Thesis Proposal

Beatriz Etchegaray García

Committee: Chad Schafer,
Peter Freeman,
Chris Genovese,
Ann Lee,
Jeffrey Newman

Department of Statistics
Carnegie Mellon University

[Draft September 8, 2011]

Abstract

Upcoming astronomical surveys will require real-time classification of sources with varying brightness. Correctly cataloguing distinct variable sources is useful for fitting models that accurately describe the underlying physics of the universe. Real-time classification is also important for resource allocation of telescope follow-up. Manual classification using spectroscopy has been the preferred method in astronomy, but the large amounts of non-spectroscopic data of future surveys makes classification of variable sources difficult. For each candidate source, we consider brightness measurements at irregularly-spaced time points (i.e., a light curve). Classification of sources based on these light curves alone is challenging because the variability in the time series is related in complex ways to the underlying physical processes that generate the data. Some attempts have been made to aid the classification of variable sources using machine learning methods, but it seems that realistic classification of these sources will require one to incorporate current understanding of the physics behind the processes that generated the observable data.

In this thesis we propose a classification scheme that will combine (1) the physical knowledge of the relationship between the type of object and spectroscopic information - in a training set - with (2) the reality of the low quality time series that we will observe. The underlying idea is that the spectra can accurately predict the type of source, thus we can hope to learn a complex structure between the light curve and features derived from the spectra. The light curves are then used to predict the value for the spectral feature, which is then used for classification. In this way, we have, in essence, performed a *transformation* of the class label. The hope is that it will be easier to ascertain a relationship between the real-valued spectral feature and the observed light curves using standard regression techniques. Also, errors in the estimate of the spectral feature could be quantified more naturally, and transforming these into uncertainties in the class label will be straightforward.

Contents

1 Introduction

Time series classification is built upon supervised learning methods to label temporally varying sequences. Consider the situation where we observe a sample of time series from each of K different classes, where the class of each observed time series is known. The goal of classification is to use the sample information to construct a rule to classify any future time series for which the class is unknown. Such a partition of the data space would have to maximize both the within-group similarity and between-group differences. However, in cases of interest, the complex nature of the data generation process, as well as excess variability and sampling effects, make the time-domain information insufficient for a clear separation when analyzed alone. Therefore, classification accuracy when using only time series data is limited.

In this thesis we propose a classification scheme for time series that will exploit data by combining (1) the knowledge of the relationship between the class type and ancillary data - in a training set - with (2) the reality of the low quality and noisy time series that we observe. The underlying idea is that the ancillary data can accurately predict the class, and thus we can hope to learn a complex structure between the time series and features derived from the ancillary data. The observed data are then used to predict the value for the ancillary data, which is then used for classification. In this way, we have, in essence, performed a *transformation* of the class label. In applications of interest, the class label is a categorical variable, while the ancillary data will be a continuous quantity. The hope is that it will be easier to ascertain a relationship between the continuous quantity and the observable data using standard regression techniques. Also, errors in the estimate of the ancillary feature could be quantified more naturally, and transforming these into uncertainties in the class label will be straightforward.

Very large quantities of data generated by modern astronomical instruments are collected and stored in databases. There is an increased need for efficient and effective automated analysis methods to process the data. There is interest in the classification of sources that have varying brightness as a function of time. The issue here is to label new candidate variable sources that are found based on the observations of brightness measurements at irregularly-spaced time points. In most cases a spectrum is the sure way to do this, but obtaining spectra is expensive, especially given the large number of transient candidates and that the time series data is readily available. Classification of variable sources based on the light curves alone is challenging because the variability in the time series is related in complex ways to the underlying physical processes that generate the data. Some attempts have been made to aid the classification of variable sources using machine learning methods, but it seems that realistic classification of these sources will require one to incorporate current understanding of the physics behind the processes that generated the observable data.

2 Background and Related Work

It is frequently of interest to study a set of objects in order to identify the underlying structure of classes. This task is known as classification, and it consists in finding some properties that characterize the differences among classes. As the observable elements of each class are typically similar, some kind of distance, based on the characterizing properties, is necessary to evaluate proximity. Proximity is evaluated between two distinct elements as well as an element and a group. Finally, a criterion is applied to determine the underlying structure.

When there is no previous knowledge about the structure, the problem is named unsupervised classification. If the structure of classes is indicated by provided labels, the problem is named supervised classification. The classification problem consists of applying a criterion to decide to which class a new element belongs to.

Classification of time series is a statistical subject with many applications. Time series can be studied from both time and frequency domains.

2.1 Time Series Analysis

Data obtained from observations collected sequentially over time are extremely common. A discrete-time stochastic process is a sequence of random variables (defined in the same probability space and taking values in the same state space, S),

$$\{X(t), t \in \mathbb{Z}\}$$

where the index t represents time. A stochastic process serves as a model for an observed time series, i.e., a time series can be interpreted as a realization of a stochastic process. Note that the variables are not simultaneous, as those in a multivariate vector. Stochastic processes are objects defined to study dynamic univariate changes in time.

Stochastic processes can be studied from the time domain or, alternatively or additionally, from the frequency domain. Time domain uses time as an index, and the autocovariance function is the natural tool for studying the evolution in the time domain. Fourier analysis allows the use of frequency as variable. Frequency domain analysis is the adaptation of Fourier analysis to deal with stochastic functions in time. In order to handle stochastic processes, sometimes additional structure is assumed under the name of stationarity. Stationarity implies homogeneity in the time domain, such that the autocovariance function is invariant under time shifts.

We will use the following notation: $\{x(t)\}$ is a stochastic process that is measured only at a set of discrete times t_i, $i = 1, \ldots, n$, yielding the time series data

$$x_i \equiv x(t_i), \quad i = 1, \ldots, n.$$

In general we will denote the time series data as $\{(t_i, x_i), i = 1, \ldots, n\}$.

2.1.1 Frequency Domain Analysis

Frequency domain or spectral analysis focuses on how signals of different frequencies are represented in a time series. The frequency spectrum of a time domain signal is a representation of that signal in the frequency domain. Frequency spectra, generated via Fourier transforms, are an important tool for the analysis of periodic variations. The main resulting value is the power spectral density (PSD), which describes how the power of a time series is distributed with frequency. Details on frequency domain analysis of time series can be found in the book by Bloomfield (2000).

Some of the problems encountered while doing spectral analysis are spectral leakage and aliasing. Spectral leakage is that for a sinusoidal signal at a given frequency f_0, the power in the PSD not only appears at f_0, but also leaks to other frequencies. Aliasing is leakage of power from high frequencies to much lower ones and it occurs if the sampling rate is not high enough to sample a signal correctly. But often we encounter time series observations that have uneven temporal sampling and/or non-uniform coverage (i.e., large gaps) (see e.g., Figure 1). This complicates the search for periodic signals, as a fast Fourier transform (FFT) algorithm cannot be employed. There are some ways to obtain evenly spaced data from irregularly spaced data (e.g. by interpolation), but these techniques introduce uncertainties and often perform poorly.

A method of spectral analysis for irregularly sampled data was developed by Lomb (1976) based on earlier work by Barning (1963) and further refined by Scargle (1982). The Lomb-Scargle (LS) periodogram is a widely used tool in period searches and frequency analysis of time series in astronomy that can handle irregularly spaced data. Like other classical methods of spectral analysis, the Lomb-Scargle methodology is based on the assumption that the analyzed time series is stationary. It is equivalent to fitting sine waves to the observed time series of the form

$$x(t) = a\cos(2\pi f t) + b\sin(2\pi f t),$$

where a and b are amplitudes, f is frequency, and t is time. The LS periodogram has the same statistical properties of classical power spectra. Zechmeister and Kurster (2009) generalize the LS periodogram to allow for an offset term to account for statistical fluctuations in the mean of the time series, and also weights to take in to account measurement errors, for example.

2.1.2 Structure Function Analysis

The goal of time domain analysis is to investigate the nature of the variability of a time series. Structure function (SF) analysis provides a method for quantifying time variability without the problems of leakage, aliasing, etc. that are encountered by Fourier analysis. It can provide information on the nature of the process that causes variation. Historically, the SF has been used to study turbulent plasmas (Kolmogorov, 1941), and it is a standard tool in geostatistics to investigate spatial correlations (Chilès and Delfiner, 1999).

The first order SF is the variance of differences between values of a random process $\{x(t)\}$ that are separated by a given time interval $\tau \geq 0$,

$$S_x(\tau) = Var\left(x(t+\tau) - x(t)\right).$$

The SF is commonly characterized in terms of its (log) slope β, where $S_x(\tau) \propto \tau^{\beta}$. If the mean function of $\{x(t)\}$, $\mu_x(t) = E(x(t))$ is independent of t then

$$S_x(\tau) = E\left(x(t+\tau) - x(t)\right)^2, \quad \tau \geq 0.$$

For a stationary random process, the SF is related to the variance, $Var(x(t)) = \sigma_x^2$, $\forall t$, and the autocorrelation function, $\rho_x(\tau)$,

$$\rho_x(\tau) = \frac{Cov\left(x(t+\tau), x(t)\right)}{\sigma_x^2}, \quad \tau \geq 0,$$

by

$$S_x(\tau) = 2\sigma_x^2(1 - \rho_x(\tau)).$$

The sample SF in the case of evenly sampled discrete time series $\{(t_i = i \times \Delta t, x_i), i = 1, \ldots, n\}$ is

$$\widehat{S}_x(k) = \frac{1}{n-k}\sum_{i=1}^{n-k}\left(x_{i+k} - x_i\right)^2, \quad k = 1, \ldots, n-1.$$

To approximate this relationship in the case where the time series is given by $\{(t_i, x_i), i = 1, \ldots, n\}$ with arbitrary t_i, we estimate the SF in a bin of width δ for a lag τ by

$$\widehat{S}_x(\tau, \delta) = \frac{1}{N(\tau, \delta)} \sum_{\{(i,j):|t_j - t_i| \leq \tau + \delta/2\}} \left(x_j - x_i\right)^2,$$

where $N(\tau, \delta) = \#\{(i,j) : |t_j - t_i| \leq \tau + \delta/2\}$. It is important to note that the SF does not describe the range of variability behavior in a time series, because it averages over variability amplitudes. If two time series have similar SF it is not clear if the distributions of variability amplitudes are truly similar, or if they only have similar average values.

If one considers a signal plus noise model for the data, $x(t) = \eta(t) + \epsilon(t)$, where the noise process $\{\epsilon(t)\}$ has mean 0 and is uncorrelated with both the signal process $\{\eta(t)\}$ and itself, one obtains

$$\begin{aligned} S_x(\tau) &= S_\eta(\tau) + S_\epsilon(\tau) = S_\eta(\tau) + Var(\epsilon(t) - \epsilon(t+\tau)), \\ \Rightarrow \quad S_\eta(\tau) &= S_x(\tau) - Var(\epsilon(t) - \epsilon(t+\tau)). \end{aligned}$$

In order for the structure function to measure only intrinsic variation in the magnitude we correct by subtracting the variability of the differenced noise process.

Model-based fitting of the SF has been used to explore variability of both ensembles and individual light curves of quasars (see, e.g., Butler and Bloom (2011)). Quasars are AGNs that exhibit correlated variability on long time scales. Kelly et al. (2009) propose a parametrization of the SF of an ensemble of quasars using the damped random walk model. In particular, a damped random walk process can be described by an exponential covariance

$$C_{ij} = \sigma^2 \exp\left\{-\frac{|t_i - t_j|}{\tau}\right\},$$

between measurements at times t_i and t_j. However, this is a purely statistical model, not a physical one, and in particular it is a good fit for flux variability of quasars. For other types of variable sources with light curves that are stochastic in nature, this model may not be a good fit.

2.2 Functional Data Analysis

Functional data analysis (FDA) refers to the statistical analysis of data in the format of curves or functions. More specifically, realizations of underlying random functions with noise. The aims of functional data analysis are the same as those of other statistical analysis: to develop ways of presenting the data that highlight interesting and important features; to investigate variability as well as mean characteristics; to build models for the data observed, including those that allow for dependence of one observation or variable on another; etc. Ramsay and Silverman (2005) is an introductory book and Ferraty and Vieu (2006) is a summary on contributions to nonparametric estimation with functional data. Because FDA treats an entire function as the unit of observation, as opposed to traditional multivariate analysis, it provides a solution for analyzing high dimension and low sample size data without dimension reduction as data pre-processing.

One application of FDA consists of regression models through which we can describe the relation between a real-valued outcome and explanatory functional variables. Functional regression models have been used extensively: for example, to predict the total annual precipitation in a sample of Canadian weather stations from the temperature curves measured during the year (see Ramsay and Silverman (2005)) and to estimate the fat content in meat samples from spectrometric measurements (see Ferraty and Vieu (2006)).

