(§5.7.2), are good choices when the shape of underlying light curve cannot be approximated with a small number of Fourier terms.

### 10.3.4. Classification of periodic light curves

As illustrated in Fig 10.17, stellar light curves often have distinctive shapes (e.g., such as skewed light curves of RR Lyrae type ab stars, or eclipsing binary stars). In addition to shapes, the period and amplitude of the light curve also represent distinguishing characteristics. With large data sets, it is desirable and often unavoidable to use machine learning methods for classification (as opposed to manual/visual classification). In addition to light curves, other data such as colors are also used in classification.

As dicussed in Chapters 6 and 9, classification methods can be divided into supervised and unsupervised. With supervised methods we provide a training sample, with labels such as "RR Lyrae", "Algol type", "cepheid" for each light curve, and then seek to assign these labels to another data set (essentially, we ask "Find me more light curves such as this one in the new sample"). With unsupervised methods, we provide a set of attributes and ask if the data set displays clustering in the multi-dimensional space spanned by these attributes. As a practical example, here we discuss unsupervised clustering of variable stars with light curves found in the LINEAR data set, augmented with photometric (color) data from the SDSS and 2MASS surveys.

The Lomb-Scargle periodogram fits a single harmonic (eq. 10.23). If the underlying time series includes higher harmonics, a more general model than a single sinusoid should be used to better describe the data and obtain a more robust period, as discussed in the preceding section. As an added benefit of the improved modeling, the amplitudes of Fourier terms can be used to efficiently classify light curves, e.g. see [40; 28]. In some sense, fitting a low-$M$ Fourier series to data represents an example of the dimensionality reduction techniques discussed in Chapter 7. Of course, it is not necessary to use Fourier series and other methods have been proposed, such as direct analysis of folded light curves using PCA, see [17]. For an application of PCA to analyze light curves measured in several bandpasses simultaneously, see [54].

Given the best period, $P = 2\pi/\omega_0$, determined from the $M$-term periodogram $P_M(\omega)$ given by eq. 10.69 (with $M$ either fixed a priori, or determined in each case using BIC/AIC criteria), a model based on the first $M$ Fourier harmonics can be fit to the data

$$y(t) = b_0 + \sum_{m=1}^{M} a_m \sin(m\omega t) + b_m \cos(m\omega t). \tag{10.74}$$

Since $\omega_0$ is assumed known, this model is linear in terms of $(2M+1)$ unknown coefficients $a_j$ and $b_j$ and thus the fitting can be performed rapidly (approximate solutions given by eq. 10.72 are typically accurate enough for classification purposes).

Given $a_m$ and $b_m$, useful attributes for the classification of light curves are the amplitudes of each harmonic

$$A_m = (a_m^2 + b_m^2)^{1/2}, \tag{10.75}$$

and phases

$$\phi_m = \text{atan}(b_m, a_m), \tag{10.76}$$

with $-\pi < \phi_m \leq \pi$. It is customary to define zero phase to correspond to the maximum, or the minimum, of a periodic light curve. This convention can be accomplished by setting $\phi_1$ to the desired value (0 or $\pi/2$), and redefining phases of other harmonics as

$$\phi_m^0 = \phi_m - m\,\phi_1. \tag{10.77}$$

It is possible to extend this model to more than one fundamental period, for example, as done by Debosscher et al. in analysis of variable stars [18]. They subtract the best-fit model given by eq. 10.74 from data and recompute the periodogram to obtain next best period, find the best-fit model again, and then once again repeat all the steps to obtain three best periods. Their final model for a light curve is thus

$$y(t) = b_0 + \sum_{k=1}^{3} \sum_{m=1}^{M} a_{km} \left[\sin(m\omega_k t) + b_{km} \cos(m\omega_k t)\right], \tag{10.78}$$

where $\omega_k = 2\pi/P_k$. With three fixed periods, there are $6M + 1$ free parameters to be fit. Again, finding the best-fit parameters is a relatively easy linear regression problem. This and similar approaches to the classification of variable stars are becoming a standard in the field [2; 43]. A multi-staged tree-like classification scheme, with explicit treatment of outliers, seems to be an exceptionally powerful and efficient approach, even in the case of sparse data [42; 2].

