On Clustering Daily Mobile Phone Density Profiles

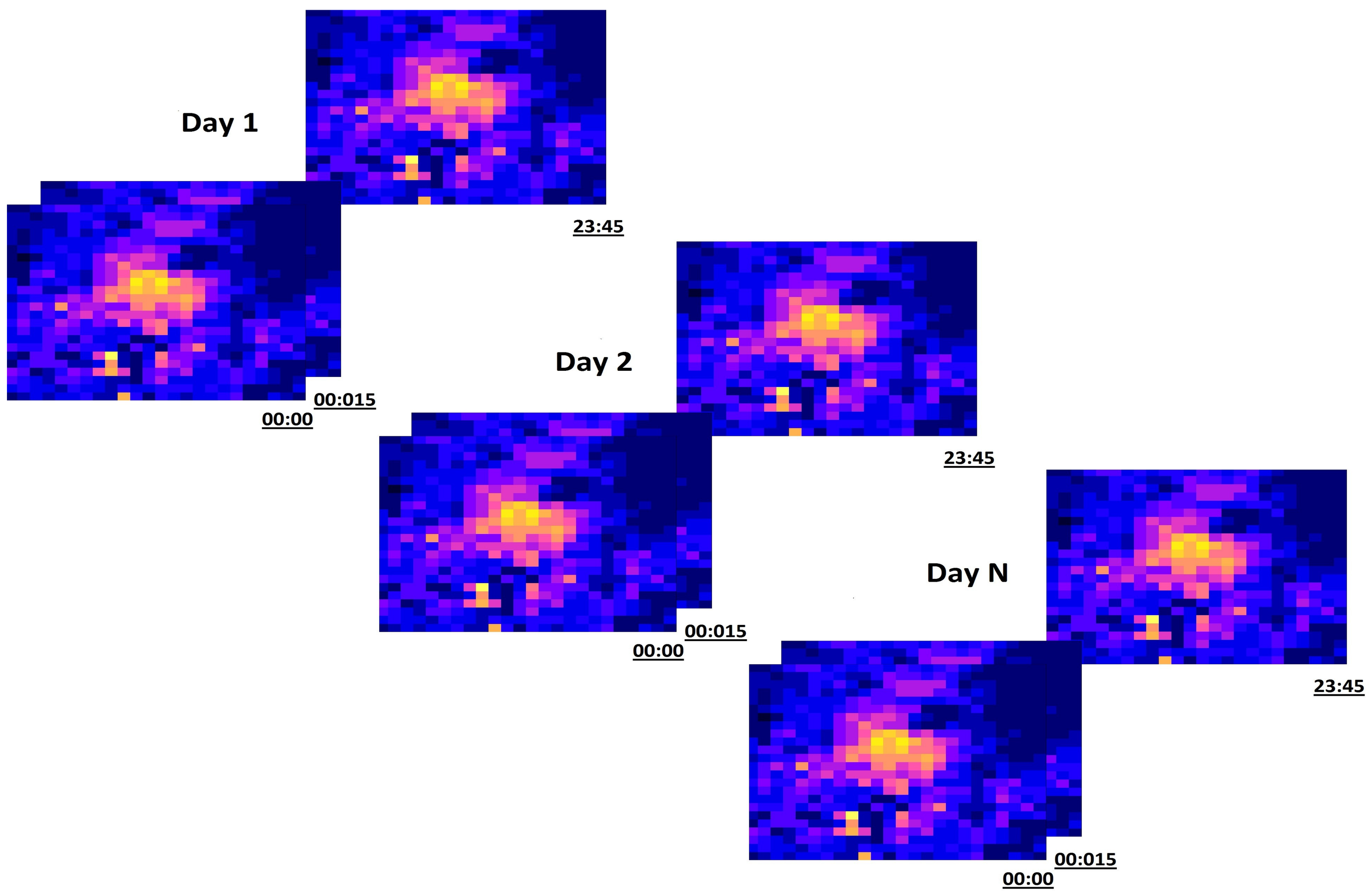
Rodolfo Metulini (Univ. Brescia, Italy) - rodolfo.metulini@unibs.it