The classical linear regression model is often of the form

$$y_i = \beta_0 + \sum_{j=1}^{p} z_{ij}\beta_j + \varepsilon_i, \quad i = 1, \ldots, N,$$

where $(y_1, \ldots, y_N)$ is the vector of responses and $(z_{i1}, \ldots, z_{ip})$ is the vector of covariates for the $i-$th response. The error terms ε_i's are usually considered to be independent and identically distributed. A functional regression model would replace at least one of the p observed scalar covariates by a functional covariate. Next, we will describe a model consisting of a single functional independent variable plus an intercept term.

We want to replace the vector of covariate observations $\boldsymbol{z}_i = (z_{i1}, \ldots, z_{ip})$, $i = 1, \ldots, N$, by a function $z_i(t)$. A first approach may be to discretize the functional covariate by choosing a set of times $t_1, \ldots, t_q$ and consider fitting the model

$$y_i = \beta_0 + \sum_{j=1}^{q} z_i(t_j)\beta_j + \varepsilon_i, \quad i = 1, \ldots, N.$$

But, how do we chose the t_j's with the additional restriction that $q < N$? Choosing a finer and finer grid of times $\{t_j, j = 1, \ldots, q\}$, the summation approaches an integral equation

$$y_i = \beta_0 + \int z_i(t)\beta(t)dt + \varepsilon_i, \quad i = 1, \ldots, N.$$

This is the standard functional linear regression (FLR) model, which relates functional predictors to a scalar response, where $\beta(t)$ is the coefficient function. With a finite number N of observations it is impossible to determine the infinite dimensional $\beta(t)$. Many approaches have been proposed to deal with this underdetermination issue.

2.3 Decision Tree Learning

In the supervised classification problem, it is known that each case in a sample belongs to one of a finite number of possible classes. Given a set of features for N observations, we want to accurately predict to which class each observation belongs to. A classifier is a rule that assigns a predicted class based on a set of features. Consider the data $(\boldsymbol{X}_1, Y_1), \ldots, (\boldsymbol{X}_N, Y_N)$ where $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip}) \in \mathcal{X}$ are the features, $\mathcal{X}$ is the feature space and $Y_i \in \mathcal{Y}$ are the class labels. Assume K classes so that $\mathcal{Y} = \{1, \ldots, K\}$. Formally, a classifier is a function $h : \mathcal{X} \to \mathcal{Y}$, and it corresponds to a partition of $\mathcal{X}$ into disjoint sets $A_1, \ldots, A_K$ such that the predicted class of an observation with features $\boldsymbol{X}$ is k if $\boldsymbol{X} \in A_k$. The goal of classification is to find a rule that makes accurate predictions. Accuracy is in general defined in terms of the true error rate,

$$L(h) = P\left(h(\boldsymbol{X}) \neq Y\right),$$

and the empirical error rate

$$\widehat{L}(h) = \frac{1}{N} \sum_{i=1}^{N} I(h(\boldsymbol{X}_i) \neq Y_i).$$

Classification trees are simple nonparametric classifiers. Classification and regression trees (Breiman et al., 1984) work by recursively partitioning the feature space, $\mathcal{X}$, into disjoint rectangular regions, $R_1, \ldots, R_M$. The node m of a tree corresponds to the region R_m of the feature space with N_m observations. The proportion of class k observations in node m is

$$\widehat{p}_{mk} = \frac{1}{N_m} \sum_{\boldsymbol{X}_i \in R_m} I(Y_i = k).$$

At each stage of the recursive partitioning, say at node m, all the possible ways of splitting $\mathcal{X}$ into subsets are considered. The subset that results in the least node impurity is chosen. Typical measures of node impurity are the Gini index $\sum_{k \neq k'} \widehat{p}_{mk} \widehat{p}_{mk'}$ and entropy function $-\sum_{k=1}^{K} \widehat{p}_{mk} \log \widehat{p}_{mk}$.

A delicate issue in creating a tree classifier is how to determine the number of nodes the tree should have (i.e., the complexity). If nodes continue to be created until each node has 1 observation, the tree will be overfitting the training sample and will not be a good classifier for future (test) cases. On the other hand, if a tree has only a few number of end or terminal nodes, then it is not using enough information from the training sample and the classification accuracy for future cases will also be poor. Initially, in the tree growing process, the predictive accuracy will improve as more nodes are created, but at some point the misclassification rate will get worse as the tree becomes more complex. A standard approach is to choose the complexity so that the true error rate is low. To estimate the error rate a standard approach is to use $B-$fold cross-validation. The first step is to split the data into B subsets. For each subset $b = 1, \ldots, B$, the classifier $h_{(-b)}$ is computed on all the data not in subset b. The classifier is then used to predict the class for observations in subset b with an empirical error rate $\widehat{L}(h_{(-b)})$. Finally the error rates are averaged to obtain an approximately unbiased estimator of $L(h)$,

$$\widehat{L}(h) = \frac{1}{B} \sum_{b=1}^{B} \widehat{L}(h_{(-b)}).$$

2.3.1 Random Forests

Decision tree learning is a popular method for nonparametric classification and regression in statistics because it is an efficient method that is also robust to noisy data. A classification tree is especially attractive for easy understanding and intuitive representation. However, classification trees yield estimates with high variance: a small change in the data set can often result in a very different tree.

Ensemble learning methods use multiple classifiers and aggregate their results. Bootstrap aggregation or bagging (Breiman, 1996) is a well know ensemble learning method that reduces the variance associated with prediction. In bagging, the idea is to obtain many bootstrap samples from the data, apply a classification tree, and then combine the results by taking a simple majority vote for prediction. Hence, successive trees do not depend on earlier trees because each is independently constructed using a bootstrap sample of the data. Random forest (Breiman, 2001) is a modification of bagging that has the effect of reducing the correlation between different trees and hence improving averaging. In addition to constructing each tree using a different bootstrap sample of the data, random forests change how the tree is constructed. Each node is split using the best among a subset of predictors randomly chosen at that node. Random forests have only three parameters that are usually not very sensitive to their values: B, the number of bootstrap samples, the number of variables in the random subset at each node, and the number of trees in the forest.

2.3.2 Classifier Performance Evaluation

In practice one may encounter problems that have imbalanced data, i.e., at least one of the classes makes up only a small minority of the data. In these cases, the focus is to correctly classify the "rare" or minority class or classes. Commonly used classification procedures aim to minimize the overall error rate, rather than focusing on the minority class.

Two common approaches to the imbalanced data problem are: unbalanced loss functions (see e.g., **??**) and sampling techniques (see e.g., **??**). With an unbalanced loss function, one assigns a high loss or cost to the misclassification of the minority class and aims to minimize the weighted overall classification error. Sampling techniques aim at down-sampling the majority class or over-sampling the minority class, or both. The random forest grows each tree from a bootstrap sample of the training data. In the case of imbalanced data, there is a high probability that a bootstrap sample contains few or even none of the minority class, resulting in a tree with poor performance for predicting the minority class. One way to fix this is to use a stratified bootstrap, i.e., sample with replacement from within each class.

In the classification of imbalanced data, the overall classification accuracy is usually not an appropriate measure of performance. The trivial classifier that predicts every case as the majority class can still achieve very high. Other measures such as the the true negative rate (TNR), the true positive rate (TPR), etc (see e.g. XXX). Based on Table 1, these measures are defined as:

$$TNR = \frac{TN}{TN + FP},$$
$$TPR = \frac{TP}{TP + FN}.$$

For any classifier, there is a trade-off between the TPR and the TNR. When the accurate classification of the minority class is of interest, it is desirable to have a classifier that gives high prediction accuracy over the minority class (TPR), while maintaining reasonable accuracy for the majority class (TNR).

One way to compare the performance of classifiers is to use the ROC curve (see e.g., **??**). The ROC curve is a graphical representation of the trade-off between the false negative and false positive rates for every possible cut off. In the random forests, we can change the votes cutoff for final prediction: as we raise the cutoff for the minority class, we can achieve a lower TPR and a higher TNR, thus yielding a set of points on the ROC diagram.

3 Previous Work

Upcoming astronomical surveys will require real-time classification of variable sources. A variable sources is an astronomical object whose brightness varies as a function of time. Correctly cataloguing distinct variable sources is useful for fitting models that accurately describe the underlying physics of the universe. Real-time classification is also important for resource allocation of telescope follow-up. Manual classification using spectroscopy has been the preferred method in astronomy, but the large amounts of non-spectroscopic data of future surveys makes classification of variable sources difficult. Furthermore, additional information, such as the spectra, is not always available. For each candidate source, we consider brightness measurements at irregularly-spaced time points (i.e., a light curve). Our goal is to classify new variable stars in a fast and accurate way using only the light curve.

3.1 Classification of Variable Sources

Figure 1 are examples of light curves of two types of variable sources: blazars and cataclysmic variables (hereafter referred to as the BL and CV data set). These data were collected by the Catalina Sky Survey[1](A. Mahabal, private communication). Blazars (BLs) or BL Lac objects are energetic objects in the extragalactic universe. They are radio sources with highly variable optical and radio emission. There are two types of blazars: blazars with flat optical spectra (i.e., no distinguishable features), and blazars with emission-line-dominated optical spectra and flat radio spectra. They belong to a class of galaxies with active galactic nuclei (AGN), which is driven by infall of matter to a supermassive black hole. Variability of blazars occurs from shocks or relativistic jets (Valtaoja, 1992).

A cataclysmic variable (CV) is a binary star system where stellar mass is transferred from a highly evolved giant star (the secondary) to a white dwarf (the primary). However, the stellar material does not fall directly into the primary, instead an accretion disc is formed around the primary, becoming the brightest part of the system. Erratic flickering in photometric measurements of CV systems can be attributed to regular eclipses of the stars during their orbit or a unstable mass transfer rate, resulting in a varying brightness of the accretion disc (Hellier, 2001).

Light curves describe how the brightness of a variable star varies as a function of time. The variability depends on the physical processes inside the source. The BL and CV data set contains light curve information for 121 BLs and 437 CVs. Without any other additional information we want to see how well the classes can be described (and separated) using only this minimal amount of information.

Several time series analysis methods have been developed to study the variability properties of astronomical sources (see Subba Rao et al. (1997) for a review). Linear and non-linear analysis methods in the time and frequency domains, adjusted to the needs of astronomical data sets, i.e. taking into account measurement errors and data gaps, can well describe the time behavior of some astronomical sources. Other tools based on running variances as for example, the SF, have been used to study the underlying variability properties of the observed sources.

3.1.1 Frequency Domain Analysis

The Lomb-Scargle is a popular technique for analyzing the behavior of light curves of periodic and non-periodic variable sources. To analyze the photometric time series in the BL and CV data set we consider only the non-censored observations. After removing outlier measurements, a possible linear trend of the form $a + bt$ is also removed, where a is the intercept, b the slope, and t the time. We fit each light curve a

[1] *http://www.lpl.arizona.edu/css/index.html*

time series model that is a harmonic sum of sinusoids:

$$x(t) = c + \sum_{i=1}^{N_f} \sum_{j=1}^{N_h} \left(a_{ij} \sin\left(2\pi f_i j t\right) + b_{ij} \cos\left(2\pi f_i j t\right) \right),$$

where N_h is the number of harmonics for N_f test frequencies. The number of harmonics and test frequencies used to best fit a light curve are unknown. These values could be determined through modeling of the process (e.g., by forward selection and $F-$tests). A model with more parameters will always fit the data at least as well as a model with fewer parameters. The question is whether the model with additional harmonics and test frequencies will significantly fit the data better.

Following Debosscher et al. (2007) we choose $N_f = 3$ and $N_h = 4$. Each of 3 test frequencies f_i, are allowed to have 4 harmonics at frequencies jf_i, $j = 1, 2, 3, 4$.. The 3 test frequencies f_i are found iteratively, by successively finding the peaks of the generalized Lomb-Scargle periodogram in the following manner:

1. Define a search range for frequencies (f_0, f_N, and Δf): the starting frequency is taken as $f_0 = 1/T_{tot}$, where T_{tot} is the total time span of the observations in light curve; the frequency step is taken as $\Delta f = 1/T_{tot}$; and the highest frequency is taken as the average of the inverse time intervals between measurements, $f_N = 0.5 \times \overline{1/\Delta t}$.

2. LS periodogram is calculated and the highest peak is selected. The corresponding frequency value f_1 is then used to calculate a harmonic fit to the light curve, via least-squares:

$$x(t) = c + \sum_{j=1}^{4} \left(a_j \sin\left(2\pi f_1 j t\right) + b_j \cos\left(2\pi f_1 j t\right) \right).$$

3. This curve is subtracted from the time series, $x_i - \widehat{x(t_i)}$, $i = 1, \ldots, n$. A new LS periodogram is computed on the residuals. This procedure is repeated until three frequencies are found.

