UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Sede Amministrativa: Università degli Studi di Padova

Dipartimento di Scienze Statistiche
Corso di Dottorato di Ricerca in Scienze Statistiche
Ciclo XXX

Finite Dirichlet mixture models for classification and detection of new classes of variable stars

Coordinatore del Corso: Prof. Nicola Sartori

Supervisore: Prof. Alessandra R. Brazzale

Co-supervisore: Prof. Maria Süveges

Dottorando/a: Prince John

31 October 2017

“And we know that God causes everything to work together for the good of those who love God and are called according to his purpose for them. ”

Romans 8:28

Abstract

The data that is being acquired by the Gaia space mission will allow us to compile a catalog of one billion stars. In the backdrop of this huge influx of data, it is crucial to have an efficient classification model. The aim of this thesis is, in particular, to develop appropriate models for the classification of variable stars based on the data that will be provided by the Gaia space survey.

The first contribution of the thesis is the development of a two-stage classification model, the Two Stage Dirichlet Mixture model (TSDM), based on finite mixtures of Dirichlet distributions. We validated this model on a well-studied subgroup of variable stars in the Hipparcos catalog analogously to what done by Dubath et al. (2011). We also propose two different transformations of the attributes used for the classification, which allow us to use the Dirichlet distribution whose support is a simplex. The adequacy of these transformations was evaluated with the selected data, highlighting an ability to correctly classify variable stars of 69.3%.

Secondly, we introduced an extension of the TSDM model, called the fixed backdrop (FB) model, whose purpose is to identify new variable star classes. Our proposal is based on the semi-supervised classification model developed by Vatanen et al. (2012) for the identification of anomalies. The FB model, in particular, combines the TSDM model, used to represent the already known classes (the so-called background), with a finite mixture of Dirichlet distributions which represent the new class. We have looked at the proposed model assuming a scenario in which the β Cephei (BCEP) class is the anomaly, achieving a sensitivity of 77%.

The third contribution of the thesis is the feasibility study for a Bayesian supervised variable stars classification using finite mixture of Dirichlet distributions. In particular, we propose a possible a priori conjugate distribution to the model.

Sommario

I dati che saranno acquisiti dalla missione spaziale Gaia consentiranno di compilare un catalogo contenente circa un miliardo di stelle. Alla luce di questo enorme afflusso di dati, è cruciale poter disporre di un modello di classificazione efficiente. L'obiettivo di questa tesi, in particolare, è sviluppare dei modelli adeguati per la classificazione delle stelle variabili in base ai dati che saranno forniti dalla missione spaziale Gaia.

Il primo contributo della tesi è lo sviluppo di un modello di classificazione a due stadi, detto modello Two Stage Dirichlet Mixture (TSDM), basato su delle misture finite di distribuzioni Dirichlet. Abbiamo validato questo modello su un sottogruppo ben studiato di stelle variabili riportate nel catalogo Hipparcos in analogia a quanto fatto da Dubath et al. (2011). Proponiamo, inoltre, due diverse trasformazioni delle caratteristiche utilizzate per la classificazione, che ci consentono di utilizzare per l'appunto la distribuzione di Dirichlet il cui supporto è un simplesso. L'adeguatezza di queste trasformazioni è stata vagliata con i dati selezionati, evidenziando una capacità di corretta classificazione delle stelle variabili considerate del 69.3%.

In secondo luogo, abbiamo introdotto un'estensione del modello TSDM, detta modello a sfondo fisso (FB), il cui scopo è identificare nuove classi di stelle variabili. La nostra proposta si basa sul modello per la classificazione semi supervisionata sviluppato da Vatanen et al. (2012) per l'identificazione di anomalie. Il modello FB, in particolare, combina il modello TSDM, usato per rappresentare le classi già note (il cosiddetto sfondo), con una mistura finita di distribuzioni di Dirichlet che rappresenta la nuova classe. Abbiamo vagliato il modello proposto assumendo uno scenario in cui la classe β Cephei (BCEP) rappresenta l'anomalia, conseguendo una sensibilità del 77%.

il terzo contributo della tesi valuta la fattiblità di una classificazione di stelle Bayesiana supervisionata tramite l'utilizzo di misture di distribuzioni di Dirichlet. In particolare, proponiamo una possibile distribuzione a priori coniugata per il modello.

Acknowledgements

I had the privilege to be cared for and guided by the best and I'm forever grateful. I would like to thank my supervisor, Professor Alessandra R. Brazzale, for her full support and skillful guidance throughout this PhD project. Her patience, encouragement and motivation has helped me immensely to overcome numerous obstacles I faced through my research. Also my co-supervisor, Dr. Maria Süveges for providing valuable insights from astronomy perspective and also guiding me in some key areas in the research, making this interdisciplinary work possible. I'm extremely grateful to them, for taking me, guiding me and nurturing me into a better researcher.

I would like to thank and appreciate the Department of Statistical Sciences for the amazing research experience and especially Prof. Monica Chiogna for the motivation provided during this journey. I would also like to thank Patrizia Piacentini for her continued love and support which made me belong from the day I walked in. The campus and people of Max Planck Institute for Astronomy will be in my heart forever for the love and for welcoming me to their wonderful campus and hosting me for those memorable ten weeks.

My heart is filled with gratitude to those behind the scenes who made this possible. My Dad and Mom, who believed in me and motivated me to pursue my dreams, showering love and always being a cushion I could fall back into. My beautiful wife, Aneesha who was by my side throughout this journey to help me and support me during the ups and downs of my research. Marrying her was the best decision I took in the last 3 years. I also thank Papa, Mummy, Dennis, Darly, Enoch, Blessy, Nitin, Rexon and Moneesha for their love and support throughout the last 3 years.

I thank my wonderful family at International Christian fellowship in Padova who continuously prayed for me and made the last three years in Padova memorable. I also thank my colleagues from the 30th cycle Adil, Leo, Kim, Ehsan, Sifiso, Dan, Umberto and Claudio. I feel blessed to have shared three years with these amazing people.

But most of all, I thank God for His grace throughout this PhD as its by is grace and calling that I am able to complete this thesis.

Contents

I Introduction **1**

1 Introduction **7**
1.1 Preface . . . 7
1.2 Preliminary concepts . . . 8
1.2.1 Mixture distributions . . . 8
1.2.2 Mixture model using categorical random variables . . . 9
1.2.3 Dirichlet distribution . . . 10
1.2.4 Random forests . . . 11
1.3 Chapter summaries . . . 13

II The Astronomy **15**

2 Variable stars **17**
2.1 Introduction . . . 17
2.2 History of variable stars . . . 18
2.3 Preliminaries . . . 19
2.4 Classification of Variable stars . . . 26
2.4.1 Intrinsic variables . . . 27
Pulsating variables . . . 27
Eruptive and cataclysmic variables . . . 31
2.4.2 Extrinsic variables . . . 32
Eclipsing binaries . . . 32
Rotating variables . . . 32
2.5 The Gaia mission . . . 35
2.6 Synopsis . . . 36

III Modeling and methodologies **37**

3 Training data-set **39**
3.1 Introduction . . . 39
3.2 Data sources . . . 40
3.2.1 Hipparcos Catalogue . . . 41
3.2.2 VSX-AAVSO . . . 41
3.3 Data . . . 42
3.4 Class attributes . . . 45

3.4.1 Attribute selection . . . 45
Towards a minimum attribute list . . . 46
Dealing with multicollinearity . . . 51
3.4.2 Selected attributes . . . 52
3.5 Synopsis . . . 52

4 TSDM model **55**
4.1 Introduction . . . 55
4.2 Model . . . 55
4.2.1 First stage . . . 58
Transformation to the probability scale . . . 58
Transformation to the simplex . . . 61
4.2.2 Second stage . . . 67
4.3 Application and Discussion . . . 68
4.3.1 Application Case 1 . . . 69
4.3.2 Application Case 2 . . . 74
4.3.3 Application : Case 3 . . . 77
4.3.4 Comparison with Dubath et al. (2011) . . . 79
4.3.5 Dirichlet vs multivariate Gaussian . . . 81
4.3.6 Are the different clusters sub-classes? . . . 83
4.4 Synopsis . . . 85

5 New class detection **87**
5.1 Overview . . . 87
5.2 FB model . . . 88
5.3 Estimation . . . 90
5.4 Application and discussion . . . 93
5.5 Synopsis . . . 98

6 Towards Bayesian classification **101**
6.1 Overview . . . 101
6.2 Model formulation . . . 101
6.2.1 Mixture of exponential families . . . 102
6.2.2 The case of Dirichlet densities . . . 103
6.3 Conjugate prior . . . 105
6.3.1 Previous work . . . 105
6.3.2 Conjugate prior for the complete data-likelihood . . . 106
6.4 Synopsis and prospectives . . . 107

7 Conclusions and future work **109**
B.1 Expectation-Maximization (EM) algorithm . . . 133
B.2 Exponential family . . . 133
B.3 Empirical distribution . . . 133
B.4 logit and inverse logit transformation . . . 134
B.5 Bayesian Information Criterion (BIC) . . . 135
B.6 Dirichlet distribution in exponential family form . . . 135

List of Figures

0.1 The need for classification methodology illustrated 4

1.1 Density plot illustrating Dirichlet distribution for different parameter values 11

2.1 Variable star light curve for Omicron Ceti star 17
2.2 Absorption and emission spectrum of hydrogen 20
2.3 Hertzsprung-Russell diagram 23
2.4 Hertzsprung-Russell diagram for the data used in the thesis . 25

3.1 Plot to determine the number of attributes to use 46
3.2 Distribution of the most important variable for each of the variable star classes-I 48
3.3 2-D plot of different combinations of attributes 49
3.4 2-D plot of different combinations of attributes 50
3.5 Correlation plot of the selected attributes 51

4.1 Illustration of the stages of a 2-component TSDM model 57
4.2 Empirical distribution of the Log Amplitude fitted to a cubic spline 59
4.3 Attribute data-vector transformed to probability scale 60
4.4 Data-set transformed from probability scale to the simplex using STT1 transformation 62
4.5 Illustration of structure of classes being distorted in the 2-D projection, for STT1 transformation 63
4.6 Data-set transformed from probability scale to the simplex using STT2 transformation 64
4.7 Dirichlet signature vectors plotted against the data for the class LPV 70
4.8 Confusion matrix for case 1 classification 72
4.9 Confusion matrix comparison for the classifications with STT1 and STT2 transformations 75
4.10 Confusion matrix for classification with uncorrelated attributes 78
4.11 Comparison of our case 1 TSDM model with Dubath et al. (2011) 80
4.12 The color index JmK of variable type class SPB in our data-set 83
4.13 Clustering of DSCT-DSCTC combined set 84

5.1 Illustration of the FB model 89
5.2 Division of data for the detection of BCEP as new class 93

5.3 The confusion matrix for the FB model, detecting the "new" BCEP class 94
5.4 Comparison of signatures of the new class BCEP against BCEP and DSCTC data 96
5.5 FB model classification results on 5 classes 97

A.1 Distribution of the important variables for each of the variable star classes-II 115
A.2 Dirichlet signature vectors plotted against the data for each class 123
B.1 Illustration of an empirical distribution function. 134
B.2 Illustration of a logit function. 135

List of Tables

2.1 List of variable star types and their acronyms used in the thesis 34

3.1 Composition of the training set used in this thesis by type . . . 43
3.2 Summary of the data used in this thesis, attribute-wise. We have listed all the 45 attributes from the data and their range, quartiles, median and mean. 44
3.3 Comparison of the attribute lists of this thesis and of Dubath et al. (2011) 53

4.1 Summary of the test and training data-set 69
4.2 Classification results for the TSDM model with 16 correlated attributes and STT1 transformed attributes. 71
4.3 Classification results for the TSDM model with 16 correlated attributes and STT2 transformed attributes. 74
4.4 Classification results for the TSDM model with 8 uncorrelated attributes and STT1 transformed attributes. 77
4.5 Classification accuracy rates : TSDM vs TSGM 82
4.6 Clustering of DSCT and DSCTC into classes and comparing it with the original classes 85

5.1 Classification accuracy of the FB model when detecting the BCEP class. FB model classification accuracy shows the classification accuracy of each of the classes when BCEP was detected as the new class. 95

A.1 The entire list of attributes in the training data-set. There are 45 attributes in the raw data. 113
A.2 Attributes in the training data divided by functionality. 114

List of Abbreviations

ACV	$\alpha - 2$ Canum Venaticorum
ACYG	α Cygni
ASAS	**A**ll **S**ky **A**utomated **S**urvey
BCEP	β Cephei
BE+GCAS	B emission-line star and γ Cassiopeiae
CEP(B)	δ Cepheid multi-mode
CoRoT	**Co**nvection **Ro**tation (and planetary) **T**ransits
CWA	W Virginis
CWB	W Virginis
DCEP	δ Cepheid
DCEPS	δ Cepheid first overtone
DSCT	δ Scuti
DSCTC	δ Scuti low amplitude
EA	Eclipsing binaries
EB	Eclipsing binaries
ELL	Ellipsoidal
ESA	**E**uropean **S**pace **A**gency
EW	Eclipsing binaries
GCVS	**G**eneral **C**atalogue (of) **V**ariable **S**tars
GDOR	γ Doradus
LPV	**L**ong **P**eriod **V**ariables
LSST	(The) **L**arge **S**ynoptic **S**urvey **T**elescope
MACHO	**MA**ssive **C**ompact **H**alo **O**bject
OGLE	**O**ptical **G**ravitational **L**ensing **E**xperiment
OOB	**O**ut **O**f **B**ox
Pan-STARRS	**Pan**oramic **S**urvey **T**elescope **A**nd **R**apid **R**esponse **S**ystem
ROTSE	**R**obotic **O**ptical **T**ransient **S**earch **E**xperiment
RRAB	RR Lyrae AB
RRC	RR Lyrae C
RS+BY	RS Canum Venaticorum and BY Draconis
RV	RV Tauri
SPB	**S**lowly **P**ulsating **B** star
SXARI	SX Arietis

Physical Constants

Speed of Light	$c_0 = 2.997\,924\,58 \times 10^8\,\mathrm{m\,s^{-1}}$ (exact)
1 Kelvin	$\mathrm{K} = -272.15$ Celsius
1 Parsec	pc = 3.26 light years
1 light year	ly = 9460730472580800 metres (exactly)
1 Solar mass	$M_\odot = (1.98855 \pm 0.00025) \times 10^{30}$ kg

Part I

Introduction

Preamble

Overview

Variable stars are stars whose brightness as seen from the Earth fluctuates. In Chapter 2, we see that there are different astrophysical classes of variable stars. But what are the criteria/factors that define these classes?. Understanding this will give us an insight as to how different these classes are, and why they need to be classified.

The fluctuations in brightness are caused by two major factors, namely the physical properties of the star and/or factors external to the star. This gives us an obvious criteria for class definition. However, the different physical processes in the background are not directly observable. The classification must be based on quantities that are straightforwardly measurable. Since we are focusing on the variability of stars, the properties of the light curve of the star, are important criteria for defining classes. Its period, regularity, amplitude and other details about its shape can be represented by numbers and are used in defining classes. But, though this will help in defining some of the stars, other properties also need to be considered to define other types of variable stars. Color and radial velocity curves of the star also help us to understand about the temperature changes or motion. These help in defining classes by providing clues and insight into the physical nature of the variability. Also, there are other factors such as population types which are detailed in Percy (2007). Together, the light curve, color, luminosity and population type have formed the official classification system[1] of variable stars, as defined in edition four of the General Catalogue of Variable Stars (GCVS) (Samus et al. (2017)).

Hence, each of the official classes are different. Each of the classes represent different physical systems and when we classify the variable stars into different classes, we are in effect filtering out different systems that behave differently. But it also needs to be noted that there are overlaps in certain physical properties across classes.

Also, understanding some of the classes of the variable stars have brought significant contributions to the Astronomy which is summarized in Chapter 2 of this thesis. Thus, its really important to classify the variable stars into different classes. There has has recently been a lot of work done in classifying variable stars from different surveys in the past which is mentioned in Chapter 1.

[1] Discussion : http://vizier.u-strasbg.fr/viz-bin/getCatFile?II/214A/./vartype.txt

However, we look into the future. Gaia, which is a survey of the European Space Agency (ESA) has been targeting over 1 billion astronomical objects and we are expecting a surge of data regarding these astronomical objects in the near future. Not surprisingly, many of these will be variable stars. There is a need of a statistical classification methodology that can classify the entire variable star data-set provided by Gaia.

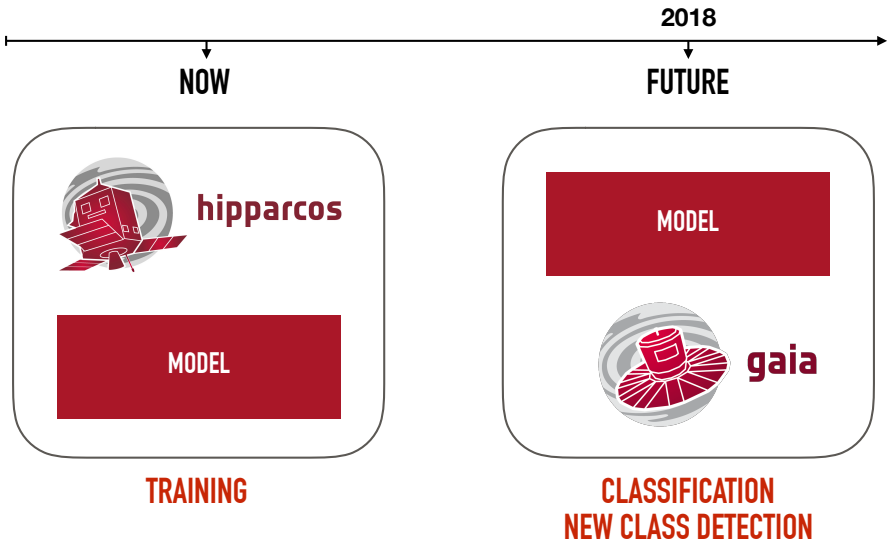


Figure 0.1: (Left) Our model which will be discussed in this thesis (Chapter 4) is trained by the Hipparcos data-set presented in Chapter 3. (Right) The trained model may be used in the future with the Gaia data for classification and new class detection.
Logo Image Credits : ESA

Also, we can never rule out the possibility of new classes in the incoming data. Since the classification technique is based on quantitative properties of stars, there is scope to divide each of the classes into subclasses. However, such sub-classifications and new classifications in some cases must be done cautiously. If we aim for homogeneity in the properties within each class and eventually form newer classes hoping that all the properties of all the stars within these classes will be the same, it can result in a lot of classes with uncertain affiliations. This will also result in many variable stars classes with incomplete information, and misclassifications will become more likely. Figure 0.1 gives an illustration of how

our model (Chapter 4) trained using the Hipparcos data-set (Chapter 3) acts like a "prior" for the classification of data-points in Gaia data-set.

In this thesis, we will present statistical methods which aim at providing substantive help in the classification of variable stars and also the detection of new stars. These methods have been tailored so as to take into account features that are relevant to this aim. It is in this regard that we use Dirichlet distributions for our modeling.

With many surveys collecting data of different stars, there is a possibility that the data will be different across surveys. For instance, the observations of the star Chi CYGNI will be different in Gaia and Hipparcos. Hence there is a need to have a common scale of reference, which is why we use transformations (detailed in Chapter 2) to transform data into probability scale. With the data in the probability scale, we see that Dirichlet density is a natural choice. More about this is discussed in Chapter 4. However let us look into the main contributions of this thesis.

Main Contributions of the thesis

Methodological contributions

- We propose a supervised classification model namely, Two stage Dirichlet mixture model for the classification of variable stars by training the model on the Hipparcos Catalogue (see Chapter 4).
- We propose the use of a semi-supervised classification model, namely Fixed background model, for the detection of new classes of variable stars (see Chapter 5).
- We provide an analysis of the unsupervised classification of select classes of Hipparcos data to a mixture of Dirichlet distribution (see Chapter 6).
- We propose a methodology to fit the data into a simplex, reasonably maintaining the structure of the data.

Astronomical contributions

- We have built a classification model which can classify the upcoming Gaia survey data into predefined classes (see Chapter 4).
- We have built a classification model which can detect any new class in the Gaia data (see Chapter 5).
- Our classification models can suggest the presence of sub-classes within each of the classes.

Chapter 1

Introduction

1.1 Preface

In the previous chapter we discussed about the need of efficient statistical methodologies for the classification of variable star types. The development and implementation of efficient classification schemes is the need of the hour. This is because large surveys are in the process of observing and collecting huge amounts of data from millions, and soon billions, of targets. The time taken to observe the changes in the physical properties of the star range from minutes to even decades, making it impossible to scrutinize them by eye. Hence, it is of prime importance to use powerful statistical and data mining tools. Automated supervised classification methods give predictions on the type or class of an object based on the values of a set of attributes that characterize the object. The dependencies of the type of the object on its attribute values are modeled using a labeled data set, which is a collection of prototype objects of known type and attribute values. After the dependencies are modeled, the types of any other unknown object with available attribute value, can be predicted. As naming schemes differ across publications, we make the following definitions throughout this thesis : *objects* (i.e. stars) are classified into *types* by modeling how the types depend on the attributes.

The classification methods based on finite mixtures of probability densities are simple to use and popular in many applications (see e.g. McLachlan and Peel (2004) for more details on mixture modeling). Mixture models provide a convenient semi-parametric framework which helps to model unknown distributional shapes. It can provide a satisfactory model as it can be useful to model any type of data, due to its flexibility. In this thesis, we have taken advantage of this flexibility of mixture distributions for the classification of variable stars. Variable stars are explained on more detail in Chapter 2 of this thesis.

Variable star classification studies have used the data from surveys such as (1) ASAS (Pojmanski (2002), Pojmanski (2003); Eyer and Blake (2002), Eyer and Blake (2005)), (2) OGLE (Debosscher et al. (2009)), (3) MACHO (Belokurov et al. (2003), Belokurov et al. (2004)) (4) CoRoT (Deleuil et al. (2009))), (5) Kepler (Blomme et al. (2010)). A number of survey projects are also in various phases of their timeline, in particular (1)

Pan-STARRS[1], (2) LSST[2], and (3) Gaia[3]. We are expected to receive data of exceptional quality from these surveys and in particular Gaia, which can be used in the study and analysis of variable stars (For abbreviations refer to the table List of Abbreviations on page xvii).

The data set which we have used in our thesis is mainly the Hipparcos data set. More details about the Hipparcos data can be found in Chapter 3. The Hipparcos mission provides accurate data for a lot of well studied stars and hence can be used as a control sample for validation of the model classifications. Also, it contains almost all types of variable stars in the solar vicinity (Dubath et al. (2011)). With an influx of data expected to arrive from the Gaia survey in the future, the data from the Hipparcos mission can be particularly useful for training models for classification of variable stars.

Some work has been done for classification of variable stars previously. Aerts et al. (1998) used multivariate discriminant analysis to isolate certain variable classes. Willemsen and Eyer (2007) present a systematic classification of variable stars using PCA and support vector machines. Blomme et al. (2011) uses multivariate Bayesian statistics and multistage approach while Debosscher et al. (2007) implemented a procedure for fast light curve analysis and for the derivation of classification parameters for variable star classification. Richards et al. (2011) presented machine learning methods for classifying variable stars with noisy time series data. The variable stars were classified using Gaussian mixture classifier (Debosscher et al. (2007),Debosscher et al. (2009)) and a Bayesian network classifier in (Sarro et al. (2009)). Dubath et al. (2011) presented an evaluation of the performance of an automated classification of the Hipparcos periodic variable stars into 26 types.

The remainder of the introduction is organized as follows: in Section 1.2 we will review mixture distributions and the Dirichlet distribution which will be used in our thesis for the formulation of our model. In Section 1.3 we will briefly outline the summaries of the upcoming chapters.

1.2 Preliminary concepts

1.2.1 Mixture distributions

Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ denote a random sample of size n, where $\mathbf{Y_i}$ is a D-dimensional random vector with probability density function $f(\mathbf{y_i})$ on $\mathbb{R}^D$. Let the entire sample be represented by $\mathbf{Y^T} = (\mathbf{Y_1}, \ldots, \mathbf{Y_n})^\mathbf{T}$, where the superscript T denotes vector transpose. Thus $\mathbf{Y}$ is an n-tuple of points in $\mathbb{R}^D$ and is an $n \times D$ matrix.

[1] http://pan-starrs.ifa.hawaii.edu/public
[2] http://www.lsst.org/lsst
[3] http://www.rssd.esa.int/Gaia

Throughout this thesis, we will denote realizations of random vectors by lower-case letters, unless mentioned otherwise. Thus, $\mathbf{y^T} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^\mathbf{T}$ denotes an observed sample where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ is the D-dimensional observed value of the random vector $\mathbf{Y_i}$.
The K-component finite mixture density $f(\mathbf{y_i})$ of $\mathbf{Y_i}$ can be written in the form

$$f(\mathbf{y_i}) = \sum_{k=1}^{K} \pi_k f_k(\mathbf{y_i}) \tag{1.1}$$

where $f_k(\mathbf{y_i})$ are densities and π_k are the such that,

$$0 \leq \pi_k \leq 1 \quad (k = 1, \ldots, K)$$

and

$$\sum_{k=1}^{K} \pi_k = 1$$

The quantities $\pi_1, \pi_2, \ldots, \pi_K$ are called the mixing proportions or weights. The $f_k(\mathbf{y_i})$ for each k are called the component densities of the mixture and K, the number of components. In this thesis, we shall refer to the finite mixture distribution as mixture distribution, unless stated otherwise. For a comprehensive discussion, please refer McLachlan and Peel (2004).

In the next section, we will see that mixture distributions can be interpreted using categorical random variables.

1.2.2 Mixture model using categorical random variables

We discussed that a K-component mixture density, $f(\mathbf{y_i})$ can be represented by Equation (1.1). However, the same can be reformulated using categorical random variables as follows.

There are different ways of representing mixtures through categorical random variables, First - through a single categorical random variable. If S_i is a single categorical random variable taking on the values $1, \ldots, K$ with probabilities $\pi_1, \ldots, \pi_K$ respectively, and if the conditional density of $\mathbf{Y_i}$ given $S_i = k$ is $f_k(\mathbf{y_i}), (k = 1, \ldots, K)$, then the marginal density of $\mathbf{Y_i}$ is given by $f(\mathbf{y_i})$ and McLachlan and Peel (2004) explains S_i to be like the component label of attribute vector $\mathbf{Y_i}$.

Second way is, to work with a D-dimensional component label vector $\mathbf{S_i}$, which is what we have used in our thesis, where the kth element of $\mathbf{S_i}$, $S_{ki} = (\mathbf{S_i})_\mathbf{k}$, is defined to be either one or zero, according to whether $\mathbf{Y_i}$, originated from k or not $(k = 1, \ldots, K)$. Hence $\mathbf{S_i}$ is distributed according to multinomial distribution with K categories and with probabilities $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$.

Thus

$$P(\mathbf{S_i} = s_i) = \pi_1^{s_{1i}} \pi_2^{s_{2i}} \dots \pi_K^{s_{Ki}} \tag{1.2}$$

and

$$\mathbf{S_i} \sim \text{Mult}_K(1, \boldsymbol{\pi})$$

where $\boldsymbol{\pi} = (\pi_1, \dots, \pi_K)$
For a comprehensive discussion on mixture distributions, please refer to McLachlan and Peel (2004). Now that we have discussed mixture distributions, lets look into the other major component of our model defined in Chapter 4 namely, Dirichlet distributions.

1.2.3 Dirichlet distribution

Before we define a Dirichlet distribution we need to define an open simplex. For the ith observed value of the random vector $\mathbf{Y_i}$ described in Section 1.2.2, i.e. $\mathbf{y_i} = (y_{i1}, y_{i2}, \dots, y_{iD})$, the $D-1$-dimensional open simplex in $\mathbb{R}^D$ is defined by,

$$\mathbb{V}_{D-1} = \left\{ (y_{i1}, y_{i2}, \dots, y_{iD})^T : y_{id} > 0, 1 \leq d \leq D-1, \sum_{d=1}^{D} y_{id} = 1 \right\} \tag{1.3}$$

A random vector $\mathbf{y_i} = (y_{i1}, y_{i2}, \dots, y_{iD})^T$ defined on a $D-1$-dimensional open simplex in $\mathbb{R}^D$, is said to have a Dirichlet distribution if the probability density of $\mathbf{y_i} = (y_{i1}, y_{i2}, \dots, y_{iD})$ is

$$\text{Dir}(\mathbf{y_i}|\boldsymbol{\alpha}) = \frac{1}{\mathbf{B}(\boldsymbol{\alpha})} \prod_{d=1}^{D} y_{id}^{\alpha_d - 1} \quad \mathbf{y_i} \in \mathbb{V}_{D-1} \tag{1.4}$$

where $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots, \alpha_D)$ is the parameter vector which we will refer to as Dirichlet parameters in this thesis. When $\mathbf{Y_i}$ follows a Dirichlet distribution, we will denote as $\mathbf{Y_i} \sim \text{Dir}(\boldsymbol{\alpha})$ or $\text{Dir}(\alpha_1, \alpha_2, \dots, \alpha_D)$ on $\mathbb{V}_{D-1}$ accordingly in this thesis. Plots of Dirichlet distribution for different values of the Dirichlet parameters are given in Figure 1.1.

It will be interesting to see some of the properties of Dirichlet distributions and lets have a look at the first two moments, namely the mean and variance of equation (1.4). If α_+ is defined as $\sum_{d=1}^{D} \alpha_d$, then,

$$\text{E}(Y_{id}) = \frac{\alpha_d}{\alpha_+}$$

and

$$\text{Var}(Y_{id}) = \frac{\alpha_d(\alpha_+ - \alpha_d)}{\alpha_+}$$

for $d = 1, \dots, D$

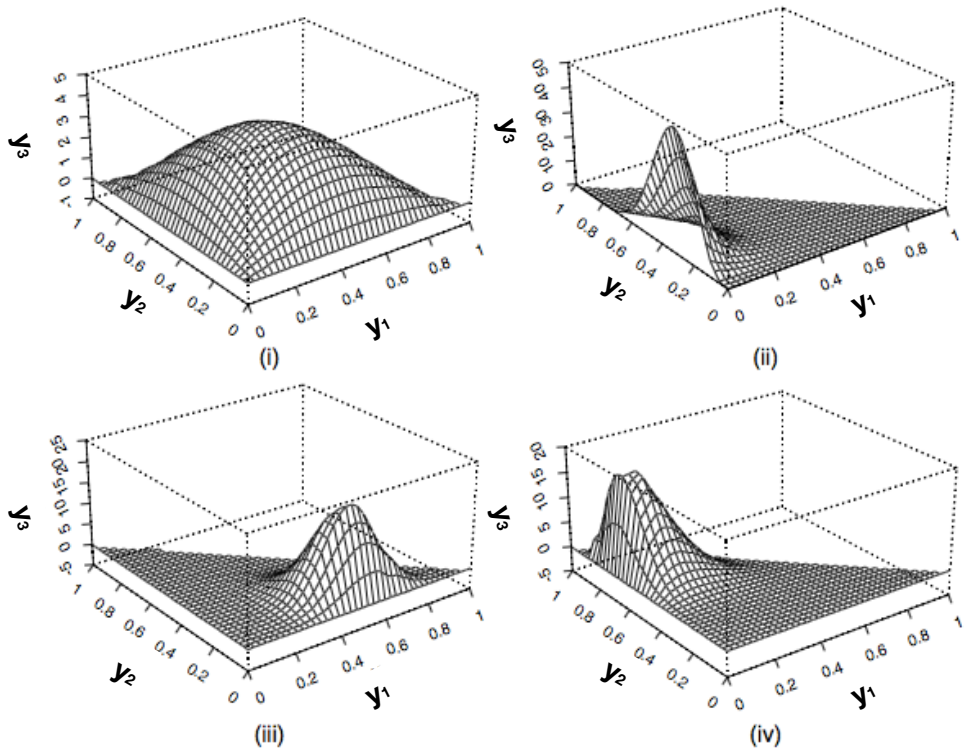


Figure 1.1: Plots of densities of $\mathbf{y_i} = (y_{i1}, y_{i2}, y_{i3})^T \sim \text{Dir}(\alpha_1, \alpha_2; \alpha 3)$ on $\mathbb{V}_2$ with various parameter values : (i) $\alpha_1 = \alpha_2 = \alpha_3 = 2$ (ii) $\alpha_1 = 1, \alpha_2 = 5, \alpha_3 = 10$ (iii) $\alpha_1 = 10, \alpha_2 = 3, \alpha_3 = 8$ (iv) $\alpha_1 = 2, \alpha_2 = 10, \alpha_3 = 4$.
Image Credits : Ng et al. (2011)

where Y_{id} is a column vector of length n, for the ith data-vector and the dth attribute.

For a comprehensive discussion on the Dirichlet distribution kindly refer to Ng et al. (2011) and also Appendix B.6, where it is shown that Dirichlet distribution is an exponential family distribution (see Appendix B.2 for discussion on exponential family distributions).

1.2.4 Random forests

In Chapter 3, we try to search for the simplest and smallest plausible subset of variables which effectively explains the population. Later in the chapter we use an attribute ranking algorithm, using Random forests. However, before we look closely into random forests, let us look into the concept of bagging.

Decision trees and Bagging

Decision trees can be problematic because they can cause high variance in predictions. Bagging is a procedure for reducing the variance of statistical learning methods; in this case, decision trees.

Say we have n independent observations $Y_1, \ldots, Y_n$, each with variance σ^2, then the variance of the mean $\bar{Y}$ of observations is given by σ^2/n. Evidently, the averaging of the set of observations has reduced the variance. Bagging takes advantage of this for classification, and first forms (say) B different bootstrapped training data-sets by taking repeated samples from the single training data set. Then, the method is trained on each of the B bootstrap samples in order to get a classifications for each of the samples. The classes predicted by each of the B trees are recorded and a majority vote is taken. The class that was predicted the most frequently among the B classes will be selected as the overall predicted class.

Out-of-bag error estimation

We discussed earlier that the key to bagging is that the trees are fit to bootstrapped subsets of the observations. However, an estimate of the error rate can be obtained from the training set. Each bagged tree uses a certain proportion of the observations and the remaining observations are defined as the out-of-bag (OOB) observations. Thus any training set data-point is OOB in a certain proportion of the trees. This will give us $B/3$ predictions for any given data-point. As we discussed earlier, in order to get one final prediction for this observation, we take a majority vote. Thus an OOB prediction can be formed this way for all the observations and consequently the OOB error, which is the classification error of the OOB predictions. This upholds the validity of the error of the bagged model while testing, since the response for each data-point is predicted using only the trees that were not fit using that data-point. This is similar to the cross validation performance estimate, but at a much lower computational cost.

From bagging to random forests

Random forest (Breiman (2001)) is a tree based classification method. Just like bagging, a number of decision trees are grown on bootstrapped training samples. At each node, divisions into two groups are considered each using a randomly selected set of attributes. Among these, the best split is selected and the process is repeated for the child nodes with a new set of attributes at each of the nodes. The procedure stops when we have attained a maximum tree, i.e. a tree with terminal nodes containing only a single type of stars.

For a more detailed discussion on Random forests, refer Friedman et al. (2001). The `randomforest` package (Liaw and Wiener (2002)) in R is used in this thesis.

1.3 Chapter summaries

In Chapter 2, we focus on the astronomy part of our thesis. We will discuss about variable stars and also briefly explain the different classes of variable stars that we aim to classify in this thesis. We will also explain briefly the physics that determine the variability of the stars and also why it is important to classify variable stars. At the end of the chapter we also give a brief introduction to the Gaia survey.

In Chapter 3, we focus on the data that is used in our thesis. As mentioned in Section I, our goal is to implement a statistical classifier that can classify the upcoming Gaia data-set. Hence we train our model to the data-set mentioned in Chapter 3 of the thesis. We reduce the dimension of the data by limiting the number of attributes. This simplifies the model, reduces time taken to train the model and avoids overfitting. We also discuss the problem of correlation and list out a set of attributes that are not so correlated which we use in our analysis in Chapter 4.

In Chapter 4, we present our model, namely the Two stage Dirichlet mixture model. We also apply two transformation techniques to transform the data to an open simplex. First we transform the raw data to the probability scale. Secondly, we transform the data to the open simplex. For this transformation, we administer two types of transformations. We compare the performances of the classification based on these two transformation techniques and discuss the results.

In Chapter 5, we focus on the second main goal of our thesis, namely a model for the detection of new classes of variable stars. We use the fixed background model which was proposed by Vatanen et al. (2012) and train it for the detection of new classes of variable stars. We also perform an analysis for the detection of new classes.

In Chapter 6, we provide an analysis of Bayesian classification and also discussion on Bayesian classification using Bayesian conjugate priors. Chapter 7 we contains the conclusion of the entire thesis and possible future directions.

Part II

The Astronomy

Chapter 2

Variable stars

2.1 Introduction

As we survey the night skies, it is easy to imagine that the skies are unchanging. Apart from the twinkling due to the effects of the atmosphere, stars appear fixed and constant to the untrained eye. However, if we try to carefully observe with the naked eye, we may see that some stars do in fact change in brightness over time. We can see a quick brightening and diminishing only to repeat the process again. The period may range from the order of hours to even years. Most of the stars exhibit a change in luminosity as they change and evolve over time. The energy emitted by the Sun, for example varies by about 0.1% over 11 years.

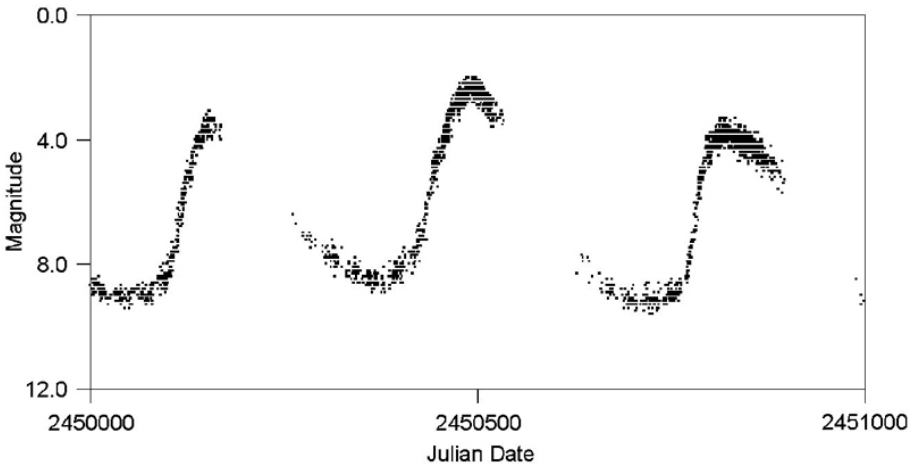


Figure 2.1: Variation in brightness of the star Omicron Ceti, plotted as a light curve. Omicron ceti is a prototype of the Long Period Variable (LPV) stars (refer to Section 2.4.1). The magnitude of the light decreases from Julian year 2450000 (1995 Oct 9) and after reaching a minimum and increasing to a local maximum, it reaches another maximum at Julian year 2450500 (1997 Feb 20) and then continues the same.
Image credit : NASA/CXC/SAO

Also, Figure 2.1 gives the variable star light curve for the star Omicron Ceti which is a Long Period Variable star that will be discussed in Section

2.4.1. The variation in the brightness of the star is plotted as a light curve. The magnitude of the light curve varies from Julian year 2450000 (1995 Oct 9) to Julian year 2450500, periodically. Such stars are termed as variable stars. Simply put, a variable star is one whose brightness as seen from the Earth fluctuates.

Astronomical surveys are the main source for discovery and accumulation of observational data of variable stars for further analysis and interpretation. There are several surveys that collect observable information of variable stars like the ASAS mission (Pojmanski (2004)), ROTSE mission[1] (Akerlof et al. (2000)) etc. This observable information is then used to deduce characteristics of the star such as mass, radius, luminosity, temperature, internal and external structure, composition. These can then be used to constrain and test our theories about stellar or galactic evolution, star and planetary system formation, or cosmology.

Observable information about variable stars in particular can yield valuable findings. Each of the variable star classes or types have distinct differences in physical properties and further research and understanding of each of these classes can give rise to a wide range of applications, among which are asteroseismology (e.g. Aerts et al. (2010)), the determination of the physical parameters of the stars (e.g. radius) in eclipsing binaries[2] (Steinfadt et al. (2010)), and the measurement of distance with standard candles like Cepheids[3] (Feast and Walker (1987)) or RR Lyrae. See Catelan et al. (2004), Sesar et al. (2010) and Castellani et al. (1993) for comprehensive accounts on some of the applications.

The discussion in this chapter has been arranged as follows. Section 2.3 briefly explains some of the preliminary physics and definitions to understand star variability, while Section 2.4 introduces the different classes of variable stars and also how they are different from each other. In Section 2.5 we introduce the Gaia survey. But now let us have a quick look into the history of variable stars and how it has progressed over the years.

2.2 History of variable stars

Ancient philosophers thought that stars were eternal and invariable. However, when in 1638 Johannes Holwards observed that the star Omicron Ceti (Mira) in the constellation Cetus varied its brightness in a regular 11 month cycle, it was a breakthrough. In 1596, David Fabricius (1564−1617), a Protestant minister in Osteel, Ostfriesland noticed that a star in the constellation Cetus became fainter over time and then disappeared after a certain time. This mustered a lot of interest and awareness in the field of star variability (Williams and Saladyga (2011)).

[1] refer table List of Abbreviations on page xvii
[2] refer Section 2.4.2 for details
[3] refer Section 2.4.1 for details

Consequently, more variable stars were observed in the period that followed. For instance, Algol, also called Beta Persei, was observed in 1669 by the Italian Geminiano Montanari, who became the first European astronomer to note the light variation, while R Hydrae was observed in 1670 by Montanari and χ Cygni in 1686 by Gottfried Kirche, who noticed that it was missing from the sky over a period of 13 months.

Variable star astronomy improved significantly with contributions from Edward Pigott and John Goodricke in the 18th century. They confirmed the variability of eclipsing binaries (which is explained in Section 2.4) such as β Persei and η Aquilae (Hoskin (1982)). They contributed also to the determination of precise periods of other known periodic stars (Furness (1915)). However, history was made in 1786, when the first catalogue of variable stars was published by the English amateur Edward Pigott, consisting of 12 variable stars and 39 suspected variables (Gilman (1978)).

Today with the advance of astronomy and of scientific methods, the detection of variable stars by means of photography has become an everyday task. The General Catalogue of Variable stars (GCVS 5.1 version) contains data for 52011 variable stars by 2015 which are located mainly in the Milky Way galaxy (Samus et al. (2017)). However the GCVS is not the only catalogue which is used in studies in astronomy. There are many more which have been detected by photometric surveys such as OGLE (Udalski et al. (2015)), Drake et al. (2014), Kim et al. (2014) and Minniti et al. (2010).

Now let us look into some of the underlying physics that causes the variability in stars. This will help us to understand the different classes of variable stars which will be briefed later in Section 2.4. However we shall not delve into an extensive astrophysics discussion of the stars, but restrict our discussion to the basic preliminaries, as this thesis is primarily meant for a statistics audience.

