Comparing Stochastic Young Stellar Object Light Curves Using Dimension Reduction and Clustering Algorithms

Benjamin Womack Ryan

School of Engineering, Mathematics and Physics

June 2025

Abstract

Terrestrial planet formation occurs within the inner regions of circumstellar discs. Areas that remain beyond the reach of direct imaging due to current resolution limitations. However, insights into the physical processes shaping these regions can be gained by analysing the photometric variability of Young Stellar Objects. This variability arises from a range of mechanisms, including accretion dynamics, variable extinction, and stellar surface inhomogeneities, each contributing valuable information about the structure and evolution of the planet-forming environment.

We present a quantitative framework for comparing light curves based on variability fingerprints. These are two-dimensional histograms encoding the probability of observing a given increase or decrease in brightness over all timescales. Applied to a refined subset of 240 highly variable young stellar objects from our dataset, these fingerprints span variability from ±0.05 to ±2.0mag over timescales of 1 day to 8.6 years, with $> 90\%$ achieving $S/N > 3$.

Dimensionality reduction via principal component analysis was found to yield a topologically stable variability landscape, in contrast to the sample-sensitive output of non-linear dimension reduction. The projections were minimally affected by the addition or removal of individual sources, enabling robust comparison between observed and model-generated light curves. Simple sinusoidal models with the cadence of an observed dataset and random phase occupied a restricted region of principal component analysis space, indicating that cadence, observing baseline, and photometric noise do not dominate the global structure.

Although principal component analysis provided a stable low-dimensional representation, neither it nor t-stochastic neighbour embedding in conjunction

with clustering algorithms revealed distinct clusters. Instead, the data formed a continuum, reflecting the overlapping physical processes driving variability of young stellar objects. Indicating that a continuous, rather than categorical, framework is more appropriate.

Analysis of the loadings matrices for the two dominant principal components revealed that the primary axis of variance corresponds to the onset timescale of significant (Δmag > 0.3) variability, with 1–3 month trends being most influential. The second component primarily encodes long-term (> 1.5 yr) variability of either increasing or decreasing brightness. By manually inspecting the light curves of objects that lie near one another in the principal component analysis projection, we confirmed that these neighbours display genuinely similar variability patterns. This shows that the principal component coordinates successfully group together stars with comparable light-curve morphology.

These results demonstrate that principle component analysis of variability fingerprints provides a statistically robust and interpretable landscape for comparing observed young stellar object light curves and constraining synthetic variability models rooted in planet formation scenarios.

Contents

1 Introduction 12

1.1 The Project 12

1.2 Star Formation - Gas Clouds to YSOs 13

1.2.1 Collapsing Gas Clouds 13

1.2.2 Classification of YSOs 14

1.3 Circumstellar Discs 17

1.4 Photometric Variability of YSOs 19

1.4.1 Accretion Variability 21

1.4.2 Circumstellar Disc Extinction 21

1.4.3 Additional Causes of Variability 22

1.5 Dimension Reduction and Clustering Algorithms an Overview . . 23

2 From Photons to Fingerprints 27

2.1 Data 27

2.2 The Observatories 28

2.3 Photometric and Colour Calibration 29

2.4 Removing Photometry Outliers 31

2.5 Identification of a Sample of Variables 34

2.6 From Light Curves to Variability Fingerprints 35

3 Comparing Stochastic Light Curves Using Clustering Algorithms 46

3.1 Algorithms and Processes Used in Clustering Workflow 47

3.1.1 Standard Scaler 47

3.1.2 Principle Component Analysis (PCA) 48

3.1.3 Variance Maintained After Dimension Reduction 52

3.2 Clustering Algorithms 54
3.3 K-means Algorithm 56
3.4 Density Based Clustering with Noise (DBSCAN) 57
3.5 t-Distributed Stochastic Neighbour Embedding (t-SNE) 59
3.6 Clustering Algorithm Parameters 61
3.7 Clustering Using DBSCAN 63
3.8 Clustering Using K-means 68
3.8.1 K-Means Stability 72
3.9 What do Principal Components Represent? 76

4 Conclusions **82**

4.1 Further Work 84

5 Appendix **86**

List of Figures

1 a) A cold, dense region within a molecular cloud, approximately 5000 AU across, begins to collapse under its own gravity. b) As collapse proceeds, material accretes onto a central protostar surrounded by a rotating disc. c) Bipolar outflows (orange arrows) emerge from the protostar, clearing material along the polar axes while accretion continues through the disc. d) The system evolves into a more compact configuration, with the disc shrinking to a radius of around 100 AU, marking the later stages of disc evolution before planet formation begins. 14

2 Class system definitions for YSOs showing the circumstellar envelopes alongside their Spectral Energy Distribution(SED) (Vogel, 2013) 16

3 The evolution of a typical disc. Gas distribution shown in blue and the dust in red. (Williams & Cieza, 2011) 18

4 “Amplitude versus timescale for various flavors of YSO variability. Blue indicates the accretion-related events, purple shows the routine variability, either brightening or fading, that is detected at longer wavelengths, which can be referred to as ’disk weather’. In addition, red indicates extinction-related behavior, yellow, stellar phenomena, and green, the variability expected from binary-related phenomena” (Fischer et al., 2023). 20

5 Light curve of the young star V350Cep situated in the NGC7129 cluster. 32

6 Left: light curve of V350Cep with outliers highlighted in cyan. Right: Example V-I vs V CMD plot of the same source. 33

7 Welch-Stetson index of our YSO sample determined from all three filter combinations, $\mathcal{I}_{V,R}$, $\mathcal{I}_{V,I}$, and $\mathcal{I}_{R,I}$ from left to right. The cutoff for highly variable objects used in this study is marked at $\mathcal{I} = 2$ with a solid line. We also indicate $\mathcal{I} = 1$ and 0.5 with a dashed and dotted line, respectively. 35

8 Smoothed variability fingerprint of V350Cep. 36

9 Plot of magnitude differences between light curve points with a positive time difference of V350Cep, V-band. 38

10 Four variants of fingerprints using V350Cep V-band data. Top left: 40x40 pixels, top right: 20x20 pixels, bottom left: 20x20 pixels with timescales less than one day excluding, bottom right: 9x16 adaptive pixels over same timescale as bottom left. 39

11 Plot of mean fingerprint pixel value against number of fingerprints in a stack for a randomly chosen pixel. 41

12 Comparison of bootstrapped (left) and Poisson (right) uncertainty maps for 20x20 pixel configuration (top) and 9x16 pixel configuration (bottom). 42

13 Comparison of uncertainty values of V350Cep V-band data, slope value of 0.83. 43

14 Comparison of uncertainty values of the young star V350Cep for low uncertainty values, slope value of 1.02. 44

15 Two directions of greatest variance belonging to a set of data with three dimensions. 51

16 Fraction of total variance captured by components after PCA. Left: 9x16 fingerprints, Right: 20x20 fingerprints. 53

17 An example of using 2D Euclidian clustering where the intra-cluster distance is minimized (Murty & Devi, 2016). 55

18 Visualization of clustering using DBSCAN with t-SNE, minimum samples = 20, epsilon = 20, perplexity = 5, on 240 variable 9x16 V-band fingerprints. 64

19 The light curves of three random pairs of objects taken from clusters shown in figure 18, top: cluster 0, middle: cluster 1, bottom: outliers. 65

20 The light curve (left) and 9x16 fingerprint (right) generated by a sine wave with a two year period and one magnitude amplitude overlaid over a random HOYS object cadence. 66

21 Left: Clustering using DBSCAN with t-SNE, epsilon = 20, perplexity = 5 and minimum number of samples 20, on 240 variable 9x16 V-band fingerprints. Right: Same as left with one fingerprint added to sample based on a sine wave light curve. 67

22 K-means clustering using 240 9x16 V-band fingerprints with PCA dimension reduction applied. 69

23 The light curves of three random pairs of objects taken from clusters shown in figure 22, top: cluster 0, middle: cluster 1, bottom: cluster 2. 70

24 Left: K-means clustering using 240 9x16 V-band fingerprints with PCA dimension reduction applied. Right:As left panel with one fingerprint derived from a sine wave light curve added to data. . . 71

25 The means (center) and 20 standard deviations (ovals) of K-means clustering using 240 9x16 V-band fingerprints one of which is bootstrapped 1000 times (orange) 72

26 The means (center) and standard deviation (ovals) of K-means clustering using 240 9x16 V-band fingerprints plus fingerprints created from sine wave light curves with varying periods and an amplitude of one magnitude whose cadence were varied approximately 960 times. 73

27 The means (center) and standard deviation (ovals) of K-means clustering using 240 9x16 V-band fingerprints plus fingerprints created from sine wave light curves with varying periods and an amplitude of half a magnitude whose cadence were varied approximately 960 times. 76

28 Left: Loadings vector heat map for principal component 1. Right: Loadings vector heat map for principal component 2. 78

29 Left: The mean fingerprint for all HOYS cadences generated from a sine wave type light cure with a period on one month and an amplitude of one magnitude. Right: As left with a period of eight years. 78

30 top left: PC1=35.50 PC2=14.04, top right: PC1=32.10 PC2=9.64, mid left: PC1=-0.19 PC2=0.15, mid right: PC1=-0.85 PC2=0.42, bottom left: PC1=-7.58 PC2=1.41, bottom right;PC1=2.66 PC2=3.75 79

31 A selection of light curves with a high PC1 value (> 20). 87

32 A selection of light curves with a low PC1 value (< -7). 88

33 A selection of light curves with a high PC2 value (> 5). 89

34 A selection of light curves with a low PC2 value (< -7). 90

List of Tables

1 Minimum Davies-Bouldin Index across the tested parameter space of Perplexity and Epsilon, for varying values of minimum samples. 62

2 Multiples of π used to define approximate variability periods. . . . 75

3 Test results for code optimization of normalizing fingerprint columns, 40x40 configuration used 1000 times 86

4 Mean slopes of comparison between Poisson and Bootstrapping error values for every object. 86

1 Introduction

1.1 The Project

Young Stellar Objects (YSOs) are stars in the early stages of formation, still accreting material from their surrounding circumstellar discs and often exhibiting variability due to dynamic processes in their immediate environments (Joy, 1945). The primary objective of this research is to develop a robust method for comparing the light curves of YSOs with the ultimate goal of evaluating planet formation simulations. These simulations begin with defined initial conditions and model the dynamical and physical processes that lead to the formation of planetary systems around low-mass stars. Of particular interest are terrestrial planets, rocky Earth-like bodies composed predominantly of silicate materials or metals. Characteristically, terrestrial planets possess solid surfaces, exhibit higher densities relative to gas giants, and in our own Solar System, are located closer to the Sun.

Planet formation occurs within a circumstellar disc of gas and dust surrounding the YSO. Terrestrial planets are believed to form within the inner regions of these discs. However, direct imaging of these inner disc regions remains beyond the resolution capabilities of most current telescopes. Instead, variability in the observed brightness of YSOs, driven by dynamic processes within the disc, provides an indirect probe of disc structure and evolution. By quantifying this variability over time, we aim to derive statistical representations of YSO light curves, which can then be used to construct a comparative framework for characterizing their behavior (Fischer et al., 2023).

For this framework to be meaningful, objects with similar light curve behaviour should appear proximate in the visual or analytical representation constructed from their variability characteristics. The framework must exhibit structural stability, that is, the overall arrangement of objects should remain consistent as new data are incorporated or existing entries are modified.

To achieve this, we employ a suite of machine learning techniques, including unsupervised dimensionality reduction and clustering algorithms. A brief overview of these methods is presented in Section 1.5.

1.2 Star Formation - Gas Clouds to YSOs

1.2.1 Collapsing Gas Clouds

The process of star formation differs fundamentally between low- and intermediate-mass stars ($M < 8M_{\odot}$) and high-mass stars ($M > 8M\odot$) (Tan et al., 2014). The formation of a low-mass star begins in cool molecular clouds often found near massive, young stars (Lizano & Shu, 1987). These clouds, which can span more than a hundred parsecs, are shaped by processes such as supernova explosions, stellar winds, or galactic interactions. Star formation is governed by the balance between self-gravity, which drives collapse, and supersonic turbulence, which provides support against it. Although molecular clouds are globally supported against collapse by supersonic turbulence, gravitational collapse proceeds locally within shock-compressed regions where the density becomes sufficiently high (Stone et al., 1998; Mac Low et al., 1998).

Figure 1 illustrates a YSO at four distinct stages in its evolution toward becoming a main sequence star. There is a substantial disparity between the density of a molecular cloud core and that of the youngest observable stars. Directly observing the gravitational collapse that bridges this gap remains ex-

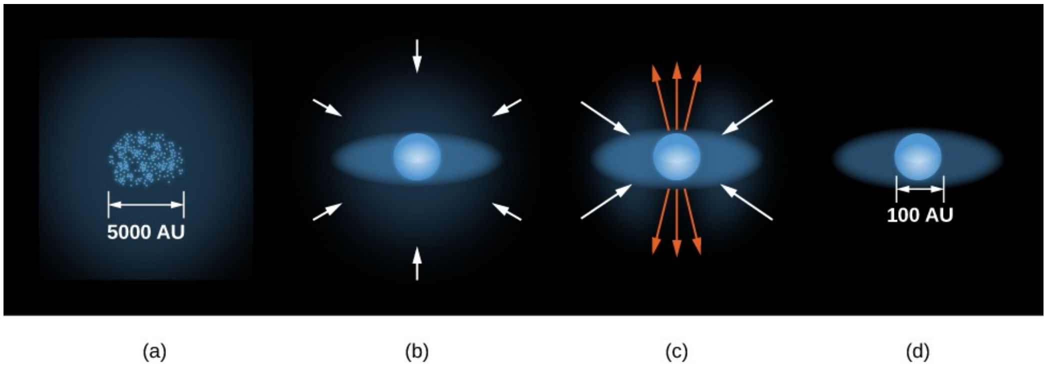

Figure 1: a) A cold, dense region within a molecular cloud, approximately 5000 AU across, begins to collapse under its own gravity. b) As collapse proceeds, material accretes onto a central protostar surrounded by a rotating disc. c) Bipolar outflows (orange arrows) emerge from the protostar, clearing material along the polar axes while accretion continues through the disc. d) The system evolves into a more compact configuration, with the disc shrinking to a radius of around 100 AU, marking the later stages of disc evolution before planet formation begins.

tremely challenging for two primary reasons. First, the interiors of molecular clouds—where star formation occurs—are heavily obscured by dust, rendering them opaque to optical wavelengths. Second, the collapse phase is both rapid, typically lasting only a few thousand years, and rare, meaning only a small fraction of stars are caught in this transient stage. Despite these observational limitations, a combination of theoretical models and indirect evidence has allowed astronomers to develop a coherent picture of the earliest phases of star formation.

1.2.2 Classification of YSOs

As the protostar continues accreting mass, the in-falling gas forms a rotating accretion disc around it, funneling material onto the growing star. The increasing pressure and temperature at the core lead to the emission of infrared radiation, signaling the emergence of a protostar. During this phase, magnetic fields and strong bipolar outflows help regulate the star's angular momentum by expelling excess material. Eventually the surrounding envelope dissipates. Jets and outflows clear part of the envelope. The rest either accreting onto the star or being blown away by stellar winds, revealing a pre-main-sequence object.

The Spectral Energy Distribution(SED), see figure 2, is used to classify YSOs. An evolutionary stage of the object can be characterized by the relation of star to disc emission. Lada (1987) defines three classes(1-3) of YSOs using SEDs in wavelengths $2 < \lambda < 25\mu$m using an infrared spectral index:

$$\alpha_{SED} = \frac{dlog(\lambda F_{\lambda})}{dlog\lambda} \tag{1}$$

Work by André et al. (1993) added an additional precursory stage to the classification of YSO evolution, class 0. Due to the YSO being entirely shrouded in its accretion envelope which is spherical at this stage, Class 0 objects are undetectable at $\lambda < 10\mu m$. This protostellar stage is very short, $\approx 10^4 yr$, (Hueso & Guillot, 2005). Class I objects, $\alpha_{SED} > 0$, are generally optically obscured by their envelope. As the cloud continues to collapse, due to the conservation of angular momentum, the envelope settles into an accretion disc. Class II objects have cleared their envelope and have $-2 < \alpha_{SED} < 0$. The protostar becomes optically visible at this point with the accretion disc contributing to excess infrared emission. Class III objects have $\alpha_{SED} < -2$. By this stage most of the disc has either been accreted onto the star or been dissipated by planet formation or photo evaporation. These stars have very little excess emission. Further work on refining the boundaries between classes has been carried out by Greene et al. (1994). The evolution of a protostar from a Class 0 to a Class III is in the order of $10Myr$. Class II objects with masses below two solar masses are known as classical T-Tauri Stars (CTTS). The prototype for this star being

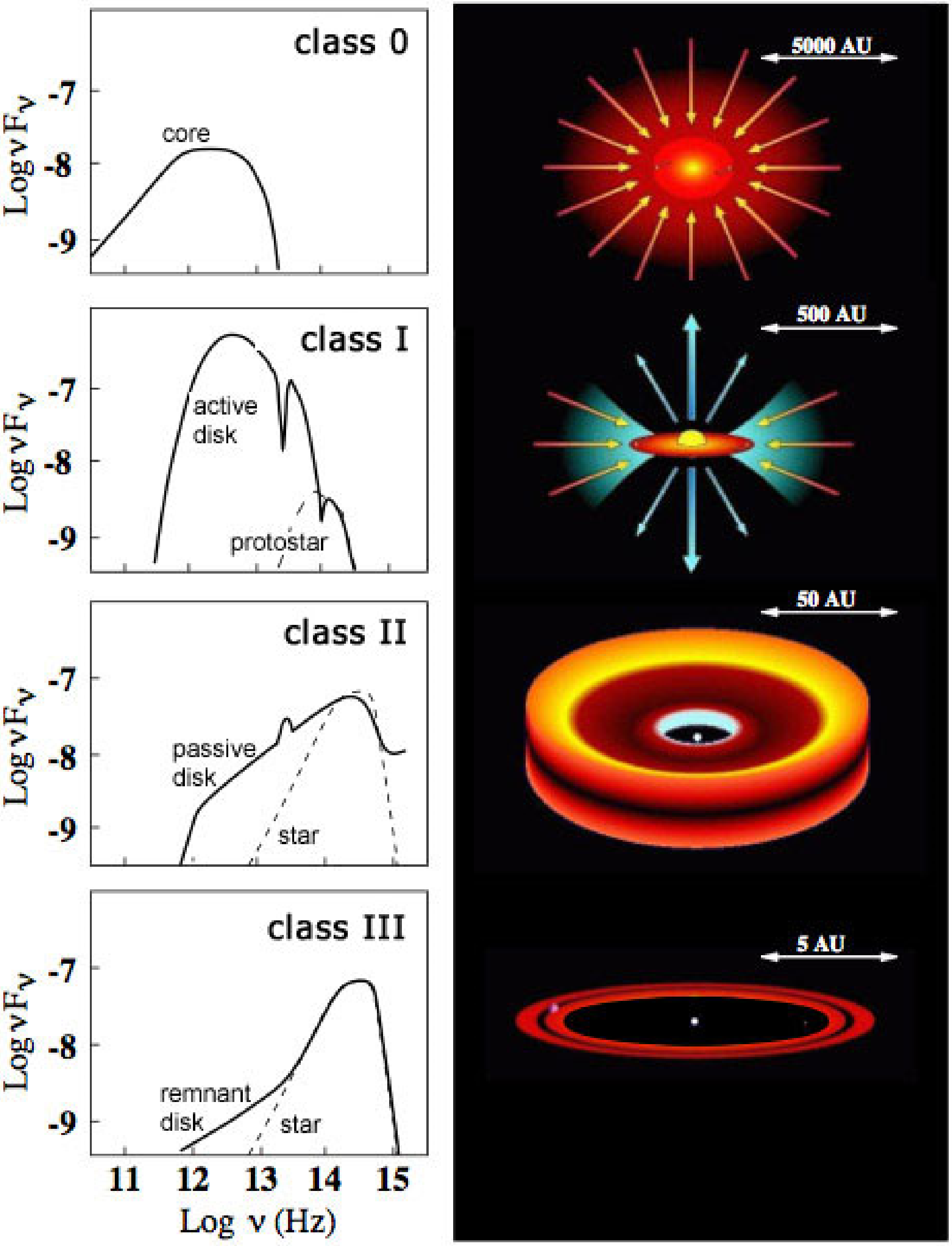

Figure 2: Class system definitions for YSOs showing the circumstellar envelopes alongside their Spectral Energy Distribution(SED) (Vogel, 2013)

a young variable star in the Taurus-Aur region (Joy, 1945). Class III objects due to their lack of strong emission line features are known as Weak-line T Tauri stars (WTTS).

As the star enters the T-Tauri phase it is characterized by strong magnetic activity, erratic luminosity variations, and powerful winds that clear out the remaining circumstellar material. T-Tauri stars are still contracting and generating energy through gravitational contraction rather than hydrogen fusion, which has not yet fully ignited in their cores. The surrounding protoplanetary disc, rich in gas and dust, may give rise to planets over time. Eventually, once core temperatures exceed 10 million K, stabilizing the star as it reaches the main sequence and enters a long period of steady nuclear burning. For stars with masses up to approximately $1.1M_{\odot}$, energy generation is dominated by the proton–proton (PP) chain. In more massive stars, the carbon–nitrogen–oxygen (CNO) cycle becomes the primary fusion process.

1.3 Circumstellar Discs

After the protostellar stages, the envelope surrounding a YSO collapses into a circumstellar disc. At the time of formation, the composition of this disc is approximately that of the interstellar medium (ISM), consisting of approximately 99% gas and 1% dust. Observing gas in circumstellar discs is more challenging, as gas emits at specific wavelengths and its emission can be obscured by dust. As a result, direct detection typically requires spectroscopic observations. In contrast, dust is more readily observed, either through its thermal emission at long wavelengths or due to its optical thickness at shorter wavelengths.

