A FLEXIBLE PROCEDURE FOR POSITIVE–UNLABELED LEARNING

&

PERIODS ESTIMATION FOR MIRAS USING MULTI-BAND LIGHT CURVES AND INVERSE PERIOD-LUMINOSITY RELATIONS

A Dissertation

by

ZHENFENG LIN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Jianhua Huang
Co-Chair of Committee,	James P. Long
Committee Members,	Lucas M. Macri
	Raymond K.W. Wong
Head of Department,	Jianhua Huang

May 2019

Major Subject: Statistics

ABSTRACT

This dissertation contains two independent projects: the first project develops a general methodology for solving the Positive–Unlabeled (PU) learning problem, and the second project creates a hierarchical Bayesian model that solves a specific astronomical problem – periods estimation for Miras.

In the first project, we deal with the PU learning which considers two samples, a positive set P with observations from only one class and an unlabeled set U with observations from two classes. The goal is to classify observations in U. Class mixture proportion estimation (MPE) in U is a key step in PU learning. Blanchard et al. (2010) show that MPE in PU learning is a generalization of the problem of estimating the proportion of true null hypotheses in multiple testing problems. Motivated by this idea, we propose a flexible framework: fistly reduce the problem to one dimension via construction of a probabilistic classifier trained on the P and U data sets, and then apply a one–dimensional mixture proportion method to the observation class probabilities. The flexibility of this framework lies in the freedom to choose the classifier and the one–dimensional MPE method. Using this framework, we propose two mixture proportion estimators: one adapts ROC technique (Storey, 2002; Scott, 2015), and another adapts isotonic regression (Patra and Sen, 2015). Theoretically we prove the consistency of these two estimators. Empirically we demonstrate that our proposed estimators have competitive performance on simulated waveform data and a protein signaling problem. And the implementations of our estimators are tuning parameter free.

The second project of this dissertation is to present an inverse Period-Luminosity relation (PLR) enhanced multi-band semi-parametric model (SP3) to efficiently recover periods for quasi-periodic variable stars such as Miras. Mira variables are promising distance indicators because the oxygen-rich type Miras follow a tight PLR in the near-infrared. However, the Mira light curves are quasi-periodic, making their period estimation significantly challenging. In recent few years, several methods have been developed to estimate period for Miras. He et al. (2016) develop a single-band semi-parametric model based on the Gaussian processes tool. Yuan et al. (2018) ex-

tend the above model to a multi-band case. These two models are designed for fitting observations for single Mira (single-band or multi-band) and do not use the PLR. To borrow the strength across light curves, our proposed SP3 model uses inverse Period-Luminosity relation (iPLR) to adaptively feed a frequency prior to each light curve. This model outperforms existing methods in various simulated data sets.

DEDICATION

To my beloved wife, Peiqi.

ACKNOWLEDGMENTS

I would like to express my deepest gratitude to my committee chair and co-chair, Dr. Huang and Dr. Long, and my committee members, Dr. Macri and Dr. Wong, for their guidance and support throughout my Ph.D. study. I would also like to wholeheartedly thank two of my collaborators Dr. Wenlong Yuan and Dr. Shiyuan He, without whom the second project of this dissertation cannot be possible.

I am also grateful to my friends, colleagues, and the department faculties and staffs for making my time at Texas A&M University a wonderful experience. This dissertation would not have been possible without their warm love, continued patience, and endless support.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supported by a dissertation committee consisting of Professor Jianhua Huang, Professor Raymond K.W. Wong of the Department of Statistics, Professor James P. Long of the Department of Biostatistics at MD Anderson Cancer Center, and Professor Lucas M. Macri of the Department of Physics & Astronomy.

The analyses depicted in the first project were conducted in part by Dr. James P. Long. The second project was contributed in part by Dr. Lucas M. Macri, Dr. Jianhua Huang, Dr. Shiyuan He (Assistant Professor of Institute of Statistics and Big Data at The Renmin University of China), and Dr. Wenlong Yuan (post-doc fellow of the Department of Physics and Astronomy at The Johns Hopkins University).

All other work conducted for the dissertation was completed by the student independently.

Funding Sources

Graduate study was supported by the assistance from Department of Statistics at Texas A&M University. The first project was partially supported by the Statistics and Applied Mathematical Sciences Institute (SAMSI) in Fall 2016.

TABLE OF CONTENTS

Page

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

TABLE Page

1. INTRODUCTION

1.1 Introduction of the first project: a flexible procedure for Positive–Unlabeled learning

Let

$$X_1, \ldots, X_n \sim F = \alpha F_0 + (1-\alpha) F_1, \tag{1.1}$$
$$X_{L,1}, \ldots, X_{L,m} \sim F_1,$$

all independent, where F_0 and F_1 are distributions on $\mathbb{R}^p$ with densities f_0 and f_1 with respect to measure μ. The goal is to estimate α and the classifier

$$C_{01}(x) = \frac{(1-\alpha) f_1(x)}{\alpha f_0(x) + (1-\alpha) f_1(x)}, \tag{1.2}$$

which can be used to separate the unlabeled data $\{X_i\}_{i=1}^n$ into the classes 0 and 1. The above problem has been termed *Learning from Positive and Unlabeled Examples*, *Presence Only Data*, *Partially Supervised Classification*, and the *Noisy Label Problem* in the machine learning literature (Elkan and Noto, 2008; Ward et al., 2009; Ramaswamy et al., 2016; Scott et al., 2013; Scott, 2015; Liu et al., 2002). In this work, we use the term PU learning to refer to Model (1.1). Here we denote the positive set $P := \{X_{L,i}\}_{i=1}^m$ and the unlabeled set $U := \{X_i\}_{i=1}^n$. This setting is more challenging than the traditional classification framework where one possesses labeled training data belonging to both classes. In particular α and C_{01} are not generally identifiable from the data $\{X_i\}_{i=1}^n$ and $\{X_{L,i}\}_{i=1}^m$. PU learning has been applied to text analysis (Liu et al., 2002), time series (Nguyen et al., 2011), bioinformatics (Yang et al., 2012), ecology (Ward et al., 2009), and social networks (Chang et al., 2016).

Several strategies have been proposed for solving the PU learning problem. Ward et al. (2009) assume α is known and use logistic regression to classify U. The SPY method of Liu et al. (2002) classifies U directly by identifying a "reliable negative set." The SPY method has prac-

tical challenges including choosing the reliable negative set. Other strategies estimate α directly. Ramaswamy et al. (2016) estimate α via kernel embedding of distributions. Scott (2015) and Blanchard et al. (2010) estimate α using the ROC curve produced by a classifier trained on P and U.

Blanchard et al. (2010) show that MPE in the PU model is a generalization of estimating the proportion of true nulls in multiple testing problems. Specifically, suppose that F_0 and F_1 are one–dimensional distributions and F_1 is known. Then the unlabeled set $X_1, \ldots, X_n$ may be interpreted as test statistics with the hypotheses:

$$H_0 : X_i \sim F_1,$$
$$H_a : X_i \sim F_0.$$

In this context, $1 - \alpha$ is the proportion of true null hypotheses and the classifier C_{01} is the local FDR (Efron et al., 2001). There are many works on addressing identifiability and estimation of α and C_{01} in this simpler setting (Patra and Sen, 2015; Efron, 2012; Genovese and Wasserman, 2004; Robin et al., 2007; Meinshausen and Rice, 2006).

FDR α estimation methods have been developed for one–dimensional MPE problems and are not directly applicable on the multidimensional PU learning problem in which $X_i \in \mathbb{R}^p$. In this work, we show that the PU MPE problem can be reduced to dimension one by constructing a classifier on the P versus U data sets followed by transforming observations to class probabilities. One dimensional MPE methods from the FDR literature can then be applied to the class probabilities. Computer implementation of this approach is straightforward because one can use existing classifier and one–dimensional MPE algorithms. We prove consistency for adaptations of two one–dimensional MPE methods: Storey (2002) based on empirical processes and Patra and Sen (2015) based on isotonic regression. These proofs use results from empirical process theory. We show that the ROC method used in Blanchard et al. (2010) and Scott (2015) is a variant of the method proposed by Storey (2002).

1.2 Introduction of the second project: periods estimation for Miras using multi-band light curves and inverse Period-Luminosity relations

The current expansion rate of the Universe, known as the Hubble constant or H_0, is one of a few fundamental parameters required to understand the contents and evolution of the cosmos. The most precise and accurate technique for estimating H_0 is based on white-dwarf supernovae and Cepheid variables (Riess et al., 2016). The latest determination of this parameter (Riess et al., 2018) has a value that is 3.8σ higher than the expectation based on the current cosmological model (Planck Collaboration et al., 2018). If this discrepancy is confirmed at a higher confidence level (i.e., $> 5\sigma$), it may be evidence for an additional component of the Universe.

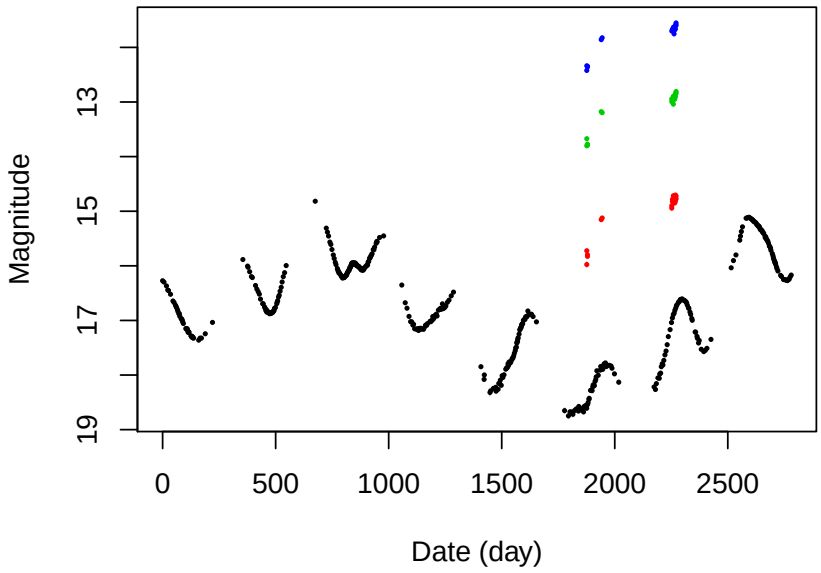
In order to reach above goals of accurately estimating H_0 and proving additional component of the Universe, it is desirable to establish different techniques that can provide equally robust but independent estimates of H_0. Mira variables (hereafter, Miras) have been shown to be promising alternatives to Cepheids (Whitelock et al., 2014; Huang et al., 2018) and can be used to calibrate additional techniques that will supplement the white-dwarf supernovae method.

Miras are long-period ($P = 100-1000$ days) pulsators with large-amplitude cyclical variations in luminosity (Kholopov et al., 1985). Owing to a different pulsation mechanism than strictly-periodic variable stars such as Cepheids, Miras exhibit quasi-periodic changes as seen in figure 1.1. Figure 1.1a shows a typical "light curve" of a Mira from the OGLE survey (Udalski et al., 2008). This particular star is located in a satellite galaxy of the Milky Way known as the Large Magellanic Cloud (hereafter, LMC). A "light curve" is a time-series data containing brightness measurements (known as "magnitudes") at a series of unevenly-spaced points. The measurement error associated with each magnitude is also provided in standard astronomical surveys. The measurement error describes an one-standard deviation uncertainty on the magnitude. Typically, the brightness of a star is measured at different ranges of wavelengths (known as "bands") using various filters that limit the light received by the detector. A light curve is said to be "quasi-periodic" in that it has an intrinsically periodic pattern but the amplitudes or phases for each cycle may be different. The LMC is relatively nearby and has been extensively imaged by the OGLE survey for more than

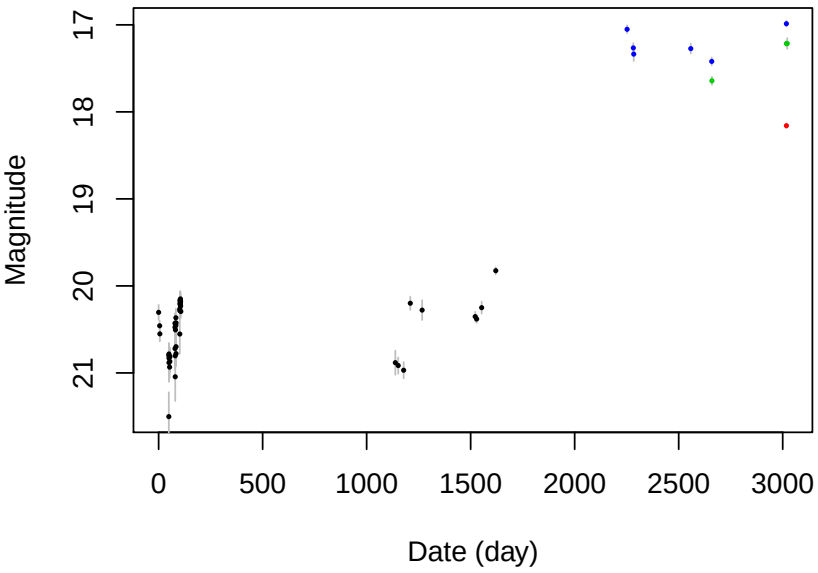
two decades, yielding very high-quality measurements as shown in the aforementioned figure. In contrast, most astronomical time-series data are significantly noisier and rather poorly sampled. Figure 1.1b shows the light curve of a Mira in the galaxy Messier 33 (M33) from a recent study (Yuan et al., 2018). M33 is located $\sim 17\times$ farther than the LMC, making the stars $\sim 300\times$ fainter and requiring much larger telescopes that can only be used infrequently.

The period of pulsation of a variable star is used to construct Period-Luminosity relations (PLRs), which are one of the techniques used to estimate H_0. PLRs are linear correlations between the logarithm of the period and the logarithm of the average luminosity for stars of a given type, first discovered for Cepheids by Henrietta Leavitt more than a century ago (Leavitt and Pickering, 1912). One subtype of Miras (known as "Oxygen-rich") also exhibit tight PLRs (Ita et al., 2005; Soszyński et al., 2007; Yuan et al., 2017b; Yuan et al., 2018). In order to realize the full potential of PLRs one must have accurate period estimations, which are particularly challenging for Miras due to their quasi-periodic behavior.

In the recent few years, two methods have been proposed to estimate the periods of Miras. He et al. (2016) use a semi-parametric model to estimate period with densely-sampled single-band data (e.g. figure 1.1a). But this model is difficult in recovering the period for sparsely-sampled light curve. Yuan et al. (2018) make an improvement over above model by taking advantage of multi-band information. This multi-band model improves the performance on spasely-sample data, but its period estimates should be followed by an *ad hoc* correction before they are used for PLRs inference. All these methods are designed for an individual light curve. Even when estimating periods for a set of stars that share a single PLR, the light curves are processed separately and their periods are estimated independently. A natural and intuitive way to borrow information from each light curve is to use PLRs, which can serve as a conductor to orchestrate the connection between light curves. However the PLRs are not easy to use, but their inverse relations – inverse Period-Luminosity relations (iPLRs) – are more straightforward to be used to enhance the performance of the aforementioned two methods. In this project, we develop an iPLR-enhanced multi-band model for quasi-periodic light curves, especially for poor quality light curves as shown in figure 1.1b.



(a)



(b)

Figure 1.1: Light curves of Miras in two nearby galaxies: (a) LMC and (b) M33. Each star has time-series data in four bands: $\cdots$: I; $\cdots$: H; $\cdots$: J; $\cdots$: K. The vertical bars around each point are two-standard deviation. . The quality and sampling of the I light curve for the LMC Mira is atypical; most astronomical data sets will be poorly sampled and noisy as the other examples.

2. A FLEXIBLE PROCEDURE FOR POSITIVE–UNLABELED LEARNING

In this chapter, we develop a flexible procedure for solving the Positive–Unlabeled learning problem. Using this procedure, two mixture proportion estimators are proposed in this project. We will demonstrate that these two estimators have competitive performance on simulated data and real data.

The rest of this chapter is organized as follows. In section 2.1 we give a sketch of the proposed procedure, which includes two proposed estimators C-patra/sen and C-roc. This section consists of three subsections. First, a motivation of the procedure from hypothesis testing community is explained. Second, identifiability of α is addressed. Third, a workflow is provided to explain how to implement the proposed procedure. In section 2.2 we show that model (1.1) can be reduced to one-dimension with a classifier. In section 2.3 we show consistency of two α estimators. In section 2.4 we numerically show that the estimators perform well in various settings.

2.1 Background and proposed procedure

2.1.1 Multiple testing, FDR, and estimating the proportion of true nulls

Suppose one conducts n tests of null hypothesis $H_0 : X_i \sim F_1$ versus alternative hypothesis $H_a : X_i \sim F_0$, $i = 1, \ldots, n$. The X_i are typically test statistics or p–values and the null distribution F_1 is assumed known (usually $Unif[0, 1]$ in the case of X_i being p–values). The distribution of the X_i are $F = \alpha F_0 + (1 - \alpha)F_1$, where $1 - \alpha$ is the proportion of true null hypotheses. The *false discovery rate* (FDR) is the expected proportion of false rejections. If R is the number of rejections and V is the number of false rejections then FDR $\equiv \mathbb{E}[\frac{V}{R}\mathbf{1}_{R>0}]$. Benjamini and Hochberg (1995) developed a linear step–up procedure which bounds the FDR at a user specified level β. In fact, this procedure is conservative and results in an FDR $\leq \beta(1 - \alpha) \leq \beta$. This conservative nature causes the procedure to have less power than other methods which control FDR at β. *Adaptive* FDR control procedures first estimate $1 - \alpha$ and then use this estimate to select a β which ensures control at some specified level while maximizing power. Many estimators of α have been proposed

(Patra and Sen, 2015; Storey, 2002; Benjamini et al., 2006; Langaas et al., 2005; Blanchard and Roquain, 2009; Benjamini and Hochberg, 2000).

There are two reasons why these procedures cannot be directly applied to the PU learning problem. First, many of the methods have no clear generalization to dimension greater than one because they require an ordering of the test statistics or p–values. Second, the distribution F_1 is assumed known where as in the PU learning problem we only have a sample from this distribution. The classifier dimension reduction procedure we outline in subsection 2.1.3 addresses the first point by transforming the PU learning problem to 1–dimension. The theory we develop in section 2.2 and 2.3 addresses the second issue.

2.1.2 Identifiability of α and C_{01}

Many works in both the PU learning and multiple testing literature have discussed the non–identifiability of the parameters α and F_0. For any given (α, F_0) pair with $\alpha < 1$, one can find a $\gamma > 0$ such that $\alpha' \equiv \alpha + \gamma \leq 1$. Define $F_0' \equiv \frac{\alpha F_0 + \gamma F_1}{\alpha + \gamma}$. Then

$$F = \alpha' F_0' + (1 - \alpha') F_1,$$

which implies (α', F_0') and (α, F_0) result in the same distributions for P and U.

To address this issue, we follow the approach taken by Blanchard et al. (2010) and Patra and Sen (2015) and estimate a lower bound on α defined as

$$\alpha_0 := \inf \left\{ \gamma \in (0, 1] : \frac{F - (1 - \gamma) F_1}{\gamma} \text{ is a c.d.f.} \right\}. \tag{2.1}$$

The parameter α_0 is identifiable. Recall the objective is to estimate

$$C_{01}(x) = \frac{(1 - \alpha) f_1(x)}{\alpha f_0(x) + (1 - \alpha) f_1(x)},$$

the probability an observation in U is from class 1. We can use α_0 to upper bound C_{01} in the

following way. Note that the classifier

$$C(x) = \frac{\pi f_1(x)}{\pi f_1(x) + (1-\pi) f(x)}$$

outputs the probability an observation is from the labeled data set at a given x. We can approximate C by training a model on the P versus U data sets. The classifiers C and C_{01} are related through α. To see this, note that after some algebra

$$\frac{f_1(x)}{f(x)} = \frac{C(x)}{1-C(x)} \frac{1-\pi}{\pi}.$$

Thus

$$C_{01}(x) = \frac{(1-\alpha) f_1(x)}{f(x)} = \frac{1-\pi}{\pi} \frac{C(x)}{1-C(x)} (1-\alpha).$$

Since α is not generally identifiable, neither is C_{01}. However the plug-in estimator using C_n (a classifier trained on P versus U) and $\widehat{\alpha}_0$ (some estimator of α_0) serves as an upper bound for C_{01}. Specifically,

$$\widehat{C}_{01}(x) = \frac{1-\pi}{\pi} \frac{C_n(x)}{1-C_n(x)} (1-\widehat{\alpha}_0).$$

We can classify an unlabeled observation X_i as being from F_1 if $\widehat{C}_{01}(X_i) > \frac{1}{2}$. The problem has now been reduced to estimation of α_0. The classifier C_n plays an important role in estimation of α_0 as well, as shown in the following section.

2.1.3 Workflow for α_0 estimation

The proposed procedure to estimate α_0 in model (1.1) is summarized in figure 2.1. The key idea of this procedure is to reduce the dimension of PU learning problem via the classifier C_n trained on P versus U and then apply a one-dimensional MPE method on the transformed data to estimate α_0. The procedure consists of three steps:

- **Step 1**. Label the P samples with pseudo label ($Y = 1$) and label the U samples with pseudo label ($Y = 0$). Hence we have $\tilde{P} := \{(X_{L,i}, Y_i = 1), i = 1, \ldots, m\}$ and $\tilde{U} := \{(X_i, Y_i =$

$0), i = 1, \ldots, n\}$.

- **Step 2**. Train a probabilistic classifier $C_n(\cdot) = \widehat{P}(Y = 1|X = \cdot)$ on $\tilde{P}$ versus $\tilde{U}$. Compute probabilistic predictions: $\mathbf{p}_1 := \{p_{1i}, i = 1, \ldots, m\}$ and $\mathbf{p}_0 := \{p_{0i}, i = 1, \ldots, n\}$, where $p_{1i} := C_n(X_{L,i})$ and $p_{0i} := C_n(X_i)$.

- **Step 3**. Apply a one-dimensional MPE method to $\mathbf{p}_1$ and $\mathbf{p}_0$ to estimate α_0.

