The Astrophysical Journal Supplement Series, 236:12 (14pp), 2018 May https://doi.org/10.3847/1538-4365/aab77c

Robust Period Estimation Using Mutual Information for Multiband Light Curves in the Synoptic Survey Era

Pablo Huijse[1,2], Pablo A. Estévez[2,3], Francisco Förster[2,4], Scott F. Daniel[5], Andrew J. Connolly[5], Pavlos Protopapas[6], Rodrigo Carrasco[3], and José C. Príncipe[7]
[1] Informatics Institute, Universidad Austral de Chile, Valdivia, Chile
[2] Millennium Institute of Astrophysics, Chile; phuijse@inf.uach.cl, pablo.huijse@gmail.com
[3] Department of Electrical Engineering, Universidad de Chile, Santiago, Chile
[4] Center for Mathematical Modeling, Universidad de Chile, Santiago, Chile
[5] Department of Astronomy, University of Washington, Seattle, WA, USA
[6] Institute for Applied Computational Science, Harvard University, Cambridge, MA, USA
[7] Computational Neuroengineering Laboratory of University of Florida, FL, USA
Received 2017 June 16; revised 2017 September 4; accepted 2017 September 5; published 2018 May 11

Abstract

The Large Synoptic Survey Telescope (LSST) will produce an unprecedented amount of light curves using six optical bands. Robust and efficient methods that can aggregate data from multidimensional sparsely sampled time-series are needed. In this paper we present a new method for light curve period estimation based on quadratic mutual information (QMI). The proposed method does not assume a particular model for the light curve nor its underlying probability density and it is robust to non-Gaussian noise and outliers. By combining the QMI from several bands the true period can be estimated even when no single-band QMI yields the period. Period recovery performance as a function of average magnitude and sample size is measured using 30,000 synthetic multiband light curves of RR Lyrae and Cepheid variables generated by the LSST Operations and Catalog simulators. The results show that aggregating information from several bands is highly beneficial in LSST sparsely sampled time-series, obtaining an absolute increase in period recovery rate up to 50%. We also show that the QMI is more robust to noise and light curve length (sample size) than the multiband generalizations of the Lomb–Scargle and AoV periodograms, recovering the true period in 10%–30% more cases than its competitors. A python package containing efficient Cython implementations of the QMI and other methods is provided.

Key words: methods: data analysis – methods: statistical – stars: variables: general

1. Introduction

The next decade will see the rise of extremely large telescopes (Tyson & Borne 2012), allowing astronomers to probe the sky with unprecedented depth, resolution, and coverage. An emblematic example of this is the Large Synoptic Survey Telescope (LSST; Ivezić et al. 2008; Abell et al. 2009). The LSST will begin operations in 2022, capturing the whole southern hemisphere in six bands (*ugrizy*) over 10 years. This translates into 500 PetaBytes of images and 50 PetaBytes in catalogs, corresponding to 37 billion astronomical objects. Robust and computationally efficient methods are needed in order to process the sheer amount of light curves (Feigelson & Babu 2012; Tyson & Borne 2012; Huijse et al. 2014). In this work we focus on the task of period estimation in multiband light curves such as those that will be produced by the LSST. In what follows we describe the problem and review some of the existing methods.

Variable stars are celestial objects whose brightnesses vary through time due to intrinsic or extrinsic reasons (Percy 2007). There are certain classes of variable stars, such as Cepheids, RR Lyrae, and eclipsing binaries, whose brightnesses vary regularly following periodical patterns. The periods of these stars are key in cosmological research, as they can be used to measure the distance to their host galaxies. The period of variable stars is also important for asteroseismology research and variable star classification (Richards et al. 2011).

The main tool to study variable stars is the *light curve*, a time-series of stellar flux or magnitude. Light curves obtained from Earth-based surveys are irregularly sampled due to observation constraints and also have data gaps of different lengths. Light curves are affected by several noise sources, e.g., photon noise, sky background noise, and scintillation, which can be modeled as uncorrelated (white) noise with variance that changes between samples, i.e., light curves have heteroscedastic errors (Akritas 1997). Additionally, light curves are affected by correlated (red) noise due to observations taken with changing air-mass and atmospheric conditions, telescope tracking, and other systematics (Pont et al. 2006). These characteristics make light curve analysis a challenging task.

Conventional methods for period estimation, such as the Fast Fourier Transform, cannot be directly applied due to the irregular sampling. Several methods have been developed by the statistics and astronomy communities to deal with the analysis of unevenly sampled time-series (Graham et al. 2013a). These methods can be broadly classified as parametric and non-parametric. The most widely used parametric method is the Lomb–Scargle (LS) periodogram (Scargle 1982), which equates to finding the best sinusoidal model fit to the light curve in a least squares sense. The LS periodogram has been generalized to take into account heteroscedastic errors (Zechmeister & Kürster 2009) and also more complex models based on Truncated Fourier series (Palmer 2009).

Phase Dispersion Minimization (Stellingwerf 1978), Minimum String Length (Clarke 2002), and the Analysis of Variance (AoV) periodogram (Schwarzenberg-Czerny 1996) are classical examples of non-parametric methods. These methods do not rely on sinusoidal models for the data. Instead they optimize a metric on the phase diagram of the light curve $\{\phi_i, m_i\}_{i=1,\ldots,N}$, where m_i are the magnitudes and the phases

ϕ_i are obtained from the time instants t_i given a certain trial period P as

$$\phi_i = \frac{\mathrm{mod}(t_i, P)}{P} \in [0, 1], \quad (1)$$

where mod(·, ·) stands for the division remainder operator. For example, in the AoV periodogram the phase diagram is binned and a ratio of the variance of the bins and the total variance is computed. By minimizing this ratio over a set of trial periods, an estimate of the true period is obtained. Non-parametric methods that rely on information theoretic criteria have also been proposed, e.g., The Conditional entropy (CE) periodogram (Graham et al. 2013b) and the Correntropy Kernelized Periodogram (Huijse et al. 2012; Protopapas et al. 2015). In Zucker (2016) a statistical criterion for independence using the cumulative distribution of the folded light curve was proposed. This criterion outperformed the LS in sparsely sampled non-sinusoidal light curves.

LSST (Ivezić et al. 2008) will produce time-series in six optical bands (ugrizy) with non-simultaneous observations, i.e., time intervals between bands will differ. The main observing strategy consists of two exposures per night for a given field. Fields will be revisited every three days on average considering all bands. Single-band average revisit times are longer, e.g., *r*-band is revisited every 15 days. This means that single-band data will be rather sparse. By the end of the first year, an average of 18.4 points will be available in the *r*-band. Bands will have different priorities, e.g., *r*-band and *i*-band will get more visits than the rest. A reliable period detection algorithm for LSST light curves must take into account non-simultaneous observations from all available bands. In recent years VanderPlas & Ivezic (2015) presented an extension of the LS periodogram to sparsely sampled multiband light curves. The multiband LS periodogram combines the single-band periodograms and also fits a term to take into account the variability shared between bands. The AoV periodogram was generalized in a similar way by Mondrik et al. (2015). The multiband AoV is a normalized weighted average of the single-band AoV periodograms. We propose a new period estimation method for multiband light curves that is based on mutual information (MI). We test this method using synthetic light curves generated using the LSST Operations Simulator (OpSim) and Catalog Simulator (CatSim) (Oluseyi et al. 2012; Connolly et al. 2014; Delgado et al. 2014). The proposed method achieves better period recovery rates than established methods, especially in low sample and low signal-to-noise data.

2. Literature Review

In this work we make extensive use of the information theoretic concept of MI. In a broad sense, MI measures the reduction of the uncertainty of a random variable (RV) given that we know a second RV. MI can also be seen as a measure of dependence although, unlike correlation, MI is able to capture nonlinear dependence between RVs. More formally, MI is posed as the divergence (statistical distance) between the joint probability density function (PDF) of the RVs and the product of their marginal PDFs. Several definitions of MI exist in the literature, Shannon's MI being the most well known (Gray 2011). Shannon's MI for continuous RVs X and Y with joint PDF $f_{X,Y}(\cdot, \cdot)$ is defined as

$$\begin{aligned} \mathrm{MI}_S(X, Y) &= D_{KL}(f_{X,Y} || f_X f_Y) \\ &= \iint f_{X,Y} \log f_{X,Y}\, dx\, dy \\ &\quad - \int f_X \log f_X\, dx - \int f_Y \log f_Y\, dy, \end{aligned} \quad (2)$$

where $D_{KL}(\cdot||\cdot)$ is the Kullback–Leibler divergence and $f_X(x) = \int f_{X,Y}(x, y)dy$, $f_Y(y) = \int f_{X,Y}(x, y)dx$ are the marginal PDFs of X and Y, respectively. Computing MI using Equation (2) is a difficult task, as it requires estimating the joint and marginal PDFs of the RVs. Ideally we want to avoid posing assumptions on the PDFs, hence we focus on non-parametric estimators. Two widely used approaches to compute MI through PDF estimation are the kernel density (KDE; Moon et al. 1995) and k-nearest neighbors (KNN; Kraskov et al. 2004) estimators. A review of these and other MI estimators using short data sets (50 samples) can be found in Khan et al. (2007).

In this work we intend to avoid the estimation of the PDF by using MI definitions arising from generalized divergences. Such MI estimators have been proposed in the information theoretic learning (ITL; Xu 1999; Principe et al. 2000; Principe 2010) literature. In what follows we present the derivation of two MI definitions for continuous RVs from the ITL framework. Starting from the Euclidean distance between PDFs

$$D_{\mathrm{ED}}(f(x)||g(x)) = \int (f(x) - g(x))^2\, dx,$$

the Euclidean distance quadratic MI (Xu 1999; Principe 2010) between RVs X and Y is defined as

$$\begin{aligned} \mathrm{QMI}_{\mathrm{ED}}(X, Y) &= D_{\mathrm{ED}}(f_{X,Y}(x, y)||f_X(x)f_Y(y)) \\ &= \iint f_{X,Y}^2\, dx\, dy - 2 \iint f_{X,Y} f_X f_Y\, dx\, dy \\ &\quad + \int f_X^2\, dx \int f_Y^2\, dy \\ &= V_J - 2V_C + V_M, \end{aligned} \quad (3)$$

where $f_{X,Y}(\cdot, \cdot)$ is the joint PDF of X and Y, while $f_X(\cdot)$ and $f_Y(\cdot)$ are the marginal PDFs, respectively.