Disc masses do not increase with time as the core collapses implying a rapid

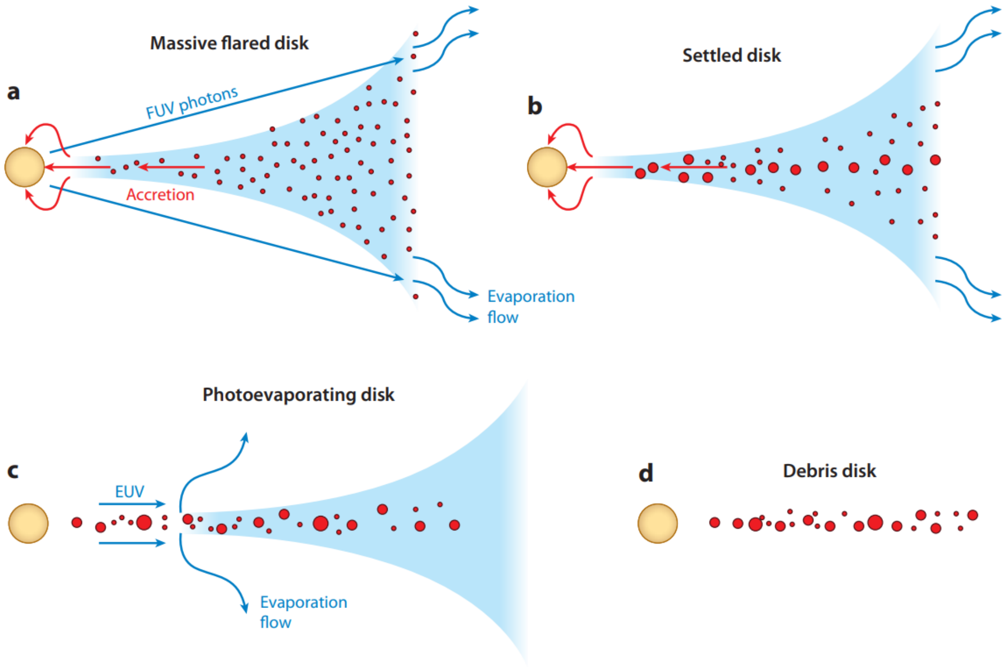

Figure 3: The evolution of a typical disc. Gas distribution shown in blue and the dust in red. (Williams & Cieza, 2011)

accretion of material from disc to the young star. The low average luminosity of YSOs, combined with the sudden accretion bursts observed in FU Orionis-type objects, suggests that young circumstellar discs are gravitationally unstable (Williams & Cieza, 2011). As a YSO evolves from Class I to Class II, the disc mass has decreased to just a few percent of the central stellar mass. At this stage, the disc is considered protoplanetary. For class II YSOs work by Mannings & Sargent (2000), Natta et al. (2000), Acke et al. (2004) and Scholz et al. (2006) shows that there is a large scatter, ~ 0.5 dex, but confirms a ratio of $\frac{M_d}{M_\star} \sim 0.1$ for low mass stars.

Looking at figure 3 we see the following:

(a) Early in its evolution, the disc loses mass through accretion onto the central star and far-ultraviolet (FUV) photoevaporation of its outer regions . The photoevaporation caused by the FUV settles the disc into a truncated form.

This limits its viscous expansion to a size of roughly several hundred AU in diameter (Gorti et al., 2009).

(b) Simultaneously dust grains grow into larger bodies. These can begin to settle towards the mid plane of the disc allowing them to form into rocks, planetesimals and larger. Hence the flared dusty disc becomes flatter. This steepens the slope of the mid and far-IR SED as a smaller fraction of the stellar radiation is intercepted by circumstellar dust (Dullemond & Dominik, 2005).

(c) As the disc mass and accretion rate continue to decline, extreme-ultraviolet (EUV) photoevaporation becomes more significant. At this point, the outer disc can no longer replenish the inner disc with material, leading the inner disc to drain on a viscous timescale ($\sim 10^5$ years). This results in the formation of an inner hole of roughly a few AU in radius. Accretion onto the star halts. This allows energetic photons from the star to impact the inner disc. The disc rapidly dissipates from the inside out leading to a rapid transition from a CTTS to a WTTS (Padgett et al., 2006; Cieza et al., 2007; Wahhaj et al., 2010).

(d) Once the remaining gas is cleared via photoevaporation, small ($r < 1\mu$m) dust grains are removed by radiation pressure. Some of the larger grains spiral inwards due to Poynting–Robertson drag. Those that reach the dust sublimation radius are evaporated. What remains is a debris disc composed primarily of large grains, planetesimals, and/or planets. This disc is of very low mass and may not always be detectable (Williams & Cieza, 2011).

1.4 Photometric Variability of YSOs

The variability of YSOs led to their first identification (Joy, 1945). Photometric variability in YSOs arises from several mechanisms, including the accretion of material from the surrounding circumstellar disc, extinction caused by the disc itself, and stellar phenomena linked to the YSO's rotation. The duration of these variability events can range from hours to centuries, depending on their physical scale and underlying cause. Figure 4 illustrates the relationship between the amplitude of variability and its timescale. This project focuses on investigating variability driven by the interaction between a YSO and its circumstellar disc material, with particular emphasis on the inner disc region where terrestrial planets are thought to form.

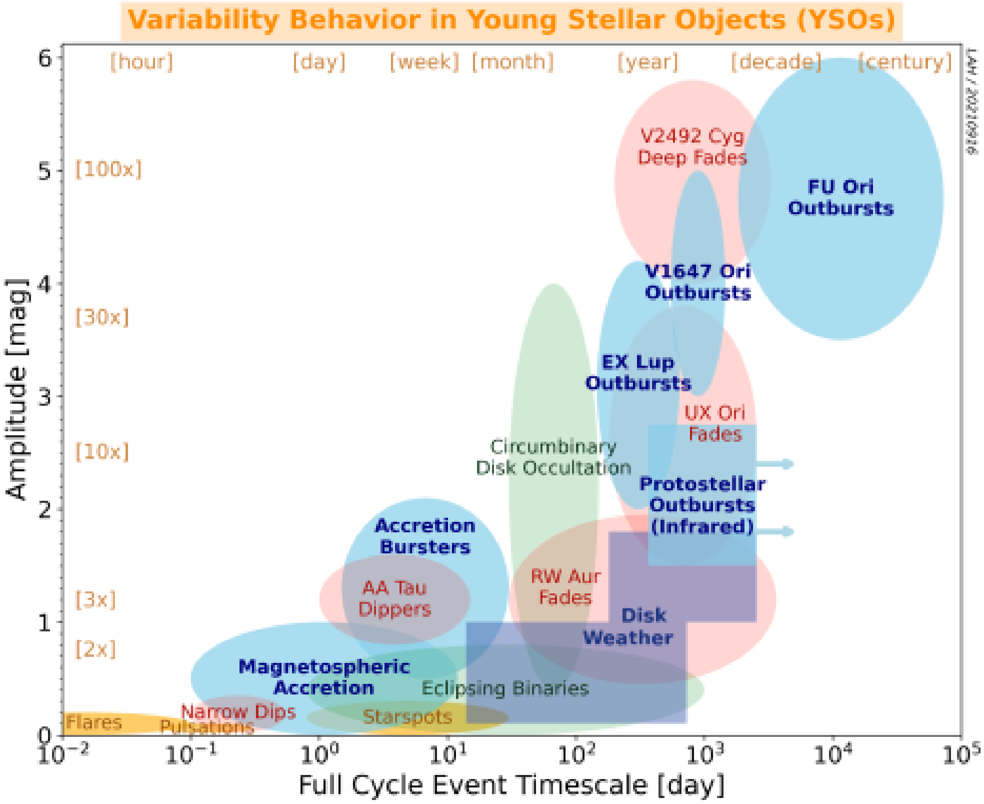

Figure 4: "Amplitude versus timescale for various flavors of YSO variability. Blue indicates the accretion-related events, purple shows the routine variability, either brightening or fading, that is detected at longer wavelengths, which can be referred to as 'disk weather'. In addition, red indicates extinction-related behavior, yellow, stellar phenomena, and green, the variability expected from binary-related phenomena" (Fischer et al., 2023).

1.4.1 Accretion Variability

As YSOs evolve toward the main sequence, they accrete material from, or otherwise disperse, their surrounding circumstellar discs. Most of a star's final mass is thought to be accumulated during short-lived episodes of high-rate accretion in the earliest, protostellar stages (Class 0/I). These so-called episodic accretion events have been proposed to resolve the 'luminosity problem'.

To form a 1 $M_\odot$ star within the typical 10^5 year protostellar phase, an average accretion rate of $\sim 10^{-5}$ $M_\odot$ yr^{-1} is required. However, the luminosities predicted from continuous accretion at this rate greatly exceed the observed luminosities of most protostars (Kenyon et al., 1990). This discrepancy suggests that a significant fraction of stellar mass is instead accumulated through brief, intense bursts of accretion. Supporting this, Evans et al. (2009) found that approximately half of a star's final mass is accreted during just 7% of its ~0.5 Myr Class I lifetime. These episodic accretion events drive variability in YSOs, increasing both their luminosity and photometric variability (see Figure 4).

1.4.2 Circumstellar Disc Extinction

Extinction caused by dust in the circumstellar disc of a YSO results in the reddening and dimming of the light observed by telescopes. The circumstellar disc that surrounds T-Tauri stars consists of mixed dust and gas which is flared towards the outer edge of the disc (e.g. Kenyon et al., 1990; Greene et al., 1994). Extinction can occur from warps or prominences in the disc as it rotates. These events can take place periodically or aperiodically. For periodic events this is determined by the radius of the warp. Assuming Keplerian rotation this radius of the rotating material is given by:

$$r_c = \left(\frac{P_{\mathrm{kep}}}{2\pi}\right)^{\frac{2}{3}} (GM_\star)^{\frac{1}{3}}. \tag{2}$$

Inner disc warps or stellar prominences near the corotation radius can lead to occultations that occur with a period close to the stellar rotation period. The prototypical object for this behaviour is AA Tau (Bouvier et al., 1999). This system exhibits periodic dimming of approximately 1.4 magnitudes in the B, V, R, and I bands, caused by a warped inner disc that regularly moves into the line of sight. The dimming is nearly achromatic, suggesting that the dust grains responsible for the extinction are larger than 1 µm.

Photometric variability in YSOs serves as a powerful diagnostic of the dynamic processes occurring in and around forming stars. Whether driven by accretion bursts or variable extinction from circumstellar disc structures, these fluctuations encode information about the physical environment and evolutionary state of the system. By characterizing and comparing such variability, we gain critical insights into disc evolution, accretion processes, and the early conditions that may influence planet formation.

1.4.3 Additional Causes of Variability

Surface spots on YSOs—regions hotter or cooler than the surrounding photosphere—produce periodic photometric variability as the star rotates, typically with amplitudes of 0.1–0.8 mag. This rotational modulation is commonly used to infer stellar rotation periods but is degenerate with spot configuration and requires multi-band data for accurate interpretation (e.g. Luger et al. (2021)). Polar spots do not contribute to modulation unless there is asymmetry relative to the rotation axis. Additional variability in YSOs includes δScuti-type pulsations ($P \sim 0.05d$, ~0.01mag) in 1.5–3.5$M_\odot$ stars and X-ray flares lasting hours to days due to plasma temperatures around 10^7K (Bedding et al., 2020), (Kirmizitas, Cavus & Aliçavuş, Kirmizitas et al.).

1.5 Dimension Reduction and Clustering Algorithms an Overview

Dimension reduction is a process used to simplify complex data by reducing the number of variables or features while retaining essential information. The goal is to make the data more manageable and easier to analyze without losing critical patterns or relationships (Murty & Devi, 2016).

In many datasets, especially those with high-dimensional data (where there are many features or variables), not all features contribute equally to understanding the underlying structure. Dimension reduction techniques aim to identify the most relevant dimensions or features that capture the majority of the variation in the data, reducing the noise and irrelevant information.

This can be particularly useful in machine learning, data visualization, and pattern recognition tasks, where working with fewer, more meaningful variables can improve performance and make the data easier to interpret.

Clustering algorithms are a type of unsupervised machine learning technique used to group similar data points together based on certain characteristics or features. The goal is to organize a dataset into clusters, where each cluster contains items that are more similar to each other than to those in other clusters (Scitovski et al., 2021).

Clustering doesn't require labeled data, as it aims to uncover inherent groupings or patterns within the data. These algorithms are widely used in fields like data mining, pattern recognition, and customer segmentation. They can help identify patterns, trends, or outliers that may not be immediately obvious.

There are different types of clustering algorithms, each using different methods to define ”similarity” and to group data points. Some focus on the distance between data points, while others might use density, probability distributions, or other criteria (Murty & Devi, 2016).

Given the scope of this project, a deliberate decision was made to focus on a limited number of techniques, with alternative clustering and dimensionality-reduction methods briefly acknowledged to contextualise the chosen approach. A range of linear and non-linear approaches are commonly applied to high-dimensional time-series data, including Independent Component Analysis, Uniform Manifold Approximation and Projection, hierarchical clustering, and Gaussian mixture models (Scitovski et al., 2021). These methods were considered in a conceptual sense, but the techniques adopted here were selected to best address the specific aim of grouping stars according to similarity in light-curve morphology.

Two complementary analysis pipelines are therefore employed. The first consists of a standard scaler followed by Principal Component Analysis (PCA) and k-means clustering. This pipeline represents a conventional, widely used approach that provides a clear and interpretable baseline, allowing clusters to be identified in a reduced space that captures the dominant sources of variance in the fingerprint representations. PCA serves to reduce dimensionality while mitigating noise and redundancy, and k-means offers a straightforward means of partitioning the data into a fixed number of groups based on global variance structure (Murty & Devi, 2016).

However, PCA is a linear technique and does not guarantee that objects with similar variability morphology will be well separated in Euclidean distance within the reduced space. In particular, clusters corresponding to subtly different light-curve behaviours may remain elongated or overlapping when projected onto the leading principal components. For this reason, a second pipeline is employed, consisting of a standard scaler followed by PCA, t-Distributed Stochastic Neighbor Embedding (t-SNE), and Density-Based Clustering with Noise (DBSCAN) (Hahsler et al., 2019).

In this second pipeline, PCA is again used as an initial step to reduce noise and compress the data, after which t-SNE is applied as a non-linear dimensionality-reduction method. t-SNE was chosen specifically for its ability to preserve local neighbourhood structure in high-dimensional data, rather than for maintaining global distances or variance. The primary objective of this work is to identify stars with genuinely similar variability patterns, making local structure more important than the faithful representation of large-scale geometry. In this context, t-SNE is well suited to revealing small-scale structure and separating populations that remain blended in linear projections.

The density-based clustering algorithm DBSCAN is then applied in the non-linearly reduced space. Although DBSCAN can in principle be applied directly to PCA outputs, its reliance on distance thresholds and local point density makes it sensitive to cluster shape and overlap. Applying DBSCAN after non-linear dimensionality reduction allows density contrasts to be more clearly defined, improving the identification of coherent groups corresponding to similar light-curve morphology. Together, these two pipelines provide complementary perspectives, one offering interpretability and a conventional baseline, and the other emphasising sensitivity to local structure without assumptions about cluster number or geometry (Hahsler et al., 2019). A more detailed description of the individual components of the two pipelines, along with the associated data analysis, is intentionally deferred to Section 3, where their application is described in context.

2 From Photons to Fingerprints

2.1 Data

The data used are provided by the Hunting Young Outbursting Stars (HOYS) citizen science project (Froebrich et al., 2018). The project is run by Dr. Dirk Froebrich (University of Kent) with science co-leads Dr. Aleks Scholz (University of St Andrews) and Justyn Campbell-White (European Southern Observatory). The aim of the HOYS project is to provide a long-term, multi-filter photometric study of young star forming clusters or star forming regions visible from the northern hemisphere. The project has been running since October 2014 until present and planned to run until at least 2040. The project observes 25 nearby star forming clusters which are d<1 kpc and an age of less than 10 Myr. The data are collected by a mixture of professional, university and amateur observatories. HOYS data are used to conduct surveys of the 25 nearby young clusters and star forming regions. Gathering statistics on the causes of variability of the YSO. Including variability caused by inner disc obscuration and disc excess emission (Froebrich et al., 2024a). By studying the variability of YSOs, it is possible to infer critical details about the structure and evolution of their circumstellar discs. Variability in YSOs can arise from several mechanisms, including changes in accretion rates, extinction due to disc inhomogeneities, and structural modifications influenced by planet formation processes (Lakeland & Naylor, 2022). Analyzing these variations across multiple wavelengths allows for the characterization of disc morphology, dust distribution, and the presence of substructures such as gaps and spirals, providing valuable insights into early stellar and planetary system development. The inner disc, within four astronomical units of the

YSO, cannot be resolved by telescopes for all but a few forming stars within 100 parsecs (Pott et al., 2010). A long term study of the photometric variability of the YSOs within these clusters can provide the statistics necessary to determine inner disc properties and insights into terrestrial planet formation.

2.2 The Observatories

The Hoys data are collected from a combination of amateur, university and professional observatories[1]. Nearly all are situated in the northern hemisphere with approximately 75% being in Europe but also some in North America. There are nine young clusters visible in the winter, nine targets for the summer as well as seven additional variable objects of interest on the HOYS target list [2]. Roughly 50% of the data are collected by three observatories. The university of Kent's Beacon observatory (BEACON), providing data totaling 44% of observing time and 10% of images. The Astrolab (IRIS) observatory contributing 10% of images and 3% of observing time and the Remote Observatory Atacama Desert (ROAD) providing 46% of images and 7% of observing time. The HOYS data is taken in six filters: Visual (V), Red (R), Infrared (I), Blue (B), Ultraviolet (U), and Hydrogen Alpha (Ha). The Hoys project is a survey designed to be carried out by small telescopes such as those used by Beacon, Iris and Road observatories. The Beacon observatory consists of a 17" Planewave Corrected Dall-Kirkham (CDK) Astrograph telescope situated at the University of Kent (51.296633° North, 1.053267° East, 69m elevation). It's CCD camera has a field of view of approximately 1° × 1°. The usable field of view of the detector is a circular area with a diameter of 1°, due to the vignetting of the corners of

[1] https://hoys.space/

[2] https://hoys.space/target-lists/

the detector (Froebrich et al., 2018). Images taken by the observatory for the HOYS project are currently taken in the following sequence: 180s integration for V filter, 120s integration for R and I filters repeated eight times. All images are dark and bias subtracted and flat-fielded using a set of 12 sky-flats. As well as the three observatories mentioned data are provided by other amateur astronomers via the HOYS website[3]. All images provided by the observatories are subject to photometric and colour calibration which is detailed in the next section.

2.3 Photometric and Colour Calibration

All images for the project are bias, dark, and flat corrected, uploaded to our publicly available web server[4], and preliminary calibrated. The astrometry in the images is solved using the `Astrometry.net` software (Hogg et al., 2008). Aperture photometry is performed on all images using the Source Extractor software (Bertin & Arnouts, 1996). A deep image obtained at photometric conditions is used as a reference for each region. The reference images for B, V, R & I filters are from the Beacon observatory whilst the U-band reference images are from the Thüringer Landessternwarte. By fitting a photo-function and 4th order polynomial (Bacher et al., 2005), (Moffat, 1969) to matching stars with accurate photometry, the calibrated magnitudes, f(m$_i$), for all images into the reference frames are obtained.

$$f(m_i) = A \cdot log(10^{B(m_i-C)} + 1) + \mathcal{P}_4(m_i) \tag{3}$$

[3] https://hoys.space/

[4] https://astro.kent.ac.uk/HOYS-CAPS/

where m_i is the instrumental magnitude and A, B and C are constants. $\mathcal{P}_4(m_i)$ denotes a fourth order polynomial and the other term is a photocurve function proposed by Bacher et al. (2005) and Moffat (1969).

Due to a significant fraction of amateur data using slightly different filters, in particular from digital single-lens reflex (DSLR) cameras, the calibration of the photometry needs to consider colour terms. Stars within the target region are chosen that are known to not vary in colour and brightness. Stars with a Stetson index I (Welch & Stetson, 1993) of less than 0.1 across the V, R & I filters were considered to be non-variable (Evitts et al., 2020). For these non-variable stars the median magnitude and colour is determined for all filters as reference for the calibration. A unique function $\mathcal{W}_N$(m,c) for each image (N) is determined to correct for any systemic magnitude offset caused by colour terms. Any filter can be chosen that the star is detectable in. For example, for V-I the functional form of the correction factor is a simple second order polynomial for both magnitude and colour with no mixed terms and a common offset $\mathcal{P}_0$, i.e:

$$\mathcal{W}_N(m, V-I) = p_0 + \mathcal{P}^2(m) + \mathcal{P}(V-I) \tag{4}$$

where $\mathcal{P}^2$ represents a second order polynomial without the offset (Evitts et al., 2020). The five free parameters for the correction function $\mathcal{W}_N$(m,c) are then determined. All non-variable stars detected in an image N are identified and the difference between their magnitude and real magnitude is determined. Any stars that have a larger than ± 0.5 mag difference and whose magnitude uncertainty is greater than 0.2 mag are removed. It is required for at least ten non-variable stars to be present in the image. The parameters for $\mathcal{W}_N$(m,c) are then determined by performing a least-squares optimization of these magnitude differences. A magnitude m$_i$ dependent weighting factor w$_i$ is introduced for each star i during the fitting process, due to the fainter non-variable stars far outnumbering the bright ones.

$$w_i = \frac{1}{(m_i - min(m_i) - 2)^2}, \tag{5}$$

where min(m$_i$) is the magnitude of the brightest star included in the fitting process. A three sigma clipping is used to ensure the fit is not influenced by misidentified objects or stars showing previously undetected variability. The median magnitude in V and I is then determined from all images taken within ±5 days of the observation to estimate a more representative uncertainty for the photometry after the correction of systematic effects. Where uncertainty is defined as the RMS scatter of the magnitude offsets of all calibration stars in the image which have the same magnitude (within ± 0.1 mag) as the star in question (Evitts et al., 2020).

2.4 Removing Photometry Outliers

Removing photometry outliers was carried out by H. Stokes-Geddes a PhD. student working alongside myself as part of the larger project (Ryan et al., 2025). As well as photometric and colour calibration our HOYS requires further data quality control. This section describes the process developed for cleaning the light curves of as many outliers as possible before the next stages of the data analysis were carried out. An example light curve of a variable YSO is shown in figure 5.

i) Given that the project is based on the statistics of variability of YSOs from a high-cadence, long-duration survey, we made a preliminary cut of the light curves with low cadences and insufficient time-series photometry data. B and

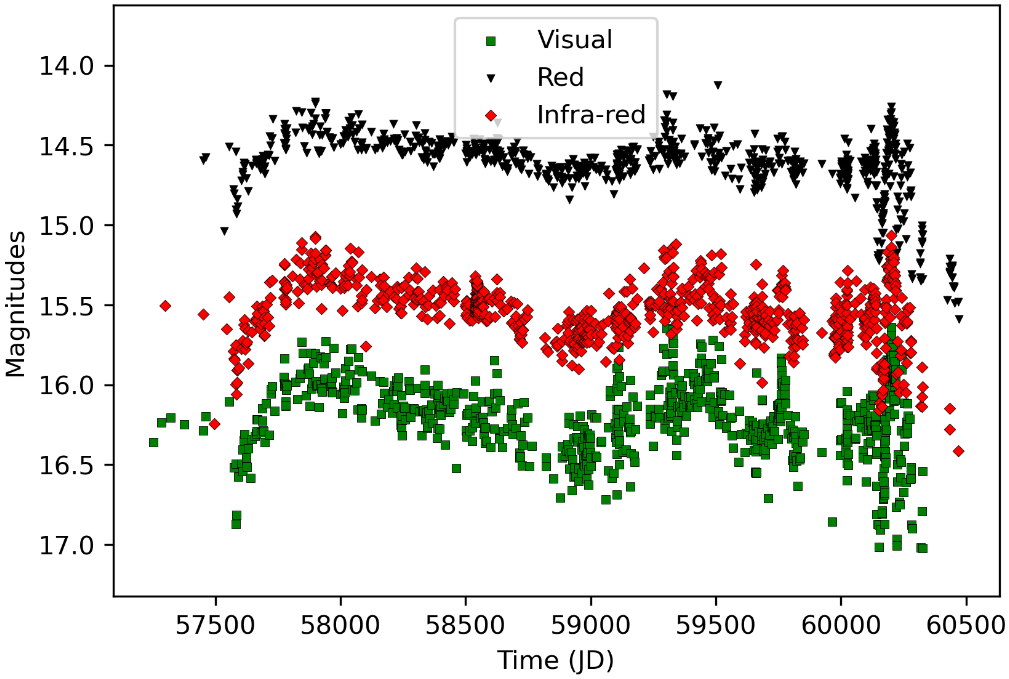

Figure 5: Light curve of the young star V350Cep situated in the NGC7129 cluster.

