Communications for Statistical Applications and Methods
2013, Vol. 20, No. 4, 271–281

DOI: http://dx.doi.org/10.5351/CSAM.2013.20.4.271

Functional Data Classification of Variable Stars

Minjeong Park[a], Donghoh Kim[b], Sinsup Cho[c], Hee-Seok Oh[1, c]

[a]Statistical Research Institute, Statistics Korea
[b]Department of Applied Mathematics, Sejong University
[c]Department of Statistics, Seoul National University

Abstract

This paper considers a problem of classification of variable stars based on functional data analysis. For a better understanding of galaxy structure and stellar evolution, various approaches for classification of variable stars have been studied. Several features that explain the characteristics of variable stars (such as color index, amplitude, period, and Fourier coefficients) were usually used to classify variable stars. Excluding other factors but focusing only on the curve shapes of variable stars, Deb and Singh (2009) proposed a classification procedure using multivariate principal component analysis. However, this approach is limited to accommodate some features of the light curve data that are unequally spaced in the phase domain and have some functional properties. In this paper, we propose a light curve estimation method that is suitable for functional data analysis, and provide a classification procedure for variable stars that combined the features of a light curve with existing functional data analysis methods. To evaluate its practical applicability, we apply the proposed classification procedure to the data sets of variable stars from the project STellar Astrophysics and Research on Exoplanets (STARE).

Keywords: Classification, functional data analysis, principal component analysis, variable star.

1. Introduction

Any star whose brightness fluctuates is called a variable star. The fluctuation of brightness, which is caused by actual luminosity changes or something that partly blocks light can be regular or irregular according to some reason. Intrinsic and extrinsic variables are two big categories for variable stars. Intrinsic variables whose luminosity actually changes include Cepheid, pulsating, cataclysmic and erupted variables. Eclipsing binary and rotating variables are categorized as extrinsic variables whose brightness changes come from variations in the amount of light that reaches Earth.

Studies on variable stars have been conducted to obtain information on the structure and composition of the star along with properties such as mass, radius, temperature and luminosity. Moreover, variable stars provide an important clue to understand the Sun, the age of the universe, distant galaxies, and an expanding universe.

Among the several issues related to variable stars studies, in this paper, we consider how to efficiently classify variable stars. We propose a light curve estimation method that is appropriate for the functional data analysis (FDA), and provide a effective classification method by coupling of the proposed the light curve estimation method and existing FDA methods. We focus on only the light curve shape among various characteristics in order to classify variable stars, which share the same

This work was supported by a grant from the Next-Generation BioGreen 21 Program (No. PJ008068), Rural Development Administration, Republic of Korea and Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2012R1A1A2039968).

[1] Corresponding author: Professor, Department of Statistics, Seoul National University, 599 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea. E-mail: heeseok@stats.snu.ac.kr

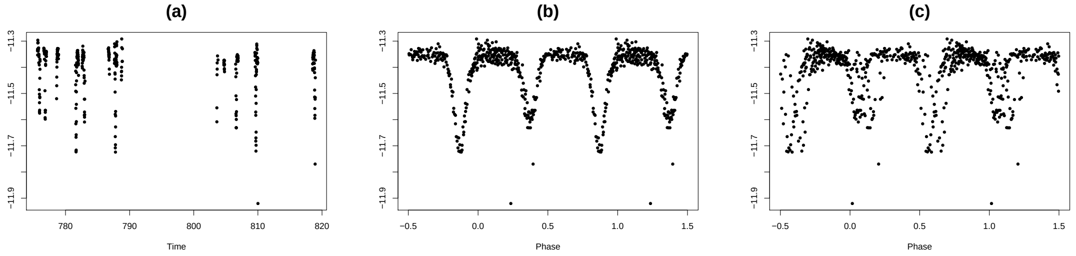

Figure 1: *Period estimation and light curves*

motivation of Deb and Singh (2009). We employ FDA in order to treat the light curve shapes properly. Since the light curve data are unequally spaced on the phase domain and have some functional properties, FDA methodology could be more appropriate than multivariate data analysis used in many existing classification methods that include Deb and Singh (2009) that are briefly reviewed in Section 2. Key features that distinguish the proposed method from most existing classification methods of variable stars are: (1) the proposed method focuses on light curve shape rather than the characteristics of the light curve and (2) we employ FDA instead of multivariate data analysis, which is capable of representing light curve data efficiently. An objective method to classify variable stars efficiently in massive surveys might be useful to identify new types of variable stars and detect unusual objects.

The rest of this paper is organized as follows. Section 2 reviews the literature of period estimation and classical classification of variable stars. We also review functional data analysis for the proposed classification method. Section 3 proposes a classification procedure based on FDA. Section 4 presents a description of the data used in this study and the classification results. Finally, Section 5 addresses the concluding remarks.

