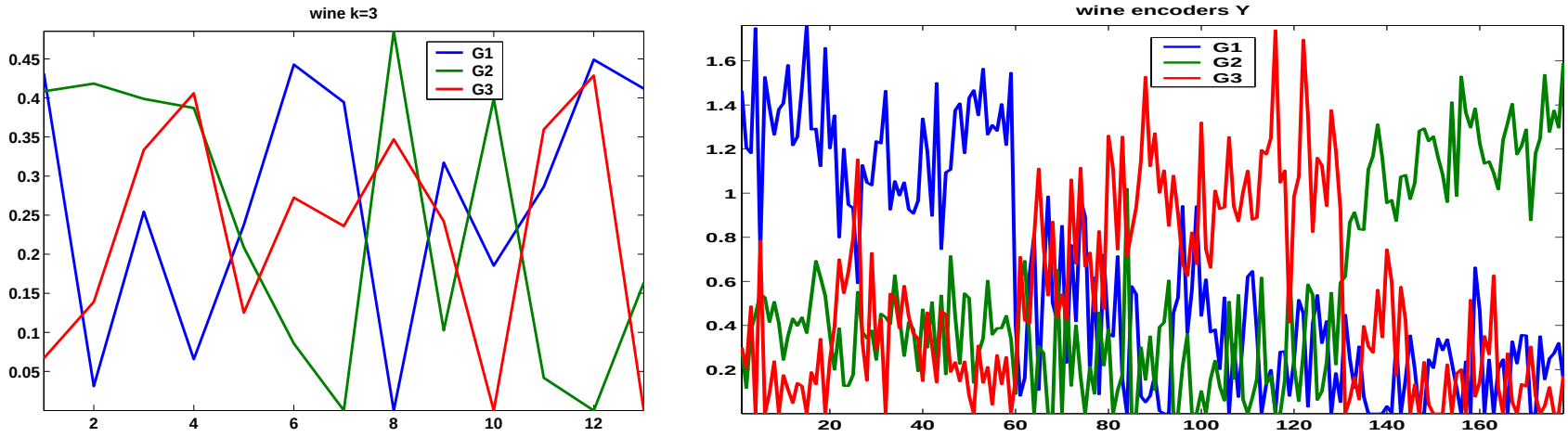


Fig. 3. caption
Profiles of basis vectors/columns in A(left) and profiles of group encoders in H displayed row-wise (right) obtained by the **standard NMF** algorithm

4. Analysis of the 'wine' data (b) fig:4

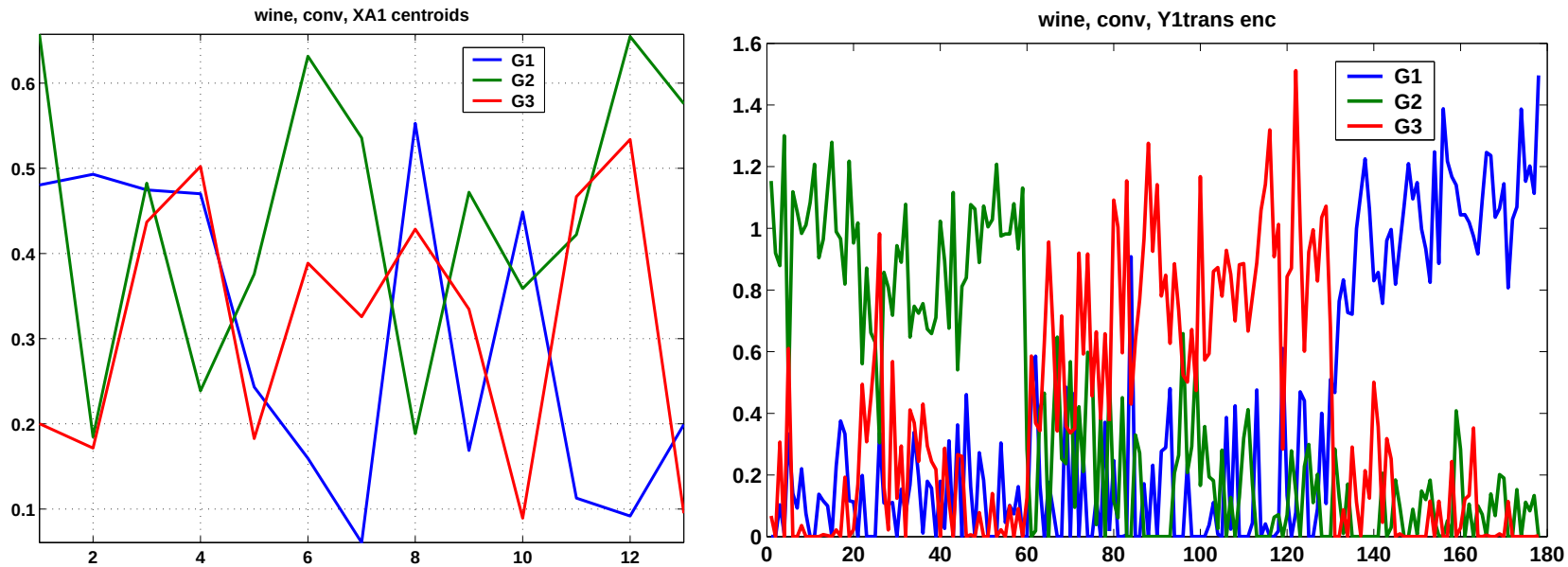
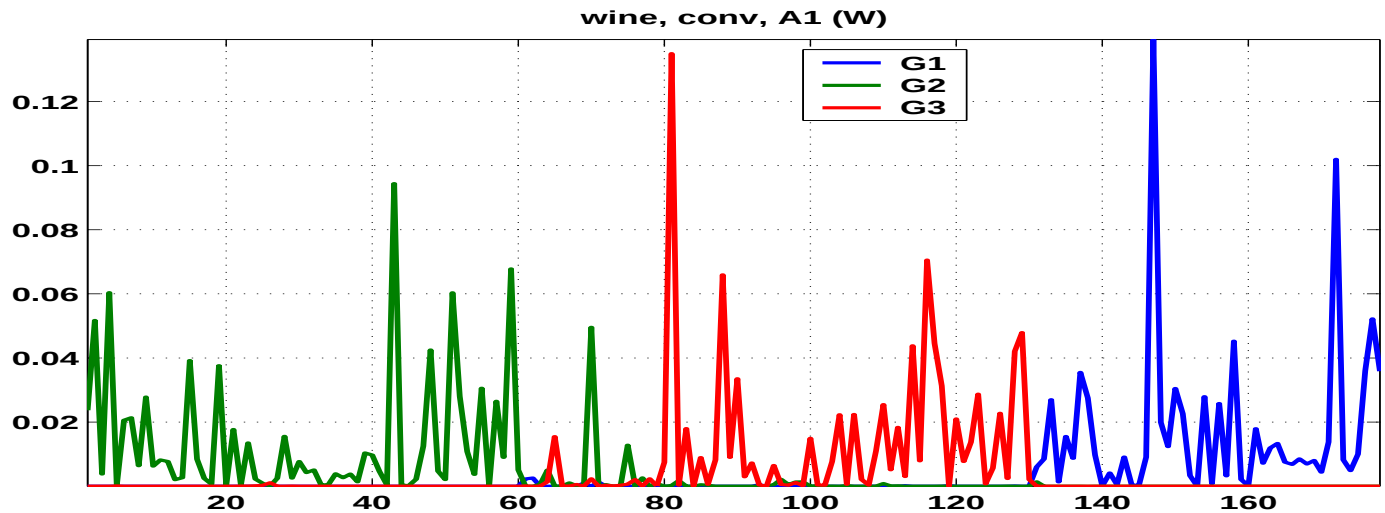