Hα filter data were discarded within our sample, as both filters do not meet the requirement for high cadence.

ii) We removed photometry within 5 arcmin of very bright stars ($Gmag <$ 6 mag) at observatories where PSF wings were found to influence the measurements, including the University of Kent's Beacon Observatory.

iii) Some stars appear to have two brightness measurements within the same image. This is likely due to either good seeing conditions revealing a binary system (apparent or real) or tracking issues during image capture. All such photometry has been removed.

iv) We are not interested in very short-term events such as flares. Thus, for each light curve and filter, we determine the mean and root mean square (RMS), see Equation 6, of all magnitude measurements. Any points deviating more than four times the RMS from the mean are removed.

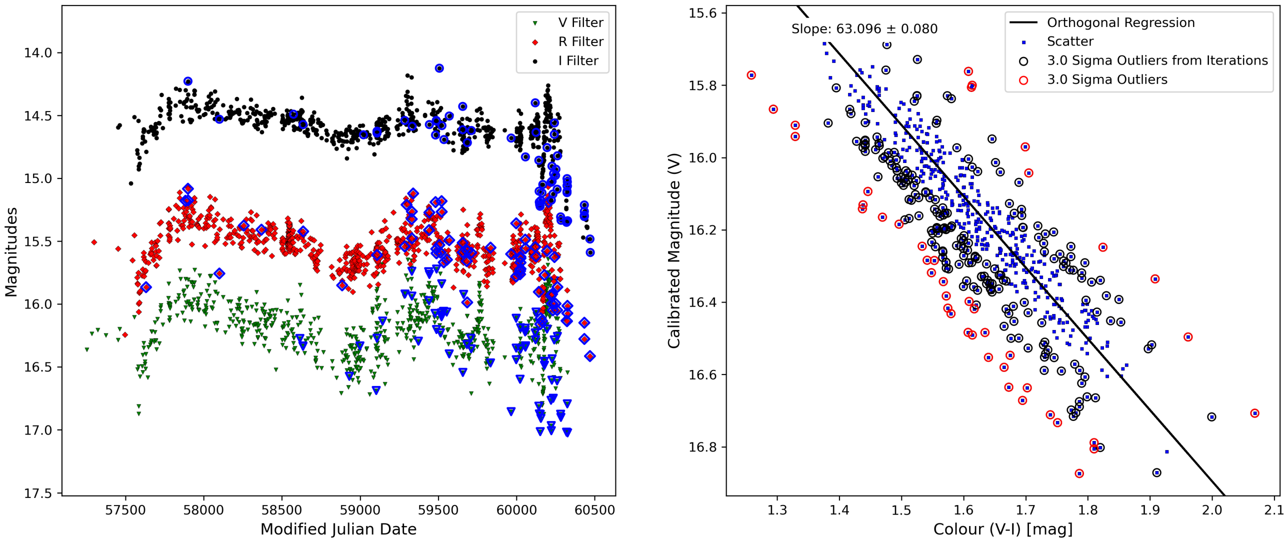

Figure 6: Left: light curve of V350Cep with outliers highlighted in cyan. Right: Example V-I vs V CMD plot of the same source.

$$\mathrm{RMS} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - \bar{m})^2} \tag{6}$$

v) For each light curve, a V vs. $R - I$ colour-magnitude diagram (CMD) is created. For each V magnitude measurement, the $R - I$ colour is computed as the difference between the average of all R and I measurements taken within two days of the V data. A linear perpendicular distance fit to the CMD is obtained, and the RMS scatter perpendicular to the best-fit line is determined iteratively. Any points deviating by more than three times this RMS value are removed from the light curve. This process is repeated for all possible combinations of the V, R, and I filters in the CMD, resulting in a total of nine iterations.

This process gave us a large set of light curves with filtered photometry data from which to select a smaller sample of variable YSOs.

2.5 Identification of a Sample of Variables

The above data extraction (performed on June 18^{th}, 2024), calibration, and 'cleaning' have been applied to a little over 3,000 light curves of potential HOYS cluster members identified in (Froebrich et al., 2024b). This includes all clusters, even if they were not analyzed in detail in that paper. For this study, we focus on a sample of clearly variable light curves to use in our analysis. We removed all objects that had fewer than 100 data points in V, R, and I after cleaning.

For each light curve, we computed the Welch-Stetson index, $\mathcal{I}$, (Stetson, 1996). This metric measures the correlation of variability between two different filters and is defined for the combination of V and I filters in equation 7. It requires N contemporaneous brightness measurements and their uncertainties in the two filters. These are calculated similarly to the colors in the CMDs over a time window of two days.

$$\mathcal{I}_{V,I} = \sqrt{\frac{1}{N(N-1)} \sum_{i=1}^{N} \left(\frac{V_i - \overline{V}}{\sigma_{V,i}} \right) \left(\frac{I_i - \overline{I}}{\sigma_{I,i}} \right)} \tag{7}$$

In this work, we focus only on data in the V, R, and I filters. Thus, we compute three variations of the Welch-Stetson index: $\mathcal{I}_{V,I}$, $\mathcal{I}_{V,R}$, and $\mathcal{I}_{R,I}$. In figure 7, we show how these values depend on the apparent magnitude of the sources. Most stars exhibit only low-level variability; however, some objects reach extreme $\mathcal{I}$ values, with the highest observed value being 1323. In our analysis, we found that 8%, 5%, and 6% of the $\mathcal{I}$ values exceeded five for $\mathcal{I}_{V,R}$, $\mathcal{I}_{V,I}$, and $\mathcal{I}_{R,I}$, respectively. To identify highly variable sources, we selected all stars with an $\mathcal{I}$ value above two in all three filter combinations. This selection results in a final sample of 240 highly variable sources.

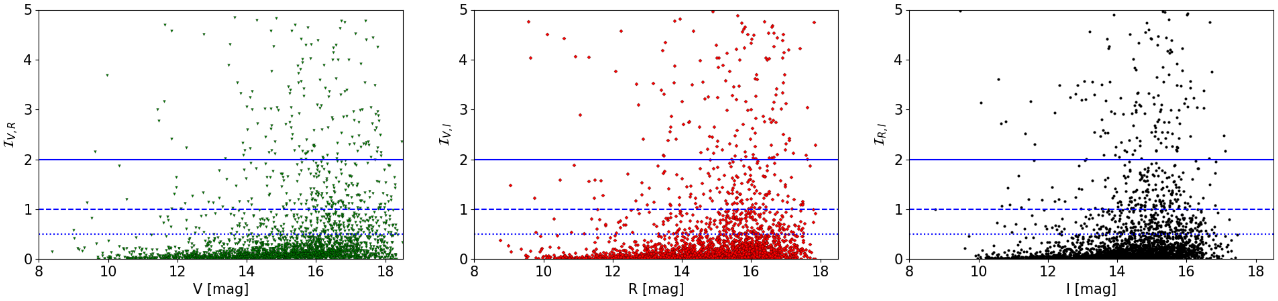

Figure 7: Welch-Stetson index of our YSO sample determined from all three filter combinations, $\mathcal{I}_{V,R}$, $\mathcal{I}_{V,I}$, and $\mathcal{I}_{R,I}$ from left to right. The cutoff for highly variable objects used in this study is marked at $\mathcal{I} = 2$ with a solid line. We also indicate $\mathcal{I} = 1$ and 0.5 with a dashed and dotted line, respectively.

2.6 From Light Curves to Variability Fingerprints

After data reduction, calibration, and the removal of photometric outliers, we now have a sample suitable for the next stage of our analysis. A light curve is a visualization of an object's brightness over time. The x-axis represents time, typically in days, while the y-axis represents brightness, measured in magnitude on a logarithmic scale.

While light curves are useful for identifying features of an object's variability across multiple filters, they are not directly comparable. They capture an object's behavior over different time spans and at varying cadences. In a project like HOYS, where data are collected from multiple observatories, observations occur at irregular cadences due to factors such as target visibility, weather, and observatory availability.

To directly compare the variability of multiple objects those objects need to have the statistics of their light curves represented in a standardized way. Work by Evitts et al. (2020) which improved upon work by Scholz & Eislöffel (2004), Findeisen et al. (2015) and Rigon et al. (2017) has shown that the statistics of light curves can be represented using a variability fingerprint. Figure 8 shows

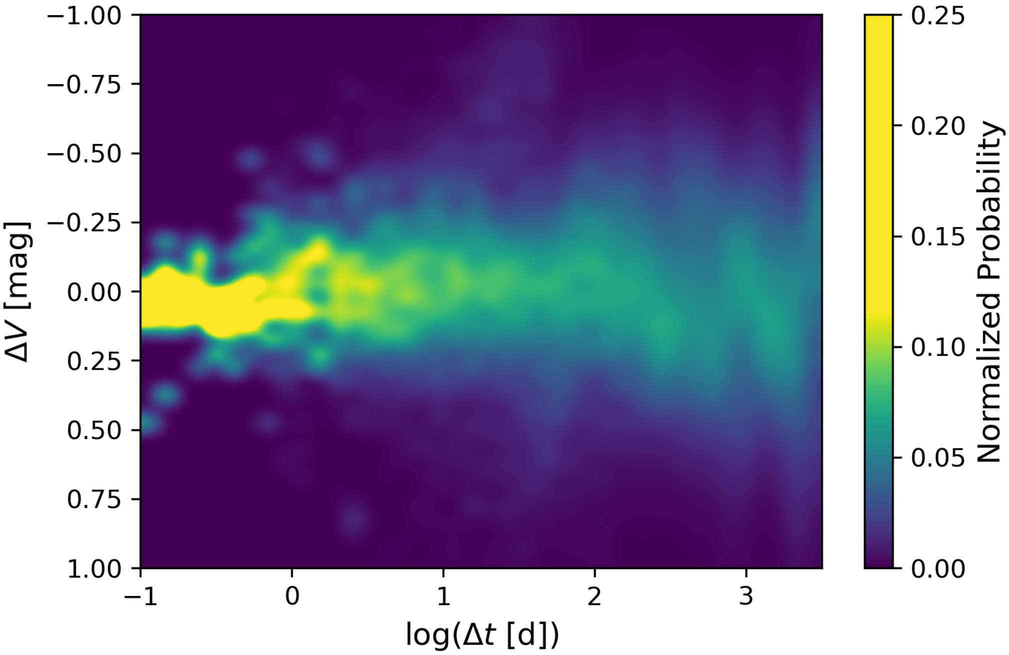

Figure 8: Smoothed variability fingerprint of V350Cep.

an example of a fingerprint of variability that is used in publications. These fingerprints provide a standardized way to visualize the probability that the object varies by a given magnitude for a set time period. Unlike figure 5 which shows the behavior of the object in three filters a fingerprint represents the variability of an object in just one filter. Although the dataset includes photometric measurements obtained using six filters, the B, U, and Hα filters were excluded at an early stage, as their low observational cadence was insufficient to provide reliable long-term statistics on variability. The V, R, and I bands were all suitable for this study in terms of data availability and cadence. For clarity and consistency, only results based on the V-band data are presented in this thesis. Preliminary tests applying the clustering workflows to the V, R, and I bands showed no significant qualitative or quantitative differences in the resulting groupings.

To create a fingerprint, for every combination of two points in figure 5 with a positive time difference, the difference between their calibrated magnitude and time is recorded and stored. These differences in calibrated magnitude are then plotted against the differences in time giving a plot with a very large number of points, see figure 9. Creating a large set of data for a given object allows statistics to be applied to the objects variability. The HOYS survey is not aimed at studying the variability of objects on timescales less than one day so although recorded in figure 9 are excluded from all further stages of creating the fingerprints. Any points that have a signal to noise ratio less than three are excluded based upon the recorded uncertainty values recorded at the time of observation.

A two-dimensional histogram is then created with the bin counts being determined from the points for that area in figure 9. The columns of the histogram are then normalized to give a probability for each bin that an object varies by a magnitude for a given time interval for one filter.

$$P(x, y) = N(x, y) / \sum_{i=1}^{\mathcal{M}} N(x, i) \tag{8}$$

No matter what the observing cadence or the number of observations made for an object a uniform way of visualizing the variability of that object is now created. Figure 10 shows how the fingerprints have been changed to attempt to capture as much of the objects variability across as much of the image area as possible while maximizing the signal to noise ratio of the fingerprint pixels. Top left is the 40x40 configuration that is usually used as a basis for the smoothed fingerprint released in previous publications shown in figure 8. Top right is the 20x20 configuration with the same time axis limits as top left. The bottom two panels in figure 10 both have a lower limit for the time axis of one day with left panel being a

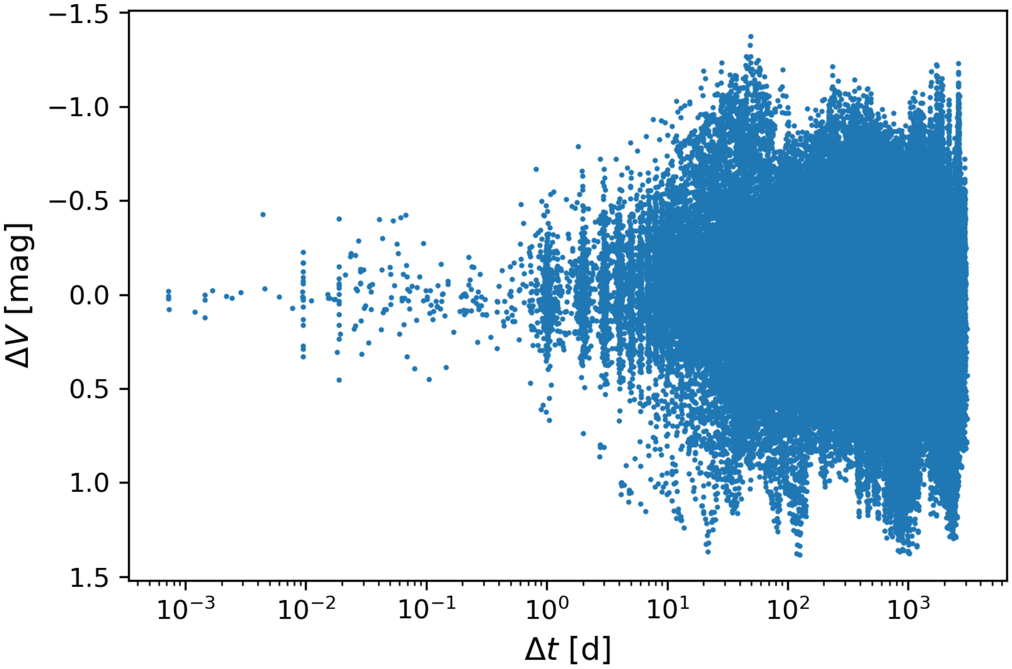

Figure 9: Plot of magnitude differences between light curve points with a positive time difference of V350Cep, V-band.

20x20 configuration and the right panel being the 9x16 pixel configuration. The columns for the 9x16 configuration fingerprint were created by starting with an interval of 0.05 magnitude either side of zero and then increasing the the amount of magnitude represented by 0.05 magnitude for each pixel away from the center of the plot. We will refer to this type of pixel configuration as adaptive pixels. A measure and representation of the uncertainty value for the fingerprint pixel was now needed. To create an error value for each pixel the magnitude values of the original light curve shown in figure 5 would be perturbed a large number of times to create the same number of fingerprints, sometimes referred to as 'bootstrapping'. These fingerprints would then be stacked so that a mean and standard deviation could be calculated for each pixel with the standard deviation being the absolute uncertainty which when divided by the mean would give you the relative uncertainty for each pixel, thus creating an uncertainty map for each

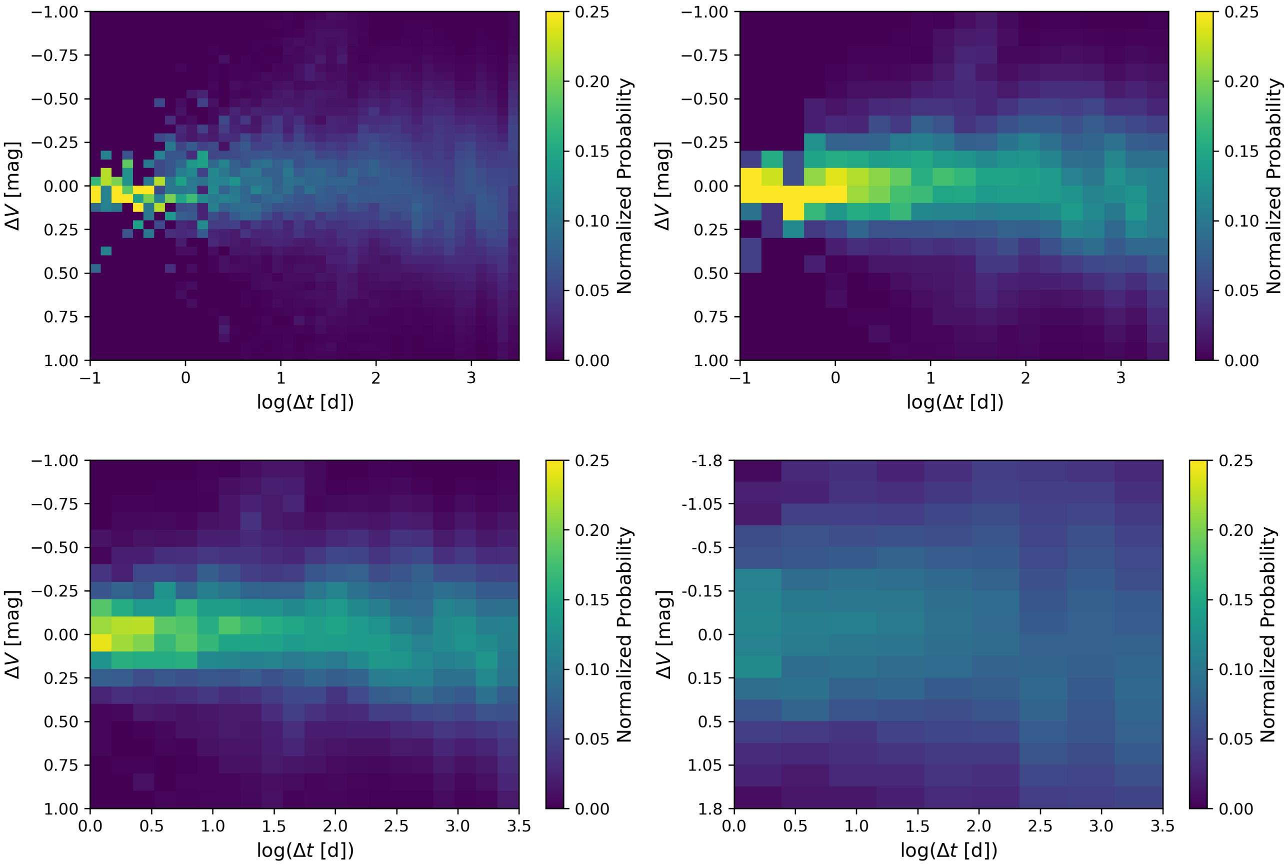

Figure 10: Four variants of fingerprints using V350Cep V-band data. Top left: 40x40 pixels, top right: 20x20 pixels, bottom left: 20x20 pixels with timescales less than one day excluding, bottom right: 9x16 adaptive pixels over same timescale as bottom left.

object.

This approach required the generation of a large number of fingerprints for each object, resulting in a significant computational cost. It was therefore necessary to minimise the number of fingerprints produced per object and to optimise the fingerprint-generation code. To improve efficiency, the fingerprints and their associated uncertainty maps were processed and stored directly as numerical arrays, enabling the workflow to scale to large sample sizes and repeated resampling. A revised, optimised version of the fingerprint-generation code was developed specifically for use within the bootstrapping pipeline. The resulting performance improvements are illustrated in Table 3, which demonstrates the time savings achieved by replacing a loop-based column normalisation with a vectorised matrix operation. To determine the optimal number of fingerprints to stack during the bootstrapping process, a test stack of 150,000 fingerprints was created for a single object. Several pixel positions were randomly selected, and the mean pixel value was computed for different stack sizes drawn from this set. The cutoff for the bootstrapping process was set at the point where the mean pixel value stabilized within one standard deviation of its final value. A random example of this convergence is shown in Figure 11, with the mean pixel value for $N = 150,000$ indicated by the red dashed line.