2. Background: Period Estimation, Classification Method, and Functional Data Analysis

2.1. Period estimation for variable stars

Brightness time series are obtained in the physical time domain as shown on panel (a) in Figure 1. Light curves in the phase domain like one on panel (b) can be obtained through a period estimation. Since the classification procedure is based on the light curves, estimating period is an important issue before classifying variable stars. Therefore, we additionally review how to estimate periods of variable stars and its importance.

Suppose that a time series $\{(t_i, x_i)\}$ with a single period p is observed from the model

$$x_i = f\left(\frac{t_i}{p}\right) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where x_i, and ϵ_i are the brightness measurement and measurement error, respectively, at the i^{th} sampling time, t_i. The f denotes a periodic function on $[0, 1]$. We extend the phase domain by half to both sides for a visually better representation of a light curve. The extension also alleviates boundary problems in analyzing light curves.

To obtain a light curve, we fold the brightness time series with a period p and obtain the light curve in the phase domain. Note that the brightness in the physical time domain is unequally obtained since stars can be observed only at certain times of day. The light curve data in the phase domain are also measured at unpredicted locations by folding and therefore are unequally spaced.

Figure 1 shows in panel (a), that the observations are unequally spaced with very long interruptions and that it is not easy to find the periodicity of the star in the time domain. By folding data over the period $p = 0.87642$ (days) of variability, we can obtain the light curve in panel (b). However, the light curve induced by an incorrect period produces a blurred pattern of brightness as shown in panel (c), which does not provide useful information for identifying the star. Thus, period estimation is an important issue for variable star studies.

Based on the periodogram, simple cosine models were considered in Deeming (1975), Lomb (1976), and Scargle (1982), respectively. However, in Lafler and Kinman (1965), Stellingwerf (1978) and Dwortesky (1983), respectively, some measures of dispersion in the light curve were used to find the period. Nonparametrically, based on cubic B-splines, Akerlof *et al.* (1994) suggested a method fitting period-folded light curves with examples in the MAssive Compact Halo Objects (MACHO) dataset. As another nonparametric method, Reimann (1994) fitted the brightness as a function of phase at a given period using the SuperSmoother of Friedman (1984) and obtained four final candidates of period. In Hall *et al.* (2000), a local linear smoother was suggested to estimate both the period and amplitude function. Finally, Oh *et al.* (2004) proposed a robust method to estimate periods and light curves, which is less sensitive to the outliers.

2.2. Previous classification methods for variable stars

Most previous studies extracted some characteristics of the light curve for the input variables of classification, and adapted classification methods such as a Bayesian classifier, support vector machine (SVM), and discriminant analysis. Hegland *et al.* (2001) studied the classification of variable stars in the MACHO dataset and developed software for astronomers. They employed various characteristics such as color index, amplitude, average of magnitude, difference and correlation between red and blue magnitude, average of frequencies, and average frequency spread. They suggested an algorithm to find variable stars with extreme values using a Bayesian classifier, *k*-means clustering, and boxing algorithm that allows astronomers to decide if each of the light curves belongs to a specific class or not. However, it does not provide a fully automatic procedure for classification.

Woźniak *et al.* (2004) and Usatov and Nosulchik (2008) selected five characteristics that include period, amplitude, and several independent color indexes; subsequently, they employed SVM method to classify the slowly varying stars in Northern Sky Variability Survey database. Willemsen and Eyer (2005) also used a SVM-based classification method with 51 characteristics such as skewness, the median subtracted 10-percentiles, and forty bins from the Fourier envelope of light curve.

Debosscher *et al.* (2008) used Fourier coefficients as characteristics and then developed a multistage tree procedure to combine Gaussian mixture classifier and SVM. Some real data analysis following their procedure was performed in Sarro *et al.* (2009) and Blomme *et al.* (2010). Most existing classification methods of variable stars focus on extracting some characteristics that can be appropriately used as input variables of multivariate data analysis.

Conversely, Deb and Singh (2009) utilized the light curve shape to classify variable stars. Deb and Singh (2009) conducted a comparison study for the classification results between the characteristics of the light curve and the light curve shape based on principal component analysis (PCA). They showed that their approach would perform better. When finding some features representing light curve shapes, PCA might be a natural approach as in Deb and Singh (2009). In their study, they selected 100 equally spaced points in each light curve by an interpolation method in order to carry out PCA directly on the light curve.

