Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **471,** 759–769 (2017)
Advance Access publication 2017 June 9

doi:10.1093/mnras/stx1422

Robust, open-source removal of systematics in *Kepler* data

S. Aigrain,[1]★ H. Parviainen,[1,2,3] S. Roberts,[4] S. Reece[4] and T. Evans[1]

[1] *Sub-department of Astrophysics, Department of Physics, University of Oxford, Oxford OX1 3RH, UK*
[2] *Instituto de Astrofísica de Canarias, Vía Láctea s/n, E-38205 La Laguna, Spain*
[3] *Departamento de Astrofísica, Universidad de La Laguna, E-38205 La Laguna, Tenerife, Spain*
[4] *Machine Learning Research Group, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ, UK*

Accepted 2017 June 6. Received 2017 June 5; in original form 2017 March 7

ABSTRACT

We present ARC2 (Astrophysically Robust Correction 2), an open-source PYTHON-based systematics-correction pipeline, to correct for the *Kepler* prime mission long-cadence light curves. The ARC2 pipeline identifies and corrects any isolated discontinuities in the light curves and then removes trends common to many light curves. These trends are modelled using the publicly available co-trending basis vectors, within an (approximate) Bayesian framework with 'shrinkage' priors to minimize the risk of overfitting and the injection of any additional noise into the corrected light curves, while keeping any astrophysical signals intact. We show that the ARC2 pipeline's performance matches that of the standard *Kepler* PDC-MAP data products using standard noise metrics, and demonstrate its ability to preserve astrophysical signals using injection tests with simulated stellar rotation and planetary transit signals. Although it is not identical, the ARC2 pipeline can thus be used as an open-source alternative to PDC-MAP, whenever the ability to model the impact of the systematics removal process on other kinds of signal is important.

Key words: methods: data analysis – techniques: photometric – planetary systems – stars: rotation.

1 INTRODUCTION

During a little more than 4 yr of operations, the *Kepler* space mission produced continuous, high-precision light curves for over 150 000 stars, with a cadence of 29.4 min. This forms a very rich data set for a wide range of exoplanet and stellar variability studies. However, the light curves also contain instrumental artefacts and systematic trends. Correcting these while preserving 'real' astrophysical variability is challenging, but very important for the community to make the most of the *Kepler* data.

The publicly available *Kepler* data products (Thompson et al. 2016) include three versions of the time series data for each target star. The target pixel files contain flux measurements in each of the individual pixels falling within a pre-defined area around each star. These have been through low-level CCD and instrument calibrations and pixel-level cosmic ray removal, but are otherwise 'raw'. The light-curve files contain two versions of the light curve, dubbed SAP (simple aperture photometry) and PDC (pre-search data conditioning). Jenkins et al. (2010) and Fanelli et al. (2011) give an overview of the data processing steps involved in producing both the target pixel files and the light curves. The SAP light curves are obtained by summing the flux falling within a subset of the pixels included in the target pixel files, and applying a correction for background flux (Twicken et al. 2010a). The PDC light curves result from additional processing steps designed to remove instrumental artefacts and systematics (Smith et al. 2012; Stumpe et al. 2012a). As its name indicates, the PDC pipeline is primarily intended to ready the data for planetary transit searches, and is not specifically optimized to preserve other forms of astrophysical variability. The PDC data are none the less widely used for both transit searches and variability studies (for example stellar rotation studies, see e.g. McQuillan, Aigrain & Mazeh 2013a; McQuillan, Mazeh & Aigrain 2013b, 2014; Nielsen et al. 2013; Reinhold, Reiners & Basri 2013), because they are the only widely available set of light curves that (a) cover the full target list and time coverage of the *Kepler* mission and (b) are free of many of the systematics, which dominate the SAP light curves. However, some versions of the PDC pipeline were prone to overcorrection (removal of real astrophysical variability and injection of additional noise; see e.g. Roberts et al. 2013). Additionally, the PDC pipeline is not in the public domain, making it difficult to reproduce results based on this pipeline independently, or to perform independent evaluations of the way in which it affects different types of stellar and planetary signals (as required, for example, for planet incidence studies). This motivated us to develop an alternative procedure to correct systematic trends and artefacts in

★ E-mail: suzanne.aigrain@physics.ox.ac.uk

the *Kepler* data, aiming to match (or improve on) the performance of the PDC pipeline, but with a specific emphasis on retaining real astrophysical variability, and a commitment to making the code publicly available.

The *Kepler* detector consists of 21 modules arranged in a 5 × 5 grid with missing corners, each containing two 4K × 2K CCDs (van Cleve et al. 2009). The two halves of each CCD are read out separately, leading to four output channels per module. Throughout this paper, we refer to each module plus output channel combination as a modout and identify it as $X.Y$, where X is the module number (2–24) and Y the output channel number (1–4). For the tests described in this paper, we focused on four modules: 2.1, 7.3, 13.1 and 17.2. These were selected because they are collectively representative of the range of systematic effects, which affect the *Kepler* data: 2.1 is located near the upper left corner of the field of view and is particularly sensitive to focus changes, 7.3 is 'atypically typical', 13.1 is at the centre of the focal plane and 17.2 contains some peculiar image artefacts (J. Smith, private communication).

Every 3 months, the *Kepler* satellite rolls by 90° about its boresight in order to keep its solar panels pointing towards Sun. Each 3-month period between rolls is known as a quarter, and we follow the usual convention of referring to each quarter as Q0, Q1, Q2, etc. As each star falls on a different modout in each quarter, the systematics observed vary from quarter to quarter, and their treatment is best carried out separately for each quarter. Each star returns to approximately the same position on the detector every four quarters (1 yr), so the systematics observed in a given star's light curve in quarters (say) 3 and 7 tend to be mutually similar.

1.1 Systematics removal in the PDC pipeline

In this subsection, we give a brief overview of the systematics removal methods implemented in the pipeline that produces the publicly available PDC light curves. This is a very simplified description, intended only to set the scene for this paper, and a number of important features have been omitted for the sake of brevity; full details are given in the relevant publications.

The standard approach for systematic trend removal in transit survey data is to model each light curve as a linear combination of systematic trends, which are derived either from ancillary engineering and meteorological data (such as telescope pointing, focus, seeing and airmass; see e.g. Bakos et al. 2007) or from the light curves themselves (Kovács, Bakos & Noyes 2005; Tamuz, Mazeh & Zucker 2005). In the latter case, variants of principal component analysis are used to construct a reduced basis from the light curves.