4. The three frequencies are used to make a harmonic best-fit to the original (detrended) time series,

$$x(t) = c + \sum_{i=1}^{3} \sum_{j=1}^{4} \left(a_{ij} \sin\left(2\pi f_i j t\right) + b_{ij} \cos\left(2\pi f_i j t\right) \right).$$

For purposes of classification, the Fourier coefficients obtained here are not unique, because they are not invariant under time translations. We translate the coefficients into amplitudes and a phases as follows:

$$A_{ij} = \sqrt{a_{ij}^2 + b_{ij}^2}, \qquad \phi_{ij} = \arctan\left(\frac{b_{ij}}{a_{ij}}\right).$$

Following Debosscher et al. (2007) we correct the phases ϕ_{ij} to relative phases with respect to the phase of the first component ($\phi_{11} = 0$). These parameters are a time-translation invariant description of the light curves and are suitable for classification purposes. A list of the periodic features is found in Table 2.

Plots of one-dimensional density estimates by class of selected period features are shown in Figure 2. By inspection of these distributions we can quickly see that there is almost no separability of the two variable classes. The more the densities are separated from each other, the better the classes are defined, and the fewer the misclassifications which will occur in the case of a supervised classification. Complementary to these one-dimensional plots, we have conducted a more detailed analysis of the statistical properties of the training set by calculating correlations between different features, and computing 2d (see Figure 3) nonparametric density estimates of features.

3.1.2 Time Domain Analysis

The first systematic description of the SF adjusted to the needs of astronomical data sets was made by Simonetti et al. (1985) to study radio source scintillation using as a reference the work of Rutman (1978) from the field of electrical and electronic engineering. It has since been used extensively used to study, for example, blazar variability (Hughes et al., 1992) and quasar selection (Butler and Bloom, 2011).

To show the range of the magnitude differences in the BL and CV data set, we first calculate for each pair of measurements (i, j), $i, j = 1, \ldots, n$ ($n(n-1)/2$ in total) a lag τ_{ij} and a one-point estimate of the SF, s_{ij},

$$\tau_{ij} = |t_i - t_j|, \quad s_{ij} = (x_i - x_j)^2.$$

From Figure 4 we can see that the distribution of the magnitude differences, $\sqrt{s_{ij}}$, are centered around 0 and have larger spread for CVs than for BLs. A one-point estimate of the corrected SF is h_{ij},

$$\tau_{ij} = |t_i - t_j|, \quad h_{ij} = (x_i - x_j)^2 - e_i^2 - e_j^2,$$

where e_i and e_j are the measurement errors for the magnitudes x_i and x_j, respectively. However, for 19.66% of the measurement pairs (22.5% for BLs and 18.76% for CVs), no variability is seen, i.e., $h_{ij} < 0$. In our following SF analysis we will not use the corrected version of the SF.

To estimate the SF we bin s_{ij} and average. The SF for the BL and CV ensembles are shown in Figure 5. Because of the particular sampling in the data, for some values of τ the SF will be completely unknown. On the the other hand, for some values of τ the number of pairs of observations per bin will be very large, and should lead to a good estimate of the SF. A clear feature can be observed in the SFs, there are no horizontal trends. When the variability of a time series is dominated by a white noise (WN) process, then the SF is constant, with a value equal to twice the variance of the WN. This is because the amplitude of a WN process is independent of the time lag between two observation.

Microvariability (variability on very short time scales) has been confirmed to be the intrinsic nature of AGNs, especially for blazars (see, e.g., Miller et al. (1989)). For longer time scales, there is a roughly linear increase (on a logarithmic scale) of the SFs with τ, with a slope of about $\beta = 0.40\ (0.37, 0.43)$ for BLs and $\beta = 0.062\ (0.041, 0.084)$ for CVs. However, these trends are valid for the ensemble of BLs and CVs, not for the individual SFs of sources (see Figure 6). For purposes of classification, the individual SFs or features of the individual SFs (e.g., slope of linear trends, see Figure 7) of light curves do not provide a clear separation between classes.

We suspect that the uneven sampling affects severely the SF estimates of individual light curves. To check this hypothesis we run simulations. There exists a functional relation between the SF and the power spectral density (PSD); when a time series whose PSD follows a power-law, the SF will also follows a power-law (under certain assumptions, see **??** for details). Via simulations, we obtained a mean SF for each light curve characterized by a power-law PSD with the estimated indices. By visual inspection, the mean SFs were clearly different to the estimated SFs.

3.1.3 Feature Extraction

The classification of light curves relies upon the ability to recognize and quantify the differences between the variability. In the previous two sections we have seen that working in the time and frequency domain does not yield additional information to separate the different classes. Following Richards et al. (2011) we take another approach by transforming each light curve into a set of features that ignore the time structure in the data. Table 3 contains the features computed. Many of these are simple statistics on the distribution of apparent magnitudes (e.g., standard deviation, skewness). Density estimates for the features by type of source are shown in Figure 8. Again, we can see that in this feature space there is no separation of the type of variable source.

If no useful time-domain information can be obtained from light curves, one can estimate the density of the time series to examine, for example, the shape of the distribution, the spread; etc. These density estimates may contain useful information for classification. Nonparametric density estimation is done using a Gaussian kernel estimator of which the asymptotic properties are well established for i.i.d. data and for time series data (see e.g., **??**). With a simple example, we will illustrate how to classify noisy time series with no useful time-domain information by incorporating ancillary information.

3.2 A Simple Example

To illustrate the main ideas of our proposed methods consider the following simple example. Consider a set of training examples of the form $\{(\boldsymbol{x}_i, y_i), i = 1, \ldots, N\}$, where the $\boldsymbol{x}$'s are input variables (evenly spaced time series data), $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$ and the y's are the categorical (class type) outputs. Assume that there are only two classes, and so $y \in \{0, 1\}$. The details of the simulation follow. To simulate a data set:

- generate the output Y such that $Y = \begin{cases} 0 & \text{with probability } 0.3 \\ 1 & \text{with probability } 0.7 \end{cases}$;
- generate σ such that $\sigma = \begin{cases} 1 & \text{if } Y = 0 \\ \text{Uniform}(0.5, 1) & \text{if } Y = 1 \end{cases}$;
- generate the input $\boldsymbol{x} = (x_1, \ldots, x_n) \overset{i.i.d.}{\sim} N(0, \sigma)$.

Because the simulated time series are i.i.d., the time series data have no useful time-domain information. The class labeled $Y = 1$ constitutes a small minority of the data. We will focus on the correct classification of this "rare" class.

3.2.1 One Data Set

We simulate one data set with 300 time series, each with 30 equally spaced values given the details above. 214 time series have label $Y = 0$ and 86 have label $Y = 1$. Figure 9 are some examples of the simulated time series. Using a random forest classifier on the time series data we obtain the classification in Table 4 with an overall error rate of 20.33% (500 trees are grown and 5 variables are randomly sampled as candidates at each split). All but 31 time series were classified in the larger class, $Y = 0$. Because 67% of the time series in class $Y = 1$ are misclassified and only 10% for class $Y = 0$, we have an imbalanced classification and we conclude that there are limitations in terms of classification accuracy when using only this data.

The random forest classifier is constructed to minimize the overall error rate and will tend to focus more on the prediction accuracy of the majority class. This results in poor accuracy of the minority class. If we down-sample the majority class, then we grow each tree on more balanced data. A majority vote is taken as usual for prediction. Figure 10 compares the performance of the random forest classifier using different sample sizes for the majority class. The size of the majority class is 214, but the class error rates are roughly equal when we use 55 bootstrap samples in the random forest. Using stratified bootstrap samples sizes of 55 and 86 for class $Y = 0$ and $Y = 1$, respectively, we obtain an overall error rate of 17.67%. The class errors are more balanced: 18.22% for class 0 and 16.28% for class 1, as shown in Table 5. At the expense of misclassifying more observations in the majority class we are able to reduce the error rate in the minority class, without with a small change in the overall error rate. By inspection of the ROC curves in Figure 11 we can see that classifier using stratified sample sizes to balance out the class error rates performs better.

In this particular example we derive the ancillary feature from the time series data. In a real situation, the ancillary feature is obtained independently from the data. In this example, the ancillary feature is the

sample standard deviation, $s = \sqrt{\sum_{j=1}^{n}(x_j - \overline{x})^2/(n-1)}$. From Figure 12 it is clear that the ancillary feature is correlated with the class type, and therefore contains valuable information to separate the classes.

If no useful time-domain information can be obtained from the time series, one estimates the densities of the time series to examine, for example, the shape of the distribution. Nonparametric density estimation is done using a Gaussian kernel estimator of which the asymptotic properties are well established for i.i.d. data and for time series data (see e.g., **??**). The individual density estimates of the simulated time series are shown in Figure 13. These densities give us a visual impression that there are two groups of time series with different variability. We propose to regress the ancillary data on the density estimates. This approach relies on the fact that the time series and its density estimates can be well separated by the ancillary feature. We consider a setting where a regression model is used to learn the complex structure between the ancillary data and the density of the noisy time series.

To clarify the notation, we have simulated one data set consisting of $N = 300$ evenly-spaced time series $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$ of length $n = 30$, $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{in})$. Furthermore, we have nonparametric density estimates for each of the time series, $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_N$ on a grid $x_1^*, \ldots, x_m^*$ with $m = 120$, such that $\boldsymbol{f}_i = (f_{i1}, \ldots, f_{im})$, where f_{ik} is the density estimate for x_k^*. Furthermore, for each time series we have an ancillary (scalar) feature, $a_1, \ldots, a_N$.

The goal now is to predict the ancillary feature from the density estimates of the time series data. We fit a functional linear model where the response is a scalar quantity. We want to predict this scalar response on the basis of one functional covariate, the density of the observations in a time series. For each time series, we consider a density estimate on a grid $\boldsymbol{f}_i = (f_{i1}, \ldots, f_{im})$, $i = 1, \ldots, N$, as a vector of discretized functional values $f_i(x_k^*) = f_{ik}$, $k = 1, \ldots, m$. The one-dimensional argument, x^* are the values in the domain of the time series. The density estimate is a function of x^* and it is observed only at discrete sampling values $x_1^*, \ldots, x_m^*$ with $m = 120$ that are equally spaced. We are considering N time series, and therefore, there are N replications of the function, indexed by $i = 1, \ldots, N$. Each replicate is referred to as an observation, since we want to treat the discrete values as a unitary whole.

We fit the following FLR model,

$$a_i = \beta_0 + \int f_i(x^*)\beta(x^*)dx^* + \varepsilon_i, \quad i = 1, \ldots, N.$$

To deal with the underdetermination issue of β we use a basis coefficient expansion of β:

$$\beta(x^*) = \sum_{k=1}^{K_\beta} b_k \phi_k(x^*) = \boldsymbol{b}'\boldsymbol{\phi}(x^*),$$

where $\{\phi_k, k = 1, \ldots, K_\beta\}$ are the basis functions. At the same time, the covariate function, $f_i(x^*)$, can also be expanded in terms of a basis expansion in $\{\theta_k, k = 1, \ldots, K_f\}$ as:

$$f_i(x^*) = \sum_{k=1}^{K_f} c_{ik}\theta_k(x^*) = \boldsymbol{c}_i'\boldsymbol{\theta}(x^*), \quad i = 1, \ldots, N.$$

Therefore, the FLR model can be expressed as

$$\begin{aligned} a_i &= \beta_0 + \int \boldsymbol{c}_i'\boldsymbol{\theta}(x^*)\boldsymbol{\phi}(x^*)'\boldsymbol{b}dx^* + \varepsilon_i \\ &= \beta_0 + \boldsymbol{c}_i'\left(\int \boldsymbol{\theta}(x^*)\boldsymbol{\phi}(x^*)'dx^*\right)\boldsymbol{b} + \varepsilon_i \\ &= \beta_0 + \boldsymbol{c}_i'\boldsymbol{J}_{\phi\theta}\boldsymbol{b} + \varepsilon_i. \end{aligned}$$

The parameters we want to estimate are β_0 and $\boldsymbol{b}$. To convert the density estimates to function form we choose a cubic B-spline basis with 58 (roughly $m/2$) equally spaced knots.