2.3. Functional data analysis

While a set of given data is treated as n length of vector in multivariate data analysis, FDA deals with the same data as a single function under the assumption that the data is a realization of a function. From the theoretical viewpoint, FDA has several merits; dealing unequally spaced data and avoiding high dimension problems (Ramsay and Silverman, 2005). In practice, data can be considered as discretized functions rather than as standard vectors. In this study, we reasonably assume that each variable star has its own light curve function, which means that each light curve is a set of functional data. Hence, it is natural to employ the FDA approach for classification of the light curves. We briefly review the basic idea of FDA. The full explanation of FDA can be found in Ramsay and Silverman (2005).

Let y_j be a realization of smooth function x at time t_j with error ϵ_j. Then we have the following model of observations

$$y_j = x(t_j) + \epsilon_j, \quad j = 1, 2, \ldots, n,$$

and represent the underlying function $x(t)$ as an expansion of some basis functions $\boldsymbol{\phi} = (\phi_1(t), \ldots, \phi_L(t))^T$,

$$x(t) = \sum_{\ell=1}^{L} c_\ell \phi_\ell(t) = \boldsymbol{c}^T \boldsymbol{\phi}, \tag{2.2}$$

where $\boldsymbol{c} = (c_1, \ldots, c_L)^T$. For estimating the underlying function, we consider a linear smoother minimizing the following least squares criterion (Ramsay and Silverman, 2005)

$$Q(\boldsymbol{c}) = (\boldsymbol{y} - \boldsymbol{\Phi c})^T (\boldsymbol{y} - \boldsymbol{\Phi c}),$$

where $\boldsymbol{y} = (y_1, \ldots, y_L)^T$ and $\boldsymbol{\Phi}$ is an n by L matrix containing $\phi_\ell(t_j)$ values. The estimator of the coefficient vector $\boldsymbol{c}$ is $\hat{\boldsymbol{c}} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{y}$ provided that $(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}$ exists, and the corresponding fitted vector at each sampling point t_j is $\hat{\mathbf{x}} = \boldsymbol{Sy}$, where $\boldsymbol{S} = \boldsymbol{\Phi}(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T$. In order to avoid too wiggle approximation, a regularization method that minimizes the penalized least squares can be used. For details, refer to Ramsay and Silverman (2005). Applications in different fields have been developed, and several examples are found in Ramsay and Silverman (2002). The software for FDA are also available (Clarkson *et al.*, 2005; Ramsay *et al.*, 2009).

3. Variable Star Classification Based on FDA

3.1. Functional data classification

For the efficient classification of functional data, we need to consider the classification method as well as the functional features used as input variables for the classification method. The functional features should effectively represent the shape of the light curve function for classification. Taking the expansion of function into consideration as in Equation (2.2), the fitted function values $\hat{\mathbf{x}}$ and the estimator of the coefficient $\hat{\boldsymbol{c}}$ of some basis function for example Fourier basis, would be proper features. However, PCA is extended to extracting important functional features from a set of functional data since most multivariate data analysis can be generalized to a functional approach (Ramsay and Silverman, 2005). Therefore, functional principal component (PC) scores could be natural candidates. Once some functional features are obtained, we consider how to functionally classify light curves.

Table 1: Functional classification methods. Note that NP density implies Bayesian classifier with nonparametrically estimated density.

Input variables	Classification method	Reference
Functional PC scores	QDA	Hall *et al.* (2001)
	NP density	Ferraty and Vieu (2003)
	FSVM	Rossi and Villa (2006)
Coefficients of Fourier basis	QDA	Biau *et al.* (2005)
	FSVM	Rossi and Villa (2006)
Fitted function values	NP density	Ferraty and Vieu (2003)
	FSVM	Rossi and Villa (2006)

In Appendix, we intensively review which functional features can be used for classification methods in the FDA approaches with describing some details of Bayesian classifier, k-nearest neighbor (k-NN) and support vector machine (SVM) aspects, respectively. We believe that this summarization is useful for potential readers who are interested in functional data classification.

According to the combination of input variable and classification method, several functional classification methods have been recently proposed. Table 1 lists the methods used in this paper. Biau *et al.* (2005) and Hall *et al.* (2001) proposed a classification procedure based on quadratic discriminant analysis (QDA). They used the coefficients of Fourier basis and functional PC scores for the analysis, respectively. Rossi and Villa (2006) introduced the functional support vector machine (FSVM). The functional PC scores, the fitted function values, and coefficients of a basis system can be utilized for FSVM (Rossi and Villa, 2006). Ferraty and Vieu (2003) proposed the Bayesian classification method. Ferraty and Vieu (2003) adapted the nonparametric density estimation for f_g. For this approach, the fitted function values and functional PC scores can be used as input for classification.

