Leibniz-Institut für
Astrophysik Potsdam

Practical application of t-distributed stochastic neighbor embedding in classifying chromospheric spectra

Meetu Verma

DFG Deutsche Forschungsgemeinschaft

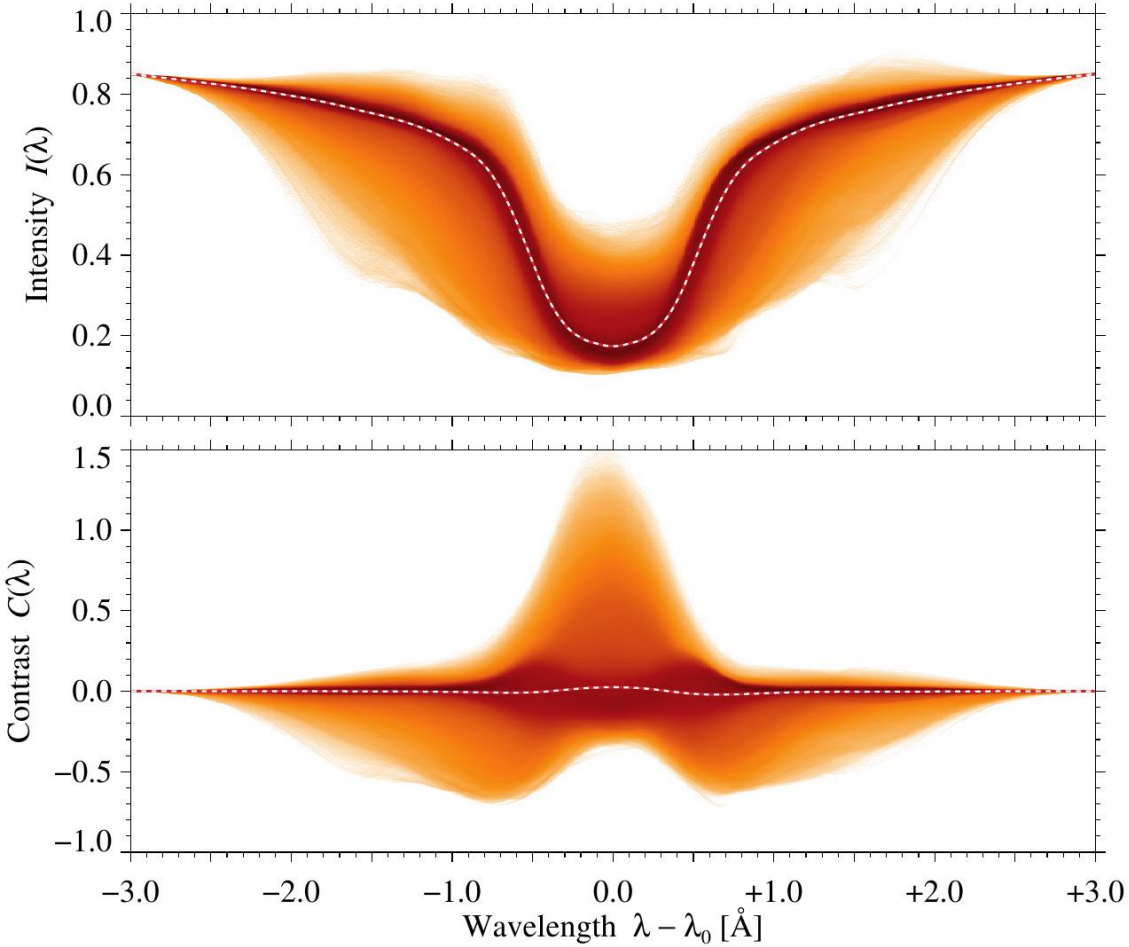
SOLARNET

ESCAPE

3D data → science in every pixel

Verma et al. 2021, ApJ 907, 54

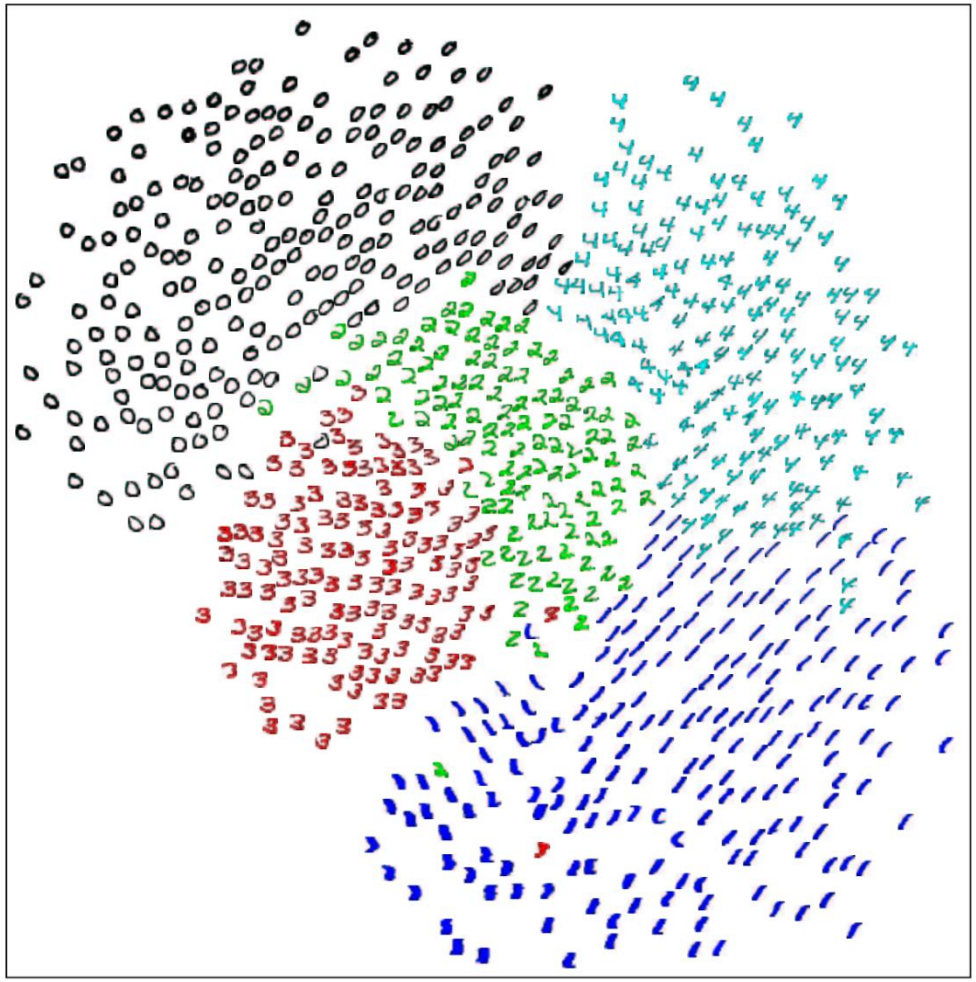
- 3-hours of observation at VTT contains about 8.7 million intensity and contrast profiles.



Classifying the spectra using t-SNE

t-SNE → appropriate tool to classify spectra

- Probabilistic approach
- Dimensionality reduction
- t-SNE result of classifying on 3000 256-dimensional grayscale images of handwritten digits.
- Classes are quite well separated even though t-SNE had no information about class labels.
- Within each class, properties like orientation, skew and stroke thickness tend to vary smoothly across the space.



van der Maaten and Hinton 2008, J. Mach. Learn. Res. 9, 2579
van der Maaten 2014, J. Mach. Learn. Res. 15, 3221