One approach to estimate the parameters is to truncate the basis such that $K_\beta < K_f$ and then use the least squares approach. We choose a cubic B-spline with 5 basis functions for the regression coefficient $\boldsymbol{b}$, and a constant function for β_0. Figure 14 is the estimated regression coefficient function. The squared multiple correlation is 0.9736 and the corresponding $F-$statistic with 5 and 29 degrees of freedom is 141.1767, suggesting a fit to the data that is better than what we would expect by chance. Figure 15 compares the fitted values for the ancillary data from a FLR model with the true values. With a random forest classifier on the predicted ancillary feature we obtain the classification in Table 6 with an overall error rate of 12.67% (500 trees are grown and 1 variable is sampled at each split). Again we have a case of unbalanced error rates: 9.35% of the observations in class 0 are misclassified while 10.93% of the observations in class1 are misclassified. Again we balance the class error rates by choosing bootstrap samples of sizes of 84 and 86 for class $Y = 0$ and $Y = 1$, respectively (see Table 7). The overall error rate for this classifier is 13%.

Another approach to estimate the parameter is to use penalized regression. We fit the FLR by minimizing the penalized sums of squares (PSSE):

$$PSSE_\lambda(\beta_0, \beta) = \sum_{i=1}^{n} \left[a_i - \beta_0 - \int f_i(x^*)\beta(x^*)dx^* \right]^2 - \lambda \int L\beta(x^*)dx^*,$$

where the second term on the right side penalizes some form of roughness in the coefficient function. We use the criterion

$$\int L\beta(x^*)dx^* = \int \left(\beta''(x^*)\right)^2 dx^*,$$

which measures the roughness of the function β by integrating the square of its second derivative, i.e., the curvature of β. The more wiggly β is, the larger this term will be. The smoothing parameter, λ, plays a key role. The larger λ, the more roughness in β is penalized. As $\lambda \to \infty$, β tends to a line, for which the second derivative is 0. On the other hand, for small λ the roughness of β matters less.

We replace our previous choice of basis for defining the β estimate by a cubic B-spline basis with 58 equally spaced knots. By cross-validation we find a smoothing parameter of $\log_{10}(\lambda) = -1.5$ (see Figure 17). The squared multiple correlation is again 0.9737. The $F-$statistic is 136.9076 with 4.8483 and 29 degrees of freedom. However, the $F-$distribution in this case is only an approximation. The predicted ancillary feature is almost identical to the least squares estimate. Therefore, the random forest classifiers are nearly identical.

3.2.2 Many Data Sets

Next, we repeat the above analysis by simulating 100 data sets, each with 300 time series, each of these time series with 30 equally spaced values given the details above. To classify the time series data and the predicted (via FLR) ancillary data in each data set we use a random forest classifier just as in the previous section. In order to balance the individual class errors we use the down-sampling method. We arbitrarily set the class 0 sample sizes to 55 and 84 when building classifiers for the time series and predicted ancillary data, respectively. From Figure 18 we can see that the overall classification accuracy of the classifiers that use the predicted ancillary feature is higher than the classifiers that use the raw time series data.

The times series in each data set is our observed data and used to predict the ancillary feature, which in turn, are used for classification. Via functional regression models we have learned the relationship between the ancillary feature and the observable data. The errors in the density estimates have propagated onto the predicted ancillary features and hence in the predicted class. The question about how to quantify these errors remains as future work.

4 Proposed Work

4.1 Regression With Distribution as Predictor

Regression is a widely studied problem in statistics. One of the key methodological contributions to be made by this dissertation will be an exploration of methods for regression in cases where the response is real-valued, but the predictor is a distribution, or an estimate of a distribution. Estimating and testing the parameters of a regression function has been well studied, however, when the predictors are measured with error, the problem becomes more complex from a statistical point of view. In particular, we will focus on how the effects of the errors in the density estimate propagate to errors in the regression. Because the predictor is estimated, this problem becomes a errors in variables problem.

In particular, we will explore existing parametric and nonparametric methods for relating a continuous response to functional predictors, i.e., the estimated density or log density. Often functions must satisfy some constraints. If, for example, the data are values that cannot be negative, then we do not want negative function values, even over regions where values are at or close to zero. Furthermore, if one considers a histogram as a density estimate, then the total area under the density is 1. The density in this particular case are non-negative proportions with unit sum, i.e., compositional data. We will investigate models with compositional covariates where the goal is to predict a real-valued response as a function of a composition. The comparisons of existing methods on simulated data sets will result in concrete recommendations for this type of regression problem.

Finally we will investigate how to assess the fit of regression models with distributions as predictors. Because the response is real-valued, standard methods of evaluation of looking a residuals should extend. However, an interesting question is approximating the number of degrees of freedom in the mode, i.e., how to assess when we may be overfitting.

Specifically, I will do the following:

To be included...

4.2 Further Exploration of Time-Domain Approaches

The SF is one of the most extensively used tools in the field of AGN variability. Conclusions are based on observed SF characteristics such as breaks and slopes and they are attributed physical meaning. Through extensive simulation, we want to study the properties of the SF. In particular we want to study the effects of the data length on the position of the SF breaks and analyze the sensitivity of SF to the presence of data gaps.

A major problem in the use of the SF is that the estimates $\widehat{S}(\tau_i)$ are not independent of each other. This affects common fitting routines, for example, least squares and maximum likelihood, methods that are commonly used in SF astronomy literature to derive the SF breaks and slopes. The estimation of the structure function via maximum likelihood usually assume a normal distribution for the observations in a light curve. The covariance structure is specified empirically or via a statistical model (e.g., a damped random walk model). These assumptions can be generalized by assuming other distributions for the light curves or other specifications of the covariance structure.

Specifically, I will do the following:

To be included...

4.3 Astronomy Application

We will develop a classification scheme to combine (1) the physical knowledge of the relationship between the type of object and spectroscopic information - in a training set - with (2) the reality of the low quality time series that we will observe. Because the spectra can accurately predict the type of source, we can learn

a complex structure between the light curve and features derived from the spectra. The learned structure is used as an input to perform classification of variable sources based on light curves in the test set. Main emphasis will be placed on finding relevant ancillary features in spectroscopic data that can separate types of variable sources.

In the classification problem we may have class labels that are derived from an underlying continuous variable (in this particular case, the ancillary feature). Hence we will explore methods where real-valued data is taken as input to perform classification. For example, it would be interesting to explore if transition classes exist between two classes. These transition classes can be objects that show properties in between two standard classifications. Furthermore, defining more and more transition classes could ultimately reveals that variable sources do not form a discrete class but rather a continuum of variability.

Specifically, I will do the following:

To be included...

5 Tables

Table 1: Confusion matrix.

		Prediction	
		Negative	Positive
Truth	Negative	True Negative (TN)	False Positive (FP)
	Positive	False Negative (FN)	True Positive (TP)

Table 2: Periodic features extracted from light curves using the generalized Lomb-Scargle periodogram.

Feature	Description
f_i	$i-$th frequency, $i = 1, 2, 3$
A_{ij}	Amplitude $i-$th frequency, $j-$th harmonic, $i = 1, 2, 3$, $j = 1, 2, 3, 4$
PH_{ij}	Phase $i-$th frequency, $j-$th harmonic, $i = 1, 2, 3$, $j = 1, 2, 3, 4$
varratio	Ratio of the variance after, to the variance before subtraction of the fit with f_1 and its 4 harmonics

Table 3: Other features calculated using the magnitude measurements of the light curves.

Feature	Description
w.mean	weighted (by photometric errors) mean of the mags
std.dev	standard deviation of the mags
skew	skewness of the mags
kurt	kurtosis of the mags
pct.beyond.1.std.dev	fraction of mags that lie above or below one std. dev. from the weighted mean
amplitude	difference between the max and the min mags
min.abs.slope	min absolute slope between two consecutive observations
med.abs.slope	median absolute slope between two consecutive observations
max.abs.slope	max absolute slope between two consecutive observations
min.abs.dev	min discrepancy of the mags from the median mag
med.abs.dev	median discrepancy of the mags from the median mag
max.abs.dev	max discrepancy of the mags from the median mag
within.20pct.ampl.from.med	Fraction of mags within 20% of the amplitude from the median mag
slope.trend.first.30	Considering the first 30 mags, the % of increasing first diffs minus the fraction of decreasing first diffs
slope.trend.last.30	Considering the last 30 mags, the % of increasing first diffs minus the fraction of decreasing first diffs
pct.ratio.mid.20	Ratio of mag percentiles (60th - 40th) over (95th - 5th)
pct.ratio.mid.50	Ratio of mag percentiles (75th - 25th) over (95th - 5th)
pct.ratio.mid.80	Ratio of mag percentiles (90th - 10th) over (95th - 5th)

Table 4: Confusion matrix of random forest classifier on the simulated time series data. 500 trees were grown and 5 variables were tried at each split.

		Prediction		Classification Error
		0	1	
Truth	0	211	3	0.01401869
	1	58	28	0.67441860

Table 5: Confusion matrix of random forest classifier on the simulated time series data, with bootstrap sample sizes of 55 and 86, for class 0 and 1, respectively. 500 trees were grown and 5 variables were tried at each split.

		Prediction		Classification Error
		0	1	
Truth	0	175	39	0.2149533
	1	14	72	0.1976744

Table 6: Confusion matrix of random forest classifier on the least squares estimate of the ancillary feature using FDA. 500 trees were grown and 1 variable was tried at each split.

		Prediction		Classification Error
		0	1	
Truth	0	194	20	0.09345794
	1	18	68	0.20930233

Table 7: Confusion matrix of random forest classifier on the least squares estimate of the ancillary feature using FDA, with bootstrap sample sizes of 84 and 86, for class 0 and 1, respectively. 500 trees were grown and 1 variable was tried at each split.

		Prediction		Classification Error
		0	1	
Truth	0	186	28	0.1308411
	1	11	75	0.1279070

6 Figures

Figure 1: Real light curves of four variable sources: cataclysmic variables (top) and blazars (bottom). The light curves have measurement errors associated to the brightness measurements. Some of the light curve data may contain (right) censored observations.

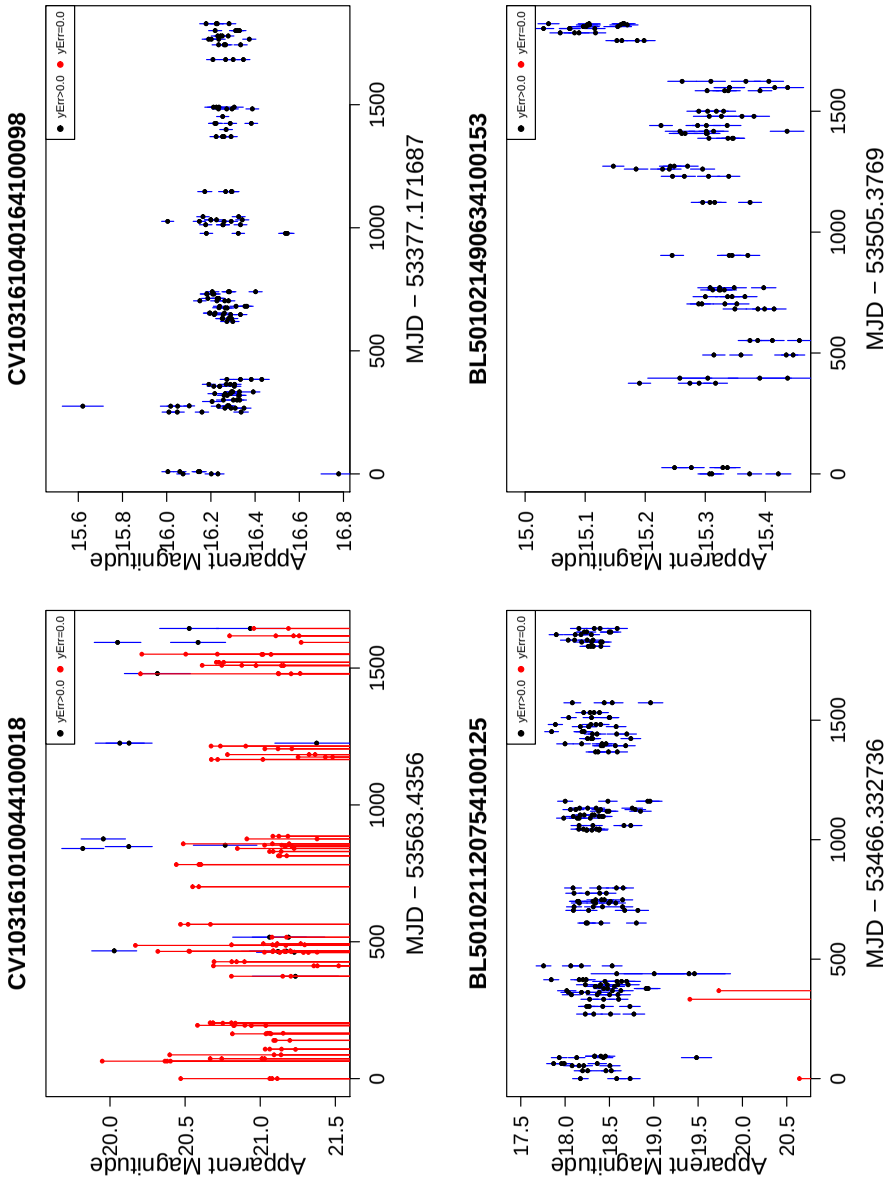


Figure 2: Nonparametric density estimates of selected periodic features of the BL and CV data set.