Maurizio Carpita (Univ. Brescia, Italy) - maurizio.carpita@unibs.it

DMS StatLab
Data Methods and Systems Statistical Laboratory
Department of Economics and Management
University of Brescia

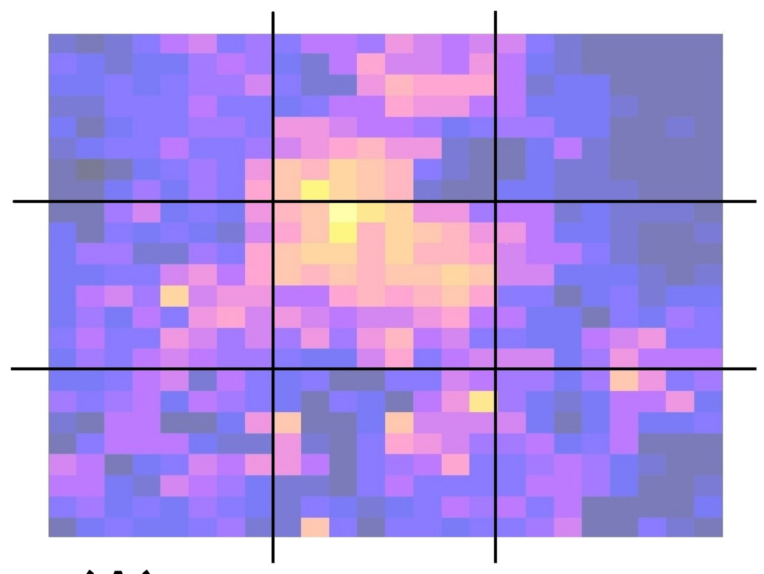
1. Context & Objective

Daily Mobile Phone Density Profiles (DMPDPs) are characterized by a 2-D spatial component (i.e. the cells of the grid) and by a temporal component (i.e. the cell has repeated values in time, for a total of 96 daily dimensions per cell).



The **Aim** is to find **regularities** and to detect **anomalies** in the flow of people's presences, by clustering similar daily profiles.

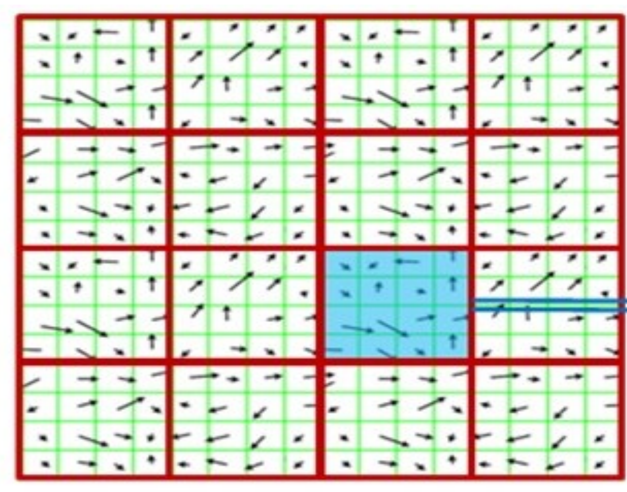
2. The Approach



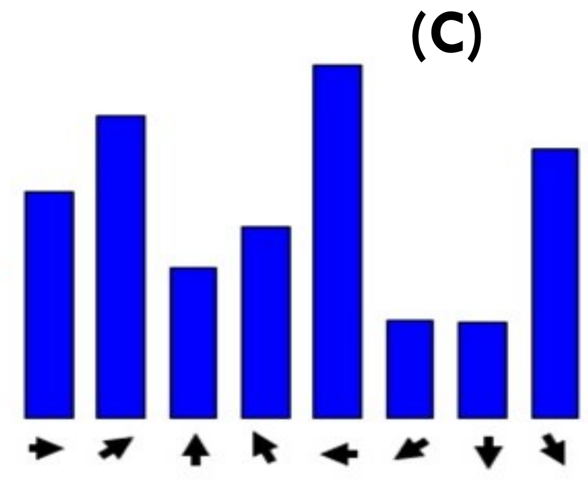
(A)

Step 1: reducing the spatial dimension (2-D to 1-D)

For each quarter (**Q**), considering the grid as a RGB color image spanning in [0,1], divide the image into **c x c** smaller grids (**fig A**).