Although bootstrapping is a reliable method for finding the uncertainty of the fingerprint pixel values it places time constraints on producing error maps for the fingerprint of any given object especially for objects that have a large amount of data points in their light curves. Working on the assumption that the counts for each pixel in the original fingerprint before normalization would adhere to Poisson counting statistics and the uncertainty of the counts could be expressed

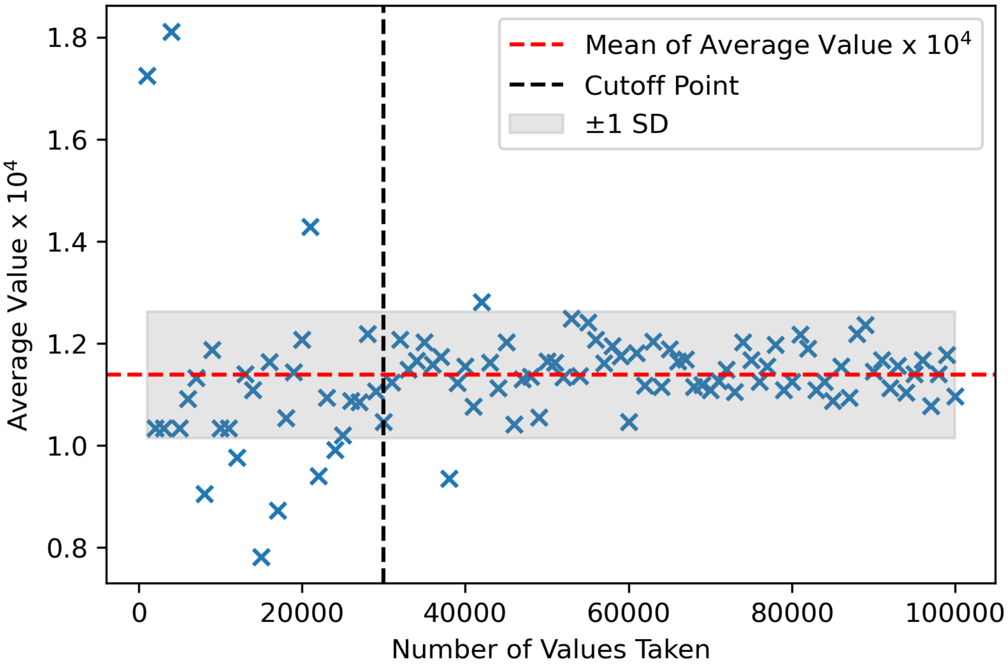

Figure 11: Plot of mean fingerprint pixel value against number of fingerprints in a stack for a randomly chosen pixel.

as $\sqrt{N_i}$ where N_i is the count of a given pixel i. An error propagation could then be carried out to give an uncertainty value for the normalized values in the fingerprints which represent the probability that an object varies in brightness by a certain magnitude for a given time.

The relative uncertainty values for each pixel in the map created using Poisson counting statistics were calculated using the following equations;

$$\frac{\Delta P(x,y)}{P(x,y)} = \sqrt{\frac{1}{N(x,y)} + \left(\sum_{i=1}^{\mathcal{M}} N(x,i)\right)^{-1}} \tag{9}$$

Where $N(x,y)$ is the count for any given pixel x, y and the column is given by i giving an absolute uncertainty of $\Delta P(x,y)$ from which we can obtain the relative uncertainty. Any pixels that return a value of infinity due to a division of zero are given a value of one for plotting while maintaining its original value

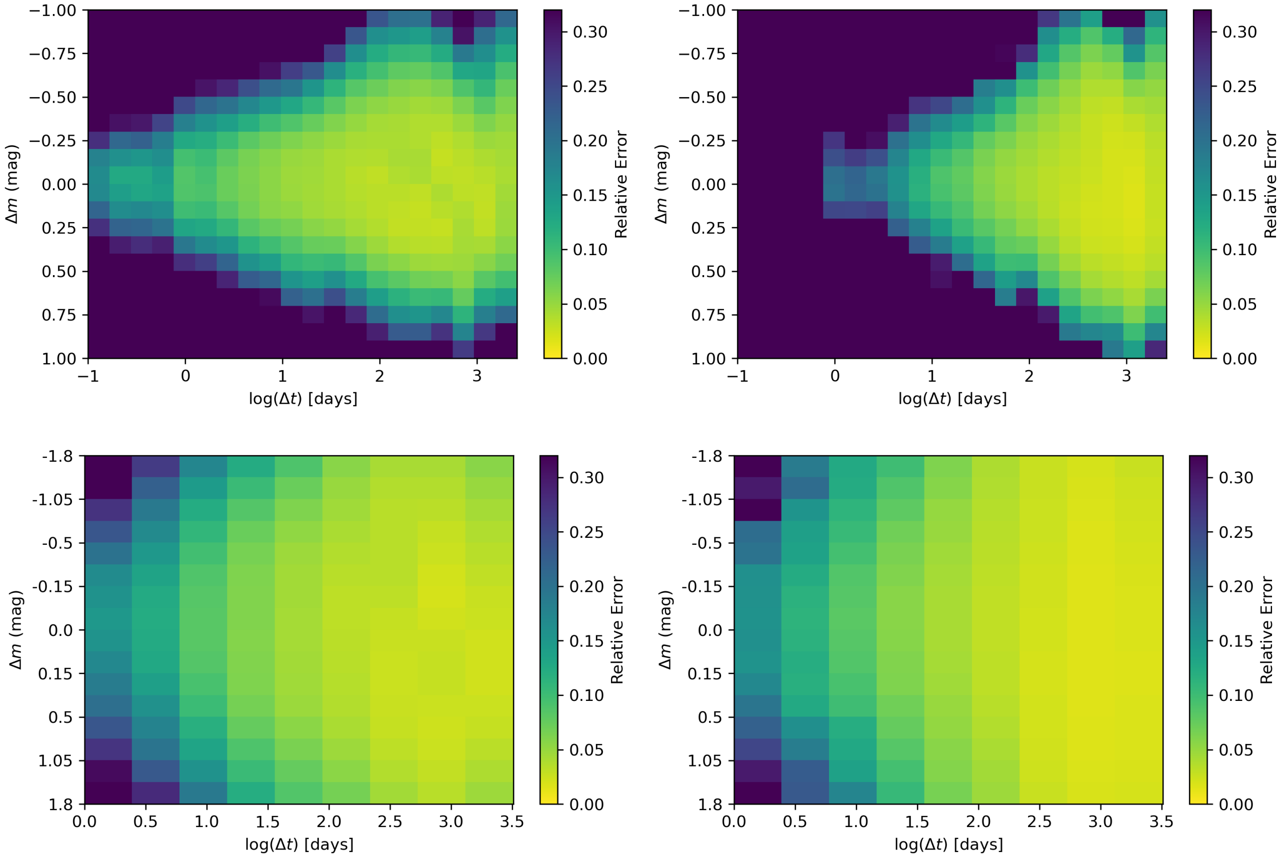

Figure 12: Comparison of bootstrapped (left) and Poisson (right) uncertainty maps for 20x20 pixel configuration (top) and 9x16 pixel configuration (bottom).

in the array. Figure 12 shows that the uncertainty maps for the Poisson and bootstrap methods are similar for both pixel configurations. Also by using the adaptive pixels the uncertainty values have been reduced for large parts of the fingerprint area with only the left hand top and bottom corners having significant uncertainty values. Given that the Poisson method would save considerable time when producing uncertainty maps it was important to know if the Poisson method could be exchanged directly for the bootstrapping method. The uncertainty value for each pixel using the Poisson method was plotted against its corresponding pixel using the bootstrapping method. Linear regression was then applied to give a slope value an example of which can be seen in figure

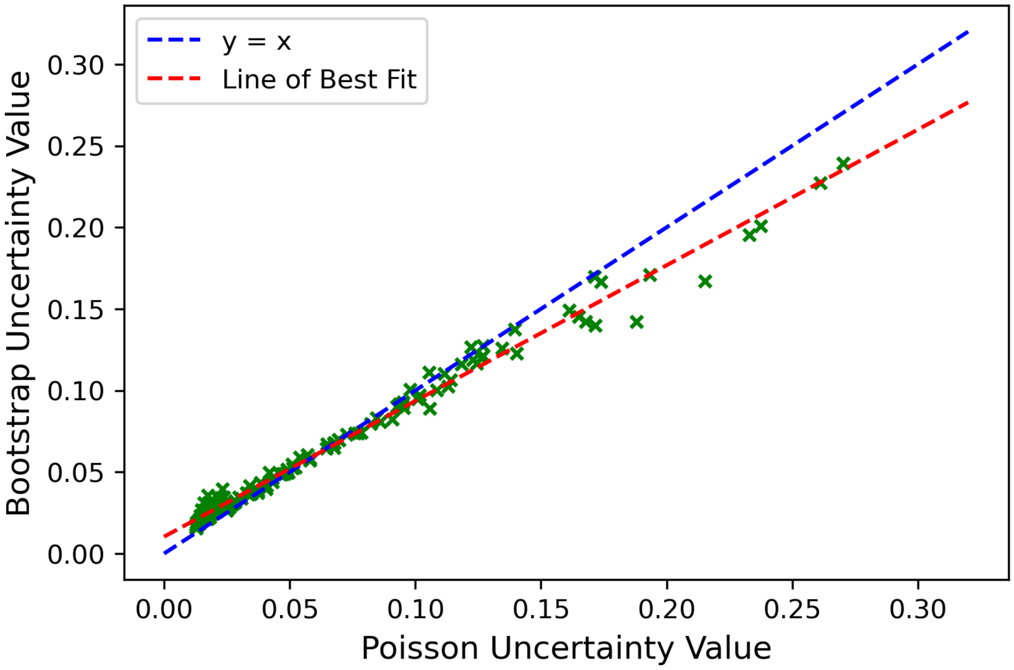

Figure 13: Comparison of uncertainty values of V350Cep V-band data, slope value of 0.83.

13. If the slope value was one or very close to one then the error calculation method described in equation 9 could be exchanged directly for the bootstrapping method. The motivation being to save computation time. As HOYS data consists of over 3000 objects in multiple filters calculating error maps using the bootstrapping method requires significant computation time. The comparison of uncertainty values was made for every object for a given configuration and filter and a mean slope and its standard deviation were calculated. Any pixel values with an uncertainty value greater than one third were excluded from the regression calculation. Any objects that had fewer than two thirds of their pixels with an uncertainty value less than one third were excluded from the calculation of the mean slope. The full results of the mean slopes for each pixel and filter configuration can be found in table 4. Given the values calculated for the average slopes and their standard deviation substituting the low compute Poisson method for the higher compute bootstrapping method would depend on the cir-

cumstances. The values are close enough that under time constraints one could be substituted for the other but the difference between the values show that the high compute method is preferable to obtain accurate uncertainty values. There is also some structure to the difference between the values. Figure 14 shows that when a plot is made of the uncertainty values below 0.025 the bootstrapped values are higher than those of the low compute Poisson method but have a slope of 1.020 which is very close to one. It is worth noting that the line of best fit does not imply anything about the uncertainty values beyond the range of the calculated values. The line of best fit is being used to obtain a gradient between the two methods and highlight this graphically. In figure 14 the line of best fit intercepts the y-axis above zero this does not imply that the bootstrapping method has an uncertainty value when the Poisson uncertainty value is zero.

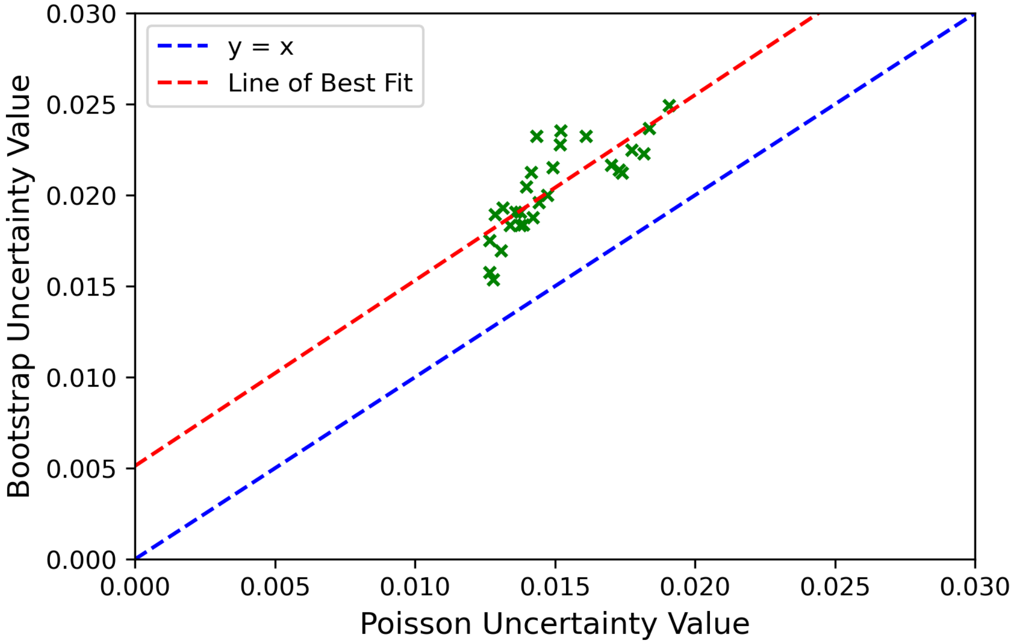

Figure 14: Comparison of uncertainty values of the young star V350Cep for low uncertainty values, slope value of 1.02.

We have now established a fingerprint configuration that yields a high signal-to-noise ratio in each pixel, together with a reliable method for calculating the corresponding pixel-level uncertainties. The variability of each object can therefore be represented using a uniform fingerprint in which a large fraction of the pixels have low uncertainty. In practice, this means that the vast majority, over 90% of the pixel values in the 9×16 adaptive fingerprints used to describe an object's variability are reliable. Constructing fingerprints in this way provides the best opportunity to apply clustering algorithms effectively, allowing meaningful insight into how changes in the variability of simulated objects influence their relationship to the observed properties of real objects.

3 Comparing Stochastic Light Curves Using Clustering Algorithms

This section presents the methodology and results of applying clustering algorithms to characterize the fingerprints developed in the previous section. The goal is to use these algorithms to generate a stable visualization of the fingerprint landscape. If a light curve can be simulated, a corresponding fingerprint can be produced and compared to the HOYS dataset. To facilitate this comparison, a model fingerprint will be created using an artificial light curve. This model fingerprint will then be incorporated into the visualization to assess its relationship to the HOYS data. This approach may be useful for testing the outputs of simulations related to planet formation and the disc structures surrounding YSOs. Insights into the HOYS data as a whole or subsets of the data and how they relate to each other can also be gained.

The two clustering algorithms tested are DBSCAN and K-means. Both are iterative methods but differ in their approach to cluster membership. K-means determines cluster membership based on proximity to centroids, whereas DBSCAN uses core and border points to define clusters. Principle component analysis (PCA) was used as a dimension reduction tool prior to using both algorithms. T-Distributed Stochastic Neighbor Embedding (t-SNE) was used as part of the DBSCAN workflow. See section 1 for details on any algorithms and processes used in this section. Before the algorithms can be tested some analysis has to be done to determine which of the fingerprint configuration is most suitable for use in the clustering process. The most important metric being how much of the data sets variability can be preserved during the dimension reduction process.

The clustering algorithms have a number of parameters that need to be determined. While the parameters can be chosen arbitrarily within a sensible range there are tests that can be done to determine the optimal value for parameters. In the case of DBSCAN the tests will return a Davies-Bouldin index which will help in determining these parameters. This index will also give a measure of success of the clustering process using DBSCAN which will form part decision on algorithm selection.

3.1 Algorithms and Processes Used in Clustering Workflow

This subsection outlines the processes and algorithms used in the clustering workflow, along with the rationale for their selection. While other methods and configurations were tested, they did not produce significantly different results, either qualitatively or quantitatively. These tests explored alternative fingerprint configurations and different sequences of dimensionality reduction during both the pre-analysis clustering phase and the final workflow.

The methods and results presented here focus on the final workflow, which employs the optimal fingerprint configuration and clustering algorithm. Additionally, this subsection includes an example of a workflow that was found to be unsuitable for comparing YSO variability.

3.1.1 Standard Scaler

Standard scaler is a process used to standardize features by removing the mean and scaling to unit variance, by transforming each component using the equation:

$$x = \frac{P(x, y) - \mu}{\sigma} \tag{10}$$

where x is the component value, μ is the mean of component value, σ is the standard deviation of component value and z is the new standardized value. This results in the new component having a mean of zero and standard deviation of one. Standard scaler is sensitive to outliers with variance more than one order of magnitude. Given our components are normalized between zero and one this makes it a suitable tool to scale our components. (Hastie et al., 2009)

3.1.2 Principle Component Analysis (PCA)

The two fingerprint configurations being considered for use in the clustering workflow are the 20x20 linearly spaced pixels and the 9x16 adaptive pixels, see figure 10. Both of these configurations have extremely high dimensionality 144 dimensions/features in the case of the 9x16 adaptive pixel and 400 dimensions/features for the 20x20 pixel fingerprint. Without reducing the number of features down to a small number of components it is not viable to compare our 240 sample of YSOs within a visualization. PCA is an unsupervised dimensionality reduction technique used on data that have high dimensionality but linear structure. It requires no prior knowledge of the dataset and will reduce the dimensionality while preserving the variance within the linear structure (Scitovski et al., 2021). It was first developed and introduced by Harold Hotelling in the September and October issue of the Journal of Educational Psychology (Hotelling, 1933a), (Hotelling, 1933b). The technique is used in clustering workflows and was a key step in the workflow used in this project. For all of the applications of PCA used in the project the built in function within the SciPy Python module was used[5]. The general method of PCA is outlined below for a

[5] https://scikit-learn.org/stable/modules/generated/sklearn.decomposition.PCA.html//

case where there are three dimensions/features reduced to two principal components.

The first process is to center the data around zero, in our case this is handled by the standard scaler described in section 3.1.1. After this variances and covariances of each feature are calculated to construct the covariance matrix. Variance measures how much a set of values deviates from the mean. The variance of a random variable X is given by:

$$\mathrm{Var}(X) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{11}$$

where X_i are the individual data points. $\bar{X}$ is the mean of X, given by:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{12}$$

and n is the total number of observations. Covariance measures how two random variables X and Y vary together. The covariance between X and Y is given by:

$$\mathrm{Cov}(X, Y) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) \tag{13}$$

where X_i and Y_i are individual data points for variables X and Y. $\bar{X}$ and $\bar{Y}$ are the respective means:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{14}$$

and n is the total number of observations.

To understand how the variables relate, we calculate the covariance matrix:

$$\Sigma = \begin{bmatrix} \text{Var}(x) & \text{Cov}(x, y) & \text{Cov}(x, z) \\ \text{Cov}(y, x) & \text{Var}(y) & \text{Cov}(y, z) \\ \text{Cov}(z, x) & \text{Cov}(z, y) & \text{Var}(z) \end{bmatrix} \tag{15}$$

This symmetric 3 × 3 matrix describes the variance along each axis and how the variables correlate.

To find the principal component directions, we solve the eigenvalue equation:

$$\Sigma v = \lambda v \tag{16}$$

where v are the eigenvectors which denote the directions of the principal components and λ are the eigenvalues which represent the amount of variance captured by each eigenvector. For data that has three dimensions/features there will be three orthogonal solutions to equation 16. In PCA eigenvalues come from the covariance matrix, which is always symmetric and positive semi-definite. Therefore all eigenvalues in PCA are non-negative. The two solutions with the largest eigenvalues are selected. These two solutions will have eigenvectors that indicate the directions of maximum variance in the data (Hotelling, 1933a) (Hotelling, 1933b). Figure 15 shows data with three dimensions/features with the principal components as vectors which indicate the directions of maximum variance across the data. Once the two eigenvectors corresponding to the largest eigenvalues are selected the loadings matrix is formed:

$$V_{\text{selected}} = \begin{bmatrix} v_1 & v_2 \end{bmatrix} \tag{17}$$

where v_1 and v_2 are the eigenvectors for PC1 and PC2. Now we have every-

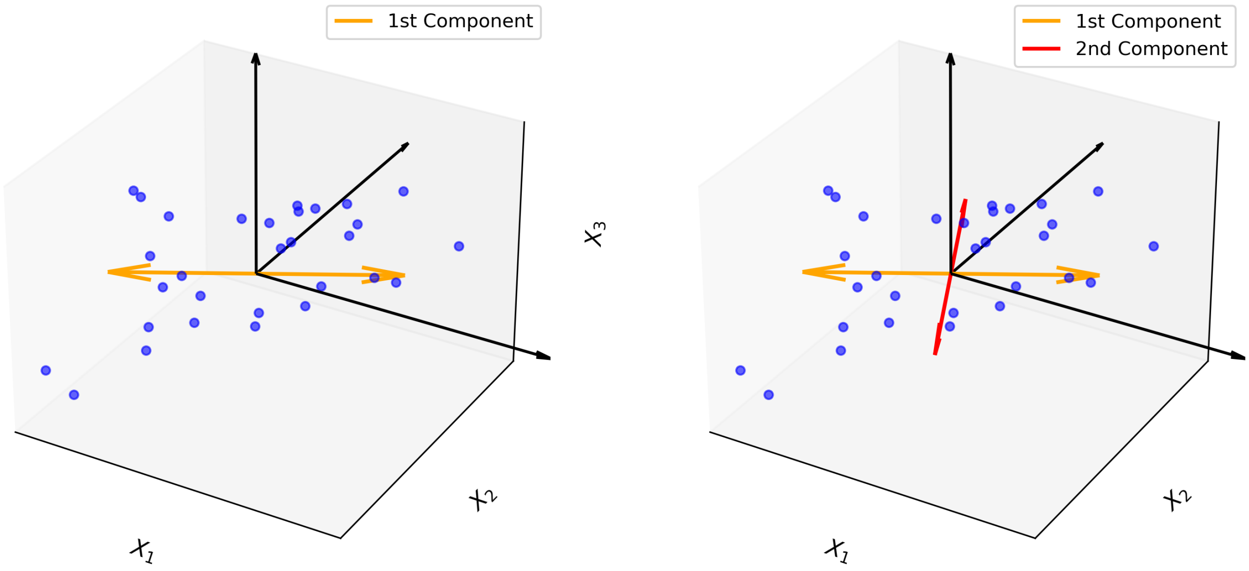

Figure 15: Two directions of greatest variance belonging to a set of data with three dimensions.

thing necessary to transform our data to a new component space where PC1 and PC2 are a linear combination of our original pixel values. In the general case of the three dimensions reduced to two given:

$$v_n = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \quad \text{and} \quad X_n = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \tag{18}$$

$$\begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \text{component value} \tag{19}$$

$$\Rightarrow \text{component value} = a_1x_1 + a_2x_2 + a_3x_3 \tag{20}$$

Through the PCA process we have effectively transformed our coordinate set, see figure 15, such that one of the axis lie across the largest extent of the data, this is component one. The orthogonal direction which indicates the next largest amount of variation forms component two. This method of dimension reduction captures the largest amount of variance in the data while preserving its linear structure (Murty & Devi, 2016). Our data has a high number of dimensions ranging from 144 for the 9x16 adaptive pixel fingerprints to 1600 for the 40x40 fingerprints. It is next required to evaluate how much of the variance can be maintained after reducing the dimensions down to just two. The generally recommended amount of variance to be captured by the first two components is 0.80-0.95 Jolliffe (2002) of the total variance. If an amount of variance of 0.80 can be maintained after PCA then the clusters should have a good separation between them and be fairly dense. Given the work outlined in previous sections regarding analysis of the fingerprints the 40x40 configuration was discarded. The variance maintained after PCA dimension reduction for the 9x16 adaptive and 20x20 evenly sized configurations is investigated in the next section.