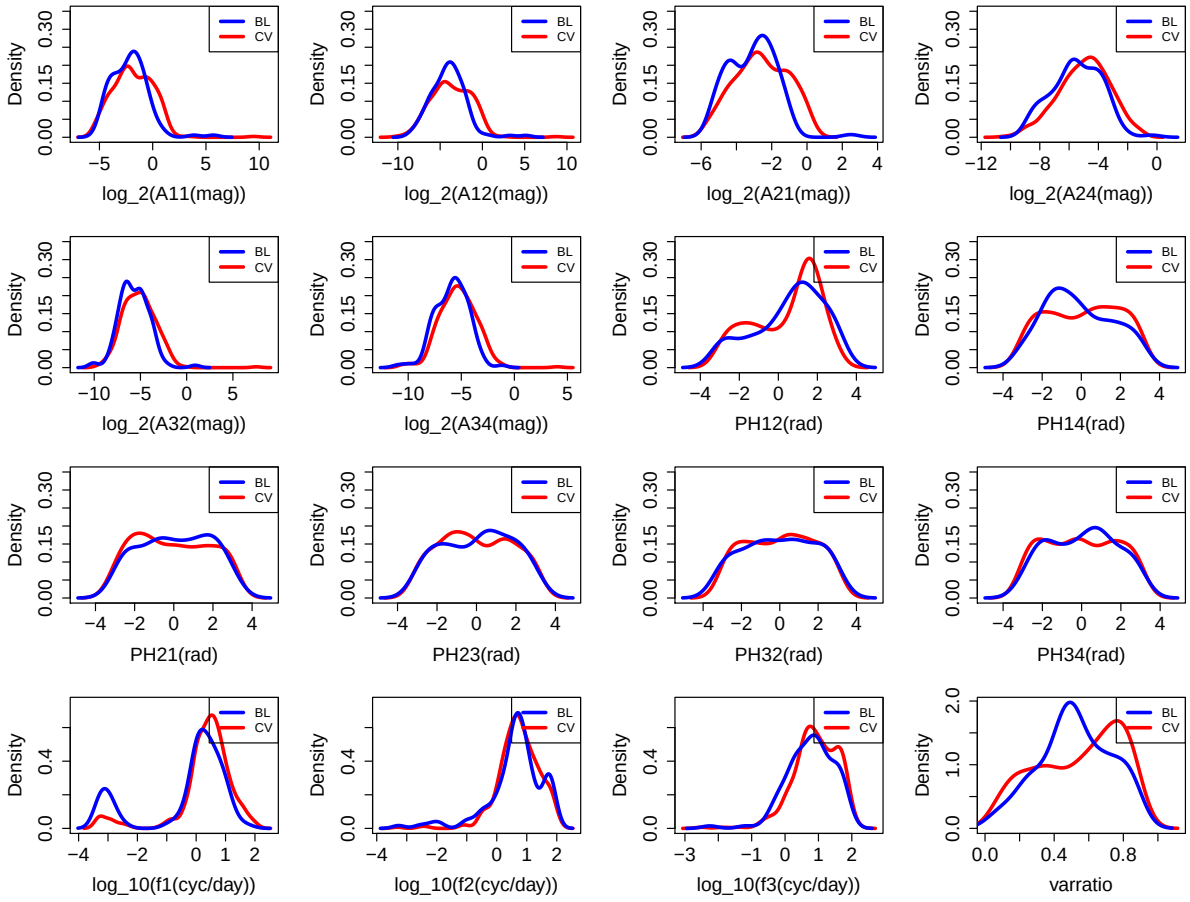


Figure 3: Bivariate nonparametric density estimates of selected pairs of periodic features of the BL and CV data set.

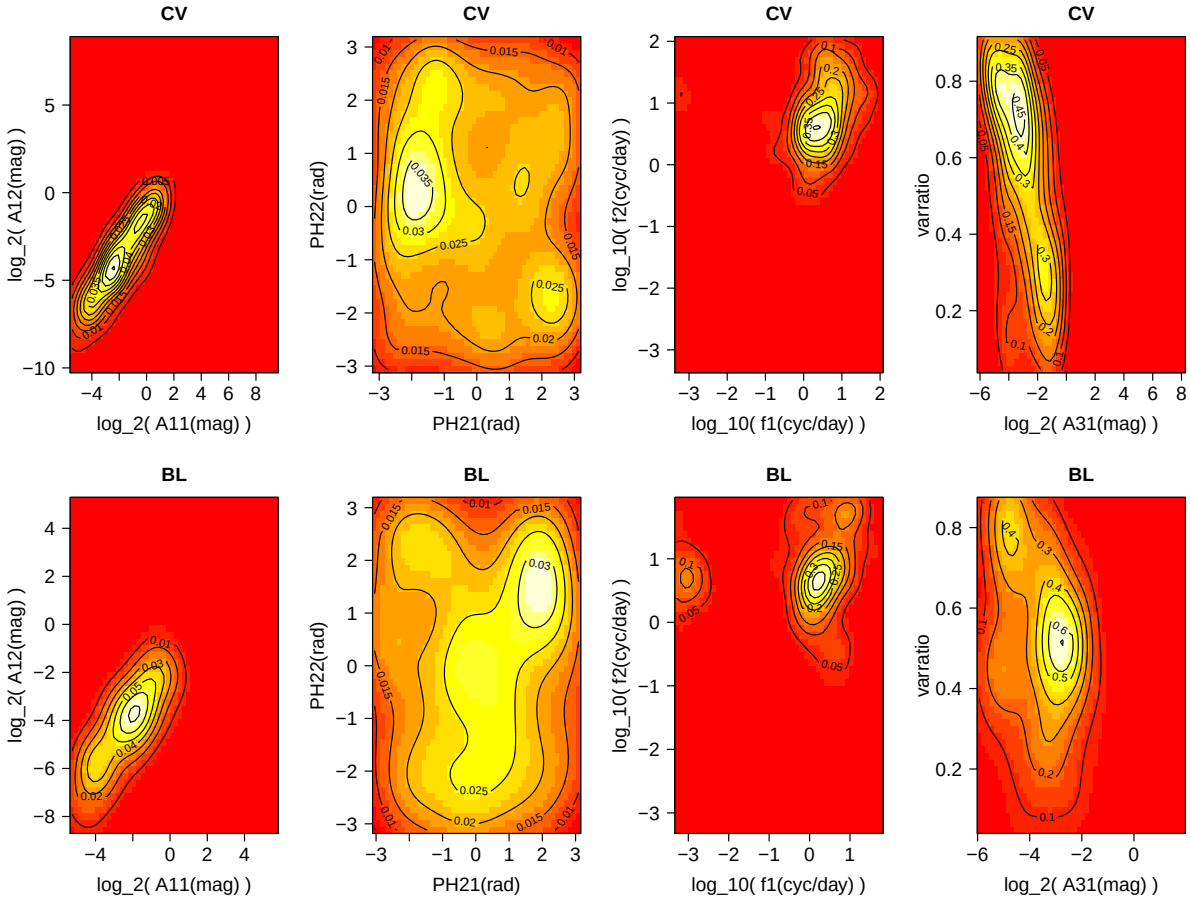


Figure 4: Magnitude differences for the BL and CV ensembles, for very short (left) and long (right) time differences. Only 10,000 points are shown in each plot.

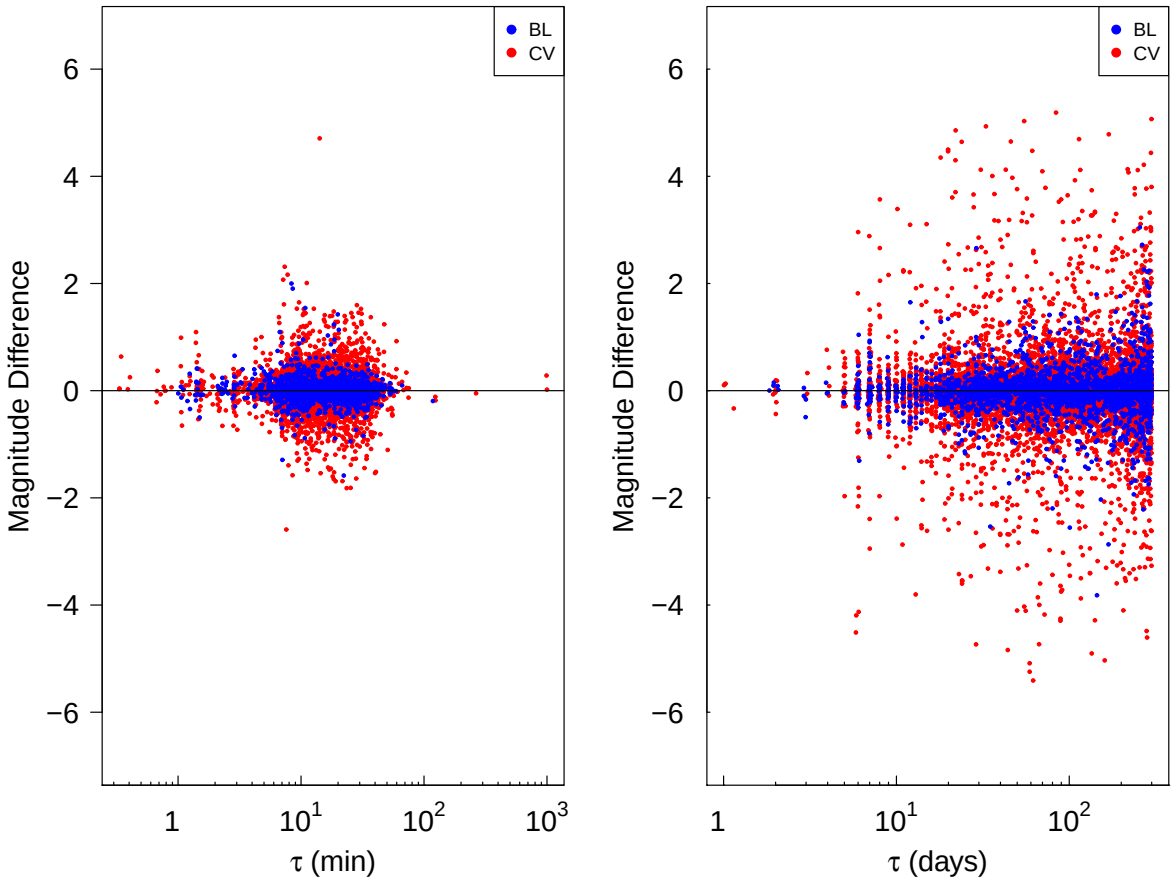


Figure 5: Ensemble SFs for the BL and CV data set. SFs for very short τ, with $\delta = 0.01$ min (left). SFs for long τ, with $\delta = 1$ day (right). The solid lines are the fitted linear regression. The size of the points is proportional to the number of observations used to calculate the particular value of the SF.