(B)



(C)

For each grid, compute oriented gradients (**fig B**);

setting number of bins, compute the **histogram of the oriented gradients** (HOG) (**fig C**);

stack into a vector the **h** HOG values of the 96 quarters of the same day, producing the matrix **X** (**fig D**)

Q	HOG	Day 1	Day 2	..	Day N
1	1	X1_1,1	X2_1,1	..	XN_1,1
1	2	X1_1,2	X2_1,2	..	XN_1,2
1	..	..	..	..	..
1	h	X1_1,h	X2_1,h	..	XN_1,h
..	..	..	..	..	..
96	h	X1_96,h	X2_96,h	..	XN_96,h

(D)

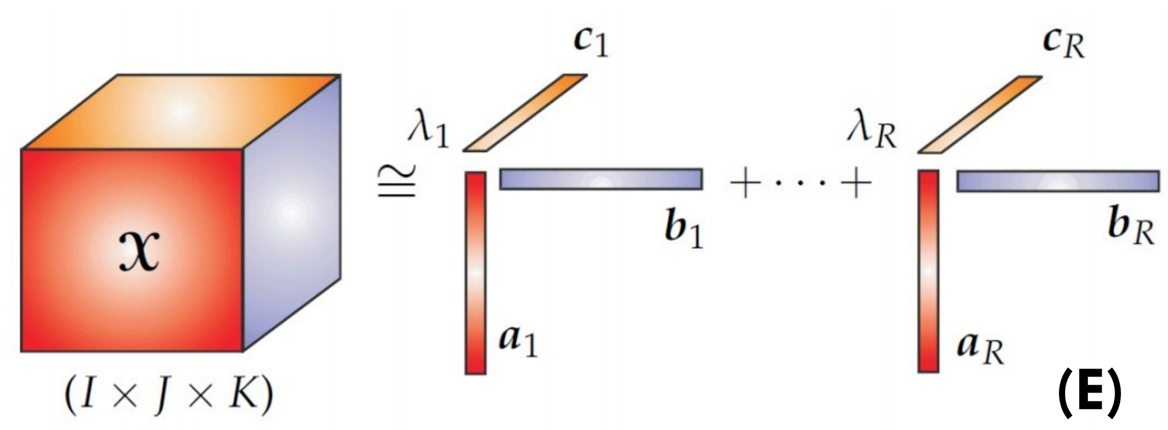
Step 2: grouping daily profiles

Apply an high dimensional **cluster analysis** to group days (**X**'s columns, objects) in terms of the HOG features (**X**'s rows, variables)

Step 3: detecting trends & outliers

For each group, consider the 3D array with dimensions ***a*** (**quarters**), ***b*** (**days**) and **c** (**space**, HOG values);

estimate the Canonical polyadic (CP) **tensor decomposition** (CANDECOMP/PARAFAC, **fig E**)



(E)

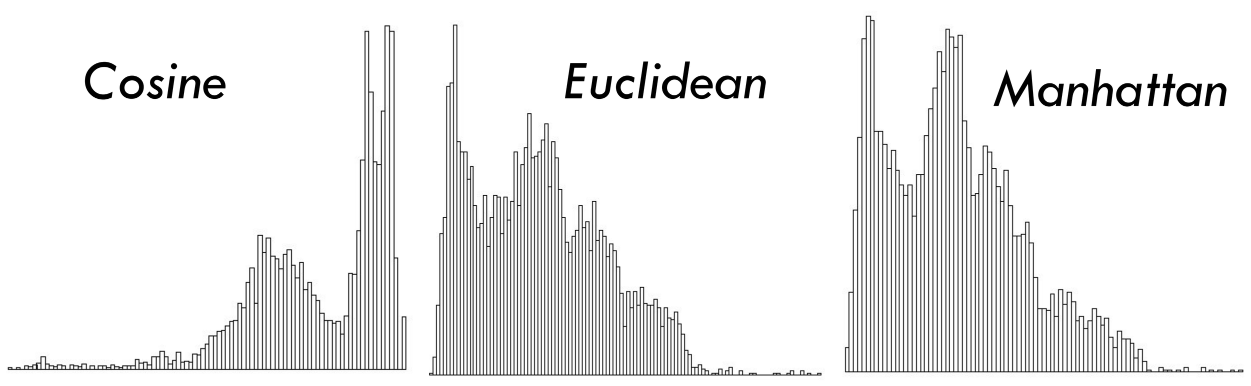
$$X = \sum_{r=1}^{R} \lambda_r * a_r \circ b_r \circ c_r$$

