Efficient Discovery of Binary Stars

Pablo Navarro Barrachina

Lund Observatory
Lund University

2020-EXA158

Degree project of 60 higher education credits (for a degree of Master)
May 2020

Supervisor: Gregor Traven

Lund Observatory
Box 43
SE-221 00 Lund
Sweden

Abstract

Purpose: even in the era of exponential increase in the amount of stellar data gathered, binaries are still often overlooked in observational data due to the special handling they require. The goal of this work is to develop a method capable of automatically and efficiently identifying and extract double-lined spectroscopic binaries (SB2) from a spectroscopic survey, while being scalable and technically successful, and to identify and optimize the parameters that influence their detection.

Method: we combine two state-of-the-art machine learning algorithms that group the spectra in the data-set in clusters based on their similarities, projecting them in a human readable manner (t-distributed Stochastic Neighbor Embedding, t-SNE), and automatically identify and retrieve those clusters that contain binary spectra (Density Based Spacial Clustering of Applications with Noise, DBSCAN). These methods are then optimized for efficient recovery of binaries from a synthetic spectroscopic data-set, where we know exactly which stars are single and which are binaries.

Results: we study the results following from 360 combinations of our method's parameters and obtain a total average of recovered binaries of 57%. We show that under optimal conditions we are able to reach a recovery of 75%. We find that bluer spectral regions (450 nm - 600 nm) are better suited to identify binary stars than redder regions (600 nm - 900 nm) with our method. Not only this, but we also show that a moderate amount of noise can be beneficial and can improve the recovery of binary stars. Furthermore, we find that the stellar parameters that most influence the final recovery are the luminosity (or mass) ratio and the radial velocity different between the two stellar components of the binary system, while some standard stellar parameters can play a major role as well.

Conclusions: we show that our method and the adopted combination of machine learning algorithms to be successful at automatically detect and retrieve binary stars from our synthetic spectroscopic data and we provide a list with guidelines for its application to real spectroscopic surveys.

Acknowledgments

I would like to express my gratitude to Gregor for his constant help and advice, to my parents for their unconditional support and to Ariane, for always being there, for encouraging me, for always believing in me, for everything.

Populärvetenskaplig beskrivning

Contrary to the popular belief that most stars are singles, around half of the stars we see in the galaxy are actually found in pairs called binary systems or simply "binaries". Their binary nature can be discovered or inferred in many different ways, such as through eclipses that occur when one star of the pair passes in front of the other, or through the characteristic features and behavior of their combined spectrum. Moreover, binaries play a major role in astrophysics. They offer scientists an insight into crucial stellar processes as well as enable accurate measurements of fundamental stellar parameters such as mass and radius through the gravitational interaction between the two components of the system.

In recent years, stellar surveys have increased exponentially in complexity and amount of stars observed, and while binaries have been shown to be abundant they are often missed in observational data. The reason for this is that in order to reveal their true nature, binary stars require a special handling besides that given by traditional methods for the analysis of stellar data. This can, in most cases, be quite time consuming. However, new approaches for discovery and characterization of binary stars have been made possible by advances in the field of machine learning and the increase of computational power. Machine learning is the name given to a set of algorithms and statistical tools used by computers to extract information from data by recognizing patterns without being explicitly programmed to do so. With it, it is possible to not only examine and study the large amounts of new data gathered by stellar surveys, but also "revisit" older data-sets in order to extract insights and patterns that were overlooked in the past.

In this project we will try to address the discovery efficiency of binary stars in an archetypical spectroscopic survey when using machine learning algorithms. By generating different types of spectra ourselves, we create a mock spectroscopic survey data-set for which the distribution of stellar parameters and the amount of binary stars are known. Unlike in real surveys, by using self-generated data the nature of each star is known beforehand. This allows us to evaluate a series of machine learning algorithms with respect to its own input parameters and the ranges of stellar parameters present in the generated data. With this evaluation, we want to determine and constrain the efficiency and limits of the used method regarding their efficiency discovering and detecting binary stars.

Our aim is to generate an automated method capable of maximizing the recovery and detection of binary systems from real spectroscopic data while being scalable and applicable to future surveys. To achieve such a goal, we combined two well-known and readily available machine learning algorithms for the automatic analysis of spectroscopic data and the retrieval of binary stars from it. One algorithm is called t-SNE, which is used to project the data onto a plane, grouping objects that are similar in clusters and separating those that are dissimilar. The groups of data-points representing spectra created by t-SNE are then automatically recovered regardless of their morphology by the second machine learning algorithm we use, which is called DBSCAN.

The combination of t-SNE and DBSCAN, whose individual implementation has been carefully chosen to minimize the computation time, allowed us to obtain results that are easy to implement and understand. Results from our study are promising, showing a mean recovery of 57%, averaged over all the 360 simulations we carried. We find that the presence of moderate noise levels in the studied spectra can help improving the detection of spectroscopic binaries, as it can smear out information from it that might throw off the machine learning analysis. Furthermore, we show that bluer spectral regions (between 450 and 650 nm) are better suited than those in redder parts of the spectrum (between 650 and 900 nm) due to the increased amount of information in the form of spectral lines present in the analyzed spectroscopic data. In the end, we provide a table of stellar parameters for binary stars that were contained in our synthetic

sample and which were successfully identified in more than 90% of our simulations and which can serve as a guide for future implementations of our method.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

When looking at the stars in the night sky, it might seem that all of them are found alone. In reality, around 40% of the stars we observe are part of associations known as multiple systems (Duchêne and Kraus 2013). These systems range from two components, known as binary stars or binary systems, to higher order systems with six or even seven components, which revolve around a common center of mass that receives the name of barycenter. In binary stars the more massive is commonly called the primary star, and the less massive the secondary[1] and it is common to find in the literature the primary star denoted by A and the secondary denoted by B. The number of multiple systems in an observed region or for a given range of masses is commonly referred to as multiplicity frequency or MF, which is a function of the mass of the system's primary star, where higher primary mass means higher multiplicity fraction (Raghavan et al. 2010) and specifically, it has also been known for almost half a century that at least half of solar-like stars are found within binary systems (Abt and Levy 1976). Moreover, stars are born in multiple systems within molecular clouds (Sadavoy and Stahler 2017; Reipurth et al. 2014) and for this reason multiplicity is a property that is fixed in the early stages of stellar evolution (Kounkel et al. 2019). It has also been shown that pre-main sequence stars have higher multiplicity frequency (Mathieu 1994; Tobin et al. 2016), although these systems can disintegrate as they age due to dynamical interactions (Lada 2006 and references therein).

The physics and evolution of binary stars and subsequent higher order systems is tightly coupled to many fields of astrophysics and their importance cannot be denied (Dorn-Wallenstein and Levesque 2018; Breivik et al. 2019). In stellar physics, accurate parameters (such as stellar mass or effective temperature) are needed for the validation of stellar evolutionary models and the constraining of formation scenarios. The desired accuracy of these measurements, however, can only be reached with a careful study of binary systems, where masses and radii can be accurately derived thanks to their motions under the influence of mutual gravity. Furthermore, binaries are arguably the main source for benchmark stellar measurements (see the review by Andersen 1991). Regarding planetary systems, undetected binarity can introduce strong biases in the parameters derived from the depth of the measured transits (Ciardi et al. 2015). Moreover, the statistics of binary populations are important even in fields such as cosmology and dark matter due to the uncertainties present in their extragalactic distributions, e.g. in dwarf galaxies (Spencer et al. 2018).

Even though the formation and evolution of single stars is broadly understood, this is not at all the case for binary stars, for which the formation and evolution channels are still a topic

[1] Assuming that both stars are located on the main sequence - which we will do throughout this thesis - and that there is no mass exchange between both stars. In the literature it is common to find that the distinction of primary and secondary is done regarding the luminosities. As we will see in section 2.3.2, it can be assumed that luminosity is a quantity that scales with the stellar mass in the main sequence regime and therefore both designations can be treated as equivalent.

of debate. Furthermore, the effect of a binary companion has a crucial effect on the evolution of that particular star which further complicates the constraining of the possible scenarios (Beccari and Boffin 2019). There are several formation mechanisms, which ordered in terms of their importance are: fragmentation of the birth stellar cloud into two or more fragments (prompt fragmentation), which forms binary systems with large separations between the two components (wide binaries), fragmentation of the proto-stellar disk (delayed fragmentation), which forms the secondary from within the disk and results in a close binary system of low separation (Offner et al. 2016), and dynamical mechanisms such as capture, however due to the large separations between objects in the field, this formation channel only becomes significant in dense environments such as clusters or galactic centers. For a more detailed overview on the formation of binary stars, see the comprehensive review on this topic by Tohline 2002 and its references.

For a long time, strong biases have been present in the measurements of binary stars. Furthermore, up until two decades ago, efforts to characterize binary population were replete of biases coming from sample incompleteness, either due to volume or luminosity (Duquennoy and Mayor 1991). This has caused strong disagreements between research publications and only recently instrumentation has allowed to obtain samples to do reliable statistics with (Duchêne and Kraus 2013). However, even with the recent samples, the obtained distributions associated with binary populations and their parameters have been far from reliable, and as we will see in subsection 2.3.3, there is for example no consensus on how their mass ratios are distributed. Even today, in the age of big technological advancements and big data, where large astronomical surveys such as *Gaia* (A. G. A. Brown et al. 2018), *Gaia*-ESO (Gilmore et al. 2012), GALAH (Galactic Archeology with Hermes, Buder et al. 2018) or SDSS-IV (Sloan Digital Sky Survey, Blanton et al. 2017), gather enormous amounts of data, binaries are often overlooked due to difficulties that arise in their detection and classification. These are intrinsic to their nature and require a careful handling apart from the bulk of single stars that are observed by these surveys, as their binary nature can be concealed quite effectively within the abundant observational data. However, due to the increasing amount of gathered data (in the order of tera- and petabytes), its manual inspection and analysis is neither a realistic nor a feasible option anymore, not only because the quantity of data to be analyzed is enormous but also because of the rapid increase in its complexity (high-dimensional data, e.g. multi-epoch measurements)(Süveges et al. 2017).

This accelerated growth and change of the paradigm in the field of astronomy and astrophysics demands deep changes in the methods used to handle the data measured by the aforementioned surveys, which in turn requires a shift from the more classical, mathematical models to more sophisticated, scalable methods for efficient solutions (Pesenson et al. 2010; Baron 2019). These changes in methodology have seen numerous practical efforts recently, of which most have a common denominator: the usage of machine learning. Machine learning is the name given to a large set of statistical tools and algorithms that are used in the computer-assisted data analysis and they rely on inference and patterns to extract useful information from it. The wide-spread of these techniques is tightly bound to the fast increase of computational power of personal computers and the existence of numerous open-source solutions. Moreover, the possible applications of machine learning in the field of astronomy are countless and increase at a rapid pace. With it, it is possible to perform tasks such as deriving stellar labels with The Cannon (Ness et al. 2015), performing chemical tagging for the reconstruction of past stellar aggregations (Kos et al. 2017) and even classifying Kepler data and detect exoplanets (Shallue and Vanderburg 2018). Furthermore, besides applying these methods to study new data-sets for the first time, it is also possible to "revisit" older samples in order to extract new insights and identify patterns that were overlooked in the past.

With machine learning growing in importance within the astronomical community, the ques-

tion of whether it can effectively be applied to identify binary stars from stellar surveys arises. Indeed it can, and it has already been done on several occasions and on different types of binaries, e.g. on eclipsing binaries (Armstrong et al. 2015), from the Kepler-2 mission (Howell et al. 2014) and also on spectroscopic binaries from the GALAH survey (Traven et al. 2016). As we will see in chapter 3, one of the benefits of machine learning is that it possible to use algorithms that require very little input from the user, contrary to more classical techniques, thus reducing the amount of time spent calibration the algorithms and avoiding biases that may arise from it.

In the following we will see the most important types of binary stars and an overview of their inner workings, why we choose to focus exclusively on double-lined spectroscopic binaries and what is special about them regarding their spectrum and their detection.

1.1 Types of binary stars

Due to their abundance, binary stars are classified into several types and subtypes. The convention for the naming of these categories has been, traditionally, given depending the method the binaries were discovered with. Due to this, it is possible for a binary star to be part of several different categories as it might be observable through different methods. In the following we will present the main types of binary stars and some details about their nature.

Visual binaries

Visual binaries refer to those binary systems that can be seen as two individual stars with the aid of optical means, such as telescopes or the bare eye. This implies not only that the stars have an angular separation that can be resolved by the instrument but also that the brightness of the primary does not completely overpower the secondary star as well. It must be noted that even if two stars appear to be close in the night sky, they may not form a binary system, which depends on whether the two objects are gravitationally bound or not. These alignments are known as optical doubles and they are only a product of chance.

Astrometric binaries

In an astrometric binary system, the presence of the companion cannot be directly observed and it has to be inferred from the proper motion of the primary (Southworth 2019). The proper motion of the primary star, if unaffected by the companion, would appear as a straight line across the sky (assuming that the annual parallax has been properly removed from the measurements). However, its proper motion exhibits periodic wobbles that are caused by the motion around a barycenter created by the invisible companion. The first astrometric binary was discovered by Bessel 1844 and is the brightest star in the sky and most representative example for this type of binaries, Sirius. In figure 1.1 below, we can see a depiction of the motion of Sirius and its companion, Sirius B. The dashed line represents the path Sirius would follow on the sky if it was a single star, whereas the thick and dotted lines represent the real proper motion of Sirius A and Sirius B, respectively.

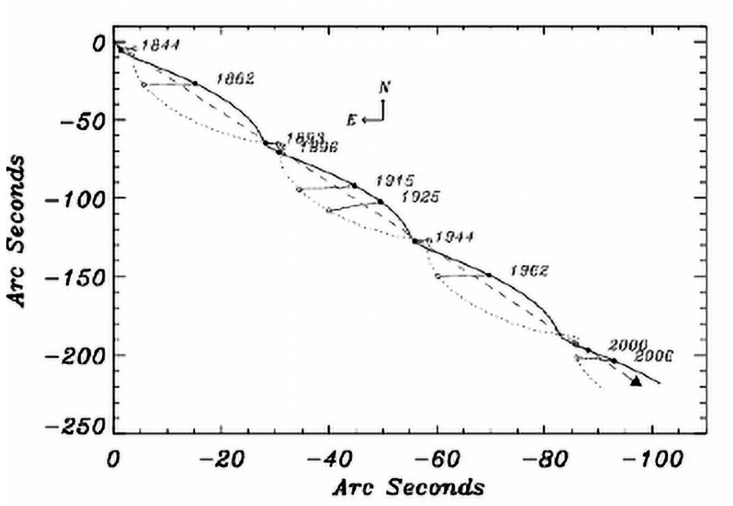

Figure 1.1: Proper motion of Sirius A and Sirius B. Image credit: Jay B. Holberg.

Eclipsing binaries

An eclipse occurs when an astronomical object passes in between another object and the viewer and thus partially or totally blocks the light coming from the other object. Eclipsing binaries occur for viewers on Earth when the observed binary system is at an inclination close to 90° (binary orbit is viewed edge-on). In this kind of systems two types of periodic eclipses occur: the primary eclipse; when the secondary star passes in front of the primary thus blocking part of its light, and the secondary eclipse; which happens when the secondary star is located behind the primary. Both eclipses have a characteristic effect on the light measured from the system, which in turn can be used for accurate measurement of stellar parameters. This can clearly be seen on the measurements presented below in figure 1.2, on a diagram called the light curve. We can see two distinct dips: the primary eclipse at phases 0 and 1; and the secondary eclipse at phase 0.5.

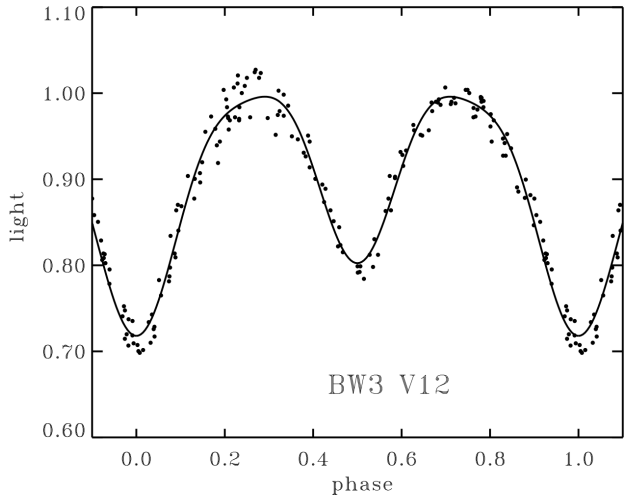

Figure 1.2: Light curve of BW3 V12, a close binary. The dots on the figure are the measurements and the black line corresponds to the fitted light curve. Figure extracted from Rucinski 1996.

Spectroscopic binaries (SB)

The first spectroscopic binary system discovered was ζ Ursa Majoris in 1889 (Pickering 1890). Interestingly, this star system, which was thought to be only an optical double star system (that is, not a real binary but only appear close in the night sky), was discovered to have a companion of its own through the, back then, novel technique of spectroscopy.

Spectroscopic binaries are divided in two categories: SB1, where no line duplicity can be seen in the measured spectrum, and SBn ($n \geq 2$), where n sets of lines corresponding to n components of the system can be seen. The detection of spectroscopic binaries is based either on the direct observation of the line duplicity or through the periodic shifts of lines in multi-epoch observations due to the motion of the visible component around the barycenter caused by the corresponding Doppler shift. A range of diverse spectroscopic binary spectra will be shown in subsection 1.2.2.

1.2 SB2 and their spectrum

Except for a few examples of very close or very large ones, stars appear through telescopes as point sources. Even binary and higher-order multiple systems do appear as such. For this reason, spectroscopy has become one of the best ways astronomers have to study stars as they do not need to be fully resolved to extract important information from. Since the first catalogue of spectroscopic binary stars was published (W. W. Campbell and Curtis 1905), the field of stellar spectroscopy has grown rapidly and has become an indispensable source of astronomical data, not only for detecting multiple stellar systems but also as a source of stellar parameters e.g. temperatures, chemical abundances and radial velocities. Some of the latest published catalogues reach numbers of more than 12000 confirmed spectroscopic binaries (Traven et al. 2020).

1.2.1 Stellar spectra

Through their spectrum and the absorption lines within, stars reveal a fingerprint that not only contains information about the corresponding chemical composition, but also about multitudes of physical parameters of the star itself. Most of this information can be found in a type of features that appears on the spectrum called absorption lines. Absorption is a phenomenon that occurs when light emitted as a black body radiation, or continuum, passes through a cooler region. This cooler regions in stars correspond to the upper layers of the atmosphere and it is there where the cooler atoms and molecules absorb the photons with energies (or wavelenghts) corresponding to those of their respective energetic levels. This absorption can be interpreted as a removal of the photons with those exact wavelengths and thus formation of absorption lines (missing light) in the observed spectrum. Stars are generally assumed to radiate as black body radiators[2] and the deviations from this approximation are caused by absorption processes, which in turn are very sensitive to the stellar parameters. It is this strong correlation between a stars' spectrum and its physical parameters that allows for their precise measurement. Using the previous black body approximation, one can measure the effective temperature according

[2]A black body radiator is an idealized, perfectly opaque object that completely absorbs the incident electromagnetic radiation. If found in thermodynamic equilibrium, it can also emit electromagnetic radiation according to Planck's law, where the radiation is only dependent on the temperature of the black body. It is possible to calculate an approximated spectrum for a star according to Planck's law of black body radiation, which is given by $B_\nu(\nu, T) = \frac{2h\nu^3}{c^2} \frac{1}{e^{\frac{h\nu}{k_B T}} - 1}$ where B_ν corresponds to the spectral irradiance of the object, h and k_B are the Planck and Boltzmann constant respectively, c is the vacuum speed of light, ν is the frequency of the radiation and T is the temperature of the emitting black body.

to Wien's displacement law[3], which is given by the wavelength corresponding to the maximum irradiance value. Other parameters that concern us in this work, the surface gravity, $log\, g$ and the metallicity determined by the iron abundance, [Fe/H] can also be measured from a stars' spectrum, albeit with different methods.

1.2.2 Double-lined spectroscopic binaries: SB2

Spectroscopic binaries have the advantage that they can be observed at larger distances, as their detection depends on the shifts on the spectrum caused by the Doppler effect and this effect does not depend on distance (Carling and Kopal 2012), assuming the quality of the measured spectrum is good enough to allow for the proper detection. Each type of spectroscopic binaries allows for different types of measurements, which in turn allow to probe different aspects of their nature. However the measurement of single- and multiple-lined binaries presents an important difference as well. Whereas in SB1 several (at least two) multi-epoch observations are needed in order to observe the periodicity in the shifts of the spectral lines due to the variation of the radial velocity of the component with visible lines (the primary), for SBn binaries one observation can be enough to allow for detection due to the presence of the multiple lines (Merle et al. 2020). As we will be working with only one spectrum for each star and because they are the most abundant of the SBn type of binaries, we will focus exclusively on SB2 for the entirety of this work. In this work two parameters (among others, Katoh et al. 2013) will be used to characterize SB2 binary systems: the radial velocity difference of the two components and their luminosity ratio.

Radial velocity difference: $\mathbf{\Delta v_{rad}}$

For objects in a circular motion, their velocity has two components: the radial, which is directed towards the center of the motion and the tangential, which is perpendicular to the former. In the case of stars, the radial velocity represents the motion towards or away from the Earth (in the literature then Sun is generally considered as the reference point). It is this motion away or towards the observer that causes the Doppler shifts on the light measured from the star. The Doppler effect[4] is a phenomenon that occurs when there is relative motion between a wave emitting source and an observer, which causes an apparent change in the frequency (and wavelength) of the emitted wave. If the source of emission is moving towards the observer, the perceived frequency will be higher (blue-shift) and vice-versa (red-shift).

A visual representation of this concept can be seen on figure 1.3, with four distinct stages of a spectroscopic binary system's orbit and its effect on the spectrum. On points 2 and 4, both stars have the same radial velocity and therefore there is no perceivable Doppler shift between them when viewed from Earth. On the contrary stages 1 and 3 do show Doppler shifts, as the radial velocities for both stars are opposite and of different value. For 1, the secondary is moving away from the observer and therefore its spectrum appears red-shifted, whereas the primary star moves on the opposite direction, blue-shifting its spectrum. On stage 3, the opposite situation occurs: the primary star is red-shifted and the secondary is blue-shifted. Moreover, it can clearly be seen that the shifts of the primary star are less strong than those of the secondary, which is a consequence of its higher mass.

[3] $\lambda_{peak} = \frac{b}{\mathrm{T}_{eff}}$ where $b = 2.897 \cdot 10^{-3}\, m\, K$

[4] Doppler 1842. Funnily enough, Doppler based his hypothesis on the distinct coloration of binary stars for the effect that carries his name.

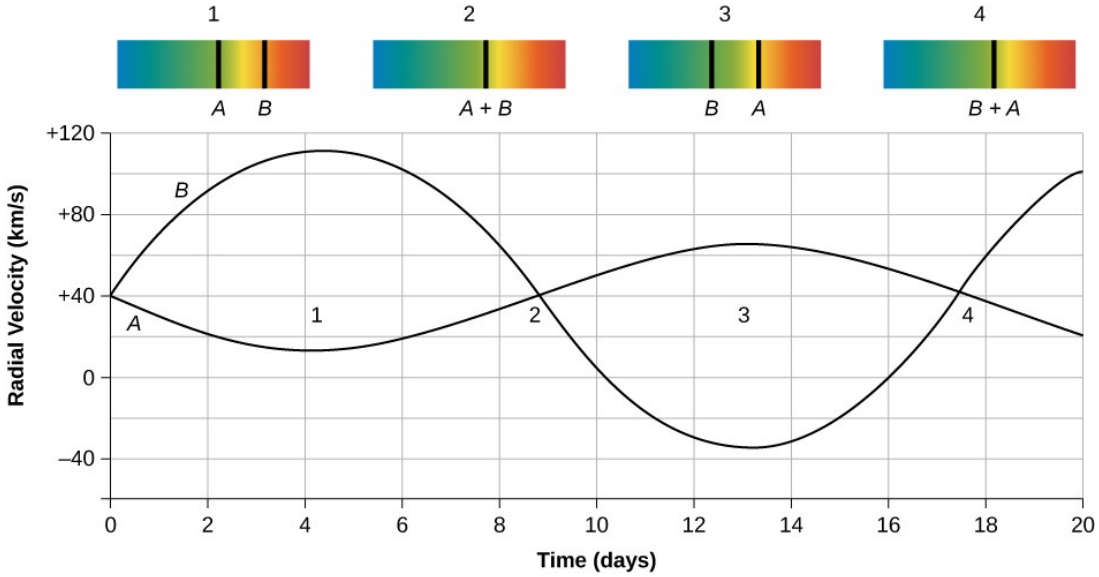

Figure 1.3: Radial velocity curves of the two components for a given spectroscopic binary system. The four boxes on top represent the changes on the measured spectrum, where the numbering represents each of the four distinct stages in the orbit. Note that the radial velocities are measured relative to the motion of the barycenter, which is 40 km/s, therefore a larger velocity than that of the barycenter would be away from the observer and inferior would be towards it. Image credit: Lumen Learning.

In a real spectroscopic survey, the secondary spectra is shifted to the rest frame of the primary and therefore only the secondary appears shifted. Thus one can define the difference between the radial velocities of the two stellar components: Δv_{rad}. This quantity can be interpreted as the net motion of the secondary star away or towards the observer if the primary was located on the barycenter of the system and therefore it would appear still as seen from the sun (this quantity is sometimes named barycentric radial velocity in the literature). This implies that the only the spectrum of the secondary star will be Doppler shifted, moving the spectral features along the wavelength axis. The values at which each line transition of the secondary star occurs are shifted by the factor s, which is derived from the equality for the non-relativistic Doppler effect. For the wavelength at rest λ_o and λ the measured wavelength, we have:

$$\frac{\Delta v}{c} = \frac{\lambda - \lambda_o}{\lambda_o} \tag{1.1}$$

where expanding and rearranging the above terms leads to:

$$\lambda = \lambda_o s, \text{ where } s = \frac{\Delta v}{c} + 1 \tag{1.2}$$

For our purposes, we set $\Delta v = \Delta v_{rad}$.

Luminosity ratio: $\mathbf{L_B/L_A}$

The luminosity ratio is defined as the ratio between the bolometric luminosities of both stellar components in the binary system. It is expressed mathematically as L_B/L_A, where the individual luminosities are given by the Stefan-Boltzmann law:

$$L = 4\pi R^2 \sigma_{SB} T_{\text{eff}}^4 \tag{1.3}$$

where R is the stellar radius, σ_{SB} the Stefan-Boltzmann constant and T_{eff} is the effective temperature of the star. From equation 1.3 it becomes clear that a slight increase in temperature or stellar radius will lead to a very large difference in luminosities due to the R^2 and T_{eff}^4 terms, respectively. As we shall see later, the luminosity ratio plays a fundamental role in the detection of binary stars.

Example of SB2 spectra

Because the features that define the spectra of SB2 binaries will be very important later during their analysis with machine learning, we exemplify a variety of possible cases (for spectra with varying values Δv_{rad}) in figure 1.4 with spectra extracted from the latest GALAH data release.

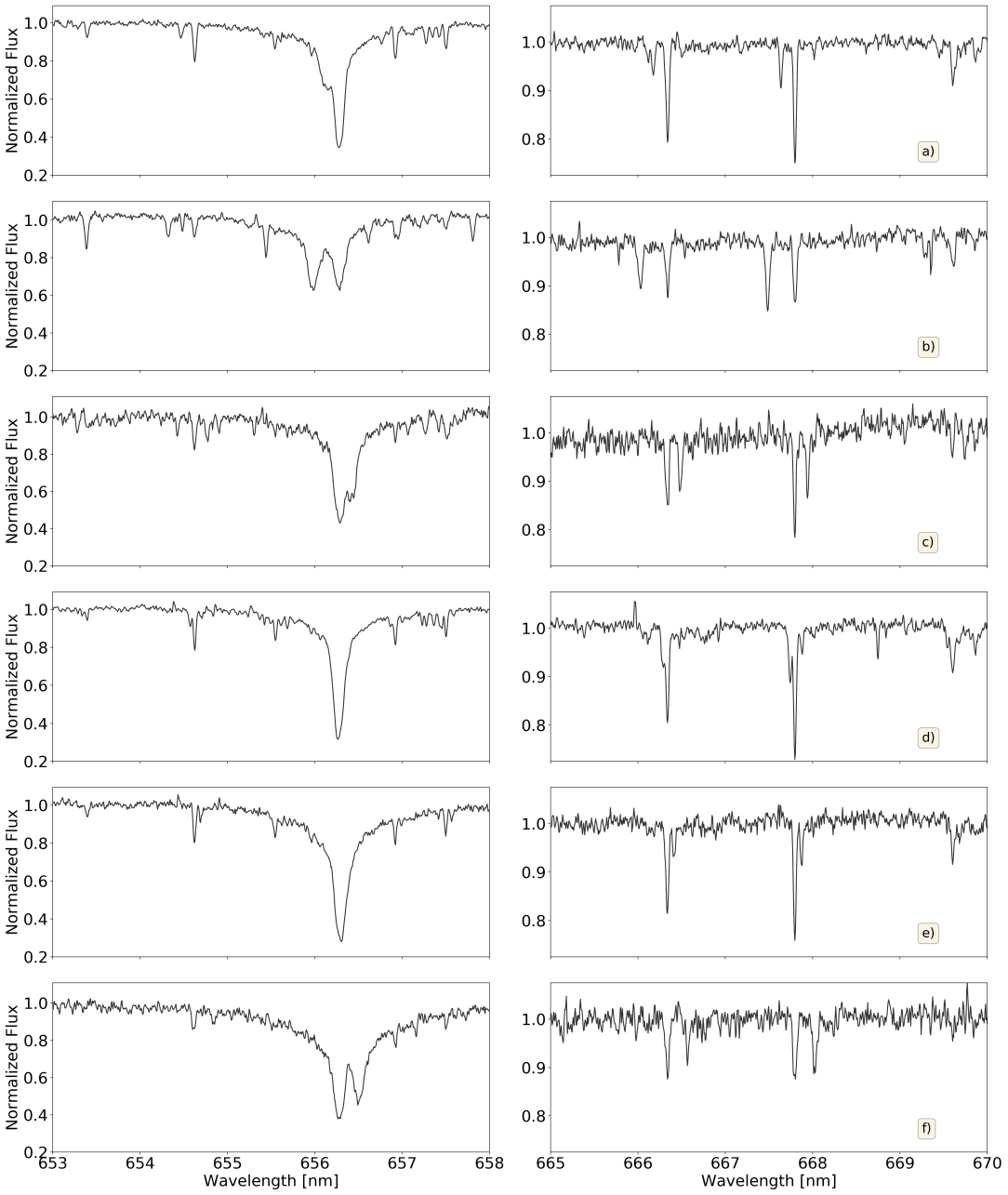

Figure 1.4: Binary star spectra from the GALAH survey. The 6 different systems are ordered following a gradient of increasing radial velocity difference Δv_{rad}. Spectra on rows a) to c) have negative Δv_{rad}, whereas spectra on figures from row d) to row f) show positive values.