t-SNE → appropriate tool to classify spectra

$$P_{j|i} = \frac{\exp(-\|p_i - p_j\|^2 / 2\sigma_i^2)}{\sum_{k \neq j} \exp(-\|p_i - p_k\|^2 / 2\sigma_i^2)}.$$

A Gaussian probability distribution centered on each point in high-dimensional space. The similarity between points p_i and p_j (two profiles) is the conditional probability $P_{j|i}$ for point p_i to pick point p_j as its neighbor.

$$Q_{j|i} = \frac{\exp(-\|q_i - q_j\|^2)}{\sum_{k \neq j} \exp(-\|q_i - q_k\|^2)},$$

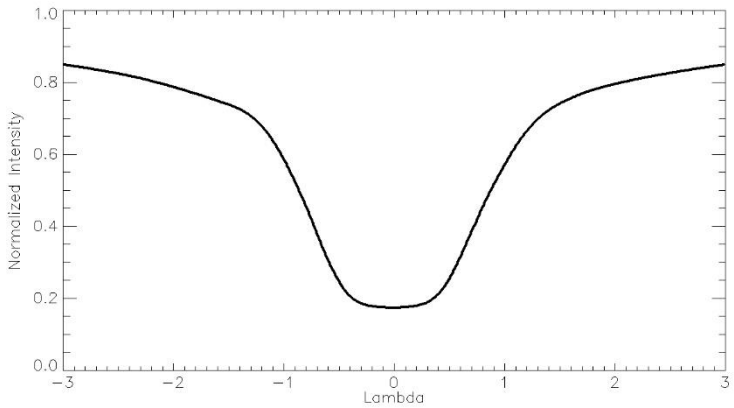
To find only pairwise similarities, the value of $P_{i|j}$ can be set to zero. For the low-dimensional counterparts q_i and q_j of the high-dimensional points p_i and p_j, similar conditional probability $Q_{i|j}$

$$C = \sum_i \sum_j P_{i|j} \log \frac{P_{i|j}}{Q_{i|j}}.$$

To minimize the mismatch between the two distributions. Done by minimizing the sum of the Kullback-Leibler divergence over all data points using a gradient descent algorithm.

