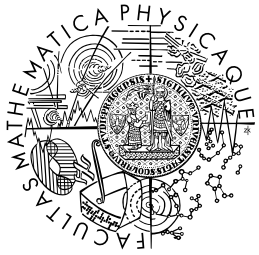

FACULTY
OF MATHEMATICS
AND PHYSICS
Charles University

DOCTORAL THESIS

Milan Pešta

Illuminating binary star evolution with observed populations and theoretical modeling

Institute of Theoretical Physics

Supervisor of the doctoral thesis: doc. Mgr. Ondřej Pejcha, Ph.D.
Study programme: Theoretical Physics, Astronomy and Astrophysics
Study branch: P4F1

Prague 2024

I declare that I carried out this doctoral thesis on my own, and only with the cited sources, literature and other professional sources. I understand that my work relates to the rights and obligations under the Act No. 121/2000 Sb., the Copyright Act, as amended, in particular the fact that the Charles University has the right to conclude a license agreement on the use of this work as a school work pursuant to Section 60 subsection 1 of the Copyright Act.

In ..... date ....

Author's signature

I dedicate this thesis to my wife Lýdia, without whom I would have never finished it, and to my daughter Samara, without whom I would have finished it much sooner. I thank my parents for their support and encouragement, and I thank my supervisor Ondřej Pejcha for his guidance and support during my studies.

Title: Illuminating binary star evolution with observed populations and theoretical modeling

Author: Milan Pešta

Institute: Institute of Theoretical Physics

Supervisor: doc. Mgr. Ondřej Pejcha, Ph.D., Institute of Theoretical Physics

Abstract: We present a method for the inference of the minimum mass ratio of contact binary stars from the observed distribution of their light curve amplitudes. We apply the method to a sample of contact binary candidates from the *Kepler* Eclipsing Binary Catalog. We find minimum mass ratios $q_{\rm min} = 0.087^{+0.024}_{-0.015}$ and $q_{\rm min} = 0.246^{+0.029}_{-0.046}$ for late-type contact binaries with periods $P > 0.3\,{\rm d}$ and $P \leq 0.3\,{\rm d}$, respectively, and $q_{\rm min} = 0.030^{+0.018}_{-0.022}$ for early-type contact binaries with $P < 1\,{\rm d}$. We also address the problem of identifying dark companion binaries—ellipsoidal variables hosting dormant black holes and neutron stars—in large photometric surveys using random forest classifiers trained on low-dimensional representations of synthetic light curves of dark companion, semidetached, and contact binaries. We find that the three classes are largely separable even under adverse conditions, such as the presence of spots and strong instrumental noise. Our method can significantly increase the purity of samples of dark companion candidates, improving the cost-efficiency of follow-up observations.

Keywords: binary stars, machine learning, photometry, contact binaries, black holes

Název práce: Důležité fáze vývoje dvojhvězd studované pomocí pozorovaných populací a teoretického modelování

Autor: Milan Pešta

Ústav: Ústav teoretické fyziky

Vedoucí disertační práce: doc. Mgr. Ondřej Pejcha, Ph.D., Ústav teoretické fyziky

Abstrakt: V této práci představujeme metodu pro odhad minimálního poměru hmotnosti složek kontaktních dvojhvězd na základě pozorované distribuce amplitud jejich světelných křivek. Metodu aplikujeme na vzorek kandidátů kontaktních dvojhvězd z katalogu *Kepler* Eclipsing Binary Catalog. Určili jsme minimální poměr hmotností $q_{\rm min} = 0.087^{+0.024}_{-0.015}$ a $q_{\rm min} = 0.246^{+0.029}_{-0.046}$ pro kontaktní dvojhvězdy pozdního spektrálního typu s periodami $P > 0.3\,{\rm d}$ a $P \leq 0.3\,{\rm d}$ a $q_{\rm min} = 0.030^{+0.018}_{-0.022}$ pro kontaktní dvojhvězdy raného spektrálního typu s $P < 1\,{\rm d}$. Dále se zabýváme problémem identifikace temných společníků—elipsoidálních proměnných, jejichž jednou složkou jsou neaktivní černé díry a neutronové hvězdy—v rozsáhlých fotometrických průzkumech pomocí random forest klasifikátorů vyškolených na nízkodimenzionálních reprezentacích syntetických světelných křivek temných společníků, polodotykových a kontaktních dvojhvězd. Zjistili jsme, že tyto tři třídy jsou v převážné míře oddělitelné i za nepříznivých podmínek, jako jsou přítomnost skvrn a silný instrumentální šum. Naše metoda může významně zvýšit čistotu vzorků kandidátů temných společníků, čímž zlepší nákladovou efektivitu následných pozorování.

Klíčová slova: dvojhvězdy, strojové učení, fotometrie, kontaktní dvojhvězdy, černé díry

Contents

1 Binary stars

Stars, especially massive ones, often occur in binaries or higher-order multiples. A binary star is a system of two stars that are gravitationally bound and orbit a common center of mass. The stars tidally interact with each other, and if their separation is small enough, they start exchanging mass, which dramatically affects their evolution and can lead to the merger of the system. In Sect. 1.1, we provide an overview of the Roche model, which is used to describe the equilibrium configuration of binary stars. We discuss the classification of binary stars based on the method of observation in Sect. 1.2, and we explore contact binaries and dark companion binaries in more detail in Sects. 1.3 and 1.4, respectively.

1.1 Roche model

In a reference frame that corotates with a binary system, three forces act on a test particle that is at rest with respect to the frame: the gravitational forces of the two stars and the centrifugal force due to the rotation of the system. The Roche model assumes that the two stars are in a circular orbit and their masses are concentrated in point masses located at their centers. Under these assumptions, the net force on the particle can be expressed as the gradient of the *Roche potential* Φ,

$$\Phi = -\frac{GM_1}{r_1} - \frac{GM_2}{r_2} - \frac{1}{2}\omega^2 r_3^2, \tag{1.1}$$

where M_1 and M_2 are the masses of the two stars, r_1 and r_2 are the distances of the particle from the centers of the two stars, r_3 is the distance of the particle from the center of mass of the system, and ω is the angular velocity of the system. If we express ω from Kepler's third law and divide Eq. (1.1) by GM_1/a, where a is the semimajor axis of the orbit, we obtain the normalized Roche potential ϕ,

$$\Phi/(GM_1/a) \equiv \phi = -\frac{1}{\tilde{r}_1} - \frac{q}{\tilde{r}_2} - \frac{1}{2}(1+q)\tilde{r}_3^2, \tag{1.2}$$

where $\tilde{r}_i = r_i/a$, $i = 1, 2, 3$, and $q = M_2/M_1$ is the mass ratio of the two stars. The normalized Roche equipotentials do not depend on the parameters a and M_1, which are absorbed in the normalization factor, and their shape is determined only by the mass ratio q. In hydrostatic equlibrium, the surface of each star coincides with a closed equipotential. We show the contours of the normalized Roche potential for $z = 0$ and $q = 0.5$ in Fig. 1.1.

The net force on a test particle at rest in the rotating frame vanishes when $\nabla\phi = 0$. In general, there are five points, known as the Lagrange points, where this condition is satisfied. We show the Lagrange points labeled as $L1$–$L5$ in Fig. 1.1. The Lagrange point $L1$ lies between the two stars and the critical equipotential passing through it defines the *Roche lobes* of the stars. When the sizes of the stars are small relative to their separation, they reside deep within the Roche lobes, where the equipotentials are nearly spherical, and the stars are said to be in a detached configuration (Fig. 1.2a). As we decrease the separation between the stars or equivalently increase their sizes, the surfaces of the stars move closer to the Roche lobes, and the equipotentials start to significantly deviate from

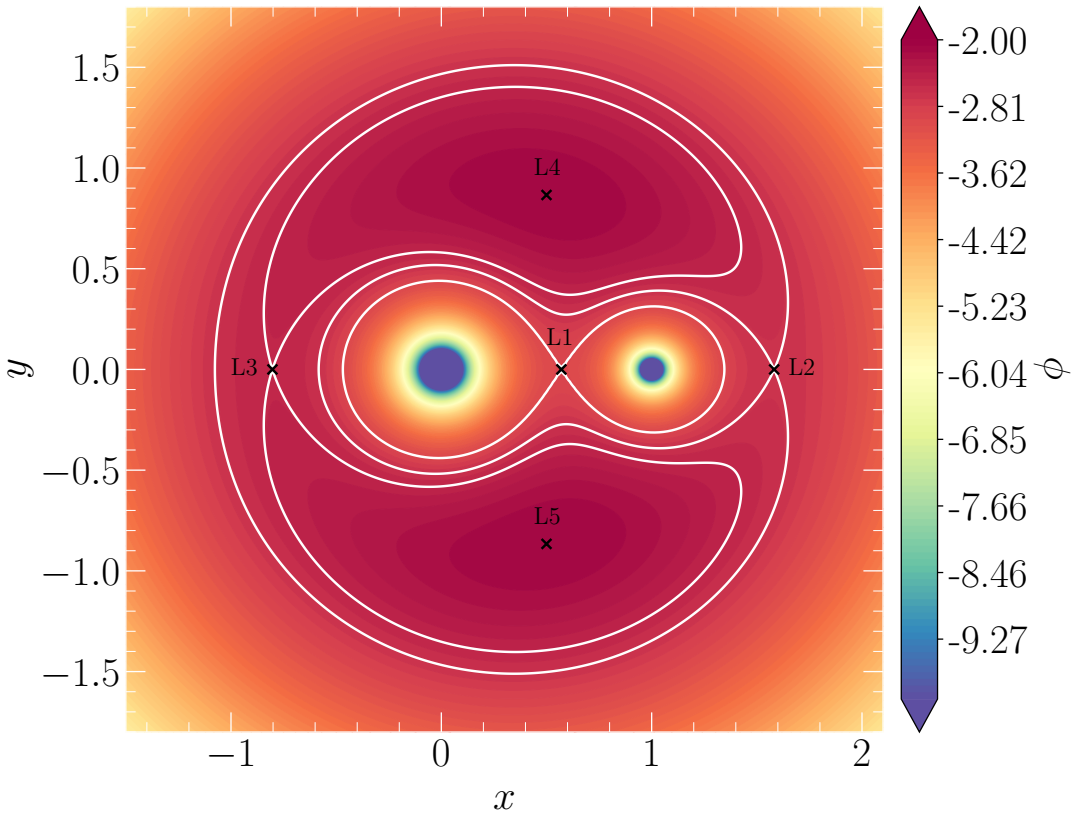

Figure 1.1 Normalized Roche potential in the orbital plane of a binary star with $q = 0.5$. The solid white lines represent the equipotentials passing through the Lagrange points $L1$–$L3$.

spherical. When one of the stars fills its Roche lobe, it starts transferring mass to the companion, and the system becomes a semidetached binary (Fig. 1.2b). If both stars overflow their Roche lobes, they form a contact binary (Fig. 1.2c). The contact configuration is stable up to the outer critical equipotential, which represents the maximum volume the system can occupy before it starts losing mass and angular momentum through the $L2$ point. The $L3$ point is located beyond the outer critical equipotential and is therefore of lesser physical significance. The $L4$ and $L5$ form equilateral triangles with the centers of the two stars and despite being maximum points of the Roche potential, they are stable due to the effect of the Coriolis force, provided $q \lesssim 1/25$ (Murray & Dermott 1999).

1.2 Binary classification

The classification of binary stars into detached, semidetached, and contact binaries introduced by Kopal (1955, 1959) is based on the physical configuration of the system with respect to the critical inner and outer Roche equipotentials. However, the physical character of the system is not always apparent, which is why classification based on the method of observation is often more practical. We distinguish between the following types of binary stars according to the way they are observed (Carroll & Ostlie 2017):

- **Visual binaries.** Visual binaries are systems which can be resolved into

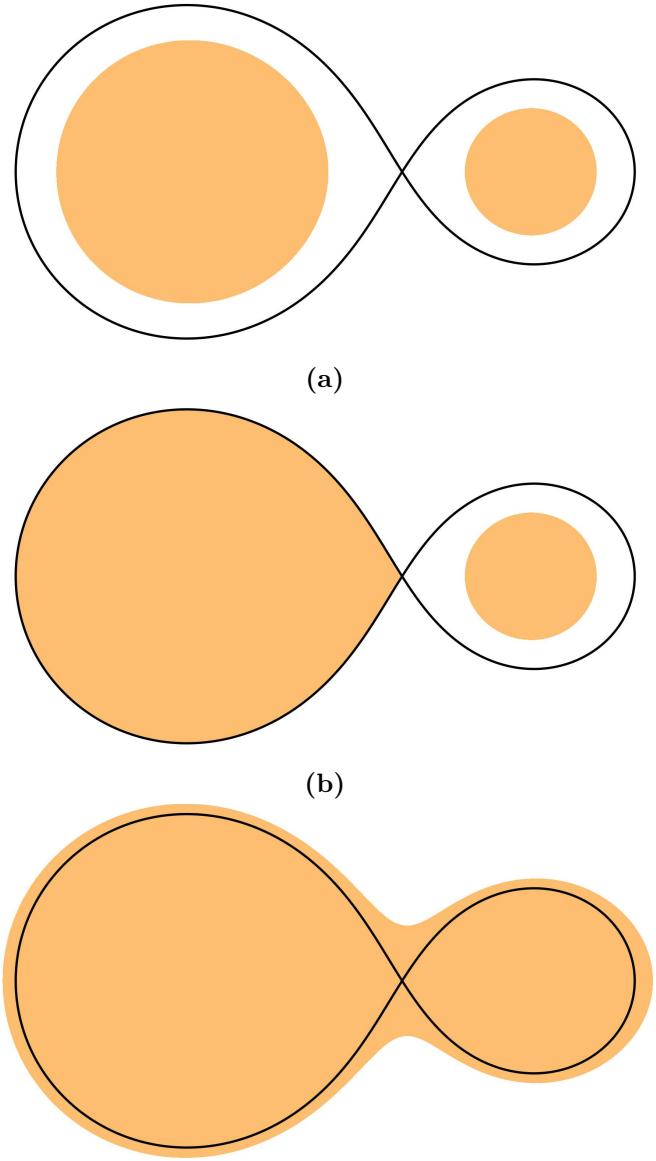

Figure 1.2 Face-on schematics of detached (a), semi-detached (b) and contact (c) binary configurations. The solid black lines represent the inner critical Roche equipotentials of the systems.

two stars with a telescope. The typical separation of the stars is of the order of arcseconds, and their periods range from a few years to centuries. Visual binaries should not be confused with optical doubles, which are not actually binaries but coincidentally appear close together in the sky with no physical connection.

- **Astrometric binaries.** If one of the stars in the system is much brighter than the other, or the companion is a compact and electromagnetically silent object, such as a black hole or a neutron star, it may not be possible to observe both components directly. In such cases, the presence of the companion can be inferred from periodic wobbles of the visible star around the center of mass of the system.

- **Spectroscopic binaries.** If the stars in a binary system are too close to be visually resolved into individual components, we can infer their binary nature from periodic Doppler shifts in their spectra. The separation between the stars in spectroscopic binaries is usually very small and their orbital periods are of the order of days. We distinguish between single-lined spectroscopic binaries, where the spectrum of only one component is observed, and double-lined spectroscopic binaries, where the spectra of both components are visible.

- **Photometric binaries.** Photometric binaries are systems which exhibit periodic variations in their light curves. The variations can be due to the mutual eclipses of the stars, tidal distortion of the stars (ellipsoidal variables), reflection of light off the surface of the stars, or Doppler beaming of light due to their orbital motion of the stars. Based on the shape of the light curve, we further distinguish between EA, EB, and EW eclipsing binaries, roughly corresponding to the detached, semidetached, and contact binary configurations, respectively.

In this thesis, we focus on photometric binaries, namely contact binaries observed as EW eclipsing binaries and ellipsoidal variables hosting electromagnetically silent black holes and neutron stars, which we refer to as *dark companion binaries*. These systems are of particular interest to us because their unique nature is imprinted in their light curves, making it possible to identify them and infer the parameters of their populations from photometry alone. We discuss these systems in more detail in the following section.

1.3 Contact binaries

The prototype of the EW variable class, W Ursae Majoris (W UMa), was first observed by Muller & Kempf (1903), who hypothesized that its variability and short period of approximately four hours are due to a rotating body with an unequal brightness distribution or deformed shape that significantly deviates from a sphere. The authors also considered the possibility that the observed light curve might be produced by two stars of equal size and luminosity in a close orbit around each other, but they concluded that the long-term stability of such configuration is uncertain. It was later revealed that the variations are caused

by equal-depth primary and secondary eclipses and the period of the system is actually twice as long, confirming the binary nature of the system (Russell et al. 1917; Adams & Joy 1919). However, it was not until the 1940s and 1950s that the existence of contact binaries was seriously considered and the connection with W UMa systems was made (Kuiper 1941; Kopal 1955).

The seminal paper by Kuiper (1941), which introduced the term contact binary, revealed that contact configurations with non-identical components are intrinsically unstable, leading to large-scale circulations carrying mass between the components as long as the masses are unequal. This is in agreement with the observed light curves of W UMa systems, which exhibit equal-depth primary and secondary minima, pointing to nearly identical mean surface brightnesses of the components. Yet, W UMa systems with unequal masses are known to exist, giving rise to the so-called *Kuiper paradox*. The paradox was resolved by Lucy (1968b), Lucy (1976), Flannery (1976), and Webbink (1976), who developed a model which assumes that the stars are separately out of thermal equilibrium but the system as a whole maintains a global thermal equilibrium. As a result, the binary undergoes a series of *thermal relaxation oscillations* (TROs)—cycles of mass transfer on the thermal timescale, during which the components move mass back and forth between each other. During the TROs, the system alternates between contact and semidetached states, and the equal effective temperatures of the components are achieved through turbulent convection driven by pressure gradients.

Despite the success of the TRO model in explaining the geometry and light curves of W UMa variables, there are some discrepancies between the model predictions and observations, such as the lack of semidetached systems with periods shorter than 0.45 days (Stępień 2011). To address these discrepancies, Stępień (2004), Stępień (2006), and Stępień (2009) proposed a model in which the components reach thermal equilibrium following a rapid mass transfer and mass ratio reversal. The difference from the TRO model is that the secondary is an evolved star with a hydrogen-depleted core, which is why it seems oversized for its mass when interpreted as a main-sequence star. The equal effective temperature of the components is achieved through large-scale circulation carrying high entropy matter from the primary to the secondary. Regardless of the exact mechanism, we observe secular evolution of the system towards smaller mass ratios until it starts losing mass through the $L2$ point or becomes unstable to the tidal Darwin instability (Darwin 1879) and merges. Due to the correlation between the parameters of contact binary systems, the Darwin instability leads to a minimum mass ratio $q_{\min}$, below which basically no contact binaries should be observed.

In this thesis, we present a Bayesian reformulation and extension of the method by Rucinski (2001) for the inference of $q_{\min}$ from the photometric amplitude distribution of W UMa systems. The method utilizes synthetic light curves for likelihood-free inference of the parameters of the mass ratio distribution of contact binary stars. We provide a high-level overview of the employed data analysis techniques—synthetic light curve generation and Bayesian inference—in Chap. 2, and we present a detailed description of our method for the inference of $q_{\min}$ in Chap. 4.

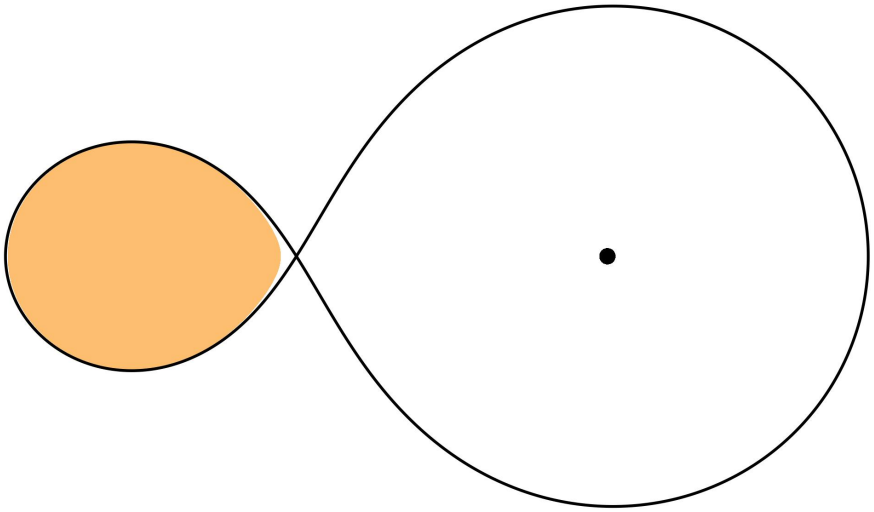

Figure 1.3 Face-on schematic of a dark companion binary with $q = 5$. The solid black line represents the inner critical Roche equipotential of the system.

1.4 Dark companion binaries

Dark companion binaries represent an interesting class of binary systems. In Fig. 1.3, we show a schematic of a dark companion binary hosting a black hole five times more massive than the visible companion. The star is close to filling its Roche lobe, resulting in large ellipsoidal variations in its light curve. If it actually overflowed the critical equipotential and started transferring mass to the black hole, the system could be observed as an X-ray binary and would no longer be considered a dark companion binary. However, black hole X-ray binaries seem to be a rare outcome of binary evolution, and only about 1000 such systems are estimated to exist in our Galaxy (Portegies Zwart et al. 1997; Corral-Santana et al. 2016). Conversely, based on binary synthesis models, a significant fraction of black hole binaries might actually be dark companion binaries (Breivik et al. 2017; Chawla et al. 2022), highlighting the importance of characterizing the dark companion binary population.

The lack of strong electromagnetic signatures of black holes and neutron stars in dark companion binaries makes it difficult to detect them. As a result, only a few such systems have been discovered so far. All these systems were identified by searching available spectroscopic and astrometric catalogs and selecting objects exhibiting peculiar patterns consistent with the presence of a dark companion, such as unusually high binary mass functions or large photocenter wobbles. There is no issue with this approach when applied to archival data, but if we want to systematically search for new dark companion binaries, it would be highly cost-inefficient to obtain high-resolution spectra and/or astrometric measurements for large numbers of randomly selected objects. It would be far more practical to preselect dark companion binary candidates based on photometry and then follow up with spectroscopic and astrometric observations only on the most promising candidates, optimizing the allocation of resources and maximizing the chances of discovery.

To facilitate the search for dark companion binaries in large photometric

surveys, Gomel et al. (2021b,a) developed a proxy for the minimum mass ratio of dark companion systems, the *modified minimum mass ratio* (mMMR), derived solely from the photometric amplitude of the ellipsoidal variations. The method is based on the idea that high values of mMMR could be indicative of the presence of a dark companion in the system. In practice, the method yields high false positive rates (Nagarajan et al. 2023), pointing to the need for a more sophisticated approach to the problem.

In this thesis, we present a novel method for the identification of dark companion binaries based on both the photometric amplitude and the shape of the light curves. The method utilizes principal component analysis (PCA) to construct low-dimensional representations of synthetic light curves of dark companion binaries and their common contaminants, such as contact binaries and semidetached binaries. We then train a random forest classifier on the PCA representations of the light curves to distinguish between the different types of binaries and identify dark companion binary candidates. We outline the techniques of light curve synthetis and random forest classification in Chap. 2, and we describe our method for the identification of dark companion binaries in large photometric surveys in Chap. 5.

2 Selected methods of data analysis

In this chapter, we provide an overview of selected methods of data analysis that we utilized in the main part of the thesis, including light curve synthesis (Sect. 2.1), Bayesian inference (Sect. 2.2), and random forest classification (Sect. 2.3). We present brief descriptions of the methods and demonstrate their applications through simple examples.

2.1 Light curve synthesis

Both contact binaries and dark companion binaries are strongly affected by tidal interactions between the components, which manifests as variations in their light curves. Although thousands of contact binary light curves have been observed, only a fraction of these systems have well-determined physical properties, yielding a heterogeneous sample of contact binaries with known parameters. The situation is even worse for dark companion binaries, where only a few systems have been observed so far (see Sect. 5.1 for an incomplete list of such detections). The lack of well-curated samples of contact and dark companion binaries hinders the application of machine learning techniques to study of these systems. However, recent advances in numerical simulations of binary systems have made it possible to synthesize light curves of contact and dark companion binaries with high accuracy, allowing us to generate large homogeneous samples of well-characterized systems.

In this thesis, we make extensive use of `PHOEBE 2`* (PHysics Of Eclipsing BinariEs, Prša et al. 2016; Conroy et al. 2020b). `PHOEBE` is an open-source modeling code in `Python` for computing theoretical light curves, radial velocity curves as well as spectral line profiles of eclipsing binary systems. The code allows the user to specify a wide range of physical and orbital parameters of the system, including but not limited to the atmosphere tables for the components, passbands, and limb-darkening and gravity-brightening coefficients. `PHOEBE` supports parallelization via MPI and offers swappable backends for computing the light curves, including `JKTEBOP`† (Southworth et al. 2004), `ellc`‡ (Maxted 2016), and native `PHOEBE` backend. The code allows for the inclusion of various advanced effects in the model, such as irradiation, spin-orbit misalignment, or reddening and extinction. As of version 2.3, `PHOEBE` includes a general framework for inverse problem solving, allowing the user to infer the physical parameters of the system from observational data. To test the robustness and accuracy of `PHOEBE` in solving the inverse problem, we tried to reproduce the results from Maxted et al. (2020), who precisely estimated the masses and radii of the stars in the binary system AI Phe using various methods, including `PHOEBE`. We present the results of our analysis in Chap. 3.

*`https://phoebe-project.org`
†`https://www.astro.keele.ac.uk/jkt/codes/jktebop.html`
‡`https://github.com/pmaxted/ellc`

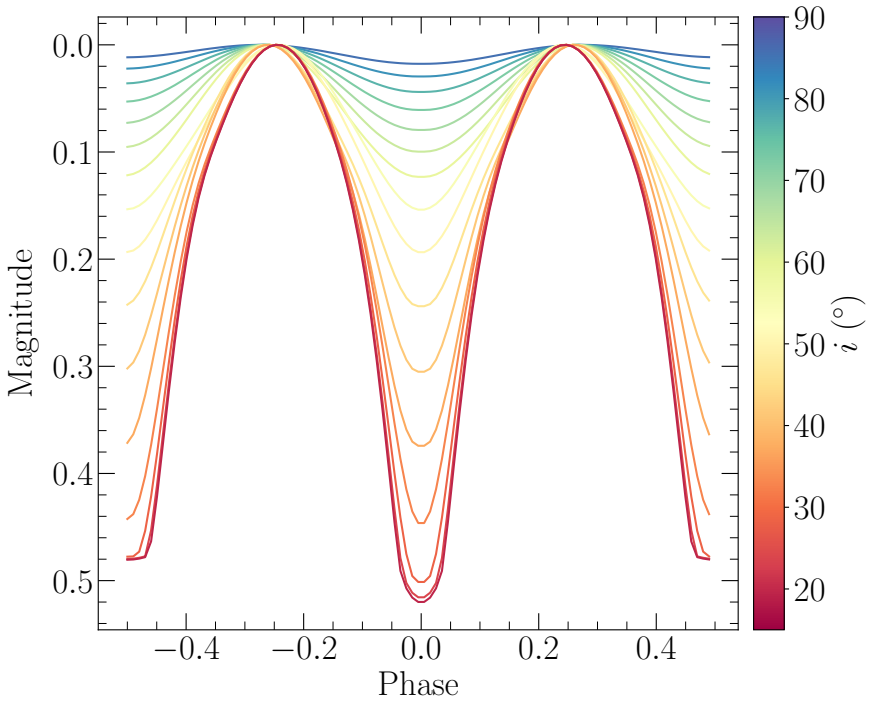

Figure 2.1 Synthetic light curves of a contact binary system with $q = 0.3$ observed at different inclinations.

To illustrate the capabilities of `PHOEBE` in generating binary light curves, we present Figs. 2.1–2.2, which show synthetic light curves of contact binary systems with $q = 0.3$ and dark companion binary systems with $q = 3$ observed at different inclinations. The light curves vary smoothly with inclination and demonstrate the similarities and differences between the two types of binary systems, especially close to the maxima. In Chap. 4, we generate a large number of contact binary light curves similar to those in Fig. 2.1 but covering a much wider range of parameters. This allows us to model the photometric amplitude distribution of contact binaries as a function of the parameters of their mass-ratio distribution, providing a novel approach for constraining the minimum mass ratio of contact binary stars. We follow a similar approach in Chap. 5, where we generate synthetic light curves of dark companion binaries, contact binaries, and semidetached binaries to investigate whether the information contained in the light curves is sufficient to reliably distinguish between these types of systems.

2.2 Bayesian inference

Traditional statistical methods of parameter estimation, such as least squares fitting or χ^2 minimization, operate within the frequentist framework, which treats the parameters as fixed but unknown quantities. In contrast, Bayesian inference treats the parameters as random variables with probability distributions, allowing for the use of Bayes' theorem to update the distributions as we collect more data. We write Bayes' theorem as

$$p(\theta|D) = \frac{\mathcal{L}(\theta|D)p(\theta)}{p(D)}, \tag{2.1}$$

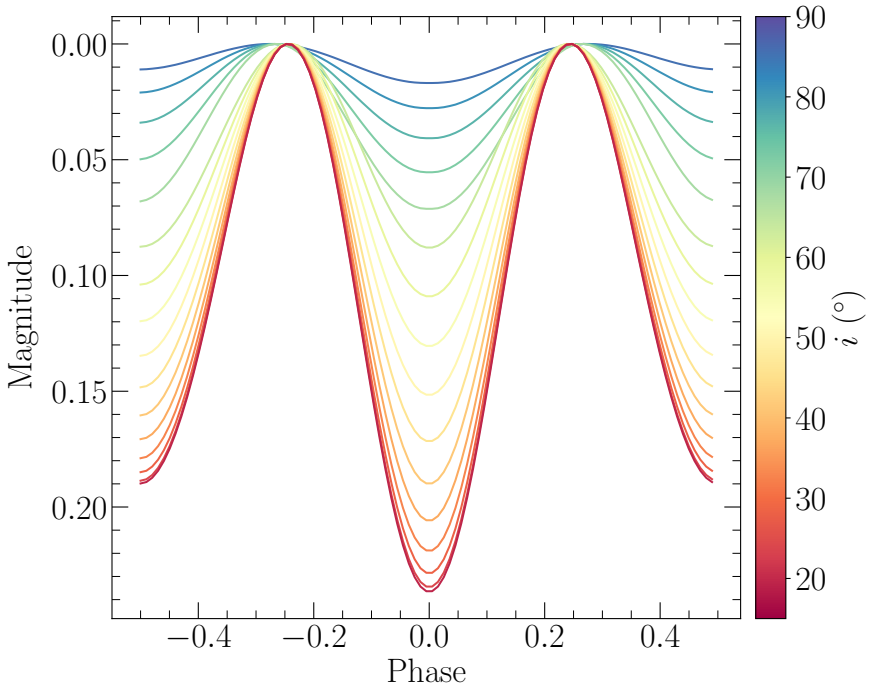

Figure 2.2 Synthetic light curves of a dark companion binary system with $q = 3$ observed at different inclinations.

where $p(\theta|D)$ is the posterior distribution of the parameters θ given the data D, $\mathcal{L}(\theta|D) \equiv p(D|\theta)$ is the likelihood of θ given D, $p(\theta)$ is the prior distribution of θ, and $p(D)$ is the marginal likelihood also known as the prior predictive or evidence. Assuming that the observations are independent and identically distributed (i.i.d.), the likelihood function can be factorized as

$$\mathcal{L}(\theta|D) = \prod_{i=1}^{N} l(d_i|\theta), \tag{2.2}$$

where N is the number of data points, d_i is the ith data point, and $l(d_i|\theta) \equiv p(d_i|\theta)$ is the single-observation likelihood of θ given d_i.

The prior distribution $p(\theta)$ encodes our knowledge about the parameters before observing the data. If we have no prior knowledge of the problem, we can use a non-informative uniform prior, assigning equal probability to all values of θ within a certain interval. The posterior distribution reflects the updated knowledge about the parameters after observing the data. In high-dimensional problems, the likelihood is often computationally expensive to evaluate, making direct calculation of the posterior distribution intractable. In such cases, Markov Chain Monte Carlo (MCMC) methods provide a way to directly sample the posterior distribution without the need for evaluating the evidence. MCMC methods operate by constructing a sequence of samples where the inclusion of each sample in the sequence depends only on the ratio of the posterior distribution at the proposed and the current position in the parameter space, eliminating the evidence from the calculation. In practice, a number of chains are generated in parallel, each starting from a different initial position in the parameter space. For more details about MCMC methods, we refer the reader to Ivezić et al. (2020).

Following the approach of Hogg et al. (2010), we illustrate the application of

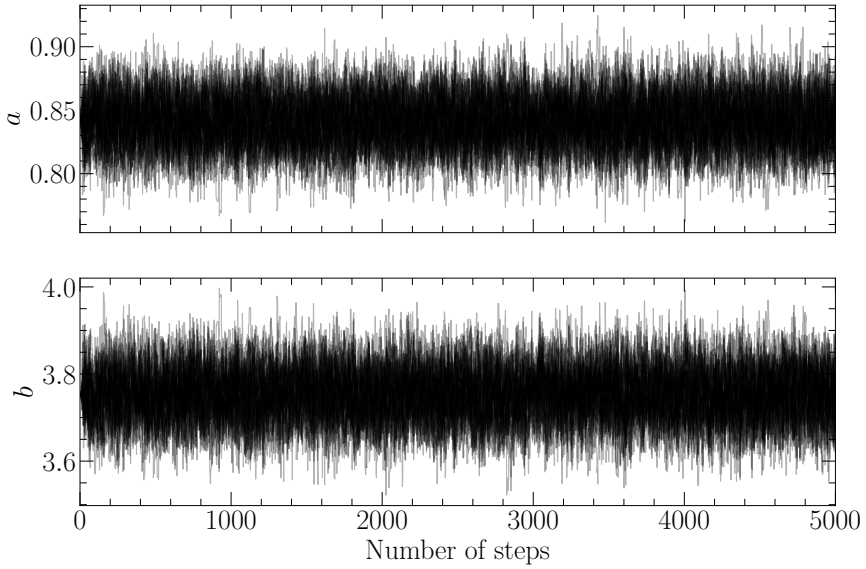

Figure 2.3 Chain plots of the parameters of the linear model.

Bayesian inference on a simple example of fitting a line to a set of data points. We considered a linear model $y = 0.84x + 3.75$, and we randomly sampled 50 values of x from the interval $[0, 5]$. We then injected y with uncorrelated Gaussian noise with standard deviation $\sigma = 0.2$ to simulate observational errors. The likelihood of the model is given by

$$\begin{aligned} \mathcal{L}(a, b|\{x_i, y_i\}_{i=1}^N) &= \prod_{i=1}^N l(a, b|x_i, y_i) \\ &= \prod_{i=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - ax_i - b)^2}{2\sigma^2}\right). \end{aligned} \tag{2.3}$$

We considered non-informative uniform priors for the parameters a and b, and we assumed that σ is known. We sampled the posterior distribution $p(a, b|\{x_i, y_i\}_{i=1}^N)$ using the `emcee`[§] implementation of MCMC (Foreman-Mackey et al. 2013). We generated 32 chains, which we initialized at the values of a and b that maximize the likelihood, and we ran the chains for 5000 steps each. We discarded the first 100 steps of each chain as burn-in to eliminate the impact of the initial conditions, and we thinned the chains by a factor of 30 to reduce the autocorrelation between the samples. We present the chain plots of the parameters a and b in Fig. 2.3, and we show the obtained posterior distributions of the parameters in Fig. 2.4.

We observe that the posterior distributions are concentrated at the true values of the parameters, demonstrating the ability of Bayesian inference to recover the true parameters of the model. In addition, by taking different percentiles of the sampled posterior distributions, we can construct credible intervals for the parameters, providing uncertainties of the estimates. Credible intervals, which represent the probability that the parameter lies within the interval, should not be confused with confidence intervals, which are frequentist constructs and describe the proportion of intervals that would contain the true parameter value if the

[§]`https://emcee.readthedocs.io`

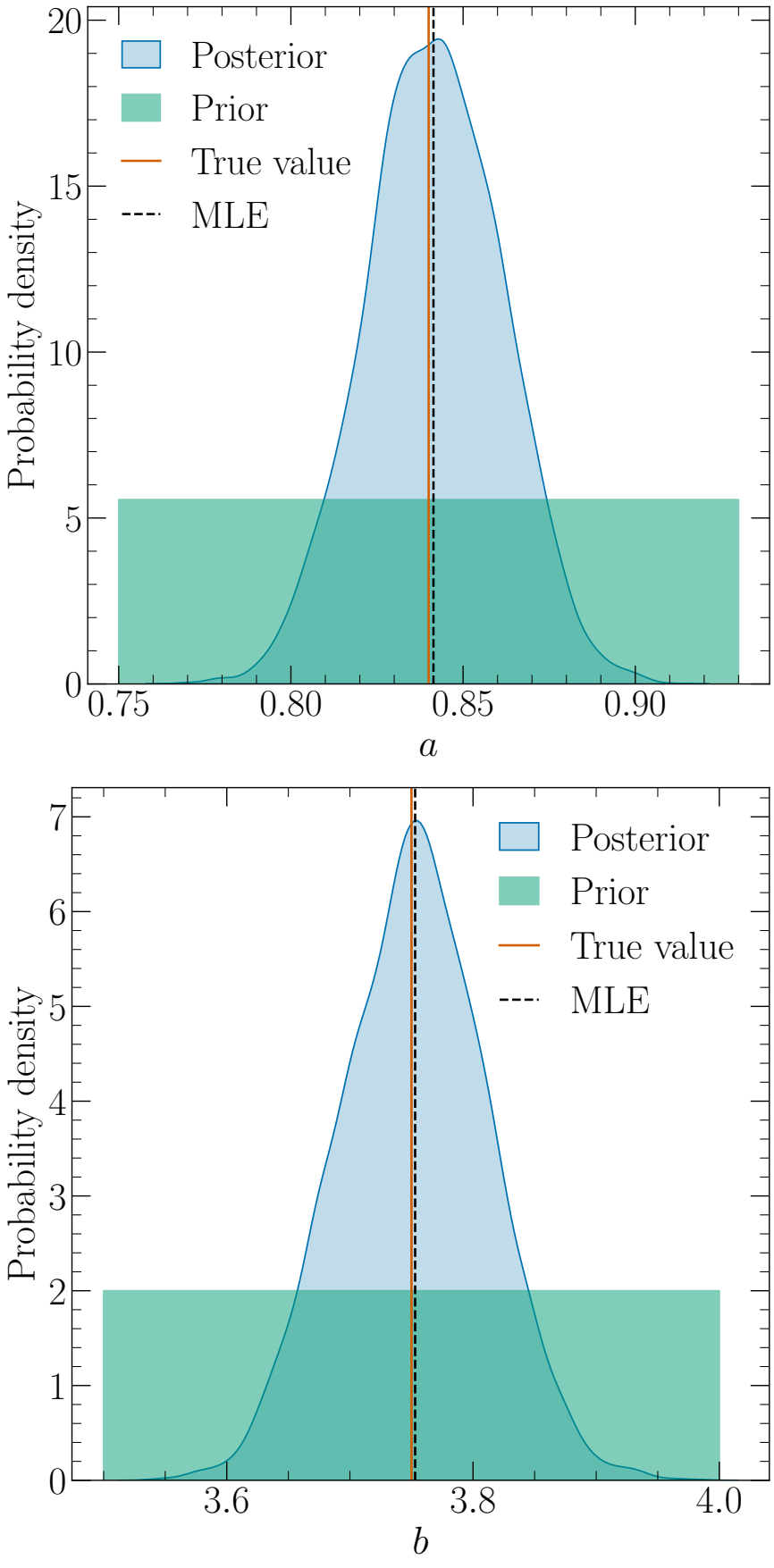

Figure 2.4 Prior and posterior distributions of the parameters a (top panel) and b (bottom panel).

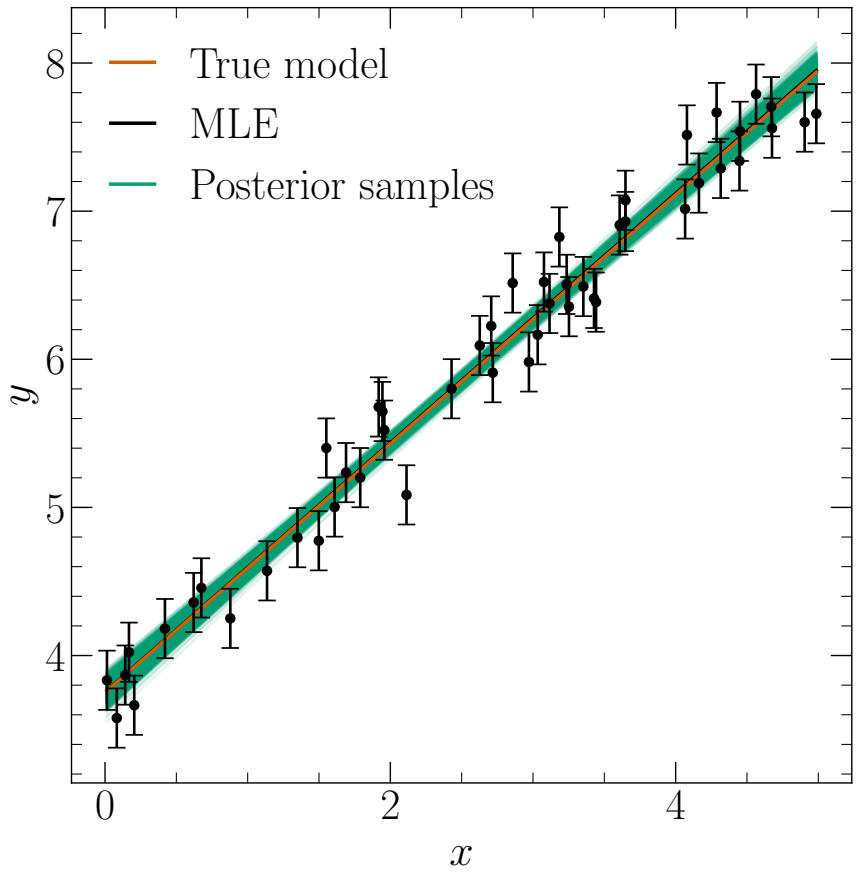

Figure 2.5 Comparison of the true model, maximum likelihood model, and models sampled from the posterior distribution of the parameters a and b.

experiment were repeated many times. We illustrate the utility of credible intervals in Fig. 2.5, where we overlay the true model and the maximum likelihood model on models sampled from the posterior distribution, allowing us to visualize the uncertainty in the estimated model parameters.

Although simplified, this example demonstrates all the essential steps of Bayesian inference. We follow the same approach in Chap. 4, where we use Bayesian mixture modeling to construct a sample of contact binary stars from the *Kepler* Eclipsing Binary Catalog (Kirk et al. 2016).

2.3 Random forest classifiers

Astronomical problems often involve identification of objects based on their observed properties, e.g., classification of variable stars based on their light curves or morphological classification of galaxies. There are many machine learning algorithms available for classification tasks, ranging in complexity from simple linear classifiers to complex deep neural networks, but random forest classifiers have proven to be particularly effective in many astronomical applications due to their robustness and scalability (e.g., Jayasinghe et al. 2018, 2019; Dubath et al. 2011; Förster et al. 2021).

Random forest classifiers are a subclass of random forests, introduced in their current form by Breiman (2001), which can be used for both classification and

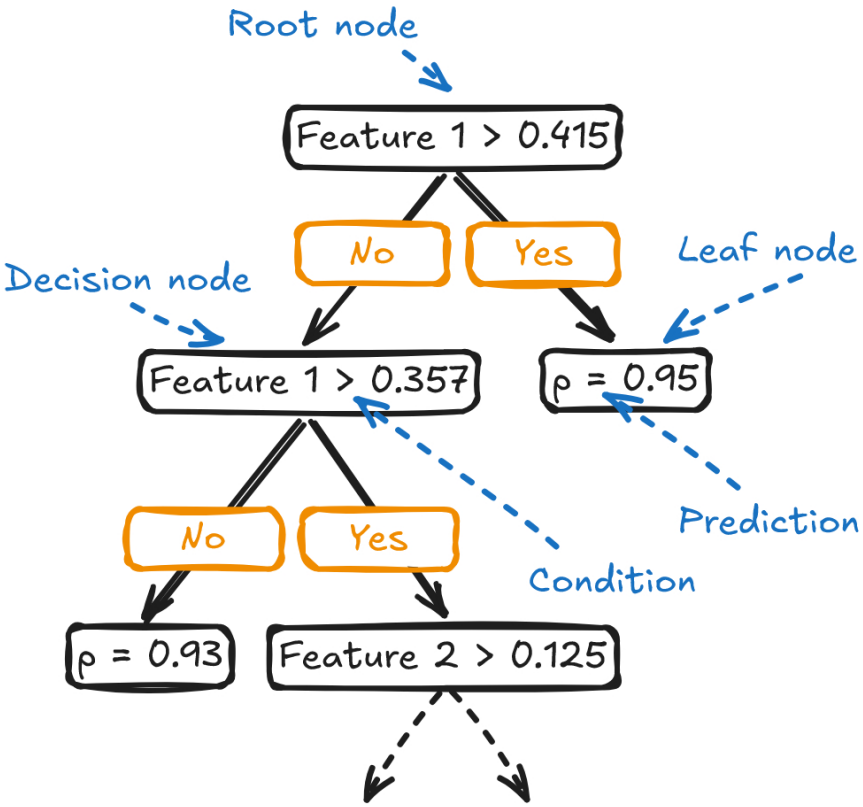

Figure 2.6 Schematic representation of a decision tree classifier.

regression. A random forest classifier consists of an ensemble of decision tree classifiers, where each tree is trained on a different random subset of the training data. A decision tree classifier is a collection of nested if-else statements that recursively split the data into subsets to achieve the best separation of the classes. We illustrate the architecture of a decision tree classifier in Fig. 2.6. The tree starts with a *root node* containing the entire dataset, which is then split into two child nodes by imposing a threshold condition on the values of the most informative feature. Each child node is either a *leaf node*, which is terminal node that contains the final prediction (class probabilities), or a *decision node*, which is further split and the process repeats. A node becomes a leaf node if it is pure (contains only one class) or if a stopping criterion is met, e.g., the number of samples in the node falls below a certain value. The tree is grown until all nodes are leaf nodes.

There are different metrics to determine what feature and threshold should be used to split the data at each node. For example, the *information gain* measures the decrease in the entropy of the node due to the split. We define the entropy $S(X)$ of a node X as

$$S(X) = -\sum_{i=1}^{n} p_i(X) \log p_i(X), \tag{2.4}$$

where n is the number of classes in the data and $p_i(X)$ is the probability that a randomly selected sample from X belongs to the class i. Denoting the entropies of the left and right child nodes of X obtained by splitting the data on the feature F with a threshold t as $S(X|F \leq t)$ and $S(X|F > t)$, respectively, we write the

information gain $IG(X, F, t)$ of the split as

$$IG(X, F, t) = S(X) - \frac{N_{\text{left}}}{N} S(X|F \leq t) - \frac{N_{\text{right}}}{N} S(X|F > t), \tag{2.5}$$

where N_{left} and N_{right} are the numbers of samples in the left and right child nodes, respectively. We obtain the best split at each node by maximizing the information gain over all possible features and thresholds. For continuous features, there are infinitely many thresholds to be considered. In practice, the search is limited to the midpoints between the sorted unique values of the features.

Gini impurity, which represents the probability that a randomly selected sample is misclassified if it is randomly labeled according to the class distribution of the node, is another metric that is often used to find the best split. We define the Gini impurity $G(X)$ of a node X as

$$G(X) = 1 - \sum_{i=1}^{n} p_i^2(X), \tag{2.6}$$

where the meanings of the symbols are the same as in Eq. (2.4). The Gini impurity reaches its minimum value of 0 when the node is pure. In analogy with the information gain, the optimal split is obtained by minimizing the weighted sum of the Gini impurities of the child nodes over all possible features and thresholds.

Decision tree classifiers have many favorable properties, such as interpretability, ability to handle both numerical and categorical data, and invariance to the scaling of the features. However, they are prone to overfitting, especially when the trees are too deep. We can mitigate the effect of overfitting by setting a maximum depth of the tree or requiring a minimum number of samples in the leaf nodes. Another option is to fully grow the tree and then prune it by removing certain branches that hurt the performance of the classifier on validation data. Despite these regularization techniques, overfitting remains a major weakness of decision tree classifiers.

Random forest classifiers address the issue of overfitting by averaging the predictions from multiple decision trees. The idea is based on the central limit theorem, which states that the squared error of the mean of N i.i.d. random variables scales as $1/N$ and tends to zero as N increases. For this to hold for the average of the predictions of an ensemble of decision trees, the trees should be independent and the predictions should be uncorrelated. Random forests achieve this by *bootstrap aggregating* (bagging) and *attribute sampling*. Bagging corresponds to training each decision tree on a different subset of the training data and aggregating the predictions of the trees by majority voting. In brief, we construct the training sets for the trees by randomly sampling the original training set with replacement until we have obtained a set of the same size, yielding 63.2% unique samples in each training set on average. Attribute sampling introduces additional randomness into the training process by considering only a random subset of the features at each split, further increasing the independence of the trees. The number of features to consider at each split is a hyperparameter of the random forest classifier, and it is typically set to $\sqrt{M}$, where M is the total number of features.

We illustrate the use of random forest classifiers on a model example. We generated a synthetic dataset of 3000 samples using `make_classification` from

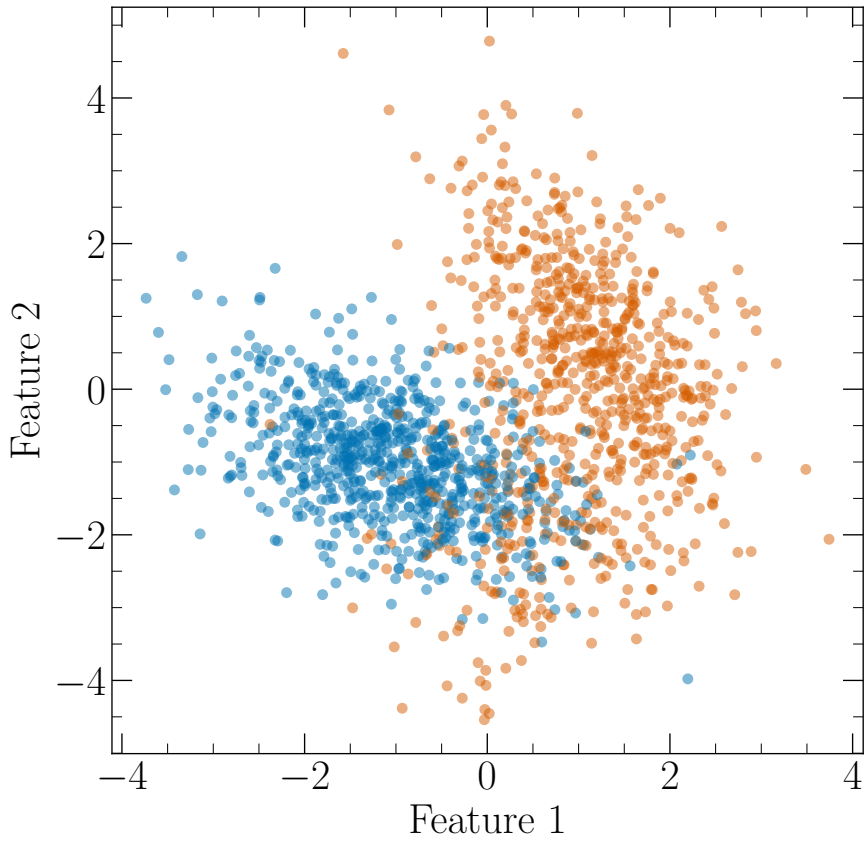

Figure 2.7 Scatter plot of the first two features of the synthetic dataset.

the `sklearn.datasets` module in `Python`. The dataset consists of two classes and ten features, three of which are informative. We randomly split the dataset into two halves corresponding to the training and test sets. We show a scatter plot of the first two features of the dataset in Fig. 2.7. We then trained the `scikit-learn` implementation of the random forest classifier on the training set using the default hyperparameters, with `n_estimators` = `100`, `max_depth` = `None`, and `min_samples_leaf` = `1`. The classifier achieved an accuracy of 1.0 on the training set and 0.954 on the test set, demonstrating the ability of random forest classifiers to generalize well to unseen data. In Fig. 2.8, we visualize the decision boundary of the classifier in the space of the first two features, setting the remaining features equal to their mean values. The decision boundary is non-linear and divides the feature space into regions predominantly occupied by either of the two classes. We also evaluated the feature importances of the classifier, which are proportional to the accumulated decrease in the Gini impurity due to the splits on the respective features. We present the obtained feature importances in Fig. 2.9, demonstrating the ability of the classifier to correctly identify the first three features as the most informative.

We leverage the power of random forest classifiers in Chap. 5, where we train them on low-dimensional representations of synthetic light curves of dark companion binaries, contact binaries, and semidetached binaries to investigate the possibility of distinguishing between these types of systems based on photometric data alone.