The terms V_J, V_M, and V_C correspond to the integrals of the squared joint PDF, the squared product of the marginal PDFs, and the product of joint PDF and marginal PDFs, respectively. The ITL framework provides an estimator of these quantities that can be computed directly from data samples. This estimator is called the information potential (IP; Principe 2010) of an RV and it corresponds to the expected value of its PDF. Note that the expected value of a PDF is equivalent to the integral of the squared PDF. Appendix A shows how the IP estimator is derived. Assuming that we have $\{x_i, y_i\}_{i=1,\ldots,N}$ independent and identically distributed (iid) realizations of RVs X and Y, and using the IP estimator, the following expressions are obtained:

$$\begin{aligned} V_M = \mathrm{IP}_X \mathrm{IP}_Y &= \left(\frac{1}{N^2} \sum_{i,j=1}^{N,N} G_{\sqrt{2}h}(x_i - x_j) \right) \\ &\quad \times \left(\frac{1}{N^2} \sum_{i,j=1}^{N,N} G_{\sqrt{2}h}(y_i - y_j) \right), \end{aligned} \quad (4)$$

$$V_J = \mathrm{IP}_{X,Y} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}h}(x_i - x_j) G_{\sqrt{2}h}(y_i - y_j), \quad (5)$$

and

$$V_C = \mathrm{IP}_{X\times Y} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}h}(x_i - x_j)\right) \times \left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2}h}(y_i - y_j)\right), \quad (6)$$

where

$$G_h(x) = \frac{1}{\sqrt{2\pi}\,h}\exp\left(\frac{\|x\|^2}{2h^2}\right) \quad (7)$$

is the Gaussian kernel with bandwidth h. Note how the integrals have been replaced by sums of pairwise differences between data samples.

The second ITL quadratic MI that we consider in this work is obtained by defining a divergence measure based on the Cauchy–Schwarz inequality:

$$D_{\mathrm{CS}}(f(x)\|g(x)) = -\log \frac{\left(\int f(x)g(x)\,dx\right)^2}{\int f(x)^2\,dx \int g(x)^2\,dx}.$$

Then, the Cauchy–Schwarz quadratic MI (Principe et al. 2000; Principe 2010) for continuous RVs X and Y is

$$\begin{aligned}\mathrm{QMI}_{\mathrm{CS}}(X, Y) &= D_{\mathrm{CS}}(f_{X,Y}(x, y)\|f_X(x)f_Y(y)) \\ &= \log \iint f_{X,Y}^2\,dx\,dy - 2\log \iint f_{X,Y} f_X f_Y\,dx\,dy \\ &\quad + \log \int f_X^2\,dx \int f_Y^2\,dy \\ &= \log V_J - 2\log V_C + \log V_M,\end{aligned} \quad (8)$$

where V_M, V_J, and V_C are computed using Equations (4), (5), and (6), respectively. In the following sections we adapt these quadratic mutual information (QMI) estimators for the case of period estimation in light curves.

3. Methods

3.1. Generating Synthetic LSST Light Curves

In this section we describe the procedure to generate synthetic light curves using the LSST Operation Simulator (OpSim) and the Catalog Simulator (CatSim) tools. In its normal operation regime the LSST will visit the same field every three nights. Six bands will be available (ugrizy). The single-visit 5σ depth in the r-band will be approximately 24.5. The actual cadence will depend on weather conditions, slew and filter-change times, and downtime due to maintenance, among other factors. The OpSim simulates these factors to produce multiband pointings that are consistent with the LSST scientific drivers.

The CatSim provides tools to generate different types of sources. In our case we are interested in generating periodic variable stars. CatSim requires the user to specify a normalizing magnitude, spectral energy distribution (SED), and a template. The template file sets the variability type and the period of the resulting light curve. Templates of Cepheids (CEPH), ab-type RR Lyrae (RRab), and c-type RR Lyrae (RRc) are available, among other models. The CatSim RR Lyrae models correspond to Sesar et al. (2010) Stripe 82 Sloan Digital Sky Survey templates. SEDs of modeled main sequence stars are also available.

The procedure we use to generate a synthetic light curve is as follows:

1. Select a variability model, e.g., CEPH, RRab or RRc.
2. Randomly select a template associated to the variability type. This defines the period.
3. Randomly select a SED profile.
4. Randomly select a normalizing magnitude by drawing from $U(20, 25)$, where $U(a, b)$ is the Uniform distribution.
5. Randomly select an MJD for the initial phase of the template by drawing from $U(59580, 60580)$.
6. Randomly select a position in the sky by drawing R.A. from $U(65, 75)$ and decl. from $U(-30, -20)$.
7. Generate the object using CatSim *StellarLightCurveGenerator* class.
8. Generate a set of multiband pointings for the synthetic object using OpSim according to its position in the sky.

As we are only interested in estimating period recovery, we do not aim to model a realistic sky distribution of these variables. A python code that executes this procedure and also the resulting light curves used in this paper can be found at github.com/phuijse/LSST_simulations. To run this code, previous installation of the LSST simulations framework is required.[7] We run this procedure to generate a set containing 1000 synthetic LSST light curves for each variability type.

This procedure generates a "clean" light curve $\{t_i, m_i, \sigma_i\}_i^b$ with $i = 1, \ldots, N_b$, where t corresponds to a time instant in MJD, m is the stellar magnitude, σ is the photometric error, and b denotes the band index. Bands may have a different number of points N_b. The photometric error is generated according to Equation (4) of (Ivezić et al. 2008).

The last step to produce a realistic light curve is to contaminate the clean magnitude values m with the photometric error σ. This is done by drawing a standard normal RV $\{r_i\}$ of length N_b and then updating the magnitudes as $\widehat{m}_i = m_i + r_i\sigma_i$. For each of the "clean" light curves we draw 10 contaminated light curves, thus increasing the size of the set to 10000 per variability type. Figure 1 shows an example of a synthetic RRab light curve before and after the contamination process.

3.2. Period Estimation by Maximizing Mutual Information

We propose to use MI estimators to detect the underlying period in variable star light curves. In this section we present the rationale behind this proposition. We start by applying the epoch-folding transformation for a certain trial period (Equation (1)) to the unevenly sampled time instants in order to obtain the phase diagram $\{\phi_i, m_i, \sigma_i\}_{i=1,\ldots,N}$. We assume that the light curve is periodic with an unknown period. The phases $\{\phi_i\}$ correspond to our non-parametric model of the periodicity, while $\{m_i\}$ correspond to our noisy observations. As usual $\{\sigma_i\}$ are the estimated errors on our observations.

If the light curve is periodic with period P_T, then folding with this period will yield the model that best explains

[7] Instructions can be found at https://confluence.lsstcorp.org/display/SIM/Catalog+Simulations+Documentation

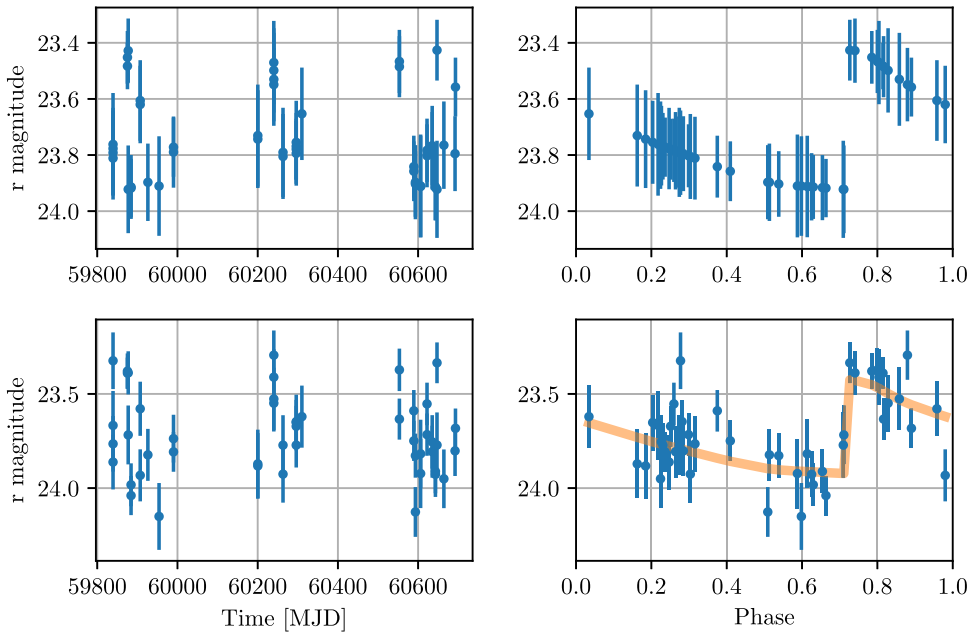

Figure 1. Synthetic ab-type RR Lyrae light curve with a period of 0.64698 days. The upper row corresponds to the clean light curve obtained using the LSST tools and its respective phase diagram. The lower row corresponds to a contaminated realization of the original light curve. In the lower right plot the line corresponds to the clean light curve.

our observations. This can be measured by calculating the MI between phases and magnitudes, i.e., the amount of information shared by the model and observations. We can test several models (foldings) and find the one that maximizes MI to detect the best period. Second-order methods (e.g., correlation) are limited to detecting linear relations. More robust periodicity detection methods can be obtained using MI, which overcomes this limitation. Note that MI requires iid realizations of the RVs. Although light curves are time-series and hence there exist serial correlations in time, these correlations are broken in the phase diagram. Appendix B refers to this issue in detail.

A second interpretation of using MI for periodicity detection relies on MI's definition as the divergence (statistical distance) between the joint PDF and the marginal PDF of the RVs. If the light curve is folded with the wrong period, the structure in the joint PDF will be almost equal to the product of the marginal PDFs, i.e., magnitudes are independent of the phases. On the other hand, if the correct period is chosen, the joint PDF will present a structure that is not captured by the product of the marginals. By maximizing MI we are maximizing the dependency between model and observations.