3.1.3 Variance Maintained After Dimension Reduction

Our data set has a high degree of dimensionality. Each fingerprint pixel represents a dimension, 144 in the case of the 9x16 adaptive pixel configuration and 400 in the case of the 20x20 configuration. Reducing those dimensions prior to clustering can improve the outcomes and is essential for effective visual representation. Dimension reduction can result in more distinct clusters and stable positions of the objects in the visualizations. Without dimension reduction a meaningful visualization of our data is not possible as it may not represent the dimensions upon which the cluster allocations were made correctly. The type of dimension reduction implemented on our data set was Principle Component Analysis (PCA), see section 3.1.2 for details on any algorithms and processes

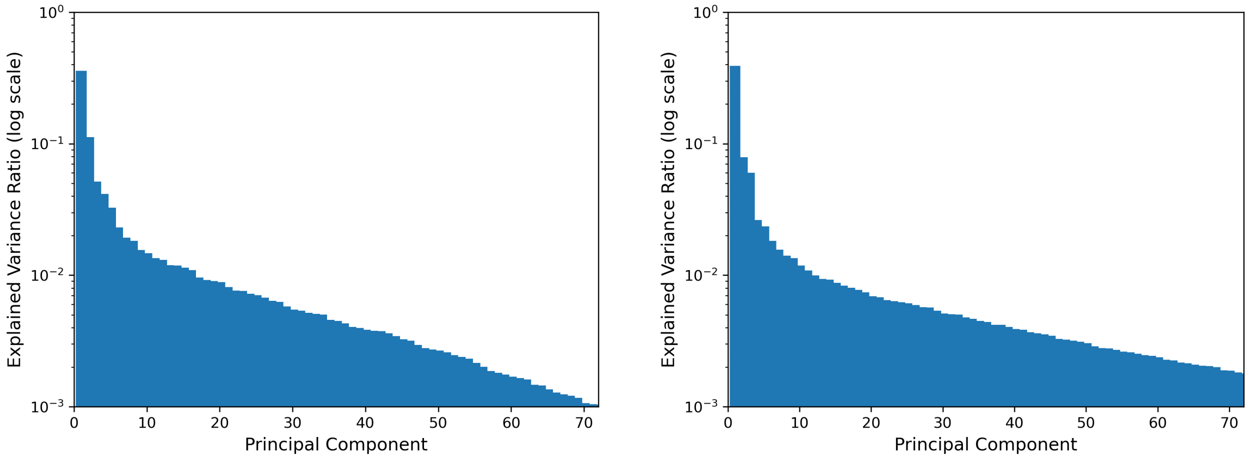

Figure 16: Fraction of total variance captured by components after PCA. Left: 9x16 fingerprints, Right: 20x20 fingerprints.

used in this section. Applying PCA to our data set for the 9x16 and 20x20 configurations we are able to determine how much of the variance of the data set can be represented by two dimensions. The pixel values of the fingerprints are scaled using the Standard Scaler then the variance captured by each pixel is calculated. The variance captured can then be plotted on a logarithmic scale, see figure 16. The first two components of the 9x16 fingerprints capture 0.48 of the total variance and of the 20x20 fingerprints captures 0.47 of the total variance. As the fraction of the variance is very similar for both configurations the clustering outcomes should not be affected by choosing one type of pixel configuration over the other. This allows us to select the pixel configuration with the lowest uncertainty values, shown in figure 13, the 9x16 adaptive pixel configuration. This amount of variance captured by the first two components is lower than would be expected to produce distinct clusters. The generally recommended amount of variance to be captured by the first two components is 0.80-0.95 Jolliffe (2002) of the total variance. Figure 16 shows the majority of the variance preserved within that 0.48 of total variance is represented by the first component. Figure 16 shows only the first 70 components for clarity of presentation. The second component preserving around 0.1 of the total variance. For further components the contribution to preserved variance reduces significantly. Using more than two components would not add a significant amount of variance in either case. Although the clustering algorithms are able to allocate cluster membership based on any number of dimensions our aim is represent the data in a 2D visualization.

3.2 Clustering Algorithms

Clustering is a way of separating different patterns into groups based upon the characteristics of the patterns. Representations or descriptions of the clusters formed are used in decision making. Clustering is carried out so that the patterns grouped together in one cluster are similar and different from the other clusters generated by some quantifiable measure. It is useful to represent patterns in 2D space to enable the clustering process, it is then possible to use a method such as the squared Euclidean distance between the points to group in clusters based on some threshold (Zollanvari, 2023). The squared Euclidean distance between two points x_i and x_j is defined as:

$$d(x_i, x_j) = \sum_{l=1}^{d} (x_{il} - x_{jl})^2 \tag{21}$$

where d is the dimensionality of the points. (Murty & Devi, 2016) Figure 17 shows an example of clustering where each pattern is clustered with its nearest neighbour, this is an example of hard clustering where there are clearly defined boundaries between the clusters. If a pattern belongs to more than one cluster

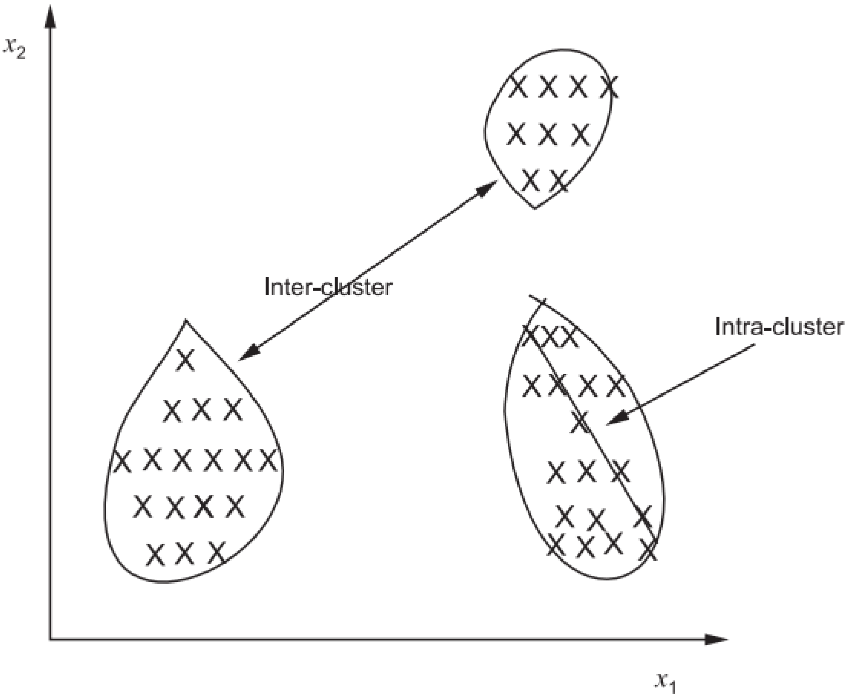

Figure 17: An example of using 2D Euclidian clustering where the intra-cluster distance is minimized (Murty & Devi, 2016).

then this is known as soft clustering. Clustering is useful for representing large amounts of data by using representatives of a cluster such as a centroid or a medoid. The centroid is defined as the sample mean of the points in the cluster and the medoid is that point in the cluster from which the sum of the distances from the points in the cluster is the minimum. Points that are far off from any other points in the cluster should be labeled as outliers. The centroid of the data can shift without any bound based on the location of an outlier or outliers whereas the medoid cannot therefore clusters that use medoids are more robust and less influenced by noisy patterns or outliers (Murty & Devi, 2016).

Clustering is a very important tool as the number of cluster representatives is smaller than the number of patterns in the original data hence there is data reduction and clusters and their descriptions can be used in decision making processes such as classification or prediction.

3.3 K-means Algorithm

A widely used clustering algorithm which can generate a hard partition of the data is the k-means algorithm, an outline of the steps involved in this algorithm are as follows

i) Select k out of the given n patterns as the initial cluster centres. Assign each of the remaining n-k patterns to one of the clusters by assigning each pattern to its closest centre.

ii) Compute the cluster centres based on the current assignment of patterns.

iii) Assign each of the n patterns to its closest centre.

iv) If there is no change in the assignment of patterns to clusters during two successive iterations then stop, else return to the second step. (Murty & Devi, 2016)

The k-means algorithm was first developed by Stuart Lloyd in 1957 for an internal Bell labs technical report but was not formally published until 1982 Lloyd (1982). Lloyd originally developed it for vector quantization but k-means was independently described and popularized for clustering by James MacQueen in 1967 (MacQueen, 1967). For all of the applications of k-means used in the project the built in function within the SciPy Python module was used[6].

K-means aims to partition data into k clusters by minimizing the sum of squared distances between each point and its assigned cluster centroid:

$$J = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \mu_i\|^2 \tag{22}$$

where k is the number of clusters, C_i is the set of points in cluster i, x is a

[6] `https://scikit-learn.org/stable/modules/clustering.html#k-means//`

data point, μ_i is the centroid (mean) of cluster i, $\|x - \mu_i\|^2$ represents the squared Euclidean distance. Each point x is assigned to the cluster whose centroid is closest:

$$c_j = \arg\min_i \|x_j - \mu_i\|^2 \tag{23}$$

where c_j is the cluster assigned to point x_j and μ_i is the centroid of cluster i. After assigning all points to clusters, the centroids are updated:

$$\mu_i = \frac{1}{|C_i|} \sum_{x \in C_i} x \tag{24}$$

where μ_i is the new centroid for cluster i and $|C_i|$ is the number of points in cluster i.

$$\|\mu_i^{(t+1)} - \mu_i^{(t)}\| < \epsilon \tag{25}$$

where ϵ is a small threshold and t is the iteration number. K-means stops when either the centroids do not change position significantly or a maximum number of iterations is reached (MacQueen, 1967).

3.4 Density Based Clustering with Noise (DBSCAN)

DBSCAN is a clustering algorithm that groups together points that are closely packed while marking outliers as noise. Unlike k-means, it does not require specifying the number of clusters beforehand. Instead, it relies on two parameters ϵ, which defines the radius around a point and the minimum number of points (MinPts) required to form a relatively dense region. Points within an ϵ radius of a dense region are added to the same cluster, while isolated points are classified as outliers (Zollanvari, 2023). This makes DBSCAN especially useful for detecting clusters of varying shapes and handling noisy data effectively. DBSCAN was first introduced in work by Ester et al. (1996). For all of the applications of DBSCAN used in the project the built in function within the SciPy Python module was used[7]. An outline of the general method of DBSCAN follows below. Defining the Neighborhood DBSCAN relies on a density-based notion of clusters. The ε-neighborhood of a point x is defined as:

$$N_\varepsilon(x) = \{y \in X \mid d(x, y) \leq \varepsilon\} \tag{26}$$

where $d(x, y)$ is the distance (usually Euclidean) between points x and y and ε is the neighborhood radius. Each point in the data is classified based on its local density as core, border or noise points. A core point must have at least (MinPts) points including itself in its ε-neighborhood:

$$|N_\varepsilon(x)| \geq \text{MinPts} \tag{27}$$

A border point is classified as a point that has fewer than MinPts points in its ε-neighborhood but is reachable from a core point. While a noise point is neither a core nor a border point. A point p is directly density-reachable from q if:

$$p \in N_\varepsilon(q) \quad \text{and} \quad |N_\varepsilon(q)| \geq \text{MinPts} \tag{28}$$

A point p is density-reachable from q if there exists a sequence of points $p_1, p_2, \ldots, p_n$ such that $p_1 = q$, $p_n = p$, and each p_{i+1} is directly density-reachable

[7] `https://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html//`

from p_i (Ester et al., 1996).

A cluster is then defined as the set of all points that are density-connected to at least one core point. DBSCAN clusters are formed as follows:

i) Select an unvisited point p.

ii) If p is a core point, a new cluster is started, and all density-reachable points are added.

iii) If p is a border point, it is assigned to an existing cluster.

iv) If p is a noise point, it remains unassigned to a cluster.

v) This is repeated until all points are visited with the stopping condition all points have assigned to clusters (Core or border points) or noise points (outliers) (Hahsler et al., 2019).

3.5 t-Distributed Stochastic Neighbour Embedding (t-SNE)

t-SNE is a nonlinear dimensionality reduction technique that preserves local structure in high-dimensional data while embedding it into a lower-dimensional space. It was first introduced in work by van der Maaten & Hinton (2008).t-SNE defines a probability distribution over pairs of points in the high-dimensional space, modeling their similarity using a Gaussian distribution. The probability that point x_j is a neighbor of point x_i is given by:

$$p_{j|i} = \frac{\exp\left(-\frac{||x_i - x_j||^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(-\frac{||x_i - x_k||^2}{2\sigma_i^2}\right)} \tag{29}$$

where σ_i is the perplexity-controlled bandwidth of the Gaussian kernel around x_i, $||x_i - x_j||^2$ is the squared Euclidean distance between points x_i and x_j. The symmetric joint probability is then defined as:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{30}$$

where N is the total number of points. In the low-dimensional space, t-SNE models pairwise similarities using a Student's t-distribution with one degree of freedom (which is a Cauchy distribution):

$$q_{ij} = \frac{(1 + ||y_i - y_j||^2)^{-1}}{\sum_{k \neq l}(1 + ||y_k - y_l||^2)^{-1}} \tag{31}$$

where $||y_i - y_j||^2$ is the squared Euclidean distance between the low-dimensional embeddings of x_i and x_j. The heavy-tailed Student's t-distribution prevents the "crowding problem." The objective of t-SNE is to minimize the difference between the probability distributions $P = \{p_{ij}\}$ and $Q = \{q_{ij}\}$. This is achieved by minimizing the Kullback-Leibler (KL) divergence:

$$C = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{32}$$

This cost function ensures that similar points in the high-dimensional space remain close in the low-dimensional space. To update the low-dimensional coordinates y_i, gradient descent is applied using the following gradient:

$$\frac{dC}{dy_i} = 4 \sum_j (p_{ij} - q_{ij})(y_i - y_j)(1 + ||y_i - y_j||^2)^{-1} \tag{33}$$

This gradient ensures that points with high p_{ij} (i.e., close neighbors) are moved closer together in the low-dimensional space.

The optimization process continues until the KL divergence converges (i.e., there is little improvement in cost function C). A maximum number of iterations is reached van der Maaten & Hinton (2008).

3.6 Clustering Algorithm Parameters

The two clustering algorithms chosen for use with our dataset require the specification of parameters prior to execution. For the k-means algorithm, the primary parameter is the number of clusters. As the k-means algorithm completes in a relatively short amount of time on our dataset (within minutes), this parameter can be determined empirically by running the algorithm with a range of cluster values and examining the resulting visualizations and cluster memberships. Given that the dataset represents variable YSOs with differing variability profiles, a value of five clusters was selected to capture the diversity of behaviours, including those that do not fit neatly into specific categories.

For DBSCAN, three key parameters must be selected. The first is ε, which defines the radius of the neighbourhood around a point. The second is the minimum number of objects required to form a cluster, referred to as *minimum samples*. These two parameters are intrinsic to the DBSCAN algorithm. The third parameter is *perplexity*, which pertains to the t-SNE dimensionality reduction technique used prior to clustering. Low perplexity values emphasize the preservation of local structure over global relationships in the data (Hahsler et al., 2019).

Optimization of ε and perplexity was achieved by evaluating the Davies-Bouldin Index (DBI) across a sensible parameter space. For ε, values between 1 and 20 were tested. The commonly accepted range for perplexity is between 5 and 50. These parameters were evaluated at multiple intervals of *minimum samples*, ranging from 5 to 50. The full results, including DBI scores and their associated parameters, are shown in Table 1. The DBI provides a quantitative assessment of clustering quality, with lower values indicating more compact and well-separated clusters. According to Davies & Bouldin (1979), a DBI greater than 2 may indicate poorly separated or diffuse clusters, a range of 0.5–1.5 suggests moderately well-separated clusters with some overlap or density variation, and values between 0.1–0.5 imply highly compact and distinct clusters.

The minimum DBI achieved was just over 0.45, for two distinct values of *minimum samples*, 20 and 35. For *minimum samples* 40–50 the requirement of forming more than one cluster was not met. This indicates a strong clustering result, with compact and well-separated clusters. However the DBI only evaluates the quality of the clusters themselves and does not consider outliers. A high proportion of uniformly distributed outliers between clusters may imply that the overall data distribution forms more of a continuum than distinct groupings, complicating the interpretation of the clustering structure.

Table 1: Minimum Davies-Bouldin Index across the tested parameter space of Perplexity and Epsilon, for varying values of minimum samples.

Min. Samples	**Davies-Bouldin Index**	**Perplexity**	**Epsilon**
5	0.607	6	14
10	0.765	13	5
15	0.582	10	7
20	0.460	5	20
25	0.836	6	16
30	0.819	9	7
35	0.463	7	16

In both cases a low value of perplexity was determined indicating there is a strong priority on the local structure during dimension reduction. This can make visualizations unstable when introducing an additional object to the data set. Some analysis was also carried out on the parameter space using a silhouette score obtaining a maximum silhouette score of 0.4 for an epsilon of 18, a perplexity of 5, and a number of samples of 20. Generally a silhouette score of above 0.5 is required for separated distinct clusters (Murty & Devi, 2016). The low silhouette score could derive from the fingerprints forming a continuum or that it is not suitable as a metric for concave clusters such as those formed by DBSCAN (Rousseeuw, 1987). The optimal set of parameters using the silhouette score as a metric again derived a low value for perplexity indicating a landscape that is potentially unstable when fingerprints are added or removed.

3.7 Clustering Using DBSCAN

DBSCAN was chosen for clustering our data set primarily because it is not sensitive to noise and will still form clusters. If there are outliers present they will be identified as such and not given cluster membership based on an arbitrary condition, number of clusters, prior to clustering. Clustering was carried out using DBSCAN and 240 9x16 V-band fingerprints. The fingerprints were created from light curves that have been cleaned of outliers using the method outlined in section 2.4. Scaling was applied using the standard scaler. Dimension reduction was applied using PCA from 144 down to ten components. The output was then passed to t-SNE and the dimensions were further reduced to two components. DBSCAN then applies a cluster label to each object to plot the visualization. The results of the clustering using DBSCAN can be seen in figure 18 which shows that clusters have been formed and a number of outliers identified.

By applying clustering algorithms to our data set we have two main objectives. To group together objects that have similar properties and to have stable landscape of objects that an artificial light curve can be introduced to give a

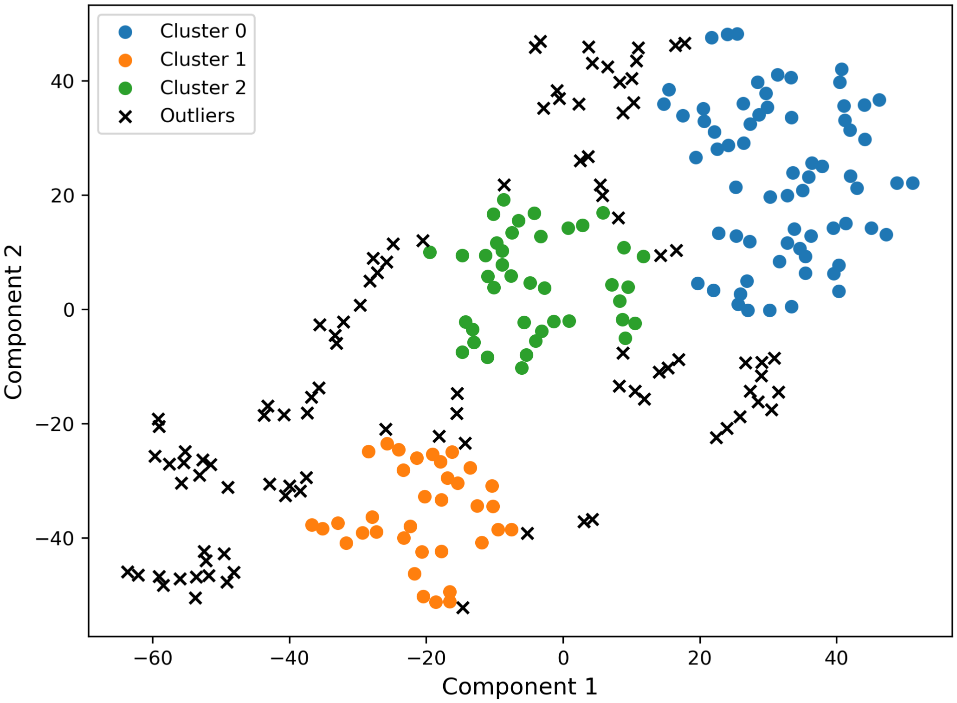

Figure 18: Visualization of clustering using DBSCAN with t-SNE, minimum samples = 20, epsilon = 20, perplexity = 5, on 240 variable 9x16 V-band fingerprints.

comparison against real objects. Figure 19 shows that there is some success in the clusters containing objects that have light curves with similar properties. The bottom panel shows in figure 19 shows two objects randomly chosen from the outliers which do not qualify for cluster membership based on the parameters of DBSCAN. The outliers are not expected to have similar properties and are rather a collection of objects that do not fit into any cluster. The top and middle pairs in figure 19 were randomly selected from two clusters and show that by visual inspection that these two pairs of light curves have partially similar features.