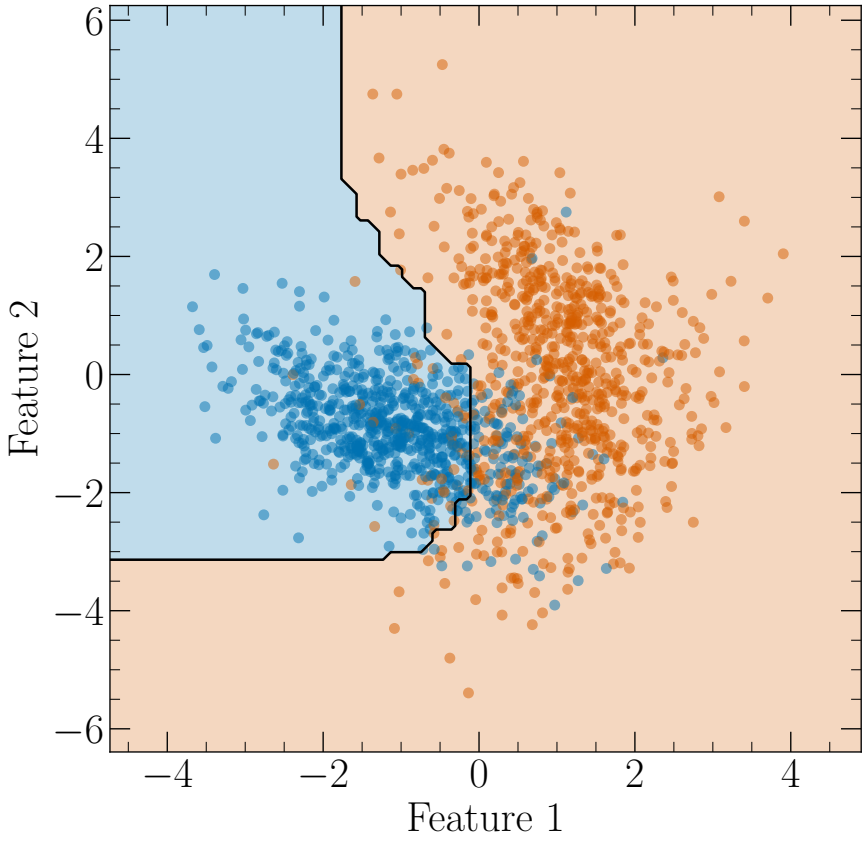

Figure 2.8 Decision boundary of the random forest classifier trained on the synthetic data. The features that are not shown are set equal to their mean values. Overplotted are the test samples colored by their true class.

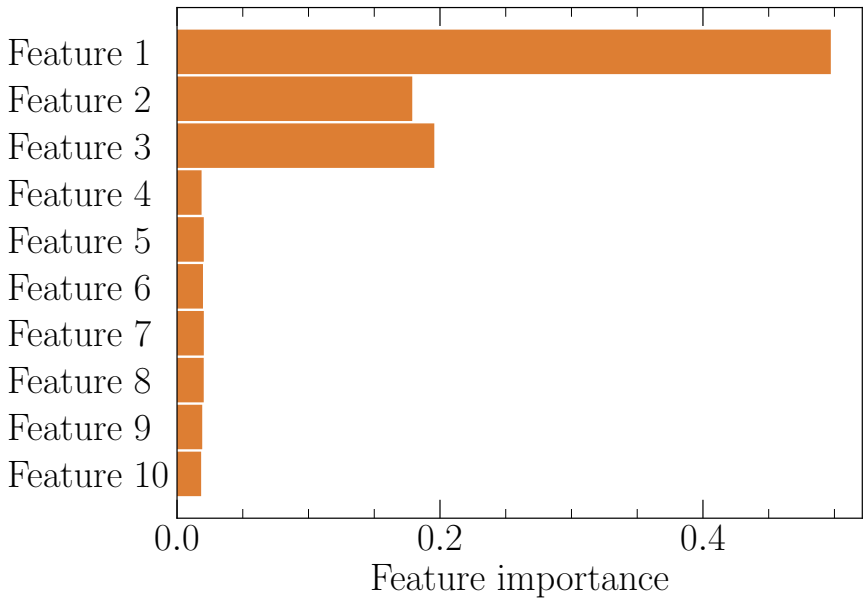

Figure 2.9 Feature importances of the random forest classifier trained on the synthetic data.

3 Consequences of parameterization choice on eclipsing binary light curve solutions*

Authors: J. Korth, A. Moharana, M. Pešta, D.R. Czavalinga, K.E. Conroy

Published in: Contributions of the Astronomical Observatory Skalnaté Pleso, vol. 51, no. 1, p. 58-67

Abstract

Eclipsing Binaries (EBs) are known to be the source of most accurate stellar parameters, which are important for testing theories of stellar evolution. With improved quality and quantity of observations using space telescopes like *TESS*, there is an urgent need for accuracy in modeling to obtain precise parameters. We use the soon to be released `PHOEBE 2.3` EB modeling package to test the robustness and accuracy of parameters and their dependency on choice of parameters for optimization.

3.1 Introduction

It is well known that eclipsing binaries (EBs) provide highly accurate observations of stellar parameters, which is important for testing theories of star evolution. Thanks to the increasingly more precise photometry of the past and recent space missions (e.g. *Kepler* (Borucki 2016) and the Transiting Exoplanet Survey Satellite (*TESS*; Ricker et al. 2014)), it is now possible to observe and study EBs in details never seen before (e.g. reflection from the companion, lensing, or Doppler beaming). To understand the observations, one needs to employ models capable of making predictions with sufficient precision, so that it is possible to compare the observations with predicted theoretical values. A popular example of such software is `PHOEBE` (Prša et al. 2016; Horvat et al. 2018; Jones et al. 2020), a robust Python package for modeling of EB systems. The latest release provides users with control over a large number of orbital and physical parameters, which allows them to generate synthetic light curves and radial velocities of the binary system. One can also take advantage of various built-in or imported solvers (e.g. `emcee`) and optimizers (e.g. Nelder-Mead) to solve the inverse problem—for a comprehensive introduction to the inverse problem using `PHOEBE` see Conroy et al. (2020a).

For the purpose of this paper, we use the soon to be released version 2.3 of `PHOEBE` to try and reproduce the results from the article by Maxted et al. (2020),

*The following text is a postprint version of an article accepted for publication in Contributions of the Astronomical Observatory Skalnaté Pleso. The first three authors contributed equally to this work and are listed in alphabetical order. The published article is available at `https://www.astro.sk/caosp/Eedition/FullTexts/vol51no1/pp58-67.pdf`.

which examines a number of various methods to accurately estimate the masses and radii for the stars in the binary system AI Phoenicis (AI Phe). This system, which contains two well-separated, sharp-lined stars of comparable luminosity, was first identified as an EB by Strohmeier (1972). It is an excellent target for model testing as it is relatively bright ($V = 8.6\,\mathrm{mag}$), has a long orbital period ($P \approx 24.59\,\mathrm{days}$), and does not show any distinct spots nor flares associated with increased magnetic activity of the components (e.g., Kirkby-Kent et al. 2016; Maxted et al. 2020).

To compare the results by Maxted et al. (2020), we first carried out a number of runs with varying underlying physical models, free parameters, and their initial values (see Section 3.2). This further motivated us to design a controlled experiment in order to systematically analyze the effect of parameterization choice on the final light curve resulting from the model (Section 3.3). Finally, in Section 3.4, we compare our results from the controlled experiment with the Maxted et al. (2020) values, and we discuss our findings regarding the precision of the employed model.

3.2 Observations and Modeling Set-up

The photometric data of AI Phe used in the subsequent analyses were obtained under the *TESS* Guest Investigator Program (G011130, P.I. Maxted; G011083, P.I. Helminiak; G011154, P.I. Prša) during Sector 2 of the *TESS* mission observed in the 2-min cadence mode (TIC 102069549). The Sector was observed for 27 days from 2458354.113259 BJD to 2458381.517643 BJD (covering both the primary and secondary eclipse), and the data were reduced by the *TESS* data processing pipeline developed by the Science Processing Operations Center (SPOC; Jenkins et al. 2016). In our analyses, we used the Pre-search Data Conditioning Simple Aperture Photometry (PDCSAP) light curve, which was additionally detrended by fitting a chain of 5^{th} order Legendre polynomials (Maxted et al. 2020, Section 2.6).

To get a sense of the effect of parameterization on the resulting values, we independently solved the inverse problem for AI Phe by using a separate model with its own set of free parameters and approximation of physical phenomena (e.g. limb-darkening law, reflection, etc.). Following the approach from Maxted et al. (2020), we initialized the parameters of the models with their estimates from Kirkby-Kent et al. (2016), which are summarized in Table 3.1.

The initialized free parameters were used as input to the Nelder-Mead algorithm (Nelder & Mead 1965) in order to refine the estimates. These estimates then served as a starting point for initial distributions (either Gaussian or uniform) of the free parameters entering the Markov Chain Monte Carlo (MCMC) algorithm implemented in the `emcee` solver (Foreman-Mackey et al. 2013), which we used to obtain posterior distributions of the relevant parameters. Furthermore, we used the following software: `Python` (Van Rossum & Drake 2009), and the `Python` libraries `Matplotlib` (Hunter 2007) and `numpy` (van der Walt et al. 2011).

Unfortunately, the individual runs did not yield satisfactory results as the obtained values showed a wide spread. Due to the different choices of parameterization and the large numbers of parameters, it was not possible to associate the observed variation with a specific parameter or a set of parameters.

Table 3.1 A list of the parameters of the binary system AI Phe that were adopted from Kirkby-Kent et al. (2016).

Parameters	Values
P (days)	24.592483
q	1.0417
e	0.1821
ω (^{o})	110.73
i (^{o})	88.502
M_1 ($M_\odot$)	1.1973
M_2 ($M_\odot$)	1.2473
R_1 ($R_\odot$)	1.835
R_2 ($R_\odot$)	2.912
T_1 (K)	6310
T_2 (K)	5237.3

Therefore, we decided to design a controlled experiment, in which we defined a "nominal" run and then we examined the effect of altering the parameters one at a time. For more information see the following section.

3.3 Controlled Experiment

The "nominal" run (which we shall denote "Run A") served as a benchmark for all the other runs ("B" through "K"), which we systematically varied from Run A in a controlled fashion—that is, for each run, we altered one aspect of the "nominal" set-up and kept the rest unchanged. For the definitions of Runs B–K, see Table 3.2.

Table 3.2 A list of the individual runs and their differences from the "nominal" run.

Run	Description
A	The "nominal" run
B	Logarithmic limb darkening law
C	Sample/interpolate in phase-space
D	Marginalization over albedos
E	Marginalization over gravity darkening parameters
F	Marginalization over gravity darkening parameters from Claret & Bloemen (2011)
G	Marginalization over noise nuissance parameter
H	Marginalization over parameters q and a using radial velocities posteriors from Gallenne et al. (2019) on q and $a \sin i$
I	Meshes on binary surfaces to estimate $L_{\rm pb}$
J	*TESS* light curve without detrending (PDCSAP)
K	Masking the out-of-eclipse points

Run A uses the binary star model ELLC (for more information see Maxted

2016) with the quadratic limb-darkening law in the "lookup" mode (automatic querying of coefficients from tables based on mean stellar values), and uses the Stefan-Boltzmann approximation in the determination of the passband luminosity, L_{pb}, which is needed to scale the fluxes and estimate the surface-brightness ratio. Similar to our initial test runs, we initialized the parameters with values from Kirkby-Kent et al. (2016), and then used the Nelder-Mead algorithm to refine the estimates. After that, we used the `emcee` solver to sample over the radii, R_1 and R_2, (for the primary and secondary component), the eccentricity, e, along with the argument of pericenter, ω_0, (parameterized as $e\sin\omega_0$ and $e\cos\omega_0$), the time of the primary eclipse, T_0, the third light, l_3, L_{pb}, the ratio of the effective temperature of the secondary and primary component, $T_{\mathrm{secondary}}/T_{\mathrm{primary}}$, and the orbital inclination, i, to get an estimate for their uncertainties. We present the obtained results in Fig. 3.1.

3.4 Discussion of Results

Analysis of Runs A–K shows a rather limited spread of the obtained values, with the individual parameters lying within each other's uncertainties (see Fig. 3.1). Seeing that the parameters have a fairly small effect on the final results, this implies that the high variation of our initial runs was most likely not caused by any specific parameter but rather by a combined effect of various parameterization choices. That said, it is still clear that the choices made in Table 3.2 do influence the final results.

As for the runs presented in Maxted et al. (2020), each of them employs a distinct combination of the underlying physical model, optimization method and parameterization (see Table 3.3). In principle, to properly compare our results with those obtained by Maxted et al. (2020), the initial set-ups should also be compared so that the effect of initial configuration choice can be distinguished from other effects. Although we compare our results with all the runs listed in Table 3.3 (see Fig. 3.1), we shall inspect in detail only the initial set-ups for Runs A and S by Maxted, as they utilize the same physical binary model (`ellc`) and optimization method (`emcee`) as our runs. Our runs use a wrapper for mapping of `PHOEBE` parameterization onto `ellc`, and thus they minimize the effect of choice (there is still the freedom of parameterization) and can serve as a "benchmark" for our results. Moreover, Runs A and S by Maxted use essentially the same initial set-ups, therefore it suffices to examine only the former (which additionally corrects for instrumental systematic variations). To avoid confusion, we shall prefix the runs by Maxted et al. (2020) with "M-" (e.g. Run A by Maxted becomes Run M-A) in the rest of the section.

For the sake of simplicity, we shall compare the parameterization of Run M-A only with our "nominal" Run A, which is assumed to be representative of the other runs (for their definitions, see Table 3.2). In Table 3.4, we compare the free parameters entering the `emcee` algorithm in the two runs. Apart from working with mostly disjunct (but correlated) sets of free parameters, the runs also differ in the limb-darkening law (power-2 for Run M-A and quadratic for Run A) and the treatment of the stellar masses. Run A adopts the values $1.1973\,\mathrm{M}_\odot$ and 1.0417 from Kirkby-Kent et al. (2016) for the mass of the primary component and the ratio of the masses of the secondary and primary component, respectively.

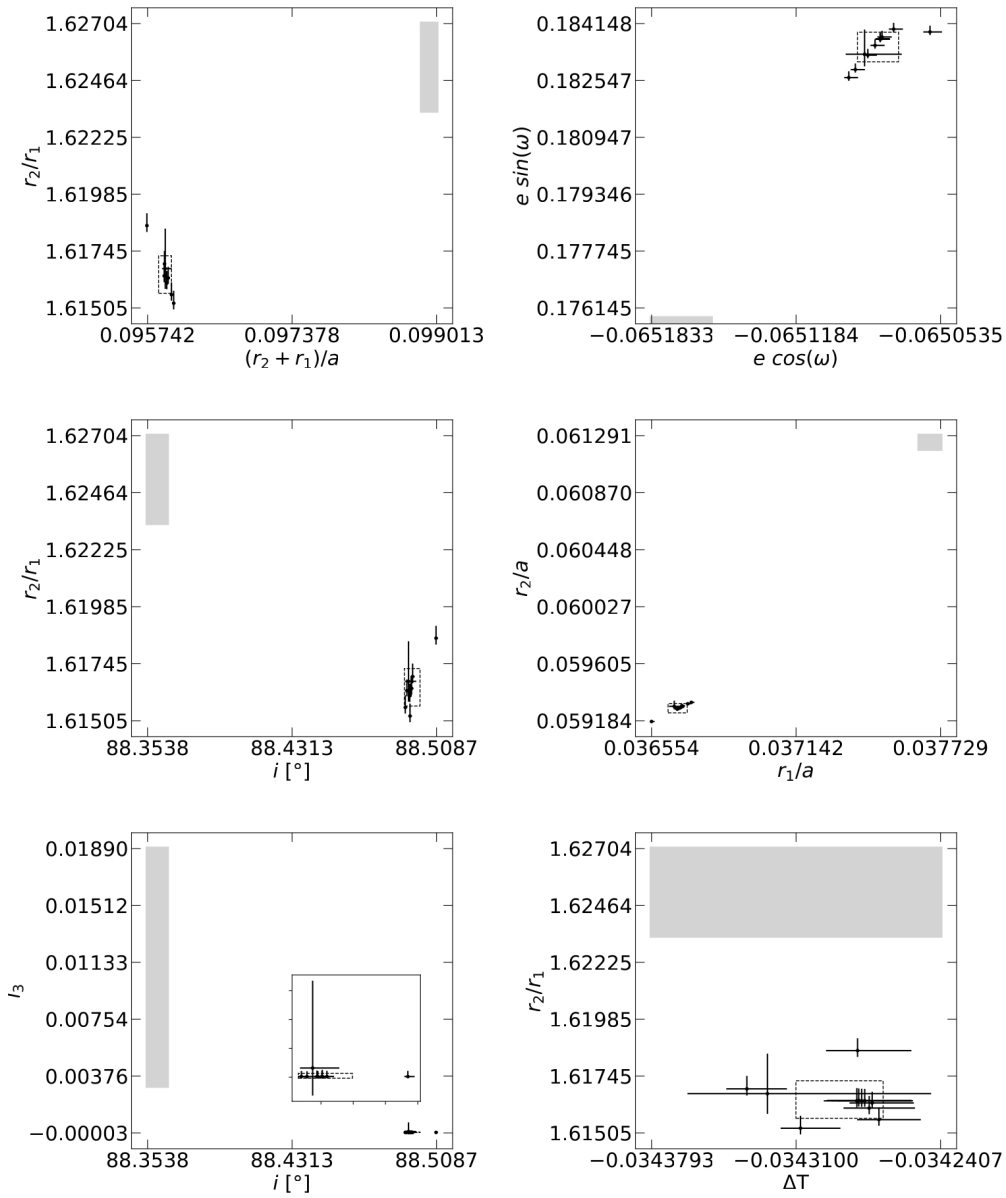

Figure 3.1 A comparison between the parameters resulting from Runs A–K and those obtained by Maxted et al. (2020). The filled rectangle represents the 1-σ spread around the average of the runs from the Maxted et al. paper (see Table 3.3), while the dotted box shows the same for our controlled runs. The inner box in the l_3 vs i plot represents a zoomed-in view on the parameter distribution.

In addition, the stellar masses are held constant at the Kirkby-Kent values as they have minimal effect on the light curve. In contrast, Run M-A uses the `emcee` posterior distributions of the free parameters together with observed radial velocities of the binary system to get an estimate for the masses. Finally, both runs hold the semi-major axis constant, with Run M-A assuming the value of 47.868 $R_\odot$ and Run A keeping it equal to 47.941 $R_\odot$.

Coming back to the general case, we expected our results to agree with those obtained by Maxted et al. (2020) within the reported uncertainties despite the differences in the initial set-ups of the runs. However, Fig. 3.1 shows that this is not the case as none of the results from Runs A–K lies in the 1-sigma spread of the

Table 3.3 An overview of the various runs analysed in the paper by Maxted et al. (2020). The table was adopted from page 6 of the mentioned paper.

Run	Investigator	Model	Optimization	Limb-darkening	Detrending	Notes
A	Maxted	`ellc`	`emcee`	power-2	`celerite`	
B	Helminiak	`JKTEPOB`	L-M	quadratic	sine+poly	Monte Carlo error estimated
C	Torres	`EB`	`emcee`	quadratic	spline	Quadratic l.d. coeffs. fixed
D	”	”	”	”	”	
E	Graczyk	`WD2007`	L-M	logarithmic	–	Fixed l.d. coefficients
F	Johnston	`PHOEBE 1.0`	`emcee`	square-root	–	
G	Prša	`PHOEBE 2.1`	MCMC	grid	legendre	
H	Orosz	`ELC`	DE-MCMC	logarithmic	polynomial	
I	Orosz	”	”	square-root	”	
J	Orosz	”	”	quadratic	”	
K	Southworth	`JKTEBOP`	L-M	quadratic	polynomial	
L	Southworth	`JKTEBOP`	L-M	cubic	polynomial	
S	Maxted	`ellc`	`emcee`	power-2	`celerite`	Same as Run A with SAP light curve

Table 3.4 A list of the free parameters entering the `emcee` algorithm in Runs M-A and A. The “Nelder-Mead” column marks the parameters which were optimized before running `emcee`, in order to speed-up the convergence of the posterior distributions.

Parameter	Description	Run M-A	Run A	
		`emcee`	Nelder-Mead	`emcee`
$e\cos\omega$			✓	✓
$e\sin\omega$			✓	✓
f	Flux scaling factor	✓		
f_c	$\sqrt{e}\cos\omega$	✓		
f_s	$\sqrt{e}\sin\omega$	✓		
$h_{1,\mathrm{F}}, h_{2,\mathrm{F}}$	Parameters of the power-2 limb darkening law for star 1	✓		
$h_{1,\mathrm{K}}, h_{2,\mathrm{K}}$	Parameters of the power-2 limb darkening law for star 2	✓		
i	Orbital inclination	✓	✓	✓
k	Ratio of the radii	✓		
l_3	Third light	✓		✓
L_{pb}	Passband luminosity			✓
r_{sum}	Sum of the fractional radii	✓		
R_1	Radius of the primary star		✓	✓
R_2	Radius of the secondary star		✓	✓
σ_f	Standard error per observation	✓		
S_{T}	Surface brightness ratio averaged over the stellar disks in the TESS band	✓		
T_0	Time of primary eclipse	✓	✓	✓
$T_{\mathrm{secondary}}/T_{\mathrm{primary}}$	Ratio of the effective temperatures		✓	✓

Maxted et al. (2020) values. As to the reason behind this discrepancy, multiple explanations present themselves. First, due to time and computational constraints, we stopped our `emcee` runs after appearing flat (converged to a single value) for about 1500 iterations compared to ∼10000 iterations for the runs in Maxted et al. (2020). Thus, there is a slight possibility that the runs might yet “jump” and converge to some other values. All of our runs, however, were treated consistently with each other, and still exhibit the influence of a number of decisions in the

fitting process on the final results. Next, the mapping between the `PHOEBE` and `ellc` parameterizations includes several assumptions and approximations, which leads to residuals between the two forward models (see Fig. 3.2). This suggests that the offset between the two sets of results might be caused by not employing the native `PHOEBE` backend.

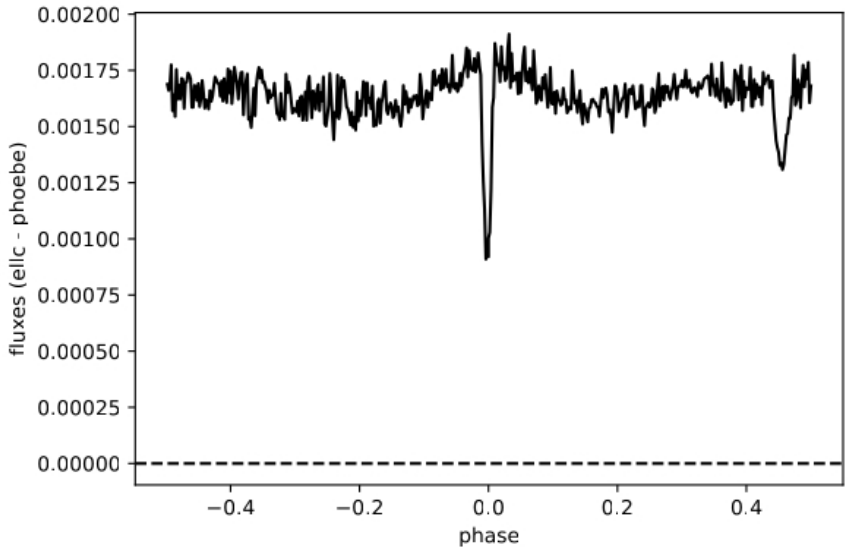

Figure 3.2 Residuals between the `ellc` and `PHOEBE` forward models for the nominal case. This shows the influence of the approximations in mapping between the parameterization of the different codes and likely explains the offset between the results from our controlled sample and those in Maxted et al. (2020).

3.5 Conclusions

In this work, we tried to fit the parameters of AI Phoenicis from the *TESS* light curve and to reproduce the values estimated by Maxted et al. (2020). First, we independently modeled the light curves with individual sets of free parameters and their approximations, e.g., difference between initialized parameters of binary masses, radius, effective temperatures of the stars, or model of limb darkening, reflections etc. Since our independent models did not lead to the same results for the system parameters of AI Phe, we designed a controlled experiment to systematically analyze the effect of the parameterization.

The parameters obtained from different runs were expected to be distributed in a parameter space as found in Maxted et al. (2020) but we find our results to be quite different in comparison to the accuracy that has been found before. We suspect the discrepancy may be reconciled by running `emcee` for substantially longer to allow further convergence and to switch to the native `PHOEBE` forward model to avoid the assumptions and approximations in the translation between parameterizations. Our results do, however, show the importance of several different choices in the fitting process on the final parameter values and their uncertainties.

The work presented here was part of the GATE Summer School, hosted (virtually) by Masaryk University Brno and sponsored by ERASMUS+ under grant number 2017-1-CZ01-KA203-035562. The authors would like to thank all involved and Marek Skarka for his organization. A.M. would like to acknowledge the support provided by the Polish National Science Center (NCN) through the grant number 2017/27/B/ST9/02727.

4 Mass-ratio distribution of contact binary stars*

Authors: M. Pešta and O. Pejcha
Published in: Astronomy & Astrophysics, Volume 672, id.A176, 27 pp.

Abstract

The mass ratio q of a contact binary star evolves through mass transfer, magnetic braking, and thermal relaxation oscillations to low values until it crosses a critical threshold $q_{\min}$. When this occurs, the binary undergoes the tidal Darwin instability, leading to a rapid coalescence of the components and to an observable brightening of the system. The distribution of q has not been measured on a sufficiently large population of contact binary stars so far because determining q for a single contact binary usually requires spectroscopy. As was shown previously, however, it is possible to infer the mass-ratio distribution of the entire population of contact binaries from the observed distribution of their light-curve amplitudes. Employing Bayesian inference, we obtained a sample of contact binary candidates from the Kepler Eclipsing Binary Catalog combined with data from *Gaia* and estimates of effective temperatures. We assigned a probability of being a contact binary of either late or early type to each candidate. Overall, our sample includes about 300 late-type and 200 early-type contact binary candidates. We modeled the amplitude distribution assuming that mass ratios are described by a power law with an exponent b and a cutoff at $q_{\min}$. We find $q_{\min} = 0.087^{+0.024}_{-0.015}$ for late-type contact binaries with periods longer than 0.3 days. For late-type binaries with shorter periods, we find $q_{\min} = 0.246^{+0.029}_{-0.046}$, but the sample is small. For early-type contact binary stars with periods shorter than one day, we obtain $q_{\min} = 0.030^{+0.018}_{-0.022}$. These results indicate a dependence of $q_{\min}$ on the structure of the components, and they are broadly compatible with previous theoretical predictions. We do not find any clear trends in b. Our method can easily be extended to large samples of contact binaries from TESS and other space-based surveys.

4.1 Introduction

A contact binary system consists of two stars that have filled their Roche lobes and started sharing a single envelope. The luminosity generated by the individual stars is efficiently distributed through the envelope, leading to a nearly constant temperature across the whole shared surface, regardless of the masses of the components (Lucy 1968b,a; Shu et al. 1976; Shu & Lubow 1981). If the orbital

*The following text is a postprint version of an article accepted for publication in Astronomy & Astrophysics. The published article is available at `https://www.aanda.org/articles/aa/pdf/2023/04/aa45613-22.pdf`.

inclination is sufficiently high, contact binaries can be observed as eclipsing W Ursae Majoris (W UMa) or EW variables. In the rest of the paper, we use the terms contact binary, W UMa variable, and EW variable interchangeably. The class of W UMa variables is characterized by equal-depth primary and secondary eclipses and by periods from approximately 0.22 days up to about one day (e.g., Rucinski 2007; Jiang et al. 2012). There are two subclasses of W UMa variables: A type (or early type) and W type (or late type). In A-type systems, both components are A or F stars, in contrast to G-K stars, which make up W-type systems. Jayasinghe et al. (2020) showed that this dichotomy of W UMa variables is most likely related to their location relative to the Kraft break, which is a sudden drop in the average rotation rate of stars in the temperature range 6200–6700 K. The drop is caused by the different efficiency of magnetic braking for stars possessing or lacking subsurface convection zones (Kraft 1967). The two subclasses also differ in the slopes of their period–luminosity–color (PLC) relations, which result from the blackbody relation applied to Roche-lobe filling stars (Rucinski 1994, 2004; Pawlak 2016). Additionally, Stępień & Gazeas (2012) argued that W-type systems with periods shorter than 0.3 days form a distinct population that is different from both A types and longer-period W types.

Observations suggest that the majority of contact binaries originate in triple systems (e.g., Pribulla & Rucinski 2006; D'Angelo et al. 2006; Hwang 2023). Some of these triples might condense directly out of the star-forming region, but they are more likely the result of dynamical interaction between independent binaries or higher multiples followed by the ejection of the excess stars (e.g., Bate et al. 2002; Tokovinin 2014; Antognini & Thompson 2016). Under the right conditions, the inner binary in the triple system is subject to the von Zeipel–Lidov–Kozai mechanism, forcing the orbital eccentricity and inclination to undergo long oscillation cycles. The cycles lead to the extraction of orbital energy through tidal friction and the inner orbit gradually shrinks (Lidov 1962; Kozai 1962; Eggleton & Kiseleva-Eggleton 2001; Fabrycky & Tremaine 2007; Naoz 2016). This process is no longer efficient when the orbital period of the inner binary reaches about 1–3 days. At that point, either magnetic braking or nuclear evolution of the more massive component takes over, and coupled with tidal friction, it reduces the period even further, leading to the formation of a contact binary system (Eggleton & Kisseleva-Eggleton 2006; Hwang & Zakamska 2020). The relative importance of the two mechanisms depends on the mass of the stars in the binary. Magnetic braking is thought to be the driving force in the formation of W-type systems, while nuclear evolution is most likely the dominant mechanism in the precontact phase of A-type systems (Yıldız 2014).

After it is formed, a contact binary evolves toward low mass ratios q on the timescale of the dominant evolutionary process. The evolution to small q is not linear, and the binary goes through a series of thermal relaxation oscillations (TROs), during which the flow of mass is temporarily reversed and the contact is broken (Lucy 1976; Flannery 1976; Robertson & Eggleton 1977; Yakut & Eggleton 2005; Paczyński et al. 2006). The cycle length of TROs is set by the thermal timescale of the secondary, which grows as q decreases. As a result, contact binaries pile up at small q (Rucinski 2001). Stępień (2006, 2011) and Stępień & Gazeas (2012) proposed an alternative to the TRO model, which assumes rapid initial mass transfer followed by a mass-ratio inversion and linear evolution toward

small q. Regardless of the actual mechanism, the trend toward unequal masses continues until the binary becomes unstable due to the tidal Darwin instability, which occurs when the spin angular momentum of the more massive component exceeds one-third of the orbital angular momentum of the system (e.g., Darwin 1879; Hut 1980). The angular momentum criterion translates into a minimum mass ratio $q_{\rm min}$ (Webbink 1976; Rasio 1995). The exact value of $q_{\rm min}$ depends on the stellar structure and masses of the components, but theoretical models generally predict values below 0.1 (Rasio 1995; Li & Zhang 2006; Arbutina 2007, 2009; Wadhwa et al. 2021). Alternatively, the binary can expand and overflow its outer critical surface before it reaches $q_{\rm min}$, leading to a rapid mass and angular momentum loss through the vicinity of the L2 point (Webbink 1977; Shu et al. 1979; Stępień & Gazeas 2012; Pejcha et al. 2016b,a; Hubová & Pejcha 2019). When the contact binary becomes unstable due to either of the two mechanisms, it enters the dynamical common envelope phase, which is accompanied by a luminous red nova transient (e.g., Tylenda et al. 2011; Ivanova et al. 2013b,a; Pejcha 2014; Pejcha et al. 2017; MacLeod et al. 2017; Blagorodnova et al. 2021) and leads to a single, rapidly rotating remnant (Paczyński et al. 2007).

The effects of all the evolutionary processes are imprinted on the mass-ratio distribution of contact binaries. Since many of these processes, such as magnetic braking and thermal and tidal instabilities, are not completely understood, we might be able to illuminate them by studying the observed mass-ratio distribution (Vilhu 1981). Surprisingly, little work has been done to observationally constrain the distribution of q on sufficiently large and homogeneous samples of contact binaries with a well-understood selection function. The reason is that to accurately estimate q of a contact binary system, spectroscopy of both components is typically required. Another option is to infer q directly from photometry, but this method does not yield reliable results due to the degeneracy of contact binary light curves with respect to q and the orbital inclination. An exception to this is the special case of totally eclipsing contact systems, for which the degeneracy is lifted and q can be reliably estimated, but precise photometry is required to resolve the shape of the minimum for the low amplitudes expected from systems with small q (Rucinski 2001; Terrell & Wilson 2005; Hambálek & Pribulla 2013). Yakut & Eggleton (2005) investigated parameters of about 100 binaries close to contact, but their systems were collected from the literature and could be a very biased representation of the actual population. More efforts have focused on the determination of $q_{\rm min}$. For several contact binaries, q is close to the theoretically predicted minimum value (e.g., Paczyński et al. 2007; Li et al. 2021; Wadhwa et al. 2021; Popov & Petrov 2022; Christopoulou et al. 2022), but it is unclear how these detections relate to the entire population. Recently, Kobulnicky et al. (2022) performed a computationally expensive Monte Carlo exploration of the light-curve parameter space for about 200 contact binaries and found that $q_{\rm min}$ increases with orbital period from 0.044 at 0.74 days to 0.15 at 2 days. However, none of these approaches scale well to the large amounts of astronomical data that have recently become available or will become available in the future.

To overcome these issues, Rucinski (2001) developed an independent method for the inference of the mass-ratio distribution using only photometric amplitudes extracted from contact binary light curves. The method does not require modeling of each contact binary system in the sample individually, but rather it exploits

the strong correlation between the shape of the mass-ratio distribution and the photometric amplitude distribution of contact binary stars. Rucinski (2001) constructed a sample of contact binaries from ground-based data and modeled the mass-ratio distribution as a power law with a cutoff at $q_{\min}$, but could not obtain any reasonable constraint on $q_{\min}$ due to the insufficient sensitivity of ground-based photometry, blending, and the limited size of their sample. Another complication is that at low photometric amplitudes, contact binary samples can be contaminated by unresolved companions, ellipsoidal variables, or various types of pulsating stars (Skarka et al. 2022). The advantage of the method of Rucinski (2001) is that it requires neither spectroscopy nor tedious modeling of individual objects, and it is therefore very well suited for current and future massive high-precision photometric surveys.

Our goal is to characterize the mass-ratio distribution and $q_{\min}$ of contact binaries with the help of high-precision photometric amplitudes that are available from space-borne telescopes. In Section 4.2 we formulate the ideas of Rucinski (2001) in the framework of Bayesian inference. In Section 4.3 we describe our initial sample of contact binary candidates from the Kepler Eclipsing Binary Catalog. In Section 4.4 we present a Bayesian model for selecting a clean sample of contact binaries using the PLC relation. In Section 4.5 we infer the mass-ratio distribution of contact binary stars, and we investigate its dependence on various parameters of our model. Finally, in Section 4.6 we summarize our findings and discuss possible future extensions and applications of the method.

4.2 Method

The light curves of contact binary stars are special because their shapes depend more strongly on the geometrical features of the system than on the intrinsic properties of the stellar components. Due to the transfer of mass and energy in the system, the two components have nearly identical effective temperatures, rendering the light curve amplitude a almost exclusively dependent on the orbital inclination i, the fill-out factor f, and the mass ratio q, that is, $a = a(i, f, q)$. The fill-out factor is usually defined by linearly relating the photospheric Roche potential to the potentials of the L1 and L2 points, giving $f = 0$ for stars barely touching at L1 and $f = 1$ for stars starting to overflow L2. The mass ratio is defined as the ratio of the less massive star to the more massive component, that is, $q \leq 1$. Higher-order physical effects such as gravitational and limb darkening do not significantly affect a, but rather influence the shape of the light curve. These effects can be used to alleviate the degeneracy between i, f, and q, but their usefulness is reduced by the necessary time-consuming modeling and the effects of stellar spots. Depending on their distribution on the surface, spots can deform the light curve and decrease or increase the observed a. The position and size of spots changes over time, which suggests that estimates of a can be improved by averaging data taken over longer periods of time. Furthermore, not the amplitude of any single system, but the overall distribution constructed from many systems is important.

In this section, we give a general overview of the procedure to obtain the distribution of q from the observed distribution of a based on Rucinski (2001) (Sect. 4.2.1). We describe the functional form of the mass-ratio distribution

(Sect. 4.2.2) and the construction of light curves and amplitude distributions (Sect. 4.2.3), and we present the Bayesian procedure for finding posteriors (Sect. 4.2.4).

4.2.1 Overview of the method

The observed distribution of light-curve amplitudes $A(a)$ for a constant f is given by

$$\begin{aligned} A(a) &= \int \delta\left(a'(i,q) - a\right) I(i) Q(q) \mathrm{d}i \mathrm{d}q = \\ &= \int I(i) Q(q(i,a)) \frac{\partial q}{\partial a}(a,i) \mathrm{d}q, \end{aligned} \tag{4.1}$$

where we assume that the joint probability distribution of (i, q, f) can be separated into individual components, specifically, $I(i)$ is the distribution of i, and $Q(q)$ is the distribution of q, which we aim to obtain. The second line of Eq. (4.1) works out an explicit form for $A(a)$ by assuming that $a(i, q)$ can be inverted to give a function $q(i, a)$. This form of $A(a)$ is similar to what was derived by Rucinski (2001). For most of this work, we suppress the dependence of $a(i, q)$ on f by assuming that all contact binaries have the same value of f. Detailed light-curve models as well as the TRO theory suggest that f is typically small, $0 \lesssim f \lesssim 0.5$, and that its distribution has a poorly defined maximum around $f \approx 0.25$ (Lucy 1973; Rucinski 1973, 1997). However, there are some indications that early-type binaries have higher f than late-type ones (Mochnacki 1981). Still, we chose $f = 0.25$ as the default value, and we investigate the sensitivity of our results to f in Sect. 4.5.4.

In Fig. 4.1 we outline our procedure for obtaining $Q(q)$ from $A(a)$. The key assumptions are that the inclinations are distributed isotropically, that is, $I(i) \propto \sin i$ (Fig. 4.1a), and that the function $a(i, f, q)$ can be calculated with a binary light-curve synthesis code (Fig. 4.1b). For a constant q, the distribution of a is peaked near the maximum a achievable for the given q (Fig. 4.1c). Following Rucinski (2001), we modeled $Q(q)$ as a power law with a cutoff at $q_{\min}$ (Fig. 4.1d, Sect. 4.5.3). The existence of $q_{\min}$ gives rise to a local maximum in $A(a)$ and prevents $A(a)$ from diverging as $a \to 0$. The value of $q_{\min}$ is directly related to the location of the local maximum, while the exponent in the power law controls the overall shape and slope of $A(a)$ (Fig. 4.1e).

By employing a specific goodness-of-fit metric, it is possible to find the optimal value of the two parameters that best match the observed data. Rucinski (2001) used χ^2-minimization applied to the binned histogram of $A(a)$. This is problematic, because binning leads to loss of information and the result might depend on the specific choice of the bins. In this work, we use Bayesian inference, which is applicable to both binned and continuous data and works even with small samples. By applying the Bayes theorem, we obtain the posterior distributions of the mass-ratio distribution parameters (Fig. 4.1f). By marginalizing over the posteriors, we smooth out the amplitude distribution and the mass-ratio distribution, obtaining their 1σ credible intervals in the process (Figs. 4.1g and 4.1h).

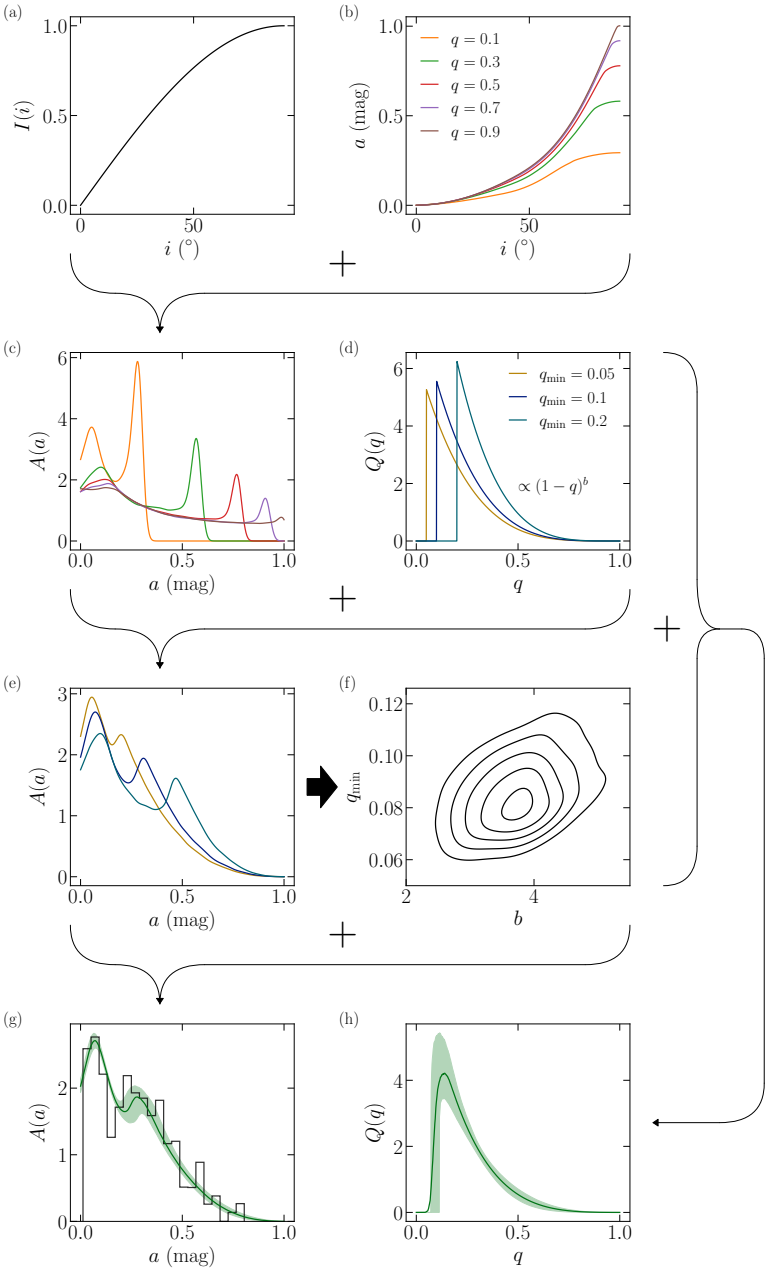

Figure 4.1 Summary of our method for the inference of the mass-ratio distribution of contact binary stars. *(a)* Assuming the orbits of contact binary systems are randomly distributed in space, the probability of observing a system with an inclination i is proportional to $\sin i$. *(b)* By using light-curve synthesis models, we derive the contact binary photometric amplitude a as a function of i for different values of the mass ratio q. *(c)* By marginalizing out the inclination, we obtain a as a function of q. *(d)* We approximate the mass-ratio distribution by a power law with index b and a sharp cutoff at the minimum mass ratio $q_{\rm min}$. *(e)* By using the power law to marginalize out q, we construct the full photometric amplitude distribution, with its shape strongly depending on the value of $q_{\rm min}$. *(f)* We apply Bayesian inference to fit the amplitude distribution to a sample of contact binary stars, yielding the posterior distribution of the parameters of the model. *(g)* We marginalize over the posterior and get the mean amplitude distribution (solid green line) and its 1σ credible interval (green band). *(h)* Repeating the same procedure, we obtain the mean mass-ratio distribution and its 1σ credible interval.

4.2.2 Mass-ratio distribution

Motivated by Rucinski (2001), we considered two different power-law prescriptions for $Q(q)$, which should capture the essential manifestations of contact binary evolution, specifically, the pile-up of objects at small q and a cutoff at $q_{\min}$. The two prescriptions are

$$Q_1(q;\Theta) = \begin{cases} \frac{1}{K} q^{-b} & \text{if } q_{\min} < q \leq 1, \\ 0 & \text{else,} \end{cases} \tag{4.2a}$$

$$Q_2(q;\Theta) = \begin{cases} \frac{1}{K} (1-q)^{b} & \text{if } q_{\min} < q \leq 1, \\ 0 & \text{else,} \end{cases} \tag{4.2b}$$

where $q_{\min}$ represents the theoretical minimum q cutoff due to the Darwin instability, b controls the slope of the power law, and $\Theta = (q_{\min}, b)$. The parameter b encodes the effect of nuclear evolution, magnetic braking, and TROs. The normalization constant K ensures that Q_1 and Q_2 integrate to unity.

4.2.3 Amplitude distribution and light-curve synthesis

After defining $I(i)$ and $Q(q;\Theta)$, we could use Eq. (4.1) to obtain $A(a;\Theta)$, but this requires inverting $a(i,q)$ to give $q(i,a)$ and calculating its derivative with respect to a. This is difficult, because $a(i,q)$ is not an analytic function of its arguments and a forward light-curve synthesis model is needed to get from (i,q,f) to a. Moreover, the derivative must be calculated numerically, which amplifies any numerical noise introduced in the calculation of $q(i,a)$. As Rucinski (2001) showed, a much simpler option is to sample the joint distribution $I(i) \times Q(q;\Theta)$ and obtain $A(a;\Theta)$ by repeated evaluation of $a(i,q)$.

We used PHOEBE version 2.3.58 (Prša & Zwitter 2005; Prša et al. 2016; Conroy et al. 2020b) to derive the functional form of a. We started by initializing the default contact binary supplied by PHOEBE. To make f unconstrained, we flipped the constraints on the potential of the envelope and the equivalent radius. Next, we changed the effective temperature of both components to 5700 K and we set the passband to "Kepler:mean". We did not change the default limb darkening and gravitational brightening coefficients. We evaluated this model on a three-dimensional grid of i, q, and f, and for each value of f, we performed linear interpolation of a as a function of i and q. The grid covers $0 \leq i \leq 90°$ with a step of $0.5°$ and $0.01 \leq q \leq 1$ with a step of 0.01. We do not expect f to have a significant impact on the shape of the light curves, therefore, we considered only six discrete values $f = (0.15, 0.25, 0.5, 0.75, \text{and } 0.99)$. We show the linearly interpolated a as a function of i and q in Fig. 4.2.

Next, we obtained the synthetic distribution of a from $Q(q;\Theta)$ by randomly sampling the joint distribution of i and q, $I(i)Q(q)$, for a given f, and we evaluated a for each sample using linear interpolation. To get from the synthetic amplitude distribution to $A(a;\Theta)$, we employed kernel density estimation (KDE), which is a nonparametric approach for estimating the probability density function from a finite sample by replacing each localized observation with a delocalized kernel function. By performing KDE with a normal kernel of bandwidth h, we obtained an analytical approximation of $A(a;\Theta)$.

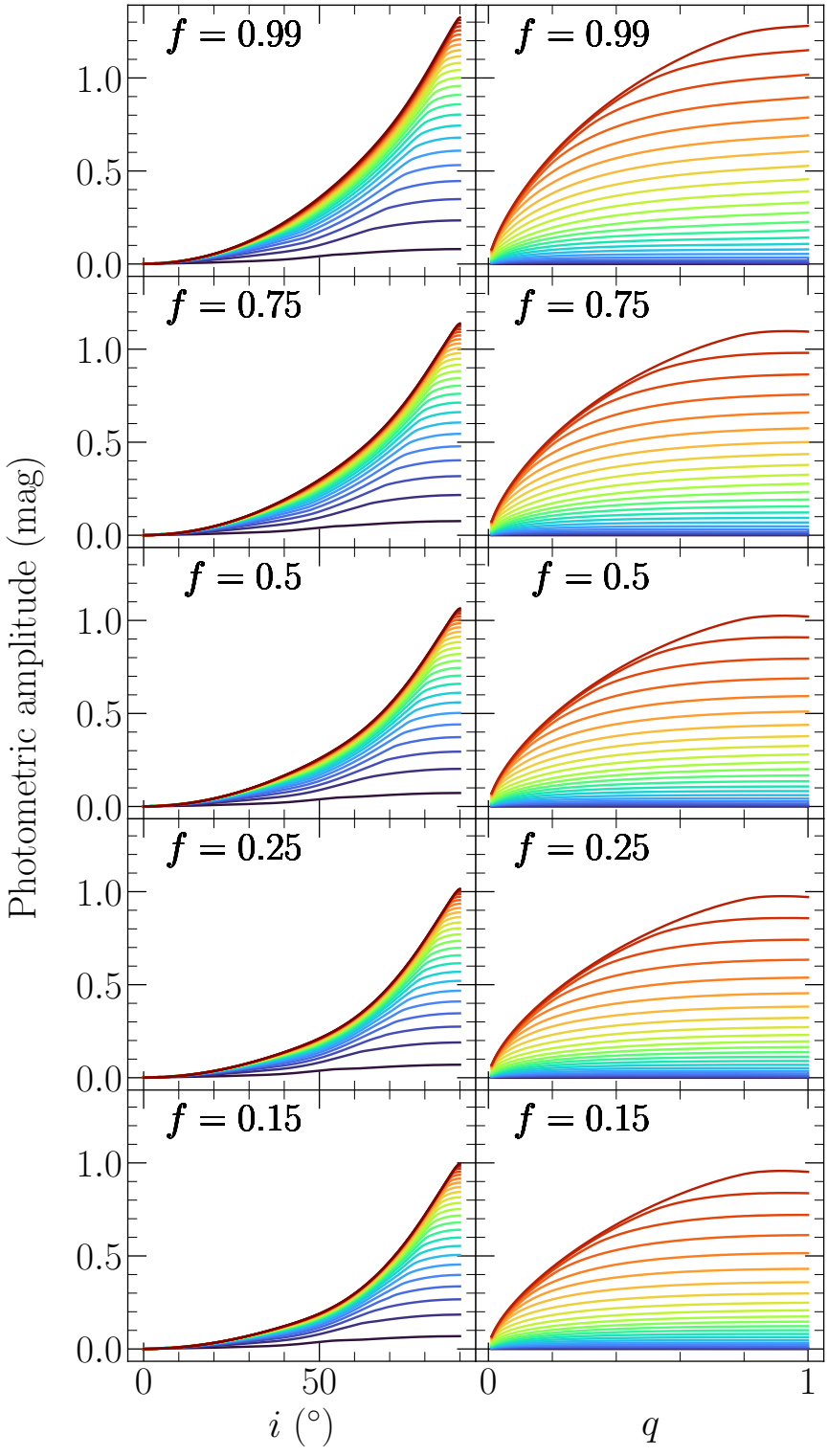

Figure 4.2 Dependence of the contact binary photometric amplitude on the orbital inclination i for different fixed values of the mass ratio q (left) and vice versa (right), conditional on distinct values of the fill-out factor f. For each panel, the transition from blue to red indicates the gradual increase in the fixed parameter from its minimum to its maximum value.

For our KDE, we chose to use a linear combination of Gaussians centered on the drawn amplitudes with standard deviations given by h. For this to converge to $A(a;\Theta)$ as the number of drawn samples goes to infinity, we ought to correct for the boundary effect, which is a downward bias near the boundaries of the KDE support. The effect is present only when the modeled distribution is nonzero near the boundaries, which is the case for $A(a \to 0)$, as demonstrated by Rucinski (2001). To mitigate the boundary effect, we followed the procedure for a first-degree boundary correction outlined by Jones (1993). We also adopted a sufficiently high value for the right boundary, so that $A(a;\Theta)$ effectively goes to zero as a approaches this value ($a \gtrsim 3$ mag). This left us with a correctly calibrated analytical approximation of $A(a;\Theta)$, which we denote by $\hat{A}(a;\Theta)$.

The stochastic process involved in the generation of the synthetic distribution entering the KDE means that $\hat{A}(a;\Theta)$ is not a deterministic function of Θ. In other words, repeated evaluation of $\hat{A}(a;\Theta)$ for the same Θ will give slightly different results depending on the number of samples entering the KDE. We explore the nondeterministic property of $\hat{A}(a;\Theta)$ in more detail in Appendix 4.A.

4.2.4 Likelihood construction

Given a sample of contact binaries, we can infer the mass-ratio distribution by assuming a specific $Q(q;\Theta)$ and fitting the resulting $A(a;\Theta)$ to the observed distribution of photometric amplitudes. In the Bayesian framework, this is achieved by evaluating the Bayes theorem and updating the prior distribution of the model parameters based on the observed data, resulting in the posterior distribution of the parameters.