Fig. 4. Caption
Wine by **convex** NMF. Left: Profiles of basis vectors/columns in A size 13×3 evaluated as $A_{13x3} = X_{13x178} * W_{178x3}$. Right: Profile of encoders H size 3×178 displayed row-wise.

4. Analysis of the 'wine' data (b) fig:5

Fig. 5. caption
Factorization using **convex** NMF. Profile of **weighting matrix W** size 178×3 entering into the formula A = X*W. . Notice that the matrix is very sparse. The sparseness conforms with the groups of the data.



Group contents in the displayed indexplot: G1: 1–59, G2: 60–130, G3: 101–178

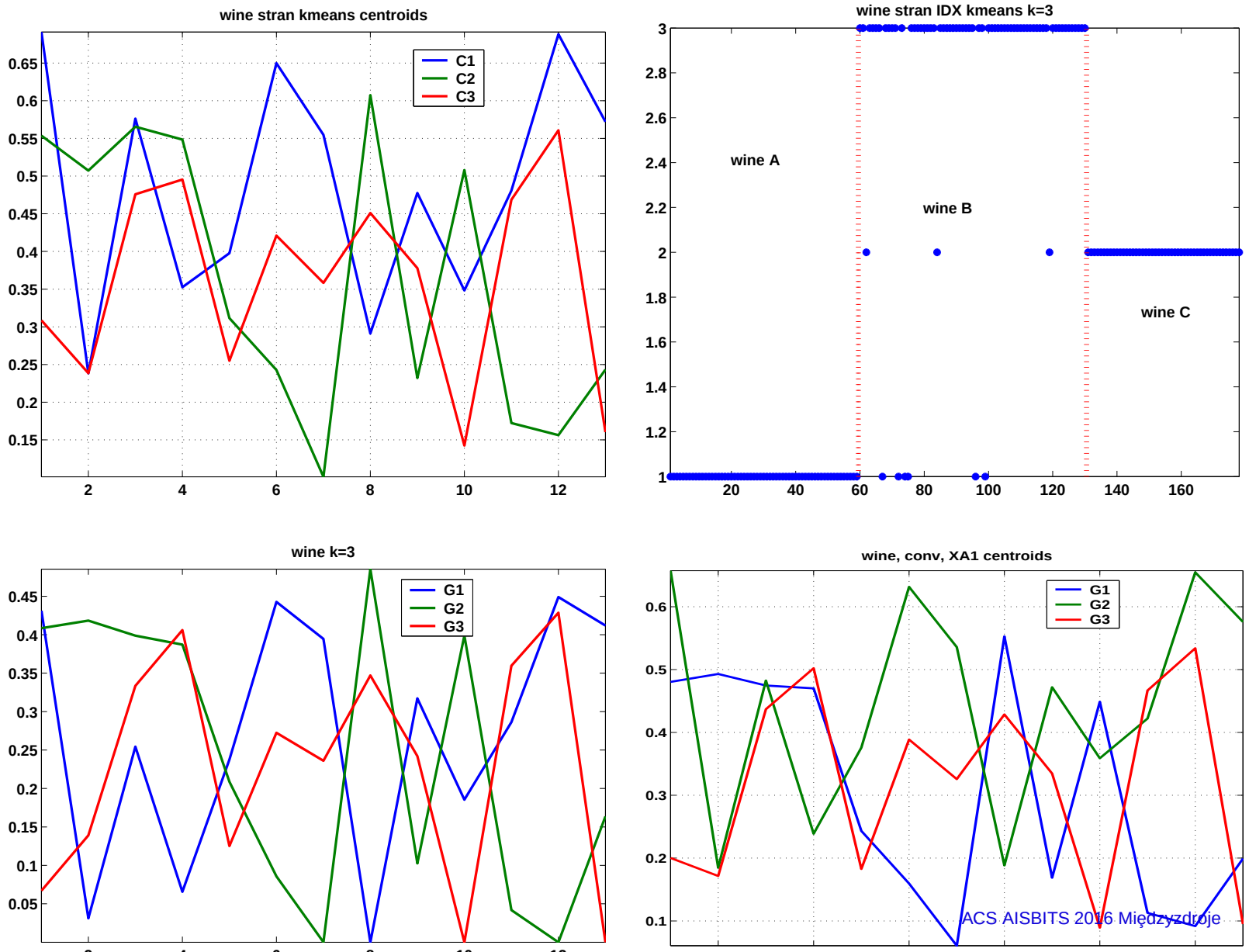
4. Analysis of the 'wine' data (c)