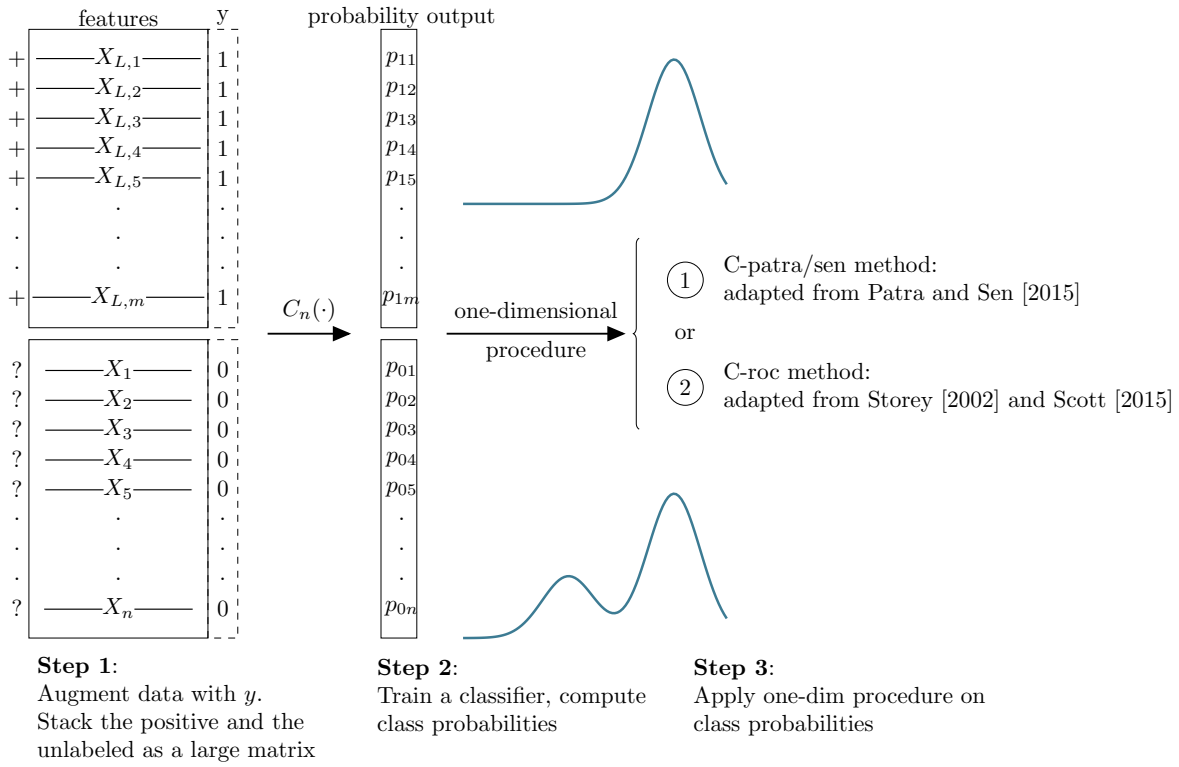


Figure 2.1: Workflow of proposed procedure. In **Step 1**, "+" denotes the positive samples, and "?" denotes the unlabeled samples whose label are unknown (can be "+" or "-"). We stack the set P and the set U together as a large matrix, and add a new column y to manually impose pseudo labels on observations: "1" for $X_{L,i}$ and "0" for X_i. In **Step 2**, a classifier $C_n(\cdot)$ is trained on the stacked matrix and the probability predictions ($y = 1$ as reference) are obtained. In **Step 3**, a one-dimensional procedure is applied to the probability output from Step 2. In this project, two methods C-patra/sen and C-roc are introduced as examples. The upper density curve is used to demonstrate that the $\mathbf{p}_1 := \{p_{1i}\}_{i=1}^m$ are from one population, while the bottom density curve shows that $\mathbf{p}_0 := \{p_{0i}\}_{i=1}^n$ are from mixture of two populations.

We augment the original data with pseudo labels in Step 1, in order to use a supervised learning classification algorithm. In Step 2 we use Random Forest (Breiman, 2001). However in principle any classifier can be used. Note that the p_{0i} and p_{1i} are scalars. Hence in Step 3 we can utilize any one-dimensional method to estimate α_0. In this work we adapt two methods – one from Storey (2002) and Scott (2015), another from Patra and Sen (2015). Note that the original theory developed for these methods assumed that the null distribution is known, but in the PU problem we need to estimate it from $\mathbf{p_1}$. Since this setting is more complex and more challenging, new theory is needed. In section 2.3, we prove the consistency of two estimators in the PU setting, using Theorems 1 and 2.

2.2 Dimension reduction via classifier

Using the P and U samples we can make probabilistic predictions, i.e. compute the probability that the observation is from distribution F_1 versus from distribution F. The true classifier is

$$C(x) = \frac{f_1(x)\pi}{f_1(x)\pi + f(x)(1-\pi)}, \tag{2.2}$$

where $\pi = \frac{m}{m+n}$ is the proportion of labeled sample within the entire data. We treat π as a known constant.

Denote the distribution of probabilistic predictions for P and U, respectively, as

$$G(t) = P(C(X) \leq t | X \sim F), \tag{2.3}$$

$$G_L(t) = P(C(X) \leq t | X \sim F_1). \tag{2.4}$$

One can consider the two-component mixture model

$$G = \alpha^G G_s + (1-\alpha^G) G_L, \tag{2.5}$$

for α^G and G_S, which are again potentially non-identifiable. Define

$$\alpha_0^G := \inf \left\{ \gamma \in (0,1] : \frac{G - (1-\gamma)G_L}{\gamma} \text{ is a c.d.f.} \right\}. \tag{2.6}$$

Theorem 1. $\alpha_0^G = \alpha_0$.

See appendix A.1.1 for a proof. Theorem 1 shows one can solve the p–dimensional MPE problem (2.1) by solving the 1–dimensional MPE problem (2.6). In what follows we use α_0 instead of α_0^G to simplify notation.

In practice, the classifier $C(X)$ is approximated by a trained model $C_n(X)$ on a given sample. For convenience, we assume the classifier $C_n(X)$ is trained using another independent sample $\mathcal{D}'_n$. The $\mathcal{D}'_n$ is omitted in the following to lighten notation. We require the approximated classifier to be a consistent estimator of the true classifier.

Assumptions 1. *We assume*

$$\mathbb{E}|C_n(X) - C(X)| = O\left(n^{-\tau}\right), \tag{2.7}$$

for some $\tau > 0$.

Such convergence results have been proven for a variety of probabilistic classifiers, including variants of Random Forest (Biau, 2012). Define

$$G_{L,n}(t) := \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{C_n(X_{L,i}) \leq t},$$
$$G_n(t) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{C_n(X_i) \leq t}.$$

Intuitively we can think of $G_{L,n}$ and G_n as approximate empirical distribution functions of G_L and G respectively. The approximation is due to the fact that C is estimated with C_n. Thus we would expect Glivenko-Cantelli and Donsker properties for $G_n(t)$ and $G_{L,n}(t)$. However problems can arise when $C(X)$ is not continuous. Essentially convergence in probability for $C(X)$, implied by

Assumptions 1, only implies convergence of distribution functions at points of continuity. By assuming G_L and G possess densities, we can obtain uniform convergence of distribution functions.

Assumptions 2. *We assume that G and G_L are absolutely continuous and have bounded density functions g and g_L.*

Theorem 2. *Under Assumption 1 and 2, for $\beta = \tau/3$*

$$n^{\beta}(G_{L,n}(t) - G_L(t)) \text{ is } O_P(1),$$
$$n^{\beta}(G_n(t) - G(t)) \text{ is } O_P(1),$$

where both $O_P(1)$ are uniform in t.

See appendix A.1.2 for a proof. The result from theorem 2 is the key step in showing consistency of our α_0 estimators in the following sections.

2.3 Estimation of α_0

We generalize a one–dimensional method of Patra and Sen (2015) to the PU learning problem. We call it C-patra/sen to emphasize the fact that the method developed in Patra and Sen (2015) is applied to the output of a classifier. Then we generalize a one dimensional method of Storey (2002) to the PU learning problem. We show that the ROC method developed in Blanchard et al. (2010) and Scott (2015) can be viewed as a variant of the Storey (2002) idea. We develop a version of ROC termed C-roc.

2.3.1 C-patra/sen

Patra and Sen (2015) remove as much of the $G_{L,n}$ distribution from G_n as possible, while ensuring that the difference is close to a valid cumulative distribution function. We briefly review the idea and provide theoretical results to support use of this procedure in the PU learning problem. See Patra and Sen (2015) for a fuller description of the method in the one–dimensional case.

For any $\gamma \in (0, 1]$ define

$$\widehat{G}^{\gamma}_{s,n} = \frac{G_n - (1-\gamma)G_{L,n}}{\gamma}.$$

If $\gamma \geq \alpha_0$, $\widehat{G}^{\gamma}_{s,n}$ will be a valid c.d.f. (up to sampling uncertainty) while the converse is true if $\gamma < \alpha_0$. Find the closest valid c.d.f. to $\widehat{G}^{\gamma}_{s,n}$, termed $\check{G}^{\gamma}_{s,n}$, and measure the distance between $\widehat{G}^{\gamma}_{s,n}$ and $\check{G}^{\gamma}_{s,n}$. Define

$$\check{G}^{\gamma}_{s,n} = \underset{\text{all c.d.f. } W(t)}{\operatorname{argmin}} \int \left(\widehat{G}^{\gamma}_{s,n} - W(t)\right)^2 dG_n(t), \tag{2.8}$$

$$d_n(g,h) = \sqrt{\int (g(t) - h(t))^2 \, dG_n(t)}.$$

Isotonic regression is used to solve Equation 2.8. If $d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \approx 0$, then $\alpha_0 \leq \gamma$ where the level of approximation is a function of the estimation uncertainty and thus the sample size. Given a sequence c_n define

$$\widehat{\alpha}_0^{c_n} = \inf\left\{\gamma \in (0,1] : \gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \leq \frac{c_n}{n^{\beta-\eta}}\right\}$$

where $\eta \in (0, \beta)$ is a constant and the rate β is from Theorem 2.

Theorem 3. *Under Assumptions 1 and 2, if $c_n = o(n^{\beta-\eta})$ and $c_n \to \infty$, then $\widehat{\alpha}_0^{c_n} \xrightarrow{p} \alpha_0$.*

The proof, contained in appendix A.1.4, is a generalization of results in Patra and Sen (2015) which accounts for the fact that both G_n and $G_{L,n}$ are estimators. While Theorem 3 provides consistency, there are a wide range of choices of c_n. Patra and Sen (2015) showed that $\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n})$ is convex, non-increasing and proposed letting $\widehat{\alpha}_0$ be the γ that maximizes the second derivative of $\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n})$. We use this implementation in our numerical work in section 2.4.

2.3.2 C-roc

Recalling the definitions of G, G_s, and G_L from Section 2.2, note

$$G(t) = \alpha G_s(t) + (1-\alpha) G_L(t) \leq \alpha + (1-\alpha) G_L(t)$$

for all t. Thus for any t such that $G_L(t) \neq 1$ we have

$$k(t) \equiv \frac{G(t) - G_L(t)}{1 - G_L(t)} \leq \alpha.$$

In the FDR literature, G_L is the distribution of the test statistic or p–value under the null hypothesis and is generally assumed known. Thus only G must be estimated, usually with the empirical cumulative distribution function. Storey (2002) proposed an estimator for $k(t)$ at fixed t (Equation 6) and determined a bootstrap method to find the t which produces the best estimates of the FDR.

The PU problem is more complicated in that one must estimate G and G_L. However the structure of G and G_L enables one to estimate the identifiable parameter α_0. Specifically with $t^* = \inf\{t : G_L(t) \geq 1\}$ we have

$$\lim_{t \uparrow t^*} k(t) = \alpha_0. \tag{2.9}$$

See Lemma 1 for a proof. This result suggests estimating α_0 by substituting the empirical estimators of G_n and $G_{L,n}$ into Equation 2.9 along with a sequence $\widehat{t}$ which is converging to the (unknown) t^*. Such a sequence $\widehat{t}$ must be chosen so that the estimated denominator $1 - \widehat{G}_{L,n}(\widehat{t})$ is not converging to 0 too fast (and hence too variable). For $\widehat{t}$ we use a quantile of the empirical c.d.f. which is converging to 1, but at a rate slower than the convergence of the empirical c.d.f.. For some $q \in (0, \beta)$, define

$$\widehat{t} = \inf\{t : G_{L,n}(t) \geq 1 - n^{-q}\} - n^{-1}.$$

The n^{-1} term in $\widehat{t}$ avoids technical complications.

Theorem 4. *Under Assumptions 1 and 2*

$$k_n(\widehat{t}) \equiv \frac{G_n(\widehat{t}) - G_{L,n}(\widehat{t})}{1 - G_{L,n}(\widehat{t})} \xrightarrow{P} \alpha_0.$$

See appendix A.1.3 for a proof.

2.3.2.1 *Connection with ROC method*

The ROC method of Scott (2015) (Proposition 2) and Blanchard et al. (2010) is a variant of the Storey (2002) method with a particular cutoff value t. Define the true ROC curve by the parametric equation

$$\{(G_L(t), G(t)) : t \in [0, 1]\}.$$

Scott (2015) showed that α_0 is the supremum of one minus the slope between (1,1) and any point on the ROC curve.[1] This is equivalent to the Storey method because

$$\begin{aligned}\alpha_0 &= \sup_t 1 - \frac{1 - G(t)}{1 - G_L(t)} \\ &= \sup_t \frac{G(t) - G_L(t)}{1 - G_L(t)} \\ &= \sup_t k(t).\end{aligned}$$

The true ROC curve is not known, so α_0 cannot be computed directly from this expression. Blanchard et al. (2010) found a consistent estimator and Scott (2015) determined rates of convergence using VC theory. For application to data, Scott (2015) splits the labeled and unlabeled data sets in half, constructs a kernel logistic regression classifier on half the data, and estimates the slope between (1,1) and a discrete set of points on the ROC curve. The α_0 estimate is the supremum of 1 minus each of these slopes. Thus we see that the ROC method and earlier methods developed in the FDR literature are in the same family of α estimation strategies. Choosing a t in the Storey approach is equivalent to choosing a point on the ROC curve.

2.3.2.2 *Practical implementation*

We consider two implementations of these ideas. The method of Scott (2015), using a kernel logistic regression classifier and a PU training–test set split to estimate tuning parameters, is referred to as "ROC." To facilitate comparison with C-patra/sen, we consider another version with a

[1] Scott (2015) estimated $\kappa = 1 - \alpha$. We have modified the ROC method notation to reflect the α notation used in this work.

Random Forest classifier using out–of–bag probabilities to construct the ROC curve. We call this method C-roc.

2.4 Numerical experiments

To illustrate the proposed methods we carry out numerical experiments on simulated *waveform* data and a real protein signaling data set *TCDB-SwissProt*. We compare the performance of the three methods (C-patra/sen, C-roc and ROC) discussed in Section 2.3 and the SPY method. With the SPY method, once the classifications ("positive" or "negative") in set U are made, we use the proportion of "negative" cases as an approximation of α_0. For the C-roc and C-patra/sen methods (Breiman, 2001), we use Random Forest to construct $C_n(\cdot)$.

2.4.1 Waveform data

We simulate observations from the *waveform* data set using the R-package *mlbench* (Leisch and Dimitriadou, 2010). The *waveform* data is a binary classification problem with 21 features. We fix $\pi = 0.5$ for all simulations.

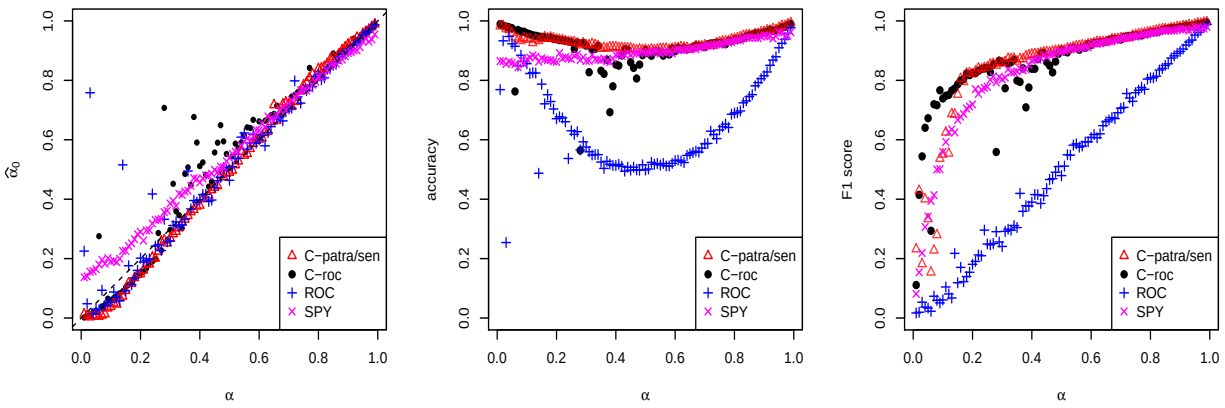


Figure 2.2: Comparison of methods with different α values. On the x-axis, α varies from 0.01 to 0.99 by step size 0.01. The left plot displays the estimates of the lower bound α_0. The middle plot displays the accuracy of classifying observations in U. The right plot displays the F1 score of the classifications.

2.4.1.1 *Varying* α

We vary α from 0.01 to 0.99 in Model (1.1) in increments of 0.01. For each α the sample sizes are fixed at $m = n = 3000$. At each α we run the methods described to estimate α and classify observations in U. Results are shown in figure 2.2. For α estimation shown in the left panel, the ROC method performs well when α is large, but overestimates α when it is near zero. If α is small, the ROC method is sensitive to the random seed used to divide samples into training and testing sets. The SPY method depends on a good choice of noise level, so with misspecified noise level it usually overestimates or underestimates α. C-roc and C-patra/sen methods are more stable with small α.

2.4.1.2 *Varying sample size*

We empirically examine consistency and convergence rates of the methods by estimating α at increasing sample sizes, keeping the number of labeled and unlabeled observations equal, i.e. $n = m$. In figure 2.3, every method is repeated 20 times for each (n, α) pair. The 20 α_0 estimates are displayed as a boxplot, which show estimator bias and variance.

We see that 1) all methods, except SPY, appear consistent under different settings ($\alpha = 0.1, 0.5, 0.9$); 2) the ROC estimator has the largest variance; 3) with larger α, the estimators have smaller variance 4) C-patra/sen and C-roc are the best methods on average.

2.4.1.3 *Single feature* α_0 *estimation*

One approach to solving the multidimensional PU learning problem is to estimate α separately using each feature. If $X_i \in \mathbb{R}^p$, this results in p estimates $\widehat{\alpha}_0^1, \ldots, \widehat{\alpha}_0^p$ of the parameter α. Each of these is an estimated lower bound on α. Thus a naive estimate of α_0 is $\max(\widehat{\alpha}_0^1, \ldots, \widehat{\alpha}_0^p)$. This approach ignores the correlation structure among features.

Using the *waveform* data, we compare this strategy to the multi–dimensional classifier approach. To make the problem challenging we select the 14 weakest features, defined as having the lowest Random Forest importance scores. We apply the Patra–Sen one–dimensional method to obtain individual feature α_0 estimates. The results are summarized in figure 2.4. Feature impor-

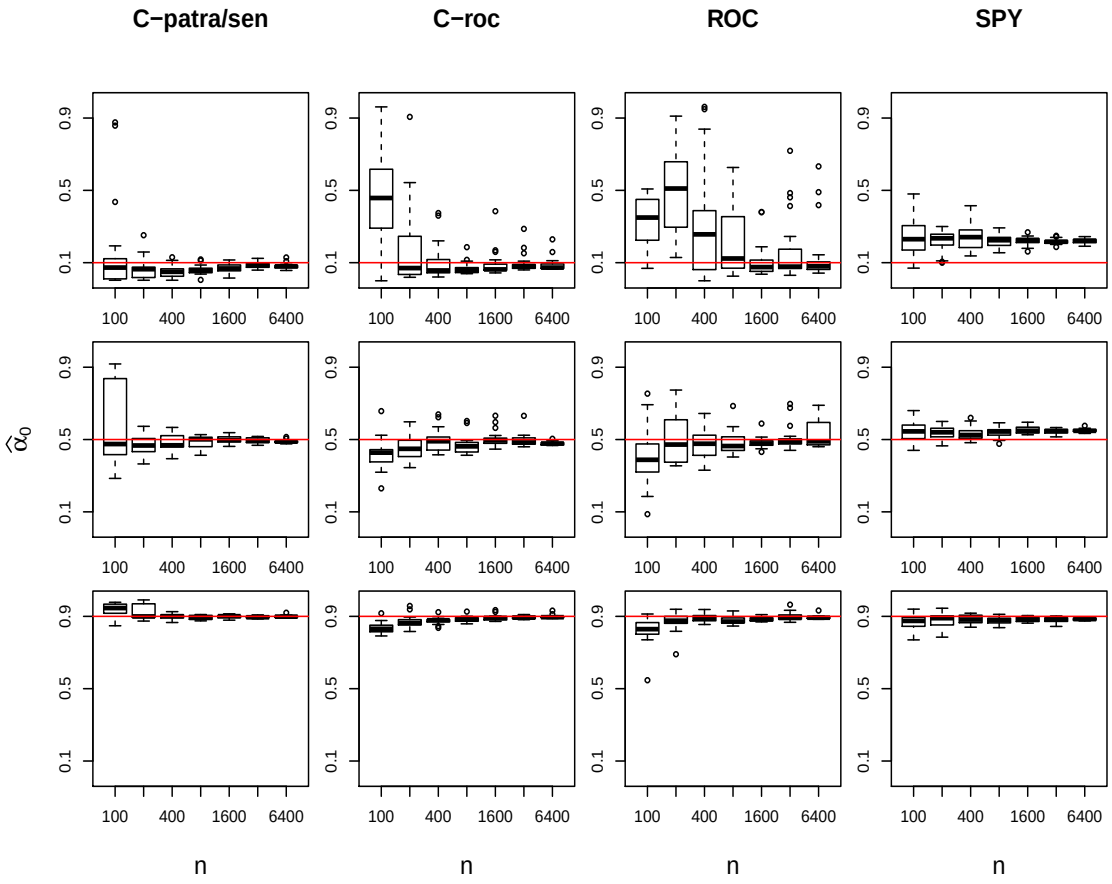


Figure 2.3: Comparison of methods with different sample sizes. The red solid line represents the true α (0.1,0.5,0.9). The range for all y-axes is [0, 1] from bottom to top. The unlabeled sample size n varies with $100 \times 2^j (j = 0, \ldots, 6)$. Each boxplot summarizes 20 repeated estimates $\widehat{\alpha}_0$ for each (n, α) pair.

tance matches well with the performance of the α estimates. On the right panels of Figure 2.4, we see that feature 5 is not useful because there is little difference between the unlabeled and labeled samples, leading to a feature based α estimate of approximately 0.012. In contrast, feature 8 is better in that it gives an alpha estimate of approximately 0.542. The SPY, C-roc, and C-patra/sen methods all perform better than the individual feature estimates (upper left of figure 2.4).