3. Application & Results

We select the grids of the **city of Brescia** (lat/long [10.2, 10.24, 45.52, 45.55], dim 24 x 24), from **March 18th to June 30th, 2015**.

We extract HOG features by dividing each grid into **9 8x8 cells**.
In each cell, gradients has been computed and **5 bins** have been selected to compute the histogram.
Each grid counts for **45 HOG features,** with a dimensionality reduction in the order of 576/45 = **12.8**.
Stacking in the same column all the quarters of the same day, the matrix **X** counts for **4320 variables** and **105 objects** (days).

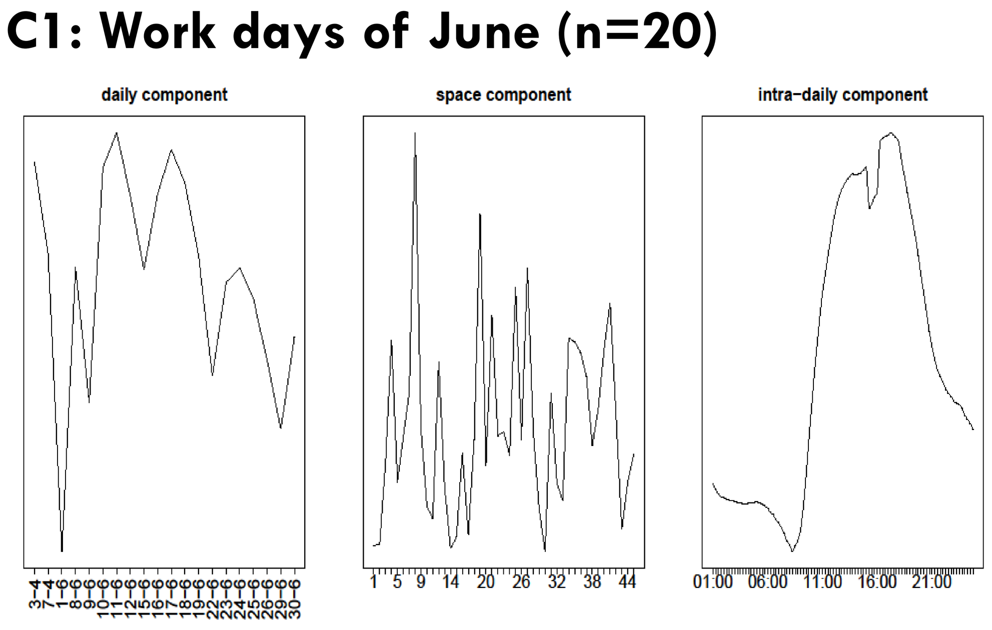
We apply a cluster analysis using k-means and k-medoids with *Manhattan*, *Euclidean* and *Cosine* distance.
The **curse of dimensionalty** does not subsist.