In figure 1.4 we show examples for binary spectra with totally blended lines on rows c) and d), spectra with lines that show a clear separation on rows a) and f), or in between examples where the line duplicity can still be seen on rows b) and e). We exemplify this further by showing two regions, one between 653 and 658 nm containing a strong spectral line such as H-α (656.28 nm) on the first column, and adjacent region with weaker lines, between 665 and 670 nm.

1.3 This work

The rapid growth and spread of machine learning methods in astrophysics has been mainly out of necessity (Pesenson et al. 2010). In an effort to mitigate the issue about insufficient amounts of properly characterized spectroscopic binary stars (and binary stars in general) and the incomplete statistics that arise from this, we intend to complement the more conventional methods for detecting of SB2 systems, such as the Cross-Correlation Function (Matijevic et al. 2010), with a combination of two state-of-the-art machine learning algorithms. For this, we will follow an approach similar to that presented in Traven et al. 2016 and Traven et al. 2020.

While the CCF method is effective at detecting binary stars on its own, it requires a series of assumptions and restrictions that have to be i mposed by the user to ensure a high degree of confidence in the results, such as a radial velocity lower limit or which templates should the spectra be compared to. On the contrary, machine learning is designed to identify patterns in the data without much user input based on the peculiarities of the given data-set, which this in turn can avoid the introduction of certain sorts of biases, such as . This has already been done on real stellar spectra, where the nature of each star is not known a priori thus hampering the proper calibration of the methods. To overcome this, we will generate a synthetic spectroscopic survey that contains both single and binary stars and is generated using the GALAH survey as a reference. We do so to stay within realistic margins. This will allow for an effective and realistic calibration of the algorithms in order to maximize the efficiency and detection of binary stars. For this, we will make use of a combination of two algorithms: t-distributed Stochastic Neighbor Embedding or t-SNE (Maaten and G. Hinton 2008), which will generate an overview of the data-set by grouping similar data-points together and separating those that are dissimilar, creating a projection where groups of binaries will be placed in clusters independent from the bulk of single stars, and a second algorithm to select those cluster of data-points that correspond to the binary spectra from our synthetic survey with the name of Density Based Clustering Of Applications with Noise, or DBSCAN (Ester et al. 1996). We expect to obtain results of a similar flavor to those presented in Matijevic et al. 2010, where a similar approach was taken to examine the detection ranges with the CCF.

For this reason it is within the scope of this project to examine and optimize the parameters that go into each of the algorithms in order to detect and extract groups of data-points that represent binary stars from our data set in an automatized manner. For this, we aim to develop a method that is capable of being not only technically successful, but also that is applicable in the context of real spectroscopic surveys such as GALAH, Buder et al. 2018 or APOGEE, Ahumada et al. 2019.

This thesis will be structured as follows: this chapter served as an introduction both to the present paradigm in the study of binary stars, and more specifically that of spectroscopic binaries and to the required theoretical concepts to understand this work. In chapter 2 we go over the synthesis of single stellar spectra and the different models and assumptions involved, as well as the pairing of single stars to form a binary population. Chapter 3 will focus exclusively on the topic of machine learning and an overview of the inner workings of the algorithms selected for this project, whereas in chapter 4 we will explain how we go about applying this methods

to the generated data set. In chapter 5 we will present the most interesting results, as well as their corresponding interpretation. Finally, we summarize everything in chapter 6, where we will expose our concluding thoughts on the feasibility of the developed method and its applicability in the real world.

Chapter 2

Spectral Synthesis

The synthesis of stellar spectra is a technique whose main utility lies in analysis of observed spectra and can be applied in several different ways. It offers a much broader range of versatility compared to traditional methods for spectral analysis, such as the possibility of selecting any combination of fundamental stellar parameters, elemental abundances and wavelength range for a modeled star in order to accurately match its spectrum and obtain the corresponding stellar parameters. However, this flexibility is limited by the quality of the used atomic data, line-lists and the assumptions taken, namely regarding the geometry of stellar atmospheres, e.g. plane-parallel or spherical, and regarding its state, which can be assumed to be in a local-thermodynamical equilibrium (LTE) or non-LTE state. (Husser et al. 2013).

One peculiarity of our work is the usage of synthetic spectroscopic data for the calibration and optimization of machine learning algorithms. The main reason for using self-generated data is that the nature of each data point is known (each star's spectrum is associated to a high-dimensional data point, where each data point contains n flux values, however this will be explained in more depth in the next chapter). To keep our sample within realistic margins, we base our data-set on a real spectroscopic survey by taking combinations of stellar parameters from stars present in it and thus generating a synthetic sub-sample of the survey. However, the data we sample from the survey is comprised only of single stars and because the main goal of this project is the detection of binaries, we need to design a process that will allow us to obtain pairs of stars that could realistically be found in the field, as well as generating the corresponding spectrum directly from the spectroscopic data present in the survey.

In this chapter we will present the steps undertaken to go from the data of a spectroscopic survey with only single stars to a synthetic subsample of it, with both single and binary stars. In section 2.1, we will describe the spectroscopic survey selected for this project and the selection of the sub-sample. On section 2.3 we briefly introduce the tools used for the synthesis of the single spectra, whereas in section 2.3 and 2.4 we elaborate on the algorithms we designed to both create a binary population and to combine the spectra of the individual components, respectively. At the end of the chapter in section 2.6, we present a small set of synthetic binaries to show that the synthesis procedure was successful.

2.1 The GALAH survey and the selection of single stars

GALactic Archeology with Hermes (Buder et al. 2018) or GALAH for short, is a spectroscopic survey with the goal of obtaining large-scale sample of high resolution spectra to serve as a complement of the *Gaia* mission (Gaia-Collaboration et al. 2016; Gaia-Collaboration et al. 2018) for which it will deliver key chemical information about the stars observed in it. The GALAH survey uses the HERMES spectrograph (High-Efficiency and high-Resolution Mercator Echelle

Spectrograph, Raskin et al. 2011), which operates in four different spectral bands with a width of roughly 25 nm, the ranges of which are shown in table 2.1 The mean resolving power of the GALAH survey is around 28000, with an estimated signal-to-noise ratio (SNR) of at least 100 (this number varies depending on the wavelength).

Band	$\lambda_{\min}$ (nm)	$\lambda_{\max}$(nm)
Blue	4718	4903
Green	5649	5873
Red	6481	6739
IR	7590	7890

Table 2.1: GALAH spectral regions.

A data-driven approach is used in the GALAH survey to estimate the stellar parameters and the elemental abundances. A training sample for the data-driven algorithm composed of high quality GALAH spectra is analyzed using the spectral synthesis code Spectroscopy Made Easy or SME for short (Valenti and Piskunov 1996; Piskunov and Valenti 2017), yielding reliable parameters for these spectra. The sample is then used as by *The Cannon* (Ness et al. 2015) to generate a spectral model, which will in the end be used to analyze the rest of the spectroscopic data captured in the survey. As a last step, the quality of results for the measured spectra is assessed using diverse statistical methods and the quality of the measured spectra is indicated using various flags, shown in Table 5 of Buder et al. 2018.

The GALAH survey was chosen because recent work shows that it presents a high proportion of detected SB2 to single stars, between 2-3% (Traven et al. 2020). This high percentage of SB2 allows for more reliable statistics and will be important source of data for future studies. Furthermore, GALAH is equipped with a relatively high resolution and high signal-to-noise ratio. This is very helpful in resolving double lines, which as we mentioned previously is essential for the detection of double-lined spectroscopic binaries.

Data selection and dwarf stars

To select the appropriate data from the GALAH survey and define the data-set this work will be based upon, we apply two filters. First, the stars with the most reliable parameters are selected, i.e those that are not marked with any quality flag (which means that they present accurate parameter measurements and have no known issues). Further filtering is done under an assumption presented in Matijevic et al. 2010 that mostly dwarf, main-sequence stars will be observable as double-lined spectroscopic binaries. For this reason, we will be focusing our study only on dwarf stars, which we assume to be main-sequence even if the number of dwarfs and giants in the GALAH survey is very similar. This assumption is further supported by several arguments:

1. For the system to be detectable regarding the luminosity difference between the two components, both must be in the same phase (main-sequence or giant) at the same time. However, for two stars to be found in the red giant phase at the same time, their masses cannot vary more than $\sim$1% from each other, i.e they must have rather similar masses, as the main-sequence lifetime of a star scales approximately as $\tau/\tau_{\odot} \propto (M/M_{\odot})^{-2.5}$.

2. The lifetime of the giant phase is much shorter than that of a dwarf on the main sequence in the mass regime for typical stars of GALAH and therefore the chances of finding a giant-dwarf system is further reduced.

3. For giant stars, the minimum size of the orbit is larger than that of dwarf stars. This translates to a smaller maximum radial velocity difference and as a consequence less separated lines, thus meaning harder detection.

Additionally, Matijevic et al. 2010 argues that even if the chances of finding a binary system with two giant components is small, and in case of doing so, the results would not be drastically altered as the spectrum of a giant star is comparable to that of a dwarf with similar effective temperature. To filter accordingly, we use equation 1 from Zwitter et al. 2018 given by:

$$\log g = g_1 + (g_2 - g_1)\,\frac{(T_1 - T_{\rm eff})}{(T_1 - T_2)} \tag{2.1}$$

where $g_1 = 3.2$, $g_2 = 4.7$, $T_1 = 6500\,\mathrm{K}$ and $T_2 = 4100\,\mathrm{K}$. The constants $g_{1,2}$ and $T_{1,2}$ determine the points where the division between dwarfs and giants is traced.

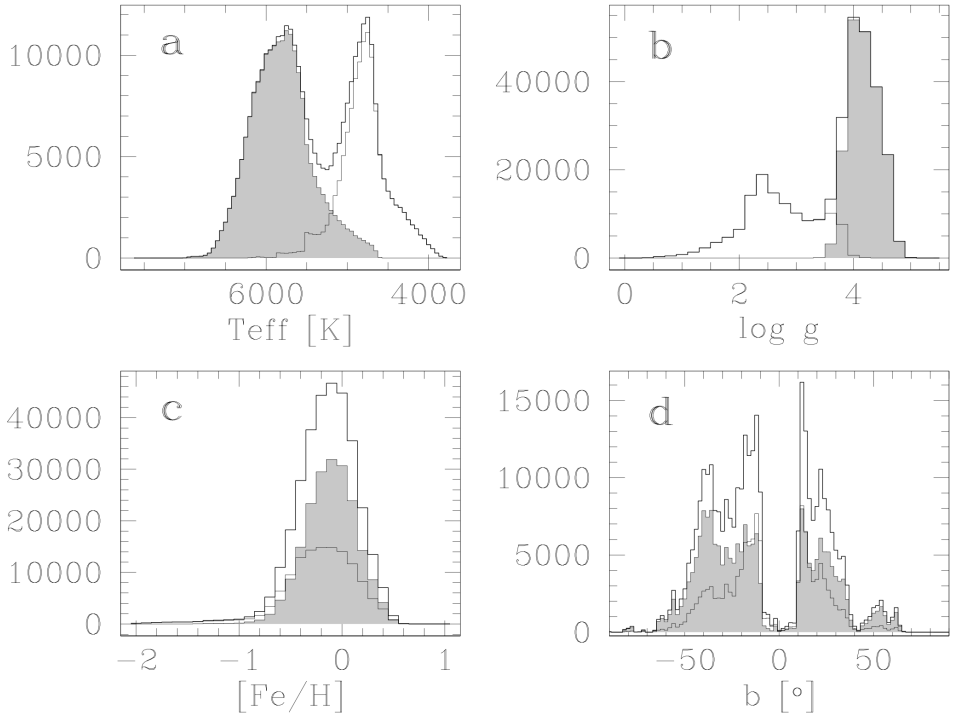

Figure 2.1: Parameter distributions for GALAH DR2. The thin grey line corresponds to the division between dwarfs (highlighted with the shaded grey histogram) and giants given by equation 2.1 and the thick grey line represents the complete sample. Figure from Zwitter et al. 2018.

On figure 2.1 we can see the parameter distributions of the stars from the GALAH survey that are not marked with any quality flag where plot a) is for the effective temperature, plot b) for the logarithmic surface gravity and c) for the metallicity (plot d) refers to the galactic latitude, which we will not be using in this work). It is interesting to note that both distributions for the effective temperature and surface gravity show a double peaked feature, which is caused by two distinct populations of stars, which according to equation 2.1 these are dwarf and giant stars. The metallicity distribution does not present this feature as both populations are centered around a value slightly lower than the solar value. A Kiel diagram of the stars without flags and already divided in dwarfs and giants can be seen on figure 2.2, where the black line is traced according to expression 2.1.

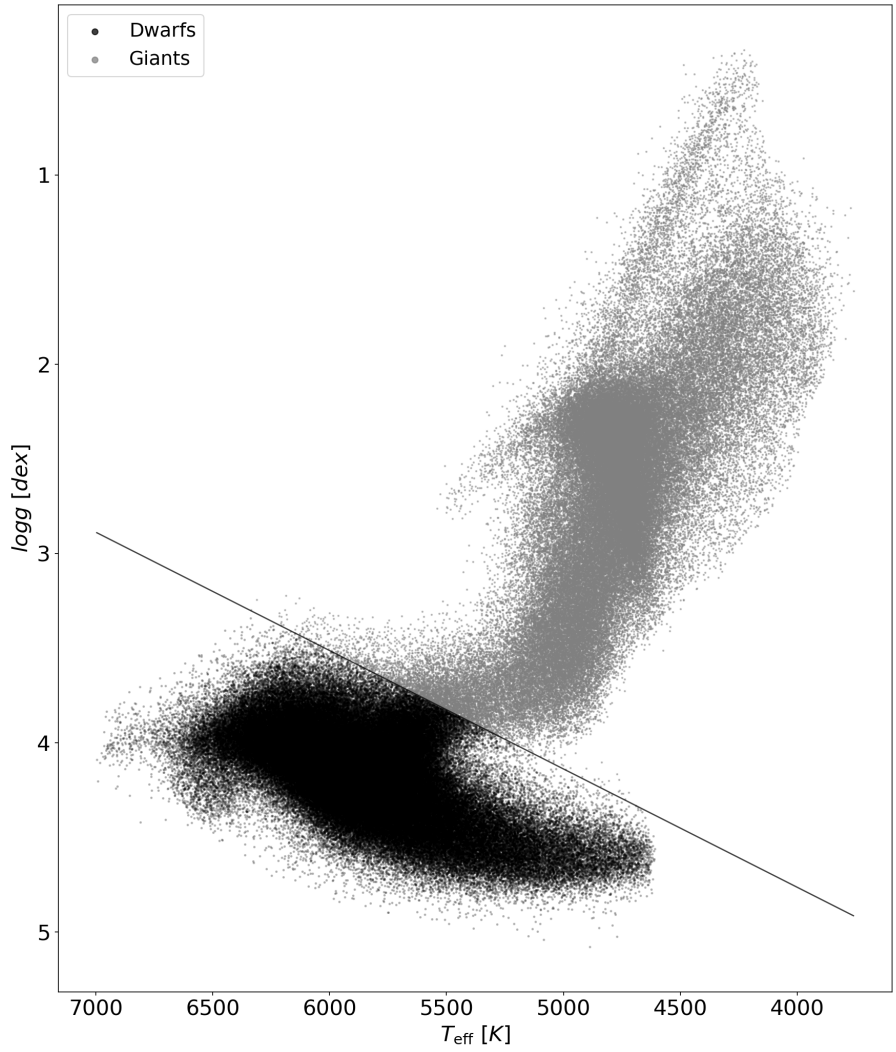

Figure 2.2: Kiel diagram of the GALAH stars not marked with flags, divided into giants (grey points) and dwarfs (black points) according to Zwitter et al. 2018.

Although the GALAH survey has a flag to mark whether a star is a binary or not, as we introduced in chapter 1, the determined binarity in stellar surveys is not always correct and can suffer from errors and false positives. For this reason, it is possible that undetected binaries might be within the stars selected for this work. Even if this is the case, we will be assuming that all of our dwarf stars are indeed singles and will use them in our later analysis indifferently and that the combination of their stellar parameters are still within realistic margins. In terms of numbers, the second public release of GALAH (GALAH DR2) contains information about 342682 stars (Buder et al. 2018), out of which 264227 were selected and shown in figure 2.2 as they were not marked with any flag. Using the prescription from equation 2.1, we found 163279 dwarf stars (and 100948 giant) in the filtered GALAH data-set from which we randomly select 100000 stars. This is a large enough amount of stars to yield statistically meaningful results but it is still manageable regarding the available computational resources.

2.2 Single spectra synthesis

The goal of spectral synthesis is, for a given set of stellar parameters, to simulate the processes that take place within the stellar atmosphere and propagate the emitted photons to obtain the corresponding spectrum, which is enabled by the theory of the radiative transfer. Although it

can be pursued through 3-D hydrodynamical simulations to a high degree of accuracy, the computational cost of this is far too large for our purposes. Consequently, we will use *turbospectrum*, by Plez 2012, which is a spectral synthesis code and coupled with it, a model under the approximation of 1-D atmospheric under the assumption of local thermodynamical equilibrium (LTE) with the name of MARCS (Gustafsson et al. 2008). Together with the code and the model, information about the spectral line transitions occurring within the model atmosphere, as well as solar abundances to extrapolate from are needed. For this purpose, we will be using the line-list from *Gaia*-ESO, GESv5 (Asplund et al. 2013) and the solar abundances from Asplund et al. 2009.

All of the above is conveniently contained within a Python wrapper called iSpec (S. Blanco-Cuaresma et al. 2014; Sergi Blanco-Cuaresma 2019) which additionally allows for the selection of diverse options of atmospheric grids, the spectral synthesis codes and line-lists. iSpec made possible to manually examine different model atmosphere and synthesis code combinations, which allowed us to pick *turbospectrum* mainly due to its speed and reasonably accurate results. For the model atmosphere and the line-list, we followed Buder et al. 2018. GESv5 was chosen because it contains information of transitions occurring between wavelengths 420nm and 920nm and thus covering the four GALAH spectral regions.

For the spectral synthesis itself we sampled randomly a subset of 100000 dwarf stars from the filtered GALAH data-set and fed their parameters into the synthesis code. These parameters are: T_{eff}, $log\, g$, [Fe/H], alpha enrichment [α/H], micro-turbulence v_{mic} and projected stellar rotational velocity $v \sin(i)$; where i is the inclination. The linear limb darkening coefficient (Schwarzschild 1906) was arbitrarily fixed to be the same for all of the considered stars at 0.6, as it is not measured by the GALAH survey and was needed as an input in *turbospectrum*. To match the GALAH survey, we use the values of 28000 for the resolving power and 0.004 nm/px for the sampling. Our synthetic spectra has flux values for wavelengths between 450 and 900 nm, which covers all four GALAH spectral regions (shown in table 2.1). This wide range of wavelengths allows for a later, more specific selection of any region, regardless if it was observed by GALAH or not, and as we will see, the analyzed spectral region plays a major role on the discovery of SB2.

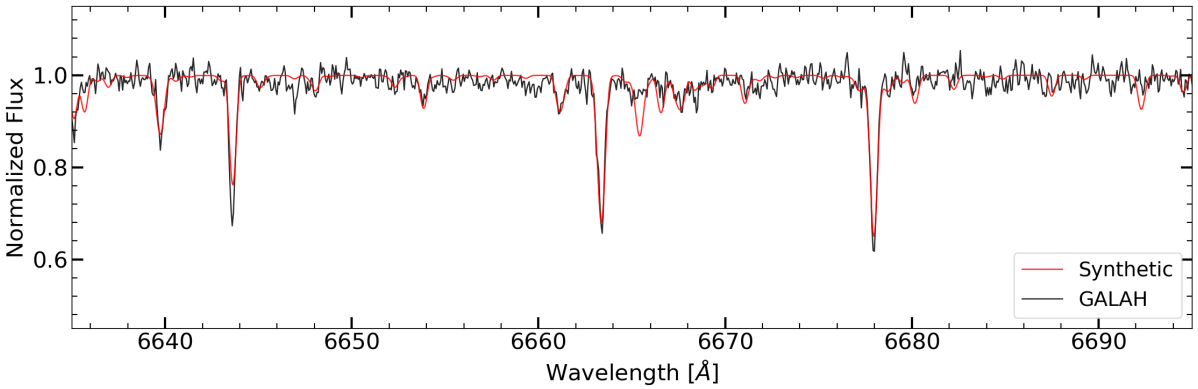

Figure 2.3: Comparison of a real spectrum from the GALAH survey to its synthetic counterpart. Both spectra correspond to a star with $T_{\text{eff}} = 5560\,K$, $log\, g = 3.85$ and [Fe/H] $= 0.25$.

On figure 2.3 two instances of spectra for the same star are shown: the real spectrum measured by the GALAH survey and the spectrum we synthesized using the tools presented before. The similarities can clearly be seen, as almost every line is properly represented, albeit their depth appears to have been slightly underestimated. This is probably because the parameters from the GALAH survey were determined using a different model than ours (SME and *The Cannon*) and also due to non-linear effects that are not contemplated by the used model. We

expect this underestimation in the line depth to not have a major influence on the later analysis, as the same behavior will be carried further on to every other synthetic spectrum. Moreover, on the figure we can also see that the synthesized spectrum does not have noise yet, as it will be added later in section 2.5. All in all we can conclude the synthesis of single star spectra was successful and that the generated spectra match well their real counterparts within margins acceptable for our future analysis.

2.3 Binary population and the pairing algorithm

2.3.1 Pairing algorithm

The generation of binary spectra requires several more steps than their single counterparts because we do not (nor can do so) sample parameters specific to binary stars from any existing data-set of spectroscopic binaries, rather we create such a population ourselves. To do so, we create an algorithm to set up pairs of single GALAH dwarf stars for which we have spectrum synthesized and once the pairs are defined, it is then possible to combine their individual spectra. The main goal behind our pairing algorithm is therefore to obtain pairs of stars that could realistically be observed. It works as follows:

1. Our pairing algorithm begins by sampling the mass of the primary star and the mass ratio between the secondary and primary star of the given system. Their distributions are given by the initial mass function or IMF (Salpeter 1955; Kroupa 2001), and diverse functional forms of the mass ratio distribution, such as those given in Hogeveen et al. 1991; Raghavan et al. 2010; Duchêne and Kraus 2013. We use the power-law from Salpeter 1955 as our IMF for the sake of simplicity, as it only differs of more modern forms such as Kroupa 2001 or Chabrier 2003 in the low mass regime (which we do not contemplate). To sample a random primary mass from this power-law, we apply the convenient formulation found in equation 1.8 of Eggleton 2006 and it is given by

$$M_A = \frac{M_0}{(1-X)^{0.75}}, \tag{2.2}$$

where X is a uniformly distributed random number and M_0 is mass at which we truncate the distribution, set to $M_0 = 0.55$. This lower boundary M_0 is chosen as such, because otherwise the star would be to faint to be detected by GALAH. For the mass ratio $q = M_B/M_A$, where M_A and M_B are the stellar masses of the primary and secondary components of the binary system respectively, we will use a fairly recent result from Duchêne and Kraus 2013 (in subsection 2.3.3 we will discuss the mass ratio in more detail). Duchêne and Kraus 2013 claims that for stars with masses between $0.7M_\odot$ and $1.3M_\odot$, q is distributed as a single power-law with the exponent of $\lambda = 0.3$, $f(q) \propto q^\lambda$ (we assume that these limits can be extended and thus applied to the whole range of masses we will be dealing with). From q, the secondary mass is defined as:

$$M_B = qM_A \tag{2.3}$$

2. It is possible to approximate the effective temperature for each of the system's components just by knowing their masses under the assumption that both are non-evolved and non-interacting main sequence stars. For this we use what is known as scaling relations, which will be presented in more depth in subsection 2.3.2. With the approximated $T_{\rm eff}$, we cross-match the GALAH survey and search for a star with a temperature within a range of $\pm$75K of the computed one. Both $log\, g$ and [Fe/H] were left undefined, as they will be part of the selected star's own set parameters. If the search for a given primary mass

and its corresponding theoretical $T_{\rm eff}$ results in several matches, we randomly select one of them. This ensures that the individual star is not only realistic regarding its parameters but also that we have a synthetic spectrum for it (because we synthesized it from the GALAH sub-sample we selected).

3. If the search for a primary star for a given mass is successful, the next step is to find a suitable companion. We assume that for $M_A \geq M_B$, the following equalities of $log\, g_A \leq log\, g_B$ and $[\mathrm{Fe/H}]_A \approx [\mathrm{Fe/H}]_B$ must hold. The theoretical $T_{\rm eff}$ of the secondary star is calculated in the same way as for the primary and the allowed ranges for its cross-match are also maintained. If the search of the selected dwarf stars with the previous conditions is successful, then the secondary star of the system for a given primary mass and mass ratio is found. If the search for a secondary star does not deliver a result, then we remove the primary and start the process again for the next one.

We do not contemplate an exact value for orbital period, eccentricity, or inclination to our binary systems as their effect on the composite binary spectrum will be effectively replaced by a more observational quantity, the difference in radial velocities of both components (see subsection 1.2.2 in chapter 1). For a comprehensive review of different binary pairing algorithms found in the literature and their consequences in the resulting binary population, the reader is referred to Kouwenhoven et al. 2008.

2.3.2 Empirical scaling relations

The term empirical scaling relations refers to a set of proportionality equations that are defined through observations. In stellar astrophysics, the most common of these equalities is the mass luminosity relation (MLR) for main-sequence stars, which has been known and studied since the beginning of the 20th century. It was discovered independently by Hertzsprung et al. 1923 and Russell et al. 1923, and shortly after by Eddington 1926. Since then, the MLR has been revisited in uncountable occasions, more recently by Moya et al. 2018 and Eker et al. 2015; Eker et al. 2018. On the contrary, the mass-radius relation (MRR) has been studied only since the second half of the 20th century such as Plaut 1953; Demircan and Kahraman 1991 and more recently also by Eker et al. 2015; Eker et al. 2018. Although we will not be using the MRR directly, it is useful to note that a mass-temperature relation (MTR) can be derived from a combination of the MLR, MRR and equation 1.3.

Classically, the MLR has been described as a power-law with the exponent $\alpha = 3.5$. For this work however, we fitted a power-law to the results presented Eker et al. 2018. Because the results also contained data relating the stellar mass to the $T_{\rm eff}$, we were able to obtain both a MLR and a MTR as power-laws with $\alpha_{MLR} = 4.5$ and $\alpha_{MTR} = 0.38$ and thus

$$\frac{L}{L_\odot} \propto \left(\frac{M}{M_\odot}\right)^{4.5} \quad \text{and} \quad \frac{T_{\rm eff}}{T_{\rm eff,\odot}} \propto \left(\frac{M}{M_\odot}\right)^{0.38} \tag{2.4}$$

2.3.3 The mass ratio distribution

The question of whether binary systems follow an IMF-like distribution in their masses and the subsequent distribution of their mass ratios, $f(q)$, has been and still is a matter of debate (Duchêne and Kraus 2013). Early efforts in this matter, such as Kuiper 1935, suggested that the observed binary populations were consistent with random pairing (both stellar masses were sampled from the same IMF and are therefore uncorrelated), although this hypothesis has been long rejected by observations (Duchêne and Kraus 2013). More recent efforts have results that include a uniform distribution (Mazeh et al. 2003) a uniform distribution with an excess of twins with $q \approx 1$(Van der Swaelmen et al. 2019), a distribution with two assymetric peaks at

$q \approx 0.2$ and $q \approx 0.8$ (Goldberg et al. 2003), a decreasing power-law (Ducati et al. 2011) or even an increasing power law (Hogeveen et al. 1991) among several others. This collection of very dissimilar distributions for q is a probable consequence of the different data-sets used in each of the studies, which could indicate a strong dependency of the stellar type, and also of the biases and selection effects present therein, where a correction has been deemed necessary on multiple occasions, e.g. Hogeveen et al. 1991.

Duchêne and Kraus 2013 present in their comprehensive review, after careful examination of binary data-sets for several masses and spectral classes , a power-law approximation for different mass bins. As previously mentioned, we will be using their results for the range corresponding to solar-type stars and assuming it is valid for all of the masses we have in our data-set. Using equation 2.4 and inverting the MTR, we get that the lowest and highest stellar masses in our data-set are $\sim 0.55 M_{\odot}$ and $\sim 1.65 M_{\odot}$ respectively.