2.3 Preliminaries

Chemical composition using spectra

Light can split up to form a spectrum using a prism. The light waves are refracted as they enter and leave the prism. There is a connection between the wavelength and the refraction with shorter wavelengths resulting in higher refraction. As a result, red light is refracted the least and violet light is refracted the most, causing the colored light to spread out to form a spectrum as shown in Figure 2.2 (top).

Chemical elements produce their own emission or absorption spectrum. For instance, If hydrogen is put in a discharge chamber and electrons are shot at it, it emits light at some characteristic wavelengths. If we analyze this light using a spectrometer, it would look like the one shown in Figure 2.2 (bottom). Instead of getting a rainbow of light, as we would

normally get when white light passes through a prism, we observe emission only at these specific wavelengths..

Similarly, when light passes through an absorption chamber, a part of the light will be absorbed and we will get a spectrum which looks like the one shown in Figure 2.2 (middle). These are the emission and absorption spectrum, respectively.

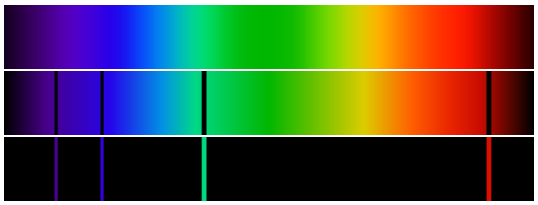


Figure 2.2: From top to bottom · (Top) Spectra that is formed when light is passed and refracted through a prism. (Middle) Absorption lines of hydrogen that are formed when the light is shone on hydrogen in an absorption chamber. (Bottom) Emission lines of hydrogen formed when hydrogen is put in a discharge chamber and electrons are shot at it.
Image credit : Wikimedia Commons

It was discovered by Neil Bohr that, as the electron jumps to a higher energy level it absorbs a photon, and as it moves to a lower energy level, it emits a photon. These photons represents the colors of light missing from the absorption spectrum or that of the lines present in the emission spectrum. These photons carry exactly the necessary energy that is required for the electron to move between energy levels. However when we irradiate it with photons of other wavelengths, nothing happens as those do not have the right amount of energy.

Since the energy levels of each of the elements are unique, their resulting spectra are unique. This helped in identifying the composition of stars. Modern day scientists receive light from a star, pass it through a prism and analyze the absorption and emission lines to understand the composition of the star.

In summary, spectral lines are these lines that form due to emission or absorption of photons in a narrow frequency range, called emission or absorption lines, respectively. Absorption lines occur when an atom, element or molecule absorbs a photon with an energy equal to the difference between two energy levels, while emission lines occur when the electrons of an excited atom, element or molecule move between energy levels.

The evolution of stars

A main-sequence star can be regarded in a zero approximation as a spherically symmetric mass of gas in a stable state. It is described by four equations: hydrostatic equilibrium, energy balance, energy transport equations and the thermodynamical equation of state. The temperature in its

core and the usual chemical composition (75% H, 25% He and traces of other elements) implies that nuclear fusion is maintained in its core, contributing to the terms of the above four equations. Mass and chemical composition determines the type of energy production and transport in the star as well as the evolution of the star.

However, stars evolve and their evolution depends primarily on the amount of hydrogen at the core. At the birth of a star, it is composed of around 75% of hydrogen and around 25% of helium and trace amounts of other elements. Every time hydrogen nuclei fuse into helium nuclei, a photon is released. There are innumerable number of fusions of hydrogen to helium happening at the core which results in the emission of photons, and ultimately, this energy production is what makes the star to shine. The light emitted by these stars is analyzed by scientists as spectrum (Figure 2.2), which we discussed earlier. It should be noted that the light is absorbed and re-emitted millions of times, and changes its energy, so this is not an easy task.

If the amount of hydrogen determines how the star evolves, the mass of a star determines the most important features in the history of a star, including how long it will live, how long it will burn and how it will die. Stars are categorized by how bright they are, called their luminosity, and by and their color which indicates their temperature. These categories are referred to as spectral classes. There are seven main spectral classes of stars, namely O,B,A,F,G,K,M with O class stars being extremely hot and bright, while M class stars are cool and dim. Each stellar class is subdivided by a number from 0 to 9, where G0 is hotter than G3 which is hotter than G8.

Pressure can broaden spectral lines. With increasing pressure in the star's outer layers more and more atoms will be disturbed during the time when emitting or absorbing a photon. This results in a change of energy of the levels of the atom. The pressure in the outer layers of a star can vary in a wide range and thus require a further component of the spectral classification, the luminosity class. The Roman numerals identify what face in the life cycle the star is in. I identifies super giant stars. II identifies bright giant stars, III identifies giant stars. These stars are near the end of their lives as they burnt all their hydrogen and were massive enough to burn their helium into massive elements. IV identifies sub-giants. These stars have burnt all their hydrogen and have started to burn all their helium. They are identified by the giant stage. The main sequence stars are identified by the numeral V. This is where most of the stars spend most of their lives burning hydrogen into helium. Our Sun is in this stage and stars that are in the main sequence are called dwarfs. The next group are the sub dwarf stars identified by VI. They are stars that are burning hydrogen but they are not as bright as they should be for their size. The last luminosity class is the white dwarf which are the stellar corpses of the large stars. These luminosity classes and spectral classes will be used

in Section 2.4 when we define the various variable star types.
The Sun is a normal main sequence star and is classified as G2V; Vega, the bright bluish white star of the constellation Lyra is classified as A0V.

Hertzsprung-Russell (H-R) diagram

In 1910, Ejnar Hertzsprung and Henry Norris Russell created a diagram which has been a major step in the understanding of stellar evolution, the Hertzsprung Russell (H-R) diagram. It is a graph of a measure of the luminosity of a star (absolute magnitude), plotted against a measure of the temperature of a star (spectral type, or colour). Figure 2.3 gives a representation of the H-R diagram.

As we discussed in the previous section, depending on its initial mass, every star evolves over time. Its internal structure varies as a function of its age. These stages of evolution which are explained later, imply changes in the temperature and luminosity of the star and hence the star moves in the H-R diagram. Consequently, the internal structure and evolutionary stage of the star can be determined by locating it in the H-R diagram. The temperature decreases as we move left to right on the H-R diagram, on the x-axis and the luminosity increases as height increases on the y-axis. Each point on the H-R diagram represents a star with that spectral type-luminosity combination. For example, the Sun is of spectral type G2 with a Luminosity if 1 solar luminosity. On the H-R diagram stars that are on the upper left are hot and luminous while stars that are on the lower right are cool and dim. The Hertzsprung-Russell diagram also provides information about stellar radii. Luminosity depends on surface temperature and size. For two stars that have the same temperature, a star has higher luminosity only if it is larger.

The Stefan-Boltzmann equation states that the luminosity of a star (L) is proportional to the square of its radius (R) and to the fourth power of its effective temperature (T).

$$L = 4\pi\sigma R^2 T^4$$

where σ is the Stefan-Boltzmann constant. As a consequence to this, stars on the Hertzsprung-Russell diagram which are cool and luminous must be very large super giants. On the other hand, stars which are hot but not very luminous must therefore be very small white dwarfs.

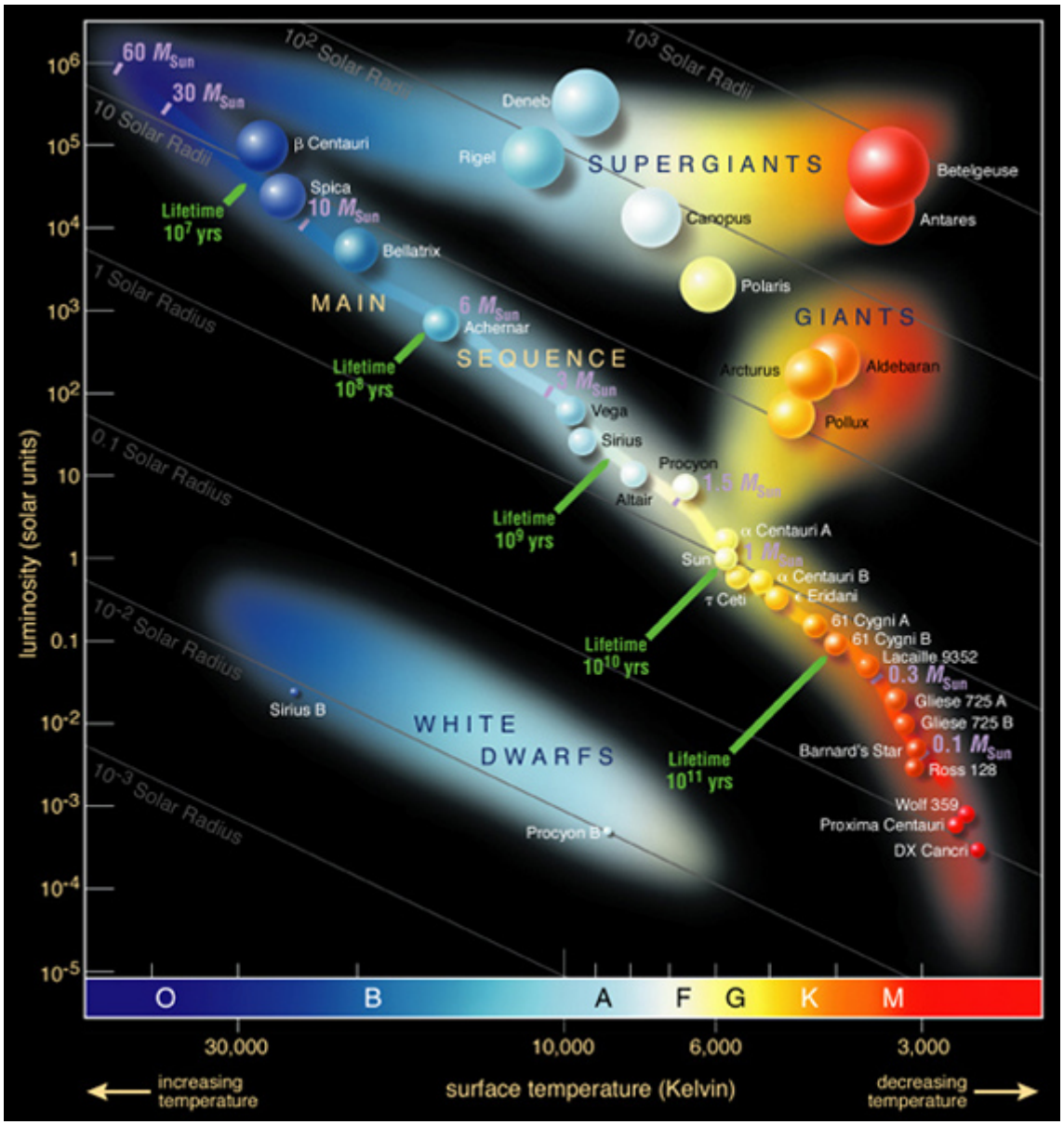


Figure 2.3: Hertzsprung-Russell diagram depicting the different stages of stellar evolution. The main sequence stars are located at the line connecting the top left to the bottom right of the diagram. The Sun is a main sequence star and it is located at spectral class G and luminosity 1 (solar units) in the figure above.
Image Credits: ESA

Star positioning in the Hertzsprung-Russell (H-R) diagram

Depending on the stage of evolution of stars, they occupy different positions in the H-R diagram. Hence, the evolution of a star on the H-R diagram can be classified into 3, as follows,

- The area of the H-R diagram covering the upper left to the bottom right accomodates the main sequence stars or dwarf stars (Powell (2006)). They form the majority of the stars in the H-R diagram. The upper left stars apparently are the hot and luminous stars and the bottom right ones the cool and faint stars. Stars spend almost 90% of their lifetime here, burning hydrogen in their cores (Unsöld (1969)). Our Sun is a main sequence star.
- The region above the main sequence stars is occupied by the Luminosity classes I, II and III, which are the red giant and the super giant stars. The high luminosity implies larger stars, by the Stefan-Boltzmann equation.
- The bottom left of the H-R diagram is occupied by the white dwarf stars, which are the low to intermediate mass stars. This is the final evolutionary stage of a star of about solar or lower mass. They can have radii as small as the Earth, having temperatures around 10,000 K.

In Figure 2.4 we have attempted to plot an approximation of the H-R diagram for the data set of variable stars used in this thesis. Since we have no attribute that records the temperature of the star, we have plotted the absolute magnitude of the variable stars against V-I color. Although V-I Color is an index which indicates the color of the star, inferences on the temperature can be made from the color, which is why we have used it. But the relationship between V-I Color and temperature is not linear, so Figure 2.4 is somewhat distorted, and only its topology can be compared. When we compare Figure 2.4 to Figure 2.3, we see that Figure 2.4 does not contain the faint red end of the main sequence, that is, the lower right part of the diagonal line. We see only the upper half of the main sequence. However the LPVs are where we would expect them: rightward from the visible upper part of the main sequence, at very red colors and relatively bright magnitudes. The leftmost thick stripe in Figure 2.4 corresponds to a small portion of the main sequence. We don't see the yellow-red part of it (no variables there, so it does not appear in our variability data set), and we don't see the hottest, biggest blue O stars either as there are no such stars in the sample. Figure 2.4 contains ACV, BE+GCAS, BCEP, SPB stars in its upper half and also eclipsing binaries in the lowest part, which corresponds to the H-R diagram. According to astronomical knowledge, the single stars and the binary stars form two different main sequences, which are somewhat shifted with respect to each other, but this shift is

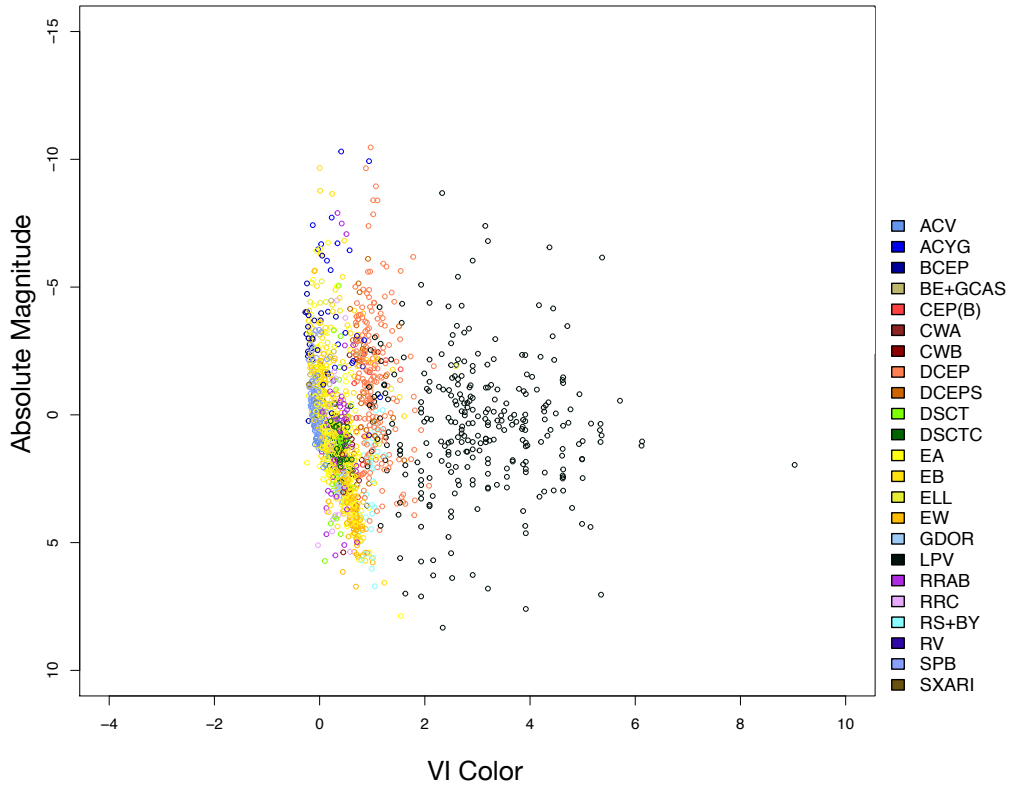


Figure 2.4: H-R diagram for the data set used in this thesis. The absolute magnitude, which is an attribute which gives the absolute luminosity values of the star, is plotted against the color indicator, namely V-I Color. As we don't have attributes that directly record temperature, colors are a good indicator of temperature. Thus the above plot is an approximation of the H-R diagram. The variable types or classes from the data-set are plotted in different colors, according to the legends on the right hand side of the plot. Refer to the table List of Abbreviations on page xvii, for the definition of the variable star type acronyms.

small, and it is not surprising that we don't see it. DSCTs, DSCTCs and GDORs are also available in the lowest part and also the Cepheids are to the right of the main sequence, that is, are redder than the visible part of the main sequence, and are relatively bright as shown in Figure 2.3 as well. RRAB and RRC are slightly more blue than the Cepheids and are positioned between the main sequence and Cepehids.

Instability strip

The instability strip is a narrow band in the Hertzsprung-Russell diagram which contains many different types of variable stars (RR Lyrae, Cepheid variables, W Virginis)(Gautschy and Saio (1996)).

Stars more massive than the Sun enter the instability strip and become variable at least once after leaving the main sequence. In the instability strip the stars become unstable which cause them to pulsate in size and vary in luminosity. We will discuss in Section 2.4 that these instabilities are due to pulsations. They are caused by the pulsations by the doubly ionized helium III (Gautschy and Saio (1996)). For stars in spectral classes A,F and G, helium is neutral on the surface of the star. But near the core, at about $25,000 - 30,000$ K, the HeI layer ionizes to HeII and HeII ionizes at about $35,000 - 50,000$ K.

The contraction of the star results in the increase in density and temperature of the He II layer. Consequently, He II ionizes to He III, resulting in an increase in the opacity of the star and consequently, a rise in temperature of the star. The star hence begins to expand. The expansion causes the He III to recombine to He II and the opacity of the star drops lowering the surface temperature of the star. The outer layers contract and the cycle repeats.

2.4 Classification of Variable stars

Broadly, there are two types of variable stars, intrinsic and extrinsic variable stars. As the name suggests, intrinsic variables are stars whose variability is caused by changes in the physical properties of the stars themselves, whereas for extrinsic variables the variability is caused by external properties like eclipses or rotation. The variable star types used here is based on the variability types found in the General Catalog of Variable stars (GCVS 5.1 version) Samus et al. (2017).

Intrinsic variable stars can be further divided into three subgroups,

1. **Pulsating variables**, i.e. stars that are variable due to the expansion and contraction of the surface layers of the stars.

2. **Eruptive variables**, i.e. stars that experience variance in brightness because of the violent flares, eruptions and mass ejections on their surface.

3. **Cataclysmic or explosive variables**,i.e. stars that undergo an irregular and cataclysmic change in their brightness like novae and supernovae, and then drop back down to dormant state.

Extrinsic variable stars can be divided into two main subgroups,

1. **Eclipsing binaries**, i.e. double stars where they, as seen from Earth's vantage point, occasionally eclipse one another as they orbit;

2. **Rotating variables**, i.e. stars whose variability is related to their rotation. This can be non-uniform surface brightness or/and ellipsoidal shape of the stars. The fluctuations in surface brightness can caused by extreme "sunspots" which affect their apparent brightness and the ellipsoidal shape can be due to fast rotation speeds or binarity (that does not cause eclipses).

As we look deeper into these classes of variable stars, we'll see that they are different physical systems with different reasons for variability, even though their definition originally was based on observable features such as the light curve, color, luminosity and population type.
To sum up, each of the official classes are different physical systems and when we classify the variable stars into different classes, we are in effect filtering out different systems that behave differently. Table 2.1 presents the acronyms (with slight abuse in nomenclature to ensure readability) of the variable classes that will be used in this thesis.

2.4.1 Intrinsic variables

Pulsating variables

These types of variables periodically expand and contract their surface layers resulting in changes in their size, effective temperature and spectral properties, which induces in the apparent change in brightness. These movements are termed pulsations.

Pulsation of stars are caused by the imbalance of two major forces namely the gravitational force and the radiation and gas pressure from the production of photons in the core by fusion processes. However, pulsation is not due to increased radiation pressure from higher rates of fusion in the core, which is constant, but instead from variations in the rate at which the radiation can escape from the star. Pulsation as described in Section 2.3 is the main variability mechanism in the instability strip of the HRD. Pulsating variability can nevertheless also occur in other regions of the HRD, such as in B-type main sequence pulsators and subclasses of white dwarfs. In these cases,the heat engine in the background is usually similar to the one in the instability strip, but the element acting as the heat valve regulating the oscillation is different.

Some of the pulsating variable classes, that have been used in our work, has been listed below.

α Cygni stars (ACYG)

Variables of the α Cygni type or ACYG, exhibit non-radial pulsations i.e. some part of the stellar surface moves inwards while others move outwards simultaneously. They are blue super giant variables of spectral types B or A. Variations in amplitudes are of the order of 0.1 magnitudes, and are often multiperiodic. This is caused by the superposition of many oscillations with close periods. Periods range from days to weeks.

β Cephei stars (BCEP)

β Cephei, or BCEP, are pulsating variables with periods of 0.1 to 0.3 days. They have spectral types B0.5$-$B2 and are found above the main sequence on the H-R diagram. They exhibit small light amplitudes and lie in the upper high temperature part of the main sequence in the H-R diagram. The light amplitudes go from 0.01 to 0.3 magnitudes, in V. The majority of these stars show radial pulsations, but some display nonradial pulsations. Most of these stars are multiperiodic.

Cepheids

Cepheids are yellow super giant variables with periods of 1 to 100 days or more. Light curves produced by Cepheids reveal a quick increase in brightness followed by a much slower decline, generating unique curves like sharp fins. Their amplitude range is typically 0.5 to 2 magnitudes. The spectral class of a Cepheid actually changes as it pulsates, being about an F at maximum luminosity and down to a G or K at minimum.

There are in fact two types of Cepheids, the original Type I or classical Cepheids of periods ranging from days to months, and the slightly dimmer type II with periods between 1 and 50 days. Both types are located in a region of the H-R Diagram called the instability strip. The difference is in the chemical composition with Type I stars containing more number of heavier elements than hydrogen and helium than the Type 2 Cepheids. The light curves of these two classes can also be very similar, but can be distinguished by fourier decomposition.

1. *Type I classical Cepheids*

   These are young stars, more massive than the Sun. This class of Cepheids have a period between 5 and 10 days and an amplitude range of 0.5 to 2.0 magnitudes in visible light. They are 1.5 to 2 magnitudes more luminous than Type II Cepheids.

One of the reasons why the classical Cepheids are extremely popular, is because of their period-luminosity relationship (Leavitt (1908) Leavitt and Pickering (1912)).

After determining the luminosity from the period observed, the inverse square law of brightness is used, which states that the apparent brightness of a source is inversely proportional to the square of its distance, and hence we can easily determine the distance of the stars and place them in the cosmic map.

Some of the Type I Cepheids in our work are the δ Cepheids, DCEP and DCEPS.

δ Cepheid stars (DCEP and DCEPS)

DCEP stars are type I Cepheids that are present in open clusters and are relatively young objects that have left the main sequence and evolved into the instability strip of the Hertzsprung-Russell diagram, while the DCEPS stars have light amplitudes less than 0.5 magnitudes in V (less than 0.7 magnitudes in B) and almost symmetrical light curves. Their periods never exceed 7 days.

2. *Type II Cepheid stars (CW)*

Type II Cepheids are older than Type I Cepheids, and they have lower mass, typically 0.5 to 0.6 solar masses. They have an amplitude range of 0.3 to 1.2 magnitudes. Population II Cepheids are usually observed further from the plane of the Milky Way Galaxy, in orbits which are not circular and in the plane. Type II Cepheids have periods that range from 1 to 50 days. There exist two sub-groups: stars with periods from 1 to 7 days are called BL Herculis stars, while longer-period stars are termed W Virginis. The General Catalogue of Variable stars (GCVS) classifies the former as CWB, and the latter as CWA.

As with the Type I Cepheids they also display a similar well-defined period-luminosity relationship and can be used for distance determination.

δ Scuti stars, (DSCT and DSCTC)

Variables of the δ Scuti type are of spectral types A0-F5 III-V displaying small light amplitudes in V and periods from 0.01 to 0.2 days. The shapes of the light curves, periods, and amplitudes usually vary highly. Many of these stars are multiperiodic. Period determination is usually difficult for these stars due to their small amplitude and multiperiodicity. They are usually found in the downward extension of the Cepheid instability strip, where it crosses the densely populated main sequence.

DSCTC are low-amplitude group of δ scuti variables (light amplitude less than 0.1 magnitudes in V). The majority of DSCTCs are stars of luminosity class V.

γ Doradus stars (GDOR)

γ Doradus stars are a homogeneous group of variables with spectral types of F0-F2. They are non-radial pulsators, dwarfs (luminosity classes IV and V) from spectral types A7 to F7 showing one or multiple frequencies of variability. Amplitudes do not exceed 0.1 magnitudes and periods usually range from 0.3 to 3 days. They lie at or just beyond the Delta Scuti instability strip in the H-R diagram.

RR Lyrae stars (RRab and RRc)

These old population II giant stars are characterized by their short periods, usually about 1.5 hours to a day and have an amplitude range of 0.3 to 2 magnitudes. Their spectral classes range from A7 to F5. RR Lyrae stars are less massive than Cepheids but they also follow their own period-luminosity relationship. They are thus useful in determining distances to the globular clusters within which they are commonly found to a distance of about 200 kilo-parsecs. Sub-types are classified according to the shape of their light curves. RR Lyraes are found on the instability strip of the H-R diagram.

RR Lyrae can be divided into several classes namely ab and c, on the basis of their light curves. The light curves of Type a have relatively long periods, wider ranges and are highly asymmetrical whereas Type b have longer periods, narrow ranges and less asymmetrical light curves. Type c are nearly symmetrical and sinusoidal with shorter periods and narrow ranges. Their amplitudes are not greater than 0.8 magnitudes in V. Nowadays, type a and b are usually grouped together, forming the class RRab. In addition, there exist double-mode RR Lyrae-type variables (denoted by RRd); however, we did not have this class in our data.

RV Tauri stars

The RV Tauri type stars are radially pulsating yellow super giants having spectral types F to G at maximum and K to M at minimum light. The

light curves are distinguished by alternating deep and shallow minima. The complete light amplitude may reach 3 to 4 magnitudes in V. Periods between two adjacent primary minima lie in the range 30 to 150 days (these are the periods that appear in the catalogues). The light curves are non-sinusoidal, and usually non-repeating. The RV Tauri stars seem to be old stars, with masses similar to that of the Sun.

Slowly pulsating B stars (SPB)

They are main sequence B2 to B9 stars with $3-9$ solar masses. Their periods may be multiple and range from 0.4 to 5 days and amplitudes are smaller than 0.1 magnitudes.

Long-period variable stars (LPVs)

Long period variables, or LPVs as the name suggests, are variables whose light fluctuations are fairly regular and long. It may require many months, or years, for the completion of a cycle. They are red giant and super giant stars. A subgroup of them is also called Mira stars after the Mira Ceti star, in the constellation Cetus, which was the first pulsating variable discovered. LPVs are cool red giants or super-giants and have periods of months to years. Their luminosity can range from 10 to 10,000 times that of the Sun.

Eruptive and cataclysmic variables

Eruptive variable classes consists of stars whose variation in brightness can be attributed to the violent and dramatic flares that occur on the stellar surface. These changes in luminosity are usually accompanied with shell events or mass outflow in the form of stellar wind, or interaction with outside interstellar medium.

For example, stars in the Be variable class show variability but no light outbursts- In the General Catalogue of Variable stars, Be stars are known in most cases as Gamma Cassiopeiae (GCAS) variables, after the prototype, but others are arbitrarily classified as Be. Be stars are very luminous variable stars.

On the other hand, cataclysmic variables, or CVs, are binary systems that consist of a normal star and a white dwarf. They are small and roughly the size of the Earth-Moon system, with an orbital period in the range of 1 to 10 hours. The companion star, a more or less normal star like our Sun, loses material onto the white dwarf by accretion, which is the accumulation of dust and gas onto larger bodies. The high density of the white dwarf star causes strong X-ray emission during the accretion process. There are probably a million of such cataclysmic variables in our Galaxy[4].

[4] refer to http://www.imagine.gsfc.nasa.gov/science/objects/cataclysmic_variables.html

2.4.2 Extrinsic variables

Eclipsing binaries

These are binary variables with the orbital planes so close to the observer's line of sight that the stars involved eclipse each other periodically. The changes of the apparent combined brightness of the system is because of the geometric effect of eclipses. The period will be the same as the period of the orbital motion of the components. The light curves produced by eclipsing binaries show distinctive periodic minima.

There are different eclipsing binaries namely EA, EB and EW. EAs are are widely separated binaries with spherical or slightly ellipsoidal components. Between the eclipses, the light remains almost constant or varies only insignificantly because of reflection effects, slight ellipsoidality of the components, or physical variations. The periods can be extremely wide-ranging from 0.2 to more than 10,000 days and we may not observe a secondary minima.

EB are eclipsing binaries with ellipsoidal components and light curves for which it is difficult to specify the exact times of onset and end of eclipses owing to a continuous change of the apparent combined brightness of the system, between eclipses. Unlike EAs, a secondary minima is observed in all cases. Periods are longer than 0.5 days and the components generally belong to the early spectral types (B to A). Light amplitudes are usually less than 2 magnitudes in V and the components generally belong to spectral types F to G and later.. These variables have light curves which are slightly rounded.

EWs are eclipsing systems with periods less than 1 day, consisting of ellipsoidal components. Like EB it is extremely difficult to to specify the exact times of onset and end of eclipses. Light amplitudes are usually less than 0.8 mag in V. However, one of the major properties of an eclipsing system is that, the observed values for color and absolute magnitude of the stars can be almost any value as its a combinations of the values of these two stars. This property of the eclipsing binaries have resulted in some confusion in classification which is discussed in Section 4.3.

Rotating variables

Stars including the Sun sometimes have conspicuous spots on its surface. These regions appear darker than the surrounding areas because they are cooler. As the Sun rotates the spots appear to move across its surface. One of the sides of the sun can have fewer sunspots than the other, which hence result in fractionally lower light output than for the other side. This principle can be extended to other stars, some of which are thought to have much stronger star-spot activity. Star-spots can be either dimmer or brighter than surrounding regions. As a star with star-spots rotates,

its brightness changes slightly. Stars exhibiting such behavior are called rotating variables.

Ellipsoidal variable stars are components of close binary systems and their variability is not caused by eclipses. The shape of each star gets distorted by the gravity of the companion, which will in turn cause its brightness to be non-uniform. In this case, the close binarity distorts them and causes temperature inhomogeneity on their surface, making them brighter at one place and fainter at others.

α^2 Canum Venaticorum stars (ACV)

These are the α^2 Canum Venaticorum variables and are main-sequence stars with spectral types B8p to A7p which display strong magnetic fields. They exhibit magnetic field and brightness changes (periods of 0.5 to 160 days or more). The amplitudes of the brightness changes are usually within 0.01 to 0.1 magnitudes in V.

BY Draconis stars (BY)

BY Draconis-type variables, which are emission-line dwarfs of dKe−dMe spectral type showing quasi-periodic light changes with periods from a fraction of a day to 120 days and amplitudes from several 0.01 to 0.5 magnitudes in V. Some of these stars also show flares similar to those of eruptive variable stars, and in those cases they also belong to the latter type and are simultaneously considered eruptive variables.

Ellipsoidal stars (ELL)

Rotating ellipsoidal variables are close binary systems with ellipsoidal components, which change combined brightness with periods equal to those of orbital motion because of changes in areas emitting, toward an observer, but showing no eclipses. Light amplitudes usually do not exceed 0.1 magnitudes in V.

RS Canum Venaticorum stars (RS)

These are binary systems, whose primary component is usually giants from late F to late K spectral type. A significant property of these systems is the presence in their spectra of strong Calcium II, H and K emission lines[5] of variable intensity, indicating increased chromospheric activity of the solar type. These systems are also characterized by the presence of radio and X-ray emission. Their light curves look like sine waves outside eclipses, with amplitudes changing slowly with time.

[5] H and K emission lines are absorption lines in the spectra of the star.

Table 2.1: List of variable star types and their acronyms that are used in this thesis (for details about the catalogue refer to Chapter 4). There are 23 variable star types.

Type	**Acronym used in this thesis**
Eclipsing binary	EA
	EB
	EW
Ellipsoidal	ELL
Long period variable	LPV
RV Tauri	RV
W Virginis	CWA
	CWB
δ Cepheid	DCEP
(first overtone)	DCEPS
(multi-mode)	CEP(B)
RR Lyrae	RRAB
	RRC
γ Doradus	GDOR
δ Scuti	DSCT
(low amplitude)	DSCTC
β Cephei	BCEP
Slowly Pulsating B star	SPB
B emission-line star and	
γ Cassiopeiae	BE+GCAS
α Cygni	ACYG
$\alpha-2$ Canum Venaticorum	ACV
SX Arietis	SXARI
RS Canum Venaticorum and	
BY Draconis	RS+BY

SX Arietis stars (SXARI)

SX Arietis-type variables. These are main-sequence B0p to B9p stars with variable Helium I and Silicon III lines and magnetic fields. They are sometimes called helium variables. Periods of light and magnetic field changes (about 1 day) coincide with rotational periods, while amplitudes are approximately 0.1 magnitudes in V. These stars are high-temperature analogs of the ACV variables.
There are various other variable star types but we have only restricted our discussion to the ones we have considered in our thesis. For further discussion on the above listed variable class types and other types refer Percy (2007).

2.5 The Gaia mission

Gaia is a cornerstone mission in the science programme of the European Space Agency (ESA) to chart a three dimensional map of our Galaxy, the Milky Way. It is a public spectroscopic survey, targeting observations about more than 1 billion celestial objects, and opening an unprecedented insight into the structure and history of the Milky Way (Süveges et al. (2017)). This amounts to about 1 per cent of the Galactic stellar population. The time series of on average will be instrumental in the detection and analysis of stars that are variable. Studies in Eyer and Cuypers (2000) predict about 18 million variable stars, including about 5 million ”classic” periodic variables. This mission aims to develop the largest and most precise 3D space catalogue ever. Each of the target stars will be scrutinized and monitored 70 times over a 5 year period.

When combined with Gaia astrometry, the survey will quantify the formation history and evolution of young, mature and ancient galactic populations. This alone will revolutionise knowledge of galactic and stellar evolution. When combined with precision astrometry, delivering accurate distances, 3D spatial distributions, 3D space motions, and improved astrophysical parameters for each star, the survey will help us to understand the formation history and evolution of young, mature and ancient Galactic populations.

Gaia was launched on 19 December 2013 and arrived at its operating point a few weeks later. We are expected to receive huge amount of data regarding the stars in the coming years. In preparation for the analysis of the data, many studies are devoted to the classification of the observed objects; our research makes part of this effort, aiming at the development of flexible semi-supervised and anomaly detection methods.

2.6 Synopsis

Variable stars are important in the field of astronomy. We briefly mentioned how each variable star class is a different physical system with properties that are not entirely the same. Gaia is expected to give us lot of data in the next few years and it is vital to astronomical research that we classify them into different classes effectively. We were able to see that these variable types have differences in period, amplitude and their light curves. This will be useful in Chapter 3, where we take the first step to the development of our classification model, the training data-set.

Part III

Modeling and methodologies

Chapter 3

Training data-set

3.1 Introduction

In Chapter 2, we discussed about the different types of variable stars and why it is important to classify these types of variables. Now, we step into statistical learning. Friedman et al. (2001) says "Statistical learning refers to a set of tools for modeling and understanding complex datasets". They are divided into supervised or unsupervised learning. Broadly speaking, supervised statistical learning is about building a statistical model for predicting (or estimating) an output based on one or more inputs. These inputs are labeled data that train the model, and are called training data. On the other hand, with unsupervised learning, the model learns the relationships and structure from the data without any preliminary knowledge. For a detailed discussion on statistical learning, refer Friedman et al. (2001).

The main prerequisite to an efficient statistical learning model in application is the quality of the data used. By quality, we mean how accurately the data represent the variable type. In supervised learning, the model is trained with a training data-set and if the data-set doesn't accurately represent or characterize the phenomenon that is being studied (i.e. variable stars, in this thesis), the model predictions will be incorrect. Similarly, in unsupervised learning, if the model is based on data that does not represent well enough the relationships and structure of interest, the estimates won't be useful for further study of the phenomenon. With the surge of world wide web and technology, modern day scientists and statisticians have access to a plethora of data. However the trustworthiness and quality of these data-sets need to be scrutinized and assessed before using them in any study.

Variable star classification studies in the past have used the data from surveys such as (1) ASAS (Pojmanski (2002), Pojmanski (2003); Eyer and Blake (2002), Eyer and Blake (2005)), (2) OGLE (Debosscher et al. (2009)), (3) MACHO (Belokurov et al. (2003), Belokurov et al. (2004)) (4) CoRoT (Deleuil et al. (2009)), (5) Kepler (Blomme et al. (2010)). Some projects like Pan-STARRS1 (PS1) have contributed heavily in astronomy. Other ambitious projects are in their advanced stages of preparation, in particular

(1) LSST[1] and (2) Gaia[2] (References are given below). These projects have different primary goals, but the observational data of unprecedented quality that they will provide will help in the study of variable stars. These data help us to characterize different types of variable stars and also help us to understand the types from different surveys which are like different vantage points. Systematic progress in this field facilitates the development and leads to the contribution of wide range of astronomical topics.

In this chapter, we will be looking closely into the data that we will be using in Chapter 4 of this thesis. We will discuss on why the data-set is most reliable (Section 3.2) as a training set and discuss its composition in detail in Section 3.3. We will list each of the attributes in Section 3.4 and devote the remainder of the hapter to find the smallest attribute set which will give the best classification results. Finally, we will look into the structure of our data and also explain the composition of the data that we have used as our training data-set.

We will be referring to each of the variable star types by their acronyms used in this thesis (Table 2.1), unless mentioned otherwise.

3.2 Data sources

In this thesis, we have selected the data which were used in Dubath et al. (2011). These authors selected a subset of stars mainly from the Hipparcos catalogue (Perryman et al. (1997)), which is an outstanding training data-set. Stars in Hipparcos catalogue have reliable types (or classes) from literature and hence it makes them perfect for training and testing our supervised classification model in Chapter 4, semi-supervised classification model in Chapter 5.

Dubath et al. (2011) revised the classes of the training data originating from Hipparcos according to more recent information, mainly from the International Variable Star Index (Watson et al. (2009)) catalogue from the American Association of Variable Star Observers (AAVSO catalogue hereafter). This was because the variability types provided in the Hipparcos periodic star catalogue were mainly taken from the literature available at the time of publication (1997). Also some types from personal communications with Lebzelter, De Cat and Romanyuk and two types from Eker et al. (2008) were also included in the training data-set, as it was based on experience by specialists.

Table 3.1 gives the list of variable star types, the number of instances for each of the types and their references. Some of the types with very few instances or types are similar from the viewpoint of the statistical learning methods were combined together to form a single class; these combinations are denoted by the + sign between the variable type names,

[1] http://www.lsst.org/lsst
[2] http://www.rssd.esa.int/Gaia

e.g. BE+GCAS, RS+BY. Before we detail our training-data in Section 3.3, let us have a closer look at the main catalogues that are used in our work, namely Hipparcos and AAVSO.

3.2.1 Hipparcos Catalogue

The Hipparcos space astrometry mission was a European project of the European Space Agency (ESA) which pinpointed the positions of more than one hundred thousand stars with high precision and more than one million stars with lesser precision. Launched in 1989, it was the first to be dedicated to measuring the positions, distances, motions, brightness and colours of stars (Perryman (2010)).

The satellite observations relied on a pre-defined list of target stars, known as the Hipparcos-Input Catalogue. This input catalogue was compiled over the period 1982−89 (Turon et al. (1992), Turon et al. (1995)), finalized pre-launch, and published both digitally and in printed form. Constraints on total observing time, and on the uniformity of stars across the celestial sphere for satellite operations and data analysis, led to an input catalogue of some 118,000 stars.

Every star in this predefined star list was scanned about 100 times over the mission's span of 3.5 years. The data collection was completed in March 1993 and the resulting Hipparcos catalogue of more than 118,200 stars, was published in 1997 (Perryman et al. (1997)).

The Hipparcos catalogue contains 118,204 entries with associated photometry, among which 11,597 sources are identified as variable: 2,712 and 5,542 of them were published in the periodic and unsolved catalogues (Perryman (1997)), respectively, and 3,343 objects remained not investigated. In December 2013, ESA launched a successor mission to the Hipparcos called the Gaia mission (Section 2.5).

Hence the Hipparcos periodic star catalogue includes some of the best studied stars and results obtained for these stars can be validated using many published information. Dubath et al. (2011) calls this catalogue a "control sample" as this catalogue is particularly useful in evaluating variable star classification methods before they are be applied in large surveys.

3.2.2 VSX-AAVSO

VSX[3] was initiated by an amateur astronomer, Christopher Watson in response to the specific desires of a group of amateur astronomers of the American Association of Variable Star Observers (AAVSO). It was an answer to the need for a globally-accessible central repository for all up-to-the-minute information on variable stars, both established and suspected.

[3]https://www.aavso.org/vsx/index.php?view=about.top

The VSX is an online medium by which variable star data are made available, maintained and revised. The validity or the authenticity of the VSX data is maintained by populating it with the latest findings known to be accurate and with approved reviewers constantly revising the metadata, citing sources for any new details and logging the justification for any changes.

This index uses information from the General Catalogue of Variable Stars (GCVS) and the New Catalogue of Suspected Variables (NSV) and it is kept up-to-date with the literature with two releases per month. The release of 13 June 2010 was adopted by Dubath et al. (2011) in their work and we will use the same in this thesis.

3.3 Data

Hipparcos catalogue and VSX-AAVSO have been used to form our training set. In Section 3.2.1 we mentioned that out of 11,597 sources that were identified as variable, 2,712 were listed periodic. Dubath et al. (2011) formed a training set, which was a subset of Hipparcos stars with most reliable types available from literature. In this regard, 171 out of 2,712 stars were removed due to incomplete photometric data. Most of them were eclipsing binaries (152 EAs) with too few Hipparcos measurements during the eclipses.

Also, each of the Hipparcos data-points were assigned types according to the type-assignment process detailed in Dubath et al. (2011):- For eclipsing binaries and ellipsoidal variables, the Hipparcos periodic star catalogue was taken as the reference. However, lists of Hipparcos stars of the types GDOR, SPB and BCEP were provided by personal communications of P. De Cat and of LPVs by T. Lebzelter, as both of them maintained up-to-date compilations of literature of these types. For ACVs and SXARIs, only the Hipparcos stars included in the list of stars provided by I.I. Romanyuk (private communication) were retained. Also, only the subset of Hipparcos RS and BY stars listed in the third edition of the catalogue of chromospherically active binary stars (Eker et al. (2008)) were included.

All the stars from the AAVSO catalogue with a type matching any of the above mentioned types were excluded. For example, a star which was listed as Mira, SR, LB or SARV in the AAVSO catalogue that was not in the Lebzelter list of LPV was discarded from the training set. The AAVSO catalogue was also used to assign a type to the remaining stars from the Hipparcos periodic star catalogue.