$\Downarrow$

Fig. 6. caption

Top: Results of **k-means** for k=3. Left: Profiles of the three centroids. Right: Group assignment by k-means. Bottom: Corresponding bases from NMF. Left: The A basis from ordinary NMF. Right: The $A = XW$ basis from convex NMF.

4. Analysis of the ’wine’ data (c) fig:6



SONAR data set (UCI)

size = 208 x 60

n= 208 = 97 + 111 instances

k = 2 groups

Analyzed similarly as wine

Results shown in Figs 7 – 12

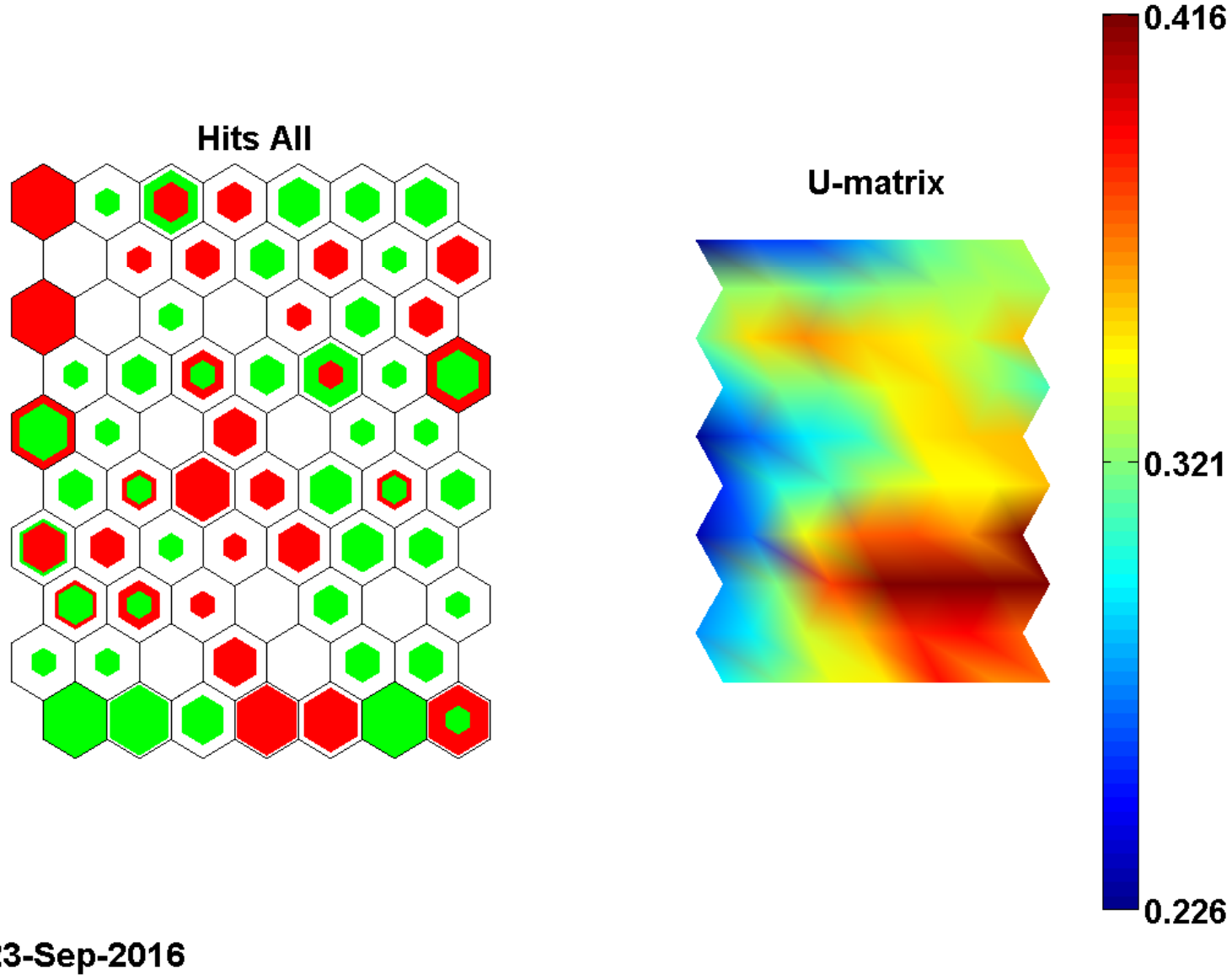
5. Analysis of the ’sonar’ data

We proceed here along the points (a), (b) and (c) announced in Section 3.
The data were obtained from spectrograms and as parameters characterizing power spectral densities they are by their nature non-negative. They are reals contained in the interval [0, 1].

`(a) Exploratory data visualization`

We visualize the data using Kohonens’ self-organizing maps, t-distributed stochastic neighbors embedding, and canonical discriminant variates.
The obtained graphical visualizations are shown in figs:7 and figs:8

5. Analysis of the ’sonar’ data (a) fig:7



5. Analysis of the ’sonar’ data (a) fig:7

Fig. 7. caption
Visualization of the sonar data using **self-organizing maps**. Left: SOM-hits of the two groups of data ’rock’ and ’mine’ marked by colors red and green. The topology of the data space where the sonar data vectors reside.