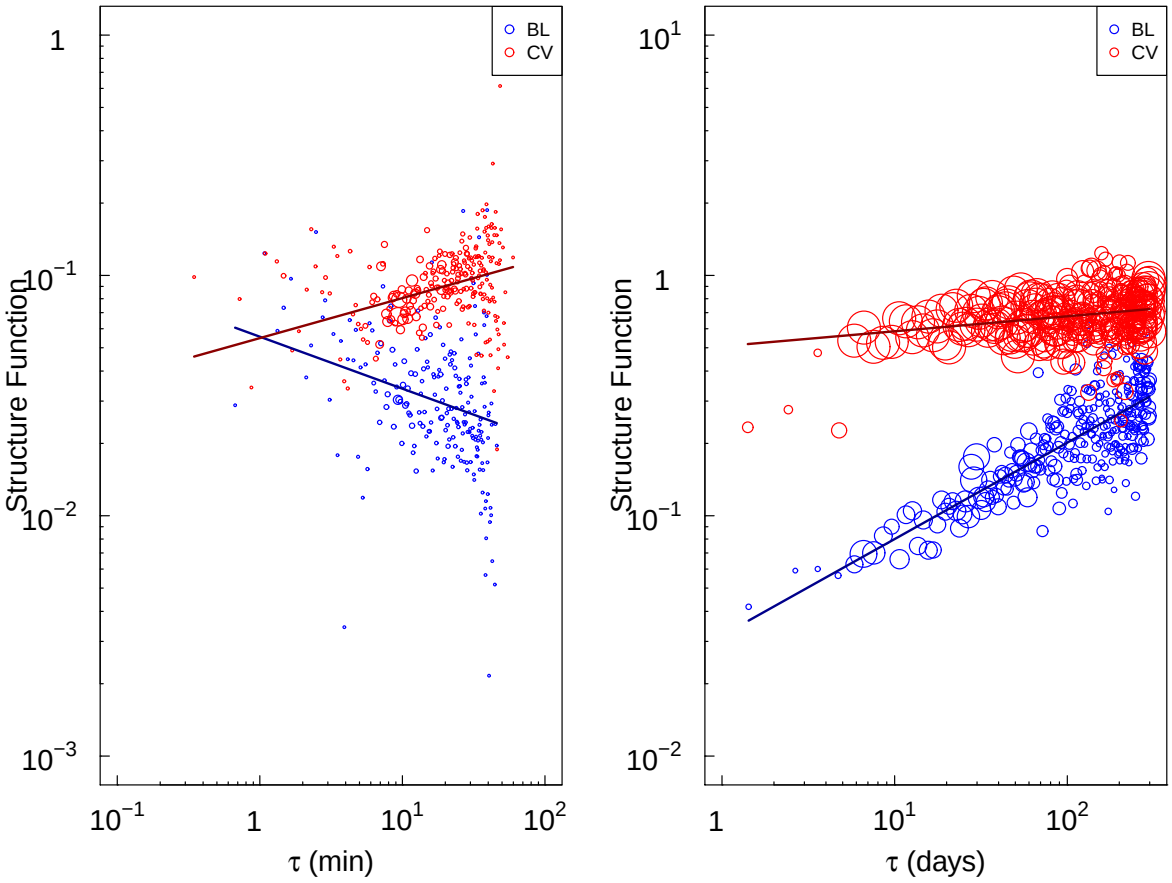


Figure 6: Individual SFs for long τ, with $\delta = 1$, for 6 BLs and 6 CVs. The colored solid lines are the fitted linear regression to the ensemble SF. The size of the points is proportional to the number of observations used to calculate the particular value of the SF. We fit a non-parametric regression to the individual SFs using regression splines and the number of spline knots is chosen by minimizing the generalized cross-validation score (black curves).

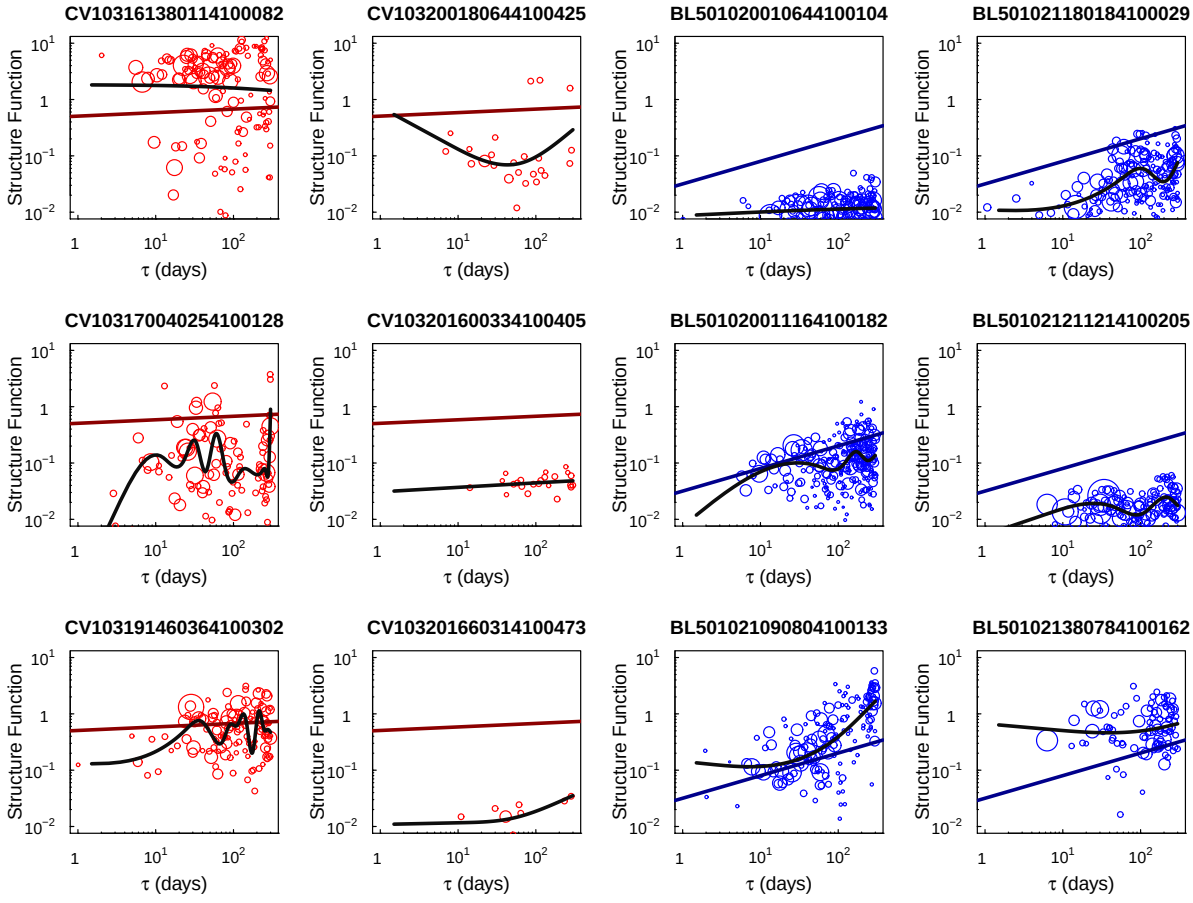


Figure 7: Density estimates of the distribution of the slope for the linear regression $\log_{10}(SF) \sim \log_{10}(\tau)$, long τ, with $\delta = 1$.

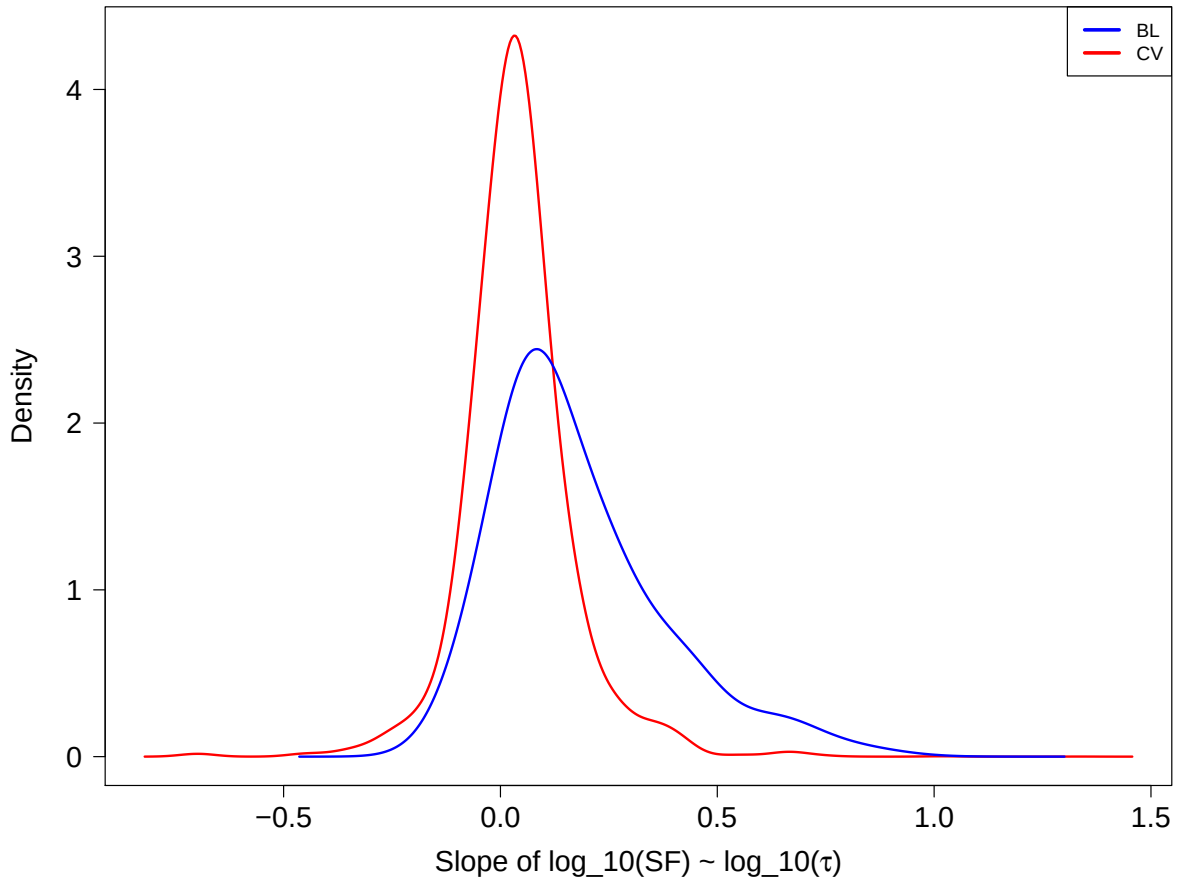


Figure 8: Density estimates of the other features calculated using the magnitude measurements of the light curves.

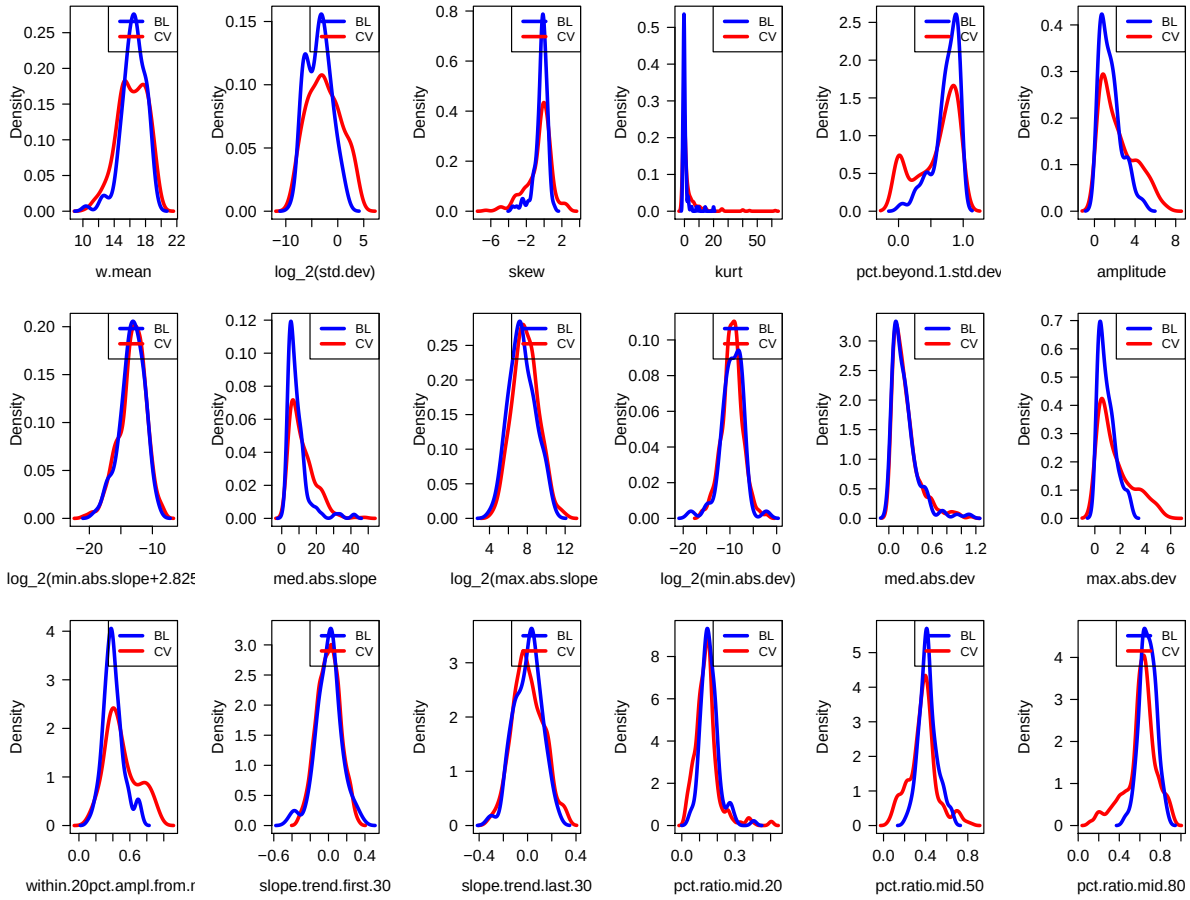


Figure 9: Four simulated time series. The red time series have labels $Y = 0$ and the blue have labels $Y = 1$.