In addition to these, 64 stars with low quality data, star types with less than 3 representatives, 92 stars with uncertain type, 32 stars lacking good-quality photometric information, 107 stars with missing color attributes, and combined types (such as an intrinsic variable included in an eclipsing binary) were also excluded. Hence we have a well-studied

Table 3.1: This is an extension of Table 2.1. Variability types included in the training set are listed together with the corresponding number of instances and references. The training set contains 1661 sources in total. These sources are based on the type-assignment by Dubath et al. (2011) mentioned in Section 3.4. The acronym p.c refers to personal communication. Each of the attributes are defined in Table A.1

Type	**Type acronym**	**Num**	**Main Reference**
Eclipsing binary	EA	228	Hipparcos
	EB	255	Hipparcos
	EW	107	Hipparcos
Ellipsoidal	ELL	27	Hipparcos
Long period variable	LPV	285	Lebzelter(p.c)
RV Tauri	RV	5	AAVSO
W Virginis	CWA	9	AAVSO
	CWB	6	AAVSO
Delta Cepheid	DCEP	189	AAVSO
(first overtone)	DCEPS	31	AAVSO
(multi-mode)	CEP(B)	11	AAVSO
RR Lyrae	RRAB	72	AAVSO
	RRC	20	AAVSO
Gamma Doradus	GDOR	27	De Cat(p.c)
Delta Scuti	DSCT	47	AAVSO
(low amplitude)	DSCTC	81	AAVSO
Beta Cephei	BCEP	30	De Cat(p.c)
Slowly Pulsating B star	SPB	81	De Cat(p.c)
B emission-line star and Gamma Cassiopeiae	BE+GCAS	13	AAVSO
Alpha Cygni	ACYG	18	AAVSO
Alpha-2 Canum Venaticorum	ACV	77	Romanyuk(p.c)
SX Arietis	SXARI	7	Romanyuk(p.c)
RS Canum Venaticorum and BY Draconis	RS+BY	35	Eker et al. (2008)
	Total	1661	

Table 3.2: Summary of the data used in this thesis, attribute-wise. We have listed all the 45 attributes from the data and their range, quartiles, median and mean.

Attributes	**Minimum**	**1st quartile**	**Median**	**Mean**	**3rd quartile**	**Maximum**
p2pScatterOnDetrendedTS	0.0000409	0.0011554	0.0111440	0.0630335	0.0451677	1.1778933
p2pScatterOnFoldedTS	0.0000166	0.0005658	0.0022588	0.0202556	0.0099132	1.7284108
scatterOnResidualTS	0.0000074	0.0002307	0.0008654	0.0105634	0.0045914	0.9795798
Raw_WeightedStdDev	0.005472	0.031716	0.109623	0.224071	0.249317	1.677473
Raw_WeightedSkewness	-5.1913	-0.1456	0.2479	0.5679	0.8847	6.3173
Raw_WeightedKurtosis	-1.8208	-0.8960	-0.3571	1.1565	0.7624	42.8390
Raw_PercentileRange10	-3.6397	-0.3245	-0.0706	-0.2796	-0.0261	-0.0051
stetsonJ	-0.1356	2.1569	5.1688	11.5894	12.3016	130.6171
stetsonJweighted	-4.001	1.789	4.362	10.936	11.631	124.777
stetsonK	0.3695	0.7743	0.8203	0.7991	0.8527	0.9637
WstetsonJ	-0.1406	2.1252	5.0878	11.3046	12.1372	119.2925
WstetsonJweighted	-3.876	1.744	4.244	10.624	11.343	118.789
WstetsonK	0.3257	0.7765	0.8241	0.7981	0.8560	0.9688
logPnonQso	-237.793	-55.727	-38.969	-46.362	-29.244	-9.962
logPqso	-261.331	-80.385	-56.359	-63.504	-38.358	-9.954
qsoVar	0.4226	3.7046	12.9549	34.3229	40.6568	1125.1889
nonQsoVar	0.6899	1.2910	2.6115	12.5887	9.9796	311.5876
LogPeriod	-1.2954	-0.3481	0.1436	0.4324	0.8983	2.7959
LogAmplitude	-2.0046	-1.1107	-0.4782	-0.5522	-0.1087	0.7696
HarmNum	0.000	0.000	1.000	2.046	3.000	20.000
A11	0.004947	0.034916	0.124919	0.328561	0.324545	2.738420
A12	0.00000	0.00000	0.02677	0.06222	0.08987	0.74621
PH12	-3.1359	-1.5627	1.5708	0.3153	1.5708	3.1415
A13	0.00000	0.00000	0.00000	0.02839	0.03751	0.49180
PH13	-3.14147	-0.59898	0.00000	-0.09385	0.00000	3.14155
A14	0.00000	0.00000	0.00000	0.01305	0.00000	0.59525
PH14	-3.1314	-1.5708	-1.5708	-0.8102	-0.3786	3.1267
A15	0.000000	0.000000	0.000000	0.006067	0.000000	0.363823
PH15	-3.124	3.142	3.142	2.491	3.142	3.142
logA11minusA	0.000000	0.000000	0.002478	0.007015	0.009518	0.134241
logA12_A11	0.00000	0.00000	0.08070	0.08734	0.14636	0.48048
logA13_A12	0.0000	0.0000	0.0000	0.0834	0.1862	0.7036
absGlat	0.0144	4.5693	15.4296	22.4868	35.8289	87.4443
Glat	-85.1703	-15.8684	-0.8604	0.1754	14.8707	87.4443
Glon	0.1603	95.0324	184.7437	185.1528	281.9354	359.5429
Parallax	-31.670	1.020	2.840	4.248	5.920	84.580
Absolute_Mag00	-10.4671	-1.1552	0.5657	0.3464	1.9949	8.3295
BV_Color	-0.3100	0.1330	0.4120	0.5992	0.9910	5.3000
VI_Color	-0.2700	0.1500	0.4800	0.8712	0.9900	9.0300
JmK	-1.2990	0.0830	0.2640	0.4545	0.6670	6.8890
JmH	-0.4990	0.0360	0.1950	0.3051	0.4860	1.8230
HmK	-1.2080	0.0300	0.0760	0.1494	0.1900	6.7670

and reliable training data-set of 1,661 stars. Refer to Table 3.1 for the number of instances of each of the types and the main source of reference of these types. Also Table 3.2 gives the summary of the attributes that we have used for our analysis. For more details on the type-assignment process refer Dubath et al. (2011).

3.4 Class attributes

There are 45 attributes in our data-set, which are listed in Table A.1 in Appendix A and Table 3.2, along with a short description. Some of the attributes describe the stellar properties, like mean color, absolute brightness, while others reflect the characteristics of the light curve such as the period of the light curve, the amplitude etc. Further attributes describe the shape of the folded light curves. Broadly, these attributes can be divided into five based on what they summarize, as listed in Table A.2. Further in-depth description of attributes from the Hipparcos catalogue are available in ESA (1997a) and ESA (1997b).

3.4.1 Attribute selection

Effective classification algorithms require the selection of a subset of attributes which will provide the most accurate classification. The search is for the simplest and smallest plausible subset of features which effectively predicts the classes in the population. The attributes chosen must be able capable of characterizing the stars as thoroughly as possible. Otherwise, the predictors will distort the classification to give biased results. Also, as the number of attributes or variables decreases, so does the CPU cost of classification and the time taken for classification. In this section our aim is to list the most effective attributes for classification.

In the following discussion, we use a measure for variable importance from the random forests literature (refer to Section 1.2.4 and Breiman (2001)) to determine the best attributes for our classification model. That is, we use the mean decrease in accuracy measure as our variable importance measure. It is computed by permuting out-of-bag (OOB) (Section 1.2.4) data. For each tree, the prediction error on the out-of-bag sample data is recorded. The same is done after permuting the data of each predictor variable. Now that the data in the predictor variable is permuted or shuffled, the accuracy is affected. The difference between the two accuracy is averaged over all the trees and normalized by the standard deviation of the differences, to get the variable importance measure. The attributes are sorted by the decreasing values of mean decrease in accuracy, to get the list of attributes ranked by importance. We will refer to this list as attribute-rank-list in this thesis. The R package `randomForest` is used to achieve this.

Towards a minimum attribute list

The procedure we described earlier helps us to get an attribute-rank-list of the most important attributes for classification. The importance value decreases but the question is where should we cut the list. In other words, what is the smallest set of attributes that will give us still reliable results?. For this we construct a scoring system as follows.

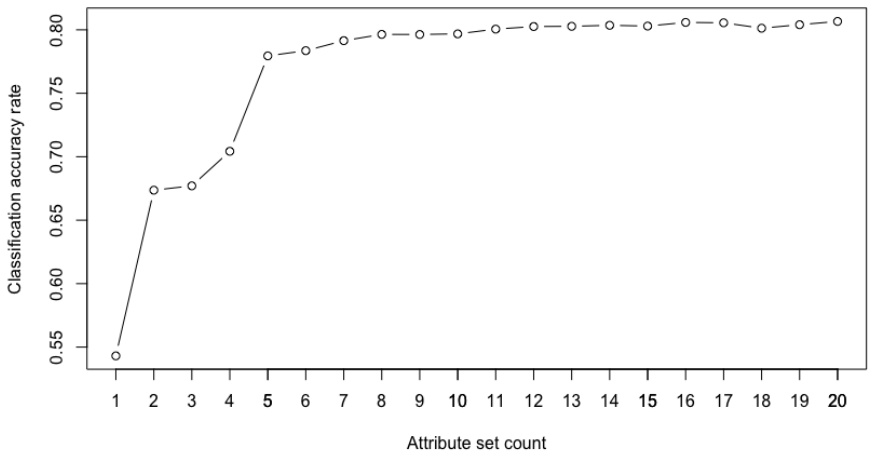


Figure 3.1: Classification accuracy plotted against the attribute-set count. For example the classification accuracy rate for the attribute-set of count six, i.e. the attribute set containing the first six attributes ranked by importance, is around 0.77. Here we see a steep rise in classification accuracy rate till attribute-set count 6 and after which it plateaus. These accuracies are plotted by the `randomForest` package.

We will apply a forward selection strategy for attribute set selection. We start with the top two ranked attributes from the attribute-rank list say, $\{A_1, A_2\}$, run the random forest classifier with these attributes and record the classification accuracy, for the entire data set. We repeat these steps again, but this time by adding the next ranked attribute to the training set, to form the attribute-set $\{A_1, A_2, A_3\}$ of attribute-set count 3 and repeat the steps again, to get the classification accuracy or the classification accuracy rates. Figure 3.1 gives the plots of the classification accuracy rates against the attribute-set count. The classification accuracy rates are plotted against the attribute-set-count in Figure 3.1. The rates go over 75 percent after the attribute set count crosses five, but plateaus around 80 percent from 8 onwards. This is a variant of the method which was adopted in Dubath et al. (2011).

However, though the classification accuracy rates are almost the same after the attribute-set count crosses 6, we have decided to select 16 as the

attribute set count. These are the 16 most important attributes according to the attribute-rank-list we mentioned earlier since some of the attributes between ranks 6 and 16 are known to be astrophysically meaningful, for instance the amplitude of the light curve (ranked 9th), the relative phase PH12 of the first two harmonic components of the light curve (ranked 14th) or the $B - V$ color (ranked 7th). For instance amplitude of the light curve is ranked 9 and PH12, the relative phase of the second harmonic term is ranked 14, B-V color index is ranked 7.

Figure 3.2 display the distribution of the most important attribute, Log (Period) for each of the variability type. The distribution of the remaining 15 of the attributes can be found in Appendix B (Figure A.2).
Let us look at some of the combinations of the 16 attributes used in our thesis. Figure 3.3 and 3.4 gives a 2-D plot of the different combinations of attributes. Each of the colors represent a variable class or type. It can be seen that when Log(Period) is plotted against Log(Amplitude) it provides much better segregation between the variable types.

If we compare the data-plots Figure 3.3 and the Log (period)-Log (amplitude) plot in Figure 3.4, we see much segregation among classes. In the former, apart from DSCT, DSCTC and LPV the other classes were relatively close in period. However when we combine the attributes, Log (period)-Log (amplitude) as in Figure 3.4, we see that many of the classes are segregated. For instance, DCEP and RS+BY classes, which have closer periods are well segregated in Figure 3.4.

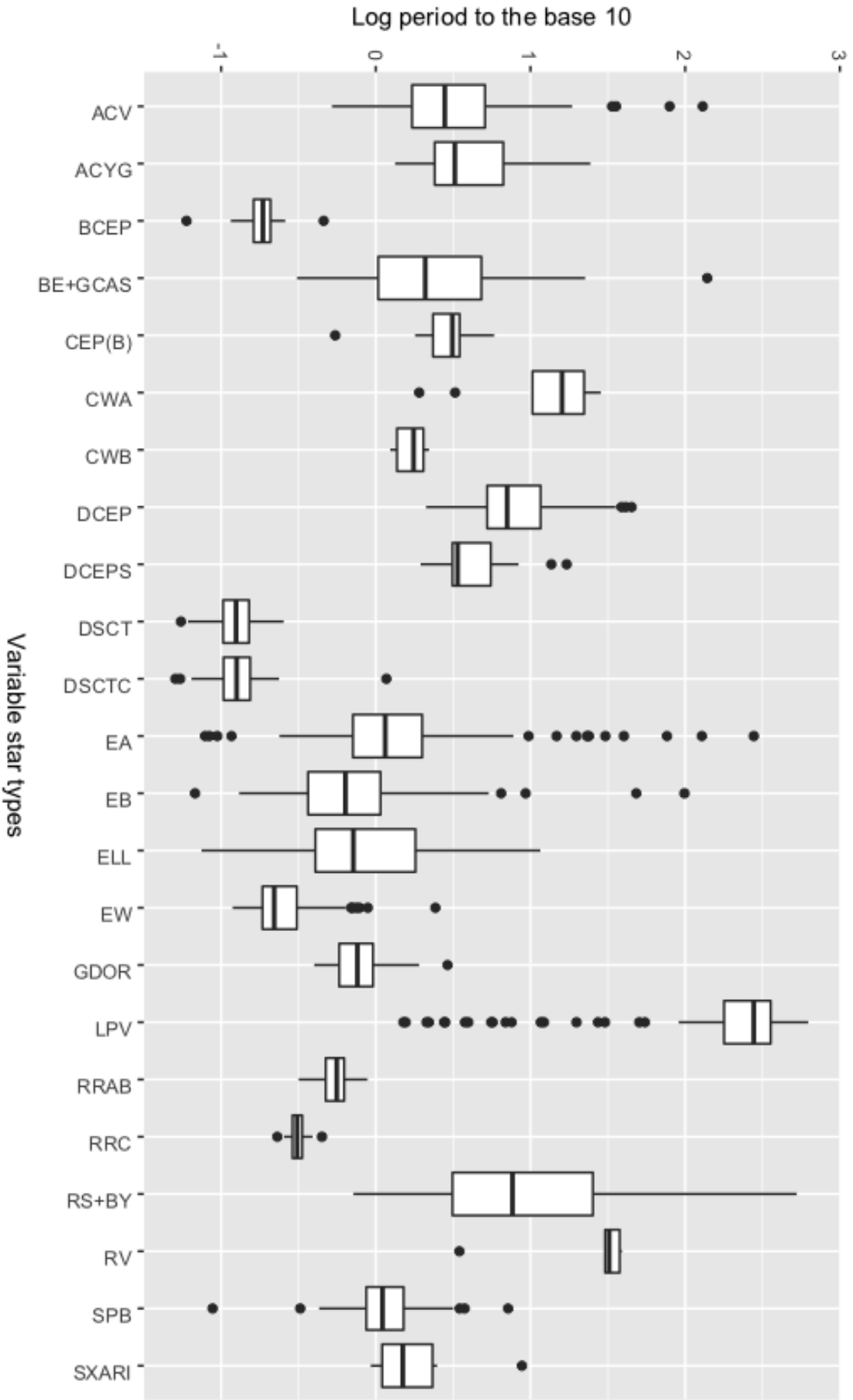


Figure 3.2: Distribution of the important variable, Log(Period) for each of the variable star classes or types. The other 15 attributes can be found in Appendix B, Figure A.2. We see that LPV, which has long periods shows high values of decadic log of the period. DSCT and DSCTC gives very low values and hence it has higher discriminative power for LPV, DSCT and DSCTC but it will be difficult to classify the other classes on the basis of Log (period) alone.

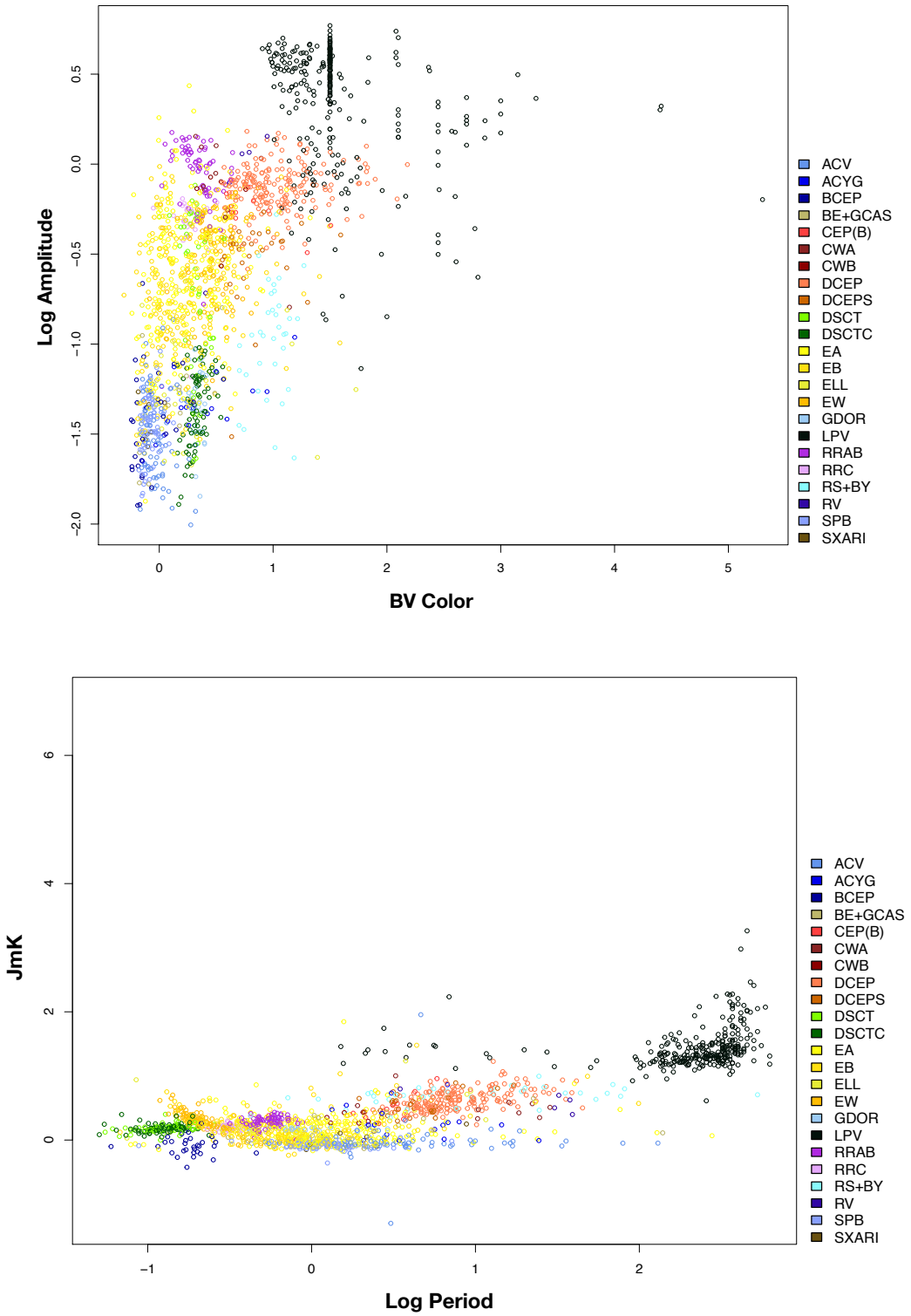


Figure 3.3: 2-D plot of different combinations of attributes. Each of the colors represent a variable star class or type. (Top) log (amplitude) plotted against the color index, B-V Color. (Bottom) JmK plotted against log (period).Also the type refers to the variable star types. Attribute descriptions are found in Table A.1

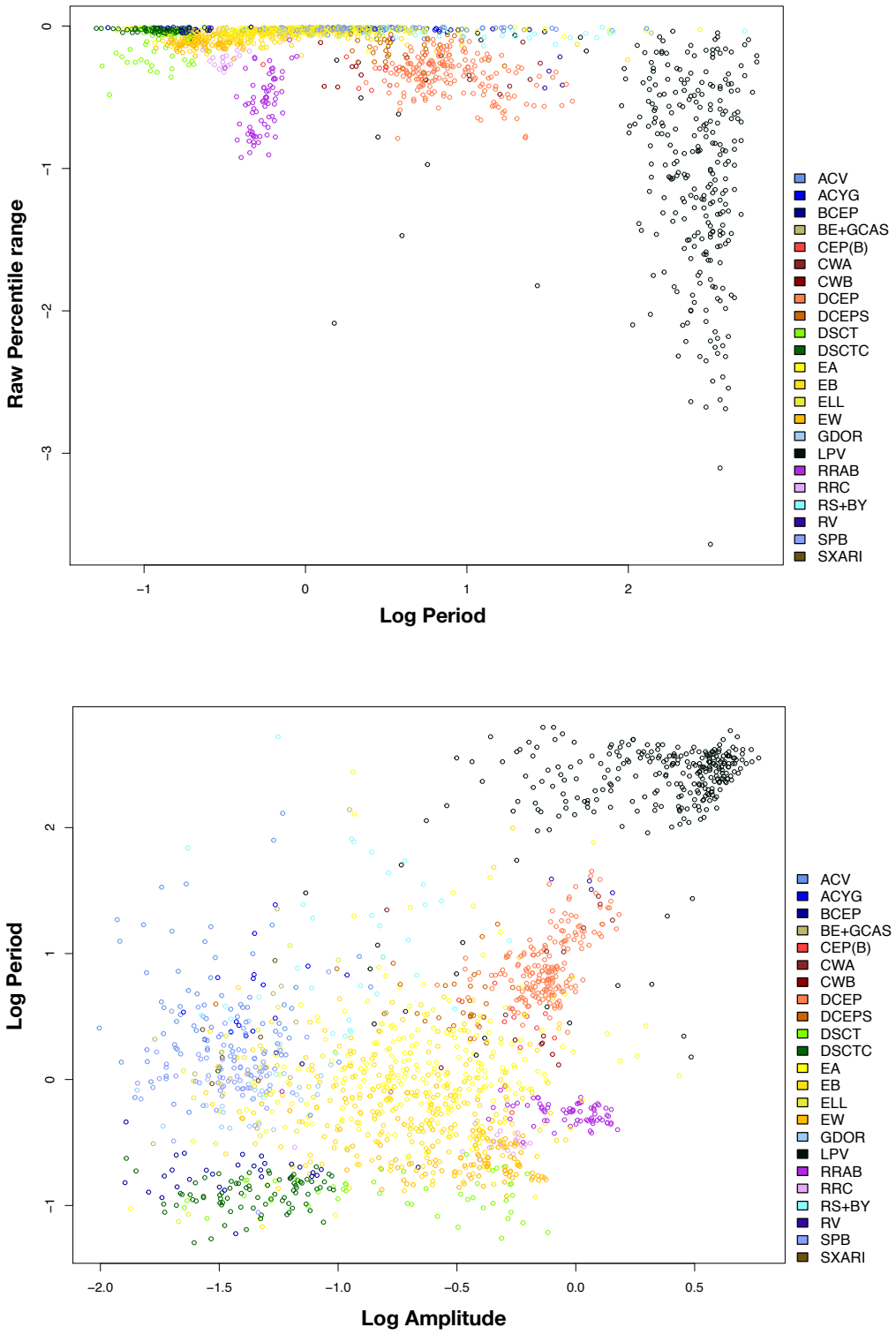


Figure 3.4: (Top) Raw percentile range plotted against the decadic log of the period. (Bottom) Decadic log of the amplitude plotted against the decadic log of the period. Here we see that the log (period) plotted against log (amplitude) provide an acceptable segregation between many of the classes. Also the type refers to the variable star types. Attribute descriptions are found in Table A.1.

Dealing with multicollinearity

Figure 3.5 gives a correlation plot of all these star attributes that we have shortlisted so far. Some of the attributes are highly correlated like JmK, JmH, V-I Color and B-V Color. These are all attributes that indicate the color of the star, and their correlation is due to the fact that the spectrum of the stars can roughly be approximated by a black-body spectrum.

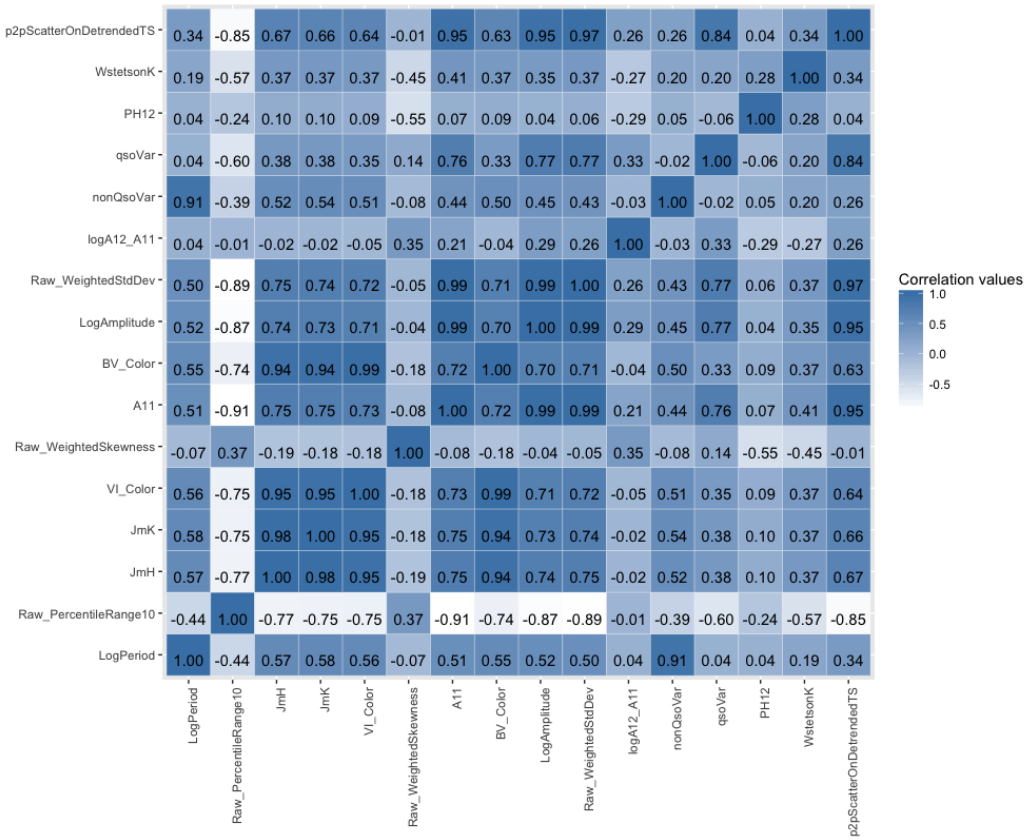


Figure 3.5: Correlation plot of the first 16 attributes selected by attribute importance.

It should be noted that correlations between some of the attributes are expected and even though they are correlated each of these attributes contain pieces of information that can be useful in the classification. However, It could also be useful to investigate the minimum set of attributes that are uncorrelated, if we ignore the astronomical insight. In this regard, we used the attribute-rank list and selected the first 8 uncorrelated attributes as follows.

- **Step 1 :** Consider the first ranked attribute A_1. It automatically becomes a part of the final set of attributes. We check the correlation of A_1 with the other attributes.

- **Step 2 :** We check the correlation for every attribute with A_1, i.e. $cor(A_1, A_d)$ where d is the attribute indicator. We delete those d for which $cor(A_1, A_d) > 0.8$. Hence we form the subset of the attributes.
- **Step 3 :** In this new list, we take the next ranked attribute and repeat **Step 2** for this attribute.
- **Step 4 : Steps 1 to 3** continues till we get the final set of uncorrelated attributes.

Our uncorrelated list of attributes are Log Period, JmK, Percentile range 10, raw weighted skewness, log A12_A11, qsoVar, wstetsonK and PH12 (Attribute descriptions are found in Table A.1). However a word of caution. If we use the above procedure to take only the uncorrelated attributes, this deletes many of the attributes that are known to be astrophysically meaningful. For instance, B-V color have been removed and it is an important attribute. Nevertheless, it will be interesting to see how the model performs with just these attributes in Chapter 4.

3.4.2 Selected attributes

Hence our 16 most important attributes are defined below in the order of importance. These are Log(Period), JmK , Raw_PercentileRange10, V_I Color. Raw_WeightedSkewness , BV_Color, A11, Log(Amplitude), Raw_WeightedStdDev, nonQsoVar, logA12_A11, qsoVar, WstetsonK, PH12 and p2pScatterOnDetrendedTS. The descriptions of these attributes are found in Table A.1.
Dubath et al. (2011) lists 14 most important attributes and it will be worthwhile to compare their list with ours.

A quick comparison of the lists of attributes chosen by Dubath et al. (2011) and the list chosen in our thesis is given in Table 3.3. Though we have used only one attribute that summarize the different steps of modeling of the light curves, we have used more attributes that describe the physics of the stars. The attributes that summarize the distribution of the observed magnitude of the light curve have been included more in the list and also the attributes that explain the stochastic variability of the light curves. The authors of Dubath et al. (2011) have added a few more attributes and more about it can be read in Dubath et al. (2011).

3.5 Synopsis

In this chapter, we emphasized the importance of quality data for statistical learning. We showed that our training data-set represents the best available knowledge about the variable star classes that will be used in

Table 3.3: Comparison of the attribute lists of this thesis and of Dubath et al. (2011)

Attribute type	**Attributes**	**Used by Dubath et al. [2011]**	**Used in this thesis**
Attributes that summarize the improvements after different steps of modelling of the light curves	p2pScatterOnDetrendedTS p2pScatterOnFoldedTS scatterOnResidualTS	All of them	Only p2pScatterOnDetrendedTS
Attributes related to astrophysics of the star	absGlat Glat and Glon Parallax Absolute_Mag00 BV_Color VI_Color JmK, JmH and HmK	Absolute_Mag00 V_I Color	Absolute_Mag00 VI_Color BV_Color JmK JmH
Attributes that summarize the distribution of the observed magnitudes	Raw_WeightedStdDev Raw_WeightedSkewness Raw_WeightedKurtosis Raw_PercentileRange10 StetsonJ stetsonJweighted stetsonK WstetsonJ WstetsonJweighted WstetsonK	Raw_WeightedSkewness	Raw_WeightedSkewness Raw_PercentileRange10 Raw_WeightedStdDev WstetsonK
Attributes that quantify the strength of a stochastic variability of the light curve	logPnonQso logPqso qsoVar nonQsoVar	None	nonQsoVar qsoVar
Attributes related to the period search and harmonic modeling of the light curve	LogPeriod LogAmplitude HarmNum A11, A12, A13, A14, A15 PH12, PH13, PH14, PH15 logA11minusA logA12_A11 logA13_A12	LogPeriod LogAmplitude PH12 logA12_A11	LogPeriod LogAmplitude PH12 logA12_A11
Other attributes	Percentile90 P2P slope	All of them	None

the next chapter. We also selected the 16 best attributes for classification by an importance measure. This set of attributes form will be used in Chapter 4 to reduce the dimension of the data.

We listed out the set of uncorrelated attributes. We will be using these attributes for classification and we will also compare the performances of classification when each of these two attribute sets are considered.

Chapter 4

TSDM model

4.1 Introduction

In Section I, we discussed why we need to classify different types of variable stars. In Chapter 3, we explained the composition of our training data-set and showed its reliability and quality. This is necessary in the development of an effective statistical model. After these preliminaries, in this chapter we present the first statistical contribution of our thesis, namely the Two Stage Dirichlet Mixture Model (TSDM model hereafter). The TSDM model is a supervised classification model. As we briefly mentioned in Chapter 3, a supervised classification model is a model that analyzes the training data and produces an inferred or a trained model which can be used to classify new data-points into classes. We will discuss this in detail below.

The discussion in this chapter is divided into sections as follows. Since the TSDM model uses is the Dirichlet distribution (Section 1.2.3), the data-points needs to be transformed into the open simplex (Section 1.2.3). We present this in Section 4.2, along with the model definition. Section 4.2 discusses about the stages of the TSDM model and parameter estimation. Section 4.3 is devoted to the data studies after classification using the TSDM model and we address some questions as in, why we decided to use the Dirichlet distribution and if we can find sub-classes with our model, in the end of Section 4.3 including comparison with the random forest classifier (Section 1.2.4) in Dubath et al. (2011).

4.2 Model

Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ denote a random sample of size n, where $\mathbf{Y_i}$ is a D-dimensional random vector with probability density function $f(\mathbf{y_i})$ on $\mathbb{R}^D$. Let the entire sample be represented by $\mathbf{Y^T} = (\mathbf{Y_1}, \ldots, \mathbf{Y_n})^\mathbf{T}$, where the superscript T denotes vector transpose. Thus $\mathbf{Y}$ is an n-tuple of points in $\mathbb{R}^D$ and is an $n \times D$-dimensional matrix and $\mathbf{y^T} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^\mathbf{T}$ denotes an observed sample where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ is the D-dimensional observed value of the random vector $\mathbf{Y_i}$.

The K-component TSDM model can be written in the form

$$f(\mathbf{y_i}) = \sum_{k=1}^{K} \rho_k \sum_{j=1}^{J_k} \frac{\pi_{kj}}{\mathbf{B}(\boldsymbol{\alpha_{kj}})} \prod_{d=1}^{D} y_{id}^{\alpha_{kjd}-1} \quad \mathbf{y}_{i-D} \in \mathbb{V}_{D-1} \tag{4.1}$$

where $\mathbf{y_i}$ and the open D-dimensional simplex $\mathbb{V}_{D-1}$ are defined as in Section 1.2.3. Also, ρ_k and π_{kj} are the such that

$$0 \leq \rho_k \leq 1 \quad (k = 1, \ldots, K)$$
$$0 \leq \pi_{kj} \leq 1 \quad (j = 1, \ldots, J_k),$$

where

$$\sum_{k=1}^{K} \rho_k = 1 \qquad \sum_{j=1}^{J_k} \pi_{kj} = 1$$

and for $\boldsymbol{\alpha_{kj}} = (\alpha_{kj1}, \alpha_{kj2}, \ldots, \alpha_{kjD})^T$

$$\mathbf{B}(\boldsymbol{\alpha_{kj}}) = \frac{\Gamma(\alpha_{kj1})\,\Gamma(\alpha_{kj2}) \cdots \Gamma(\alpha_{kjD})}{\Gamma(\alpha_{kj1} + \alpha_{kj2} + \cdots + \alpha_{kjD})}.$$

We use a D-dimensional labeled data-set $\mathbf{Y_i}$ as our training data-set which has K classes or types of variable stars (according to Table 3.1). Then we fit the training-data to our model specified in Equation (4.1) in two stages as the name suggests, using maximum likelihood estimation.

In the first stage, each of these K types of variable stars of the training data-set is fit with a finite mixture of Dirichlet densities i.e. a J_k-component mixture of Dirichlet densities for the kth variable star type, $k = 1, \ldots, K$. Unsupervised classification is used to fit the data of each of the variable types which is detailed in Section 4.2.1 of this chapter. The component densities of these mixtures are called inner mixture components and their mixing proportions are called inner mixing proportions or weights π_{kj} for the kth variable star type and the jth inner mixture component of the class.

In the second stage, these K J_k-component mixture of Dirichlet densities, become mixture components of an ensemble (or mixture) of all these mixtures, which we call the outer mixture. The entire model is called the two stage Dirichlet mixture model (TSDM) model, which is represented by Equation (4.1). The mixing proportions of these mixtures are called outer mixing proportions or weights and the components of the mixture are called outer mixture components (ρ_k is the outer mixture proportion or weight for the kth variable star class or outer mixture). An illustration of a 2-component TSDM model with $J_1 = 4$ and $J_2 = 3$ can be found in Figure 4.1. It should be noted that even though in Equation (4.1) we

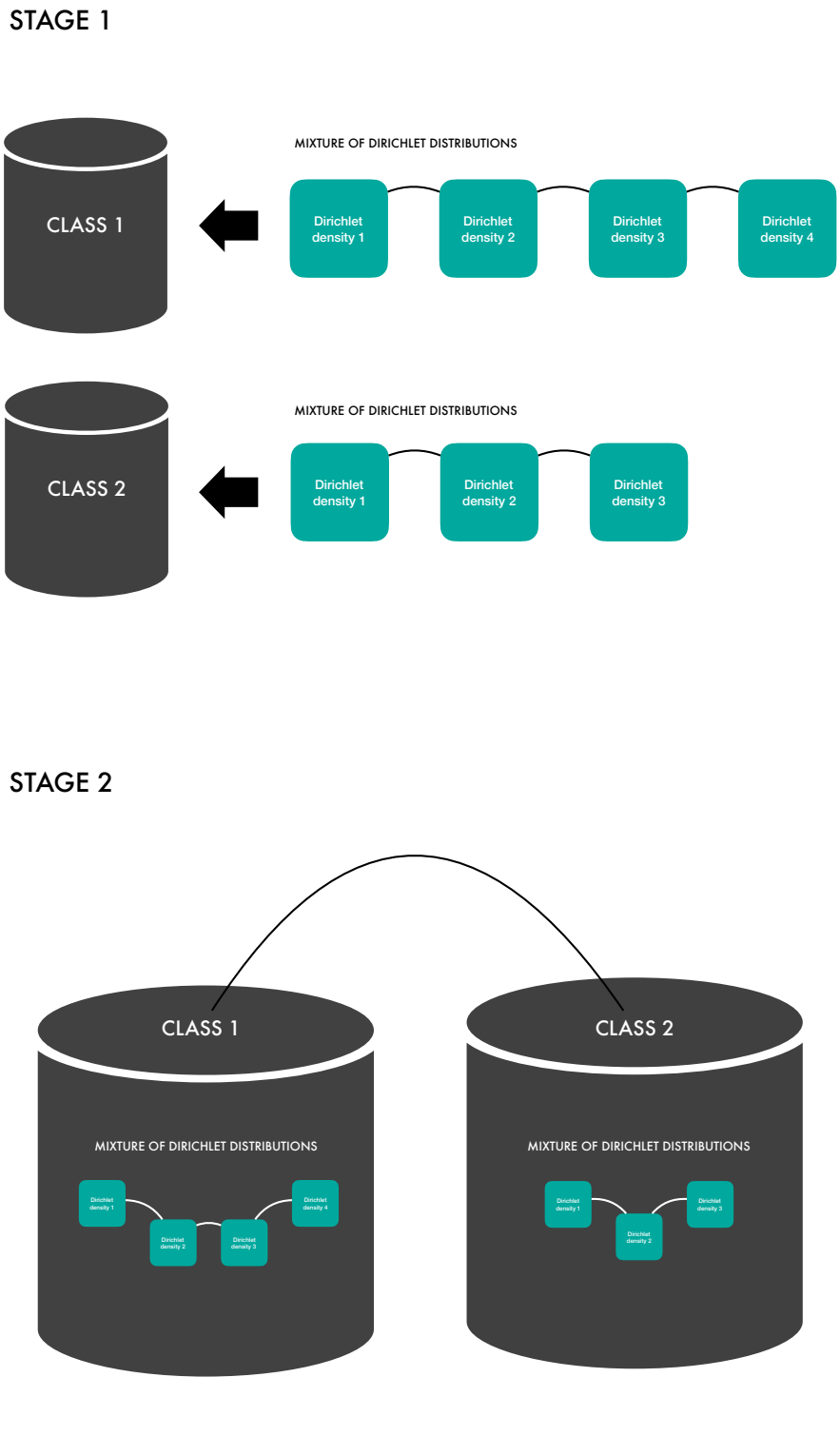


Figure 4.1: Illustration of the two stages of the TSDM model with the first stage positioned at the top and the second stage positioned at the bottom of the figure. In the first stage, class 1 and class 2 are modeled into a mixture of 4 Dirichlet densities and a mixture of 3 Dirichlet densities respectively. In the second stage, a mixture of class 1 and class 2 is formed to get the TSDM model.

have used the density of Dirichlet distribution to suit the needs of variable star classification, any multivariate density can be used in practice. Later in Section 4.3.5, we will use multivariate Gaussian densities as the components of the inner mixture and compared it with Dirichlet distributions. Also, as mentioned in Chapter I we decided to use the probability scale so as to transform the data into a common scale of reference and Dirichlet distributions can be considered as a natural choice for modeling probabilities.

We present now the stages of the TSDM model.

4.2.1 First stage

We discussed earlier that each of the variable star types are fit to a mixture distribution, in the first stage. However, since Dirichlet densities form the components of the mixture distribution, we need to transform the data to to take values in the open simplex (Section 1.2.3), before fitting it to the mixture.

This transformation is done in two steps. In the first step, the training data-set is transformed into the probability scale and in the second step the resulting transformed data-set is once again transformed to belong to an open-simplex, $\mathbb{V}_{D-1}$. We refer to the final transformed data-set as the simplex-transformed data-set, in our thesis. Let us look at these steps in detail.

Transformation to the probability scale

Consider our data-set $\mathbf{Y}$, which is an n-tuple of points in $\mathbb{R}^D$ as defined in Section 4.2. The entire data-set can be treated as D uni-variate column vectors of length n. These vectors will be referred to as attribute-vectors in this thesis.
Each attribute-vectors are first fit to to an empirical distribution (Appendix B.3), which is a step function that jumps $1/n$ in height for every observation. We then logit transform the empirical distribution, which is a step function that jumps $1/n$ in height for every observation, and plot the sorted attribute-vector data against the logit-transformed jumps. We fit a cubic smoothing spline to the resulting plot and the transformed attribute-vector is obtained by taking the inverse-logit transform on the y-coordinate values of the predicted values of the spline fit.

More about the logit transform, the inverse logit transform are found in Appendix B.4. An illustration of the transformation into the probability scale for the attribute, LogAmplitude is given in Figure 4.2. The reason why it is important to transform the data into the probability scale is briefly discussed in Section 4.3.5.

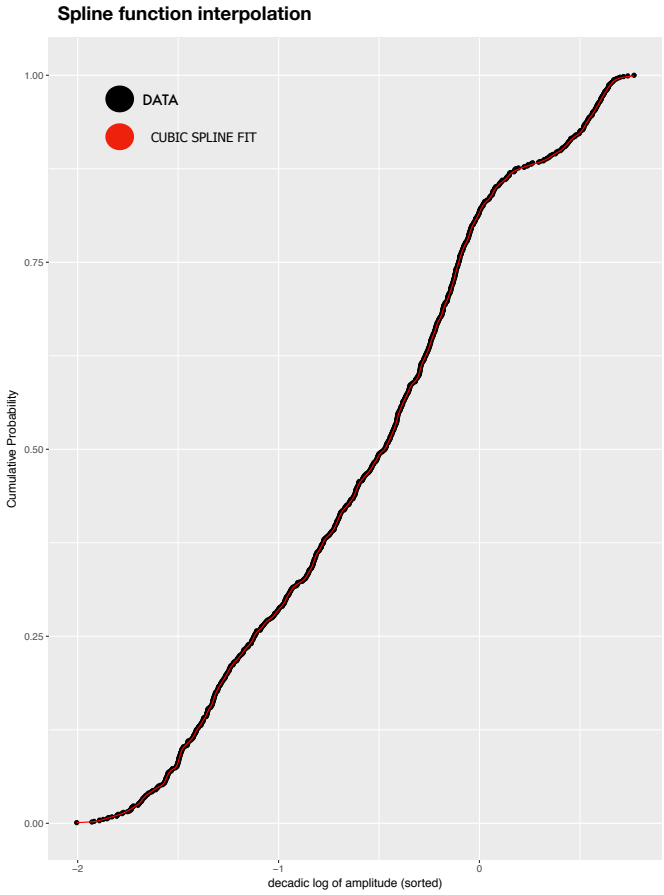


Figure 4.2: The empirical distribution of the sorted LogAmplitude (decadic log of the amplitude, refer to Table 3.1) attribute data. A cubic spline is fit to the empirical distribution indicated by the red line passing through the data plot.