3.2. The proposed functional data classification procedures for variable stars

Before classifying variable stars based on light curve shapes, the correct form of light curve must be estimated. There are several challenges to deal with light curve data: (1) huge amount of data, (2) large errors, (3) unequally spaced observations, and (4) existence of big gaps between observations even in the phase domain. The original light curve time series are obtained on the physical time domain. Light curves in the phase domain can be obtained by folding the time series with its period. To better represent a light curve, we usually extend the phase domain by half to each left and right side, respectively. The extension can alleviate boundary problems in analyzing light curves. Note that the brightness on the physical time domain is unequally obtained since stars can be observed only at certain times of day. The light curve data in the phase domain are also measured at unpredicted locations by folding and are therefore unequally spaced. The light curve should be registered to be suitable for FDA. Coping with several challenges and considering the aspects of FDA, we propose the efficient light curve estimation procedure. The detailed preprocessing procedure for functional classification is as follows:

1. Fold the given brightness time series by its period, and obtain an initial light curve on the phase domain $[0, 1]$.

2. Delete outliers using quantiles on a moving window from the light curve data.

3. Register the light curve data to be located from phase 0 (basic registration).

4. Standardize the curve by migrating minimum value to 0 and maximum value to 2.

5. Extend the light curve on the phase domain $[-0.5, 1.5]$.
6. Obtain the best fit by FDA with roughness penalty for each light curve data.
7. Refit the fitted light curves with a set of common basis functions.
8. Perform landmark registration, so that each estimated light curve function has its minimum at phase 0 and 1 (functional registration).

Here, we remark some details of the above steps.

Step 2 (Removing outliers): For a given phase t, consider a moving window $B_t = [t - 0.05, t + 0.05]$ with width 0.1. If an observation is outside of the 10% and 90% quantiles of the data on the window B_t, then we regard that observation as an outlier and delete it.

Step 6 (Obtaining the best fit): In this step, we estimate the light curve as a function. The quality of the fit is closely related to the number of basis and the smoothing parameter. In this study, we use 100 spline basis functions to capture all the variations in a light curve, and then provide a roughness penalty to impose further smoothness on the resulting fit. Moreover, for fitting the shape appropriately, we select knots from the quantiles of phase, which can select the basis functions data-adaptively. In this manner, we obtain the best fit for each light curve.

Step 7 (Refitting): For the use of functional classification methods and landmark registration in Step 8, it is required to use the same basis functions for all light curves in Step 7. Therefore we refit the independently estimated curves with a set of common basis functions, whose knots are equally spaced.

Figure 2 shows an example for the preprocessing steps. Figure 2(a) shows an initial light curve from Step 1. The red lines represent 10% and 90% quantiles obtained on the moving window. After deleting outliers through Step 2, we obtain Figure 2(b). Next, by following Step 3, we relocate the minimum point of a light curve at phase 0; subsequently, we can register the light curve. Then we standardize the light curve at Step 4. Finally, we extend it on the phase domain $[-0.5, 1.5]$, and obtain a light curve shown in Figure 2(d). Figure 2(c) shows the result of the preprocessing without using Step 2; the resultant light curve data is contaminated by outliers.

Figure 3 shows an example for Steps 6–8. In Figure 3(a), the blue line is the best fit in Steps 6 and 7. However, from Figure 3(a), we observe that the minimum of estimated function do not locate on the phases 0 and 1, and the minimum points of light curve data are not identical to those of the estimated light curve function (blue line). By using the warping function shown in Figure 3(b), we obtain the red registered light curve function shown in Figure 3(c), which will be used for further analysis.

Once we obtain the registered light curve functions of all candidate variable stars by the above preprocessing procedure, we apply the three functional classification procedures listed in Table 2 to the light curve functions. Here, we take a cross-validated classification step. Specifically, (i) first select a curve and treat it as a new one; (ii) determine the class by applying the above classification methods to the remaining curves; and (iii) obtain the predicted class (PC) and then compare it with the observed one (OC). In this manner, all curves are applied for classification.

4. Data and Classification Results

There have been numerous photometric surveys such as MACHO (Hegland *et al.*, 2001), Optical Gravitational Lensing Experiment (OGLE) (Sarro *et al.*, 2009), and STARE. Each survey has its