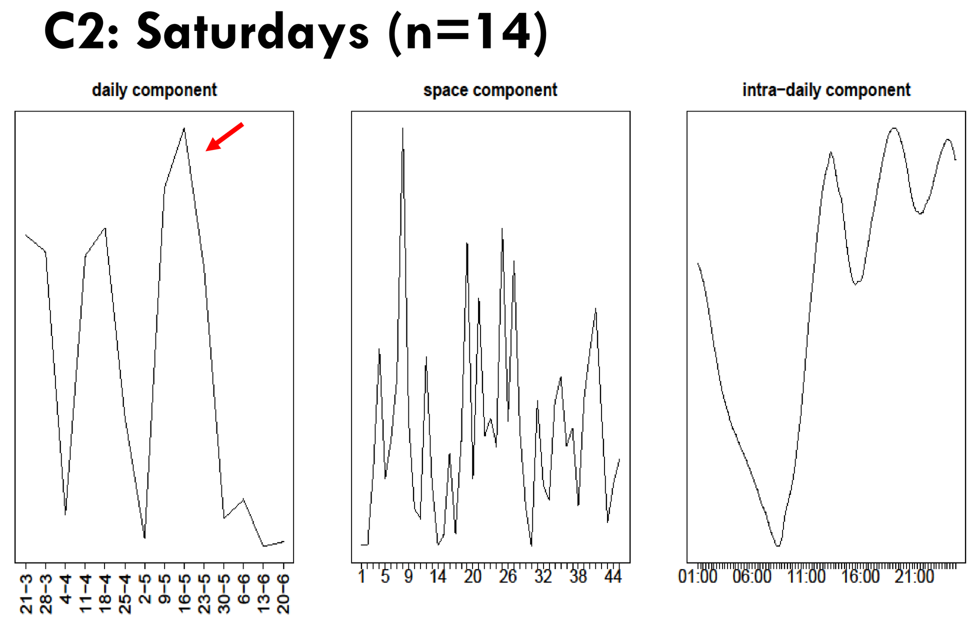
Bi-modal pair distances distributions

For each cluster, we plot the first tensor (**r=1**) component to display regularities and outliers.

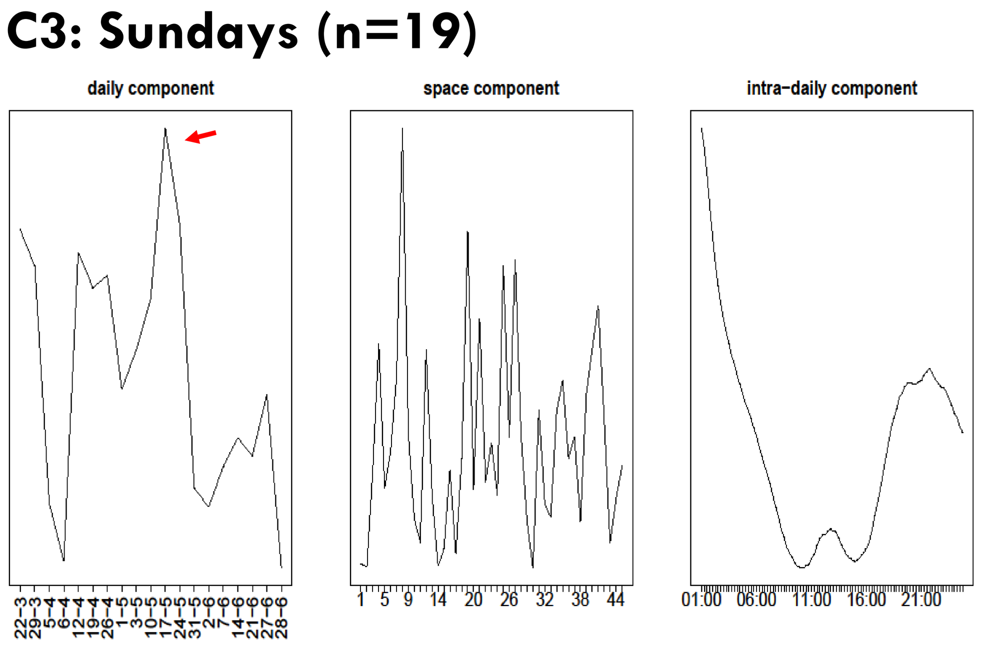
C1: Work days of June (n=20)