Let's denote M and Φ as the RVs associated with magnitude and phase, respectively. We can estimate the PDF of M, given its realizations using KDE, as follows:

$$f_M(m) = \frac{1}{N}\sum_{i=1}^{N} G_{\sqrt{\sigma_i^2+h_m^2}}(m - m_i) = \frac{1}{N}\sum_{i=1}^{N} \frac{1}{\sqrt{2\pi(\sigma_i^2 + h_m^2)}} \exp\left(-\frac{1}{2}\frac{(m-m_i)^2}{(\sigma_i^2 + h_m^2)}\right), \quad (9)$$

where each sample m_i has a bandwidth that incorporates the KDE bandwidth h_m and its given uncertainty σ_i. As Φ is a periodic RV, we need a periodic kernel to appropriately estimate its PDF. We consider a kernel arising from the Wrapped Cauchy (WC) distribution (Jammalamadaka et al. 2001) and estimate Φ's PDF as

$$f_\Phi(\phi) = \frac{1}{N}\sum_{i=1}^{N} \mathrm{WC}_{h_\phi}(\phi - \phi_i) = \frac{1}{2\pi N}\sum_{i=1}^{N} \frac{1 - e^{-2h_\phi}}{1 + e^{-2h_\phi} - 2e^{-h_\phi}\cos(2\pi(\phi - \phi_i))}, \quad (10)$$

where $h_\phi \in (0, \infty)$ is the scale of the Cauchy distribution. For $h_\phi \to \infty$ the WC kernel behaves like the circular uniform distribution, while for $h_\phi \to 0$ it concentrates on its mean. The WC kernel is symmetric, translation-invariant, and closed under convolution[8] (Jammalamadaka et al. 2001). Being closed under convolution is desirable because it allow us to compute information potentials efficiently. The joint PDF of Φ and M is estimated as

$$f_{\Phi,M}(\phi, m) = \frac{1}{N}\sum_{i=1}^{N} G_{\sqrt{\sigma_i^2+h_m^2}}(m - m_i) \cdot \mathrm{WC}_{h_\phi}(\phi - \phi_i), \quad (11)$$

because the multiplication of valid kernel functions is also a kernel.

Figure 2 shows the estimated joint and product of marginal PDFs of a synthetic ab-type RR Lyrae for three different trial periods, its real period (0.682 days), the sidereal day (0.9973 days), and a random period. By inspecting the PDFs we can see that the difference between the joint (middle column) and marginals (right column) is greater when folding with the true period (first row).

In Section 2 we reviewed the quadratic MI estimators based on the Euclidean distance (Equation (3)) and the Cauchy–Schwarz (CS) divergence (Equation (8)). These estimators

[8] The convolution of two WC kernels is a WC kernel.

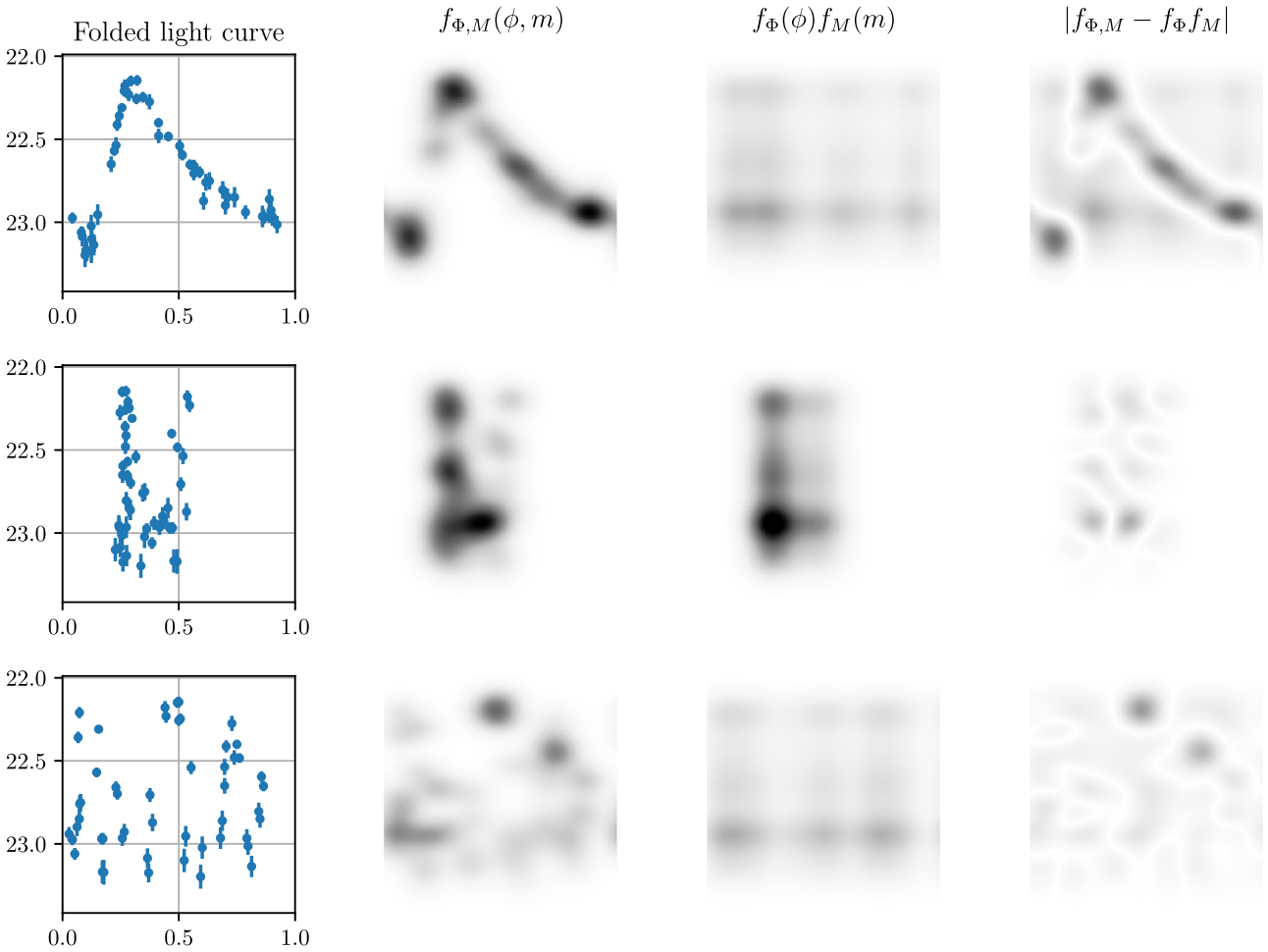

Figure 2. Folded light curve of a synthetic ab-type RR Lyrae light curve (first column). The light curve is folded using its real period (first row), the sidereal day (second row), and a random period (third row). The second and third columns show the joint PDF and the product of the marginal PDFs of the phases and magnitudes of the light curve. The fourth column is the absolute value of the difference between the joint and marginals. A correctly folded light curve produces a large statistical distance between the joint and marginals.

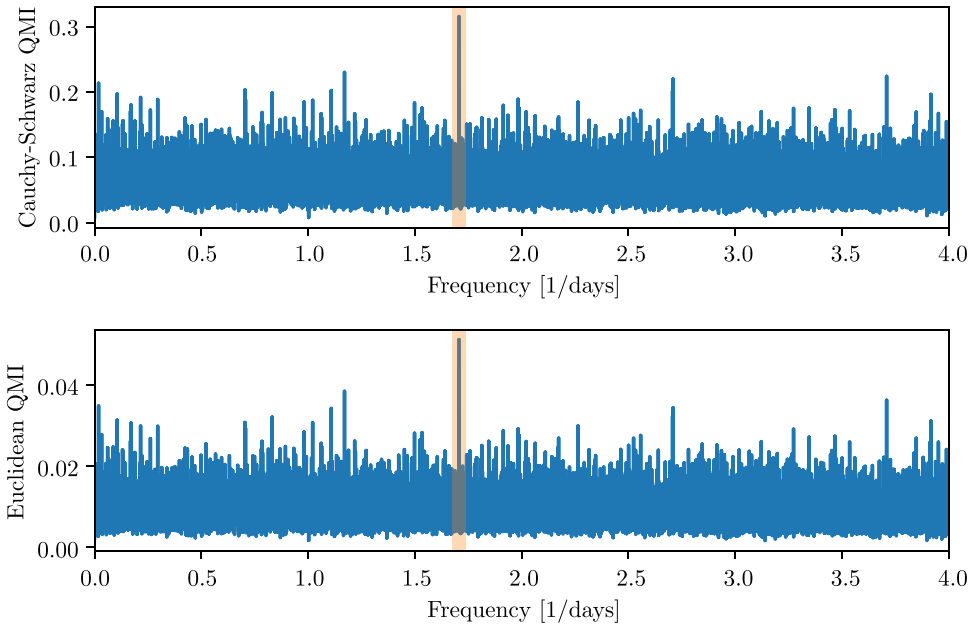

Figure 3. Cauchy–Schwarz and Euclidean QMI estimators as a function of the trial frequency. The underlying period of the light curve is shaded and it corresponds to the highest peak in both cases.