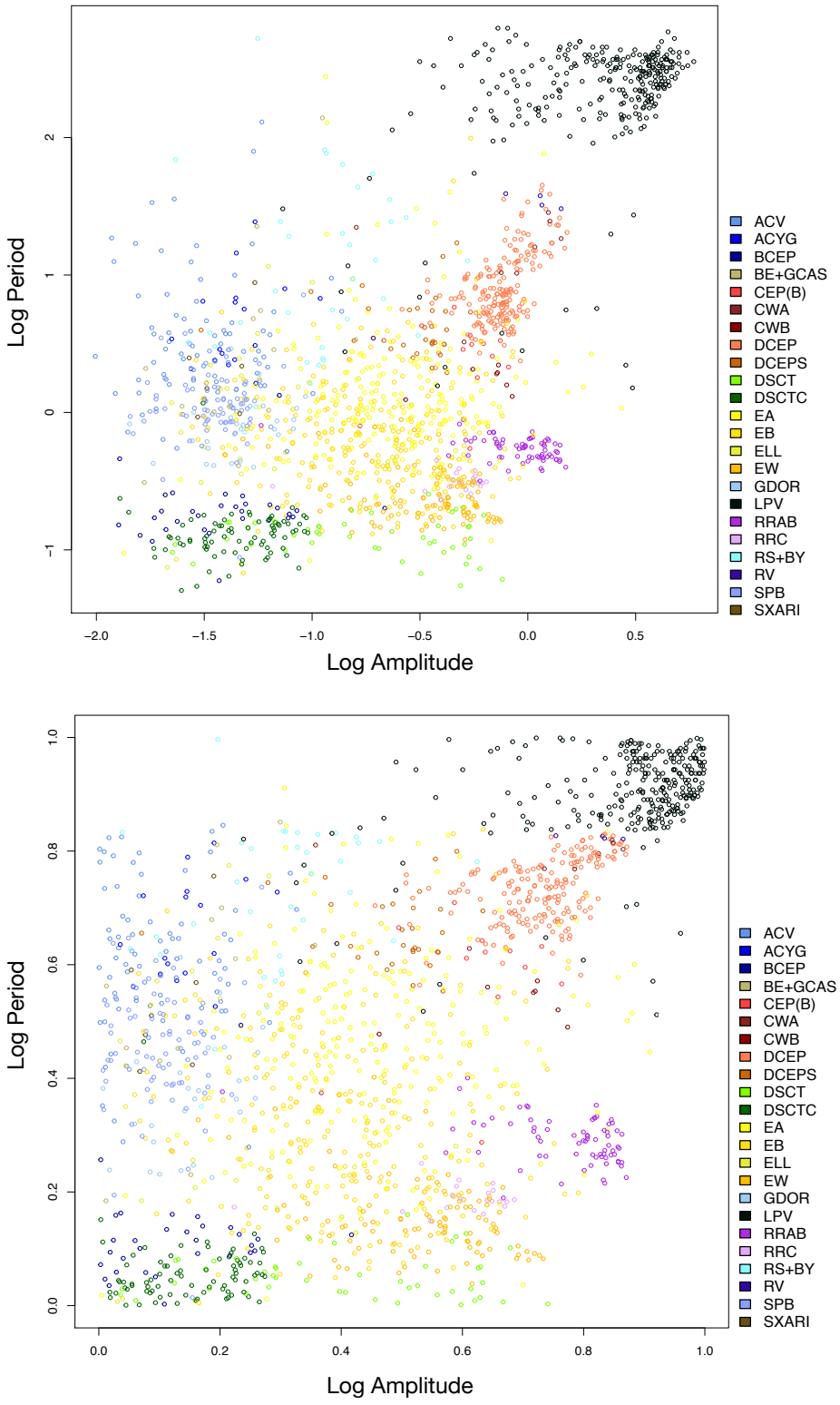


Figure 4.3: (top) The raw data projected onto 2 dimension (decadic logs of period and amplitude) shows reasonable segregation between the classes. (bottom) The data transformed according as Section 4.2.1, into the probability scale. The transformation of log period is not shown here but follows the same procedure.

Transformation to the simplex

Once we transform the attribute-vectors into the probability scale, we have the transformed D dimensional data-set $\mathbf{y} = (\mathbf{y_1^T}, \ldots, \mathbf{y_n^T})^\mathbf{T}$ where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$, for each $i = 1, \ldots, n$.
However, as we discussed earlier, we need to transform the data into a simplex (or simplex transform hereafter), as Dirichlet distributions form the components of our TSDM model. In this thesis we will be using two types of transformations to transform the data into a simplex, which is detailed below. There are two types of transformations that will be considered, namely STT1 and STT2. STT1 is a straightforward and natural choice for transformation while STT2 is a modification of the STT1 transformation. In the below discussions we have shown that STT2 holds a slight advantage over STT1. Let us start our discussion by looking at STT1 first and the advantages STT2 has over STT1.

Simplex transformation type-1 (STT1)

Consider the ith data-vector, $\mathbf{y_i}$. We take the sum of the vector components $(y_{i1}, y_{i2}, \ldots, y_{iD})$ as follows

$$y_i{}^{(\text{sum})} = \sum_{d=1}^{D} y_{id}$$

Each of the vector components is then divided by $y_i{}^{(\text{sum})}$ to get the simplex transformed data-vector. The same is done for the other $n-1$ points to get the transformed data-set. The transformed data-set will be denoted by $\mathbf{Y}$ as well, in this thesis. Figure 4.4 compares the STT1 transformed data against the probability transformed data, in a 2-D projection. In STT1, we see that the variable star class with high values of the decadic logs of period and amplitude namely LPV (top right of the blue-shaded region in Figure 4.4) is well segregated in the raw data. However when the data is simplex transformed, we notice a distortion in the data, when projected onto the two dimension, and this is because of the following reason.
STT1 transformation ensures that the sum of the components of each of the data-vectors is equal to one, i.e. $y_{i1} + y_{i2} + \cdots + y_{iD} = 1$, for the ith STT1 transformed data-vector. The two dimensional projection of the same would be $y_{i1} + y_{i2} = 1$, which is a straight line as shown in Figure 4.4 (In the figure the line is represented as $x + y = 1$). So all the transformed data-vectors fall into the region $y_{i1} + y_{i2} < 1$, as we are considering the projection of a 16-dimensional simplex into two dimensions. In effect, the data-vectors in a unit square gets transformed into a right-angled isosceles triangle and hence gets distorted. In Figure 4.5, where the STT1 transformed data is shown in the right, is an enlargement of the STT1 transformed data in Figure 4.4. We see this point illustrated in Figure 4.4 as the data-points (x, y) such that $x + y > 1$ are found in classes in

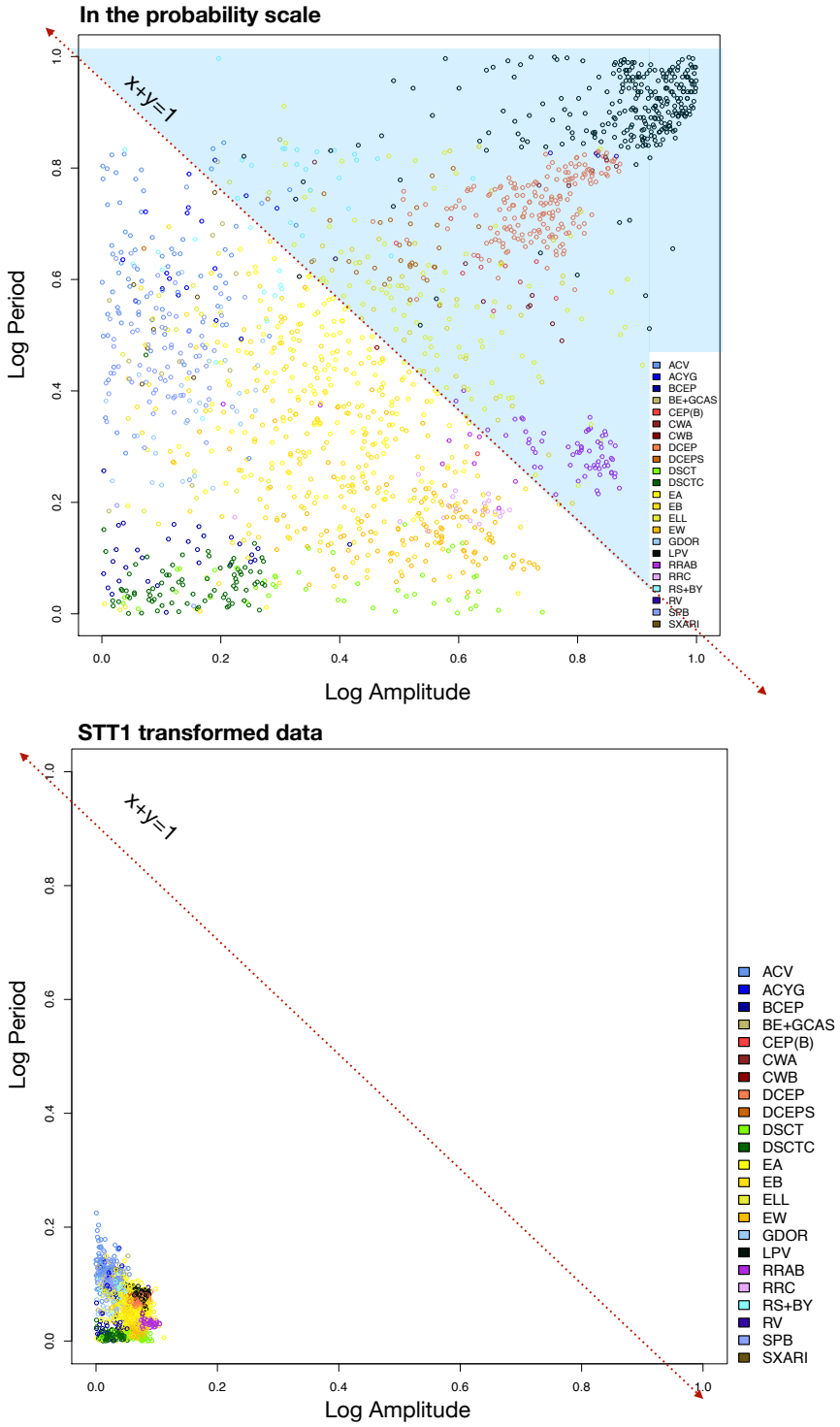


Figure 4.4: Plot which shows the distortion due to compression of our work, with a 2D projection of the 16 dimensional transformed data.(Top) Two-dimensional projection of the data-set transformed to probability scale. The 2-dimensional projection of transforming the data into the simplex using STT1 would be to confine the data from the blue-shaded region to the unshaded area. (Bottom) This is the simplex transformed STT1 data and apparently the structure of the data is distorted due to the compression). The compression in data is shown in Figure 4.5.

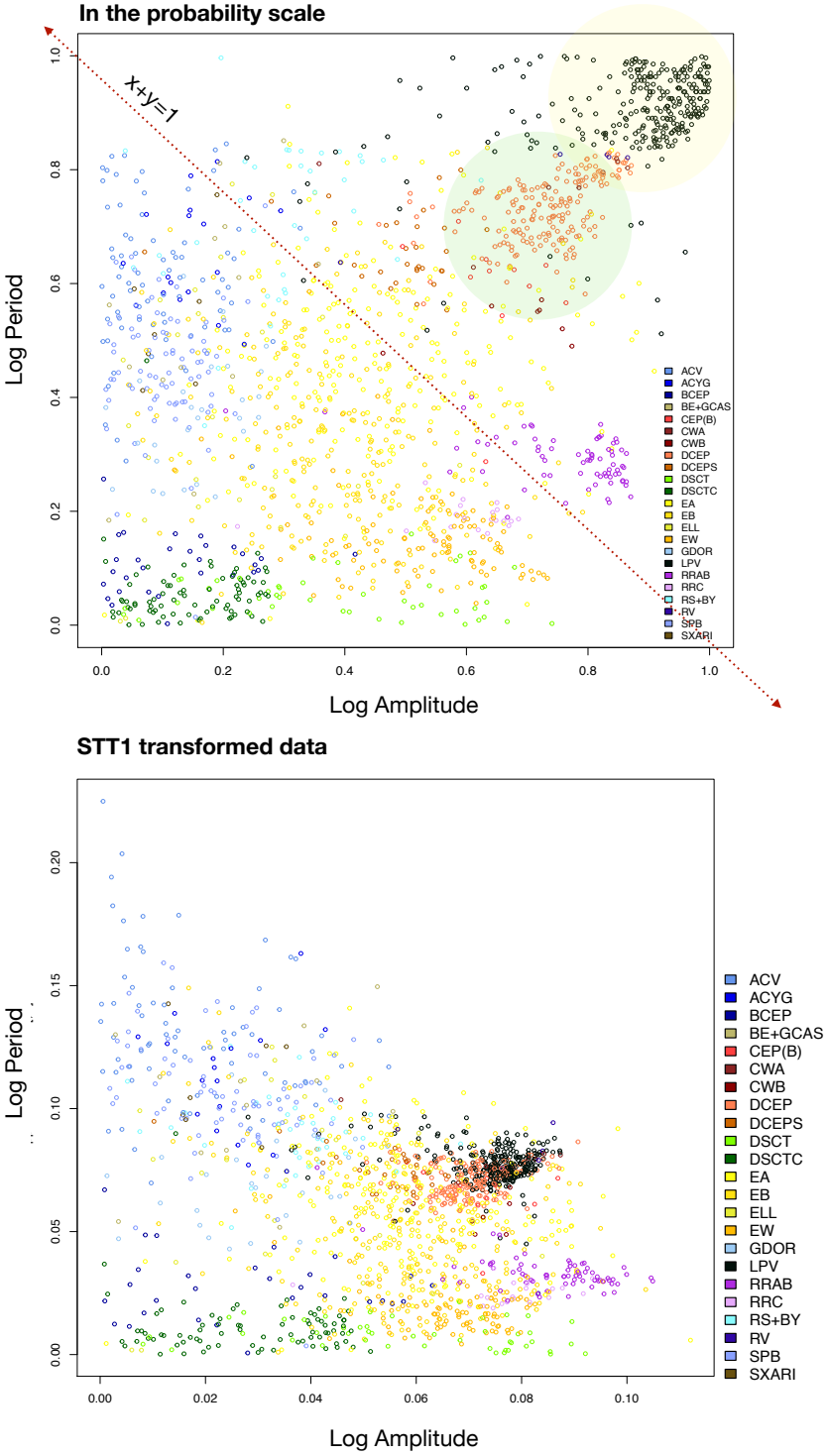


Figure 4.5: (Top) The probability scale transformed data needs to be transformed to the simplex and for the 2-D projection in this case, within $x + y < 1$. The blue and green shaded region are variable type classes which are clearly outside $x + y < 1$. For data within the region $x + y < 1$, the data is transformed proportionally, while for data outside the region, it gets compressed. Hence the classes shaded in green in blue seem distorted. This is shown in the plot on the (Bottom)

the yellow and green shaded region. However, as they are transformed, the structure of these classes seem distorted due to the compression.

Simplex transformation type-2 (STT2)

We observe the problem of distortion when we transform the data to the simplex. Though this wouldn't necessarily mean that the classes aren't well segregated in the 16 dimensional STT1 transformed data, it would be worthwhile to transform the data in such a way that the structure is preserved. This is the motivation for the simplex transformation type-2 (STT2 hereafter). The transformation steps are as follows.

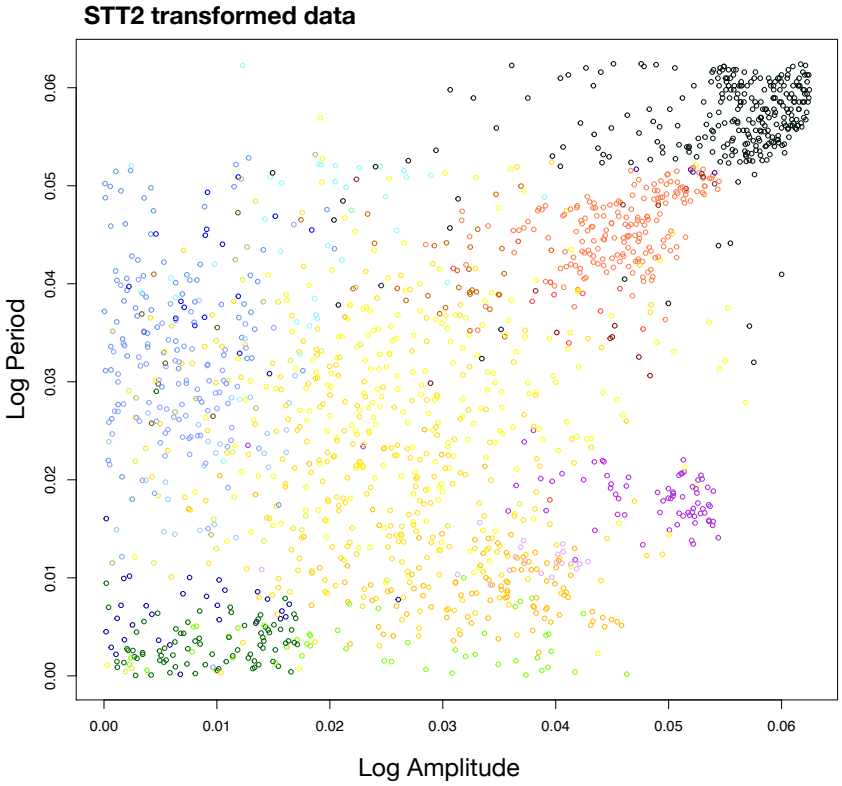


Figure 4.6: For the 2-D projection of our data-set, the structure of the data is preserved by the STT2 transformation. Unlike STT1 the high period and high amplitude classes are not distorted.

Consider the ith data-vector, $\mathbf{y_i}$ and the sum of its components $y_i^{(\text{sum})}$, as defined in the definition of STT1. First $y_i^{(\text{sum})}$ is subtracted from the dimension D, to form a new attribute as follows, $y_i^{(\text{Dummy})} = D - y_i^{(\text{sum})}$. This new attribute is appended to the data-vector to get a $D+1$ dimensional data-vector. Finally, we transform the $D+1$ dimensional data-vector by dividing it by D, $\mathbf{y_i^{appended}}$, by projecting the $D+1$ dimensional data-vector back to D as given below by excluding $y_i^{(\text{Dummy})}/D$.

That is,$\mathbf{y_i^{appended}} = \left(y_{i1}/D, y_{i2}/D, \ldots, y_{iD}/D, y_i^{(\text{Dummy})}/D)\right)$ has been constructed to be on the simplex, and does contain the sum of all probabilities, avoiding distortion.
Like we mentioned earlier, the advantage of STT2 over STT1 is that the sum of the components of the data-vector is preserved and hence the segregation between the classes will not be distorted as in STT1. Figure 4.6 gives the plot of the STT2 transformed data-set, projected in two dimension. In Section 4.3, we compare the results obtained from these two transformations.

Now that the data-set has been simplex-transformed, we continue with the TSDM model. As we will use the simplex-transformed data-set hereafter, we will refer it simply as data-set from now onwards, unless stated otherwise. As we discussed in Section 4.2, we will use unsupervised classification methods to fit each of the K classes of variable stars with a mixture of Dirichlet densities using the Expectation Maximization (EM) algorithm (refer to Appendix B.1 for details).

Estimation of parameters

Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ denote our labeled data-set of size n, and let n_k be the number of data-vectors or points in the kth variable type ($k = 1, \ldots, K$) and $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^{\mathbf{T}}$ be the observed sample where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ is the D-dimensional observed value of the random vector $\mathbf{Y_i}$. Hence for the kth class, our data-set has $n_k \times D$-dimensional data-points and we use the following steps of the EM algorithm.

- **Step 1** : First we set the number of components of mixtures, say J_k (for $J_k = 1$ it is not a mixture as it means only 1 component). Thus in effect, we are using the EM algorithm to fit the data-set to an J_k-component mixture of Dirichlet densities.
- **Step 2** : We then form an initial value vector of the parameters of Dirichlet densities. As we are trying to fit the data-set to an J_k-component mixture of Dirichlet densities, we have J_k Dirichlet parameters and J_k mixing proportions or inner mixture probabilities. They are as follows.

$$\boldsymbol{\alpha_k^0} = (\boldsymbol{\alpha_{k1}^0}, \boldsymbol{\alpha_{k2}^0}, \ldots, \boldsymbol{\alpha_{kJ_k}^0}),$$

where

$$\boldsymbol{\alpha_{kj}^0} = (\alpha_{kj1}^0, \alpha_{kj2}^0, \ldots, \alpha_{kjD}^0)^T$$

and

$$\boldsymbol{\Pi_k^0} = (\pi_{k1}^0, \pi_{k2}^0, \ldots, \pi_{kJ_k}^0).$$

- **Step 3** : We update the mixing proportions as follows. (This is the same as the E step of the EM algorithm (Appendix B.1).)

$$\pi_{kj}^{1} = \frac{f(\mathbf{y_i}|\boldsymbol{\alpha_k^0}, \boldsymbol{\Pi_k^0})\pi_{kj}^{0}}{\sum_{j=1}^{J_k} f(\mathbf{y_i}|\boldsymbol{\alpha_k^0}, \boldsymbol{\Pi_k^0})\pi_{kj}^{0}}$$

- **Step 4** : In this step, the Q function, $Q(\theta_k|\theta_k^{(1)})$ (defined in (Appendix B.1)) is maximized where $\theta_k = \left(\boldsymbol{\alpha_k}, \boldsymbol{\Pi_k}\right)$ and

$$\boldsymbol{\alpha_k^1} = \arg\max_{\boldsymbol{\alpha_k}} Q(\theta_k|\theta_k^{(1)})$$

- **Step 5 : Termination step** : We conclude the iterative process if $Q(\theta_k|\theta_k^{(t)}) \leq Q(\theta_k|\theta_k^{(t-1)}) + \epsilon$, for epsilon below some preset threshold. If the termination step holds, then we say that the EM algorithm has converged. We proceed the steps 1-6 for different initial value sets till convergence and choose the best model from these as follows. Please note that (t) denotes the iteration number

- **Step 6** : Proceed to step 2 and repeat till **Termination step** holds.

- **Step 7 : Computation of BIC** : At convergence, the optimum values, $\boldsymbol{\alpha^{(t)}}_k$ and $\Pi_k^{(t)}$ are used to compute the Bayesian Information Criterion (BIC hereafter) (refer to Appendix B.5).

Also, after convergence for J_k, we choose other values for J_k to and proceed from steps 1 to 7 till convergence, and finally compare the BIC values of the best models (according to Step 7) for each value of J_k and select the lowest BIC values among the different J_k values. The values of J_k is increased till a certain maximum value which is set according to intuition, but it is roughly directly proportional to the number of data-points in class k, i.e. if $n_3 > n_2$ then $\max(J_3) > \max(J_2)$. It has to be noted that we haven't imposed any restrictions on the π_{jk}'s while the EM algorithm searches for the maximum, and we evaluate the maximum for different initial values (Step 2) and J_k values to avoid the algorithm to converge to a local minima.

The same is done for each of the K classes and at the conclusion of the first stage of the TSDM model, we have fit each of the K classes of variable types into a finite mixture of Dirichlet densities. Let us proceed to the second stage of the TSDM model.

4.2.2 Second stage

In the second stage, an ensemble (or mixture) of each of the K mixtures of the first stage is formed by taking a mixture of these mixtures to form the TSDM model as shown in Equation (4.1) with outer mixture probabilities $\boldsymbol{\rho} = \{\rho_1, \rho_2, \ldots, \rho_K\}$. In this stage we update the outer mixture probabilities using a prior distribution on these proportions, explained as follows.

Consider the data-set, $\mathbf{y} = (\mathbf{y_1^T}, \ldots, \mathbf{y_n^T})^\mathbf{T}$ $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$. Let $\mathbf{S} = (\mathrm{S_1}, \mathrm{S_2}, \ldots, \mathrm{S_n})$ be defined as $\mathrm{S_i} = k \Rightarrow \mathbf{y_i} \in \mathrm{k^{th}}$ variable type. Let furthermore $\mathbf{I}$: $(\mathbf{Y}, k) \rightarrow \{0, 1\}$ be the indicator function defined as follows

$$\mathbf{I}(\mathrm{S_i} = k) := \begin{cases} 1 & \text{if } \mathbf{y_i} \in k^{th} \textit{ variable class}, \\ 0 & \text{if } \mathbf{y_i} \notin k^{th} \textit{ variable class}. \end{cases}$$

Finally let $\Theta = \big(\theta_1, \theta_2, \ldots, \theta_K\big)$ with $\theta_k = \big(\boldsymbol{\alpha_k}, \boldsymbol{\Pi_k}\big)$ as defined in the previous section. Then to specify $f(\mathbf{y}, \mathbf{S}|\Theta, \boldsymbol{\rho})$ we have

$$f(\mathbf{y}, \mathbf{S}|\Theta, \boldsymbol{\rho}) = f(\mathbf{y}|\mathbf{S}, \Theta, \boldsymbol{\rho}) p(\mathbf{S}|\Theta, \boldsymbol{\rho}) = \prod_{i=1}^{n} f(\mathbf{y_i}|\mathrm{S_i}, \Theta, \boldsymbol{\rho}) p(\mathrm{S_i}|\Theta, \boldsymbol{\rho})$$

Because $f(\mathbf{y_i}|\mathrm{S_i} = k, \Theta, \boldsymbol{\rho}) = f(\mathbf{y_i}|\theta_k)$ and $\mathrm{P}(\mathrm{S_i} = \mathrm{k}|\Theta, \boldsymbol{\rho}) = \rho_k$, the complete-data likelihood function reads :

$$\begin{aligned} f(\mathbf{y}, \mathbf{S}|\Theta, \boldsymbol{\rho}) &= \prod_{i=1}^{n} \prod_{k=1}^{K} \big[f(\mathbf{y_i}|\theta_k)\rho_k\big]^{\mathbf{I}(\mathrm{S_i}=k)} \\ &= \prod_{k=1}^{K} \big[\prod_{i;\mathrm{S_i}=k} f(\mathbf{y_i}|\theta_k)\big] \prod_{k=1}^{K} \rho_k^{n_k} \\ &= \left\{ \prod_{k=1}^{K} \big[\prod_{i;\mathrm{S_i}=k} f(\mathbf{y_i}|\theta_k)\big] \right\} \times \left\{ \prod_{k=1}^{K} \rho_k^{n_k} \right\} \\ &\propto \left\{ \prod_{k=1}^{K} \big[\prod_{i;\mathrm{S_i}=k} f(\mathbf{y_i}|\theta_k)\big] \right\} \times \mathrm{L}(\boldsymbol{\rho}|\mathbf{S}) \end{aligned} \tag{4.2}$$

Due to the constraint $\sum_k \rho_k = 1$, the group sizes, n_k, $k = 1, \ldots, K$, are not independent. The complete data-likelihood, when regarded as a function of $\boldsymbol{\rho} = \{\rho_1, \rho_2, \ldots, \rho_K\}$, is the density of a Dirichlet distribution and thus the conjugate prior distribution family is again the Dirichlet-distribution, according to Frühwirth-Schnatter (2006) and Bernardo and Girón (1988). Thus, we define the conjugate prior for the outer mixture probabilities $p^{\mathrm{prior}}(\boldsymbol{\rho})$ as follows,

$$p^{\text{prior}}(\boldsymbol{\rho}) \sim \text{Dir}(e_1, e_2, \ldots, e_K), \quad e_i \in \mathbb{V}_{K-1}$$

where the e_i's are set after inputs from scientific collaborators. For more on this, the reader is referred to Chapter 7. In our model we used a non-informative prior which is discussed in Section 4.3.1.
Then the posterior for the outer-mixture probabilities, $\boldsymbol{\rho}$ is given by,

$$\begin{aligned} p^{(posterior)}(\boldsymbol{\rho}|\mathbf{S}) &\propto \text{L}(\boldsymbol{\rho}|\mathbf{S}) \times p^{\text{prior}}(\boldsymbol{\rho}) \\ &\propto \prod_{k=1}^{K} \rho_k^{n_k} \times \text{Dir}(e_1, e_2, \ldots, e_K) \\ &\propto \prod_{k=1}^{K} \rho_k^{n_k + e_k - 1} \end{aligned}$$

Therefore,

$$\boldsymbol{\rho}^{(est)} = \boldsymbol{\rho}|S \sim \text{Dir}(e'_1, e'_2, \ldots, e'_K)$$

where $e'_k = e_k + n_k$ with $k = 1, \ldots, K$
Thus our estimated $\boldsymbol{\rho}$ is $\boldsymbol{\rho}^{(est)} = (\rho_1^{(est)}, \rho_2^{(est)}, \ldots, \rho_K^{(est)})$

Any new data vector $\mathbf{y}_+ \in \mathbb{R}^D$, can be classified to the variable type C_+ where,

$$\text{C}_+ = \arg\max_k \text{P}(\text{S}_+ = k)$$

where $\text{P}(\text{S}_\text{i} = \text{k})$ is such that

$$\text{P}(\text{S}_\text{i} = k) = \frac{f(\mathbf{y}_i|\Theta_k, \text{S}_\text{i} = k)\rho_k^{(est)}}{\sum_{k=1}^{K} f(\mathbf{y}_i|\Theta_k, \text{S}_\text{i} = k)\rho_k^{(est)}}$$

Now let us look into the application of the TSDM model.

4.3 Application and Discussion

In the previous sections, we presented the TSDM model and how each of the stages are estimated. In this section we use this model to fit the data-set we discussed in Chapter 3 — the training-set developed from the Hipparcos catalogue. Table 4.1 gives us a summary of how the test and the training data-set is divided, and the number of variable types under consideration. 70% of the data was considered as training data and 30% of the data as test-data.

Table 4.1: Summary of the test and training data-set

Type	Notation	Freq
Number of data-points in training-set	n	1162
Number of variable types	K	23
Number of data-points in test data-set		499

4.3.1 Application Case 1

We discussed in the previous sections on the different attribute sets that can be considered, like correlated and uncorrelated (Section 3.4.1) and the different types of simplex-transformations, STT1 and STT2 (Section 4.2.1). Lets see how our model works for each of these cases.

Below we list out the model conditions for the first case,

Model to which the data are fit	TSDM model
Attributes chosen	Attributes correlated : 16 most important attributes of Section 3.4
Prob scale to simplex transformation	STT1

In the first stage of the TSDM model, the STT1 transformed data for each class are fit to a finite mixture of Dirichlet densities using the unsupervised learning methods. But how good are these fits? One way to check the goodness of the fits, is by comparing the mean values of the estimated inner Dirichlet components with the data, which we have used in our thesis. In this thesis we refer to these mean values of estimated inner Dirichlet components as Dirichlet signatures.
For the kth variable type and jth inner mixture component, the Dirichlet signature vector of the fit is defined to be the D-dimensional vector

$$\left(\frac{\alpha_{kj1}}{\sum_{d=1}^{D}\alpha_{kjd}}, \frac{\alpha_{kj2}}{\sum_{d=1}^{D}\alpha_{kjd}}, \ldots, \frac{\alpha_{kjD}}{\sum_{d=1}^{D}\alpha_{kjd}}\right).$$

Figure 4.7 shows the Dirichlet signatures of the fits (for each component of the mixture) for the class LPV. The Dirichlet signature vector plots for the remaining classes can be found in Appendix A.2. We see the data plotted in grey and the Dirichlet signature vectors plotted in purple. Though with the Dirichlet signature vector plot it is difficult to determine how good the fit is, it can be inferred if it is not a good fit. Figure 4.7 doesn't give any indication that the fits are not good, as it seems to represent the mean of the data well.
In the second stage, we use the frequency proportion of $\boldsymbol{\rho}$ defined as below, as the estimate of outer mixture probabilities

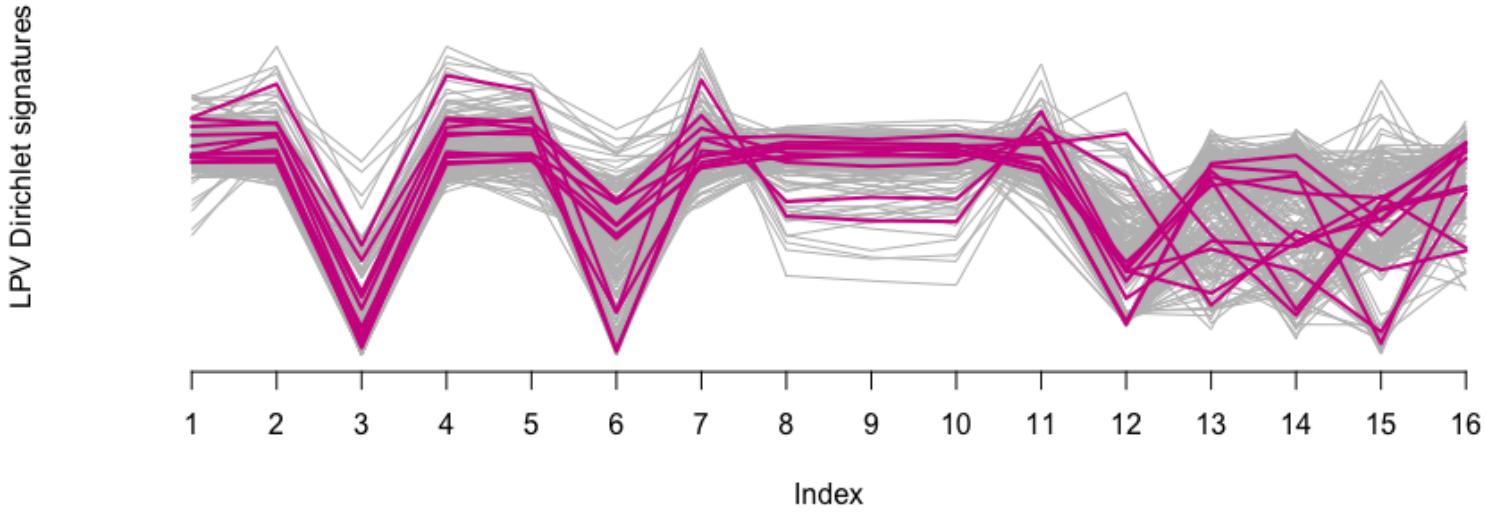


Figure 4.7: The Dirichlet signature vectors for LPV is plotted against its data. We see that the Dirichlet signatures pass through the data-well and hence the Dirichlet parameters for each component, seems to represent the data well

$$\boldsymbol{\rho}^{\boldsymbol{est}} = \left(n_1 / \sum_{k=1}^{K} n_k, \ldots, n_K / \sum_{k=1}^{K} n_k \right)$$

where $n_1, n_2, \ldots, n_K$ are the variable-type frequencies of the labeled data-set for the K classes.

For updating the outer-mixture probabilities, we have used a non-informative prior, $\mathrm{Dir}(1/16, 1/16, \ldots, 1/16)$ (refer to Section 4.2.2) in our analysis. However, the hyper-parameters for the prior set for the outer-mixture probabilities can be set based on expert scientific advice.

Before we present the classification results, we define a few terms, which we shall be using in our thesis, while presenting the results, namely sensitivity and specificity. Sensitivity and specificity are statistical measures of the performance of a binary classification. For each variable type class and each data-vector, the classification can be considered to be a binary classifier, with either the data-vector classified as a member of the class or not. If the data-vector actually belongs to the variable type class then it is called a positive and if not it is a negative. For each class, sensitivity measures the proportion of positive data-vectors that are correctly classified, while specificity measures the proportion of negative data-vectors that are correctly classified as such.

Table 4.2: Classification results for the TSDM model with 16 correlated attributes and STT1 transformed attributes.

Variable type	**Frequency in the training data-set**	**Frequency in the test data-set**	**Estimated number of inner mixtures**	**Sensitivity**	**Specificity**
LPV	201	84	12	0.9047619	0.9638554
EA	162	66	6	0.8333333	0.9422633
EW	74	33	5	0.7878788	0.9763948
DCEP	138	51	5	0.8627451	0.9754464
DSCT	33	14	2	0.4285714	0.9876289
ACV	54	23	3	0.6956522	0.9852941
GDOR	22	5	3	1.0000000	0.9919028
SPB	62	19	3	0.7894737	0.9708333
RRAB	50	22	4	0.9090909	0.9958071
DSCTC	60	21	4	0.8095238	0.9769874
RRC	15	5	4	1.0000000	0.9919028
EB	163	92	9	0.4565217	0.9434889
BCEP	23	7	3	0.5714286	0.9939024
DCEPS	20	11	3	0.8181818	0.9836066
RS+BY	21	14	2	0.3571429	0.9896907
ELL	18	9	3	0.1111111	0.9918367
CWA	6	3	1	0	1
ACYG	12	6	1	0	1
SXARI	3	4	1	0	1
BE+GCAS	8	5	1	0	1
CEP(B)	7	4	1	0	1
RV	4	1	1	0	1
CWB	6	0	1	0	1
Classification accuracy				0.6934	

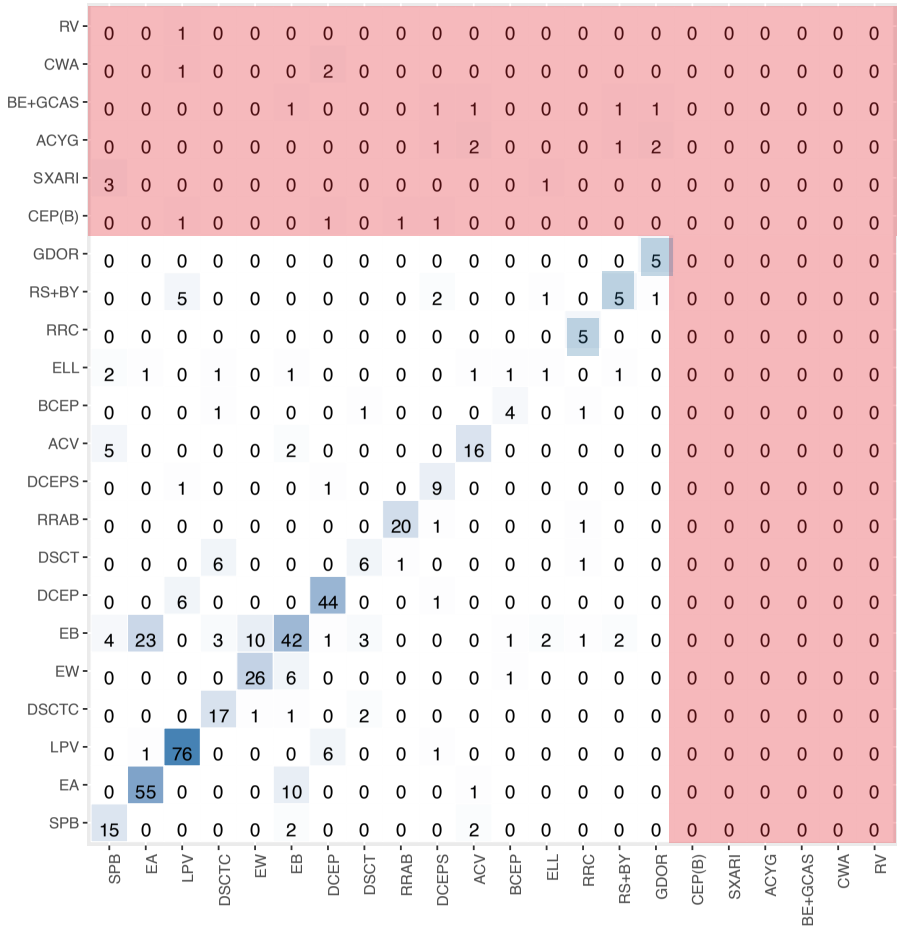


Figure 4.8: The overall classification accuracy is 69.3% but, as we can see, for the classes RV, CWA, BE+GCAS, ACYG, SXARI and CEP(B) (shaded in red), the number of data-points in the training and test data-set is too small, which is why they have high misclassification. However, there is a confusion case which we can observe in the case of EA, EB and EW where they are misclassified among each other. This was also observed in Dubath et al. (2011).

Table 4.2 gives us the sensitivity and the specificity of the classifications, for each class. We tried with different samples to assess the sensitivity of the latter to the choice of the training set: the classification accuracy varied by a mere 3%. We have a classification accuracy of 69.3% for the TSDM model with specifications given earlier. However, for smaller classes (with class frequency of the training-set < 15), the sensitivity is 0. This is primarily because the frequency of the training-set is too low for these classes to be trained efficiently. Also, the number of test-points are also low (< 6). If we exclude these classes, we can correct the classification accuracy as follows,

$$\frac{\left[initial\ overall\ classification\ accuracy \times frequency\ of\ test\ data\right]}{frequency\ of\ test\ data - \sum_{k;n_k<15} n_k}$$

$$= \frac{0.6934 \times 499}{499 - 23} = 0.7268$$

Thus the TSDM has a corrected-classification accuracy of 72.68%.

Classification error analysis

Figure 4.8 gives the confusion matrix for the classification results of the 23 classes. The matrix rows indicate the reference types resulting from the literature survey, or the original data-set, while the columns represent the predicted types.

As we mentioned earlier, the classification accuracy rate is 69.3%. However the model performs better than this, as the confusions or misclassifications within groups of similar stars are less problematic than others. The most important confusion case is that of the eclipsing binaries (Section 2.4.2).

The types EA, EB, EW are eclipsing binaries and there exists a significant confusion within these stars. Dubath et al. (2011) calls them a "difficult case". About 25% of EB data has been misclassified to EA and about 10% to EW. Similarly around 20% of EW has been misclassified to EB and 16% of EA classified to EB in our classification. There are several reasons for this misclassification which are given in Dubath et al. (2011). They note that the light curves of some of the EB and EW stars are quite symmetrical and is similar to that of other variability types. Also the values of the color and absolute magnitude of these stars are the combination of the properties of the two stars, and hence can take almost any value. Hence it is a challenge to correctly classify these types solely on the basis of photometric attributes.

In addition there seems to be misclassifications among other similar classes like DSCT and DSCTC, where about 40% of the DSCT classes were misclassified as DSCTC, the low-amplitude type of DSCT (Section 2.4.1).

4.3.2 Application Case 2

We discussed in Section 4.2.1 that the STT2 transformation maintains the structure of the data, as opposed to STT1 which distorted the data structure in the two dimensional projection of the data. Now, lets see how the TSDM performs when the probability-scaled data is STT2 transformed.

Below we list out the model conditions for the second case.