The main objective of the project is to introduce a light curve generated by a simulation of a YSO with a planetary disc. For this the landscape of objects shown in figure 18 should have a high degree of stability. That is to say if an object is removed or added to the data set the original objects should

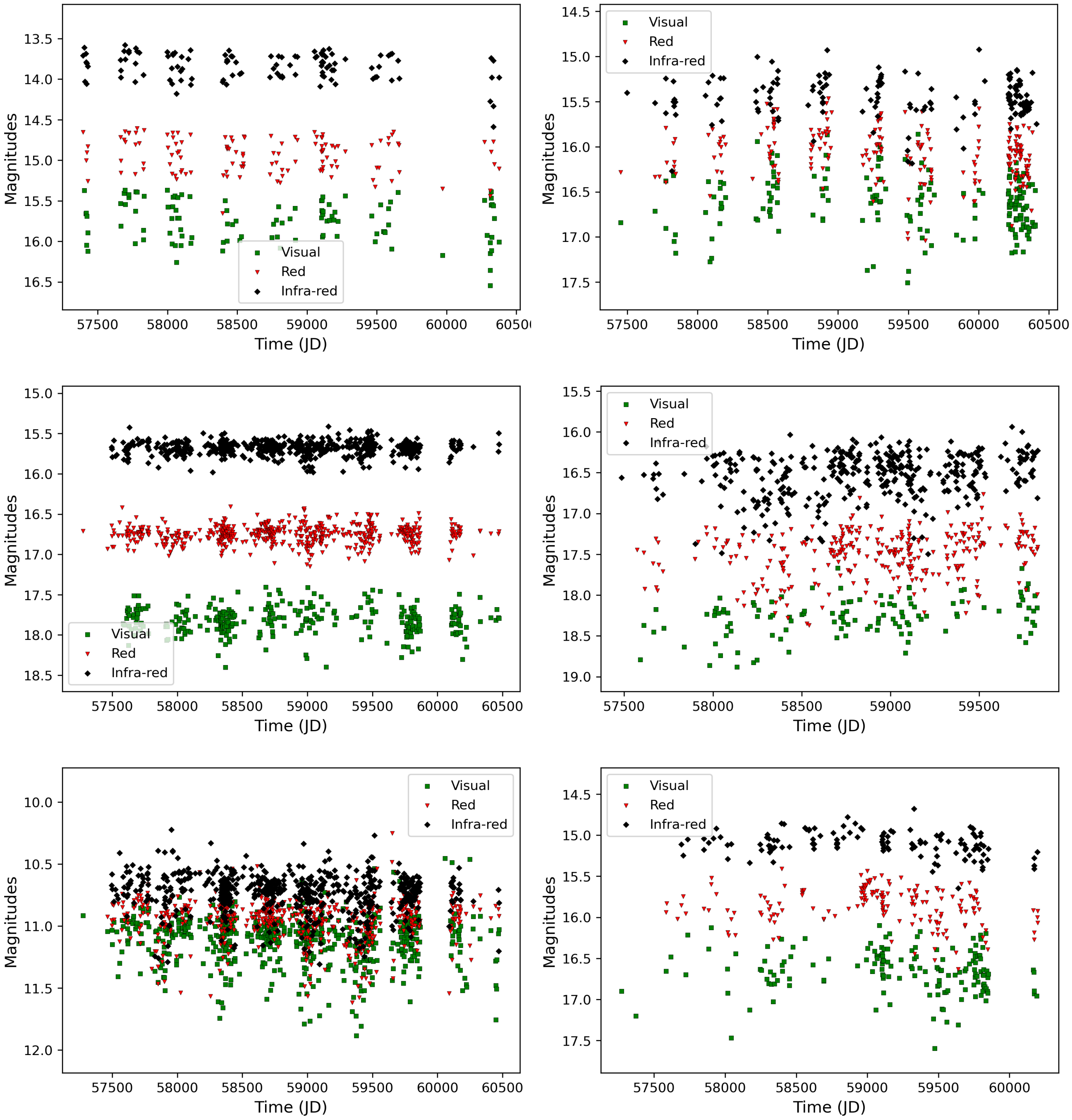

Figure 19: The light curves of three random pairs of objects taken from clusters shown in figure 18, top: cluster 0, middle: cluster 1, bottom: outliers.

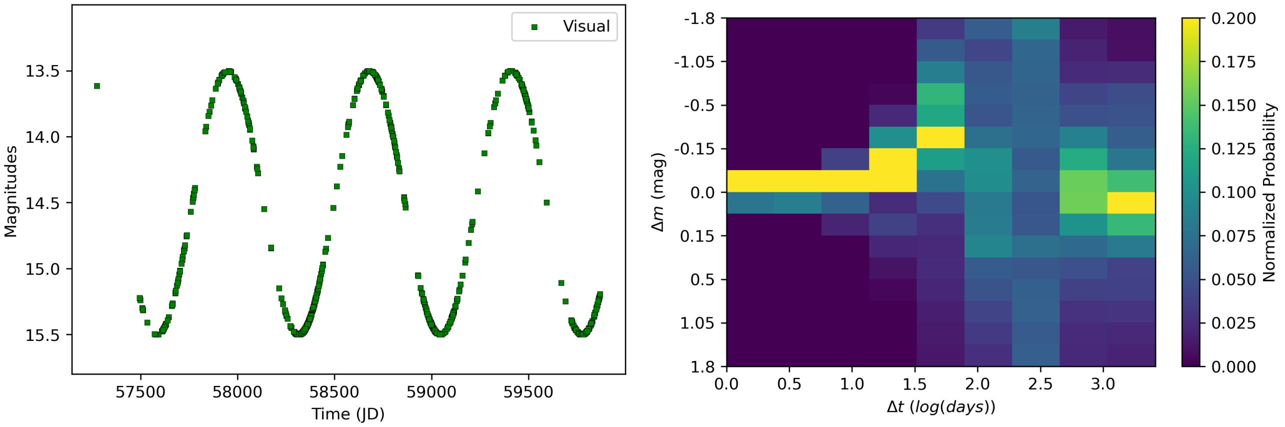

Figure 20: The light curve (left) and 9x16 fingerprint (right) generated by a sine wave with a two year period and one magnitude amplitude overlaid over a random HOYS object cadence.

maintain their position on the visualization and largely maintain the same cluster membership. Using the cadence of a random light curve selected from the HOYS data a sine wave was plotted as a light curve to be used as a simple analogue for a simulated light curve. A sine wave with a period of two years and an amplitude of one magnitude was chosen Although some stars would produce similar light curve as a perfect sine wave a YSO with a surrounding disc would not. The fingerprint was then created for sine wave light curve and then added to our HOYS data for clustering.

Using the same DBCSAN Figure 21 shows the visualization of the DBSCAN clustering after adding one fingerprint generated from a sine wave added to the HOYS data. The sine wave has been given membership to the outliers which was expected as its light curve properties are not analogous to those of a variable YSO. Figure 21 shows that the data as a whole is now visualized very differently from figure 20. Adding one object has prevented any clusters from being formed with all objects being assigned as outliers. The structure of the data has also been altered significantly highlighting the instability of this method of PCA then t-SNE dimension reduction. The cluster membership is a secondary goal of the

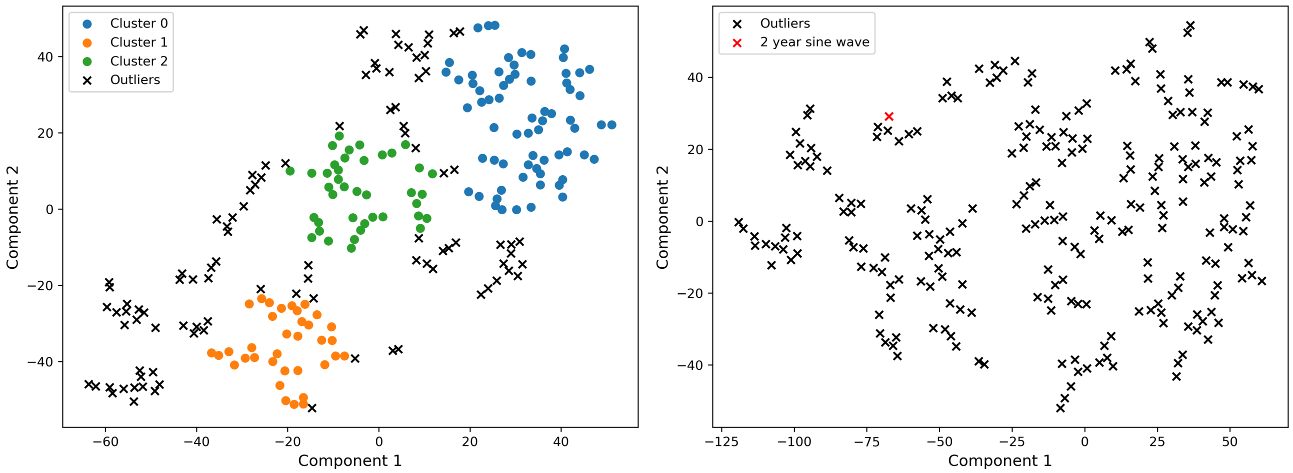

Figure 21: Left: Clustering using DBSCAN with t-SNE, epsilon = 20, perplexity = 5 and minimum number of samples 20, on 240 variable 9x16 V-band fingerprints. Right: Same as left with one fingerprint added to sample based on a sine wave light curve.

project with the stability of the landscape of objects being the primary objective of the project.

A comparison of of the two clustering attempts, see figure 21 shows that the landscape of objects has significantly changed. The change in the positions of the HOYS objects would not allow a simulated light curve to be modified and then added to the HOYS data to investigate the effects of those changes. The cause of the instability in the landscape of objects is driven by the low values of perplexity during the t-SNE process required for successful clustering. Typical values for perplexity range from 5-50. For low values of perplexity, towards five the algorithm focuses on preserving the local structure, highlighting small clusters or local similarities. For high values of perplexity, the algorithm tries to preserve more of the global structure showing larger clusters or overall data trends. While DBSCAN has some utility in grouping light curves with similar properties into clusters the instability driven by the low perplexity necessary for clustering makes it unsuitable for our purpose of adjusting the properties of a simulated light curve for comparison with the HOYS data.

3.8 Clustering Using K-means

K-means is a less sophisticated algorithm than DBSCAN, which converges quickly often within a few iterations. It has only one parameter that has to be defined before use which is the number of clusters. This will determine the number of centroids around which the clusters are formed. K-means is sensitive to outliers and noisy data. Outliers can skew the clustering results since they pull the outliers towards them. These properties make k-means an ideal algorithm to test if a stable landscape can be created into which an additional fingerprint can be introduced, possibly an outlier, without disturbing that landscape. Given that the number of clusters formed is less important for developing a stable landscape the number of clusters chosen was arbitrarily chosen as five clusters. Clustering was carried out using k-means and 240 9x16 V-band fingerprints. The fingerprints were created from light curves that have been cleaned of outliers using the method outlined in section 2.4. Scaling was applied using the standard scaler. Dimension reduction was applied using PCA, reducing the dimensions from 144 to two. The k-means algorithm was then applied and the visualization can be seen in figure 22.

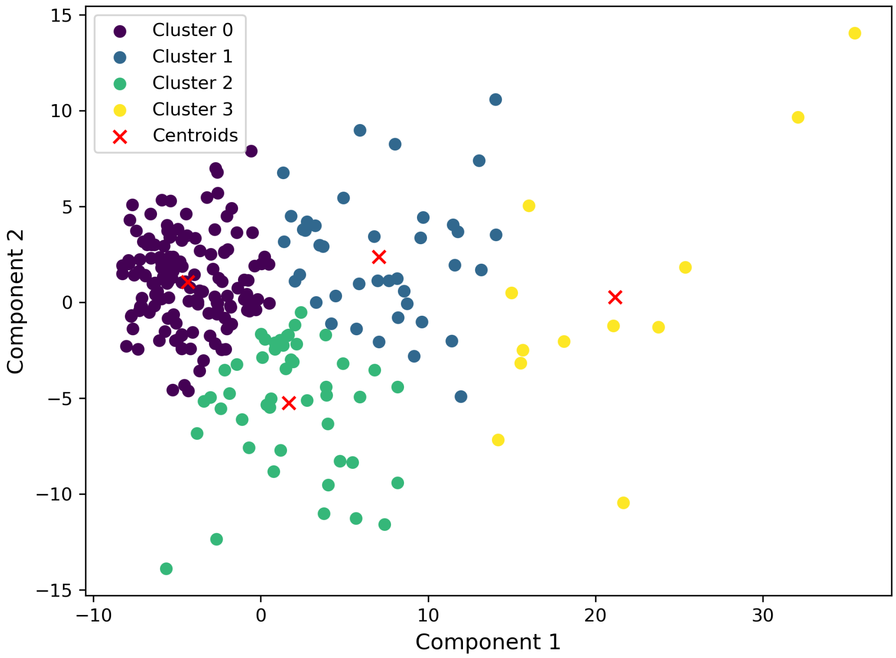

Figure 22: K-means clustering using 240 9x16 V-band fingerprints with PCA dimension reduction applied.

K-means does not assign fingerprints to outliers and as shown in figure 22 all fingerprints have been assigned to one of five clusters based on their cartesian distance to the cluster centroid. There is no separation between the clusters although there is increased density on the left hand side of figure 22 with a lower density of the fingerprints on the right hand side of the visualization.

Unlike the clusters shown in figure 18 those formed by k-means shown in figure 22 have less separation between the clusters. Again three pairs of light curves were selected from the most populous clusters zero, one and two. The light curves shown in figure 23 show that light curves plotted from a particular cluster do have similarity. This indicates that specific areas of the visualizations do represent properties of the fingerprints being clustered. The landscape created by clustering seems to be more of a continuum than distinct clusters. The

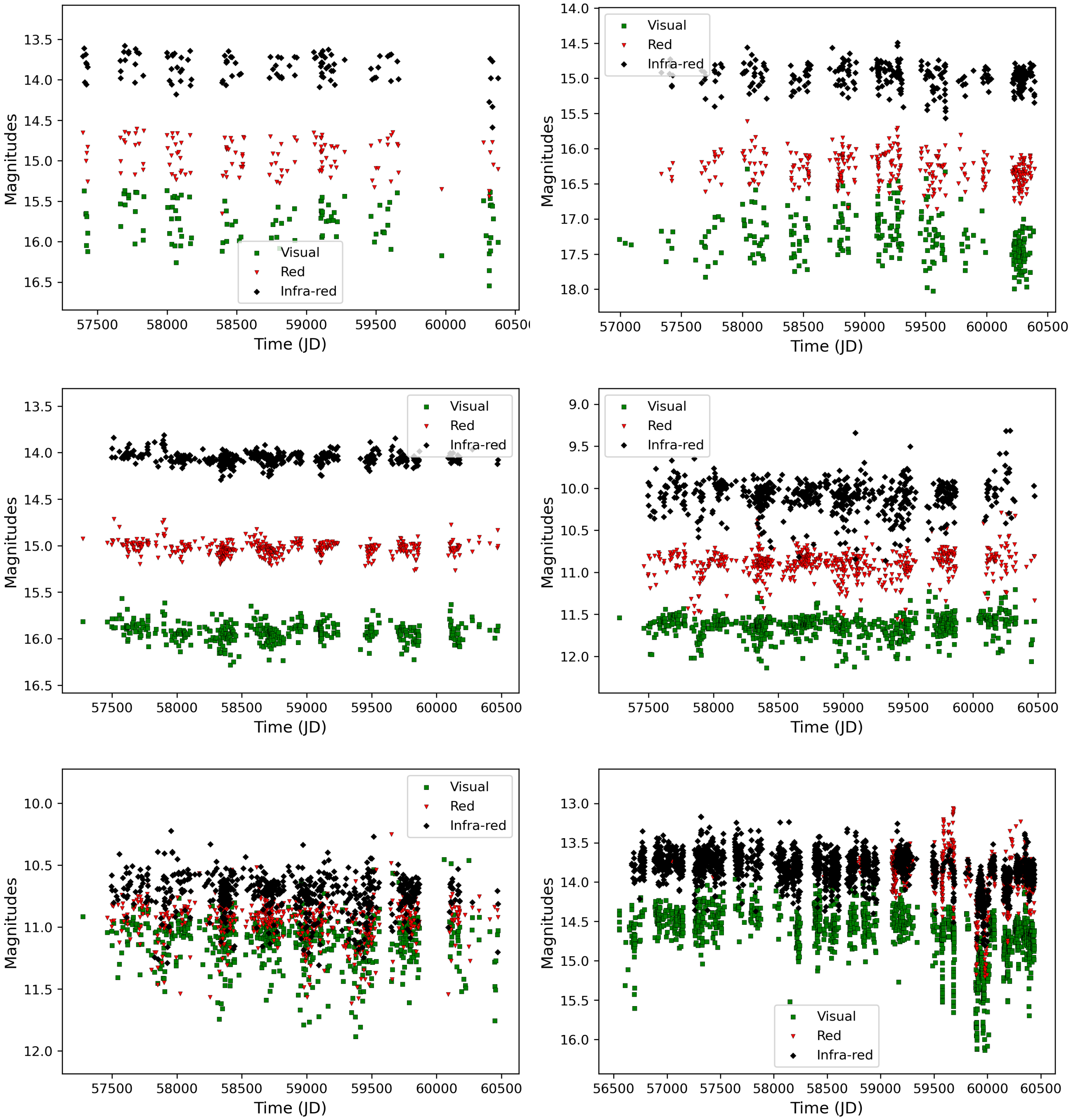

Figure 23: The light curves of three random pairs of objects taken from clusters shown in figure 22, top: cluster 0, middle: cluster 1, bottom: cluster 2.

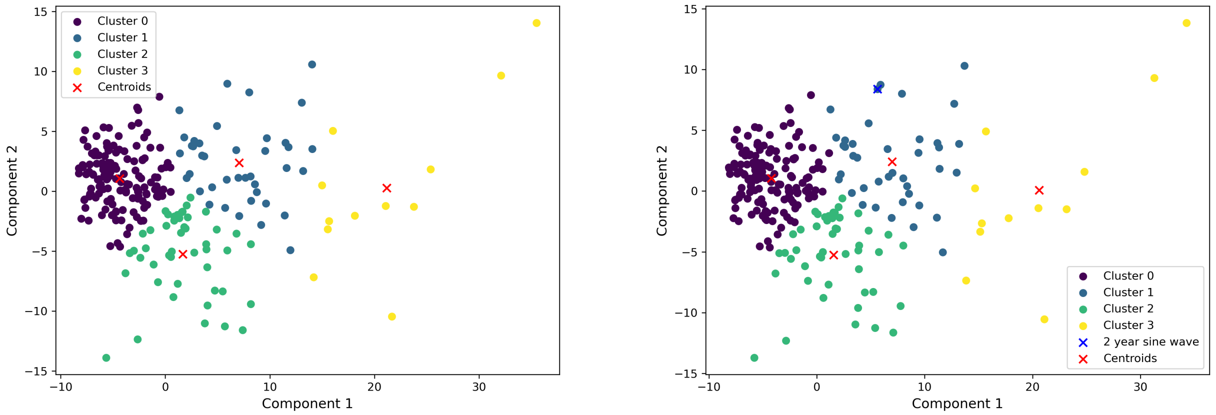

Figure 24: Left: K-means clustering using 240 9x16 V-band fingerprints with PCA dimension reduction applied. Right:As left panel with one fingerprint derived from a sine wave light curve added to data.

difference in outcomes between DBSCAN and k-means is due to the differences in the processes between these two algorithms.

K-means is only able to form convex cluster shapes around its centroid. DBSCAN also makes use of the t-SNE process as part of its dimension reduction which helps mitigate the effect of noise in a high dimensional data set. This process is not compatible with k-means and will lead to an outcome that is not useful. DBSCAN has used a combination of t-SNE with a low perplexity parameter and concave cluster shapes to force clusters onto the data. However the large number of outliers between the clusters in the DBSCAN visualization and the application of k-means shows it is a continuum. A continuum would still be useful for the purpose of comparison against a fingerprint derived from a simulated light curve as it has been shown that different areas of the visualization represent fingerprints of a differing profile provided it has stability. To be considered high stability firstly the landscape of objects should not alter their positions on the visualization when one object is added.

Figure 24 shows a comparison between two k-means clustering outcomes, the

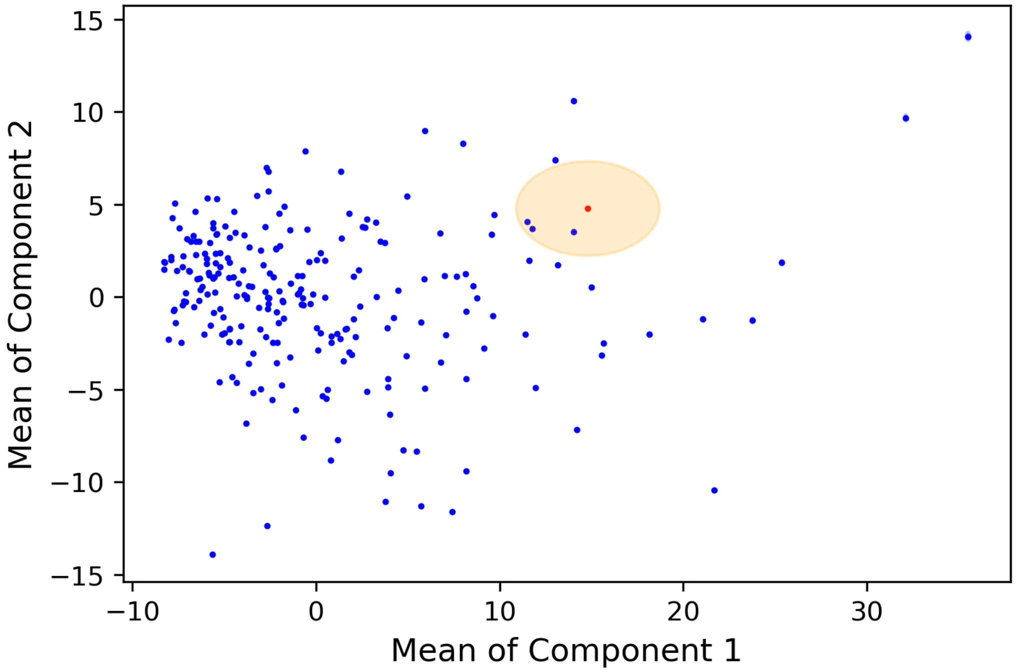

Figure 25: The means (center) and 20 standard deviations (ovals) of K-means clustering using 240 9x16 V-band fingerprints one of which is bootstrapped 1000 times (orange)

left with 240 fingerprints from our HOYS data and the right the HOYS data along with a fingerprint created from a sine wave overlaid on the cadence of a random HOYS data light curve with a period of two years and an amplitude of one magnitude. Although this fingerprint represents an object which is unlike a variable YSO it falls within the set of fingerprints. Its position is shown by a black cross in the left hand panel of figure 24 in one of the least dense areas of the the visualization. It can be seen from figure 24 that by adding the extra fingerprint to the HOYS data does not seem to change the landscape although a more rigorous test of this was required.

3.8.1 K-Means Stability

Using the bootstrapping process described in section 2.6 the magnitudes of one light curve were perturbed to create a stack of 1000 fingerprints. K-means clustering was then applied to the HOYS data set including one layer of the stack of perturbed fingerprints each time for 1000 runs. The mean of the principal components were then plotted against each other along with an oval. The axes of each oval being 20 standard deviations of the corresponding fingerprints principle component value. It was necessary to use 20 standard deviations to highlight that when a fingerprint is perturbed the landscape in blue is many times more stable than the perturbed fingerprint in orange. Although successful this test was based on perturbing the magnitude values by small amounts and to test the stability more rigorously a more significant change to the fingerprints was needed.