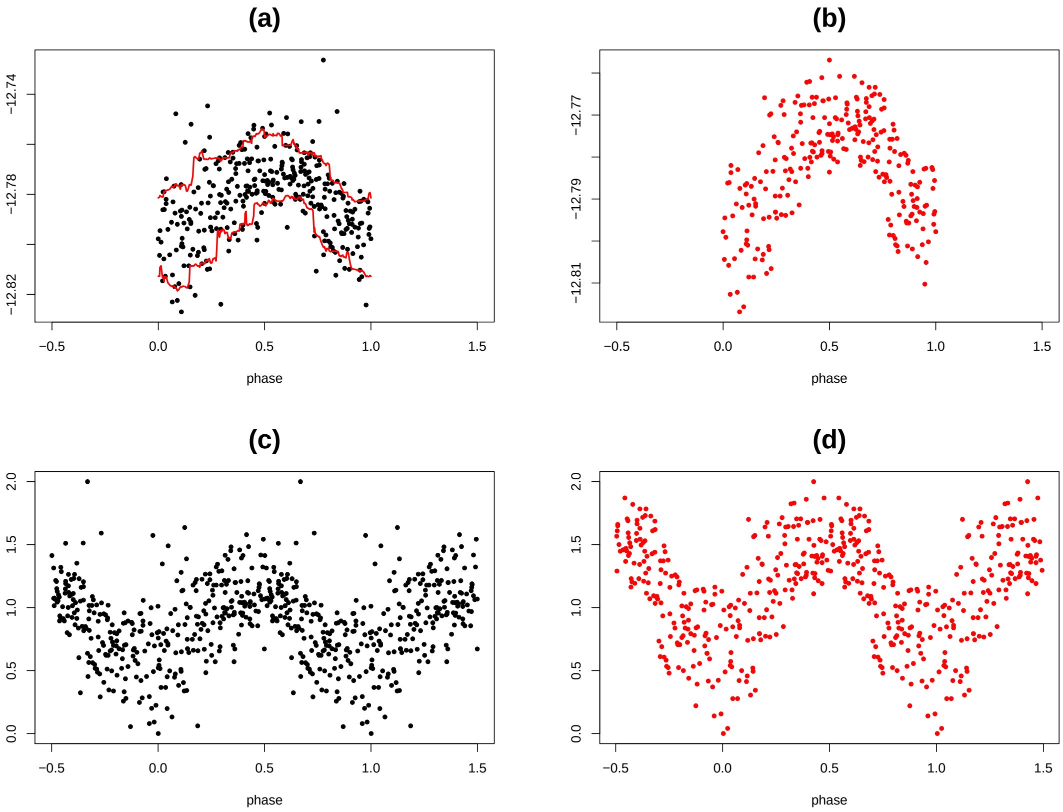

Figure 2: *An example of preprocessing for obtaining light curves in the group of BCEP*

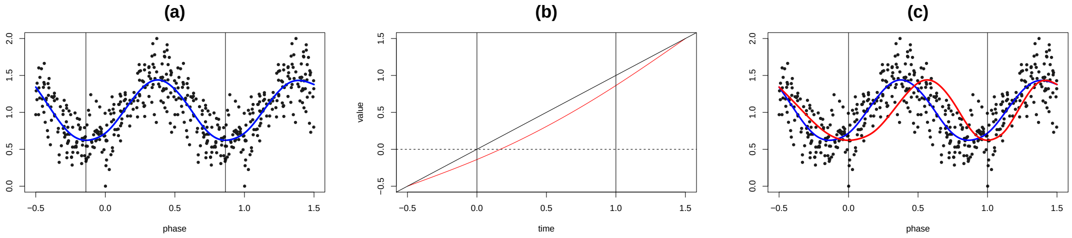

Figure 3: *An example of functional registration*

own specific goals and enormous amount of brightness time series for variable stars can be produced subordinately. The STARE project provides complete survey of variable stars. Since the classification results can be verified, we apply our proposed method to the data in the STARE project.

The STARE project contains 209 relatively well identified stars and 57 suspected variables. For each time series of brightness, the estimated period is also suggested. The details of variable types can be found in the website of the project (http://www.hao.ucar.edu/research/stare/stare.html). Variable stars are noted with some types, which are EA, EB, EW, and E (indeterminate type) in eclipsing binaries; DCEP, BCEP, CEP (indeterminate type), DSCT, and SRV in pulsating stars; Ellip, El|Sp,

Table 2: Variable star classes in the dataset

Pulsating stars		Eclipsing binaries	
type	number	type	number
BCEP	4	EA	14
DSCT	12	EB	13
		EW	14

Table 3: Classification matrices of the three functional classification methods according to different input variables.