Model to which the data is fit	TSDM model
Attributes chosen	Attributes correlated : 16 most important attributes of Section 3.4
Prob scale to simplex transformation	STT2

The training and test data-set are the same as the STT1 transformed case (refer to Table 4.1). Table 4.3 gives the classification results, for each variable type.

Table 4.3: Classification results for the TSDM model with 16 correlated attributes and STT2 transformed attributes.

Variable type	Frequency in the training data-set	Frequency in the test data-set	Estimated number of inner mixtures	Sensitivity	Specificity
LPV	201	84	11	0.9643	0.9855
EA	162	66	4	0.8485	0.9492
EW	74	33	5	0.72727	0.96352
DCEP	138	51	5	0.92157	0.97545
DSCT	33	14	2	0.50000	0.99588
ACV	54	23	2	0.60870	0.99160
GDOR	22	5	2	0.800000	0.987854
SPB	62	19	4	0.84211	0.96042
RRAB	50	22	2	0.77273	0.99790
DSCTC	60	21	2	0.95238	0.96862
RRC	15	5	4	1.00000	0.98785
EB	163	92	5	0.45652	0.95332
BCEP	23	7	2	0.571429	0.995935
DCEPS	20	11	3	0.63636	0.98975
RS+BY	21	14	3	0.85714	0.98969
ELL	18	9	2	0	0.993878
CWA	6	3	1	0	1
ACYG	12	6	1	0	1
SXARI	3	4	1	0	1
BE+GCAS	8	5	1	0	1
CEP(B)	7	4	1	0	1
RV	4	1	1	0	1
CWB	6	0	1	0	1
Classification accuracy				0.7134	

The classification accuracy has improved by 2% to 71.34% from the STT1 case. The sensitivity of some of the classes has shown a minor increase, while for some others a minor decrease, but nothing to affirm that STT2 is better than STT1, classwise. The only significant change (rise by 50%) is that of RS+BY, for which the classification accuracy increased

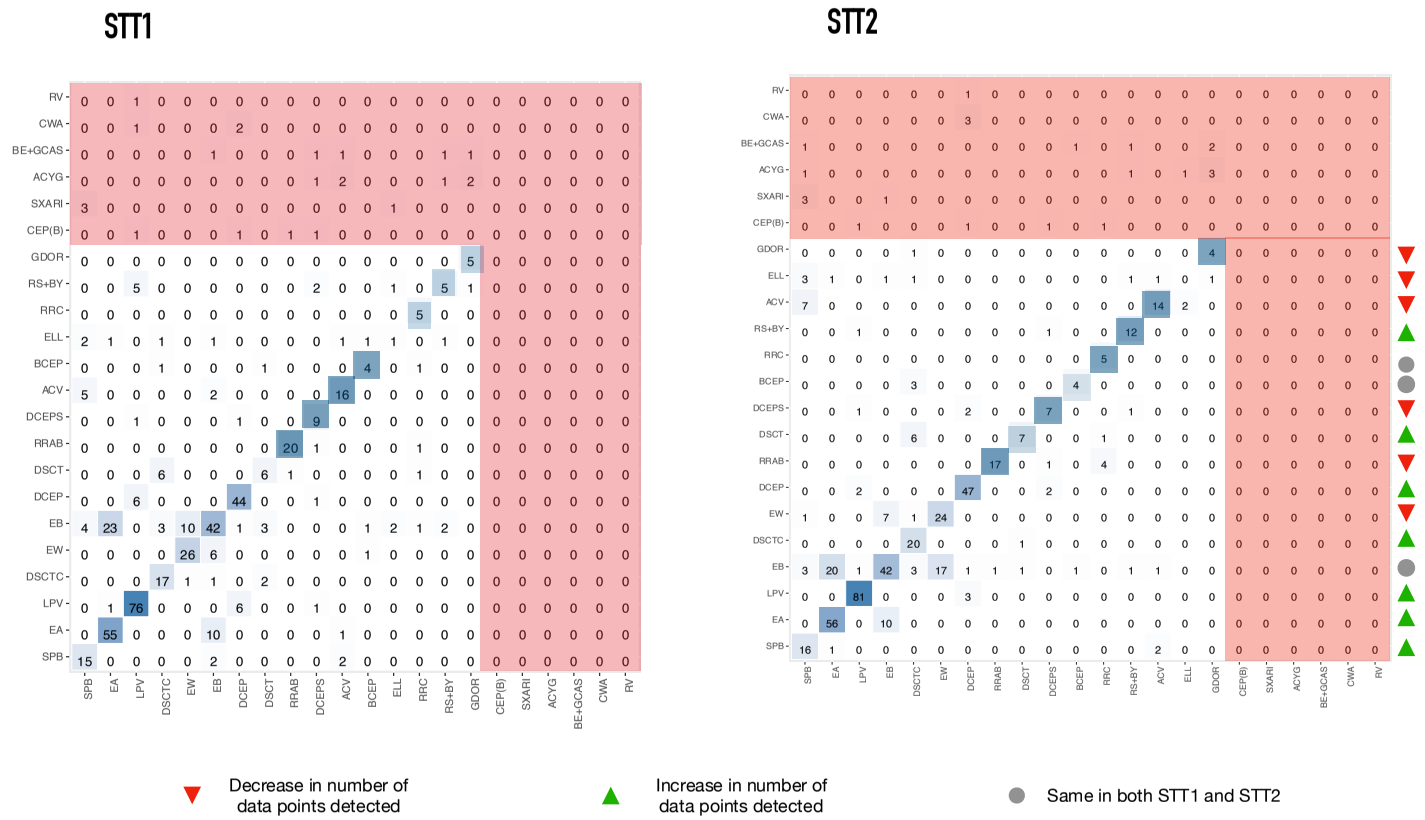


Figure 4.9: (Left) The confusion matrix for the TSDM model with STT1 transformed data. (Right) The confusion matrix for the TSDM model with STT2 transformed data. We can also see the indicators of increase, decrease and similarity in classification accuracies for each class. The classification accuracy for most of the classes or variable types are similar with slight decrease or increase but RS+BY shows an increase of over 50%. But similar to case 1, the eclipsing binaries are misclassified.

from 36% to 86%. Also, the problem of eclipsing binaries which was noticed in the STT1 transformation still exists for this transformation. Figure 4.9 gives a comparison of the confusion matrices of the STT1 transformed and STT2 transformed data-set. The score corrected for the too small classes like Application Case 1 is, 74.78%.

4.3.3 Application : Case 3

In Section 3.4.1, we presented the uncorrelated list of attributes and stated our interest to see how they would perform with the TSDM model. The model conditions have been listed as follows.

Model to which the data is fit	TSDM model
Attributes chosen	Uncorrelated attributes : 8 most important attributes of Section 3.4.1
Prob scale to simplex transformation	STT1

Table 4.4 gives the classification accuracy rate for all the classes with the uncorrelated attributes.

Table 4.4: Classification results for the TSDM model with 8 uncorrelated attributes and STT1 transformed attributes.

Variable type	Frequency in the training data-set	Frequency in the test data-set	Estimated number of inner mixture components	Sensitivity	Specificity
LPV	201	84	10	0.8928571	0.9734940
EA	162	66	3	0.8484848	0.9422633
EW	74	33	4	0.8484848	0.9763948
DCEP	138	51	5	0.9411765	0.9687500
DSCT	33	14	2	0.4285714	0.9917526
ACV	54	23	3	0.6956522	0.9726891
GDOR	22	5	3	1.0000000	0.9919028
SPB	62	19	3	0.6315789	0.9770833
RRAB	50	22	3	0.7727273	0.9979036
DSCTC	60	21	2	0.7619048	0.9874477
RRC	15	5	3	1.0000000	0.9898785
EB	163	92	6	0.5000000	0.9459459
BCEP	23	7	4	0.7142857	0.9939024
DCEPS	20	11	2	0.6363636	0.9918033
RS+BY	21	14	2	0.6428571	0.9835052
ELL	18	9	2	0	0.9897959
CWA	6	3	1	0	1
ACYG	12	6	1	0	1
SXARI	3	4	1	0	1
BE+GCAS	8	5	1	0	1
CEP(B)	7	4	1	0	1
RV	4	1	1	0	1
CWB	6	0	1	NA	0.9979960
		Classification accuracy		70.34%	

The total classification accuracy rate is 70.34% which is 1% more than the classification using correlated attributes. The classification accuracy rates for each of the classes have shown just minor differences, but nothing significant to justify the use of uncorrelated attributes in our analysis. Moreover like we discussed in Section 3.4.1, it is not advisable from a scientific point of view to remove important attributes like amplitude and color indexes. Hence we suggest the use of the correlated list of 16 attributes which we listed in Section 3.4.1. Figure 4.10 gives the confusion matrix for the the classification with these uncorrelated attributes. The score corrected for the too small classes like Application Case 1 is, 73.74%.

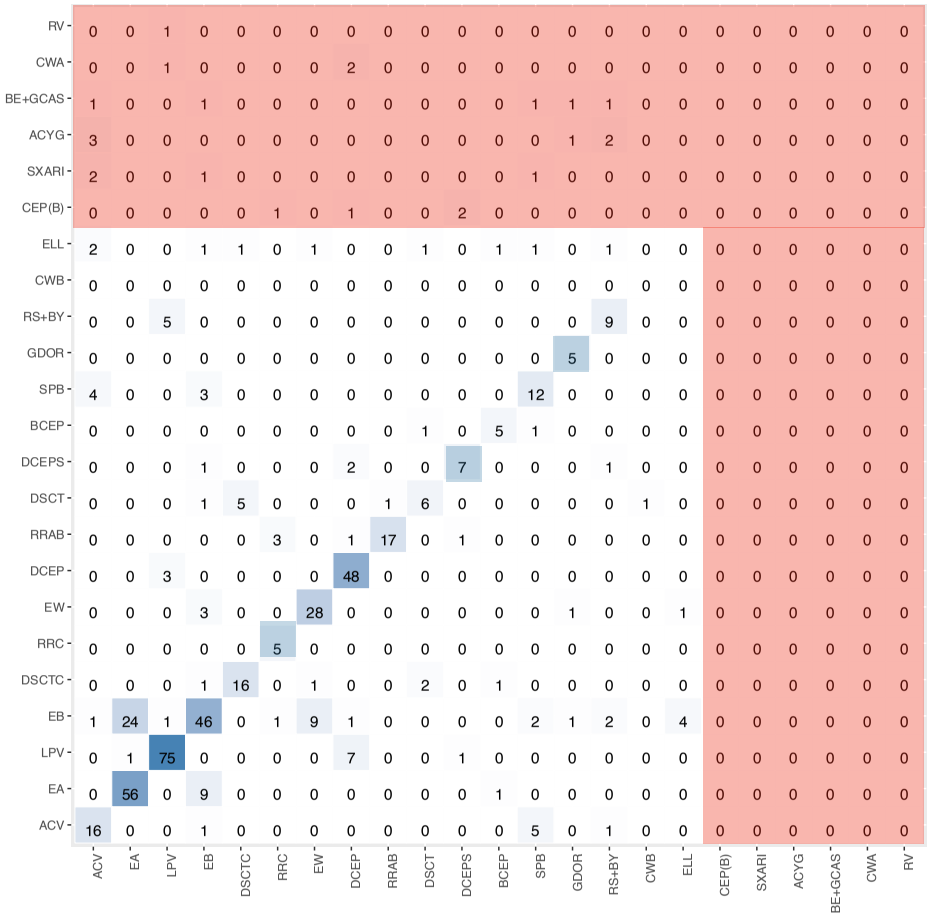


Figure 4.10: The confusion matrix for classification with uncorrelated attributes is similar to Case 1 and 2 (compare with Figure 4.9). The area shaded in red indicates the classes with very low representations and hence has low classification accuracy rates like in Case 1 and Case 2. Also the problem of confusion among the eclipsing binaries exists in this case as well.

4.3.4 Comparison with Dubath et al. (2011)

Dubath et al. (2011) used random forests (Section 1.2.4) for classification based on the 14 attributes which are listed in Table 3.3. Without getting into the details of the attributes, the most important ones which were used in their work were the period, amplitude, V-I color index, absolute magnitude, the residual around the folded light curve model, the magnitude distribution skewness and the amplitude of the second harmonic of the Fourier series model relative to the fundamental frequency. Figure 4.11 gives us the comparison between the confusion matrix of the TSDM model against the model defined in Dubath et al. (2011).

- The overall classification accuracy rate was 84.3% as opposed to 69.3% by the TSDM model.
- Dubath et al. (2011) discusses the difficulty in classifying eclipsing binaries and ellipsoidal variables. In the TSDM model classifications, we encountered the same problem with both STT1 and STT2.
- The DCEP and DCEPS were classified much better with the TSDM model than by the random forest classifier by Dubath et al. (2011).
- With the TSDM model, we have a possibility to extend it to detect new classes which will be discussed in Chapter 5.

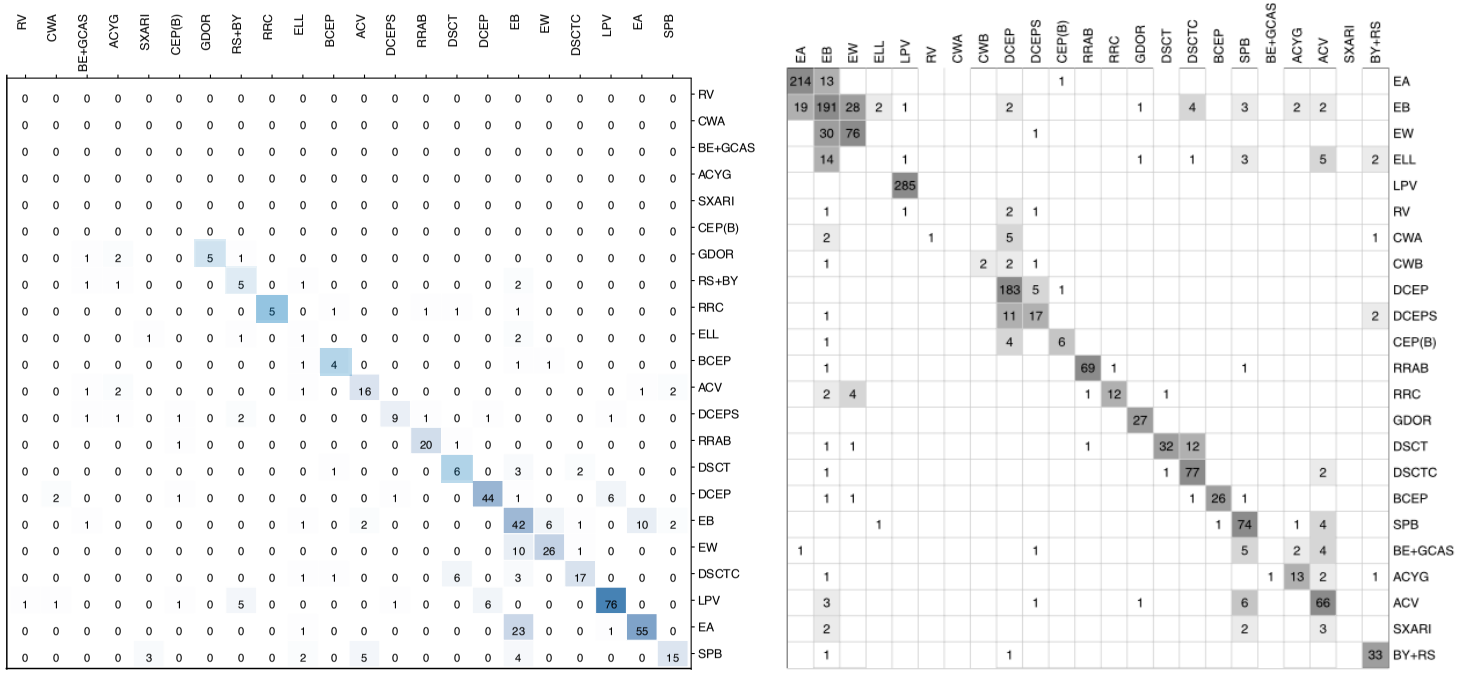


Figure 4.11: Though there is a difference in the number of data-points classified in both the confusion matrices, Dubath et al. (2011) (on the right) seems to have classified some of the classes much better than TSDM (on the left). However the DCEP class was classified much better with the TSDM model. We can see that the confusion case of eclipsing binaries exist for both models.

4.3.5 Dirichlet vs multivariate Gaussian

One might wonder why the Dirichlet distribution is used as opposed to the much widely used and preferred multivariate Gaussian distribution. To understand this, lets fit the data to a two stage Gaussian mixture (TSGM) model which is our two stage model with the multivariate Gaussian as the base distribution. Obviously, since we are dealing with the multivariate Gaussian, we don't have to transform the data to a simplex.
Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ denote a random sample of size n, where $\mathbf{Y_i}$ is a D-dimensional random vector with probability density function $f(\mathbf{y_i})$ on $\mathbb{R}^D$. Let the entire sample be represented by $\mathbf{Y^T} = (\mathbf{Y_1}, \ldots, \mathbf{Y_n})^\mathbf{T}$, where the superscript T denotes vector transpose. Thus $\mathbf{Y}$ is an n-tuple of points in $\mathbb{R}^D$ and is an $n \times D$-dimensional matrix and $\mathbf{y^T} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^\mathbf{T}$ denotes an observed sample where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ is the D-dimensional observed value of the random vector $\mathbf{Y_i}$.

The K-component TSGM model can be written in the form

$$f(\mathbf{y_i}) = \sum_{k=1}^{K} \rho_k \sum_{j=1}^{J_k} \pi_{kj} \frac{\exp\left(-\frac{1}{2}(\mathbf{y_i} - \boldsymbol{\mu_{kj}})^{\mathrm{T}} \boldsymbol{\Sigma_{kj}}^{-1} (\mathbf{y_i} - \boldsymbol{\mu_{kj}})\right)}{\sqrt{|2\pi \boldsymbol{\Sigma_{kj}}|}} \tag{4.3}$$

where $\boldsymbol{\mu_{kj}}$ and $\boldsymbol{\Sigma_{kj}}$ are the D-dimensional mean vector and $D \times D$ covariance matrix of the kth variable type and jth component of the inner mixture. Also π represents the mathematical constant and $|\boldsymbol{\Sigma}| \equiv \det \boldsymbol{\Sigma}$ is the determinant of Σ. The parameters, ρ_k and π_{kj} are the such that

$$0 \leq \rho_k \leq 1 \quad (k = 1, \ldots, K)$$
$$0 \leq \pi_{kj} \leq 1 \quad (j = 1, \ldots, J_k),$$

where

$$\sum_{k=1}^{K} \rho_k = 1 \qquad \sum_{j=1}^{J_k} \pi_{kj} = 1$$

We have used the same data-set as mentioned in Table 4.1, without any transformation. For the first stage, we have used the R package `mclust` to fit each of the variable type classes to a finite mixture of Dirichlet distributions. Table 4.5 gives the comparison of the sensitivities of the TSDM and TSGM models.

The classification accuracy of the TSGM model is 65.53% as opposed to the 69.34% accuracy by the TSDM model with STT1 transformation and 70.34% with the STT2 transformation. Though the overall classification accuracy rate doesn't show a clear advantage for TSDM over TSGM, the classification accuracy for each of the classes does suggest that the TSGM

Table 4.5: Classification accuracy rates : TSDM vs TSGM

Variable type	**Frequency in the training data-set**	**Frequency in the test data-set**	**TSDM sensitivity**	**TSGM Sensitivity**
LPV	201	84	0.9047619	0.9880952
EA	162	66	0.8333333	0.7121212
EW	74	33	0.7878788	0.6969697
DCEP	138	51	0.8627451	0.9607843
DSCT	33	14	0.4285714	0
ACV	54	23	0.6956522	0.7391304
GDOR	22	5	1	0.6
SPB	62	19	0.7894737	0
RRAB	50	22	0.9090909	0.9090909
DSCTC	60	21	0.8095238	0.9523810
RRC	15	5	1	1
EB	163	92	0.4565217	0.5869565
BCEP	23	7	0.5714286	0
DCEPS	20	11	0.8181818	0.18
RS+BY	21	14	0.3571429	0.29
ELL	18	9	0.1111111	0
CWA	6	3	0	0
ACYG	12	6	0	0
SXARI	3	4	0	0
BE+GCAS	8	5	0	0
CEP(B)	7	4	0	0
RV	4	1	0	0
CWB	6	0	NA	NA
	Classification accuracy		69.34%	65.53%

model tends to classify the larger classes (with more than 50 data-points in the training data-set) slightly better like LPV, DCEP, DSCTC, ACV, EB with the exception of some of the eclipsing binaries like EA, EW. On the other hand, 5 classes are not classified at all which were classified well by the TSDM model.

The results which suggests that the TSGM model shows an inclination to better classify larger classes is a disadvantage for our new class detection model in Chapter 5. With the TSGM model having a poor detection history of the smaller classes, this model might fail to effectively detect new classes in the Gaia data-set, because it is likely that the new classes are smaller classes. Also, some of the important variable classes like SPB and DSCT were not classified at all.

This also reaffirms the use of Dirichlet densities over multivariate Gaussian densities, as the base distribution. Also, there are differences in photometric observations across surveys. For instance, the photometric observations of Chi-CYGNI taken from OGLE survey could be different from the Gaia survey. Hence we need to transform the data to a common domain such as the probability scale. Dirichlet densities are a natural choice

to model data in the probability scale and hence, the TSDM model is a viable choice.

4.3.6 Are the different clusters sub-classes?

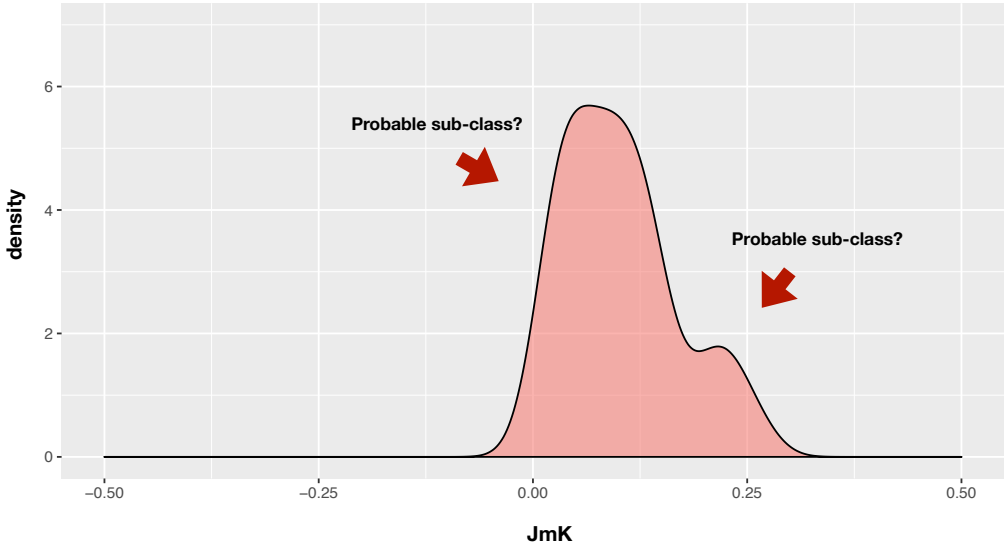


Figure 4.12: This is the plot of the density of the attribute vector JmK of the variable type class SPB. We see that the density plot shows bimodality. Do these suggest sub-classes?.

In the first stage of the TSDM model (Section 4.2), we used unsupervised classification or clustering to classify each of the classes into a mixture of Dirichlet distributions. Table 4.2 lists the number of mixtures (or inner mixtures) for each class. Are these sub-classes?

For instance, Figure 4.13 gives us a plot of the density of the color attribute JmK for the variable type class SPB. The plots suggest a bimodal distribution, possibly. But are these different classes? The SPB class was clustered to three components in our application earlier (Case 1 application, refer Table 4.2).

It is of prime interest for the astronomers to know whether the clustering using Dirichlet mixtures in Stage 1, effectively singles out sub-classes as clusters. To check this, we combined two very similar classes DSCT and DSCTC together and used our unsupervised clustering methods to single each of the classes out. Table 4.6 and Figure 4.13 gives an illustration of the clustering.

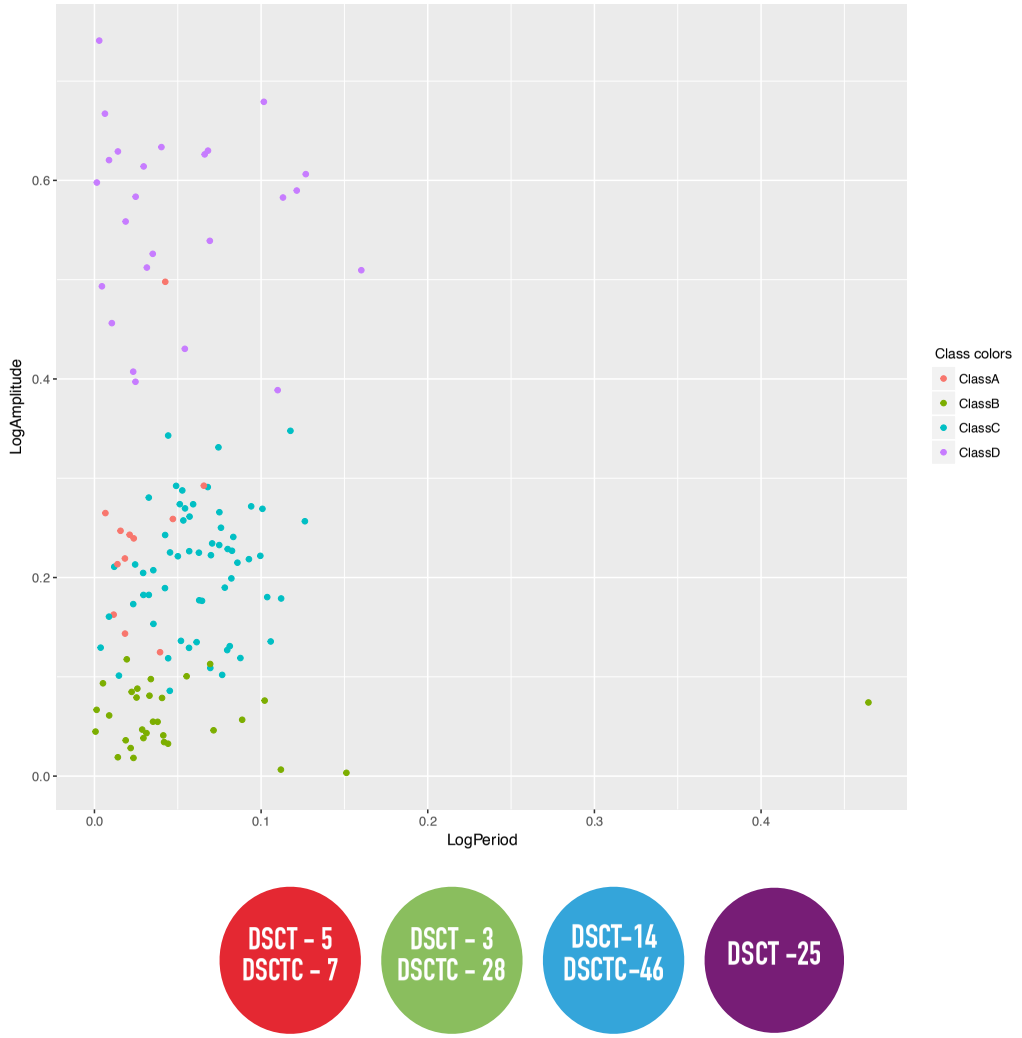


Figure 4.13: The clustering of a combined data-set of delta scutis which have very similar photometric features or attributes. We have clustered the entire combined data-set to 4 clusters.

Table 4.6: Clustering of DSCT and DSCTC into classes. The following table shows how a combined data-set of Delta Scuti's i.e., DSCT and DSCTC combined together performed when the unsupervised classification of Stage 1 was done on this data-set. The algorithm detected a total of 4 classes Class A,B,C and D. The actual class gives in which class the data-vector originally belonged to and assigned class shows us which the cluster assigned by the algorithm.

Actual Class	**Assigned Class**	**Frequency**
DSCT	Class A	5
DSCTC	Class A	7
DSCT	Class B	3
DSCTC	Class B	28
DSCT	Class C	14
DSCTC	Class C	46
DSCT	Class D	25
DSCTC	Class D	0

We see that there are less impurities in each of the clusters. By impurities, we mean the proportion of the classes of lesser frequency in each cluster. In class A, there seems to be about 50% of DSCT and DSCTC approximately, and has high levels of impurity. However, class B detects DSCTC with less than 10% DSCT impurity. Class C detects DSCTC with less than 24% DSCT impurity and Class D detects DSCT with no impurities. Certainly, we need more scientific input to suggest these as subclasses but it is a good point to start.

4.4 Synopsis

We developed a Two Stage Dirichlet Mixture (TSDM) model and used different data transformation techniques to transform the data into a simplex. We found that STT2 performs slightly better than STT1 and also that the uncorrelated attributes perform only 2% better than the correlated attributes. We got a classification accuracy of about 69.3% and compared our results with the random forests classifier in Dubath et al. (2011). Though the classifier of Dubath et al. (2011) had better classification results in the next chapter we will discuss about the extension of our model to detect new classes, which is why it holds an advantage over the random forest methodology. We used the multivariate Gaussian to form a similar TSGM model but found out that TSDM is better than the TSGM model in terms of classifying smaller classes and will also be useful in detecting new classes in Chapter 5.

Now we move to the part where our model holds a clear advantage over the random forest methodology by Dubath et al. (2011), namely new class detection.

Chapter 5

New class detection

5.1 Overview

In Section I, we mentioned that we are using our model as a "prior" (not in the Bayesian sense, but rather a classification rule) for the classification of periodic variable stars (refer to Figure 0.1) in the Gaia data-set. The Gaia survey is expected to provide data from 1 billion objects and will have a time series of photometric data, on average 70 points for each object (Süveges et al. (2017)). With an influx of such a high number of data-points expected from Gaia, it is important to consider the possibility of new classes of variable stars. In the recent years, a lot of research has been done on the possibility of new classes of variable stars. For instance, in recent times more than dozen, previously unknown, short-period variable stars have been discovered (blue large amplitude pulsators, refer Pietrukowicz et al. (2017)). It is likely that there are new variable star types out there waiting to be found. Our training set (Chapter 3) has 23 classes and 1,661 data-points. It will be interesting to extend our model for the detection of new classes.

This leads to the second main contribution of our thesis, the extension of the TSDM model using the semi-supervised new class detection model of Vatanen et al. (2012) to detect outliers and anomalies. Anomaly detection (outlier detection, novelty detection) refers to the problem of detecting patterns in the data that deviate from the usual behavior so much that the arouse suspicion of having been generated by a different mechanism. In our case we consider these anomalies that occur in clusters of reasonable size to be a new class. Though the usual treatment of anomalies is like an error, we treat it is a possible new class.

The discussions in this chapter is divided into different sections as follows. In Section 5.2 we introduce the model for semi-supervised new class detection and its components. In Section 5.3 we explain the expectation-maximization algorithm that is used to estimate the parameters of the model. We verify the performance of the model with an application in Section 5.4, i.e the detection of a new class from our data-set, before concluding the chapter. First, lets look into our model in detail.

5.2 FB model

The fixed background model (FB model hereafter) is a mixture distribution with two components. The first component is called the background model, while the second component is called the new class model.

Once the data is fit to a model with say K classes, we fix the estimated parameters of the model. This model is called the background model and the data is called the background data. Once the model is fit, ideally any data-point that is a representative of any of the K classes will be fit to that model. They look at individual observations as an anomaly if it seems unlikely to have been produced by the process corresponding to the background model (Vatanen et al. (2012)). In other words, the data-points are classified as new class data-points should they fall in the low probability density regions of the data space (Markou and Singh (2003)). Lets discuss each of the components of the model in detail.

Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ denote a random sample of size n, where $\mathbf{Y_i}$ is a D-dimensional random vector with probability density function $f(\mathbf{y_i})$ on $\mathbb{R}^D$. Let the entire sample be represented by $\mathbf{Y^T} = (\mathbf{Y_1}, \ldots, \mathbf{Y_n})^\mathbf{T}$, where the superscript T denotes vector transpose. Thus $\mathbf{Y}$ is an n-tuple of points in $\mathbb{R}^D$ and is an $n \times D$-dimensional matrix and $\mathbf{y^T} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^\mathbf{T}$ denotes an observed sample where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ is the D-dimensional observed value of the random vector $\mathbf{Y_i}$. The detection of new classes among the background can be done in two steps. First, we use parametric density estimation to learn the background model with density $f(\mathbf{y_i})$ using our background data-set. Secondly, we model the unlabeled data with a fixed background model or FB model, $f_{FB}(\mathbf{y_i})$, which is a mixture of the background model and the new class model $f_{NC}(\mathbf{y_i})$.

The fixed background (FB) model can be represented as,

$$f_{FB}(\mathbf{y_i}) = (1 - \lambda) f(\mathbf{y_i}) + \lambda f_{NC}(\mathbf{y_i}) \tag{5.1}$$

where $0 \leq \lambda \leq 1$ is the new-class mixture probability.

The components of the fixed background model or FB model are the background model $f(\mathbf{y_i})$ and the new class model, $f_{NC}(\mathbf{y_i})$. The background model parameters are fixed after they are fit to the background data. When a new data-set is classified, the data-vectors that belong to the classes in the background model will be classified into the background model while any deviation from the distribution of the background will be captured as an anomaly. Thus the FB model is fitted to the unlabeled data by maximizing the likelihood under the constraint that the background model is fixed. Hence the new class model should be able to capture any unexpected deviation from the distribution of the background.

But how do we use it with our model for new class detection? In this thesis, we use the TSDM model (Chapter 4) as the background model and the Hipparcos data (Chapter 3) as the background data. For the new class model, we can use any density or mixture densities, but we have considered a mixture of Dirichlet densities for the same reason that was discussed in the previous chapter (Section 4.3.5). Figure 5.1 gives an illustration of how the FB model works.

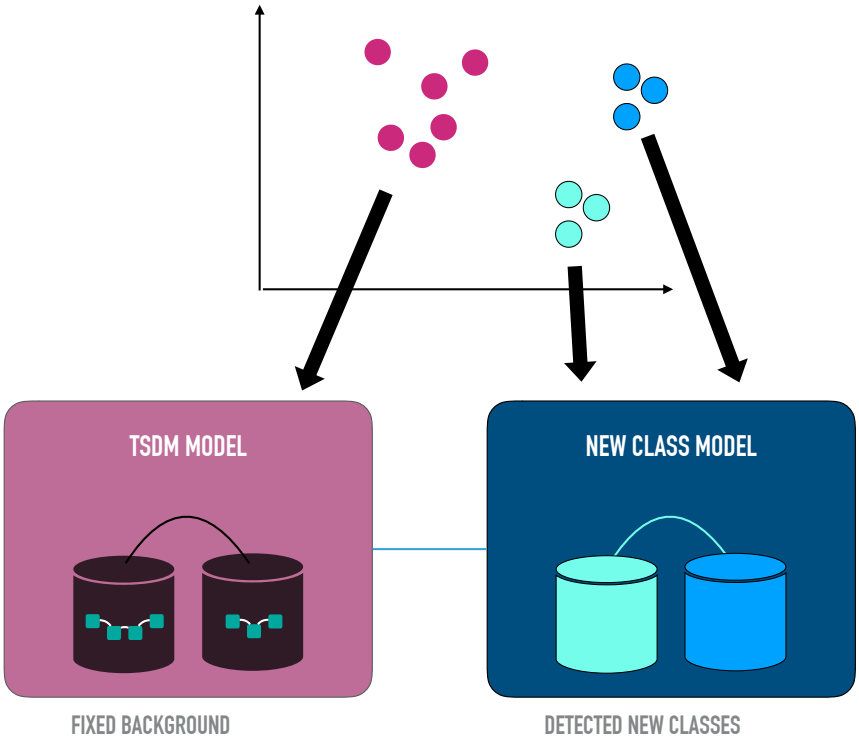


Figure 5.1: The FB model is illustrated as follows. The TSDM model which we discussed in Figure 4.1 are used as the background model and the Hipparcos data is used as the background data. For any new data-set, the data that belongs to any of the classes in the TSDM model is classified into one of the classes in the TSDM model while the data that deviates from the TSDM model is captured in the new class model. They form the new classes.

Thus for detecting new classes in the Gaia data, the FB model with TSDM model as the background model, and a mixture of Dirichlet densities as the new class model can be represented as follows. For the kth component of the new class model

$$\mathrm{Dir}(\mathbf{y_i}|\boldsymbol{\beta}_k) = \frac{1}{\mathbf{B}(\boldsymbol{\beta}_k)} \prod_{d=1}^{D} y_{id}^{\beta_{kd}-1}$$

Thus

$$
\begin{aligned}
f_{FB}(\mathbf{y_i}) &= (1-\lambda)\sum_{k=1}^{K}\rho_k\sum_{j=1}^{J_k}\frac{\pi_{kj}}{\mathbf{B}(\boldsymbol{\alpha_{kj}})}\prod_{d=1}^{D}y_{id}^{\alpha_{kjd}-1} + \lambda\sum_{k=K+1}^{K+Q}\kappa_k \mathrm{Dir}(\mathbf{y_i}|\boldsymbol{\beta_k}) \\
&= \lambda^B\sum_{k=1}^{K}\rho_k\sum_{j=1}^{J_k}\frac{\pi_{kj}}{\mathbf{B}(\boldsymbol{\alpha_{kj}})}\prod_{d=1}^{D}y_{id}^{\alpha_{kjd}-1} + \sum_{k=K+1}^{K+Q}\lambda_k^{NC}\mathrm{Dir}(\mathbf{y_i}|\boldsymbol{\beta_k}) \\
&= \lambda^B f(\mathbf{y_i}) + \sum_{k=K+1}^{K+Q}\lambda_k^{NC}\mathrm{Dir}(\mathbf{y_i}|\boldsymbol{\beta_k}) \qquad (5.2)
\end{aligned}
$$

where $\mathbf{y_{i-D}}$ and the open D dimensional simplex. In our thesis we have used the TSDM model as the choice for $f(\mathbf{y_i})$. Also, ρ_k, π_{kj} and κ_k are the such that

$$
\begin{gathered}
0 \le \rho_k \le 1 \quad (k = 1, \ldots, K) \\
0 \le \pi_{kj} \le 1 \quad (j = 1, \ldots, J_k) \\
\lambda^B = 1 - \lambda, \quad \lambda_k^{NC} = \lambda\kappa_k \\
0 \le \kappa_k \le 1 \quad (k = K+1, \ldots, K+Q) \\
0 \le \lambda_k^{NC}, \lambda^B \le 1 \quad (k = K+1, \ldots, K+Q)
\end{gathered}
$$

where

$$
\sum_{k=1}^{K}\rho_k = 1 \quad \sum_{j=1}^{J_k}\pi_{kj} = 1 \quad \sum_{k=K+1}^{K+Q}\kappa_k = 1 \quad \sum_{k=K+1}^{K+Q}\lambda_k^{NC} + \lambda^B = 1
$$

Note that $\boldsymbol{\alpha_{kj}} = (\alpha_{kj1}, \alpha_{kj2}, \ldots, \alpha_{kjD})^T$ and $\boldsymbol{\beta_k} = (\beta_{k1}, \beta_{k2}, \ldots, \beta_{kD})^T$ are the Dirichlet parameters of the TSDM model and the new class model respectively. For more details on the model refer Vatanen et al. (2012).

5.3 Estimation

In the previous section we discussed how the FB model is fitted. In effect, the model is fitted in two steps. In the first step, the labeled data-set is fitted to the TSDM model like we discussed in Section 4.2. Hence the estimated parameters of the TSDM model are kept constant in the second step.

In the second step, we use the Expectation-Maximization (EM) algorithm to estimate the parameters of the new class model, $\lambda_k^{NC}, \boldsymbol{\beta}_k, \lambda^B$ for $k = 1, \ldots, K$. All the other parameters, $\rho_k, J_k, \pi_{kj}, \boldsymbol{\alpha_{kj}}$ are fixed to the estimated values in the first step. Lets discuss the EM algorithm steps for the second step.

The EM updates for the model in Equation (5.2) are easily found by straightforward analogy to EM algorithm steps for the standard mixture model.

- **TSDM Steps** : Since in this thesis we use the TSDM model as the background model, we first fit the model to the background data, i.e, the Hipparcos data. We estimate the inner mixture probabilities and Dirichlet parameters, outer mixture probabilities and this ends the preliminary step before we move onto the steps of the FB model (refer to Sections 4.2.1 and 4.2.2)

- **FB Step 1** : In all the FB steps (**FB Step 1** to **FB Step 9**) the parameters of the TSDM model will be fixed. First we set the number of components of mixtures, say Q. Thus in effect, we are using the EM algorithm to fit the data-set to a mixture of Q Dirichlet densities.

- **FB Step 2** : We then form an initial value vector of the parameters of Dirichlet densities. As we are trying to fit the data-set to a $Q-$component mixture of Dirichlet densities, we have Q Dirichlet parameters and Q mixing proportions or mixture probabilities and they are as follows.

$$\boldsymbol{\beta}^{(\mathbf{0})} = (\boldsymbol{\beta}^{(\mathbf{0})}_{\boldsymbol{K+1}}, \boldsymbol{\beta}^{(\mathbf{0})}_{\boldsymbol{K+2}}, \ldots, \boldsymbol{\beta}^{(\mathbf{0})}_{\boldsymbol{K+Q}})$$

  and the initial values of the mixture probabilities,

$$\boldsymbol{\lambda}^{(0)} = (\lambda^{B(0)}, \lambda_1^{NC(0)}, \lambda_2^{NC(0)}, \ldots, \lambda_K^{NC(0)})$$

- **FB Step 3** : In the E-step, for the tth iteration (denoted by the suffix (t)), the posterior probabilities of the background model and the components of the new class model, mixture of Dirichlet densities, are updated as follows. Here $\boldsymbol{\lambda} = (\lambda^B, \lambda_1^{NC}, \lambda_2^{NC}, \ldots, \lambda_K^{NC})$

$$\mathrm{P}(\mathrm{S_i} = B|\mathbf{y_i}, \boldsymbol{\beta_k}^{(t)}, \boldsymbol{\lambda}^{(t)}) = f(\mathbf{y_i})\lambda^{B(t)} \Big/ \left\{ f(\mathbf{y_i})\lambda^{B(t)} + \sum_{k=K+1}^{K+Q} f(\mathbf{y_i})\lambda_k^{NC(t)} \right\} \equiv \gamma_{iB}^{(t)}$$

$$\mathrm{P}(\mathrm{S_i} = k|\mathbf{y_i}, \boldsymbol{\beta_k}^{(t)}, \boldsymbol{\lambda}^{(t)}) = \mathrm{Dir}(\mathbf{y_i}|\boldsymbol{\beta_k})\lambda_k^{NC(t)} \Big/ \left\{ f(\mathbf{y_i})\lambda_B^{(t)} + \sum_{k=K+1}^{K+Q} f(\mathbf{y_i})\lambda_k^{NC(t)} \right\} \equiv \gamma_{ik}^{(t)}$$

  for $k = K+1, \ldots, K+Q$

- **FB Step 4** : In the subsequent M-step, the parameters are updated using the following update analogous to the standard EM algorithm.

$$\lambda_k{}^{NC(t+1)} = \frac{1}{n}\sum_{i=1}^{n} \gamma_{ik}^{(t)}$$

- **FB Step 5** : We continue the same steps as Step 4 in Section 4.2 with $\psi = (\boldsymbol{\beta}^{(\mathbf{0})}, \boldsymbol{\lambda})$.