2.3.4 Parameter distribution of the synthetic binary population

We set the number of synthetic binaries to be 5% of their single counterparts, although recent estimates of recovered SB2 systems in spectroscopic datasets (e.g. around 2-3% of the whole dataset in Traven et al. 2020) are lower. We did this for two reasons: first, a higher percentage of binaries with respect to single stars means that we are able to better cover the parameter space of the binary systems and second, because during the analysis some of the binary stars will not be recovered and we need a fairly large number of recovered binaries to do reliable statistics with, even in the worst case scenarios.

On figure 2.4 we show the distributions of $T_{\rm eff}$, $log\, g$, mass and both the mass and luminosity ratios for both the primary and secondary stars of all the 5000 synthesized binary system. These distributions differ from those belonging to the original sample of selected dwarfs from GALAH shown in figure 2.1 due to the conditions imposed to the pairing algorithm from subsection 2.3.1. The surface temperature for the secondary stars peaks at lower temperatures than that for the primaries due to the imposed condition of $T_{\rm eff,A} > T_{\rm eff,B}$. A similar situation is seen on the distribution for the stellar masses of both components. There, the mass of the secondary stars peaks at the lower values while the primary masses present a peak at slightly sub-solar values, again due to the condition $M_A > M_B$. The fact that the panels for the temperatures and the masses are very similar is due to the scaling relations from subsection 2.3.2 used to compute the temperature from the mass. The distribution for the surface gravity of the secondary components shows a peak at higher values than that of the primary. On the contrary, the distribution of the primary star $log\, g$ shows a double peaked feature, with one peak at lower values and one at higher. This is most probably due to both the random sampling from the GALAH dwarfs, which present two somewhat distinct populations, one of pure main-sequence stars and one of on the turnoff as seen in figure 2.2, and due to the interplay between the conditions imposed in the pairing algorithm and the tendency of it to produce twin pairs (binary systems of components with similar parameters). The luminosity ratios of the binary systems are almost flat distributed and resembles the one used by Matijevic et al. 2010 (however, in our case is a consequence of the pairing algorithm and not a distribution we sample from). The mass ratio distribution $f(q)$ is a power-law selected from Duchêne and Kraus 2013, although there are no stars with mass ratios below 0.3 because the range of parameters from the sampled dwarf stars cannot reach lower values. The effects of the conditions from the pairing algorithm are further explored in figure 2.5, where we can clearly see the effects the imposed conditions from subsection 2.3.1 have on the synthesized binary population.

It is interesting to note that although we tried several combinations of initial primary mass and mass ratio distributions for the pairing algorithm, all converged to results very similar to

those presented in figures 2.4 and 2.5. This suggests that because we do not use random sampling for the binary pair generation, the shape of the resulting distributions is strongly influenced by the dependency of the secondary star's parameters on those of the corresponding primary and so not only by the sampling from the distributions of M_A and q.

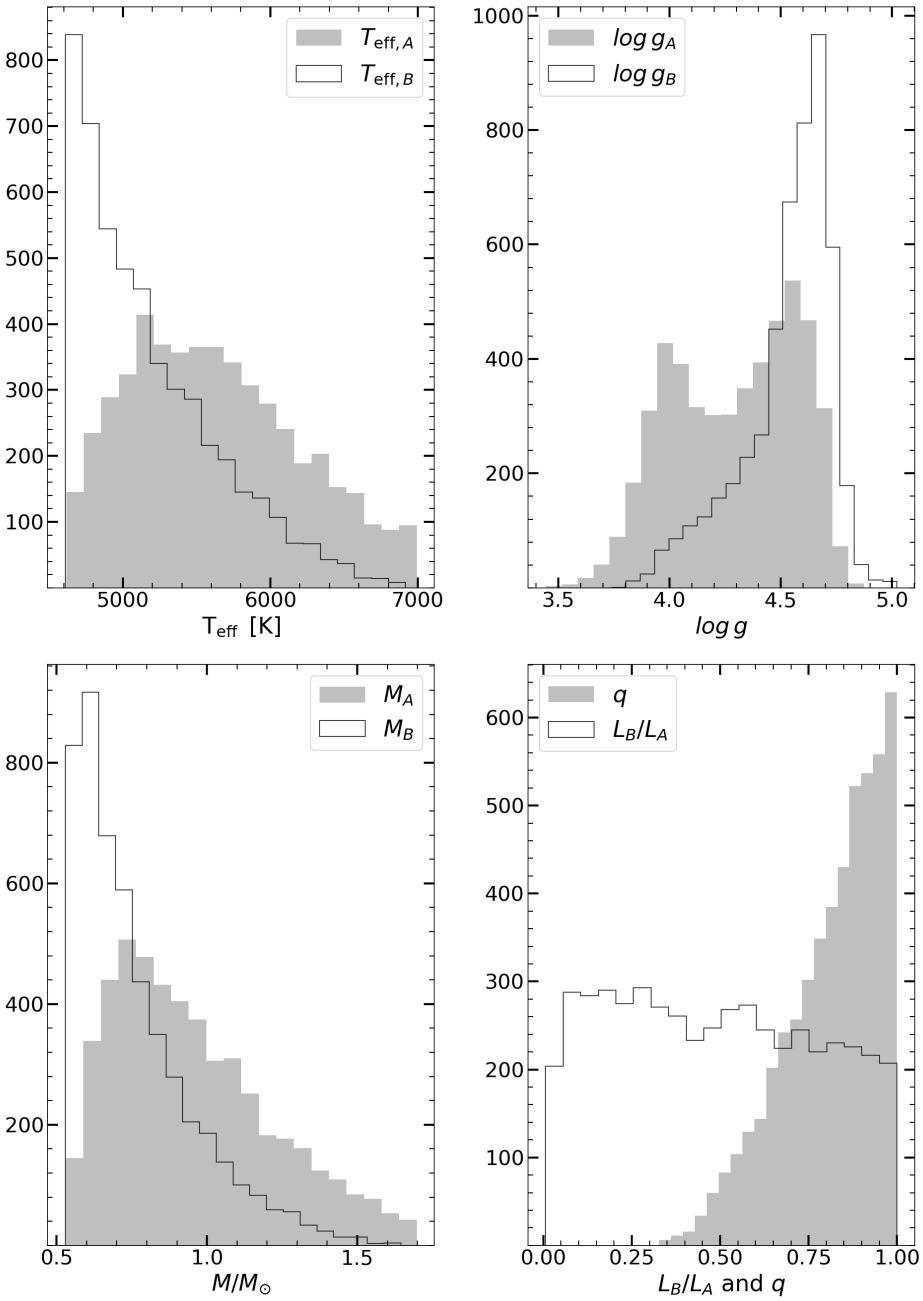

Figure 2.4: Parameter distributions of the synthesized binary population. The grey shaded histogram corresponds to the primary star and the black line corresponds to histogram of the secondary, with the exception of the last panel where both display the properties of the system.

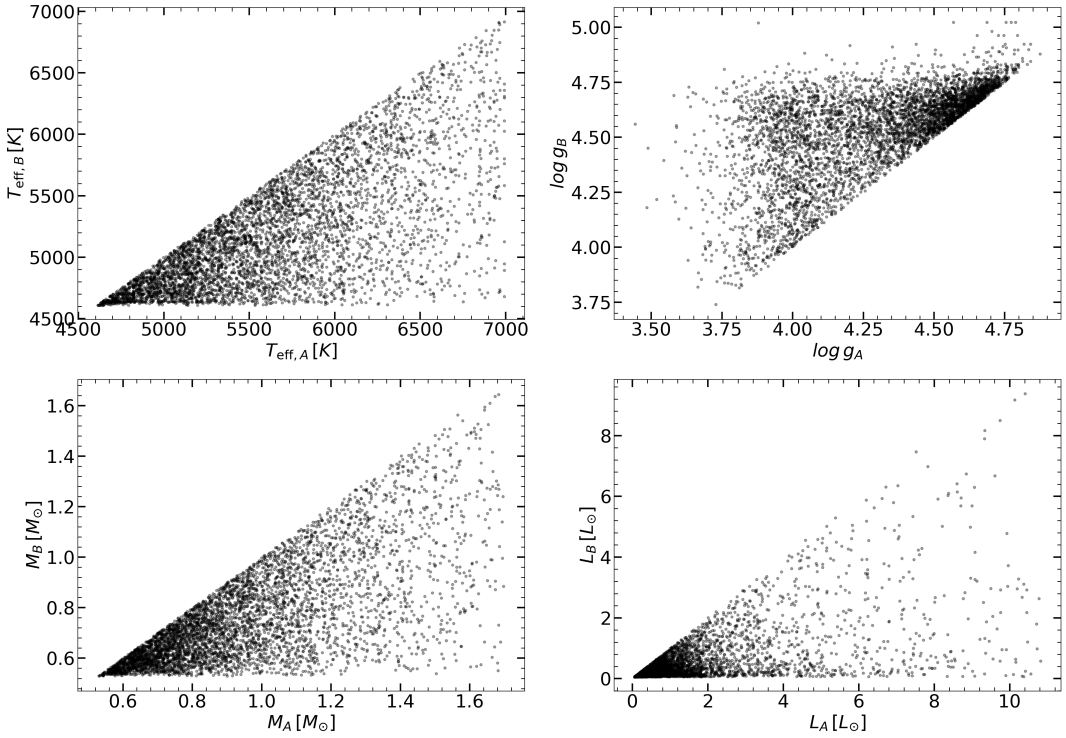

Figure 2.5: Stellar parameters of the primary star against those of the secondary.

2.4 Synthetic binary spectra

2.4.1 Combination of single spectra

With the synthetic binary population, the only thing left to complete the synthetic spectroscopic survey is to combine the spectrum of a primary and secondary stars from the pairs defined in 2.3. For this, we created an algorithm with the following steps:

1. First, a value of Δv_{rad} is assigned to the secondary component of each one of the defined pairs. The value is sampled from a Gaussian distribution that is defined to approximately match the one given in Matijevic et al. 2010 (which in turn corresponds to that what is mostly seen in observations), with the parameters $\mu = 0$ and $\sigma = 30$ km/s. The radial velocity difference is directly responsible for the Doppler shift experienced by the spectrum of the secondary star. The shift is given by the factor s, defined in equation 1.2, and it is applied by direct multiplication of the wavelength range the single spectra was synthesized for. The primary star is fixed at the zero-velocity frame of reference. For a wavelength range $\boldsymbol{\lambda} = \{\lambda_1, \lambda_2, ... \lambda_n\}$ where n is the length of the range in nm times the sampling value used in the synthesis, the resulting shifted wavelength range is $\boldsymbol{\lambda}_s = s\boldsymbol{\lambda}$.

2. After the wavelength shift, the wavelength ranges on which the spectrum of the primary and secondary star are defined will have different start and end points. To avoid having extrema with the contribution of only one star, both spectra are interpolated onto a new, common wavelength grid.

3. For the last step, we combine the spectrum of both system's components using a weighted sum. We use for the combination of the fluxes the luminosities we previously computed with the MLR from equation 2.4 and a set of coefficients found in equation 4 from Cotar et al. 2019, which serve as the weights and are given by

$$a = \frac{1}{1 + r_{1,2}} \quad \text{and} \quad b = \frac{r_{1,2}}{1 + r_{1,2}} \tag{2.5}$$

where $r_{12} = L_B/L_A$. The combination of the two single spectra can be then expressed as:

$$f_{\text{binary}} = af_A + bf_B \tag{2.6}$$

where f_A and f_B are vectors containing the flux values of primary and secondary stars' spectrum and f_{binary} the vector containing resulting synthetic spectrum of the binary system.

2.5 Noise

The signal-to-noise ratio, or SNR, is a measure of the strength of the signal against the background noise (Welvaert and Rosseel 2013). In the literature, it is assumed very often that the spectral noise can be defined as Gaussian distributed (that is, noise is a random fluctuation governed by a Gaussian distribution). Assuming the real signal can be represented by its expectation value and, then the noise can be quantified by the standard deviation of the distribution. With this, the signal-to-noise ratio can be defined as:

$$\text{SNR} = \frac{\mu}{\sigma} \tag{2.7}$$

Adding noise to all the spectra we created for our synthetic survey is necessary in order to achieve certain resemblance with that found on real spectroscopic surveys. To do so, we sample the noise values from a Gaussian distribution by choosing some value of SNR, setting the mean value at the level of continuum ($\mu = 1$), and thus obtaining the σ from equation 2.7. These sampled values are then added to the corresponding vector of flux values. The result from this can be seen on figure 2.6, where the same spectra shown in figure 2.3 can be seen, this time with noise.

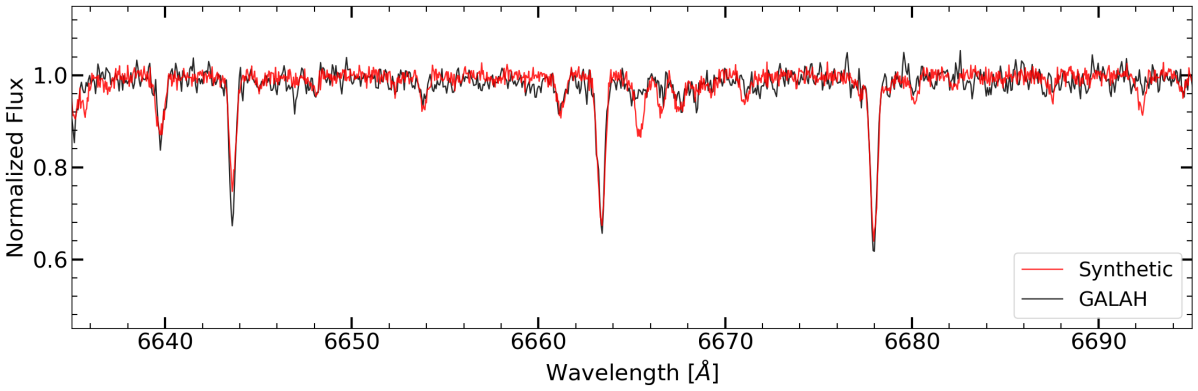

Figure 2.6: Comparison of real spectrum to its synthetic counterpart. The both spectra are the same as those shown in figure 2.3, with the only difference that here a SNR of 100 was used to match the average value used in GALAH.

2.6 Synthetic spectroscopic survey

The synthesis process described in this chapter results in a synthetic spectroscopic survey composed of single and binary star spectra. An example of a synthetic single spectra was already shown in figure 2.6. In figure 2.7 we show a sample of 6 different synthetic binary systems with

SNR of 100 with two different spectral ranges for each, between 535 and 540 nm and between 680 and 685 nm. The spectra are ordered according to an increasing radial velocity difference. The luminosity ratios were selected to be larger than 0.85, so the line duplicity could be properly shown.

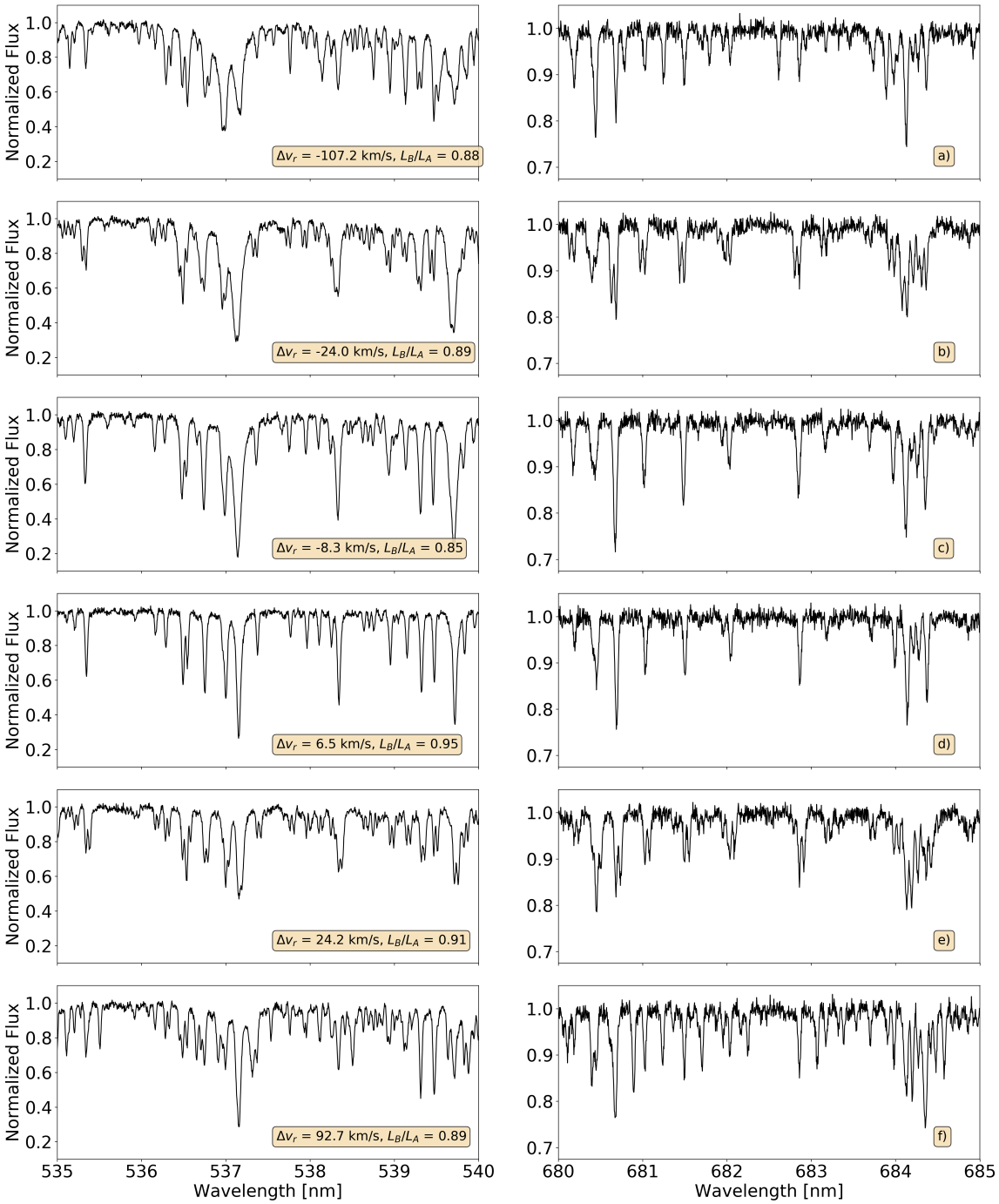

Figure 2.7: Sample set of synthetic binary spectra with SNR of 100 and increasing value of radial velocity difference from top to bottom. Each spectra (row) is shown as two separate regions (columns) in spectral regions between 535 - 540 nm and 680 - 685 nm.

Chapter 3

Machine Learning

Machine learning is a name given to a set of techniques and algorithms implemented by computers to extract useful information from data and make predictions based on its properties. It is considered a sub-field of artificial intelligence. Machine learning methods methods are designed to automatically generate an analytical, mathematical model in order to perform actions "without being explicitly programmed to do so" (Samuel 1959). Although the vast majority of the algorithms that are used within the realm of machine learning were created decades ago, its growth in popularity and importance is tightly bound to the increase of computational power and the decrease of the price per unit, which was not sufficient until recently (following Moore's law). This has made them readily available for the average user. Machine learning methods are traditionally separated in two major groups (Ayodele 2010):

- Supervised methods: represent the most popular category within machine learning. Algorithms are trained using labeled data, which contains examples of data pairs with the correct labels (inputs and their corresponding output values). By examining the training data-set, the computer is able to derive an approximate relation between the training data and their corresponding labels, which allows it to predict labels for new, unexplored data. However, this approach has several downsides such as the danger of overtraining (the algorithm is capable of high accuracy predictions when analyzing the training set but it does not generalize well to unseen data) or the introduction of biases caused by a poorly chosen or wrongfully labeled training set. Linear regression and neural networks are very common examples for this category of algorithm.

- Unsupervised methods: contrary to the supervised methods, these do not require any type of training data-set to perform predictions. Instead, they rely on finding patters and relationships that are present on the data without having been previously exposed to it. Nevertheless, their output cannot predict any type of new labels in the same way a supervised method could and the interpretation of the results is left for a human to interpret. Unsupervised machine learning algorithms are mainly used for data mining, clustering and classification with common examples for this category including k-means clustering and Principal Component Analysis (PCA).

As we first mentioned in chapter 1, that due to the increasing complexity and size of the data-sets generated by surveys in the recent years, there is a need for faster and more scalable methods to handle it. In our case, dealing with spectroscopic data means studying vast amounts of objects (stars), each with a large amount of information (one flux value per sampled wavelength unit per each star). For this reason, it would not be possible to manually compare all of the flux values with each other in order to find and extract the patterns that distinguish SB2 spectra. Regarding the two main categories of machine learning, in this work we will be using a combination of two unsupervised methods. The reason for this is that when working with real spectroscopic

surveys such as GALAH, the nature of the spectrum of each star is not known a priori (whether it was measured from a multiple system or single star), in which case one would not have a representative set of training data for a supervised method to built a proper predictive model. The usage of an unsupervised machine learning algorithm would further benefits our analysis by reducing the amount of biases that could be introduced during binary star detection.

3.1 High dimensional data and dimensionality reduction algorithms

Spectroscopic data is high-dimensional in nature. Each spectrum can be considered as a data point, which in turn is a vector of dimensionality d (one dimension per each wavelength-flux pair). Each flux value is, in turn, measured at the corresponding point on a wavelength grid given by

$$\frac{\lambda_{top} - \lambda_{bottom}}{\varsigma} \tag{3.1}$$

where λ_{top} is the higher limit of the wavelength grid, λ_{bottom} the lower limit and ς is the sampling, which gives the amount of flux measurements per unit of wavelength. With this, it is possible to define our high-dimensional set of spectra as a matrix of flux values, given by the matrix

$$\boldsymbol{\mathcal{F}} = \begin{bmatrix} f_{11} & f_{12} & \cdots & f_{1d} \\ f_{21} & f_{22} & \cdots & f_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ f_{n1} & f_{n2} & \cdots & f_{nd} \end{bmatrix} \tag{3.2}$$

that contains n vectors of d dimensions, $\boldsymbol{f}_n = \{\boldsymbol{f}_1, \boldsymbol{f}_2, ..., \boldsymbol{f}_d\}$. Because in reality both the amount of data vectors n and dimensions d are going to be very large, with $n >> d$ and the total number of numerical values $N = nd$, it would not be feasible to examine the whole data-set by hand. However, there is a type of unsupervised machine learning algorithms called dimensionality reduction algorithms, that can help by reducing the amount of dimensions in the data-set to a few ones. These methods are able to simplify high-dimensional data such as the spectral matrix $\boldsymbol{\mathcal{F}}$ while extracting important information from it and presenting it in a human readable manner.

Dimensionality reduction methods are widely used in astronomy and science in general and they can be used as mapping tools for feature extraction (dimensions reduced to $d > 3$) or as visualization tools (for $d \leq 3$). A very important example for this is Principal Component Analysis or PCA, an algorithm capable of creating a linear mapping of the data into a space of lower dimensions. An introduction to PCA can be found in Francis and Wills 1999. However, for the purpose of this work and in order to deal properly with spectroscopic data (which is highly non-linear due to the processes within within stellar atmospheres) we need a reduction method capable of handling non-linear data.

3.2 t-SNE

t-distributed Stochastic Neighbor Embedding or t-SNE is a non-linear dimensionality reduction algorithm developed by Maaten and G. Hinton 2008. It has proven to be one of the leading machine learning choices for visualization and dimensionality reduction of high-dimensional data, widely used in recent years, mostly in the field of biology but it has found its way into astrophysics as well with many successful applications (Traven et al. 2016; Valentini et al. 2017;

Lochner et al. 2016; Kos et al. 2017; Jofré et al. 2017; Anders et al. 2018). For our purposes, t-SNE performs better than other non-linear techniques such as ISOMAP (Tenenbaum et al. 2000), Locally Linear Embedding or LLE (Saul and Roweis 2000) or even its predecessor, Stochastic Neighbor Embedding or SNE (G. E. Hinton and Roweis 2003) as it is easier to optimize and its ability to solve the crowding problem[1]. Furthermore, t-SNE not only successfully alleviates the crowding problem but it also utilizes the whole low-dimensional space for the projection, making its result much easier to understand than those produced by the other alternatives available. The main idea behind t-SNE is to create a projection or a map of all the data in such a way, that it can be inspected and understood by a human. In the map, the similar points are clustered together while the dissimilar points are located further apart. Due to the differences in the spectrum of binary and single stars, it is expected that most of the binaries will be separated from the groups of single stars, making their identification possible.

In the following paragraphs we present an introduction to the main concepts of t-SNE, adapted from Maaten and G. Hinton 2008 and Van Der Maaten 2014, although for a more in depth explanation and theoretical derivation of t-SNE we refer the reader to Linderman and Steinerberger 2019. The goal of t-SNE is to achieve a low-dimensional embedding or projection of a high-dimensional data-set containing N numerical values from n objects in d dimensions. This is done by modeling the similarities between each pair of data-points within the data-set by using two symmetric joint probability distributions: P, that represents the similarities of the high-dimensional data-points $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_n\}$ and Q, that does this for the low-dimensional set $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_n\}$ where $\boldsymbol{y}_i \in \mathbb{R}^s$ with s commonly being 2 or 3. The distribution P is given by:

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2} \tag{3.3}$$

where the individual conditional probabilities are defined as:

$$p_{j|i} = \frac{\exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2}{2\sigma_i^2}\right)}{\sum_{k \neq i} \exp\left(\frac{-\|\boldsymbol{x}_i - \boldsymbol{x}_k\|^2}{2\sigma_i^2}\right)} \tag{3.4}$$

with $p_{i|i} = p_{j|j} = 0$ and σ_i is the variance of a Gaussian probability distribution located at $\boldsymbol{x}_i$, which is accounts for the density of data-points around $\boldsymbol{x}_i$. If assumed that the close neighbours of $\boldsymbol{x}_i$ are defined proportionally to the Gaussian probability density kernel around $\boldsymbol{x}_i$, then the conditional probability $p_{j|i}$ defined in equation 3.4 is the similarity between points $\boldsymbol{x}_j$ and $\boldsymbol{x}_i$. Similarly, Q is defined for the low-dimensional space as:

$$q_{ij} = \frac{\left(1 + \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2\right)^{-1}}{\sum_{k \neq l} \left(1 + \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2\right)^{-1}} \tag{3.5}$$

with $q_{ii} = q_{jj} = 0$. Unlike in P, the similarities in Q between points $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ are determined using a normalized Student's t-distribution with a single degree of freedom. The heavier tails of the Student's t-distribution compared to a Gaussian allow for a more accurate modeling of spatial distances, setting more space between data-points that are somewhat dissimilar and leaving more space available for the local structure to be modeled accurately (the local structure corresponds to the smallest pairwise distances).

[1]Maaten and G. Hinton 2008 provides some insight into the crowding problem: "[...] the area of the low-dimensional map that is available to accommodate moderately distant data-points will not be nearly large enough as compared to the area available to accommodate nearby data-points."

To obtain the positions of the points in the low-dimensional space $\boldsymbol{\mathcal{Y}}$ in which Q optimally reflects the behavior from P, t-SNE minimizes the difference between these two distributions by minimizing the cost function given by the Kullback-Leibler divergence:

$$C = KL\left(P \,||\, Q\right) = \sum_{i \neq j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \tag{3.6}$$

where C is the cost function to be minimized. The Kullback-Leibler divergence or KLD is a quantity that measures the difference between two probability distributions (Kullback and Leibler 1951). The KLD of two functions is equal to 0 only if both distributions are exactly the same. The minimization of the cost function is accomplished using iteratively a gradient descent to find the minimum (which corresponds to the most optimal embedding), with the gradient defined as:

$$\frac{\partial C}{\partial \boldsymbol{y}_i} = 4 \sum_{i \neq j} \left(p_{ij} - q_{ij}\right) q_{ij} Z \left(\boldsymbol{y}_i - \boldsymbol{y}_j\right) \tag{3.7}$$

where $Z = \sum_{k \neq l} \left(1 + \|\boldsymbol{y}_k - \boldsymbol{y}_l\|^2\right)^{-1}$ is the normalization term from equation 3.5. The gradient descent is initialised by randomly placing data-points in space $\boldsymbol{\mathcal{Y}}$. As a consequence, two runs of t-SNE can have different outputs even if the input parameters are the same. Above equation 3.6 shows that the computational complexity of the algorithm is $\mathcal{O}\left(N^2\right)$, as the computation of similarities between data-points requires $N\left(N-1\right)$ numerical operations.

3.2.1 t-SNE: an example

The optimal result of a t-SNE analysis is given in the form of a visual representation or map from the high-dimensional data in a low-dimensional space (most frequently 2-D) displaying data-points grouped in clusters (also named islands of data-points). As an example, we show the results from t-SNE applied to MNIST data-base (LeCun et al. 2010), a widely used benchmarking data-set used in machine learning. MNIST contains images of handwritten digits in a bitmap (image), composed by a matrix of 28x28 pixels where the position of each pixel in the bitmap matrix represents a dimension containing one greyscale value, so that the dimensionality of the MNIST dataset is 784. An extract of the raw data is shown in figure 3.1, where we can see that each data-point (bitmap image) represents a number from 0 to 9 and that those images representing the same digit show slight differences between each other.