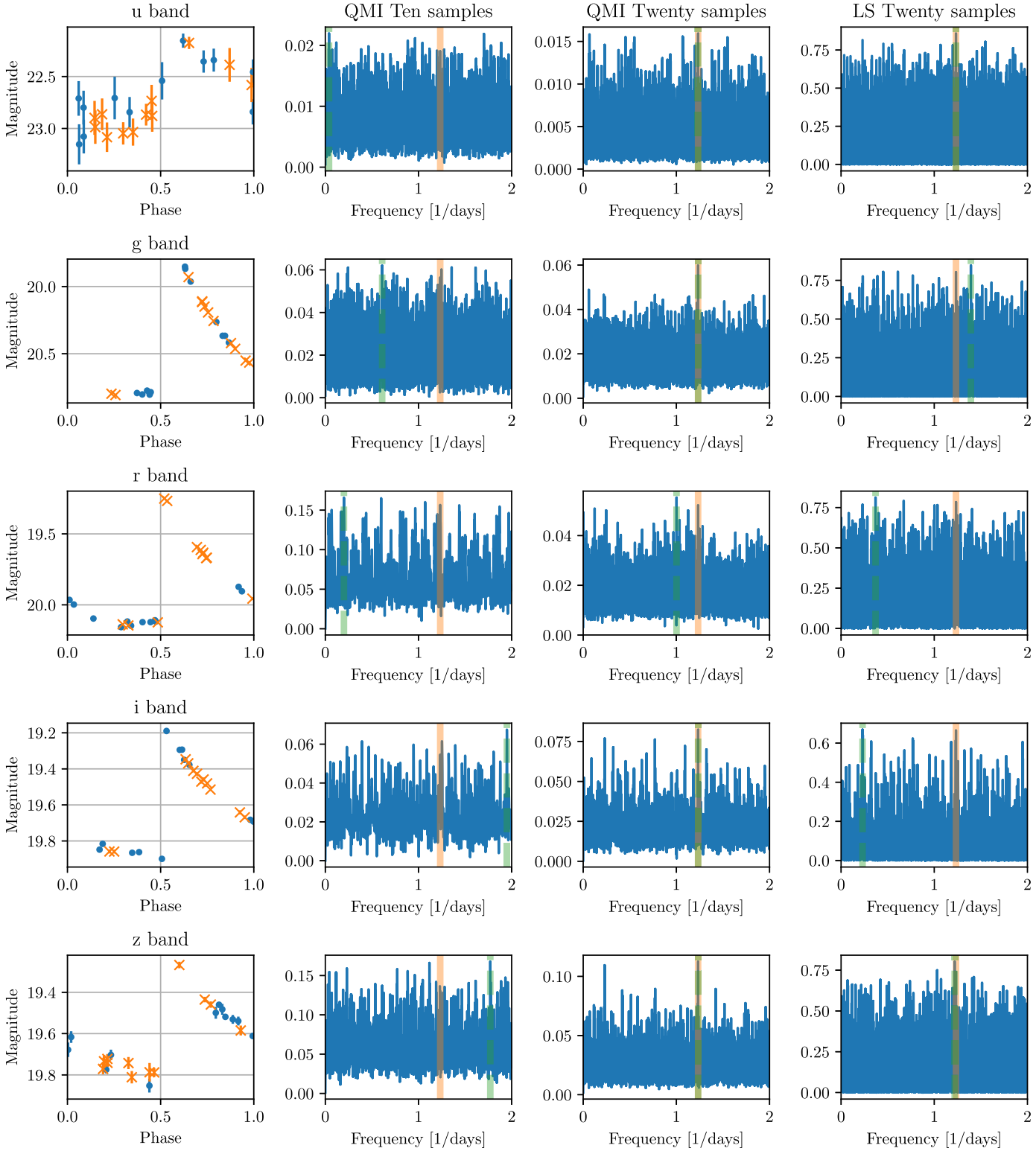

Figure 4. Rows corresponds to different bands (*ugriz*) of the same synthetic LSST RR Lyrae light curve. The first column shows the light curve folded with its true period, where blue dots and orange crosses correspond to the first 10 and second 10 samples per band, respectively. The second and third columns correspond to the $\mathrm{QMI}_{\mathrm{ED}}$ using 10 samples (dots), and 20 samples (dots and crosses), respectively. For reference we include the Lomb–Scargle periodogram (twenty samples) in the fourth column. The true period is shaded with a solid orange line, while the maximum of the periodogram is shaded with a dashed green line.

interest us because they are robust dependency measures and are computed directly from the data, bypassing the estimation of PDFs. Computing these estimators requires calculating the information potentials given by Equations (4)–(6). If we use the Gaussian kernel for the magnitudes and the WC kernel for phases we obtain

$$\mathrm{IP}_M = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2h_m^2+\sigma_i^2+\sigma_j^2}}(m_i - m_j), \tag{12}$$

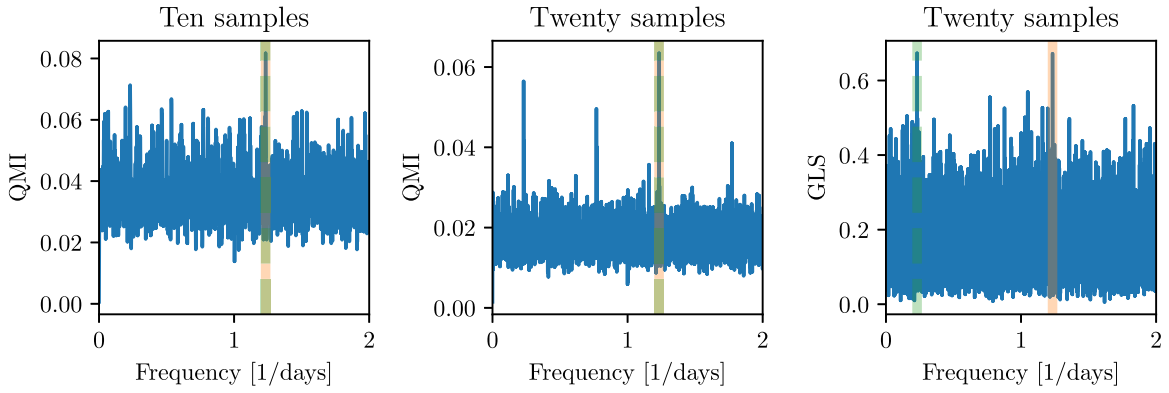

Figure 5. Average QMI across five bands for the example shown in Figure 4. In both cases the true period is correctly detected. For reference we include the multiband LS periodogram for the 20-sample case.

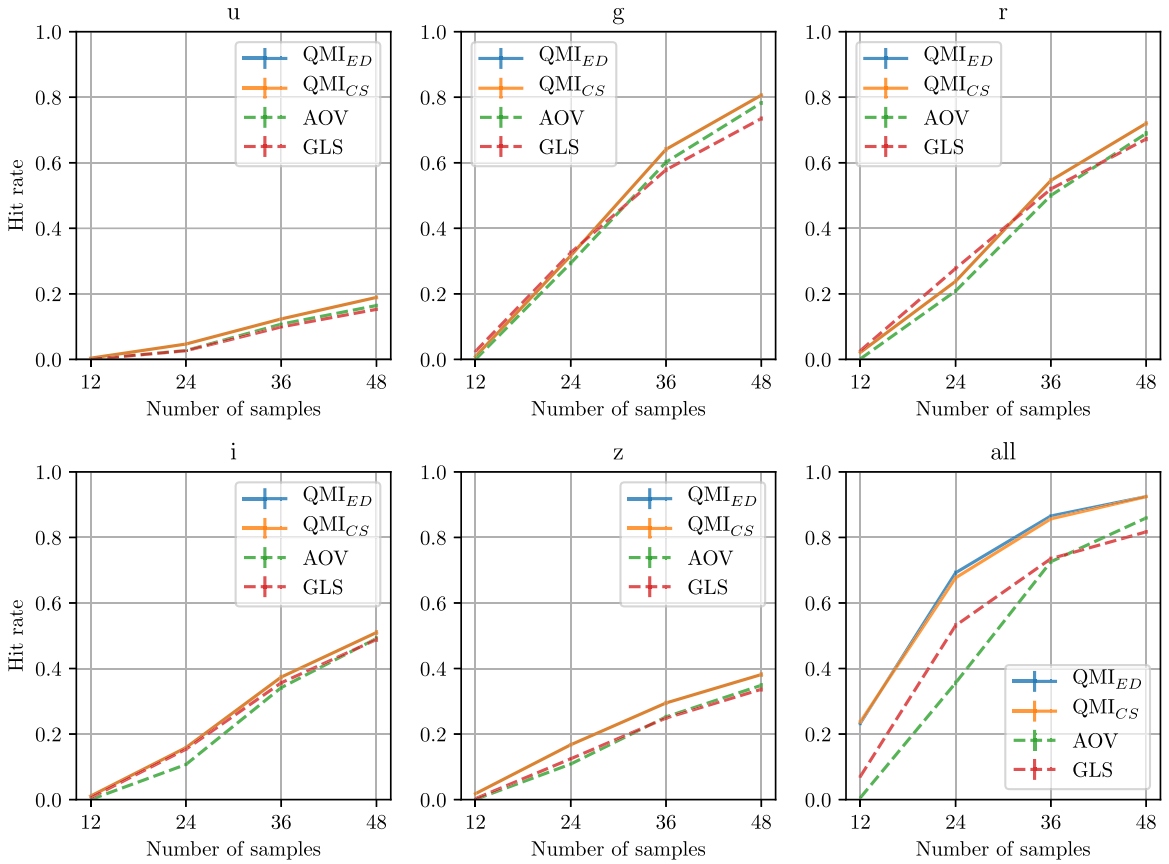

Figure 6. Average hit rate for different period detection methods on each band (*ugriz*) as a function of the number of samples per band. The lower right plot corresponds to the periodograms using all the bands. Each dot is an average of 10,000 synthetic ab-type RR Lyrae light curves.

$$\mathrm{IP}_{\Phi} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N}\mathrm{WC}_{2h_{\phi}}(\phi_i - \phi_j), \tag{13}$$

$$\mathrm{IP}_{\Phi,M} = \frac{1}{N^2}\sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2h_m^2+\sigma_i^2+\sigma_j^2}}(m_i - m_j)\mathrm{WC}_{2h_{\phi}}(\phi_i - \phi_j), \tag{14}$$

and

$$\mathrm{IP}_{\Phi\times M} = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{N}\sum_{j=1}^{N} G_{\sqrt{2h_m^2+\sigma_i^2+\sigma_j^2}}(m_i - m_j)\right) \times\left(\frac{1}{N}\sum_{j=1}^{N}\mathrm{WC}_{2h_{\phi}}(\phi_i - \phi_j)\right), \tag{15}$$

where the Gaussian kernel is used for the magnitudes and the wrapped Cauchy kernel is used for the phases. Through these potentials we restate the QMI estimators as

$$\mathrm{QMI}_{\mathrm{ED}}(\Phi, M) = \mathrm{IP}_{\Phi,M} - 2\mathrm{IP}_{\Phi\times M} + \mathrm{IP}_{\Phi}\mathrm{IP}_{M}, \tag{16}$$

and

$$\mathrm{QMI}_{\mathrm{CS}}(\Phi, M) = \log \mathrm{IP}_{\Phi,M} - 2\log \mathrm{IP}_{\Phi\times M} + \log \mathrm{IP}_{\Phi} + \log \mathrm{IP}_{M}, \tag{17}$$

respectively.