FB Steps 6, 7, 8 and **9** are same as Steps 5, 6, 7 and 8 of the TSDM model discussed in Section 4.2. Refer to Vatanen et al. (2012) for more details. Now let us use our model to detect new classes.

5.4 Application and discussion

In this section, we test the FB model for new class detection as follows. We remove the variable type BCEP (Beta Cephei) from the training data-set and train our TSDM model to estimate the parameters. Then in the test data-set, we add the BCEP class as a new class. It will be interesting to see if BCEP is detected by our model. BCEP variable type has 30 representatives in the test-data. We chose BCEP as the number of data-points in BCEP is 30 and hence it is a relatively small class with $J_k = 4$ in the TSDM model. This division of data is illustrated in Figure 5.2. Below, we list out the model conditions for the new class detection.

Model to which the data is fit	TSDM model
Attributes chosen	16 correlated attributes
Number of data-points in training set	1,141
Number of data-points in the test set	520
Prob scale to simplex transformation	STT1

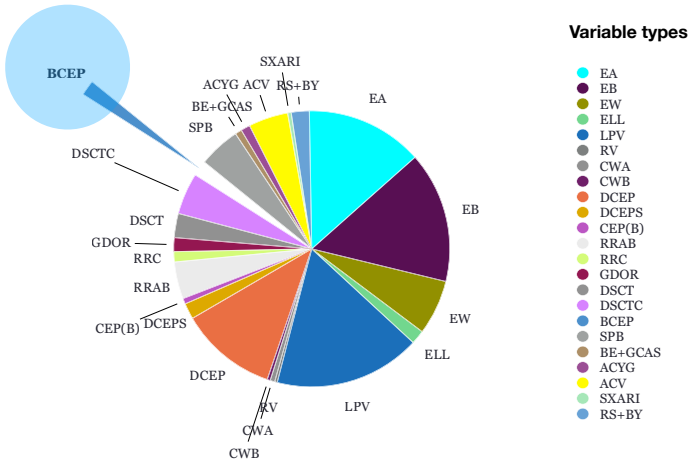


Figure 5.2: In the FB model, we train the model after removing the variable type BCEP from the training data-set. While testing the model, we use the test data-set which includes BCEP variable class with 30 data-points.

The classification results for the FB model while classifying the BCEP class as a new class are given in Table 5.1. The confusion matrix for this classification is given in Figure 5.3. The FB model gives a classification accuracy of 71.15% and detects our "new class" BCEP with 76.67% classification accuracy. This is a good result as BCEP is a small data-set and amounts to 6% of the training data-set and 3% of the test data-set. The fact that our model is able to detect such a small class is encouraging.

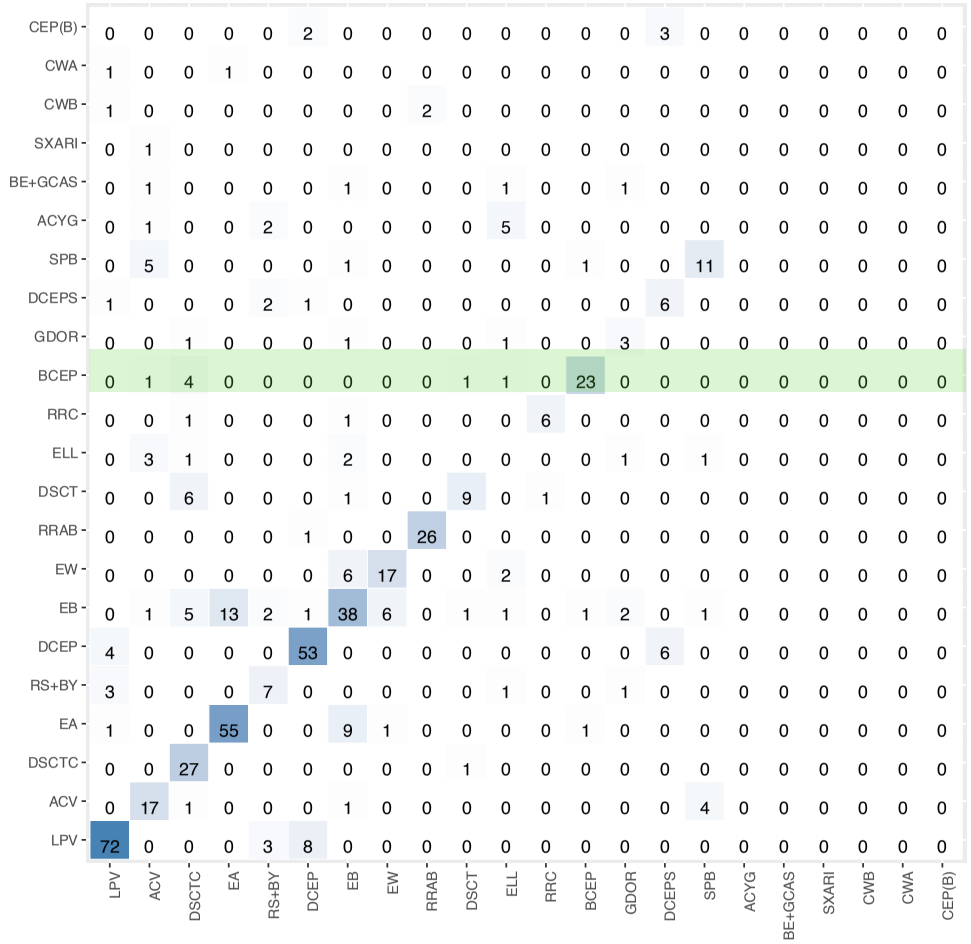


Figure 5.3: The confusion matrix for the FB model, detecting the "new" BCEP class as discussed in Section 5.4. BCEP has been detected with an accuracy of approx 77%. The columns give the predicted classes while the rows (shaded in green for BCEP) give the original classes in the training data-set. We see some confusion among BCEP and DSCTC.

Table 5.1: Classification accuracy of the FB model when detecting the BCEP class. FB model classification accuracy shows the classification accuracy of each of the classes when BCEP was detected as the new class.

Variable type	**Frequency in the training data-set**	**Frequency in the test data-set**	**FB model classification accuracy**
LPV	202	83	0.8675
EA	161	67	0.8209
EW	82	25	0.68000
DCEP	126	63	0.8413
DSCT	30	17	0.4285714
ACV	54	23	0.73913
GDOR	21	6	0.500000
SPB	63	18	0.61111
RRAB	45	27	0.96296
DSCTC	53	28	0.96429
RRC	12	8	0.75000
EB	183	72	0.52778
BCEP	0	30	0.76667
DCEPS	21	10	0.60000
RS+BY	23	12	0.58333
ELL	19	8	0
CWA	7	2	0
ACYG	10	8	0
SXARI	6	1	0
BE+GCAS	9	4	0
CEP(B)	6	5	0
RV	5	0	NA
CWB	3	3	0
	Classification accuracy		71.15%

Figure 5.3 gives the confusion matrix of the classification of the test-data (as shown in Figure 5.2) using the FB model for classifying the variable type BCEP. We see that 23/30 data-points were classified correctly. However, we see that 4 BCEP data-points were misclassified as DSCTC though there is not such a significant similarity between the classes. We can see that such a misclassification occurred even when the TSDM model was used in Chapter 4 which can be seen in Figure 4.8. Hence the DSCTC-BCEP misclassification is because of the TSDM model as the background model.

The classification accuracy rates of the other classes are similar to that of TSDM classification which was given in Section 4.3. This is ideal as the data-points that were classified into any of the 22 classes of the TSDM will be classified into the background model component of the FB model classification.

Though some of the changes in the classification accuracy is due to the fact that the test data-set and the training data-set are not entirely

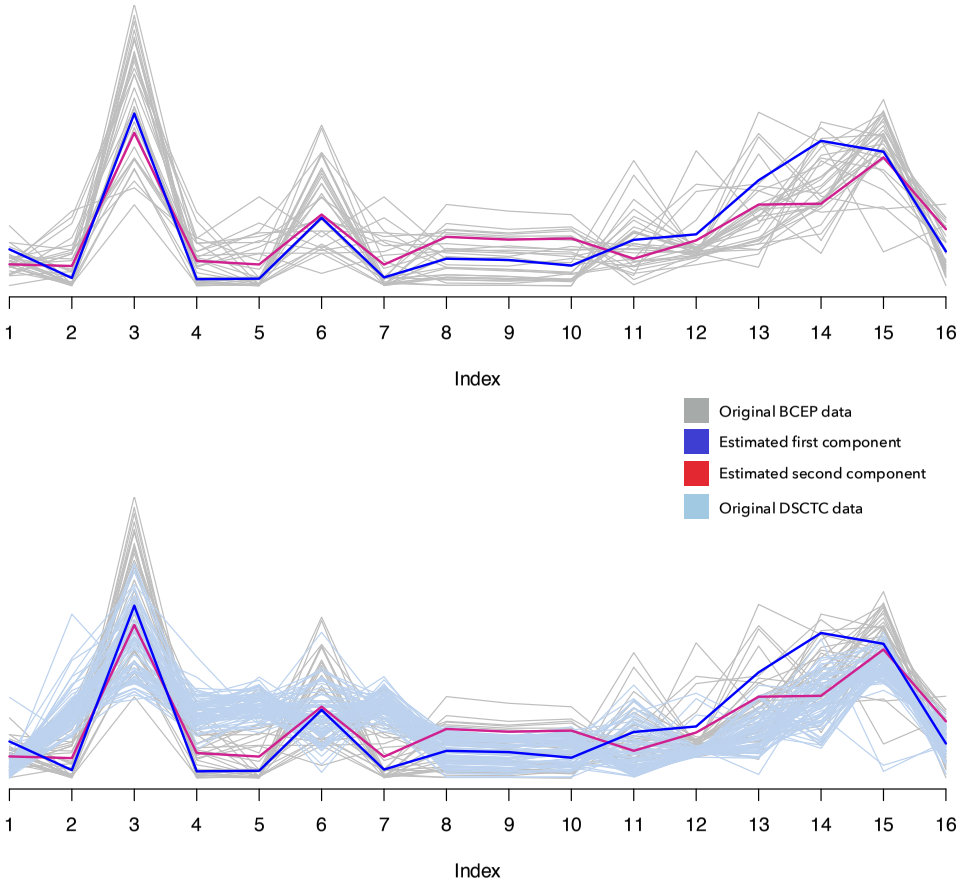


Figure 5.4: The above figure gives the signatures of the new class BCEP against the BCEP and DSCTC data. We see that the signatures of BCEP seem to suggest a good fit except on 3rd and 6th attribute (which denote the attribute ranks as well, refer to Section 3.4.1) where it seems that the data is away from the signatures. However the 3rd attribute seems to suggest a DSCTC data, which is why the confusion.

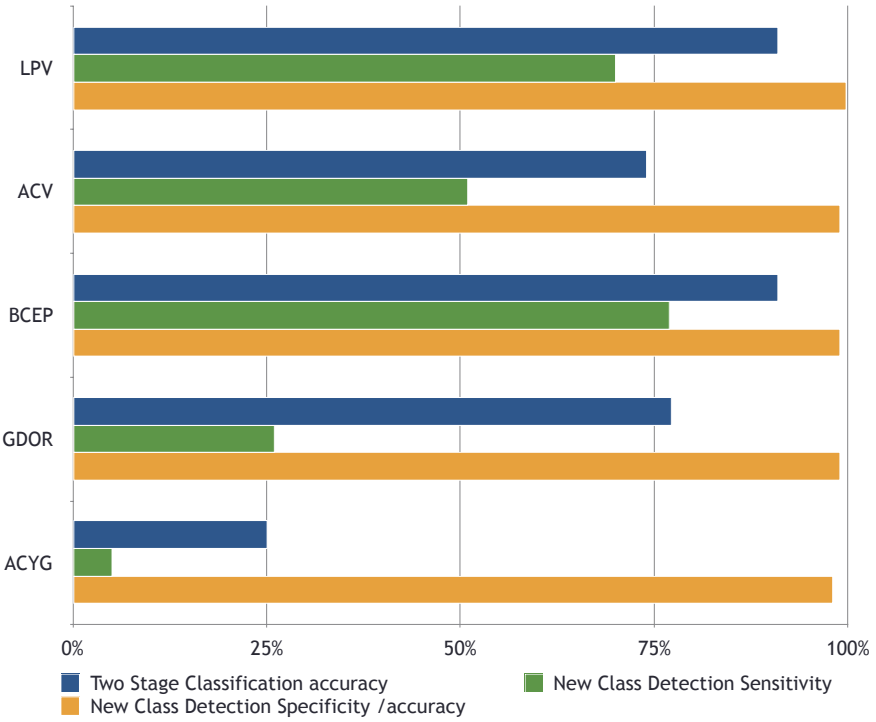


Figure 5.5: These are the model classification results when detecting 5 variable types as "new" classes (one at a time). The variable types are sorted by decreasing number of data-points with LPV having the highest number of data-points and ACYG having the lowest. The blue bars give us the Chapter 4 classification accuracy, while the green bars give us the classification accuracy while that particular class is detected as new class. The orange bars give us the specificity.

the same as in Chapter 4, in some extremely rare cases some of the data-points have been misclassified into similar classes because of the following reason. In the absence of BCEP, when the data-transformation is done on the training data, the probability scale transformation, and the subsequent simplex transformation gets affected as we discussed in Section 4.2. Hence the values of these data-points in the probability scale gets slightly altered and hence they get misclassified into some similar classes.

Now that we were able to detect the BCEP class as a new class, lets check the classification accuracy for other classes as well. Figure 5.5 gives a chart that shows how the variable types LPV, ACV, BCEP, GDOR and ACYG performed in the FB classification when they were considered as new classes and were detected. The variable types have been ordered by decreasing frequency in the training data-set and test data-set as given in Figure 5.5. We see that LPV class gets classified with an accuracy of around 88% while the sensitivity of LPV in the TSDM classification in Chapter 4 was just below 75%. We see reasonable performances of over 70% for ACV and GDOR as well, but we see that the sensitivity of ACYG is low. This is understandable given that the TSDM model classification accuracy of ACYG in Chapter 4 was around 25% which suggests that ACYG is difficult to be detected.

Analysis of Dirichlet signatures

Our FB model detected some data as anomalies or new classes which we know are the BCEP class data. Lets plot the Dirichlet signatures (Section 4.3.1) of the detected new class, BCEP against the original BCEP data.
Figure 5.4 plots the Dirichlet signatures against the BCEP data, and also against the BCEP and DSCTC combined.
The Dirichlet signatures gives us an insight into probably why 4 BCEPs were classified as DSCTC. In the figure, though the signatures of the estimated Dirichlet parameters represent the data well, we see that for the 3rd and 6th attribute namely raw percentile range and raw weighted skewness of the distribution of the curve, the Dirichlet signatures seem to represent DSCTC data instead of BCEP data. Given that these two attributes are highly ranked attributes according to our discussion in Section 3.4.1, this has caused the apparent BCEP-DSCTC confusion.

5.5 Synopsis

In this chapter we were able to build a model for the detection of new classes, namely the fixed background (FB) model. We used the TSDM model as our background model and a mixture of Q Dirichlet densities as our new class model, in the FB model. We found that our model was able to detect a relatively small class of 30 data-points with 77% accuracy

and this will be extremely useful when we use the model to detect new variable types in the Gaia data-set in the future.

Chapter 6

Towards Bayesian classification

6.1 Overview

In Chapter 4 we built the TSDM model for supervised classification using a frequentist approach predominantly, though we filled it with priors in the outer-mixtures. In this chapter we would like to expand our modeling strategy into the Bayesian territory as well. We will take advantage of the fact that Dirichlet densities are an exponential family distribution (refer to Appendix B.6). Our aim is to build a Bayesian supervised classification model to fit the data-set (Chapter 3) to a mixture of Dirichlet densities.

This chapter has been divided as follows. In Section 6.2 we discuss our model and the particular case of Dirichlet densities, while in Section 6.2.2 we present our model as an exponential family. We also present the conjugate prior and posterior density in Section 6.3.

6.2 Model formulation

Let $\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_n}$ denote a random variable of size n, where $\mathbf{Y_i}$ is a D-dimensional random vector with probability density function $g(\mathbf{y_i})$ on $\mathbb{R}^D$. Let the entire sample be represented by $\mathbf{Y} = (\mathbf{Y_1^T}, \ldots, \mathbf{Y_n^T})^\mathbf{T}$, where the superscript T denotes vector transpose. Thus $\mathbf{Y}$ is an n-tuple of points in $\mathbb{R}^D$ and $\mathbf{y} = (\mathbf{y_1^T}, \ldots, \mathbf{y_n^T})^\mathbf{T}$ denotes an observed sample where $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ is the D-dimensional observed value of the random vector $\mathbf{Y_i}$. The entire data-set $\mathbf{Y}$ is $n \times D$-dimensional. Then the mixture of K Dirichlet densities can be given as,

$$f(\mathbf{y_i}) = \sum_{k=1}^{K} \frac{\rho_k}{\mathbf{B}(\boldsymbol{\alpha_k})} \prod_{d=1}^{D} y_{id}^{\alpha_{kd}-1} \quad \mathbf{y_i} \in \mathbb{V}_{D-1} \tag{6.1}$$

where $\mathbf{y_i}$ and the open D-dimensional simplex, $\mathbb{V}_{D-1}$ are defined as in Section 1.2.3. Also, ρ_k is the mixture proportion or probability such that,

$$0 \leq \rho_k \leq 1 \quad (k = 1, \ldots, K)$$

and

$$\sum_{k=1}^{K} \rho_k = 1$$

Also for $\boldsymbol{\alpha_k} = (\alpha_{k1}, \alpha_{k2}, \ldots, \alpha_{kD})^T$

$$\mathbf{B}(\boldsymbol{\alpha_k}) = \frac{\Gamma(\alpha_{k1})\,\Gamma(\alpha_{k2}) \cdots \Gamma(\alpha_{kD})}{\Gamma(\alpha_{k1} + \alpha_{k2} + \cdots + \alpha_{kD})}.$$

6.2.1 Mixture of exponential families

A mixture of K exponential families can be represented as,

$$f(\mathbf{y_i}) = \sum_{k=1}^{K} \rho_k g_k(\mathbf{y_i}) \tag{6.2}$$

where each of the $g_k(\mathbf{y_i})$ belongs to the exponential family and is defined as (for exponential family definition refer to Appendix B.2)

$$g_k(\mathbf{y_i}) = \frac{b(\mathbf{y_i})}{a(\theta_k)} \exp\left(\theta_k^{\mathrm{T}} \mathrm{T}(\mathbf{y_i})\right), \quad \mathbf{y_i} \in \mathbb{V}_{D-1}.$$

Further including categorical random variables or missing data random variables as defined in Sections 1.2.2 and 4.2.2, we have $\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S}_\mathrm{n}$ for n data-points where $\mathrm{S_i} = k$ means $\mathbf{y_i}$ belongs to k. Thus Equation (6.2) becomes,

$$f(\mathbf{y_i}) = \sum_{k=1}^{K} h(\mathbf{y_i}, \mathrm{S_i} = k), \tag{6.3}$$

where $h(\mathbf{y_i}, \mathrm{S_i} = k) = \rho_k f_k(\mathbf{y_i})$
and thus

$$\log h(\mathbf{y_i}, \mathrm{S_i} = k) = \log \rho_k + \log g_k(\mathbf{y_i}).$$

In Boldi (2004) it was shown that when all components of a mixture come from the same exponential family and are equal upto the parameter, then the complete data log-likelihood takes also the form of an exponential family. That is,

$$\log g_k(\mathbf{y_i}) = \log b(\mathbf{y_i}) + \theta_k^T \mathrm{T}(\mathbf{y}_i) - \log a(\theta_k), \qquad \mathrm{S_i} = 1, \ldots, \mathrm{K}$$

where $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$.

Thus the complete data log-likelihood, $\log h(\mathbf{y_i}, \mathrm{S_i})$ can be represented as follows

$$\begin{aligned}\log h(\mathbf{y_i}, \mathrm{S_i}) &= \sum_{k=1}^{K} \mathbf{I}(\mathrm{S_i} = k)\big\{\log b(\mathbf{y_i}) + \theta_k^T \mathrm{T}(\mathbf{y}_i) - \log a(\theta_k) + \log \rho_k\big\} \\ &= \log b(\mathbf{y_i}) + \sum_{k=1}^{K-1} \mathbf{I}(\mathrm{S_i} = k)\left\{\theta_k^T \mathrm{T}(\mathbf{y}_i) - \log \frac{a(\theta_k)}{a(\theta_K)} + \log \frac{\pi_k}{\pi_K}\right\} \\ &+ \ \mathbf{I}(\mathrm{S_i} = K)\theta_K^T \mathrm{T}(\mathbf{y}_i) - \log a(\theta_K) + \log \rho_K\end{aligned}$$

In Boldi (2004), it was shown that the complete-data likelihood function comes from an exponential family by setting,

$$\phi = \begin{pmatrix} \log \frac{\rho_1}{\rho_K} - \log \frac{a(\theta_1)}{a(\theta_K)} \\ \log \frac{\rho_2}{\rho_K} - \log \frac{a(\theta_2)}{a(\theta_K)} \\ \cdots \\ \log \frac{\rho_{K-1}}{\rho_K} - \log \frac{a(\theta_{K-1})}{a(\theta_K)} \\ \theta_1 \\ \theta_2 \\ \cdots \\ \theta_K \end{pmatrix} \quad \text{and} \quad \mathrm{T}(\mathbf{y_i}, \mathrm{S_i}) = \begin{pmatrix} \mathbf{I}(\mathrm{S_i} = 1) \\ \mathbf{I}(\mathrm{S_i} = 2) \\ \cdots \\ \mathbf{I}(\mathrm{S_i} = K-1) \\ \mathbf{I}(\mathrm{S_i} = 1)\mathrm{T}(\mathbf{y_i}) \\ I(\mathrm{S_i} = 2)\mathrm{T}(\mathbf{y_i}) \\ \cdots \\ I(\mathrm{S_i} = K)\mathrm{T}(\mathbf{y_i}) \end{pmatrix}$$

where $\log b(\mathbf{y_i}, \mathrm{S_i}) = \log b(\mathbf{y_i})$ and $\log a(\phi) = \log a(\theta_K) - \log \rho_K$. Also ϕ and $\mathrm{T}(\mathbf{y_i}, \mathrm{S_i})$ are matrices with dimension $(2K-1) \times 1$

6.2.2 The case of Dirichlet densities

Dirichlet distribution is a member of the exponential family (refer Appendices B.2 and B.6) and hence the complete data-log-likelihood of Equation (6.1) can be written in exponential family form, as discussed in Section 6.2.1, by setting

$$\phi = \begin{pmatrix} \log \frac{\rho_1}{\rho_K} - \log \frac{\mathbf{B}(\boldsymbol{\alpha_1})}{\mathbf{B}(\boldsymbol{\alpha_K})} \\ \log \frac{\rho_2}{\rho_K} - \log \frac{\mathbf{B}(\boldsymbol{\alpha_2})}{\mathbf{B}(\boldsymbol{\alpha_K})} \\ \cdots \\ \log \frac{\rho_{K-1}}{\rho_K} - \log \frac{\mathbf{B}(\boldsymbol{\alpha_{K-1}})}{\mathbf{B}(\boldsymbol{\alpha_K})} \\ \alpha_{11} \\ \alpha_{12} \\ \cdots \\ \alpha_{KD} \end{pmatrix} \quad \text{and} \quad \mathrm{T}(\mathbf{y_i}, \mathrm{S_i}) = \begin{pmatrix} \mathbf{I}(\mathrm{S_i} = 1) \\ \mathbf{I}(\mathrm{S_i} = 2) \\ \cdots \\ \mathbf{I}(\mathrm{S_i} = K-1) \\ \mathbf{I}(\mathrm{S_i} = 1)\log y_{i1} \\ I(\mathrm{S_i} = 2)\log y_{i2} \\ \cdots \\ I(\mathrm{S_i} = K)\log y_{iD} \end{pmatrix}$$

and $\log b(\mathbf{y_i}, \mathrm{S_i}) = \log b(\mathbf{y_i}) = 1 \Big/ \Big\{ \prod_{d=1}^{D} y_{id} \Big\}$ and $\log a(\phi) = \log \mathbf{B}(\boldsymbol{\alpha_K}) - \log \rho_K$. Also ϕ and $\mathrm{T}(\mathbf{y_i}, \mathrm{S_i})$ are matrices with dimension $(2K-1) \times 1$.

The complete data-log-likelihood of a finite mixture of K Dirichlet densities, for a single data-vector $\mathbf{y_i}$ is a member of the exponential family and can be represented as follows,

$$\mathbf{L}(\phi|\mathbf{y_i}, \mathrm{S_i}) = \left[\frac{\rho_K}{\mathbf{B}(\boldsymbol{\alpha_K}) \prod_{d=1}^{D} y_{id}} \right] \exp \left[\sum_{k=1}^{K} \mathbf{I}(\mathrm{S_i} = k) \left\{ \sum_{d=1}^{D} \log y_{id} - \log \frac{\mathbf{B}(\boldsymbol{\alpha_k})}{\mathbf{B}(\boldsymbol{\alpha_K})} + \log \frac{\rho_k}{\rho_K} \right\} \right]$$

For $\mathbf{y} = (\mathbf{y_1}, \ldots, \mathbf{y_n})^{\mathbf{T}}$ and $\mathbf{S} = (\mathrm{S}_1, \mathrm{S}_2, \ldots, \mathrm{S_n})$ the likelihood for these n points can be written as

$$\begin{aligned} \mathbf{L}(\phi|\mathbf{y}, \mathbf{S}) &= \prod_{i=1}^{n} \mathbf{L}(\phi|\mathbf{y_i}, \mathrm{S_i}) \\ &= \prod_{i=1}^{n} \left[\frac{\rho_K}{\mathbf{B}(\boldsymbol{\alpha_K}) \prod_{d=1}^{D} y_{id}} \right] \exp \left[\sum_{k=1}^{K} \mathbf{I}(\mathrm{S_i} = k) \left\{ \sum_{d=1}^{D} \log y_{id} - \log \frac{\mathbf{B}(\boldsymbol{\alpha_k})}{\mathbf{B}(\boldsymbol{\alpha_K})} + \log \frac{\rho_k}{\rho_K} \right\} \right] \end{aligned} \tag{6.4}$$

Thus Equation (6.4) represents the complete data likelihood for the finite mixture of K Dirichlet densities for n data-points which as we discussed earlier, is a member of the exponential family. We'll discuss on how we can progress further in this, but before that lets discuss about conjugate priors and how conjugate priors have been used in McLachlan and Peel (2004) for the estimation of the parameters, mixing proportions and Dirichlet parameters of all the components of the mixture.

6.3 Conjugate prior

6.3.1 Previous work

For the component densities $g_k(\mathbf{y_i})$ for $k = 1, 2, \ldots, K$ being part of the same exponential family, McLachlan and Peel (2004) discussed the use of conjugate priors for the estimation of the parameters as follows.

The component densities can be expressed in exponential family form. That is for the kth component, we may write,

$$f(\mathbf{y_i}) = \exp\left(\theta_k^T \mathrm{T}(\mathbf{y_i}) - a(\theta_k) + c(\mathbf{y_i})\right),$$

where $c(\mathbf{y_i}) = \log b(\mathbf{y_i})$ relates to the definition of the exponential family given in Section 6.2.1. There exists a conjugate prior for an exponential family distribution, which has the form

$$p^{(\text{prior})}(\theta_k|\omega_k, \gamma_k) \propto \exp\left(\theta_k^T \omega_k - \gamma_k a(\theta_k)\right)$$

where ω_k, γ_k are hyper-parameters.

These conjugate priors are distinct for each component k, the hyper-parameters, ω_k is a real valued vector of constants and γ_k is a scalar constant $(k = 1, 2, \ldots, K)$. Also for the vector of mixing proportions, $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_K)^T$, the conjugate prior is the Dirichlet distribution $\mathrm{Dir}(q_1, q_2, \ldots, q_K)$ which has density of the form,

$$p^{(prior)}(\boldsymbol{\rho}|q_1, q_2, \ldots, q_K) = \Gamma(\sum_{k=1}^{K} q_k - K) \prod_{k=1}^{K} \rho_k^{q_k - 1} / \Gamma q_k$$

For $\Psi = (\Theta, \boldsymbol{\rho})$ where $\Theta = (\theta_1, \theta_2, \ldots, \theta_K)$ and $\boldsymbol{\rho} = (\rho_1, \rho_2, \ldots, \rho_K)^T$, then the posterior for Ψ is proportional to the product of the posteriors of Θ and $\boldsymbol{\rho}$, because we assume Θ and $\boldsymbol{\rho}$ to be a priori independent. That is,

$$\begin{aligned} p^{(post)}(\Psi|\mathbf{y}) &= \sum_{S_i} p^{(post)}(\boldsymbol{\rho}|\mathbf{y}) \prod_{k=1}^{K} p^{(post)}(\theta_k|\mathbf{y}) \\ &= \sum_{S_i} p^{(prior)}(\boldsymbol{\rho}|q_1 + n_1, \ldots, q_K + n_K) \prod_{k=1}^{K} p^{(prior)}(\theta_k|\omega_k + n_k \bar{\mathbf{y}_\mathbf{k}}, \gamma_k + n_k)) \end{aligned}$$

where $n_k = \sum_{i=1}^{n} \mathbf{I}(\mathrm{S_i} = k)$ and $\bar{\mathbf{y}_\mathbf{k}} = \sum_{i=1}^{n} \mathbf{I}(\mathrm{S_i} = k)\mathbf{y_i}/n_k$

Even though though the posterior for Ψ has been expressed in closed form, McLachlan and Peel (2004) states that the time taken to compute the above is too high, since we have to sum over all the values of $i, i = 1, 2, \ldots, n$ and over all the possible configurations of $\mathbf{S}$. This results in a

very high computational cost even if it has to be applied for moderate sample sizes.

6.3.2 Conjugate prior for the complete data-likelihood

In Section 6.2.2, we saw that the complete data likelihood $\mathbf{L}(\phi|\mathbf{y},\mathbf{S})$ is a member of the exponential family. Rewriting Equation 6.4, we have,

$$\begin{aligned}
\mathbf{L}(\phi|\mathbf{y},\mathbf{S}) &= \prod_{i=1}^{n}\left[\frac{\rho_K}{\mathbf{B}(\boldsymbol{\alpha_K})\prod_{d=1}^{D} y_{id}}\right]\exp\left[\sum_{k=1}^{K}\mathbf{I}(\mathrm{S_i}=k)\left\{\sum_{d=1}^{D}\log y_{id} - \log\frac{\mathbf{B}(\boldsymbol{\alpha_k})}{\mathbf{B}(\boldsymbol{\alpha_K})} + \log\frac{\rho_k}{\rho_K}\right\}\right] \\
&= \prod_{i=1}^{n}\left\{\left[\frac{\rho_K}{\mathbf{B}(\boldsymbol{\alpha_K})\prod_{d=1}^{D} y_{id}}\right]\exp\left[\sum_{k=1}^{K}\mathbf{I}(\mathrm{S_i}=k)\left\{\sum_{d=1}^{D}\log y_{id} - \log\frac{\mathbf{B}(\boldsymbol{\alpha_k})}{\mathbf{B}(\boldsymbol{\alpha_K})} + \log\frac{\rho_k}{\rho_K}\right\}\right]\right\} \\
&= \frac{\prod_{i=1}^{n} b(\mathbf{y_i},\mathrm{S_i})}{[a(\phi)]^n}\exp\left[\phi^T\sum_{i=1}^{n}\mathrm{T}(\mathbf{y_i},\mathrm{S_i})\right] \\
&= \frac{b'(\mathbf{y},\mathbf{S})}{a'(\phi)}\exp\left[\phi^T\mathrm{T}'(\mathbf{y},\mathbf{S})\right] \qquad (6.5)
\end{aligned}$$

where $b'(\mathbf{y},\mathbf{S}) = \prod_{i=1}^{n} b(\mathbf{y_i},\mathrm{S_i})$, $a'(\phi) = [a(\phi)]^n$ and $\mathrm{T}'(\mathbf{y},\mathbf{S}) = \sum_{i=1}^{n}\mathrm{T}(\mathbf{y_i},\mathrm{S_i})$ and is in the exponential family where,

$$\phi = \begin{pmatrix}
\log\frac{\rho_1}{\rho_K} - \log\frac{\mathbf{B}(\boldsymbol{\alpha_1})}{\mathbf{B}(\boldsymbol{\alpha_K})} \\
\log\frac{\rho_2}{\rho_K} - \log\frac{\mathbf{B}(\boldsymbol{\alpha_2})}{\mathbf{B}(\boldsymbol{\alpha_K})} \\
\cdots \\
\log\frac{\rho_{K-1}}{\rho_K} - \log\frac{\mathbf{B}(\boldsymbol{\alpha_{K-1}})}{\mathbf{B}(\boldsymbol{\alpha_K})} \\
\alpha_{11} \\
\alpha_{12} \\
\cdots \\
\alpha_{KD}
\end{pmatrix} \quad \text{and} \quad \mathrm{T}(\mathbf{y_i},\mathrm{S_i}) = \begin{pmatrix}
\mathbf{I}(\mathrm{S_i}=1) \\
\mathbf{I}(\mathrm{S_i}=2) \\
\cdots \\
\mathbf{I}(\mathrm{S_i}=K-1) \\
\mathbf{I}(\mathrm{S_i}=1)\log y_{i1} \\
I(\mathrm{S_i}=2)\log y_{i2} \\
\cdots \\
I(\mathrm{S_i}=K)\log y_{iD}
\end{pmatrix}$$

and $\log b(\mathbf{y_i},\mathrm{S_i}) = \log b(\mathbf{y_i}) = 1\Big/\left\{\prod_{d=1}^{D} y_{id}\right\}$ and $\log a(\phi) = \log\mathbf{B}(\boldsymbol{\alpha_K}) - \log\rho_K$. Also ϕ and $\mathrm{T}(\mathbf{y_i},\mathrm{S_i})$ are matrices with dimension $(2K-1)\times 1$.

Now for the likelihood mentioned in Equation (6.5), there exists a conjugate prior for ϕ since $\mathbf{L}(\phi|\mathbf{y},\mathbf{S})$ belongs to an exponential family. The conjugate prior can be expressed as follows,

$$p^{prior}(\phi|\chi,\nu) \propto a(\phi)^{-\nu}\exp(\phi^T\chi)$$

where χ and ν are hyper-parameters with ν being a scalar constant and χ being a real values vector of constants. Also $a(\phi)$ is such that $\log a(\phi) = \log\mathbf{B}(\boldsymbol{\alpha_K}) - \log\rho_K$.

The posterior for ϕ has the following distribution

$$\begin{aligned} p^{(post)}(\phi|\mathbf{y},\mathbf{S}) &\propto \mathbf{L}(\phi|\mathbf{y},\mathbf{S}) \times p^{prior}(\phi|\chi,\nu) \\ &\propto a(\phi)^{-(\nu+n)} \exp\left[\phi^T\left(\chi + \sum_{i=1}^{n} \mathrm{T}(\mathbf{y_i},\mathrm{S_i})\right)\right] \\ &= a(\phi)^{-\tilde{\nu}} \exp\left[\phi^T \tilde{\chi}\right] \end{aligned} \tag{6.6}$$

where

$$\begin{aligned} \tilde{\chi} &= \chi + \mathrm{T}(\mathbf{y},\mathbf{S}) \\ &= \chi + \sum_{i=1}^{n} \mathrm{T}(\mathrm{y_i},\mathrm{S_i}) \end{aligned}$$

and $\tilde{\nu} = \nu + n$ are the posterior updates. The parameters ϕ, the statistic $\mathrm{T}(\mathbf{y_i},\mathrm{S_i})$ and function $a(\phi)$ are defined in Section 6.2.2.

6.4 Synopsis and prospectives

We discussed the Bayesian model to classify different classes of variable stars. The advantages of the Bayesian approach as compared to the frequentist framework is that it allows us to effectively integrate new evidence/information as it arrives and tune the model by updating the posterior class probabilities. We also discussed that the computational cost of the posterior distribution by McLachlan and Peel (2004) is really high. In our work, we computed the posterior distribution using the result that the full data-likelihood is an exponential family distribution. Our methodology has an advantage that there are no a priori independence assumptions unlike the methodology described by McLachlan and Peel (2004) in Section 6.3.1. However there are a further steps which can be taken up as future work. The conjugate prior is for the canonical parameters of the exponential family, ϕ, which is a function of our parameters of interest. The result from Diaconis et al. (1979) can be used for computing this. Also the computational challenges can be dealt with by solving using the variational Bayesian learning algorithms proposed by Ghahramani and Beal (2000) for conjugate-exponential families. Conjugate exponential families are models that satisfy the following conditions. (1) the complete data likelihood is in the exponential family and (2) the parameter prior is conjugate to the complete data likelihood. Our model easily qualifies the criteria and it will be interesting to proceed further with this algorithm.

Chapter 7

Conclusions and future work

With an influx of large amount of data expected from the Gaia survey, an efficient classification model is the need of the hour. Our first statistical methodological contribution in the thesis was the formulation of a supervised classification model, TSDM (two stage Dirichlet mixture) model, for classification. In connection with this we proposed two transformation methodologies, STT1 and STT2 to transform the data to a simplex. We got a corrected classification accuracy of 72.68% for the STT1 classification as opposed to 74.78% by the STT2 classification. However we found out that STT2 classification doesn't hold a huge advantage over STT1 classification.

Some of the attributes we used were highly correlated. We studied the performance of the classification model with a subset of attributes that have correlations less than 0.8 among each other. The findings reaffirmed our use of Dirichlet distribution as not only was the classification accuracy lower, but also the two stage Gaussian mixture model (TSGM) model showed tendencies to accurately classify only the larger classes, as the smaller classes were misclassified heavily. We compared the TSDM model with the random forest classifier studied in Dubath et al. (2011). Random forest classifier does perform relatively better when the task is to classify the data to known classes. However, Dubath et al. (2011) does not tackle the problem of detecting new classes with random forests.

In Chapter 3 we broadly discussed the volume of data that will be provided by the Gaia mission. This is where our model holds an upperhand over the random forests classifier. In this thesis, we extended our model to a new class detection model (FB model). Our model was able to detect small and large data-sets of new classes with promising accuracy. The detection of a class as small as 30 data points with 77% accuracy is promising. This was our second methodological contribution. We also presented a feasibility study of Bayesian supervised classification by fitting our data to a mixture of Dirichlet densities. We proposed a conjugate prior for the canonical parameters of the exponential family form of our model. We also presented the closed form for the posterior distribution of our model. Also, in the first stage of the TSDM model, we used unsupervised classification on each of the variable types and cluster them. However many open problems await future research and we mention some of them here.

- *Subjective priors in the second stage of TSDM :* In the second stage of the TSDM model for classification (Section 4.2.2), we discussed the priors that can be used in the second stage of our model. Though in our applications we used a non-informative prior, we can set a subjective prior that will represent the prior beliefs about the distribution of our parameters. This will of course require a close collaboration with domain experts but will improve our model and increase the flexibility.

- *Copulas vs Dirichlet distributions :* We used Dirichlet densities as a natural choice for modeling data in the probability scale. However, further investigation needs to be done to see how a two stage Copulas mixture model performs, as opposed to the TSDM model. Copulas hold an advantage over Dirichlet distributions that dependencies among random variables can be modeled as well (Embrechts et al. (2001)).

- *Prior on the number of components in the FB model* : In the new class model component of the FB model, we fixed the number of components in the beginning before fitting the new class to the data. In practice, it would be worthwhile to not fix it in the beginning by allowing more flexibility. The next step could be to add a prior on the number of components so that the FB model will be able to choose the number of components from the data.

- *Maximizing the posterior and inference on the parameters of interest*: In Section 6.4, we proposed methods to maximize the posterior of the complete data likelihood. The mixture of Dirichlet distributions is a member of the conjugate-exponential family since (1) the complete data likelihood is in the exponential family and (2) the parameter prior is conjugate to the complete data likelihood. Variational Bayesian learning algorithms can be used to maximize the posterior of our model. Also, conjugate prior which we have proposed for our model in Chapter 6 is for the canonical parameters of the exponential family. We can use the results of Diaconis et al. (1979) to infer on the parameters of interest.

- *Sub-classes of the variable types :* Also one of the outputs of our model which we haven't analyzed further is in the astrophysical significance of the clusters in the first stage of the TSDM model (Section 4.2.1). The next step is to collaborate with astrophysicists to study if these sub-classes have any significance. Our study of clustering on a combined data-set of δ Scuti's gave promising results that needs to be studied further.

Appendix A
Tables and Figures

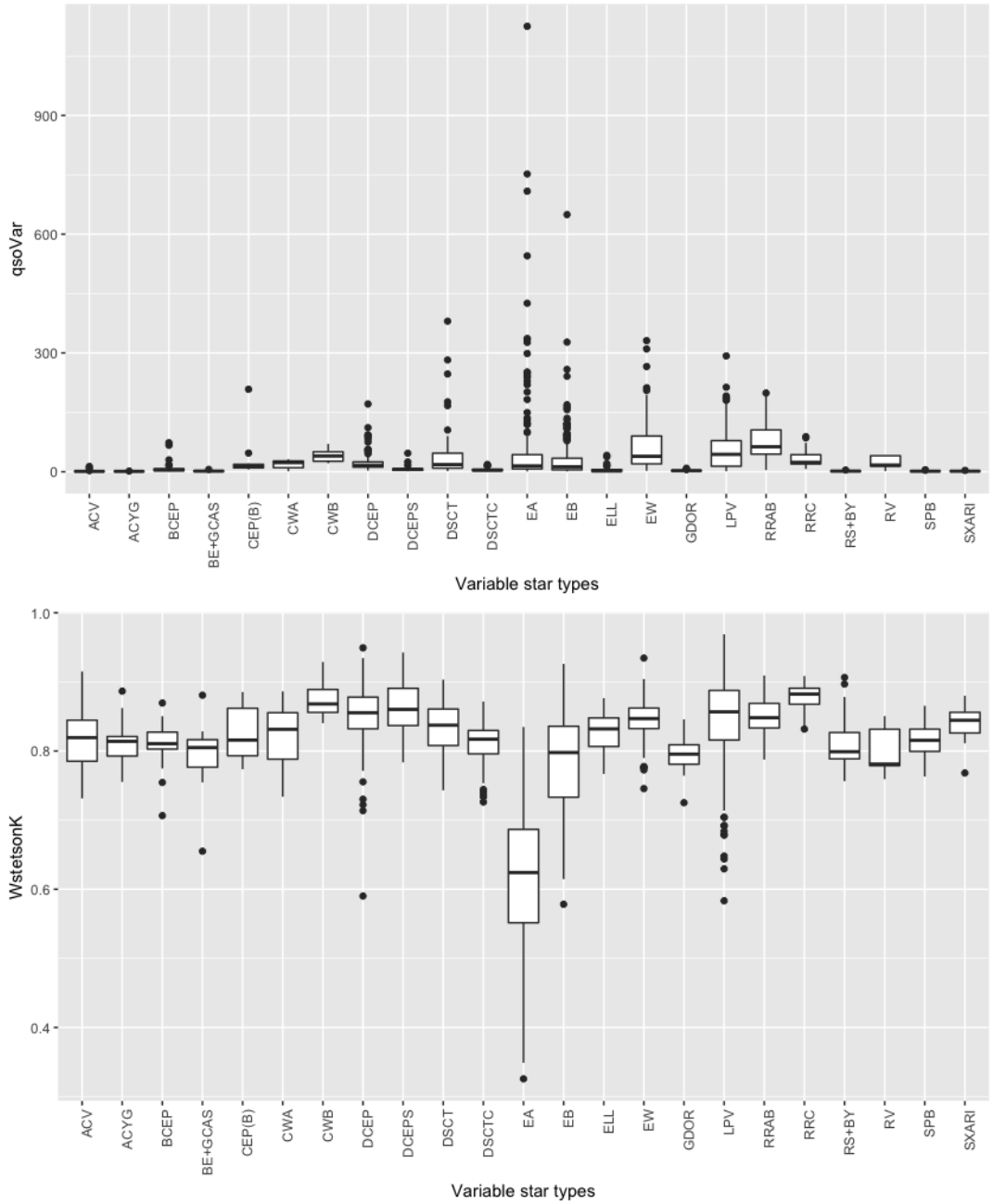
Table A.1: The entire list of attributes in the training data-set.
There are 45 attributes in the raw data.