The PDC pipeline broadly follows this paradigm. In its original version (Twicken et al. 2010a,b), known as PDC-LS (for least squares), the basis was constructed from ancillary engineering data and the coefficient relating each systematic trend (each term in the basis) to each light curve was derived by least-squares (or maximum likelihood) fitting. However, this method suffered from two problems: overfitting (removal of real astrophysical variability) and injection of noise into the light curves (a side effect of overfitting combined with noisy basis vectors).

To address these issues, a new version of the PDC was introduced, known as PDC-MAP (for maximum a posteriori; Smith et al. 2012). In the PDC-MAP, the systematic trends (which are known, in *Kepler* jargon, as co-trending basis vectors, or CBVs) are computed from the light curves themselves, by applying singular value decomposition (SVD) to 50 per cent of the light curves that show the strongest mutual correlation. This is done separately for each of the four output channels in each of the 21 modules composing the *Kepler* detector. The coefficients linking each CBV to each light curve are then evaluated in a two-step process. First, preliminary estimates are computed using the same least-squares method as in PDC-LS. The resulting coefficients are used to construct prior distributions that are parametrized according to stellar magnitude and position on the sky, and the final coefficients are found by maximizing the marginal likelihood for each star, subject to these priors.

The PDC-MAP pipeline performs significantly better than the PDC-LS one, and it is less susceptible to overfitting and noise injection, although these effects do remain apparent in some of the light curves. To address the small number of cases where the PDC-MAP pipeline performance was not entirely satisfactory, a third version, known as PDC-msMAP ('ms' stands for 'multiscale'), was developed by Stumpe et al. (2014). This version uses a wavelet transform to split the light curves into three channels, or bands, which are processed separately, resulting in a more effective separation of intrinsic and common-mode signals, and improved removal of the systematics. The final version of the *Kepler* data (Data Release 25), now available at MAST, uses the PDC-msMAP pipeline, and will be used for comparison throughout this paper, though for the sake of brevity, we hereafter refer to the PDC-msMAP pipeline and data using the shorthand 'PDC'.

Fig. 1 shows a few representative examples of SAP light curves from Q3 (in grey), together with the correction applied by the PDC pipeline, and the resulting PDC light curve (in cyan). (The magenta curves refer to the new ARC2 correction, which is presented later in this paper.) The top two rows show a relatively quiet star and a short time-scale (<5 d) variable, respectively. In those cases the PDC pipeline performs as well as one could expect. The bottom two rows show two longer time-scale (>5 d) variables. Those illustrate the limitations of even the latest version of the PDC pipeline. The correction adds broad-band noise into the light curves, and overcorrects the intrinsic variability. Of course, we do not know the ground truth, since these are real examples, but we can say this with some confidence, because the correction is unlike that applied to the vast majority of other light curves, but instead mimics variations in the individual light curves, which are typical of evolving, rotating starspots. One of the design considerations for our own pipeline was to try to reduce the incidence of these problems even further, without sacrificing the otherwise excellent overall performance.

The overall performance of both the PDC and our own pipeline is illustrated in a statistical manner in the top panel of Fig. 2. This figure shows various estimates of the light curve scatter on different time-scales as a function of magnitude, in the form of ratios between the PDC, SAP and our own correction, for four representative channels in Q3. The left-hand column shows the point-to-point (p2p) scatter σ, measured as the standard deviation of the first difference of each light curve. Using the first difference ensures that any long-term trends are taken out of the equation, giving an estimate of the high-frequency noise level alone. The middle column shows the range R, first introduced by Basri et al. (2010, 2011) as a relatively noise-independent measure of light-curve amplitude, and defined as the interval between the 5th and the 95th percentiles of the normalized[1] flux values. R captures variations on a wide range of time-scales, including both systematics and intrinsic variability. Finally, the right-hand column shows the

[1] Throughout this paper, all light curves are assumed to be normalized by dividing them by the median flux value.

