HOW TO ANALYZE MILLIONS OF SPECTRAL PROFILES OF AN X1.0 FLARE EFFICIENTLY ...

...AND STILL FEEL GOOD AFTERWARD

A. Sainz Dalda\textsuperscript{1} & L. Kleint\textsuperscript{2}

1. Stanford-Lockheed Institute for Space Research, Stanford, CA USA
2. Fachhochschule Nordwestschweiz (FHNW), Windisch, Switzerland
WHAT DO WE HAVE?...

- We observed an **X1.0 flare** with an unprecedented number of instruments and thus we have **more than 1 million spectra every minute** in the spectral lines Fe I 6302 Å, Ca II 8542 Å, Hα 6563 Å, He I 10830 Å, and Mg II h & k.

- That data come from different instruments: DST/IBIS, DST/FIRS, Hinode-SOT/SP, IRIS.

- Full-polarimetry spectral data in the **photosphere** and the **chromosphere**. Intensity spectral data in the **chromosphere** and the **transition region**.
MOTIVATION

- If Facebook, Google, Amazon...make money using Machine Learning techniques, can we use them for making...papers?

- YES! But carefully!!!

- They have been successfully used in Solar Physics:

IN THIS PRESENTATION

- k-means classification: how does work? Pros & cons? Do we need PCA? Why?

- Application of k-means classification to spectral(polarimetric) data.
GOAL:
Cluster the data into $k$ groups where $k$ is predefined.

ALGORITHM (Steinhaus, 1957; McQueen, 1967; Lloyd, 1957)
0. Select randomly $k$ points as cluster centers or centroids.

Repeat steps 1 and 2 until the same points are assigned to each cluster in consecutive rounds

{ 
1. Assign elements to their closest cluster center according to the Euclidean distance function.

2. Calculate the centroid or mean of all objects in each cluster.

}
WHAT k-MEANS CODE DOES

The objective of k-means clustering is to minimize total within-cluster variance, or, the sum squared error function:

\[ J = \sum_{j=1}^{k} \sum_{i=1}^{n} \| x_{i}^{(j)} - c_{j} \|^2 \]
WHAT k-MEANS CODE DOES
WHAT A k-MEANS CODE DOES

K=2

Random Init

Final k-mean

Random Init

Final k-mean

HEIGHT

WEIGHT

IAUS 305 – Polarimetry: The Sun to Stars and Stellar Environments
Punta Leona, 30 Nov – 5 Dec, 2014
WHAT A k-MEANS CODE DOES

K=3
WHAT A k-MEANS CODE DOES

K=4

HEIGHT

WEIGHT

S

M

XL

L
# OTHER CLUSTERING METHODS


### A comparison of the clustering algorithms in scikit-learn

<table>
<thead>
<tr>
<th>Method name</th>
<th>Parameters</th>
<th>Scalability</th>
<th>Use case</th>
<th>Geometry (metric used)</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>number of clusters</td>
<td>Very large, medium n_samples, medium n_clusters with MiniBatch code</td>
<td>General-purpose, even cluster size, flat geometry, not too many clusters</td>
<td>Distances between points</td>
</tr>
<tr>
<td>Affinity propagation</td>
<td>damping, sample preference</td>
<td>Not scalable with n_samples</td>
<td>Many clusters, uneven cluster size, non-flat geometry</td>
<td>Graph distance (e.g. nearest-neighbor graph)</td>
</tr>
<tr>
<td>Mean-shift</td>
<td>bandwidth</td>
<td>Not scalable with n_samples</td>
<td>Many clusters, uneven cluster size, non-flat geometry</td>
<td>Distances between points</td>
</tr>
<tr>
<td>Spectral clustering</td>
<td>number of clusters</td>
<td>Medium n_samples, small n_clusters</td>
<td>Few clusters, even cluster size, non-flat geometry</td>
<td>Graph distance (e.g. nearest-neighbor graph)</td>
</tr>
<tr>
<td>Ward hierarchical</td>
<td>number of clusters</td>
<td>Large n_samples and n_clusters</td>
<td>Many clusters, possibly connectivity constraints</td>
<td>Distances between points</td>
</tr>
<tr>
<td>Agglomerative clustering</td>
<td>number of clusters, linkage type, distance</td>
<td>Large n_samples and n_clusters</td>
<td>Many clusters, possibly connectivity constraints, non Euclidean distances</td>
<td>Any pairwise distance</td>
</tr>
<tr>
<td>DBSCAN</td>
<td>neighborhood size</td>
<td>Very large n_samples, medium n_clusters</td>
<td>Non-flat geometry, uneven cluster sizes</td>
<td>Distances between nearest points</td>
</tr>
<tr>
<td>Gaussian mixtures</td>
<td>many</td>
<td>Not scalable</td>
<td>Flat geometry, good for density estimation</td>
<td>Mahalanobis distances to centers</td>
</tr>
</tbody>
</table>
Hinode-SOT/SP