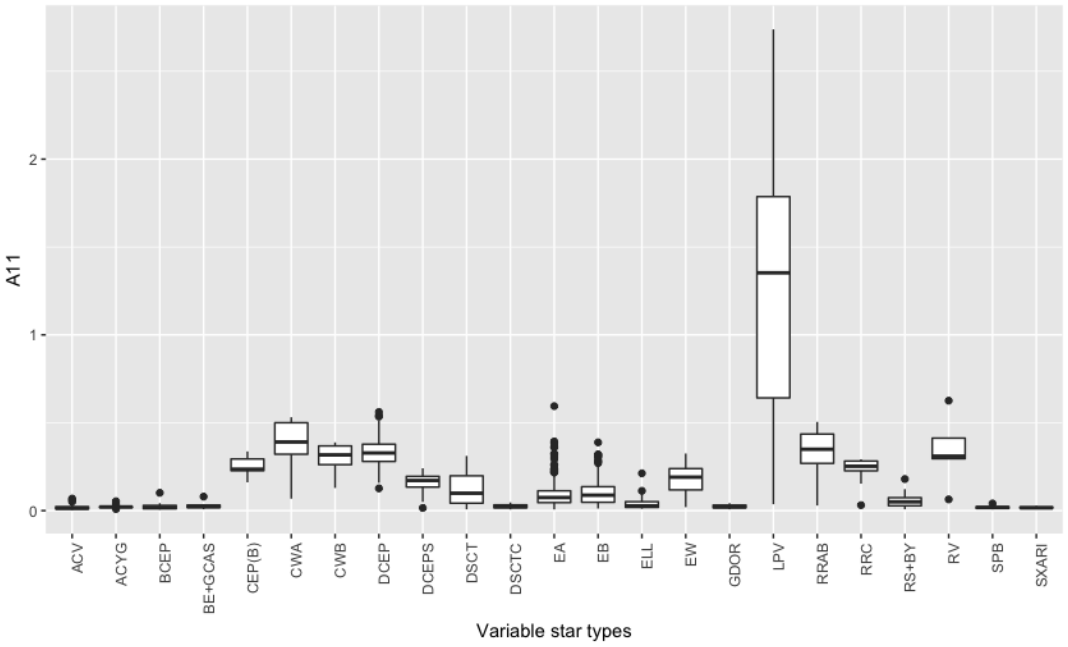
Attribute names	**Attribute description**
hip	Hipparcos identifier
cat	P = periodic, well-known,X = "unsolved"", problematic light curves: class uncertain or unknown, irregular or small-amplitude variability, very bad time sampling
type	The true class of the star
p2pScatterOnDetrendedTS	point-to-point scatter of the time series of brightness after removing a slow polynomial trend
p2pScatterOnFoldedTS	point-to-point scatter of the sequence of brightnesses after finding a period and phase-folding
scatterOnResidualTS	square root of variance of the residuals after modelling
Raw_WeightedStdDev	weighted standard deviation of the brightness
Raw_WeightedSkewness	weighted skewness of the brightness
Raw_WeightedKurtosis	weighted kurotsis of the brightness
Raw_PercentileRange10	0.1-quantile minus the median of the raw brightnesses
stetsonJ	measures of correlation between closely spaced brightness values
stetsonJweighted	measures of correlation between closely spaced brightness values
stetsonK	measures of correlation between closely spaced brightness values
WstetsonJ	measures of correlation between closely spaced brightness values
WstetsonJweighted	measures of correlation between closely spaced brightness values
WstetsonK	measures of correlation between closely spaced brightness values
logPnonQso	measure of stochastic variability,components in the light curves
logPqso	measure of stochastic variability,components in the light curves
qsoVar	measure of stochastic variability,components in the light curves
nonQsoVar	measure of stochastic variability,components in the light curves
LogPeriod	base 10 logarithm of the period in days
LogAmplitude	base 10 logarithm of the peak-to-peak amplitude (from harmonic model fit and reconstruction of the fitted light curve)
HarmNum	the highest significant order of harmonic terms in a least squares harmonic model fit
A11	amplitude of the first harmonic term
A12	amplitude of the second harmonic term
PH12	Relative phase of the second harmonic term
A13	amplitude of the third harmonic term
PH13	Relative phase of the third harmonic term
A14	amplitude of the fourth harmonic term
PH14	Relative phase of the fourth harmonic term
A15	amplitude of the fifth harmonic term
PH15	Relative phase of the fifth harmonic term
logA11minusA	$\log_{10}(1 + abs(A11 - \sqrt{(\sum(A1j^2))}))$
logA12_A11	$\log_{10}(1 + A12/A11)$
logA13_A12	$\log_{10}(1 + A13/A12)$
absGlat	absolute value of galactic latitude
Glat	galactic latitude
Glon	galactic longitude
Parallax	the parallax of the object (in milliarcsec) equivalent to distance
Absolute_Mag00	estimate of the absolute brightness of the object
BV_Color	Colors computed with visual brightness. B can be taken as "blue", V is "visual", I is a red-near infrared filter
VI_Color	Colors computed with visual brightness. B can be taken as "blue", V is "visual"
JmK	J, H, K are all infrared filters
JmH	J, H, K are all infrared filters
HmK	J, H, K are all infrared filters

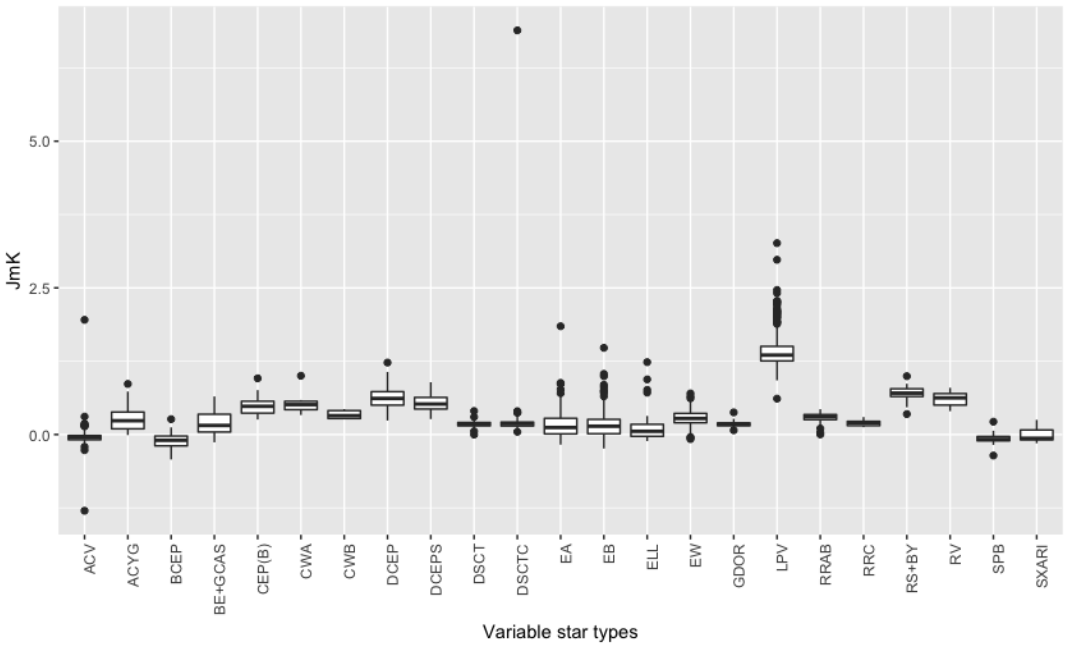
Table A.2: Attributes in the training data divided by functionality.

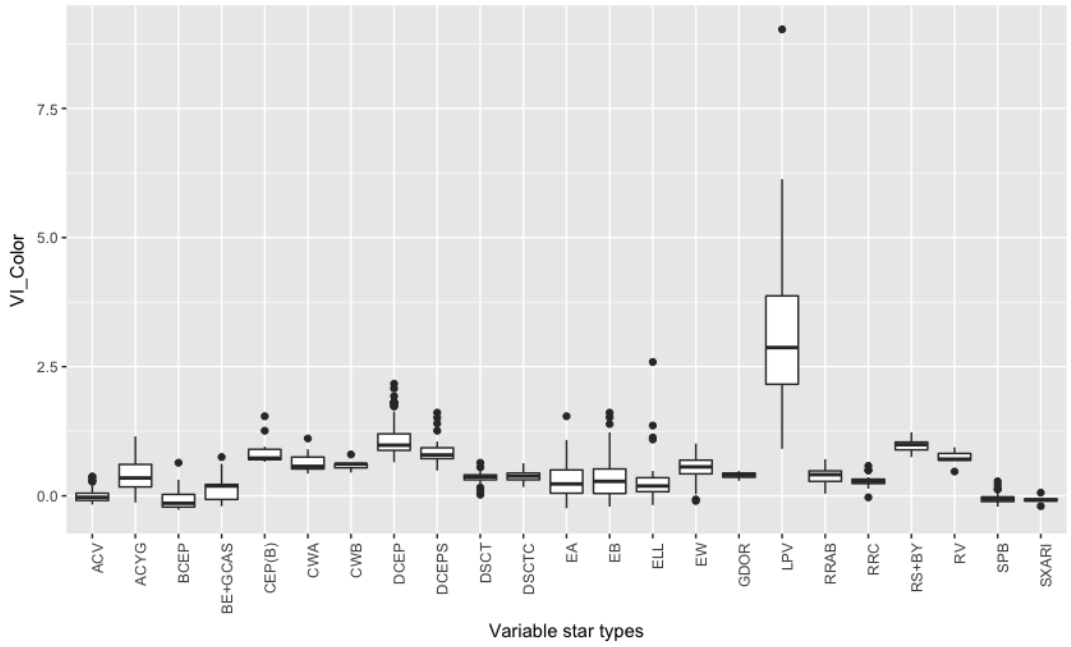
Type of Attributes	**Attribute names**
Attributes that summarize the improvements after different steps of modelling.	p2pScatterOnDetrendedTS p2pScatterOnFoldedTS scatterOnResidualTS
Attributes related to astrophysics.	absGlat Glat and Glon Parallax Absolute_Mag00 BV_Color VI_Color JmK, JmH and HmK
Attributes that summarize the distribution of the observed magnitudes.	Raw_WeightedStdDev Raw_WeightedSkewness Raw_WeightedKurtosis Raw_PercentileRange10 StetsonJ stetsonJweighted stetsonK WstetsonJ WstetsonJweighted WstetsonK
Attributes that quantify the strength of a stochastic variability.	logPnonQso logPqso qsoVar nonQsoVar.
Attributes related to the period search and harmonic modeling.	LogPeriod LogAmplitude HarmNum A11, A12, A13, A14, A15 PH12, PH13, PH14, PH15 logA11minusA

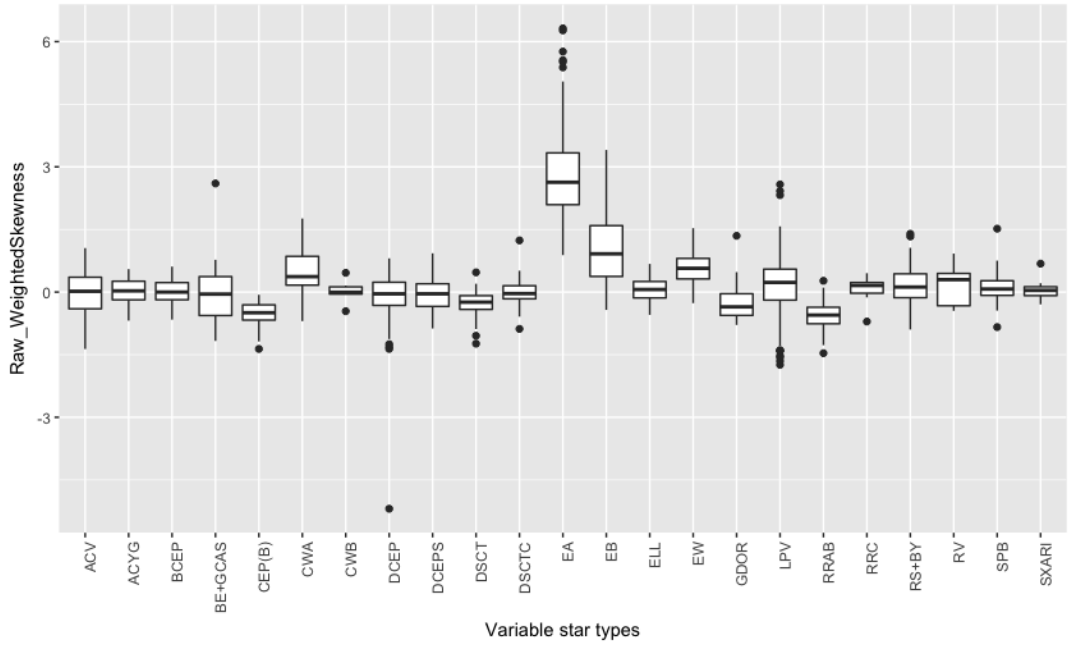
Figure A.1: Distribution of the important variables for each of the variable star classes or types, which we have chosen in our thesis. By important variables, we mean the 16 attributes selected according to the importance measure (mean decrease in accuracy) in Chapter 3.

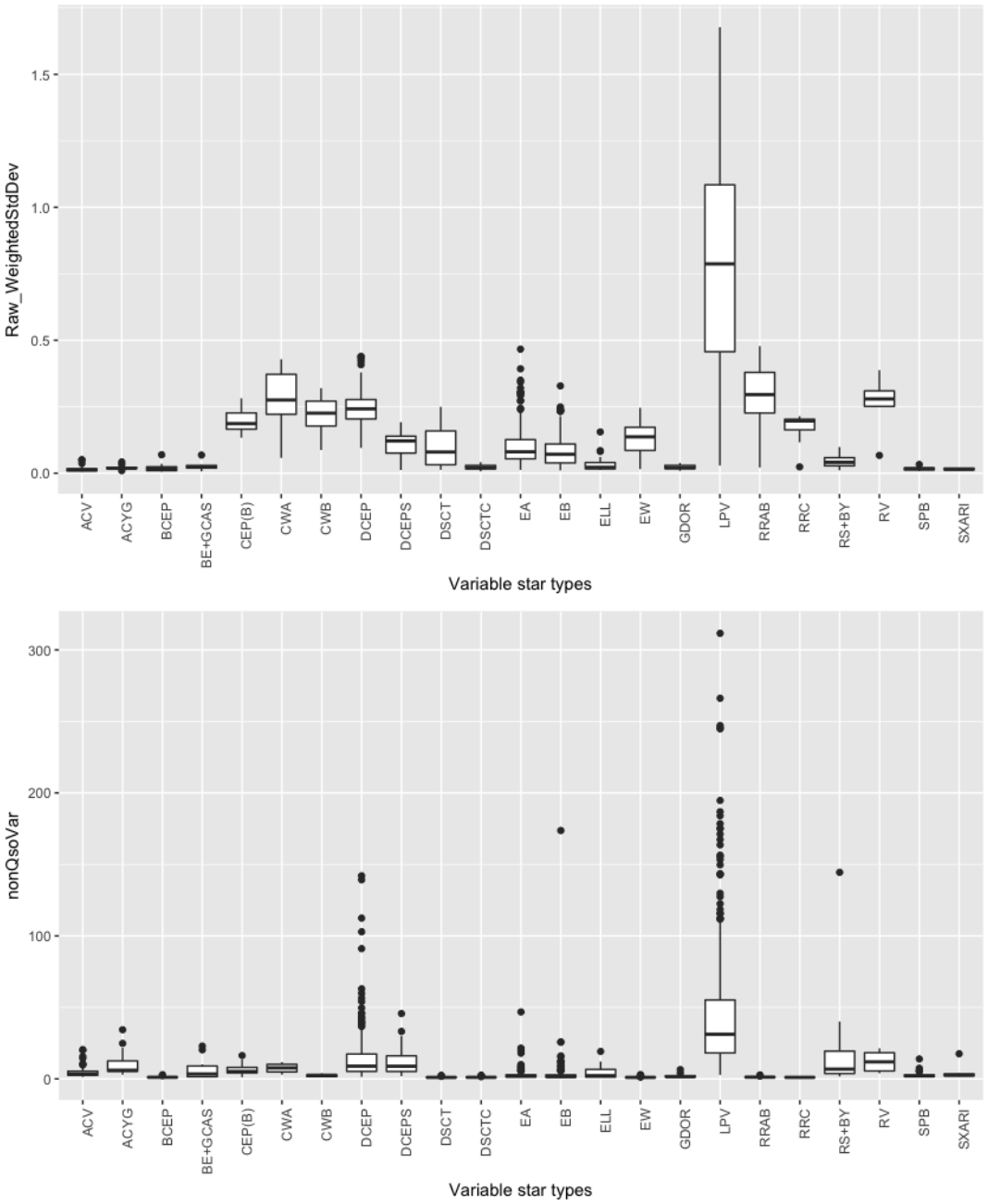


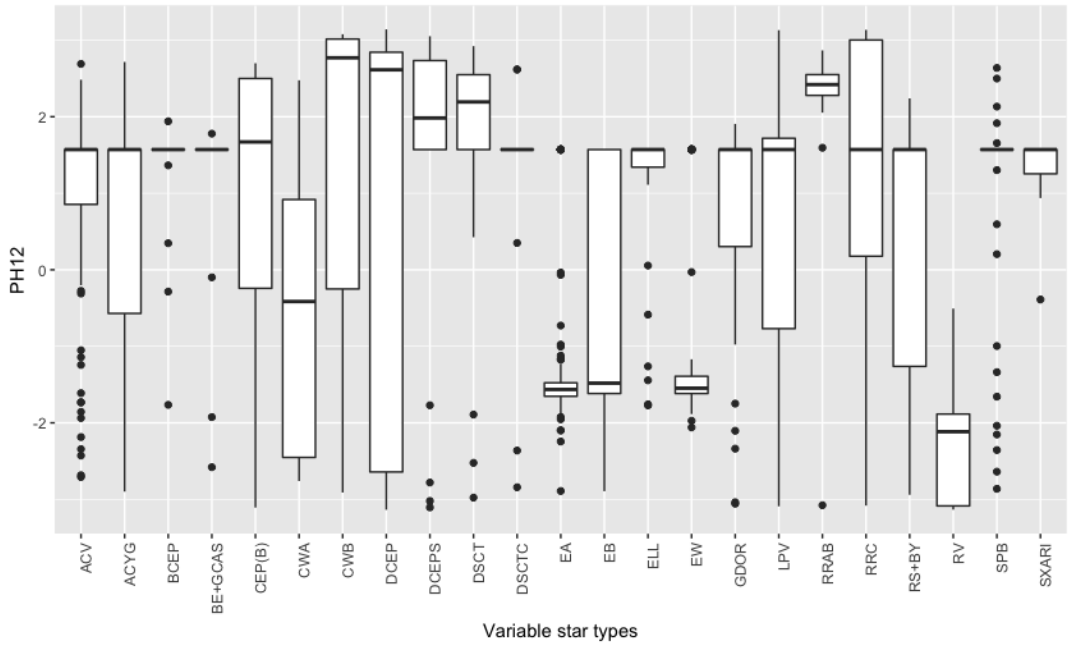

A11
2
1
0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

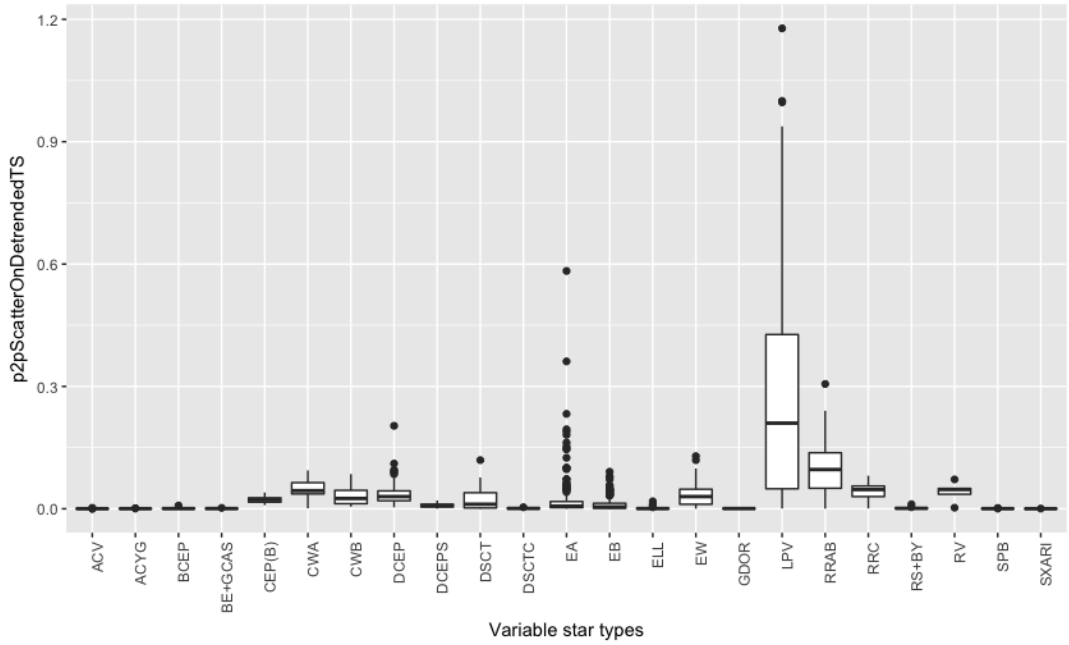

JmK
5.0
2.5
0.0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

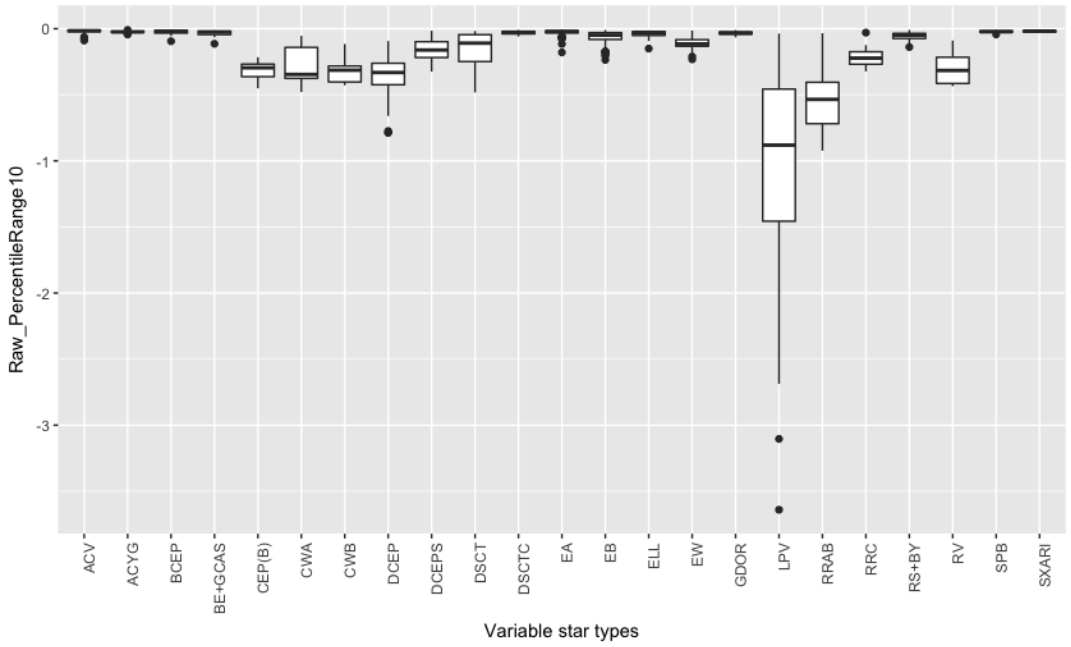

VI_Color
7.5
5.0
2.5
0.0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

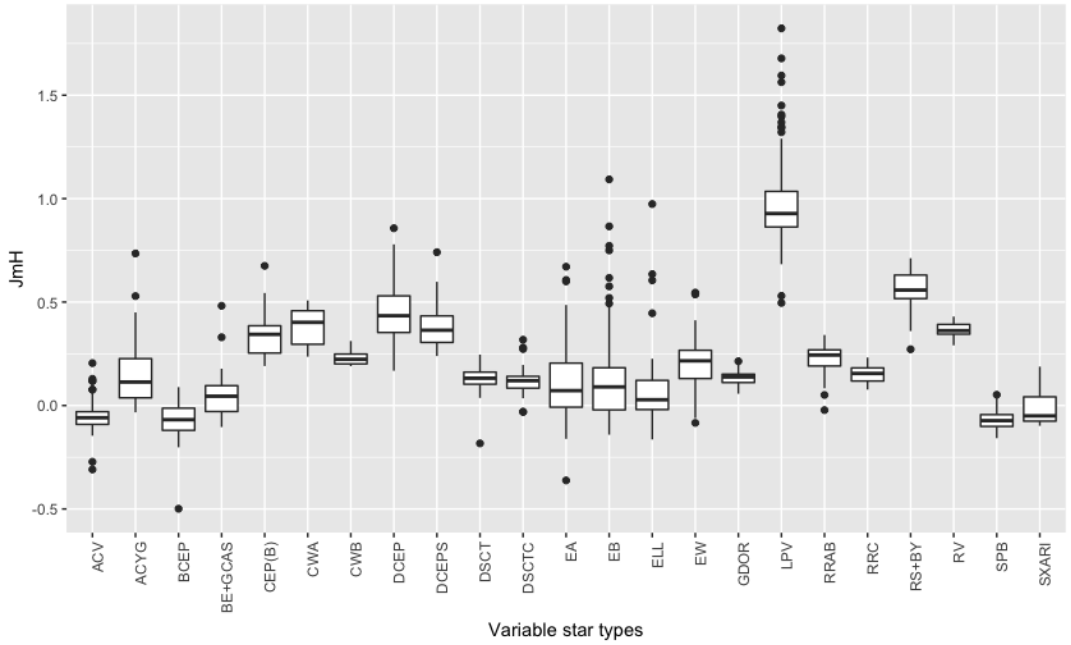

Raw_WeightedSkewness
6
3
0
-3
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

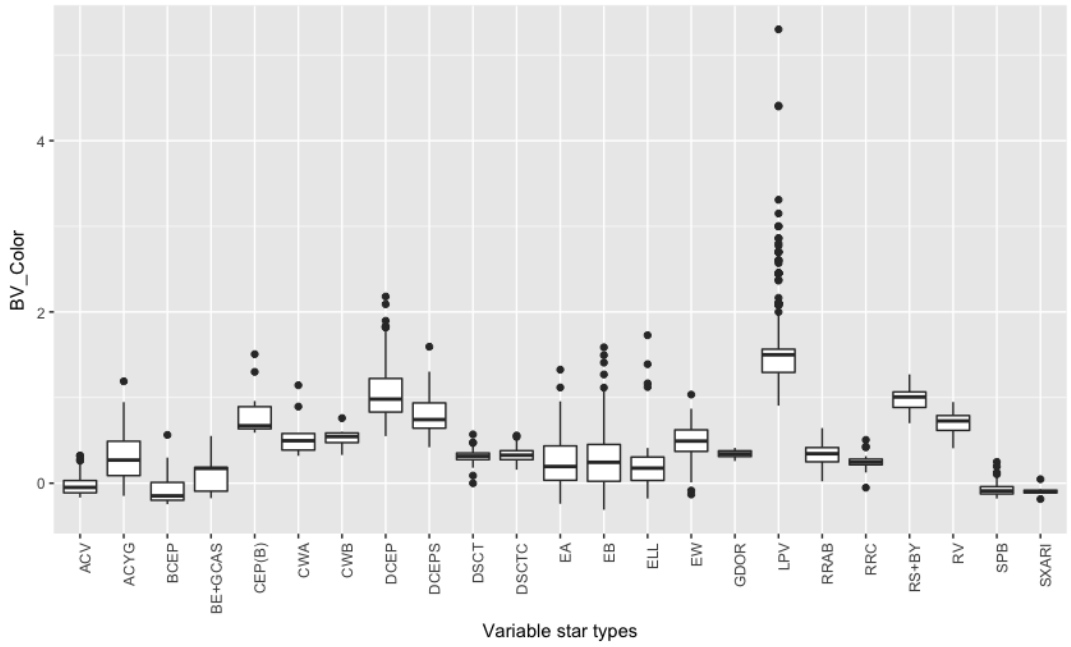

Raw_WeightedStdDev
1.5
1.0
0.5
0.0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types
nonQsoVar
300
200
100
0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

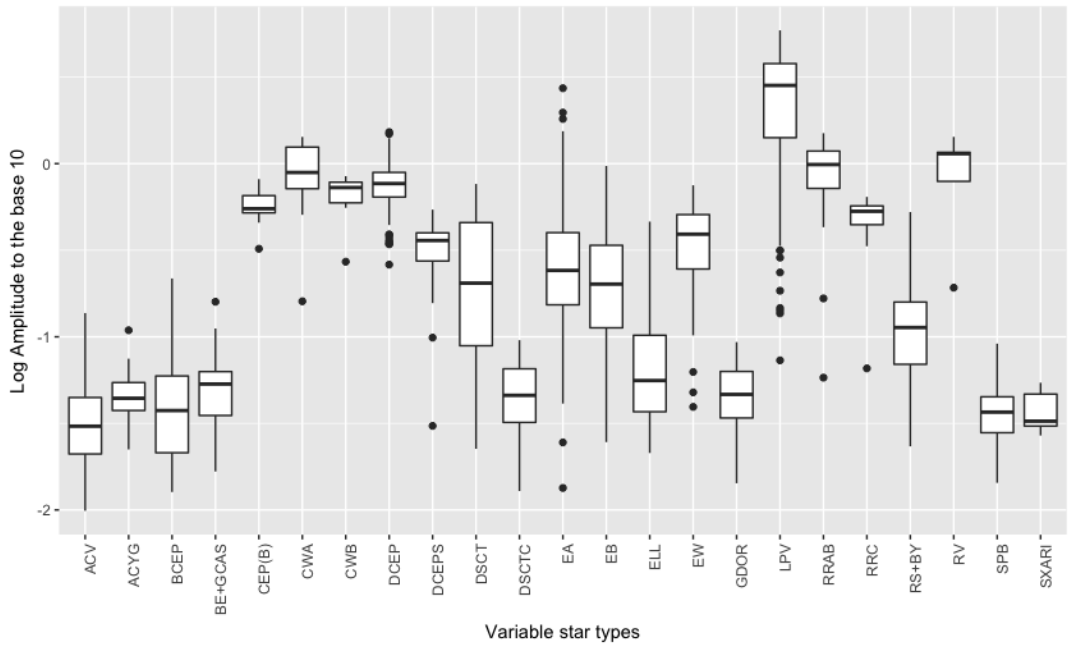

PH12
2
0
-2
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

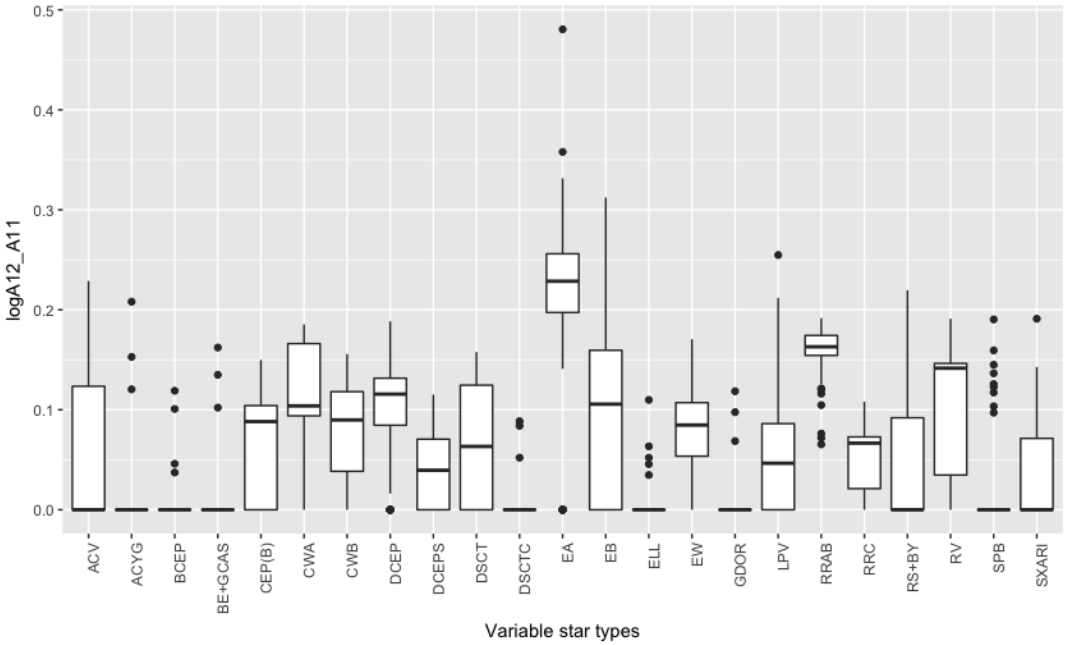

p2pScatterOnDetrendedTS
1.2
0.9
0.6
0.3
0.0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types


Raw_PercentileRange10
0
-1
-2
-3
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types


JmH
1.5
1.0
0.5
0.0
-0.5
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types


BV_Color
4
2
0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types


Log Amplitude to the base 10
0
-1
-2
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types


logA12_A11
0.5
0.4
0.3
0.2
0.1
0.0
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

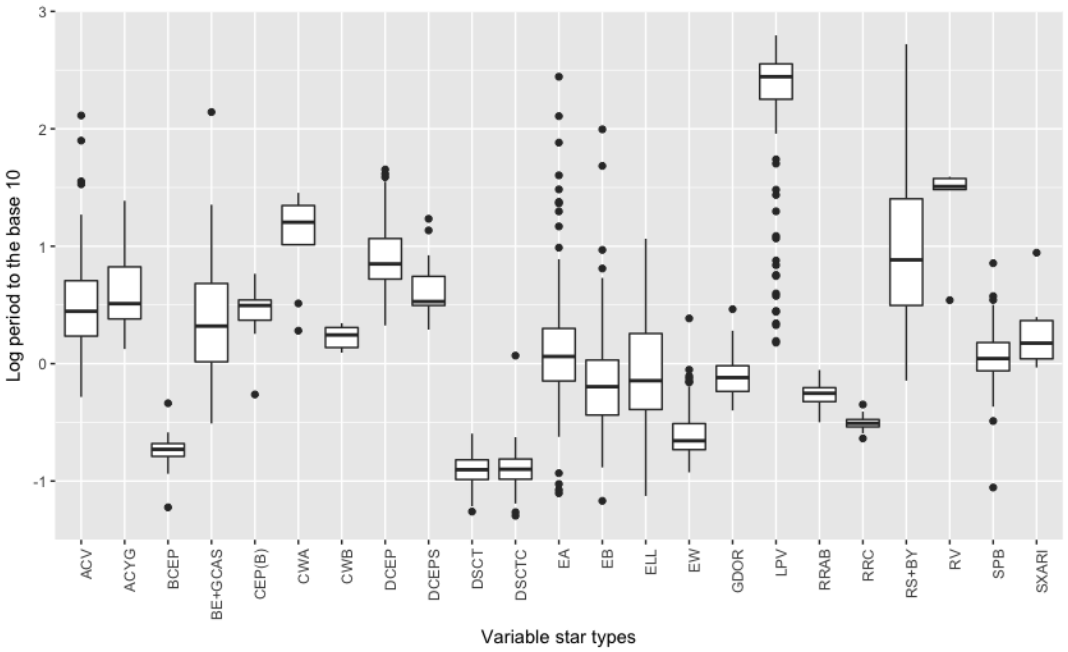
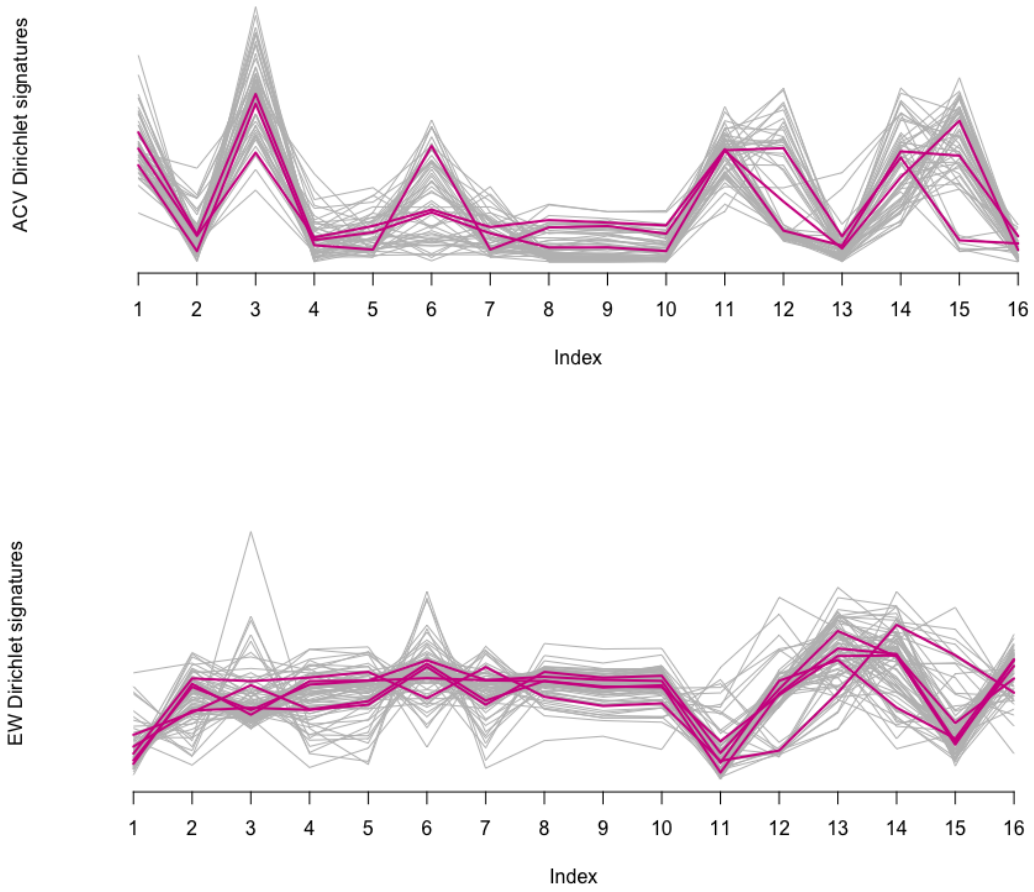
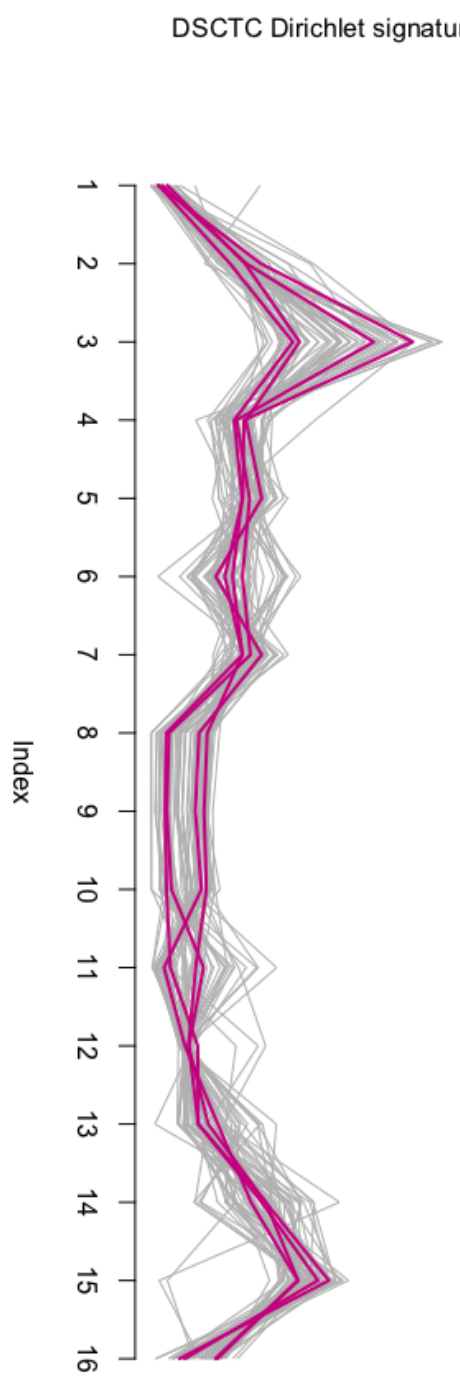
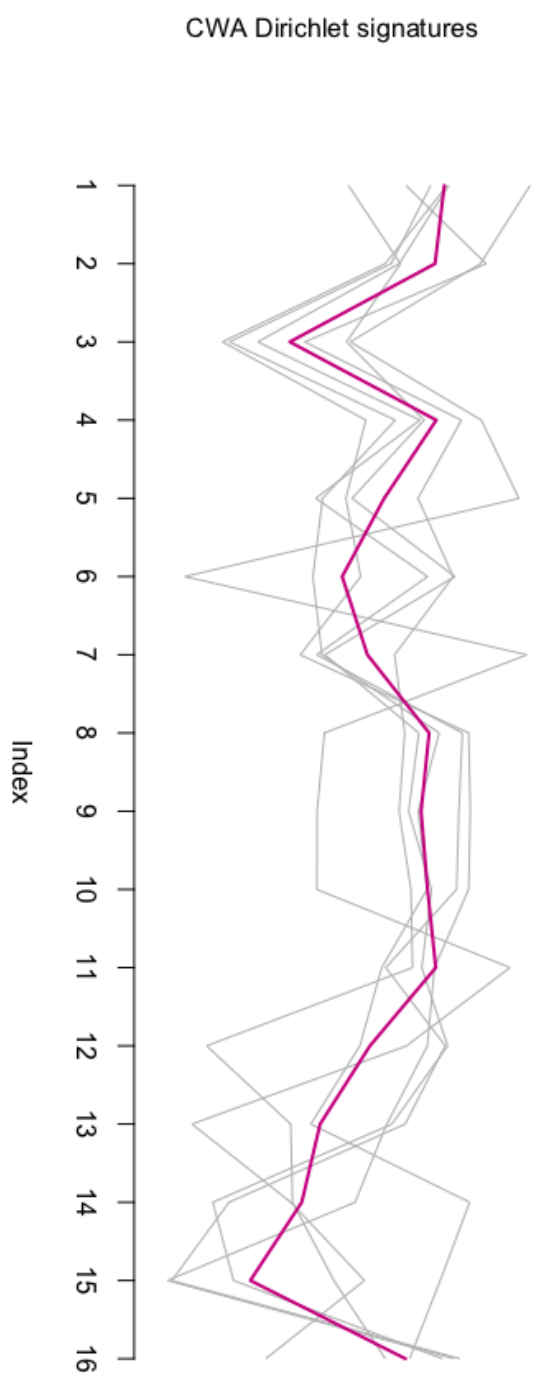