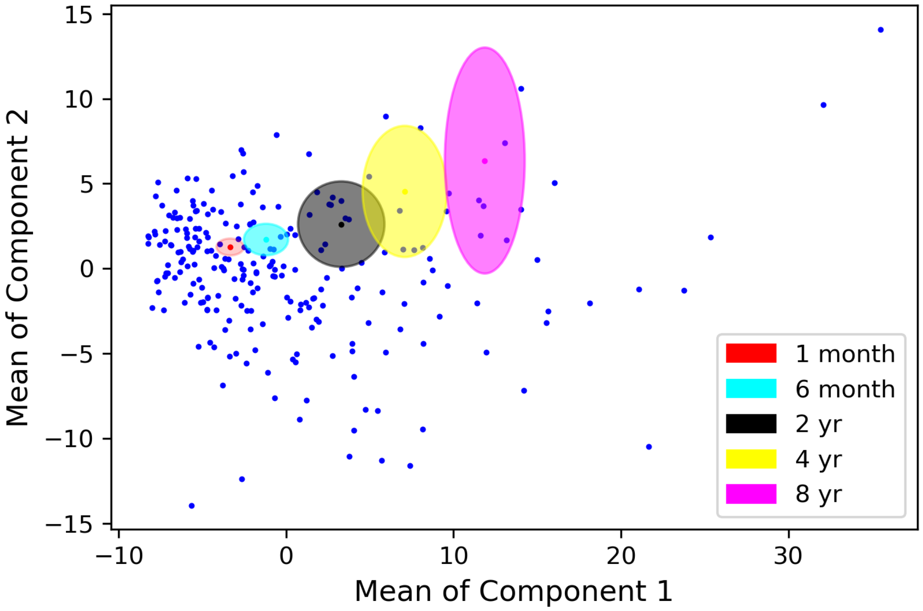

Figure 26: The means (center) and standard deviation (ovals) of K-means clustering using 240 9x16 V-band fingerprints plus fingerprints created from sine wave light curves with varying periods and an amplitude of one magnitude whose cadence were varied approximately 960 times.

Figure 26 presents a more rigorous test of the stability of the principal components, in which no multiplier of the principal component standard deviation was required. Using the cadences from our HOYS sample of 240 light curves, a stack of 960 artificial light curves was generated by randomly shifting the phase of each cadence four times and applying it to a sine wave of a given period. This stack was then converted into a corresponding stack of fingerprints.

K-means clustering was applied to the HOYS dataset, including one layer of the artificially generated fingerprints at a time, across 1000 runs. The mean principal component values of each fingerprint were plotted against one another, along with an ellipse whose axes correspond to the standard deviations of the principal component values of those fingerprints. This procedure was repeated for each period shown in Figure 26, and the resulting values were recorded.

The main plot in Figure 26 was constructed from a single 960-run process using one specific period, while additional periods were plotted manually using their recorded mean and standard deviation values. This manual plotting was chosen to better represent how a single fingerprint might be compared against a broader PCA landscape of objects. The periods used in Figure 26 are approximate and were selected as multiples of π, scaled to 365 days, in order to avoid resonances with the observing schedule. These periods correspond approximately to one month, six months, two years, four years, and eight years. Using multiples of π ensures the periods are not harmonics of the typical observing cadence.

The specific multiples of π used in Figure 26 are summarised in Table 2. These were chosen to represent approximate values corresponding to familiar durations while avoiding resonances with the observing schedule.

Table 2: Multiples of π used to define approximate variability periods.

Approximate period	**π multiples**	**Period (days)**
One month	$\frac{1}{3.8\pi} \times 365$	≈ 30.5
Six months	$\frac{1}{0.66\pi} \times 365$	≈ 176
Two years	$\frac{1}{0.15\pi} \times 365$	≈ 775
Four years	$\frac{1}{0.08\pi} \times 365$	≈ 1450
Eight years	$\frac{1}{0.04\pi} \times 365$	≈ 2900

These selected periods are representative of the range found in the HOYS survey, although they do not capture the full diversity of variability timescales observed in YSOs.

Figure 27 shows the same analysis performed using fingerprints generated from sine wave light curves with an amplitude of 0.5 mag. While some differences between the two plots are apparent, there is a clear correlation between the relative positions of the periods in the PCA space. Specifically, shorter-period signals tend to exhibit lower principal component 1 (PC1) values, while longer-period signals are associated with higher PC1 values.

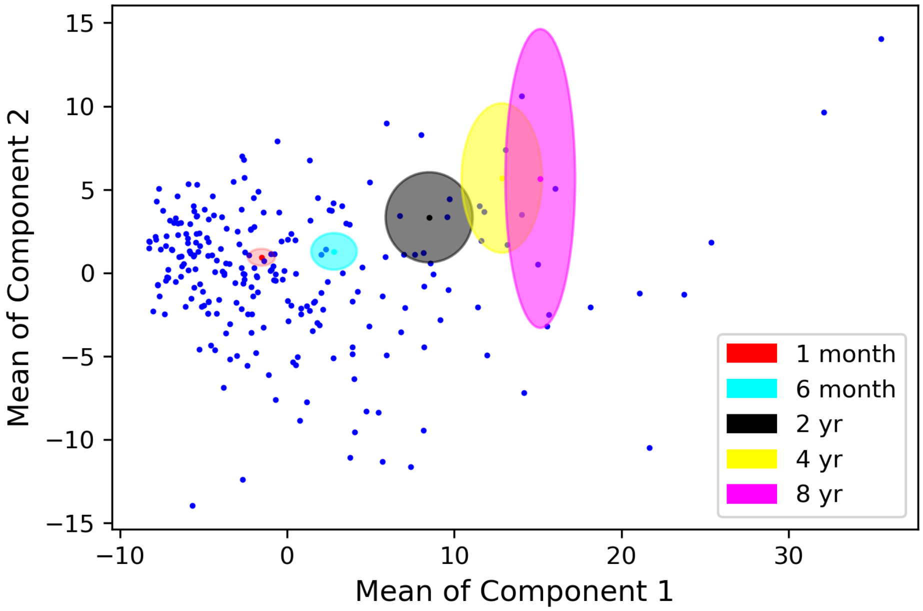

Figure 27: The means (center) and standard deviation (ovals) of K-means clustering using 240 9x16 V-band fingerprints plus fingerprints created from sine wave light curves with varying periods and an amplitude of half a magnitude whose cadence were varied approximately 960 times.

3.9 What do Principal Components Represent?

As discussed in Section 3.1.2, a principal component is a linear combination of all input features, given by

$$\text{Component} = a_1x_1 + a_2x_2 + \ldots + a_nx_n \tag{34}$$

where x_n is the value of the n-th pixel in a fingerprint, and a_n is the corresponding coefficient from the loading matrix derived via PCA. These coefficients are chosen such that the principal components are orthogonal and can be visualized in a two-dimensional space. A principal component does not directly represent any specific physical property of the fingerprints or the light curves they are derived from. However, since the fingerprints encode the probability that a YSO varies in brightness by a given amount over a given time period, the components are influenced by the variability characteristics of the underlying YSOs—particularly their periods—as demonstrated in Figures 26 and 27.

Figure 28 shows the loading vectors of the first two principal components as a heat map, where each pixel corresponds directly to a pixel in the fingerprint. Red pixels contribute positively to the component, while blue pixels contribute negatively. Figures 31, 32, 33, and 34 show representative light curves with high and low values for each component, revealing that regions of the PCA space correspond to distinct light curve properties.

Light curves with low PC1 values tend to be highly stochastic, while those with high PC1 values are typically long-period, near-constant light curves. These high-PC1 light curves occupy the sparsest regions of the PCA landscape, corresponding with our initial variability threshold of Stetson $J > 2$ across all three filters. PC1 explains a substantial fraction of the variance in the sample (36%), much more than PC2 (11%), which results in a more pronounced difference between the light curves in Figures 31 and 32 than between those in Figures 33 and 34.

Light curves with higher PC2 values tend to exhibit a long-term dimming trend, while those with lower PC2 values tend to brighten over time. The area with the highest magnitude in the PC2 loading map (Figure 28, right panel) corresponds to the longest timescales. This aligns with the long-term brightness trends, although Figure 26 also shows that the central rows of pixels on the y-axis influence PC2 significantly. For sine-wave fingerprints, the mean PC2 value increases with period (Figure 26) because longer-period signals exhibit higher probabilities in the central two rows of pixels. Since these central rows contribute

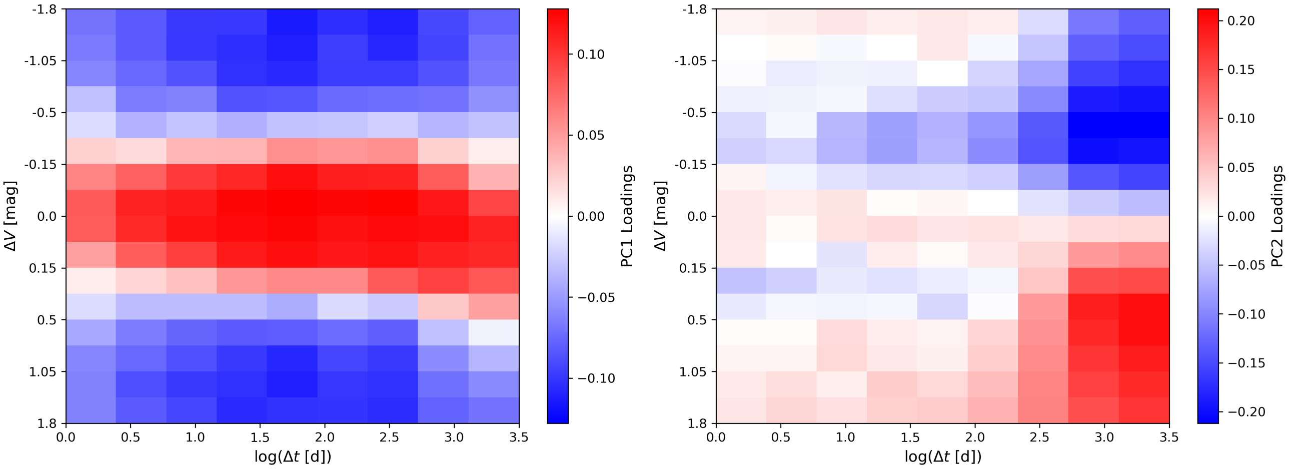

Figure 28: Left: Loadings vector heat map for principal component 1. Right: Loadings vector heat map for principal component 2.

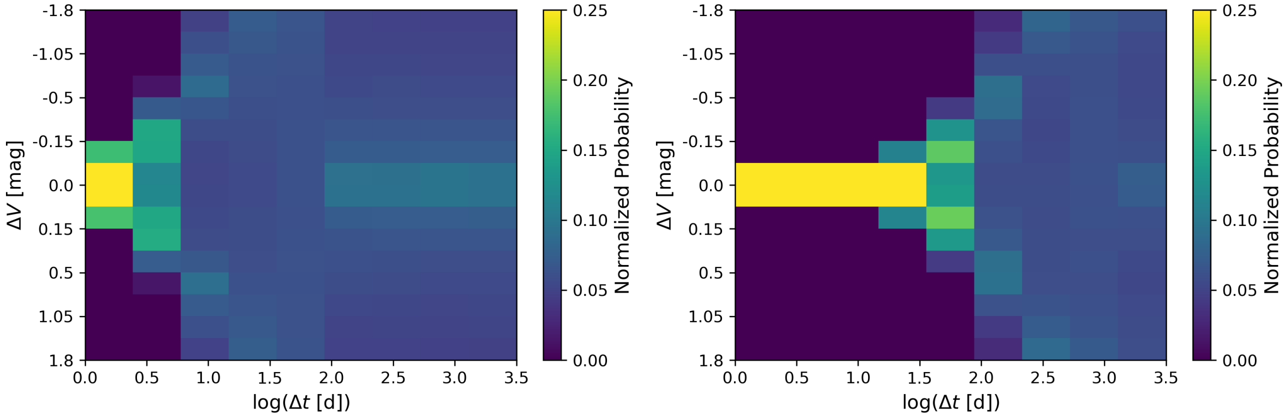

Figure 29: Left: The mean fingerprint for all HOYS cadences generated from a sine wave type light cure with a period on one month and an amplitude of one magnitude. Right: As left with a period of eight years.

positively to the PC2 value (as seen in Figure 28, right panel), the overall mean PC2 increases with period up to roughly 300 days. On short timescales, however, the sine-wave fingerprints resemble those of non-variable light curves.

An example of the mean fingerprint for the shortest and longest simulated periods is shown in Figure 29. This illustrates how the higher pixel values within the fingerprint are transformed by the loading matrix, resulting in fingerprints associated with highly stochastic light curves exhibiting low PC1 values. As demonstrated in Figure 29, once the period becomes sufficiently long, the finger-

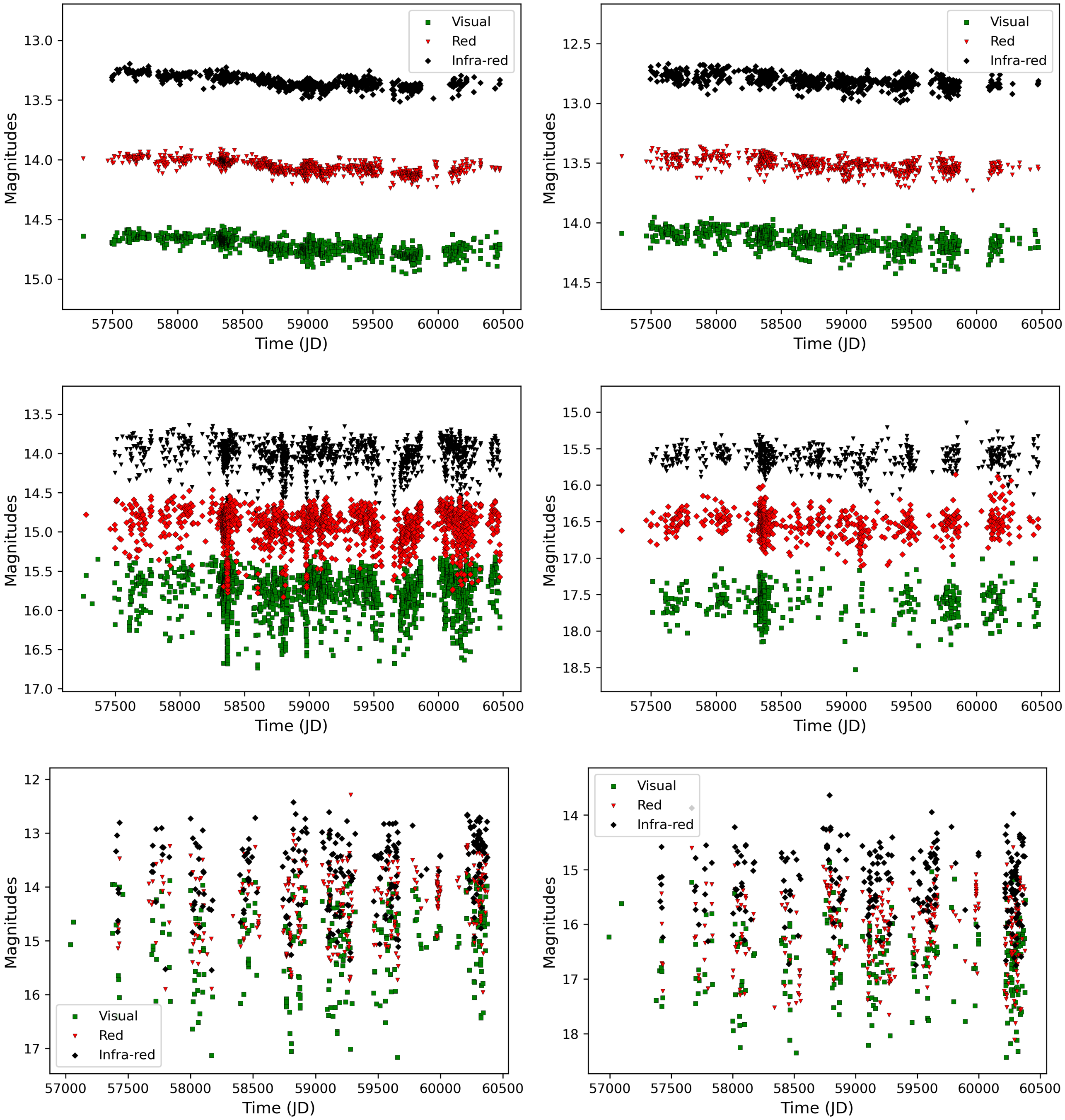

Figure 30: top left: PC1=35.50 PC2=14.04, top right: PC1=32.10 PC2=9.64, mid left: PC1=-0.19 PC2=0.15, mid right: PC1=-0.85 PC2=0.42, bottom left: PC1=-7.58 PC2=1.41, bottom right;PC1=2.66 PC2=3.75

print becomes dominated by pixel values concentrated in the central two rows. This shift leads to higher PC1 values, as those central pixels contribute more strongly and positively to the first principal component.

A final test was conducted to assess the consistency of light curve morphology among sources with nearly identical PC values. Figure 30 presents three pairs of light curves whose corresponding PC1 and PC2 values are as similar as possible within the dataset. The top pair, originating from the sparsely populated, high-PC1 region of the PCA space, exhibits the greatest numerical difference between their PC values among the selected pairs. Despite this, their overall light curve morphology remains qualitatively consistent.

This comparison highlights that even in regions of PCA space where it is not possible to select pairs of sources with nearly identical principal component values, light curves with relatively close PC coordinates still exhibit strong morphological resemblance. In the more densely populated areas, where many sources share similar variability characteristics, the selected light curve pairs show near-identical PC values and highly consistent morphology. This quantitative validation—when combined with the earlier qualitative assessments of PCA clustering and fingerprint structure—further supports the robustness and interpretability of the low-dimensional PCA landscape in capturing meaningful and continuous variability trends among YSOs.

An analysis of the PCA loading matrices, already discussed, reveals that the primary source of variance among the variability fingerprints is closely tied to the timescale at which significant photometric variability—defined here as changes greater than 0.3 mag—begins to manifest. The fingerprints demonstrate that variability on intermediate timescales, particularly between one and three months, contributes most strongly to the first principal component. This finding is consistent with earlier results linking PC1 to a continuum from stochastic behaviour to long-period, low-amplitude variation, underscoring the dominant role played by intermediate-scale variability in shaping the global structure of the dataset in PCA space.

The second principal component, although accounting for a smaller proportion of the overall variance, encapsulates long-term trends in brightness, typically spanning periods greater than 1.5 years. These trends include steady dimming or brightening, which may arise from evolutionary processes in the circumstellar environment or changes in the accretion regime.

Together, these components offer a compact but powerful summary of the diversity of variability behaviour within the sample. The PCA framework, combined with the fingerprinting methodology developed in earlier chapters, enables an efficient mapping of complex light curve morphology onto a low-dimensional space, where distinct physical regimes and variability mechanisms can be more easily interpreted and compared.

4 Conclusions

The photometric treatment of our inhomogeneous HOYS dataset has been refined. Building on the standard calibration established by Froebrich et al. (2018) and Evitts et al. (2020), unreliable photometry was excluded near bright stars and from images exhibiting tracking issues. Furthermore, potential photometric outliers in colour–magnitude space were identified and removed from the analysis. Based on the previously identified sample of approximately 3000 members in the monitored young clusters (Froebrich et al., 2024b), and using long-term V, R, and I-band photometry, a subset of 240 highly variable YSOs was identified (Ryan et al., 2025).

Variability fingerprints were constructed from the light curves of the sources, mapping the probability that an object varies by a given amount over a given timescale. This approach enables a quantitative comparison of the variability statistics of stochastically varying sources with those derived from randomly sampled light curves. By quantitatively comparing the uncertainties of fingerprints at varying resolutions, the 9×16 adaptive fingerprint was identified as producing pixels with the highest S/N ratio. Variability was probed over a range from ±0.05mag to ±2.0mag on timescales spanning 1d to 8.6yr. A signal-to-noise ratio exceeding three was achieved in over 90 percent of the fingerprints, with low-S/N regions confined to short timescales and large amplitude variations.

Two methods were tested for calculating the uncertainty of fingerprint pixel values. Bootstrapping of the light curves and uncertainty propagation based on Poisson counting statistics. The fingerprint uncertainties were found to closely follow Poisson statistics, with a regression coefficient of approximately 0.9 for the 9×16 fingerprint configuration in the V-band. Although bootstrapping produced the most accurate uncertainty estimates, it was computationally intensive and impractical for generating uncertainty maps across all filters and configurations.

Dimensionality reduction using t-SNE was found to produce an unstable landscape, attributable to its non-linear nature. The positions of individual objects shifted significantly upon modification of the sample, although the overall large-scale structure remained recognizable. In contrast, PCA generated a highly stable landscape in which the addition or alteration of a single object led to only marginal displacements of the remaining points. With their positions relative to each other remaining almost constant. This stability permitted the incorporation of model-generated fingerprints and the assessment of their relative placement with respect to the observed data. A further test demonstrated positional consistency within the PCA landscape. Pairs of objects with closely matching principal component values were selected from three distinct regions of the landscape. Upon comparison, the light curves of each pair were found to exhibit a strong subjective resemblance.

To evaluate this, simple sinusoidal light curves were simulated using varying HOYS cadences and randomized phase shifts. These models occupied a limited region within the observed landscape, indicating that observing time, cadence, and photometric uncertainties do not strongly influence an object's position.

Examination of the PCA loadings matrices revealed that the greatest variance among the fingerprints was associated with the timescale at which significant ($>$0.3mag) variability commenced, with timescales of 1–3 months identified as most influential. The second most prominent factor was associated with long-term ($>$1.5yr) trends, such as sustained fading or brightening behaviour in the light curves.

Although PCA yielded a stable and interpretable low-dimensional representation of the fingerprint data, neither PCA nor t-SNE in conjunction with k-means and DBSCAN resulted in the formation of distinct, well-separated clusters. Instead, the fingerprints formed a continuous distribution in the reduced-dimensionality space, suggesting that the variability behaviour of YSOs spans a spectrum rather than falling into discrete categories. This outcome is consistent with the expectation that multiple overlapping physical mechanisms—such as accretion variability, variable extinction, and rotational modulation—contribute to the observed light curves, leading to a continuum of variability characteristics rather than sharply defined classes. Consequently, clustering methods did not identify statistically significant groupings, underscoring the importance of a continuous representation for understanding and comparing YSO variability.

These findings demonstrate that PCA applied to variability fingerprints provides a robust and quantitative framework for comparing the variability characteristics of observed YSO light curves with those derived from theoretical models of the underlying physical mechanisms.

4.1 Further Work

A natural extension of this study involves the incorporation of more complex simulated light curves to further refine and validate the fingerprinting and PCA-based classification methodology. While simple sinusoidal models have proven useful in tracing broad trends within the PCA landscape, they do not capture the full diversity of variability observed in YSOs, such as accretion bursts, quasi-periodic dipping, or complex stochastic behavior. Simulating light curves based on detailed physical models including magnetospheric accretion and disc occultation will allow the construction of a library of physically motivated synthetic fingerprints. These could be used as templates to more rigorously interpret regions of PCA space, allowing a more direct link between observed variability and underlying physical processes.

In addition, comparing the PCA distributions of fingerprints from multiple YSO samples—spanning different star-forming regions, ages, or environments may reveal consistent structural features or key differences in variability behavior. Quantifying the degree of overlap between observed samples and simulated fingerprints could lead to the development of a variability based index to evaluate the realism of planet formation simulations. If a simulation produces light curves whose fingerprints do not occupy the same PCA regions as real YSOs, this would suggest a discrepancy in the modeled physical conditions. Ultimately, this methodology could serve as a diagnostic framework, enabling researchers to constrain star and planet formation models by comparing their synthetic outputs to the statistical and structural properties of observed variability.