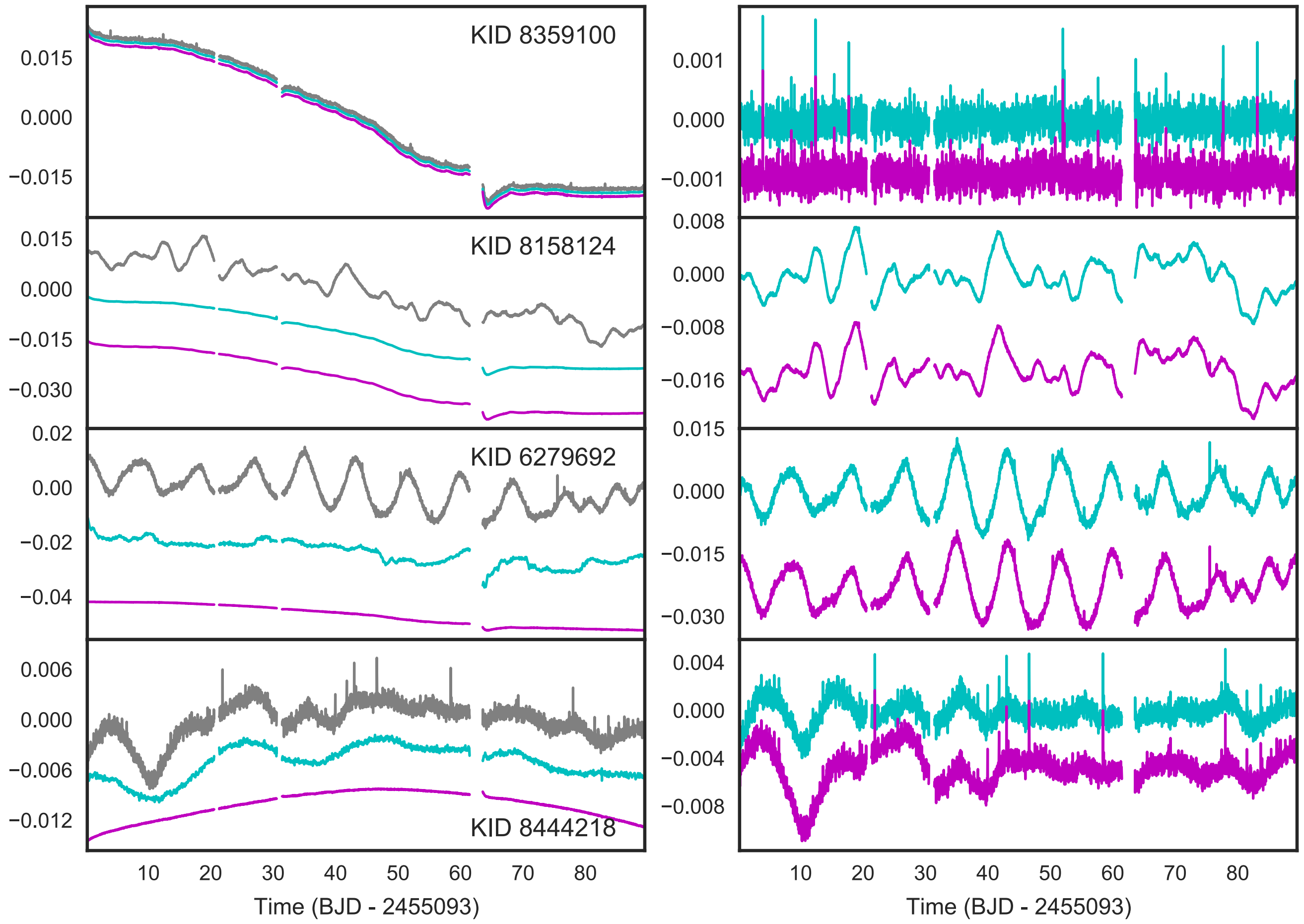

Figure 1. Example light curves before and after systematics correction (random selection from Q3, modout 7.3). The left column shows the SAP light curve for each object in grey, with the corrections applied by the PDC-MAP and by our own pipeline in cyan and magenta, respectively. The right column shows the corrected light curves (PDC-MAP in cyan, this work in magenta). In both columns, arbitrary vertical offsets have been applied to separate the different curves. Both corrections are almost identical in the top-two cases, but the PDC-MAP correction adds considerably more broad-band noise and overcorrects the intrinsic variability in the bottom two examples (see the text for details).

6.5-h combined differential photometric precision (CDPP)[2], which is widely used to evaluate the noise level on transit time-scales.

At the time this work was initiated, the PDC-msMAP results were not yet available for most quarters, and it was noticeable that the PDC-MAP pipeline introduced significant amounts of high-frequency noise in a relatively large fraction of the light curves. The top-left panel of Fig. 2 shows that this problem has been significantly reduced in the PDC-msMAP pipeline. On average, the latter leaves σ unchanged, and even reduces it in certain channels, with only a few individual exceptions. As most light curves are initially systematics-dominated, so the PDC correction should – and does – reduce R significantly in most cases, as illustrated in the top-middle panel, but it is hard to tell whether any of the reduction in R results from the removal of real variability. Finally the top-right panel shows that the 6.5 h CDPP is more or less unaffected on average, and even reduced slightly in some channels, but it is interesting to note that it is significantly increased (by 10 per cent or more) for a non-negligible fraction of the cases (about 10 per cent for these four particular channels in Q3).

1.2 An open-source systematics removal pipeline matching the PDC performance

As explained above, we seek to develop an alternative systematics-correction pipeline for *Kepler*, which comes as close as possible to matching the excellent performance of the PDC pipeline, but for which the code and the input data are all publicly available, to enable independent estimates of the impact of this pipeline on different kinds of astrophysical signals of interest. Furthermore, it would be desirable to reduce the extent to which longer term astrophysical variability is affected, and further minimize the injection of short-term noise into the corrected light curves.

In Roberts et al. (2013, hereafter Paper I), we presented an alternative method to identify and correct common-mode systematics (trends present, to a greater or lesser extent, in the majority of light curves), which we called the ARC (astrophysically robust correction) method, as it was specifically designed to minimize the risk

[2] Throughout this paper, we use our own proxy estimate of the CDPP, as defined in Aigrain, Parviainen & Pope (2016). The true CDPP is estimated by the *Kepler* transit search pipeline and published together with the PDC light curves, but we have chosen to compute our own proxy for it so that it can be compared directly to the same quantity for the light curves processed with our own pipeline.

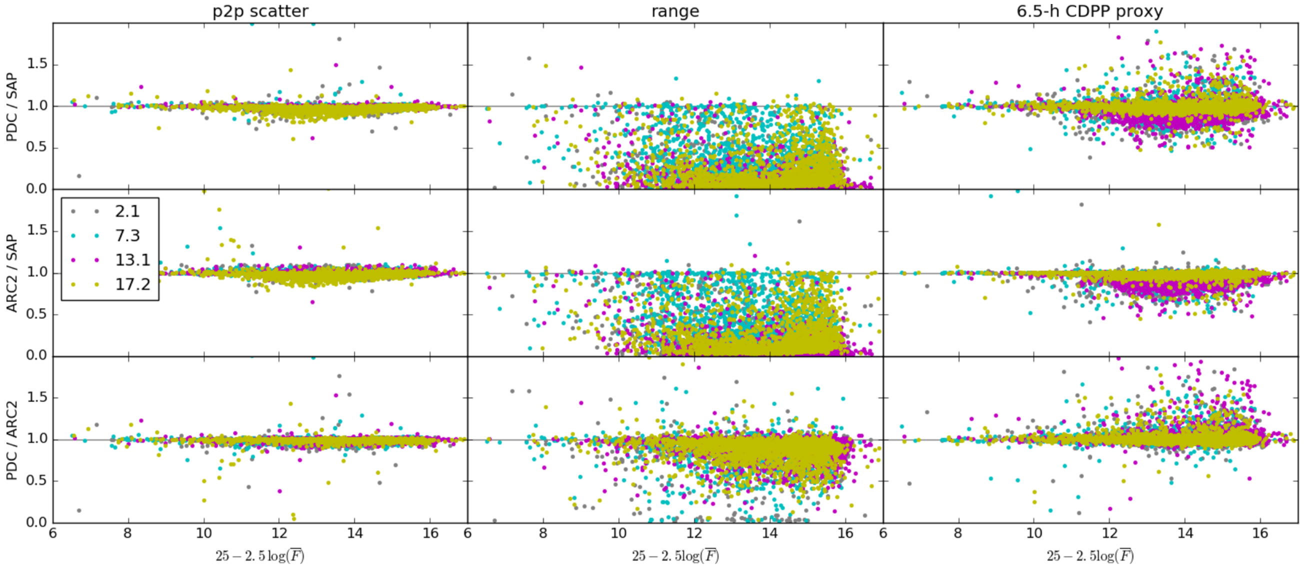

Figure 2. Statistical comparison of the SAP, PDC-MAP and ARC2 light curves for Q3. The left, middle and right columns show the short-term scatter, range and 6.5-h CDPP proxy, respectively, while the top, middle and bottom rows show the ratio of these quantities between PDC and SAP, ARC2 and SAP, and PDC and ARC2 light curves, respectively. In all cases, the x-axis shows a magnitude-like quantity based on the median flux in each light curve, but with a very approximate zero-point. The different colours correspond to different modouts. The detailed behaviour of the different modouts is discussed in Section 4.2.