Log period to the base 10
3
2
1
0
-1
ACV
ACYG
BCEP
BE+GCAS
CEP(B)
CWA
CWB
DCEP
DCEPS
DSCT
DSCTC
EA
EB
ELL
EW
GDOR
LPV
RRAB
RRC
RS+BY
RV
SPB
SXARI
Variable star types

Figure A.2: The Dirichlet signature vectors are plotted against the data for each class. We see that the Dirichlet signatures pass through the data-well and hence the Dirichlet parameters for each component, seems to represent the data well

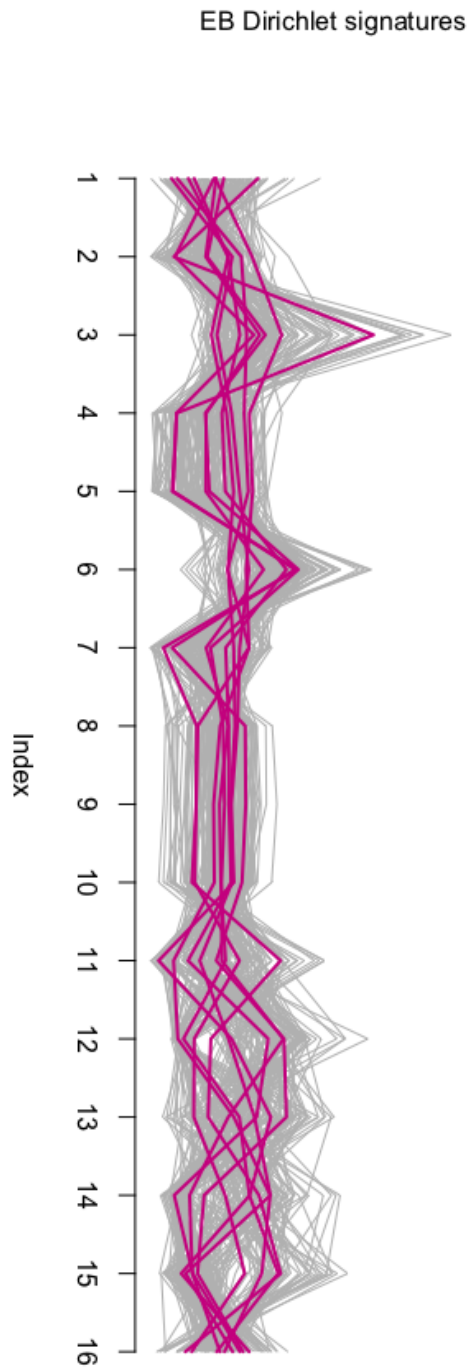


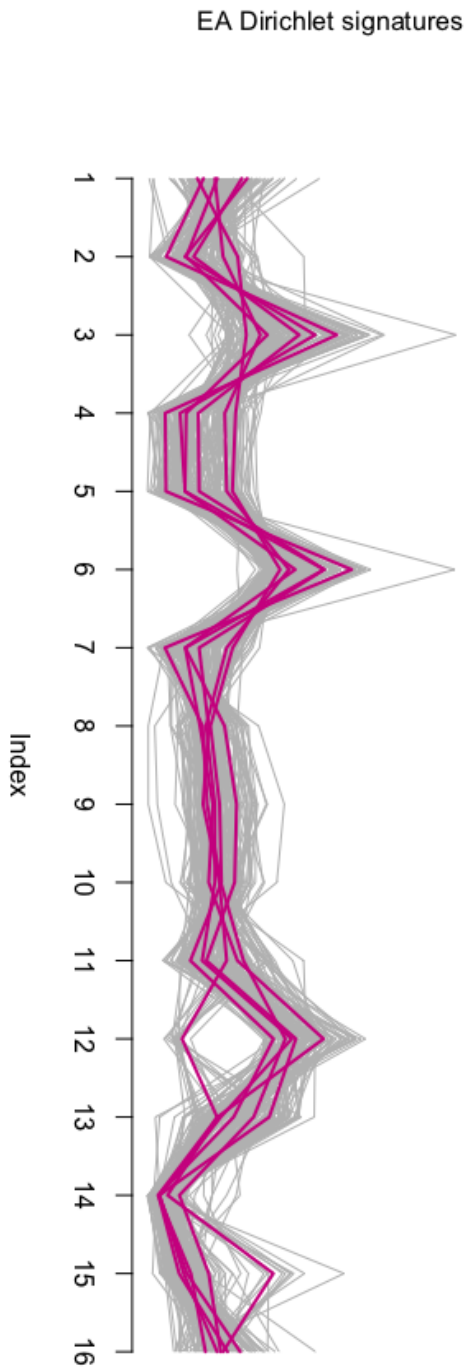

DSCTC Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


CWA Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

ELL Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

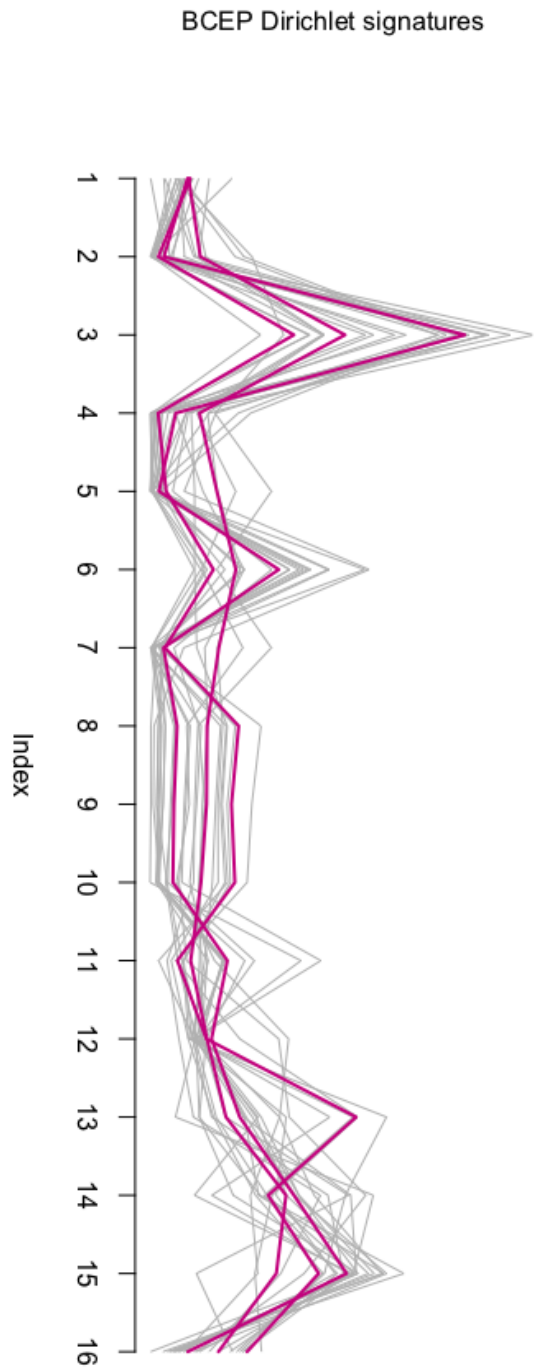
DSCT Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

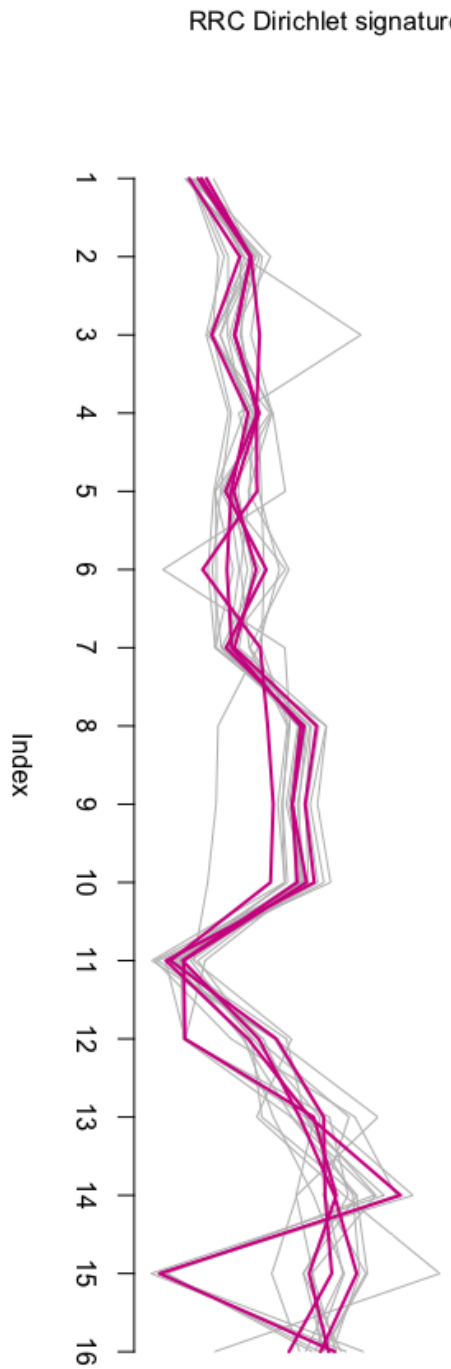

EB Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

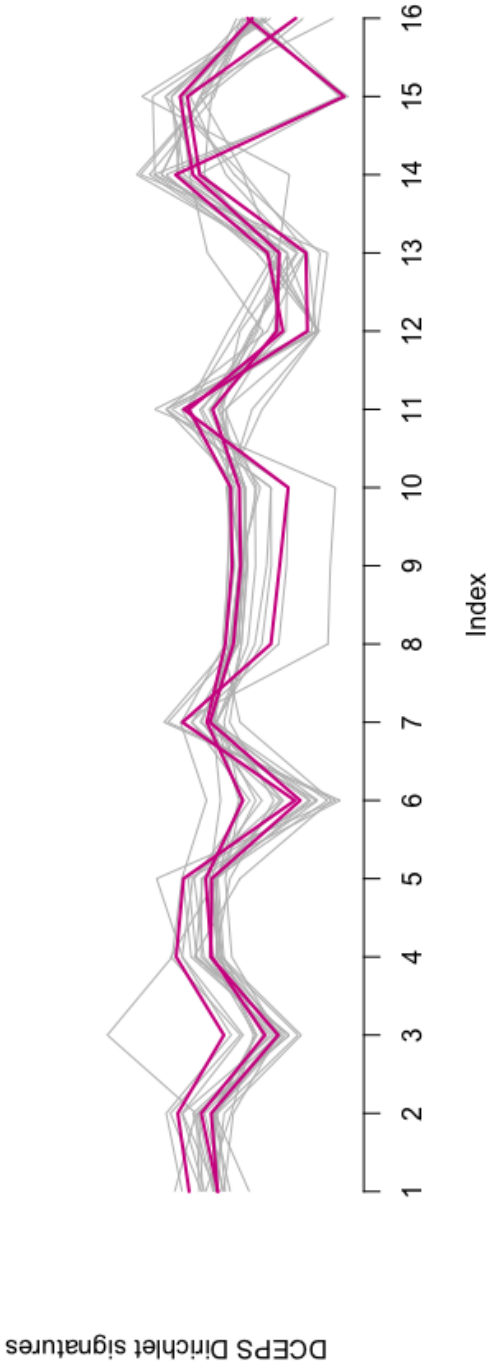

EA Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

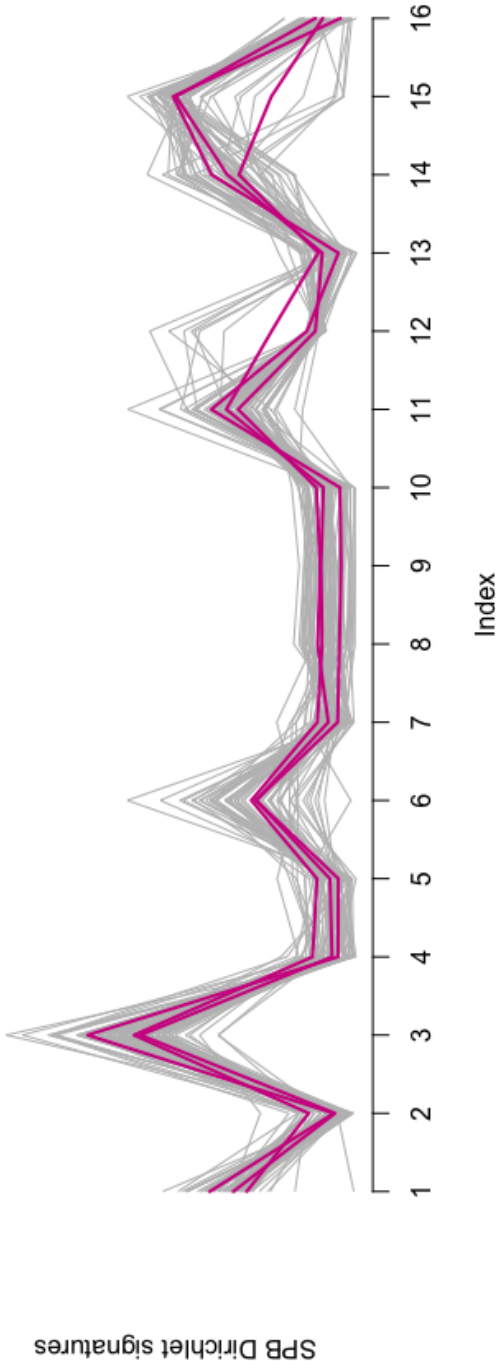
RRAB Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

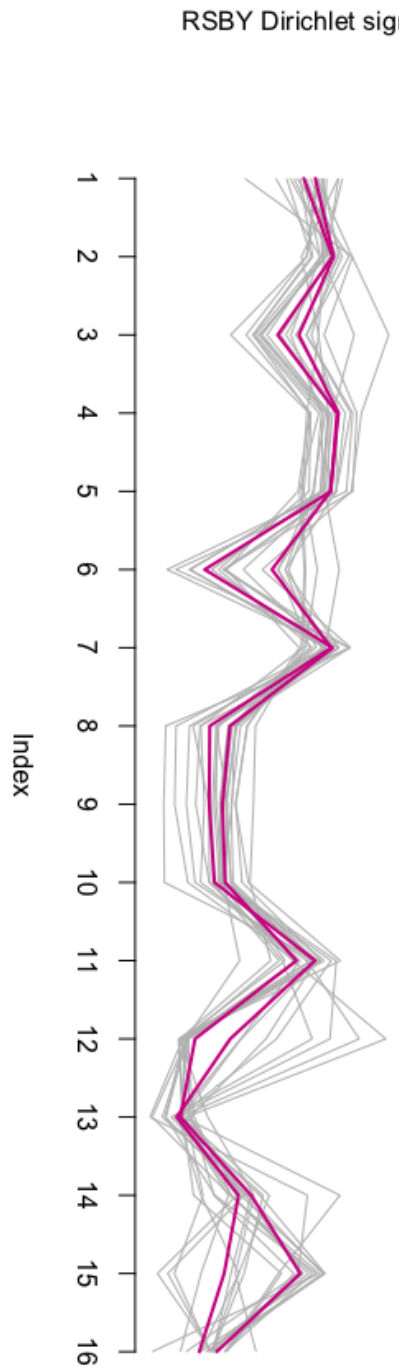
DCEP Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

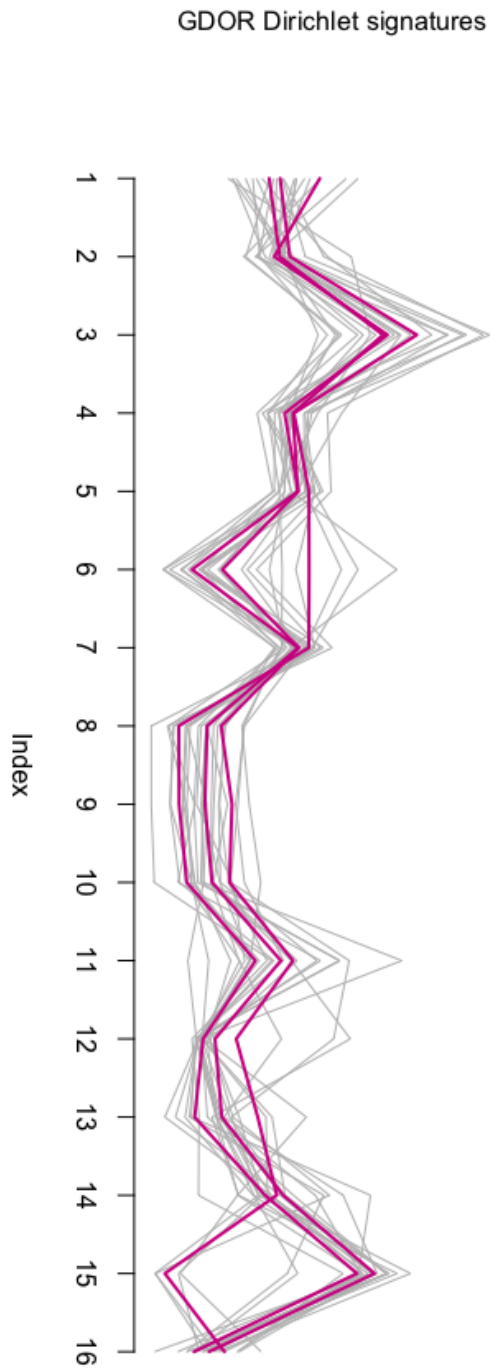

BCEP Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


RRC Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


DCEPS Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


SPB Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


RSBY Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16


GDOR Dirichlet signatures
Index
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16

Appendix B
Useful Basics

B.1 Expectation-Maximization (EM) algorithm

Expectation–Maximization (EM) algorithm is an iterative method to find maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step.

Consider the set of observed data $\mathbf{Y}$ and a set of unobserved latent data or missing values S and a vector of unknown parameters say, Θ, along with the likelihood function $\mathbf{L}(\theta; \mathbf{Y}) = f(\mathbf{y}|\theta) = \int f(\mathbf{y}, \mathbf{S}|\theta) d\mathbf{S}$. But since this is intractable, we use a method to iteratively find the maximum likelihood estimates of the marginal likelihoods by applying these two steps.

- **Expectation step :** The expected value of the log-likelihood, with respect to the conditional distribution of $\mathbf{S}$ given $\mathbf{Y}$ under the current estimate of the parameters θ

$$Q(\theta|\theta^{(t)}) = \mathrm{E}_{\mathbf{S}|\mathbf{Y},\theta^{(t)}} \left[\log L(\theta; \mathbf{Y}, \mathbf{S})\right]$$

- **Maximization step** : To find the values of the parameter that maximizes the Q function.

$$\theta^{(t+1)} = \arg\max_{\theta} Q(\theta|\theta^{(t)})$$

B.2 Exponential family

Exponential family with parameter Θ is a set of probability distributions of a certain form, specified below.

$$f(\mathbf{y_i}) = \frac{b(\mathbf{y_i})}{a(\theta)} \exp\left(\boldsymbol{\theta}^T T(\mathbf{y_i})\right)$$

where $T(\mathbf{y}), a(\theta)$ are vector of sufficient statistics and the cumulant generating function and $b(\mathbf{y_i})$ is a function of $\mathbf{y_i}$.

B.3 Empirical distribution

Let $\mathbf{y_i} = (y_{i1}, y_{i2}, \ldots, y_{iD})$ be independent, identically distributed real random variables with the common cumulative distribution function $F(t)$. Then the empirical distribution function is defined as

$$\hat{F}_n(t) = \frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{I}(y_i \leq t,)$$

where $\mathbf{I}A$ is the indicator of event A. Figure B.1 gives an illustration of the Empirical distribution.

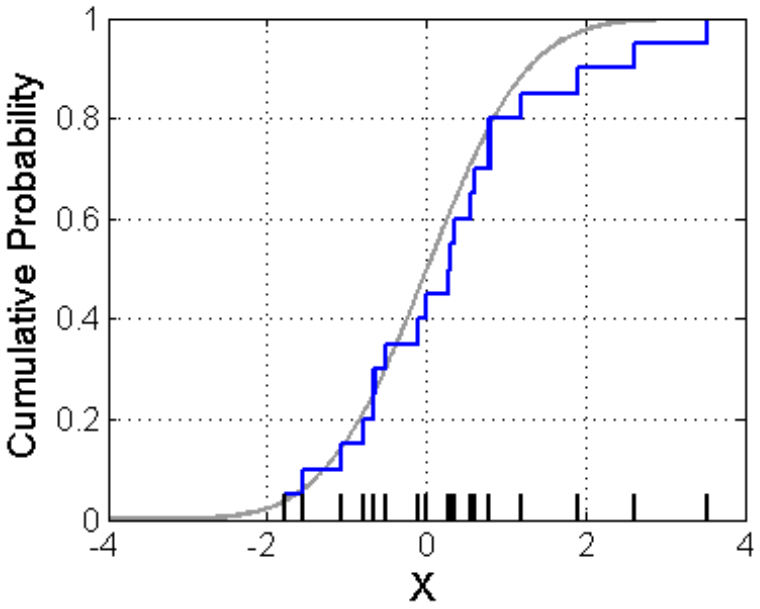


Figure B.1: The above plot is an illustration of the empirical distribution function. The black bars represent the samples corresponding to the empirical distribution function and the gray curve is the true cumulative distribution function. The blue line represents the empirical distribution function

B.4 logit and inverse logit transformation

The logit of a number p between 0 and 1 is given by the formula:

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \log(p) - \log(1-p) = -\log\left(\frac{1}{p} - 1\right).$$

The base of the logarithm function can be 10 or e but in our thesis we have used base *e*. Figure B.2 gives an illustration of the logit function.

The inverse logit is defined by $\exp(y)/(1+\exp(y))$ for $y \in \mathbb{R}$. Values in y of -Inf or Inf return logits of 0 or 1 respectively.

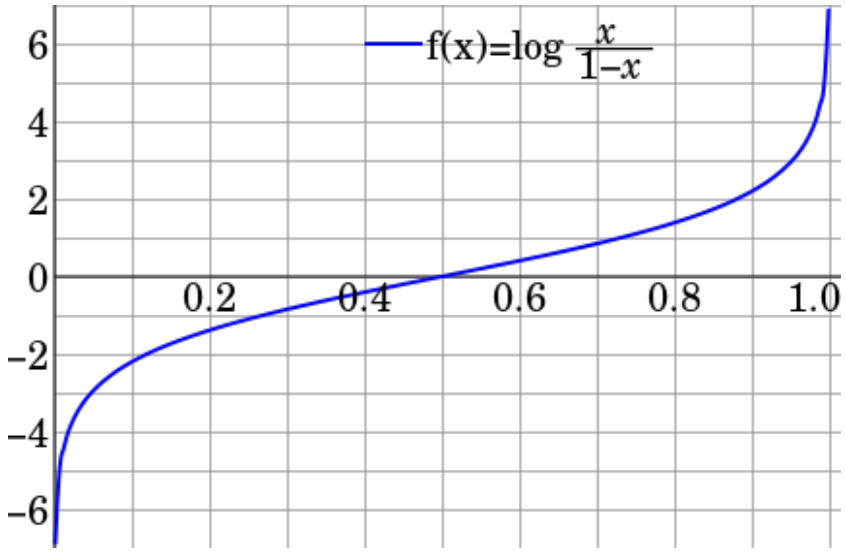


Figure B.2: Plot of logit(p) in the domain of 0 to 1, where the base of logarithm is *e*

B.5 Bayesian Information Criterion (BIC)

Bayesian information criterion (BIC) or Schwarz criterion (also SBC, SBIC) is a criterion for model selection among a finite set of models; the model with the highest BIC is preferred if the BIC is defined as below. It is based, in part, on the likelihood function.

$$\mathrm{BIC} = 2\ln(\hat{L}) - \ln(n)w.$$

where $\hat{L}$ is the maximized value of the likelihood function of the model, $\mathbf{y}$ is the observed data, n is the number of observations, or equivalently, the sample size; w is the number of free parameters to be estimated.

B.6 Dirichlet distribution in exponential family form

For deriving the exponential family form for a Dirichlet distribution of the form,

$$f_k(\mathbf{y_i}) = \frac{1}{\mathbf{B}(\boldsymbol{\alpha_k})} \prod_{d=1}^{D} y_{id}^{\alpha_{kd}-1} \quad \mathbf{y_i} \in \mathbb{V}_{D-1}$$

we have

$$\begin{aligned}\log f_k(\mathbf{y_i}) &= \sum_{d=1}^{D}(\alpha_{kd}-1)\log y_{kd} - \log \mathbf{B}(\boldsymbol{\alpha_k}) \\ &= \sum_{d=1}^{D}\alpha_{kd}\log y_{kd} - \sum_{d=1}^{D}\log y_{kd} - \log \mathbf{B}(\boldsymbol{\alpha_k}) \\ &= -\sum_{d=1}^{D}\log y_{kd} + \sum_{d=1}^{D}\alpha_{kd}\log y_{kd} - \log \mathbf{B}(\boldsymbol{\alpha_k})\end{aligned}$$

is of the form

$$\log f(\mathbf{y_i}) = \log b(\mathbf{y_i}) + \theta^T \mathbf{T}(\mathbf{y_i}) - \log a(\theta)$$

where

$$\theta = \begin{pmatrix}\alpha_{k1}\\ \alpha_{k2}\\ \vdots \\ \alpha_{kD}\end{pmatrix} \qquad \mathbf{T}(\mathbf{y_i}) = \begin{pmatrix}\log y_{i1}\\ \log y_{i2}\\ \vdots \\ \log y_{iD}\end{pmatrix} \tag{1}$$

$$b(\mathbf{y_i}) = \frac{1}{\prod_{d=1}^{D} y_{id}}$$

and

$$a(\theta) = \mathbf{B}(\boldsymbol{\alpha_k})$$

Bibliography

Aerts, C., Christensen-Dalsgaard, J., and Kurtz, D. W. (2010). *Asteroseismology*. Springer Science & Business Media.

Aerts, C., Eyer, L., and Kestens, E. (1998). The discovery of new gamma doradus stars from the hipparcos mission. *Astronomy and Astrophysics*, 337:790–796.

Akerlof, C., Amrose, S., Balsano, R., Bloch, J., Casperson, D., Fletcher, S., Gisler, G., Hills, J., Kehoe, R., Lee, B., et al. (2000). Rotse all-sky surveys for variable stars. i. test fields. *The Astronomical Journal*, 119(4):1901.

Belokurov, V., Evans, N. W., and Du, Y. L. (2003). Light-curve classification in massive variability surveys—i. microlensing. *Monthly Notices of the Royal Astronomical Society*, 341(4):1373–1384.

Belokurov, V., Evans, N. W., and Le Du, Y. (2004). Light-curve classification in massive variability surveys–ii. transients towards the large magellanic cloud. *Monthly Notices of the Royal Astronomical Society*, 352(1):233–242.

Bernardo, J. and Girón, F. (1988). A bayesian analysis of simple mixture problems. *Bayesian statistics*, 3(3):67–78.

Blomme, J., Debosscher, J., De Ridder, J., Aerts, C., Gilliland, R. L., Christensen-Dalsgaard, J., Kjeldsen, H., Brown, T. M., Borucki, W. J., Koch, D., et al. (2010). Automated classification of variable stars in the asteroseismology program of the kepler space mission. *The Astrophysical Journal Letters*, 713(2):L204.

Blomme, J., Sarro, L., O'Donovan, F., Debosscher, J., Brown, T., Lopez, M., Dubath, P., Rimoldini, L., Charbonneau, D., Dunham, E., et al. (2011). Improved methodology for the automated classification of periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 418(1):96–106.

Boldi, M.-O. (2004). Mixture models for multivariate extremes. *Thèse École polytechnique fédérale de Lausanne EPFL, n. 3098, Section de mathématiques, Faculté des sciences de base, Institut de mathématiques, Chaire de statistique*, page URL: https://infoscience.epfl.ch/record/33567.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1):5–32.

Castellani, V., Degl'Innocenti, S., and Fiorentini, G. (1993). The pp reaction in the sun and solar neutrinos. *Physics Letters B*, 303(1-2):68–74.

Catelan, M., Pritzl, B. J., and Smith, H. A. (2004). The rr lyrae period-luminosity relation. i. theoretical calibration. *The Astrophysical Journal Supplement Series*, 154(2):633.

Debosscher, J., Sarro, L., Aerts, C., Cuypers, J., Vandenbussche, B., Garrido, R., and Solano, E. (2007). Automated supervised classification of variable stars-i. methodology. *Astronomy & Astrophysics*, 475(3):1159–1183.

Debosscher, J., Sarro, L., López, M., Deleuil, M., Aerts, C., Auvergne, M., Baglin, A., Baudin, F., Chadid, M., Charpinet, S., et al. (2009). Automated supervised classification of variable stars in the corot programme-method and application to the first four exoplanet fields. *Astronomy & Astrophysics*, 506(1):519–534.

Deleuil, M., Meunier, J., Moutou, C., Surace, C., Deeg, H., Barbieri, M., Debosscher, J., Almenara, J., Agneray, F., Granet, Y., et al. (2009). Exodat: an information system in support of the corot/exoplanet science. *The Astronomical Journal*, 138(2):649.

Diaconis, P., Ylvisaker, D., et al. (1979). Conjugate priors for exponential families. *The Annals of statistics*, 7(2):269–281.

Drake, A., Graham, M., Djorgovski, S., Catelan, M., Mahabal, A., Torrealba, G., Garcia-Alvarez, D., Donalek, C., Prieto, J., Williams, R., et al. (2014). The catalina surveys periodic variable star catalog. *Astrophysical Journal Supplement Series*, 213(1):Art–No.

Dubath, P., Rimoldini, L., Süveges, M., Blomme, J., López, M., Sarro, L., De Ridder, J., Cuypers, J., Guy, L., Lecoeur, I., et al. (2011). Random forest automated supervised classification of hipparcos periodic variable stars. *Monthly Notices of the Royal Astronomical Society*, 414(3):2602–2617.

Eker, Z., Filiz-Ak, N., Bilir, S., Dogru, D., Tuysuz, M., Soydugan, E., Bakis, H., Ugras, B., Soydugan, F., Erdem, A., et al. (2008). Vizier online data catalog: Chromospherically active binaries. third version (eker+, 2008). *VizieR Online Data Catalog*, 5128:0.

Embrechts, P., Lindskog, F., and McNeil, A. (2001). Modelling dependence with copulas. *Rapport technique, Département de mathématiques, Institut Fédéral de Technologie de Zurich, Zurich*.

Eyer, L. and Blake, C. (2002). Automated classification of variable stars for asas data. In *IAU Colloq. 185: Radial and Nonradial Pulsationsn as Probes of Stellar Physics*, volume 259, page 160.

Eyer, L. and Blake, C. (2005). Automated classification of variable stars for all-sky automated survey 1–2 data. *Monthly Notices of the Royal Astronomical Society*, 358(1):30–38.

Eyer, L. and Cuypers, J. (2000). Predictions on the number of variable stars for the gaia space mission and for surveys such as the ground-based international liquid mirror telescope. In *International Astronomical Union Colloquium*, volume 176, pages 71–72. Cambridge University Press.

Feast, M. and Walker, A. (1987). Cepheids as distance indicators. *Annual review of astronomy and astrophysics*, 25(1):345–375.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer Science & Business Media.

Furness, C. E. (1915). *An Introduction to the Study of Variable Stars*. Houghton Mifflin.

Gautschy, A. and Saio, H. (1996). Stellar pulsations across the hr diagram: Part ii. *Annual Review of Astronomy and Astrophysics*, 34(1):551–606.

Ghahramani, Z. and Beal, M. J. (2000). Variational inference for bayesian mixtures of factor analysers. In *Advances in neural information processing systems*, pages 449–455.

Gilman, C. (1978). John goodricke and his variable stars. *Sky and Telescope*, 56.

Hopkins, J. (1976). Glossary of astronomy and astrophysics. *Chicago, University of Chicago Press, 1976. 174 p.*

Hoskin, M. (1982). Stellar astronomy. historical studies. *Chalfont, St. Giles: Science History Publication, 1982.*

Kim, D.-W., Protopapas, P., Bailer-Jones, C. A., Byun, Y.-I., Chang, S.-W., Marquette, J.-B., and Shin, M.-S. (2014). The epoch project-i. periodic variable stars in the eros-2 lmc database. *Astronomy & Astrophysics*, 566:A43.

Leavitt, H. S. (1908). 1777 variables in the magellanic clouds. *Annals of Harvard College Observatory*, 60:87–108.

Leavitt, H. S. and Pickering, E. C. (1912). Periods of 25 variable stars in the small magellanic cloud. *Harvard College Observatory Circular*, 173:1–3.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomforest. *R News*, 2(3):18–22.

Markou, M. and Singh, S. (2003). Novelty detection: a review—part 1: statistical approaches. *Signal processing*, 83(12):2481–2497.

McLachlan, G. and Peel, D. (2004). *Finite mixture models*. John Wiley & Sons.

Minniti, D., Lucas, P., Emerson, J., Saito, R., Hempel, M., Pietrukowicz, P., Ahumada, A., Alonso, M., Alonso-Garcia, J., Arias, J. I., et al. (2010). Vista variables in the via lactea (vvv): The public eso near-ir variability survey of the milky way. *New Astronomy*, 15(5):433–443.

Ng, K. W., Tian, G.-L., and Tang, M.-L. (2011). *Dirichlet and related distributions: Theory, methods and applications*, volume 888. John Wiley & Sons.

Percy, J. R. (2007). *Understanding variable stars*. Cambridge University Press.

Perryman, M. (1997). Esa 1997, the hipparcos and tycho catalogues. astrometric and photometric star catalogues derived from the esa hipparcos space astrometry mission. *ESA SP*, 1200.

Perryman, M. (2010). *The Making of History's Greatest Star Map*. Springer Science & Business Media.

Perryman, M. A., Lindegren, L., Kovalevsky, J., Hoeg, E., Bastian, U., Bernacca, P., Crézé, M., Donati, F., Grenon, M., Grewing, M., et al. (1997). The hipparcos catalogue. *Astronomy and Astrophysics*, 323.

Pietrukowicz, P., Dziembowski, W. A., Latour, M., Angeloni, R., Poleski, R., di Mille, F., Soszynski, I., Udalski, A., Szymanski, M. K., Wyrzykowski, L., et al. (2017). Blue large-amplitude pulsators as a new class of variable stars. *arXiv preprint arXiv:1706.07802*.

Pojmanski, G. (2002). The all sky automated survey. catalog of variable stars. i. 0 h-6 hquarter of the southern hemisphere. *Acta Astronomica*, 52:397–427.

Pojmanski, G. (2003). The all sky automated survey. the catalog of variable stars. ii. 6ˆ h-12ˆ h quarter of the southern hemisphere. *Acta Astronomica*, 53:341–369.

Pojmanski, G. (2004). The all sky automated survey. the catalog of variable stars. ii. 6h-12h quarter of the southern hemisphere. *arXiv preprint astro-ph/0401125*.

Powell, R. (2006). Hertzsprung russell diagram. *An Atlas of the Universe*.

Richards, J. W., Starr, D. L., Butler, N. R., Bloom, J. S., Brewer, J. M., Crellin-Quick, A., Higgins, J., Kennedy, R., and Rischard, M. (2011). On machine-learned classification of variable stars with sparse and noisy time-series data. *The Astrophysical Journal*, 733(1):10.

Samus, N., Kazarovets, E., and Durlevich, O. (2017). General catalogue of variable stars. *Odessa Astronomical Publications*, 14:266–269.

Sarro, L., Debosscher, J., López, M., and Aerts, C. (2009). Automated supervised classification of variable stars-ii. application to the ogle database. *Astronomy & Astrophysics*, 494(2):739–768.

Sesar, B., Vivas, A. K., Duffau, S., and Ivezić, Ž. (2010). Halo velocity groups in the pisces overdensity. *The Astrophysical Journal*, 717(1):133.

Steinfadt, J. D., Kaplan, D. L., Shporer, A., Bildsten, L., and Howell, S. B. (2010). Discovery of the eclipsing detached double white dwarf binary nltt 11748. *The Astrophysical Journal Letters*, 716(2):L146.

Süveges, M., Barblan, F., Lecoeur-Taïbi, I., Prša, A., Holl, B., Eyer, L., Kochoska, A., Mowlavi, N., and Rimoldini, L. (2017). Gaia eclipsing binary and multiple systems. supervised classification and self-organizing maps. *Astronomy & Astrophysics*, 603:A117.

Turon, C., Crézé, M., Egret, D., Gómez, A., Grenon, M., Jahreiß, H., Réquieme, Y., Argue, A., Bec-Borsenberger, A., Dommanget, J., et al. (1992). The hipparcos input catalogue. *Bulletin d'Information du Centre de Donnees Stellaires*, 41:9.

Turon, C., Requieme, Y., Grenon, M., Gomez, A., Morin, D., Crifo, F., Arenou, F., Froeschle, M., Mignard, F., Perryman, M., et al. (1995). Properties of the hipparcos input catalogue. *Astronomy and Astrophysics*, 304:82.

Udalski, A., Szymański, M., and Szymański, G. (2015). Ogle-iv: fourth phase of the optical gravitational lensing experiment. *arXiv preprint arXiv:1504.05966*.

Unsöld, A. (1969). Colour-magnitude diagrams of galactic and globular clusters. stellar evolution and abundances of the elements. In *The New Cosmos*, pages 259–280. Springer.

Vatanen, T., Kuusela, M., Malmi, E., Raiko, T., Aaltonen, T., and Nagai, Y. (2012). Semi-supervised detection of collective anomalies with an application in high energy particle physics. In *Neural Networks (IJCNN), The 2012 International Joint Conference on*, pages 1–8. IEEE.

Watson, C., Henden, A., and Price, A. (2009). Vizier online data catalog: Aavso international variable star index vsx (watson+, 2009). *VizieR Online Data Catalog*, 1:02027.

Willemsen, P. and Eyer, L. (2007). A study of supervised classification of hipparcos variable stars using pca and support vector machines. *arXiv preprint arXiv:0712.2898*.

Williams, T. R. and Saladyga, M. (2011). *Advancing variable star astronomy: The centennial history of the american association of variable star observers*. Cambridge University Press.

Glossary

absolute magnitude Absolute magnitude is defined to be the apparent magnitude an object would have if it were located at a distance of 10 parsecs. 24

asteroseismology Science that studies the internal structure of stars by the interpretation of their oscillation modes, determined from the frequency spectra of their light curves. 18

categorical random variables Random variable that can take on one of a limited, and usually fixed, number of possible values, assigning each individual or other unit of observation to a particular group or nominal category on the basis of some criterion. The probability distribution associated with a categorical random variable is called a categorical distribution. 9

energy level Discrete values of energy of electrons bound to the nucleus of an atom. 20

luminosity the amount of electromagnetic energy emitted by a body per time unit. Hopkins (1976). 17

photometry Photometry is the science of the measurement of light, in terms of its perceived brightness to the human eye or to a detector. 41

photon Fundamental particle of visible light. 20

prism This is a block of glass with a triangular cross-section. 19

refracted The change in direction of a wave passing from one medium to another caused by its change in speed. 19

spectrum the electromagnetic spectrum is the energy density of the electromagnetic radiation per unit frequency or unit wavelength. 19

Prince John

CURRICULUM VITAE

Contact Information

University of Padova
Department of Statistics
via Cesare Battisti, 241-243
35121 Padova. Italy.

Tel. +39 388 1799 031
e-mail: prince@stat.unipd.it

Current Position

Since November 2014; (expected completion: March 2018)
PhD Student in Statistical Sciences, University of Padova.
Thesis title: Finite Dirichlet mixture models for classification and detection of new classes of variable stars
Supervisor: Prof. Alessandra R. Brazzale
Co-supervisor: Dr. Maria Süveges

Research interests

- Astrostatistics
- Statistical Modeling of multivariate data
- Bayesian classification

Education

July 2010 – July 2012
Master (*laurea specialistica/magistrale***) degree in Mathematics**.
Indian Institute of Technology, Faculty of Mathematics
Title of dissertation: Non Parametric Partial Sequential test for Location at an unknown time point
Supervisor: Prof. Amitava Mukherjee
Final mark: 6.93/10 CGPA

June 2005 – June 2008
Bachelor degree (*laurea triennale***) in Mathematics**.
Mahatma Gandhi University, Faculty of Mathematics
Final mark: 89.3%

Oct 2015 – Aug 2017
Diploma degree in Bible and Doctrine.
Global University, Berean Bible School
Final mark: 3.94/4 GPA

Visiting periods

Feb 2017 – April 2017
Max Planck Institute of Astronomy,
Heidelberg, Germany.
Supervisor: Dr. Maria Süveges

Further education

Aug 2012
Foundations of Business analysis
ESI International
Organizer: Scope International

Conference, Seminars and workshop presentations

Sep 2017	Junior speed session at Astro@Stats : Finite Dirichlet Mixture Models for Classification and detection of new classes of Variable Stars
Sep 2017	Poster presentation at Astro@Stats Workshop Finite Dirichlet Mixture Models for Classification and detection of new classes of Variable Stars
Oct 2017	3MT presentation competition winner : Finite Dirichlet Mixture Models for Classification and detection of new classes of Variable Stars

Work experience

July 2012 – May 2014
Scope International, Chennai, India.
Analyst

Oct 2008 – Oct 2009
Wipro Technologies, Bangalore, India.
SAP MM Functional consultant

Awards and Scholarship

Oct 2017
Winner of 3 minute thesis presentation competition held in University of Padova.

Nov 2014- Nov 2017
PhD scholarship, University of Padova.

Feb 2008
Luminary Award for the best outgoing student of St. Berchman's College, India.

Computer skills

- R
- Visual Basic 6.0
- SAP MM and SAP ABAP
- SQL
- MS Excel

Language skills

Malayalam: native; English: fluent; Hindi: good ; Tamil: basic ; Italian : basic.

Other Interests

Worship leading
Guitaring
Blogging

References

Prof. Alessandra Brazzale
Department of Statistical Sciences
University of Padova
Via Cesare Battisti, 241
35121, Padova, Italy
alessandra.brazzale@unipd.it

Dr. Amitava Mukherjee
XLRI Jamshedpur
Circuit House Area, Sonari,
Jamshedpur, Jharkhand 831001, India
amitmukh2@gmail.com