The period of a light curve is estimated by maximizing the QMI for a range of trial periods. This yields a QMI periodogram. As an example, in Figure 3 we compute the

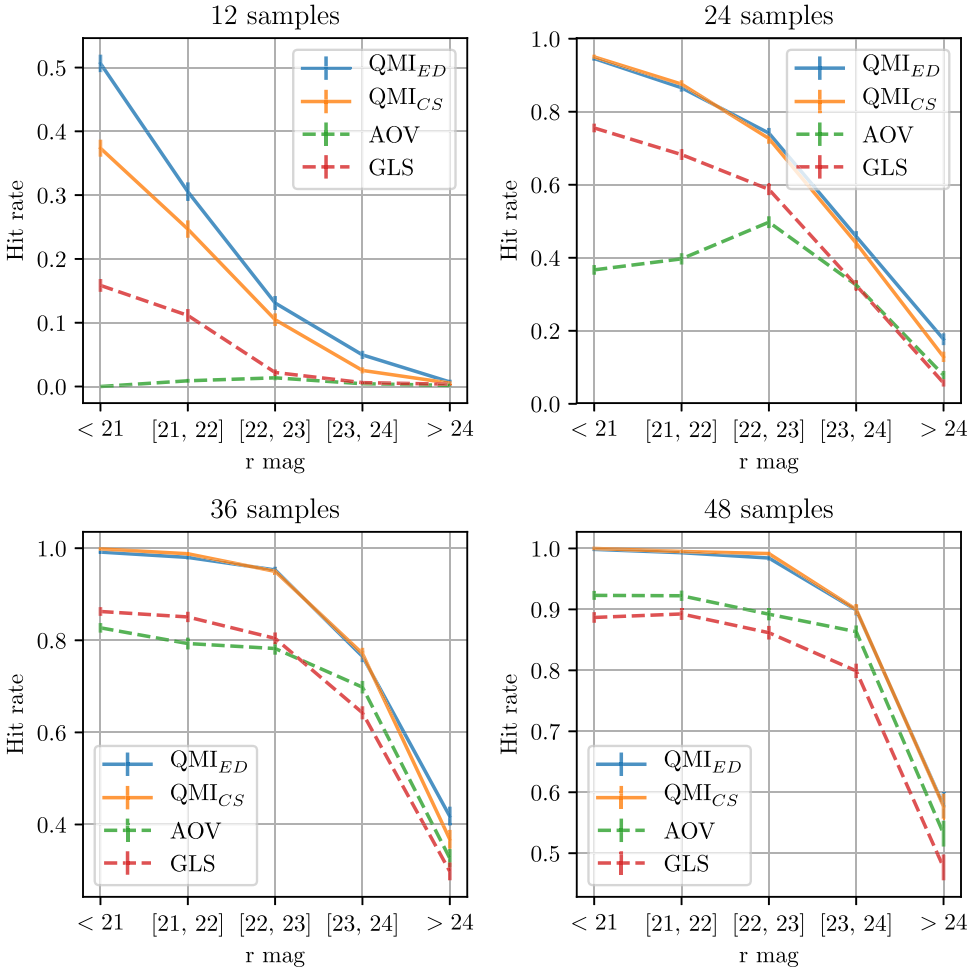

Figure 7. Average hit rate for different period detection methods as a function of the *r*-band magnitude. Each plot corresponds to a different light curve length. All light curves correspond to ab-type RR Lyrae. LSST will produce an average of 18.4 samples per year in the *r*-band.

$\mathrm{QMI_{CS}}$ and $\mathrm{QMI_{ED}}$ and plot them as a function of frequency for the same light curve used to obtain Figure 2. In both cases the underlying period corresponds to the global maximum. In the following section we discuss how to apply the QMI in multiband light curves.

3.3. Period Estimation in Multiband LSST Light Curves Using Quadratic MI

LSST light curves are characterized for being randomly and sparsely sampled. Methods for period detection that do not aggregate data from all the available bands are likely to fail, especially when few samples per band are available. In this section we show that the QMI periodogram can be easily extended to the case of multiband light curves. An efficient way to take advantage of the multiple bands is simply to combine the QMI obtained for every single band. But, an average QMI periodogram requires the individual periodograms to be on the same scale. As explained in Principe (2010) the QMI lacks a consistent absolute interpretation because it depends on its parameters, the kernel bandwidths. From Principe (2010) we extract the following conditions regarding comparisons between QMI values: (a) the kernel bandwidth has to be selected to be proportional to the dynamic range of the data and (b) the kernel bandwidth has to be a function of the number of samples and the QMI has to be normalized by its upper bound. The upper bound of the QMI estimators will be studied in the future. In the following experiments we use the same number of samples per band, hence the upper bounds can be ignored.

In our case we have two parameters h_ϕ and h_m. The former is associated to the phases that are always constrained to [0, 2π], i.e., the dynamic range of this variable is fixed. QMI is not too sensitive to h_ϕ as long as it is not extremely small or large. We have found empirically that $h_\phi = 1$ is a good choice and we keep it constant to make comparisons between QMI values easier. The second bandwidth h_m is more difficult to set, as the dynamic range of the magnitudes is not known a priori. We consider the plug-in rule from Silverman (1986),

$$h_m = 0.9 \cdot \min(\sqrt{\mathrm{VAR}[m]},\ \mathrm{IQR}[m]/1.349) \cdot N^{-1/5}, \quad (18)$$

where VAR[m] is the variance of the magnitudes, IQR[m] is the interquartile range of the magnitudes, and N is the number of samples. To avoid overestimation of h_m we use the weighted versions of variance and IQR, with weights $w_i = \sigma_i^{-2}$, $i = 1, \ldots, N$. Equation (18) complies with the conditions mentioned before. We also explored the Sheather–Jones recurrent estimator (Sheather & Jones 1991) and the more recently proposed diffusion estimator (Botev et al. 2010), but their performance was not better than Equation (18) and their

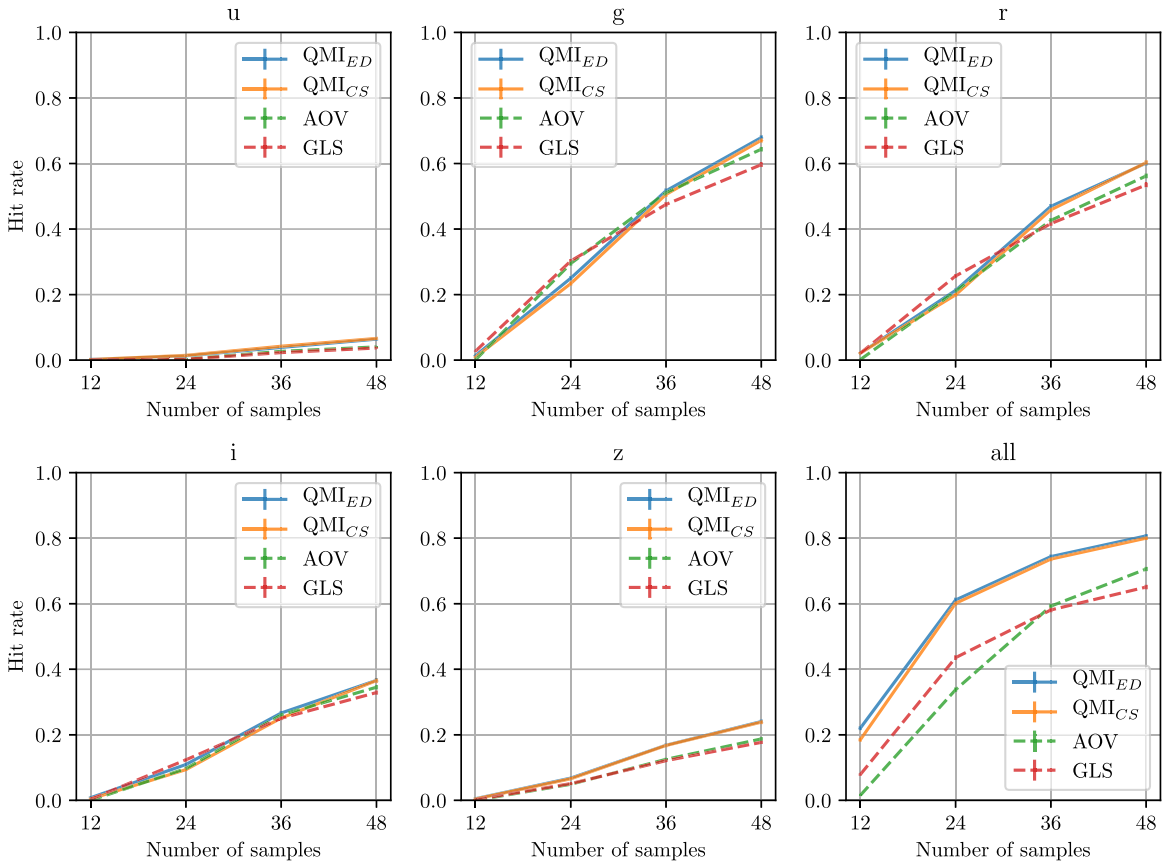

Figure 8. Average hit rate for different period detection methods on each band (*ugriz*) as a function of the number of samples per band. The lower right plot corresponds to the periodograms using all the bands. Each dot is an average of 10,000 synthetic c-type RR Lyrae light curves.

computational cost is higher. In the future, local plug-in estimators will be tested.

If Equation (18) is used for every single band, then the QMI periodograms can be averaged. Figure 4 shows the QMI periodograms for the *ugriz* bands of a synthetic RR Lyrae light curve. The first column corresponds to the folded light curve. The second and third columns are the Euclidean QMI periodograms using the first 10 points and the first 20 points, respectively. Note how, using 10 points, the true period cannot be found in any of the bands. When using twenty points the period is found in the *g* and *r* bands only. For reference, we show the LS periodogram for the case of 20 points in the fourth column. Figure 5 shows the averaged Euclidean QMI for 10 and 20 samples per band. In both cases the true period corresponds to the global maximum of the periodogram. This shows that even if the period is not the maximum in the single-band periodograms it can still be found in the combination. The explanation for this is that the true period is likely to appear in all bands, although not necessarily as a high peak. Spurious periods due to sampling will not be shared between bands and are de-emphasized in the average periodogram. In the next section we show through extensive experiments that a large gain in performance can be obtained by combining the QMI periodograms.

