"A Performance Comparison of Unsupervised Clustering Techniques for Classification of Spitzer Space Telescope Infrared Spectra"

Relevance:
→ this paper discusses the implementation of hierachical clustering to infrared telescope data for the purpose of classification.

Abstract: ↗ refresh knowledge on this
→ principal component analysis (PCA) was applied to the scaled spectral data prior to classification to reduce data dimensionality.
↳ consider taking purely spectral data from JWST, scaling it, applying PCA, and then HCA.
→ study tried different scaling methods.
→ least classification error:
- HCA with average linkage
- spectrums scaled by their maximum amplitude

} implement in future attempts.

Introduction:
→ infrared telescopes are useful in observing red giants' photospheric emissions as the stars themselves may be concealed by optically thick and dusty envelopes.
- emissions are absorbed and re-radiated in mid- and far-infrared range.

→ Study enlists the Spitzer Space Telescope Infrared Spectrograph (IRS) to investigate mass-losing AGB (Asymptotic Giant Branch) stars in Large Magellanic Cloud (LMC).
- LMC is a good candidate for observation:
  i) nearest neighbor galaxy
  ii) low metallicities & high redshift rates mimic those of more distant high-redshift galaxies.
  iii) contains many IR-luminous mass-losing objects at a similar distance, alleviating distance ambiguities.

consider using LMC data of protostellar disks? maybe even from Spitzer?

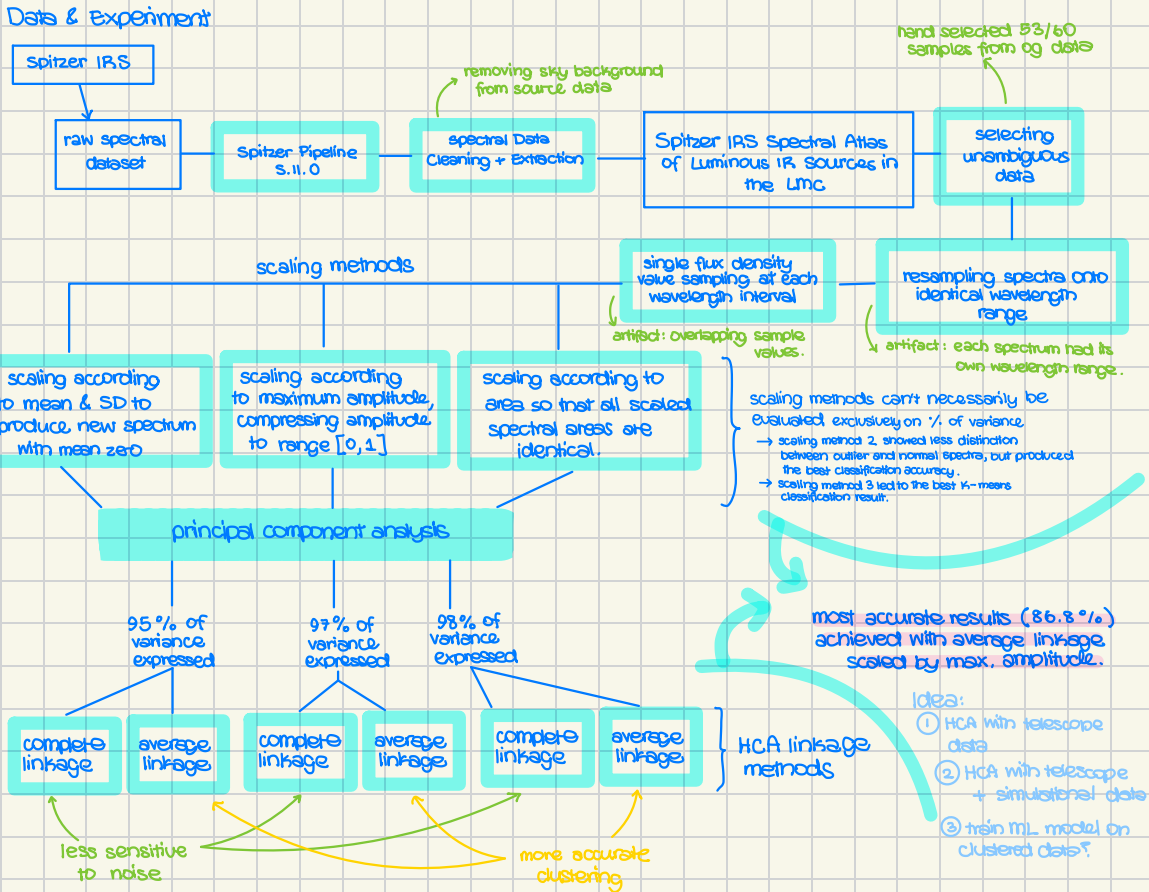
→ unsupervised spectral clustering facilitates identifying empirical similarities / dissimilarities in the absence of detailed physical knowledge.
- seeks natural groupings in dataset without predefined target info.

→ K-means Clustering:
- an iterative approach to find clusters and their centers such that the within-cluster sums of squared distance are minimized.

→ this study uses agglomerative HCA to classify LMC spectral dataset.

→ PCA can be applied to a high dimensional spectral dataset to reduce the computational complexity and simplify the visualization of the data.
- transforms related variables to a set of uncorrelated variables by applying the single value decomposition (SVD) technique to the covariance matrix of the data set.
- patterns in data can then be found & dim. of data can be reduced by mapping into a lower-dim vector space.

review.
→ Given n dimensions for each spectrum of a set of objects, m ($m \leq n$) new uncorrelated dims can be constructed via PCA s.t. each of their corresponding eigenvalues accounts for as much of the variance of the data as possible.
- projection of spectral dataset into uncorrelated vector space yields underlying patterns.



				Overall clustering accuracy
PCA	Hierarchic clustering	Average Linkage	Scaled data 1	73.6%
			Scaled data 2	84.9%
			Scaled data 3	67.9%
		Complete Linkage	Scaled data 1	64.2%
			Scaled data 2	79.2%
			Scaled data 3	75.5%
	K-means		Scaled data 1	77.4%
			Scaled data 2	71.7%
			Scaled data 3	81.1%
Without PCA	Hierarchic clustering	Average Linkage	Scaled data 1	81.1%
			Scaled data 2	86.8%
			Scaled data 3	75.5%
		Complete Linkage	Scaled data 1	71.7%
			Scaled data 2	83.0%
			Scaled data 3	83.0%
	K-means		Scaled data 1	77.4%
			Scaled data 2	71.7%
			Scaled data 3	81.1%

Takeaways:
→ include evaluation of different scaling and linkage methods + with/without PCA on simulational data to identify most optimal (accurate and efficient) clustering method.
- no PCA → HCA → avg. linkage → amplitude scaling was most accurate here.

outperforms in both cases.

→ can we obtain spectral data from simulated cases to emulate this process?