of removing or altering astrophysical variability along with the instrumental systematics. Like the PDC, the ARC is based on decomposing each light curve into a linear superposition of systematic trends, plus a unique vector representing intrinsic variability and random noise. Like in the PDC-MAP and msMAP versions, the systematic trends are identified from the light curves themselves, for each output channel, although the procedure for doing so is slightly different (the ARC uses an information entropy criterion to identify genuinely systematic trends). Once identified, the trends are smoothed before they are applied to the individual light curve. Finally, adaptive, zero-mean (shrinkage) priors were used to evaluate the coefficients linking each systematic trend to each light curve, in an effort to reduce the risk of overfitting.

Since Paper I, which included tests on Q1 data only, we have applied the ARC to later quarters, and it has become apparent that the ARC trend identification method results in basis vectors that are usually extremely similar to the CBVs. We thus concluded that it may be more expedient to use the published CBVs than to derive our own (especially as the trend identification is by far the most CPU intensive part of the ARC pipeline). Using the same basis vectors as the PDC also maximizes the chances that the light curves corrected with our pipeline are as similar as possible to the PDC version, whenever that is desirable. Furthermore, the smoothing method used in Paper I (empirical mode decomposition; Huang et al. 1998) cannot adequately reproduce the discontinuities and sharp decays that affect essentially all light curves, following each (approximately monthly) data down-link event. On the other hand, during our initial tests of the ARC on quarters 2–5, we noted that both the first few CBVs, and the raw ARC basis vectors, contain some high-frequency structure, which appears to be real, in the sense that it is also present in the light curves of bright stars (where the photon noise does not mask these effects). The origin of this high-frequency structure is not clear, but its existence implies that smoothing the basis vectors may not be a good idea after all.

This leaves us with the trend removal section of the ARC, which is essentially an alternative to the MAP portion of the PDC-MAP process. Both pipelines use priors over the coefficients relating each basis vector to each light curve, but instead of constructing those by parametrizing the distribution of these coefficients as a function of (e.g.) a star's magnitude and location on the detector, the ARC approach consists in starting with a zero-mean Gaussian prior over each coefficient, and applying a Gamma hyper-prior on the width of that Gaussian. Unless there is particularly strong support in the data for a particular basis vector, the prior for the corresponding coefficient will rapidly shrink to zero (hence the name 'shrinkage prior'). This is intended to minimize the risk of overfitting, which leads to the noise-injection and variability suppression problems we noted in some PDC light curves. This trend removal procedure is implemented in a variational inference framework (see the appendix of Paper I for full details), which is extremely fast: given a set of basis vectors, correcting a given light curve for a given quarter takes order of a second, and all the light curves (for all channels) less than half an hour on our 96-core linux cluster.

Our new pipeline is thus essentially a stripped-down version of the original ARC, retaining the variational Bayes (VB) framework and shrinkage priors, but using the pre-computed CBV basis vectors. Prior to the CBV correction, however, we must include a jump-correction step. Many light curves contain one or two isolated jumps, or discontinuities, which are suspected to be due to individual pixel malfunction. They are identified and corrected as part of the PDC pipeline, but the published data products do not provide enough information to separate this from the systematic trend correction, so it was necessary for us to develop our own jump correction. Together, our new jump correction and CBV correction method form the ARC2 pipeline.

The rest of this paper is structured as follows. In Section 2, we introduce and test our new jump correction correction. In Section 3, we describe our method for applying the CBVs to the light curves, and present a simple way of deciding, for each light curve, how many CBVs should be used in the correction. This section also includes links to the corrected data and the code used to produce them. Finally, in Section 4, we evaluate the performance of our method using injection tests, and compare it to the PDC method.

2 DETECTION AND CORRECTION OF INDIVIDUAL DISCONTINUITIES

Prior to correction using CBVs, it is vital to remove and correct for the numerous discontinuities present in the light curves. This sudden pixel sensitivity dropout (SPSD) correction is divided into (a) detection, (b) classification and (c) correction of discontinuities. The detection phase identifies discontinuities in the light curve, the classification phase identifies the type of the discontinuity (SPSD, flare or transit) and the correction phase corrects the SPSDs but leaves flares and transits untouched. Each of these steps is detailed below.

Simple approaches to discontinuity detection such as convolving a step function with the light curve (matched filter) or looking for outliers in the difference between consecutive data points (or between every other data point) perform very well in white noise and provided there are very few data gaps. However, we were keen to develop a method that performed well even for light curves containing significant amounts of short-term variability. We therefore opted for an approach based on Gaussian Process (GP) regression with a change-point built into the covariance matrix. Early in the process of developing this method, we carried out simple tests on example light curves, which confirmed that this approach outperforms the matched filter and first- or second-difference approaches for variable stars, while performing equivalently well for quiet stars.

The discontinuity detection is based on a moving window likelihood ratio test between with- and without-discontinuity GP models. The light curve, $\boldsymbol{f}$, is modelled as a zero-mean GP with a covariance matrix K, that is

$$\boldsymbol{f} \sim \mathcal{N}(0, \mathsf{K}). \tag{1}$$

The covariance matrix elements are defined by a covariance function (kernel) with the light-curve cadences $\boldsymbol{c}$ as the input parameter. We use two kernels, without and with a breakpoint (K_0 and K_1, respectively), defined as

$$K_{0,ij} = a^2 \exp\left(-\frac{|c_i - c_j|}{\lambda}\right) + \sigma^2\delta_{ij}, \tag{2}$$

$$K_{1,ij} = a^2 \exp\left(-\frac{|c_i - c_j|}{\lambda}\right) \times B(\beta, c_i, c_j) + \sigma^2\delta_{ij}. \tag{3}$$

The kernels represent the light curve as a sum of an exponential kernel with an output scale a and input scale λ and an average white noise term. The with-breakpoint kernel, K_1, includes a breakpoint function

$$B(\beta, c_i, c_j) = \begin{cases} 1 & \text{if } \left(c_i \leq \beta \wedge c_j \leq \beta\right) \vee (c_i \geq \beta \wedge c_j \geq \beta) \\ 0 & \text{else} \end{cases}$$

that forces the covariances between the points on the different sides of a given breakpoint cadence, β, to zero.