Matijevič, et al. 2017, A&A, 603,A19
Anders, et al. 2018, A&A, 619,A125
Panos & Kleint 2020, ApJ, 891, 17

t-SNE → from profiles to classification



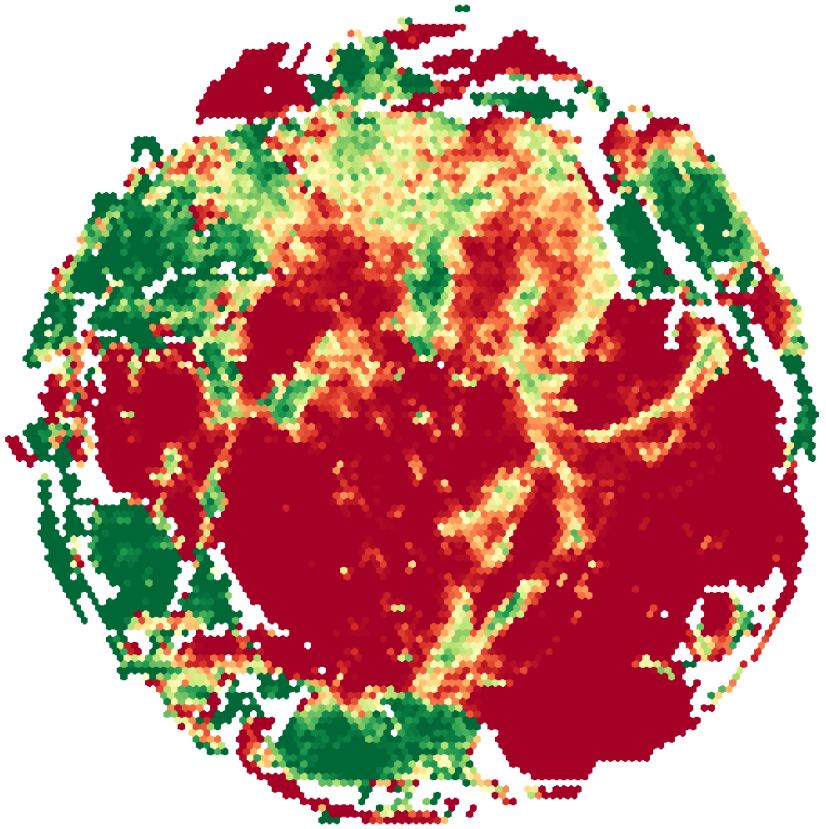
t-SNE projection of 630 x 660 spectral profiles with 601 wavelength points.

The choice of parameters perplexity = 50, theta = 0.5, number of iterations = 1000

Q1 Is the default choice ok?

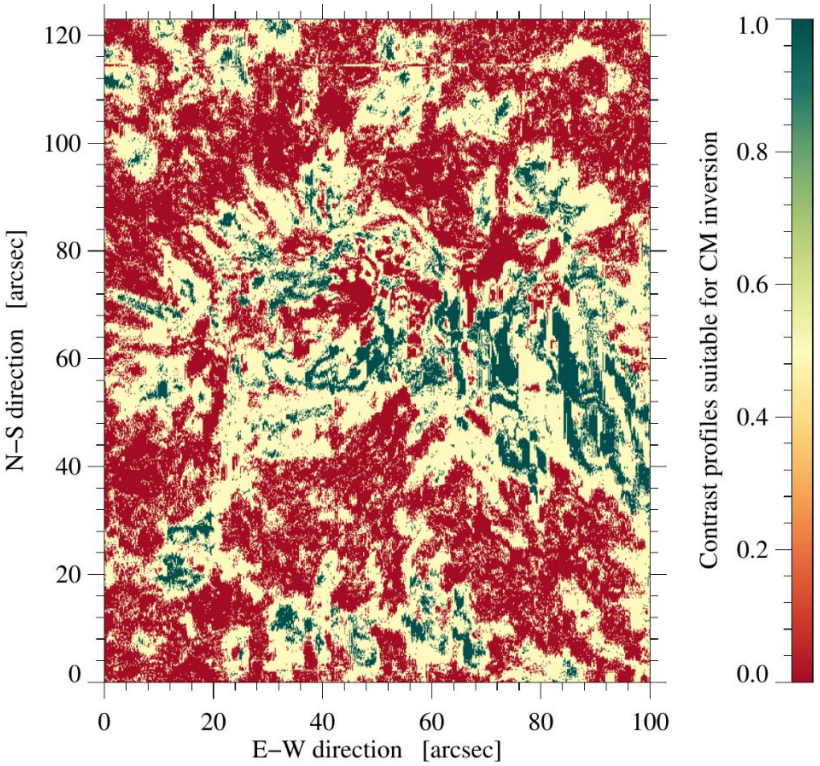
Q2 Is the projection different for profiles and PCA coefficients?