4. Results

In this section we test the proposed method using synthetic multiband LSST light curves. We generate variable star light curves of type ab RR Lyrae (RRab), type c RR Lyrae (RRc), and Cepheids (CEPH) following the procedure described in Section 3.1. We consider five bands (*ugriz*) in order to match the original bands of the variability templates. Ten noisy realizations are obtained per generated light curve. This yields a total of 10,000 light curves per variability type.

The QMI estimators are compared to the multiband generalizations of the LS and AoV periodograms. The multiband QMI and AoV methods are implemented in Cython[9] and distributed as a python package called `P4J`.[10] For the multiband LS we use the `gatspy`[11] python package. All periodograms are run from 0.0 to 4.0 [1/days], with a step size of 10^{-4} [1/days]. The kernel bandwidth h_m is set using Equation (18) and $h_\phi = 1$ in all the experiments. The AOV and generalized LS (GLS) implementations allow for multiharmonic models. We present results using three harmonics,[12] as this configuration obtains a higher hit rate. For the multiharmonic GLS, a conservative regularization term was considered to avoid singularities. All routines are single-core and parallelization is done at a time-series level. Details on how to set the periodograms using `P4J` are given in Appendix C.

We consider the period associated with the global maximum of the periodogram as the detected period P_D. The ability to recover the true period P_T is measured in terms of hit rate (HR).

[9] http://cython.org/

[10] Available at github.com/phuijse/P4J and through PyPI.

[11] http://www.astroml.org/gatspy/

[12] Truncated Fourier series model with fundamental frequency f_0, 2 times f_0, and 3 times f_0 terms.

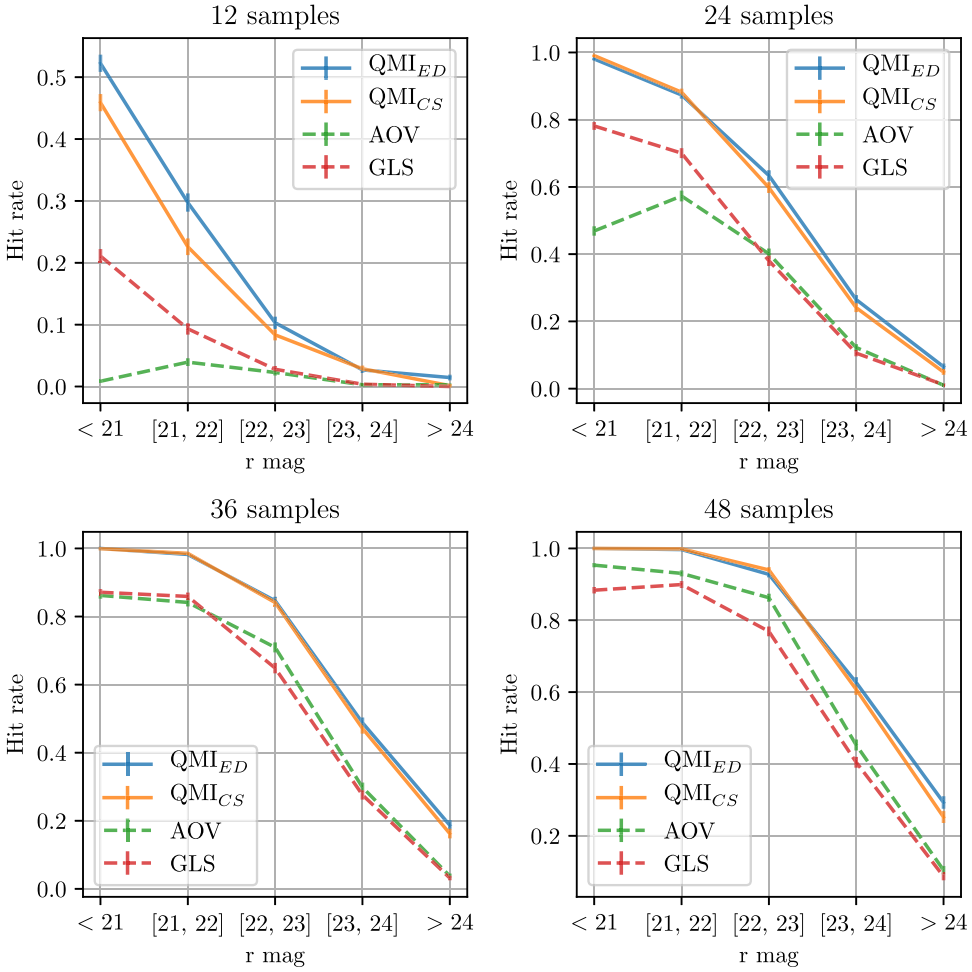

Figure 9. Average hit rate for different period detection methods as a function of the *r*-band magnitude. Each plot corresponds to a different light curve length. All light curves correspond to c-type RR Lyrae.

We follow Oluseyi et al. (2012) and define HR as the number of cases where

$$e_{\text{rel}} = \frac{|P_D - P_T|}{P_T^2} < \text{tol},$$

divided by the total number of light curves. Tolerance (tol) decreases as a function of P_T. Detecting a harmonic or an alias of the true period is considered a failure. In all experiments the tolerance is set to $1e-3$.

In addition to the multiband periodograms, we also evaluate the results on each of the five *ugriz* bands independently. The robustness against the length of the light curves, i.e., the amount of samples required to detect the period, is also studied. Each light curve is evaluated using its first 12, 24, 36, and 48 samples, respectively. Figure 6 shows the results of this experiment in the case of RRab templates. Each plot corresponds to one of the *ugriz* bands, while the lower right plot corresponds to the multiband result. In the single-band tests all methods yield a similar performance, but in the multiband test the two information theoretic estimators outperform second-order methods. In all tests the difference between the Euclidean and Cauchy–Schwarz QMI hit rates is less than 1% (their difference is barely noticed in the plots). All methods benefit when aggregating data from the five bands with respect to the best single-band result. Information theoretic (IT) methods yield the largest absolute increase in hit rate when aggregating the data (up to 40%). Single-band best results are obtained in *g*, which is expected, as RR Lyrae are inherently more variable in this filter.

Figure 7 shows the multiband hit rates averaged over different ranges of the magnitude in the *r*-band. Each plot corresponds to a different light curve length (sample size). The signal-to-noise ratio (S/N) decreases with magnitude. As expected, HR increases with light curve length and decreases with magnitude for all methods. Both QMI estimators have a similar performance, except in the 12-sample case where the Euclidean QMI performs better than CS QMI, suggesting that the former might be more robust to low sample size. QMI estimators outperform second-order methods in all cases. This is more noticeable for shorter light curves, with the absolute HR margin growing from 10% to 30% (Euclidean QMI versus multiband GLS). This shows that QMI estimators can detect the true period faster (in survey time) than second-order methods. The multiband AoV performs slightly better than the multiband GLS in the 48-sample case. On the other hand, GLS performs considerably better than AoV at shorter light curve lengths and brighter magnitudes. In the 24-sample case the performance of AoV decreases with S/N, and AoV tends to recover the harmonics of the true period more frequently in this regime. In all cases, the difference in hit rate between methods decreases when approaching the *r*-band 5σ limit of 24.5.

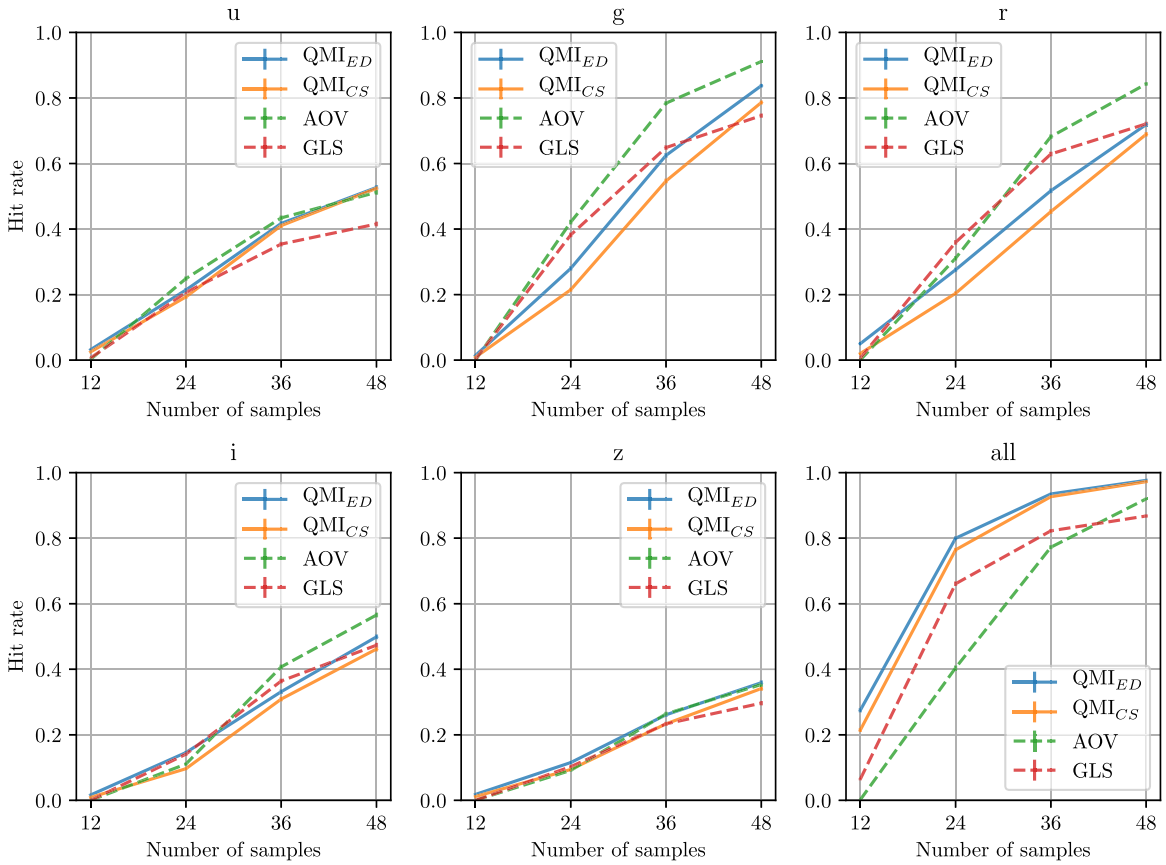

Figure 10. Average hit rate for different period detection methods on each band (*ugriz*) as a function of the number of samples per band. The lower right plot corresponds to the periodograms using all the bands. Each dot is an average of 10,000 synthetic Cepheid light curves.