Applying t-SNE on the MNIST data-set results in a projection that can be seen in figure 3.2, where similar numbers are clustered together. Even though there are some outliers, such as poorly written 4 that resemble the number 9, or some samples of the number 5 that appear very similar to number 6 (and thus are located relatively close on the map), the algorithm did a great job at separating the numbers and grouping them on clusters that are easily recognizable visually. Furthermore, a closer inspection reveals local structure within each number's cluster, showing the smallest variations even between numbers that were properly grouped, such as a gradient from thinner to fatter or different writing orientations. This ability to show both large and small structure at the same time without excessive cluttering is one of the strengths of t-SNE and is what sets this algorithm apart from the other non-linear dimensionality reduction algorithms.

Figure 3.1: Raw data from the MNIST database, from C.-L. Liu et al. 2003. Each of the shown digits corresponds to a 28x28 pixels bitmap (image) containing grayscale values.

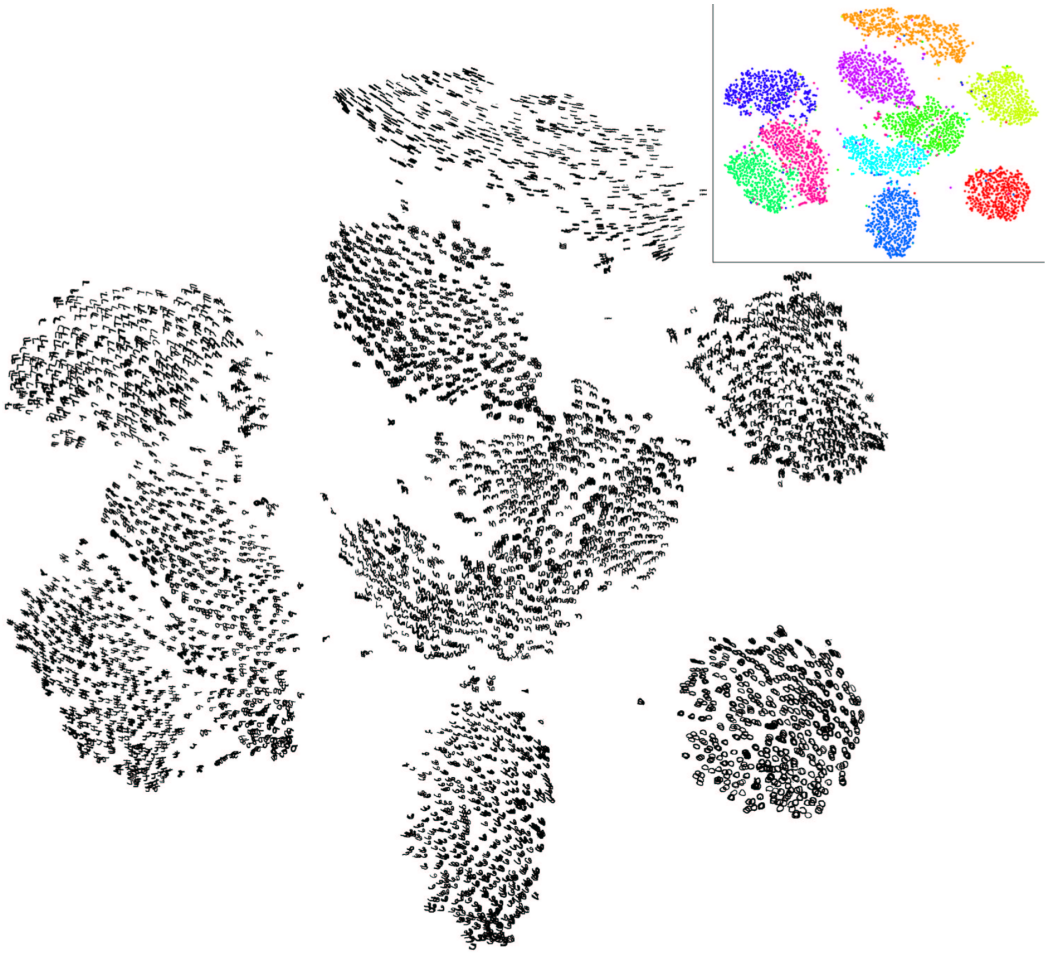

Figure 3.2: t-SNE applied on the MNIST data-set, from Maaten and G. Hinton 2008. Here are shown only 6000 digits from the 60.000 used in the t-SNE analysis for easier visualization.

It is important to remind that t-SNE does not preserve the densities nor exact distances between the data-points of the original high-dimensional space in its low-dimensional representaton, but rather groups data-points according to similarities. The separation between data-points in the projection map cannot therefore be directly related to some metric (e.g. Euclidean). This is a consequence of the highly non-linear nature of the algorithm.

3.2.2 Perplexity

From all the hyperparameters that go into t-SNE, perplexity is the one that has the largest impact on the end result. It essentially defines the scale of the projection, setting whether t-SNE is more sensitive to local or global structure, which consequently determines the morphology of the embedding (Cao and Wang 2017). The set value of perplexity is used by t-SNE to compute the standard deviation σ_i of the Gaussian distribution which effectively sets the number of neighbours for each point $\boldsymbol{x}_i$ in the high-dimensional space $\boldsymbol{\mathcal{X}}$, as given in equation 3.4. The value of σ_i is computed for each point $\boldsymbol{x}_i$ according to:

$$Perp\left(\boldsymbol{x}_i\right) = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} \tag{3.8}$$

which must then be equal to the pre-specified input value $Perp$ for perplexity, so that $Perp\left(\boldsymbol{x}_i\right) = Perp$ for every high-dimensional point $\boldsymbol{x}_i$. Perplexity must be fixed by hand before starting the t-SNE computations, with recommended values between 5 and 50 (Van Der Maaten 2014). For an interactive visualization of the behavior of t-SNE under different choices of the perplexity, we refer the reader to Wattenberg and Johnson 2016.

3.3 Clustering algorithms

Clustering algorithms are a group of machine learning methods commonly used for data mining. These algorithms are designed to identify and retrieve clusters of data-points from an input set by defining the boundaries between them. Furthermore, clustering algorithms are able to mark points that do not belong to any cluster (outliers) as noise. For a comprehensive review on the different types of clustering algorithms and an overview of their functionality we refer the reader to D. Xu and Tian 2015. These methods are used in the field of data mining to extract information and explore data-sets by detecting points close to each other and labeling those groups as clusters. This can be used in combination with the output of t-SNE, which groups similar data in clusters as we saw previously in figure 3.2 and thus a clustering algorithm could automatically identify and separate them from the rest of the data. This will allow for an automatic detection and separation of the major binary-star clusters from their single-star counterparts in the t-SNE projection.

3.3.1 DBSCAN

Density-Based Spatial Clustering of Applications with Noise or DBSCAN (Ester et al. 1996), is a very well-known and widely used unsupervised clustering algorithm. DBSCAN stands out from other clustering methods by being able to identify clusters with different morphologies based on their density. Further advantages from DBSCAN include: its input of only two parameters without the need of extensive knowledge of the data-set, its scalability, its ability to recognize noise and that no guess or knowledge about the number of clusters is needed, which removes the necessity of a manual examination of the data-set.

The two input parameters of DBSCAN are: *minPts*, which gives the minimum amount of data-points (hereafter points) per retrieved cluster, and ϵ, which defined the neighborhood of a given point and it is a measure of the maximum distance between two points so that they are

considered neighbors. DBSCAN defines three kind of points using a combination of *minPts* and ϵ, which is named mode. The following definitions of the point properties and categories they are classified in by the algorithm were adapted from Ester et al. 1996. For a given mode and a randomly chosen point p from the data-set D, one can define the following situations:

- p is a core point if there are $N_{eps}(p) \geq minPts$ within its neighborhood given by the specified epsilon value, which can be expressed as:

$$N_{eps}(p) = \{q \in D \,|\, \text{dist}(p, q) \leq \epsilon\} \tag{3.9}$$

- A point q is said to be directly reachable from the core point p if $q \in N_{eps}(p)$. In other words, q is located within a distance ϵ of p.

- A point q_n is density reachable from the core point $p = q_1$, if there is a succession of points $q_1, q_2, \ldots, q_n$ with the condition that q_{i+1}, $i \leq n$ is directly reachable from q_i. Note that due to the previous definition of direct reachability, every point of the succession must be a core point with the possible exception of q_n itself. If q_n is not a core point, it is considered a border point. The fact that q_n is density reachable from q_1 does not imply that the reverse case is true. However, two points q_1 and q_2 are said to be density connected if there is a point x such that both q_1 and q_2 are both directly reachable from x and in this case, both q_1 and q_2 are mutually density reachable.

- A cluster C is defined as a non-empty subset D and is formed by all points that are either density connected and/or density reachable from a given point within C.

- If the point u is neither directly or density reachable from a core point p it is marked as noise and it therefore does not belong to any of the detected clusters C_i.

A visual example of this can be seen in figure 3.3. The parameters used in the figure are $minPts = 4$ and ϵ is the radius of the circumference drawn around each point. In figure 3.3 there is one cluster formed by six red core points and by the two yellow border points. There is also one blue point in the upper part of the figure which is marked as noise by the algorithm. The red points are labeled by DBSCAN as core points because the algorithm was able to find Because it is possible to find at least 4 points within the circumference given by ϵ (this includes the point itself). These form part of the cluster as they are density connected. The yellow points belong to the detected cluster as well as they are density reachable from a core point. The blue point, as it is not density reachable from a cluster core point nor it is a core point itself (no neighbors within its epsilon neighborhood) is marked as noise.

DBSCAN begins by sampling a random point p from the data-set D and then tries to recover all points that are density reachable from p. This procedure can have two different outcomes: either p is a core point and the search for the density reachable points forms a cluster, or p is a border point and DBSCAN moves on to examine the next point in D. From this, it is possible to infer that the right selection of the mode is remarkably important. Indeed, for example if ϵ is selected to be too small, then only the densest regions will be clustered, on the contrary, on the other hand if ϵ is too large, then most of the data-points will be part of the same cluster. Besides the dependency of the end result on the mode selection, there are two major disadvantages of DBSCAN: first, the proper determination of both ϵ and *minPts* is not straightforward, and second, it has trouble with data-sets of varying densities as a single DBSCAN mode is only able to effectively target a single density value (or narrow range of densities). In the next chapter we will explore the solutions we designed to overcome the drawbacks presented by this algorithm. Regarding computation times, DBSCAN has a complexity of $\mathcal{O}(N^2)$, as it has to go once over each point in the data-set and calculate distances to all other points.

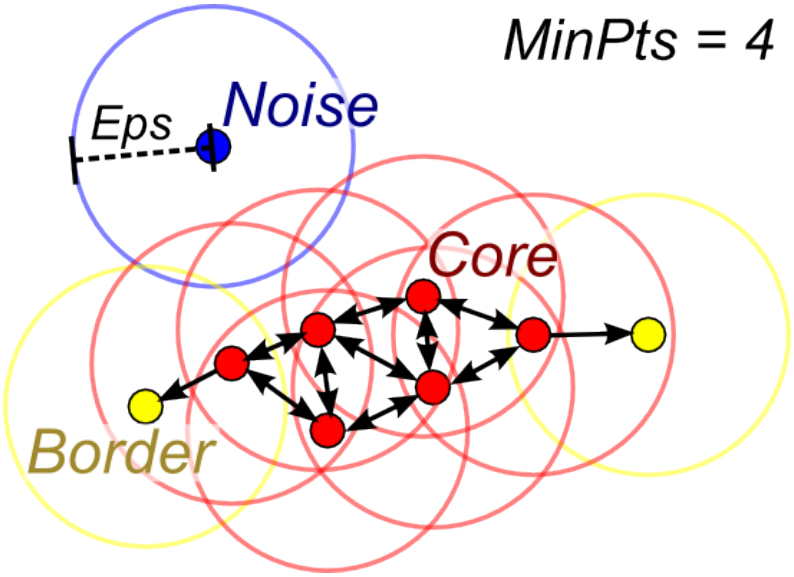

Figure 3.3: Visual example of DBSCAN. Image from Wikipedia.

In figure 3.4 we can see an example of a possible input and output for a DBSCAN analysis. In the left panel we see the raw input data-set for DBSCAN. Numerically, the input for DBSCAN will, in this case, be the location of each points as a pair of x and y coordinates. From this information, DBSCAN is able to detect the clusters present on the left panel and separate them accordingly. The result from the analysis is shown in the right panel, where each detected cluster is marked using a different color. As we can see, all of the clusters were properly recognized regardless of their morphology (shape) and the outlying points were marked as noise.

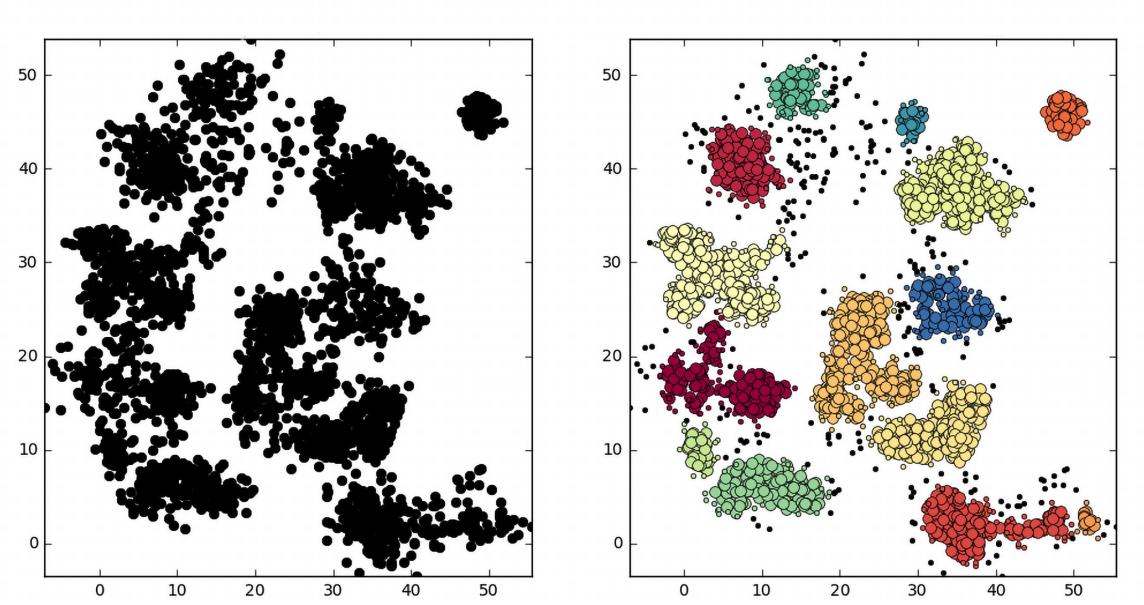

Figure 3.4: Non-uniform raw data can be seen on the left panel, clusters found by DBSCAN are shown on the right. Image by Christ Wersnt.

Chapter 4

Method

In the previous chapter we explained that the combined usage of t-SNE and DBSCAN can be useful to automatically separate binary and single stars (among other types). This has already been tried and proven successful on GALAH data in Traven et al. 2016 and more recently in Traven et al. 2020. In this work we will be, however, focusing on optimizing the combined use of these two machine learning algorithms so that together they can efficiently detect SB2 systems from a collection of spectra. This combination of methods constructs a two-step analysis method, where first t-SNE reduces the complexity of the data-set to be analyzed allowing for a manual examination if needed, and then DBSCAN automatically detects and recovers the different clusters obtained with the t-SNE projections, labeling them and allowing for a more exhaustive classification of the different categories that might be present in the data-set.

We perform a parameter space search to identify the parameter combinations that work best regarding the recovery efficiency of binaries, which in turn allows for a thorough analysis of the effect these parameters have on the machine learning algorithms and how they affect the recovery. We can divide the parameters that will influence the final recovery result in two different categories: variable and implicit.

Implicit parameters

Implicit parameters are those whose alteration would require a new synthesis procedure and therefore are fixed during the analysis; they are implicit in each stars' spectrum. These are:

- Stellar parameters used in the synthesis such as T_{eff}, $log\, g$ and [Fe/H].
- Parameters that define the binary systems: q, M_A, Δv_{rad}, L_B/L_A.
- The wavelength region of the spectral synthesis (450 to 900 nm in this work)
- Resolving power and sampling (nm/px) of the resulting synthetic spectroscopic survey.

For example, a new subset of stellar data from GALAH would be needed to change the parameters of the used stars, different q, M_A (use a different type of IMF) and Δv_{rad} distributions would be needed to change the properties of the resulting binaries or even new atomic line-list to shift the wavelength range of the synthesized spectra.

Variable parameters

Variable parameters, on the contrary, are those that can be varied for the current synthetic spectroscopic data and allow for a comprehensive study during the parameter space exploration for a given data-set. They are:

- The selected spectral range, which is a small portion selected from the entire spectral region the spectra was synthesized in.
- Signal-to-noise value.
- Machine learning parameters: perplexity for t-SNE and ϵ and $minPts$ for DBSCAN.

4.1 Optimization of machine learning algorithms for SB2 detection

The optimization procedure we designed consists of three distinct phases: the pre-processing of the data, the analysis using t-SNE and the automatic cluster recovery using DBSCAN. To maintain a certain hierarchy, the parameters will be altered in a cyclic fashion according to the diagram shown in figure 4.1, where for a given spectral range, all possible combinations of SNR, perplexity and DBSCAN mode will be tested. We perform this recursively for all possible combinations of variable parameters within a set of pre-defined ranges. This will allow not only to maintain a certain order during the analysis but it also has a positive effect on the computation efficiency.

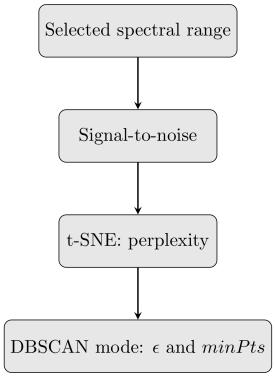

Figure 4.1: Hierarchy used in the parameter space exploration.

The ranges of parameters that for each of the variable parameters shown in figure 4.1 are specified in 4.1. Both the machine learning parameters for DBSCAN and t-SNE as well as SNR were chosen to encompass the range of most common values found in the literature. The spectral range intervals were chosen to be 25 nm in width for two reasons: they are almost as large as the average width of a single GALAH band and with this division, we were able to separate important spectral ranges such as that containing H-α.

Parameter	**Range**
Spectral range	25 nm intervals between 450 and 900 nm
Perplexity	5, 15, 30 and 100
SNR	10, 25, 50, 100 and 500
ϵ	10 equally spaced values between 0.1 and 0.75
$minPts$	10 equally spaced values between 25 and 125

Table 4.1: Ranges of the studied variable parameters.

4.1.1 Pre-processing

Before the analysis, we perform a pre-processing of the synthetic data. It includes: extracting the corresponding flux data for the selected wavelength ranges from the synthetic spectra, adding

noise to it according to an input SNR value and cutting the extrema of the spectra. The last step is crucial as after the shift of the secondary spectra due to the Δv_{rad}, the new wavelength grid the spectra were interpolated onto was too wide. This caused unwanted features to appear on the resulting binary spectrum, which could be interpreted by t-SNE as very distinct features and therefore throw off the validity of the resulting t-SNE projection.

4.1.2 t-SNE projection

In this phase we apply t-SNE to the pre-processed data for a given perplexity value. Because the standard version of t-SNE is computationally very expensive for such a large data-set as ours, we use a multi-core version of t-SNE by G. C. Linderman et al. 2019 (based on an earlier accelerated version by Van Der Maaten 2014), conveniently wrapped in a Python package called FIt-SNE (short for fast Fourier inverse transform Interpolation based t-SNE). The usage of FIt-SNE sped up the computations by a factor of 10 compared to other implementations, which allowed us to increase the range of the studied parameters.

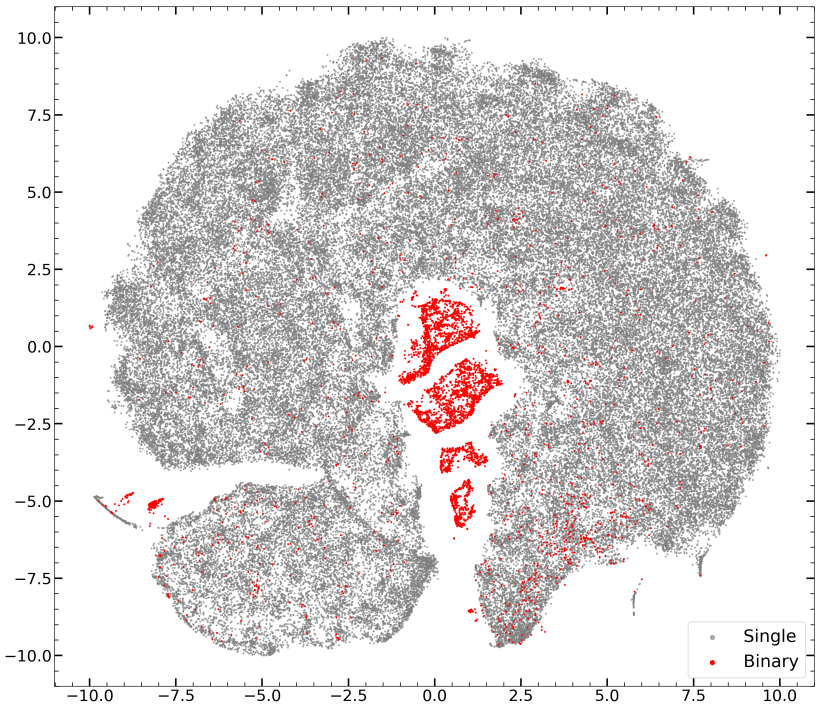

Figure 4.2: Example of a t-SNE result from spectra in the wavelength range between 800 and 825 nm for perplexity 30 and SNR 100. The binaries are colored in red, while the singles are shown in grey.

Figure 4.2 shows one of the maps created by t-SNE for spectroscopic data between the wavelength range of 800 and 825 nm, where we can see the clustered features we mentioned in the previous chapter. Data-points representing binary star spectra are mostly grouped in four distinct clusters with some outliers mixed with the single stars, probably due to their binary-defining features not being strong enough to allow for proper isolation. As we advanced in section 3.2, perplexity plays a major role in the final morphology of the projection, making it denser or sparser for lower and higher values, respectively. This high dependency plays a major role in the final recovery of binary clusters from a given t-SNE embedding and its proper handling will prove crucial for this.

4.1.3 Interpreting the t-SNE results

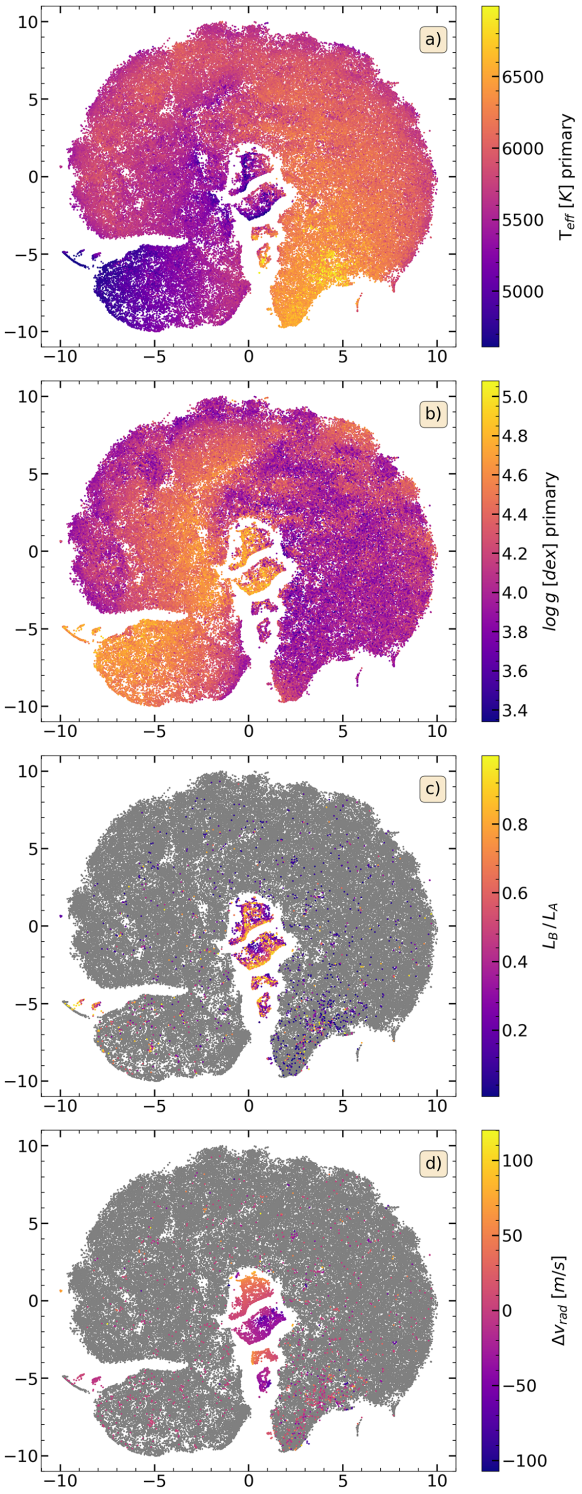

Figure 4.3: t-SNE projection from figure 4.2 with different color codings. Projection a) is colored according to the effective temperature for the singles and the effective temperature of the primary for the binary systems, b) according to the surface gravity of the singles and that of the primary component of the binaries, c) and d) are colored with respect to the luminosity ratio and radial velocity difference for each one of the pairs, respectively.

In figure 4.3 we show the same t-SNE projection as in figure 4.2, where the points representing the synthetic spectra analyzed by t-SNE have been color coded according to diverse stellar parameters that shape the spectra they represent. In both plots a) and b), we see the projection color coded according to the $T_{\rm eff}$ and $log\, g$ of the singles and primaries (for the binary systems) respectively.

There is a clear gradient from higher $T_{\rm eff}$ on the right towards cooler values on the left along the arch of the main single star islands. A gradient is seen as well on the projection color coded according to $log\, g$ but from lower values on the right to larger values on the left along the main single island. The four islands of points containing binary spectra in the middle of the projection follow a similar pattern, as the two upper clusters show lower values of $T_{\rm eff}$ and larger values of $log\, g$ whereas the lower two islands show the inverse behavior, higher temperature and lower surface gravity figures. Furthermore, binary clusters seem to show an internal substructure on plot c) with respect to the luminosity ratio, which appears as a gradient from top to bottom in each one of the clusters. On the last plot d), we see that binary clusters are presented from top to bottom with radial velocity differences of alternating signs. This in the spectra is shown as a positive and negative Doppler shifts of the secondary star. As for the binaries that are not recognized by t-SNE as such and are therefore not separated in an independent clusters, they are mainly located on the lower right corner of the main single star island. We can see from plots c) and d) that they show both low luminosity ratio and radial velocity difference values, which means that the spectra found there present a strong blending of their spectral lines and a gradual disappearance of the secondary lines. However, they still present differences from the single spectra large enough for t-SNE to place those stars apart from the bulk of single stars, but those differences are not large enough for t-SNE to consider them a special class of data-points and are therefore not separated in a single cluster. The same behavior is seen in Traven et al. 2020 as well.

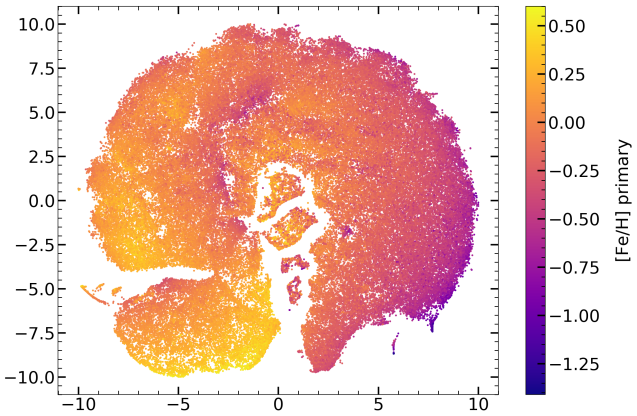

Figure 4.4: t-SNE projection from figure 4.2 color coded according to the metallicity of the primary star.

In figure 4.4 we can see the same t-SNE projection from figures 4.2 and 4.3 color coded according to the metallicity of the primary component. In this case, we can see that although there is also a gradient across the arch of the main island of single stars, this is not the case for the binary islands, whose metallicities seem to vary only from island to islands, where the two upper binary islands appear to be slightly metal-richer than the two bottom ones. Therefore, due to this and the small range of considered values, metallicity appears to have a less noticeable effect on the final identification of SB2 in our study. With all of this, we can see how the groups of data-points presented on the t-SNE projections in both figures 4.2, 4.3 and 4.4 are not a consequence of only one property but a combination of several, all of which seem to have been

considered with varying degrees of importance by the algorithm regarding the positioning of that given data-point, either with respect to the local or large scale structure. This shows that t-SNE is capable of identifying binarity with great success, even if some of the essential feature for its identification are suppressed or lessened up to some degree.

4.1.4 DBSCAN mode selection and recovery ratio

With the t-SNE projection of spectroscopic data for a given SNR and perplexity values and within a given wavelength range, we now can apply DBSCAN to the embedding for automatic cluster recovery. To do so we need a mode capable of targeting the specific density of the projection to be analyzed. However, in subsection 3.3.1 we argued that one of the downsides of DBSCAN is its intrinsic difficulty when selecting the appropriate mode. To overcome this issue we perform an exploration of the DBSCAN parameter space based on the recovery ratio of binary stars from the t-SNE clusters. For this purpose, we designed an algorithm that analyzes the output of DBSCAN to find the cluster each data-point (star) was assigned to and then it compares this to the real nature of the data-point, which is either binary or single. With this, the algorithm inspects all data-points that belong to each one of the detected clusters and it computes the ratio of binary stars to all stars in that given cluster as:

$$\Delta_B = \frac{N_B}{N_T} \tag{4.1}$$

where Δ_B is the ratio, N_B is the amount of binaries and $N_T = N_S + N_B$ the total amount of stars (both single and binary) in the cluster. We can use equation 4.1 as a filtering for the cluster identification, where a pre-defined value of Δ_B serves as a threshold to define a binary cluster. If the computed value of Δ_B for a given cluster is larger or equal than the specified, then the analyzed cluster and all of the stars within are labeled as binaries (even the single stars that might be there). On the contrary, if the calculated value is lower than the pre-defined Δ_B, then all of the stars within the examined cluster are marked as single stars. Doing so for each one of the detected clusters allows us to quantify the quality of the investigated mode.