	OC \ PC	Functional PC scores			Fourier coefficients			Function values		
		Puls	EA	EBW	Puls	EA	EBW	Puls	EA	EBW
QDA	Puls	16	0	0	13	0	3			
	EA	0	10	4	0	4	10			
	EBW	0	3	24	0	5	22			
NP density	Puls	16	0	0				16	0	0
	EA	1	11	2				1	13	0
	EBW	2	3	22				1	1	25
FSVM	Puls	16	0	0	16	0	0	16	0	0
	EA	0	9	5	0	10	4	0	11	3
	EBW	0	2	25	1	1	25	0	2	25

Table 4: Number of misclassified variable stars

	Functional PC scores	Fourier coefficients	Function values
QDA	7	18	·
NP density	8	·	3
FSVM	7	6	5

and El&Sp in rotating variables; and I (indeterminate type) for poorly studied irregular ones. We leave 45 SRV type stars out of the analysis because these are obviously classified by their period. A total of 54 indeterminate types are removed for classification; in addition, we exclude 27 variable stars whose maximum gap between observations is over 0.1 on the phase domain $[0, 1]$. We believe that the limit of observation gap 0.1 is fairly generous choice since it is corresponding to 10% of the whole domain. We exclude the rotating stars in the analysis since the rotating variable stars seem to be strong attractors for all types of stars. When some types have a common overlapping pattern (or have small number of data in each group) classification methods are not applicable. Thus, for the verification purpose of the proposed procedure, EB and EW in eclipsing binaries are regarded as the same type denoted by EBW, and BCEP and DSCT are combined as the same type denoted by Puls. We conduct the proposed functional classification method with three distant classes of Puls, EA, and EBW in Table 2, and evaluate the practical performance.

To evaluate its performance, we take a cross-validated classification step. Specifically, (i) select a curve and treat it as a new one; (ii) determine the class by applying the aforementioned classification methods to the remaining curves; and (iii) obtain the predicted class and then compare it with the observed one. In this manner, all curves are classified. Table 3 shows the classification results of the three functional classification methods according to different input variables. The $(i, j)^{th}$ entry of the matrix denotes frequency of the i^{th} observed class classified to the j^{th} predicted class. Thus, the off-diagonal entry indicates the misclassified frequency. The numbers of incorrectly classified curves are summarized in Table 4 according to the functional classification methods. We observe that Bayesian classifier using nonparametric density estimation with fitted function values as input variable produces the smallest number of misclassification.

5. Concluding Remarks

In this paper, we have proposed classification procedures of variable stars based on functional data analysis with the estimation of light curve. For the FDA-based approach, we have considered three classification methods with three different input variables of function features representing light cure shape. From the analysis with variable stars in STARE database, we may be able to draw the conclusions that the proposed method is capable of efficiently classifying stars, and provide an automatic procedure that can handle massive database.

Appendix: Functional Classification Method

We first consider a Bayesian classification method. Let Y be a categorical response valued in $\bar{G} = \{1, \ldots, G\}$, X be a functional random variable, and x denote a realization of X. In the viewpoint of regression, the classification procedure for a signal x to be in a group g can be described as follows: (1) Suppose that there exist density functions f_g for the different G groups with prior π_g, $g = 1, \ldots, G$. Then we calculate the posterior probability that a signal x belongs to the g^{th} class given as

$$p(g|x) = \frac{f_g(x)\pi_g}{\sum_{k=1}^{G} f_k(x)\pi_k}.$$

(2) We classify the signal x to the group g if $f_g(x)\pi_g > f_h(x)\pi_h$ for all $h \neq g$, that is, the quantity $f_g(x)\pi_g$ is the largest.

There are several approaches to determine the density $f_g(x)$ for functional data. First, for X valued in a normed vector space $(S, \|\cdot\|)$, x can be expressed as $x = \sum_{j=1}^{\infty} \xi_j \psi_j$ with a complete orthonormal basis $\{\psi_1, \psi_2, \ldots\}$, and the coefficients $(\xi_1, \xi_2, \ldots)$ serve as surrogates of x for purposes of density estimation and classification (Hall *et al.*, 2001). In order to work in m-dimensional Euclidean space effectively, the expansion $x = \sum_{j=1}^{\infty} \xi_j \psi_j$ should be truncated as $x = \sum_{j=1}^{m} \xi_j \psi_j$, and the choice of an m-dimensional subspace should capture the greatest part of X. Hall *et al.* (2001) showed that Karhunen-Loéve expansion of X achieves this goal and that the basis derived by principal components analysis in FDA is the empirical counterpart of Karhunen-Loéve expansion. Hence, we can use the functional PC scores/coefficients for purposes of density estimation and classification. Finally, a kernel estimator of the density of $X^{(m)}$ at $x^{(m)} = (\xi_1, \ldots, \xi_m)$ is given by

$$\hat{f}_m\left(x^{(m)}\right) = \frac{1}{n}\sum_{i=1}^{n} K\left(h^{-1}\left\|x^{(m)} - X_i^{(m)}\right\|\right), \tag{A.1}$$

where $X^{(m)}$ denotes an m-dimensional vector of coefficients for X, h is a bandwidth, K is a kernel and each datum X_i, $i = 1, \ldots, n$ is independently and identically distributed (*i.i.d.*) as X. With the dimension reduction by functional principal components analysis (FPCA) and by applying the nonparametric density estimator in (A.1) to each group, curves can be classified. Instead of nonparametric density estimation, quadratic discriminant analysis (QDA) can be used for the computational reason as in Hall *et al.* (2001).