5 Appendix

Method	Execution Time (s)
Using loop	9.301
Using matrix operation	0.007

Table 3: Test results for code optimization of normalizing fingerprint columns, 40x40 configuration used 1000 times

Configuration / Filter	Mean of Slopes	Standard Deviation of Slopes
9x16 / V-band	0.864	0.084
9x16 / R-band	0.866	0.111
9x16 / I-band	0.832	0.086
20x20 / V-band	0.830	0.080
20x20 / R-band	0.832	0.128
20x20 / I-band	0.804	0.085

Table 4: Mean slopes of comparison between Poisson and Bootstrapping error values for every object.

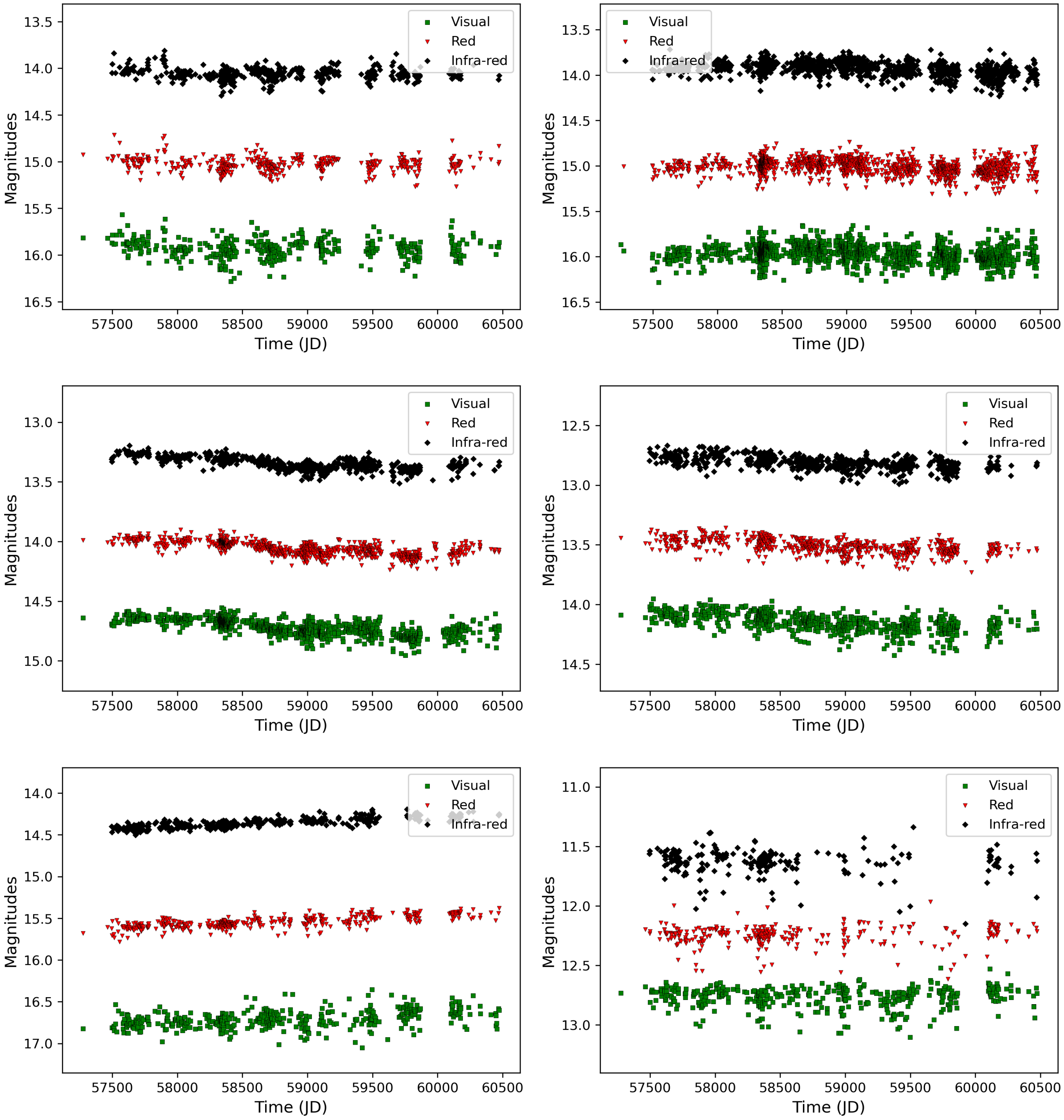

Figure 31: A selection of light curves with a high PC1 value (> 20).

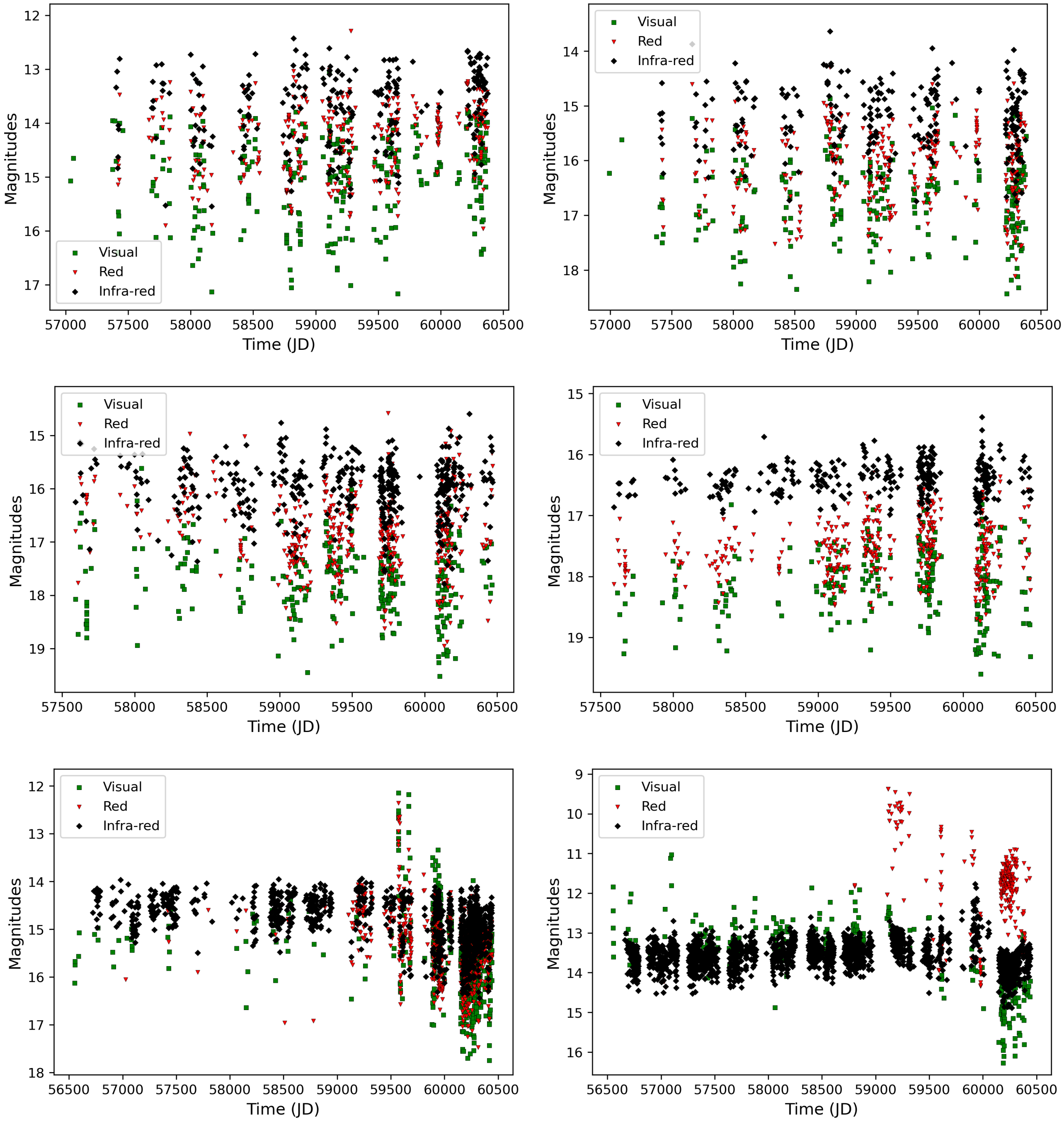

Figure 32: A selection of light curves with a low PC1 value (< -7).

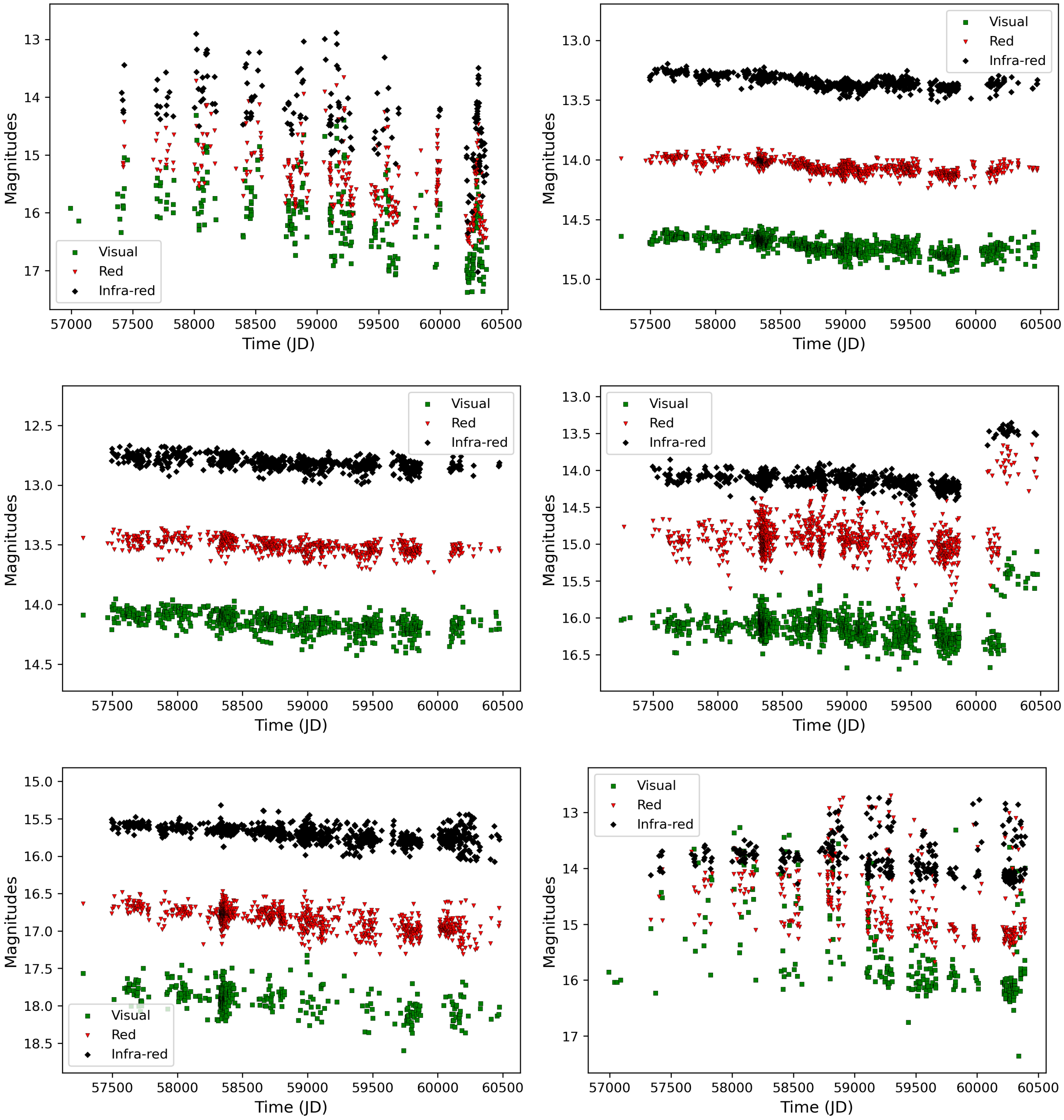

Figure 33: A selection of light curves with a high PC2 value (> 5).

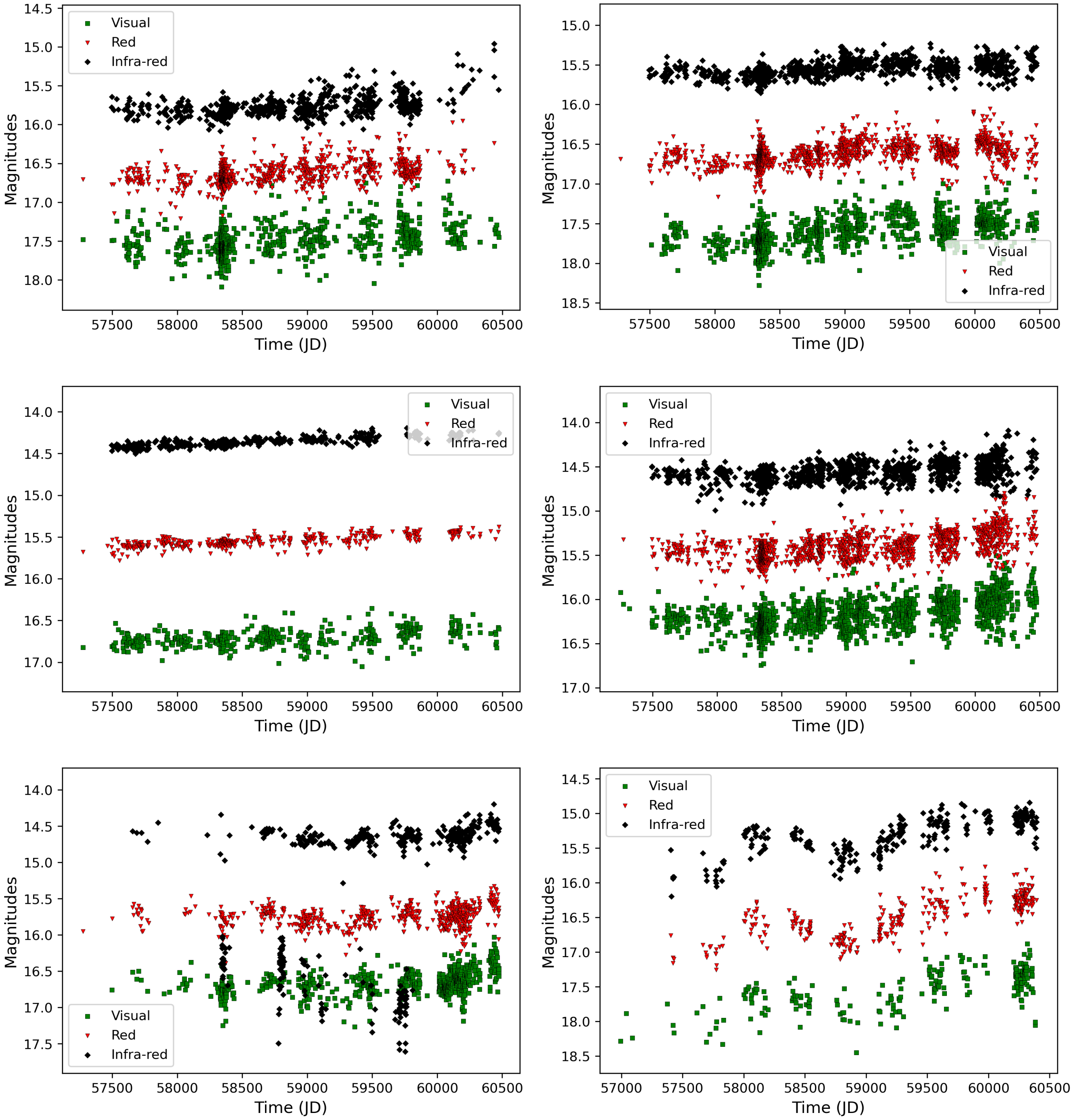

Figure 34: A selection of light curves with a low PC2 value (< -7).

References

Acke B., van den Ancker M. E., Dullemond C. P., van Boekel R., Waters L. B. F. M., 2004, Astronomy and Astrophysics, 422, 621

André P., Ward-Thompson D., Barsony M., 1993, The Astrophysical Journal, 406, 122

Bacher A., Kimeswenger S., Teutsch P., 2005, Monthly Notices of the Royal Astronomical Society, 362, 542

Bedding T. R., et al., 2020, Nature, 581, 147–151

Bertin E., Arnouts S., 1996, Astronomy and Astrophysics Supplement Series, 117, 393

Bouvier J., et al., 1999, Astronomy and Astrophysics, 349, 619

Cieza L., et al., 2007, The Astrophysical Journal, 667, 308

Davies D. L., Bouldin D. W., 1979, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, 224

Dullemond C. P., Dominik C., 2005, Astronomy and Astrophysics, 434, 971

Ester M., Kriegel H.-P., Sander J., Xu X., 1996, in Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96). AAAI Press, pp 226–231, `https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf`

Evans N. J., et al., 2009, The Astrophysical Journal Supplement Series, 181, 321–350

Evitts J. J., et al., 2020, Monthly Notices of the Royal Astronomical Society, 493, 184

Findeisen K., Cody A. M., Hillenbrand L., 2015, The Astronomical Journal, 798, 89

Fischer W. J., Hillenbrand L. A., Herczeg G. J., Johnstone D., Ágnes Kóspál Dunham M. M., 2023, Accretion Variability as a Guide to Stellar Mass Assembly (arXiv:2203.11257)

Froebrich D., et al., 2018, Monthly Notices of the Royal Astronomical Society, 478, 5091–5103

Froebrich D., et al., 2024a, Monthly Notices of the Royal Astronomical Society, 529, 1283

Froebrich D., et al., 2024b, Monthly Notices of the Royal Astronomical Society, 529, 1283

Gorti U., Dullemond C. P., Hollenbach D., 2009, The Astrophysical Journal, 705, 1237

Greene T. P., Wilking B. A., André P., Young E. T., Lada C. J., 1994, The Astrophysical Journal, 434, 614

Hahsler M., Piekenbrock M., Doran D., 2019, Journal of Statistical Software, 91

Hastie T., Tibshirani R., Friedman J., 2009, The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer series in statistics, Springer, https://books.google.co.uk/books?id=eBSgoAEACAAJ

Hogg D. W., Blanton M., Lang D., Mierle K., Roweis S., 2008, in Argyle R. W., Bunclark P. S., Lewis J. R., eds, Astronomical Society of the Pacific Conference Series Vol. 394, Astronomical Data Analysis Software and Systems XVII. p. 27

Hotelling H., 1933a, Journal of Educational Psychology, 24, 417

Hotelling H., 1933b, Journal of Educational Psychology, 24, 498

Hueso R., Guillot T., 2005, Astronomy & Astrophysics, 442, 703

Jolliffe I. T., 2002, Principal Component Analysis, 2nd edn. Springer, New York, doi:10.1007/b98835

Joy A. H., 1945, The Astrophysical Journal, 102, 168

Kenyon S. J., Hartmann L. W., Strom K. M., Strom S. E., 1990, Astronomical Journal, 99, 869

Kirmizitas O., Cavus S., Aliçavuş F. K., , Discovery of new Delta Scuti Stars

Lada C. J., 1987, in Star Forming Regions. p. 1, https://ui.adsabs.harvard.edu/abs/1987IAUS..115....1L

Lakeland B. S., Naylor T., 2022, Monthly Notices of the Royal Astronomical Society, 514, 2736–2755

Lizano S., Shu F. H., 1987, Rev. Mex. Astron. Astrofis., 14, 587

Lloyd S., 1982, IEEE Transactions on Information Theory, 28, 129

Luger R., Foreman-Mackey D., Hedges C., Hogg D. W., 2021, The Astronomical Journal, 162, 123

Mac Low M.-M., Klessen R. S., Burkert A., Smith M. D., 1998, Phys. Rev. Lett., 80, 2754

MacQueen J., 1967, in Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics. pp 281–297

Mannings V., Sargent A. I., 2000, The Astrophysical Journal, 529, 391

Moffat A. F. J., 1969, Astronomy and Astrophysics Supplement Series, 3, 455

Murty M., Devi V., 2016, Pattern Recognition, An Algorithmic Approach. Springer

Natta A., Grinin V., Mannings V., 2000, in Mannings V., Boss A. P., Russell S. S., eds, Protostars and Planets IV. pp 559–588

Padgett D. L., et al., 2006, arXiv e-prints, pp astro–ph/0603370

Pott J.-U., Perrin M. D., Furlan E., Ghez A. M., Herbst T. M., Metchev S., 2010, The Astrophysical Journal, 710, 265–278

Rigon L., Scholz A., Anderson D., West R., 2017, Monthly Notices of the Royal Astronomical Society, 465, 3889

Rousseeuw P. J., 1987, Journal of Computational and Applied Mathematics, 20, 53

Ryan B. W., Stokes-Geddes H., Froebrich D., 2025, Monthly Notices of the Royal Astronomical Society, 543, 1133

Scholz A., Eislöffel J., 2004, Astronomy and Astrophysics Supplement Series, 419, 249

Scholz A., Jayawardhana R., Wood K., 2006, The Astrophysical Journal, 645, 1498

Scitovski R., Sabo K., Martinez-Alvarez F., Ungar S., 2021, Cluster Analysis and Applications. Springer

Stetson P. B., 1996, Publications of the Astronomical Society of the Pacific, 108, 851

Stone J. M., Ostriker E. C., Gammie C. F., 1998, The Astrophysical Journal, 508, L99

Tan J. C., Beltrán M. T., Caselli P., Fontani F., Fuente A., Krumholz M. R., McKee C. F., Stolte A., 2014, in Beuther H., Klessen R. S., Dullemond C. P., Henning T., eds, Protostars and Planets VI. pp 149–172 (arXiv:1402.0919), doi:10.2458/azu˙uapress˙9780816531240-ch007

Vogel R., 2013, Telescopes and Deep Sky, https://www.reinervogel.net/index_e.html?/YSO/YSO_e.html

Wahhaj Z., et al., 2010, The Astrophysical Journal, 724, 835

Welch D. L., Stetson P. B., 1993, The Astronomical Journal, 105, 1813

Williams J. P., Cieza L. A., 2011, Annual Review of Astronomy and Astrophysics, 49, 67

Zollanvari A., 2023, Clustering. Springer International Publishing, Cham, pp 319–349, doi:10.1007/978-3-031-33342-2˙12, https://doi.org/10.1007/978-3-031-33342-2_12

van der Maaten L., Hinton G., 2008, Journal of Machine Learning Research, 9, 2579