Figure 8 shows the results obtained using the RRc templates. Again we can see that single-band results are only marginally different between methods. As with RRab, single-band best results are obtained in the *g*-band. The lower right plot shows the multiband results. In the multiband case the QMI based methods see an increase in hit rate between 15% and 40% with respect to single-band best results, outperforming second-order methods. Figure 9 shows the multiband results in more detail. The QMI methods perform better than second-order methods in all cases, and the difference in HR grows for shorter light curves. Both QMI estimators perform similarly, except in the 12-sample case where the Euclidean estimator performs better.

Figure 10 shows the results obtained with the CEPH templates. QMI estimators outperform their competitors in the multiband case, but perform worse in the single *g*- and *r*-band cases. Interestingly, there is little gain when aggregating bands for the AoV periodogram. On the other hand, QMI methods see an absolute increase in HR from 10% to 50%. The Euclidean QMI performs slightly better than the CS QMI in all tests. Figure 11 shows the multiband results in greater detail. The Euclidean QMI performs better than the CS QMI when sample size decreases. QMI methods perform better than second-order methods, and again this is more evident when sample size decreases (shorter light curves). The GLS performs better than AoV, except in the 48-sample case. Once again, we note a strong tendency of AoV to recover a harmonic of the true period for brighter magnitudes and smaller sample sizes.

Table 1 shows the computational time required to calculate a complete periodogram using our library, on time-series of different lengths (time is an average of 100 repetitions). Computational time is measured on an Intel i5-4460 CPU at 3.20 GHz. Computational time is on the same order of magnitude, but due to the increased computational complexity of QMI estimators, they scale worse with the number of samples. We are working on approximations of the information potential estimator to reduce the computational time in the case of dense light curves.

5. Conclusions and Future Work

In this paper we have proposed an estimator of the QMI for estimating periods in light curves. By maximizing the QMI between the phases and the magnitudes of a light curve, the underlying period can be estimated. Contrary to second-order methods, the QMI extracts information from the whole PDF, is not restricted to linear relations between variables and is more robust to non-Gaussian (heavy-tailed) noise. Efficient Cython implementations of the methods presented in this paper are freely available through `github` and `PyPI`.

We have applied the QMI for period estimation in sparse multiband light curves of variable stars generated with the LSST simulation tools. The OpSim and CatSim tools allow us to build a database of realistic synthetic light curves with multiband pointings, cadence, and noise distribution as expected by the LSST. Our results show that the QMI outperforms the multiband generalizations of the LS and AoV periodograms for all variability types. The QMI is efficient at aggregating data from several sparsely sampled bands, presenting an absolute hit rate increase up to 50% with respect to the best single-band results. We have

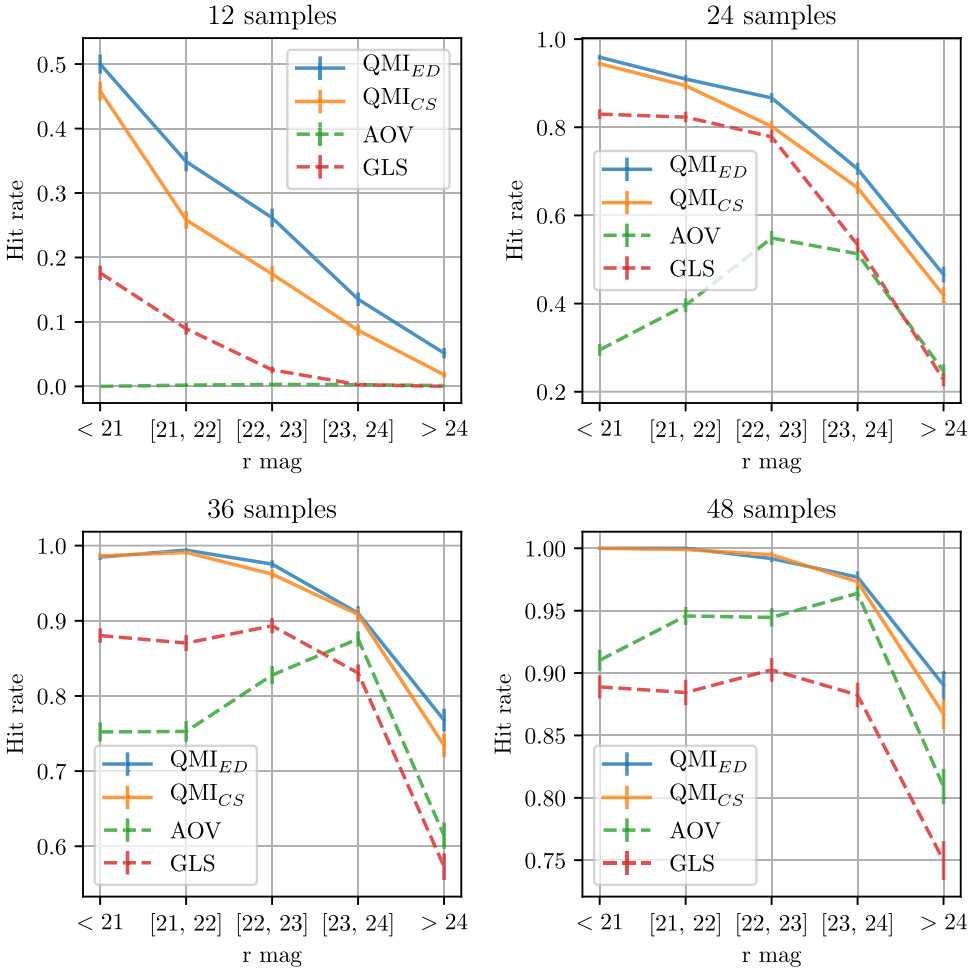

Figure 11. Average hit rate for different period detection methods as a function of the *r*-band magnitude. Each plot corresponds to a different light curve length. All light curves correspond to Cepheids.

Table 1
Computational Time (seconds) Using `P4J` to Compute a Periodogram on a Single Time-series (Averaged over 100 Repetitions)

Number of Samples	Euclidean QMI	AOV
12	0.074 ± 0.012	0.078 ± 0.011
24	0.185 ± 0.013	0.123 ± 0.017
36	0.298 ± 0.031	0.167 ± 0.018
48	0.492 ± 0.028	0.225 ± 0.025

Note. AoV is computed with three harmonics.

observed that the performance gap with second-order methods is more noticeable when the length of the light curve decreases (smaller sample sizes). The multiband QMI is more robust to noise and it can detect the true period faster (survey time) than second-order methods.

In our proposition we combine the single-band QMI periodograms, allowing us to detect the period even when individual periodograms cannot. However, this does not exploit the interaction between bands directly. We recognize an extension to this proposition that involves calculating QMI cross-products between bands. Although we note that this would increase the computational complexity, we have preliminary results that show that using these cross-products allows for even more robustness against low sample size and noise.

One weakness of the proposed estimator is that it scales quadratically with the number of samples, making it expensive to compute for dense data (more than 100 samples per band). We plan to include a Fast-Gauss transform implementation of the information potential estimator in our library in the near future to partially solve this. We will also study new ways to estimate MI that might be more efficient, such as the propositions in Giraldo et al. (2015).

Future work should also include a more profound analysis of the differences between the Euclidean and Cauchy–Schwarz QMI, and should study the upper bounds of these estimators. We expect to develop relative QMI estimators that allow us to compare results between different light curves, which is key to developing statistical criteria based on the QMI distribution to avoid the cost of case-by-case bootstrap analysis. In this work we focused on quadratic (order 2) entropy and MI estimators. In the future we will test MI estimators of different orders and study their properties.

Pablo Huijse (P.H.) acknowledges support from FONDECYT through grant No. 1170305 and CONICYT through grant PAI79170017. Pablo A. Estévez (P.E.) acknowledges support from FONDECYT through grant No. 1171678. Francisco Förster (F.F.) acknowledges support from FONDECYT through grant No. 3110042 and from Basal Project PFB-03. P.H., P.E., and F.F. acknowledge support from CONICYT through the Programme of International Cooperation project

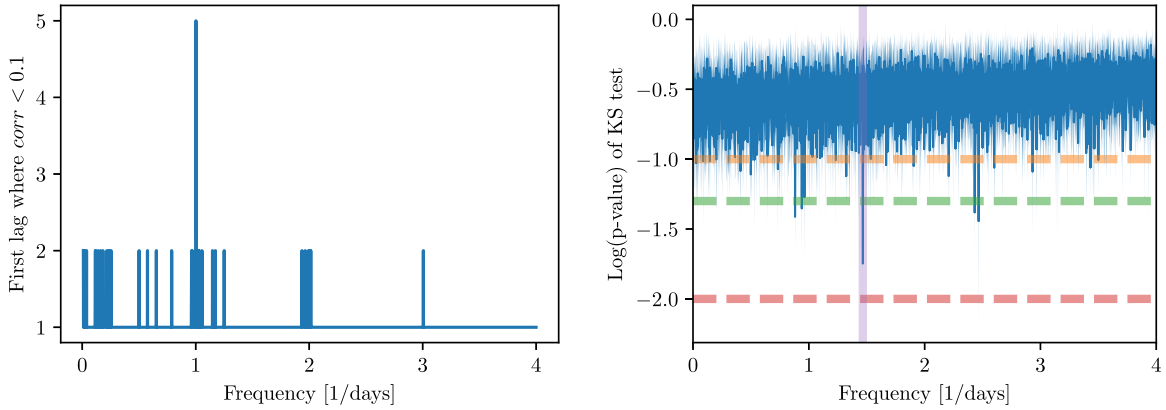

Figure 12. (Left) First lag for which the slotted autocorrelation is less than 0.1 as a function of frequencies (foldings). In most cases autocorrelation has reached this value at the first lag. The harmonics of the sidereal day are a notable exception of this. (Right) Average logarithm of the *p*-value of a KS test across data partitions for different frequencies (foldings). The shaded area corresponds to the error bars. The dotted line marks the 10%, 5%, and 1% significance. In most cases the null hypothesis cannot be rejected with a significance smaller than 10%. A notable exception is the period of the light curve (shaded), which rejects the null with a significance smaller than 5%.