Second, let X take its values in the semi-metric vector space (S, d), then the nonparametric density estimator of group g can be given by

$$\hat{f}_{g,h}(x) = \frac{\sum_{i=1}^{n} 1_{[Y_i=g]} K\left(h^{-1} d(X_i, x)\right)}{\sum_{i=1}^{n} K\left(h^{-1} d(X_i, x)\right)}, \tag{A.2}$$

where $(X_1, Y_1), \ldots, (X_n, Y_n)$ are *i.i.d.* (Ferraty and Vieu, 2003). The dimension reduction and density estimator in Hall *et al.* (2001) can be considered as special cases of those in Ferraty and Vieu (2003). Two semi-metrics, which are based on FPCA and successive derivatives, are introduced with theoretical advances and computational issues in Ferraty and Vieu (2003). Through this semi-metric approach, the estimated function values or derivative values can be used as input for classification as well as functional PC scores.

Besides Bayesian classifier, we consider several other classification methods, such as k-NN classification or SVM. The functional usage with theoretical supports can be summarized as follows: Let χ be an infinite-dimensional, separable Hilbert space with inner product $\langle\,\cdot\,,\cdot\,\rangle$ and norm $\|\cdot\|$. When X takes values in χ, it may be expressed as $X = \sum_{j=1}^{\infty} \Xi_j \psi_j$, where $\{\psi_j\}_{j=1}^{\infty}$ is a complete, orthonormal system of χ and $\Xi_j = \langle X, \psi_j\rangle$. Since any infinite-dimensional, separable Hilbert space χ is isomorphic with $\ell_2 = \{(x_1, x_2, \ldots) : \sum_{j=1}^{\infty} x_j^2 < +\infty\}$, knowing X is equivalent to knowing $(\Xi_1, \Xi_2, \ldots)$. In Biau *et al.* (2005), hence, a classification procedure by coupling $X^{(m)} = (\Xi_1, \ldots, \Xi_m)$ with k-NN or QDA is suggested with discussing the selection of the dimension m and the consistency of estimator. They used Fourier basis as ψ for data analysis. For simplicity, we consider only the classification based on QDA in this paper.

In the same Hilbert space, the functional SVM is introduced in Rossi and Villa (2006) with theoretical discussion of consistency. We consider an orthonormal basis $\{\psi_j\}_{j=1,\ldots,d}$ and the d-dimensional subspace V_d of χ. The transformation P_{V_d} is defined as the orthogonal projection on V_d, that is, $P_{V_d}(x) = \sum_{j=1}^{d} \langle x, \psi_j\rangle \psi_j$. Since $(V_d, \langle\,\cdot\,,\cdot\,\rangle_\chi)$ is isomorphic to $(R^d, \langle\,\cdot\,,\cdot\,\rangle_{R^d})$, a standard R^d SVM can be used on the vector of coefficients $(\langle x, \psi_1\rangle, \ldots, \langle x, \psi_d\rangle)$. Hence, the functional principal components, the orthogonal Fourier or wavelet basis can be used for SVM. However, as in nonlinear SVM for multivariate data analysis, the original data can be transformed from the given Hilbert space χ into another Hilbert space H, where a linear SVM can be constructed, using a feature map ϕ. The kernel in the process of SVM can be represented as $K(x_i, x_j) = \langle \phi(x_i), \phi(x_j)\rangle_H$ for a given pair (H, ϕ), but we do not need to know (H, ϕ) directly and it is enough to only provide the kernel function K, which is well known in SVM theory. Hence, as various Kernel can be used with the orthogonal Fourier or wavelet basis, the common B-spline basis can also be used in this approach. Practically, the evaluated function or derivative values obtained by a set of common B-spline basis can be input in functional SVM (Rossi and Villa, 2006).

References

Akerlof, C., Alcock, C., Allsman, R., Axelrod, T., Bennett, D. P., Cook, K. H., Freeman, K., Griest, K., Marshall, S., Park, H.-S., Perlmutter, S., Peterson, B., Quinn, P., Reimann, J., Rodgers, A., Stubbs, C. W. and Sutherland, W. (1994). Application of cubic splines to the spectral analysis of unequally spaced data, *The Astrophysical Journal*, **436**, 787–794.