In this work we perform the DBSCAN parameter space exploration for a range of 10 values that span between 0.1 and 0.75 for ϵ and 10 values between 25 and 125 for $minPts$, accounting for a total of 100 combinations. The parameter ranges were selected upon manual examination of t-SNE maps with the data-points enclosed between (-10, 10) on each axis of the 2D projection space. For this reason, all of the t-SNE embeddings that DBSCAN will analyze are normalized to fit within that range, ensuring that the parameter space exploration is as targeted as possible without the need of human interaction.

The results of the optimal mode search can be seen on figure 4.5, where all 100 combinations of explored ϵ and $minPts$ are shown as a heat-map, color-coded according to the recovery ratio each mode yielded. The results from this type of figure are interesting because they encompass information from both the t-SNE projection and the DBSCAN performance in the recovery. If the t-SNE projection was not well suited for binary detection (no defined binary islands), the sharp recovery seen on figure 4.5 would appear smeared out over a large portion of the shown parameter space. On the contrary, and as we can see on figure 4.5, the heat-map presents several peaks of high recovery located on a line with sharp edges, which correspond to other modes that could be suitable for the final analysis. This is a consequence of the t-SNE map shown in figure 4.2, which presents four well-defined islands corresponding to the binaries and which are properly recovered by the selected mode. While the parameter space search chooses the mode with the absolute best recovery efficiency, other $minPts$ and ϵ combinations would also yield optimal binary cluster recovery, however, the way the cluster detection perform varies greatly depending on the chosen parameters.

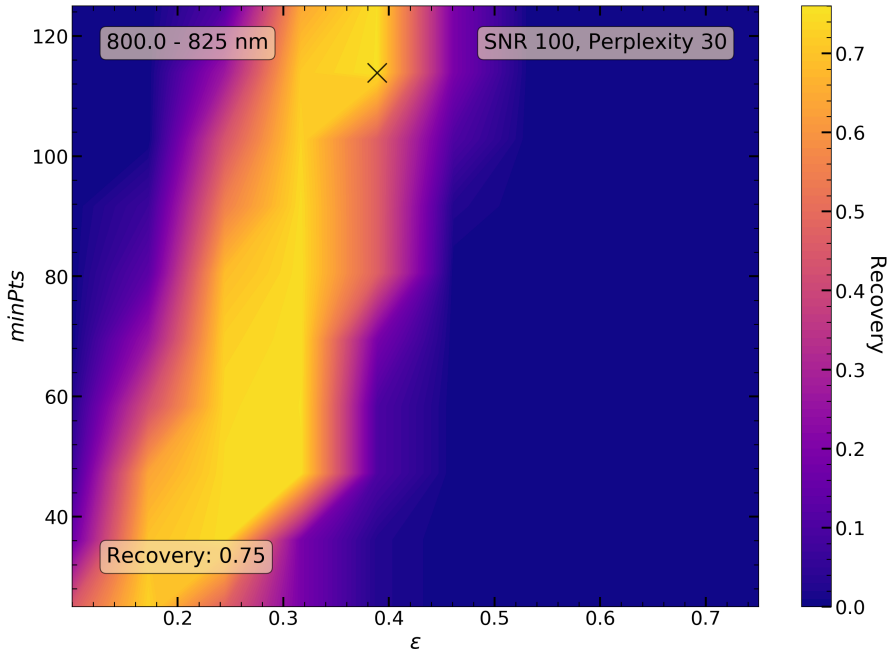

Figure 4.5: Heat map corresponding to the parameter space exploration of the possible DBSCAN modes for the t-SNE projection shown in figure 4.2. The selected mode is marked with a cross at $\epsilon = 0.39$ and $minPts = 113.9$, with a recovery of 0.75.

Furthermore, the selection of the ranges of the DBSCAN parameters is not arbitrary, as for low values of $minPts$ and ϵ, almost every small group of stars would form a cluster and therefore the number of technically recovered binary stars would be much higher. However, this would be a misleading result, as we strive for a practically useful number and size of identified DBSCAN clusters of data-points. For this reason, we tried to avoid the lowest values to ensure that even if the highest recovery occurs at the lowest value of the interval, it is still large enough so that the DBSCAN clusters can be efficiently explored by a human in a real data-set where our findings might be applied.

With the mode selected according to the point of largest recovery in figure 4.5, we can analyze the t-SNE projection shown in figure 4.2. The result of this analysis can be seen as two t-SNE embeddings in figure 4.6. The top t-SNE map is color coded according to the clusters that DBSCAN is able to recognize using the input mode ($\epsilon = 0.39$ and $minPts = 113.9$) and as we can see, it manages to properly recognize the islands as clusters with a total recovery of 75%. On the bottom projection from figure 4.6, we show in red the clusters that were labeled as binary from those detected by the selected DBSCAN mode and in blue, those the binaries that were missed. By carefully examining the results from figure 4.6 (bottom) and comparing that t-SNE map to the one shown in figure 4.2, we can see that although it recognizes 5 binary clusters, it still misses the left-most cluster, probably because the density of data-points is lower than that the selected DBSCAN mode is able to detect. This is an example of the second disadvantage for DBSCAN we mentioned in chapter 3, as every DBSCAN mode targets a single density level and for a recovery of all the binary clusters several modes targeting the different levels of density present in the analyzed t-SNE projection would be needed.

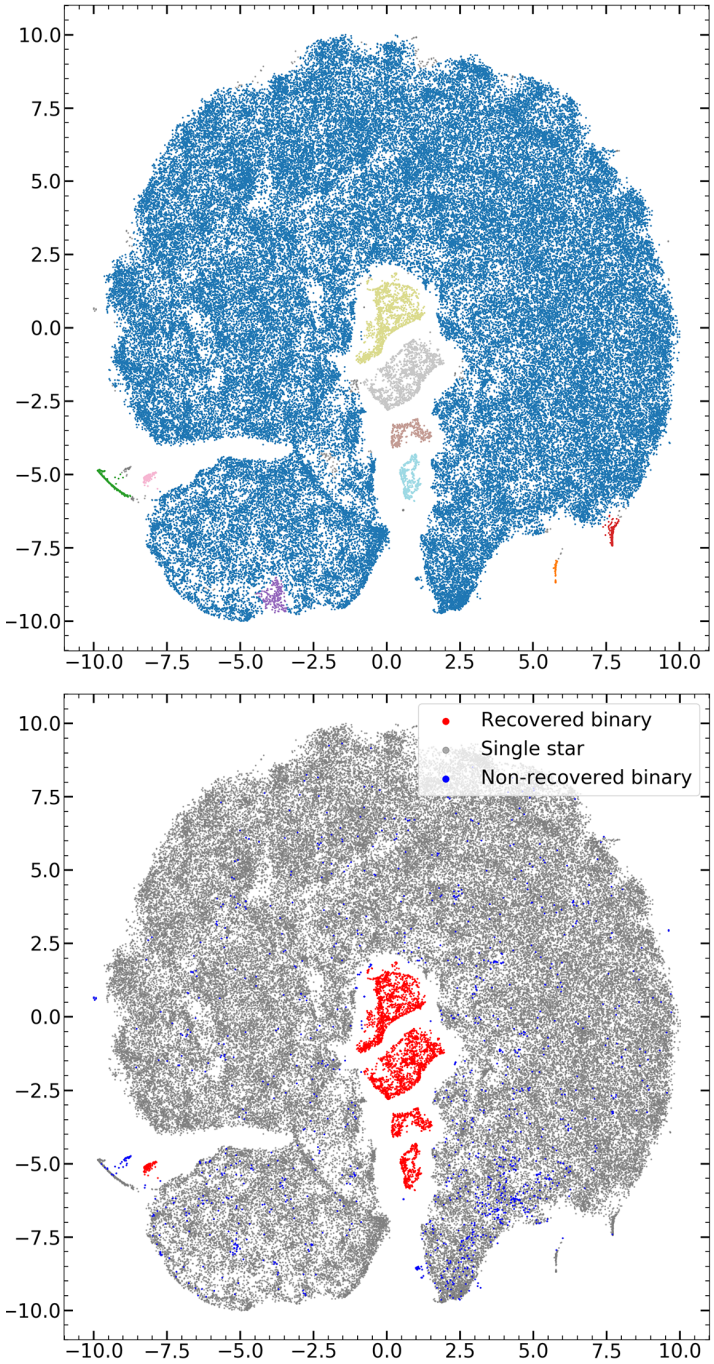

Figure 4.6: Example of the DBSCAN clustering (top) with each detected cluster marked by a different color, and the recovered binary stars using equation 4.1 (bottom).

Chapter 5

Results

In the following we present the results from the described method when applied to the previously synthesized spectroscopic survey composed of 99986 synthetic single[1] and 5000 synthetic binary stars, represented by spectra spanning over a wavelength range between 450 and 900 nm. We study every possible parameter combination within the ranges as given in table 4.1, which accounts for a total of 360 different configurations (assuming the optimal DBSCAN mode has already been selected). Furthermore, we set the threshold for the ratio of binaries compared to all stars in a DBSCAN cluster to be 0.9, as given by equation 4.1.

For some of the results we will be dividing the analyzed spectral regions in two groups groups, blueward and redward of the region containing H-α (650 - 675 nm), as it conveniently lies in the middle of the wavelength range of our synthetic spectra and corresponds to a relatively important part of the spectrum due to the presence of the H-α line.

5.1 Diagnostics

Before we present an in-depth analysis of the different trends and behaviors observed on the data from the parameter space exploration, we explore the results corresponding to a default or baseline models, with parameters that have been previously used in the literature and that we deem as standard.

5.1.1 Baseline model: parameters

As a first approach to analyze the large amount of results from all of the parameter combinations, we define a baseline model based on typical values found in the literature: 30 for perplexity (Maaten and G. Hinton 2008; Van Der Maaten 2014) and 100 for the SNR as it represents the average value intended for the GALAH survey (Buder et al. 2018).

5.1.2 Baseline mode: results

The analysis of the baseline model in general terms shows an average recovery of ∼70%, which translates to an SB2 fraction of ∼3.5%. The results are presented in figure 5.1 with their t-SNE projections and the binaries detected with the automatic recovery using DBSCAN and in figure 5.2 as well, which shows the heat-maps corresponding to the mode selection of DBSCAN as well as the recovery each mode yielded for each one of the studied parameter combinations. A close inspection of both figures reveals two major trends:

- We can see clear differences in the amount of local structure from the t-SNE plots in figure 5.1: spectral ranges bluer than the region containing H-α show a larger degree of internal

[1] Originally 100000 stars, but some were discarded due to a defective synthesis

structure, where both binaries and single stars are shown on more fragmented embeddings than those for the regions redder than H-α, whose projection's structure appears to be more uniform for the same value of perplexity due to a lower level of internal complexity. As we will see later in more depth, this is a consequence of the higher abundance of lines in the blue parts of the investigated spectra, which allow for a finer classification with t-SNE.

- Regarding the parameter space of the DBCSCAN modes in figure 5.2, we can see clearly that those spectral regions that achieved a high recovery (> 0.65) present sharper features in the corresponding heat maps than those that recovered fewer binary stars, which present a more diffuse parameter space, with features that are very faint and appear smeared out through the whole parameter space.

  Moreover, when looking at the point of maximum recovery per region for the baseline model (marked with "**x**" on the individual plots from 5.2), we can see that when the recovery was low, DBSCAN had trouble finding an appropriate mode for that given t-SNE projection and tried to find as many small clusters as possible in order to achieve the best possible overall recovery fraction.

With the exploration of the baseline model we wanted to show that a simple exploration with default parameters can already lead to trends representative for the type of data we analyze in this work.

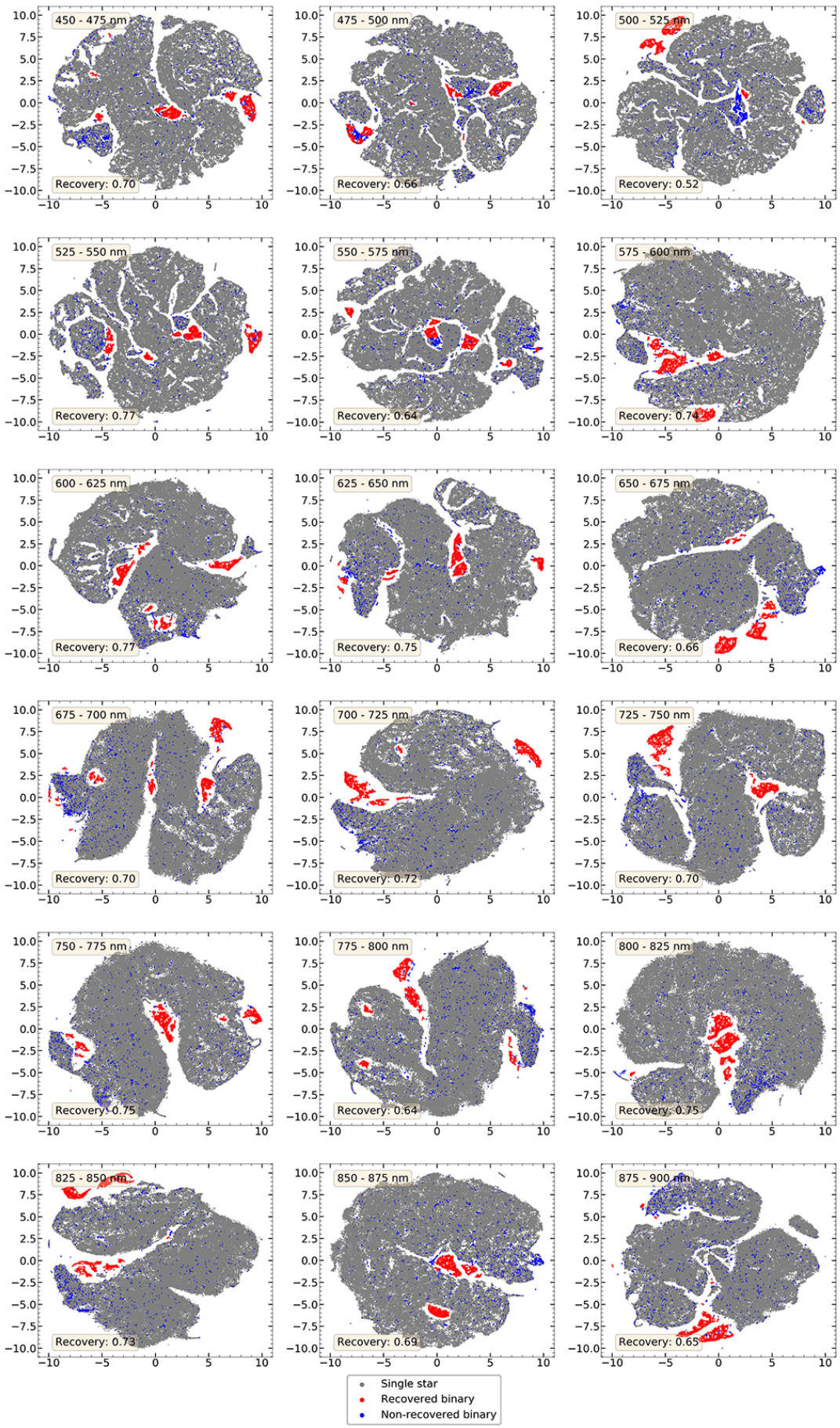

Figure 5.1: t-SNE map for every spectral region of the baseline model.

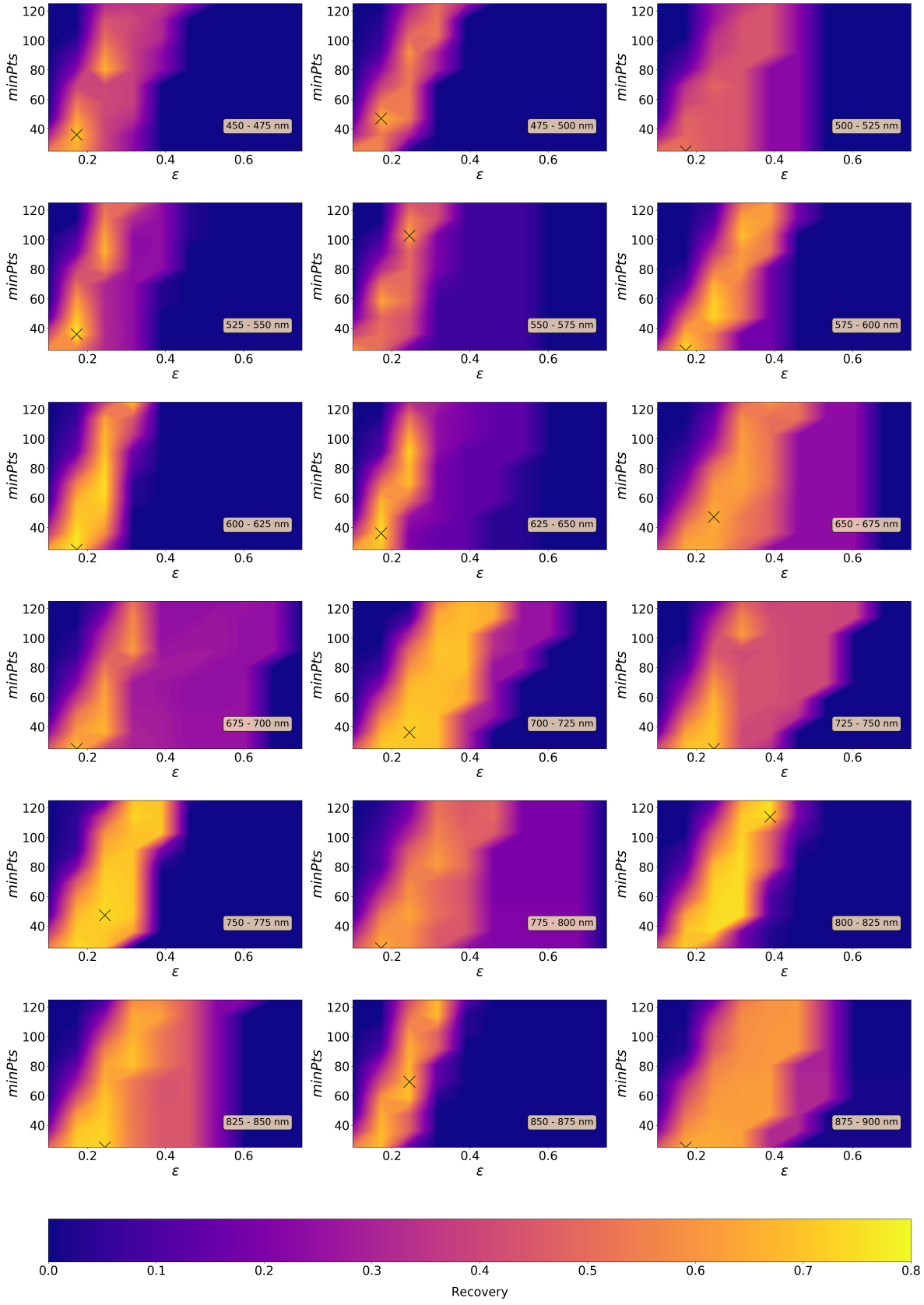

Figure 5.2: Results from the DBSCAN parameter space exploration from the data corresponding to the baseline model, with a perplexity of 30 and SNR 100, color coded by recovery fraction of SB2. The black cross on each plot marks the point of highest recovery and therefore the chosen method (in some the marker is barely visible as it is located on the lowest *minPts* value, on the very same x-axis)

5.2 General performance

So far we have presented the results from the baseline model with the default parameters from the literature and show that it is already possible to extract conclusions from it. In the following points, we will explore the combined results from all of the 360 parameter combinations possible with the parameter ranges we defined in table 4.1. We will first show the general results using simple statistics with focus on the general performance for each different wavelength region studied, which will give a better overview of the performance differences before diving into the effects of each of the studied parameters.

In table 5.1 we show the main numerical results for each of the investigated spectral ranges. All of values in the table are extracted for a given wavelength region across all perplexity and SNR variations. For a clearer visualization, the results from 5.1 are shown in figure 5.3, where the red line represents the average recovery per spectral range regardless of perplexity and SNR, and the upper and lower black lines correspond to the maximum and minimum recovery value achieved for that wavelength region. By carefully studying table 5.1 and figure 5.3 we can observe a gradient that goes from the short to the long wavelengths that shows an increasing difference between the extremal values of the recovery. We can attribute this to the decreasing complexity gradient we observed already for the baseline model in the corresponding t-SNE projections in figure 5.1. From the global results, we can extract two further observations:

- There appears to be an absolute maximum value of recovery (between 0.75 and 0.8) that any spectral region can achieve within the parameter ranges defined in table 4.1. This can be seen as the approximately flat line in figure 5.3.
- In terms of absolute numbers, the spectral region containing the H-α line presents the lowest maximum recovery, albeit a mean recovery slightly larger than the lowest achieved average value (0.471 in adjacent region between 675 and 700 nm).
- The maximum and minimum values in most spectral ranges are not symmetric with respect to their mean recovery. That is, both extremal values are not equidistant from the red line, with a tendency for the average recovery to be closer to the maximum value than to the minimum.

Spectral Range	**Recovery max.**	**Recovery min.**	**Max. difference**	**Mean recovery**
450 - 475 nm	0.7252	0.4622	0.263	0.639
475 - 500 nm	0.7264	0.4628	0.264	0.595
500 - 525 nm	0.7580	0.4574	0.301	0.640
525 - 550 nm	0.7794	0.4670	0.312	0.668
550 - 575 nm	0.7596	0.4250	0.335	0.623
575 - 600 nm	0.7464	0.3210	0.425	0.600
600 - 625 nm	0.7724	0.3504	0.422	0.604
625 - 650 nm	0.7548	0.3586	0.396	0.608
650 - 675 nm	0.6942	0.2392	0.455	0.508
675 - 700 nm	0.7378	0.0132	0.725	0.471
700 - 725 nm	0.7634	0.2344	0.529	0.554
725 - 750 nm	0.7194	0.1774	0.542	0.527
750 - 775 nm	0.7464	0.2034	0.543	0.558
775 - 800 nm	0.7158	0.0644	0.651	0.485
800 - 825 nm	0.7856	0.1740	0.612	0.552
825 - 850 nm	0.7626	0.1876	0.575	0.542
850 - 875 nm	0.7304	0.2294	0.501	0.558
875 - 900 nm	0.7192	0.1952	0.524	0.513

Table 5.1: Table of the most important values from the analysis of all the parameter combinations presented in table 4.1

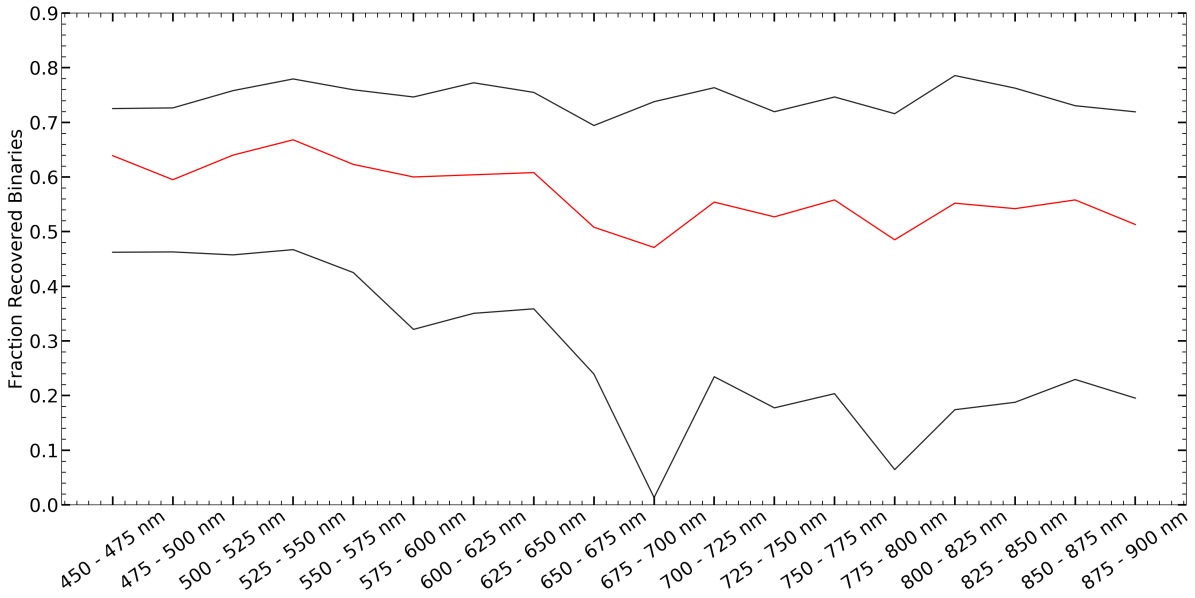

Figure 5.3: The red line represents the average recovery achieved for each spectral region, while the upper and lower black lines correspond to the maximum and minimum recovery numbers achieved.

Moreover, because the values shown in both table 5.1 and figure 5.3 encompass all of the possible parameter combinations within our given ranges, the gradient of increasing difference between largest and smallest recovery value per given spectral range can also be interpreted as measure of the stability under the variation of the aforementioned parameters. The stability against changes is the most important feature that can we extract from the exploration of the general results because it gives the best overview of the performance from each spectral range, regardless of the machine learning parameters and SNR value used.

An example of this using, the results from the baseline model in figure 5.2 can be seen in the heat map corresponding to the region the region 800 - 825 nm. Its DBSCAN parameter space presents the sharpest recovery features (two distinct yellow peaks) and also the highest recovery value of 0.75. Furthermore, we can see that DBSCAN did not have any troubles finding an appropriate mode, as both parameters $minPts$ and ϵ are relatively large, meaning that it was able to recognize the clusters as a whole entity and did not have to rely on finding many small ones. This same region also presents an almost ideal t-SNE projection on figure 5.1, which shows four distinct binary islands surrounded by the single stars and only one fourth of the studied binaries are not recovered either because they are mixed with the single star spectra or due to the DBSCAN mode not being able to recover the big and the small islands at the same time (due to the varying density between both). However, regarding its global performance, the aforementioned region shows a strong variation against parameter changes as given by table 5.1. This example already gives a sense of the different performances the studied parameter space can lead to.

5.3 Effect of parameters

One of the main trends we found for each explored spectral range is their stability (or instability) to parameter variation. Exploring this in more depth is crucial, as it will give us a better understanding on why the studied parameter space has such large differences in the recovery of SB2 spectra for a one part of the studied wavelength range, but enables the other to perform with much more stability. In the following we will explore the different choices of both perplexity

and SNR have on the recovery for each studied spectral range.

5.3.1 Variation of perplexity

The theoretical effects of perplexity were already explained in section 3.2 however, the effect it has on synthetic spectroscopic data is not as straightforward and requires a more careful examination.

We investigate the effect of varying perplexity with the help of 5.4 where columns of different color represent the mean recovery for different perplexity values, and the error bars correspond to the minimum and maximum recovery achieved for that spectral range and perplexity value regardless of the SNR. By looking at both panels, two observations become clear:

- The average recovery of the first panel, without counting the H-α region, is larger than 0.6, clearly higher than that of the second panel, whose average recovery is around 0.5.
- The error bars in the first panel, again without including the region containing the H-α line, are clearly smaller than those in the second panel. This further proves what we previously mentioned, that the regions in the bluer side of the studied wavelength range are less susceptible to parameter changes than those in the red. The gradient that we saw on figure 5.3 can also be seen here represented as the varying width of the error bars.

Furthermore, from figure 5.4, it becomes clear that the optimal range of perplexity values lies somewhere between 15 and 30 depending on the internal complexity of the analyzed spectra. This suggests that to achieve the highest possible recovery, a proper balance between the amount of local and large scale structure that is examined by t-SNE is needed.

We can explore these observations in more detail with figure 5.5. In it, we show the t-SNE maps of two regions with very different recovery ratios for the baseline SNR value of 100. Although in terms of absolute numbers for the recovery, the region with wavelengths between 725 - 750 nm has values clearly larger than those from the spectral range 475 - 500 nm, even if in terms of the stability regarding good recovery the region 475 - 500 nm is better (as shown in figure 5.3). However, the reason behind the stability of the bluer regions against the variation of perplexity is the larger amount of small scale structure that is present in the spectra for that region. This is further illustrated using the t-SNE plots. With increasing perplexity, and thus larger number of effective neighbors considered when grouping spectra, t-SNE produces smoother and larger islands of points by smearing out the local structure. We can see this for the blue region, where the embedding done using perplexity 5 clearly shows many small islands, each one containing a smaller group of data-points representing spectra with very concise similarities. Even at perplexity 100, where the amount of local structure that t-SNE accounts for is minimal, there are still sub-structures present on the projection. On the contrary, for the redder region this is not the case, as for perplexity 5 the t-SNE embedding already looks quite smooth and this behavior only increases. This is, again, due to the larger amount of spectral lines that are generally present on the bluer regions of the spectrum for the type of stars that we consider in this work.