The most essential ingredient entering the Bayes theorem is the likelihood function $\mathscr{L}$(model parameters|data), which is the probability of observing the data given the parameters of the model, viewed as a function of these parameters. Assuming the observed amplitudes are drawn independently from $A(a;\Theta)$, the likelihood $\mathscr{L}(\Theta|\{a_k\}_{k=1}^N)$ of Θ given a sample of N amplitudes $\{a_k\}_{k=1}^N$ is the product of the individual generative distributions weighted by the probability of each object being a contact binary star,

$$\mathscr{L}(\Theta|\{a_k\}_{k=1}^N) = \prod_{k=1}^{N} p_{\mathrm{CB},k} \int \hat{A}(a;\Theta)\mathscr{N}(a;a_k,\sigma_{a_k})\mathrm{d}a, \tag{4.3}$$

where $p_{\mathrm{CB},k}$ is the weight of the k-th object and we use the KDE approximation $\hat{A}(a;\Theta)$ instead of the true distribution $A(a;\Theta)$. In addition to the sampling noise of $\hat{A}(a;\Theta)$, each observed a_k also comes with its own uncertainty σ_{a_k}, as we discuss in more detail in Sect. 4.3.3. To factor the uncertainties into the model, we convolved $\hat{A}(a;\Theta)$ with normalized Gaussians $\mathscr{N}(a;a_k,\sigma_{a_k})$, which we used to model the uncertainty of the observed amplitudes. This procedure smears the likelihood in Eq. (4.3) even further. Fortunately, the specific form of $\hat{A}(a;\Theta)$ given by a sum of Gaussians makes it relatively fast and straightforward to perform the convolutions.

We used *emcee* (Foreman-Mackey et al. 2013) to sample the posterior distribution of the parameters. As a compromise between accuracy and computational cost, most runs were carried out with a KDE smoothing bandwidth $h = 0.02$ and KDE number of Gaussians $n = 10000$. Running on 16 logical cores

in parallel, a typical run with 16 walkers and 2500 steps takes 30 to 50 minutes to complete. We discarded the first 500 steps of each chain as burn in. We also thinned the chains by a factor of 20, which is higher than the autocorrelation time of most of our runs (Table 4.C.1).

4.3 Data

Our method for the inference of the mass-ratio distribution requires highly precise photometric measurements of contact binary light curves. At the time of writing, no catalog of contact binaries is available that would satisfy this requirement. For this reason, we constructed our own sample of contact binaries based on the photometry from *Kepler* (Borucki et al. 2010). First, we took the Kepler Eclipsing Binary Catalog (KEBC; Prša et al. 2011; Kirk et al. 2016; Abdul-Masih et al. 2016) (Sect. 4.3.1) and combined it with data from *Gaia* and other catalogs (Sect. 4.3.2). Finally, we determined the photometric amplitude of each object in the sample using detrended *Kepler* fluxes (Sect. 4.3.3).

4.3.1 Kepler Eclipsing Binary Catalog

The third revision of the KEBC contains 2920 eclipsing and ellipsoidal systems in the primary mission field of view of *Kepler*. The online version of the catalog* also includes the data from the K2 mission (K2 Engineering and C1–C5; Howell et al. 2014), increasing the total number of observed systems to 3584. The selection function of stars observed by *Kepler* is well understood (Batalha et al. 2010), and we mitigated its effects by considering late- and early-type contact binaries separately (Sect. 4.4). The construction of the KEBC involved manual filtering of objects that might affect the selection efficiency as a function of amplitude. However, the photometric precision of *Kepler* allows comfortable detection of signals with $a \ll 0.01\,\mathrm{mag}$, which is much smaller than $a \approx 0.2\,\mathrm{mag}$, where we expect the local maximum of the amplitude distribution (Fig. 4.1). We therefore did not perform any correction of the sample, effectively assuming that the selection efficiency is 100% in the range of amplitudes of our interest.

In addition to basic astrometric and photometric data, the catalog also contains output from the Kepler Eclipsing Binary Pipeline, which is a collection of several methods that are used to extract additional information from the data. For instance, locally linear embedding (LLE; Matijevič et al. 2012) is used to obtain the *morph* parameter that quantifies the detachedness of binary systems. Values of *morph* between 0 and 0.5 indicate a detached binary system, while over-contact systems usually have values from 0.7 to 0.8. The interval from 0.5 to 0.7 is occupied by semidetached systems, and values above 0.8 but below 1 correspond to ellipsoidal variables. The pipeline also includes polyfit, which is a polynomial-chain approximation used for light-curve fitting (Prša et al. 2008). Polyfit yields normalized fluxes, which can be used to calculate light-curve amplitudes.

*`http://keplerEBs.villanova.edu`

4.3.2 Cross-match with other catalogs

Using the *CDS XMatch Service* with a matching radius of 5″, we combined the KEBC with *Gaia* DR2 (Gaia Collaboration et al. 2018) and *Gaia* EDR3 (Brown et al. 2021). *Gaia* DR3 was not available at the time when we finalized the data we used for this study. After matching, we had multiple *Gaia* objects for some of the KEBC objects. To ensure uniqueness of the match, we calculated the relative fluxes of the individual *Gaia* sources and retained only the objects with relative *Gaia* EDR3 fluxes higher than 99%. We obtained luminosities from *Gaia* DR2 (field *lum_val*). These luminosities are based on the Apsis–FLAME pipeline, which assumes single stars. This is not entirely appropriate for contact binaries, and we discuss possible improvements in Sect. 4.6. Next, we cross-matched the sample with a catalog of stellar effective temperatures for objects in *Gaia* DR2 constructed by Bai et al. (2019). They obtained the temperatures by performing regression on stars from four spectroscopic surveys: the Large Sky Area Multi-Object Fiber Spectroscopic Telescope, the Sloan Extension for Galactic Understanding and Exploration, the Apache Point Observatory Galactic Evolution Experiment, and the Radial Velocity Extension. Bai et al. (2019) found that the temperatures estimated in this way are precise to about 200 K. Finally, we excluded any system for which information about its period, luminosity, or effective temperature was lacking, which reduced our sample to 2353 objects.

4.3.3 Determination of amplitudes

The KEBC does not specify the photometric amplitudes of the systems, but it is possible to calculate them from polyfit, which is available for most objects in the catalog. However, polyfit does not return amplitude uncertainties resulting from time variation of light curves, and in some cases, it even yields incorrect amplitudes. For this reason, we chose to estimate the amplitudes directly from the detrended fluxes included in the catalog. For each object observed during the primary mission of *Kepler*, we took the long-cadence data and divided them into blocks corresponding to 18 *Kepler* quarters (Q0–Q17). Most quarters represent ∼90 days of observation. For systems observed during the K2 mission, we divided the data into ten equal-size blocks. Before we tried to estimate the amplitudes, we ran some checks to ensure the completeness of the phase-folded light curves observed during the individual blocks, and we excluded the blocks with very few data around phase 0.

Within each block, we estimated the mean minimum flux. First, we sorted the data by ascending flux. Then, we iterated over the individual data points and applied local sigma clipping, taking only a close neighborhood of each data point in the phase and flux space into account. When the flux of the data point was not within three standard deviations from the mean of its neighbors, we excluded it as an outlier. In the opposite case, we stopped the iteration and proceeded to the next step.

Second, based on the distribution of the fluxes in the neighborhood of the selected data point, we distinguished between light curves with wide, narrow, and sharp minima. For each type of light curve, we adopted a different method for the calculation of the minimum flux. In the case of wide minima, which are characteristic of continuously varying light curves relevant to our analysis, we

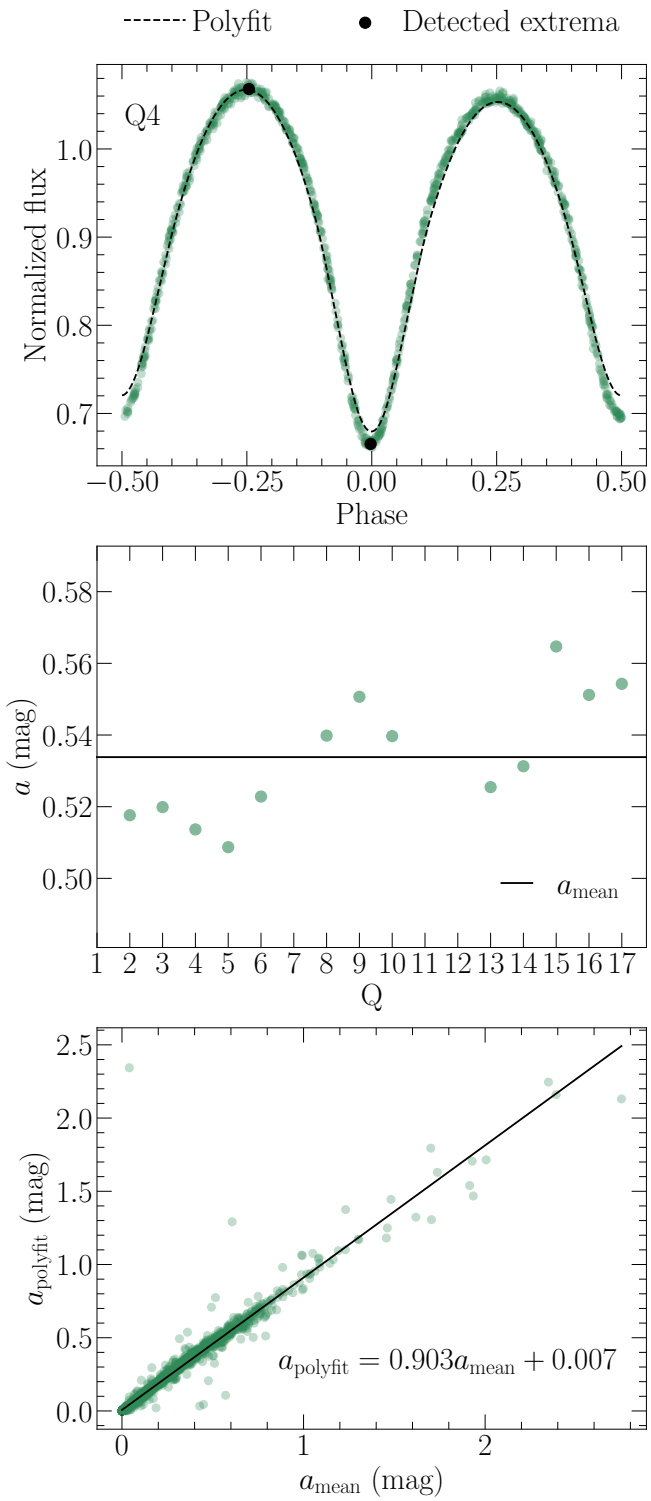

Figure 4.3 Procedure for estimating light-curve amplitudes. Top panel: Detrended *Kepler* light curve of the contact binary KIC 7871200 observed during quarter Q4. The dashed line is the result of polyfit. The data points labeled "Detected extrema" correspond to the minimum and maximum normalized fluxes resulting from the procedure described in Sect. 4.3.3. Middle panel: Photometric amplitudes extracted from the light curves observed during the individual *Kepler* quarters. The solid black line represents the arithmetic average of the values. Bottom panel: Comparison of photometric amplitudes resulting from our procedure (x-axis) and polyfit (y-axis). The slope of the line indicates a downward bias in the estimates from polyfit, most likely resulting from its tendency to underestimate the depth of light-curve minima.

approximated the vicinity of the data point selected in the previous step with a second-degree polynomial and we localized its minimum.

We estimated the mean maximum flux in a similar manner, with the exception of some systems, for which we took the median of the observed fluxes as the maximum. The top panel in Fig. 4.3 illustrates the procedure on the light curve of contact binary KIC 7871200. Using the Pogson equation, we then calculated the amplitude within each block and obtained the mean amplitude and its standard deviation over all blocks (Fig. 4.3, middle panel). For some systems, only one or two blocks passed the completeness checks, making it difficult to reliably estimate the amplitude uncertainty. For these systems, we calculated the amplitude on the full data set and assumed an uncertainty of 4%, which is comparable to the median uncertainty of $\sim$3.6% obtained from the systems with more than two complete blocks.

After we determined the mean amplitude for each object in the sample, we compared our method with polyfit (Fig. 4.3, bottom panel), and we fit the relation between the two with a straight line. The slope of the line indicates a downward bias in the estimates from polyfit. The origin of the bias is not obvious, but visual inspection of randomly chosen light curves reveals the tendency of polyfit to underestimate the depth of light-curve minima.

4.4 Identification of contaminants

When a classification based on light-curve morphology is attempted, a sample of contact binaries can become contaminated by various types of pulsating variable stars, rotating spotted stars, or ellipsoidal variables. By ellipsoidal variables we mean binaries with at least one tidally deformed star, but without a Roche-overflowing shared envelope. This contamination is especially prominent at low amplitudes, where a clean sample is crucial for determining $q_{\rm min}$. The KEBC has a disproportionately large number of objects with $a < 0.01\,$mag, and it is unlikely that most of them are true contact binaries.

To obtain a clean sample of contact binaries, we employed additional information in the form of a PLC relation, which is a combined constraint based on the Roche geometry, the third Kepler law, and the Stefan–Boltzmann law. For example, ellipsoidal variables at a given color and period will appear as underluminous compared to contact binaries because the area of their stellar surface is smaller. Rotating spotted stars at a given luminosity and temperature can have a range of rotational periods that are often longer than the corresponding Keplerian orbital period. Similarly, genuine contact binaries with bright unresolved companions that contaminate the *Kepler* photometry and reduce the observed amplitude also appear as outliers to the PLC relation. This is important because a large fraction of contact binaries should have a companion (Pribulla & Rucinski 2006; D'Angelo et al. 2006). Because of the steepness of the mass–luminosity relation on the main sequence, only companions with masses similar to or higher than the mass of the contact binary influence the amplitude and cause the object to deviate from the PLC relation. By identifying these outliers and removing them from our sample, we mitigated the effect of third light on the observed distribution of photometric amplitudes. We already removed the stars with bright companions resolved by *Gaia* from our sample in Sect. 4.3.2.

By modeling the population of contact binaries as a tube in the PLC space, we filtered out most of the contaminants. We followed the general approach for Bayesian data fitting and mixture modeling outlined in Hogg et al. (2010). That is, we viewed our sample as a mixture of a genuine signal (contact binaries) and background noise (everything except contact binaries). We modeled the components using different generative models, with each model conditional on its own set of parameters. Our analysis was based on luminosities from *Gaia* DR2, which are obtained under the assumption that the sources are single stars. Despite this drawback, we show in the following sections that our filtering method works even with these data.

We give a general overview of the method in Sect. 4.4.1, and we construct the generative model of the problem in Sect. 4.4.2. In Sect. 4.4.3 we obtain the posteriors of the model parameters, and we calculate the probability of being a contact binary of either late or early type for each object in our sample in Sect. 4.4.4. Finally, we present our clean sample of contact binaries in Sect. 4.4.5.

4.4.1 Intrinsic scatter of the PLC relation

We model the PLC relation as a straight line in the $\pi\lambda\tau$-space, where $\pi = \log(P/\mathrm{d})$, $\lambda = \log(L/L_\odot)$, and $\tau = \log(T_\mathrm{eff}/\mathrm{K})$. The relation is not exact, but rather has an intrinsic scatter, which means that the data points can depart from the relation even if we were able to observe all variables with perfect accuracy. Intrinsic scatter is generically present whenever some additional unmeasured quantities affect the measurements but are not accounted for in the relation.

The standard practice is to model the observed data as a single realization of a sequence of independent and identically distributed random variables, in which case the probability distribution of the whole sample is simply the product of the probability distributions from which the individual data points are drawn. In reality, noise is also present, and each data point is measured with a finite uncertainty. This implies that unlike the idealized noise-free data, the actually observed data are not distributed according to the intrinsic-scatter distribution, but rather each observed data point is drawn from an effective distribution given by the convolution of the intrinsic scatter with the uncertainty distribution of the data point. This is true if the individual data points are drawn independently, which we assumed implicitly.

We modeled the effective distribution in the framework of Bayesian inference. In principle, the effective distribution is different for each data point, but when we assume that all data points share the same uncertainty distribution, then formally, they are all drawn from the same effective distribution. In the absence of uncertainty measurements or a physically motivated prescription for the intrinsic scatter, it is more practical to directly model the effective distribution than the two convolution components individually. This allows us to view the originally noisy data generated from the intrinsic scatter as though they were noise free, but generated from the effective distribution, which implicitly reflects the uncertainty properties of the data.

4.4.2 Generative model

We constructed the generative model of the signal as the product of the effective distributions evaluated for each data point, but since the data are effectively noise free, we may, at least formally, fix the coordinates of the data points along a chosen dimension and instead construct a generative model conditional on these values. The difference is that the conditional generative model yields the probability of drawing a random sample with the same observed coordinates along the chosen dimension as the original sample, while the full generative model imposes no such constraint. This reduces the complexity of our problem significantly, as it is much easier to parametrically model a sequence of normalized cuts through the effective distribution than the effective distribution itself. However, this is efficient only when the parameters of the normalized cuts vary continuously with the independent variable, as otherwise the number of the parameters of the conditional generative model would scale with the size of the sample.

Traditionally, the variable that is known with the smallest uncertainty is treated as the independent variable. If we were to follow this approach for the signal, we would model the effective distribution of λ and τ conditional on π because π can be measured with the highest accuracy. However, the choice of the independent variable is not that essential when the effective distribution rather than the intrinsic scatter is modeled directly. Motivated by this, we constructed the generative model for the signal conditional on λ instead of π. We emphasize that our point here is not an accurate characterization of the PLC relation, but rather efficient distinction between contact binaries and contaminants.

In Fig. 4.4 we show the Hertzsprung–Russell diagram of our sample, which reveals that most data points lie on the main sequence. We excised the data points that are located above the line $\lambda = 9.09\tau - 33.18$ (solid black line in Fig. 4.4). The periods of most of the removed objects are longer than a few days (bottom panel in Fig. 4.4), which indicates subgiant or giant components. In addition, visual inspection of the light curves reveals that many of the removed objects are similar to heartbeat stars or have other light curve peculiarities. This reduces the number of the objects in the sample to 2172 and allows us to adopt a particularly simple conditional generative model for the background, where the background data points are generated from a plane with a nonzero Gaussian thickness in τ.

We modeled the conditional effective distribution of the signal as a two-dimensional uncorrelated Gaussian distribution in π and τ centered on the PLC relation. The total conditional effective distribution $p(\pi, \tau|\lambda, \theta)$ of a single data point is a weighted sum of the conditional effective distributions for the signal $p_{\mathrm{S}}(\pi, \tau|\lambda, \theta)$ and the background noise $p_{\mathrm{B}}(\pi, \tau|\lambda, \theta)$,

$$p(\pi, \tau|\lambda, \theta) = X p_{\mathrm{S}}(\pi, \tau|\lambda, \theta) + (1 - X) p_{\mathrm{B}}(\pi, \tau|\lambda, \theta), \tag{4.4}$$

where X is the weight parameter, and θ is a vector of all the model parameters. We write the conditional effective distributions as

$$p_{\mathrm{S}}(\pi, \tau|\lambda, \theta) = \mathscr{N}(\pi; \mu_{\mathrm{S}\pi}, \sigma_{\mathrm{S}\pi}) \mathscr{N}(\tau; \mu_{\mathrm{S}\tau}, \sigma_{\mathrm{S}\tau}), \tag{4.5}$$

$$p_{\mathrm{B}}(\pi, \tau|\lambda, \theta) = \begin{cases} \frac{\mathscr{N}(\tau; \mu_{\mathrm{B}\tau}, \sigma_{\mathrm{B}\tau})}{\pi_{\max} - \pi_{\min}} & \text{if } \pi_{\min} \leq \tau \leq \pi_{\max}, \\ 0 & \text{else}, \end{cases} \tag{4.6}$$

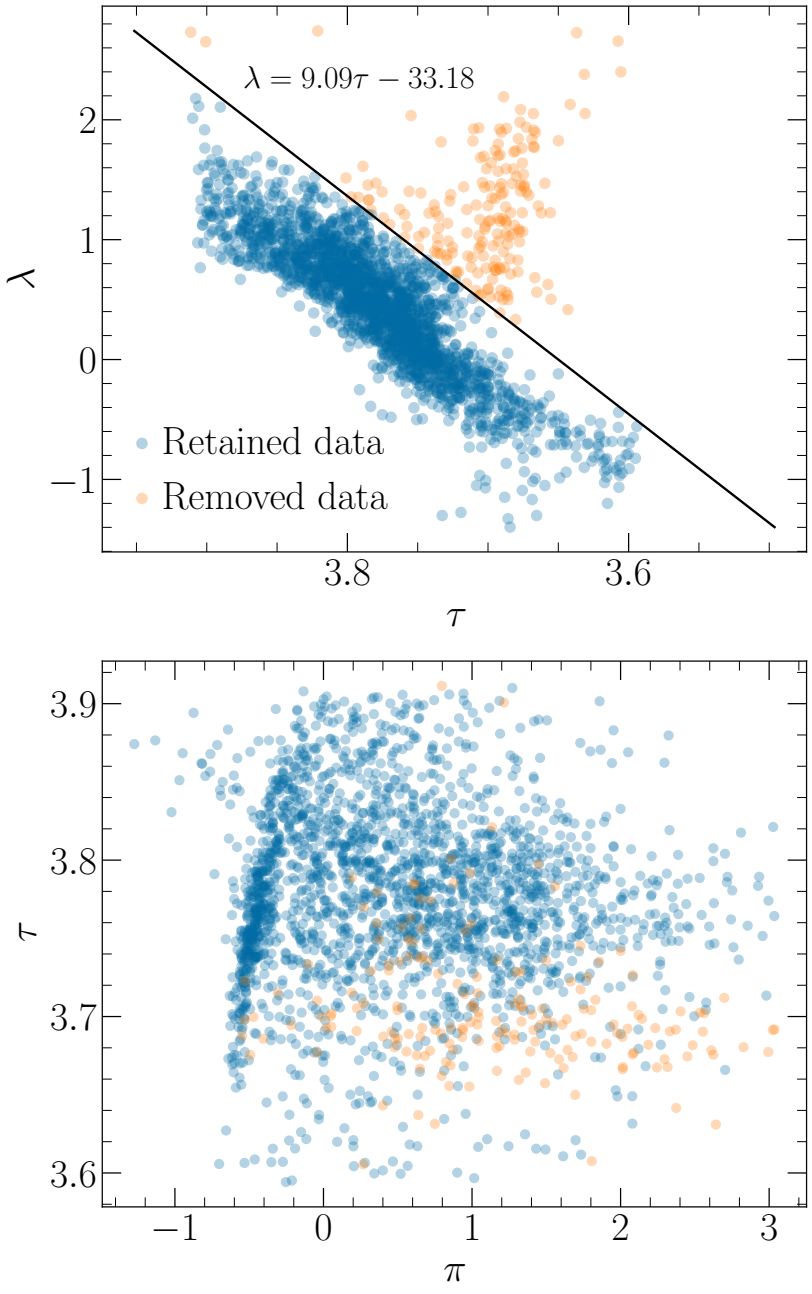

Figure 4.4 Projection of our sample to the $\tau\lambda$-plane (Hertzsprung–Russell diagram, top panel) and $\pi\tau$-plane (bottom panel). Most objects in the KEBC are constrained to the main sequence perpendicular to the $\tau\lambda$-plane. We remove objects lying above the solid black line drawn in the top panel. The bottom panel shows that the periods of most excised objects are inconsistent with the object being a typical contact binary star.

where we suppressed the dependence of the parameters of the Gaussians on λ. The specific form of Eq. (4.6) comes from the assumption that the background noise is distributed uniformly between the minimum and maximum observed values of π, denoted by $\pi_{\min}$ and $\pi_{\max}$.

To account for a possible change in the slope of the PLC relation, we considered a separate effective distribution for late-type and early-type contact binaries. The two distributions are formally the same, but each has its own set of parameters and is applicable only to a subset of the signal, disjunct from the other subset. We suspect that the dividing line between the samples will be along the Kraft break, but instead of using previously determined dividing lines such as the one from Jayasinghe et al. (2020), we included the location of the break in our model as one of the parameters. The only constraint was that the transition in the PLC relation from late types to early types is smooth. From the modeling point of view, it is more convenient to model the break along the λ-axis rather than τ-axis, as λ acts as the independent variable. This is possible because the PLC relation provides us with a unique mapping between τ and π. In general, we need two separate models for the background as well, because the parameters of the noise may also change at the Kraft break. However, in the case of the noise, we did not require that the transition is continuous.

Using subscripts 1 and 2 to refer to the parameters of the models below and above the break, we assumed the following functional dependencies for the parameters of the Gaussians:

$$\mu_{\mathrm{S}\pi} = \begin{cases} \alpha_{\pi 1} + \beta_{\pi 1}\lambda & \text{if } \lambda \leq \lambda_{\mathrm{K}}, \\ \alpha_{\pi 1} + (\beta_{\pi 1} - \beta_{\pi 2})\lambda_{\mathrm{K}} + \beta_{\pi 2}\lambda & \text{if } \lambda > \lambda_{\mathrm{K}}, \end{cases} \tag{4.7}$$

$$\sigma_{\mathrm{S}\pi} = \begin{cases} \alpha_{\sigma_{\pi 1}} + \beta_{\sigma_{\pi 1}}\lambda & \text{if } \lambda \leq \lambda_{\mathrm{K}}, \\ \alpha_{\sigma_{\pi 2}} + \beta_{\sigma_{\pi 2}}\lambda & \text{if } \lambda > \lambda_{\mathrm{K}}, \end{cases} \tag{4.8}$$

$$\mu_{\mathrm{S}\tau} = \begin{cases} \alpha_{\tau 1} + \beta_{\tau 1}\lambda & \text{if } \lambda \leq \lambda_{\mathrm{K}}, \\ \alpha_{\tau 1} + (\beta_{\tau 1} - \beta_{\tau 2})\lambda_{\mathrm{K}} + \beta_{\tau 2}\lambda & \text{if } \lambda > \lambda_{\mathrm{K}}, \end{cases} \tag{4.9}$$

$$\sigma_{\mathrm{S}\tau} = \begin{cases} \alpha_{\sigma_{\tau 1}} + \beta_{\sigma_{\tau 1}}\lambda & \text{if } \lambda \leq \lambda_{\mathrm{K}}, \\ \alpha_{\sigma_{\tau 2}} + \beta_{\sigma_{\tau 2}}\lambda & \text{if } \lambda > \lambda_{\mathrm{K}}, \end{cases} \tag{4.10}$$

$$\mu_{\mathrm{B}\tau} = \begin{cases} m_1 + l_1\lambda & \text{if } \lambda \leq \lambda_{\mathrm{K}}, \\ m_2 + l_2\lambda & \text{if } \lambda > \lambda_{\mathrm{K}}, \end{cases} \tag{4.11}$$

$$\sigma_{\mathrm{B}\tau} = \begin{cases} w_1 & \text{if } \lambda \leq \lambda_{\mathrm{K}}, \\ w_2 & \text{if } \lambda > \lambda_{\mathrm{K}}, \end{cases} \tag{4.12}$$

where λ_{K} denotes the location of the break along λ. The prescriptions for $\mu_{\mathrm{S}\pi}$ and $\mu_{\mathrm{S}\tau}$ above the break ($\lambda > \lambda_{\mathrm{K}}$) derive from the requirement that the PLC relation is continuous at the transition.

Since we employed conditional generative models instead of full generative models, we only required that the total conditional effective distribution is normalized within a given slice of constant λ and not in the whole parameter space. This allowed us to model the weight parameter X as a function of λ, and

the simplest nontrivial choice is to assume linear dependence,

$$X = \begin{cases} \alpha_{X1} + \beta_{X1}\lambda & \text{if } \lambda \leq \lambda_{\rm K}, \\ \alpha_{X2} + \beta_{X2}\lambda & \text{if } \lambda > \lambda_{\rm K}. \end{cases} \tag{4.13}$$

We emphasize that this choice does not imply that the probability of being a contact binary for each individual object depends linearly on λ. Instead, X sets the relative weight of the signal and the noise for the objects within a given slice, and the linear dependence on λ is just the simplest nontrivial model that can be employed. In total, our model has 25 parameters, which are listed together with their definitions in Table 4.B.1.

Denoting the jth object in our sample of M objects ($M \geq N$) with the subscript j, we can write the likelihood of the total conditional generative model as

$$\mathscr{L}(\theta|\{\pi_j, \tau_j\}_{j=1}^{M}) = \prod_{j=1}^{M} p(\pi_j, \tau_j|\lambda_j, \theta), \tag{4.14}$$

where the curly brackets are shorthand for iterating over all objects in the sample.

4.4.3 Posterior sampling

Following the Bayesian approach, we modeled the parameters of the model as random variables. We assumed that the prior distribution $p(\theta)$ is separable, and we assigned a uniform prior to each parameter. Using the Bayes theorem,

$$p(\theta|\{\pi_j, \tau_j\}_{j=1}^{M}) = \frac{\mathscr{L}(\theta|\{\pi_j, \tau_j\}_{j=1}^{M})p(\theta)}{p(\{\pi_j, \tau_j\}_{j=1}^{M})}, \tag{4.15}$$

we then arrive at the joint posterior distribution $p(\theta|\{\pi_j, \tau_j\}_{j=1}^{M})$.

We used *emcee* to sample the posterior distribution. With a total of 50 walkers, we ran the sampler for 160 000 steps to ensure that the chains converge. The steps in the chains were generated via differential evolution, and we discarded the first 10000 steps as burn-in. To reduce autocorrelation, we only considered every 300th sample in the chains. We optimized the efficiency by sampling the distribution in two steps. First, we prescribed rather broad priors for the parameters and ran the sampler for a few thousand steps. Then we restricted the priors based on the results from the initial run, and we ran the sampler again for the full 160 000 steps. Fig. 4.B.1 and 4.B.2 show the chain plots and corner plots resulting from the run. We present the median values of the parameters together with their $16th$ and $84th$ percentiles in Table 4.B.1.

4.4.4 Probability calculation

After we obtained the posterior distribution, we assigned the probability of being a contact binary star to each data point in our sample. The idea is that given a sample in a slice of constant λ, the probability of being a contact binary system is simply the conditional probability of being drawn from the conditional effective distribution of the signal. In general, the probability of the jth data point being a contact binary depends on the value of θ, and a straightforward derivation

Table 4.1 Photometric amplitudes and probabilities of being a contact binary of either late or early type for the objects in our sample.

KIC/EPIC	a (mag)	σ_a (mag)	p_{CBL}	p_{CBE}
1026032	0.085360	0.001293	0.0000	0.0000
1026957	0.001115	0.000429	0.0000	0.0000
1432214	0.098101	0.000707	0.0000	0.0000
1571511	0.022127	0.000116	0.0000	0.0000
1572353	0.116464	0.004660	0.9474	0.0000
⋮	⋮	⋮	⋮	⋮
212163353	0.098400	—	0.0000	0.0000
212175535	0.080279	—	0.9402	0.0000

Notes. The full table is available online. For some systems, we were not able to reliably estimate their amplitude uncertainties. In our analysis, we adopted an uncertainty of 4% for these systems (Sect. 4.3.3).

for a mixture of two distributions with a fixed θ within a slice of constant λ gives us

$$p_{\mathrm{CB},j}(\theta) = \frac{X_j p_{\mathrm{S}}(\pi_j, \tau_j|\lambda_j, \theta)}{X_j p_{\mathrm{S}}(\pi_j, \tau_j|\lambda_j, \theta) + (1 - X_j) p_{\mathrm{B}}(\pi_j, \tau_j|\lambda_j, \theta)}, \tag{4.16}$$

where $X_j \equiv X(\lambda_j, \theta)$. When we replace the total contact binary effective distribution $p_{\mathrm{S}}(\pi_j, \tau_j|\lambda_j, \theta)$ with the part purely below or above the break and assume that the distribution is zero for the data points on the opposite side of the break, we obtain the probabilities $p_{\mathrm{CBL},j}(\theta)$ and $p_{\mathrm{CBE},j}(\theta)$ of being a contact binary of either late or early type.

To remove the dependence on θ, we calculated the average probability using the thinned posterior sample that we obtained from *emcee*. This amounts to marginalizing out the parameters of the conditional generative model using the posterior probability distribution, that is,

$$p_{\mathrm{CB},j} = \int p_{\mathrm{CB},j}(\theta) p(\theta|\{\pi_j, \tau_j\}_{j=1}^{M}) \mathrm{d}\theta. \tag{4.17}$$

Again, if we replaced $p_{\mathrm{CB},j}$ with $p_{\mathrm{CBL},j}$ or $p_{\mathrm{CBE},j}$, we would obtain the marginalized probabilities of being a contact binary of either late or early type.

4.4.5 Clean sample of contact binaries

In Table 4.1 we show the photometric amplitudes and calculated probabilities for the objects in our sample. In Fig. 4.5 we show a projection of our sample to the $\pi\tau$-plane, where the color of each point indicates the probability of being a contact binary of either late or early type. Our model assigns late-type contact binaries to a tight locus around the PLC relation, while for early-type binaries the scatter is significantly larger. This is consistent with previous results (e.g., Jayasinghe et al. 2020). Similarly, the position of the break in the PLC relation matches the results of Jayasinghe et al. (2020) remarkably well, even though no prior information other than the existence of a break along the PLC relation

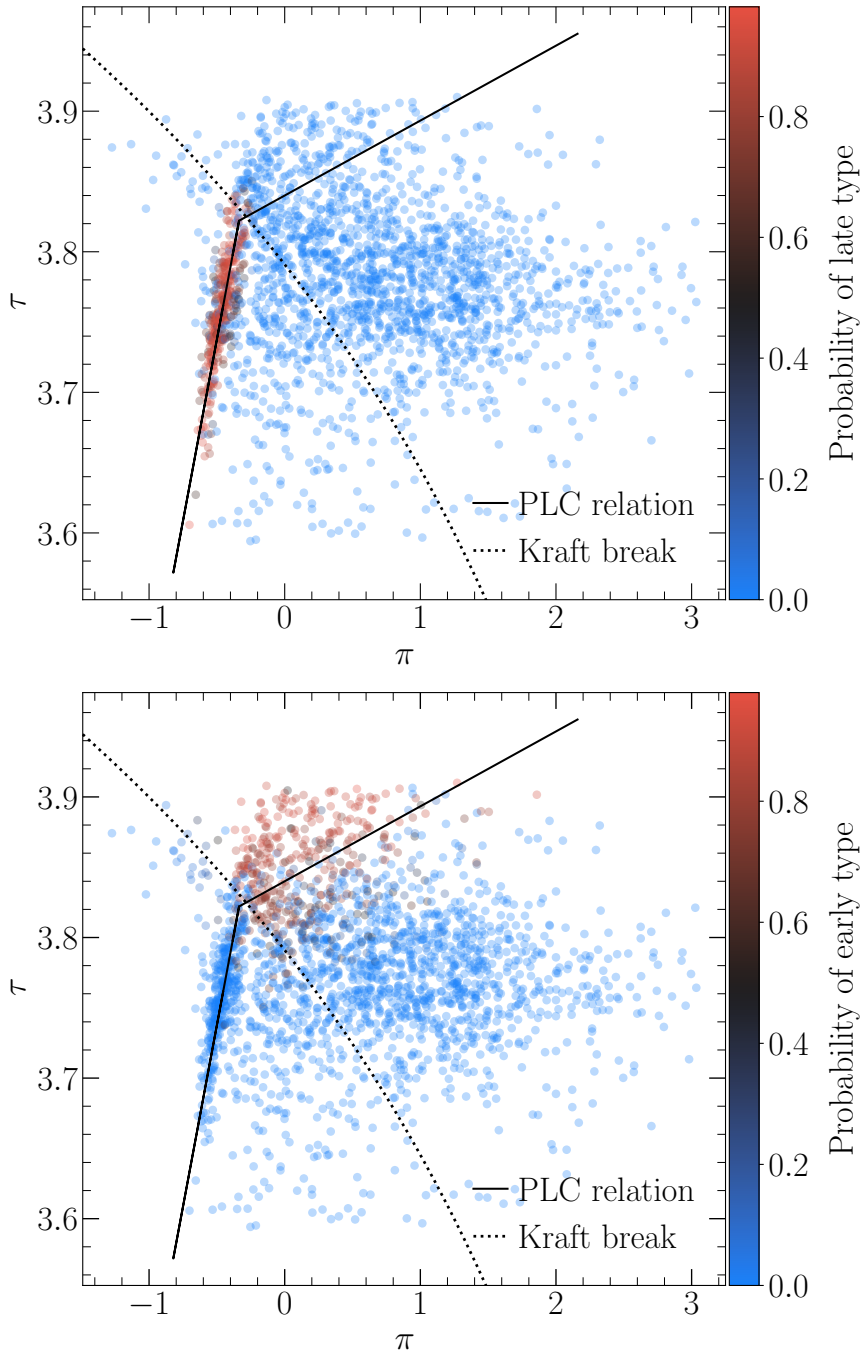

Figure 4.5 Projection of our sample to the $\pi\tau$-plane. The color of each object corresponds to the probability of being a late-type (top panel) or early-type (bottom panel) contact binary star. The solid lines show our best-fit PLC relations, and the dotted line shows the position of the Kraft break from Jayasinghe et al. (2020).

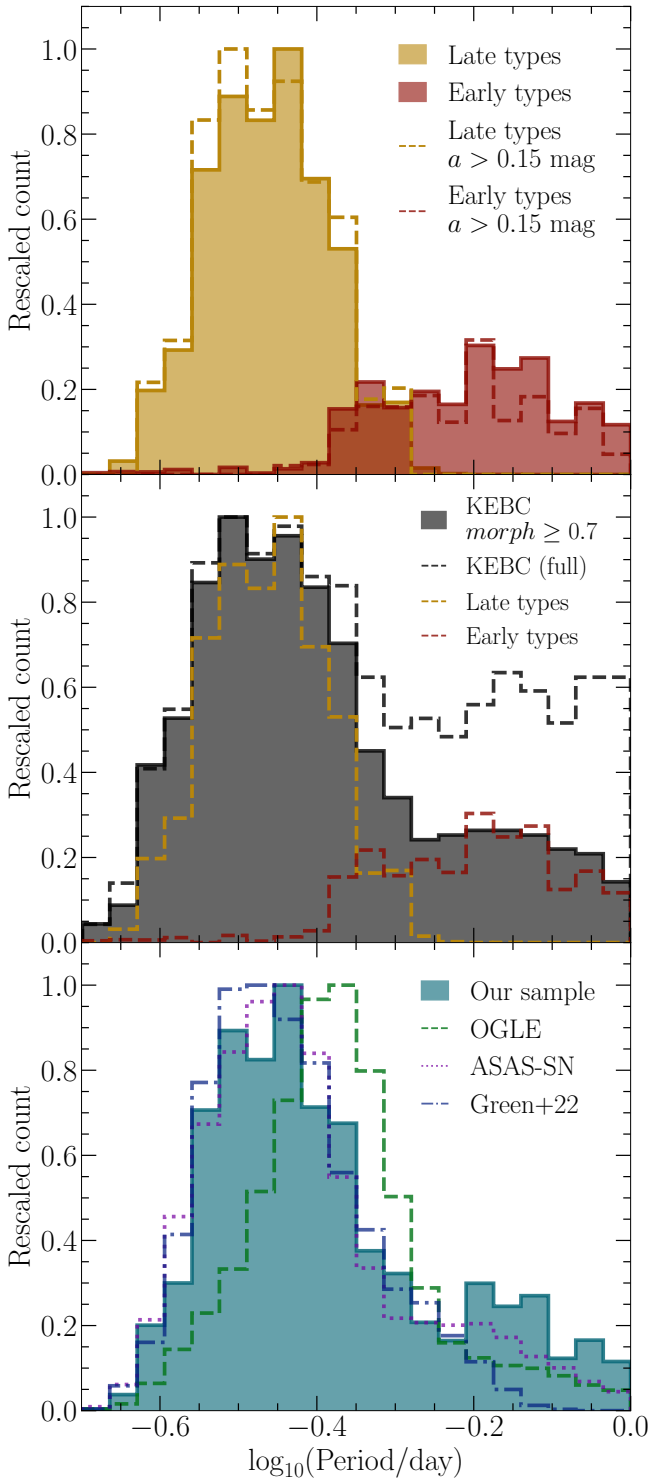

Figure 4.6 Period distribution of contact binaries in our sample. Top panel: Our full late-type and early-type samples (filled histogram) and their high-amplitude subsamples (dashed lines). Middle panel: Comparison of our samples with the KEBC, where the dashed black line marks the full KEBC, and the filled histogram shows the period distribution of objects with $morph \geq 0.7$. Bottom panel: Our sample in comparison to the samples from OGLE (Soszyński et al. 2016), ASAS-SN (Jayasinghe et al. 2018, 2019, 2020), and the recent catalog by Green et al. (2022). The histograms in all panels are appropriately rescaled and, where available, weighted by the probability of being a contact binary of the respective type.

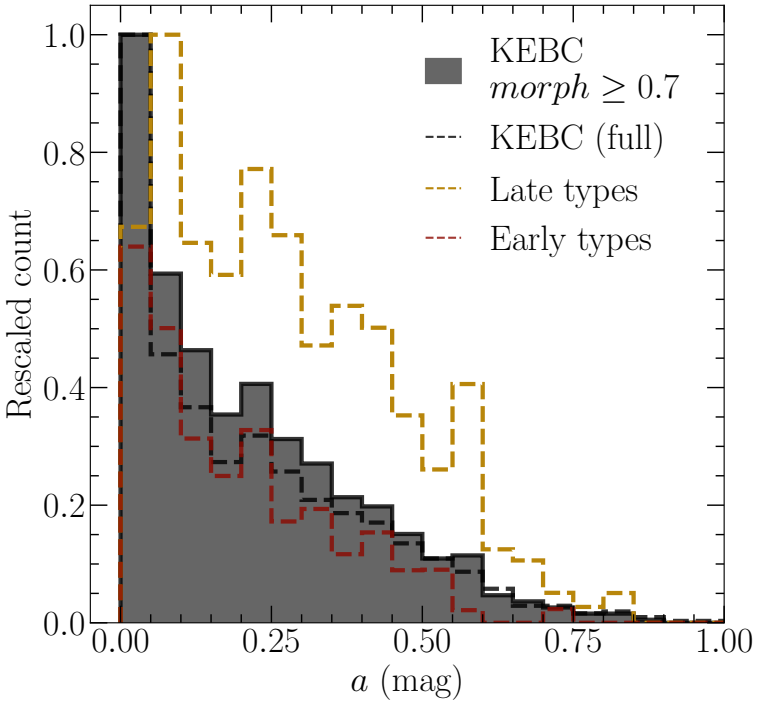

Figure 4.7 Amplitude distribution of the late-type (dashed orange line) and early-type (dashed red line) contact binaries in our cleaned sample. The dashed black line shows the amplitude distribution of the full KEBC, and the solid black line shows the distribution of the KEBC with *morph* ≥ 0.7. The late- and early-type histograms are weighted by the probability of being a contact binary of the respective type.

enters our model. This confirms our suspicion that the slope of the PLC relation changes at the Kraft break.

With these results, we can assess the quality of our sample and compare it to other existing samples of contact binaries. In Fig. 4.6 we show the period distribution of our clean sample of contact binaries. In the top panel, we compare the period distributions of our late- and early-type samples with their high-amplitude subsamples ($a > 0.15$ mag). The distributions of high-amplitude objects closely follow the distributions of the full samples. This implies that low-amplitude objects must also follow similar distributions, and therefore, they are consistent with being contact binaries. In the middle panel, we compare our samples with the unprocessed KEBC. The full KEBC substantially differs for periods longer than about 0.5 days, which is due to the presence of detached binaries. These can be efficiently removed by considering a cut on the *morph* parameter. However, the modified KEBC still differs from our sample, especially at short periods. In the bottom panel of Fig. 4.6, we compare our combined late- and early-type sample with other contact binary samples from the literature. The Galactic bulge contact binary sample of Soszyński et al. (2016) peaks at noticeably longer periods than what we find. Since Soszyński et al. (2016) also reported some short-period objects, it is not clear whether the shift is entirely due to the greater distance of the Galactic bulge compared to the *Kepler* field or if the population is truly different. The period shift is in contrast with the contact binary sample from ASAS-SN (Jayasinghe et al. 2018, 2019, 2020; Pawlak et al. 2019), which peaks at around the same periods as our sample, and agrees

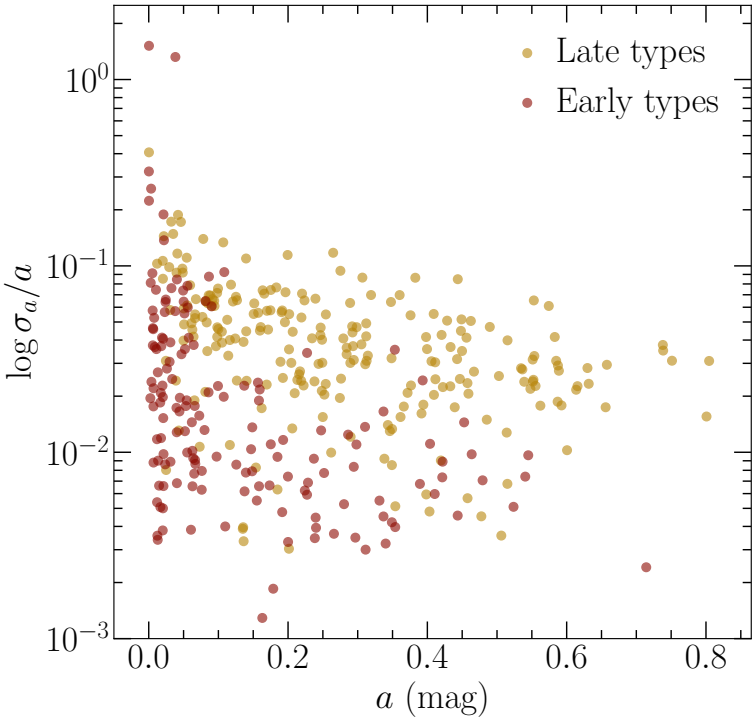

Figure 4.8 Relative amplitude uncertainty σ_a/a as a function of amplitude a for our late- and early-type samples. We show only objects with amplitude estimates in at least five *Kepler* quarters, *morph* ≥ 0.7, and a probability of belonging to either type greater than 0.5.

relatively well with our sample even for $P \gtrsim 0.6$ d, which are typical periods for early-type contact binaries. Recently, Green et al. (2022) published a sample of ellipsoidal and contact binaries, and their period distribution closely follows the distribution of our combined sample for $\log P \lesssim -0.2$. For longer periods, their sample appears to have fewer early-type objects than what we find. This is most likely due to a decreased efficiency of their selection algorithm for these systems.

In Fig. 4.7 we show the amplitude distributions of our two samples compared to the KEBC catalog. The figure shows that both the full KEBC and its *morph* ≥ 0.7 subsample have a high fraction of objects with very small $a < 0.05$ mag, while our samples show a smaller fraction at these amplitudes. We suggest that this is due to the contaminants in the KEBC that were removed by our model. We emphasize that correctly obtaining the low-amplitude part of the distribution is important to model the mass-ratio distribution of contact binary stars properly.

In Fig. 4.8 we show the relative amplitude uncertainty σ_a/a of our two samples. For the vast majority of objects, $\sigma_a/a < 0.1$. The relative uncertainty is higher only for several objects with very small amplitudes, which does not significantly affect our results. It is interesting to note that the late-type sample shows systematically higher amplitude uncertainties than the early-type sample. We can explain this observation by the appearance, disappearance, or migration of spots on the surfaces of late-type stars, which causes variations of the amplitude over time. We show an example of these amplitude variations in the middle panel of Fig. 4.3. Unless spots on contact binaries exhibit strong variability on timescales longer than the duration of the *Kepler* mission, our results in Fig. 4.8 imply that

Table 4.2 List of samples constructed from our Bayesian model for the identification of contact binary stars.

Sample	Type	Prob. cutoff	P (days)	Eff. size
CB1p50	Late	0.5	—	258.99
CB1p60	Late	0.6	—	256.27
CB1p70	Late	0.7	—	249.00
CB1p80	Late	0.8	—	228.04
CB2p10	Late	0.1	≤ 0.3	62.40
CB2p20	Late	0.2	≤ 0.3	61.96
CB2p30	Late	0.3	≤ 0.3	61.38
CB2p40	Late	0.4	≤ 0.3	61.07
CB2p50	Late	0.5	≤ 0.3	60.59
CB3p50	Late	0.5	> 0.3	198.41
CB3p60	Late	0.6	> 0.3	197.30
CB3p70	Late	0.7	> 0.3	192.69
CB3p80	Late	0.8	> 0.3	177.76
CB4p10	Early	0.1	< 1	106.42
CB4p20	Early	0.2	< 1	105.56
CB4p30	Early	0.3	< 1	105.56
CB4p40	Early	0.4	< 1	104.91
CB4p50	Early	0.5	< 1	104.42
CB5p10	Early	0.1	—	162.62
CB5p20	Early	0.2	—	161.55
CB5p30	Early	0.3	—	160.8
CB5p40	Early	0.4	—	159.49
CB5p50	Early	0.5	—	158.58

Notes. We require $morph \geq 0.7$ for all samples.

our method is not significantly affected by stellar spots.

4.5 Results

In this section, we present the results of our method for the inference of the mass-ratio distribution of contact binary stars. In Sect. 4.5.1 we define various populations of contact binaries and provide an overview of all models that we investigated. Next, we compare Bayes factors of the individual models and select a fiducial model for each population (Sect. 4.5.2). We present and discuss the mass-ratio distributions of contact binaries for the fiducial set of parameters in Sect. 4.5.3. Finally, we discuss the dependence of our results on the probability cutoffs distinguishing contact binaries from contaminants (Sect. 4.5.6), on our choice of the default fill-out factor (Sect. 4.5.4), on the splitting period for late-type binaries (Sect. 4.5.5), and on the hyperparameters of our model (Sect. 4.5.7).

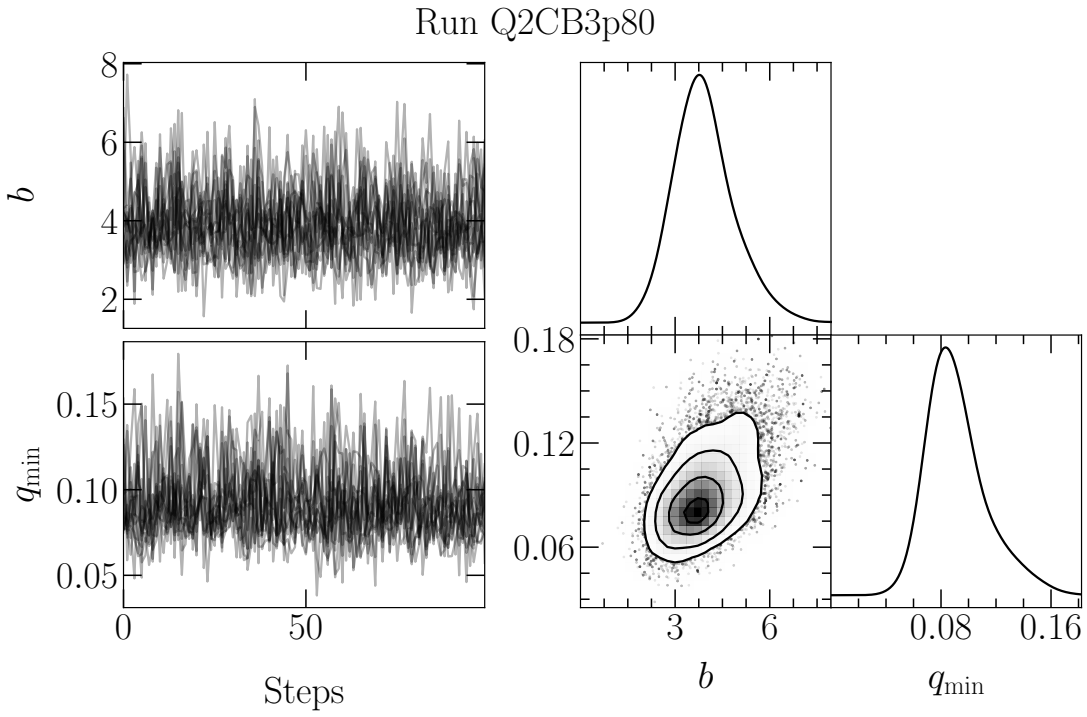

Figure 4.9 Chain plots and corner plots resulting from the run Q2CB3p80. We ran the sampler for a total of 2500 steps, but we discarded the first 500 steps as burn-in, and we thinned the chains by a factor of 20.

4.5.1 Populations of contact binary stars

We are interested in estimating the mass-ratio distribution for different populations of contact binary stars. For late-type binaries, we distinguish between the full population (CB1) and the populations of binaries with periods shorter (CB2) or longer (CB3) than $P_{\rm split} = 0.3\,$days (see Stępień & Gazeas 2012, for a justification of this period division). For early-type binaries, we distinguish between a population of early-type contact binaries with $P < 1\,$day (CB4) and an extended population containing all early-type contact binaries without a constraint on the period (CB5). These populations are separate, but overlap. By imposing various probability cutoffs on the individual populations, we obtain a number of samples with varying levels of contamination. In Table 4.2 we list the samples together with their definitions and their sizes. The effective sample size is calculated as $\sum_k p_{{\rm CB},k}$ for k belonging to the given sample and having $p_{{\rm CB},k}$ higher than the probability cutoff. To maximize the size of the samples while keeping the contamination as low as possible, we limited the maximum probability cutoffs for populations CB2, CB4, and CB5 to 0.5. In contrast, the high number of late-type contact binaries with periods around 0.37 days (Paczyński et al. 2006) allowed us to consider cutoffs up to 0.8 for populations CB1 and CB3.