Q3 Is the projection affected by seeing conditions?

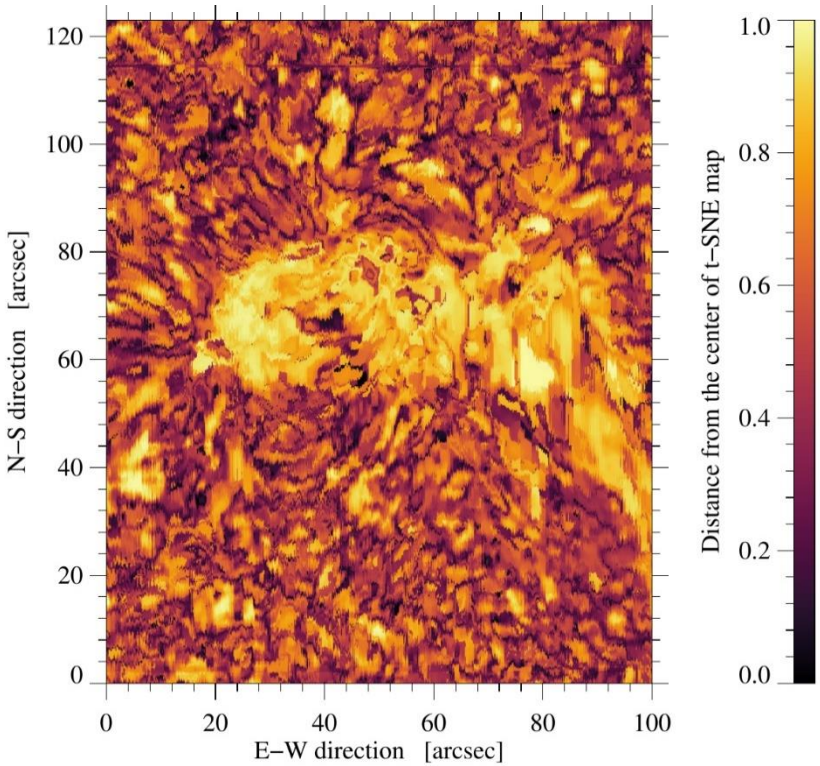


https://distill.pub/2016/misread-tsne/

Backmapping



- Binary parameter – unity if linear and rank order correlation (ρ = 0.95) between observed and CM inverted profiles.
- Number of profiles per hexagonal bin – binning and taking average suitability parameter becomes floating-point number