DPI20140090 and from the Chilean Ministry of Economy, Development, and Tourism's Millennium Science Initiative through grant IC12009, awarded to The Millennium Institute of Astrophysics, MAS. Andrew J. Connolly acknowledges partial support by the U.S. Department of Energy, Office of Science, under award No. DE-SC-0011635, from the DIRAC Institute and the LSST, and from the NSF through awards AST-1409547 and AST-1715122. This research was partially supported by the supercomputing infrastructure of the NLHPC (ECM-02). Part of this work was done under the Harvard-Chile data science school.

Appendix A
Information Theoretic Learning

ITL (Principe et al. 2000; Principe 2010) is a framework to bring information theoretic criteria into machine-learning (ML) methods. Traditionally, ML methods are trained via optimizing a second-order loss function, e.g., the mean square error and correlation. In ITL these quantities are replaced by information theoretic criteria that describe the PDF, e.g., entropy and MI. ITL criteria extract more information from data, improving the performance of the methods. By going beyond the second-order moment ITL criteria gain robustness in realistic scenarios where the Gaussianity assumption does not hold, e.g., under the presence of heavy-tailed noise and outliers.

In ITL a strong emphasis is given to the estimation of these quantities directly from data in a non-parametric way. As an example, consider the ITL estimation of Renyi's second-order generalization H_2 (X) of Shannon's entropy (Principe 2010) of a continuous RV defined as

$$H_2(X) = -\log \int f_X(x)^2 \, dx, \tag{19}$$

where $f_X(x)$ is the RV's PDF. Assuming that we have $\{x_i\}_{i=1,\ldots,N}$ realizations of the RV, its PDF can be computed using a kernel density estimator (KDE)

$$f_X(x) = \frac{1}{N}\sum_{i=1}^{N} G_h(x - x_i) = \frac{1}{N\sqrt{2\pi}\,h}\sum_{i=1}^{N} \exp\left(\frac{\|x - x_i\|^2}{2h^2}\right), \tag{20}$$

where $G_h(\cdot)$ is the Gaussian kernel with bandwidth h. By replacing Equation (20) in Equation (19) and then using the Gaussian convolution property[13] we obtain

$$\begin{aligned} H_2(X) &= -\log \frac{1}{N^2} \int \sum_{i=1}^{N}\sum_{j=1}^{N} G_h(x - x_i) G_h(x - x_j)\, dx, \\ &= -\log \frac{1}{N^2} \sum_{i=1}^{N}\sum_{j=1}^{N} G_{\sqrt{2}h}(x_i - x_j) = -\log \mathrm{IP}_X, \end{aligned} \tag{21}$$

where IP_X is the Information Potential (IP), an estimator of the expected value of the PDF of X (Principe 2010), although it is estimated directly from the data samples, bypassing the estimation of the PDF. Other symmetric and translation-invariant kernels can be used, but it is convenient to use kernels that are closed under convolution.

Appendix B
Testing the iid Hypothesis in the Phase Diagram

For the estimation of MI we assume that the samples from our joint PDF $f(\Phi, M)$ are independent and identically distributed, i.e., all realizations come from the same continuous distribution and no serial correlations exist between realizations. Light curves are time-series, so we expect to find temporal correlations, although sampling is pseudo-random and does not obey Nyquist's theorem (Eyer & Bartholdi 1999). Phase is a function of time and period, and several periods are tested per light curve. If the period is not related to the underlying periodicity of the data, the phase diagram is filled uniformly and serial correlations in the joint space are broken. This is shown in Figure 12 for a periodic light curve. The plot on the left shows that for most frequencies (foldings), the slotted autocorrelation drops very fast. Assuming that the light curve is stationary (no trends) we can partition the phase-magnitude space in equally sized bins and compute a two-dimensional Kolmogorov-Smirnov (KS) test (Fasano & Franceschini 1987)[14] to consider the null hypothesis that the binned distributions are equal. Figure 12 (right) shows the logarithm of the average *p*-value as a function of frequency. In the majority of cases we cannot reject the null at 10% significance. We evaluate a subset of 1000 light curves with different periods and found these results to be consistent. This

[13] The convolution of two Gaussian functions is also a Gaussian.
[14] Implemented at github.com/syrte/ndtest.

explains why MI works well in practice when applied to the folded data.

Appendix C
Using the P4J Python Library

Listing 1 demonstrates how to compute the QMI periodogram using the python P4J library (github.com/phuijse/P4J). The data used for this paper and multiband evaluation scripts based on P4J can be found at github.com/phuijse/LSST_simulations.

Listing 1. "P4J demonstration"

```
import P4J
# Assuming that mjd, mag and err are N-length numpy arrays
# Using the Euclidean QMI periodogram
my_per=P4J.periodogram(method="QMIEU", debug=False)
# By default silverman's rule is used and hp=1
my_per.set_data(mjd, mag, err, whitten=False)
my_per.frequency_grid_evaluation(fmin=0.0, fmax=4.0, fresolution=1e-4)
# You may want to finetune the estimations around the maxima
my_per.finetune_best_frequencies(fresolution=1e-5, n_local_optima=10)
# If you want the whole periodogram
freq, per~=~my_per.get_periodogram()
# If you only want to retrieve the best frequencies
fbest, pbest~=~my_per.get_best_frequencies()
# Other available methods are QMICS, MHAOV, PDM1 and LKSL
```

ORCID iDs

Pablo Huijse https://orcid.org/0000-0003-3541-1697
Pablo A. Estévez https://orcid.org/0000-0001-9164-4722
Francisco Förster https://orcid.org/0000-0003-3459-2270
Andrew J. Connolly https://orcid.org/0000-0001-5576-8189

References

Abell, P. A., Allison, J., Anderson, S. F., et al. 2009, arXiv:0912.0201
Akritas, M. G. 1997, in Statistical Challenges in Modern Astronomy II (Berlin: Springer), 105
Botev, Z. I., Grotowski, J. F., Kroese, D. P., et al. 2010, AnSta, 38, 2916
Clarke, D. 2002, A&A, 386, 763
Connolly, A. J., Angeli, G. Z., Chandrasekharan, S., et al. 2014, Proc. SPIE, 9150, 14
Delgado, F., Saha, A., Chandrasekharan, S., et al. 2014, Proc. SPIE, 9150, 915015
Eyer, L., & Bartholdi, P. 1999, A&AS, 135, 1
Fasano, G., & Franceschini, A. 1987, MNRAS, 225, 155
Feigelson, E. D., & Babu, G. J. 2012, Significance, 9, 22
Giraldo, L. G. S., Rao, M., & Principe, J. C. 2015, ITIT, 61, 535
Graham, M. J., Drake, A. J., Djorgovski, S., et al. 2013a, MNRAS, 434, 3423
Graham, M. J., Drake, A. J., Djorgovski, S., Mahabal, A. A., & Donalek, C. 2013b, MNRAS, 434, 2629
Gray, R. M. 2011, Entropy and Information Theory (New York: Springer Science & Business Media)
Huijse, P., Estevez, P. A., Protopapas, P., Principe, J. C., & Zegers, P. 2014, IEEE Computational Intelligence Magazine, 9, 27
Huijse, P., Estevez, P. A., Protopapas, P., Zegers, P., & Principe, J. C. 2012, ITSP, 60, 5135
Ivezić, V., Tyson, J. A., Acosta, E., et al. 2008, arXiv:0805.2366
Jammalamadaka, S., Sengupta, A., & Sengupta, A. 2001, Topics in Circular Statistics, Series on Multivariate Analysis (Singapore: World Scientific) https://books.google.cl/books?id=sKqWMGqQXQkC
Khan, S., Bandyopadhyay, S., Ganguly, A. R., et al. 2007, PhRvE, 76, 026209
Kraskov, A., Stögbauer, H., & Grassberger, P. 2004, PhRvE, 69, 066138
Mondrik, N., Long, J. P., & Marshall, J. L. 2015, ApJL, 811, L34
Moon, Y.-I., Rajagopalan, B., & Lall, U. 1995, PhRvE, 52, 2318
Oluseyi, H. M., Becker, A. C., Culliton, C., et al. 2012, AJ, 144, 9
Palmer, D. M. 2009, ApJ, 695, 496
Percy, J. R. 2007, Understanding Variable Stars (Cambridge: Cambridge Univ. Press)
Pont, F., Zucker, S., & Queloz, D. 2006, MNRAS, 373, 231
Principe, J. C. 2010, Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives (New York: Springer Science & Business Media)
Principe, J. C., Xu, D., Zhao, Q., & Fisher, J. W. 2000, The Journal of VLSI Signal Processing, 26, 61
Protopapas, P., Huijse, P., Estévez, P. A., et al. 2015, ApJS, 216, 25
Richards, J. W., Starr, D. L., Butler, N. R., et al. 2011, ApJ, 733, 10
Scargle, J. D. 1982, ApJ, 263, 835
Schwarzenberg-Czerny, A. 1996, ApJL, 460, L107
Sesar, B., Ivezic, Z., Grammer, S. H., et al. 2010, ApJ, 708, 717
Sheather, S. J., & Jones, M. C. 1991, Journal of the Royal Statistical Society. Series B (Methodological), 53, 683
Silverman, B. W. 1986, Density Estimation for Statistics and Data Analysis, Vol. 26 (Boca Raton, FL: CRC Press)
Stellingwerf, R. 1978, ApJ, 224, 953
Tyson, J. A., & Borne, K. D. 2012, in Future Sky Surveys: New Discovery Frontiers (London: Chapman and Hall), 161
VanderPlas, J. T., & Ivezic, Ž. 2015, ApJ, 812, 18
Xu, D. 1999, PhD thesis, Univ. Florida
Zechmeister, M., & Kürster, M. 2009, A&A, 496, 577
Zucker, S. 2016, MNRAS Letters, 457, L118