The discontinuity search starts with the removal of strong individual outliers identified using a narrow running median filter. Next, we learn the K_0 hyperparameters a, λ, and σ by maximizing the GP likelihood for a subset of the light curve. Then, we fix the hyperparameters to the optimized values, and calculate K_1 log likelihoods, $\log \mathcal{L}_1$, for each $\beta = [0...N]$ using a n_p cadence-wide moving window centred around β (where $n_\mathrm{p} = 150$ by default). We also calculate K_0 log likelihoods, $\log \mathcal{L}_0$, for each window, and subtract these from $\log \mathcal{L}_1$ to obtain a series of log likelihood ratios, $\log(\mathcal{L}_1/\mathcal{L}_0)$. Finally, we identify discontinuities as positive outliers in the log likelihood ratio, as illustrated in Fig. 3.

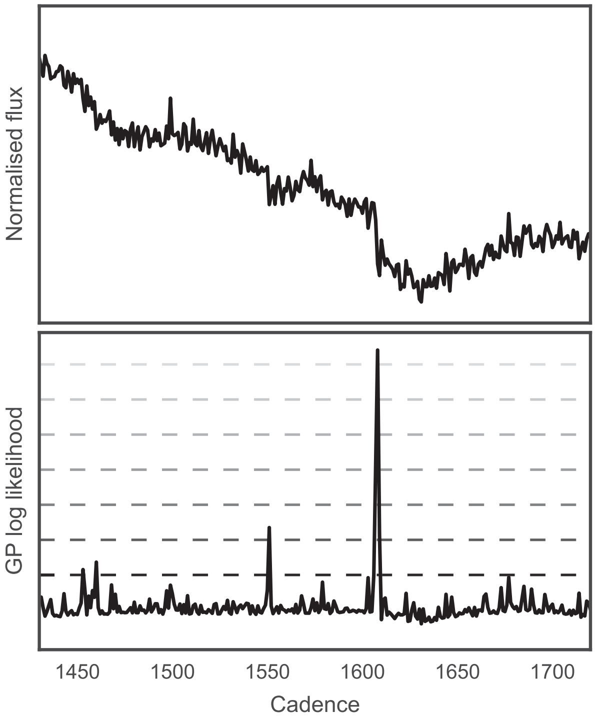

Figure 3. An example illustrating the discontinuity search: a light curve containing long-period variability, white noise, correlated noise, outliers and two SPSD signals (top), and the GP log likelihood difference between the with-breakpoint and without-breakpoint GP kernels, and the 10–80 MAD (median absolute deviation) levels for the log likelihood distribution marked as horizontal slashed lines (bottom).

The *Kepler* light curves contain several types of signals that can cause rapid changes in the flux. Besides the instrumental SPSDs that we want to find and correct, we also have astrophysical signals – such as planetary transits, binary eclipses and stellar flares – that we do not want to remove. Thus, an automated SPSD correction routine needs a way to distinguish an SPSD from these other discontinuities. The discontinuity classification uses a Bayesian information criterion (BIC)-based model selection approach to select between a set of possible discontinuity models. The current code implements five models: (a) false alarm modelled by a low-order polynomial, (b) transit-like, (c) flare, (d) SPSD and (e) SPSD followed by an exponential drift to a new level. The discontinuity classification phase fits the models to each discontinuity, modelling the flux baseline with either a low-order polynomial or a GP, and selects the model with the lowest BIC value as the true model, as shown in Fig. 4.

Finally, in phase (c), we correct the identified SPSDs based on the fitted discontinuity model and save the information about the identified discontinuities for later use.

3 OPTIMIZED USE OF THE CBVS

3.1 CBV fitting using variational Bayes

We fit each light curve using the standard linear basis model:

$$F_j^{(i)} = \sum_{k=1}^{K} w_k^{(i)} C_{kj} + \epsilon_j^{(i)} \tag{4}$$

where $F_j^{(i)}$ is the flux measured for star i in observation j, C_{kj} is the value of the kth basis vector (systematic trend, or CBV) in

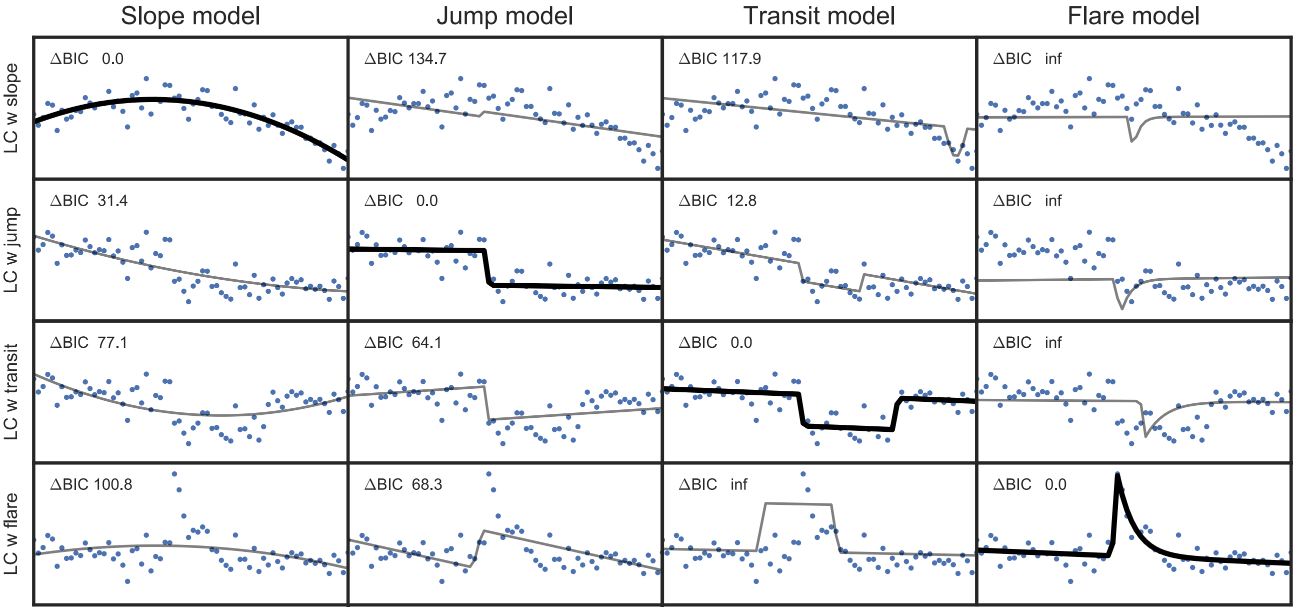

Figure 4. Discontinuity classification. Each row shows a light curve with a discontinuity (blue dots), and each column shows a discontinuity model fitted to the light curve (black line) with the model's BIC value relative to the lowest BIC value of the fitted models. The model with the lowest BIC value (thick black line) is chosen to represent the discontinuity type. The infinite Δ BIC values arise when the priors for the model parameters reject the model directly, such as with flares with negative amplitude.