5. Analysis of the 'sonar' data (a) fig:8

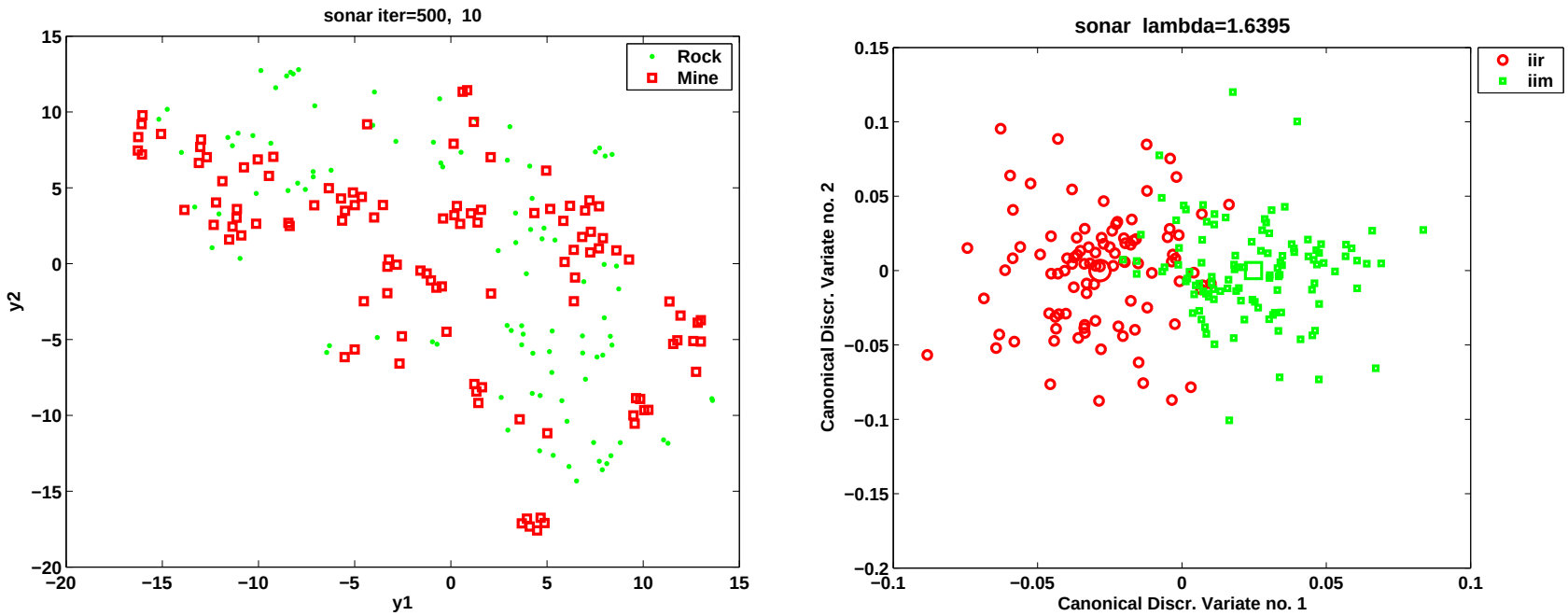


Fig. 8. Caption

Sonar data. Left: Visualization using **tsne**. Points from the two groups are intermixed. Right: Visualization using **canonical discriminant functions**. Is much better, as at left. The two groups are in large part separated.

5. Analysis of the 'sonar' data (b) fig:9

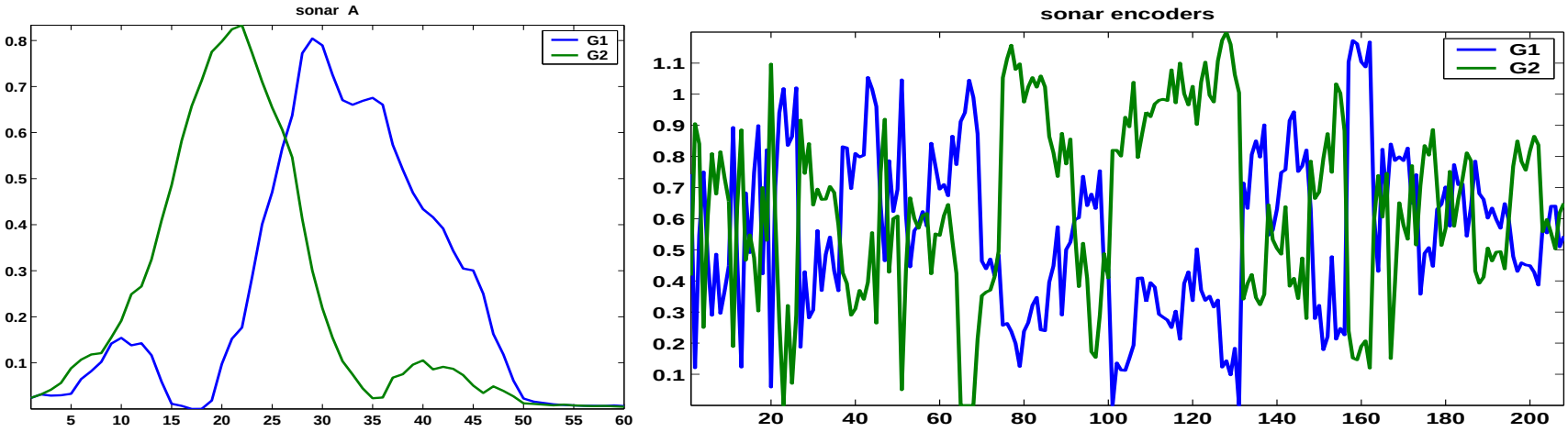


Fig. 9. Caption
Sonar with **standard nmf**. Profiles of basic vectors (left) and and profiles of group encoders (right)