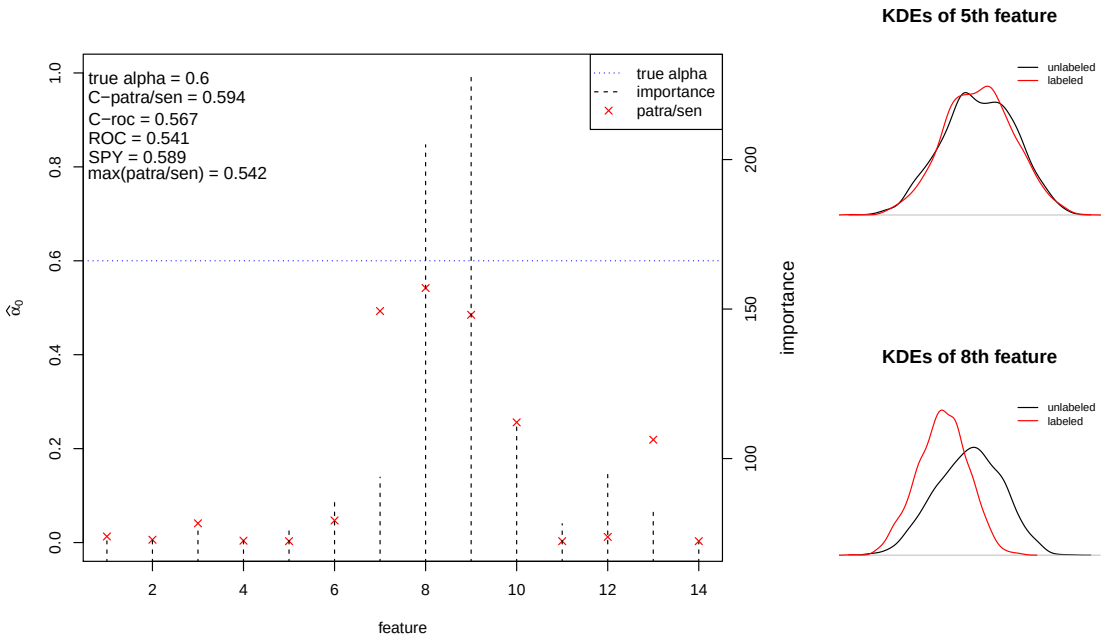


Figure 2.4: Estimation of α_0 using individual features. In the left panel the horizontal blue dash line is the true α ($= 0.6$), the vertical black dashed lines are the feature importances (right y–axis), and the red cross symbol are the α_0 estimate using the Patra/Sen procedure on a single feature (left y–axis). The right panels are kernel density estimates of "unlabeled" and "labeled" data for features 5 and 8.

2.4.2 Protein signaling

The transporter classification database (TCDB) (Saier et al., 2006), here the P set, consists of 2453 proteins involved in signaling across cellular membranes. It is desirable to add proteins to this database from unlabeled databases which contain a mixture of membrane transport and non–transport proteins. Elkan and Noto (2008) and Das et al. (2007) manually identified 348 of the 4906 proteins as being related to transport in the SwissProt (Boeckmann et al., 2003) database. We treat the SwissProt data as the unlabeled set U for which we have ground truth $\alpha = (4906-348)/4906 \approx 0.929$. Information from protein description documents are used as features including function, subcellular location, alternative products, and disease. In total there are 741 features. PCA is used on 741 features to obtain 200 new features that explain about 94% of variation.

Table 2.1 contains results of applying the four methods to estimate α and classify unlabeled observations. In the *ideal* column, $\alpha = 0.93$ is the true proportion of negative samples within the unlabeled set. The accuracy (0.99) and F1 score (0.92) in the *ideal* column are calculated using 10–fold cross-validation with all the of positive examples (in TCDB and SwissProt) against only the negative examples in SwissProt. This represents an upper bound on the performance one could expect for the PU learning methods. The second through fifth columns (C-patra/sen through SPY) contain results of the four methods discussed in this work. C-roc has the best performance among the methods.

	ideal	C-patra/sen	C-roc	ROC	SPY
alpha	0.93	0.96	0.94	0.96	0.89
accuracy	0.99	0.95	0.96	0.89	0.94
F1 score	0.92	0.54	0.68	0.04	0.66

Table 2.1: Comparison of methods for protein signaling data.

3. PERIODS ESTIMATION FOR MIRAS USING MULTI-BAND LIGHT CURVES AND INVERSE PERIOD-LUMINOSITY RELATIONS

In this chapter, we introduce a hierarchical Bayesian model to estimate periods for Miras using multi-band light curves and inverse Period-Luminosity relations. Our proposed model generalizes existing period estimation methods including those in He et al. (2016) and Yuan et al. (2018). Current existing models for period estimation are only designed for an individual single-band or multi-band light curve. In contrast, our model uses the inverse Period-Luminosity relations to borrow strength across different light curves that share common Period-Luminosity relations. We will demonstrate the power of our proposed model in two simulation experiments.

The rest of this chapter is organized as follows. Section 3.1 reviews the existing models to estimate period of Mira. Section 3.2 presents an iPLR-enhanced multi-band model for a set of light curves. Section 3.3 compares our model with other methods in two simulation experiment. An application to a set of real Miras light curves is also available in section 3.4.

3.1 A review of existing methods for period estimation

In this section, we review some popular existing methods for period estimation. In summary, most existing methods follow the procedure below: firstly, build a regression model $y = f(t, \sigma|\omega, \theta)$, where y is a vector of magnitude, t is a vector of observation time, σ is a vector of measurement error, ω is the frequency (i.e. reciprocal of period p) parameter and θ is a vector contains other parameters; secondly, do a grid search on ω, i.e. for each fixed candidate ω_j in the model, $\widehat{\theta}$ is estimated and a log-likelihood l_j based on $(\omega_j, \widehat{\theta})$ is calculated; thirdly, a periodogram is plotted in a way log-likelihood l_j against on frequency ω_j; finally, a period estimate is given $\widehat{p} = 1/\widehat{\omega}$, where $\widehat{\omega}$ is a frequency that maximizes the log-likelihood in the periodogram plot.

The reason that a grid search on frequency is performed is that there is no closed-form expression for the frequency and the multi-modality in periodogram makes it difficult to directly optimize on the frequency parameter. We will see the multi-modality property in later sections. All methods

fall into two categories: single-band and multi-band.

3.1.1 Single-band models

In this subsection, we denote a single-band light curve data as $\{t_i, y_i, \sigma_i\}_{i=1}^n$, where t_i is the observation time, y_i is the magnitude, σ_i is the measurement error associated to the magnitude.

3.1.1.1 Lomb-Scargle and generalized Lomb-Scargle

The most famous single-band period estimation approach is Lomb-Scargle (LS) (Lomb, 1976; Scargle, 1982). The LS model firstly centralizes the mean of magnitudes to be zero and then models the centralized magnitude as a sinusoid plus measurement error:

$$\tilde{y}_i = a\sin(2\pi\omega t_i + \phi) + \sigma_i\epsilon_i, \tag{3.1}$$

where $\tilde{y}_i = y_i - \frac{1}{n}\sum_{i=1}^n y_i$ is the mean-centered magnitude, $\epsilon_i \sim N(0,1)$ is the standard Gaussian variable, ω is the frequency, a is the amplitude and ϕ is the phase. Zechmeister and Kürster (2009) create a generalized Lomb-Scargle (GLS) approach by adding a floating mean to the equation (3.1) and model on the original magnitudes:

$$y_i = m + a\sin(2\pi\omega t_i + \phi) + \sigma_i\epsilon_i, \tag{3.2}$$

where m is the floating mean. The LS and GLS models can be solved straightforwardly via a series of least-squares fitting along a set of candidate frequencies. We illustrate the idea of the GLS model here. Let $\beta_1 = a\cos(\phi)$ and $\beta_2 = a\sin(\phi)$, then the equation (3.2) becomes

$$y_i = m + \beta_1\sin(2\pi\omega t_i) + \beta_2\cos(2\pi\omega t_i) + \sigma_i\epsilon_i. \tag{3.3}$$

The related Error Sum of Squares (SSE) function is

$$l(\omega, m, \beta_1, \beta_2) = \sum_{i=1}^n \left(\frac{y_i - m - \beta_1\sin(2\pi\omega t_i) - \beta_2\cos(2\pi\omega t_i)}{\sigma_i}\right)^2. \tag{3.4}$$

For a fixed ω, we can minimize the SSE function with respect to parameters (m, β_1, β_2):

$$l(\omega) = \min_{m,\beta_1,\beta_2} l(\omega, m, \beta_1, \beta_2). \tag{3.5}$$

Hence the MLE for ω is

$$\widehat{\omega} = \arg\min_{\omega} l(\omega). \tag{3.6}$$

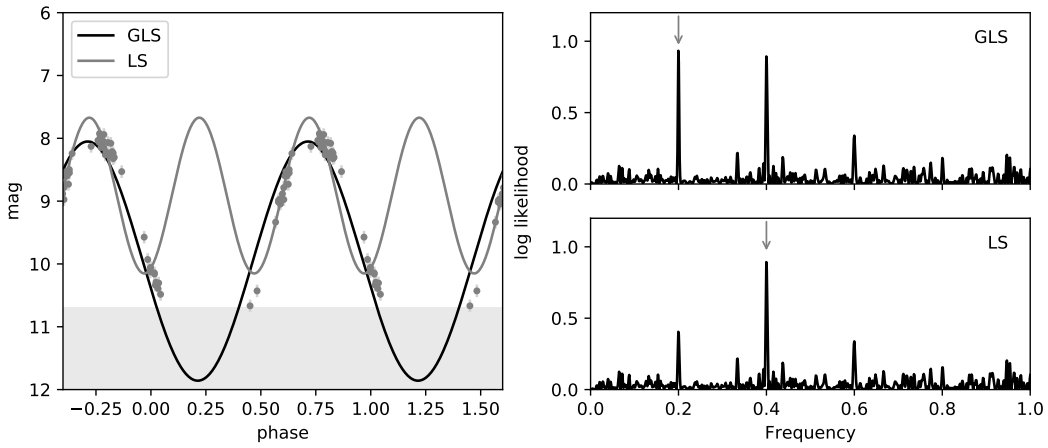


Figure 3.1: A comparison of the LS and GLS methods for data with a true frequency of 0.2 and a selection filter which removes faint observations (mag > 10.7). The GLS approach correctly recovers the true frequency of 0.2, while the LS method fails to recover frequency with estimate frequency as 0.4. This figure is adapted based on the work of VanderPlas (2018).

The difference between the LS and GLS models is illustrated in figure 3.1: the data contains 52 noisy observations of a sinusoidal signal whose faintest observations (i.e. mag > 10.7) have been omitted. In this case, the mean (used to center data) estimated from the LS approach is not close to the true mean (mag = 10), making the LS model to fail to recover the true frequency (lower panel of figure 3.1). The GLS model is more flexible than the LS model. Therefore, when comparing models in section 3.3, we only use the GLS model.

Notice that the sinusoid in the LS and GLS models could be replaced by other shapes like

splines, wavelets (Foster, 1996), Nadaraya-Watson estimator (Hall et al., 2000), Gaussian process with periodic kernel (Wang et al., 2012) or more complicated templates (Sesar et al., 2010, 2017). The template-based approach usually has good performance for stars like RR Lyrae because those templates are particularly obtained for that type of star. This template-based method is beyond the range of this project, and it is usually computationally inefficient and has no versatility.

3.1.1.2 Semi-parametric

He et al. (2016) model the quasi-periodic light curves directly using semi-parametric Gaussian process model (hereafter, SP1). The idea of the SP1 model is quite similar to that by Wang et al. (2012) but the main difference is that: the later paper uses a periodic kernel like sinusoid to serve as a flexible curve shape to the light curve, while the SP1 model uses a periodic term to model the global periodic signal and plus a Gaussian process term to model the stochastic variation. The stochastic variation is the fundamental challenge for quasi-periodic light curves.

The SP1 approach models the magnitude as

$$y_i = m + \beta_1 \sin(2\pi\omega t_i) + \beta_2 \cos(2\pi\omega t_i) + h(t_i) + \sigma_i \epsilon_i, \tag{3.7}$$

where $h(t)$ is a smooth function to model stochastic variation. They assume $h(t)$ belongs to a reproducing kernel Hilbert space $\mathcal{H}$ with norm $\|\cdot\|_{\mathcal{H}}$ and reproducing kernel $K(\cdot,\cdot)$. The negative log-likelihood function is

$$l(\omega, m, \beta_1, \beta_2) = \sum_{i=1}^{n} \left(\frac{y_i - m - \beta_1 \sin(2\pi\omega t_i) - \beta_2 \cos(2\pi\omega t_i) - h(t_i)}{\sigma_i} \right)^2, \tag{3.8}$$

To model y_i in a semi-parametric way, a penalized term of $h(t)$ should be added to prevent over-fitting. Then the object function is

$$l(\omega, m, \beta_1, \beta_2) + \lambda \|h(\cdot)\|_{\mathcal{H}}^2, \tag{3.9}$$

where λ is a regularization parameter. From the perspective of Bayesian statistics, the above regu-

larization approach (equation (3.9)) is equivalent to imposing a Gaussian process prior on the function $h(t)$ (Rasmussen and Williams, 2005). A real-valued continuous stochastic process $\{h(t)\}$ is a Gaussian process ($\mathcal{GP}$) if vector $(h(t_1), \ldots, h(t_n))$ is a multivariate Gaussian distribution for every finite set of positions $(t_1, \ldots, t_n)$. Hence the final model of SP1 is

$$y_i|m, \boldsymbol{\beta}, h(t_i), \omega \sim \mathcal{N}(m + \mathbf{b}_\omega(t)^T\boldsymbol{\beta} + h(t_i), \sigma_i^2), \tag{3.10}$$

$$m \sim \mathcal{N}(m_0, \sigma_m^2), \tag{3.11}$$

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma_\beta^2\mathbf{I}), \tag{3.12}$$

$$h(t)|\boldsymbol{\theta} \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}}(t, t')), \tag{3.13}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$ and ω are fixed parameters, kernel in equation (3.13) takes a squared exponential form $k_{\boldsymbol{\theta}}(t, t') = \theta_1 \exp\left(-\frac{(t-t')^2}{2}\theta_2\right)$, $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$ and $\mathbf{b}_\omega(t) = (\sin(2\pi\omega t), \cos(2\pi\omega t))^T$.

The SP1 model (3.10)–(3.13) is solved via a hybrid of Bayesian and frequentist approaches. $\widehat{\omega}$ is obtained with a profile likelihood method. For fixed ω, hyper-parameter $\boldsymbol{\theta}$ is estimated as $\widehat{\boldsymbol{\theta}}_\omega$ with MLE, then with plug-in $\widehat{\boldsymbol{\theta}}_\omega$ in the likelihood function, a profile MLE $\widehat{\omega}$ is found by a grid search. For details of the algorithm, we refer the interested readers to He et al. (2016) and Rasmussen and Williams (2005).

3.1.2 Multi-band models

In this subsection, we denote a multi-band light curve data as $\{\{t_{bi}, y_{bi}, \sigma_{bi}\}_{i=1}^{n_b}\}_{b=1}^{B}$, where t_{bi} is the observation time, y_{bi} is the magnitude, σ_{bi} is the measurement error associated to the magnitude, the subscript b represents which band the data point belongs to.

3.1.2.1 Multi-band GLS and penalized GLS

Modern astronomical surveys are paying increasing attention to combining multi-band information, but few methods have been developed to do so. Typically and naively, practitioners apply single-band procedures individually to each band and then determine which period estimates to use based on some *ad hoc* criteria (Watkins et al., 2009). Süveges et al. (2012) use principal com-

ponent analysis (PCA) to combine bands together and then apply the GLS approach. Their method requires the observation time points to be the same across different bands. Long et al. (2016) and VanderPlas and Ivezić (2015) develop a multi-band generalized Lomb-Scargle (MGLS) approach that systematically combines different bands in one model. The MGLS approach models magnitudes as:

$$y_{bi} = m_b + a_b \sin(2\pi\omega t_i + \phi_b) + \sigma_{bi}\epsilon_{bi}, \tag{3.14}$$

and log likelihood function is

$$l(\omega, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\phi}) = \sum_{b=1}^{B}\sum_{i=1}^{n_b}\left(\frac{y_{bi} - m_b - a_b\sin(2\pi\omega t_i + \phi_b)}{\sigma_i}\right)^2, \tag{3.15}$$

where $\boldsymbol{m} = (m_1, \ldots, m_B)^T$, $\boldsymbol{a} = (a_1, \ldots, a_B)^T$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_B)^T$. The frequency estimate is given as $\widehat{\omega} = \arg\min_\omega l(\omega)$, where $l(\omega) = \min_{\boldsymbol{m},\boldsymbol{a},\boldsymbol{\phi}} l(\omega, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\phi})$. In summary, the MGLS model integrates multi-band information by assuming period is shared across different bands and it's solved by weighted least squares.

Long et al. (2016) also propose a penalized generalized Lomb-Scargle (PGLS) method. They penalize the amplitude and phase on the log-likelihood function (3.15) as

$$l_p(\omega, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\phi}|\gamma_1, \gamma_2) = l(\omega, \boldsymbol{m}, \boldsymbol{a}, \boldsymbol{\phi}) + \gamma_1 J_1(\boldsymbol{a}) + \gamma_2 J_2(\boldsymbol{\phi}), \tag{3.16}$$

where J_1 and J_2 are some L_2-norm functions. In (3.16), parameters γ_1 and γ_2 are fixed and need to be specified before fitting.

3.1.2.2 Multi-band semi-parametric

Yuan et al. (2018) generalize the SP1 model to a multi-band model via a series of amplitude- and phase- relations between different bands. We call their model as SP2 for later reference. The

SP2 model is

$$y_{bi}|m_b, \boldsymbol{\beta}_b, h_b(t_{bi}), \omega \sim \mathcal{N}(m_b + \mathbf{b}_\omega^b(t)^T \boldsymbol{\beta}_b + h_b(t_{bi}), \sigma_{bi}^2), \tag{3.17}$$

$$m_b \sim \mathcal{N}(m_{b0}, \sigma_{bm}^2), \tag{3.18}$$

$$\boldsymbol{\beta}_b \sim \mathcal{N}(\mathbf{0}, \sigma_{b\beta}^2 \mathbf{I}), \tag{3.19}$$

$$h_b(t)|\boldsymbol{\theta}_b \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}_b}(t,t')\delta(t,t')), \tag{3.20}$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2)$, $k_{\boldsymbol{\theta}}(t,t') = \theta_1 \exp\left(-\frac{(t-t')^2}{2}\theta_2\right)$, $\boldsymbol{\beta}_b = (\beta_{b1}, \beta_{b2})^T$, $\mathbf{b}_\omega^b(t) = \frac{A_b}{A_0}(\sin(2\pi\omega t - \Delta\Phi_{b0}), \cos(2\pi\omega t - \Delta\Phi_{b0}))^T$, and $\delta(t,t')$ is defined as

$$\delta(t_{bi}, t_{b'i'}) = \begin{cases} 1 & \text{if } b = 0 \text{ and } b' = 0, \\ 1 & \text{if } b \neq 0 \text{ and } b' \neq 0, \\ 0 & \text{if otherwise.} \end{cases} \tag{3.21}$$

Here band $b = 0$ is considered as the reference. So A_b/A_0 is the amplitude ratio between bth band and the reference band, while $\Delta\Phi_{b0}$ is the phase difference between bth band and the reference band. Hyper-parameters A_b/A_0 and $\Delta\Phi_{b0}$ are learnt from other well-studied surveys of galaxies like the LMC. The role of Equation (3.21) is to force non-reference bands to share a covariance structure. The application of the SP2 model in M33 Miras is quite successful, but it is not straightforward to adapt their model to a general case.

Table 3.1 summarizes the aforementioned methods. The main contribution of this project is to propose a model that makes an improvement over the SP1 and SP2 models by using the information of (inverse) PLR. Note that for a set of multi-band Mira light curves, the SP1 model only fits one band for each light curve, while the SP2 model fits all bands data but each light curve is fitted independently. Our proposed model can fit all light curves simutaneously and hence the information borrowed across light curves can help greatly improving the performance in recovering the periods.

type	band	name	reference
strict-periodic	single-	LS	Lomb (1976); Scargle (1982)
		GLS	Zechmeister and Kürster (2009)
	multi-	MGLS	Long et al. (2016); VanderPlas and Ivezić (2015)
		PGLS	Long et al. (2016)
		Watkins	Watkins et al. (2009)
		Süveges	Süveges et al. (2012)
quasi-periodic	single-	SP1	He et al. (2016)
	multi-	SP2	Yuan et al. (2018)

Table 3.1: Some existing period estimation methods.

3.2 Inverse Period-Luminosity relation enhanced multi-band semi-parametric model

In this section, we aim at developing a multi-band semi-parametric model for a set of light curves which share the same PLR. To use the information of PLR, we propose the inverse Period-Luminosity Relations which treat period as a function of luminosity. The luminosity works as a covariate to the light curve, and hence the period can be roughly guided by luminosity via iPLR. The multi-band model enhanced by the iPLR is called iPLR-enhanced multi-band semi-parametric model, which is called SP3 for short.

3.2.1 The SP3 model

Denote a set of light curves as $\mathcal{D} = \{D_l, m_l\}_{l=1}^{L}$, where $D_l = \{\{(t_{lbi}, y_{lbi}, \sigma_{lbi})\}_{i=1}^{n_{lb}}\}_{b=1}^{B}$ is a light curve and m_l is luminosity of the interested band for D_l. Suppose we are interested in K_s band PLR, then approximately $m_l \approx \frac{1}{2}\left(\min_i(y_{lK_si}) + \max_i(y_{lK_si})\right)$ as discussed in Yuan et al. (2017a). Here we treat the luminosity m_l as a covariate to D_l. In astronomy, a PLR for Miras could take different forms. For example, Ita et al. (2005) use linear relations with break at about period equals to 400 days, while Yuan et al. (2017a) use a quadratic form:

$$m_l = a_0 + a_1 \log_{10} p_l + a_2 \log_{10}^2 p_l + \sigma\epsilon_l, \tag{3.22}$$

where p_l is the period of l-th light curve, σ is standard deviation, and $l = 1, \ldots, L$. In this project we assume such kind of PLR exists for the set of light curves $\mathcal{D}$. Additionally, we assume the PLR of interest has a quadratic form. Those PLRs of linear form can be easily adapted in our model.

Directly fitting equation (3.22) is difficult due to the large uncertainty of period estimates. Note that frequency is the inverse of period and the function of PLR is monotonic, from (3.22) we see that PLR can be used to control the location of frequency in the magnitude-frequency space. Instead of using (3.22) to govern a series of light curves, we use iPLR:

$$\log \omega_l = a_0 + a_1 m_l + a_2 m_l^2 + \tau \epsilon_l = a^T d(m_l) + \tau \epsilon_l, \tag{3.23}$$

where $a := (a_0, a_1, a_2)^T$, $d(m_l) := (1, m_l, m_l^2)^T$ and τ is the standard deviation. We call it "inverse" since it treats frequency as a function of luminosity. In astronomical domain, it is natural to use base of 10 for the logarithm of period, but we will use the base of e in this project since it is more convenient to use in statistical computation.