Biau, G., Bunea, F. and Wegkamp, M. H. (2005). Functional classification in Hilbert spaces, *IEEE Transactions on Information Theory*, **51**, 2163–2172.

Blomme, J., Debosscher, J., De Ridder, J., Aerts, C., Gilliand, R. L., Christensen-Dalsgaard, J., Kjeldsen, H., Brown, T. M., Borucki, W. J., Koch, D., Jenkins, J. M., Kurtz, D. W., Stello, D., Stevens, I. R. and Suran, M. D. (2010). Automated classification of variable stars in the asteroseismology program of the *Kepler space mission*, *The Astrophysical Journal Letters*, **713**, L204–L207.

Clarkson, D., Fraley, C., Gu, C. C. and Ramsay, J. O. (2005). *S+ Functional Data Analysis: User's Manual for Windows*, Springer.

Deb, S. and Singh, H. P. (2009). Light curve analysis of variable stars using Fourier decomposition and principal component analysis, *Astronomy & Astrophysics*, **507**, 1729–1737.

Debosscher, J., Sarro, L. M., Aerts, C., Cuypers, J., Vandenbussche, B., Garrido, R. and Solano, E. (2008). Automated supervised classification of variable stars I. Methodology, *Astronomy & Astrophysics Manuscript No.* 7638.

Deeming, T. J. (1975). Fourier analysis with unequally-spaced data, *Astrophysical and Space Science*, **36**, 137–158.

Dwortesky, M. M. (1983). A period-finding method for sparse randomly spaced observations or "How long is a piece of string ?", *Monthly Notices of the Royal Astronomical Society*, **203**, 917–924.

Ferraty, F. and Vieu, P. (2003). Curves discrimination: A nonparametric functional approach, *Computational Statistics and Data Analysis*, **44**, 161–173.

Friedman, J. H. (1984). A variable span smoother, *Technical report* No. 5. Laboratory for Computational Statistics, Department of Statistics, Stanford University.

Hall, P., Reimann, J. and Rice, J. (2000). Nonparametric estimation of a periodic function, *Biometrika*, **87**, 545–557.

Hall, P., Poskitt, D. S. and Presnell, B. (2001). A functional data-analytic approach to signal discrimination, *Technometircs*, **43**, 1–9.

Hegland, M., Clarke, W. and Kahn, M. (2001). Mining the MACHO dataset, *Computer Physics Communications*, **142**, 22–28.

Lafler, J. and Kinman, T. D. (1965). An RR Lyrae survey with the Lick 20-inch astrograph II. The calculation of RR Lyrae periods by electronic computer, *Astrophysical Journal Supplement Series*, **11**, 216–222.

Lomb, N. R. (1976). Least-squares frequency analysis of unequally spaced data, *Astrophysical and Space Science*, **39**, 447–462.

Oh, H.-S., Nychka, D., Brown, T. and Charbonneau, P. (2004). Period analysis of variable stars by robust smoothing, *Journal of the Royal Statistical Society Series C*, **53**, 15–30.

Ramsay, J. O. and Silverman, B. W. (2002). *Applied Functional Data Analysis*, Springer, New York.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, 2nd ed, Springer, New York.

Ramsay, J. O., Hooker, G. and Graves, S. (2009). *Functional Data Analysis with R and MATLAB*, Springer, Dordrecht.

Reimann, J. D. (1994). *Frequency Estimation Using Unequally-Spaced Astronomical Data*, Ph.D. thesis, Department of Statistics, University of California at Berkeley.

Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification, *Neurocomputing*, **69**, 730–742.

Sarro, L., Debosscher, J. M., López, M. and Aerts, C. (2009). Automated supervised classification of variable stars II. Application to the OGLE database, *Astronomy & Astrophysics*, **494**, 739–768.

Scargle, J. D. (1982). Studies in astronomical time series analysis II. Statistical aspects of spectral analysis of unevenly spaced data, *Astrophysical Journal*, **263**, 835–853.

Stellingwerf, R. F. (1978). Period determination using phase dispersion minimization, *Astro-physical Journal*, **224**, 953–960.

Usatov, M. and Nosulchik, A. (2008). The extended catalog of red AGB variable stars found in the NSVS database, *Open European Journal of Variable Stars*.

Willemsen, P. G. and Eyer, L. (2005). A study of supervised classification of Hipparcos variable stars using PCA and support vector machines, *Manuscript*.

Woźniak, P. R., Williams, S. J., Vestrand, W. T. and Gupta, V. (2004). Identifying red variables in the northern sky variability survey, *The Astronomical Journal*, **128**, 2965–2976.

Received February 27, 2013; Revised July 9, 2013; Accepted July 12, 2013