observation j, $w_k^{(i)}$ is the coefficient, or weight, relating basis vector k to light curve i and $\epsilon_j^{(i)}$ represents the residuals of the correction for star i in observation j. In the remainder of this section, we omit the superscript $^{(i)}$ for simplicity – the analysis is done separately for each light curve. Note that ϵ contains both intrinsic variability and noise, representing the total residual for the purposes of the systematics correction.

In a simple least-squares framework (such as PDC-LS), one seeks the set of ws that minimizes the total squared residuals, $\sum_j \epsilon_j^2$. If the residuals are assumed to be drawn independently from a Gaussian distribution with known precision (inverse variance) β, this is equivalent to maximizing the likelihood of the model:

$$p(\mathbf{F}|\mathbf{w}, \mathbf{C}) = \mathcal{N}(\boldsymbol{\epsilon}; 0, \beta^{-1}\mathbf{I}). \tag{5}$$

(Note that strict equivalence only holds if β is known, if it is a free parameter then the full likelihood expression must be used.)

Maximum likelihood linear basis models are notoriously prone to overfitting. Better results can be obtained in a Bayesian framework, by using priors over the ws to encapsulate any external information available over the expected values for the weights, and maximize the posterior distribution instead of the likelihood.

$$p(\mathbf{w}|\mathbf{F}, \mathbf{C}) = \frac{p(\mathbf{F}|\mathbf{w}, \mathbf{C})\, p(\mathbf{w})}{\int p(\mathbf{F}|\mathbf{w}, \mathbf{C})\, p(\mathbf{w})\, \mathrm{d}\mathbf{w}}, \tag{6}$$

where the normalization constant in the denominator is the model evidence $p(\mathbf{F}|\mathbf{C})$. In the PDC-MAP pipeline, the priors over the ws are based on the distribution of the coefficients derived in the maximum likelihood case (i.e. in the absence of priors), parametrized as a function of star position and magnitude. This reflects the belief that stars that are near each other on the detector and have similar brightnesses should also display similar systematics. As we have seen, it does reduce the overfitting problems that had been noted in PDC-LS, but does not entirely do away with them. One plausible explanation for this is that the PDC-MAP priors themselves are affected by overfitting in the initial, maximum likelihood step. Furthermore, a fixed prior, as in the MAP model, does not guarantee model shrinkage as required to avoid overfitting.

To reduce the risk of overfitting further, we perform inference using Bayesian learning, which allows us to regularize the model using priors that specifically penalize larger weights, and make it less likely that one basis vector will compensate for another. A natural choice for this is to use zero-mean Gaussian priors:

$$p(w_j|\alpha_j) = \frac{\alpha_j}{\sqrt{2\pi}} \exp\left(-\alpha_j w_j^2/2\right) \tag{7}$$

for each j (the individual prior weights are treated as mutually independent). Furthermore, we do not fix the priors, but instead treat the inverse variances, $\boldsymbol{\alpha} = \{\alpha_j\}$, as parameters themselves, subject to their own prior $p(\boldsymbol{\alpha})$, for which we use a Gamma distribution.[3] Unless there is strong evidence for a non-zero weight for a particular light curve/basis vector combination, the Gamma prior over α will tend to make the distribution over w collapse close to a delta function centred on zero, so most basis vectors will have zero weight in most light curves. This is often referred to as automatic relevance determination or shrinkage.

We then seek to evaluate the posterior distribution over the weights $\mathbf{w}$, marginalized over the prior precisions $\boldsymbol{\alpha}$:

$$p(\mathbf{w}|\mathbf{F}, \mathbf{C}) \propto \int p(\mathbf{F}|\mathbf{w}, \mathbf{C})\, p(\mathbf{w}|\boldsymbol{\alpha})\, p(\boldsymbol{\alpha})\, \mathrm{d}\boldsymbol{\alpha}, \tag{8}$$

and to maximize it with respect to $\mathbf{w}$. In general, the posterior distribution is unknown and is not analytically tractable. Numerically evaluating and optimizing the posterior would require evaluating the likelihood over a very large number of $(\mathbf{w}, \boldsymbol{\alpha})$ combinations, which is unfeasible, especially as it needs to be done for every *Kepler* light curve.

[3] Our choice of priors over $\mathbf{w}$ and $\boldsymbol{\alpha}$ is also mathematically convenient, because they are conjugate with each other and with the likelihood (which is Gaussian), enabling a number of the integrals involved in the inference to be performed analytically.

An elegant workaround consists in approximating the posterior with a proposal distribution that is analytically tractable, and iteratively refining the latter so that it approaches the true posterior. This class of methods is known as an approximate inference. More specifically, one can restrict oneself to proposal distributions that belong to the exponential family: refining the proposal then consists in optimizing an integral with respect to a functional, which is typically done using the calculus of variations. This approach is thus known as a variational inference, or VB. A detailed description of the VB method as applied to our linear basis model was given in the appendix of Paper I, so we do not repeat it here. The algorithm essentially consists in cycling through a set of update equations for the weights $\mathbf{w}$, the prior precisions $\boldsymbol{\alpha}$ and the noise precision β. Importantly, the model is guaranteed to improve each iteration, thus providing robust convergence that typically occurs after just a few iterations. The computational requirement scales as JK^2, where J is the number of observations and K is the number of basis vectors. Our PYTHON implementation of the method runs in a fraction of a second per light curve per quarter ($\sim$4300 observations) for up to eight CBVs.

3.2 How many CBVs?

Despite the measures described in the previous section to minimize the risk of overfitting, the results still depend, in some cases, on the number of basis vectors used. If our idealized model was correct (i.e. if each light curve was just a linear combination of the CBVs used, plus white Gaussian noise, and contained no intrinsic variability), this should not happen: the weights of any irrelevant CBVs included in the calculation should shrink to zero automatically. This does not happen in practice because the intrinsic variability of many of the stars, which is not included in our systematics model (or in that of the PDC-MAP, of course), is significant at the *Kepler* precision. To our knowledge, there is no simple way of overcoming this problem, and one is forced to resort to more ad hoc criteria.