Full-polarimetry at Fe I 6301 & 6302 Å
2 large maps observed before the flare happened
446976 x 112-vector ➞ 71068 x 112-vector

IAUS 305 – Polarimetry: The Sun to Stars and Stellar Environments
Punta Leona, 30 Nov – 5 Dec, 2014
**ADVICE:** Feed your K-means code with healthy data, e.g.: we have just used normalized Stokes V profiles with mean ($S_{lobe}$) $> \sigma$.

- **Binary Mask**
- **Masked Stokes V map**
- **Masked Stokes I map**

63% of 71068 x 112-vector $\rightarrow$ 44772 x 112
30 clusters $\rightarrow$ 30 k-means Stokes V-like profiles

**My code:** It takes about 5min

**IDL code:** It takes < 2min, 2 LINES of CODE !!!

```idl
wc=clust_wts(data_sel, n_clusters=k_clusters, n_iterations=50, /doub)
aidl=cluster(data_sel, wc, n_c=k_clusters)
```
Hinode-SOT/SP

K=10

k-mean cluster identifier number

Number of elements in the cluster

% of elements in the cluster with respect to the total of analyzed elements in the dataset
<table>
<thead>
<tr>
<th>Arcsec</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>293</td>
<td>0.7%</td>
</tr>
<tr>
<td>344</td>
<td>0.9%</td>
</tr>
<tr>
<td>350</td>
<td>0.9%</td>
</tr>
<tr>
<td>604</td>
<td>1.5%</td>
</tr>
<tr>
<td>693</td>
<td>1.8%</td>
</tr>
<tr>
<td>293</td>
<td>1.9%</td>
</tr>
<tr>
<td>782</td>
<td>2.0%</td>
</tr>
<tr>
<td>825</td>
<td>2.1%</td>
</tr>
<tr>
<td>1019</td>
<td>2.6%</td>
</tr>
<tr>
<td>1049</td>
<td>2.7%</td>
</tr>
<tr>
<td>00</td>
<td>6.2%</td>
</tr>
<tr>
<td>2379</td>
<td>6.2%</td>
</tr>
<tr>
<td>1100</td>
<td>2.8%</td>
</tr>
<tr>
<td>1712</td>
<td>4.5%</td>
</tr>
<tr>
<td>1714</td>
<td>4.5%</td>
</tr>
<tr>
<td>1880</td>
<td>4.9%</td>
</tr>
<tr>
<td>2379</td>
<td>6.2%</td>
</tr>
<tr>
<td>2709</td>
<td>7.1%</td>
</tr>
<tr>
<td>3747</td>
<td>9.8%</td>
</tr>
<tr>
<td>4134</td>
<td>10.1%</td>
</tr>
<tr>
<td>4999</td>
<td>13.0%</td>
</tr>
<tr>
<td>5630</td>
<td>14.5%</td>
</tr>
</tbody>
</table>
Hinode-SOT/SP

V strong asymmetries and multi/single-lobes -> 20.1%
Hinode-SOT/SP

- Analysis of the k-mean profiles in the region where one of ribbon of the X1 flare appears (~ 20 min later).

- Two components atmosphere, i.e.: inversions based on Milne-Eddington approximation will not reproduce successfully the physics.