For each population CB1–CB5, we also investigated the dependence on the mass-ratio prescription, fill-out factor, and various model hyperparameters. In Table 4.C.1 we give a complete list of our model runs. The runs labeled Q1 and Q2 used the power-law prescriptions Q_1 and Q_2 , respectively, as defined in Eq. (4.2). All Q1 and Q2 runs were carried out with the default values of $f = 0.25$, $h = 0.02$, and $n = 10000$. The runs labeled F investigated the dependence of the

results on the fill-out factor (Sect. 4.5.4), while the runs starting with S examined how the choice of $P_{\rm split}$ affects the mass-ratio distribution of the two late-type subpopulations (Sect. 4.5.5). Finally, the runs labeled H studied the dependence on the hyperparameters h and n (Sect. 4.5.7).

4.5.2 Fiducial models and Bayes factors

For each model, we determined the posterior distribution of $q_{\rm min}$ and b. In Fig. 4.9 we illustrate our results by showing the chain plots and posterior distributions for the run Q2CB3p80. The remaining Q1 and Q2 posteriors can be found in Figure 4.C.1. The number of steps is sufficient for the chains to converge and the parameters $q_{\rm min}$ and b show no significant correlation.

To select a fiducial model for each population, we need to determine which power-law prescription for the mass-ratio distribution fits the observed data better. We achieved this by calculating the posterior Bayes factors (Aitkin 1991) for each pair of Q1 and Q2 runs defined on the same sample. In other words, for each sample, we compared the goodness-of-fit of the two power laws by taking the ratio of the posterior average of the corresponding model likelihoods. We calibrated the Bayes factors according to the scale proposed by Aitkin (1991), which suggests that posterior Bayes factors of 20, 100, or 1000 constitute a strong, very strong, or an overwhelming weight of sample evidence in favor of the model with the higher value. We present the results of the comparison in Table 4.3.

In most cases, the second power-law prescription Q_2 is preferred over Q_1, or the comparison is inconclusive. The exception is population CB5, where Q_1 performs better than Q_2. However, the weight of sample evidence is reversed when the Bayes factor is evaluated on population CB4 (the evidence varies from very strong to strong, depending on the employed probability cutoff), suggesting that the preference of Q_1 is most likely due to an increased contamination of the CB5 samples in the long-period tail of the contact binary distribution. For populations CB1 and CB3, the Bayes factors are between 10 and 16, which gives substantial but not strong evidence in favor of Q_2. The analysis is inconclusive for population CB2, where the Bayes factors are very close to unity for all probability cutoffs. Overall, our results indicate a general preference for Q_2, which agrees with the results reported by Rucinski (2001). Based on these results, the fiducial models for our three populations are Q2CB2p50, Q2CB3p80, and Q2CB4p50.

Now we address the issue of whether our separate treatment of short-period contact binaries is supported by the data. In other words, we wish to quantify whether fitting the two populations separately and doubling the number of free parameters gives better results than fitting the entire population with a single model. For late-type binaries, we calculated the posterior Bayes factor comparing the combined Q2CB2p50+Q2CB3p50 model with model Q2CB1p50. This is justified because the two samples CB2p50 and CB3p50 are disjoint. The calculation yields a Bayes factor of about 71, providing strong evidence in favor of treating short-period and long-period late-type contact binaries separately. For early-type binaries, the limited size of the CB5 samples unfortunately prevents us from performing a similar analysis for periods shorter and longer than one day. Taking also the increased contamination of our CB5 population into account, which likely occurs due to the strongly decreasing frequency of contact binaries with period,

Table 4.3 Posterior Bayes factors for the two power-law prescriptions Q_1 and Q_2, different contact binary populations (CB1–CB5), and different probability cutoffs.

Models	Posterior Bayes factor
Q2CB1p50 vs. Q1CB1p50	15.02
Q2CB1p60 vs. Q1CB1p60	10.76
Q2CB1p70 vs. Q1CB1p70	14.87
Q2CB1p80 vs. Q1CB1p80	12.33
Q2CB2p10 vs. Q1CB2p10	1.36
Q2CB2p20 vs. Q1CB2p20	1.17
Q2CB2p30 vs. Q1CB2p30	1.04
Q2CB2p40 vs. Q1CB2p40	1.00
Q2CB2p50 vs. Q1CB2p50	1.03
Q2CB3p50 vs. Q1CB3p50	10.55
Q2CB3p60 vs. Q1CB3p60	11.44
Q2CB3p70 vs. Q1CB3p70	16.53
Q2CB3p80 vs. Q1CB3p80	15.67
Q2CB4p10 vs. Q1CB4p10	178.49
Q2CB4p20 vs. Q1CB4p20	49.31
Q2CB4p30 vs. Q1CB4p30	47.32
Q2CB4p40 vs. Q1CB4p40	32.58
Q2CB4p50 vs. Q1CB4p50	30.44
Q2CB5p10 vs. Q1CB5p10	0.00
Q2CB5p20 vs. Q1CB5p20	0.01
Q2CB5p30 vs. Q1CB5p30	0.02
Q2CB5p40 vs. Q1CB5p40	0.05
Q2CB5p50 vs. Q1CB5p50	0.08

we completely discarded CB5 from our analysis. From this point on, we only consider three distinct contact binary populations: CB2, CB3, and CB4.

4.5.3 Mass-ratio distribution of contact binary stars

We now present our main results. In Fig. 4.10 we show the inferred amplitude and mass-ratio distributions for the three distinct contact binary populations CB2, CB3, and CB4. The distributions were obtained by marginalizing over the posteriors of the model parameters. We show the posterior distributions of $q_{\min}$ and b for the three populations in Fig. 4.11. In Fig. 4.12 we compare the fiducial values of $q_{\min}$ and b with the values obtained from models with different choices for some parameters, specifically, the mass-ratio distribution prescription (Q_1 vs. Q_2), fill-out factor, and hyperparameters h and n. We show the full posterior distributions of all models in Appendix 4.C. The fiducial models give for the minimum mass ratio

$$q_{\min} = \begin{cases} 0.246^{+0.029}_{-0.046} & \text{CB2 (late-type binaries with } P \leq 0.3\,\text{d}), \\ 0.087^{+0.024}_{-0.015} & \text{CB3 (late-type binaries with } P > 0.3\,\text{d}), \\ 0.030^{+0.018}_{-0.022} & \text{CB4 (early-type binaries with } P < 1\,\text{d}), \end{cases} \tag{4.18}$$

and for the slope of the mass-ratio distribution

$$b = \begin{cases} 7.66^{+4.45}_{-3.15} & \text{CB2 (late-type binaries with } P \leq 0.3\,\text{d}), \\ 3.84^{+0.96}_{-0.80} & \text{CB3 (late-type binaries with } P > 0.3\,\text{d}), \\ 5.82^{+1.52}_{-1.30} & \text{CB4 (early-type binaries with } P < 1\,\text{d}). \end{cases} \tag{4.19}$$

Figs. 4.10 and 4.11 show that $q_{\min}$ varies noticeably between our populations. There is a clear trend that $q_{\min}$ decreases with increasing P. The same holds for the mean values of q calculated from the marginalized mass-ratio distributions, which go from $q_{\text{mean}} = 0.33^{+0.21}_{-0.19}$ for population CB2 to $q_{\text{mean}} = 0.25^{+0.06}_{-0.06}$ for population CB3 and $q_{\text{mean}} = 0.16^{+0.04}_{-0.04}$ for population CB4. Using the PLC relation, we can translate the trend in $q_{\min}$ into effective temperatures and luminosities: higher temperatures, luminosities, and larger radii imply lower values of $q_{\min}$. The shape of the CB4 fiducial posterior indicates that $q_{\min}$ for this populations is also consistent with being zero, but the limited size and the relatively high contamination of the CB4 fiducial sample prevent us from performing further tests of this hypothesis. We do not observe any clear trend in the values of the power-law exponent b. We discuss the astrophysical implications of our findings in Sect. 4.6.

4.5.4 Dependence on fill-out factor

Following Rucinski (2001), we carried out all our Q1 and Q2 runs with $f = 0.25$. To analyze the impact of f on the fiducial models, we performed the runs FCB2–FCB4, which considered five different values of f: 0.15, 0.25, 0.5, 0.75, and 0.99. Fig. 4.12 shows that the value of $q_{\min}$ strongly depends on f in populations CB2 and CB3, with larger f pushing $q_{\min}$ to lower values. Although the credible

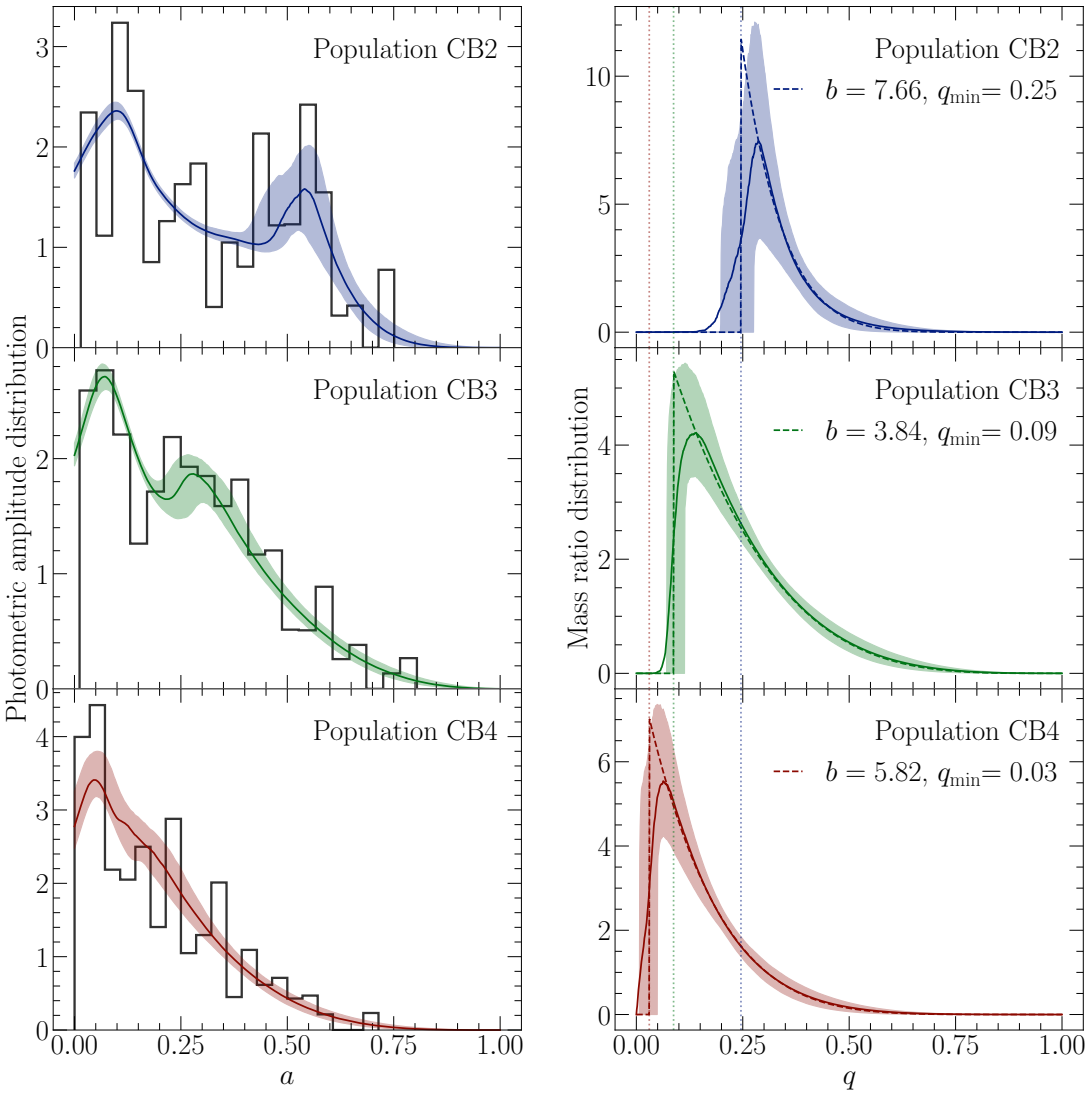

Figure 4.10 Amplitude (left panel) and mass-ratio (right panel) distributions for our three populations CB2, CB3, and CB4. The solid black lines in the left panel show weighted histograms of the observed data. The solid blue, green, and red lines in both panels are obtained by marginalizing out the functional form of the Q_2 power law, and the dashed lines show Q_2 evaluated for the median values of b and $q_{\rm min}$. The colored bands represent the 1σ credible intervals around the marginalized amplitude and mass-ratio distributions. The vertical dotted lines in the right panel compare the median values of $q_{\rm min}$ between the three populations.

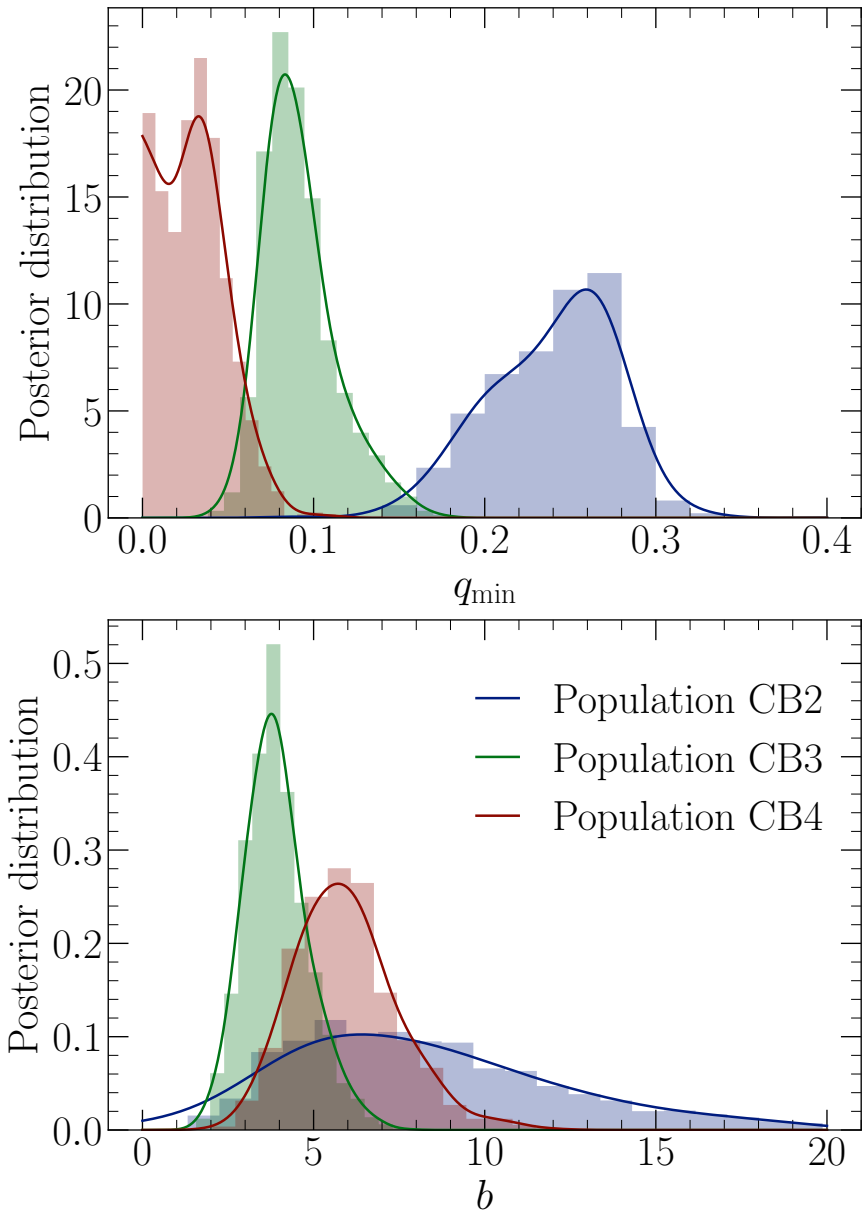

Figure 4.11 Comparison of the posteriors of $q_{\min}$ (top panel) and b (bottom panel) resulting from the fiducial models for populations CB2, CB3, and CB4.

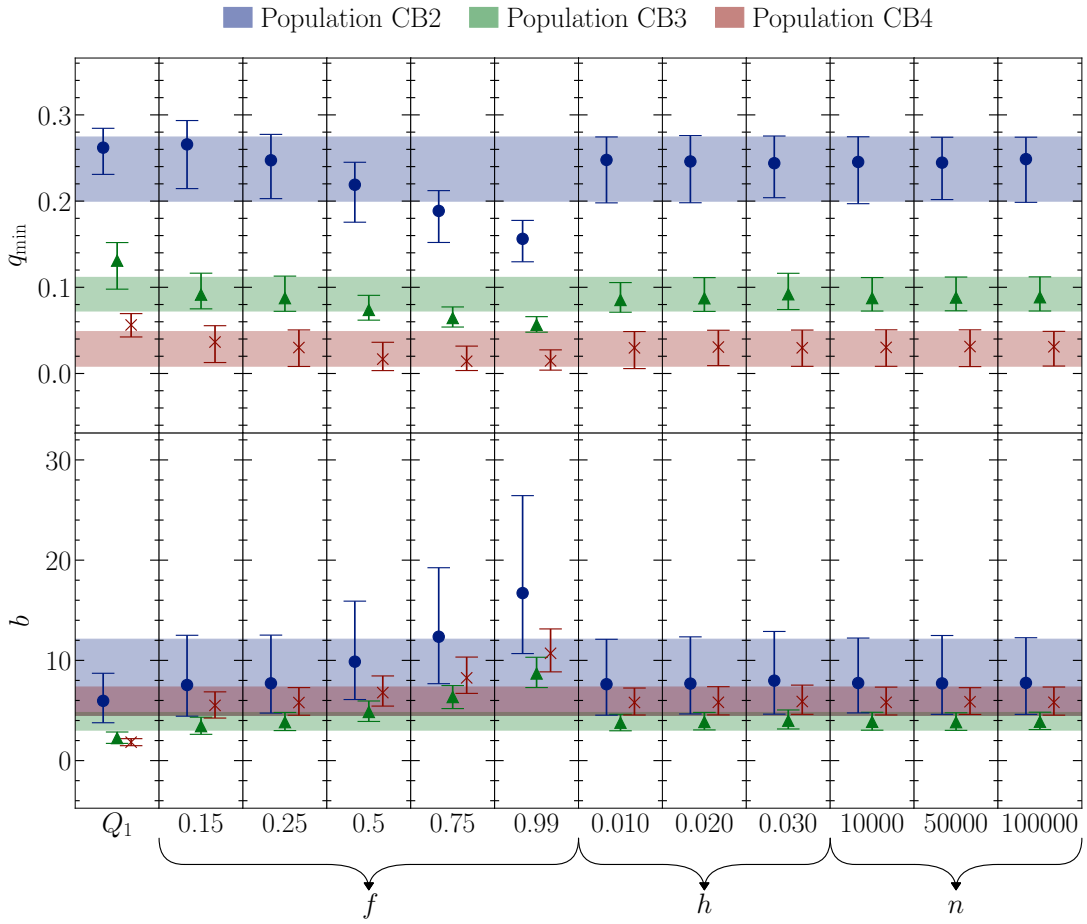

Figure 4.12 Dependence of $q_{\rm min}$ and b on the mass-ratio prescription Q_1, fill-out factor f, and hyperparameters of the model h and n. The colored bands represent the 1σ credible intervals resulting from the fiducial models for the three populations. We show the full posterior distributions in Figs. 4.C.1, 4.C.2, and 4.C.5.

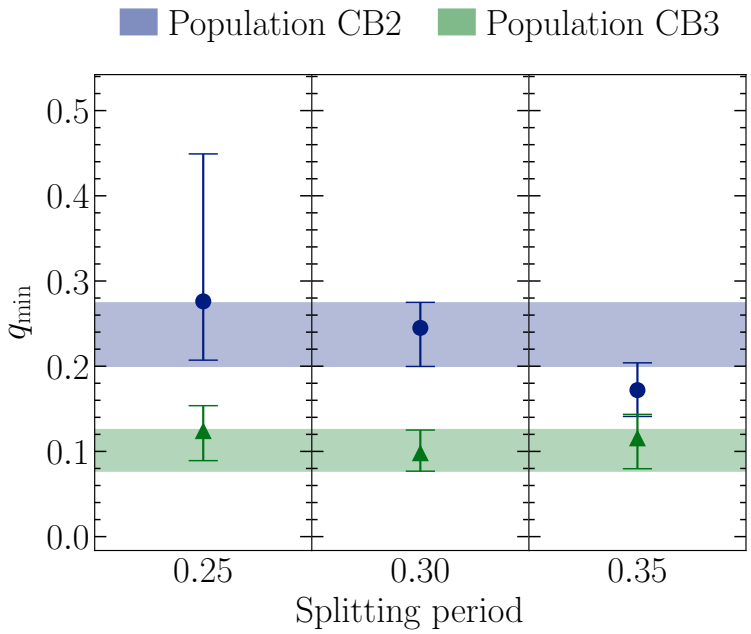

Figure 4.13 Dependence of $q_{\min}$ on the splitting period for populations CB2 and CB3. The colored bands represent the 1σ credible intervals of the fiducial models. For the purpose of this plot, we lowered the probability cutoff of the fiducial model for population CB3 from 0.8 to 0.5. The full posterior is plotted in Fig. 4.C.3.

intervals overlap for $f \leq 0.75$, the trend is clear. For population CB4, the values of $q_{\min}$ are consistent within the 1σ credible intervals across the whole range of f. Moreover, the power-law index b grows with f for all three populations.

In principle, we should be able to obtain the best-fitting value of f by evaluating the posterior Bayes factors of the models. In this specific case, all Bayes factors are below 5, rendering the analysis inconclusive. Consequently, we are not able to infer the optimal value of f from our data and we kept $f = 0.25$ based on previous detailed models and theoretical considerations (Sect. 4.2.1).

4.5.5 Dependence on splitting period

The choice to distinguish between the populations of late-type contact binaries with $P \leq 0.3\,\mathrm{d}$ and $P > 0.3\,\mathrm{d}$ is motivated by Stępień & Gazeas (2012), who argued that binaries with $P \lesssim 0.3\,\mathrm{d}$ do not live long enough to evolve to small q , but instead merge at moderate q due to the L2 overflow. Realistically, we expect a smooth transition between the two populations at around $P_{\mathrm{split}} \approx 0.3$ d. If this is the case, $q_{\min}$ of population CB2 should gradually shift to lower values with increasing P_{split} and increase or remain unchanged for $P_{\mathrm{split}} < 0.3\,\mathrm{d}$. Conversely, $q_{\min}$ of population CB3 should not significantly change when P_{split} is increased, but it should shift to higher values for $P_{\mathrm{split}} < 0.3\,\mathrm{d}$.

To investigate this hypothesis, we carried out runs SCB2 and SCB3, which examine how the fiducial results for populations CB2 and CB3 change when we shift P_{split} from 0.30 d to 0.25 d or 0.35 d. To increase the size of the CB3 sample, we performed the analysis with a probability cutoff of 0.5 instead of the fiducial value 0.8. We summarize the output from the runs in Fig. 4.13, where we compare

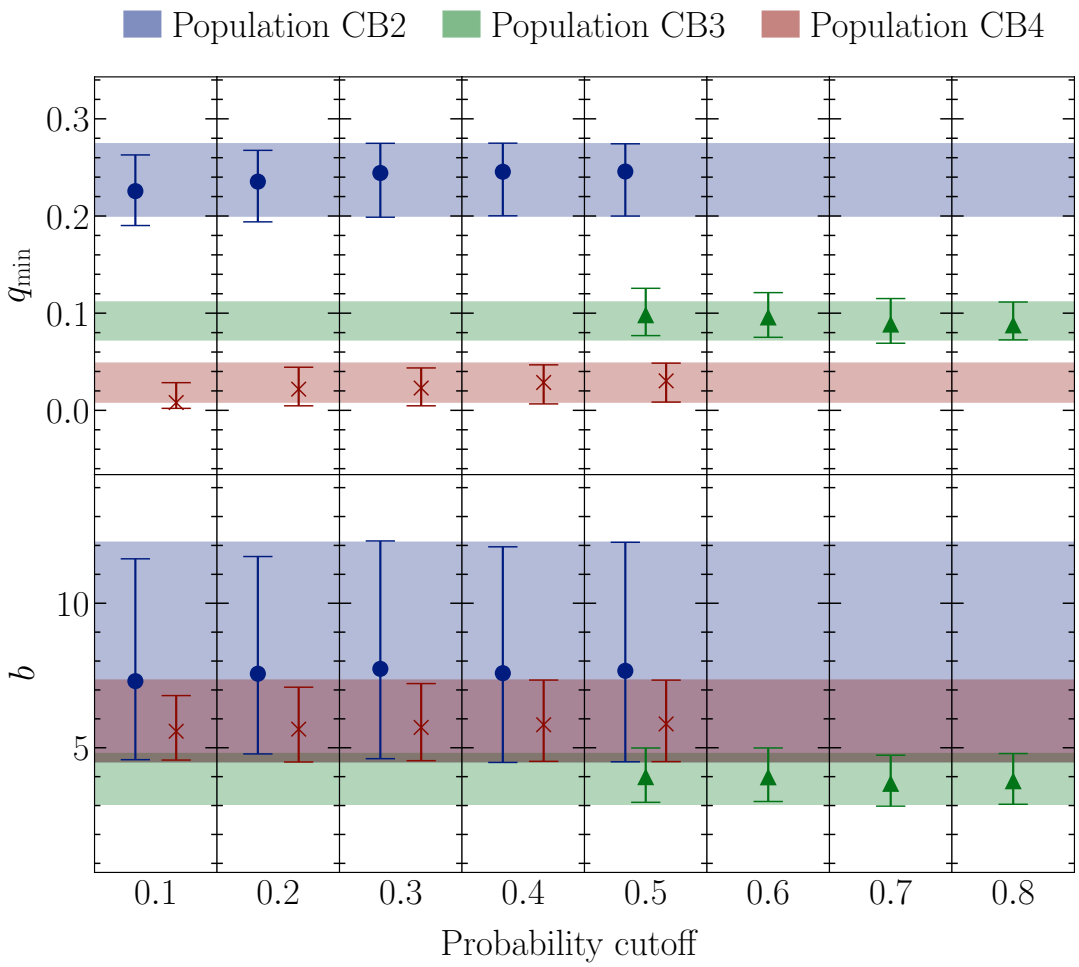

Figure 4.14 Dependence of $q_{\min}$ and b for populations CB2, CB3, and CB4 on different probability cutoffs separating contact binaries from contaminants. The colored bands represent the 1σ credible intervals of the fiducial models. We show the full posterior distributions in Fig. 4.C.4.

the resulting values of $q_{\min}$. The trends of $q_{\min}$ are consistent with our hypothesis overall, indicating that the observed difference between $q_{\min}$ of the two late-type populations is genuine and not just an artifact of the choice of P_{split}.

4.5.6 Dependence on probability cutoffs

In Fig. 4.14 we show how the fiducial values of $q_{\min}$ and b change when we gradually decrease the probability cutoff that separates contact binaries from contaminants. The values are consistent with each other, and we do not observe any jumps or discontinuous changes in the full posterior distributions (Fig. 4.C.4). In most cases, as the probability cutoff increases, the posteriors simply become more concentrated toward the central point, which is the desired behavior. Population CB4 is an exception to this rule, with its $q_{\min}$ posteriors peaking close to zero or exactly at zero, depending on the employed probability cutoff. The CB4 posteriors of $q_{\min}$ appear to be cut off from the left, which is typically seen when the parameter is actually zero, but the prior enforces that it takes non-negative values.

4.5.7 Dependence on hyperparameters

In defining the fiducial models, we had to make specific choices for the hyperparameters. To ensure that the obtained results are robust with respect to these choices, we carried out a number of runs with different values of the hyperparameters. In Fig. 4.12 we investigate the dependence of $q_{\min}$ and b on h and n. We tried three different values of h (runs H1CB2–CB4 with $h = 0.01$, 0.02, and 0.03) and three different values of n (runs H2CB2–CB4 with $n = 10000$, 50000, and $n = 100000$). The resulting $q_{\min}$ and b are consistent with each other and almost perfectly overlap within their 1σ credible intervals. Detailed investigation of the posterior distributions in Fig. 4.C.5 shows that the overall shapes and positions remain very similar. Our results indicate that $n = 10000$ is already enough for the runs to converge to the correct solution, which justifies our fiducial choice of this value.

4.6 Discussions and conclusions

We have extended and reformulated the method for the estimation of the mass-ratio distribution of contact binary stars developed by Rucinski (2001), which exploits the simplicity of contact binary light curves. Setting the fill-out factor to $f = 0.25$ and assuming that binary orbits are randomly oriented in space, we obtained a semi-parametric mapping between the mass-ratio distribution and the photometric amplitude distribution (Sect. 4.2). We approximated the mass-ratio distribution as a power law with a slope b and a sharp cutoff at $q_{\min}$, and using Bayesian inference, we obtained the posterior distributions of these parameters. This is possible because the position of the local maximum in the amplitude distribution is strongly correlated with the value of $q_{\min}$ (Fig. 4.1). For the method to work, a sufficiently large sample of contact binaries is required that is complete for amplitudes $a \gtrsim 0.1$ mag or less. Such data sets have only recently become available from *Kepler* and other space-based telescopes. The advantage of the method is that it infers $q_{\min}$ and b purely from photometry, while other methods typically require much more costly spectroscopic observations or exhaustive modeling of stars one by one.

We constructed our sample from the Kepler Eclipsing Binary Catalog (Prša et al. 2011; Abdul-Masih et al. 2016; Kirk et al. 2016), which we combined with luminosities from *Gaia* DR2 (Gaia Collaboration et al. 2018) and effective temperatures from Bai et al. (2019) (Sect. 4.3). To filter out detached and semidetached binaries as well as other types of contaminating variable stars, we made use of the PLC relation (Sect. 4.4). We distinguished between late- and early-type contact binaries, and we assumed that both types follow their own PLC relations, with a continuous transition between the two relations. We further assumed that the individual contact binaries are scattered around their respective PLC relations, and we modeled the contaminating noise as Gaussian (Sect. 4.4). Employing Bayesian inference, we assigned a probability of being a contact binary of either late or early type to each object in the sample (Fig. 4.5). Late-type contact binaries have systematically larger amplitude scatter than early-type objects (Fig. 4.8), which is most likely due to the presence of time-varying stellar spots in the atmospheres of late-type stars. Seeing that the relative

amplitude uncertainty remains below 10% for most objects, we conclude that this phenomenon does not significantly affect our method for the inference of the mass-ratio distribution.

Using different period cutoffs, we constructed five separate but overlapping populations of contact binary stars (Sect. 4.5.1): all late-type contact binaries (CB1), late-type contact binaries with $P \leq 0.3\,\mathrm{d}$ (CB2), late-type contact binaries with $P > 0.3\,\mathrm{d}$ (CB3), early-type contact binaries with $P < 1\,\mathrm{d}$ (CB4), and all early-type contact binaries with no constraint on the period (CB5). For each population, we defined multiple samples by imposing different cutoffs on the probability of being a contact binary of either type (Table 4.2). We fit each sample with two different power-law prescriptions for the mass-ratio distribution, as defined in Eq. (4.2), and for each sample, we calculated the posterior Bayes factor comparing the goodness-of-fit of the two prescriptions (Sect.4.5.2). In most cases, the second prescription $Q_2 \propto (1-q)^{\mathrm{b}}$ yields a better fit than the first prescription $Q_1 \propto q^{-b}$ (Tab. 4.3), but the evidence is not strong, with most Bayes factors at or below 20. A notable exception is the CB5 population, where Q_1 outperformed Q_2. This result is most likely affected by the increased contamination of the CB5 samples. The population of late-type contact binaries with $P \leq 0.3\,\mathrm{d}$ (CB2) does not favor either prescription. Only the CB4 population gives conclusive evidence in support of Q_2 against Q_1. In conclusion, we observe a marginally strong evidence in support of Q_2, which agrees with the previous results of Rucinski (2001).

Adopting Q_2 as the mass-ratio distribution of contact binary stars, we justified the separate treatment of populations CB2 and CB3 by calculating the posterior Bayes factor of the combined CB2+CB3 model and the model for CB1. We found very strong evidence in support of the combined model (Sect. 4.5.2 and 4.5.5). Consequently, we investigated the two populations CB2 and CB3 separately, and we discarded the combined population CB1. We also discarded CB5 due to the increased contamination of its samples. In summary, we were left with three distinct contact binary populations CB2, CB3, and CB4.

Our results for q_{min} and b are summarized in Sect. 4.5.3. We find that q_{min} decreases with increasing orbital period. For late-type binaries with $P \leq 0.3\,\mathrm{d}$, we find a relatively high $q_{\mathrm{min}} = 0.246^{+0.029}_{-0.046}$. For normal late-type binaries, we find $q_{\mathrm{min}} = 0.087^{+0.024}_{-0.015}$. For early-type binaries with $P < 1\,\mathrm{d}$, we find $q_{\mathrm{min}} = 0.030^{+0.018}_{-0.022}$. Our results are compatible with theoretical predictions of q_{min}. Specifically, our q_{min} for late-type binaries with $P > 0.3\,\mathrm{d}$ agrees with theoretical values for solar-type stars, where Rasio (1995) predicted $q_{\mathrm{min}} = 0.08$ for an $n = 3$ polytrope. It is also known that q_{min} scales with the stellar gyration radius, which is relatively small for early-type stars (Rasio 1995; Wadhwa et al. 2021; Blagorodnova et al. 2021). This agrees with our very small q_{min} for this population. However, given the credible interval of our result, we cannot definitely claim detection of the signature of the Darwin instability in the early-type population.

The trend of decreasing q_{min} with increasing orbital period agrees with the conclusions of Stępień & Gazeas (2012), who argued that this is due to the different timescales of mass transfer and angular momentum loss in low-mass contact binaries and more massive systems. The relatively moderate mass transfer in low-mass (short-period) contact binaries is insufficient to make the binary unstable to the Darwin instability, but instead, it leads to the overflow of the outer Roche lobe, resulting in the loss of mass angular momentum through the L2 point

and culminating with merger at comparatively larger q than in the case of the Darwin instability. In contrast, Kobulnicky et al. (2022) argued for an opposite trend, where $q_{\min}$ increases with period for $P \gtrsim 0.8\,\mathrm{d}$. Their model assumed that new contact binary systems form with $q \approx 1$ and conservatively evolve toward longer periods and smaller q until the onset of the Darwin instability.

We find different values of the power-law index b for different populations, but unlike for $q_{\min}$, we do not observe a clear trend with the orbital period. For Q_2, Rucinski (2001) reported $b = 6 \pm 2$, which is consistent with our results for all three populations (CB2: $b = 7.66^{+4.45}_{-3.15}$, CB3: $b = 3.84^{+0.96}_{-0.80}$, and CB4: $b = 5.82^{+1.52}_{-1.30}$). We note that Rucinski (2001) did not distinguish between late- and early-type contact binaries and that their sample is complete only for $a \gtrsim 0.3\,\mathrm{mag}$. Our relative uncertainties in b for populations CB3 and CB4 are only mildly smaller than those reported by Rucinski (2001), which is understandable given the similar sample sizes. Larger samples of contact binaries are required to better constrain b. This could be quite rewarding because b encodes physical processes such as nuclear evolution, magnetic braking, and thermal relaxation oscillations (Vilhu 1981). Rucinski (2001) indeed suggested that b is related to the thermal timescale of the secondary star and thus to the exponent of its mass–luminosity relation.

Our results show that $q_{\min}$ noticeably depends on the value of the fill-out factor f (Sect. 4.5.4 and Fig. 4.12), but our analysis of the posterior Bayes factors was inconclusive due to the insufficient evidence in favor of any specific f (all factors were below 5). Consequently, we were not able to constrain f from our data. Nonetheless, thermal relaxation oscillations theory suggests that f should be small and similar to our default value $f = 0.25$ (Lucy 1973; Rucinski 1973, 1997; Paczyński et al. 2006). Still, there is some evidence that f is different for late- and early-type binaries (Mochnacki 1981). In addition to the fill-out factor, we also verified that our estimates of b and $q_{\min}$ are fairly robust with respect to the splitting period P_{split} between populations CB2 and CB3 (Sect. 4.5.5 and Fig. 4.13), the probability cutoff (Sect. 4.5.6 and Fig. 4.14), and the KDE bandwidth h and number of Gaussians n involved in the construction of the amplitude distribution (Sect. 4.5.7 and Fig. 4.12).

The method presented here can easily be extended to the large samples of contact binaries expected from TESS and other space-based telescopes. In addition to giving better estimates for the parameters of the current model, these samples will enable characterization of more complex models that better capture the underlying mass-ratio distribution of contact binaries. One way to improve the current model is to include the splitting period as a parameter in the generative distribution constructed in Sect. 4.4.2. With this modification, we could fit the mass-ratio distributions of populations CB2 and CB3 simultaneously, and by marginalizing out the exact location of the split, we would obtain P_{split}-free estimates of $q_{\min}$ for the two populations.

A straightforward improvement of our approach would come from using more precise values for effective temperatures and luminosities. The recently released *Gaia* DR3 (Gaia Collaboration et al. 2022) provides a significant improvement over DR2, but unfortunately, the physical parameters continue to be based on single-star models (Creevey et al. 2022). We showed here that this assumption does not significantly affect our results, but improvements in this area could provide better distinction of contact binaries from various contaminants.

Another exciting possibility comes from combining space-borne all-sky photometry from TESS or *Gaia* with data from massive spectroscopic surveys such as SDSS-V (Kollmeier et al. 2019), WEAVE (Dalton et al. 2012), 4MOST (de Jong et al. 2019), LAMOST (Zhao et al. 2012), or *Gaia* RVS (Katz et al. 2022). These spectroscopic surveys often secure several spectra of each object. Although obtaining complete orbital and physical solution is still hard with these data alone (e.g., Price-Whelan et al. 2018), even a constraint with a low signal-to-noise ratio of the radial velocity amplitude or the flux ratio of the two components might greatly increase the statistical power of our model by excluding ranges of possible inclinations for each binary. Operationally, we would simultaneously fit the model to the observed amplitude and mass-ratio distributions, effectively yielding a nonuniform prior on the parameters of the power law. Ultimately, the scalability and flexibility of our method make it a powerful tool for the inference of the mass-ratio distribution and the minimum mass ratio of contact binary stars.

We thank Matthew Green for sharing their sample of contact binaries and our referee, Panagiota-Eleftheria Christopoulou, for her helpful comments. This work has been supported by INTER-EXCELLENCE grant LTAUSA18093 from the Ministry of Education, Youth, and Sports. The research of OP has been supported also by Horizon 2020 ERC Starting Grant 'Cat-In-hAT' (grant agreement no. 803158). This research made use of the cross-match service provided by CDS, Strasbourg.

4.A Evaluation of likelihood

When the amplitude distribution in Sect. 4.2 was evaluated, we were only able to construct $\hat{A}(a;\Theta)$, which is an analytical approximation to $A(a;\Theta)$. The approximation involves performing KDE on a finite number of randomly drawn amplitudes, which introduces a stochastic element into the process, transforming $\hat{A}(a;\Theta)$ into a random variable. This means that the likelihood in Eq. (4.3) does not yield a unique value for a given Θ and $\{a_k\}_{k=1}^N$, but is actually a random variable itself.

As illustrated in Fig. 4.A.1, the noise in $\hat{A}(a;\Theta)$ can be significantly reduced by employing a sufficiently large number of samples, but for the stochastic behavior to completely disappear, we would have to use the same input for KDE in each evaluation of $\hat{A}(a,\Theta)$. Unfortunately, none of these options are feasible; increasing the number of drawn amplitudes comes at huge computational cost due to the repeated log-likelihood evaluation during an MCMC run, and fixing the KDE input requires an analysis of which amplitude sample leads to the most accurate representation of $A(a)$, which cannot be achieved in any practical way.

Instead of trying to minimize the stochastic effect, we fully embraced the nondeterministic nature of $\hat{A}(a;\Theta)$ and modeled it as a sampling noise in $A(a;\Theta)$. We note that the scatter in $A(a;\Theta)$ is different from the scatter in the PLC relation that we investigated in Sect. 4.4.1. The scatter in $A(a;\Theta)$ smears the distribution itself, while the scatter in the PLC relation affects an originally exact relation and transforms it into a distribution. In principle, the extent to which $A(a;\Theta)$ is smeared depends on the value of Θ, which further adds to the complexity of the problem. The smearing can be equivalently viewed as an implicit dependence of $\hat{A}(a;\Theta)$ on an additional $2n$ parameters corresponding to the (i,q) positions of the n samples entering the KDE algorithm. Denoting the individual parameters by Y_l, with l going from 1 to $2n$, we can write the likelihood as

$$\hat{\mathscr{L}}(\Theta,\{Y_l\}_{l=1}^{2n}|\{a_k\}_{k=1}^N)=\prod_{k=1}^N p_{\mathrm{CB},k}\int \hat{A}(a_k;\Theta,\{Y_l\}_{l=1}^{2n})\mathscr{N}(a;a_k,\sigma_{a_k})\mathrm{d}a. \quad (4.\mathrm{A}.1)$$

By including these parameters, we remove the stochasticity and the likelihood becomes deterministic again. The additional parameters are distributed according to the joint distribution of i and q, which is given by $I(i)\times Q(q;\Theta)$ and serves as the conditional prior for these parameters. Since we are only interested in the posterior of Θ, we did not actively sample the additional parameters and their prior did not directly enter the Bayes theorem. Instead, in each step of the MCMC run, we updated Θ according to the chosen step-proposal strategy (e.g., stretch move or differential evolution) and the additional $2n$ parameters are simply drawn from the prior. This is equivalent to sampling the full posterior,

$$p(\Theta,\{Y_l\}_{l=1}^{2n}|\{a_k\}_{k=1}^N)=\frac{\mathscr{L}(\Theta,\{Y_l\}_{l=1}^{2n}|\{a_k\}_{k=1}^N)p(\{Y_l\}_{l=1}^{2n}|\Theta)p(\Theta)}{p(\{a_k\}_{k=1}^N)}, \quad (4.\mathrm{A}.2)$$

and marginalizing out the additional parameters, yielding the marginalized posterior probability distribution of Θ, or $p(\Theta|\{a_k\}_{k=1}^N)$.

This approach does not yield the posterior for the additional parameters, which is needed for the marginalization of $\hat{A}(a;\Theta,\{Y_l\}_{l=1}^{2n})$ or the calculation of the Bayes factors. To reconstruct the full posterior, we substituted the posterior

of the additional parameters with the prior. This is justifiable because in the limit of $N \to \infty$, the stochastic amplitude distribution $\hat{A}(a; \Theta, \{Y_l\}_{l=1}^{2n})$ converges to $A(a; \Theta)$, causing the posterior of the additional parameters to converge to the prescribed prior.

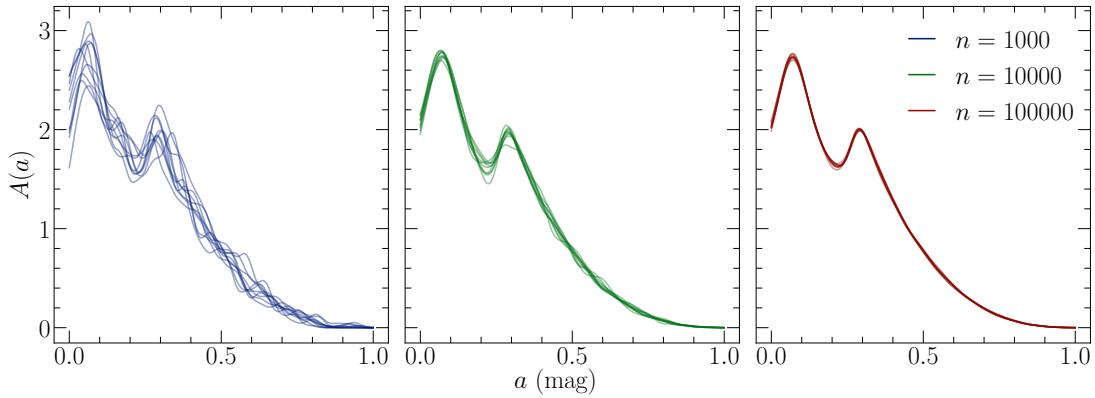

Figure 4.A.1 Comparison of synthetic contact binary amplitude distributions resulting from the kernel density estimation performed on samples of different sizes. The distribution is rather noisy for $n = 1000$, but the noise is already significantly reduced for $n = 10000$. When we increase the number of samples to 100000, the distribution effectively converges to the correct shape.

4.B Additional tables and figures for the identification of sample contamination

Here, we present a list of the parameters of our Bayesian model for removing contaminants in the sample of contact binaries (Tab. 4.B.1), plots of the MCMC chains resulting from the model (Fig. 4.B.1), and visualization of the posterior distribution of the model parameters (Fig. 4.B.2).

Table 4.B.1 List of the parameters of our Bayesian model for the identification of contact binary stars.

Parameter	Definition	Value
	Global parameters	
$\lambda_{\rm K}$	λ-location of the Kraft break along the PLC relation	$0.7631^{+0.0021}_{-0.0017}$
α_{X1}	$X = \alpha_{X1} + \beta_{X1}\lambda, \quad \lambda \leq \lambda_{\rm K}$	$0.1727^{+0.0086}_{-0.0086}$
β_{X1}		$0.0591^{+0.0219}_{-0.0225}$
α_{X2}	$X = \alpha_{X2} + \beta_{X2}\lambda, \quad \lambda > \lambda_{\rm K}$	$-0.0538^{+0.0537}_{-0.0520}$
β_{X2}		$0.4750^{+0.0262}_{-0.0283}$
	PLC parameters	
$\alpha_{\pi1}$	$\mu_{{\rm S}\pi} = \alpha_{\pi1} + \beta_{\pi1}\lambda, \quad \lambda \leq \lambda_{\rm K}$	$-0.5077^{+0.0015}_{-0.0015}$
$\beta_{\pi1}$		$0.2243^{+0.0043}_{-0.0044}$
$\beta_{\pi2}$	$\mu_{{\rm S}\pi} = \alpha_{\pi1} + (\beta_{\pi1} - \beta_{\pi2})\lambda_{\rm K} + \beta_{\pi2}\lambda, \quad \lambda > \lambda_{\rm K}$	$1.2614^{+0.0490}_{-0.0484}$
$\alpha_{\tau1}$	$\mu_{{\rm S}\tau} = \alpha_{\tau1} + \beta_{\tau1}\lambda, \quad \lambda \leq \lambda_{\rm K}$	$3.7337^{+0.0010}_{-0.0010}$
$\beta_{\tau1}$		$0.1159^{+0.0023}_{-0.0023}$
$\beta_{\tau2}$	$\mu_{{\rm S}\pi} = \alpha_{\tau1} + (\beta_{\tau1} - \beta_{\tau2})\lambda_{\rm K} + \beta_{\tau2}\lambda, \quad \lambda > \lambda_{\rm K}$	$0.0672^{+0.0033}_{-0.0031}$
$\alpha_{\sigma_{\pi1}}$	$\sigma_{{\rm S}\pi} = \alpha_{\sigma_{\pi1}} + \beta_{\sigma_{\pi1}}\lambda, \quad \lambda \leq \lambda_{\rm K}$	$0.0277^{+0.0013}_{-0.0013}$
$\beta_{\sigma_{\pi1}}$		$0.0206^{+0.0033}_{-0.0033}$
$\alpha_{\sigma_{\pi2}}$	$\sigma_{{\rm S}\pi} = \alpha_{\sigma_{\pi2}} + \beta_{\sigma_{\pi2}}\lambda, \quad \lambda > \lambda_{\rm K}$	$-0.1754^{+0.0986}_{-0.1081}$
$\beta_{\sigma_{\pi2}}$		$0.3961^{+0.0891}_{-0.0820}$
$\alpha_{\sigma_{\tau1}}$	$\sigma_{{\rm S}\tau} = \alpha_{\sigma_{\tau1}} + \beta_{\sigma_{\tau1}}\lambda, \quad \lambda \leq \lambda_{\rm K}$	$0.0159^{+0.0007}_{-0.0007}$
$\beta_{\sigma_{\tau1}}$		$-0.0032^{+0.0016}_{-0.0017}$
$\alpha_{\sigma_{\tau2}}$	$\sigma_{{\rm S}\tau} = \alpha_{\sigma_{\tau2}} + \beta_{\sigma_{\tau2}}\lambda, \quad \lambda > \lambda_{\rm K}$	$0.0470^{+0.0045}_{-0.0048}$
$\beta_{\sigma_{\tau2}}$		$-0.0153^{+0.0039}_{-0.0035}$
	Background noise parameters	
m_1	$\mu_{{\rm B}\tau} = m_1 + l_1\lambda, \quad \lambda \leq \lambda_{\rm K}$	$3.7370^{+0.0006}_{-0.0006}$
l_1		$0.0855^{+0.0013}_{-0.0013}$
m_2	$\mu_{{\rm B}\tau} = m_2 + l_2\lambda, \quad \lambda > \lambda_{\rm K}$	$3.7496^{+0.0077}_{-0.0076}$
l_2		$0.0738^{+0.0077}_{-0.0078}$
w_1	$\sigma_{{\rm B}\tau} = w_1, \quad \lambda \leq \lambda_{\rm K}$	$0.0268^{+0.0004}_{-0.0004}$
w_2	$\sigma_{{\rm B}\tau} = w_2, \quad \lambda > \lambda_{\rm K}$	$0.0307^{+0.0010}_{-0.0010}$

Notes. We assumed that contact binaries are scattered around the PLC relation, parametrically expressed as $(\lambda, \mu_{{\rm S}\pi}, \mu_{{\rm S}\tau})$ in the log-luminosity λ vs. log-period π vs. log-effective temperature τ space. The $\sigma_{\rm S}$ parameters control the level of scatter around the relation. We modeled the background noise as though it were generated from a thick plane $(\lambda, \pi, \mu_{\rm B})$ with its thickness controlled by the $\sigma_{\rm B}$ parameters. We present the values of the parameters with their 1σ credible intervals.

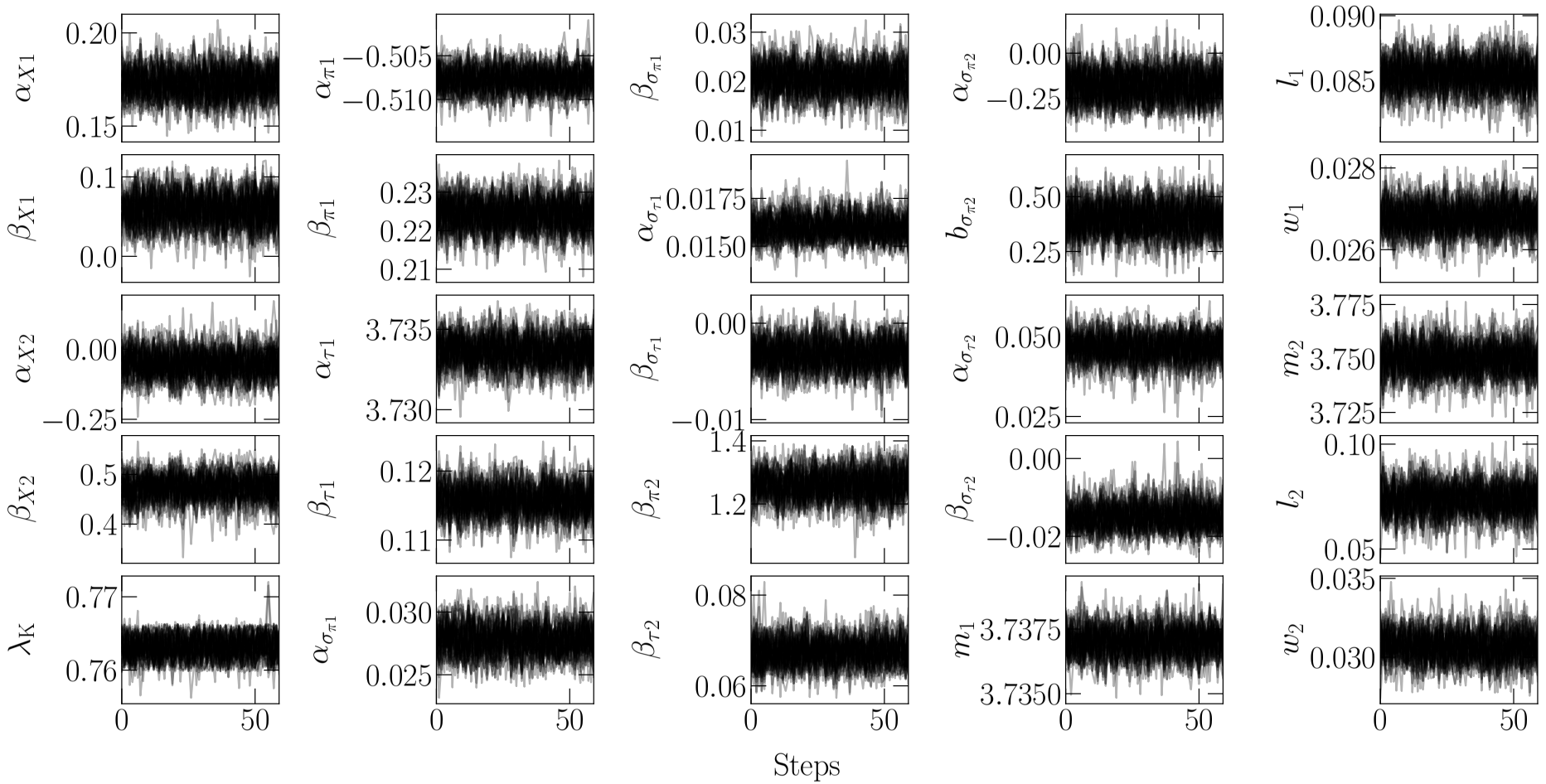

Figure 4.B.1 Chain plots resulting from the *emcee* run of our Bayesian model for the identification of contact binary stars. We ran the sampler for a total of 160 000 steps, but we discarded the first 10000 as burn-in, and we thinned the chains by a factor of 300. Visual inspection of the plot confirms that the number of steps was sufficient for the chains to converge.