The CBVs result from an SVD of a subset of the light curves, and therefore the first CBV represents a larger fraction of the overall variance of the light curves, and so on. The PDC pipeline saves 16 CBVs but only uses at most 8 to perform the correction. Additionally, a signal-to-noise ratio criterion is used to exclude CBVs that are deemed to contribute more noise than useful information (see Smith et al. 2012 for details), so the actual number of CBVs used ranges from 5 to 8. As the VB method is very fast, it is feasible to run it for every plausible number of CBVs, K (from 1 to 8). We first did this for a few representative quarters (3–6) and output channels, and performed a visual comparison of the results on a random selection of light curves.

We plotted the original (SAP) light curve, the correction applied and the corrected light curve, for a few tens of stars selected at random in each modout. A few examples are shown in Fig. 5. In most cases, the overall shape of the correction is fairly insensitive to the number of CBVs used, so long as it is at least 2 or 3. On the other hand, as the CBVs are increasingly noisy, using more CBVs introduces more noise into the light curves particularly for the brighter stars. It is therefore important to use the smallest number of CBVs that gives an adequate correction. Importantly, the examples we examined suggest that this number differs from star to star: in some cases, including the fourth and fifth CBVs removes features that appear systematic in nature, and which were not removed when

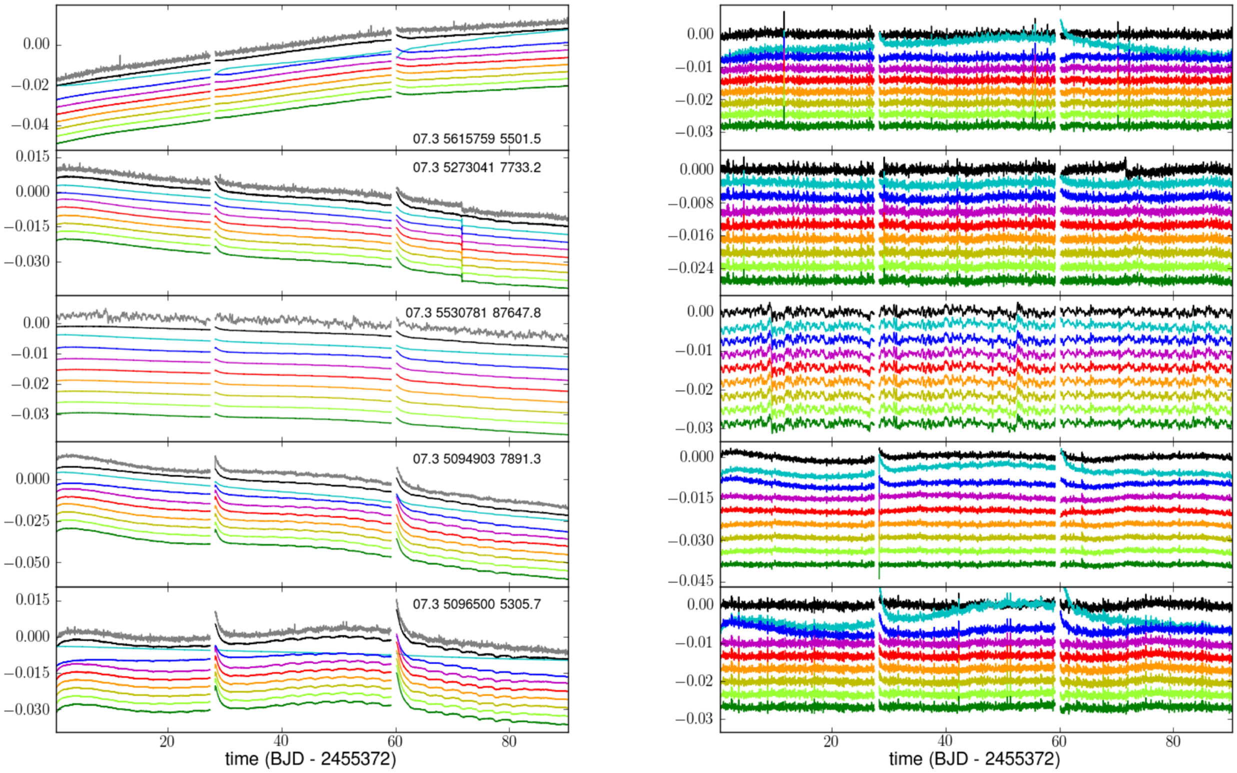

Figure 5. Comparison of the systematics correction using one to eight CBVs, for a random selection of stars in Q3, modout 17.3. The left-hand panel shows the original (SAP) light curve in grey and the corrections applied using one to eight CBVs in different colours (cyan to green, and top to bottom). The PDC-MAP correction is also shown in black, for comparison. The right-hand panel shows the corrected light curves, using the same colour coding. In both panels, the different light-curve versions have been offset by an arbitrary amount, for clarity.

using only the first two. In other cases, even using the second CBV appears to have a detrimental effect.[4]

Rather than specifying an overall value for K for, say, each quarter and modout, we therefore decided to select the optimal number of CBVs to use a posteriori, on a light curve by light curve basis, based on a statistical comparison of the light-curve properties before and after correction, using the simple statistics R and σ. If most light curves are initially dominated by systematics, the dependence of R on K tracks the extent to which the correction of systematics is improved by adding more CBVs. Typically, R decreases rapidly for low K but then reaches a plateau for larger K. Within this plateau regime, increasing K further does not usually improve the correction significantly but it does introduce more noise, so one might simply opt for the lowest value of R at which the plateau is reached. We initially define $K_{\rm opt}$ as the smallest value of K for which $R(K) < \langle R \rangle + 3\sigma_R$, where $\langle R \rangle$ and σ_R are the median and standard deviation of R over all values of K, for a given light curve.

However, when the initial amplitude of systematics is small relative to that of the intrinsic variability, this criterion alone is insufficient: R can decrease with increasing K not because more systematics are being removed, but because real variations are actually being removed. This is particularly frequent for bright, variable stars. One way of testing for this would be to inject realistic simulated variability signals into the light curve before correction, and check how well they are recovered post-correction. However, doing this for every light curve would be prohibitively expensive. We do use such injection tests to evaluate the overall ability of our correction method to preserve astrophysical signals, but only on a subset of the data (see Section 4.1).