5. Analysis of the 'sonar' data (b) fig:10

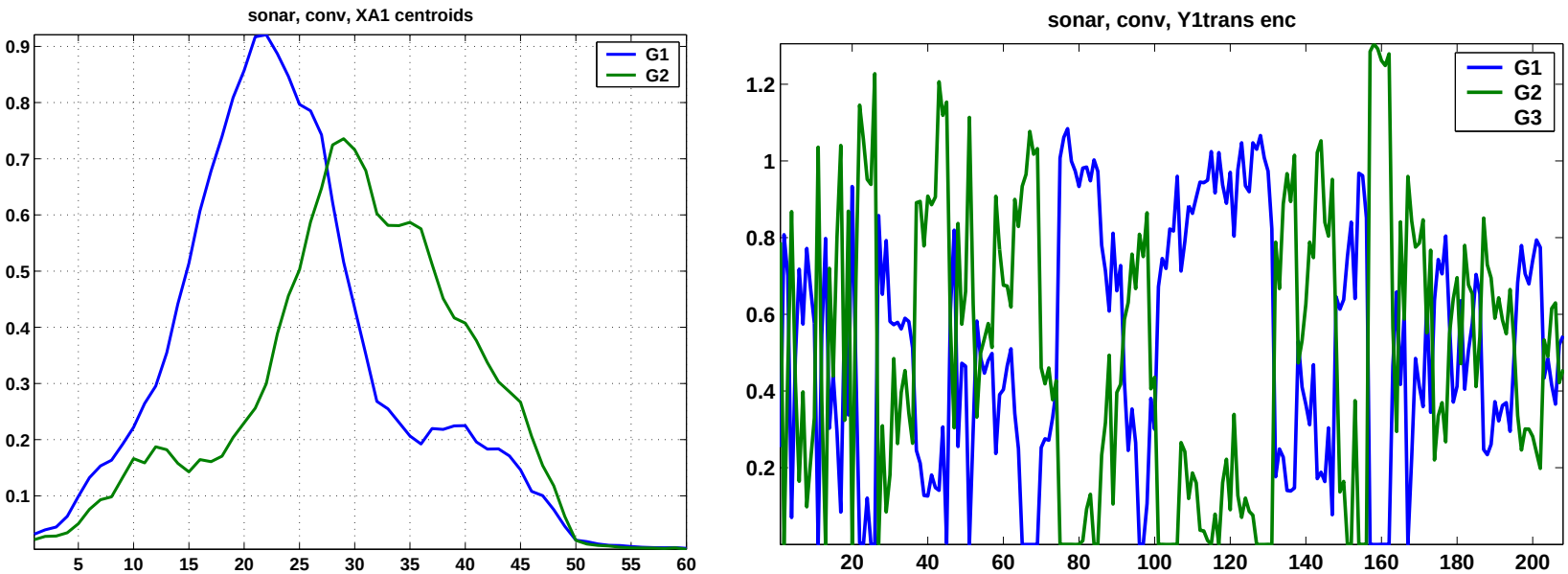
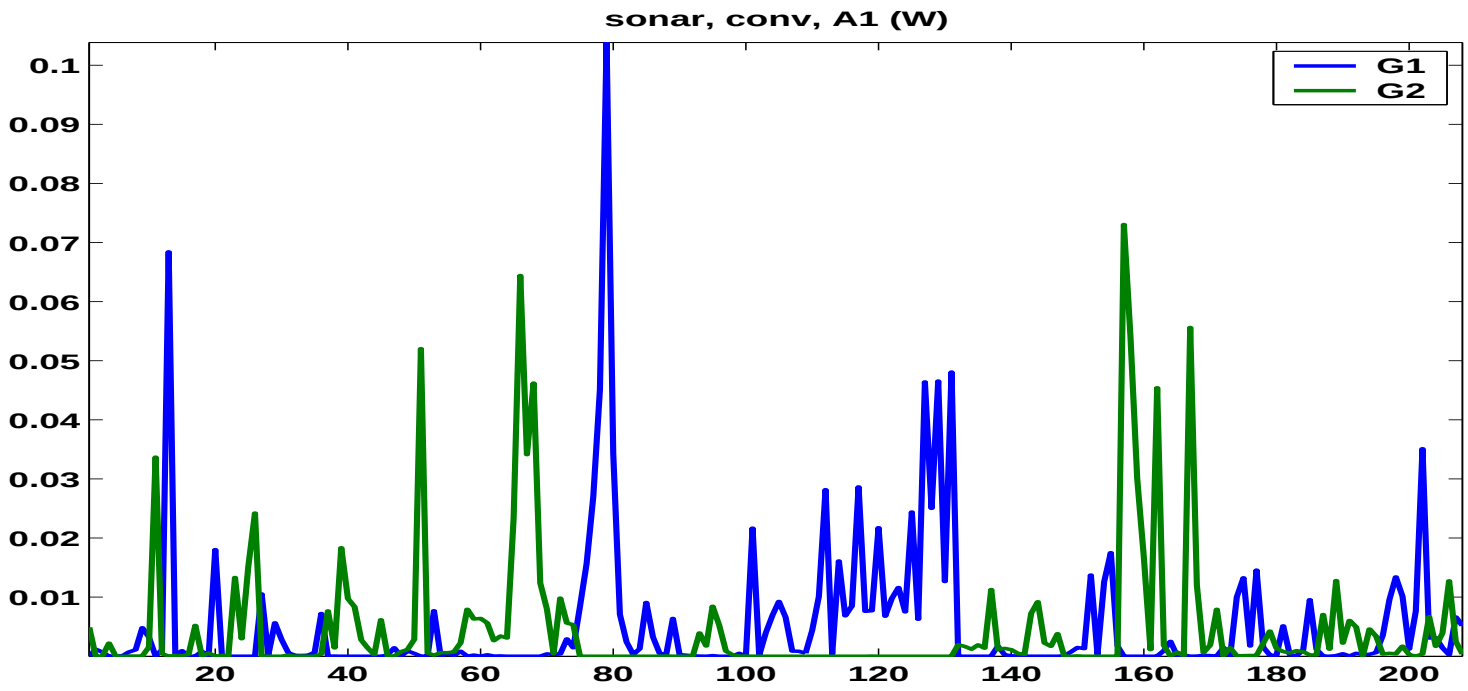