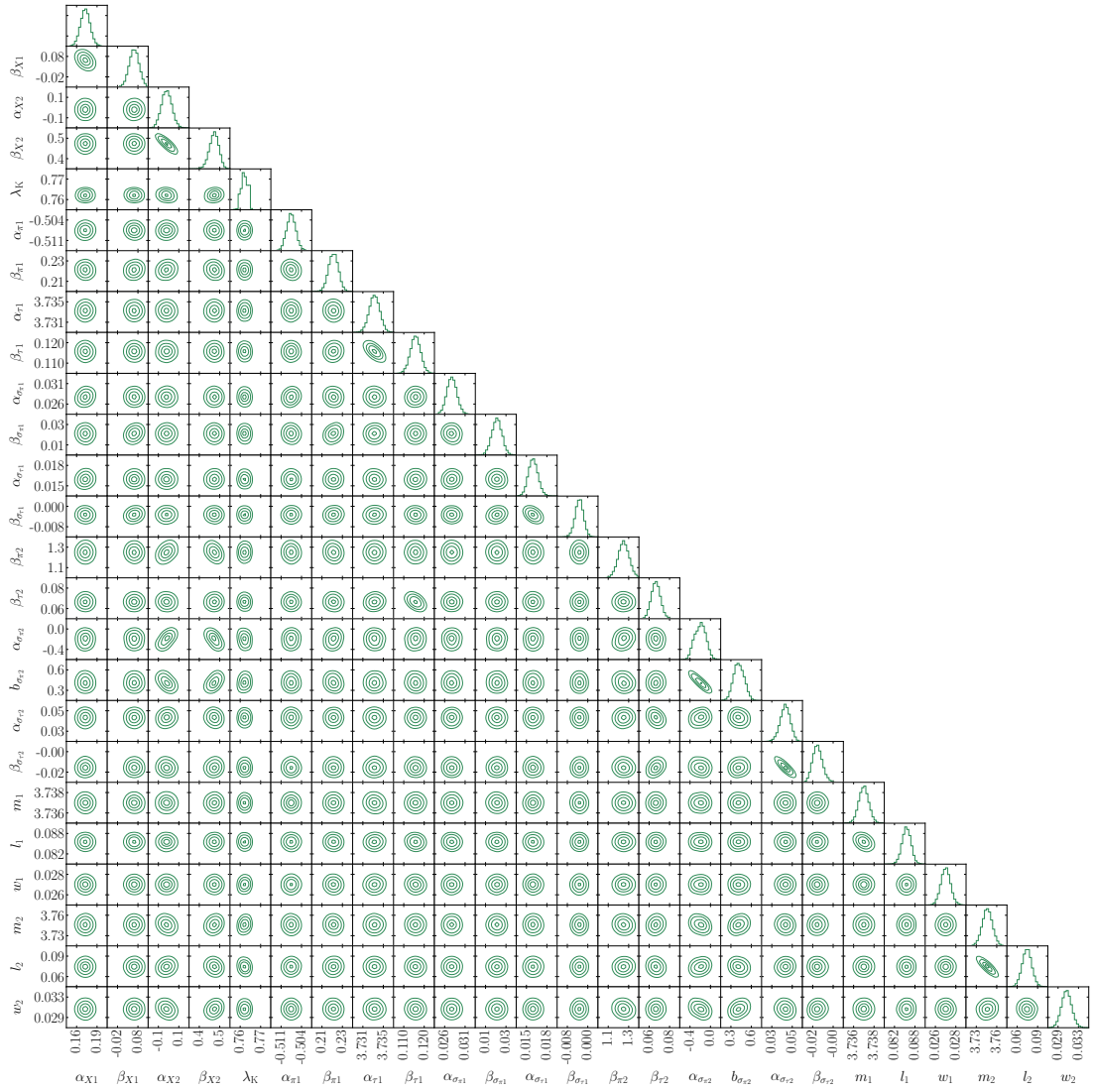

Figure 4.B.2 Corner plot resulting from the *emcee* run of our Bayesian model for the identification of contact binary stars. We ran the sampler for a total of 160 000 steps, but we discarded the first 10000 as burn-in, and we thinned the chains by a factor of 300. Visual inspection of the plot confirms that the number of steps was sufficient for the chains to converge.

4.C Additional tables and figures for the mass-ratio distribution

We show how the posterior distributions of b and $q_{\min}$ depend on the two mass-ratio distribution parameterizations (Fig. 4.C.1), fill-out factors (Fig. 4.C.2), splitting periods (Fig. 4.C.3), probability cutoffs (Fig. 4.C.4), and model hyperparameters (Fig. 4.C.5). We also present a complete list of all our *emcee* runs together with the resulting values of b and $q_{\min}$ (Tab. 4.C.1).

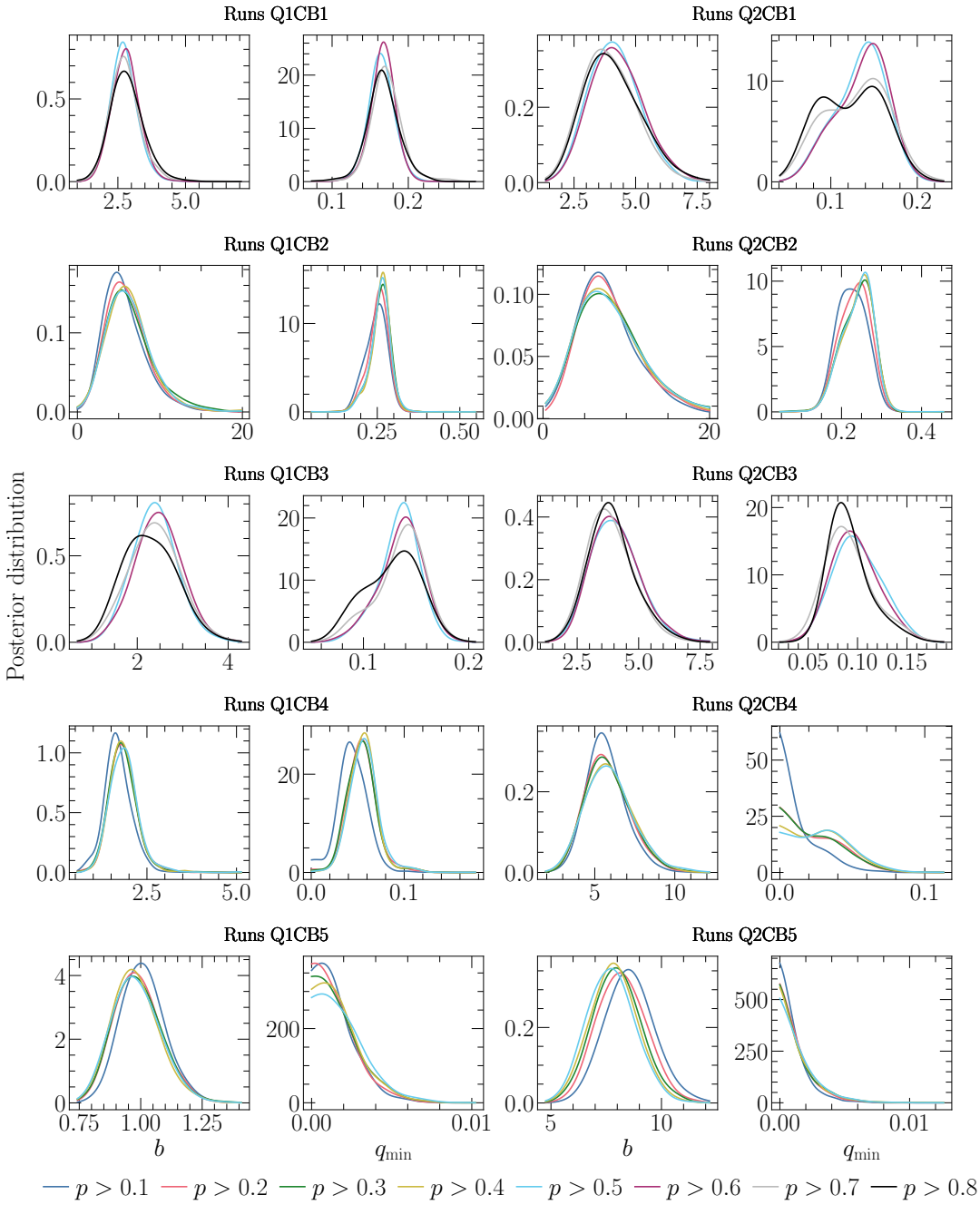

Figure 4.C.1 Posterior distributions of b and $q_{\min}$ conditional on Q_1 (left) and Q_2 (right) for populations CB1–CB5 and different probability cutoffs.

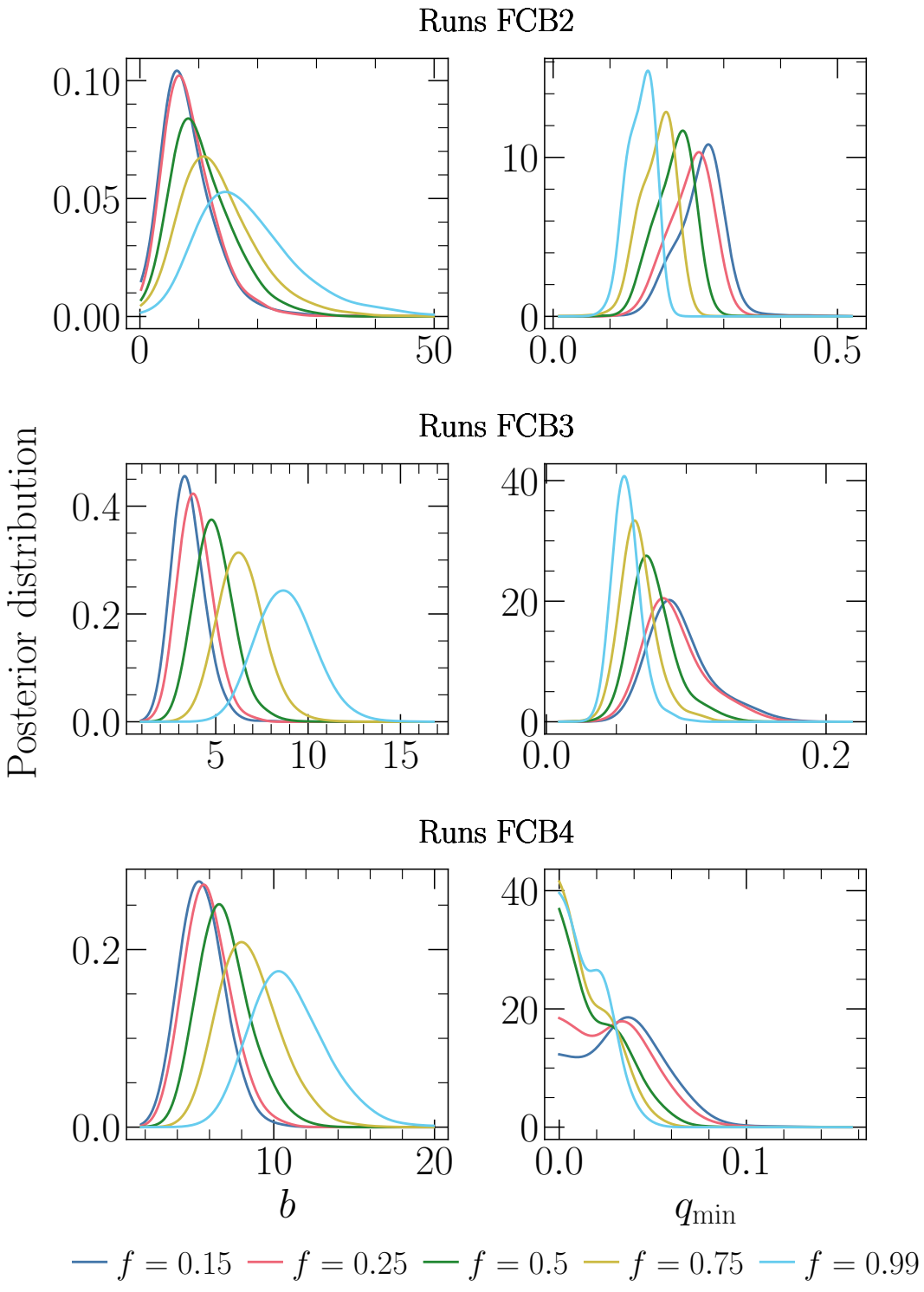

Figure 4.C.2 Dependence of the fiducial posterior distributions of b (left) and $q_{\min}$ (right) for populations CB2–CB4 on different fill-out factors.

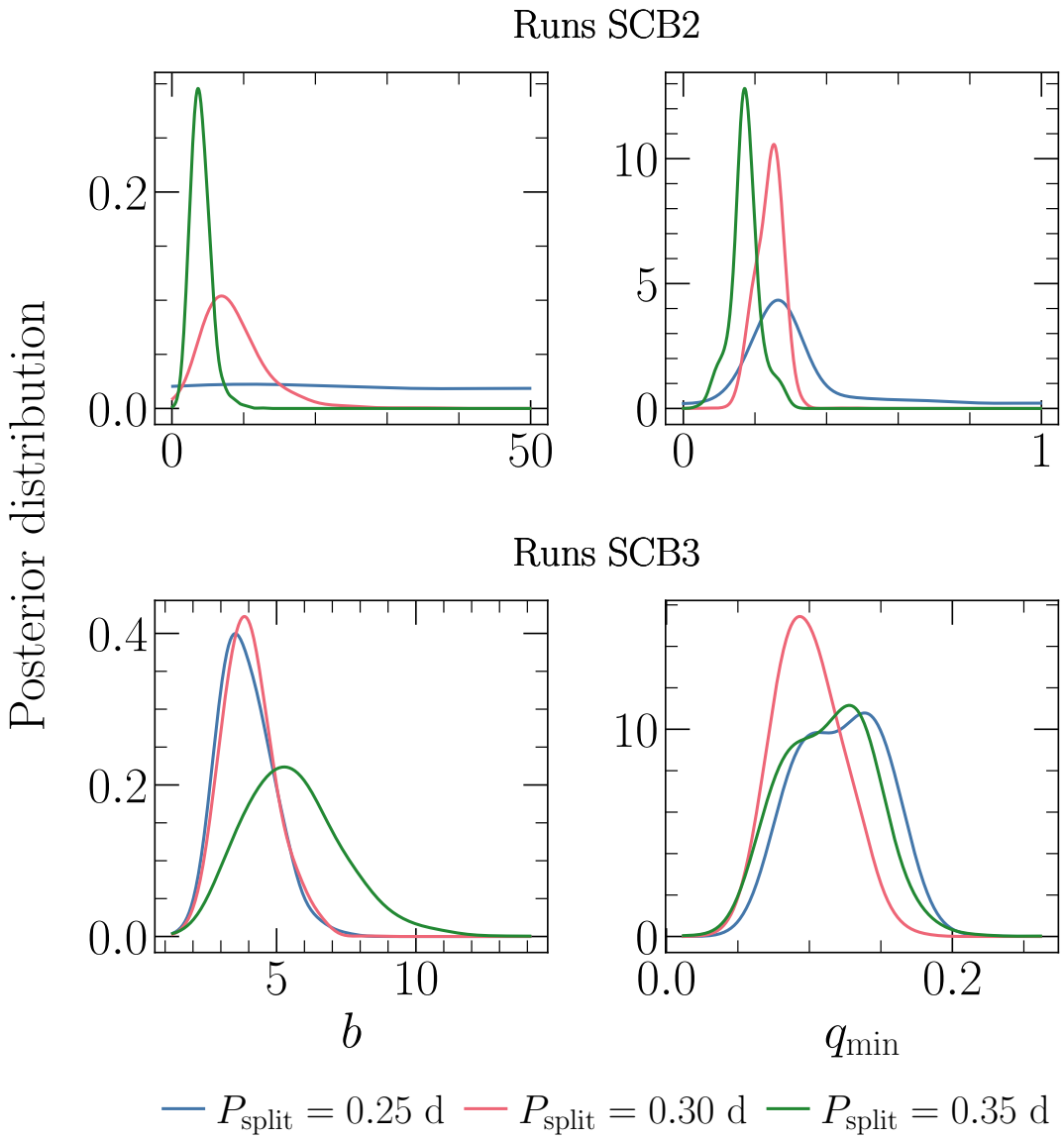

Figure 4.C.3 Dependence of the fiducial posterior distributions of b (left) and $q_{\min}$ (right) for populations CB2 and CB3 on different splitting periods between the two populations.

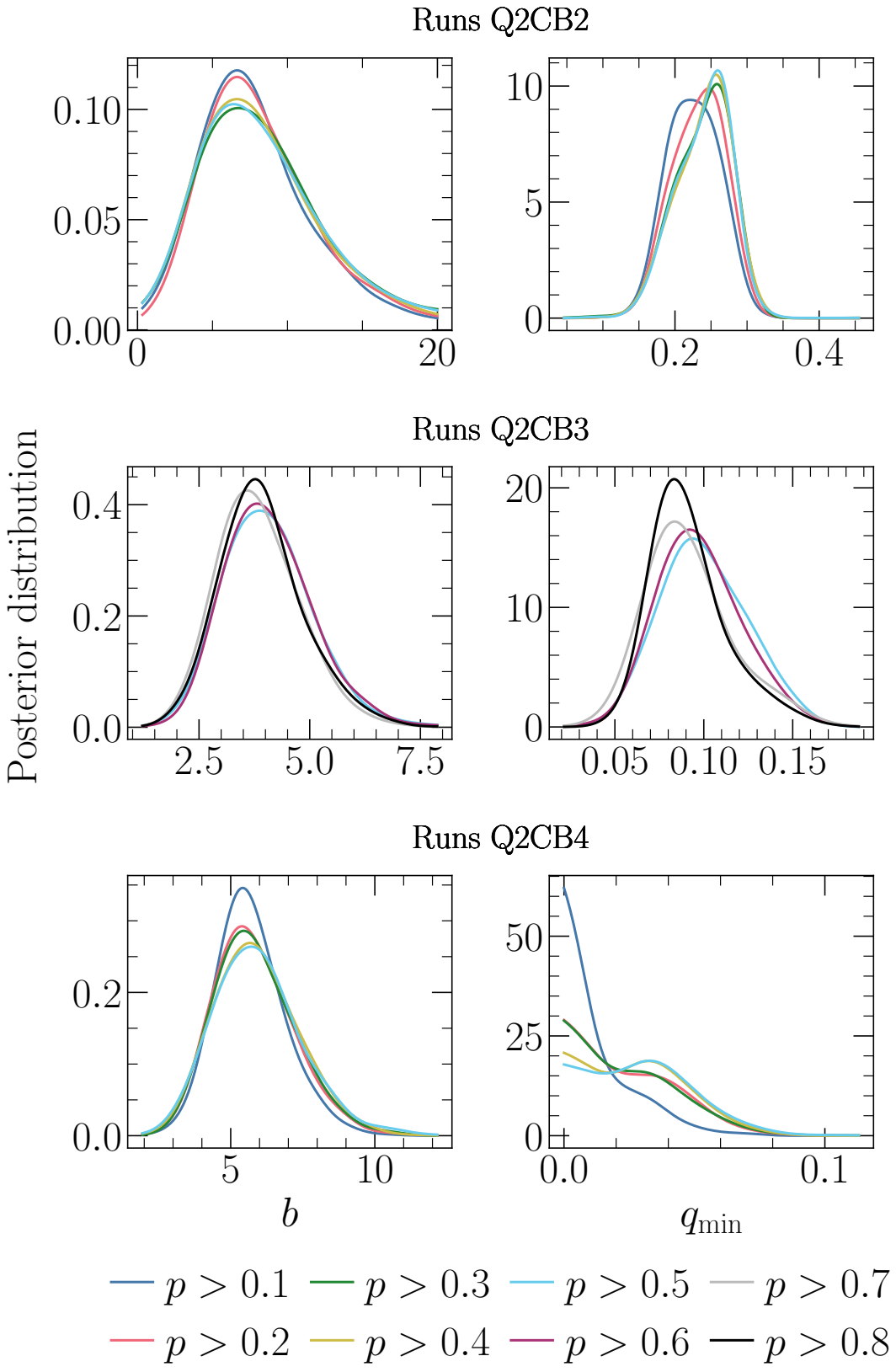

Figure 4.C.4 Dependence of the fiducial posterior distributions of b (left) and $q_{\min}$ (right) for populations CB2–CB4 on different probability cutoffs.

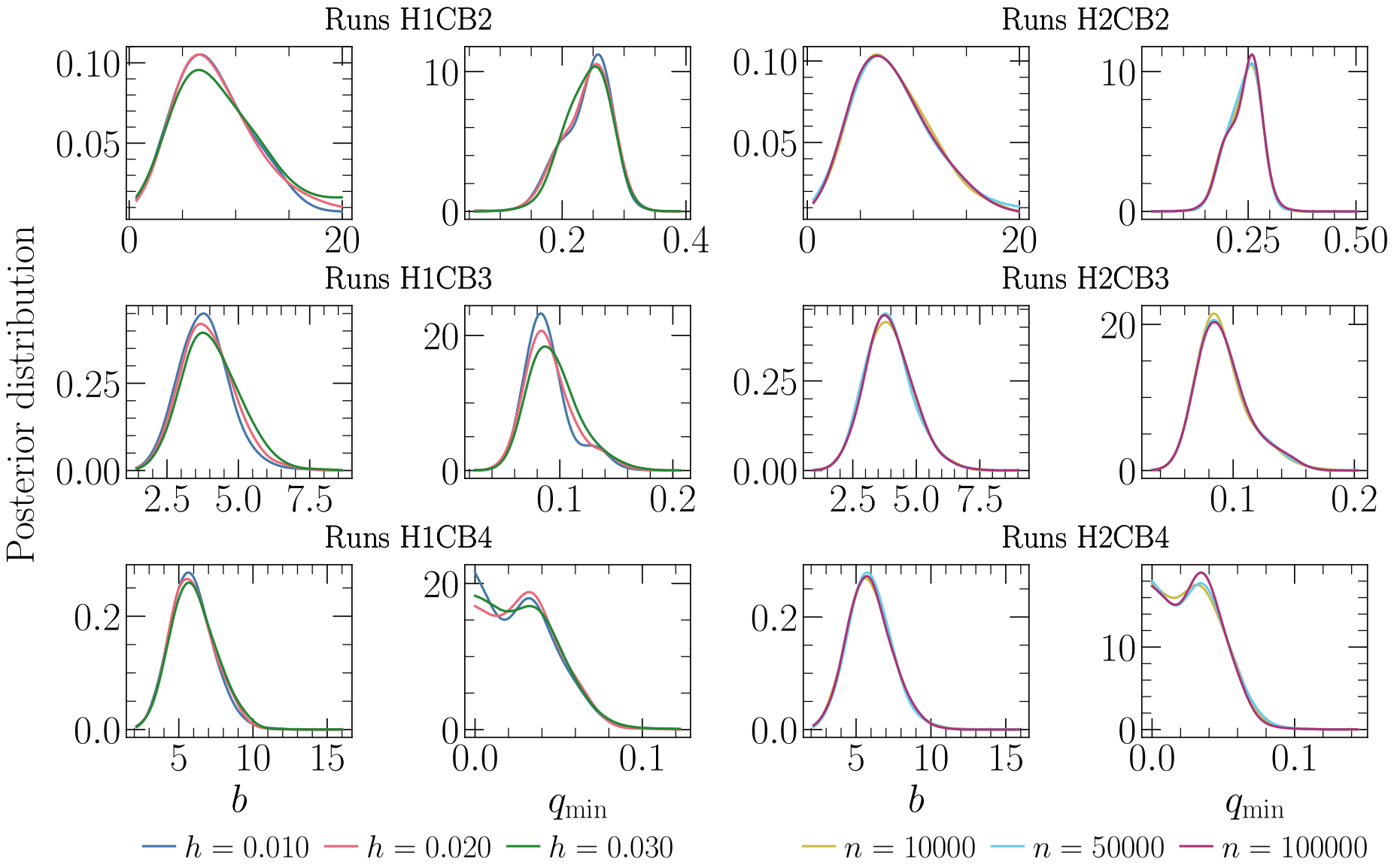

Figure 4.C.5 Dependence of the fiducial posterior distributions of b and $q_{\min}$ for populations CB2–CB4 on different values of h (left) and n (right).

Table 4.C.1 Full list of all *emcee* runs sampling the posterior distributions of the parameters of the mass-ratio distribution for different power-law prescriptions, probability cutoffs, fill-out factors, and hyperparameters.

Run	Q	f	h	n	Sample	Type	Prob. cutoff	P (days)	Eff. size	Autocorr. time	b	$q_{\min}$
Q1CB1p50	Q_1	0.25	0.02	10000	CB1p50	Late	0.5	—	258.99	11.25	$2.69^{+0.48}_{-0.43}$	$0.164^{+0.016}_{-0.015}$
Q1CB1p60	Q_1	0.25	0.02	10000	CB1p60	Late	0.6	—	256.27	11.47	$2.79^{+0.46}_{-0.48}$	$0.167^{+0.014}_{-0.015}$
Q1CB1p70	Q_1	0.25	0.02	10000	CB1p70	Late	0.7	—	249.00	10.94	$2.74^{+0.53}_{-0.48}$	$0.170^{+0.019}_{-0.017}$
Q1CB1p80	Q_1	0.25	0.02	10000	CB1p80	Late	0.8	—	228.04	1.11	$2.79^{+0.62}_{-0.52}$	$0.166^{+0.020}_{-0.017}$
Q1CB2p10	Q_1	0.25	0.02	10000	CB2p10	Late	0.1	≤ 0.3	62.40	13.24	$5.29^{+2.70}_{-1.83}$	$0.250^{+0.027}_{-0.037}$
Q1CB2p20	Q_1	0.25	0.02	10000	CB2p20	Late	0.2	≤ 0.3	61.96	15.8	$5.69^{+2.55}_{-1.99}$	$0.256^{+0.024}_{-0.032}$
Q1CB2p30	Q_1	0.25	0.02	10000	CB2p30	Late	0.3	≤ 0.3	61.38	16.97	$5.90^{+3.04}_{-2.12}$	$0.263^{+0.023}_{-0.033}$
Q1CB2p40	Q_1	0.25	0.02	10000	CB2p40	Late	0.4	≤ 0.3	61.07	14.61	$6.01^{+2.59}_{-2.22}$	$0.263^{+0.020}_{-0.032}$
Q1CB2p50	Q_1	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	15.69	$5.96^{+2.75}_{-2.18}$	$0.262^{+0.022}_{-0.031}$
Q1CB3p50	Q_1	0.25	0.02	10000	CB3p50	Late	0.5	> 0.3	198.41	1.9	$2.38^{+0.46}_{-0.47}$	$0.136^{+0.015}_{-0.020}$
Q1CB3p60	Q_1	0.25	0.02	10000	CB3p60	Late	0.6	> 0.3	197.30	9.88	$2.46^{+0.50}_{-0.48}$	$0.138^{+0.017}_{-0.023}$
Q1CB3p70	Q_1	0.25	0.02	10000	CB3p70	Late	0.7	> 0.3	192.69	10.79	$2.38^{+0.53}_{-0.55}$	$0.139^{+0.017}_{-0.028}$
Q1CB3p80	Q_1	0.25	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	10.39	$2.25^{+0.61}_{-0.53}$	$0.131^{+0.021}_{-0.033}$
Q1CB4p10	Q_1	0.25	0.02	10000	CB4p10	Early	0.1	< 1	106.42	10.55	$1.66^{+0.38}_{-0.29}$	$0.044^{+0.015}_{-0.013}$
Q1CB4p20	Q_1	0.25	0.02	10000	CB4p20	Early	0.2	< 1	105.56	13.19	$1.81^{+0.36}_{-0.32}$	$0.054^{+0.013}_{-0.015}$
Q1CB4p30	Q_1	0.25	0.02	10000	CB4p30	Early	0.3	< 1	105.56	11.9	$1.80^{+0.36}_{-0.33}$	$0.054^{+0.013}_{-0.015}$
Q1CB4p40	Q_1	0.25	0.02	10000	CB4p40	Early	0.4	< 1	104.91	9.56	$1.82^{+0.35}_{-0.31}$	$0.055^{+0.012}_{-0.014}$
Q1CB4p50	Q_1	0.25	0.02	10000	CB4p50	Early	0.5	< 1	104.42	11.77	$1.84^{+0.35}_{-0.35}$	$0.056^{+0.013}_{-0.014}$
Q1CB5p10	Q_1	0.25	0.02	10000	CB5p10	Early	0.1	—	162.62	25.9	$1.01^{+0.09}_{-0.08}$	$0.001^{+0.001}_{-0.001}$
Q1CB5p20	Q_1	0.25	0.02	10000	CB5p20	Early	0.2	—	161.55	24.71	$0.98^{+0.10}_{-0.08}$	$0.001^{+0.001}_{-0.001}$
Q1CB5p30	Q_1	0.25	0.02	10000	CB5p30	Early	0.3	—	160.80	29.96	$0.98^{+0.10}_{-0.09}$	$0.001^{+0.002}_{-0.001}$
Q1CB5p40	Q_1	0.25	0.02	10000	CB5p40	Early	0.4	—	159.49	30.56	$0.97^{+0.09}_{-0.08}$	$0.002^{+0.002}_{-0.001}$
Q1CB5p50	Q_1	0.25	0.02	10000	CB5p50	Early	0.5	—	158.58	28.57	$0.98^{+0.10}_{-0.09}$	$0.002^{+0.002}_{-0.001}$
Q2CB1p50	Q_2	0.25	0.02	10000	CB1p50	Late	0.5	—	258.99	10.61	$4.09^{+1.01}_{-0.95}$	$0.138^{+0.023}_{-0.035}$

Table 4.C.1 continued.

Run	Q	f	h	n	Sample	Type	Prob. cutoff	P (days)	Eff. size	Autocorr. time	b	$q_{\min}$
Q2CB1p60	Q_2	0.25	0.02	10000	CB1p60	Late	0.6	—	256.27	11.86	$4.17^{+1.09}_{-0.95}$	$0.141^{+0.022}_{-0.038}$
Q2CB1p70	Q_2	0.25	0.02	10000	CB1p70	Late	0.7	—	249.00	12.11	$3.84^{+1.16}_{-0.92}$	$0.135^{+0.029}_{-0.047}$
Q2CB1p80	Q_2	0.25	0.02	10000	CB1p80	Late	0.8	—	228.04	12.55	$3.96^{+1.23}_{-0.98}$	$0.127^{+0.033}_{-0.044}$
Q2CB2p10	Q_2	0.25	0.02	10000	CB2p10	Late	0.1	≤ 0.3	62.40	14.76	$7.30^{+4.23}_{-2.72}$	$0.226^{+0.037}_{-0.035}$
Q2CB2p20	Q_2	0.25	0.02	10000	CB2p20	Late	0.2	≤ 0.3	61.96	13.47	$7.56^{+4.05}_{-2.78}$	$0.235^{+0.032}_{-0.041}$
Q2CB2p30	Q_2	0.25	0.02	10000	CB2p30	Late	0.3	≤ 0.3	61.38	15.25	$7.73^{+4.43}_{-3.11}$	$0.244^{+0.031}_{-0.046}$
Q2CB2p40	Q_2	0.25	0.02	10000	CB2p40	Late	0.4	≤ 0.3	61.07	13.95	$7.58^{+4.37}_{-3.09}$	$0.246^{+0.029}_{-0.045}$
Q2CB2p50	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	12.93	$7.66^{+4.45}_{-3.15}$	$0.246^{+0.029}_{-0.046}$
Q2CB3p50	Q_2	0.25	0.02	10000	CB3p50	Late	0.5	> 0.3	198.41	10.25	$3.98^{+1.01}_{-0.87}$	$0.098^{+0.028}_{-0.021}$
Q2CB3p60	Q_2	0.25	0.02	10000	CB3p60	Late	0.6	> 0.3	197.30	9.51	$3.98^{+1.02}_{-0.84}$	$0.096^{+0.026}_{-0.020}$
Q2CB3p70	Q_2	0.25	0.02	10000	CB3p70	Late	0.7	> 0.3	192.69	9.75	$3.75^{+0.99}_{-0.77}$	$0.088^{+0.027}_{-0.019}$
Q2CB3p80	Q_2	0.25	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	9.66	$3.84^{+0.96}_{-0.80}$	$0.087^{+0.024}_{-0.015}$
Q2CB4p10	Q_2	0.25	0.02	10000	CB4p10	Early	0.1	< 1	106.42	20.35	$5.57^{+1.23}_{-1.00}$	$0.008^{+0.020}_{-0.006}$
Q2CB4p20	Q_2	0.25	0.02	10000	CB4p20	Early	0.2	< 1	105.56	16.96	$5.64^{+1.45}_{-1.14}$	$0.022^{+0.023}_{-0.017}$
Q2CB4p30	Q_2	0.25	0.02	10000	CB4p30	Early	0.3	< 1	105.56	15.21	$5.70^{+1.52}_{-1.15}$	$0.023^{+0.021}_{-0.018}$
Q2CB4p40	Q_2	0.25	0.02	10000	CB4p40	Early	0.4	< 1	104.91	13.68	$5.80^{+1.54}_{-1.27}$	$0.029^{+0.018}_{-0.022}$
Q2CB4p50	Q_2	0.25	0.02	10000	CB4p50	Early	0.5	< 1	104.42	13.28	$5.82^{+1.52}_{-1.30}$	$0.030^{+0.018}_{-0.022}$
Q2CB5p10	Q_2	0.25	0.02	10000	CB5p10	Early	0.1	—	162.62	25.37	$8.52^{+1.08}_{-1.02}$	$0.001^{+0.001}_{-0.001}$
Q2CB5p20	Q_2	0.25	0.02	10000	CB5p20	Early	0.2	—	161.55	23.00	$8.17^{+1.10}_{-1.06}$	$0.001^{+0.001}_{-0.001}$
Q2CB5p30	Q_2	0.25	0.02	10000	CB5p30	Early	0.3	—	160.80	30.26	$7.98^{+1.03}_{-1.06}$	$0.001^{+0.002}_{-0.001}$
Q2CB5p40	Q_2	0.25	0.02	10000	CB5p40	Early	0.4	—	159.49	25.45	$7.85^{+1.01}_{-1.01}$	$0.001^{+0.002}_{-0.001}$
Q2CB5p50	Q_2	0.25	0.02	10000	CB5p50	Early	0.5	—	158.58	24.71	$7.71^{+1.05}_{-1.01}$	$0.001^{+0.002}_{-0.001}$
FCB2f015	Q_2	0.15	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	20.83	$7.54^{+4.96}_{-3.11}$	$0.266^{+0.028}_{-0.051}$
FCB2f025	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	16.90	$7.71^{+4.81}_{-2.97}$	$0.247^{+0.030}_{-0.044}$
FCB2f050	Q_2	0.50	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	18.91	$9.87^{+6.04}_{-3.78}$	$0.219^{+0.026}_{-0.043}$

Table 4.C.1 continued.

Run	Q	f	h	n	Sample	Type	Prob. cutoff	P (days)	Eff. size	Autocorr. time	b	$q_{\min}$
FCB2f075	Q_2	0.75	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	19.59	$12.36^{+6.88}_{-4.69}$	$0.189^{+0.024}_{-0.037}$
FCB2f099	Q_2	0.99	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	18.10	$16.72^{+9.72}_{-6.04}$	$0.156^{+0.021}_{-0.027}$
FCB3f015	Q_2	0.15	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	11.34	$3.44^{+0.89}_{-0.81}$	$0.092^{+0.025}_{-0.016}$
FCB3f025	Q_2	0.25	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	10.79	$3.84^{+0.97}_{-0.84}$	$0.087^{+0.026}_{-0.015}$
FCB3f050	Q_2	0.50	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	10.39	$4.85^{+1.09}_{-0.95}$	$0.074^{+0.017}_{-0.012}$
FCB3f075	Q_2	0.75	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	1.12	$6.34^{+1.14}_{-1.15}$	$0.064^{+0.013}_{-0.010}$
FCB3f099	Q_2	0.99	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	11.95	$8.69^{+1.62}_{-1.40}$	$0.057^{+0.009}_{-0.009}$
FCB4f015	Q_2	0.15	0.02	10000	CB4p50	Early	0.5	< 1	104.42	12.25	$5.51^{+1.35}_{-1.27}$	$0.037^{+0.019}_{-0.024}$
FCB4f025	Q_2	0.15	0.02	10000	CB4p50	Early	0.5	< 1	104.42	12.64	$5.79^{+1.49}_{-1.26}$	$0.030^{+0.020}_{-0.022}$
FCB4f050	Q_2	0.15	0.02	10000	CB4p50	Early	0.5	< 1	104.42	20.48	$6.78^{+1.66}_{-1.34}$	$0.017^{+0.020}_{-0.013}$
FCB4f075	Q_2	0.15	0.02	10000	CB4p50	Early	0.5	< 1	104.42	19.77	$8.26^{+2.07}_{-1.55}$	$0.014^{+0.018}_{-0.011}$
FCB4f099	Q_2	0.15	0.02	10000	CB4p50	Early	0.5	< 1	104.42	14.56	$10.71^{+2.43}_{-1.85}$	$0.015^{+0.013}_{-0.011}$
H1CB2h001	Q_2	0.25	0.01	10000	CB2p50	Late	0.5	≤ 0.3	60.59	16.76	$7.62^{+4.48}_{-3.09}$	$0.248^{+0.027}_{-0.050}$
H1CB2h002	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	14.10	$7.68^{+4.66}_{-3.03}$	$0.246^{+0.030}_{-0.048}$
H1CB2h003	Q_2	0.25	0.03	10000	CB2p50	Late	0.5	≤ 0.3	60.59	14.74	$7.98^{+4.91}_{-3.34}$	$0.244^{+0.032}_{-0.040}$
H1CB3h001	Q_2	0.25	0.01	10000	CB3p80	Late	0.8	> 0.3	177.76	1.12	$3.78^{+0.84}_{-0.81}$	$0.085^{+0.020}_{-0.014}$
H1CB3h002	Q_2	0.25	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	1.10	$3.85^{+0.96}_{-0.79}$	$0.087^{+0.024}_{-0.015}$
H1CB3h003	Q_2	0.25	0.03	10000	CB3p80	Late	0.8	> 0.3	177.76	10.39	$3.98^{+1.07}_{-0.83}$	$0.092^{+0.024}_{-0.018}$
H1CB4h001	Q_2	0.25	0.01	10000	CB4p50	Early	0.5	< 1	104.42	15.33	$5.80^{+1.44}_{-1.25}$	$0.030^{+0.019}_{-0.024}$
H1CB4h002	Q_2	0.25	0.02	10000	CB4p50	Early	0.5	< 1	104.42	13.40	$5.82^{+1.56}_{-1.26}$	$0.031^{+0.019}_{-0.022}$
H1CB4h003	Q_2	0.25	0.03	10000	CB4p50	Early	0.5	< 1	104.42	11.51	$5.90^{+1.63}_{-1.28}$	$0.030^{+0.021}_{-0.021}$
H2CB2n10k	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.3	60.59	17.3	$7.74^{+4.49}_{-2.98}$	$0.245^{+0.029}_{-0.048}$
H2CB2n50k	Q_2	0.25	0.02	50000	CB2p50	Late	0.5	≤ 0.3	60.59	14.76	$7.70^{+4.78}_{-3.09}$	$0.245^{+0.029}_{-0.043}$
H2CB2n100k	Q_2	0.25	0.02	100000	CB2p50	Late	0.5	≤ 0.3	60.59	14.73	$7.75^{+4.51}_{-3.14}$	$0.249^{+0.025}_{-0.050}$
H2CB3n10k	Q_2	0.25	0.02	10000	CB3p80	Late	0.8	> 0.3	177.76	1.11	$3.87^{+0.95}_{-0.84}$	$0.087^{+0.024}_{-0.015}$

Table 4.C.1 continued.

Run	Q	f	h	n	Sample	Type	Prob. cutoff	P (days)	Eff. size	Autocorr. time	b	$q_{\min}$
H2CB3n50k	Q_2	0.25	0.02	50000	CB3p80	Late	0.8	> 0.3	177.76	9.22	$3.84^{+0.94}_{-0.82}$	$0.088^{+0.024}_{-0.015}$
H2CB3n100k	Q_2	0.25	0.02	100000	CB3p80	Late	0.8	> 0.3	177.76	9.75	$3.88^{+0.95}_{-0.78}$	$0.088^{+0.024}_{-0.016}$
H2CB4n10k	Q_2	0.25	0.02	10000	CB4p50	Early	0.5	< 1	104.42	12.42	$5.81^{+1.52}_{-1.26}$	$0.030^{+0.020}_{-0.022}$
H2CB4n50k	Q_2	0.25	0.02	50000	CB4p50	Early	0.5	< 1	104.42	11.28	$5.87^{+1.40}_{-1.27}$	$0.031^{+0.019}_{-0.023}$
H2CB4h100k	Q_2	0.25	0.02	100000	CB4p50	Early	0.5	< 1	104.42	11.26	$5.82^{+1.52}_{-1.28}$	$0.031^{+0.018}_{-0.022}$
SCB2P025	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.25	9.96	39.71	$23.37^{+18.37}_{-15.40}$	$0.276^{+0.173}_{-0.069}$
SCB2P030	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.30	60.59	20.52	$7.87^{+4.47}_{-3.15}$	$0.245^{+0.030}_{-0.045}$
SCB2P035	Q_2	0.25	0.02	10000	CB2p50	Late	0.5	≤ 0.35	135.36	11.7	$3.86^{+1.38}_{-1.14}$	$0.172^{+0.032}_{-0.031}$
SCB3P025	Q_2	0.25	0.02	10000	CB3p50	Late	0.5	> 0.25	249.04	1.11	$3.79^{+1.09}_{-0.80}$	$0.124^{+0.030}_{-0.035}$
SCB3P030	Q_2	0.25	0.02	10000	CB3p50	Late	0.5	> 0.30	198.41	10.51	$3.94^{+0.95}_{-0.83}$	$0.098^{+0.027}_{-0.021}$
SCB3P035	Q_2	0.25	0.02	10000	CB3p50	Late	0.5	> 0.35	123.63	12.50	$5.45^{+1.83}_{-1.57}$	$0.115^{+0.028}_{-0.036}$

5 Distinguishing between light curves of ellipsoidal variables with massive dark companions, contact binaries, and semidetached binaries using principal component analysis*

Authors: M. Pešta and O. Pejcha
Submitted to: Astronomy & Astrophysics

Abstract

Photometric methods for identifying dark companion binaries – binary systems hosting quiescent black holes and neutron stars – operate by detecting ellipsoidal variations caused by tidal interactions. The limitation of this approach is that contact and semidetached binaries can produce similarly looking light curves. In this work, we address the degeneracy of ellipsoidal light curves by studying the differences between synthetically generated light curves of dark companion, semidetached, and contact binary systems. We inject the light curves with various levels of uncorrelated and correlated Gaussian noise to simulate the effects of instrumental noise and stellar spots. Using principal component analysis (PCA) and Fourier decomposition, we construct low-dimensional representations of the light curves. We find that the first two to five PCA components are sufficient to explain 99% of variance in the data. The PCA representations are generally more informative than the Fourier representation for the same number of coefficients as measured by both the silhouette scores of the representations and the macro recalls of random forest classifiers trained on the representations. The random forest classifiers reach macro recalls from 0.97 in the complete absence of noise to 0.70 in the presence of spots and strong instrumental noise, indicating that the classes remain largely separable even under adverse conditions. We find that instrumental noise significantly impacts the class separation only when its standard deviation exceeds 10^{-3} mag, whereas the presence of spots can markedly reduce the class separation even when they contribute as little as 1% of the light curve amplitude. We discuss the application of our method to real ellipsoidal samples, and we show that we can increase the purity of a sample of dark companion candidates by a factor of up to 27 if we assume a prior purity of 1%, significantly improving the cost-efficiency of follow-up observations.

*The following text is a preprint version of an article submitted for publication in Astronomy & Astrophysics. The preprint is available at `https://arxiv.org/pdf/2408.11100`.

5.1 Introduction

Most stellar-mass black holes (BHs) are discovered in binary systems, where their presence is revealed either through high-energy emission from accretion processes (e.g., Remillard & McClintock 2006; Corral-Santana et al. 2016) or by gravitational waves radiated during mergers with companions (e.g., Abbott et al. 2016; Abbott et al. 2023). Many binaries with neutron stars (NSs) have been discovered in the same way. In reality, only a small fraction of BH and NS binary configurations are expected to yield observable X-ray or gravitational wave signatures. This suggests the existence of a large population of *dark companion binaries*, which host electromagnetically silent BHs and NSs orbited by normal luminous stars. Given that a significant fraction of BH binaries might actually be wide-orbit binaries (Breivik et al. 2017; Chawla et al. 2022), characterizing the dark companion population is crucial for enhancing our understanding of the evolution of massive stars and the formation of compact objects.

Without accretion or mergers, the presence of a dark companion in the system can only be infered from subtle photometric, spectroscopic, and astrometric effects that it induces in the companion star. For this reason, only a few dark companion binaries have been discovered so far. A non-exhaustive list of dark companion detections includes two BHs and one BH candidate in the globular cluster NGC 3201 identified using spectroscopy from MUSE (Giesers et al. 2018; Giesers et al. 2019), one BH or NS identified in data from APOGEE (Thompson et al. 2019), two BHs found in catalogs of single-lined spectroscopic binaries (Mahy et al. 2022; Shenar et al. 2022), and three BHs, one NS, and 20 NS candidates detected in *Gaia* astrometry (El-Badry et al. 2023b,a, 2024b; Gaia Collaboration et al. 2024; El-Badry et al. 2024a). In all these studies, a small number of candidates were selected based on criteria derived from available spectroscopic or astrometric data. The most promising candidates were then followed up with high-resolution spectroscopy, if not already available, to confirm the presence of the dark companion. The limitation of this approach is that spectroscopy and astrometry are available only for a small fraction of stars, significantly reducing the pool of candidates for follow-up analysis. While photometry is available for a much larger number of stars, the challenge lies in identifying the most promising candidates based solely on photometric signatures of dark companions.

In a close binary system consisting of a star and a massive dark companion, the gravitational pull of the companion tidally distorts the star, inducing ellipsoidal variations in the light curve of the system. In principle, by sifting through large photometric surveys and identifying stars that exhibit ellipsoidal variations, we can select candidates for follow-up analysis that are likely to harbor dark companions. Recent examples of such work include Green et al. (2023), who identified over 15 000 ellipsoidal variables in data from *TESS*, Gomel et al. (2023), who presented over 6 000 dark companion candidates from *Gaia* DR3, and Gomel et al. (2021c), who studied over 10 000 ellipsoidal variables from OGLE. The problem with this method is that besides dark companion binaries, ellipsoidal samples typically contain large numbers of contact binaries, semidetached binaries, and possibly other types of objects that produce similar light curves. In fact, dark companion binaries most likely make up only a small fraction of ellipsoidal variables, making it extremely cost-inefficient to follow up on all candidates with high-resolution

spectroscopy.

To increase the fraction of dark companion binaries in ellipsoidal samples, further filtering is required. For example, we could filter objects based on the quality of their spectral energy distribution fits assuming a single-star model (Kapusta & Mróz 2023) or we could consider only objects with high binary mass functions (Rowan et al. 2024). However, the former method requires multi-band photometry, which is not always available, while the latter relies on radial velocity measurements, thus defeating the goal of avoiding the need for spectroscopy in the candidate selection process. In our previous work (Pešta & Pejcha 2023), we used Bayesian mixture modeling to isolate a sample of contact binaries from the Kepler Eclipsing Binary Catalog. In principle, we could use the same method to exclude contact binaries from samples of ellipsoidal variables, but the method requires estimates of effective temperatures and luminosities for all objects in the sample, limiting its applicability.

A particularly attractive way of selecting dark companion binary candidates using only information contained in their broadband photometric light curves was developed by Gomel et al. (2021a,b), who introduced a proxy for the minimum mass ratio of dark companion binaries derived from the observed ellipsoidal amplitude of the system. This proxy, which they termed the modified minimum mass ratio (mMMR), is always strictly lower than the actual mass ratio of the system, and its large values can be indicative of the presence of a massive dark companion. The method is most sensitive to dark companion binaries with primaries close to filling their Roche lobes and inclinations close to 90°. Conversely, low-inclination systems or systems with a primary that did not yet evolve to fill its Roche lobe will show small mMMR even for large mass ratios. The method assumes that and all variability comes from the tidal deformation of the primary induced by the dark companion. When this assumption is violated, the method yields spurious results, resulting in high false-positive rates. For example, Nagarajan et al. (2023) spectroscopically followed up on the 14 most promising candidates obtained using the mMMR method by Gomel et al. (2023) and found that all harbor a low-mass non-degenerate star instead of a dark companion, with spotted contact binaries being the most likely culprits behind the false positives. Consequently, the efficiency of the mMMR method hinges on the purity of the ellipsoidal sample, which is typically low due to the prominent presence of contaminants (e.g., Green et al. 2023).

A proper way to address the issue of photometric identification of dark companion binaries would be to train a machine learning classifier on a large sample of ellipsoidal light curves for which we know the true nature of the systems, allowing the classifier to learn the differences between the classes and automatically detect dark companion binaries in new data. Many have followed this approach in the wider context of automatic classification of periodic variables, e.g., Paczyński et al. (2006); Pawlak et al. (2016); Soszyński et al. (2016); Jayasinghe et al. (2019); Cheung et al. (2021), etc. However, this method requires a well-curated training sample in which all classes are sufficiently represented. This is generally not an issue in variable star classification, where large samples of studied objects are readily available, but a representative sample of confirmed dark companion binaries is currently lacking, preventing us from following this approach in the context of dark companion binary identification.

Even in the absence of a well-defined sample of dark companion binaries, it is still possible to study the degeneracy of ellipsoidal light curves using theoretical models. In this work, we investigate the similarities and differences between synthetically generated light curves of dark companion binaries, semidetached binaries, and contact binaries, which we consider to be to the most likely dark companion binary impostors. We inject the light curves with various levels of uncorrelated and correlated Gaussian noise to simulate the effects of instrumental noise and stellar spots (Sect. 5.2), allowing us to study the separation of the classes under adverse observing conditions. To better visualize the light curves and make the differences between the classes more pronounced, we reduce the light curves using principal component analysis (PCA) and Fourier decomposition. We compare the informativeness of the PCA and Fourier representations using the silhouette score, and we quantify the separation of the classes in each representation using the macro recall of random forest classifiers trained on the representations. We describe the methodological details of our analysis in Sect. 5.3, and we present the results of our study in Sect. 5.4. We summarize and discuss the implications of our findings in Sect. 5.5.

5.2 Synthetic data

The small number of confirmed dark companion binaries prevents us from using real observations to systematically study the degeneracy of ellipsoidal light curves. To overcome this limitation, we generated synthetic light curves of dark companion binaries and their common contaminants (Sect. 5.2.1), which we further modified with correlated and uncorrelated noise to account for the effects of instrumental noise and stellar spots (Sec. 5.2.2).

5.2.1 Physical models

We used PHOEBE v2.4.10 (Prša et al. 2016; Conroy et al. 2020b) to generate synthetic light curves of dark companion binaries, semidetached binaries, and contact binaries. We started by initializing the default detached, semidetached or contact binary system, conditional on the type of the variable we wanted to generate. In all cases, we set the passband to `TESS:T`, the number of triangles of the stellar components to 10 000, and we kept the default limb darkening calculation settings, with the coefficients interpolated directly from either the `PHOENIX` or the `ck2004` model atmosphere tables, depending on the effective temperatures of the stellar components. In dark companion systems, we set `distortion_method` = `none` for the dark companion, allowing us to isolate the variations caused by the tidal deformation of the star without accounting for the presence of eclipses. In all other cases, we kept the default `distortion_method` = `roche`.

For each binary class, we defined a grid of physical and orbital parameters that affect the shape of the light curve. In the case of dark companion binaries, the light curve does not significantly depend on the mass M nor the effective temperature $T_{\rm eff}$ of the stellar component but rather on the mass ratio q, the inclination i, and the semi-major axis a of the system. We therefore fixed $M = 1\,M_\odot$ and $T_{\rm eff} = 6\,000\,\mathrm{K}$. We varied q from 0.05 to 10 with a step of 0.05 for $0.05 \leq q \leq 1$

and a step of 1 for $1 < q \leq 10$, i from 5° to 90° with a step of 5°, and a from $1\,R_\odot$ to $10\,R_\odot$ with ten evenly spaced steps in the logarithmic scale. We fixed the equivalent radius R of the stellar component at $1\,R_\odot$. Due to the scaling properties of the Roche potential, varying R has the same impact on the shape of the light curve as varying a.