The classical way to calculate PLRs is a two-step procedure, in which frequency ω_l is estimated for each light curve D_l, and then a quadratic regression is fitted to estimate a and the overall residual standard deviation σ. The quality of PLRs or iPLRs calculation highly depends on the accuracy of frequency estimations. This procedure has a major limitation that the frequency ω_l is estimated by setting a wild-guess initial value or doing a grid search over a large range. Either way, a wild-guess or a grid search, would be a waste of time calculating "alias" the fake frequencies. Suppose we have obtained frequency estimates $\{\widehat{\omega}_l\}_{l=1}^L$ from the first step, then the second step to calculate iPLR is to fit the model $\log \widehat{\omega}_l \sim \mathcal{N}(a^T d(m_l), \tau^2)$, i.e.

$$\widehat{\omega}_l \sim \text{LogNormal}(a^T d(m_l), \tau^2). \tag{3.24}$$

The equation (3.24) gives us the insight how to specify the prior for frequency of each light curve.

Based on the SP1 and SP2 models, we propose the SP3 model as

$$y_{lbi}|m_{lb}, \boldsymbol{\beta}_{lb}, h_{lb}(t_{lbi}), \omega_l \sim \mathcal{N}(m_{lb} + \mathbf{b}_{\omega_l}(t)^T \boldsymbol{\beta}_{lb} + h_{lb}(t_{lbi}), \sigma^2_{lbi}), \tag{3.25}$$

$$m_{lb} \sim \mathcal{N}(m_{lb0}, \sigma^2_{lbm}), \tag{3.26}$$

$$\boldsymbol{\beta}_{lb} \sim \mathcal{N}(\mathbf{0}, \sigma^2_{lb\beta}\mathbf{I}), \tag{3.27}$$

$$h_{lb}(t)|\boldsymbol{\theta}_{lb} \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}_{lb}}(t, t')), \tag{3.28}$$

$$\theta_{lb1} \sim \text{LogNormal}(0, \theta_0), \tag{3.29}$$

$$\theta_{lb2} \sim \text{LogNormal}(0, \theta_0), \tag{3.30}$$

$$\omega_l \sim \text{LogNormal}(a^T d(m_l), \tau^2), \tag{3.31}$$

where a and τ^2 both have non-informative prior s.t. $p(a) \propto 1$ and $p(\tau^2) \propto 1/\tau^2$, respectively. Equations (3.25)–(3.28) are just the multi-band version of the SP1 model; equations (3.29) and (3.30) are convenient priors that make computation easy, and the reason will be explained in next subsection; equation (3.31) is based on equation (3.24). Figure 3.2 shows the diagram of the SP3 model.

3.2.2 Computation of the SP3 model

There is a computation issue in directly sampling a from the SP3 model. The posterior can be represented as

$$p(a, \tau, \boldsymbol{\omega}, \boldsymbol{m}, \boldsymbol{\beta}, \boldsymbol{\theta}|\mathcal{D}), \tag{3.32}$$

where $\boldsymbol{\omega} = (\omega_l : l = 1, \ldots, L)$, $\boldsymbol{m} = (m_{lb} : l = 1, \ldots, L; b = 1, \ldots, B)$, $\boldsymbol{\beta} = (\beta_{lbj} : l = 1, \ldots, L; b = 1, \ldots, B; j = 1, 2)$, and $\boldsymbol{\theta} = (\theta_{lbj} : l = 1, \ldots, L; b = 1, \ldots, B; j = 1, 2)$. In total there are $5BL + L + 4$ parameters in (3.32). For example, if we use the information from two bands (i.e. $B = 2$) and 1000 light curves (i.e. $L = 1000$), then we have 11004 parameters. The MCMC for the SP3 model would require large memory and long computation time. Below we will introduce an algorithm that avoids a large MCMC sampling.

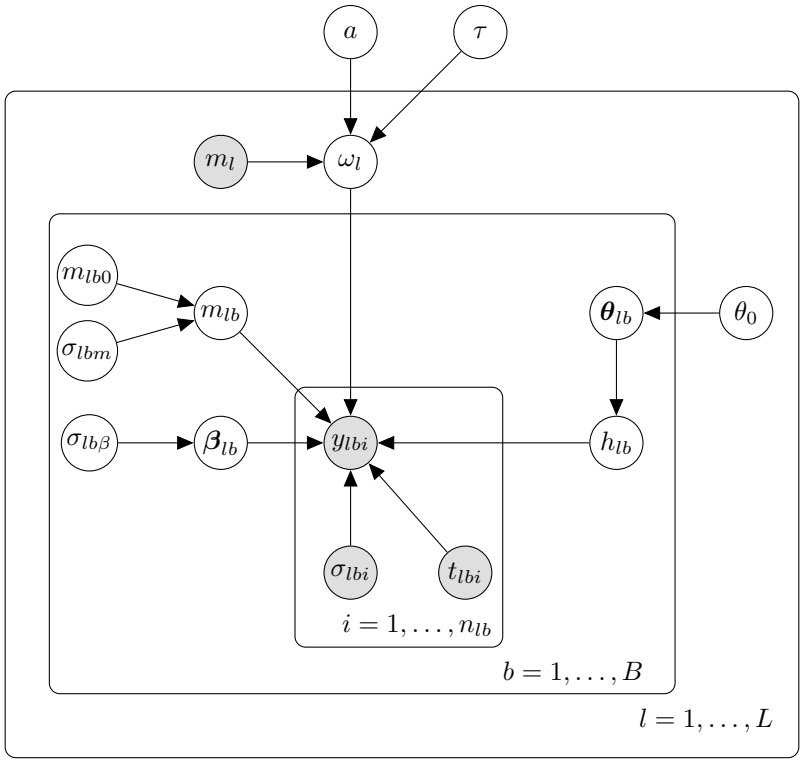


Figure 3.2: Diagram of the SP3 model.

Note that by integrating out m_b and $\boldsymbol{\beta}_b$, we can simplify equations (3.25)–(3.28) to

$$\boldsymbol{y}_{lb}|\omega_l, \boldsymbol{\theta}_{lb} \sim \mathcal{N}(m_{lb0}\mathbf{1}, K_{lb}(\boldsymbol{t}_{lb}, \boldsymbol{t}_{lb}|\omega_l, \boldsymbol{\theta}_{lb})), \tag{3.33}$$

where $\boldsymbol{y}_{lb} := (y_{lb1}, \ldots, y_{lbn_{lb}})^T$ and

$$K_{lb}(\boldsymbol{t}_{lb}, \boldsymbol{t}_{lb}|\omega_l, \boldsymbol{\theta}_{lb}) = (\sigma^2_{lbm} + \sigma^2_{lb\beta}\mathbf{b}_{\omega_l}(t_{lbi})^T\mathbf{b}_{\omega_l}(t_{lbj}) + k_{\boldsymbol{\theta}_{lb}}(t_{lbi}, t_{lbj}) + \sigma^2_{lbi}\delta_{ij})_{n_{lb}\times n_{lb}}, \tag{3.34}$$

where $\boldsymbol{t}_{lb} := (t_{lb1}, \ldots, t_{lbn_{lb}})^T$.

Treat (a, τ) as the global parameters and $(\boldsymbol{\omega}, \boldsymbol{m}, \boldsymbol{\beta}, \boldsymbol{\theta})$ as local parameters. When a and τ have been fixed, then the SP3 model is reduced to be a series of independent models. Conditional on parameters a and τ, the model estimation for (3.29)–(3.34) can be carried out by Elliptical Slice Sampler (ESS) (Murray et al., 2010). ESS is widely applicable in cases where we have normal

priors. To use ESS technique, we firstly make some transformations on the priors (3.29)–(3.31):

$$\tilde{\theta}_{lbj} = \log \theta_{lbj}, \quad j = 1, 2 \tag{3.35}$$

$$\tilde{\omega}_l = \log \omega_l, \tag{3.36}$$

where $b = 1, \ldots, B$. Denote the parameters as $\Theta_l = (\tilde{\theta}_{l11}, \tilde{\theta}_{l12}, \ldots, \tilde{\theta}_{lB1}, \tilde{\theta}_{lB2}, \tilde{\omega}_l)$. Then Θ_l has distribution $\mathcal{N}(\mu(a^T d(m_l)), \Sigma(\tau))$, where $\mu(a^T d(m_l)) = (0, \ldots, 0, a^T d(m_l))$ and $\Sigma(\tau) = \text{diag}(\theta_0, \ldots, \theta_0, \tau^2)$. Denote the likelihood function as $L(\Theta_l)$. The detailed ESS algorithm for the Θ_l is given in the algorithm 1.

Algorithm 1: ESS update for Θ_l given (a, τ)

input : current state Θ_l
output: new state Θ_l'

$\nu \sim \mathcal{N}(\mu(a^T d(m_l)), \Sigma(\tau))$;
$u \sim$ Unif$[0, 1]$;
Let $LL = \log L(\Theta_l) + \log u$;
$\eta \sim$ Unif$[0, 1]$;
$[\eta_{min}, \eta_{max}] \leftarrow [\eta - 2\pi, \eta]$;
$\Theta_l' \leftarrow (\Theta_l - \mu(a^T d(m_l))) \cos\eta + (\nu - \mu(a^T d(m_l))) \sin\eta + \mu(a^T d(m_l))$;
while $\log L(\Theta_l') < LL$ **do**
    if $\eta < 0$ **then**
        $\eta_{min} \leftarrow \eta$
    else $\eta_{max} \leftarrow \eta$;
    $\eta \sim$ Unif$[0, 1]$;
    $\Theta_l' \leftarrow (\Theta_l - \mu(a^T d(m_l))) \cos\eta + (\nu - \mu(a^T d(m_l))) \sin\eta + \mu(a^T d(m_l))$;

On the other hand, when frequency estimates $\widehat{\boldsymbol{\omega}}$ have been obtained from each light curve, then estimation of (a, τ) is trivial. Based on this insight, we propose an iterative two-step procedure (algorithm 2) for estimating periods of the SP3 model.

In algorithm 2, initial values $a^{(0)}$ and $\tau^{(0)}$ can be specified in two ways. The first way is to obtain their values from another survey. For example, we can use the iPLR of LMC Miras as the initial

Algorithm 2: Conditional maximization algorithm for the SP3 model

1 Initialize $a^{(0)}$ and $\tau^{(0)}$;
2 Let $M = \left[d(m_1)^T; \cdots ; d(m_L)^T\right]$ be a $L \times 3$ design matrix;
3 Flag $\leftarrow$ True, $j \leftarrow 1$;
4 **while** Flag **do**
5 **for** $l = 1, \ldots, L$ **do**
6 Estimate $\omega_l^{(j-1)}$ by ESS algorithm 1 on light curve D_l given $(d(m_l)^T a^{(j-1)}, \tau^{(j-1)})$;
7 Let $W = \left(\log \omega_1^{(j-1)}, \ldots, \log \omega_L^{(j-1)}\right)^T$;
8 $a^{(j)} \leftarrow \left(M^T M\right)^{-1} M^T W$;
9 $\tau^{(j)} \leftarrow \sqrt{\frac{1}{L-3} W^T \left(I - M(M^T M)^{-1} M^T\right) W}$;
10 **if** $\tau^{(j-1)} < \tau^{(j)}$ or $\tau^{(j-1)} - \tau^{(j)} < \epsilon_0$ **then**
11 Flag = False;
12 $j \leftarrow j + 1$;

iPLR for M33 Miras. The second way is to use simple models like the LS method to calculate the initial iPLR. In this project we use the second way, since the LS method is computationally cheap and we do not need to rely on additional information. The M in Line 2 of algorithm 2 is the design matrix for the quadratic regression. In lines 5–6, each light curve is independently fitted by Algorithm 1. There is a huge computation advantage in lines 5–6, which can be done with parallel or distributed computing techniques. Lines 7–9 are the updates for global parameters a and τ by quadratic regression. Lines 10–11 are used as a criteria when to stop updating global parameters. The criteria is when either (1) the updated iPLR is not as tight as the previous one or (2) the tightness of iPLR can only be improved by a small value, then we stop updating. In practice, we set $\epsilon_0 = 0.01$ for the Miras.

This method, the Algorithm 2, is also called *conditional maximization* (Gelman et al., 2013): given initial a and τ, maximizing function (3.32) w.r.t. parameters Θ_l; given updated Θ_l, maximizing function (3.32) w.r.t. parameters (a, τ); alternately updating two sets of parameters for enough times. From the perspective of frequentist, this method is also termed as *block coordinate ascent*.

For the l-the light curve, it is easy to make prediction on new observation time vector $\boldsymbol{t}^*$ on

b-band, given estimated $\widehat{m}_{lb0}$, $\widehat{\boldsymbol{\theta}}_{lb}$ and $\widehat{\omega}_l$. Suppose the prediction is $\boldsymbol{y}^*$ of length n, then

$$\begin{pmatrix} \boldsymbol{y}^* \\ \boldsymbol{y}_{lb} \end{pmatrix} |\widehat{m}_{lb0}, \widehat{\boldsymbol{\theta}}_{lb}, \widehat{\omega}_l \sim \mathcal{N}\left(\begin{pmatrix} \widehat{m}_{lb0}\mathbf{1}_n \\ \widehat{m}_{lb0}\mathbf{1}_{n_{lb}} \end{pmatrix}, \begin{pmatrix} K_{lb,11} & K_{lb,12} \\ K_{lb,21} & K_{lb,22} \end{pmatrix} \right),$$

where

$$\begin{aligned} K_{lb,11} :=& K_{lb}(\boldsymbol{t}^*, \boldsymbol{t}^* | \widehat{\omega}_l, \widehat{\boldsymbol{\theta}}_{lb}), \\ K_{lb,12} =& K_{lb,21}^T := K_{lb}(\boldsymbol{t}^*, \boldsymbol{t}_{lb} | \widehat{\omega}_l, \widehat{\boldsymbol{\theta}}_{lb}), \\ K_{lb,22} :=& K_{lb}(\boldsymbol{t}_{lb}, \boldsymbol{t}_{lb} | \widehat{\omega}_l, \widehat{\boldsymbol{\theta}}_{lb}). \end{aligned}$$

Therefore prediction and variance are

$$E\left[\boldsymbol{y}^* | \boldsymbol{y}_{lb}, \widehat{m}_{lb0}, \widehat{\boldsymbol{\theta}}_{lb}, \widehat{\omega}_l\right] = \widehat{m}_{lb0}\mathbf{1}_n + K_{lb,12}K_{lb,22}^{-1}\left(\boldsymbol{y}_{lb} - \widehat{m}_{lb0}\mathbf{1}_{n_{lb}}\right), \tag{3.37}$$

$$Var\left[\boldsymbol{y}^* | \boldsymbol{y}_{lb}, \widehat{m}_{lb0}, \widehat{\boldsymbol{\theta}}_{lb}, \widehat{\omega}_l\right] = K_{lb,11} - K_{lb,12}K_{lb,22}^{-1}K_{lb,21}. \tag{3.38}$$

3.3 Simulations

In this section, two simulation experiments are conducted to compare the performance of our proposed method (SP3 using algorithm 2) with other existing methods including: generalized Lomb-Scargle (GLS, Zechmeister and Kürster (2009)), single-band semi-parametric method (SP1, He et al. (2016)) and its multi-band extension (SP2, Yuan et al. (2018)).

3.3.1 Simulation I: 90 sets of Mira light curves at distance of LMC

The third phase of OGLE survey (Udalski et al., 2008) results in the discovery of 1663 Miras (Soszyński et al., 2009) with a median of 466 photometric measurements per object. The observation time points of these light curves and their precision are quite high, and hence making their period estimation relatively easy. An example of such light curve is in figure 1.1(a). The Miras in the LMC galaxy of the OGLE survey have good period estimation (Ita et al., 2005; Soszyński

et al., 2007; Yuan et al., 2017b), so we can use their light curves to generate simulated light curves with different qualities.

He et al. (2016) use the semi-parametric Gaussian process model (called SP1.5 for short) to fit the OGLE-III light curves using I band. The SP1.5 model is slightly more complex than the SP1 model: the SP1 model only uses a squared exponential kernel, while the SP1.5 model uses an additional independent squared exponential kernel and a periodic kernel. The additional kernels of the SP1.5 model make the templates of I band more adaptive to the observational data. And thanks to the dense sampling of the I band data, the resulted templates have high quality. Yuan et al. (2017b) use above I band templates to generate other JHK_s bands templates. Take the J band for example. $(I(t) - J(t))$ is firstly predicted via a regression model with a group of light curve features. Then the J band template is taken as $J(t) = I(t) - (I(t) - J(t))$, where the first $I(t)$ term is the I band template. He et al. (2016) and Yuan et al. (2017b) also generate their own simulated Mira light curves. The simulated light curves in He et al. (2016) only have I band data, and hence they do not meet our need for the multi-band models. The simulated light curves in Yuan et al. (2017b) will be discussed in next subsection.

In this simulation, we only consider I and K_s bands of Oxygen-rich Miras. Assuming number of I band points (n_I) is no less than that of K_s band points (n_{K_s}), we consider 10 pairs of (n_{K_s}, n_I): (5,5), (5,10), (5,20), (5,30), (10,10), (10,20), (10,30), (20,20), (20,30) and (30,30). To mimic real observation time pattern, we have 3 different scenarios to specify $\{t_{bi}, b = I \text{ or } K_s, i = 1, \ldots, n_b\}$ as follows:

- Pattern 1: M33 observation times as show in figure 1.1 (b);
- Pattern 2: LMC seasonal observation times as show in figure 1.1 (a);
- Pattern 3: The observation times are uniformly distributed.

For pattern 1 and pattern 2, t_{bi} is drawn from real light curves with replacement and is added with a random shift in order to make every set of samples unique. With the observation time points being

fixed, the light curve magnitudes are generated from Mira templates (see He et al. (2016) and Yuan et al. (2017b)).

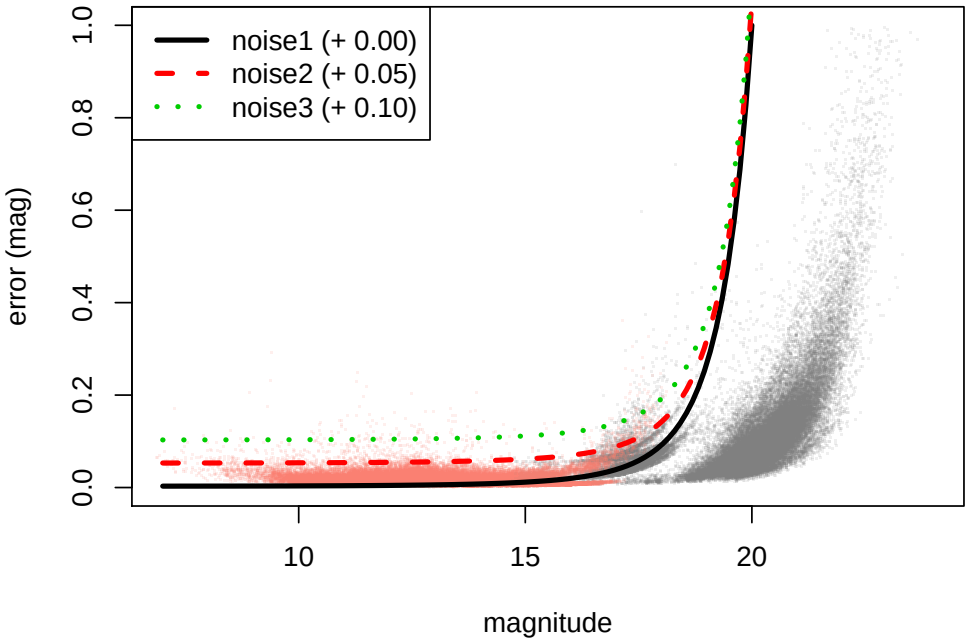


Figure 3.3: 3 noise levels for simulation I.

Another important consideration is measurement error or noise level. Note that the real noise level could be related to many factors like distance of the star, size of camera, exposure time, weather condition, light pollution, dust and so on. To make the simulation easy, we only use a simple exponential model, $\sigma(m) = \exp(a \cdot m^c - b)$, to simulate noise levels. As shown in figure 3.3, a base-line curve (black solid line) is $\sigma(m) = \exp\left(1.82 \times 10^{-6} m^5 - 5.84\right)$, which goes through points $(7, 0.003)$ and $(20, 1)$. The magnitude range from 7 to 20 roughly covers the magnitudes of LMC templates (salmon-color dots). This base line noise level curve is called "noise1" for later reference. To simulate other noise levels, we shift the "noise1" upward by 0.05 mag to create "noise2" (red dash line in figure 3.3) and by 0.1 mag to create "noise3" (green dot line in figure 3.3). The magnitude-noise pattern of M33 light curves (gray dots in figure 3.3) roughly follows this simple exponential model.

In total, we produce 90 ($= 10$ pairs of sample size $\times 3$ time patterns $\times 3$ noise levels) sets of simulated Oxygen-rich Miras, and each set consists of $L = 1000$ light curves with I and K_s bands. Figures B.1–B.3 each contains 9 simulated light curves for different patterns and different noise levels, with $(n_I = 20, n_{K_s} = 5)$, $(n_I = 20, n_{K_s} = 10)$ and $(n_I = 30, n_{K_s} = 30)$ respectively. More detail of the simulation procedure, code as well as the simulated light curves are all available on GitHub at https://github.com/zflin/mira.

3.3.1.1 Model comparisons

To evaluate the models, we can see how accurate a model recovers the frequencies for the set of light curves. Suppose $\{\omega_l\}$ are the ground truth frequencies and $\{\widehat{\omega}_l\}$ are the estimates. Then we can use root mean square error (RMSE) and "accuracy" (ACC) to compare different models. The RMSE is calculated as $\sqrt{\frac{1}{L}\sum_{l=1}^{L}\left(\widehat{\omega}_l - \omega_l\right)^2}$. The "accuracy" is another metric used in astronomy for variable stars. The ACC is calculated as $\frac{1}{L}\sum_{l=1}^{L}\text{Ind}\left(|\widehat{\omega}_l - \omega_l| \leq \lambda\right)$, where $\text{Ind}(\cdot)$ is index function and threshold λ is chosen for different purposes. For Miras, $\lambda =$ 2.7e-04, which is the average half distance to the sidelobes in the frequency spectra (He et al., 2016).