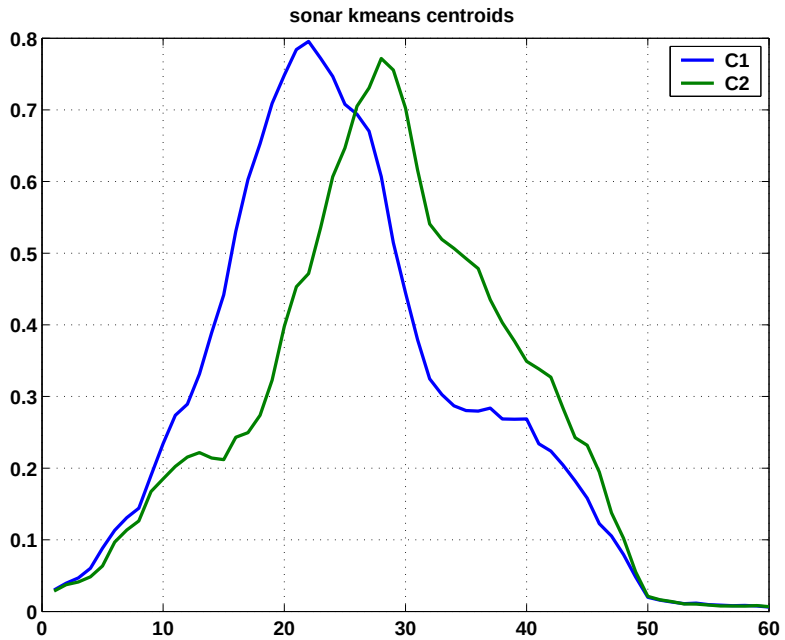
Fig. 10. Caption
Sonar with convex nmf. Profiles of Basic vectors (left) and and profiles of group encoders (right)

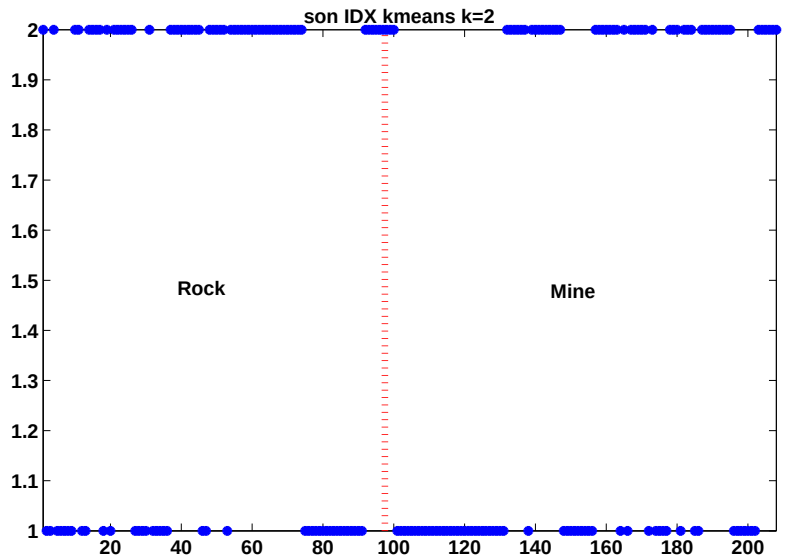
5. Analysis of the 'sonar' data (b) fig:11

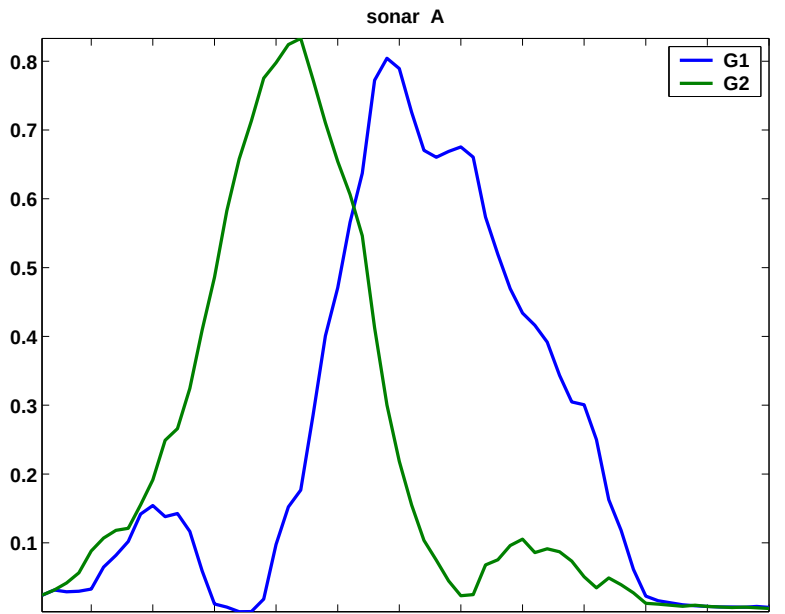
Fig. 5. caption
Sonar. Convex nmf. Profile of **weighting matrix matrix W** contributing for constructing the basis A=XW. size $m \times r = 60 \times 2$