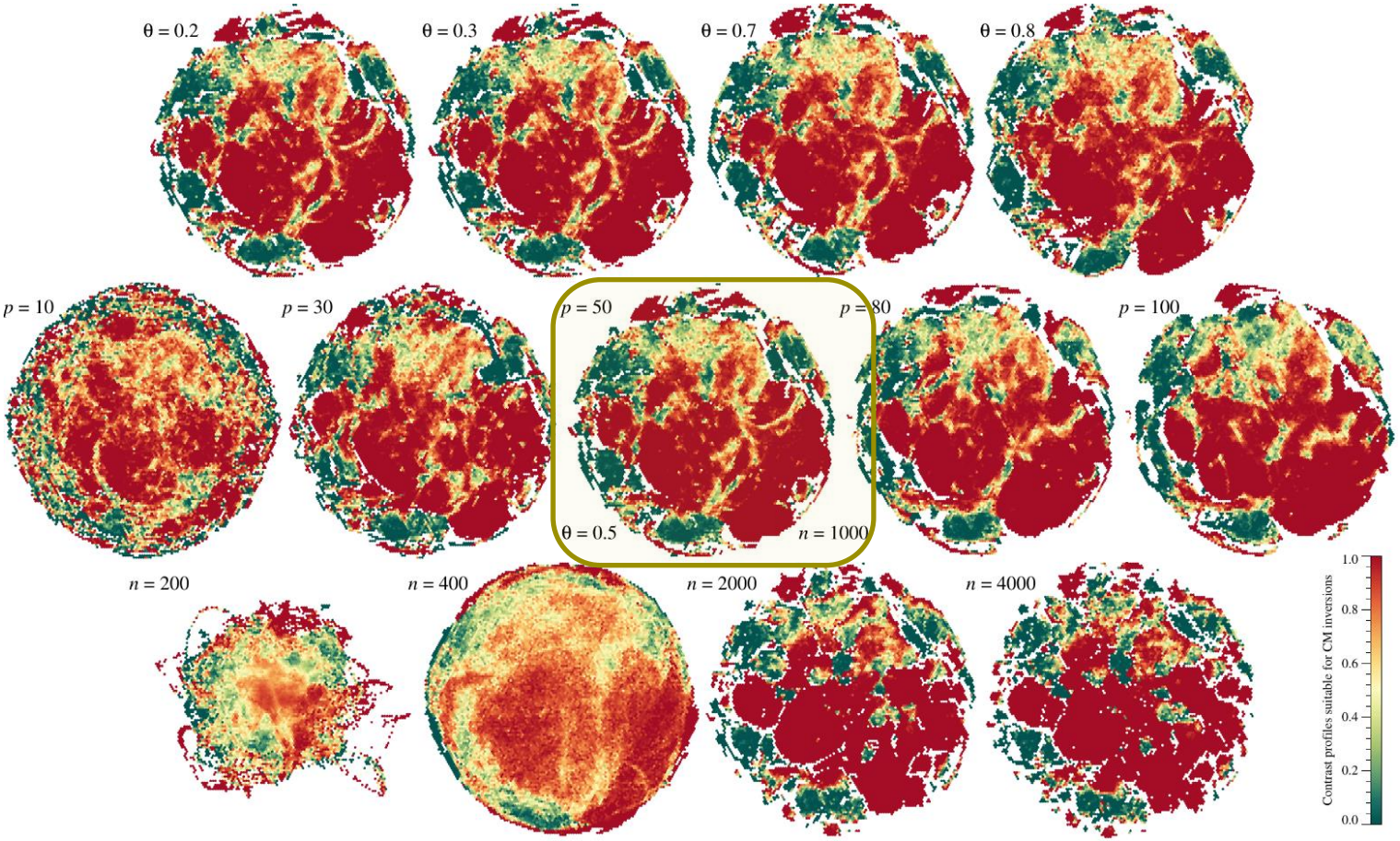


- Normalized distance from the center of t-SNE map
- t-SNE concentrates the very similar quiet-Sun profiles in center
- All other broad profiles are pushed to periphery

Is the default choice ok?