Three other models – GLS, SP1 and SP2 – are used to make comparison with our proposed SP3 model. The GLS and SP1 models are applied only on I band, which has more observation points than K_s band. Note that the GLS method is computationally fast, and hence we use the GLS result to give the initial value of $\left(a^{(0)}, \tau^{(0)}\right)$ for algorithm 2 to compute the SP3 model. Figure B.4 compares RMSE values of these four methods over 90 different settings. Similarly, figure B.5 compares ACC values. Overall, the proposed SP3 model outperforms others in all 9 different pattern-noise combinations.

3.3.2 Simulation II: a set of Mira light curves at distance M33

In subsection 3.3.1 we simulate 90 different sets of Mira light curves. In each set, every light curve roughly has same quality. In this subsection, we consider another simulated Mira data set, which consists of 5,000 O-rich Mira light curves and is constructed in the work of Yuan et al. (2018). These light curves are designed to mimic the real observed Mira stars in M33 galaxy.

Hence light curves in this set could have different quality, i.e. sample sizes and noise levels could be different from each other. As again, we only use I and K_s bands data and we are interested in the K_s band PLR, i.e. m_l represents the luminosity of K_s band. The performance of the GLS, SP1, SP2 and SP3 methods is summarized in table 3.2. The GLS and SP1 methods are applied on I band data, while the SP2 and SP3 methods are applied on I and K_s band data. Figure 3.4 compares the true frequencies with estimated frequencies in more details. By borrowing information from each light curve, the SP3 model efficiently excludes alias (fake frequency estimates) in a series of iterations as demonstrated by 3 bottom panels in figure 3.4.

	GLS	SP1	SP2	SP3
RMSE ($\times 10^{-4}$)	12.56	11.09	7.46	1.83
ACC (%)	72.02	78.66	91.26	96.02

Table 3.2: Simulation II. Comparison of methods for simulated 5,000 Mira light curves at the distance of M33 galaxy.

The iPLR is updated and improved at each iteration for the algorithm 2. This updating process for the Simulation II is demonstrated in figure 3.5. The results from the GLS model are used to calculate the initial iPLR, which has bias and wide confidence band. The iPLR becomes tighter and tighter, until 3rd updates when the criteria in algorithm 2 (line 10) holds. The improved iPLR at each iteration would in turn help to estimate periods more accurately.

3.3.2.1 An example of a light curve: prediction and local periodogram

To offer more details of the SP3 model, in this subsection we take a light curve (id=00080 in Simulation II) for example. Given global parameters a and τ estimated, each light curve can be modeled independently. For each light curve, we can do prediction using equation (3.37), as shown in figure 3.6. The gray band represents the 95% confidence band of the prediction. For this light curve, the estimated period is 171.20 compared to the true period 171.88 (days).

The periodogram is an important way to do diagnosis for the GLS, SP1 and SP2 models.

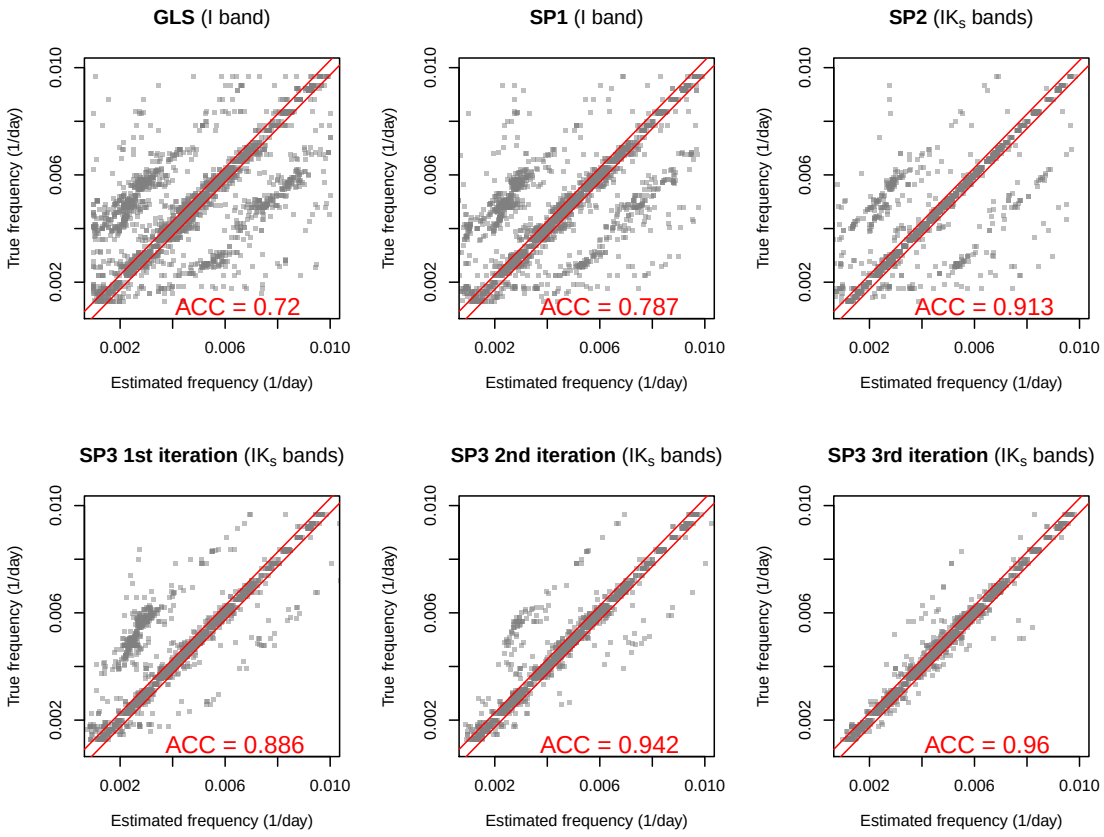


Figure 3.4: Simulation II. Comparison of the GLS, SP1, SP2 and SP3 models in periods recovering for 5,000 simulated M33 light curves.

Similarly, we can also have the periodogram for the SP3 model. But since we only softly constrain the grid search for frequency with a prior distribution, we would use the term *local periodogram* to distinguish from the traditional one. The periodogram from the SP1 model is given in figure B.6 the top-left panel. In the figure, the true frequency is marked by the vertical dash black line, while the estimated frequency is marked by the vertical dash green line for the model. Each blue line is corresponding to a green dot in the iPLR plot right next to each (local) periodogram plot. Without iPLR information, the SP1 method would incorrectly recover the period with an alias (= 352.11 days). After first iteration, the iPLR then forces the frequency prior (green solid line) to move towards the current frequency location.

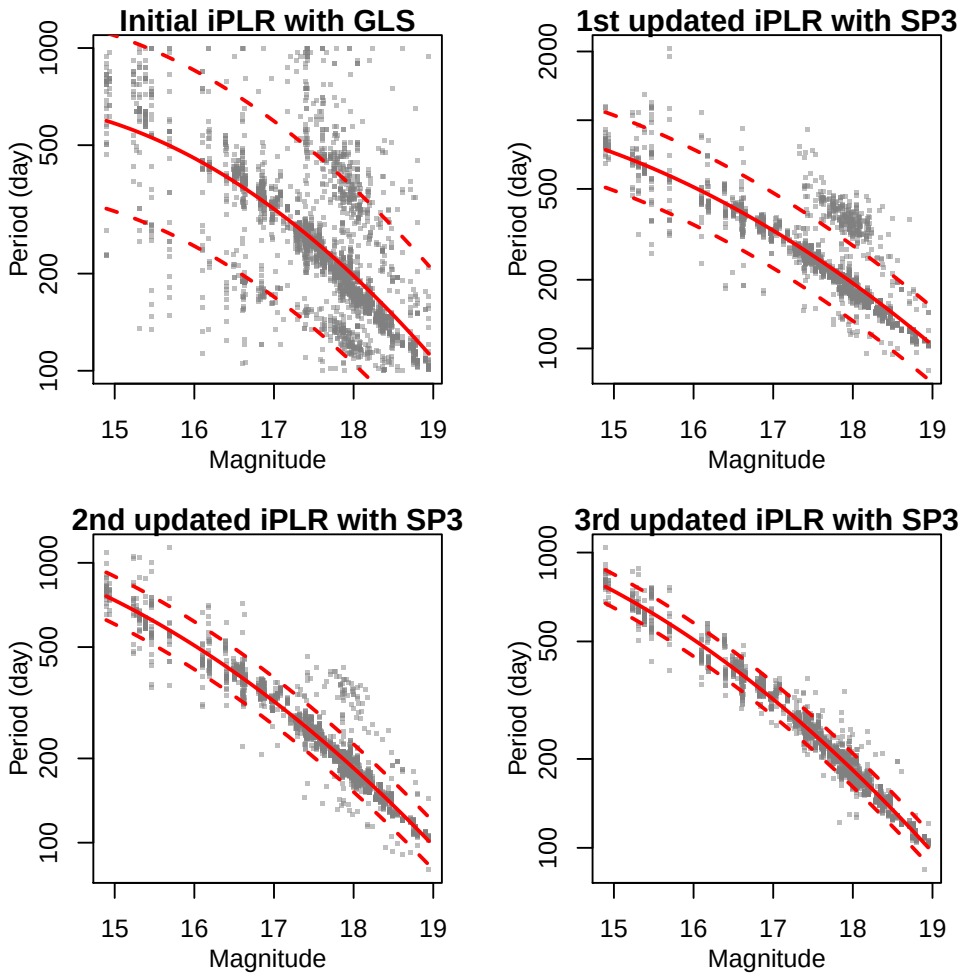


Figure 3.5: Simulation II. iPLR updates of the SP3 model for 5,000 simulated M33 light curves. In each panel, the red solid line is iPLR and the two red dash lines represent the 95% confidence interval band of the iPLR. The y-axis is for the estimated periods. The values of $\tau^{(j)}$ ($j = 0, 1, 2, 3$) for each iteration are 0.32, 0.19, 0.10 and 0.07 respectively.

3.4 Application to a set of real M33 Mira light curves

We apply the SP3 model to a real M33 dataset, which consists of 1,265 candidate O-rich Miras (Yuan et al., 2018). This dataset is collected from several sources: UKIRT (Khosroshahi et al., 2015), Kitt Peak National Observatory KPNO and DIRECT (Macri et al., 2001; Pellerin and Macri, 2011). Each star is measured in I, J, H and K_s bands, with the median number of observation each band being 68, 5, 6, 11, respectively. Yuan et al. (2018) use these Miras to determine the distance modulus for the M33 galaxy. Their periods are estimated by the SP2 method on all four bands,

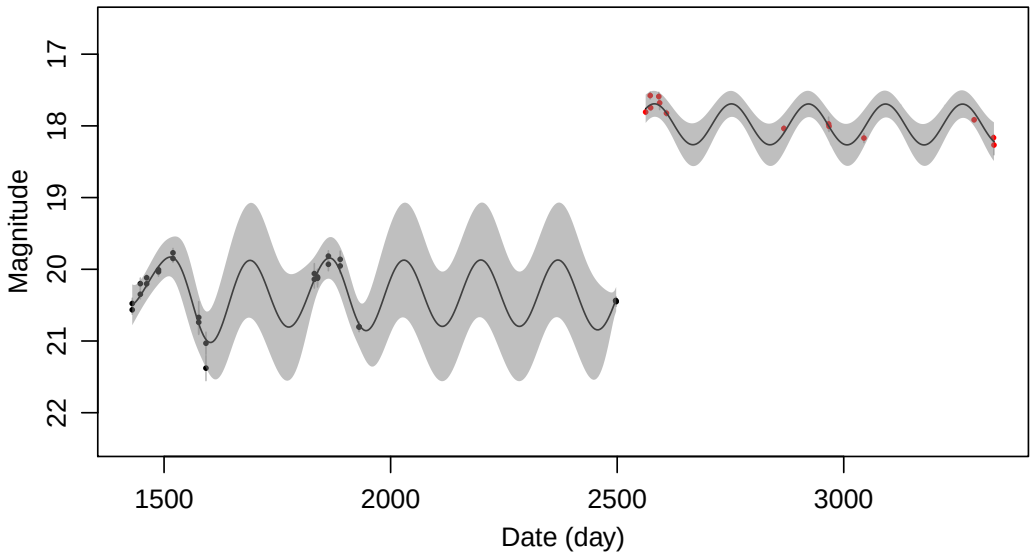


Figure 3.6: Simulation II. An example of fitted curve for a light curve (id=00080). The period estimated is 171.20 (days) compared to true period 171.88 (days).

and a further correction is made if the estimated period is far away from the PLR. The correction step is important to get tight PLRs in their work. Their estimated periods (after correction) are compared against the periods estimated by our proposed method in figure 3.7. The vertical error bars are calculated by bootstrapping for the SP2 model; the horizontal error bars are obtained by the standard deviation of ESS samples without outliers. An outlier is defined as an ESS sample point outside the range $[Q_1 - 1.5 \times \text{IQR}, Q_3 + 1.5 \times \text{IQR}]$, where Q_1 is first quantile, Q_3 is third quantile, and $\text{IQR} = Q_3 - Q_1$ is the interquartile range.

Following the convention of Yuan et al. (2018), a linear PLR is fitted for periods less than 400 days,

$$m_l = c_0 + c_1 \times (\log_{10} p_l - 2.3),$$

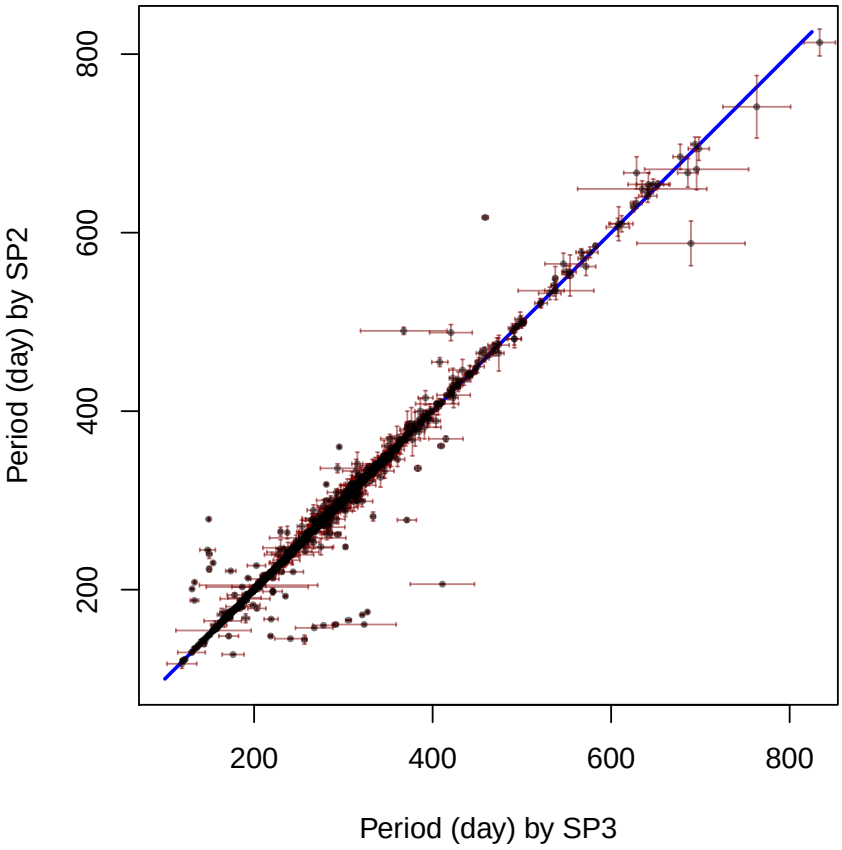


Figure 3.7: Estimated periods for M33 O-rich Miras. The horizontal axis is the period estimated by the SP3 model. The vertical axis is the period of the SP2 model with *ad hoc* correction (Yuan et al., 2018).

while a quadratic PLR is fitted for all Miras,

$$m_l = c_0 + c_1 \times (\log_{10} p_l - 2.3) + c_2 \times (\log_{10} p_l - 2.3)^2.$$

Since the LMC survey has high quality light curves and the PLRs inference is reasonably trustable, we fix c_1 (and c_2) to the values determined by Yuan et al. (2017b) and then solve for c_0. The fitted coefficients of PLRs are summarized in table 3.3. The LMC is used as a reference with distance modulus $\mu_{LMC} = 18.493 \pm 0.048$. The relative distance modulus between M33 and LMC is the bias-corrected difference between the intercepts of PLRs, i.e. $\Delta\mu = \Delta c_0 + \Delta A_\lambda + \Delta ct$, where Δc_0 is the difference of PLR intercepts, ΔA_λ is the interstellar extinction and Δct is the color term

bias in photometric calibration. Here the information of ΔA_λ and Δct can be obtained from Yuan et al. (2018) and are available in table 3.4. Then the derived M33 distance modulus for J, H, K_s bands are (1) 24.81 ± 0.06, 24.79 ± 0.06, 24.76 ± 0.06, respectively, using the linear PLR, and (2) 24.81 ± 0.06, 24.77 ± 0.06, 24.75 ± 0.06, respectively, using the quadratic PLR. These distance moduli are consistent with the results of Yuan et al. (2018).

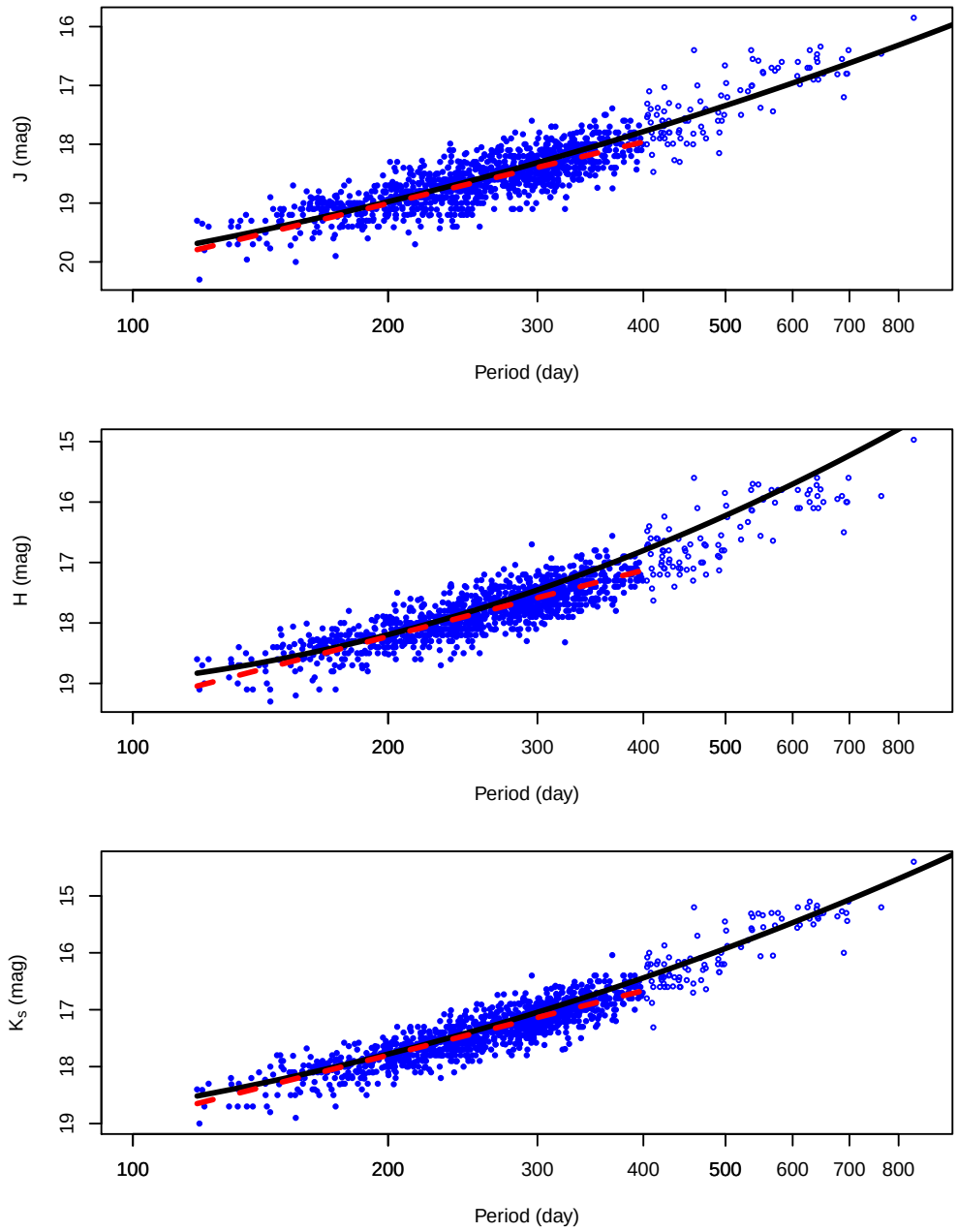


Figure 3.8: M33 O-rich Mira PLRs in J (top), H (middle), and K_s (bottom). The solid points represent stars with estimated period $\leq$ 400d, while the open circles are stars with estimated period $>$ 400d. The dash (red) and solid (black) lines indicate the PLR fits to the linear and quadratic forms, respectively.