Figures 10.20 and 10.21 show the results of a Gaussian mixture clustering analysis of the LINEAR data (see §6.3). The top panel of Figure 10.20 shows a 12-component Gaussian mixture fit to two data attributes, the $g-i$ color and $\log(P)$, the base-10 logarithm of the best Lomb-Scargle period in days. Out of the 12 clusters, five are significant (as measured by their weight and density), while the rest describe the background. Three clusters can be identified with various types of RR Lyrae stars (those with the longest periods), while the elongated sequence is mostly populated by various types of eclipsing binary stars, and the cluster with $\log(P) < -1$ is dominated by the so-called $\delta$ Scu and SX Phe variable stars [24]. The upper right panel shows the clusters in a different projection, $\log(P)$ vs. light curve amplitude. The top five clusters are still fairly well localized in this projection due to $\log(P)$ carrying significant discriminative power, but there is some mixing between the background and the clusters.

In order to better locate the clusters, we use all the parameters available in the data file. The clustering attributes include four photometric colors based on SDSS and 2MASS measurements ($u-g$, $g-i$, $i-K$, $J-K$) and three parameters determined from the LINEAR light curve data ($\log(P)$, amplitude, and light curve skewness). Fitting a 20-component Gaussian mixture to this seven-dimensional data set yields the clusters shown in the bottom panels of Figure 10.20. The clusters derived from all seven features are remarkably similar to the clusters derived from just two features: this shows that the additional data adds very little new information. Nevertheless, note that the fifth significant cluster is not $\delta$ Scu and SX Phe stars, but a horizontal feature that is likely due to unreliable period values. Figure 10.21 shows the locations of these clusters in the space of other attributes.

The means and widths of the distribution of points assigned to each cluster for the 7-attribute clustering are shown in the following table:

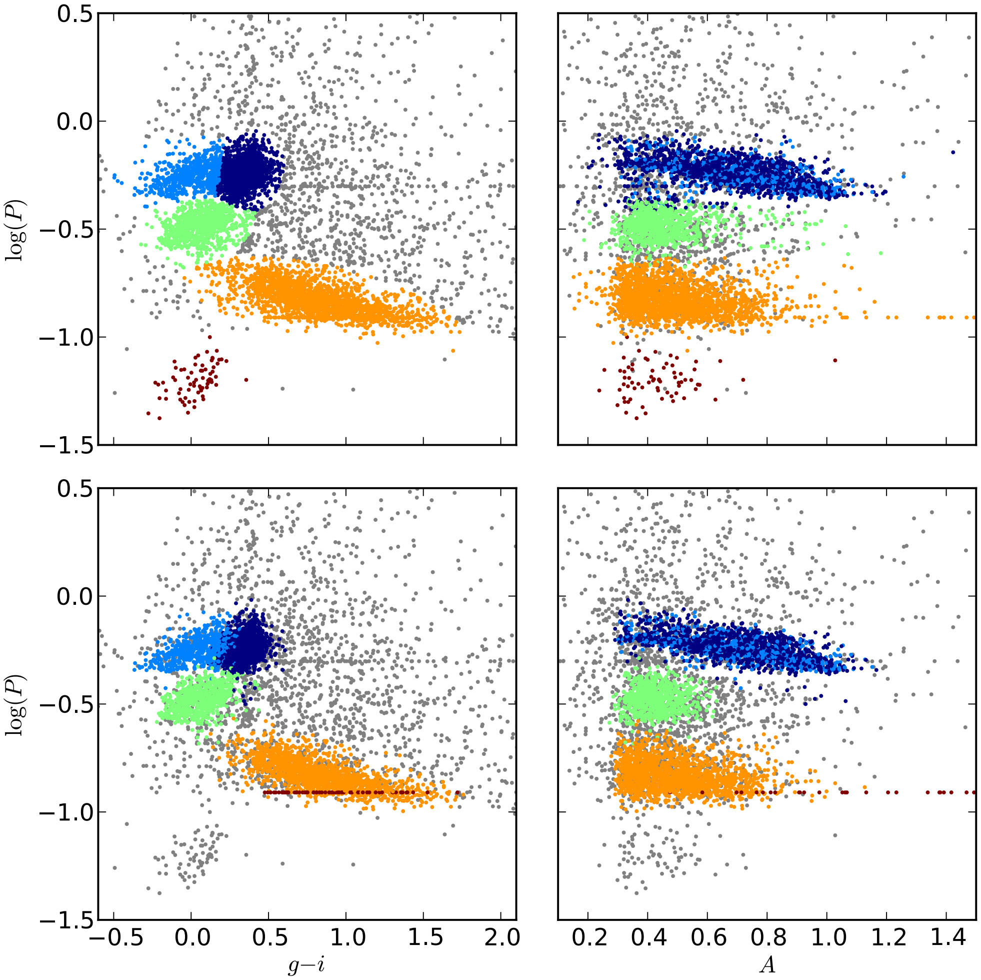

Figure 10.20.: Clustering analysis of periodic variable stars from the LINEAR data set. The top row shows clusters derived using two attributes ($g-i$ and $\log(P)$) and a mixture of 12 Gaussians. The colorized symbols mark the five most significant clusters. The bottom row shows analogous diagrams for clustering based on seven attributes (colors $u-g$, $g-i$, $i-K$, and $J-K$, $\log(P)$, light curve amplitude, and light curve skewness), and a mixture of 20 Gaussians. See Figure 10.21 for data projections in the space of other attributes for the latter case.

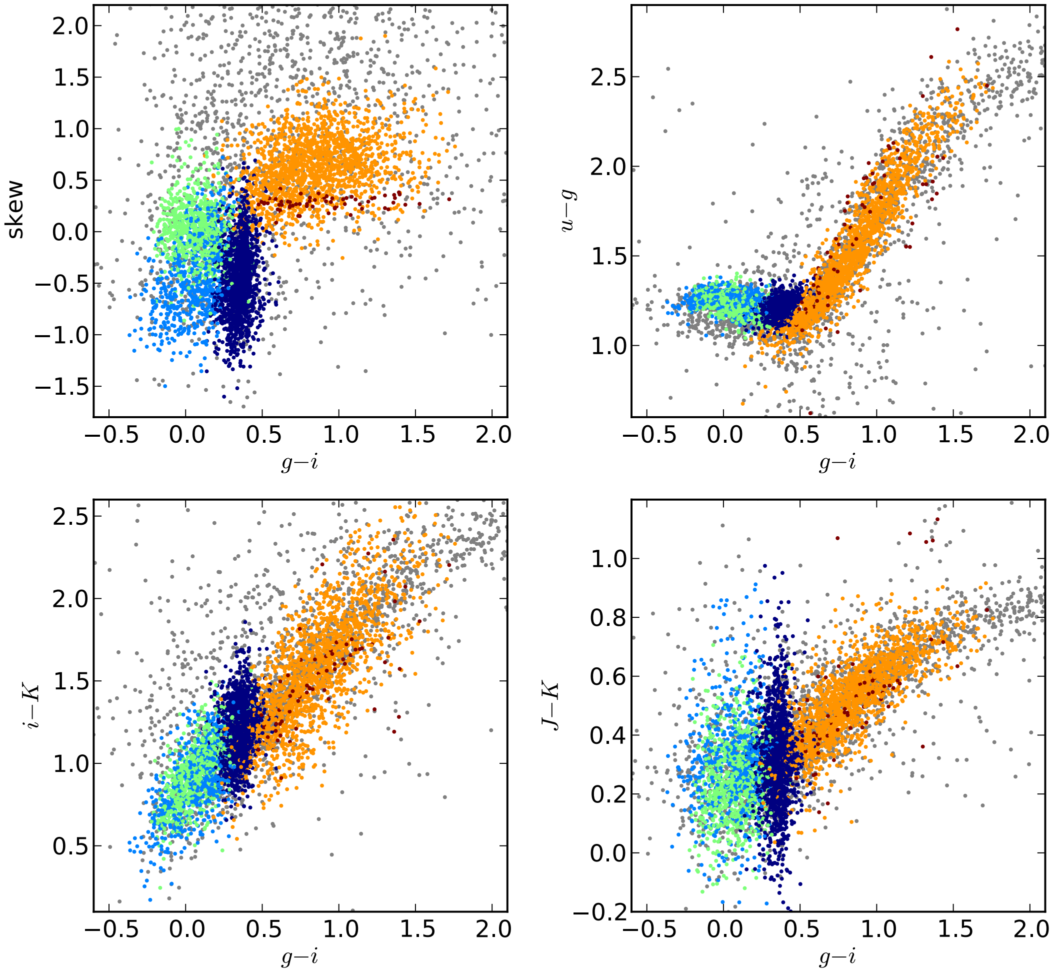

Figure 10.21.: Clustering analysis of periodic variable stars from the LINEAR data set. Clusters are derived using seven attributes (colors $u - g$, $g - i$, $i - K$, and $J - K$, $\log(P)$, light curve amplitude, and light curve skewness), and a mixture of 20 Gaussians. The $\log(P)$ vs. $g - i$ diagram and $\log(P)$ vs. light curve amplitude diagram for the same clusters are shown in the lower panels of Figure 10.20.