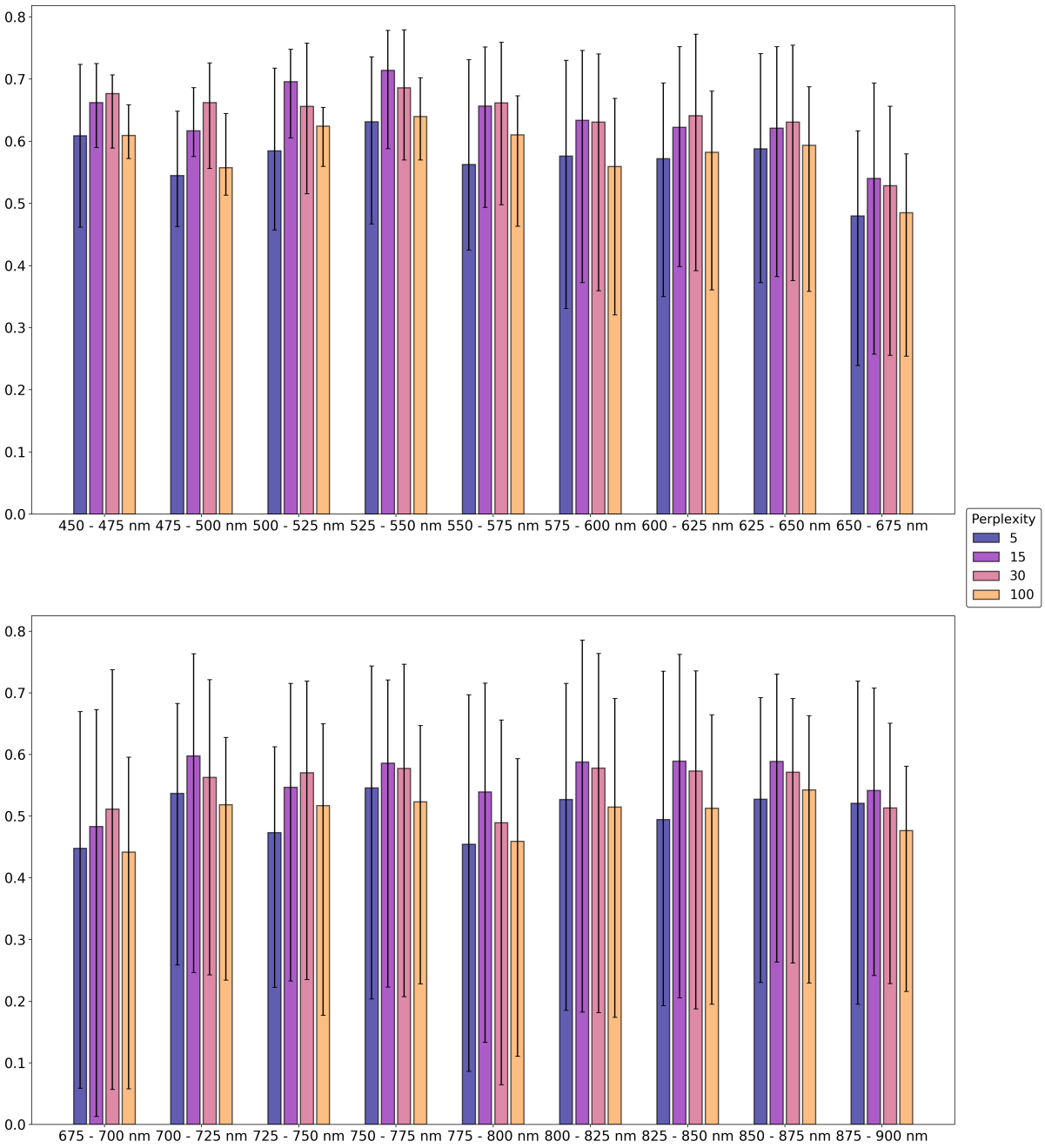

Figure 5.4: Bar plot of the mean recovery per perplexity. Each bar is color coded according to one of the examined perplexity values and it shows the recovery achieved averaging over all SNR values. The error bars represent the maximum and minimum value achieved for the given combination of perplexity and spectral range.

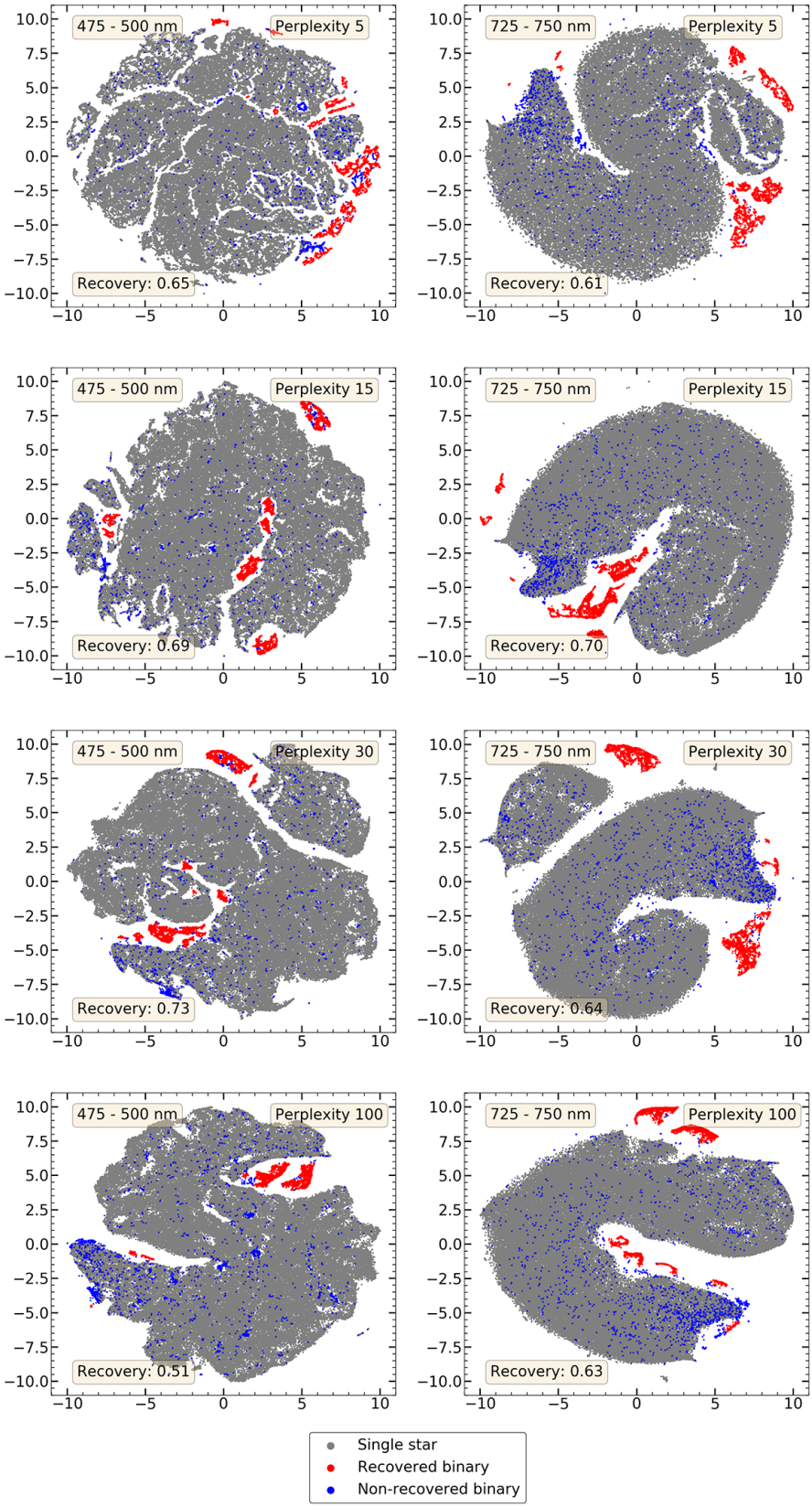

Figure 5.5: t-SNE maps of both regions between 475 - 500 nm and 725 - 750 nm with a fixed SNR of 100. The perplexity of each panel, as well as the spectral range and achieved recovery is indicated on the panel itself.

5.3.2 Variation of noise

Although SNR of spectra generally depends on specific observations, here, because we are working with clean, synthetic spectra, we can analyze the effects it has on spectroscopic binary recovery. We do so in a similar fashion as we did previously in section 5.3.1 for the perplexity. On figure 5.6 we show a similar plot to figure 5.4 but analyzing the effect of varying other parameters while keeping SNR fixed.

From figure 5.6 we observe that:

- Up until 575 nm, our method is capable of recovering almost 50% of the binary stars regardless of the noise present in the studied spectra, even under values as low as 10. Furthermore, in the spectral range between 450 - 475 nm, the average recovery under SNR of 10 is larger than the average recovery for the same spectral range for very clean spectra with SNR 500.
- The gradient we have so far seen in figures 5.3 and 5.4 is presented in figure 5.6 as a decreasing recovery when using SNR 10 and also as a larger difference between the average value of the low SNR values and the higher ones.
- The range of variation between maximum and minimum values when varying the SNR is much lower than that shown in figure 5.4 for the perplexity variation. Interestingly, the SNR value that shows overall the largest variation is the largest, 500. Furthermore, the average recovery for SNR 500 is not the highest in any of the 18 shown spectral ranges.

The strongest takeaway from 5.6 is that the highest average recovery are achieved when there is a moderate level of noise and not at lower amounts, as intuition would suggest. This is the case for SNR 50 and 100, for which we obtain the highest average recovery in all of the 18 examined spectral ranges. From this, we see that the effects of noise in removing information from the spectra can be beneficial under some circumstances for the detection and recovery of data-points representing SB2.

In figure 5.7 we show the same spectral regions as in 5.5 but with varying SNR and fixed perplexity. As we explained before, some amount of noise appears to be beneficial, mostly in the bluer regions with higher information content and many spectral lines. This is behavior is seen in the first column of figure 5.7, which corresponds to the spectral range between 500 and 525 nm and where the recovery peaks at SNR 50 and then decreases slightly. The spectral range between 800 and 825 nm shows, on the contrary, a recovery that increases with increasing SNR due to the presence of fewer spectral lines compared to the bluer region.

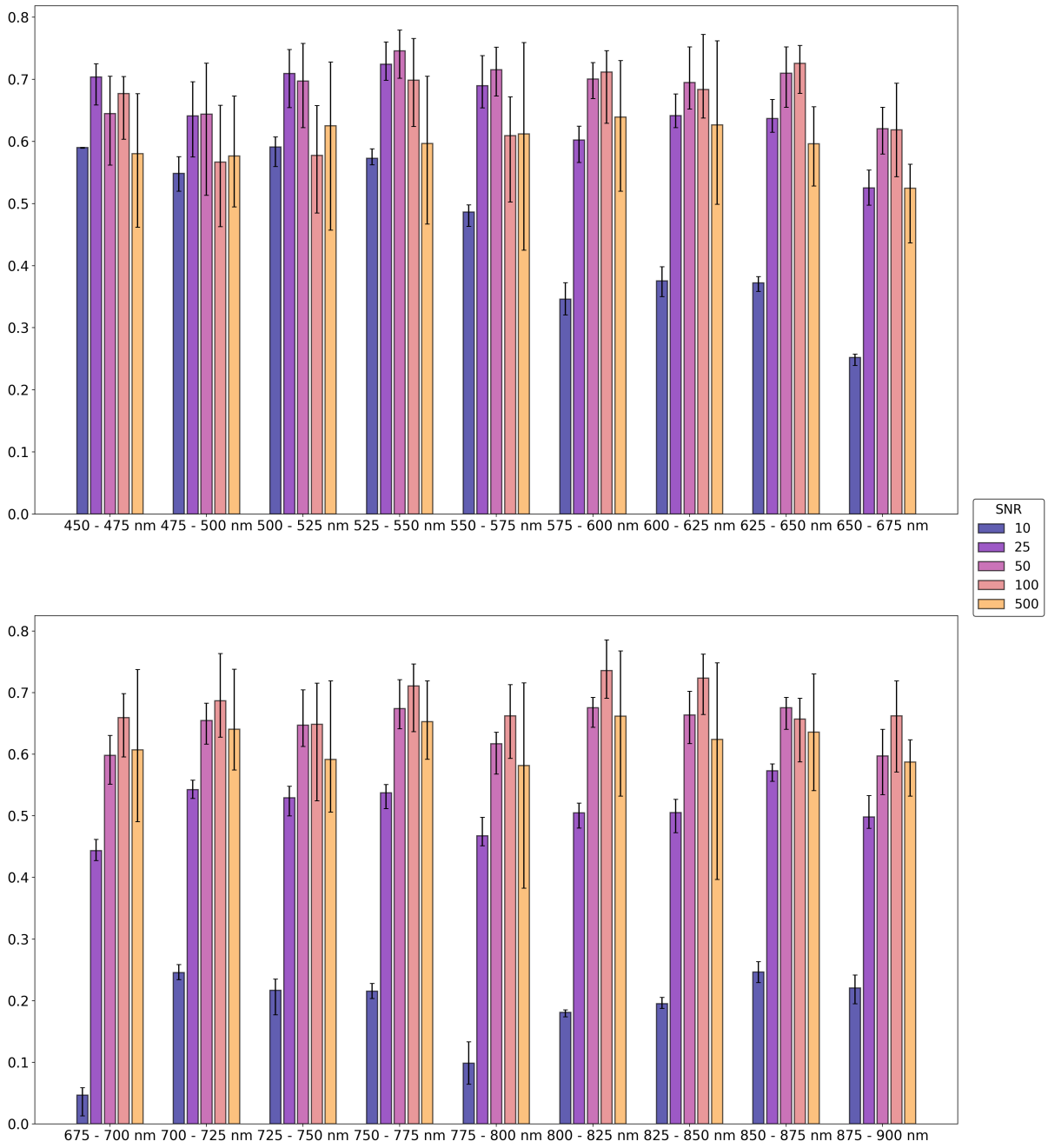

Figure 5.6: Bar plot of the mean recovery per SNR value. Each bar is color coded according to one of the examined SNR values and it shows the recovery achieved averaging over the all perplexity values. The error bars represent the maximum and minimum value achieved for the given combination of SNR and spectral range.

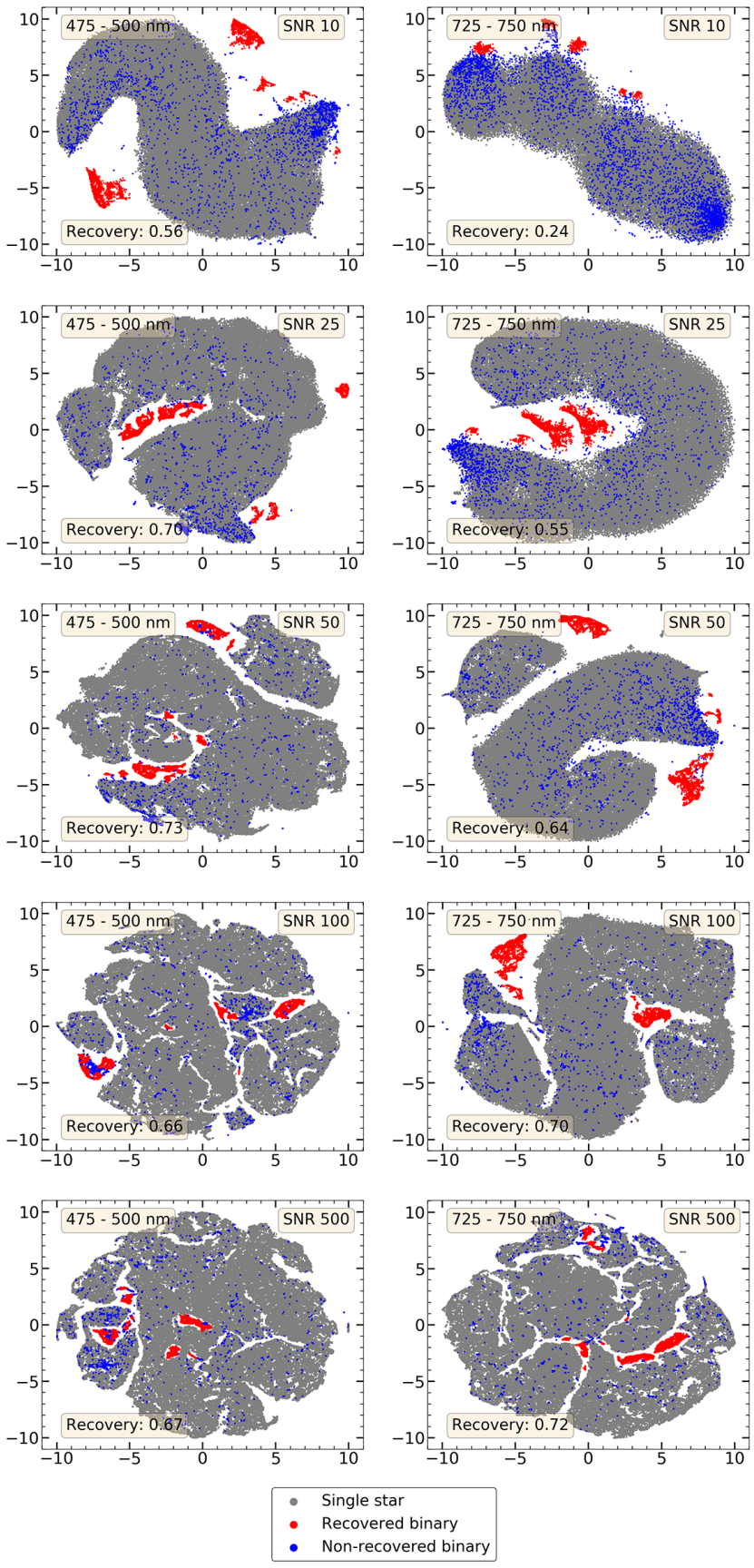

Figure 5.7: t-SNE maps of both regions between 475 - 500 nm and 725 - 750 nm with a fixed perplexity of 30. The SNR of each panel, as well as the spectral range and achieved recovery is indicated on the panel itself.

5.3.3 Effect of the stellar parameters

The synthesis process explained in chapter 2 and the subsequent stellar parameters from the selected GALAH sub-sample that were used in combination with *turbospectrum*, as well as the sampled binary parameters we used to generate the binary population (and the corresponding binary spectra) do play a major role in the analysis. In the following, we present two distinct groups of histograms on figures 5.8 and 5.9, which correspond to a parameter combination that showed low recovery and to one that resulted in a larger fraction of recovered binaries, respectively. The histograms, which are shown with the same perplexity value, present the stellar parameters that play a major role in the shape of the spectra corresponding to each individual component of the system and as well as those that shape the resulting combined spectrum. The data that was used to create the histograms shown in this section is that belonging only to the actual binary systems, thus we do not show here any of the possible false positives that may have been introduced during the DBSCAN recovery of the clusters, even if some of the binary systems were recovered alongside other single stars.

The scenario with low recovery ratio shown in figure 5.8 presents a low value of 25 for the SNR. This will cause many spectral lines to be smeared out, making it harder for t-SNE to identify differences between. From the histograms we observe that:

- The binary systems that are not recovered are those with low radial velocity difference and/or a low luminosity ratio (the two are not correlated). In this case, the spectrum of the binary shows either a very small separation on the wavelength axis or the spectrum from the primary star completely dominates the combined spectrum due to the high difference in luminosities. The difficulties in recognition are additionally amplified by the high level of noise.

- Regarding the primary component of the binary systems, a large amount of those presenting the lower values of $log\, g$ are missed during the recovery. This effect is also seen on the T_{eff} histogram for the distribution of the recovered binaries, where no primaries with high temperatures were recovered. This correlation is due to the assumption that all of the stars we synthesized are main-sequence, unevolved dwarfs and therefore a lower value of $log\, g$ means a higher value of T_{eff}.

  There are two reasons for t-SNE and DBSCSAN missing these stars. The first reason is a direct consequence of the pairing algorithm we designed. For a hotter primary, the range of parameters its secondary star could have is larger than that for a colder primary, which in turn means that hot primaries have a higher probability of being in a system with small luminosity ratios and/or small radial velocity differences. The second reason for missing the hot, low $log\, g$ stars is due to spectral lines being fewer and weaker and at the same time hydrogen lines dominating most of the spectrum as they are stronger and wider (hydrogen lines get more prominent with increasing effective temperature up to 10000 K - for main sequence stars). The combination of these two factors affects the ability of t-SNE to properly identify the features related with binarity in the resulting projections.

- The distributions for the parameters of the recovered secondary components seem to follow the distribution of the whole sample of secondaries and only on the metallicity there is a slight shift towards a worse recovery of more metal-poor spectra, which can be explained with the same reasoning that applies for the primary stars of the binary systems.

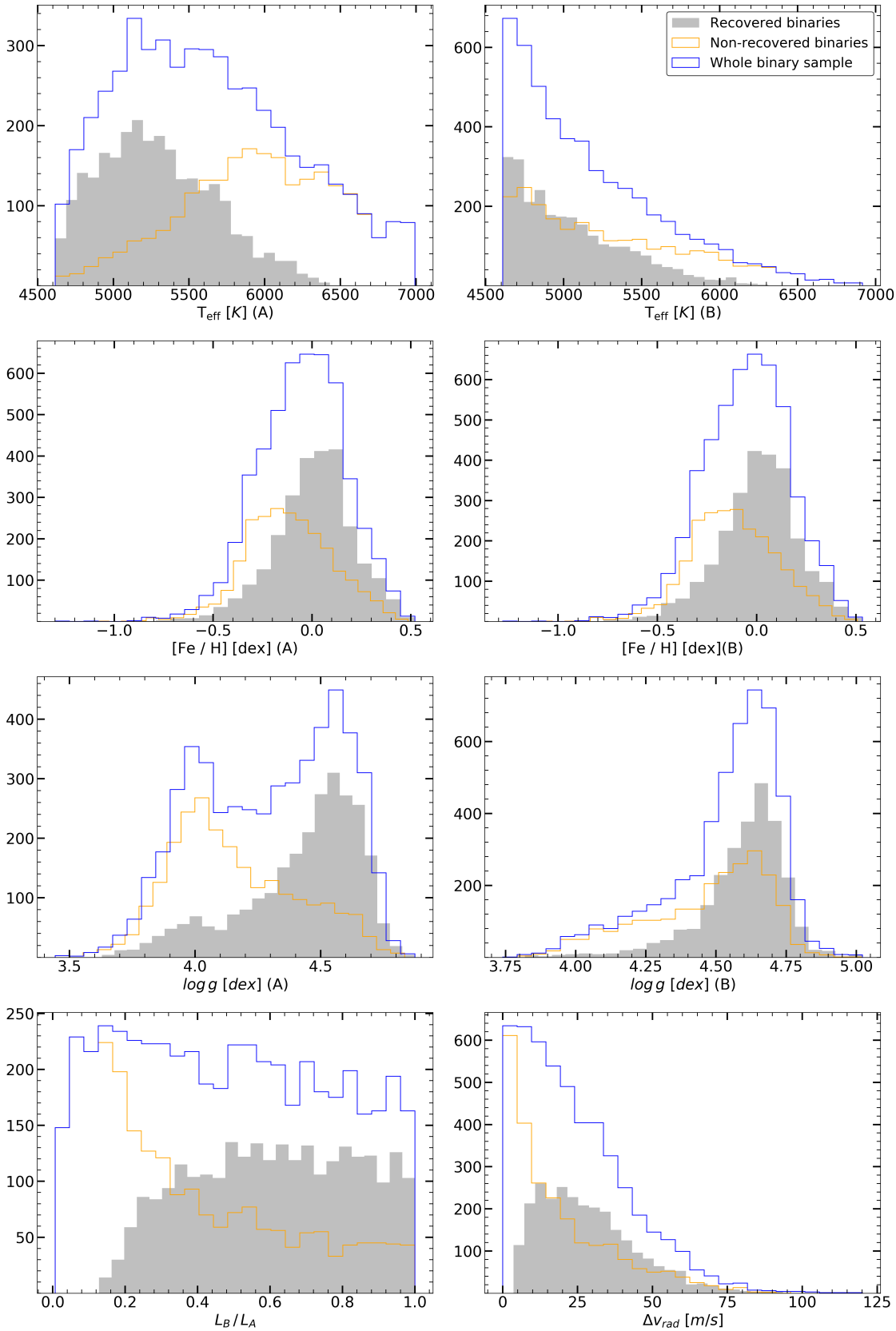

Figure 5.8: Histograms for the main stellar parameters of both the primary (A) and the secondary (B) of each of the analyzed synthetic binary systems, corresponding to 650 - 675, SNR 25 and perplexity 30.

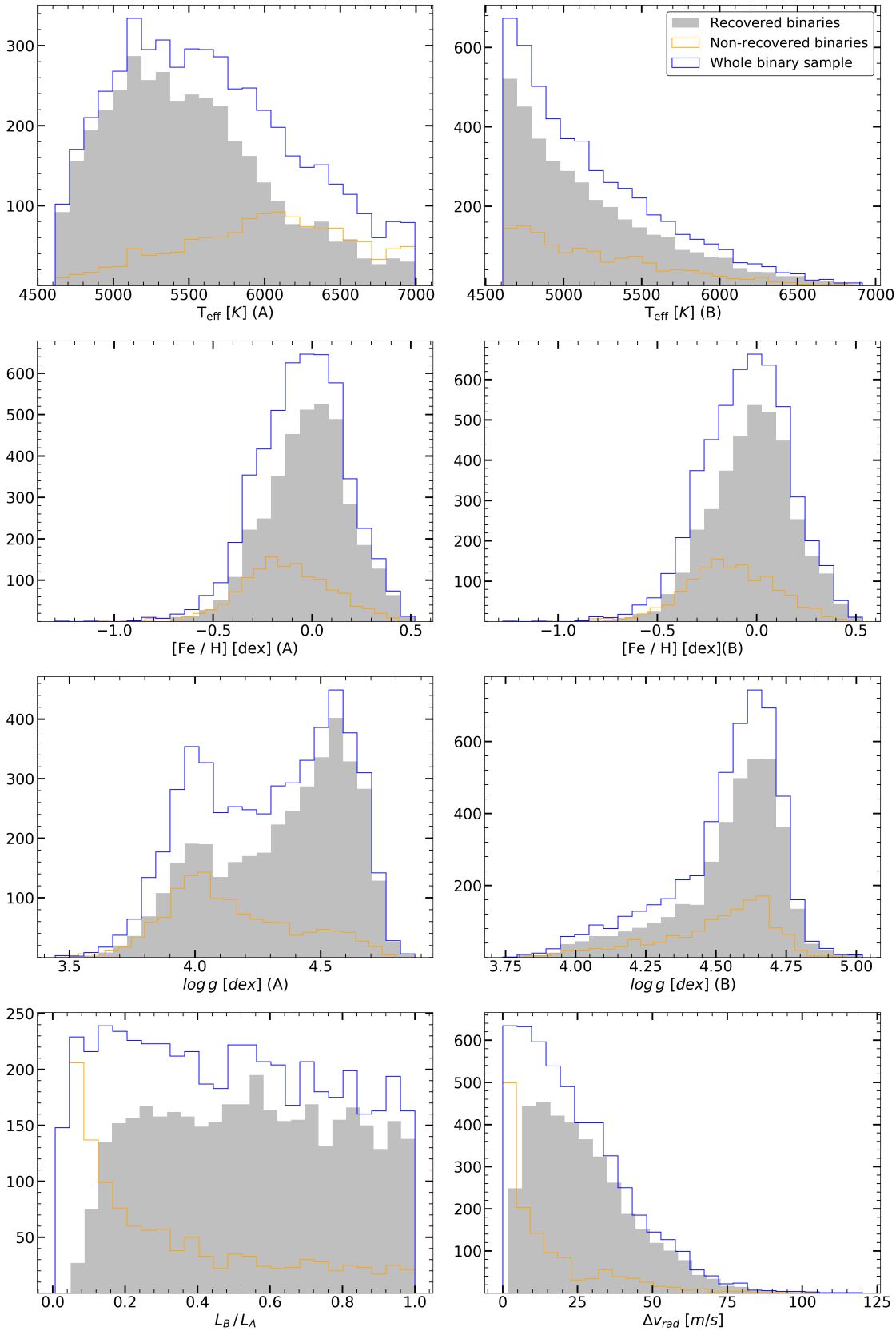

Figure 5.9: Histograms for the main stellar parameters of both the primary (A) and the secondary (B) of each of the analyzed synthetic binary systems, corresponding to 500 - 525, SNR 100 and perplexity 30.

In the high-recovery situation of figure 5.9, there appears to be a more homogeneous recovery, as all distributions for the recovered stars (for both components) resemble the distribution of the whole sample, with the exception of the surface gravity of the primary, which still presents a small dip on the low $log\,g$ primaries. This is due to the same reason we mentioned for the histograms in figure 5.8, as these primaries have the highest chances of being in a system with low luminosity ratios and therefore are the ones with an increased probability of being missed by t-SNE. We clearly see from figures 5.8 and 5.9 that the luminosity ratio and radial velocity difference are the parameters that overall most significantly influence the SB2 recovery as they present hard dips at lower values, while in a non-ideal situation (e.g. the case in figure 5.8), other stellar parameters of both components seem to influence the general performance as well.

5.4 Individual examination of the binaries

The study done on our data-set of synthetic stellar spectra can be complemented with an analysis of the individual binaries. By studying the amount of times each binary system is recovered under the given parameter combinations in our study, we can analyze with more accuracy the influence of the stellar parameters in the final recovery of the binary systems as well as give confidence intervals regarding their final detection and classification. In the following, we also only the results for the actual binary systems, thus the results presented here do not show any of the possible false positives that may have been introduced in the analysis.