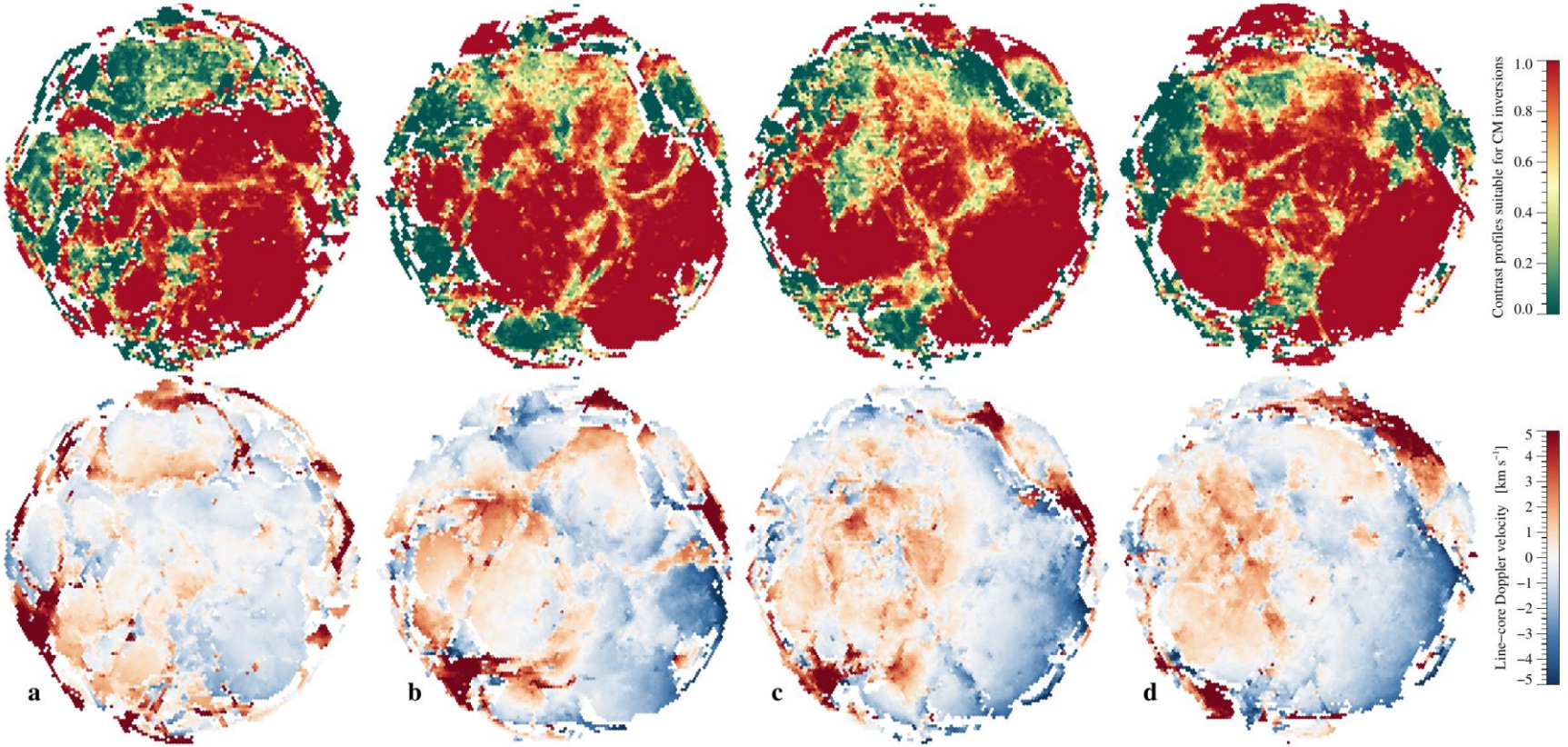
Parameter study – careful selection

A1 The default parameters are fine, maybe the number of iterations has to be increased for large datasets



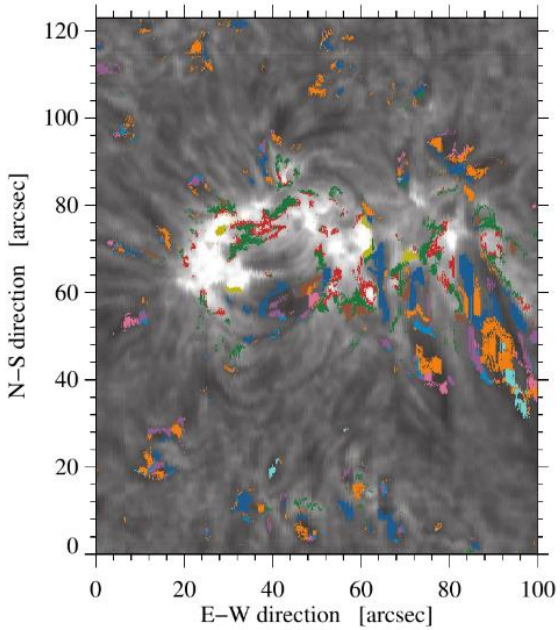
Is the projection different for different input data?

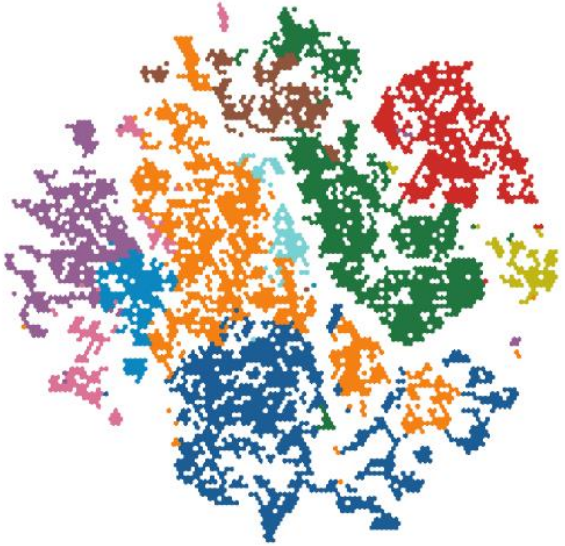
A2/3 Subtle differences, same computing time