- One magnetic component is showing strong downflows, while the other one is static and with opposite polarity.

- In the footpoints of the ribbon, the blue-only single-lobe Stokes V profiles could be explained with two static magnetic components, one of them with a gradient in $B_{\text{los}}$. 
IAUS 305 – Polarimetry: The Sun to Stars and Stellar Environments
Punta Leona, 30 Nov – 5 Dec, 2014
DST/FIRS

- Analysis of the k-mean profiles in the region where one of the ribbons of the X1 flare appears, exactly **during the maximum** of the flare!

- The LOS magnetic field component in the photosphere is opposite to the one in the chromosphere.

- How to explain the emission above the continuum of He I 10830 Å during the flare in the solar disk, i.e. in Non-LTE optically thick conditions? Ideas are very welcome!
Lucia has applied k-means algorithm to the Ca II 8542 Å.

After analyzing the k-mean profiles we realized they need different multi-Gaussian fitting and bi-sector approach.

We created dopplergrams from the k-mean profiles.
DST/IBIS

- Dopplergrams based on the K-mean clusters are spatially coherent

- Lucia says: “the dopplergrams in the chromosphere are a mess!”

- Alberto says: “maybe the lower chromosphere is a mess itself during a flare!!”

- k-means algorithm says: “Hey, guys! I just do my job!!!”
We have applied the k-means algorithm to the IRIS data.

The analysis can be done either in the same raster (i.e. in 8 slit positions) or in a fix position evolving in time.

We have selected Mg II k line in a fix slit position at 7 takes during the flare.
IRIS

DO WE NEED PCA?

- IRIS NUV Mg II k & h profiles span for 500 px. We have investigated the reduction of their dimensions using Principal Components Analysis.

- Due to the high variability of the profiles, the reduction factor might be from 5 to 20, i.e.: 500 to 100, 500 to 25.
Do we need PCA?

Mg II k

Mg II h

# of Coefs.: 5 --- Retained Variance: 0.991356
# of Coef.: 7 --- Retained Variance: 0.995995
Do we need PCA?
# of Coefs.: 26 --- Retained Variance: 0.999511
DO WE NEED PCA?

Mg II k

Mg II h

# of Coefs.: 100 --- Retained Variance: 0.999835
DO WE NEED PCA?

Mg II k

Mg II h

# of Coefs.: 5 --- Retained Variance: 0.991356
IRIS

DO WE NEED PCA?

Mg II k

Mg II h

# of Coefs.: 14 --- Retained Variance: 0.999057
# of Coefs.: 26 --- Retained Variance: 0.999511

Mg II k

Mg II h
# of Coefs.: 100 --- Retained Variance: 0.999835

Mg II k

Mg II h
Locations of k-mean #15

IRIS Slitjaw Mg - 2014.03.29 17:46
Locations of k-mean #15

IRIS Slitjaw Mg - 2014.03.29 17:48
Locations of k-mean #15

IRIS Slitjaw Mg - 2014.03.29 17:49
 Locations of k-mean #15

IRIS Slitjaw Mg - 2014.03.29 17:51
Locations of k-mean #15

IRIS Slitjaw Mg - 2014.03.29 17:53
Locations of k-mean #15

IRIS Slitjaw Mg - 2014.03.29 17:46

(arcsec)

(arcsec)
Orange dashed line: k-mean #15 profile

White line: actual profile
Locations of k-mean #15

Orange dashed line: k-mean #15 profile
White line: actual profile
Orange dashed line: k-mean #15 profile
White line: actual profile
WARNING!
Comparison between observations and numerical simulation in the quiet Sun!!!
SUMMARY

➢ k-means algorithm is a powerful tool to classify spectral profiles. It has to be properly initiated.

➢ The code has cons that can be easily quantified and, eventually, optimized.

➢ Based on the k-mean profiles we can have a fast, reliable proxy of the physical event that produces them.

➢ Machine Learning techniques may be very helpful, but they have to be used carefully: Science is done by scientists.

Many thanks for your attention.

asdaldas@stanford.edu