In figure 5.10 we show 6 plots of different binary stellar parameters and relate them to the total amount of times each individual binary system was recovered in our analysis. The first two plots of the left column, a) and c), present the same data as the two first plots shown in 2.5 but color coded according to different bins that correspond to the absolute recovery. We can see that the binary systems that are more easily recovered are those where both components have very similar $T_{\rm eff}$ and $log\,g$ values, mostly showing sub-solar values. In plot a), we can also see that even for twin systems, the amount of times the system was recovered tends to decrease with increasing temperature, consistent with our findings in the previous sections. On plot e) we show the luminosity ratio against the absolute radial velocity difference. We can see that systems that were recovered most lie in a zone between 20 and 50 km/s for the radial velocity difference and span even to luminosity ratios of ~ 0.6. Furthermore, we show that for radial velocity differences lower than ~ 10 km/s there seems to be a strip of binary systems that were recovered at most on 5% of the simulations. This suggests the presence of a hard limit for the recovery of binaries. On the contrary, an increasing radial velocity difference value does not seem to guarantee an easier recovery, although the lack of systems with such high values can cause t-SNE to not isolate them in a distinct island due to a lack of similar closest neighbors, and even if such a small island exists, it might not be identified by DBSCAN due to the $minPts$ parameter.

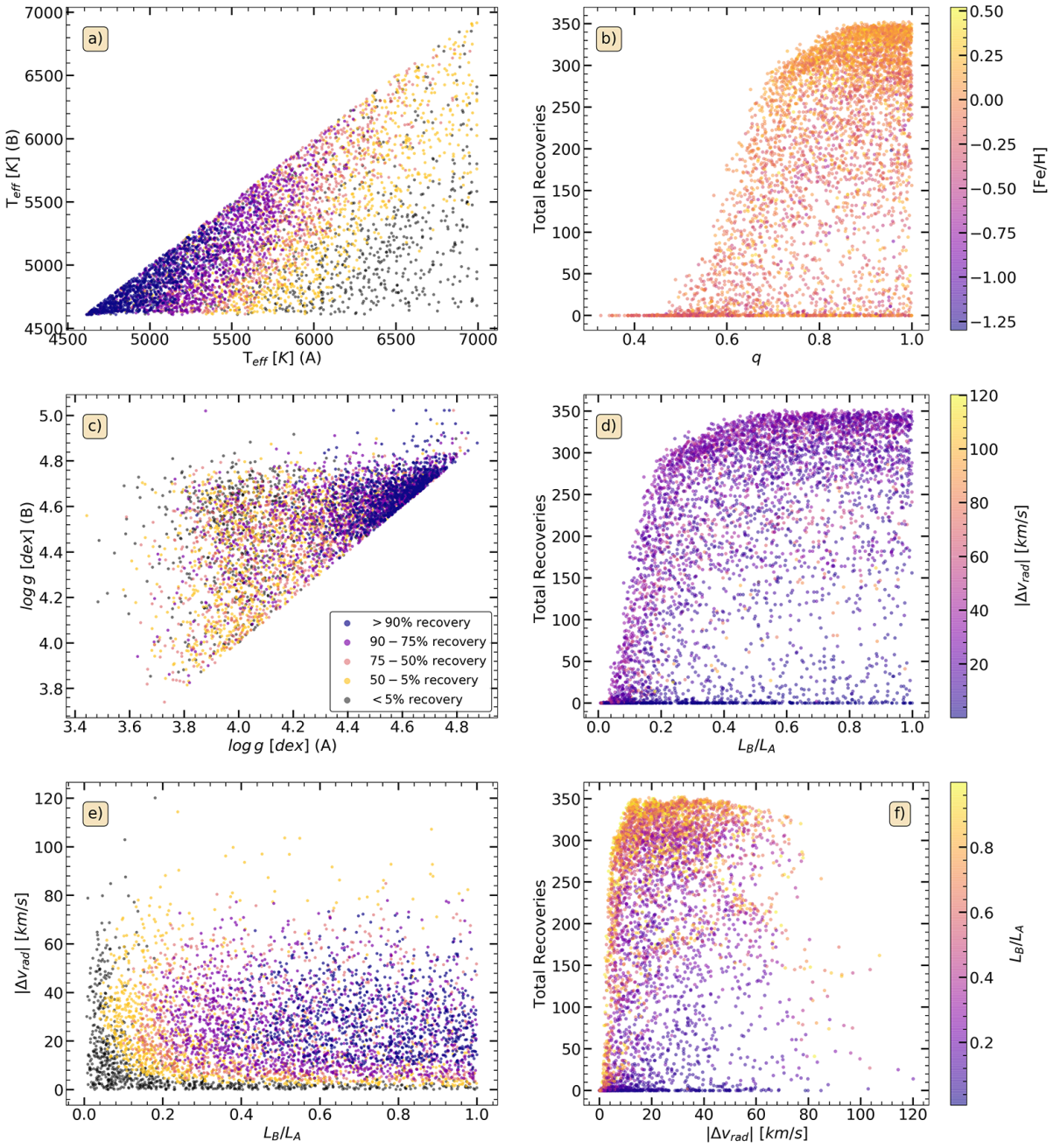

Figure 5.10: Recovery of the individual binaries. The first column (plots a), c) and e)) shows the effective temperature and surface gravity of the primary against that of the secondary respectively and the luminosity ratio against the absolute value of the radial velocity difference, color coded according to the total amount of recoveries per system. The legend on plot c) serves for all three plots on the first column. On the second column (plots b), d) and f)) we show total number of recoveries against the binary parameters mass ratio, luminosity ratio and absolute value of radial velocity difference.

In second column of figure 5.10 we show the total number of times a binary system was recovered with respect to three parameters of binary systems: q, L_B/L_A and $|\Delta v_{rad}|$. Plot b) shows the recoveries per system according to their mass ratio. This plot is color coded regarding the average metallicity of each system to further exemplify what was shown in figure 4.4, that the metallicity seems to have a less strong influence in the final shape of the t-SNE projection and therefore does not seem to influence the recovery strongly, and is seen on plot b) as an homogeneous coloring and that it hardly shows any gradient. In plot b) we can also see that were able to achieve large percentages of recovery for systems with mass ratios down to ~ 0.65. In plot d) we can see that although there is a strip of binary systems that show little to no recovery for all luminosity ratio values (which correspond to the strip of black points seen on plot e)), a large fraction of the binary systems with $|\Delta v_{rad}| > 15$ km/s are recovered in more than 80% of the simulations down to $L_B/L_A \approx 0.2$. A similar behavior can be seen on plot f), where the black points from plot e) form the band of no recovery regardless of their radial velocity difference value. Furthermore, we can see from plot f) that t-SNE is able to detect binary systems with $|\Delta v_{rad}|$ values down to 8 - 10 km/s for more than two thirds of the simulations, at which point in the plot the amount of recoveries per system falls abruptly (this radial velocity difference value accounts for a shift on the secondary spectrum of 0.015 nm with respect to the primary component).

Chapter 6

Discussion

In this work we presented a method capable of automatically identifying and extracting binary stars from a data-set of stellar spectra as well as showed an in-depth examination and optimization procedure of the parameters that play a major role in its performance. As we showed in the previous chapter, the absolute numbers regarding the recovery ratio of the binaries is promising. However, it is important to keep in mind that even if the numbers of recovery are high, the resulting projections and chosen DBSCAN modes might vary when applied to a real spectroscopic survey. We have shown that the quality of the different t-SNE maps and the amount of detected clusters by DBSCAN play a major role on the scalability and applicability of this method and it needs to be assessed properly, regardless of the absolute numbers presented in this work. The method we developed appears to be very sensitive to the level of complexity and internal structure (intrinsic dimensionality) of the data used. We have also seen that not all the examined possible parameters yield acceptable results, and for this reason in the following points, we will explore the caveats of the method, some best practices for the user to determine the best starting point and what could be done in future iterations of this work.

Previous work

To the best of our knowledge, a similar set of simulations to characterize and optimize machine learning methods for the detection of binary stars using synthetic spectra has not yet been undertaken. Matijevic et al. 2010 presents a study related to our work, where they simulate the SB2 detection capabilities of the CCF method on a synthetic subset of dwarf stellar spectra from the Radial Velocity Experiment (RAVE) survey (Steinmetz et al. 2006). Although some assumptions are similar, such as the use of dwarfs and a common metallicity for both stars in a binary system, they use spectra from only the near-infrared (between 840 and 880 nm) with a resolving power of 7500, almost a fourth of the one we use in this work, and they make use of an essentially different procedure.

6.1 Variable parameters

6.1.1 Spectral range

In figure 5.3 we showed that almost all of the inspected spectral regions yield recovery ratios ≥ 0.7 with an appropriate choice of the other variable parameters. However, from figure 5.3 it becomes clear that the regions located on the blueward of the H-α segment (650-675 nm) are more stable against parameter variations than those at redder wavelengths. This stability ensures that even under sub-optimal conditions (high noise levels or a poorly chosen perplexity), a relatively high degree of recovery can be achieved. The stability of the different spectral regions is further explored and confirmed in figures 5.4 and 5.6. According to our results, we conclude that the spectral region of the GALAH survey from those given in table 2.1 that is

most suited for binary detection would be blue (471 - 490 nm) and in a lower degree green (564 - 587 nm), however, this will be dependent on the quality of the captured spectra and the processing pipeline.

The region between 650 and 675 nm, which was specifically chosen to be centered around the H-α absorption line, showed the lowest maximum recovery percentage of all the examined spectral ranges. This behavior was expected and confirmed what was previously found by Traven et al. 2020, who by masking the H-α and also H-β (at 483.15 nm) lines on the spectral templates used by the CCF method was able to improve the detection rates of SB2 spectra. However, we find that H-β line does not have any noticeable effect on the final fraction of recovered binaries in the case of our work, most likely due to the abundance of other spectral lines in its wavelength segment and the fact that it is generally less prominent than H-α.

Furthermore, one possible prospect for future research would be to study the effects that a very broad spectral region, such as the whole visual range (between 380 and 740 nm), would have on the detection of binarity with the simulations presented in this work. We have shown that an increased amount of information in the form of spectral lines has the ability to yield higher, more stable recovery numbers (as well as better suited t-SNE projections). However, the inclusion of many more lines and thus new information about the corresponding stellar atmospheres that would come with an spectral range of increased length, could potentially result in an higher amount of local structure present in the t-SNE projection, which would in turn create unnecessary sub-divisions in the binary islands. Although this might be of interest when addressing the general categories and different morphologies present in the unexplored spectral data, we believe that the detection of binarity demands a much more careful approach in order to balance the amount of local and large scale structure explored by t-SNE. Even though this can be achieved by using different perplexity values, the increased computational costs of analyzing such high dimensional data (again, due to the wider wavelength range) might prove unrewarding. For this reason, we argue that narrower wavelength regions, such as the ones used in this work, are sufficient and well-suited for the detection of binaries together with the presented method.

6.1.2 Perplexity

The effect of perplexity on the resulting t-SNE projections cannot be ignored as it arguably is the most important parameter that drives t-SNE. Maaten and G. Hinton 2008 argues that the typical values for perplexity should lie in a range between 5 and 50 and Van Der Maaten 2014 defaults this value at 50. However, we find from our analysis that the given range is too broad and the results are not acceptable for some values.

Although there are several other parameters that can be input in t-SNE to fine-tune the processes occurring in the algorithm, their impact on the final projection is rather minimal when compared to t-SNE. For this reason, we used the default values in the used t-SNE implementation. For more information regarding these hyperparameters and their theoretical impact on the algorithm, we refer the reader to Linderman and Steinerberger 2019.

From the projections we have shown in figured 5.5, 5.7 and those from the baseline model analysis in figure 5.1, we observe that the best recovery values are associated with a t-SNE projection that has a clear separation between its single and binary clusters as well as a smoother and more uniform single star island. This indicates that a proper balancing between the local and global structure is necessary for t-SNE to be able to extract the features that highlight their binarity and allow for a successful recovery of SB2 spectra, although under the most stable spectral ranges more extreme values of perplexity are usable. For this reason, we argue that

a perplexity value between 15 and 30 should be used for this purpose. However, depending on the intrinsic dimensionality of the investigated data-set, these numbers might change as one would prefer a more localized analysis and therefore a lower value of perplexity or vice-versa. We suggest as a starting point for higher values of intrinsic dimensionality the use of a perplexity value of 30 and for the opposite situation of lower values of intrinsic dimensionality, we recommend the value for perplexity to be closer to 15.

6.1.3 SNR

In subsection 5.3.2, we studied the effects that SNR has on the other variable parameters and contrary to intuition, we found that moderate noise levels between SNR 50 and 100 can potentially have a beneficial impact on the final binary spectra recovery, with simulations on spectra with SNR values within the aforementioned range showing better recovery fractions in average, than those simulations on spectra with a very large SNR value of 500. The presence of noise can conceal undesired features that might impact the recognition of binarity by t-SNE and can reduce the amount of substructure to be analyzed, thus smoothing the resulting t-SNE embedding and making it more appropriate for binary detection. This smoothing of the embedding is beneficial for the DBSCAN analysis as well as we showed in figure 5.7, where both examined regions present almost ideal maps for recovery between 50 and 100 SNR with proper separation between the binary islands and those containing single stars. On the contrary, we also saw that in some occasions having almost noiseless spectra (SNR 500) can be counterproductive for binary star detection when applying the described method. These effects are furthermore more dominant on the bluer spectral regions from the range we examined, as there is amount of information in the form of spectral lines is larger than for those with redder wavelengths. Some of this information is, however, not relevant for binarity detection and the effects noise has on concealing superfluous features in the spectrum can be, up to some degree, beneficial for the detection of binarity.

The stability of the investigated wavelength ranges against different SNR values is also much higher than in the case of varying perplexity (excluding SNR 10, whose recovery values are only acceptable for wavelengths until $\sim$ 550 nm). However in reality, the amount of noise in real spectra depends on the wavelength range (among other factors) and is not constant throughout the whole observed range, as we assumed for our analysis. Furthermore, for real dwarf spectra (mostly from FGK stars), noise at bluer wavelengths will be higher due to those stars emitting less photons with energies within those ranges. For this reason, we expected the shown behavior will even out when dealing with real data.

We suspect that the role of noise in diminishing the influence of spectral lines will become much more important the larger the analyzed sections are, as the internal complexity and the information content of the spectra is related to the wavelength range it belongs to (Ruchti et al. 2016). For this reason, we believe that knowing the effects of SNR on this type of analysis is a must for future surveys such as 4MOST (De Jong et al. 2012), where the analyzed spectral ranges will be at least double of those studied in GALAH and the increased complexity of the investigated spectra could have a strong influence on the binarity detection.

6.1.4 DBSCAN modes and their selection

The method we developed to recover binaries based on two machine learning algorithms that work in tandem, depends strongly on the ability of DBSCAN to automatically extract the clusters where t-SNE grouped the binary stars. Its performance is also strongly tied to the used hyperparameters. However, we showed that the combination of ϵ and $minPts$ (DBSCAN mode), whose selection, as we mentioned in subsection 3.3.1, is not straightforward. We overcome this issue with an iterative testing of all the possible modes for a given range of ϵ and $minPts$ and

evaluating them regarding their binary recovery efficiency. For the exploration presented in the previous chapter, this method did not pose a large increase in computation times (100 studied modes). However, this iterative mode testing is not optimal when the search grid is made finer and the amount the modes to be analyzed is consequently larger. Every iteration is a whole new DBSCAN computation and due to the large number of points to be analyzed, the time per iteration is not negligible.

As seen in figure 5.2, there exists a linear relation between ϵ and $minPts$ and the zones of higher recovery, which approximately holds for all spectral regions. This relation is directly proportional to the density of the resulting t-SNE projection and we believe it could be related to the amount of points on the given data-set. If properly quantified, it would be possible to restrict the ranges of the parameter space exploration, as an approach that would target the specific modes that have a higher probability of detecting clusters. Targeting these specific mode combinations that show the highest SB2 recovery efficiency would be key for an efficient DBSCAN usage and would allow to extract the best possible performance while using the least amount of computational resources.

We believe that because performance of DBSCAN is directly tied to the quality of the projection generated by t-SNE, as it depends greatly on the data-point density in the final projection, the best solution for scalability of the DBSCAN mode selection method would the appropriate resizing of the t-SNE map (as we did in chapter 4) and as we explain later in section 6.4 to maintain a certain density for which the appropriate modes are known.

6.2 Data selection and spectral synthesis

In this study we chose to work only with dwarf stars according to the dwarf/giant division from Zwitter et al. 2018 under the assumption that all of the synthesized spectra belong to unevolved main-sequence stars. This allowed us to use the scaling relations from subsection 2.3.2 do define approximated stellar properties for the synthesized spectra. However, we believe that to extend the utility of the study presented in this work, the optimization procedure should be extended to include a data-set formed both by giant and dwarf stars. We used only dwarf stars as most of the SB2 systems seen in GALAH correspond to dwarf-dwarf pairs (Traven et al. 2020) and also for the reasons explained in section 2.1.

A computationally feasible spectral synthesis involves many approximations that, while reasonable, diverge from real spectra in several ways. For this work we used MARCS, a 1-D hydrostatic model of stellar atmospheres under the assumption of LTE, in combination with the spectral synthesis code *turbospectrum* and an atomic transition list with hyperfine splitting information. The resulting spectra yielded by this model and code combination had an accuracy high enough to be used in the simulations presented in this work, albeit some discrepancies in the depth of some spectral lines that were evident. However, the spectral synthesis we carried out did not account for effects single stars might exhibit, such as asymmetric line profiles due to pulsations or stellar spots, which could mimic the effect of strongly blended binary spectra and could produce a false positive binary identification. Furthermore we did not simulate any effects that might occur in very close binaries that undergo interactions (as we assumed that our binaries are all pairs of non-interacting dwarf-dwarf stars), whose measurable effects on the spectra could manifest the binary nature of the system and which could be used to more accurately classify them as SB2. This is, nevertheless, a rare occurrence. With respect to defects on the spectra caused by the measurement process and instrumentation, only the noise through the SNR parameter was modeled. For further research, we believe that other effects, such as e.g. continuum normalization effects should be included, in order to make synthetic spectra more

realistic.

In figures 5.8 and 5.9 we show the consequences from our pairing algorithm reflected on the histograms of the recovered and non-recovered binary stars. The two parameters that saw the largest impact on the histograms were binary-defining parameters: the luminosity ratio (which is directly related to the primary mass and mass ratio) and the radial velocity difference. This was to be expected, as the binary nature of systems with small velocity separations and/or large brightness differences will be harder to detect due to the blending of both spectral components and/or due to the disappearance of the secondary's lines in observational noise. Most of the effects seen on the histograms corresponding to the whole sample are due to the conditions we imposed in section 2.3 and whose effect is shown in the plots contained in figure 2.5. We did not separately investigate the effects of inclination, eccentricity, or period of binary orbits, however, these manifest themselves in the radial velocity separation distribution that we adopted for our binary population based on observational studies (see e.g. Traven et al. 2020). Even if the results we obtained were not unrealistic, a more sophisticated pairing algorithm or even a stellar population simulation would have certainly yielded more accurate pairs of stars and therefore histograms that better resemble a real binary population, such as the one shown in Traven et al. 2020.

In figure 5.10 we saw limiting values for the luminosity ratio and radial velocity difference, at ~ 0.2 and $8-10$ km/s respectively, at which the recovery fell abruptly. We believe that while the luminosity ratio is due to the partial or total disappearance of the secondary spectrum by the much larger luminosity of the primary star (see Hogeveen et al. 1991 for a discussion on the visibility of the secondary lines), the limiting value for Δv_{rad} is probably connected to the resolving power of our synthetic spectra, where a higher resolving power will in principle enable recovery down to lower values of radial velocity difference due to sharper spectral lines and thus less severe blending (Traven et al. 2020).

In this work we studied the general effects of the optimization procedure on the whole binary population. Furthermore, we investigated the effect of varying the parameters on each of the individual synthetic binary systems, the results of which are shown in figure 5.10. This allowed us to extract confidence ranges of binary parameters for which the recovery was almost guaranteed under almost any of the 360 analysis variations that we performed. With this information, we put together a table of sub-sample of 40 bona fide binary stars that were recovered in more than 90% of the performed analysis. This table is shown in Appendix A to serve as a guideline for future works and when tracer binaries are needed in investigations of binarity for future spectroscopic data-sets.

6.3 Machine learning algorithms and the detection method

We showed that the combined use of t-SNE and DBSCAN allows for an efficient recovery of binary stars from a set of stellar spectra, as previously done in Traven et al. 2016; Traven et al. 2020. As explained in chapter 3, t-SNE is one of the leading machine learning algorithms for dimensionality reduction. Even though t-SNE has performed remarkably for the purpose of this work, it still presents some drawbacks, such as the poor handling of data with a large number of intrinsic dimensions (Maaten and G. Hinton 2008), its slow computation times (mostly in its non-approximated implementation) or the strong dependency of the final projection of the selected perplexity value. One of the possible alternatives to t-SNE, a recently developed dimensionality reduction method called UMAP (McInnes et al. 2018), short of Uniform Manifold Approximation and Projection, is claimed to have all the benefits that t-SNE presents, mainly the proper modeling of similarities as distances and the clear separation between the clusters

in the low-dimensional embedding. UMAP focuses on solving some of the difficulties that t-SNE presents, such as its difficulties with the high number of intrinsic dimensions or its slow computation (UMAP is on average one order of magnitude faster than t-SNE). Furthermore, McInnes et al. 2018 claims that UMAP is capable of a better preservation of the local and large scale structure of the high-dimensional data-set on the lower-dimensional representation. For a comprehensive testing and review of UMAP, we refer the reader to Becht et al. 2018 (the testing is done using data from single-cell transcriptomics). We believe that using UMAP in combination with DBSCAN could yield at least equally good results, while being faster and easier to implement.

DBSCAN has been widely used and tested in many areas of research as well as astronomy (Tramacere and Vecchio 2013; Traven et al. 2016; Shou-kun et al. 2019) and has been highly regarded as very useful in a wide range of situations and purposes due to its ability to detect clusters of varying morphologies. However, it has some drawbacks. Even if we circumvented the difficulties present in the mode selection, there is still an issue with the DBSCAN modes: they target a specific density of data-points to determine the clusters. Moreover, as we have seen on figure 4.6, even the mode that yields the best recovery misses a cluster due to its lower density of data-points. This is an issue with the current implementation of DBSCAN based on Ester et al. 1996, which does not contemplate a data-set with clustering at different density levels. Yet this would be required to fully recover all of the binary clusters that have been properly separated in individual islands by t-SNE, as shown in the resulting projections from the simulations. In the recent years there have been, however, improvements on the original DBSCAN that specifically targeted this issue. One of this efforts is called Varied Density Based Spatial Clustering of Applications with Noise or VDBSCAN for short (P. Liu et al. 2007). It finds all of the necessary modes by computing first the distance of each point to its k-th nearest neighbor, which in turn allows the algorithm to distinguish the different levels of density present in the input data and it is able to target and detect all clusters by applying DBSCAN with the selected modes on the data-set. Other solutions to the density issue involve different algorithms, such as OPTICS (Kriegel et al. 2011), a density based algorithm which has been designed to overcome the aforementioned varying-density problem or Mean-Shift clustering (Fukunaga and Hostetler 1975), which follows a different approach for the cluster detection while keeping the most of the benefits of DBSCAN.

In subsection 4.1.4 we described the algorithm we designed to determine whether the clusters detected by DBSCAN were formed by binary or single stars. For this purpose we introduced the binary ratio in equation 4.1, which is a simple prescription that uses a user-specified threshold value which we set as 0.9 in our analysis. If the ratio of binaries in a given clusters is larger than the threshold value, then all of the stars in that DBSCAN cluster are automatically marked as binaries. For such a high value the behavior of the method is expected to be quite stable, although in some occasions where the binary clusters are not properly isolated by t-SNE or there is a density miss-match with the neighboring points, it could happen that the method introduces false positives in the detection. However, even if that is the case, their amount will be equal or lower than 10% of the stars.

Our chosen combination of machine learning methods works remarkably well for the discovery of binary spectra in spectroscopic data. However, as shown on the examples from chapter 4, figures 4.2 and 4.6, we still miss those binaries that are mixed with the single stars. Traven et al. 2020 partially solved this issue on real data by using the Cross-Correlation Function method, recovering more binaries than with the combination of t-SNE and DSBSCAN alone. These binaries lie on the edges of large clusters in their projection, probably due to them being somewhat different from the stars in that cluster, but not enough to be placed on a separate cluster. This behavior is clearly seen in our analysis as well, e.g. on the t-SNE analysis of

the baseline configuration in figure 5.1, where fairly large groups of undetected binaries lie on the edges of the main single star clusters. Nevertheless, it would not be possible or at least very difficult to detect all of the binaries in the data-set, as some of them have spectra that are almost indistinguishable due to a combination of e.g. low luminosity ratio or small radial velocity differences. This is also one of the reasons why, in order to simulate a realistic spectroscopic data-set, we use a larger relative number of input binary stars compared to what can be extracted as SB2 from a real survey, as we know that some of this input binaries will not be possible to recover. However, we are still hugely underestimating the overall number of multiple systems among FGK stars, which under some circumstances could amount to the majority of the observed stellar population according to some estimates (Raghavan et al. 2010; Duchêne and Kraus 2013).

Performance improvements

Because the developed method and extracted conclusions are intended for use on real spectroscopic data, we suggest several improvements and further explorations that could alleviate such an effort. The computations presented in this work took approximately 1500 CPU hours on an Intel(R) Core(TM) i7 3.40GHz 8-core machine. Of the total time, around 80% was spent on the t-SNE projections and the other 20% was used to prepare the spectral sample for the analysis and for the DBSCAN parameter space search. To achieve a faster and more efficient implementation of the presented method, we suggest two main changes to improve the performance: firstly, to examine the relation between the DBSCAN modes and the number of data-points and quantify it, either by exploring the consequences of resizing the t-SNE projection or by finding an analytical expression to relate $minPts$, ϵ and the total amount of data points. This would reduce the number of modes to be tested to a few instead of 100 or more. Secondly, although the used t-SNE implementation, FIt-SNE, served our purposes well, real surveys manage amounts of data that are several times the size of the data-set used in this work, and will only grow in size in the future. For that reason, we believe that the next step could be to use a GPU accelerated implementation of t-SNE, already available in Python as a package called TSNE-CUDA (Chan et al. 2018). This accelerated implementation promises performance improvements up to 4500x when compared to the standard t-SNE implementation or up to 18x when compared to FIt-SNE.

6.4 Best practices

For a great starting point when analyzing a spectral sample of an amount and wavelength range comparable to the one used in this work, we suggest the usage of the following parameters:

- Spectral region: we have showed that the regions between 450 and 550 nm present the best and most stable recovery figures, even under high levels of noise. For this reason, we suggest their usage in a comparable analysis.
- Perplexity: although the default range in the literature suggests a value between 5 and 50, for the purposes of binary star detection, we suggest values between 15 and 30, which we believe yield the best results.
- Signal-to-noise ratio: a high SNR regime is not a necessity and could even be disadvantageous under some parameter combinations. We have proven that spectra with SNR values between 50 and 100 are ideal for SB2 spectra identification.
- DBSCAN modes: even if the values for ϵ and $minPts$ are dependent on the density of each individual cluster found in the examined t-SNE projection, for a properly scaled t-SNE projection we suggest the usage of values $\epsilon \in [0.2, 0.3]$ and $minPts$ around 1% of the total amount of binaries in the data-set. We recommend that for a given dataset, the t-SNE

projection axes from this work ([-10, 10]) should be scaled by a factor $\sqrt{F}$, where the total number of data-points in a given data-set is equal to by $N = F \cdot 10^5$, where 10^5 is the number of data-points used in this work.

Analysis of unexplored data

The end goal of the method developed in this work is to extract binary stars from unexplored spectral data. One should, however, consider that when applying this method to real data, the number of recovered binaries could be considerably lower. Real spectra are plagued with problems and issues from the measurement and data reduction: continuum normalization issues, telluric lines, wavelength-dependent noise, wavelength calibration issues etc.

This analysis on real data with the method presented in this work could be complemented or upgraded by inserting artificially created binary spectra based on measured single-star spectra found in the data-set to serve as tracers. Using the study we performed on the individual binaries presented in section 5.4, we showed which binary pairs are recovered almost under any circumstance and which are those that mark the limit of the detectability ranges for the chosen variable parameters (spectral range, perplexity and SNR). Furthermore, we can avoid simulating any kind of issues and defects on the spectra by simply creating artificial SB2 spectra through combination of carefully selected real single star spectra contained in the data-set to be studied. This automatically introduces the imprint of observations and instrumentation on the created binary spectra and allow for them to be very similar to the real ones in the data-set. Because their nature is known, they can be easily traced in t-SNE projection and therefore one would, in the ideal case, be certain that the stars located around the tracer binaries in the corresponding well-isolated islands (or even within the main single star islands if chosen appropriately) are indeed binary stars as well.

6.5 Conclusions

This work presented a method capable of detecting and automatically identifying binary stars from a sample of stellar spectra by using a combination of two state-of-the-art machine learning algorithms: the dimensionality reduction technique t-SNE and density-based clustering algorithm DBSCAN. We selected a sub-sample of only dwarf stars from GALAH DR2, who were further assumed to be single, unevolved, main-sequence stars. Basing our data-set on the stars from the GALAH survey, allowed us to synthesize stellar spectra with realistic stellar configurations. The final synthetic spectroscopic sample we generated is comprised of 100000 single stars and 5000 binaries (whose spectra were constructed by combining single-star spectra). The generated synthetic spectroscopic survey was used to test and optimize our identification method and the reason behind the usage of self-generated data in this work is that, because the single/binary nature of each synthesized spectrum was known a priori, contrary to a real spectroscopic survey. Through variation and testing of 360 different combinations of machine learning parameters, signal-to-noise ratio and the spectral range, we investigated their effects on the t-SNE projections and subsequent automatic identification performed by DBSCAN on our synthetic data-set. Furthermore, we performed a secondary optimization procedure through a self-designed algorithm to maximize the recovery ratio of binary stars in each of the examined t-SNE projections, which allowed us to overcome one of the major difficulties regarding the usage of DBSCAN.