Table 3.3: LMC and M33 PLRs coefficients

galaxy	band	linear (period $<$ 400d)				quadratic				
		c_0	c_1	σ	N	c_0	c_1	c_2	σ	N
LMC[1]	J	12.67 $\pm$ 0.01	-3.48 $\pm$ 0.01	0.15	158	12.70 $\pm$ 0.01	-3.49 $\pm$ 0.09	-1.54 $\pm$ 0.23	0.15	178
M33	J	18.98 $\pm$ 0.008	$\cdots$	0.26	1169	19.01 $\pm$ 0.008	$\cdots$	$\cdots$	0.27	1265
LMC[1]	H	11.91 $\pm$ 0.01	-3.64 $\pm$ 0.01	0.16	163	11.96 $\pm$ 0.01	-3.59 $\pm$ 0.10	-3.40 $\pm$ 0.31	0.12	173
M33	H	18.20 $\pm$ 0.007	$\cdots$	0.23	1169	18.23 $\pm$ 0.007	$\cdots$	$\cdots$	0.24	1265
LMC[1]	K_s	11.56 $\pm$ 0.01	-3.77 $\pm$ 0.01	0.12	158	11.59 $\pm$ 0.01	-3.77 $\pm$ 0.08	-2.23 $\pm$ 0.20	0.12	176
M33	K_s	17.78 $\pm$ 0.006	$\cdots$	0.22	1169	17.80 $\pm$ 0.006	$\cdots$	$\cdots$	0.23	1265

Table 3.4: Derived distance moduli for M33

	band	ΔA_λ[1]	Δct[1]	Δc_0	$\Delta\mu$	μ_{LMC}	μ
	J	0.029 $\pm$ 0.008	0.016 $\pm$ 0.036	6.309 $\pm$ 0.013	6.354 $\pm$ 0.039	18.493 $\pm$ 0.048	24.85 $\pm$ 0.062
linear	H	0.018 $\pm$ 0.005	0.010 $\pm$ 0.040	6.287 $\pm$ 0.012	6.315 $\pm$ 0.042	18.493 $\pm$ 0.048	24.81 $\pm$ 0.064
	K_s	0.012 $\pm$ 0.003	-0.007 $\pm$ 0.032	6.223 $\pm$ 0.011	6.228 $\pm$ 0.034	18.493 $\pm$ 0.048	24.72 $\pm$ 0.059
	J	0.029 $\pm$ 0.008	0.016 $\pm$ 0.036	6.311 $\pm$ 0.013	6.356 $\pm$ 0.039	18.493 $\pm$ 0.048	24.85 $\pm$ 0.062
quadratic	H	0.018 $\pm$ 0.005	0.010 $\pm$ 0.040	6.300 $\pm$ 0.012	6.295 $\pm$ 0.042	18.493 $\pm$ 0.048	24.79 $\pm$ 0.064
	K_s	0.012 $\pm$ 0.003	-0.007 $\pm$ 0.032	6.220 $\pm$ 0.012	6.219 $\pm$ 0.034	18.493 $\pm$ 0.048	24.71 $\pm$ 0.059

[1]Data from Yuan et al. (2018)

4. CONCLUSIONS

4.1 First project: a flexible procedure for Positive–Unlabeled learning

In this project we proposed a flexible framework for estimating the mixture proportion and classifier in the PU learning problem. We implemented this framework using two estimators from the FDR literature, C-patra/sen and C-roc. The framework has the power to incorporate other one-dimensional MPE procedures, such as Meinshausen and Rice (2006), Genovese and Wasserman (2004), Langaas et al. (2005), Efron (2007), Jin (2008), Cai and Jin (2010) or Nguyen and Matias (2014). More generally we have strengthened connections between the classification–machine learning literature and the multiple testing literature by constructing estimators using ideas from both communities.

4.2 Second project: periods estimation for Miras using multi-band light curves and inverse Period-Luminosity relations

Accurate period estimation is important in modern astronomy. However, current methods like GLS, MGLS, PGLS are designed for strict-periodic light curves. The SP1 method can only deal with single-band data. The SP2 method is designed for multi-band Mira light curves. None of the existing approaches can nicely handle multi-band quasi-periodic light curves. To fill the gap between application and methodology, we develop the SP3 model, which is designed to model a set of light curves which share the same PLR. The SP3 method has substantial improvement over other methods in the Mira application. Besides computational advantage, the SP3 method may have other advantages: (1) from posterior samples an highest posterior density interval can be obtained for period estimation; (2) "alias" can be efficiently excluded out; (3) it is more flexible to incorporate prior information from other previous well-studied surveys.

The SP3 method proposed in this project currently cannot deal with this situation: the set of light curves contains two or more different types of variable stars, i.e. there are several groups of light curves and different groups share different PLRs or iPLRs. In this scenario, a possible

solution would be to make the prior on frequency a mixture of log normal distributions, but this will leave for future research.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society. Series B (Methodological), 57(1):289–300, 1995. ISSN 00359246.

Yoav Benjamini and Yosef Hochberg. On the adaptive control of the false discovery rate in multiple testing with independent statistics. Journal of educational and Behavioral Statistics, 25(1):60–83, 2000.

Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. Adaptive linear step-up procedures that control the false discovery rate. Biometrika, 93(3):491–507, 2006.

Gérard Biau. Analysis of a random forests model. J. Mach. Learn. Res., 13(1):1063–1095, April 2012. ISSN 1532-4435.

Gilles Blanchard and Étienne Roquain. Adaptive false discovery rate control under independence and dependence. Journal of Machine Learning Research, 10(Dec):2837–2871, 2009.

Gilles Blanchard, Gyemin Lee, and Clayton Scott. Semi-supervised novelty detection. J. Mach. Learn. Res., 11:2973–3009, December 2010. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=1756006.1953028`.

Brigitte Boeckmann, Amos Bairoch, Rolf Apweiler, Marie-Claude Blatter, Anne Estreicher, Elisabeth Gasteiger, Maria J Martin, Karine Michoud, Claire O'Donovan, Isabelle Phan, Sandrine Pilbout, and Michel Schneider. The swiss-prot protein knowledgebase and its supplement trembl in 2003. Nucleic Acids Research, 31(1):365–370, 2003.

Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

T. Tony Cai and Jiashun Jin. Optimal rates of convergence for estimating the null density and proportion of nonnull effects in large-scale multiple testing. Ann. Statist., 38(1):100–145, 02 2010.

Shiyu Chang, Yang Zhang, Jiliang Tang, Dawei Yin, Yi Chang, Mark A. Hasegawa-Johnson, and Thomas S. Huang. Positive-unlabeled learning in streaming networks. In Proceedings of the

22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pages 755–764, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4232-2.

Sanmay Das, Milton H. Saier, and Charles Elkan. Finding Transport Proteins in a General Protein Database, pages 54–66. Springer Berlin Heidelberg, Berlin, Heidelberg, 2007.

Bradley Efron. Size, power and false discovery rates. Ann. Statist., 35(4):1351–1377, 08 2007.

Bradley Efron. Large-scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction, volume 1. Cambridge University Press, 2012.

Bradley Efron, Robert Tibshirani, John D Storey, and Virginia Tusher. Empirical bayes analysis of a microarray experiment. Journal of the American statistical association, 96(456):1151–1160, 2001.

Charles Elkan and Keith Noto. Learning classifiers from only positive and unlabeled data. In Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 213–220. ACM, 2008.

G. Foster. Wavelets for period analysis of unevenly sampled time series. Astronomical Journal, 112:1709, October 1996. doi: 10.1086/118137.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. Bayesian Data Analysis. CRC Press, 2013.

Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. Annals of Statistics, pages 1035–1061, 2004.

Peter Hall, James Reimann, and John Rice. Nonparametric estimation of a periodic function. Biometrika, 87(3):545–557, 2000. ISSN 00063444. URL http://www.jstor.org/stable/2673629.

Shiyuan He, Wenlong Yuan, Jianhua Z. Huang, James Long, and Lucas M. Macri. Period estimation for sparsely-sampled quasi-periodic light curves applied to miras. The Astronomical Journal, 152(6):164, 2016. URL http://stacks.iop.org/1538-3881/152/i=6/a=164.

Caroline D. Huang, Adam G. Riess, Samantha L. Hoffmann, Christopher Klein, Joshua Bloom,

Wenlong Yuan, Lucas M. Macri, David O. Jones, Patricia A. Whitelock, Stefano Casertano, and Richard I. Anderson. A near-infrared period–luminosity relation for miras in ngc 4258, an anchor for a new distance ladder. The Astrophysical Journal, 857(1):67, 2018. URL http://stacks.iop.org/0004-637X/857/i=1/a=67.

Y. Ita, T. Tanabe, N. Matsunaga, Y. Nakajima, C. Nagashima, T. Nagayama, D. Kato, M. Kurita, T. Nagata, S. Sato, M. Tamura, H. Nakaya, and Y. Nakada. VizieR Online Data Catalog: OGLE Variables in Magellanic Clouds (Ita+, 2004). VizieR Online Data Catalog, 735, March 2005.

Jiashun Jin. Proportion of non-zero normal means: Universal oracle equivalences and uniformly consistent estimators. Journal of the Royal Statistical Society. Series B (Statistical Methodology), 70(3):461–493, 2008. ISSN 13697412, 14679868.

P. N. Kholopov, N. N. Samus, E. V. Kazarovets, and N. B. Perova. The 67th Name-List of Variable Stars. Information Bulletin on Variable Stars, 2681, March 1985.

Habib Khosroshahi, Atefeh Javadi, Jacco Th. van Loon, Maryam Saberi, Mohammad Taghi Mirtorabi, and Najmeh Golabatooni. The UK Infrared Telescope M33 monitoring project âĂŞ IV. Variable red giant stars across the galactic disc. Monthly Notices of the Royal Astronomical Society, 447(4):3973–3991, 01 2015. ISSN 0035-8711. doi: 10.1093/mnras/stu2637. URL https://dx.doi.org/10.1093/mnras/stu2637.

Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 67(4):555–572, 2005. ISSN 1467-9868.

H. S. Leavitt and E. C. Pickering. Periods of 25 Variable Stars in the Small Magellanic Cloud. Harvard College Observatory Circular, 173:1–3, March 1912.

Friedrich Leisch and Evgenia Dimitriadou. mlbench: Machine Learning Benchmark Problems, 2010. R package version 2.1-1.

Bing Liu, Wee Sun Lee, Philip S Yu, and Xiaoli Li. Partially supervised classification of text documents. In ICML, volume 2, pages 387–394. Citeseer, 2002.

N. R. Lomb. Least-squares frequency analysis of unequally spaced data. Astrophysics and Space

Science, 39:447–462, February 1976. doi: 10.1007/BF00648343.

James P. Long, Eric C. Chi, and Richard G. Baraniuk. Estimating a common period for a set of irregularly sampled functions with applications to periodic variable star data. Ann. Appl. Stat., 10(1):165–197, 03 2016. doi: 10.1214/15-AOAS885. URL https://doi.org/10.1214/15-AOAS885.

L. M. Macri, K. Z. Stanek, D. D. Sasselov, M. Krockenberger, and J. Kaluzny. DIRECT distances to nearby galaxies using detached eclipsing binaries and cepheids. VI. variables in the central part of m33. The Astronomical Journal, 121(2):870–890, feb 2001. doi: 10.1086/318773. URL https://doi.org/10.1086%2F318773.

Nicolai Meinshausen and John Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. The Annals of Statistics, 34(1):373–393, 2006.

Iain Murray, Ryan Prescott Adams, and David J. C. MacKay. Elliptical slice sampling. 9:541–548, 2010.

Minh Nhut Nguyen, Xiao-Li Li, and See-Kiong Ng. Positive unlabeled learning for time series classification. In Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence - Volume Volume Two, IJCAI'11, pages 1421–1426. AAAI Press, 2011. ISBN 978-1-57735-514-4.

Van Hanh Nguyen and Catherine Matias. On efficient estimators of the proportion of true null hypotheses in a multiple testing setup. Scandinavian Journal of Statistics, 41(4):1167–1194, 2014. ISSN 1467-9469.

Rohit Kumar Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 2015.

Anne Pellerin and Lucas M. Macri. THE m 33 SYNOPTIC STELLAR SURVEY. i. CEPHEID VARIABLES. The Astrophysical Journal Supplement Series, 193(2):26, mar 2011. doi: 10.1088/0067-0049/193/2/26. URL https://doi.org/10.1088%2F0067-0049%

2F193%2F2%2F26.

Planck Collaboration, N. Aghanim, Y. Akrami, M. Ashdown, J. Aumont, C. Baccigalupi, M. Ballardini, A. J. Banday, R. B. Barreiro, N. Bartolo, S. Basak, R. Battye, K. Benabed, J. P. Bernard, M. Bersanelli, P. Bielewicz, J. J. Bock, J. R. Bond, J. Borrill, F. R. Bouchet, F. Boulanger, M. Bucher, C. Burigana, R. C. Butler, E. Calabrese, J. F. Cardoso, J. Carron, A. Challinor, H. C. Chiang, J. Chluba, L. P. L. Colombo, C. Combet, D. Contreras, B. P. Crill, F. Cuttaia, P. de Bernardis, G. de Zotti, J. Delabrouille, J. M. Delouis, E. Di Valentino, J. M. Diego, O. Doré, M. Douspis, A. Ducout, X. Dupac, S. Dusini, G. Efstathiou, F. Elsner, T. A. Enßlin, H. K. Eriksen, Y. Fantaye, M. Farhang, J. Fergusson, R. Fernandez-Cobos, F. Finelli, F. Forastieri, M. Frailis, E. Franceschi, A. Frolov, S. Galeotta, S. Galli, K. Ganga, R. T. Génova-Santos, M. Gerbino, T. Ghosh, J. González-Nuevo, K. M. Górski, S. Gratton, A. Gruppuso, J. E. Gudmundsson, J. Hamann, W. Handley, D. Herranz, E. Hivon, Z. Huang, A. H. Jaffe, W. C. Jones, A. Karakci, E. Keihänen, R. Keskitalo, K. Kiiveri, J. Kim, T. S. Kisner, L. Knox, N. Krachmalnicoff, M. Kunz, H. Kurki-Suonio, G. Lagache, J. M. Lamarre, A. Lasenby, M. Lattanzi, C. R. Lawrence, M. Le Jeune, P. Lemos, J. Lesgourgues, F. Levrier, A. Lewis, M. Liguori, P. B. Lilje, M. Lilley, V. Lindholm, M. López-Caniego, P. M. Lubin, Y. Z. Ma, J. F. Macías-Pérez, G. Maggio, D. Maino, N. Mandolesi, A. Mangilli, A. Marcos-Caballero, M. Maris, P. G. Martin, M. Martinelli, E. Martínez- González, S. Matarrese, N. Mauri, J. D. McEwen, P. R. Meinhold, A. Melchiorri, A. Mennella, M. Migliaccio, M. Millea, S. Mitra, M. A. Miville-Deschênes, D. Molinari, L. Montier, G. Morgante, A. Moss, P. Natoli, H. U. Nørgaard-Nielsen, L. Pagano, D. Paoletti, B. Partridge, G. Patanchon, H. V. Peiris, F. Perrotta, V. Pettorino, F. Piacentini, L. Polastri, G. Polenta, J. L. Puget, J. P. Rachen, M. Reinecke, M. Remazeilles, A. Renzi, G. Rocha, C. Rosset, G. Roudier, J. A. Rubiño-Martín, B. Ruiz-Granados, L. Salvati, M. Sandri, M. Savelainen, D. Scott, E. P. S. Shellard, C. Sirignano, G. Sirri, L. D. Spencer, R. Sunyaev, A. S. Suur-Uski, J. A. Tauber, D. Tavagnacco, M. Tenti, L. Toffolatti, M. Tomasi, T. Trombetti, L. Valenziano, J. Valiviita, B. Van Tent, L. Vibert, P. Vielva, F. Villa, N. Vittorio, B. D. Wandelt, I. K. Wehus, M. White, S. D. M. White, A. Zacchei, and A. Zonca. Planck 2018 results. VI.

Cosmological parameters. ArXiv e-prints, art. arXiv:1807.06209, July 2018.

Harish Ramaswamy, Clayton Scott, and Ambuj Tewari. Mixture proportion estimation via kernel embedding of distributions. PMLR, pages 2052–2060, 2016.

Carl Edward Rasmussen and Christopher K. I. Williams. Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning). The MIT Press, 2005. ISBN 026218253X.

A. G. Riess, L. M. Macri, S. L. Hoffmann, D. Scolnic, S. Casertano, A. V. Filippenko, B. E. Tucker, M. J. Reid, D. O. Jones, J. M. Silverman, R. Chornock, P. Challis, W. Yuan, P. J. Brown, and R. J. Foley. A 2.4% Determination of the Local Value of the Hubble Constant. ApJ, 826:56, July 2016. doi: 10.3847/0004-637X/826/1/56.

A. G. Riess, S. Casertano, W. Yuan, L. Macri, B. Bucciarelli, M. G. Lattanzi, J. W. MacKenty, J. B. Bowers, W. Zheng, A. V. Filippenko, C. Huang, and R. I. Anderson. Milky Way Cepheid Standards for Measuring Cosmic Distances and Application to Gaia DR2: Implications for the Hubble Constant. The Astrophysical Journal, 861:126, July 2018. doi: 10.3847/1538-4357/aac82e.

Stéphane Robin, Avner Bar-Hen, Jean-Jacques Daudin, and Laurent Pierre. A semi-parametric approach for mixture models: Application to local false discovery rate estimation. Computational Statistics & Data Analysis, 51(12):5483–5493, 2007.

Milton H. Saier, Jr, Can V. Tran, and Ravi D. Barabote. Tcdb: the transporter classification database for membrane transport protein analyses and information. Nucleic Acids Research, 34(suppl_1):D181–D186, 2006.

J. D. Scargle. Studies in astronomical time series analysis. II - Statistical aspects of spectral analysis of unevenly spaced data. The Astrophysical Journal, 263:835–853, December 1982. doi: 10.1086/160554.

Clayton Scott. A rate of convergence for mixture proportion estimation, with application to learning from noisy labels. In AISTATS, 2015.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In COLT, pages 489–511, 2013.

Branimir Sesar, Zeljko Ivezić, Skyler H. Grammer, Dylan P. Morgan, Andrew C. Becker, Mario Jurić, Nathan De Lee, James Annis, Timothy C. Beers, Xiaohui Fan, Robert H. Lupton, James E. Gunn, Gillian R. Knapp, Linhua Jiang, Sebastian Jester, David E. Johnston, and Hubert Lampeitl. Light curve templates and galactic distribution of rr lyrae stars from sloan digital sky survey stripe 82. The Astrophysical Journal, 708(1):717, 2010. URL http://stacks.iop.org/0004-637X/708/i=1/a=717.

Branimir Sesar, Morgan Fouesneau, Adrian M. Price-Whelan, Coryn A. L. Bailer-Jones, Andy Gould, and Hans-Walter Rix. A probabilistic approach to fitting periodâĂŞluminosity relations and validating gaia parallaxes. The Astrophysical Journal, 838(2):107, 2017. URL http://stacks.iop.org/0004-637X/838/i=2/a=107.

I. Soszyński, W. A. Dziembowski, A. Udalski, M. Kubiak, M. K. Szymanski, G. Pietrzynski, L. Wyrzykowski, O. Szewczyk, and K. Ulaczyk. The Optical Gravitational Lensing Experiment. Period–Luminosity Relations of Variable Red Giant Stars. AcA, 57:201–225, September 2007.

I. Soszyński, A. Udalski, M. K. Szymański, M. Kubiak, G. Pietrzyński, Ł. Wyrzykowski, O. Szewczyk, K. Ulaczyk, and R. Poleski. The Optical Gravitational Lensing Experiment. The OGLE-III Catalog of Variable Stars. III. RR Lyrae Stars in the Large Magellanic Cloud. ACTA ASTRONOMICA, 59:1–18, March 2009.

John D. Storey. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64(3):479–498, 2002. ISSN 1467-9868.

M. Süveges, B. Sesar, M. Váradi, N. Mowlavi, A. C. Becker, Z. Ivezić, M. Beck, K. Nienartowicz, L. Rimoldini, P. Dubath, P. Bartholdi, and L. Eyer. Search for high-amplitude Îť scuti and rr lyrae stars in sloan digital sky survey stripe 82 using principal component analysis. Monthly Notices of the Royal Astronomical Society, 424(4):2528–2550, 2012. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2012.21229.x. URL http://dx.doi.org/10.1111/j.1365-2966.2012.21229.x.

Andrzej Udalski, Michal K. SzymaÅĎski, Igor Soszyński, and Radoslaw Poleski. The optical gravitational lensing experiment. final reductions of the ogle-iii data. ACTA ASTRONOMICA,

58:69–87, 2008.

Jacob T. VanderPlas. Understanding the lomb–scargle periodogram. The Astrophysical Journal Supplement Series, 236(1):16, may 2018. doi: 10.3847/1538-4365/aab766. URL https://doi.org/10.3847%2F1538-4365%2Faab766.

Jacob T. VanderPlas and Z. Ivezić. Periodograms for multiband astronomical time series. The Astrophysical Journal, 812(1):18, 2015. URL http://stacks.iop.org/0004-637X/812/i=1/a=18.

Yuyang Wang, Roni Khardon, and Pavlos Protopapas. Nonparametric bayesian estimation of periodic light curves. The Astrophysical Journal, 756(1):67, 2012. URL http://stacks.iop.org/0004-637X/756/i=1/a=67.

Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R Leathwick. Presence-only data and the em algorithm. Biometrics, 65(2):554–563, 2009.

L. L. Watkins, N. W. Evans, V. Belokurov, M. C. Smith, P. C. Hewett, D. M. Bramich, G. F. Gilmore, M. J. Irwin, S. Vidrih, ÅĄ. Wyrzykowski, and D. B. Zucker. Substructure revealed by rr lyraes in sdss stripe 82. Monthly Notices of the Royal Astronomical Society, 398(4):1757–1770, 2009. ISSN 1365-2966. doi: 10.1111/j.1365-2966.2009.15242.x. URL http://dx.doi.org/10.1111/j.1365-2966.2009.15242.x.

P. A. Whitelock, J. W. Menzies, M. W. Feast, F. Nsengiyumva, and N. Matsunaga. VizieR Online Data Catalog: JHKs photometry of AGB stars in NGC 6822 (Whitelock+, 2013). VizieR Online Data Catalog, 742, January 2014.

Peng Yang, Xiao-Li Li, Jian-Ping Mei, Chee-Keong Kwoh, and See-Kiong Ng. Positive-unlabeled learning for disease gene identification. Bioinformatics, 28(20):2640–2647, 2012.

W. Yuan, S. He, L. M. Macri, J. Long, and J. Z. Huang. The M33 Synoptic Stellar Survey. II. Mira Variables. The Astronomical Journal, 153:170, April 2017a. doi: 10.3847/1538-3881/aa63f1.

W. Yuan, L. M. Macri, S. He, J. Z. Huang, S. M. Kanbur, and C.-C. Ngeow. Large Magellanic Cloud Near-infrared Synoptic Survey. V. Period-Luminosity Relations of Miras. The Astronomical Journal, 154:149, October 2017b. doi: 10.3847/1538-3881/aa86f1.

Wenlong Yuan, Lucas M. Macri, Atefeh Javadi, Zhenfeng Lin, and Jianhua Z. Huang. Near-infrared mira period–luminosity relations in m33. The Astronomical Journal, 156(3):112, 2018. URL http://stacks.iop.org/1538-3881/156/i=3/a=112.

M. Zechmeister and M. Kürster. The generalised Lomb-Scargle periodogram. A new formalism for the floating-mean and Keplerian periodograms. Astronomy and Astrophysics, 496:577–584, March 2009. doi: 10.1051/0004-6361:200811296.