Compared to the dark companion case, the light curves of semidetached variables additionally depend on the ratios of $T_{\rm eff}$ and R of the stellar components. We considered $T_{\rm eff}$ ratios of 0.5, 1, and 2, and we varied $T_{\rm eff}$, the gravity brightening coefficients, and the bolometric reflection coefficients of the system accordingly so that both the primary and the secondary were covered by the available atmosphere tables and the system passed all PHOEBE internal checks. We considered four different R ratios: 0.1, 0.5, 2, and 5, with R of the primary fixed by the condition that the primary fills its Roche lobe. We sampled q, i, and a in the same way as in the dark companion case.

The simplest is the case of contact binary stars, whose light curves depend primarily on q, i, and the fill-out factor f of the system. We sampled i and q in the same way as in the previous cases, with the only difference being that we limited q to $0.05 \leq q \leq 1$. We considered $f = 0.15$, 0.25, 0.5, and 0.75, and we assumed that both components share a common atmosphere with $T_{\rm eff} = 6000\,{\rm K}$.

Not all parameter combinations produced a valid light curve. Some setups were not covered by either atmosphere table or yielded a configuration that was incompatible with the assumed binary class (e.g., Roche overflow in dark companion binaries). We excluded these configurations from the analysis. We also excluded any setup that yielded a light curve with a photometric amplitude smaller than 0.01 mag. By performing these cuts, we obtained samples of 492 dark companion, 37 386 semidetached, and 1 302 contact binary synthetic light curves. The observed disparity in the sample sizes does not reflect the relative occurrence rates of the three binary classes but rather the relative extents of their parameter spaces. To counter this imbalance, we randomly undersampled the semidetached binary class by a factor of 20, resulting in a total of 1 846 semidetached light curves. Finally, we randomly split the data into training, validation, and test sets, with 20% of each variable class going to the test set and 20% going to the validation set.

We generated the light curves in the magnitude space with a resolution of 100 points per orbit, covering phase from -0.5 to 0.5. We shifted the light curves so that the phase 0 corresponds to the primary minimum, and we discarded the rightmost point of each light curve to avoid redundancy at phase 0.5. We normalized the light curves by vertically shifting and rescaling them in such a way that the maximum and the minimum were equal to 1 and 0, respectively. Hereafter, we shall refer to this sample of normalized synthetic light curves as S0.

5.2.2 Addition of noise and oversampling

To account for the effects of instrumental noise and stellar spots, we injected the sample S0 with various levels of uncorrelated and correlated noise. We modeled the instrumental noise as uncorrelated Gaussian noise with standard deviations $\sigma_{\rm WN} = 10^{-4}$, 10^{-3}, and 10^{-2} mag, covering almost the entire range of the *TESS* noise characteristic curve (Ricker et al. 2015). We modeled the effects of

spots as correlated Gaussian noise, which we generated using the `scikit-learn` implementation of Gaussian processes with a periodic `ExpSineSquared` kernel. We considered correlated noise with standard deviations $\sigma_{\rm CN} = 0.01$, 0.05, and 0.10 of the unperturbed light curve amplitude and correlation length scales $l_{\rm CN} = 0.25$, 0.50, and 1.00 of the orbital period. The standard deviation of the injected correlated noise is proportional to the amplitude of the unperturbed light curves, because we want to simulate a scenario in which stellar spots account for a specific fraction of the overall light curve amplitude, having the same relative effect on all light curves. Also, since we assume that all variability in the light curves before noise injection comes from eclipses and ellipsoidal variations, the light curve amplitude should be zero for a system observed exactly face-on irrespective of the presence of spots, which would not be the case if we injected correlated noise with an absolute standard deviation.

For each combination of the levels of correlated and uncorrelated noise, including the complete absence of noise, we generated multiple realizations of each light curve in the sample S0, with the oversampling factor inversely proportional to the occurrence rate of the corresponding binary class in the sample. We oversampled each semidetached and contact binary light curve 10 times, and each dark companion binary light curve 30 times, resulting in 40 synthetic samples with well-sampled noise distributions. To prevent data leakage, we oversampled the training, validation, and test sets separately, so that each light curve and its noisy realizations were present in only one of the sets. After injecting the noise, we normalized the light curves by: i) fitting each light curve with a fourth-order Fourier series, ii) horizontally shifting the light curves so that the primary minimum (corresponding to the maximum magnitude) of the Fourier fit is at phase 0, iii) vertically shifting and rescaling the light curves so that the Fourier fit has a minimum and maximum of 0 and 1, respectively. We present a list of all synthetic samples and their noise characteristics in Table 5.1.

5.3 Methods

Light curves can be viewed as vectors in a high-dimensional space, with the dimension of the space given by the total number of data points in the light curve. Current space-based photometric surveys have typical sampling frequencies of the order of seconds to minutes, resulting in densely sampled light curves with thousands of points. Intuitively, the denser the sampling, the more information the light curve contains and the easier it should be to construct a classifier that can distinguish between different types of objects generating the light curves. In practice, the high-dimensional nature of the data might actually hurt the performance of the classifier. The reason is that as the dimension of the data increases, the number of samples required to evenly cover the space grows exponentially, and the available data become increasingly sparse. This phenomenon, known as the *curse of dimensionality* (Bellman 1957), is further exacerbated by the fact that real-life light curves are often contaminated by noise and outliers, which in combination with overfitting can lead to poor generalization to previously unseen data.

Training an accurate light curve classifier that generalizes well to new observations requires a training sample which is representative of the true

Table 5.1 List of synthetic samples of dark companion, semidetached, and contact binary light curves. Each sample was generated with a different combination of uncorrelated noise standard deviation σ_{WN}, correlated noise standard deviation σ_{CN} (proportional to the unperturbed light curve amplitude), and correlation length scale l_{CN} (in units of the orbital period). All samples contain the same number of light curves, with 14 760 coming from the dark companion class, 18 460 from the semidetached class, and 13 020 from the contact class, resulting in a total of 46 240 light curves.

Sample	σ_{WN} (mag)	σ_{CN}	l_{CN}
W0C0	–	–	–
W0C1L25	–	0.01	0.25
W0C1L50	–	0.01	0.50
W0C1L100	–	0.01	1.00
W0C5L25	–	0.05	0.25
W0C5L50	–	0.05	0.50
W0C5L100	–	0.05	1.00
W0C10L25	–	0.10	0.25
W0C10L50	–	0.10	0.50
W0C10L100	–	0.10	1.00
W1C0	10^{-4}	–	–
W1C1L25	10^{-4}	0.01	0.25
W1C1L50	10^{-4}	0.01	0.50
W1C1L100	10^{-4}	0.01	1.00
W1C5L25	10^{-4}	0.05	0.25
W1C5L50	10^{-4}	0.05	0.50
W1C5L100	10^{-4}	0.05	1.00
W1C10L25	10^{-4}	0.10	0.25
W1C10L50	10^{-4}	0.10	0.50
W1C10L100	10^{-4}	0.10	1.00
W10C0	10^{-3}	–	–
W10C1L25	10^{-3}	0.01	0.25
W10C1L50	10^{-3}	0.01	0.50
W10C1L100	10^{-3}	0.01	1.00
W10C5L25	10^{-3}	0.05	0.25
W10C5L50	10^{-3}	0.05	0.50
W10C5L100	10^{-3}	0.05	1.00
W10C10L25	10^{-3}	0.10	0.25
W10C10L50	10^{-3}	0.10	0.50
W10C10L100	10^{-3}	0.10	1.00
W100C0	10^{-2}	–	–
W100C1L25	10^{-2}	0.01	0.25
W100C1L50	10^{-2}	0.01	0.50
W100C1L100	10^{-2}	0.01	1.00
W100C5L25	10^{-2}	0.05	0.25
W100C5L50	10^{-2}	0.05	0.50
W100C5L100	10^{-2}	0.05	1.00
W100C10L25	10^{-2}	0.10	0.25
W100C10L50	10^{-2}	0.10	0.50
W100C10L100	10^{-2}	0.10	1.00

distribution of the data. However, in our synthetic dataset, the relative frequencies of the binary classes and their within-class parameter distributions are the result of our choices and do not reflect the actual occurrence rates of the different binary configurations. In addition, we injected the synthetic data with the correlated and uncorrelated noise to systematically study the separation of the classes under various noise levels rather than to simulate the noise characteristics of real data, which are specific to the instrument and observing conditions.

Despite the synthetic data not being representative, we can still use it to quantify our ability to distinguish between the three binary classes. This can be achieved by constructing a discriminative low-dimensional representation of the data and investigating the separation of the classes in this representation. The idea is based on the observation that classification tasks often include a dimensionality reduction preprocessing step where the data is projected to a low-dimensional space in a way that preserves most of the information contained in the original high-dimensional data. In the absence of a representative training sample, we can separate dimensionality reduction from the classification task, allowing us to focus on the quality of the data representation rather than on the performance of the classifier. Apart from alleviating the effects of the curse of dimensionality and overfitting, dimensionality reduction also makes data easier to visualize and interpret, resulting in a more compact and informative representation. Once we have collected a representative sample, we can project the data to the learned low-dimensional space and train a classifier on the reduced data, ensuring robust generalization to new observations.

There are many dimensionality reduction methods, ranging in complexity from simple summary statistics and direct encodings to sophisticated latent representations learned directly from the data through optimization. In this work, we utilized PCA, a simple linear method that is easy to interpret and requires next to no hyperparameter tuning. By performing PCA separately on the three binary classes, we obtained three distinct latent representations of the synthetic data, each optimized to capture the underlying structure of the respective class (Sect. 5.3.1). As a baseline for comparison, we also expanded the light curves into Fourier coefficients, which is a standard practice in time series analysis (Sect. 5.3.2). We used two metrics to quantify the separation of the classes in the PCA and Fourier representations: the silhouette score, which compares the average intra-class distance to the average inter-class distance (Sect. 5.3.3), and the macro recall of random forest classifiers, which can be interpreted as a measure of the mean non-overlap of the classes in the feature space (Sect. 5.3.4).

5.3.1 PCA representations

PCA (Pearson 1901; Hotelling 1933) is an orthogonal affine transformation of the feature space – a Euclidean vector space with the dimensions corresponding to the data features and each data point represented as a vector in the space – to a new basis in which the features are uncorrelated. The transformation is achieved by centering and projecting the data in the directions of the unit eigenvectors (i.e., principal components) of the covariance matrix of the data, with the eigenvalues given by the variances of the data along the eigenvectors. When ordered by their eigenvalues, the eigenvectors maximize the projected variance in the orthogonal

complement of the preceeding components, resulting in an orthonormal basis in which the first principal component is oriented in the direction of the highest variance in the data, the second principal component points in the direction of the highest variance in the subspace perpendicular to the first component, and so on for higher-order components. By keeping only the principal components whose eigenvalues add up to a certain fraction of the total variance, we can reduce the dimensionality of the data while retaining most of the information contained in the original high-dimensional space.

Our motivation for using PCA in this work is multifold: (i) PCA is a well-tested, widely adopted method that is easy to use and computationally very efficient. (ii) PCA is a simple yet powerful enough method to gain intuition with dimensionality reduction, allowing us to illustrate the idea of using data-driven methods for feature extraction and classification before moving on to more sophisticated methods. (iii) With only one tunable hyperparameter, the number of retained principal components, PCA does not require extensive tuning, making it an ideal starting point in any dimensionality reduction task. (iv) The linear character of PCA representations makes them robust to noise, meaning that small perturbations in light curves do not significantly alter their representations. Consequently, we can perform PCA on noiseless light curves and then project noisy light curves using the obtained principal components, knowing that similar light curves will have similar representations, which is not necessarily the case with non-linear methods. (v) When interpreted in the original magnitude space, the principal components can be interpolated, yielding a set of continuous functions that are orthogonal under the standard L^2 inner product. In analogy with Fourier decomposition, these continuous principal components can then be used to generalize the PCA representation to light curves with arbitrary sampling.

In our analysis of the synthetic light curves, we utilized the `scikit-learn` implementation of PCA. The implementation returns the unit eigenvectors and the eigenvalues of the covariance matrix of the centered data, meaning that the mean is subtracted from each vector before the computation. Using the noiseless sample S0 as input, we performed PCA separately on the light curves of each binary class, yielding three distinct orthonormal bases of principal components. The synthetic light curves have a resolution of 100 points per orbit, with the last point removed for reasons related to the point (v) above, resulting in vectors of length $N_{\mathrm{grid}} = 99$. PCA preserves the dimensionality of the feature space, so the number of principal components in each basis is also N_{grid}.

We performed PCA on the normalized light curves instead of the original light curves in the magnitude space to ensure that the principal components reflect the intrinsic variations in the shapes of the light curves rather than the variations in amplitude. Since the amplitudes vary significantly within the classes, performing PCA on the original light curves would yield principal components that are dominated by amplitude, thereby obscuring the shape variations and resulting in suboptimal data representation. By factoring out the absolute scale before performing PCA and treating amplitude as a separate feature, we ensure that the morphology of the light curves is properly captured by the principal components, maximizing the overall information content of the representation.

If we denote the ith principal component of the class K as e_i^K, where K = DC, SD, C for the dark companion, semidetached, and contact binary

classes, respectively, we can expand any normalized light curve $\mathbf{v}$ evaluated on the same grid as the principal components as

$$\mathbf{v} = \mathrm{e}_0^K + \sum_{j=1}^{N_{\mathrm{grid}}} c_j^K \mathrm{e}_j^K, \tag{5.1}$$

where e_0^K is the mean normalized light curve of the class K and the vector of coefficients $\mathbf{c}_{N_{\mathrm{grid}}}^K = (c_1^K, c_2^K, \ldots, c_{N_{\mathrm{grid}}}^K)$ gives the coordinates of $\mathbf{v}$ in the PCA basis of the class K. If we consider the full vector of coefficients, the representation is perfect with no information loss. If we keep only the first $n < N_{\mathrm{grid}}$ coefficients, we can write the reconstruction of $\mathbf{v}$ as

$$\mathbf{v}_n^K = \mathrm{e}_0^K + \sum_{j=1}^{n} c_j^K \mathrm{e}_j^K, \tag{5.2}$$

with the reduced vector of coefficients $\mathbf{c}_n^K = (c_1^K, c_2^K, \ldots, c_n^K)$ constituting the n-dimensional PCA representation of the light curve in the basis of the class K. The superscript K on the left side of Eq. (5.2) emphasizes that the light curve reconstructed from the first n principal components is class-dependent, unlike the fully reconstructed light curve in Eq. (5.1).

The coefficients $\mathbf{c}_n^K$ can be obtained either by projecting the light curve onto the principal components or equivalently by fitting the light curve with a linear combination of the principal components using least squares. The equivalence of the two methods allows us to easily generalize the PCA representation to light curves with arbitrary sampling by interpolating the principal components in the normalized magnitude space and fitting the light curve with the interpolated principal components. By allowing arbitrary non-uniform sampling, we are no longer guaranteed the orthogonality of the principal components when evaluated on the new grid. Consequently, the PCA coefficients obtained from least squares fitting can change as we increase the number of the components in the fit. However, for densely sampled light curves, we expect the most informative coefficients to converge for $n \ll N_{\mathrm{grid}}$, yielding a representation that is robust to the light curve sampling. With a typical sampling frequency of the current space-based photometric surveys of the order of seconds to minutes, this condition is satisfied for a vast majority of light curves.

There is one issue with defining the PCA representation using principal components with a unit norm. If we doubled the resolution of the synthetic light curves and performed PCA on this finer grid, the photometric amplitude of the principal components would be approximately $\sqrt{2}$ times smaller than the amplitude of the original principal components. This is because the principal components are normalized to have a unit norm, and the finer grid contains twice as many points as the original grid. To avoid this issue, we fixed the scaling of the principal components to have a unit amplitude in the normalized magnitude space. This scaling ensures that the PCA coefficients are directly comparable between representations obtained from grids with different resolutions. Denoting the rescaled principal components as $\tilde{\mathrm{e}}_j^K$, we can write the projection of the light curve $\mathbf{v}$ onto the first n rescaled principal components of the class K as

$$\mathbf{v}_n^K = \mathrm{e}_0^K + \sum_{j=1}^{n} \tilde{c}_j^K \tilde{\mathrm{e}}_j^K, \tag{5.3}$$

where the vector of rescaled coefficients $\tilde{\mathbf{c}}_n^K = (\tilde{c}_1^K, \tilde{c}_2^K, \ldots, \tilde{c}_n^K)$ forms the n-dimensional PCA representation of $\mathbf{v}$ in the rescaled PCA basis of the class K. The generalization of the rescaled PCA representation to light curves with arbitrary sampling is achieved in the same way as in the case of the original unit PCA representation.

Both the unit and the rescaled PCA representations operate on normalized light curves, which are vertically shifted and rescaled so that the amplitude of the fourth-order Fourier fit is equal to unity. By normalizing the light curves, we lose the information about their absolute scaling. To recover this information, we prefix the vector of PCA coefficients in both representations with the amplitude obtained from the Fourier fit, increasing the dimensionality of the representations by one. We shall denote the amplitude as c_0 and the extended unit and rescaled PCA representations as capital $\mathbf{C}_n^K = (c_0, c_1^K, c_2^K, \ldots, c_n^K)$ and $\tilde{\mathbf{C}}_n^K = (c_0, \tilde{c}_1^K, \tilde{c}_2^K, \ldots, \tilde{c}_n^K)$, respectively. This way, the amplitude of the light curve is encoded as the zeroth element of the extended PCA representations, allowing us to rescale the normalized light curve back to the original magnitude space by multiplying the PCA coefficients and the mean light curve with c_0.

5.3.2 Fourier representation

Historically, expansion to Fourier coefficients has been the most popular method for dimensionality reduction of time series data. In discrete Fourier series, we decompose a uniformly sampled normalized light curve $\mathbf{v}$ of length N_{grid} into a linear combination of harmonics of increasing order up to the Nyquist frequency, totalling N_{grid} coefficients. For an odd N_{grid}, this can be expressed as

$$\mathbf{v} = a_0 \mathbb{1} + \sum_{j=1}^{(N_{\mathrm{grid}}-1)/2} a_j \mathbf{cos}_j + b_j \mathbf{sin}_j, \tag{5.4}$$

where $\mathbb{1}$ is a constant N_{grid}-dimensional vector of ones and $\mathbf{cos}_j$ and $\mathbf{sin}_j$ are vectors of the jth-order cosine and sine harmonics sampled on the same grid as the light curve. The base period of the harmonics is given by the period of the light curve, which is equal to one for phased light curves.

To emphasize the similarity between Fourier decomposition and PCA, we can factor out the mean light curve and express $\mathbf{v}$ as

$$\mathbf{v} = \mathbf{e}_0^K + \sum_{j=1}^{N_{\mathrm{grid}}} \tilde{c}_j^{\mathrm{F}} \tilde{\mathbf{e}}_j^{\mathrm{F}}, \tag{5.5}$$

where $\mathbf{e}_0^K$ is the mean normalized light curve of the class K and

$$\tilde{\mathbf{e}}_j^{\mathrm{F}} = \begin{cases} \mathbb{1} & \text{if } j = 1, \\ \mathbf{cos}_{j/2} & \text{if } j > 1 \text{ is even}, \\ \mathbf{sin}_{(j-1)/2} & \text{if } j > 1 \text{ is odd}. \end{cases} \tag{5.6}$$

We use the tilde notation from the previous section to emphasize the fixed scaling of the discretized Fourier basis elements. The mean light curve $\mathbf{e}_0^K$ can belong to any class K = DC, SD, C. In this work, we are mainly interested in searching for dark companion binaries, so we choose K = DC. By keeping only the first

n coefficients, we can reduce the dimensionality of the data while capturing the information about frequencies up to the harmonic preceded by the nth coefficient. We write the reconstruction of $\mathbf{v}$ using the first n coefficients as

$$\mathbf{v}_n^F = \mathbf{e}_0^{\mathrm{DC}} + \sum_{j=1}^{n} \tilde{c}_j^{\mathrm{F}} \tilde{\mathbf{e}}_j^{\mathrm{F}}. \tag{5.7}$$

Motivated by the analogy between Eqs. (5.7) and (5.2), we define the n-dimensional Fourier representation of $\mathbf{v}$ as the vector of coefficients $\tilde{\mathbf{c}}^F = (\tilde{c}_1^{\mathrm{F}}, \tilde{c}_2^{\mathrm{F}}, \ldots, \tilde{c}_n^{\mathrm{F}})$. Consequently, all the considerations from the previous section regarding the generalization of the PCA representations to light curves with arbitrary samplings apply to the Fourier representation as well.

In addition to the standard "rescaled" Fourier representation (the discretized harmonics have fixed scaling), we also define the unit Fourier representation $\mathbf{c}_n^{\mathrm{F}} = (c_1^{\mathrm{F}}, c_2^{\mathrm{F}}, \ldots, c_n^{\mathrm{F}})$, where the coefficients are obtained with respect to the normalized Fourier basis elements with unit norms. By analogy with the PCA representations, we define the extended Fourier representation as $\tilde{\mathbf{C}}_n^F = (c_0, \tilde{c}_1^F, \tilde{c}_2^F, \ldots, \tilde{c}_n^F)$, where c_0 is the amplitude of the light curve defined in the previous section. The extended unit Fourier representation is defined analogously as $\mathbf{C}_n^F = (c_0, c_1^F, c_2^F, \ldots, c_n^F)$.

Hereafter, we collectively refer to the PCA and Fourier representations as the *latent representations*, and we refer to the vector spaces spanned by the coefficients of the latent representations as *the latent spaces*. We further distinguish between rescaled and unit latent representations, which differ in the scaling of the basis vectors. The extended latent representations, be they rescaled or unit, include the amplitude of the light curves as the zeroth element, ensuring that the information about the absolute scale is preserved. We omit the lower index n in the notation when we refer to the latent representations in general, without reference to a specific dimension.

5.3.3 Silhouette score

There are various ways to assess the separation of clusters in a dataset, with the silhouette score being one of the most popular clustering measures (Rousseeuw 1987). The silhouette score quantifies how similar an object is to its own class compared to the other classes. The score ranges from -1 to 1, with higher values indicating better separation of the classes. The silhouette score of the ith object in the class K is calculated as

$$s_i^K = \frac{b_i^{K'} - a_i^K}{\max(a_i^K, b_i^{K'})}, \tag{5.8}$$

where a_i^K is the average Euclidean distance of the ith object to all other objects in the class K, and $b_i^{K'}$ is the average distance of the ith object to all objects in the closest neighboring class $K' \neq K$. In our analysis, we utilized the `scikit-learn` implementation of the silhouette score, namely the `silhouette_samples` function, which returns the silhouette score of each object in the dataset. We calculated the overall silhouette score of the dataset as the average of the individual silhouette scores weighted by the inverse of the respective class sizes, yielding a robust measure of clustering quality that is insensitive to class imbalance.

In general, higher values of the silhouette score indicate better separation of the classes in the dataset. However, the silhouette score is not invariant under independent rescaling of the features, meaning that the score can be artificially inflated by rescaling each feature with a different factor. Without a physically motivated scaling of the features, the absolute value of the silhouette score is not meaningful. Still, the silhouette scores of two competing representations can be directly compared to asses which representation separates the classes better, provided the bases of the representations are brought to the same (arbitrary) scale. The simplest way to achieve this is to normalize the basis vectors to have unit norms, which is the approach we followed in our analysis.

To compare the class separation in the PCA and Fourier representations, we calculated the weighted silhouette scores of the representations $\mathbf{c}_n^{\mathrm{DC}}$, $\mathbf{c}_n^{\mathrm{SD}}$, $\mathbf{c}_n^{\mathrm{C}}$, and $\mathbf{c}_n^{\mathrm{F}}$ as a function of the dimension of the representation $n = 1$–9. We evaluated the silhouette scores on the unit representations instead of the extended unit representations, because we want to maximize the class separation with respect to the shapes of the light curves independent of their amplitudes. The amplitude is a robust discriminative feature and, in the presence of strong noise, it can skew the silhouette scores of low-dimensional representations towards artificially high values, potentially obscuring the true number of coefficients that maximize the separation of the classes. Also, the amplitude affects the silhouette scores of all latent representations in roughly the same way, and since we are only interested in the difference between the silhouette scores of different representations and not their absolute values, we can safely omit the amplitude from the calculation. To provide a baseline, we also calculated the silhouette score of the full representation consisting of 99-dimensional vectors of normalized light curves in the magnitude space. We performed the calculation on the validation sets of the samples W0C0, W100C0, W0C10L50, and W100C10L50, which are the synthetic samples with the lowest and the highest levels of uncorrelated and correlated noise, either separately or in combination. Hereafter, we shall refer to these samples as the corner cases. The corner cases provide the most extreme conditions for the separation of the classes, and the conclusions drawn from the them are generally applicable to the intermediate cases as well.

5.3.4 Macro recall and random forest classifiers

There are several downsides to using the silhouette score as a measure of class separation in the latent space. First, due to the silhouette score not being invariant under independent rescaling of the features, the absolute value of the silhouette score is meaningless, only the difference between the silhouette scores of different representations is informative. Second, we calculated the silhouette scores as a function of the number of coefficients in the representation. However, not all coefficients are equally informative, meaning that the first n coefficients can yield a lower silhouette score that the same number of non-consecutive but more informative coefficients. Third, the silhouette score as a measure of separation is best suited for convex clusters, which is not necessarily the case for the binary classes in the latent representations, not to mention the full representation. For concave overlapping and/or nested clusters, the silhouette score can be close to zero even if the clusters are perfectly separated. For these reasons, we turn to a

more robust measure of class separation: the macro recall.

In a classification task, the recall R_K of a classifier for the class K is defined as

$$R_K = \frac{TP_K}{TP_K + FN_K}, \tag{5.9}$$

where TP_K is the number of class-K true positives (correctly predicted objects in the class K) and FN_K is the number of class-K false negatives (objects in the class K that were incorrectly predicted as belonging to a different class). The macro recall R_M of the classifier is calculated as a simple arithmetic average of the recalls of the N_classes individual classes,

$$R_\mathrm{M} = \frac{1}{N_\mathrm{classes}} \sum_{K=1}^{N_\mathrm{classes}} R_K. \tag{5.10}$$

The benefit of the macro recall is that it is not sensitive to the class sizes, which makes it an ideal measure of class separation for imbalanced datasets or datasets with unknown class frequencies, such as our synthetic data. The macro recall should not be confused with the accuracy A of the classifier, which is calculated as

$$A = \frac{\sum_{K=1}^{N_\mathrm{classes}} TP_K}{\sum_{K=1}^{N_\mathrm{classes}} (TP_K + FN_K)} = \sum_{K=1}^{N_\mathrm{classes}} f_K R_K, \tag{5.11}$$

where f_K are the relative frequencies of the classes in the dataset. By comparing Eqs. (5.11) and (5.10), we see that the macro recall coincides with the accuracy only if the classes are balanced, i.e., if $f_K = 1/N_\mathrm{classes}$ for all K. In the case of imbalanced classes, the accuracy is skewed towards the recall of the majority class, while the macro recall treats all classes with equal importance. However, if we train the classifier directly on the imbalanced data, the macro recall can also become skewed, provided the classes are not well-separated in the latent space. For the macro recall to be a robust measure of the mean non-overlap of the classes in the latent space, we need to artificially balance the data by weighing the samples with the inverse of their class sizes prior to training the classifier. This way, we can ensure that the classifier is not biased towards the majority class and the class contours in the latent space are not affected by the class sizes.

The macro recall is specific to the classifier, which means that it can change when we train a different classifier on the same data or when we use the same classifier with different hyperparameters. Choosing the optimal classifier that yields the best macro recall is a non-trivial task that requires hyperparameter tuning and cross-validation to select the best performing model. Without any prior knowledge of the problem, the best approach is to start with a simple and robust classifier that does not require excessive tuning, such as the random forest classifier (Breiman 2001), which is known to perform well on a wide range of problems and is not too sensitive to the choice of hyperparameters.

Random forests operate by constructing a large number of decision trees that are trained on random subsets of the data and features. The final prediction is made by taking the average of the individual tree predictions, making the method robust to overfitting and noise. In our analysis, we used the `scikit-learn` implementation of the random forest classifier. For each synthetic sample (Table 5.1), we trained a number of random forest classifiers with different

hyperparameter configurations on the extended rescaled representations $\tilde{\mathbf{C}}_n^K$, K = DC, SD, C, F and n = 1–9. To provide a baseline, we also considered the $(1 + 99)$-dimensional extended full representation consisting of photometric amplitudes + normalized light curves, and the one-dimensional representation consisting of photometric amplitudes only. In addition, we augmented each representation except the extended full representation with the variances of the coefficients obtained from the least squares fits of the photometric amplitude and the latent coefficients, and we retrained the random forest classifiers on the augmented representations. We did this to examine whether the uncertainties of the coefficients contain useful information that could help us better separate the classes. We calculated the uncertainty of the photometric amplitude as the variance of the residuals from the fourth-order Fourier fit of the light curve. Given an extended rescaled representation $\tilde{\mathbf{C}}_n^K$, we denote its augmented version as $\tilde{\mathbf{C}}_n^{K+\mathrm{V}}$, where K = DC, SD, C, F, and n is the number of coefficients in the representation.

We trained the random forest classifiers on the extended rescaled representations to ensure that absolute scale of the light curves is taken into account when separating the classes. We did not include the amplitude in the calculation of the silhouette scores, because it could bias the optimal number of coefficients towards lower values in the presence of strong noise, but the properties of the random forest classifier make it possible to include the amplitude in the input without obscuring the discriminative patterns in the coefficients of the representations. In addition, the macro recall is an absolute measure of class non-overlap, which means that we are actually interested in its values, not just its differences between different representations. To achieve the best possible macro recall, it is necessary we consider all the information contained in the light curves, including the amplitude.

We obtained the optimal hyperparameters of the random forest classifiers trained on each representation of each synthetic sample by performing a basic grid search for selected hyperparameters. Namely, we considered: the number of trees in the forest `n_estimators` = 100, 500; the minimum number of samples required to be at a leaf node `min_samples_leaf` = 1, 10; and the method for selecting the number of features at each split `max_features` = `sqrt`, `log2`, `None`. We used the default values for the remaining hyperparameters. In all cases, we trained the random forest classifiers on the training sets, performed the hyperparameter tuning on the validation sets, and evaluated the best performing classifiers on the test sets of the synthetic samples.

5.4 Results

In this section, we present the results of our analysis of the synthetic light curves of dark companion, semidetached, and contact binary systems. In Sect. 5.4.1, we provide an overview of the PCA models of the three binary classes. We visually inspect the PCA and Fourier representations of the light curves in Sect. 5.4.2, and we compare the informativeness of the representations using the silhouette score in Sect. 5.4.3. In Sect. 5.4.4, we quantify the separation of the classes in the latent representations using the macro recall of random forest classifiers trained on the representations, and we assess the impact of the coefficient variances on the macro

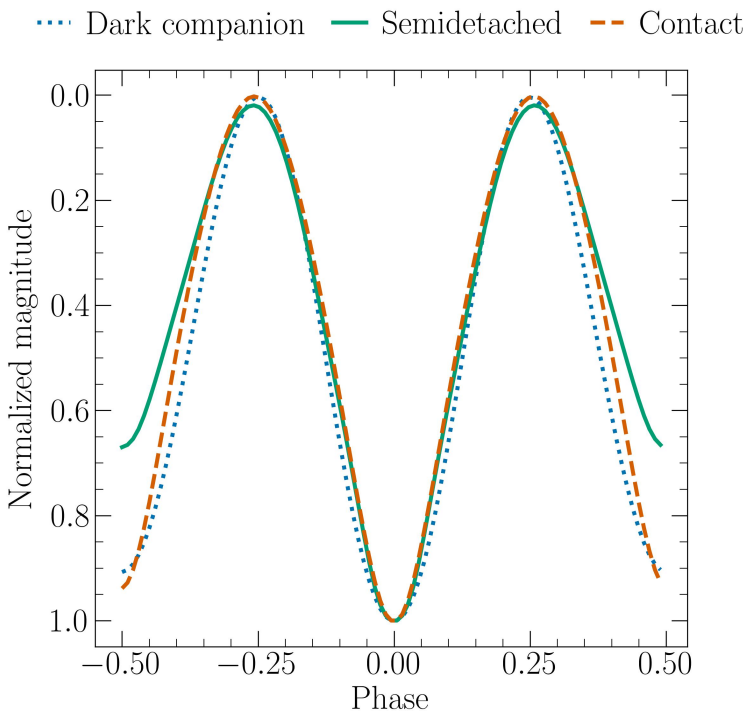

Figure 5.1 Mean normalized light curves of dark companion, semidetached, and contact binary systems. The mean light curves are subtracted from the light curves of the corresponding class before performing PCA.

recall in Sect. 5.4.5. Finally, is Sect. 5.4.6, we obtain the expected precision of the random forest classifiers on previously unseen data.

5.4.1 PCA models of synthetic light curves

We show the mean normalized light curves of dark companion, semidetached, and contact binary systems in Fig. 5.1. The light curves exhibit a remarkable similarity, particularly between the mean dark companion and contact binary light curves, highlighting the importance of the finer details captured by principal components in distinguishing between the classes. In Fig. 5.2, we show the first nine rescaled principal components of the three binary classes. The principal components have a unit amplitude and are unique up to a sign change. Due to the normalization of the synthetic light curves, all components coincide at phase 0, corresponding to the primary minimum of the light curves. Compared to the mean light curves, we observe significant differences between the classes already in the first few principal components, with the differences becoming more pronounced as we move away from the primary minimum towards the secondary minimum at phases -0.5 and 0.5. Higher-order principal components are progressively more oscillatory, making it harder to interpret the differences between the classes. Still, we observe that the dark companion components are generally more similar to the contact components than to the semidetached components, revealing increased levels of degeneracy between these two classes.

Starting with the fifth component, the dark companion components become increasingly affected by numerical noise, basically becoming pure noise by the

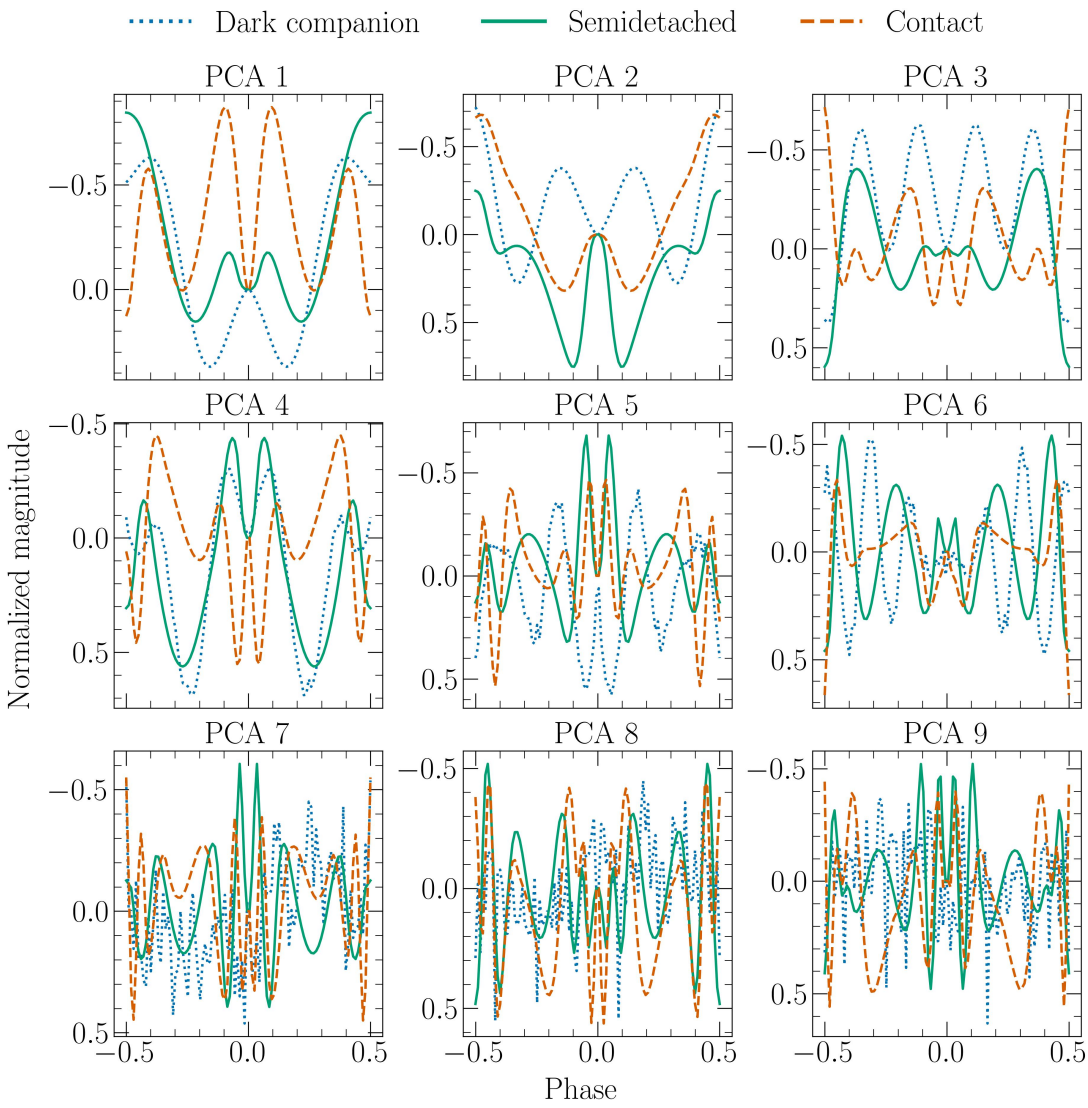

Figure 5.2 First nine principal components of dark companion, semidetached, and contact binary light curves, ordered by explained variance in descending order.

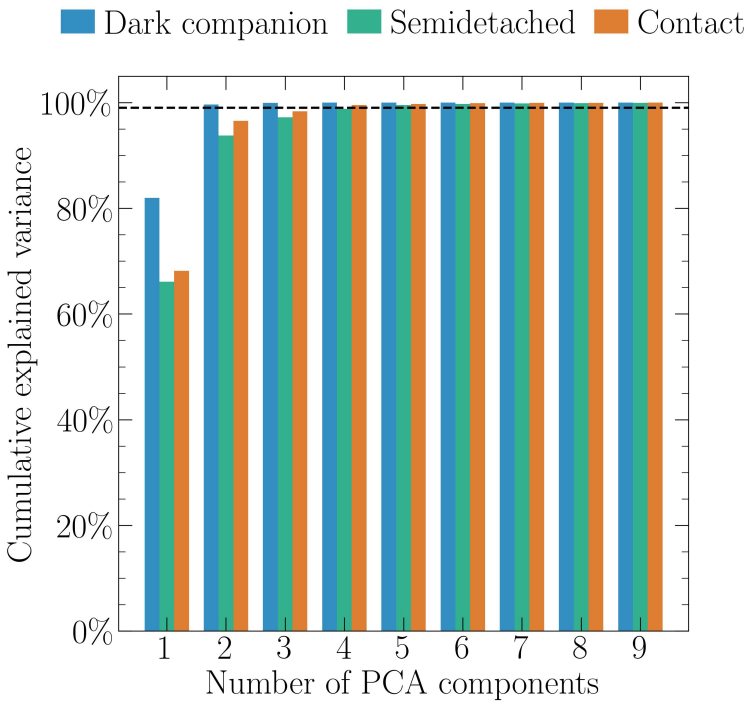

Figure 5.3 Cumulative explained variance of the principal components of dark companion, semidetached, and contact binary light curves. The dashed line indicates the threshold of 99% explained variance.

ninth component. Consequently, the first four to five principal components capture virtually all the variance in the dark companion binary light curves. This is not the case for the semidetached and contact binary light curves, which require more components to capture the same level of variance. The disparity in the informativeness of the principal components can be also seen in Fig. 5.3, which shows the cumulative explained variance as a function of the number of retained PCA components for the three binary classes. The dark companion binary light curves require only two principal components to explain more than 99% of the variance, while the contact and semidetached binary light curves require four and five components, respectively, to cross the 99% threshold.

5.4.2 Latent representations of synthetic light curves

Utilizing the PCA models and the discretized Fourier basis, we constructed the representations $\tilde{\mathbf{c}}_3^{\mathrm{DC}}$, $\tilde{\mathbf{c}}_3^{\mathrm{SD}}$, $\tilde{\mathbf{c}}_3^{\mathrm{C}}$, and $\tilde{\mathbf{c}}_3^{\mathrm{F}}$ of the validation sets of the four corner cases: samples W0C0, W100C0, W0C10L50, and W100C10L50. In Fig. 5.4, we show the scatter plots of the first and third coefficients of the latent representations for these samples. Panel (a) illustrates the rich structure of the PCA representations of the noiseless sample W0C0, where the dark companion and contact binaries form relatively well-separated clusters, while the semidetached binaries are scatter all over the latent space. In contrast, the Fourier representation is collapsed along the third coefficient and does not show a clear separation between the classes.

Panel (b) of Fig. 5.4 displays the latent representations of the sample W100C0, which was injected with uncorrelated Gaussian noise at $\sigma_{\mathrm{WN}} = 0.01\,\mathrm{mag}$. The light curves from this sample exhibit much greater scatter in the latent spaces

compared to the noiseless light curves, making it more challenging to differentiate between the classes. Despite this, the PCA representations still retain some of the original structure, especially in the cases of the representations $\tilde{\mathbf{c}}_3^{\mathrm{DC}}$ and $\tilde{\mathbf{c}}_3^{\mathrm{C}}$. In the Fourier representation, the classes are almost completely mixed, with the exception of the semidetached class, which protrudes from the main intermixed cluster. The situation becomes even worse when we introduce correlated noise, intended to simulate the effects of surface spots. In panel (c) of Fig. 5.4, we present the latent representations of the sample W0C10L50, which was injected with correlated Gaussian noise at $\sigma_{\mathrm{CN}} = 0.1$ and $l_{\mathrm{CN}} = 0.5$. The Fourier representation of the light curves is practically featureless, with the classes clumped together in a single cluster. The representation $\tilde{\mathbf{c}}_3^{\mathrm{SD}}$ is slightly more informative, but the classes are still thoroughly mixed. We observe the best separation of the classes in the representations $\tilde{\mathbf{c}}_3^{\mathrm{DC}}$ and $\tilde{\mathbf{c}}_3^{\mathrm{C}}$, but the separation is still far from ideal, with most of the structure present in the sample W0C0 lost due to the correlated noise.

Panels (b) and (c) of Fig. 5.4 demonstrate the independent effects of uncorrelated and correlated noise on the representations of the synthetic light curves. In panel (d), we show the latent representations of the sample W100C10L50, which incorporates the combined noise from the samples W100C0 and W0C10L50. The cumulative effect of the two types of noise is remarkably similar to the effect of correlated noise alone, with the classes only slightly more mixed in all representations. We conclude that correlated noise, such as the one arising from surface spots, affects the structure of the latent representations more severely than uncorrelated noise, disrupting the patterns present in the absence of noise and effectively mixing the classes.

Although our discussion is based on the visual inspection of the first and third coefficients of the latent representations, the projections to the remaining coefficients yield similar results (Figs. 5.A.1–5.A.2). The only difference is that for some projections, the representation $\tilde{\mathbf{c}}^{\mathrm{SD}}$ seems to be the most informative, while for others, $\tilde{\mathbf{c}}^{\mathrm{DC}}$ or $\tilde{\mathbf{c}}^{\mathrm{C}}$ separate the classes better. In all projections, the Fourier representation $\tilde{\mathbf{c}}^{\mathrm{F}}$ yields worse visual separation than the most informative PCA representation, demonstrating the superiority of the PCA representations in capturing the latent structure of the synthetic light curves.

5.4.3 Silhouette scores

While visual inspection of the latent representations can provide qualitative insight into the separation of the dark companion, semidetached, and contact binary light curves, the silhouette score allows for a more quantitative and systematic approach to assessing the separation of the classes in the different representations. In Fig. 5.5, we show the silhouette scores for the unit representations $\mathbf{c}_n^{\mathrm{DC}}$, $\mathbf{c}_n^{\mathrm{SD}}$, $\mathbf{c}_n^{\mathrm{C}}$, and $\mathbf{c}_n^{\mathrm{F}}$ of the validation sets of the corner cases, evaluated as a function of the number of coefficients in the representation $n = 1$–9. The solid black lines represent the silhouette scores for the full representations of the samples, which vary from approximately 0.085 in the case of the sample W0C0 (Fig. 5.5a) to about -0.04 for the sample W100C10L50 (Fig. 5.5d). Taken at face value, the silhouette scores of the full representations are low, indicating that the classes are not well separated in the original high-dimensional space. However, as we discuss in Sect. 5.3.3, the absolute value of the silhouette score

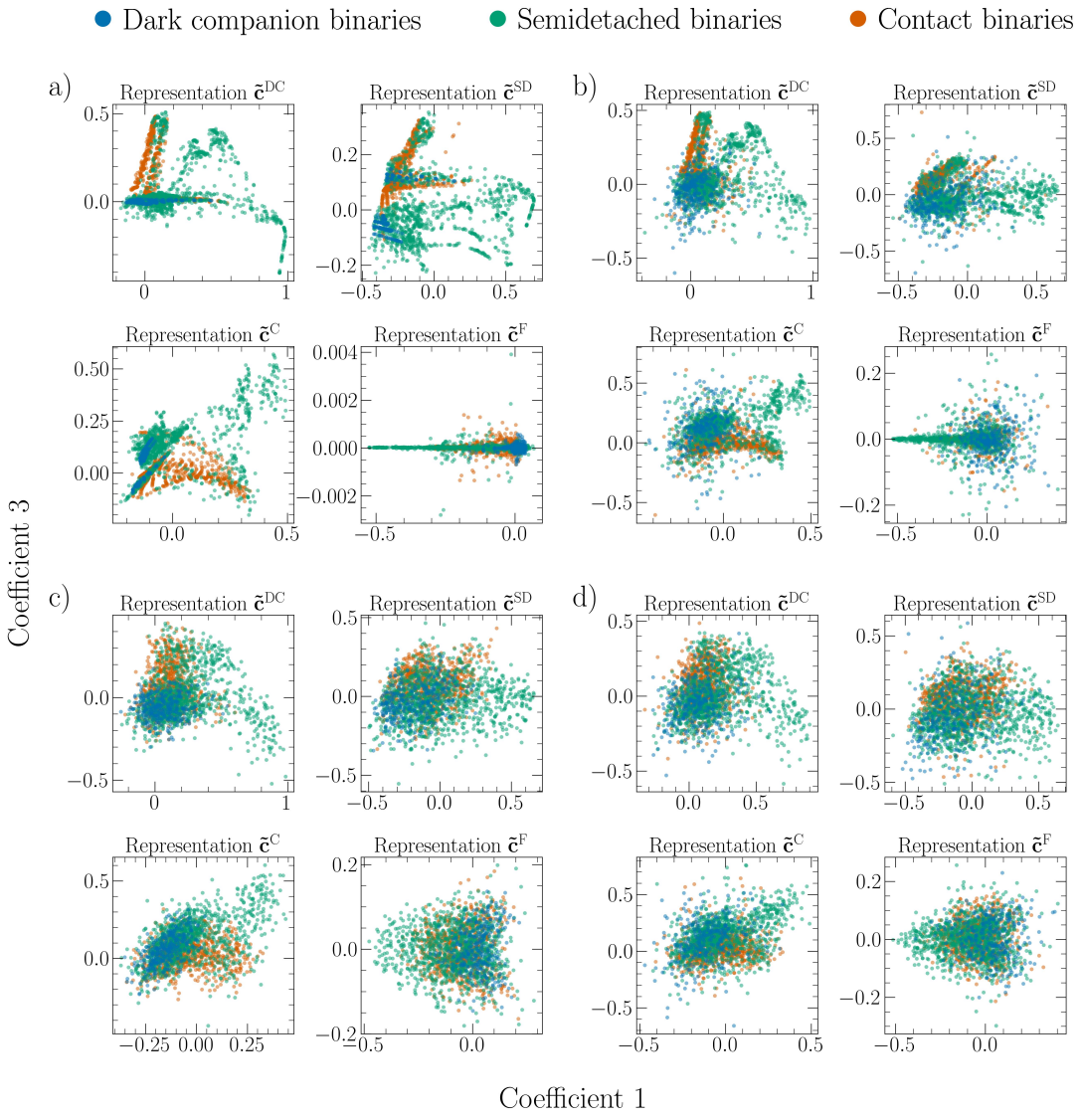

Figure 5.4 Scatter plots of the first and third coefficients of the representations $\tilde{\mathbf{c}}^{\mathrm{DC}}$, $\tilde{\mathbf{c}}^{\mathrm{SD}}$, $\tilde{\mathbf{c}}^{\mathrm{C}}$, and $\tilde{\mathbf{c}}^{\mathrm{F}}$ of the dark companion, semidetached, and contact binary light curves in the validation sets of the synthetic samples W0C0 (a), W100C0 (b), W0C10L50 (c), and W100C10L50 (d). We describe the synthetic samples in Sect. 5.2, and we provide the definitions of the representations in Sects. 5.3.1–5.3.2.

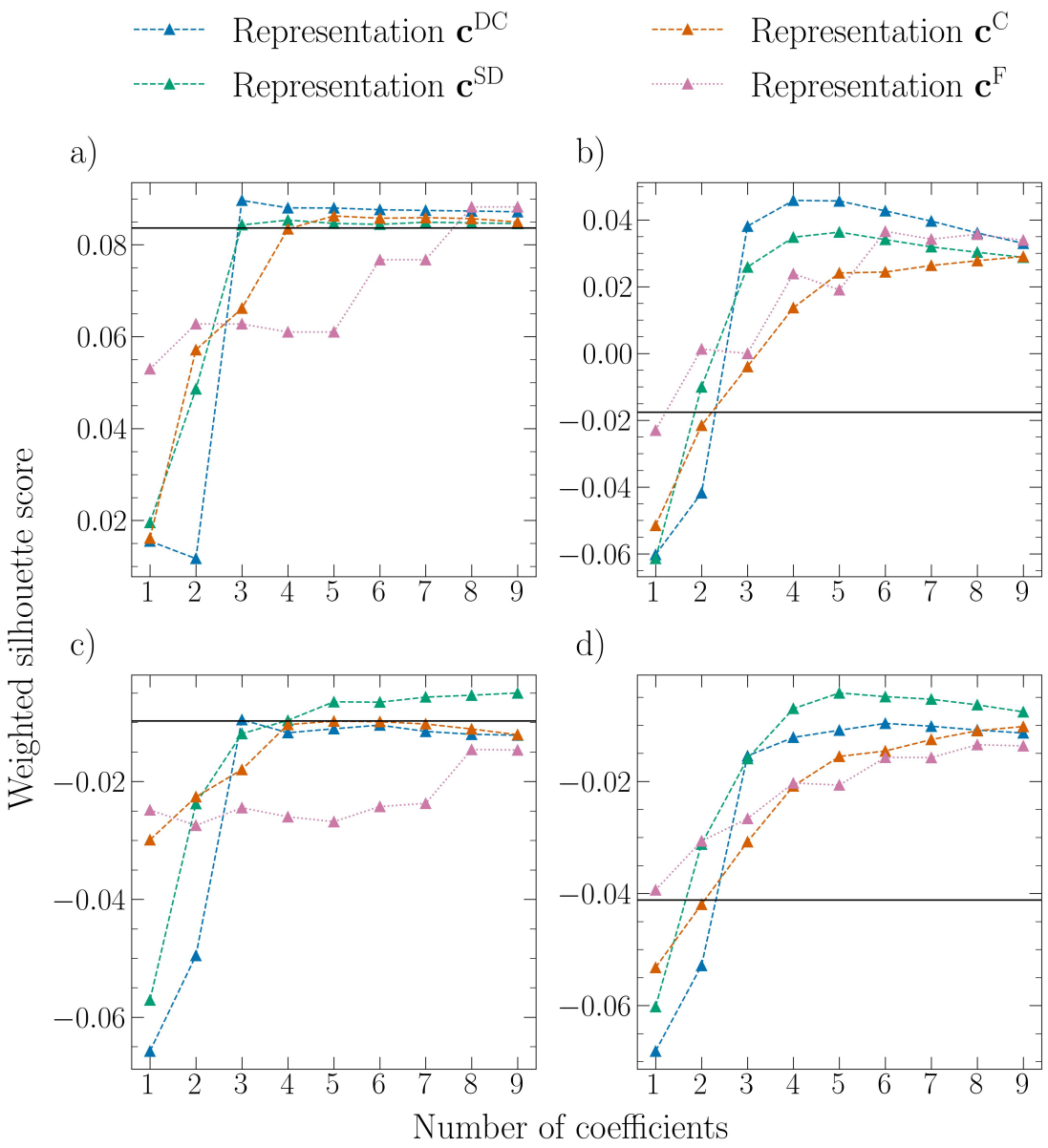

Figure 5.5 Weighted silhouette scores for the representations $\mathbf{c}^{\mathrm{DC}}$, $\mathbf{c}^{\mathrm{SD}}$, $\mathbf{c}^{\mathrm{C}}$, and $\mathbf{c}^{\mathrm{F}}$ of the validation sets of the synthetic samples W0C0 (a), W100C0 (b), W0C10L50 (c), and W100C10L50 (d), evaluated as a function of the number of coefficients in the representation. The solid black lines represent the weighted silhouette scores calculated for the full normalized light curves in the validation sets. We describe the synthetic samples in Sect. 5.2, and we define the representations in Sects. 5.3.1–5.3.2.

is not important, only the difference between the silhouette scores of different representations is informative. Thus, we use the silhouette score of the full representation as a benchmark against which we compare the silhouette scores of the latent representations to see whether projection to a lower-dimensional space can improve the separation of the classes.