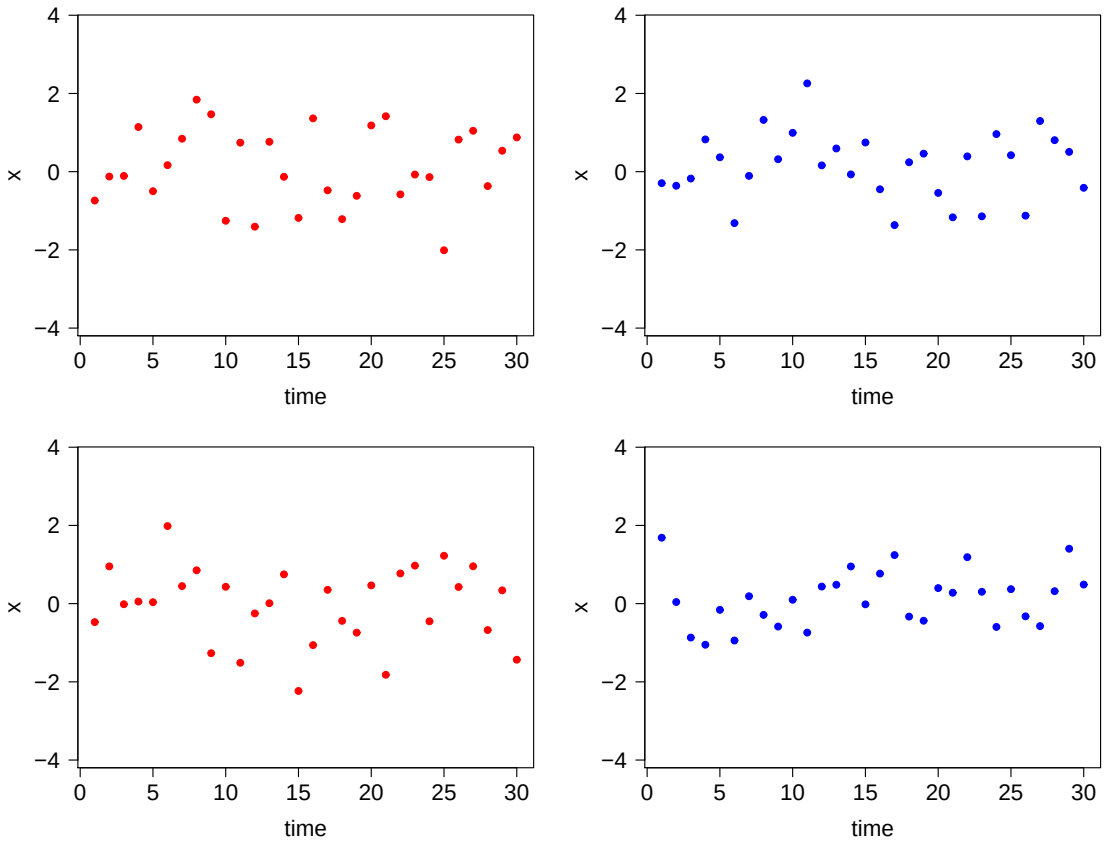


Figure 10: Smoothed class and overall error rates for different stratified bootstrap sample sizes of the majority class ($Y = 0$). The class 0 and class 1 error rates are roughly equal when the stratified bootstrap sample sizes are 55 for $Y = 0$ and 86 for $Y = 1$.

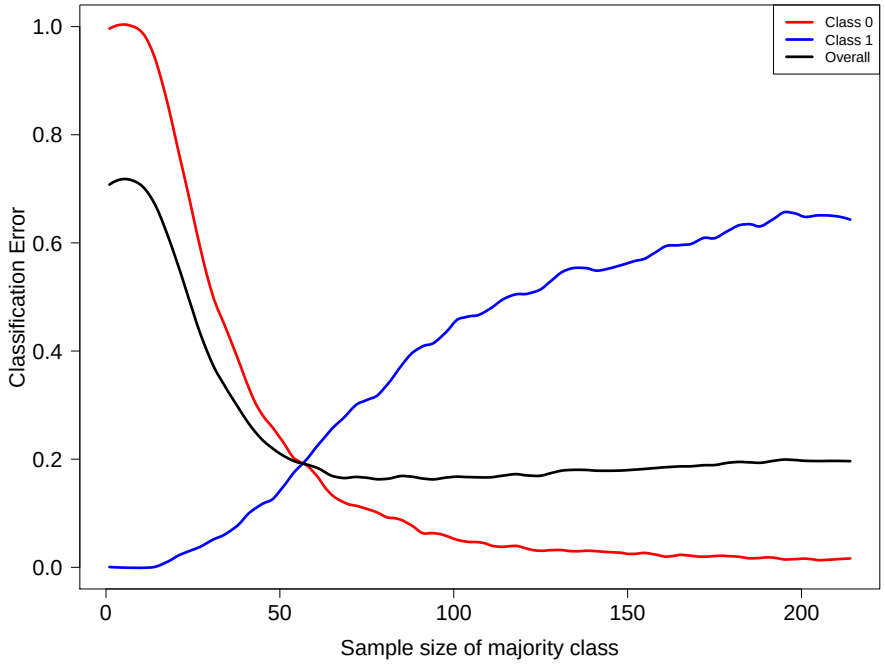


Figure 11: ROC curves for the imbalanced and balanced random forest classifiers on the time series data of Table 4 (curve A) and Table 5 (curve B).

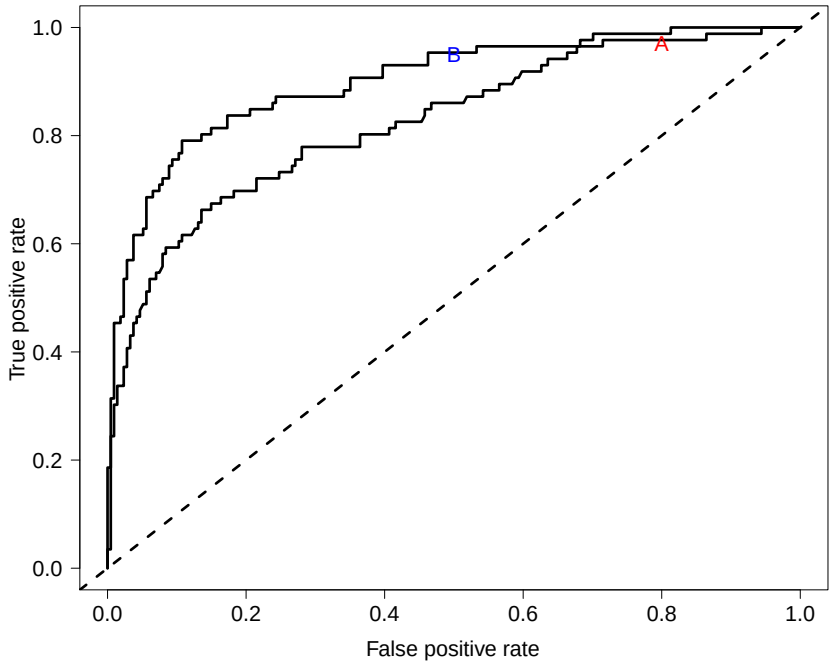


Figure 12: Distribution of the ancillary feature on one simulated data set. The ancillary feature is the sample standard deviation of the time series.

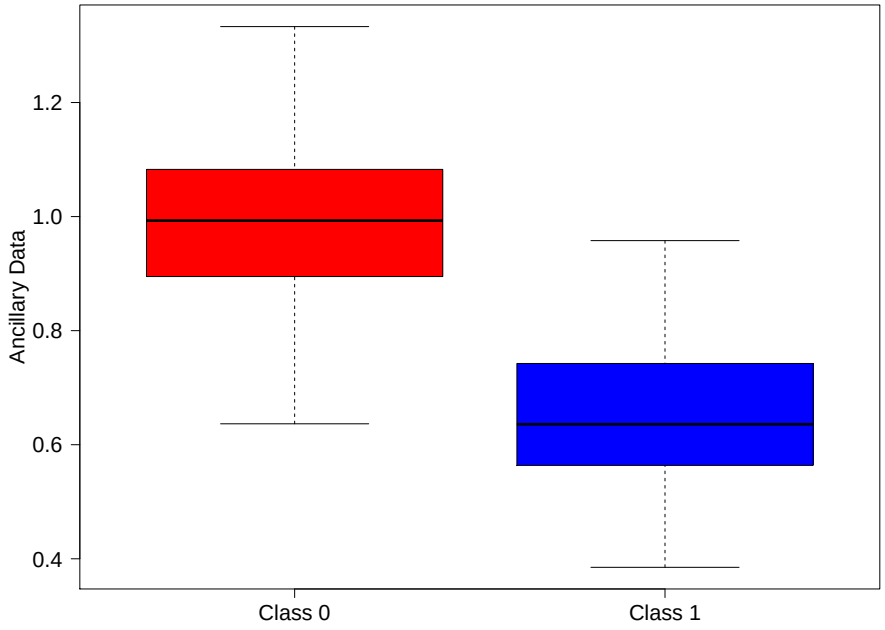


Figure 13: Nonparametric density estimates on one simulated data set. The red time series have labels $Y = 0$ and the blue have labels $Y = 1$.

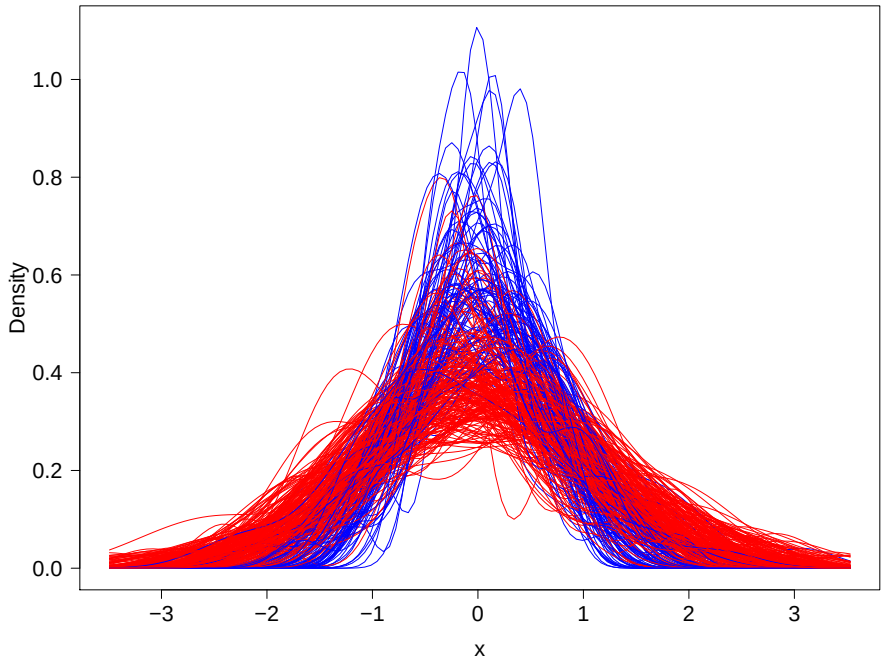


Figure 14: Estimate of $\beta(t)$ for predicting the ancillary feature from the density estimates via least squares. The dashed lines indicate pointwise 95% confidence limits for values of $\beta(t)$.

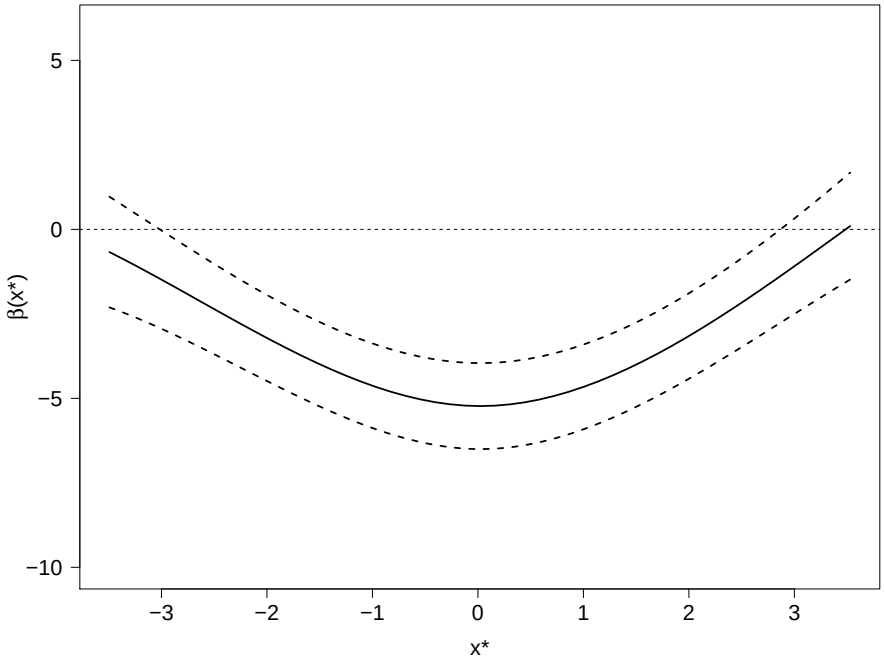


Figure 15: Observed ancillary feature values plotted against values predicted by FLR on the density estimates using least squares. The red points have labels $Y = 0$ and the blue have labels $Y = 1$. The black line is $y = x$.