Using a quality measure of our method in the form of the ratio of recovered binary stars we were able to determine how the different parameters influence the amount and type of successfully identified binary stars. We conclude that an optimal combination of perplexity, SNR and DBSCAN mode can be found for each of the studied spectral ranges such that they always yield

a recovery of $\sim 70\%$ or higher, with a promising average absolute recovery of binary systems of 57%. Furthermore, while we showed that the best perplexity values lie within the range recommended in the literature, we found that that is not the case for the noise, as a moderate amount of it (corresponding to SNR between 50 and 100) can have a beneficial effect in the detection of binarity especially in the bluest spectral regions with higher abundance of spectral lines.

We also showed that spectral ranges lying blueward of the H-α line present a higher stability against the variation of the parameters used in this method, with small variations in the yielded recovery. This suggests a benefit of using bluer spectral regions for binary detection regardless of the other parameters used.

In our analysis, we saw that those binaries that get more easily avoid detection are those that present heavily blended lines and/or whose secondary spectrum is concealed by noise due to low values of radial velocity difference and luminosity ratio, respectively. Furthermore, we studied the effect of the most important parameters on the individual binary systems, for which we found that systems of twins with presenting stellar parameters corresponding to a sub-solar mass for $T_{\rm eff}$ and $log\, g$ were recovered in more than 90% of the simulations.

Based on our results, we provided a list of recommended starting values for a similar analysis and suggestions on how to apply our method to a real spectroscopic survey. Moreover, we give in Appendix A a list of 40 binaries that were successfully detected in 90% or more of our simulations to serve as suggestions for possible tracers when using our approach for detection of binary stars in spectroscopic data-sets.

References

[1] Abt, Helmut A and Levy, Saul G. "Multiplicity among solar-type stars". In: *The Astrophysical Journal Supplement Series* 30 (1976), pp. 273–306.

[2] Ahumada, Romina et al. "The Sixteenth Data Release of the Sloan Digital Sky Surveys: First Release from the APOGEE-2 Southern Survey and Full Release of eBOSS Spectra". In: *arXiv preprint arXiv:1912.02905* (2019).

[3] Anders, Friedrich et al. "Dissecting stellar chemical abundance space with t-SNE". In: *Astronomy & Astrophysics* 619 (2018), A125.

[4] Andersen, J. "Accurate masses and radii of normal stars". In: 3.2 (Jan. 1991), pp. 91–126. DOI: 10.1007/BF00873538.

[5] Armstrong, DJ et al. "K2 variable catalogue–II. Machine learning classification of variable stars and eclipsing binaries in K2 fields 0–4". In: *Monthly Notices of the Royal Astronomical Society* 456.2 (2015), pp. 2260–2272.

[6] Asplund, Martin et al. "GES linelist V5". In: (2013).

[7] Asplund, Martin, Grevesse, Nicolas, Sauval, A Jacques, and Scott, Pat. "The chemical composition of the Sun". In: *Annual Review of Astronomy and Astrophysics* 47 (2009), pp. 481–522.

[8] Ayodele, Taiwo Oladipupo. "Types of machine learning algorithms". In: *New advances in machine learning* (2010), pp. 19–48.

[9] Baron, Dalya. "Machine learning in astronomy: A practical overview". In: *arXiv preprint arXiv:1904.07248* (2019).

[10] Beccari, Giacomo and Boffin, Henri MJ. *The Impact of Binary Stars on Stellar Evolution*. Vol. 54. Cambridge University Press, 2019.

[11] Becht, Etienne et al. "Evaluation of UMAP as an alternative to t-SNE for single-cell data". In: *BioRxiv* (2018), p. 298430.

[12] Bessel, Friedrich Wilhelm. "On the variations of the proper motions of Procyon and Sirius". In: *Monthly Notices of the Royal Astronomical Society* 6 (1844), pp. 136–141.

[13] Blanco-Cuaresma, S., Soubiran, C., Heiter, U., and Jofré, P. "Determining stellar atmospheric parameters and chemical abundances of FGK stars with iSpec". In: *Astronomy & Astrophysics* 569 (Sept. 2014), A111. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201423945. URL: http://dx.doi.org/10.1051/0004-6361/201423945.

[14] Blanco-Cuaresma, Sergi. "Modern stellar spectroscopy caveats". In: *Monthly Notices of the Royal Astronomical Society* 486.2 (2019), pp. 2075–2101.

[15] Blanton, Michael R et al. "Sloan digital sky survey IV: Mapping the Milky Way, nearby galaxies, and the distant universe". In: *The Astronomical Journal* 154.1 (2017), p. 28.

[16] Breivik, Katelyn et al. "Stellar multiplicity: an interdisciplinary nexus". In: (Mar. 2019). eprint: 1903.05094. URL: https://arxiv.org/pdf/1903.05094.pdf.

[17] Brown, A. G. A. et al. "Gaia Data Release 2". In: *Astronomy & Astrophysics* 616 (Aug. 2018), A1. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201833051. URL: http://dx.doi.org/10.1051/0004-6361/201833051.

[18] Buder, S. et al. "The GALAH Survey: Second Data Release". In: (Apr. 2018). DOI: 10.1093/mnras/sty1281. eprint: 1804.06041. URL: https://arxiv.org/pdf/1804.06041.pdf.

[19] Campbell, William Wallace and Curtis, Heber Doust. "First catalogue of spectroscopic binaries". In: *Lick Observatory Bulletin* 3 (1905), pp. 136–146.

[20] Cao, Yanshuai and Wang, Luyu. "Automatic selection of t-SNE Perplexity". In: *arXiv preprint arXiv:1708.03229* (2017).

[21] Carling, Ellen B and Kopal, Zdenek. *Photometric and Spectroscopic Binary Systems: Proceedings of the NATO Advanced Study Institute held at Maratea, Italy, June 1–14, 1980*. Vol. 69. Springer Science & Business Media, 2012.

[22] Chabrier, Gilles. "Galactic stellar and substellar initial mass function". In: *Publications of the Astronomical Society of the Pacific* 115.809 (2003), p. 763.

[23] Chan, David M, Rao, Roshan, Huang, Forrest, and Canny, John F. "t-SNE-CUDA: GPU-Accelerated t-SNE and its Applications to Modern Data". In: *2018 30th International Symposium on Computer Architecture and High Performance Computing (SBAC-PAD)*. IEEE. 2018, pp. 330–338.

[24] Ciardi, David R, Beichman, Charles A, Horch, Elliott P, and Howell, Steve B. "Understanding the effects of stellar multiplicity on the derived planet radii from transit surveys: implications for Kepler, K2, and TESS". In: *The Astrophysical Journal* 805.1 (2015), p. 16.

[25] Cotar, Klemen et al. "The GALAH survey: unresolved triple Sun-like stars discovered by the Gaia mission". In: (Apr. 2019). DOI: 10.1093/mnras/stz1397. eprint: 1904.04841. URL: https://arxiv.org/pdf/1904.04841.pdf.

[26] De Jong, Roelof S et al. "4MOST: 4-metre multi-object spectroscopic telescope". In: *Ground-based and Airborne Instrumentation for Astronomy IV*. Vol. 8446. International Society for Optics and Photonics. 2012, 84460T.

[27] Demircan, Osman and Kahraman, Goksel. "Stellar Mass / Luminosity and Mass / Radius Relations". In: 181.2 (July 1991), pp. 313–322. DOI: 10.1007/BF00639097.

[28] Doppler, Christian. *Über das farbige Licht der Doppelsterne und einiger anderer Gestirne des Himmels*. 1842.

[29] Dorn-Wallenstein, Trevor Z and Levesque, Emily M. "Stellar Population Diagnostics of the Massive Star Binary Fraction". In: *The Astrophysical Journal* 867.2 (2018), p. 125.

[30] Ducati, Jorge Ricardo, Penteado, Eduardo Monfardini, and Turcati, Rodrigo. "The mass ratio and initial mass functions in spectroscopic binaries". In: *Astronomy & Astrophysics* 525 (2011), A26.

[31] Duchêne, Gaspard and Kraus, Adam. "Stellar Multiplicity". In: (Mar. 2013). DOI: 10.1146/annurev-astro-081710-102602. eprint: 1303.3028. URL: https://arxiv.org/pdf/1303.3028.pdf.

[32] Duquennoy, A. and Mayor, M. "Multiplicity among solar-type stars in the solar neighbourhood. II - Distribution of the orbital elements in an unbiased sample." In: 500 (Aug. 1991), pp. 337–376.

[33] Eddington, A. S. *The Internal Constitution of the Stars*. 1926.

[34] Eggleton, Peter. *Evolutionary processes in binary and multiple stars*. Vol. 40. Cambridge University Press, 2006.

[35] Eker, Z. et al. "Interrelated Main-Sequence Mass-Luminosity, Mass-Radius and Mass-Effective Temperature Relations". In: (July 2018). DOI: 10.1093/mnras/sty1834. eprint: 1807.02568. URL: https://arxiv.org/pdf/1807.02568.pdf.

[36] Eker, Z. et al. "Main-Sequence Effective Temperatures from a Revised Mass-Luminosity Relation Based on Accurate Properties". In: (Jan. 2015). DOI: 10.1088/0004-6256/149/4/131. eprint: 1501.06585. URL: https://arxiv.org/pdf/1501.06585.pdf.

[37] Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, Xu, Xiaowei, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise." In: *Kdd*. Vol. 96. 34. 1996, pp. 226–231.

[38] Francis, Paul J and Wills, Beverley J. "Introduction to principal components analysis". In: *arXiv preprint astro-ph/9905079* (1999).

[39] Fukunaga, Keinosuke and Hostetler, Larry. "The estimation of the gradient of a density function, with applications in pattern recognition". In: *IEEE Transactions on information theory* 21.1 (1975), pp. 32–40.

[40] Gaia-Collaboration et al. "The Gaia mission". In: *arXiv preprint arXiv:1609.04153* (2016).

[41] Gaia-Collaboration et al. "Gaia Data Release 2 Summary of the contents and survey properties". In: *Astronomy & Astrophysics* 616.1 (2018).

[42] Gilmore, Gerry et al. "The Gaia-ESO public spectroscopic survey". In: *The Messenger* 147 (2012), pp. 25–31.

[43] Goldberg, Dorit, Mazeh, Tsevi, and Latham, David W. "On the mass-ratio distribution of spectroscopic binaries". In: *The Astrophysical Journal* 591.1 (2003), p. 397.

[44] Gustafsson, B. et al. "A grid of MARCS model atmospheres for late-type stars". In: *Astronomy & Astrophysics* 486.3 (May 2008), pp. 951–970. ISSN: 1432-0746. DOI: 10.1051/0004-6361:200809724. URL: http://dx.doi.org/10.1051/0004-6361:200809724.

[45] Hertzsprung, Ejnar et al. “On the relation between mass and absolute brightness of components of double stars”. In: *Bulletin of the Astronomical Institutes of the Netherlands* 2 (1923), p. 15.

[46] Hinton, Geoffrey E and Roweis, Sam T. “Stochastic neighbor embedding”. In: *Advances in neural information processing systems*. 2003, pp. 857–864.

[47] Hogeveen, Sake Jogchum et al. “The mass-ratio distribution of binary stars”. PhD thesis. Sterrenkundig Instituut'Anton Pannekoek', 1991.

[48] Howell, Steve B et al. “The K2 mission: characterization and early results”. In: *Publications of the Astronomical Society of the Pacific* 126.938 (2014), p. 398.

[49] Husser, T-O et al. “A new extensive library of PHOENIX stellar atmospheres and synthetic spectra”. In: *Astronomy & Astrophysics* 553 (2013), A6.

[50] Jofré, P. et al. “Climbing the cosmic ladder with stellar twins in RAVE with Gaia”. In: *Monthly Notices of the Royal Astronomical Society* 472.3 (Aug. 2017), pp. 2517–2533. ISSN: 1365-2966. DOI: 10.1093/mnras/stx1877. URL: http://dx.doi.org/10.1093/mnras/stx1877.

[51] Katoh, Noriyuki, Itoh, Yoichi, Toyota, Eri, and Sato, Bun'ei. “Determination of Orbital Elements of Spectroscopic Binaries Using High-dispersion Spectroscopy”. In: *The Astronomical Journal* 145.2 (2013), p. 41.

[52] Kos, Janez et al. “The GALAH survey: Chemical Tagging of Star Clusters and New Members in the Pleiades”. In: (Sept. 2017). DOI: 10.1093/mnras/stx2637. eprint: 1709.00794. URL: https://arxiv.org/pdf/1709.00794.pdf.

[53] Kounkel, Marina et al. “Close companions around young stars”. In: *The Astronomical Journal* 157.5 (2019), p. 196.

[54] Kouwenhoven, M. B. N. et al. “Exploring the consequences of pairing algorithms for binary stars”. In: (Nov. 2008). DOI: 10.1051/0004-6361:200810234. eprint: 0811.2859. URL: https://arxiv.org/pdf/0811.2859.pdf.

[55] Kriegel, Hans-Peter, Kröger, Peer, Sander, Jörg, and Zimek, Arthur. “Density-based clustering”. In: *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 1.3 (2011), pp. 231–240.

[56] Kroupa, Pavel. “On the variation of the initial mass function”. In: *Monthly Notices of the Royal Astronomical Society* 322.2 (2001), pp. 231–246.

[57] Kuiper, GP. “Problems of double-star astronomy. II”. In: *Publications of the Astronomical Society of the Pacific* 47.277 (1935), pp. 121–150.

[58] Kullback, Solomon and Leibler, Richard A. “On information and sufficiency”. In: *The annals of mathematical statistics* 22.1 (1951), pp. 79–86.

[59] Lada, Charles J. “Stellar multiplicity and the initial mass function: most stars are single”. In: *The Astrophysical Journal Letters* 640.1 (2006), p. L63.

[60] LeCun, Yann, Cortes, Corinna, and Burges, CJ. “MNIST handwritten digit database”. In: (2010).

[61] Linderman, George C. et al. “Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data”. In: *Nature Methods* 16.3 (Feb. 2019), pp. 243–245. ISSN: 1548-7105. DOI: 10.1038/s41592-018-0308-4. URL: http://dx.doi.org/10.1038/s41592-018-0308-4.

[62] Linderman and Steinerberger. “Clustering with t-SNE, provably”. In: *SIAM Journal on Mathematics of Data Science* 1.2 (2019), pp. 313–332.

[63] Liu, Cheng-Lin, Nakashima, Kazuki, Sako, Hiroshi, and Fujisawa, Hiromichi. “Handwritten digit recognition: benchmarking of state-of-the-art techniques”. In: *Pattern recognition* 36.10 (2003), pp. 2271–2285.

[64] Liu, Peng, Zhou, Dong, and Wu, Naijun. “VDBSCAN: varied density based spatial clustering of applications with noise”. In: *2007 International conference on service systems and service management*. IEEE. 2007, pp. 1–4.

[65] Lochner, Michelle et al. “Photometric supernova classification with machine learning”. In: *The Astrophysical Journal Supplement Series* 225.2 (2016), p. 31.

[66] Maaten, Laurens van der and Hinton, Geoffrey. “Visualizing data using t-SNE”. In: *Journal of machine learning research* 9.Nov (2008), pp. 2579–2605.

[67] Mathieu, Robert D. “Pre-main-sequence binary stars”. In: *Annual Review of Astronomy and Astrophysics* 32.1 (1994), pp. 465–530.

[68] Matijevic, G. et al. “Double-lined Spectroscopic Binary Stars in the Radial Velocity Experiment Survey”. In: (June 2010). DOI: 10.1088/0004-6256/140/1/184. eprint: 1006.2517. URL: https://arxiv.org/pdf/1006.2517.pdf.

[69] Mazeh, T et al. "The Mass Ratio Distribution in Main-Sequence Spectroscopic Binaries Measured by Infrared Spectroscopy". In: *The Astrophysical Journal* 599.2 (2003), p. 1344.

[70] McInnes, Leland, Healy, John, and Melville, James. "Umap: Uniform manifold approximation and projection for dimension reduction". In: *arXiv preprint arXiv:1802.03426* (2018).

[71] Merle, Thibault et al. "The Gaia-ESO Survey: detection and characterisation of single-line spectroscopic binaries". In: *Astronomy & Astrophysics* 635 (2020), A155.

[72] Moya, Andy, Zuccarino, Federico, Chaplin, William J, and Davies, Guy R. "Empirical relations for the accurate estimation of stellar masses and radii". In: *The Astrophysical Journal Supplement Series* 237.2 (2018), p. 21.

[73] Ness, Melissa et al. "The Cannon: A data-driven approach to stellar label determination". In: *The Astrophysical Journal* 808.1 (2015), p. 16.

[74] Offner, Stella SR et al. "The turbulent origin of outflow and spin misalignment in multiple star systems". In: *The Astrophysical Journal Letters* 827.1 (2016), p. L11.

[75] Pesenson, Meyer Z, Pesenson, Isaac Z, and McCollum, Bruce. "The data big bang and the expanding digital universe: high-dimensional, complex and massive data sets in an inflationary epoch". In: *Advances in Astronomy* 2010 (2010).

[76] Pickering, Edward C. "On the spectrum of zeta Ursae Majoris". In: *The Observatory* 13 (1890), pp. 80–81.

[77] Piskunov, Nikolai and Valenti, Jeff A. "Spectroscopy Made Easy: Evolution". In: *Astronomy & Astrophysics* 597 (2017), A16.

[78] Plaut, Lukas. "An Investigation of the Eclipsing Binaries Brighter than Photographic Magnitude 8.50 at Maximum". In: *Publications of the Kapteyn Astronomical Laboratory Groningen* 55 (1953), pp. 1–62.

[79] Plez, B. "Turbospectrum: Code for spectral synthesis". In: *Astrophysics Source Code Library* (2012).

[80] Raghavan, Deepak et al. "A survey of stellar families: multiplicity of solar-type stars". In: *The Astrophysical Journal Supplement Series* 190.1 (2010), p. 1.

[81] Raskin, Gert et al. "HERMES: a high-resolution fibre-fed spectrograph for the Mercator telescope". In: *Astronomy & Astrophysics* 526 (2011), A69.

[82] Reipurth, Bo et al. "Multiplicity in early stellar evolution". In: *Protostars and Planets VI* (2014), pp. 267–290.

[83] Ruchti, GR et al. "A new algorithm for optimizing the wavelength coverage for spectroscopic studies: Spectral Wavelength Optimization Code (SWOC)". In: *Monthly Notices of the Royal Astronomical Society* 461.2 (2016), pp. 2174–2191.

[84] Rucinski, Slavek M. "Eclipsing Binaries in the OGLE Variable Star Catalog. I. W UMa-type Systems as Distance and Population Tracers in Baade's Window". In: *arXiv preprint astro-ph/9607009* (1996).

[85] Russell, HN, Adams, WS, and Joy, AH. "A Comparison of Spectroscopic and Dynamical Parallaxes". In: *Publications of the Astronomical Society of the Pacific* 35.206 (1923), pp. 189–193.

[86] Sadavoy, Sarah I and Stahler, Steven W. "Embedded binaries and their dense cores". In: *Monthly Notices of the Royal Astronomical Society* 469.4 (2017), pp. 3881–3900.

[87] Salpeter, Edwin E. "The luminosity function and stellar evolution." In: *The Astrophysical Journal* 121 (1955), p. 161.

[88] Samuel, Arthur L. "Some studies in machine learning using the game of checkers". In: *IBM Journal of research and development* 3.3 (1959), pp. 210–229.

[89] Saul, Lawrence K and Roweis, Sam T. "An introduction to locally linear embedding". In: *unpublished. Available at: http://www. cs. toronto. edu/~ roweis/lle/publications. html* (2000).

[90] Schwarzschild, K. "On the equilibrium of the sun's atmosphere". In: *Nachrichten von der Königlichen Gesellschaft der Wissenschaften zu Göttingen. Math.-phys. Klasse, 195, p. 41-53* 195 (1906), pp. 41–53.

[91] Shallue, Christopher J and Vanderburg, Andrew. "Identifying exoplanets with deep learning: A five-planet resonant chain around kepler-80 and an eighth planet around kepler-90". In: *The Astronomical Journal* 155.2 (2018), p. 94.

[92] Shou-kun, Xu, Chao, Wang, Li-hua, Zhuang, and Xin-hua, Gao. "DBSCAN Clustering Algorithm for the Detection of Nearby Open Clusters Based on Gaia-DR2two". In: *Chinese Astronomy and Astrophysics* 43.2 (2019), pp. 225–236. ISSN: 0275-1062. DOI: https://doi.org/10.1016/j.chinastron.2019.04.001. URL: http://www.sciencedirect.com/science/article/pii/S0275106219300244.

[93] Southworth, John. "Binary stars: a cheat sheet". In: *arXiv preprint arXiv:1912.13400* (2019).

[94] Spencer, Meghin E et al. "The Binary Fraction of Stars in Dwarf Galaxies: The Cases of Draco and Ursa Minor". In: *The Astronomical Journal* 156.6 (2018), p. 257.

[95] Steinmetz, Matthias et al. "The radial velocity experiment (RAVE): first data release". In: *The Astronomical Journal* 132.4 (2006), p. 1645.

[96] Süveges, M et al. "Gaia eclipsing binary and multiple systems. Supervised classification and self-organizing maps". In: *Astronomy & Astrophysics* 603 (2017), A117.

[97] Tenenbaum, Joshua B, De Silva, Vin, and Langford, John C. "A global geometric framework for nonlinear dimensionality reduction". In: *science* 290.5500 (2000), pp. 2319–2323.

[98] Tobin, John J et al. "The VLA nascent disk and multiplicity survey of perseus protostars (VANDAM). II. Multiplicity of protostars in the perseus molecular cloud". In: *The Astrophysical Journal* 818.1 (2016), p. 73.

[99] Tohline, Joel E. "The origin of binary stars". In: *Annual Review of Astronomy and Astrophysics* 40.1 (2002), pp. 349–385.

[100] Tramacere, A. and Vecchio, C. "γ-ray DBSCAN: a clustering algorithm applied toFermi-LATγ-ray data". In: *Astronomy & Astrophysics* 549 (Jan. 2013), A138. ISSN: 1432-0746. DOI: 10.1051/0004-6361/201220133. URL: http://dx.doi.org/10.1051/0004-6361/201220133.

[101] Traven et al. "The Galah Survey: Classification and diagnostics with t-SNE reduction of spectral information". In: (Dec. 2016). DOI: 10.3847/1538-4365/228/2/24. eprint: 1612.02242. URL: https://arxiv.org/pdf/1612.02242.pdf.

[102] Traven et al. "The GALAH survey: Multiple stars and our Galaxy. I. A comprehensive method for deriving properties of FGK binary stars. (in prep.)" In: *Astronomy & Astrophysics* (2020).

[103] Valenti, Jeff A and Piskunov, Nikolai. "Spectroscopy made easy: A new tool for fitting observations with synthetic spectra". In: *Astronomy and Astrophysics Supplement Series* 118.3 (1996), pp. 595–603.

[104] Valentini, M et al. "RAVE stars in K2-I. Improving RAVE red giants spectroscopy using asteroseismology from K2 Campaign 1". In: *Astronomy and Astrophysics* 600 (2017), A66.

[105] Van Der Maaten, Laurens. "Accelerating t-SNE using tree-based algorithms". In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 3221–3245.

[106] Van der Swaelmen, M, Merle, T, Van Eck, S, and Jorissen, A. "Spectroscopic binaries with Gaia and large spectroscopic surveys". In: *sf2a* (2019), p. Di.

[107] Wattenberg, FVM and Johnson, I. "How to use t-SNE effectively". In: *distill.pub/2016/misread-tsne* (2016).

[108] Welvaert, Marijke and Rosseel, Yves. "On the definition of signal-to-noise ratio and contrast-to-noise ratio for fMRI data". In: *PloS one* 8.11 (2013).

[109] Xu, Dongkuan and Tian, Yingjie. "A comprehensive survey of clustering algorithms". In: *Annals of Data Science* 2.2 (2015), pp. 165–193.

[110] Zwitter et al. "The GALAH survey: accurate radial velocities and library of observed stellar template spectra". In: *Monthly Notices of the Royal Astronomical Society* 481.1 (Aug. 2018), pp. 645–654. ISSN: 1365-2966. DOI: 10.1093/mnras/sty2293. URL: http://dx.doi.org/10.1093/mnras/sty2293.

Appendix A

Bona fide binaries

In this Appendix we present a table containing the parameters belonging to a sub-sample of 40 binaries from those that were recovered on 90% of the carried analysis or more. We hope that these bona fide binary systems can serve as a guide for the investigation of binarity in future spectral surveys because as we have shown, the stars presented in table A.1 can be identified by the method described in this work under almost any circumstance.

The given parameters are, in order: the total amount of times the given system was recovered, the effective temperature of the primary and secondary star, the surface gravity of the primary and secondary, the average metallicity of the system, the masses of the primary and secondary components, the mass ratio of the system and the radial velocity difference.

Recoveries	$T_{\mathrm{eff},A}$	$T_{\mathrm{eff},B}$	$log\, g_A$	$log\, g_B$	[Fe/H]	M_A	M_B	q	Δv_{rad}	L_B/L_A
341	4615.43	4609.46	4.49	4.56	0.08	0.53	0.53	1.0	20.31	1.0
346	4650.33	4613.36	4.4	4.62	-0.04	0.54	0.53	0.98	40.2	0.92
344	4676.69	4651.03	4.58	4.69	-0.06	0.57	0.54	0.95	-17.87	0.8
338	4706.82	4634.72	4.46	4.61	0.15	0.57	0.54	0.94	-12.63	0.77
346	4720.22	4635.04	4.63	4.64	0.01	0.57	0.56	0.98	41.47	0.91
336	4738.58	4666.63	4.65	4.67	0.04	0.59	0.55	0.92	64.78	0.69
347	4754.69	4628.1	4.35	4.64	-0.01	0.58	0.55	0.95	28.2	0.8
344	4772.53	4745.07	4.7	4.71	0.01	0.58	0.58	0.99	-17.3	0.96
342	4787.51	4640.41	4.56	4.65	0.1	0.63	0.56	0.9	52.08	0.61
340	4802.85	4662.61	4.6	4.7	-0.04	0.61	0.58	0.95	-19.11	0.78
345	4821.33	4709.23	4.56	4.57	0.19	0.61	0.6	1.0	-13.9	0.99
344	4829.79	4760.62	4.55	4.57	0.05	0.62	0.58	0.94	45.92	0.74
344	4842.09	4680.72	4.59	4.61	0.05	0.62	0.57	0.91	25.66	0.67
347	4855.82	4763.21	4.48	4.76	-0.03	0.65	0.6	0.92	-36.34	0.68
350	4863.69	4845.76	4.62	4.67	0.09	0.65	0.65	1.0	20.37	0.99
336	4879.74	4696.59	4.57	4.72	0.11	0.64	0.57	0.89	15.32	0.58
352	4896.55	4870.3	4.59	4.66	0.11	0.65	0.64	0.98	-31.57	0.92
337	4911.40	4785.0	4.71	4.73	0.13	0.65	0.6	0.93	-24.38	0.73
336	4921.55	4875.57	4.46	4.54	0.14	0.64	0.62	0.96	11.87	0.84
336	4930.10	4770.48	4.75	5.02	0.15	0.64	0.6	0.94	-18.62	0.76
336	4946.34	4713.78	4.61	4.71	-0.15	0.65	0.58	0.9	41.02	0.61
338	4959.97	4765.1	4.6	4.6	0.04	0.69	0.61	0.88	-27.21	0.57
347	4968.62	4884.65	4.51	4.55	0.15	0.67	0.62	0.92	-39.05	0.69
341	4983.76	4934.27	4.63	4.68	0.23	0.7	0.64	0.92	-23.62	0.69
345	5004.17	4870.3	4.57	4.66	0.12	0.68	0.65	0.96	-39.64	0.84
349	5020.97	4889.46	4.43	4.58	0.15	0.7	0.65	0.93	-38.84	0.71
345	5036.75	4682.72	4.7	4.76	0.12	0.68	0.59	0.86	32.73	0.51
349	5051.39	4887.56	4.41	4.51	0.19	0.7	0.65	0.93	40.11	0.72
336	5063.18	4719.14	4.64	4.65	-0.12	0.69	0.59	0.86	48.96	0.51
336	5080.18	4918.84	4.28	4.74	0.15	0.73	0.67	0.91	-13.95	0.67
339	5096.98	5080.52	4.57	4.57	0.17	0.75	0.73	0.98	-12.68	0.91
347	5109.97	5013.21	4.65	4.75	0.44	0.74	0.71	0.97	-32.19	0.86
345	5127.10	4878.08	4.47	4.77	0.06	0.71	0.63	0.89	31.53	0.58
347	5142.68	5108.14	4.49	4.68	0.28	0.76	0.71	0.94	37.16	0.75
336	5156.28	4814.42	4.46	4.53	0.39	0.72	0.62	0.85	48.38	0.48
342	5193.73	4983.34	4.44	4.62	0.18	0.75	0.68	0.91	50.83	0.65
349	5222.96	5130.48	4.41	4.48	0.25	0.76	0.76	1.0	32.68	1.0
339	5268.27	5224.45	4.52	4.59	0.31	0.8	0.76	0.95	-36.77	0.79
337	5332.19	5224.56	4.59	4.79	0.19	0.81	0.74	0.91	14.04	0.67
337	5572.42	5455.86	4.3	4.39	0.46	0.89	0.88	0.98	-32.01	0.93

Table A.1: Stellar parameters of a sub-sample of 40 binary systems that were recovered in 90% or more of our simulations.