Due to the orthogonal character of the unit PCA and Fourier representations, the silhouette scores of the representations converge to the silhouette score of the full representation as the number of coefficients n goes to N_{grid}. The reason is that the distance in the definition of the silhouette score is calculated using the dot product of the difference vector with itself, which is preserved under orthogonal transformations. In the case of the synthetic samples W0C0 and W0C10L50 (Fig. 5.5a and c), the silhouette scores of the PCA representations quickly plateau slightly above or below the benchmark limit value, with the exception of the representation $\mathbf{c}^{\text{SD}}$, whose silhouette score for the sample W0C10L50 is still increasing at $n = 9$. Most likely, the silhouette score of $\mathbf{c}^{\text{SD}}$ reaches maximum at $n > 9$, but we did not investigate this further. The situation is different for the samples W100C0 and W100C10L50 (Fig. 5.5b and d), where the silhouette scores of all latent representations peak high above the benchmark score and then gradually start to decrease towards the limit value, or they continue to increase up to $n = 9$, reaching maximum at $n \geq 9$ (representation $\mathbf{c}^{\text{C}}$). In all four corner cases, the silhouette score of the representation $\mathbf{c}^{\text{F}}$ follows similar trends to the silhouette scores of the PCA representations, but it requires more coefficients to reach the same levels of class separation, with the exception of the first one to two coefficients, which seem to be more informative than the PCA coefficients.

The general trends in the silhouette scores of the latent representations under different noise conditions can be explained by the properties of the injected noise or the lack thereof. In the absence of noise (Fig. 5.5a), the PCA representations require only a few coefficients to almost perfectly reconstruct the light curves, leaving only negligible unexplained variance to be captured by higher-order coefficients. Consequently, the higher-order coefficients are close to zero and do not significantly contribute to the silhouette score, which explains the quick plateauing of the silhouette scores. In other words, the first three to four coefficients, depending on the PCA representation, capture effectively all the information that is present in the full light curves and account for most of the separation between the classes. Conversely, the representation $\mathbf{c}^{F}$ requires at least eight coefficients to reach the plateau, pointing to the poor alignment of the Fourier basis with the data.

The power spectrum of a light curve injected with noise is the sum of the power spectrum of the signal, the power spectrum of the noise, and an additional interference term arising from the interaction of the signal and the noise. While the power spectrum of the signal is skewed towards low frequencies, the power spectrum of uncorrelated noise is flat, which means that, in relative terms, low-order coefficients of the PCA representations are less affected by the noise than higher-order coefficients. Given that most of the signal is contained in the first four to five coefficients, we are able to extract useful information from light curves even in the presence of strong uncorrelated noise. This can be seen in Fig. 5.5b, where the silhouette scores of the the representations $\mathbf{c}^{\text{DC}}$ and $\mathbf{c}^{\text{SD}}$ peak at $n = 4$ and 5, respectively, and then start to decrease as we keep adding coefficients

that are increasingly more affected by the noise. Compared to the case with no noise, the contribution of the higher-order coefficients to the silhouette score is not negligible, because they capture the high-frequency noise that is absent in the former case. Given that the high-frequency noise affects all classes equally, the classes become gradually more mixed together as we increase n, resulting in the low benchmark silhouette score of the full representation. The silhouette scores of the representations $\mathbf{c}^{\mathrm{C}}$ and $\mathbf{c}^{\mathrm{F}}$ follow a similar trend, but they require more coefficients to reach the same levels of class separation as $\mathbf{c}^{\mathrm{DC}}$ and $\mathbf{c}^{\mathrm{SD}}$, revealing a suboptimal alignment of their bases with the data. While the silhouette score of $\mathbf{c}^{\mathrm{F}}$ peaks at $n = 6$, with its value below the maximum achieved by $\mathbf{c}^{\mathrm{DC}}$, the silhouette score of $\mathbf{c}^{\mathrm{C}}$ is still increasing at $n = 9$, possibly attaining global maximum at $n > 9$. However, this is not likely, considering the negligible explained variance of the higher-order coefficients.

The situation is different when we inject the light curves with strong correlated noise, where the power spectra of both the signal and the noise are skewed towards low frequencies (Fig. 5.5c). In this case, the noise effectively masks the signal in the low-order coefficients of the PCA representations, preventing us from recovering the original signal. Not being able to distinguish between the signal and the noise, the representations treat the noise as part of the signal, resulting in a scenario analogous to the noiseless case. The first few PCA coefficients are enough to capture virtually all variance in the light curves, pushing higher-order coefficients to zero. Consequently, the silhouette scores of the representations $\mathbf{c}^{\mathrm{DC}}$ and $\mathbf{c}^{\mathrm{C}}$ quickly settle slightly below the benchmark score, whose value itself is significantly lower than in the noiseless case. The Fourier representation $\mathbf{c}^{\mathrm{F}}$ is affected even more strongly, with its silhouette score approaching the benchmark from well below and requiring more than nine coefficients to reach the benchmark, if at all. The only representation that decidedly outperforms the benchmark is $\mathbf{c}^{\mathrm{SD}}$, whose silhouette score is still increasing at $n = 9$, demonstrating the resilience of the semidetached representation to correlated noise.

The combined effect of strong correlated and uncorrelated noise is to decrease the separation between the classes even further, pushing the benchmark score to clearly negative values. (Fig. 5.5d). Despite the low benchmark score, the silhouette scores of the latent representations are comparable to the values obtained for the sample W0C10L50, revealing that in the presence of both types of noise i) the latent representations still manage to disentangle low-frequency signal from high-frequency noise and ii) the level of the correlated noise is the decisive factor in determining the structure of the latent representations and the separation of the classes. The latter is consistent with our visual inspection of the corner cases in the previous section, where the latent representations of the samples W0C10L50 and W100C10L50 exhibited similar features.

Based on our analysis of the silhouette scores, the representation $\mathbf{c}^{\mathrm{SD}}$ seems to be the most flexible and robust to noise, achieving the best separation of the classes for the samples W0C10L50 and W100C10L50 and almost matching the best performing representation for the samples W0C0 and W100C0. Conversely, the representation $\mathbf{c}^{\mathrm{C}}$ seems to be the least informative of the PCA representations, underperforming in the presence of strong uncorrelated noise, but yielding comparable silhouette scores to $\mathbf{c}^{\mathrm{DC}}$ for the samples W0C0 and W100C10L50. For $n > 2$, the representation $\mathbf{c}^{\mathrm{F}}$ generally achieves worse silhouette scores than

the best performing PCA representation for the same number of coefficients, possibly with the exception of the samples W0C0 and W100C0, where it performs comparably to $\mathbf{c}^{\mathrm{DC}}$ at $n = 8$–9. Still, $\mathbf{c}^{\mathrm{F}}$ does not achieve the maximum silhouette score in any of the corner cases, which is consistent with our visual inspection of the latent representations in the previous section, where the Fourier representation consistently yielded worse class separation than the most discriminative PCA representation.

5.4.4 Macro recalls and random forest hyperparameters

The silhouette score is useful for comparing the separation of the dark companion, semidetached, and contact binary classes across different latent representations, but it does not provide an absolute measure of class separation. To address this, we trained random forest classifiers on the extended representations $\tilde{\mathbf{C}}_n^{\mathrm{DC}}$, $\tilde{\mathbf{C}}_n^{\mathrm{SD}}$, $\tilde{\mathbf{C}}_n^{\mathrm{C}}$, and $\tilde{\mathbf{C}}_n^{\mathrm{F}}$ of the training sets of all synthetic samples (Table 5.1) for $n = 1$–9, and we conducted a basic hyperparameter search on the validation sets of the samples to find the configurations that yield the best macro recalls. In Fig. 5.6, we present the obtained validation macro recalls of the random forest classifiers as a function of n for the four corner cases. The errorbars show the minimum and the maximum validation macro recalls achieved by the classifiers during the hyperparameter tuning. The solid black lines represent the best macro recalls achieved by the classifiers trained on the extended full representations. We also trained random forest classifiers on the one-dimensional representations of the samples, but we do not show the results in the plots, because they make them less readable. The best validation macro recalls achieved by the classifiers trained on the one-dimensional representations are: $R_{\mathrm{M}}^{\mathrm{V}} = 0.46$ for the noiseless sample W0C0 (Fig. 5.6a) and $R_{\mathrm{M}}^{\mathrm{V}} = 0.55$ for the remaining corner cases (Fig. 5.6b–d).

In all corner cases, the macro recalls of the PCA representations quickly reach a saturation point somewhere between $n = 3$ and 6, and then they level off with a slight increase or decrease. We observed similar trends in the silhouette scores of the PCA representations, but the saturation points of the macro recalls are generally shifted towards higher n compared to the peaks and plateaus in the silhouette scores. The representation $\tilde{\mathbf{C}}^{\mathrm{DC}}$ of the samples W0C0 and W0C10L50 exhibits the most pronounced shift, with the macro recall saturation points at $n = 5$–6 and the silhouette score plateaus at $n = 3$. Another difference between the silhouette scores and the macro recalls of random forest classifiers is that the latter are not as sensitive to the presence of uninformative features. This can be seen by comparing the trends in the silhouette scores and the macro recalls for the PCA representations of the samples W100C0 and W100C10L50. The silhouette scores of the representations $\mathbf{c}^{\mathrm{DC}}$ and $\mathbf{c}^{\mathrm{SD}}$ peak at $n = 4$–6 and then start to decrease, eventually reaching the benchmark silhouette score well below the maximum at $n = N_{\mathrm{grid}}$ (Figs. 5.5b and d), while the macro recalls of the representations $\tilde{\mathbf{C}}^{\mathrm{DC}}$ and $\tilde{\mathbf{C}}^{\mathrm{SD}}$ take the same number of coefficients to reach a plateau located slightly above or below the limit value of the best macro recall achieved for the extended full representation, allowing only for a marginal increase or decrease with additional coefficients (Figs. 5.6b and d).

Both the shift of the saturation points towards higher n and the generally non-decreasing trends in the macro recalls of the PCA representations with increasing

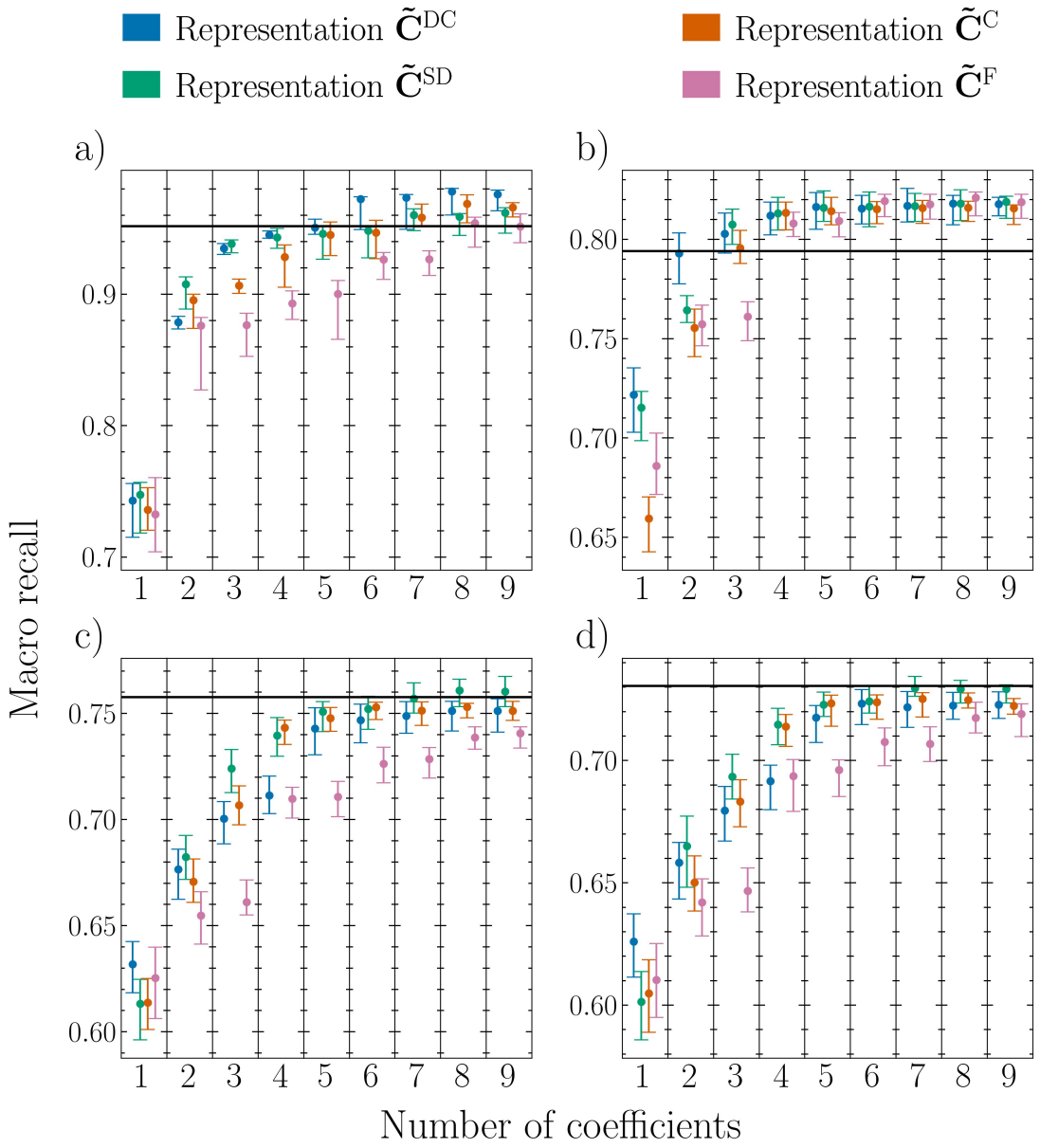

Figure 5.6 Validation macro recalls of the random forest classifiers trained on the representations $\tilde{\mathbf{C}}^{\mathrm{DC}}$, $\tilde{\mathbf{C}}^{\mathrm{SD}}$, $\tilde{\mathbf{C}}^{\mathrm{C}}$, and $\tilde{\mathbf{C}}^{\mathrm{F}}$ of the synthetic samples W0C0 (a), W100C0 (b), W0C10L50 (c), and W100C10L50 (d), evaluated as a function of the number of coefficients in the representation. The errorbars show the minimum and the maximum validation macro recalls for a given number of coefficients. The solid black lines represent the best validation macro recalls achieved by the classifiers trained on the extended full representations of the samples. We describe the synthetic samples in Sect. 5.2, and we provide the definitions of the representations in Sects. 5.3.1–5.3.2.

n can be explained by the nonlinear nature of the random forest classifier and its robustness to noise and overfitting. The nonlinear nature allows the classifier to learn complex decision boundaries that are not necessarily convex nor connected, making it possible to extract useful information even from coefficients that decrease the silhouette score of the representation. The robustness to noise and overfitting means that even if we include an uninformative feature, the classifier can learn to ignore it and focus on the informative features, leaving the macro recall

essentially unchanged. These arguments also extend to the macro recall of the representation $\tilde{\mathbf{C}}^{\mathrm{F}}$, which follows the same non-decreasing trend with increasing n, but consistently requires more coefficients to reach the same levels as the macro recalls of the PCA representations. The only exception is the sample W100C0, where $\tilde{\mathbf{C}}^{\mathrm{F}}$ performs comparably to the PCA representations. Unlike the PCA representations, $\tilde{\mathbf{C}}^{\mathrm{F}}$ does not show a significant shift in the macro recall trends towards higher n compared to the trends in the silhouette score, indicating that the silhouette score is relatively well-aligned with how the data is organized in the Fourier latent space.

In Table 5.2, we present the optimal representations and random forest hyperparameters that yielded the best validation macro recalls for each synthetic sample. We also present the class and macro recalls of the best performing classifiers evaluated on the validation and test sets of the synthetic samples, demonstrating the generalization of the recalls to previously unseen data. The optimal number of trees in the forest is `n_estimators` = `500` in the majority of cases, but for some samples, `n_estimators` = `100` yields better results. We do not observe a clear pattern between the optimal value of `n_estimators` and the level or type of noise, indicating that the random forest classifier is robust to the choice of this hyperparameter in the context of the synthetic samples. The optimal minimum number of samples at a leaf node alternates between `min_samples_leaf` = `1` and `10`, but in the presence of moderately strong to strong uncorrelated noise ($\sigma_{\mathrm{WN}} \gtrsim 10^{-3}$ mag) and/or correlated noise ($\sigma_{\mathrm{CN}} \gtrsim 0.05$ and $l_{\mathrm{CN}} \leq 0.5$), the optimal value is preferentially `min_samples_leaf` = `10`, reducing the risk of overfitting. In the majority of noise conditions, the optimal method for selecting the number of features at each split is `max_features` = `sqrt`, but in the limit of strong uncorrelated and correlated noise, `max_features` = `log2` is also a valid choice. Overall, the setup with `n_estimators` = `500`, `min_samples_leaf` = `10`, and `max_features` = `sqrt` seems to be the most robust, yielding the best macro recalls for 11 out of 40 synthetic samples across a wide range of noise conditions.

To visualize the optimal representations of the synthetic samples, we present Fig. 5.7. The figure shows the representations that yielded the best validation macro recalls for each synthetic sample, along with the corresponding dimensions at which these recalls were attained. The samples are color-coded according to the highest macro recall achieved on their validation sets. In most cases, the best macro recalls are achieved for the representation $\tilde{\mathbf{C}}^{\mathrm{SD}}$. The dominance of $\tilde{\mathbf{C}}^{\mathrm{SD}}$ is most pronounced in the regime of strong correlated noise, where it consistently outperforms the other representations. This is in agreement with what we observed in our analysis of the silhouette scores of the latent representations, where the representation $\mathbf{c}^{\mathrm{SD}}$ yielded the highest silhouette scores for the samples W0C10L50 and W100C10L50. The superior performance of $\tilde{\mathbf{C}}^{\mathrm{SD}}$ and $\mathbf{c}^{\mathrm{SD}}$ in the presence of correlated noise is most likely due to the increased variance of semidetached binary light curves compared to dark companion and contact binary light curves (Fig. 5.3), resulting in more informative higher-order principal components that are less affected by numerical noise and are able to better capture the effects of correlated noise. As we decrease the level of correlated noise in the synthetic samples, first $\tilde{\mathbf{C}}^{\mathrm{C}}$ and then $\tilde{\mathbf{C}}^{\mathrm{DC}}$ become more competitive, occasionally outperforming $\tilde{\mathbf{C}}^{\mathrm{SD}}$. In the absence of correlated noise, $\tilde{\mathbf{C}}^{\mathrm{DC}}$ is generally superior to the other representations. Irrespective of the noise conditions, the best performing PCA representation

Table 5.2 Optimal representations and random forest hyperparameters that yielded the best validation macro recalls ($R_{\mathrm{M}}^{\mathrm{V}}$) on the synthetic samples. We also present the validation class recalls for the dark companion ($R_{\mathrm{DC}}^{\mathrm{V}}$), semidetached ($R_{\mathrm{SD}}^{\mathrm{V}}$), and contact ($R_{\mathrm{C}}^{\mathrm{V}}$) binary light curves as well as the test class recalls ($R_{\mathrm{DC}}^{\mathrm{T}}$, $R_{\mathrm{SD}}^{\mathrm{T}}$, $R_{\mathrm{C}}^{\mathrm{T}}$) and the test macro recall ($R_{\mathrm{M}}^{\mathrm{T}}$) of the best performing classifiers. We describe the synthetic samples in Sect. 5.2 and provide the definitions of the representations in Sects. 5.3.1–5.3.2. The definitions of the hyperparameters are given in Sect. 5.3.4.

Sample	Representation	`n_estimators`	`min_samples_leaf`	`max_features`	$R_{\mathrm{DC}}^{\mathrm{V}}$	$R_{\mathrm{SD}}^{\mathrm{V}}$	$R_{\mathrm{C}}^{\mathrm{V}}$	$R_{\mathrm{M}}^{\mathrm{V}}$	$R_{\mathrm{DC}}^{\mathrm{T}}$	$R_{\mathrm{SD}}^{\mathrm{T}}$	$R_{\mathrm{C}}^{\mathrm{T}}$	$R_{\mathrm{M}}^{\mathrm{T}}$
W0C0	$\tilde{\mathbf{C}}_8^{\mathrm{DC}}$	500	1	sqrt	0.98	0.98	0.98	0.98	0.96	0.98	0.96	0.97
W0C1L25	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	100	1	sqrt	0.91	0.91	0.97	0.93	0.92	0.90	0.95	0.92
W0C1L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	100	10	sqrt	0.94	0.92	0.96	0.94	0.95	0.91	0.97	0.94
W0C1L100	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	500	1	sqrt	0.96	0.97	0.98	0.97	0.96	0.97	0.97	0.97
W0C5L25	$\tilde{\mathbf{C}}_9^{\mathrm{C}}$	500	10	sqrt	0.86	0.76	0.88	0.83	0.83	0.71	0.84	0.80
W0C5L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	10	sqrt	0.84	0.79	0.90	0.84	0.83	0.74	0.86	0.81
W0C5L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	500	1	sqrt	0.90	0.91	0.95	0.92	0.93	0.90	0.92	0.91
W0C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	500	10	sqrt	0.81	0.69	0.76	0.75	0.79	0.66	0.73	0.72
W0C10L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	10	sqrt	0.80	0.73	0.77	0.77	0.79	0.69	0.75	0.74
W0C10L100	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.86	0.83	0.89	0.86	0.87	0.82	0.87	0.85
W1C0	$\tilde{\mathbf{C}}_8^{\mathrm{DC}}$	100	1	sqrt	0.96	0.96	0.98	0.96	0.96	0.95	0.97	0.96
W1C1L25	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.90	0.91	0.97	0.93	0.92	0.91	0.95	0.93
W1C1L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.93	0.93	0.97	0.94	0.93	0.93	0.96	0.94
W1C1L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	500	1	sqrt	0.96	0.96	0.98	0.97	0.96	0.95	0.97	0.96
W1C5L25	$\tilde{\mathbf{C}}_6^{\mathrm{SD}}$	500	10	sqrt	0.85	0.75	0.88	0.83	0.83	0.70	0.86	0.79
W1C5L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	10	sqrt	0.84	0.78	0.88	0.84	0.83	0.74	0.86	0.81
W1C5L100	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.91	0.90	0.94	0.92	0.91	0.88	0.92	0.91
W1C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	500	10	sqrt	0.81	0.69	0.77	0.75	0.79	0.64	0.71	0.71
W1C10L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	10	sqrt	0.80	0.71	0.80	0.77	0.78	0.67	0.75	0.73
W1C10L100	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.86	0.83	0.90	0.86	0.85	0.81	0.86	0.84
W10C0	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.91	0.91	0.96	0.93	0.94	0.89	0.93	0.92
W10C1L25	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.90	0.89	0.95	0.91	0.91	0.86	0.92	0.90
W10C1L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	500	1	sqrt	0.91	0.89	0.95	0.92	0.92	0.87	0.93	0.91
W10C1L100	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	500	1	sqrt	0.91	0.90	0.96	0.92	0.93	0.89	0.93	0.92
W10C5L25	$\tilde{\mathbf{C}}_7^{\mathrm{C}}$	500	10	log2	0.85	0.78	0.87	0.83	0.83	0.71	0.84	0.80
W10C5L50	$\tilde{\mathbf{C}}_9^{\mathrm{C}}$	100	10	sqrt	0.86	0.77	0.88	0.84	0.82	0.72	0.85	0.80
W10C5L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	100	10	sqrt	0.88	0.84	0.94	0.88	0.88	0.81	0.91	0.87
W10C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	100	10	log2	0.81	0.70	0.75	0.75	0.79	0.64	0.72	0.72
W10C10L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD}}$	100	10	sqrt	0.80	0.72	0.79	0.77	0.78	0.68	0.74	0.73
W10C10L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	500	10	sqrt	0.85	0.78	0.90	0.84	0.84	0.75	0.85	0.81
W100C0	$\tilde{\mathbf{C}}_7^{\mathrm{DC}}$	100	10	log2	0.90	0.74	0.83	0.83	0.88	0.71	0.78	0.79
W100C1L25	$\tilde{\mathbf{C}}_7^{\mathrm{DC}}$	100	10	sqrt	0.90	0.74	0.82	0.82	0.87	0.70	0.77	0.78
W100C1L50	$\tilde{\mathbf{C}}_7^{\mathrm{DC}}$	500	10	log2	0.91	0.75	0.82	0.83	0.87	0.70	0.77	0.78
W100C1L100	$\tilde{\mathbf{C}}_5^{\mathrm{SD}}$	500	10	sqrt	0.90	0.73	0.84	0.82	0.88	0.70	0.79	0.79
W100C5L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	100	10	log2	0.87	0.70	0.80	0.79	0.84	0.66	0.76	0.75
W100C5L50	$\tilde{\mathbf{C}}_5^{\mathrm{SD}}$	500	10	sqrt	0.86	0.72	0.80	0.79	0.83	0.67	0.77	0.76
W100C5L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD}}$	100	10	sqrt	0.87	0.72	0.82	0.80	0.83	0.69	0.78	0.77
W100C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	500	10	log2	0.83	0.67	0.70	0.73	0.79	0.62	0.68	0.70
W100C10L50	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	100	10	log2	0.81	0.68	0.72	0.73	0.77	0.64	0.69	0.70
W100C10L100	$\tilde{\mathbf{C}}_7^{\mathrm{SD}}$	500	10	log2	0.83	0.71	0.80	0.78	0.79	0.68	0.76	0.74

consistently outperforms $\tilde{\mathbf{C}}^{\mathrm{F}}$ as well as the one-dimensional representation and the extended full representation, even if marginally in the presence of correlated noise.

Regarding the optimal number of coefficients in the latent representations, we observe that the best macro recalls are generally achieved for $n = 7$–9, but as few as five coefficients are enough in the case of the samples W100C1L100

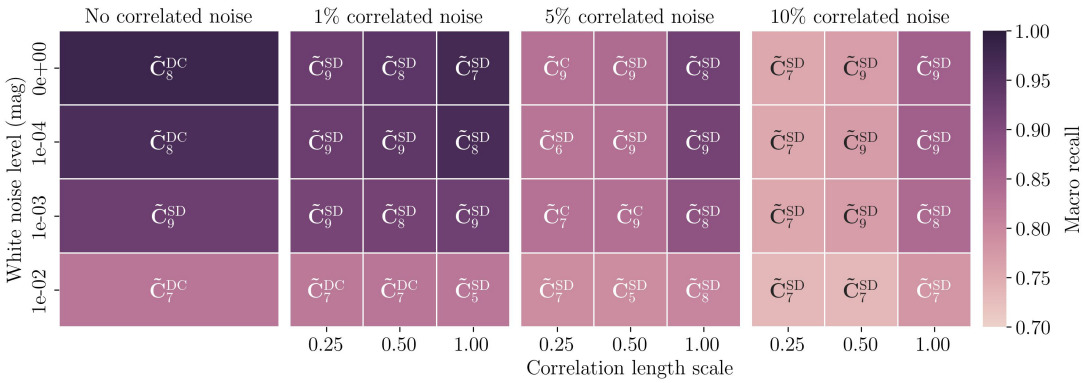

Figure 5.7 Latent representations that yielded the best macro recalls on the validation sets of the synthetic samples. We describe the synthetic samples in Sect. 5.2, and we provide the definitions of the representations in Sects. 5.3.1–5.3.2.

and W100C5L50. The relatively high numbers of coefficients required to achieve the best macro recalls can be explained by the non-linear nature of the random forest classifier and its robustness to overfitting. These properties ensure that the classifier generally does not perform significantly worse when trained on a representation with more coefficients, even if the increase in the macro recall is only marginal and the classifier would perform comparably well with fewer coefficients. It is possible that in some cases, the optimal number of coefficients is even greater than nine, but we did not explore this possibility further. Even if that was the case, the marginal increase in the macro recall beyond the saturation points (Fig. 5.6) makes the analysis largely redundant. Consequently, the random forest classifier is robust to the choice of the number of coefficients in the latent representations, provided we are in the saturated macro recall regime. As a rule of thumb, seven to nine coefficients should be sufficient to capture all the relevant information while avoiding overfitting, but in the presence of strong noise, fewer coefficients may be more appropriate.

In Fig. 5.8, we show the best macro recalls achieved by the random forest classifiers on the validation sets of the synthetic samples (top panel) and the macro recalls of the best performing classifiers on the test sets (bottom panel). By evaluating the macro recalls on the test sets, we obtain a more realistic estimate of the class overlap in the latent representations. We observe that the test macro recalls are generally lower than the validation macro recalls, but the difference is relatively low, with a maximum decrease of 0.05 in absolute terms, demonstrating reliable generalization to previously unseen data. In the presence of purely uncorrelated noise, the test macro recalls vary from $R_{\mathrm{M}}^{\mathrm{T}} = 0.96$ to 0.79, while for purely correlated noise, the macro recalls range from $R_{\mathrm{M}}^{\mathrm{T}} = 0.97$ to 0.75, depending on σ_{CN} and l_{CN}. The general trend is that for fixed σ_{WN} and σ_{CN}, the test macro recalls decrease with decreasing l_{CN}, revealing that the effect of correlated noise is more adverse for shorter correlation length scales. Overall, the test macro recalls of the best performing random forest classifiers range from $R_{\mathrm{M}}^{\mathrm{T}} = 0.97$ (noiseless sample W0C0) down to $R_{\mathrm{M}}^{\mathrm{T}} = 0.70$ (samples W100C10L25 and W100C10L50), indicating a fairly low overlap of the classes in the latent space even in the presence of high levels of uncorrelated and correlated noise.

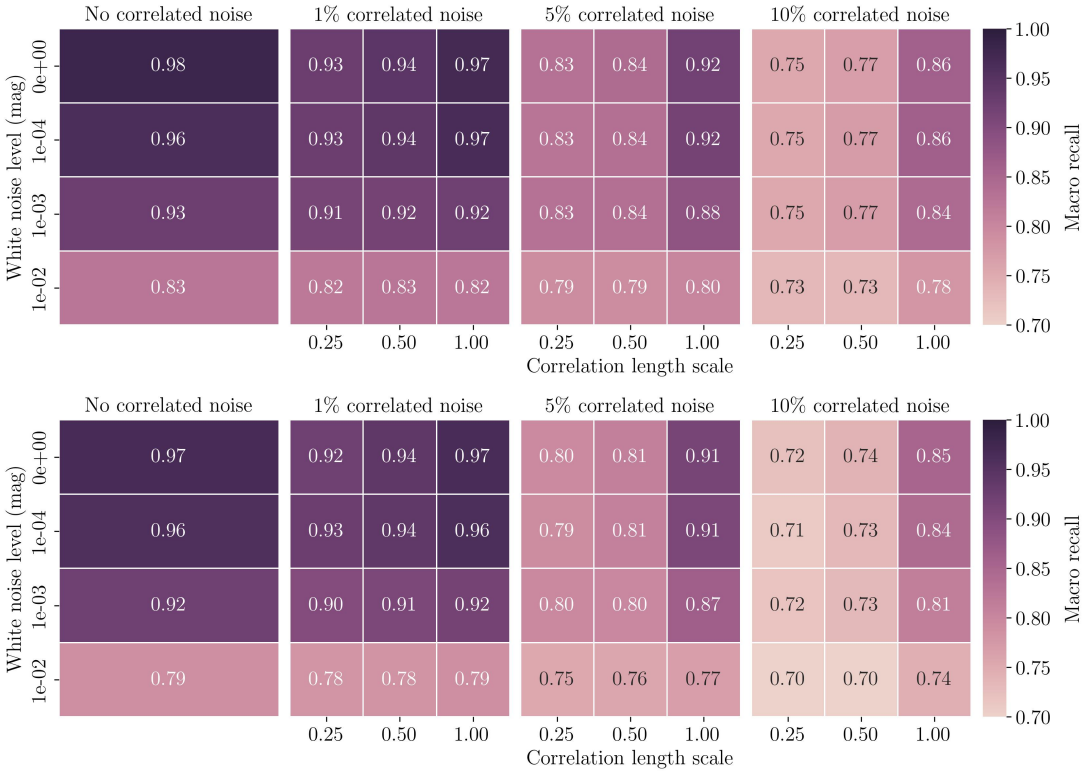

Figure 5.8 Best validation macro recalls achieved by the random forest classifiers trained on the latent representations of the synthetic samples (top panel) and the macro recalls of the best performing classifiers evaluated on the test sets of the synthetic samples (bottom panel).

5.4.5 Impact of variances on macro recalls

To assess the additional information contained in the uncertainties of the coefficients of the latent representations, we repeated the analysis, including the optimization of the hyperparameters of the random forest classifiers, on the extended latent representations augmented with the variances obtained from the least squares fits of the photometric amplitude and the coefficients. In Fig. 5.9, we show the obtained validation macro recalls as a function of n for the four corner cases. The errorbars and the solid black lines have the same meaning as in Fig. 5.6. For readibility reasons, we do not show the results for the one-dimensional representations. The best validation macro recalls achieved by the classifiers trained on the one-dimensional representations augmented with variances are: $R_{\mathrm{M}}^{\mathrm{V}} = 0.59$ for the sample W0C0 (Fig. 5.9a), $R_{\mathrm{M}}^{\mathrm{V}} = 0.62$ for the samples W100C0 and W0C10L50 (Fig. 5.9b–c), and $R_{\mathrm{M}}^{\mathrm{V}} = 0.61$ for the sample W100C10L50 (Fig. 5.9d). This represents an absolute increase of 0.13 to 0.17 in the macro recalls of the one-dimensional representations compared to the unaugmented case. We observe a similar shift to higher macro recalls for all augmented latent representations across all synthetic samples, but the shift becomes less pronounced with increasing n. The increase in the macro recalls is most prominent in the noiseless sample W0C0, where the performance of the PCA representations is considerably improved for up to $n = 6$–7, and the macro recall of the representation $\tilde{\mathbf{C}}^{\mathrm{F+V}}$ is positively affected all the way up to $n = 9$, outperforming even the PCA representations. In the presence of noise, the macro recalls of low-dimensional

latent representations are still higher when augmented with variances, but the improvement practically vanishes beyond the saturation points of the unaugmented representations.

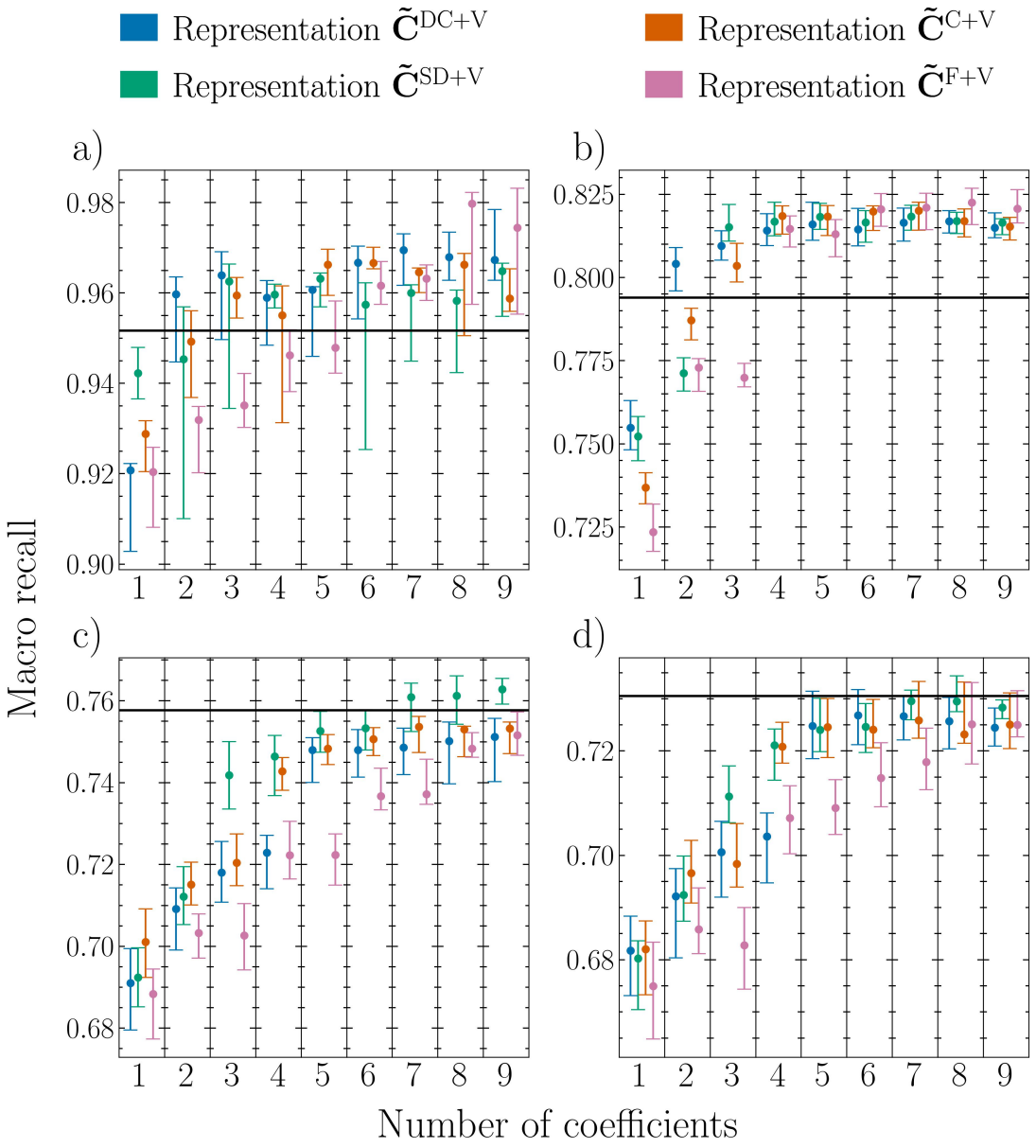

Figure 5.9 Validation macro recalls of the random forest classifiers trained on the augmented representations $\tilde{\mathbf{C}}^{\mathrm{DC+V}}$, $\tilde{\mathbf{C}}^{\mathrm{SD+V}}$, $\tilde{\mathbf{C}}^{\mathrm{C+V}}$, and $\tilde{\mathbf{C}}^{\mathrm{F+V}}$ of the synthetic samples W0C0 (a), W100C0 (b), W0C10L50 (c), and W100C10L50 (d), evaluated as a function of the number of coefficients in the representation. The errorbars show the minimum and the maximum validation macro recalls for a given number of coefficients. The solid black lines represent the best validation macro recalls achieved by the classifiers trained on extended full representations of the samples. We describe the synthetic samples in Sect. 5.2 and provide details about the augmented representations in Sect. 5.3.4.

We present the optimal augmented representations and hyperparameter setups for each synthetic sample in Table 5.3. The optimal setups are remarkably similar to the setups obtained for the unaugmented representations, with the most significant difference being the preference of `max_features` = `None`

instead of `max_features` = `log2` in the limit of strong noise. Other than that, the optimal setups remain largely unchanged, with `n_estimators` = `500`, `min_samples_leaf` = `10`, and `max_features` = `sqrt` leading to the best results for 13 out of 40 synthetic samples. Similarly, the optimal representations are mostly unaffected by the inclusion of the variances, with $\tilde{\mathbf{C}}^{\mathrm{SD+V}}$ yielding the best macro recalls for the majority of the samples and absolutely dominating in the presence of strong correlated noise (Fig. 5.10). Perhaps suprisingly, the Fourier representation seems to benefit the most from the inclusion of the variances in the input of the random forest classifiers, replacing $\tilde{\mathbf{C}}^{\mathrm{DC+V}}$ as the most informative representation in the absence of correlated noise and outperforming $\tilde{\mathbf{C}}^{\mathrm{SD+V}}$ on several occasions, especially in the presence of strong uncorrelated noise. Similar to the representations without the variances, the augmented representations achieve the best results for $n = 7$–9. This is not surprising, given the marginal increase in the macro recalls beyond the saturation points of the unaugmented representations at $n = 3$–6 (Fig. 5.9). Consequently, the best validation macro recalls achieved by the random forest classifiers and the test macro recalls of the best performing classifiers are almost identical whether we augment the representations with the variances or not (Figs. 5.8 and 5.11). Even in the noiseless case, where the impact of the variances is the most pronounced, the improvement is only marginal. In some cases, the best macro recall even decreases when we include the variances (e.g., samples W10C1L25–100).

We conclude that while the inclusion of variances can significantly improve the macro recall of random forest classifiers trained on low-dimensional latent representations ($n \lesssim 3$), the positive effect diminishes with increasing dimension of the representation, becoming negligible once we cross the macro recall saturation points of the unaugmented representations at $n = 3$–6. Consequently, augmenting the latent representations with variances is generally not necessary, provided the dimension of the representation is sufficiently high to capture all the relevant information. To achieve the best results, we recommend training the classifier on latent representations with $n = 7$–9.

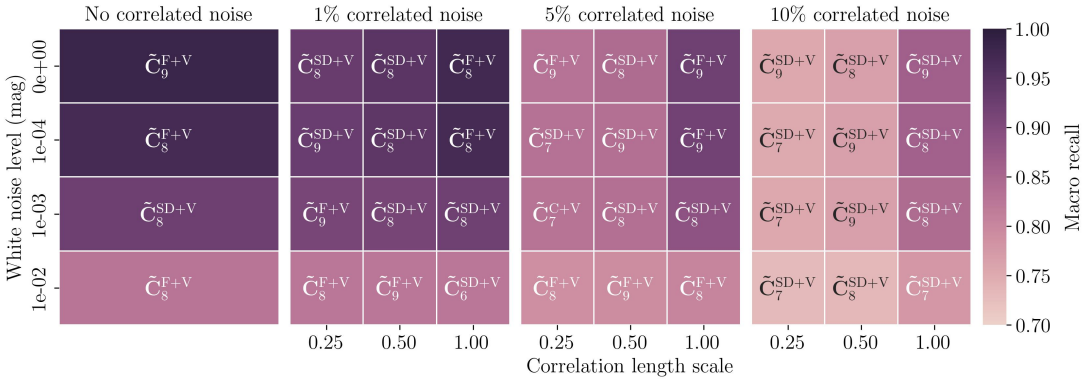

Figure 5.10 Augmented latent representations that yielded the best macro recalls on the validation sets of the synthetic samples. See Sects. 5.2 and 5.3.4 for the descriptions of the synthetic samples and the augmented representations, respectively.

Table 5.3 Optimal augmented representations and random forest hyperparameters that yielded the best validation macro recalls ($R_{\mathrm{M}}^{\mathrm{V}}$) on the synthetic samples. We also provide the validation class recalls for the dark companion ($R_{\mathrm{DC}}^{\mathrm{V}}$), semidetached ($R_{\mathrm{SD}}^{\mathrm{V}}$), and contact ($R_{\mathrm{C}}^{\mathrm{V}}$) binary light curves as well as the test class recalls ($R_{\mathrm{DC}}^{\mathrm{T}}$, $R_{\mathrm{SD}}^{\mathrm{T}}$, $R_{\mathrm{C}}^{\mathrm{T}}$) and the test macro recall ($R_{\mathrm{M}}^{\mathrm{T}}$) of the best performing classifiers. See Sect. 5.2 for the description of the synthetic samples and Sect. 5.3.4 for the definitions of the augmented representations and the hyperparameters.

Sample	Representation	`n_estimators`	`min_samples_leaf`	`max_features`	$R_{\mathrm{DC}}^{\mathrm{V}}$	$R_{\mathrm{SD}}^{\mathrm{V}}$	$R_{\mathrm{C}}^{\mathrm{V}}$	$R_{\mathrm{M}}^{\mathrm{V}}$	$R_{\mathrm{DC}}^{\mathrm{T}}$	$R_{\mathrm{SD}}^{\mathrm{T}}$	$R_{\mathrm{C}}^{\mathrm{T}}$	$R_{\mathrm{M}}^{\mathrm{T}}$
W0C0	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	100	10	sqrt	0.98	0.99	0.98	0.98	0.97	0.98	0.96	0.97
W0C1L25	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.92	0.90	0.97	0.93	0.92	0.90	0.95	0.92
W0C1L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.94	0.93	0.97	0.95	0.94	0.93	0.96	0.94
W0C1L100	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	1	sqrt	0.98	0.97	0.98	0.97	0.96	0.98	0.97	0.97
W0C5L25	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	500	10	sqrt	0.87	0.75	0.87	0.83	0.83	0.71	0.83	0.79
W0C5L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	10	sqrt	0.85	0.78	0.90	0.84	0.82	0.74	0.86	0.81
W0C5L100	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	500	1	sqrt	0.92	0.91	0.94	0.92	0.94	0.89	0.91	0.91
W0C10L25	$\tilde{\mathbf{C}}_9^{\mathrm{SD+V}}$	500	10	sqrt	0.81	0.70	0.75	0.75	0.78	0.66	0.71	0.72
W0C10L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	10	sqrt	0.80	0.72	0.77	0.77	0.78	0.69	0.74	0.74
W0C10L100	$\tilde{\mathbf{C}}_9^{\mathrm{SD+V}}$	500	1	sqrt	0.86	0.83	0.90	0.86	0.86	0.81	0.87	0.85
W1C0	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	10	sqrt	0.96	0.97	0.98	0.97	0.96	0.96	0.97	0.96
W1C1L25	$\tilde{\mathbf{C}}_9^{\mathrm{SD+V}}$	500	10	sqrt	0.93	0.88	0.97	0.93	0.92	0.87	0.96	0.92
W1C1L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.93	0.93	0.97	0.94	0.93	0.93	0.96	0.94
W1C1L100	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	1	sqrt	0.96	0.96	0.98	0.97	0.96	0.96	0.96	0.96
W1C5L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD+V}}$	100	10	sqrt	0.85	0.76	0.88	0.83	0.82	0.70	0.85	0.79
W1C5L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD+V}}$	500	1	sqrt	0.83	0.81	0.86	0.84	0.82	0.77	0.83	0.81
W1C5L100	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	100	10	sqrt	0.93	0.89	0.94	0.92	0.94	0.86	0.91	0.90
W1C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD+V}}$	100	10	None	0.80	0.70	0.75	0.75	0.78	0.64	0.71	0.71
W1C10L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD+V}}$	500	10	sqrt	0.80	0.71	0.79	0.77	0.78	0.68	0.74	0.73
W1C10L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.86	0.83	0.90	0.86	0.85	0.80	0.86	0.84
W10C0	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.90	0.91	0.96	0.92	0.94	0.90	0.93	0.92
W10C1L25	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	500	10	sqrt	0.91	0.87	0.96	0.91	0.90	0.84	0.93	0.89
W10C1L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.90	0.89	0.95	0.91	0.91	0.87	0.92	0.90
W10C1L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	1	sqrt	0.91	0.90	0.96	0.92	0.92	0.89	0.93	0.91
W10C5L25	$\tilde{\mathbf{C}}_7^{\mathrm{C+V}}$	500	10	sqrt	0.85	0.77	0.87	0.83	0.82	0.71	0.84	0.79
W10C5L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	10	sqrt	0.85	0.77	0.88	0.83	0.83	0.73	0.85	0.80
W10C5L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	100	10	sqrt	0.88	0.84	0.94	0.89	0.86	0.81	0.91	0.86
W10C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD+V}}$	500	10	sqrt	0.81	0.71	0.74	0.75	0.78	0.65	0.71	0.72
W10C10L50	$\tilde{\mathbf{C}}_9^{\mathrm{SD+V}}$	500	10	sqrt	0.80	0.72	0.78	0.77	0.78	0.69	0.73	0.73
W10C10L100	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	500	10	None	0.85	0.77	0.91	0.84	0.84	0.75	0.85	0.81
W100C0	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	10	sqrt	0.90	0.75	0.83	0.83	0.89	0.71	0.78	0.79
W100C1L25	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	10	None	0.89	0.75	0.83	0.83	0.86	0.70	0.78	0.78
W100C1L50	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	100	10	None	0.89	0.76	0.83	0.82	0.87	0.70	0.78	0.78
W100C1L100	$\tilde{\mathbf{C}}_6^{\mathrm{SD+V}}$	100	10	sqrt	0.90	0.74	0.83	0.82	0.87	0.70	0.78	0.79
W100C5L25	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	10	None	0.86	0.72	0.79	0.79	0.84	0.67	0.74	0.75
W100C5L50	$\tilde{\mathbf{C}}_9^{\mathrm{F+V}}$	500	10	None	0.86	0.73	0.79	0.79	0.83	0.68	0.76	0.76
W100C5L100	$\tilde{\mathbf{C}}_8^{\mathrm{F+V}}$	500	10	None	0.86	0.74	0.82	0.80	0.83	0.70	0.78	0.77
W100C10L25	$\tilde{\mathbf{C}}_7^{\mathrm{SD+V}}$	100	10	sqrt	0.83	0.67	0.69	0.73	0.79	0.63	0.67	0.70
W100C10L50	$\tilde{\mathbf{C}}_8^{\mathrm{SD+V}}$	100	10	None	0.81	0.68	0.71	0.73	0.78	0.64	0.69	0.70
W100C10L100	$\tilde{\mathbf{C}}_7^{\mathrm{SD+V}}$	100	10	None	0.83	0.71	0.79	0.78	0.80	0.68	0.75	0.74

5.4.6 Expected precision of random forest classifiers

The relatively high macro recalls of the random forest classifiers on the validation and test sets of the samples of dark companion, semidetached, and contact binary light curves must be interpreted in the context of the synthetic data, which we deliberately balanced to avoid bias towards a particular class. Consequently, the macro recalls should be understood as a measure of the mean

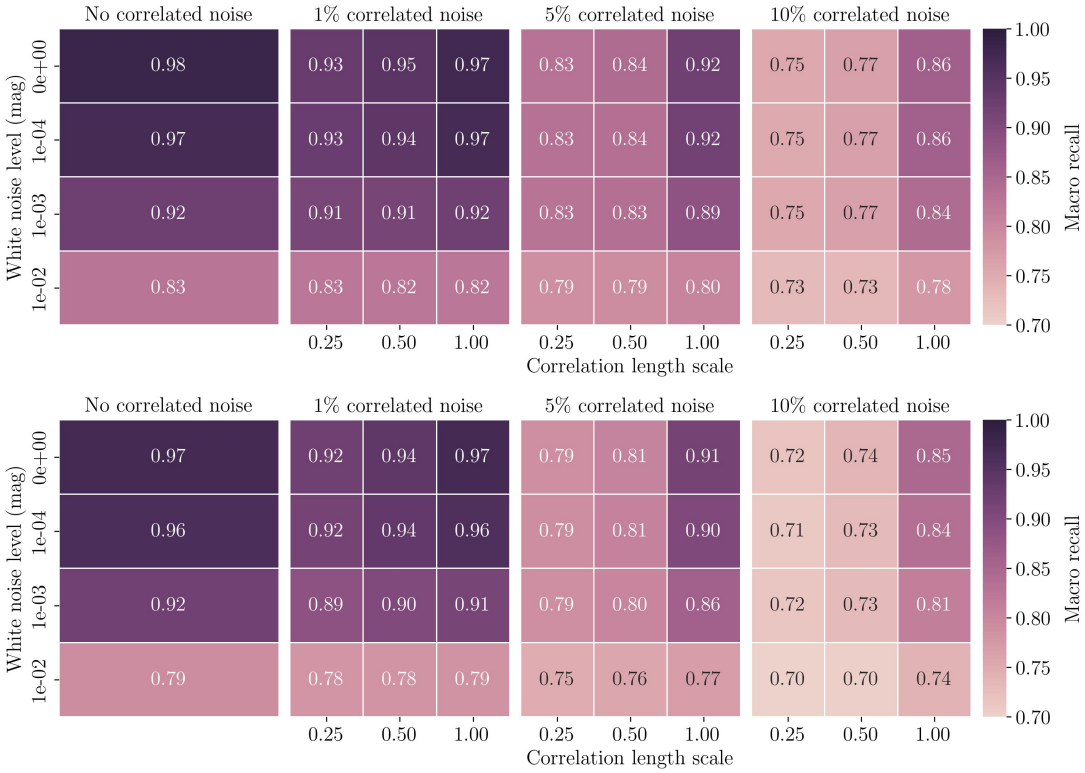

Figure 5.11 Highest validation macro recalls achieved by the random forest classifiers trained on the augmented latent representations of the synthetic samples (top panel) and the macro recalls of the best performing classifiers evaluated on the test sets of the synthetic samples (bottom panel).

non-overlap of the classes in the latent space rather than the expected accuracy of the classifier on real data. Despite that, we can still use the class recalls obtained for the test sets of the synthetic samples (Table 5.2) to estimate the *expected precision* of the classifier on previously unseen data, provided we know the relative frequencies of the classes in the sample. Assuming the worst case scenario, where all misclassifications fall into the dark companion class, we can estimate the expected precision P_{DC} of the classifier for the dark companion class as

$$P_{\mathrm{DC}} = \frac{f_{\mathrm{DC}} R^{\mathrm{T}}_{\mathrm{DC}}}{f_{\mathrm{DC}} R^{\mathrm{T}}_{\mathrm{DC}} + f_{\mathrm{SD}}(1 - R^{\mathrm{T}}_{\mathrm{SD}}) + f_{\mathrm{C}}(1 - R^{\mathrm{T}}_{\mathrm{C}})}, \tag{5.12}$$

where f_{DC}, f_{SD}, and f_{C} are the relative frequencies of the dark companion, semidetached, and contact binary classes in the sample, respectively, and $R^{\mathrm{T}}_{\mathrm{DC}}$, $R^{\mathrm{T}}_{\mathrm{SD}}$, and $R^{\mathrm{T}}_{\mathrm{C}}$ are the class-specific test recalls of the classifier trained on data with the same noise characteristics as the sample. The expected precision of the classifier for the semidetached and contact binary classes can be calculated analogously. The expected precision P_{DC} tells us what fraction of the objects in the sample classified as dark companions can we expect to be actual dark companions, as opposed to $R^{\mathrm{T}}_{\mathrm{DC}}$, which tells us what fraction of the actual dark companions in the sample can we expect to be classified as such. The expected precision allows us to estimate the purity of the refined sample of objects classified as dark companions, which is crucial for assessing the cost-efficiency of follow-up observations. Alternatively, if the relative frequencies are unknown, we can

estimate the prior purity of the sample that is required to achieve a predefined level of purity in the refined sample, helping us decide whether the sample is worth pursuing.