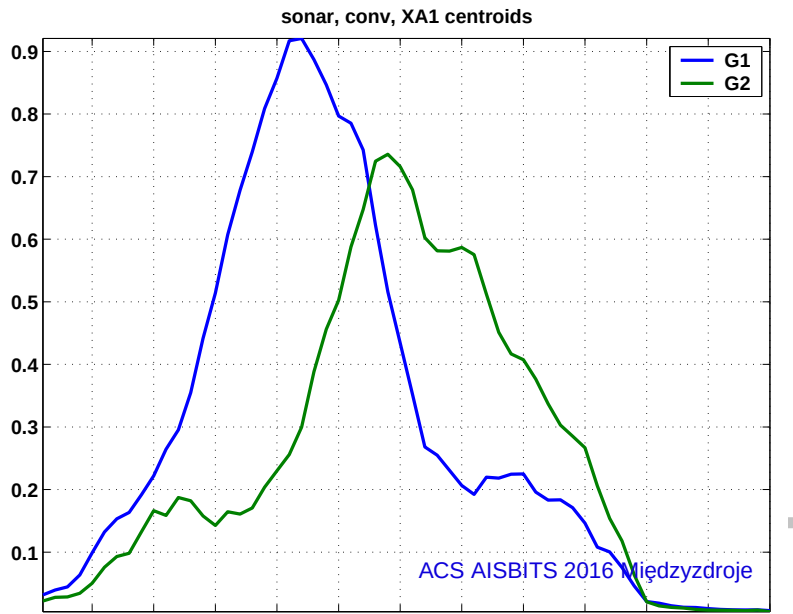


5. Analysis of the ’sonar’ data (c) fig.12.









6. Discussion fig.13. mesh X

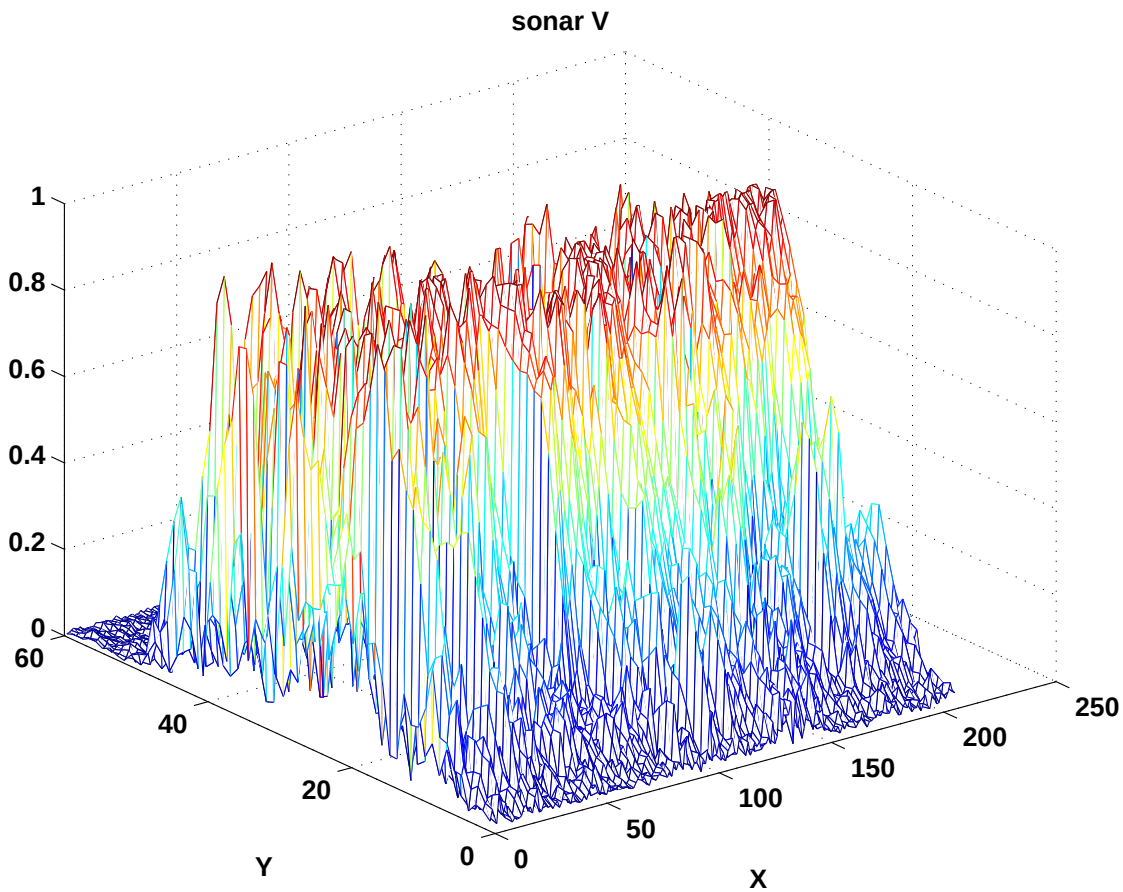


Fig. 13. SONAR. A holistic view of analyzed data matrices using the Matlab mesh plot function: The **original data matrix** $\mathbf{X}$ of size m=60 n=208; Group no.s: G1: 1–97, G2: 98–208