APPENDIX A

PROOFS IN CHAPTER 2

A.1 Theorems

A.1.1 Proof of Theorem 1

Equivalently, we are trying to prove

$$\frac{G-(1-\gamma)G_L}{\gamma} \text{ is a CDF} \Leftrightarrow \frac{F-(1-\gamma)F_1}{\gamma} \text{ is a CDF.} \tag{A.1}$$

Sufficient to show

$$G-(1-\gamma)G_L \text{ non-decreasing} \Leftrightarrow f-(1-\gamma)f_1 \geq 0 \text{ with probability 1.} \tag{A.2}$$

First we show $\Leftarrow$. Consider any $t_2 > t_1$. Then

$$(G(t_2)-(1-\gamma)G_L(t_2)) - (G(t_1)-(1-\gamma)G_L(t_1)) = \int_{\{x:C(x)\in(t_1,t_2]\}} \underbrace{f(x)-(1-\gamma)f_1(x)}_{\geq 0 \text{ by assumption}} d\mu(x)$$

$$\geq 0.$$

Now we show $\Rightarrow$ by proving the contrapositive. By assumption there exists

$$A = \{x : f(x)-(1-\gamma)f_1(x) < 0\}$$

such that $P(A) > 0$. Further we have

$$A = \left\{ x : (1-\gamma)\frac{(1-\pi)}{\pi} > \frac{f(x)}{f_1(x)}\frac{(1-\pi)}{\pi} \right\}$$
$$= \left\{ x : \underbrace{\frac{1}{1+(1-\gamma)\frac{(1-\pi)}{\pi}}}_{\equiv t^*} < C(x) \right\}.$$

So

$$(G(1)-(1-\gamma)G_L(1)) - (G(t^*)-(1-\gamma)G_L(t^*)) = \int_{A=\{x:C(x)>t^*\}} f(x)-(1-\gamma)f_1(x)d\mu(x)$$
$$< 0.$$

A.1.2 Proof of Theorem 2

$$n^\beta(G_n(t)-G(t)) = \frac{n^\beta}{n^{1/2}}\underbrace{n^{1/2}\left(G_n(t)-\mathbb{E}[\mathbb{1}_{C_n(X)\le t}|C_n]\right)}_{\equiv \mathbb{R}_n(t)}$$
$$+\underbrace{n^\beta\left(\mathbb{E}[\mathbb{1}_{C_n(X)\le t}|C_n]-G(t)\right)}_{\equiv \mathbb{Q}_n(t)}$$

We now show that $\mathbb{R}_n(t)$ and $\mathbb{Q}_n(t)$ are $O_P(1)$ uniformly in t. Together these facts show the expression is $O_P(1)$ uniformly in t.

$\underline{\mathbb{R}_n(t)}$**:** Note

$$\mathbb{R}_n(t) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n \mathbb{1}_{C_n(X_i)\le t} - \mathbb{E}[\mathbb{1}_{C_n(X)\le t}|C_n]\right).$$

By the DKW inequality

$$P(||\mathbb{R}_n||_\infty > x\big|C_n) \le 2e^{-2x^2}.$$

Thus $||\mathbb{R}_n||_\infty$ is $O_P(1)$.

$\underline{\mathbb{Q}_n(t)}$**:** We have

$$\begin{aligned}
\mathbb{Q}_n(t) &= \mathbb{E}[\overbrace{(\mathbb{1}_{C_n(X)\leq t} - \mathbb{1}_{C(X)\leq t})}^{\equiv T_n} \big| C_n] \\
&\leq \underbrace{|E[T_n \mathbb{1}_{|C(X)-t|\leq \epsilon_n} \big| C_n]|}_{B_1} \\
&+ \underbrace{|E[T_n \mathbb{1}_{|C(X)-t|>\epsilon_n} \mathbb{1}_{|C(X)-C_n(X)|<\epsilon_n} \big| C_n]|}_{B_2} \\
&+ \underbrace{|E[T_n \mathbb{1}_{|C(X)-t|>\epsilon_n} \mathbb{1}_{|C(X)-C_n(X)|>\epsilon_n} \big| C_n]|}_{B_3}
\end{aligned}$$

Noting that $|T_n| \leq 1$ and C_n is independent of $C(X)$, we have

$$B_1 \leq P(|C(X) - t| \leq \epsilon_n) \leq 2\epsilon_n \sup_t g(t)$$

where g is the density of $C(X)$, which exists and is bounded by Assumptions 2. B_2 is 0 because $T_n = 0$ whenever the indicator functions in B_2 are both 1. Finally noting $B_3 \leq \mathbb{1}_{|C(X)-C_n(X)|>\epsilon_n}$ and using Markov's inequality twice, we have

$$\begin{aligned}
P(B_3 > r_n) &\leq P(\mathbb{E}[\mathbb{1}_{|C(X)-C_n(X)|>\epsilon_n} | C_n] > r_n) \\
&\leq \frac{P(|C(X) - C_n(X)| > \epsilon_n)}{r_n} \\
&\leq \frac{E[|C_n(X) - C(X)|]}{\epsilon_n r_n}.
\end{aligned}$$

If we choose $\epsilon_n = n^{-\tau/3}$ and $r_n \sim n^{-\tau/3}$, then we can set $\beta = \tau/3$ and achieve the desired result. Identical arguments hold for showing $n^\beta(G_{L,n}(t) - G_L(t))$ is $O_P(1)$ uniform in t.

A.1.3 Proof of Theorem 4

Since $\widehat{t} = \inf\{t : G_{L,n}(t) \geq 1 - n^{-q}\} - n^{-1}$ and $0 < q < \beta$, we have

$$(n^\beta(1 - G_{L,n}(\widehat{t})))^{-1} = \frac{n^q}{n^\beta} = o(1).$$

Recall by Theorem 2 we have

$$n^{\beta}(G_{L,n}(t) - G_L(t)) \equiv d_L(t) = O_P(1)$$
$$n^{\beta}(G_n(t) - G(t)) \equiv d(t) = O_P(1)$$

where this and subsequent O_P and o_P are uniform in t. We have

$$\begin{aligned}\frac{G_n(\widehat{t}) - G_{L,n}(\widehat{t})}{1 - G_{L,n}(\widehat{t})} &= \frac{G(\widehat{t}) - G_L(\widehat{t})}{1 - G_{L,n}(\widehat{t})} + \frac{n^{-\beta}(d_L(\widehat{t}) - d(\widehat{t}))}{1 - G_{L,n}(\widehat{t})} \\ &= \underbrace{\left(\frac{1 - G_L(\widehat{t})}{1 - G_{L,n}(\widehat{t})}\right)}_{\equiv A} \underbrace{\left(\frac{G(\widehat{t}) - G_L(\widehat{t})}{1 - G_L(\widehat{t})}\right)}_{\equiv k(\widehat{t})} + \underbrace{\frac{d_L(\widehat{t}) - d(\widehat{t})}{n^{\beta}(1 - G_{L,n}(\widehat{t}))}}_{o_P(1)}.\end{aligned}$$

Note that

$$A = 1 + \frac{d_L(\widehat{t})}{n^{\beta}(1 - G_{L,n}(\widehat{t}))} = 1 + o_P(1).$$

Thus it is sufficient to show that $k(\widehat{t}) \to \alpha_0$. By Lemma 1, $k(t) \uparrow \alpha_0$ as $t \uparrow t^*$. We show that for any $\epsilon > 0$

$$P(\widehat{t} \in (t^* - \epsilon, t^*)) \to 1.$$

Thus by the continuous mapping theorem, the estimator is consistent.

Part 1: We show $P(t^* - \widehat{t} > \epsilon) \to 0$. By the definition of t^*, there exists $\gamma > 0$ such that $G_L(t^* - \epsilon/2) = 1 - \gamma$. We have

$$\begin{aligned}P(t^* - \widehat{t} > \epsilon) &= P(G_{L,n}(t^* - \epsilon + n^{-1}) > G_{L,n}(\widehat{t} + n^{-1})) \\ &\leq P(G_{L,n}(t^* - \epsilon + n^{-1}) > 1 - n^{-q}) \\ &\leq \underbrace{P(G_L(t^* - \epsilon + n^{-1}) > 1 - n^{-q} - \gamma/2)}_{\equiv A} \\ &\quad + \underbrace{P(|G_{L,n}(t^* - \epsilon + n^{-1}) - G_L(t^* - \epsilon + n^{-1})| > \gamma/2)}_{\to 0 \text{ by Theorem 2}}.\end{aligned}$$

$A \to 0$ because for sufficiently large n, $G_L(t^* - \epsilon + n^{-1}) \leq G_L(t^* - \epsilon/2) = 1 - \gamma < 1 - n^{-q} - \gamma/2$.

Part 2: We show $P(\widehat{t} \geq t^*) \to 0$. We have

$$\begin{aligned} P(\widehat{t} \geq t^*) &= P(G_{n,L}(\widehat{t} + n^{-1}) \geq G_{n,L}(t^* + n^{-1})) \\ &= P(1 - n^{-q} \geq G_{n,L}(t^* + n^{-1})) \\ &= P(1 - G_{n,L}(t^* + n^{-1}) \geq n^{-q}) \\ &= P(\underbrace{n^{\beta}(G_L(t^* + n^{-1}) - G_{n,L}(t^* + n^{-1}))}_{O_P(1) \text{ by Theorem 2}} \geq n^{\beta - q}). \end{aligned}$$

Since $\beta > q$ we have the result.

A.1.4 Proof of Theorem 3

Proof. $\forall \epsilon > 0$, we need to show $P(|\widehat{\alpha}_0^{c_n} - \alpha_0| > \epsilon) \to 0$. Note

$$P(|\widehat{\alpha}_0^{c_n} - \alpha_0| > \epsilon) = P(\widehat{\alpha}_0^{c_n} < \alpha_0 - \epsilon) + P(\widehat{\alpha}_0^{c_n} > \alpha_0 + \epsilon).$$

First we show that $P(\widehat{\alpha}_0^{c_n} < \alpha_0 - \epsilon) \to 0$. If $\alpha_0 \leq \epsilon$, then

$$P(\widehat{\alpha}_0^{c_n} < \alpha_0 - \epsilon) \leq P(\widehat{\alpha}_0^{c_n} < 0) = 0.$$

If $\alpha_0 > \epsilon$, suppose we have $\widehat{\alpha}_0^{c_n} < \alpha_0 - \epsilon$, then by Lemma 6,

$$d_n(\widehat{G}_{s,n}^{\alpha_0 - \epsilon}, \check{G}_{s,n}^{\alpha_0 - \epsilon}) \leq \frac{c_n}{n^{\beta - \eta}(\alpha_0 - \epsilon)}.$$

The LHS of above converges to positive constant by Lemma 5, while the RHS converges to zero by the choice of c_n, hence $P(\widehat{\alpha}_0^{c_n} < \alpha_0 - \epsilon) \to 0$.

Now we show that $P(\widehat{\alpha}_0^{c_n} > \alpha_0 + \epsilon) \to 0$. Suppose we have $\widehat{\alpha}_0^{c_n} > \alpha_0 + \epsilon$, then by Lemma 6,

$$n^{\beta - \eta} d_n(\widehat{G}_{s,n}^{\alpha_0 + \epsilon}, \check{G}_{s,n}^{\alpha_0 + \epsilon}) > \frac{c_n}{(\alpha_0 - \epsilon)}.$$

The LHS of above converges to zero by Lemmas 5 and 4, while the RHS converges to infinity by the choice of c_n, hence $P(\widehat{\alpha}_0^{c_n} > \alpha_0 + \epsilon) \to 0$. □

A.2 Lemmas

Lemma 1. $\lim_{t\uparrow t^*} k(t) = \alpha_0$.

Proof. Define $\alpha_0' = \lim_{t\uparrow t^*} k(t)$.

Show $\alpha_0' \leq \alpha_0$: By the definition of α_0 there exists c.d.f G_{α_0} such that

$$\begin{aligned} G(t) &= \alpha_0 G_{\alpha_0}(t) + (1-\alpha_0) G_L(t) \\ &\leq \alpha_0 + (1-\alpha_0) G_L(t). \end{aligned}$$

Thus

$$k(t) = \frac{G(t) - G_L(t)}{1 - G_L(t)} \leq \alpha_0$$

for all t. Thus $\alpha_0' = \lim_{t\uparrow t^*} k(t) \leq \alpha_0$.

Show $\alpha_0' \geq \alpha_0$: Consider any $\gamma < \alpha_0$. We show $\gamma < \alpha_0'$. Since $\gamma < \alpha_0$,

$$\frac{G - (1-\gamma) G_L}{\gamma}$$

is not a c.d.f. Thus there exists $t_1 < t_2$ such that

$$\frac{G(t_1) - (1-\gamma) G_L(t_1)}{\gamma} > \frac{G(t_2) - (1-\gamma) G_L(t_2)}{\gamma}. \tag{A.3}$$

Since the left hand side is bounded above by 1 and $G(t) = 1 \forall t \geq t^*$, $t_2 < t^*$. From Equation (A.3) we have

$$G(t_1) - G(t_2) > (1-\gamma)(G_L(t_1) - G_L(t_2))$$

which implies (since $G_L(t_1) - G_L(t_2) < 0$) that

$$\frac{G(t_2) - G(t_1)}{G_L(t_2) - G_L(t_1)} < (1 - \gamma). \tag{A.4}$$

From Lemma 3 we have

$$\frac{1 - G_L(t_2)}{1 - G(t_2)} = \frac{G_L(1) - G_L(t_2)}{G(1) - G(t_2)} \geq \frac{G_L(t_2) - G_L(t_1)}{G(t_2) - G(t_1)}$$

Combining this result with Equation (A.4) we obtain

$$\frac{1 - G(t_2)}{1 - G_L(t_2)} \leq 1 - \gamma$$

which implies

$$\gamma \leq \frac{G(t_2) - G_L(t_2)}{1 - G(t_2)} = k(t_2)$$

Since $k(t) \uparrow$ as $t \uparrow t^*$ (see Lemma 2), we have the result. □

Lemma 2. *$k(t)$ is increasing on $t \in [0, t^*)$.*

Proof. Recall $Q(p) = \inf\{t \in (0, 1] : G_L(t) \geq p\}$ and $t^* = Q(1)$. Note that with $a, b, c, d > 0$ and $a/b < c/d$,

$$\frac{a + c}{b + d} > \frac{a}{b}.$$

Next note that by Lemma 3, for $t^* > t_2 > t_1$,

$$\frac{G(t_2) - G(t_1)}{G_L(t_2) - G_L(t_1)} > \frac{1 - G(t_2)}{1 - G_L(t_2)}.$$

Thus we have

$$\begin{aligned}
1 - k(t_1) &= \frac{1 - G(t_1)}{1 - G_L(t_1)} \\
&= \frac{1 - G(t_2) + G(t_2) - G(t_1)}{1 - G_L(t_2) + G_L(t_2) - G_L(t_1)} \\
&\geq \frac{1 - G(t_2)}{1 - G_L(t_2)} \\
&= 1 - k(t_2).
\end{aligned}$$

□

Lemma 3 (Ratio)**.** *For all* $0 \leq t_1 < t_2 \leq 1$ *where* $G(t_2) - G(t_1) > 0$ *we have*

$$\frac{1-\pi}{\pi}\frac{t_1}{1-t_1} < \frac{G_L(t_2) - G_L(t_1)}{G(t_2) - G(t_1)} \leq \frac{1-\pi}{\pi}\frac{t_2}{1-t_2}$$

where $1/0 \equiv \infty$.

Proof. The classifier is

$$C(x) = \frac{\pi f_L(x)}{\pi f_L(x) + (1-\pi) f(x)} = \frac{1}{1 + \frac{1-\pi}{\pi}\frac{f(x)}{f_L(x)}}$$

Define $A_t = \{x : C(x) \leq t\} = \{x : \frac{1-t}{t}\frac{\pi}{1-\pi} f_L(x) \leq f(x)\}$. Therefore on the set $A_{t_2} \cap A_{t_1}^C$ we have

$$\frac{1-t_2}{t_2}\frac{\pi}{1-\pi} f_L(x) \leq f(x) < \frac{1-t_1}{t_1}\frac{\pi}{1-\pi} f_L(x)$$

So

$$\frac{G_L(t_2) - G_L(t_1)}{G(t_2) - G(t_1)} = \frac{\int_{A_{t_2} \cap A_{t_1}^C} f_L(x)}{\int_{A_{t_2} \cap A_{t_1}^C} f(x)} > \frac{\int_{A_{t_2} \cap A_{t_1}^C} f_L(x)}{\frac{1-t_1}{t_1}\frac{\pi}{1-\pi}\int_{A_{t_2} \cap A_{t_1}^C} f_L(x)} = \frac{t_1}{1-t_1}\frac{1-\pi}{\pi}.$$

We can obtain the upper bound in an identical manner. □

Lemma 4.

$$n^{\beta-\eta} d_n(G, G_n) = o_P(1),$$
$$n^{\beta-\eta} d_n(G_L, G_{L,n}) = o_P(1).$$

Proof.

$$n^{\beta-\eta} d_n(G, G_n) = \sqrt{\int \left[\underbrace{n^{-\eta}}_{=o_P(1)} \underbrace{n^{\beta} \left(G_n(t) - G(t)\right)}_{=O_P(1)} \right]^2 dG_n(t)},$$

where $n^{\beta} \left(G_n(t) - G(t)\right) = O_P(1)$ uniformly, and then $n^{-\eta} n^{\beta} \left(G_n(t) - G(t)\right) = o_P(1)$ uniformly. Therefore

$$n^{\beta-\eta} d_n(G, G_n) \leq \sup_t |n^{-\eta} n^{\beta} \left(G_n(t) - G(t)\right)| = o_P(1).$$

The G_L, $G_{L,n}$ case can be proved in an identical manner. □

Lemma 5. *For* $1 \geq \gamma \geq \alpha_0$,

$$\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \leq d_n(G, G_n) + (1-\gamma) d_n(G_L, G_{L,n}).$$

Thus,

$$\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \to \begin{cases} 0 & \text{if } \gamma \geq \alpha_0, \\ > 0 & \text{if } \gamma < \alpha_0. \end{cases}$$

Proof. Let

$$G^{\gamma}_s = \frac{G - (1-\gamma) G_L}{\gamma}.$$

If $\gamma \geq \alpha_0$, then

$$\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \leq \gamma d_n(\widehat{G}^{\gamma}_{s,n}, G^{\gamma}_{s}) \leq d_n(G, G_n) + (1-\gamma) d_n(G_L, G_{L,n}).$$

The first inequality holds by the definition of $\check{G}^{\gamma}_{s,n}$ due to the fact that G^{γ}_s is a valid CDF when $1 \geq \gamma \geq \alpha_0$, and the second inequality is due to triangle inequality.

Now we prove the limit property of $\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n})$. If $\gamma \geq \alpha_0$, then $\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \to 0$ since $d_n(G, G_n) \to 0$ and $d_n(G_L, G_{L,n}) \to 0$ by Lemma 4. If $\gamma < \alpha_0$, by the definition of α^G_0, G^{γ}_s is not a valid c.d.f.. Pointwisely, $\widehat{G}^{\gamma}_{s,n} \to G^{\gamma}_s$. So for large n, $\widehat{G}^{\gamma}_{s,n}$ is not valid c.d.f., while $\check{G}^{\gamma}_{s,n}$ is always a c.d.f.. So $\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n})$ would converge to some positive constant. □

Lemma 6. *$B_n := \{\gamma \in [0,1] : n^{\beta-\eta}\gamma d_n(\widehat{G}^{\gamma}_{s,n}, \check{G}^{\gamma}_{s,n}) \leq c_n\}$ is convex. Thus, $B_n = (\widehat{\alpha}^{c_n}_0, 1]$ or $B_n = [\widehat{\alpha}^{c_n}_0, 1]$.*

Proof. Obviously, $1 \in B_n$. Assume $\gamma_1 \leq \gamma_2$ from B_n, let $\gamma_3 = \xi\gamma_1 + (1-\xi)\gamma_2$, where $\xi \in [0,1]$. Then by definition of $\widehat{G}^{\gamma}_{s,n}$,

$$\xi\gamma_1 \widehat{G}^{\gamma_1}_{s,n} + (1-\xi)\gamma_2 \widehat{G}^{\gamma_2}_{s,n} = \gamma_3 \widehat{G}^{\gamma_3}_{s,n}.$$

Note that $\frac{1}{\gamma_3}\left(\xi\gamma_1 \check{G}^{\gamma_1}_{s,n} + (1-\xi)\gamma_2 \check{G}^{\gamma_2}_{s,n}\right)$ is a valid c.d.f. We have $\gamma_3 \in B_n$ because

$$\begin{aligned}
d_n(\widehat{G}^{\gamma_3}_{s,n}, \check{G}^{\gamma_3}_{s,n}) &\leq d_n\left(\widehat{G}^{\gamma_3}_{s,n}, \frac{1}{\gamma_3}\left(\xi\gamma_1 \check{G}^{\gamma_1}_{s,n} + (1-\xi)\gamma_2 \check{G}^{\gamma_2}_{s,n}\right)\right) \\
&= d_n\left(\frac{1}{\gamma_3}\left(\xi\gamma_1 \widehat{G}^{\gamma_1}_{s,n} + (1-\xi)\gamma_2 \widehat{G}^{\gamma_2}_{s,n}\right), \frac{1}{\gamma_3}\left(\xi\gamma_1 \check{G}^{\gamma_1}_{s,n} + (1-\xi)\gamma_2 \check{G}^{\gamma_2}_{s,n}\right)\right) \\
&\leq \frac{\xi\gamma_1}{\gamma_3} d_n(\widehat{G}^{\gamma_1}_{s,n}, \check{G}^{\gamma_1}_{s,n}) + \frac{(1-\xi)\gamma_2}{\gamma_3} d_n(\widehat{G}^{\gamma_2}_{s,n}, \check{G}^{\gamma_2}_{s,n}) \\
&\leq \frac{\xi\gamma_1}{\gamma_3} \frac{c_n}{n^{\beta-\eta}\gamma_1} + \frac{(1-\xi)\gamma_2}{\gamma_3} \frac{c_n}{n^{\beta-\eta}\gamma_2} = \frac{c_n}{n^{\beta-\eta}\gamma_3}.
\end{aligned}$$

□

APPENDIX B

CHAPTER 3

B.1 Examples of simulated light curves in Simulation I

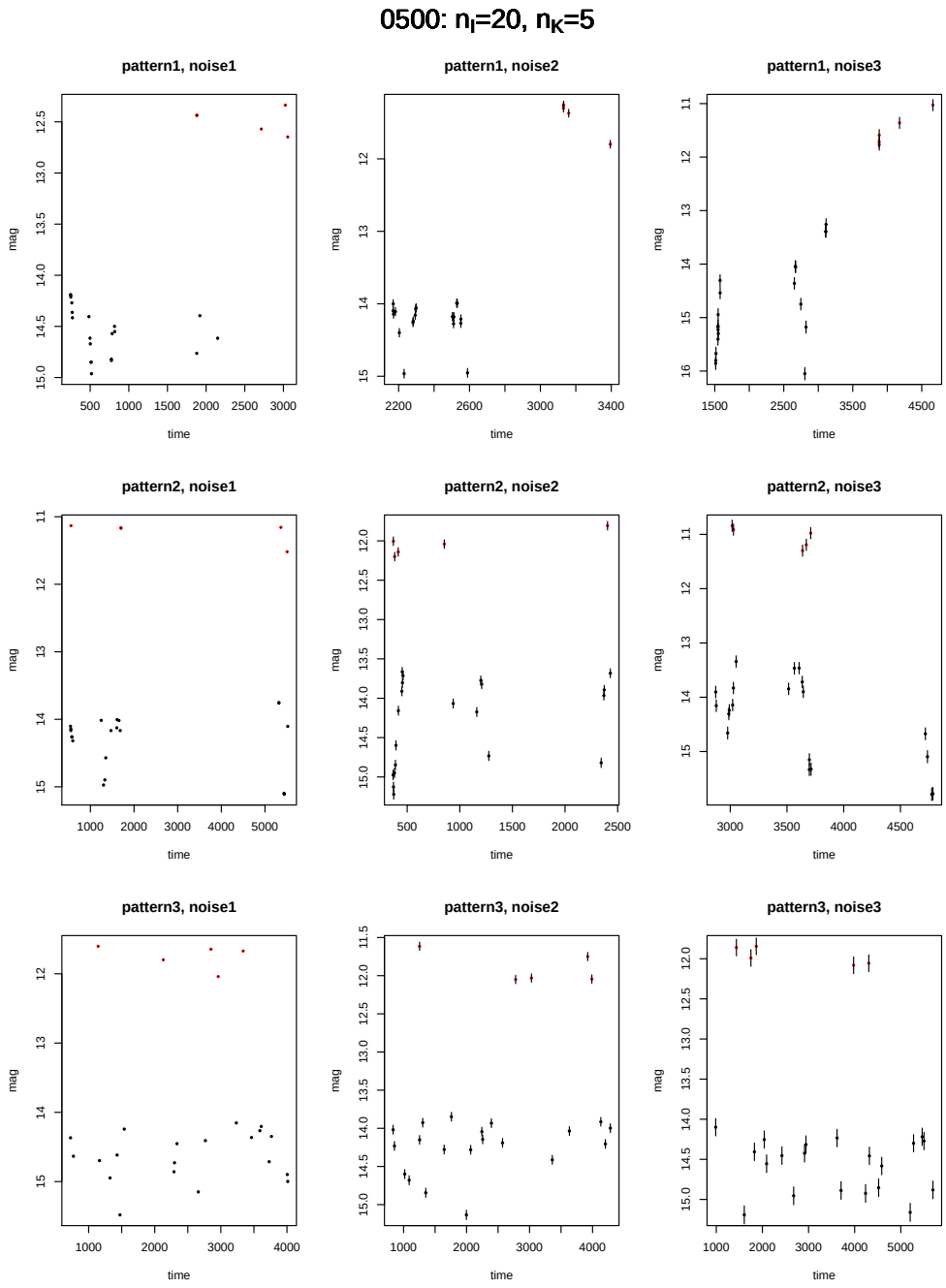


Figure B.1: 9 examples of light curve with different time patterns and noise levels, when $n_I = 10$ and $n_K = 5$. "0500" is the id in each set of light curves.