We illustrate the calculation of P_{DC} with a model example: consider a sample of ellipsoidal variables with $f_{\mathrm{DC}} = 0.01$, $f_{\mathrm{SD}} = 0.66$, and $f_{\mathrm{C}} = 0.33$, and a classifier with $R^{\mathrm{T}}_{\mathrm{DC}} = 0.96$, $R^{\mathrm{T}}_{\mathrm{SD}} = 0.98$, and $R^{\mathrm{T}}_{\mathrm{C}} = 0.96$, corresponding to the best-case scenario in Table 5.2. If we plug these values into Eq. 5.12, we obtain $P_{\mathrm{DC}} \approx 0.27$, yielding a refined sample of objects classified as dark companions with an expected purity of approximately 27%, which is a significant improvement over the 1% prior purity of the full sample. If we decrease the class recalls to $R^{\mathrm{T}}_{\mathrm{DC}} = 0.79$, $R^{\mathrm{T}}_{\mathrm{SD}} = 0.62$, and $R^{\mathrm{T}}_{\mathrm{C}} = 0.68$, corresponding to the worst-case scenario in Table 5.2, we obtain a refined sample with $P_{\mathrm{DC}} \approx 0.02$, merely doubling the purity of the full sample. This example demonstrates that even low-purity samples can yield significantly improved results if the classifier produces high enough recalls for all classes.

In practice, we cannot reasonably expect that the noise characteristics of real data will exactly match the characteristics of one of our synthetic samples. When designing the synthetic samples, our goal was not to simulate realistic observing conditions but rather to systematically study the effects of correlated and uncorrelated noise on our ability to distinguish between the three binary classes. Consequently, the classifiers trained on the synthetic samples are not directly applicable to real data. However, if we somehow manage to transfer the noise characteristics of real observations to the synthetic data, including the effects of spots and other phenomena that introduce variations from the synthetic models, we can train a classifier on the augmented data and obtain class recalls that are specific to the target sample. In general, this is a challenging task, requiring a detailed understanding of the instrumental noise of the survey and the physical processes affecting the shapes of the light curves, such as flares, pulsations, etc. As a first approximation, we can fit the light curves in the target sample using Gaussian process regression, modeling the mean as a Fourier series of sufficiently high order. We can then inject the residuals from the Fourier fit of a randomly selected light curve into a preselected noiseless synthetic light curve (before normalization), independently repeating the process several times for all light curves in the synthetic data. By training, validating, and testing the classifier on the augmented data, we can obtain class recalls that roughly reflect the noise characteristics of the target sample, allowing us to estimate the expected precision of the classifier on real data.

5.5 Discussion and conclusions

In this work, we addressed the issue of whether it is possible to identify non-interacting black holes and neutron stars in close binary systems based solely on the effects they induce in the broadband photometric light curves of their companion stars. A massive compact companion in a close binary system can tidally deform the primary star into a teardrop shape, causing periodic changes in the area of the star that is visible to the observer and giving rise to ellipsoidal variations in its light curve. By searching for stars that exhibit ellipsoidal variations, we can potentially identify binary systems that host electromagnetically silent black

holes and neutron stars, which we collectively refer to as dark companions. The problem with this approach is that other types of objects, such as contact binaries and semidetached binaries, can also exhibit ellipsoidal variations, making the identification of dark companions challenging.

One way to distinguish dark companion binaries from contaminants is to train a machine learning classifier on a well-curated sample of their observed light curves in which each class properly represented. However, the limited number of known dark companion binaries prevents us from following this approach. Instead, we generated a large number of synthetic light curves of dark companion binaries, semidetached binaries, and contact binaries, covering a wide range of physical and orbital parameters of the systems (Sect. 5.2.1). To account for the effects of instrumental noise and stellar spots, we injected the light curves with various levels of correlated and uncorrelated Gaussian noise, resulting in 40 synthetic samples (Table 5.1). While injecting the noise, we oversampled each binary class by a different oversampling factor, mitigating the bias towards the majority class and resulting in a more balanced representation of the classes (Sect. 5.2.2). We normalized the light curves in the synthetic samples by fitting them with a fourth-order Fourier series, realigning them to have the primary minimum at phase 0, and vertically shifting and rescaling them so that their Fourier fits have a minimum and maximum of 0 and 1, respectively.

To uncover the underlying discriminative patterns in the high-dimensional synthetic data, we reduced the light curves using PCA and discrete Fourier series – two linear methods that are well-suited for the decomposition of discrete periodic signals. We performed PCA separately on the noiseless normalized light curves of each binary class, yielding three distinct PCA bases, with the expansion coefficients in the bases forming the PCA representations of the light curves. We also constructed a discretized Fourier basis by sampling the Fourier basis functions on the same grid as the synthetic light curves. For the Fourier representation to be directly comparable with the PCA representations, we subtracted the mean dark companion binary light curve from all light curves, including the semidetached and contact binary light curves, prior to the decomposition. In all four bases, we distinguished between unit and rescaled latent representations, as well as extended unit and rescaled latent representations, where the extended representations contain the amplitude of the light curve as the zeroth element. We provide a detailed description of the representations and their notations in Sects. 5.3.1–5.3.2.

Our analysis of the noiseless synthetic sample S0 revealed that the mean light curves of dark companion binaries, semidetached binaries, and contact binaries are very similar (Fig. 5.1), demonstrating the difficulty of distinguishing between the classes using photometric data alone. The finer details of the light curves are captured by the principal components, which differ between the classes and become progressively more oscillatory with increasing order (Fig. 5.2). In all cases, the cumulative explained variance of the PCA components exceeds 99% somewhere between the second and fifth component (Fig. 5.3), indicating that the light curves are effectively confined to low-dimensional hyperplanes in the original high-dimensional space and justifying the use of PCA for dimensionality reduction.

Visual inspection of the latent representations of the synthetic samples W0C0, W100C0, W0C10L50, and W100C10L50, which are the corner cases with the

lowest and the highest levels of correlated and uncorrelated noise, revealed major differences between the PCA and Fourier representations (Fig. 5.4). We found that the first three coefficients of the PCA representations exhibit much richer structure in the latent space than the coefficients of the Fourier representation, in which the distributions of the classes are collapsed along the third coefficient, mixing the classes together and decreasing their separation. This is true for all corner cases, but the differences are more pronounced in the absence of correlated noise. We found that correlated noise affects the structure of the latent space on a more fundamental level than uncorrelated noise, making it more difficult to visually separate between the classes. Still, even in the presence of strong correlated and uncorrelated noise, the PCA representations retain some of their original structure, while the Fourier representation becomes almost featureless, demonstrating the superiority of the PCA representations.

To compare the class separation in the unit PCA and Fourier representations under different noise conditions, we calculated the silhouette scores of the four corner cases as a function of the number of coefficients in the representation $n = 1$–9 (Sect. 5.4.3). We provide the definition of the silhouette score and details of its calculation in Sect. 5.3.3. Our findings from the analysis of the silhouette scores are largely consistent with the conclusions drawn from the visual inspection of the corner cases. We found that the PCA representations generally yield better class separation than the Fourier representation for a given n, with the exception of the first few coefficients which seem to be more informative in the Fourier representation (Fig. 5.5). This holds true across all corner cases, demonstrating the robustness of the PCA representations to noise. The silhouette scores of the PCA representations typically peak or plateau at $n = 3$–6, depending on the level and type of injected noise. The only exception is the contact representation, which exhibits a more gradual increase of the silhouette score with n, similar to what we observe for the Fourier representation. By comparing the silhouette scores of the latent representations with the benchmark silhouette score of the full representation, we found that the latent representations are much more immune to uncorrelated noise than correlated noise, confirming what we observed in Figs. 5.4c–d. The semidetached representation proved to be the most robust to correlated noise, yielding the best class separation for the synthetic samples W0C10L50 and W100C10L50. We found that in the absence of correlated noise, the dark companion representation outperforms the other representations, achieving the highest silhouette scores for the synthetic samples W0C0 and W100C0.

To assess the mean overlap of the classes in the latent space, we trained random forest classifiers on the extended rescaled latent representations of the synthetic samples and analyzed their macro recalls as a function of $n = 1$–9 (Sect. 5.4.4). We describe the calculation of the macro recalls and the hyperparameter setups of the random forest classifiers in Sect. 5.3.4. We observed that the macro recalls of the PCA representations start to saturate at $n = 3$–6, depending on the representation and the noise level, and then slightly increase or decrease with increasing n (Fig. 5.6). The saturation points of the PCA representations are slightly shifted to higher n compared to the plateaus and peaks of the silhouette scores, especially in the case of the dark companion representation of the samples W0C0 and W0C10L50. We observed no significant shift in the saturation points of the Fourier representation, indicating good alignment of the silhouette score

with the Fourier representation.

Consistent with the results of our analysis of the silhouette scores, we observed that the semidetached representation generally yields the best macro recalls in the presence of correlated noise, while the dark companion representation outperforms the other representations in the complete absence of noise and in the presence of strong uncorrelated noise (Fig. 5.7). However, the best macro recalls are typically obtained for $n = 7$–9, where the differences between the macro recalls are marginal and all PCA representations perform comparably (Fig. 5.6). In all corner cases except W100C0, the Fourier representation consistently yields lower macro recalls than the PCA representations for the same n. In addition, the Fourier representation does not achieve the best macro recalls for any of the synthetic samples, suggesting that the classes are generally more intermixed in the Fourier latent space and pointing to the superiority of the PCA representations in terms of class separation and robustness to noise.

We obtained the best validation macro recalls of the random forest classifiers on the synthetic samples by taking the maximum across all representations and hyperparameter setups (top panel of Fig. 5.8). By evaluating the best performing random forests on the test sets of the synthetic samples, we verified that the validation macro recalls generalize well to previously unseen data, with a typical decrease of 1–5% in absolute terms, depending on the synthetic sample (bottom panel of Fig. 5.8). The test macro recalls of the best performing random forest classifiers vary from $R_{\mathrm{M}}^{\mathrm{T}} = 0.97$ in the absence of noise to $R_{\mathrm{M}}^{\mathrm{T}} = 0.70$ in the presence of strong correlated and uncorrelated noise, manifesting low to medium overlap of the classes in the latent space. We found that in the presence of moderate levels of uncorrelated noise ($10^{-4}\,\mathrm{mag} \leq \sigma_{\mathrm{WN}} \leq 10^{-3}\,\mathrm{mag}$), the overlap of the classes is largely determined by the level of correlated noise, with shorter correlation lengths generally yielding worse class separations. Uncorrelated noise starts to significantly affect the macro recalls only at higher levels ($\sigma_{\mathrm{WN}} > 10^{-3}\,\mathrm{mag}$), whereas correlated noise can considerably increase the class overlap even at low to moderate levels ($0.01 \leq \sigma_{\mathrm{CN}} \leq 0.05$). This contrast reveals a more fundamental effect of correlated noise on the separation of the classes in the latent space. Nevertheless, even in the presence of strong correlated noise ($\sigma_{\mathrm{CN}} = 0.1$), which can amount to significant surface coverage with stellar spots, the classes remain largely separated in the latent space, with test macro recalls reaching $R_{\mathrm{M}}^{\mathrm{T}} = 0.70$–$0.72$.

We retrained the random forest classifiers on the extended representations augmented with the variances of the photometric amplitude and the latent coefficients to investigate whether the inclusion of the variances in the input of the classifiers improves the separation of the classes (Sect. 5.4.5). We found that while the macro recalls of low-dimensional PCA representations ($n \lesssim 3$) significantly increase, the improvement is only marginal beyond the saturation points of the unaugmented representations at $n = 3$–6. The positive effect of the variances is the most pronounced in the case of the Fourier representation, where the macro recalls are considerably improved for $n = 1$–9 across all noise conditions (Fig. 5.9), even surpassing the macro recalls of the PCA representations in the complete absence of noise and in the presence of strong uncorrelated noise (Fig. 5.10). However, the best macro recalls of the random forest classifiers trained on the augmented representations (Fig. 5.11) remain largely unchanged compared to the macro recalls obtained for the unaugmented representations, pointing to

the limited benefit of including the variances in the representations.

Using the obtained test class recalls (Table 5.2), we showed that it is possible to estimate the expected precision of the classifier on real data, assuming we have a rough estimate of the relative frequencies of the classes in the sample (Sect. 5.4.6). We illustrated the calculation of the expected precision on a model example, and we showed that in the best-case scenario, we can increase the purity of a dark companion sample by a factor of up to 27, assuming a prior purity of 1%.

There are several limitations to our study that need to be addressed. First, we generated the synthetic light curves using PHOEBE binary models, which are not perfect representations of reality. As a result, the synthetic light curves may not capture all the complexities of real light curves, such as Doppler beaming and boosting (Loeb & Gaudi 2003; Zucker et al. 2007), which are not supported as of version 2.2. In addition, we made several simplifying assumptions in the generation of the synthetic light curves, such as the circularity of the orbits or the default limb darkening calculation settings. However, we expect that these effects are secondary to the main features of the light curves and become negligible in the presence of noise. Consequently, we do not expect the main findings of our study, such as the superiority of the PCA representations over the Fourier representation, to be significantly affected by these assumptions.

Second, we injected the synthetic light curves with various levels of uncorrelated and correlated Gaussian noise to account for the effects of instrumental noise and stellar spots, respectively. While the uncorrelated noise is a good approximation of the instrumental noise, the correlated noise is a very simplified model of the effects of stellar spots, which can be more complex in reality. Also, we limited our analysis to $l_{\mathrm{CN}} = 0.25$, 0.5, and 1, which may not cover the full range of possible timescales of the spots. A more realistic treatment of spots could be achieved either by simulating the spots directly in PHOEBE, which would greatly increase the size and complexity of the synthetic data as well as the computational cost of its generation, or by using spot models of single stars (e.g., Luger et al. 2019, 2021b,a) instead of correlated Gaussian noise. In addition, we did not consider the effects of other sources of noise, such as the intrinsic variability of the stars. All these effects can potentially complicate the separation of the classes in the latent space and decrease the macro recalls on real data. To avoid modeling the noise altogether, we can transfer the noise characterics directly from the target sample as we described in Sect. 5.4.6, taking into account all the complexities and idiosyncracies of real astronomical observations and yielding a more accurate estimate of the expected precision of the classifier on real data. There are several samples of ellipsoidal variables in the literature that are suitable for this purpose, e.g., Green et al. (2023); Gomel et al. (2023, 2021c). We plan to investigate these samples using our method in future work.

Third, in our analysis, we implicitly assumed that ellipsoidal samples contain only dark companion binaries, semidetached binaries, and contact binaries, which we regard as the most challenging classes to separate. In reality, low-inclination detached binaries are also prominently present in ellipsoidal samples, and other types of objects, such as pulsating stars and spotted rotating stars, can be found in the samples as well. We excluded detached binaries from our analysis, because we assumed that their light curves are sufficiently similar to those of

semidetached binaries to be treated as such by the classifier, effectively increasing the relative frequency of the semidetached class in the sample. However, further analysis is required to confirm the validity of this assumption. If needed, our method can be easily extended to accommodate detached binaries as a separate class, allowing for a more detailed differentiation between the classes. As for pulsating stars and spotted rotating stars, these can be efficiently removed from the sample by performing suitable cuts on period and amplitude (Green et al. 2023). Consequently, we assume that the fraction of these objects in ellipsoidal samples is negligible, and their effect on the performance of the classifiers is minimal. Lastly, in our definition of dark companion binaries, the dark companion is either a black hole or a neutron star. In practice, it is difficult to distinguish ellipsoidal variations induced by a neutron star from those induced by a massive white dwarf (WD), leading to contamination of the dark companion class by these objects. Since the true nature of the dark companion can only be reliably determined through high-resolution spectroscopy, we did not attempt to separate dark companion binaries from binaries hosting massive WDs. Instead, we treated them as a single class in our analysis, leaving their separation up to follow-up observations. To assess the level of contamination by massive WDs, further analysis of the candidates identified by our method is required.

Fourth, we trained the classifiers on the extended latent representations, which encode the absolute scales and the morphologies of the light curves, but do not take into account their periods. While we expect that most of the discriminative information is contained in the shapes of the light curves, the periods can provide additional information about the physical characters of the systems, potentially improving the separation of the classes. We plan to investigate the impact of including the periods in the latent representations on the performance of the classifiers in future work.

Fifth, to obtain the boundaries and quantify the overlap between the classes, we trained the random forest classifiers on the oversampled synthetic data, which we further balanced by weighting the objects with the inverse of the class size. This is a valid approach, provided the distributions of the objects within the classes are representative. That is, the objects of a given class populate the parameter space in a way that is representative of real data. However, this is not the case in our synthetic data, which we generated on uniform and log-uniform grids of physical and orbital parameters of the systems (Sect. 5.2.1). Our motivation was to cover a wide range of parameters in order to capture the full diversity of the light curves rather than to mimic the real distributions of the objects. Consequently, the boundaries of the classes in the latent space may be distorted with respect to real data, leading to biased estimates of the macro recalls to either side. This can only be avoided by generating synthetic light curves with representative distributions of the parameters, which is a challenging task given the complexity of the parameter space. Representative distributions of at least some physical parameters could be obtained using binary population synthesis (e.g., Weller & Johnson 2023; Chawla et al. 2023). Until we train the classifiers on data with representative parameter distributions, we cannot quantify the impact of parameter sampling strategy on the macro recalls and the expected precision of the classifiers. However, assuming that the objects within the classes do not accumulate close to the decision boundaries in the latent space, we expect our

estimates to generalize well to real data.

Our method of projecting light curves to PCA components learned from synthetic data can be easily extended to multi-band photometry, which can provide additional information about the physical properties of the systems. For example, ultraviolet photometry from upcoming satellites such as QUVIK (Werner et al. 2024; Krtička et al. 2024) or ULTRASAT (Shvartzvald et al. 2024) could be used to constrain the nature of semidetached binaries, including binaries with stripped-envelope stars (Rowan et al. 2024). Precise ultraviolet observations could also be used to further break the degeneracy between contact binaries and dark companion binaries using the different limb darkening properties of these systems in UV. We plan to pursue this direction in future work.

Another possibility of improving our method is to explicitly model the correlated noise in the light curves using Gaussian processes. We found that correlated noise affects the light curves on a more fundamental level than uncorrelated noise, preventing the latent representations from disentangling the signal from the noise. This is not suprising – by fitting the light curves using least squares, we implicitly assume homoscedastic uncorrelated Gaussian noise, which is clearly not justified in the presence of correlated noise. We can avoid this assumption by modeling the light curves using Gaussian process regression with a nonzero mean given by a linear combination of Fourier components. We expect that this approach will yield more robust latent representations of the light curves and better separation of the classes in the latent space. The recovered parameters of the correlated noise can possibly also be informative about the physical properties of the systems. We intend to explore this approach in future work.

MP thanks Yuan-Sen Ting for helpful discussions. We acknowledge the support of the Czech Science Foundation Grant No. 24-11023S. The work of OP was supported by the Charles University Research program No. UNCE24/SCI/016. This work made use of the following software packages: `PHOEBE` (Prša et al. 2016; Conroy et al. 2020b), `numpy` (Harris et al. 2020), `scikit-learn` (Pedregosa et al. 2011), `matplotlib` (Hunter 2007), `pandas` (Wes McKinney 2010).

5.A Scatter plots of the coefficients of the latent representations

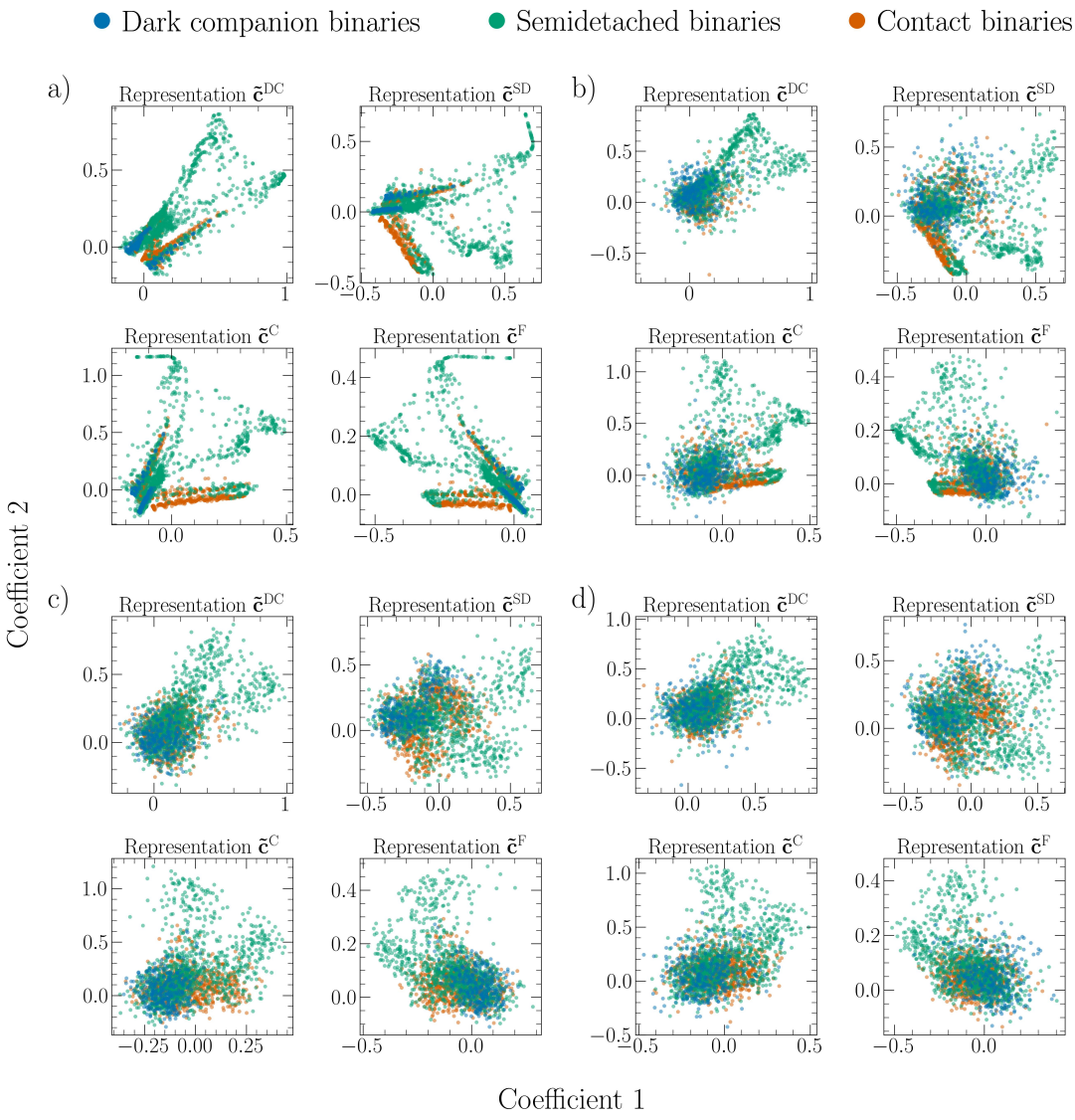

Figure 5.A.1 Scatter plots of the first and second coefficients of the representations $\tilde{\mathbf{c}}^{\mathrm{DC}}$, $\tilde{\mathbf{c}}^{\mathrm{SD}}$, $\tilde{\mathbf{c}}^{\mathrm{C}}$, and $\tilde{\mathbf{c}}^{\mathrm{F}}$ of the dark companion, semidetached, and contact binary light curves in the validation sets of the synthetic samples W0C0 (a), W100C0 (b), W0C10L50 (c), and W100C10L50 (d). We describe the synthetic samples in Sect. 5.2 and provide the definitions of the representations in Sects. 5.3.1–5.3.2.

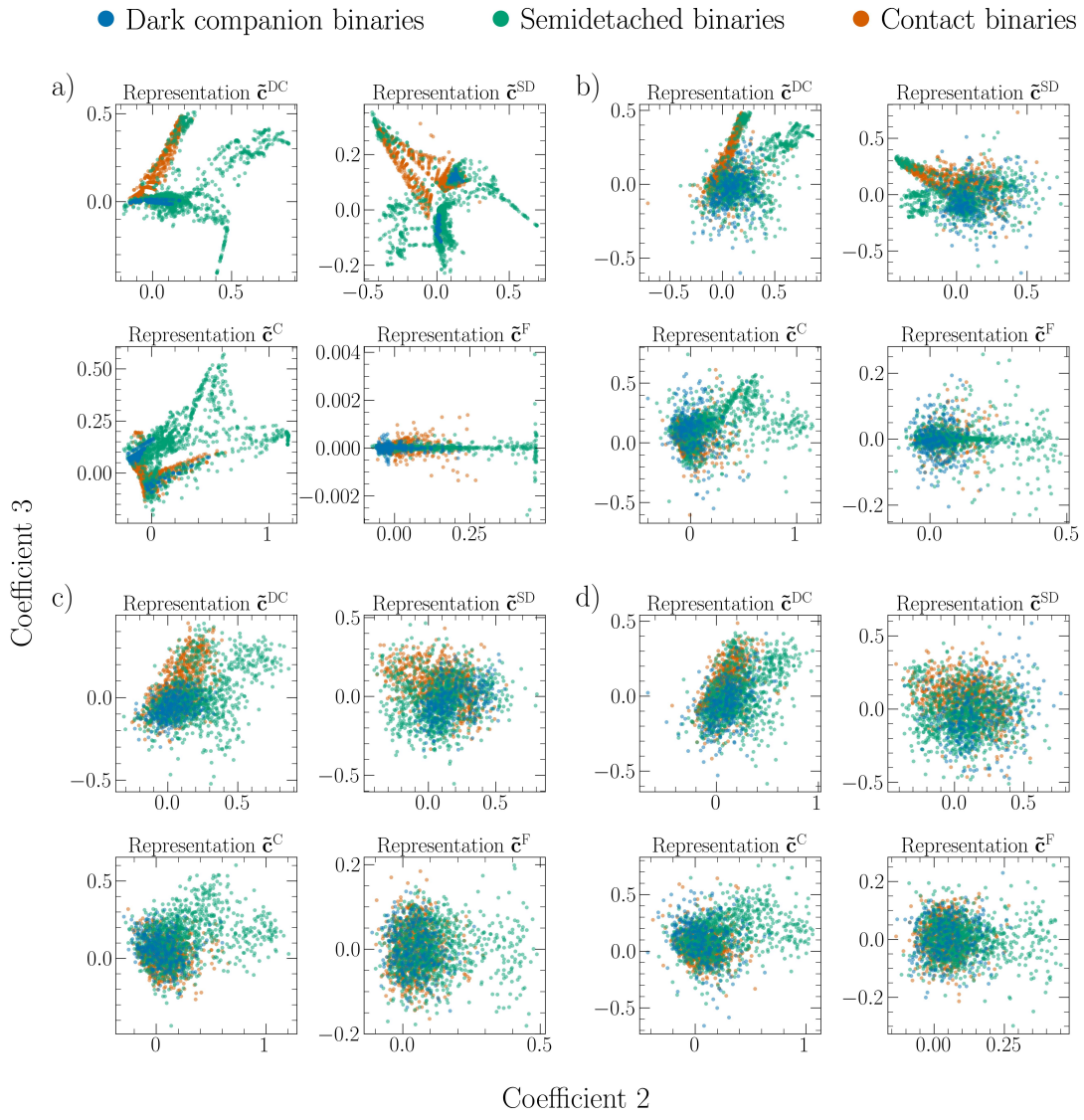

Figure 5.A.2 Scatter plots of the second and third coefficients of the representations $\tilde{\mathbf{c}}^{\mathrm{DC}}$, $\tilde{\mathbf{c}}^{\mathrm{SD}}$, $\tilde{\mathbf{c}}^{\mathrm{C}}$, and $\tilde{\mathbf{c}}^{\mathrm{F}}$ of the dark companion, semidetached, and contact binary light curves in the validation sets of the synthetic samples W0C0 (a), W100C0 (b), W0C10L50 (c), and W100C10L50 (d). We describe the synthetic samples in Sect. 5.2 and provide the definitions of the representations in Sects. 5.3.1–5.3.2.

Conclusion

The unifying theme of this thesis is the application of machine learning methods to study binary star populations through photometric observations. Until recently, this would not have been possible due to the absence of suitable data analysis techniques and the lack of sufficiently large high-precision photometric samples. However, with the recent advancements in the fidelity of numerical models of binary star systems and the advent of large-scale space-based photometric surveys, such as the *Kepler* and *TESS* missions, it has become possible to study binary star populations in unprecedented detail.

To effectively extract information from binary star light curves, we require a precise physical model of the binary system capable of reproducing observed light curves with high accuracy. In this thesis, we utilized `PHOEBE`, a state-of-the-art binary star modeling software. Before we applied `PHOEBE` on a large scale, we wanted to validate its accuracy and robustness to parameterization on an example of a single binary system. By robustness to parameterization, we mean the impact of the choice of the free parameters of the model and the parametric prescriptions of the physical processes included in the model, such as reflection and limb darkening, on the inferred physical parameters of the system. To this end, we tried to recover the physical and orbital parameters of the eclipsing binary system AI Phoenicis, a relatively bright ($V = 8.6$) detached system with well-constrained parameters from the literature (Chap. 3). To study the effects of parameterization on the results, we defined a nominal model, and we systematically varied the free parameters and hyperparameters of the model one at a time, inferring the parameters of the system for each variation. We found that parameterization has little effect on the results, with the inferred parameters mostly consistent with each other and showing only a limited spread around the nominal values. However, we found that the our results are not consistent with the values reported in the literature. While there are several possible explanations for the discrepancy, we conclude that the most likely cause is the assumptions and approximations in the wrapper that `PHOEBE` uses to interface with the `ellc` forward model, which we used instead of the native `PHOEBE` backend to speed up the computation. Further investigation revealed a systematic offset in the light curves produced by the two backends, possibly explaining the offset in the inferred parameters. To avoid such discrepancies in the future, we switched to the native `PHOEBE` backend for the subsequent analyses in this thesis.

Next, we turned our attention to the problem of the estimation of the minimum mass ratio $q_{\min}$ of contact binary stars. The value of $q_{\min}$ has a significant impact on the evolution of contact binary stars, as it determines the onset of the tidal Darwin instability, leading to the rapid coalescence of the components. The minimum mass ratio is difficult to measure directly, because it requires detailed spectroscopic observations for a large sample of contact binary stars, which is difficult to obtain in practice, especially for systems close to the tidal instability. In this thesis, we presented a method for the inference of $q_{\min}$ from the observed distribution of light curve amplitudes of contact binary stars (Chap. 4). We applied this method to a sample of contact binary candidates, which we obtained by modeling the *Kepler* Eclipsing Binary Catalog as a Bayesian mixture of contact

binary stars and contaminants. This way, we assigned a probability of being a contact binary of either late or early type to each object in the sample, with the probabilities serving as weights in the subsequent analysis. We modeled the mass-ratio distribution of contact binary stars as a power-law $Q \propto q^b$ with a cutoff at $q_{\min}$. We utilized `PHOEBE` to construct the amplitude distribution of contact binaries as a function of the parameters of Q, allowing us to infer $q_{\min}$ by fitting the model to the observed distribution of light curve amplitudes from *Kepler*. Performing the analysis separately for late-type contact binaries with periods $P \leq 0.3\,\mathrm{d}$ and $P > 0.3\,\mathrm{d}$, and for early-type contact binaries with $P < 1\,\mathrm{d}$, we found $q_{\min} = 0.246^{+0.029}_{-0.046}$, $0.087^{+0.024}_{-0.015}$, and $0.030^{+0.018}_{-0.022}$, respectively, indicating a dependence of $q_{\min}$ on the structure of the components. The method that we developed can be easily extended to large samples of contact binaries from *TESS* and other space-based surveys.

Lastly, we addressed the issue of identifying dark companion binaries—ellipsoidal variables hosting electromagnetically silent black holes and neutron stars—in large photometric surveys. This is a challenging task because the light curves of dark companion binaries can be similar to those of semidetached and contact binaries, yielding high false-positive rates when not properly accounted for. To systematically study the differences between these classes, we generated synthetic light curves of dark companion, semidetached, and contact binaries (Chap. 5). We injected the light curves with various levels of uncorrelated and correlated Gaussian noise to simulate the effects of instrumental noise and starspots. We then reduced the light curves using PCA and Fourier decomposition, resulting in low-dimensional representations of the light curves. We found that the first two to five PCA components are typically enough to explain 99% of the variance in the data. We used two metrics to compare the informativeness of different representations: the silhouette score, which quantifies how similar and object is to its own class compared to other classes in a given representation, and the macro recall of random forest classifiers trained on the representations, which we interpret as the mean non-overlap of the classes in the representation space. We found that the PCA representations are generally more informative than the Fourier representation for the same number of coefficients as measured by both metrics. The macro recalls achieved by the random forest classifiers trained on the representations range from 0.97 in the complete absence of noise to 0.70 in the presence of spots and strong instrumental noise, indicating that the classes remain largely separable even under adverse observing conditions. We found that the effect of instrumental noise on our ability to distinguish between the classes is not that severe, provided that its standard deviation does not exceed $10^{-3}\,\mathrm{mag}$. In contrast, the presence of spots can significantly reduce the class separation even when they contribute as little as 1% of the light curve amplitude. Finally, we proposed a way to apply the random forest classifiers trained on synthetic data to real ellipsoidal samples, showing that we can increase the purity of a sample of dark companion candidates by a factor of up to 27 if we assume a modest prior purity of 1%. Our method is easily extendable to multi-band photometric light curves and can be further improved by explicitly modeling the correlated noise in the light curves using Gaussian processes, paving the way for a systematic search for dark companion binaries in large photometric surveys.

Bibliography

Abbott, B. P., Abbott, R., Abbott, T. D., Abernathy, M. R., & Zweizig, J. 2016, Phys. Rev. Lett., 116, 061102

Abbott, R., Abbott, T. D., Acernese, F., et al. 2023, Physical Review X, 13, 041039

Abdul-Masih, M., Prša, A., Conroy, K., et al. 2016, AJ, 151, 101

Adams, W. S. & Joy, A. H. 1919, ApJ, 49, 186

Aitkin, M. 1991, Journal of the Royal Statistical Society. Series B (Methodological), 53, 111

Antognini, J. M. O. & Thompson, T. A. 2016, MNRAS, 456, 4219

Arbutina, B. 2007, MNRAS, 377, 1635

Arbutina, B. 2009, MNRAS, 394, 501

Bai, Y., Liu, J., Bai, Z., Wang, S., & Fan, D. 2019, The Astronomical Journal, 158, 93

Batalha, N. M., Borucki, W. J., Koch, D. G., et al. 2010, ApJ, 713, L109

Bate, M. R., Bonnell, I. A., & Bromm, V. 2002, MNRAS, 336, 705

Bellman, R. 1957, Dynamic Programming (Princeton University Press)

Blagorodnova, N., Klencki, J., Pejcha, O., et al. 2021, A&A, 653, A134

Borucki, W. J. 2016, Reports on Progress in Physics, 79, 036901

Borucki, W. J., Koch, D., Basri, G., et al. 2010, Science, 327, 977

Breiman, L. 2001, Machine Learning, 45, 5

Breivik, K., Chatterjee, S., & Larson, S. L. 2017, The Astrophysical Journal Letters, 850, L13

Brown, A. G. A., Vallenari, A., Prusti, T., et al. 2021, Astronomy & Astrophysics, 650, C3

Carroll, B. W. & Ostlie, D. A. 2017, An introduction to modern astrophysics, Second Edition

Chawla, C., Chatterjee, S., Breivik, K., et al. 2022, The Astrophysical Journal, 931, 107

Chawla, C., Chatterjee, S., Shah, N., & Breivik, K. 2023, arXiv e-prints, arXiv:2310.16891

Cheung, S.-H., Villar, V. A., Chan, H.-S., & Ho, S. 2021, Research Notes of the American Astronomical Society, 5, 282

Christopoulou, P.-E., Lalounta, E., Papageorgiou, A., et al. 2022, MNRAS, 512, 1244

Claret, A. & Bloemen, S. 2011, A&A, 529, A75

Conroy, K. E., Kochoska, A., Hey, D., et al. 2020a, arXiv e-prints, arXiv:2006.16951

Conroy, K. E., Kochoska, A., Hey, D., et al. 2020b, ApJS, 250, 34

Corral-Santana, J. M., Casares, J., Muñoz-Darias, T., et al. 2016, A&A, 587, A61

Creevey, O. L., Sordo, R., Pailler, F., et al. 2022, arXiv e-prints, arXiv:2206.05864

Dalton, G., Trager, S. C., Abrams, D. C., et al. 2012, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 8446, Ground-based and Airborne Instrumentation for Astronomy IV, ed. I. S. McLean, S. K. Ramsay, & H. Takami, 84460P

D'Angelo, C., van Kerkwijk, M. H., & Rucinski, S. M. 2006, AJ, 132, 650

Darwin, G. H. 1879, Proceedings of the Royal Society of London Series I, 29, 168

de Jong, R. S., Agertz, O., Berbel, A. A., et al. 2019, The Messenger, 175, 3

Dubath, P., Rimoldini, L., Süveges, M., et al. 2011, MNRAS, 414, 2602

Eggleton, P. P. & Kiseleva-Eggleton, L. 2001, ApJ, 562, 1012

Eggleton, P. P. & Kisseleva-Eggleton, L. 2006, Ap&SS, 304, 75

El-Badry, K., Rix, H.-W., Cendes, Y., et al. 2023a, MNRAS, 521, 4323

El-Badry, K., Rix, H.-W., Latham, D. W., et al. 2024a, The Open Journal of Astrophysics, 7, 58

El-Badry, K., Rix, H.-W., Quataert, E., et al. 2023b, MNRAS, 518, 1057

El-Badry, K., Simon, J. D., Reggiani, H., et al. 2024b, The Open Journal of Astrophysics, 7, 27

Fabrycky, D. & Tremaine, S. 2007, ApJ, 669, 1298

Flannery, B. P. 1976, ApJ, 205, 217

Foreman-Mackey, D., Hogg, D. W., Lang, D., & Goodman, J. 2013, Publications of the Astronomical Society of the Pacific, 125, 306

Förster, F., Cabrera-Vives, G., Castillo-Navarrete, E., et al. 2021, AJ, 161, 242

Gaia Collaboration, Brown, A. G. A., Vallenari, A., et al. 2018, A&A, 616, A1

Gaia Collaboration, Panuzzo, P., Mazeh, T., et al. 2024, A&A, 686, L2

Gaia Collaboration, Vallenari, A., Brown, A. G. A., et al. 2022, arXiv e-prints, arXiv:2208.00211

Gallenne, A., Pietrzyński, G., Graczyk, D., et al. 2019, A&A, 632, A31

Giesers, B., Dreizler, S., Husser, T.-O., et al. 2018, Monthly Notices of the Royal Astronomical Society: Letters, 475, L15

Giesers, B., Kamann, S., Dreizler, S., et al. 2019, A&A, 632, A3

Gomel, R., Faigler, S., & Mazeh, T. 2021a, MNRAS, 504, 2115

Gomel, R., Faigler, S., & Mazeh, T. 2021b, MNRAS, 501, 2822

Gomel, R., Faigler, S., Mazeh, T., & Pawlak, M. 2021c, MNRAS, 504, 5907

Gomel, R., Mazeh, T., Faigler, S., et al. 2023, A&A, 674, A19

Green, M. J., Maoz, D., Mazeh, T., et al. 2022, arXiv e-prints, arXiv:2211.06194

Green, M. J., Maoz, D., Mazeh, T., et al. 2023, MNRAS, 522, 29

Hambálek, Ľ. & Pribulla, T. 2013, Contributions of the Astronomical Observatory Skalnate Pleso, 43, 27

Harris, C. R., Millman, K. J., van der Walt, S. J., et al. 2020, Nature, 585, 357

Hogg, D. W., Bovy, J., & Lang, D. 2010, arXiv e-prints, arXiv:1008.4686

Horvat, M., Conroy, K. E., Pablo, H., et al. 2018, The Astrophysical Journal Supplement Series, 237, 26

Hotelling, H. 1933, Journal of Educational Psychology, 24, 498

Howell, S. B., Sobeck, C., Haas, M., et al. 2014, PASP, 126, 398

Hubová, D. & Pejcha, O. 2019, MNRAS, 489, 891

Hunter, J. D. 2007, Computing in Science and Engineering, 9, 90

Hut, P. 1980, A&A, 92, 167

Hwang, H.-C. 2023, MNRAS, 518, 1750

Hwang, H.-C. & Zakamska, N. L. 2020, MNRAS, 493, 2271

Ivanova, N., Justham, S., Avendano Nandez, J. L., & Lombardi, J. C. 2013a, Science, 339, 433

Ivanova, N., Justham, S., Chen, X., et al. 2013b, A&A Rev., 21, 59

Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2020, Statistics, Data Mining, and Machine Learning in Astronomy. A Practical Python Guide for the Analysis of Survey Data, Updated Edition

Jayasinghe, T., Kochanek, C. S., Stanek, K. Z., et al. 2018, MNRAS, 477, 3145

Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2019, MNRAS, 486, 1907

Jayasinghe, T., Stanek, K. Z., Kochanek, C. S., et al. 2020, MNRAS, 493, 4045

Jenkins, J. M., Twicken, J. D., McCauliff, S., et al. 2016, in Society of Photo-Optical Instrumentation Engineers (SPIE) Conference Series, Vol. 9913, Software and Cyberinfrastructure for Astronomy IV, 99133E

Jiang, D., Han, Z., Ge, H., Yang, L., & Li, L. 2012, MNRAS, 421, 2769

Jones, D., Conroy, K. E., Horvat, M., et al. 2020, The Astrophysical Journal Supplement Series, 247, 63

Jones, M. C. 1993, in Simple boundary correction for density estimation kernel

Kapusta, M. & Mróz, P. 2023, Acta Astron., 73, 197

Katz, D., Sartoretti, P., Guerrier, A., et al. 2022, arXiv e-prints, arXiv:2206.05902

Kirk, B., Conroy, K., Prša, A., et al. 2016, AJ, 151, 68

Kirkby-Kent, J. A., Maxted, P. F. L., Serenelli, A. M., et al. 2016, A&A, 591, A124

Kobulnicky, H. A., Molnar, L. A., Cook, E. M., & Henderson, L. E. 2022, ApJS, 262, 12

Kollmeier, J., Anderson, S. F., Blanc, G. A., et al. 2019, Bulletin of the AAS, 51, https://baas.aas.org/pub/2020n7i274

Kopal, Z. 1955, Annales d'Astrophysique, 18, 379

Kopal, Z. 1959, Close binary systems

Kozai, Y. 1962, AJ, 67, 591

Kraft, R. P. 1967, ApJ, 150, 551

Krtička, J., Benáček, J., Budaj, J., et al. 2024, Space Sci. Rev., 220, 24

Kuiper, G. P. 1941, ApJ, 93, 133

Li, K., Xia, Q.-Q., Kim, C.-H., et al. 2021, ApJ, 922, 122

Li, L. & Zhang, F. 2006, MNRAS, 369, 2001

Lidov, M. L. 1962, Planet. Space Sci., 9, 719

Loeb, A. & Gaudi, B. S. 2003, ApJ, 588, L117

Lucy, L. B. 1968a, ApJ, 153, 877

Lucy, L. B. 1968b, ApJ, 151, 1123

Lucy, L. B. 1973, Ap&SS, 22, 381

Lucy, L. B. 1976, ApJ, 205, 208

Luger, R., Agol, E., Foreman-Mackey, D., et al. 2019, AJ, 157, 64

Luger, R., Foreman-Mackey, D., & Hedges, C. 2021a, AJ, 162, 124

Luger, R., Foreman-Mackey, D., Hedges, C., & Hogg, D. W. 2021b, AJ, 162, 123

MacLeod, M., Macias, P., Ramirez-Ruiz, E., et al. 2017, ApJ, 835, 282

Mahy, L., Sana, H., Shenar, T., et al. 2022, A&A, 664, A159

Matijevič, G., Prša, A., Orosz, J. A., et al. 2012, AJ, 143, 123

Maxted, P. F. L. 2016, A&A, 591, A111

Maxted, P. F. L., Gaulme, P., Graczyk, D., et al. 2020, MNRAS[arXiv:2003.09295]

Mochnacki, S. W. 1981, ApJ, 245, 650

Muller, G. & Kempf, P. 1903, ApJ, 17, 201

Murray, C. D. & Dermott, S. F. 1999, Solar System Dynamics

Nagarajan, P., El-Badry, K., Rodriguez, A. C., van Roestel, J., & Roulston, B. 2023, MNRAS, 524, 4367

Naoz, S. 2016, ARA&A, 54, 441

Nelder, J. A. & Mead, R. 1965, The Computer Journal, 7, 308

Paczyński, B., Sienkiewicz, R., & Szczygieł, D. M. 2007, MNRAS, 378, 961

Paczyński, B., Szczygieł, D. M., Pilecki, B., & Pojmański, G. 2006, MNRAS, 368, 1311

Pawlak, M. 2016, MNRAS, 457, 4323

Pawlak, M., Pejcha, O., Jakubčík, P., et al. 2019, MNRAS, 487, 5932

Pawlak, M., Soszyński, I., Udalski, A., et al. 2016, Acta Astron., 66, 421

Pearson, K. 1901, Philosophical Magazine, 2, 559

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, Journal of Machine Learning Research, 12, 2825

Pejcha, O. 2014, ApJ, 788, 22

Pejcha, O., Metzger, B. D., & Tomida, K. 2016a, MNRAS, 461, 2527

Pejcha, O., Metzger, B. D., & Tomida, K. 2016b, MNRAS, 455, 4351

Pejcha, O., Metzger, B. D., Tyles, J. G., & Tomida, K. 2017, ApJ, 850, 59

Pešta, M. & Pejcha, O. 2023, A&A, 672, A176

Popov, V. A. & Petrov, N. I. 2022, New A, 97, 101862

Portegies Zwart, S. F., Verbunt, F., & Ergma, E. 1997, A&A, 321, 207

Pribulla, T. & Rucinski, S. M. 2006, AJ, 131, 2986

Price-Whelan, A. M., Hogg, D. W., Rix, H.-W., et al. 2018, AJ, 156, 18

Prša, A., Conroy, K. E., Horvat, M., et al. 2016, The Astrophysical Journal Supplement Series, 227, 29

Prša, A., Guinan, E. F., Devinney, E. J., et al. 2008, The Astrophysical Journal, 687, 542

Prša, A., Batalha, N., Slawson, R. W., et al. 2011, AJ, 141, 83

Prša, A., Conroy, K. E., Horvat, M., et al. 2016, ApJS, 227, 29

Prša, A. & Zwitter, T. 2005, ApJ, 628, 426

Rasio, F. A. 1995, ApJ, 444, L41

Remillard, R. A. & McClintock, J. E. 2006, ARA&A, 44, 49

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2015, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 014003

Ricker, G. R., Winn, J. N., Vanderspek, R., et al. 2014, Journal of Astronomical Telescopes, Instruments, and Systems, 1, 1

Robertson, J. A. & Eggleton, P. P. 1977, MNRAS, 179, 359

Rousseeuw, P. J. 1987, Journal of Computational and Applied Mathematics, 20, 53

Rowan, D. M., Thompson, T. A., Jayasinghe, T., Kochanek, C. S., & Stanek, K. Z. 2024, The Open Journal of Astrophysics, 7, 24

Rucinski, S. M. 1973, Acta Astron., 23, 79

Rucinski, S. M. 1994, PASP, 106, 462

Rucinski, S. M. 1997, AJ, 113, 1112

Rucinski, S. M. 2001, AJ, 122, 1007

Rucinski, S. M. 2004, New A Rev., 48, 703

Rucinski, S. M. 2007, MNRAS, 382, 393

Russell, H. N., Fowler, M., & Borton, M. C. 1917, ApJ, 45, 306

Shenar, T., Sana, H., Mahy, L., et al. 2022, Nature Astronomy, 6, 1085

Shu, F. H. & Lubow, S. H. 1981, ARA&A, 19, 277

Shu, F. H., Lubow, S. H., & Anderson, L. 1976, ApJ, 209, 536

Shu, F. H., Lubow, S. H., & Anderson, L. 1979, ApJ, 229, 223

Shvartzvald, Y., Waxman, E., Gal-Yam, A., et al. 2024, ApJ, 964, 74

Skarka, M., Žák, J., Fedurco, M., et al. 2022, A&A, 666, A142

Soszyński, I., Pawlak, M., Pietrukowicz, P., et al. 2016, Acta Astron., 66, 405

Southworth, J., Maxted, P. F. L., & Smalley, B. 2004, MNRAS, 351, 1277

Stępień, K. 2004, in IAU Symposium, Vol. 219, Stars as Suns : Activity, Evolution and Planets, ed. A. K. Dupree & A. O. Benz, 967

Stępień, K. 2009, MNRAS, 397, 857

Stępień, K. 2011, in Magnetic Stars, 86–103

Stępień, K. 2006, Acta Astron., 56, 199

Stępień, K. 2011, Acta Astron., 61, 139

Stępień, K. & Gazeas, K. 2012, Acta Astron., 62, 153

Strohmeier, W. 1972, Information Bulletin on Variable Stars, 665, 1

Terrell, D. & Wilson, R. E. 2005, Ap&SS, 296, 221

Thompson, T. A., Kochanek, C. S., Stanek, K. Z., et al. 2019, Science, 366, 637

Tokovinin, A. 2014, AJ, 147, 87

Tylenda, R., Hajduk, M., Kamiński, T., et al. 2011, A&A, 528, A114

van der Walt, S., Colbert, S. C., & Varoquaux, G. 2011, Computing in Science Engineering, 13, 22

Van Rossum, G. & Drake, F. L. 2009, Python 3 Reference Manual (Scotts Valley, CA: CreateSpace)

Vilhu, O. 1981, Ap&SS, 78, 401

Wadhwa, S. S., De Horta, A., Filipović, M. D., et al. 2021, MNRAS, 501, 229

Webbink, R. F. 1976, ApJ, 209, 829

Webbink, R. F. 1977, ApJ, 211, 881

Weller, M. K. & Johnson, J. A. 2023, MNRAS, 520, 935

Werner, N., Řípa, J., Thöne, C., et al. 2024, Space Sci. Rev., 220, 11

Wes McKinney. 2010, in Proceedings of the 9th Python in Science Conference, ed. Stéfan van der Walt & Jarrod Millman, 56 – 61

Yakut, K. & Eggleton, P. P. 2005, ApJ, 629, 1055

Yıldız, M. 2014, MNRAS, 437, 185

Zhao, G., Zhao, Y.-H., Chu, Y.-Q., Jing, Y.-P., & Deng, L.-C. 2012, Research in Astronomy and Astrophysics, 12, 723

Zucker, S., Mazeh, T., & Alexander, T. 2007, ApJ, 670, 1326

List of Publications

1 **Pešta, M.**, Pejcha, O., 2024, "Distinguishing between light curves of ellipsoidal variables with massive dark companions, contact binaries, and semidetached binaries using principal component analysis", 2024, submitted to A&A

2 **Pešta**, M., Pejcha, O., 2023, "Mass-ratio distribution of contact binary stars", A&A, 672, A176

3 Pejcha, O., Cagaš, P., Landri, C., Fausnaugh, M. M., De Rosa, G., Prieto, J. L., Henzl, Z., **Pešta, M.**, 2022, "The complex dynamical past and future of double eclipsing binary CzeV343: Misaligned orbits and period resonance", A&A, 667, A53

4 Korth, J., Moharana, A., **Pešta, M.**, Czavalinga, D. R., Conroy, K. E., 2021, "Consequences of parameterization choice on eclipsing binary light curve solutions", CAOSP, 51, 58-67