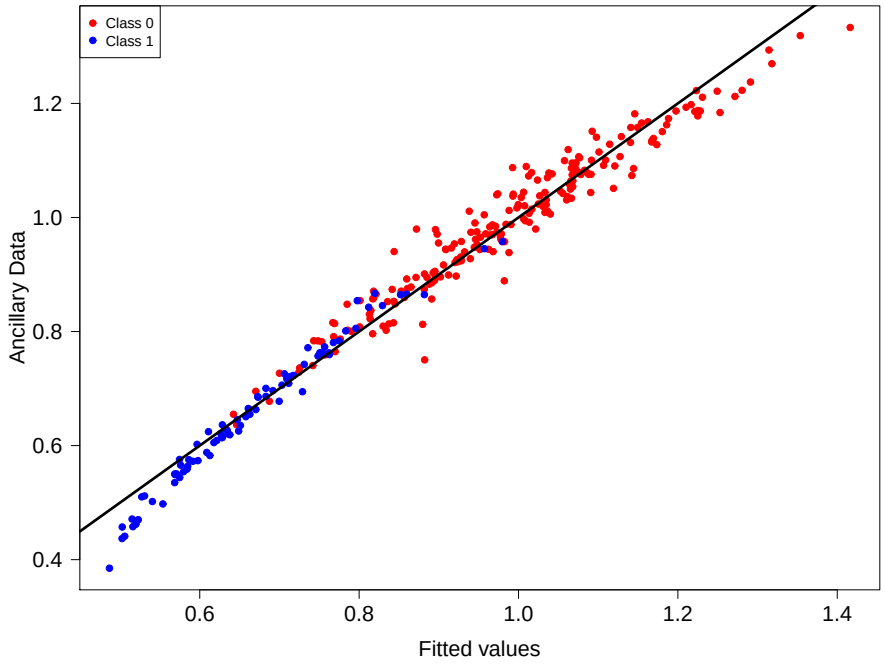


Figure 16: ROC curves for the imbalanced and balanced random forest classifiers on the predicted ancillary data of Table 6 (curve C) and Table 7 (curve D) in addition to the ROC curves in Figure 11.

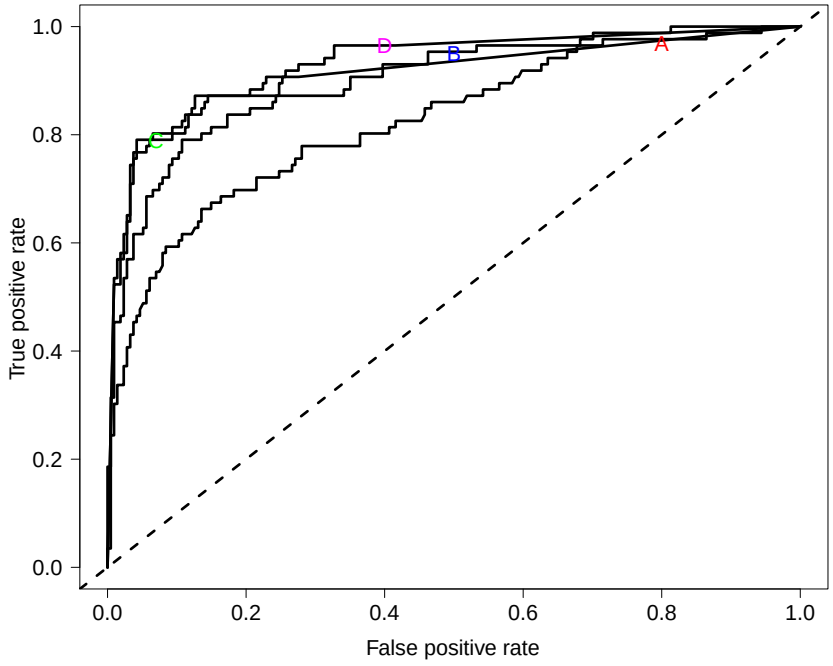


Figure 17: Cross-validation scores $CV(\lambda)$ for fitting the ancillary feature by the density estimates, with a penalty on the curvature of the coefficient function, $\beta(x^*)$.

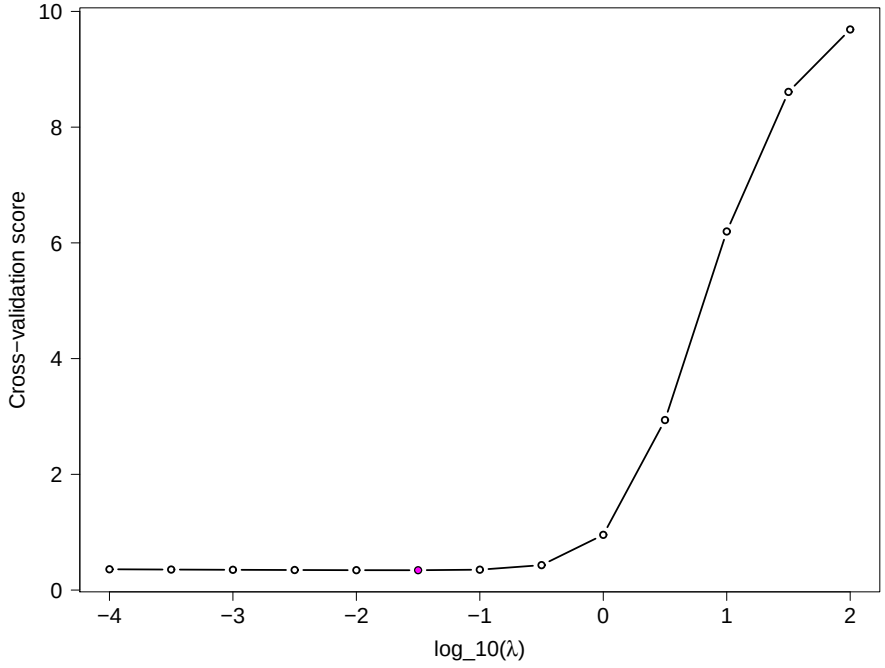
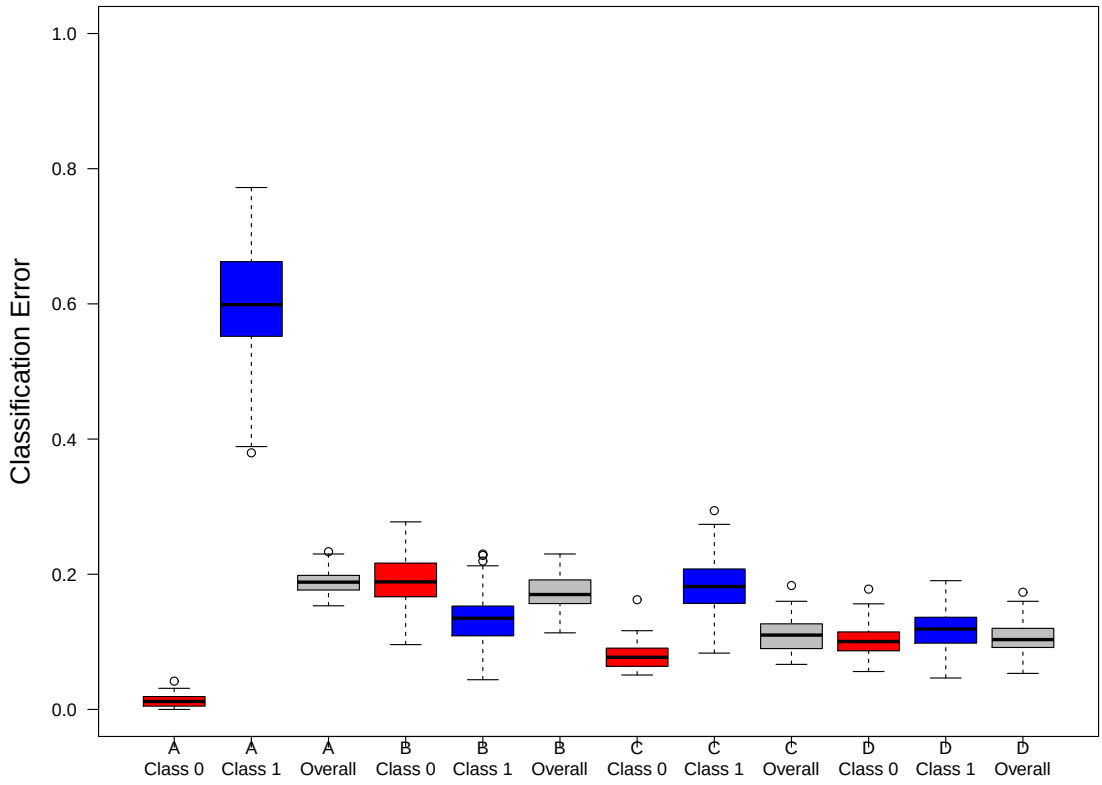


Figure 18: Classification error rates for 100 simulated data sets using random forest classifiers on: (A) the time series data; (B) the time series data, down-sampling class 0 to 55; (C) the estimated ancillary data; (D) the estimated ancillary data, down-sampling class 0 to 84.



7 References

References

Barning, F. (1963), "The numerical analysis of the light-curve of 12 Lacertae," *Bulletin of the Astronomical Institutes of the Netherlands*, 17, 22–28.

Bloomfield, P. (2000), *Fourier analysis of time series: An introduction*, Wiley-Interscience, 2 edn.

Breiman, L. (1996), "Bagging predictors," *Machine Learning*, 24, 123–140.

Breiman, L. (2001), "Random forests," *Machine Learning*, 45, 5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and regression trees*, Wadsworth and Brooks.

Butler, N. and Bloom, J. (2011), "Optimal time-series selection of quasars," *The Astronomical Journal*, 141, 93–103.

Chilès, J. and Delfiner, P. (1999), *Geostatistics: Modeling spatial uncertainty*, Wiley-Interscience, 1 edn.

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 and Astrophysics*, 475, 1159–1183.

Ferraty, F. and Vieu, P. (2006), *Nonparametric functional data analysis: Theory and practice*, Springer Science and Business Media.

Hellier, C. (2001), *Cataclysmic variable stars: how and why they vary*, Springer-Praxis books in astronomy and space sciences, Praxis.

Hughes, P., Aller, H., and Aller, M. (1992), "The University of Michigan radio astronomy data base. I - Structure function analysis and the relation between BL Lacertae objects and quasi-stellar objects," *The Astrophysical Journal*, 396, 469–486.

Kelly, B., Bechtold, J., and Siemiginowska, A. (2009), "Are the variations in quasar optical flux driven by thermal fluctuations?" *The Astrophysical Journal*, 698, 895–910.

Kolmogorov, A. (1941), "The local structure of turbulence in incompressible viscous fluid for very large Reynolds' numbers," *Akademiia Nauk SSSR Doklady*, 30, 301–305.

Lomb, N. (1976), "Least-squares frequency analysis of unequally spaced data," *The Astrophysical Journal*, 39, 447–462.

Miller, H., Carini, M., and Goodrich, B. (1989), "Detection of microvariability for BL Lacertae objects," *Nature*, 337, 627–629.

Ramsay, J. and Silverman, B. (2005), *Functional Data Analysis*, Springer.

Richards, J., Starr, D., Butler, N., Bloom, J., Brewer, J., 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, 10–30.

Rutman, J. (1978), "Characterization of phase and frequency instabilities in precision frequency sources: Fifteen years of progress," *Proceedings of the IEEE*, 66, 1048 – 1075.

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

Simonetti, J., Cordes, J., and Heeschen, D. (1985), "Flicker of extragalactic radio sources at two frequencies," *The Astrophysical Journal*, 296, 46–59.

Subba Rao, T., Priestley, M. B., and Lessi, O. (1997), *Applications of time series analysis in astronomy and meteorology*, Chapman and Hall.

Valtaoja, E. (1992), *Variability of blazars*, Cambridge University Press.

Zechmeister, M. and Kurster, M. (2009), "The generalised Lomb-Scargle periodogram," *Astronomy and Astrophysics*, 496, 577–584.