| # | $u-g$ | $g-i$ | $i-K$ | $J-K$ | $\log(P)$ | amplitude | skewness |
|---|---|---|---|---|---|---|---|
| 1 | $1.23 \pm 0.05$ | $0.06 \pm 0.12$ | $0.94 \pm 0.19$ | $0.24 \pm 0.14$ | $-0.48 \pm 0.06$ | $0.43 \pm 0.07$ | $0.01 \pm 0.28$ |
| 2 | $1.20 \pm 0.05$ | $0.35 \pm 0.07$ | $1.20 \pm 0.17$ | $0.31 \pm 0.15$ | $-0.24 \pm 0.06$ | $0.70 \pm 0.18$ | $-0.47 \pm 0.35$ |
| 3 | $1.24 \pm 0.05$ | $0.06 \pm 0.14$ | $0.89 \pm 0.21$ | $0.32 \pm 0.17$ | $-0.24 \pm 0.05$ | $0.70 \pm 0.18$ | $-0.42 \pm 0.39$ |
| 4 | $1.53 \pm 0.33$ | $0.83 \pm 0.27$ | $1.54 \pm 0.33$ | $0.52 \pm 0.14$ | $-0.83 \pm 0.07$ | $0.50 \pm 0.13$ | $0.62 \pm 0.29$ |
| 5 | $1.65 \pm 0.39$ | $0.90 \pm 0.29$ | $1.54 \pm 0.32$ | $0.59 \pm 0.28$ | $-0.91 \pm 0.00$ | $1.80 \pm 0.68$ | $0.28 \pm 0.06$ |

As was evident from the visual inspection of Figures 10.20 and 10.21, the most discriminative attribute is period. Clusters #2 and #3, which have very similar period distributions, are separated by the $g-i$ and $i-K$ colors, which are a measure of the star's effective temperature [11]. The problematic cluster #5 has a very narrow distribution of periods, and also unusually large amplitudes. Once it has been identified by clustering analysis, light curves and observing metadata can be searched for the causes of unreliable periods.

### 10.3.5. Analysis of arrival time data

Discussion of periodic signals in the preceding sections assumed that the data included a set of $N$ data points $(t_1, y_1), \ldots, (t_N, y_N)$, $j = 1 \cdots N$, with known errors for $y$. An example of such a data set is the optical light curve of an astronomical source where many photons are detected and the measurement error is typically dominated by photon counting statistics and background noise. Very different data sets are collected at X-ray and shorter wavelengths where individual photons are detected and background contamination is often negligible. In such cases, the data set consists of the arrival times of individual photons, $t_1, \ldots, t_N$, $j = 1 \cdots N$, where it can be typically assumed that errors are negligible. Given such a data set, how do we search for a periodic signal, and more generally, how do we test for any type of variability?

The best known classical test for variability in arrival time data is the Rayleigh test, and it bears some similarity to the analysis of periodograms (its applicability goes far beyond this context). Given a trial period, the phase $\phi_j$ corresponding to each datum is evaluated using eq. 10.21 and the following statistic is formed

$$R^2 = \left(\sum_{j=1}^{N} \cos(2\pi\phi_j)\right)^2 + \left(\sum_{j=1}^{N} \sin(2\pi\phi_j)\right)^2. \qquad (10.79)$$

This expression can be understood in terms of a random walk, where each angle $\phi_j$ defines a unit vector, and $R$ is the length of the resulting vector. For random data $R^2$ is small, and for periodic data $R^2$ is large when the correct period is chosen. Similarly to the analysis of the Lomb-Scargle periodogram, $R^2$ is evaluated for a grid of $P$, and the best period is chosen as the value that maximizes $R^2$. For $N > 10$, $2R^2/N$ is distributed as $\chi^2$ with two degrees of freedom (this easily follows from the random walk interpretation), and this fact can be used to assess the significance of the best-fit period (i.e., the probability that a value that large would happen by chance when the signal is stationary). A more detailed discussion of classical tests can be found in [14].

An alternative solution to this problem was derived by Gregory and Loredo [27] and here we will retrace their analysis. First, we divide the time interval $T = t_N - t_1$ into many arbitrarily small