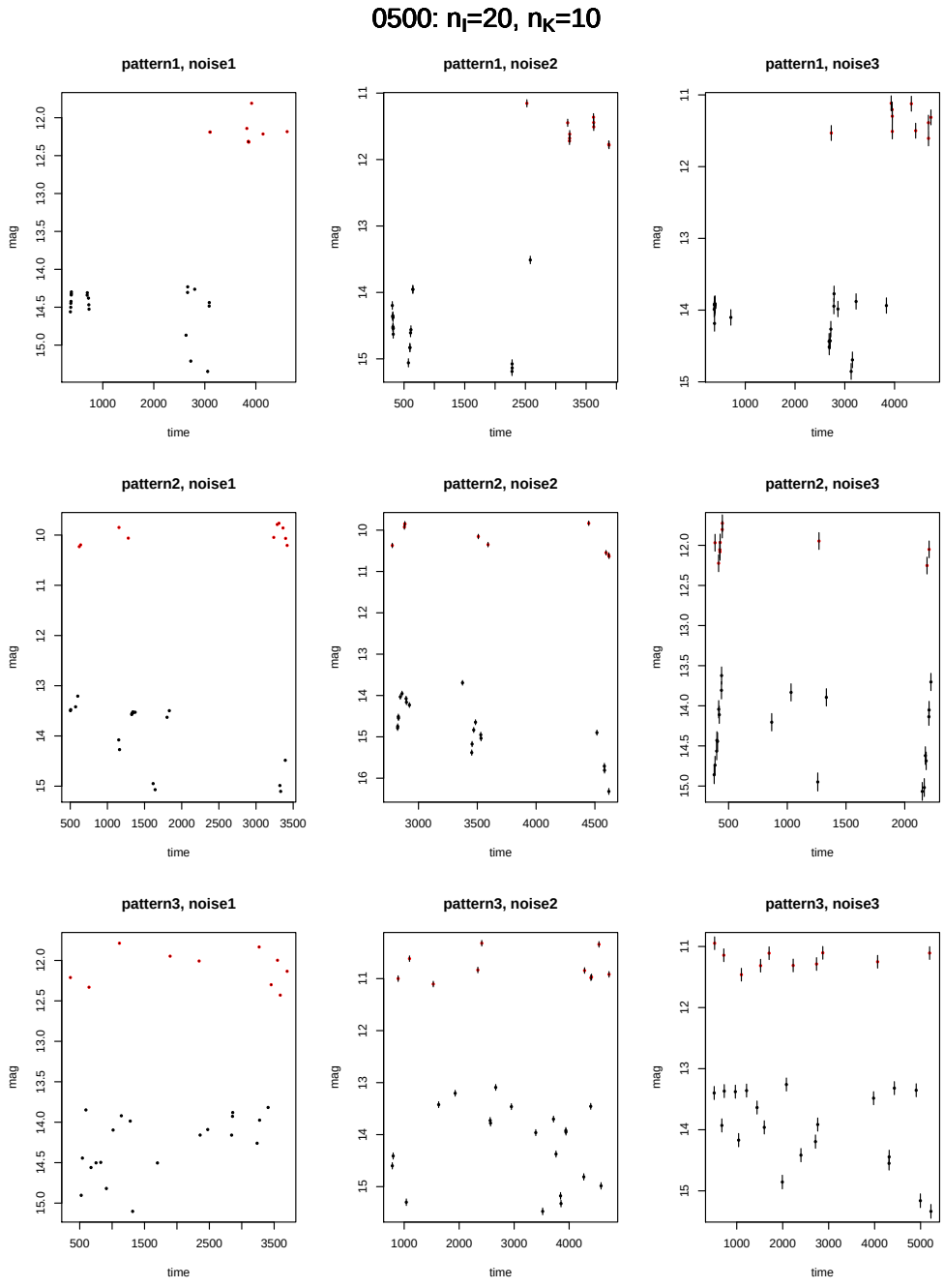


Figure B.2: 9 examples of light curve with different time patterns and noise levels, when $n_I = 20$ and $n_K = 10$. "0500" is the id in each set of light curves.

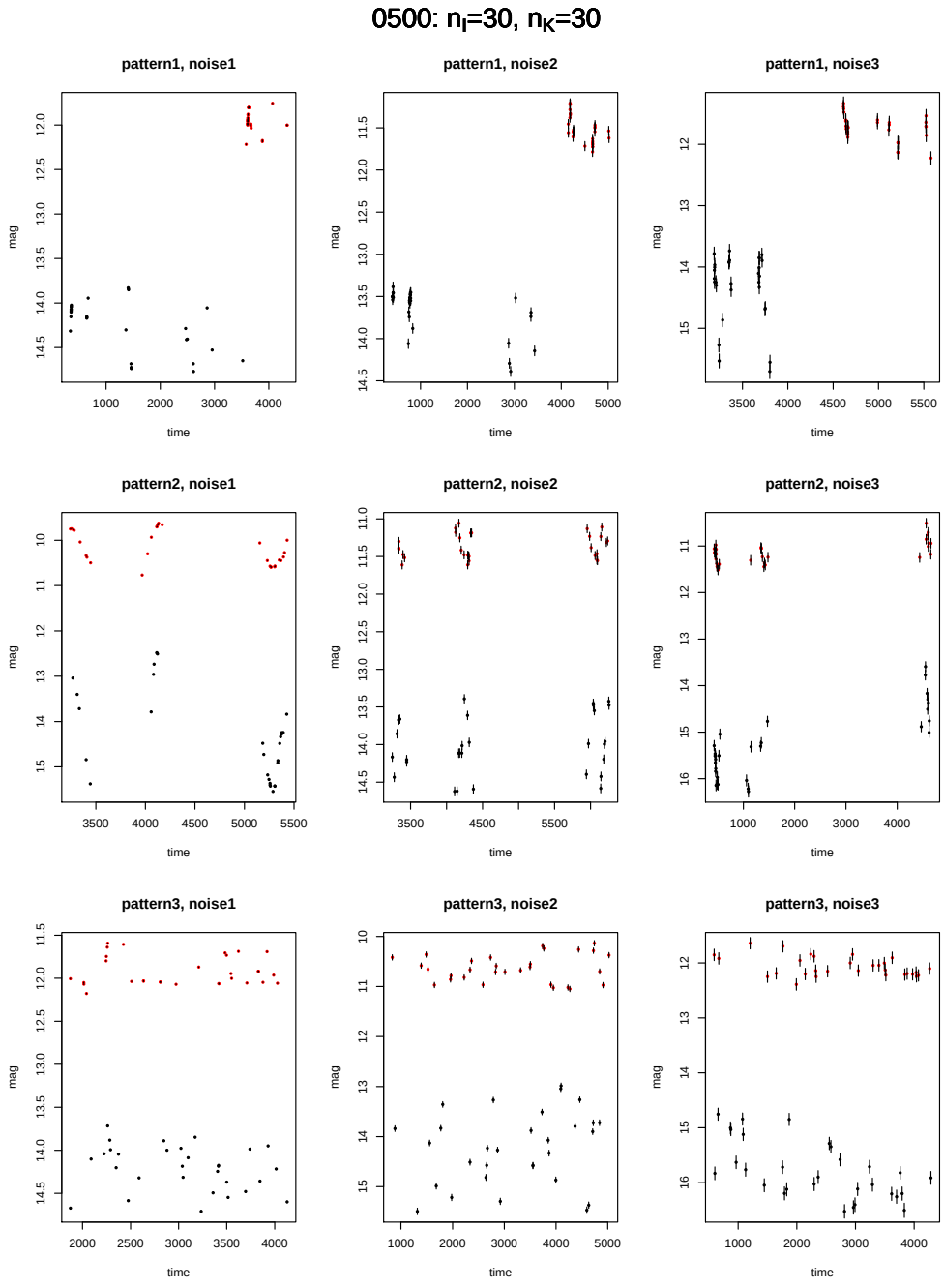


Figure B.3: 9 examples of light curve with different time patterns and noise levels, when $n_I = 30$ and $n_K = 30$. "0500" is the id in each set of light curves.

B.2 Performance comparisons of models in Simulation I

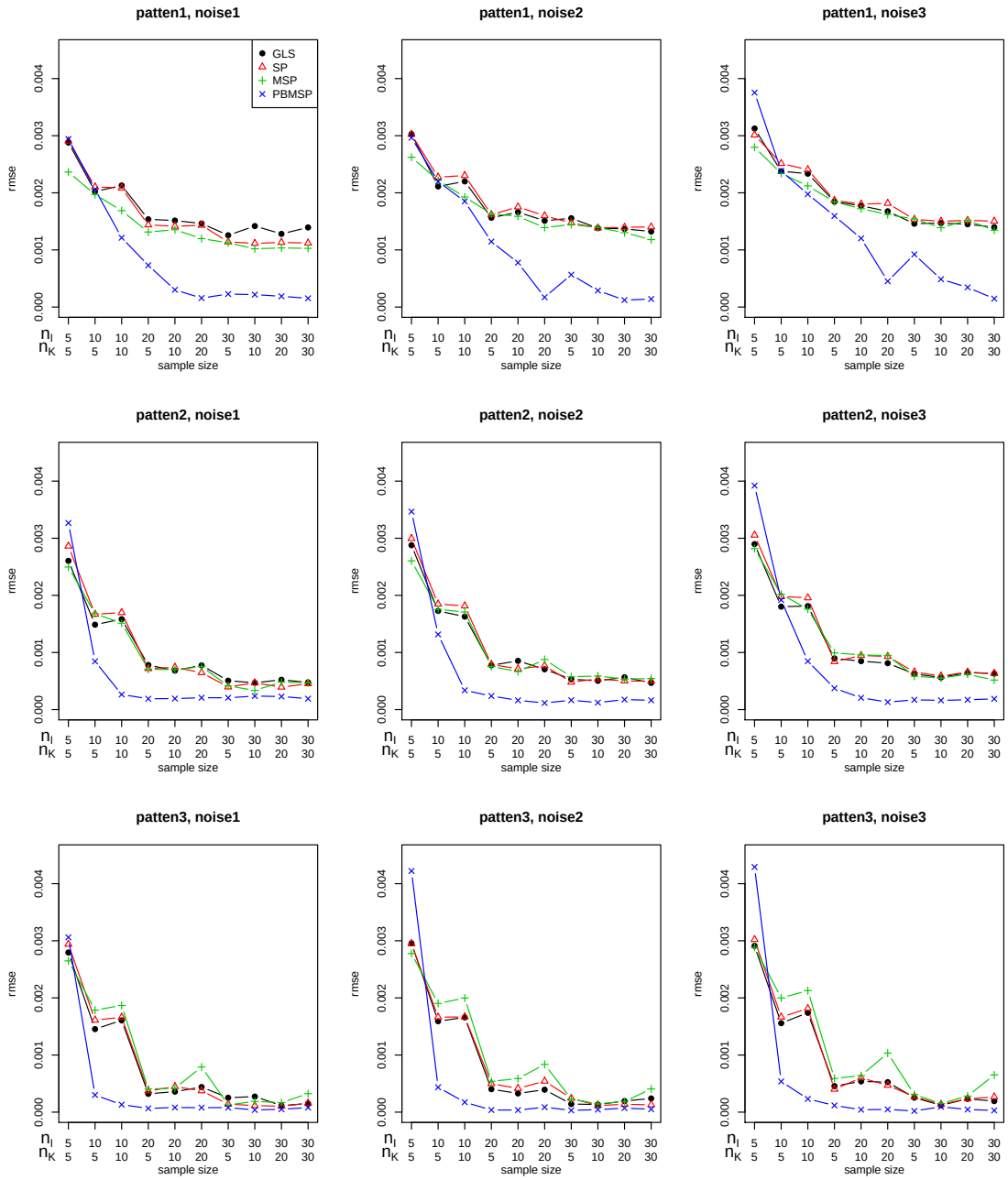


Figure B.4: Simulation I: Performance comparison of GLS, SP, MSP and PBMSP models with RMSE metric.

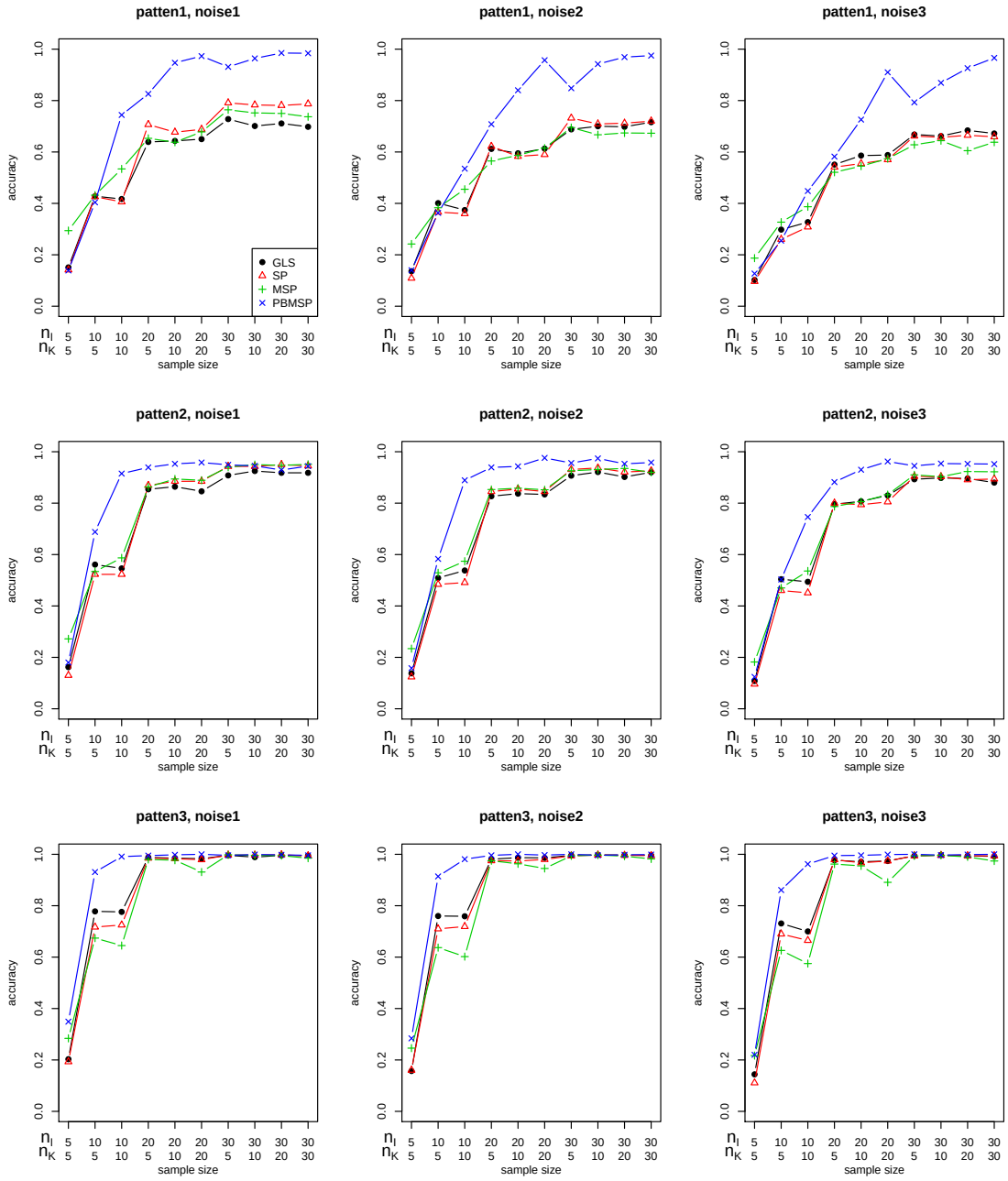


Figure B.5: Simulation I: Performance comparison of GLS, SP, MSP and PBMSP models with ACC metric.

B.3 Local periodogram for a light curve in Simulation II

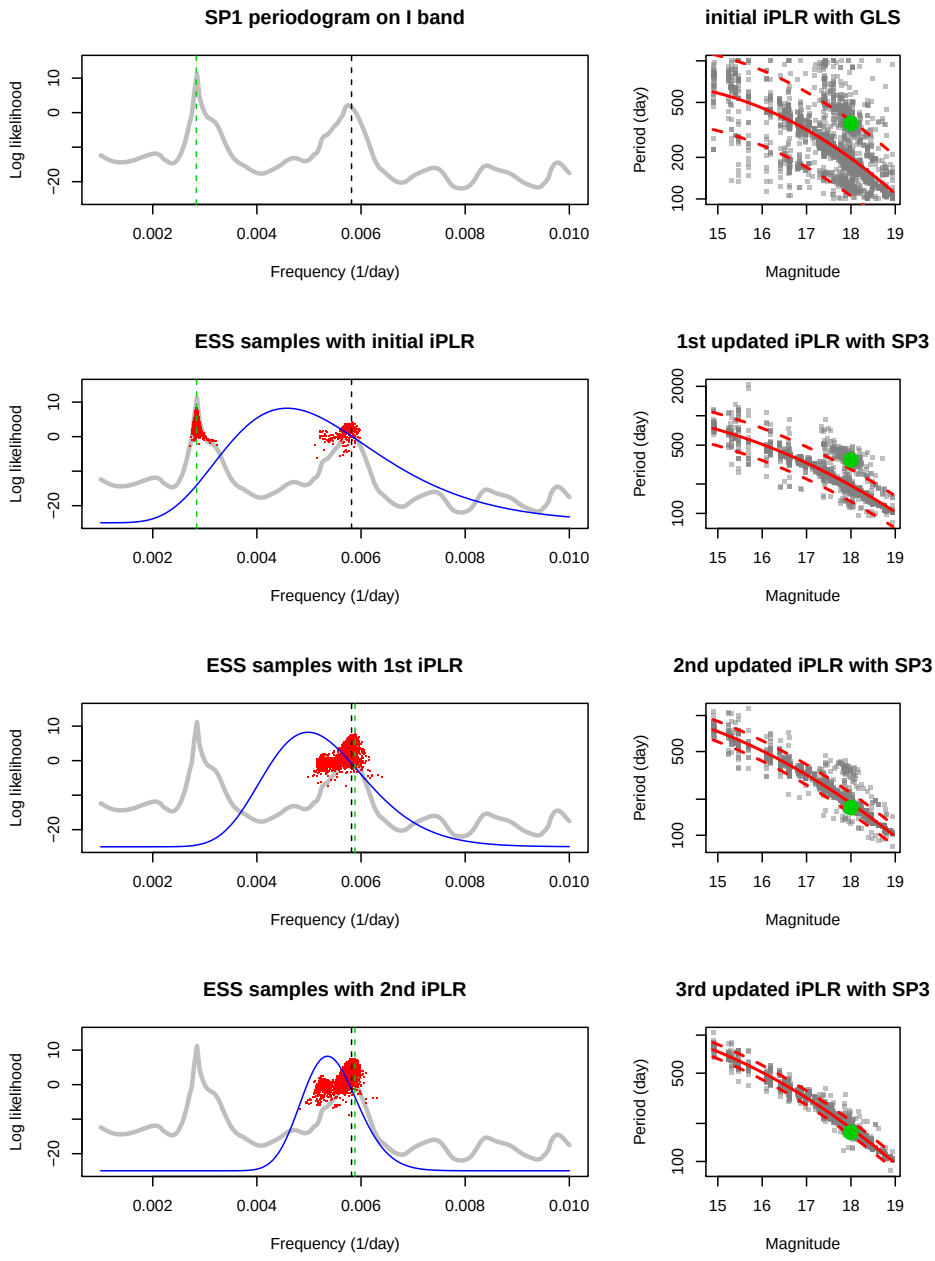


Figure B.6: An example of periodogram for a light curve (id=00080). True period = 171.88 (days), denoted as vertical black dash line in left panels. The left top panel is periodogram of the SP1 model on I band. The left panels in other rows are local periodogram with ESS samples (red dots) and frequency priors (blue solid lines). Vertical green dash lines in all left panels represent the estimated frequencies, which are also marked as green dots in all right panels respectively.