6. Discussion fig.14. mesh

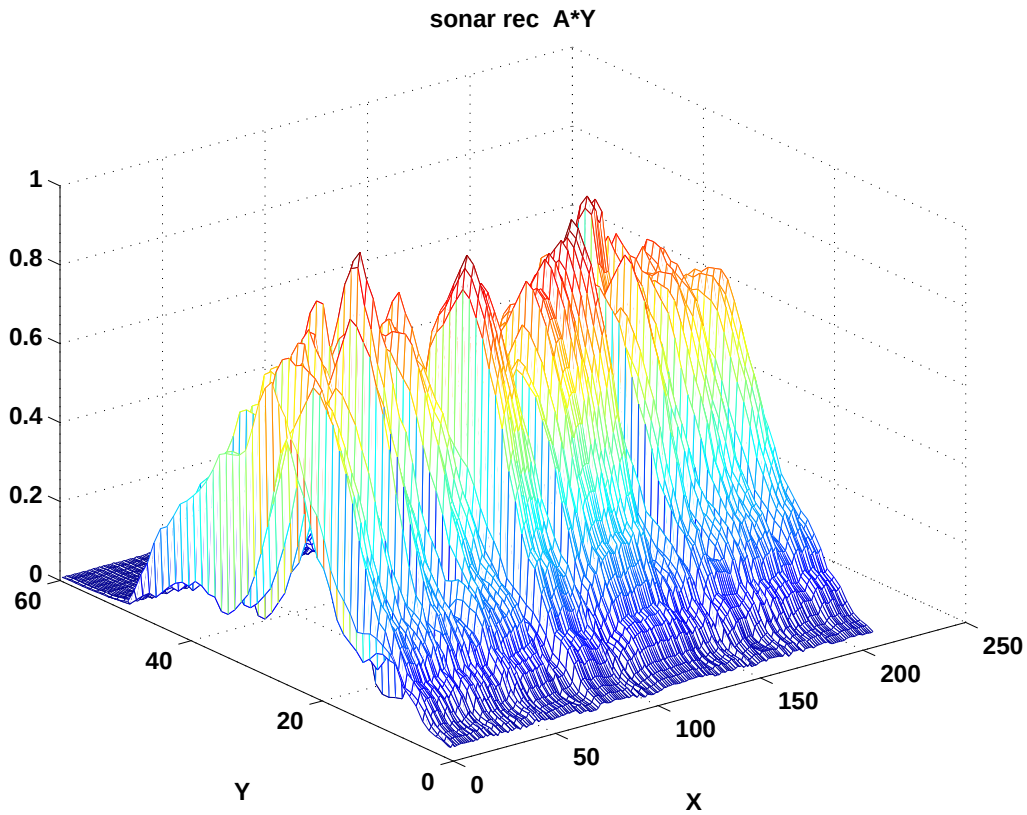


Fig. 14. SONAR. Data matrix **XRec reconstructed from standard NMF** rank 2 factorizing

6. Discussion fig.15. mesh

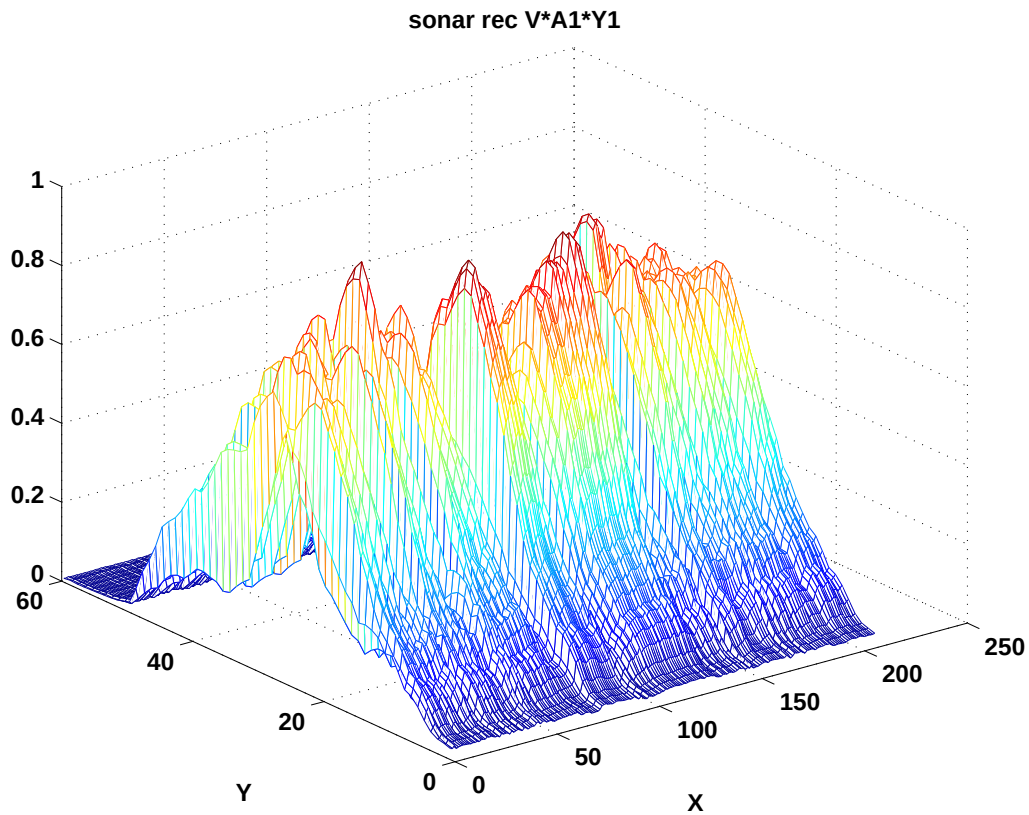


Fig. 15. SONAR. Data matrix **XRec reconstructed from convex NMF** rank 2 factorizing

References

SeungNat Lee D.D., Seung H.S.: Learning the parts of objects by nonnegative matrix factorization. Nature 401, 788–791, 1999.

DingJord Ding C., Li T., Jordan MI: Convex and semi-nonnegative matrix factorizations. IEEE Trans. on Pattern Analysis and Machine Intelligence (TPAMI) 32, pp. 45-55, 2010.

Cichocki A. Cichocki, R. Zdunek, A.H. Phan, Sh. Amari: Nonnegative matrix and tensor factorizations. Applications to exploratory multi-way data analysis and blind source separation. Wiley, Chichester U.K. 2009.

Lachen Lachenbruch P.: Discriminant Analysis, Hafner Press, London 1975.

Koh95 Kohonen T.: Self-organizing Maps, Springer, Heidelberg, 1995, 1997, Third Extended Edition 2001, 501 pages.

MaaHinton van der Maaten L., Hinton G., Visualizing data using t-SNE, Journal of Machine Learning Research, 1, pp. 1-48, 2008.

Thanks for attention !