- Noise stripped contrast profiles for bad seeing (a), PCA coefficients for good seeing (b), observed contrast profiles for best seeing (c), noise stripped intensity profiles (d)

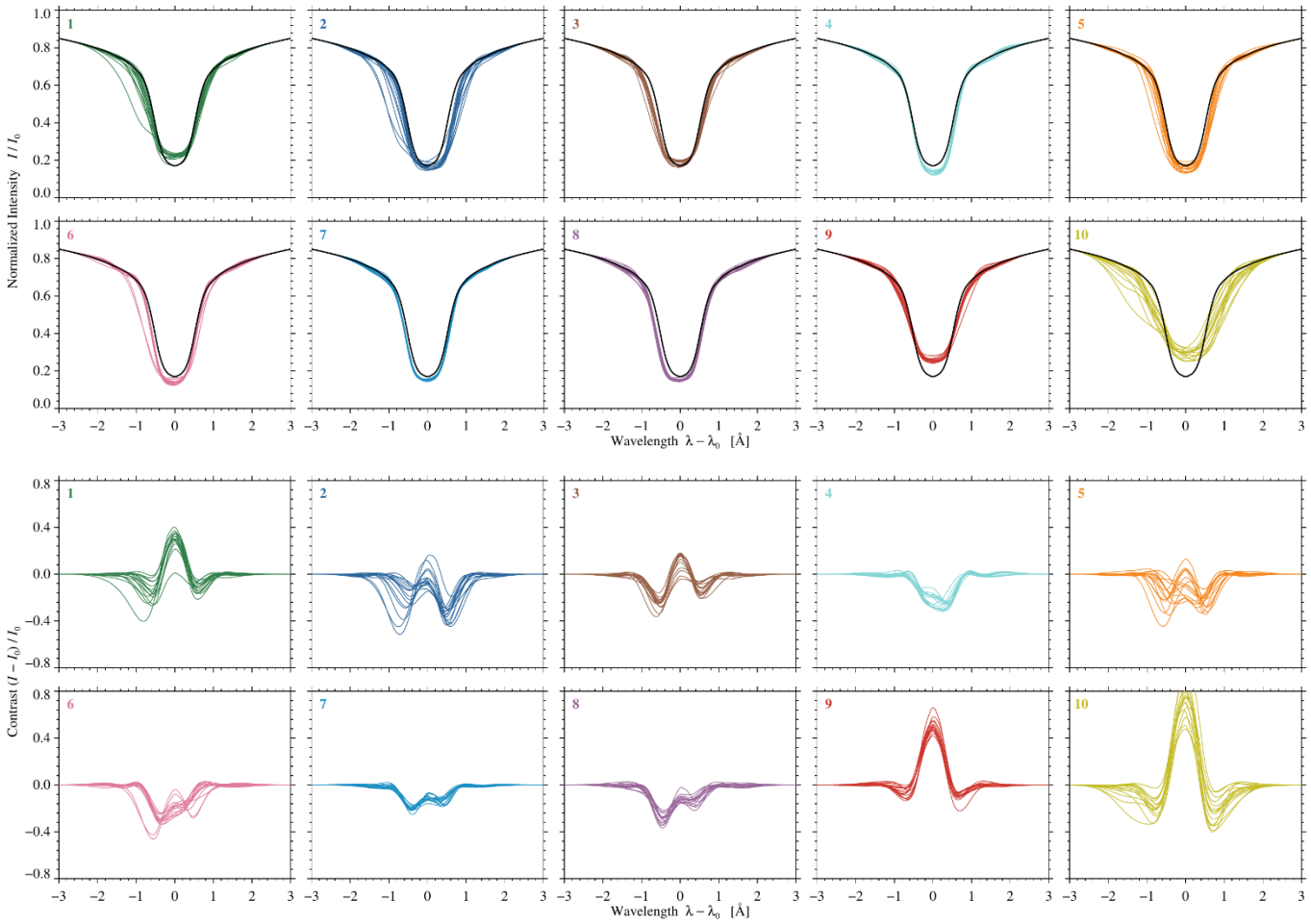
Interpreting clustered spectral profiles





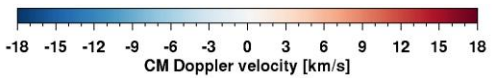
- Clustering ability of t-SNE and identify the ten largest clusters with suitability parameter exceeds 0.9
- Four are isolated remaining in pairs (2 & 3, 4 & 5, 8 & 9)
- Back-mapped to line-core intensity map – all cluster associated with dark features.
- 1 & 9: arch-filaments – 2: footpoint regions – 4, 5, & 8: upper, middle, and side edges of surge
- Sanity check: reprojection – initial clusters are present
- However, pairwise clusters remains for 4 & 5, new pair 1 & 3 – human inference needed