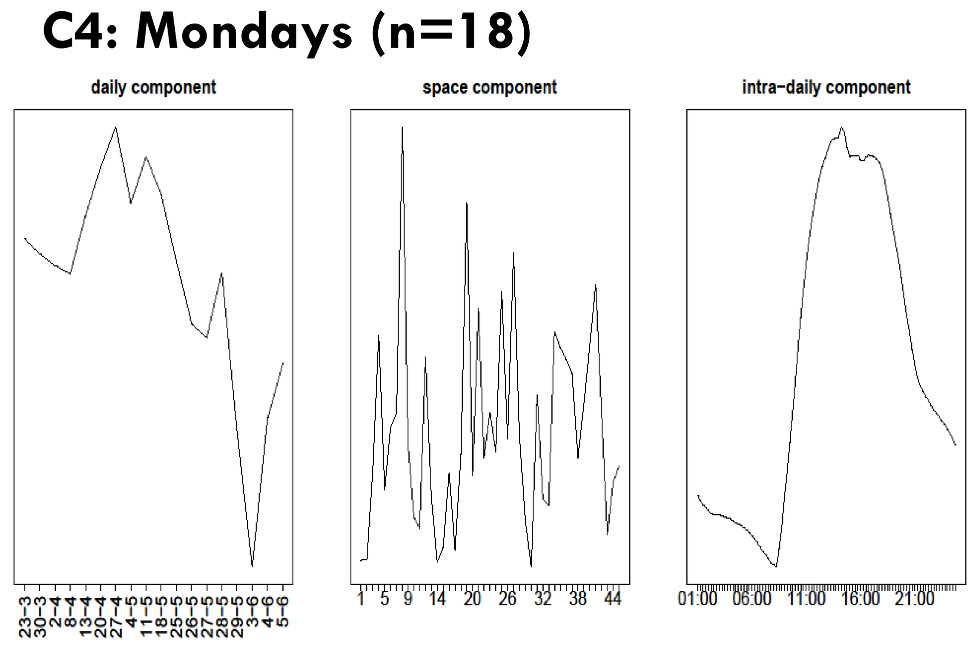
C2: Saturdays (n=14)



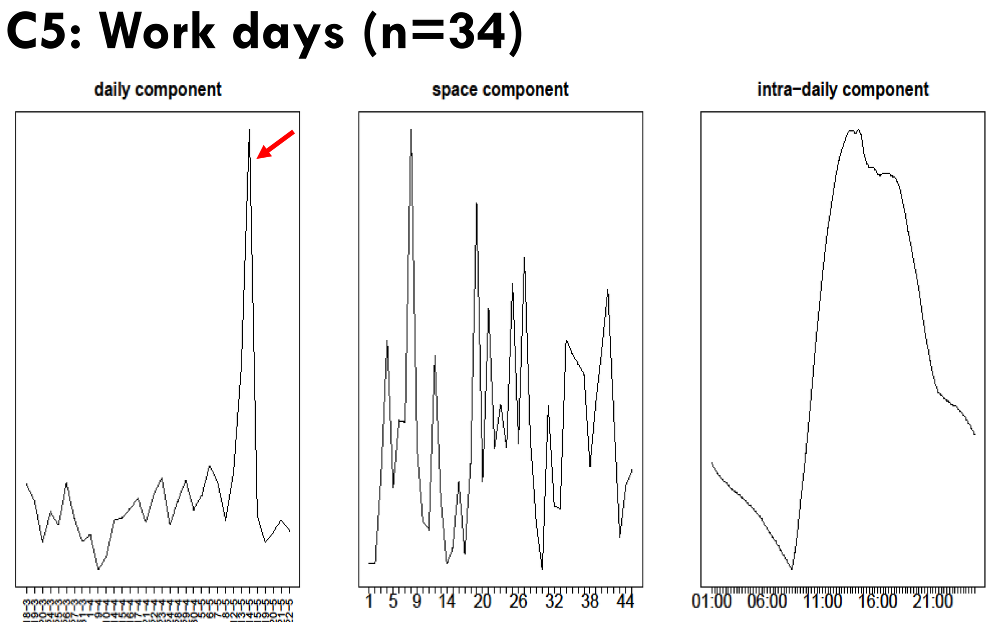
C3: Sundays (n=19)



C4: Mondays (n=18)



C5: Work days (n=34)



References

1. Carpita, M., Simonetto, A. (2014). Big Data to Monitor Big Social Events: Analysing mobile phone signals in the Brescia Smart City. Electronic Journal of Applied Statistical Analysis: Decision Support Systems, Volume 5, Issue 1, pp. 31-41
2. Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(4), 340-350.
3. Tomasi, C. (2012). Histograms of oriented gradients. Computer Vision Sampler, 1-6.
4. Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455-500.