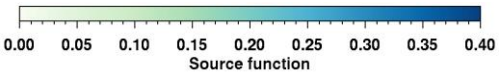
Profiles belonging to ten clusters

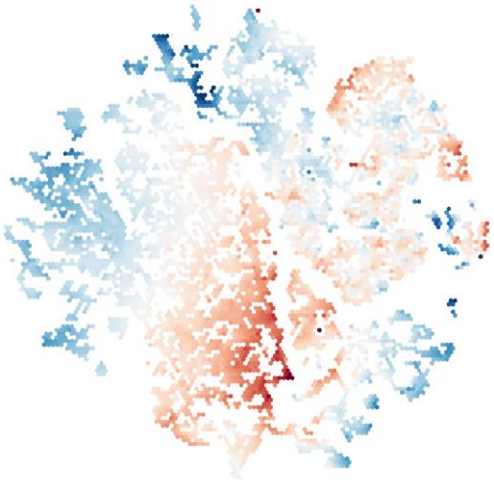


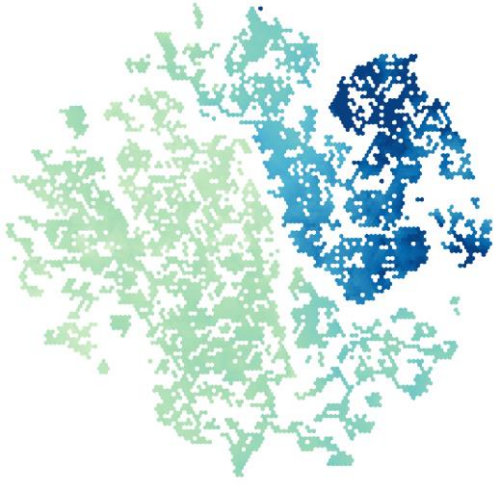
- Variations in all 10 clusters
- 10 clusters in 3 classes
- Contrast profiles with pronounced central component (1, 3, 9, & 10)
- Broad and deep profiles with similar amplitude of central maxima and neighboring minima (2 & 5)
- Contrast profiles where central maximum is less pronounced and the contrast is almost everywhere negative (4, 6,7, & 8)

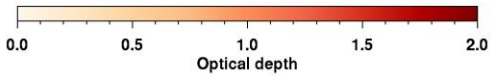
Re-projection of ten clusters

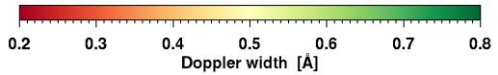


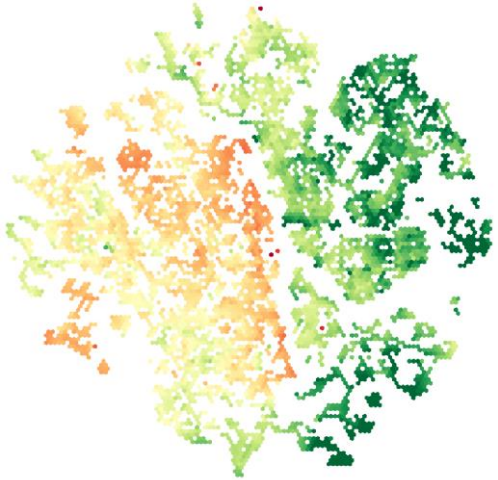












- Reprojection of 10 clusters in terms of CM parameters
- Vertical line with small Doppler width separating largest values to the right and moderately large values to the left
- Projection can be separated in two large clusters with high (Class 1) and low (Class 2 and 3) values
- Difference in Class 2 and 3 is mainly due to difference in the cloud velocities
- From 10 clusters to two or three classes associated with chromospheric absorption features

What did we learn?

What did we learn?

- As an unsupervised machine learning algorithm t-SNE is capable, without any a priori knowledge, to identify Hα profiles suitable for CM inversions.
- Profiles characteristic of active chromosphere are mainly pushed to periphery.
- Default parameters yield already very good results while being computationally efficient.
- Projections are comparable for various input data, however, both noise-stripped contrast profiles and their PCA decomposition based on 10 eigenvectors performed best.
- Clustering based on suitability parameter
- Spectral classes can be defined based on CM parameters.
- Human inference is an essential part of classification.
- Framework presented is particularly relevant for Big Data.

https://arxiv.org/pdf/2011.13214.pdf