Fortunately, there is an easier way to identify these problematic cases: visual inspection shows that when the correction removes what looks like real variability, it also introduces significant amounts of high-frequency noise into the light curves. This can be diagnosed by tracking the dependence of σ on K. Specifically, if $\sigma(K_{\rm opt})/\sigma_0 > 1.1$, where $K_{\rm opt}$ is determined from the behaviour of R as described above, and σ_0 is the short-term scatter before correction, $K_{\rm opt}$ is decreased further until the scatter ratio falls below the 1.1 threshold. The choice of threshold is somewhat arbitrary, but its exact value is not critical: when it is exceeded, it is usually by a fairly wide margin.

Fig. 6 shows histograms of the number of CBVs used for the four representative output channels from Q3 that featured in Fig. 2. Note that the number of CBVs used is never less than 1 or more than 4, and is fairly evenly distributed between those two values, but the distribution does vary noticeably from one channel to the next – as do the CBVs themselves.

3.3 Publicly available code

The code is distributed as a GPL-licenced PYTHON package OXKEPLERSC.[5] The package can be installed using standard PYTHON package installation procedures, and will be made available through PIP in the future. The code is released as an open source to encourage contributions and adaptations to specific use-cases.

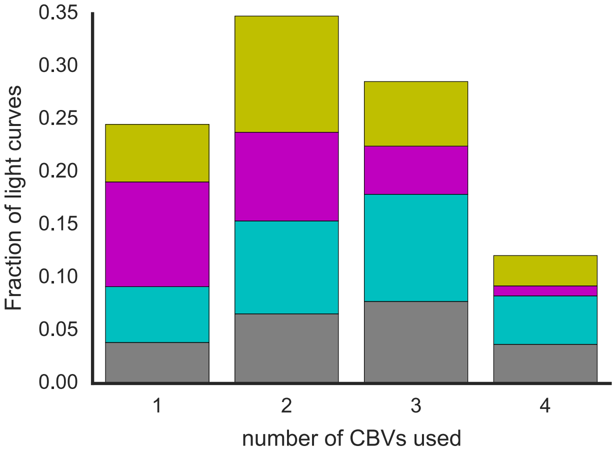

Figure 6. Distribution of the number of CBVs used for the four channels from Q3 shown in Fig. 2 (same colour-coding).

After installation, the jump detection and detrending steps can be carried out for a single file or a whole directory as

```
keplerjc path_to_file_or_dir

keplersc path_to_file_or_dir cbvdir
```

where, if given a directory, the user can optionally select a subset of modules, outputs or their combinations to process using command line arguments. The discontinuity detection and correction take several tens of seconds per light curve, and the detrending seconds. Batch processing of files is parallelized using MPI automatically if the scripts are called with MPIRUN or MPIEXEC.[6]

4 PERFORMANCE TESTS

4.1 Injection tests

Our prescription for selecting the number of CBVs used, as described in the previous section, is intended to remove systematics as effectively as possible while minimizing the risk of overfitting, i.e. removing astrophysical variability, and of introducing extra noise into the light curves due to the noisy nature of the CBVs themselves. To test the extent to which our systematics removal method affects stellar variability or transit signals, we now perform a series of injection tests. Specifically, we are interested in variability caused by rotational modulation of surface inhomogeneities such as star-spots, since this is a powerful diagnostic of stellar rotation rates and hence angular momentum evolution, as well as simulated planetary transits.

Rotation-like signals

We simulate rotation-like signals, consisting of between one and five co-added sinusoidal variations, with periods randomly drawn from a log uniform distribution ranging from 5 to 60 d, and amplitudes drawn from a log normal distribution with mean 10^{-3} and standard deviation 0.5 dex. These were added to 500 randomly selected light curves in each channel and quarter, which were then corrected for

[4] We also experimented briefly with using the CBVs in a different order than the one in which they are supplied, which results from the SVD process. However, this did not lead to any clear performance improvement. As our method is already designed to suppress irrelevant CBVs if the basis set supplied contains any, so the order of the CBVs should not matter, provided that all the relevant ones are included.

[5] Available from https://github.com/OxES/OxKeplerSC

[6] The parallelization depends on the MPI4PY package, and will be changed to use a higher level implementation with IPYTHON Parallel in the future.

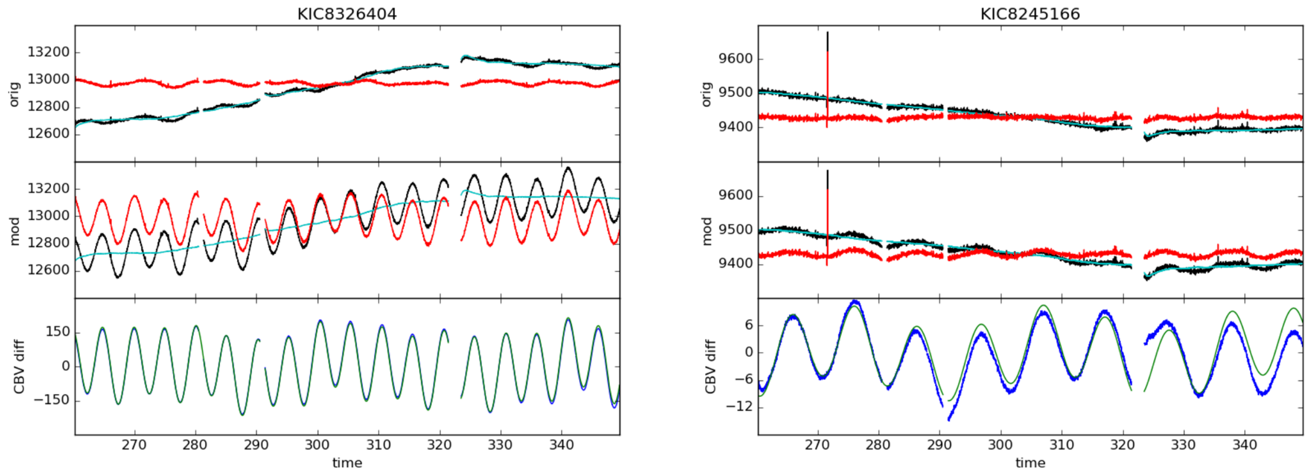

Figure 7. Injection test examples (Q3, modout 17.2). Top: original and corrected light curve (black and red, respectively) with the correction shown in cyan. Middle: same after injecting the simulated signal. Bottom: injected and recovered signals (green and blue, respectively).

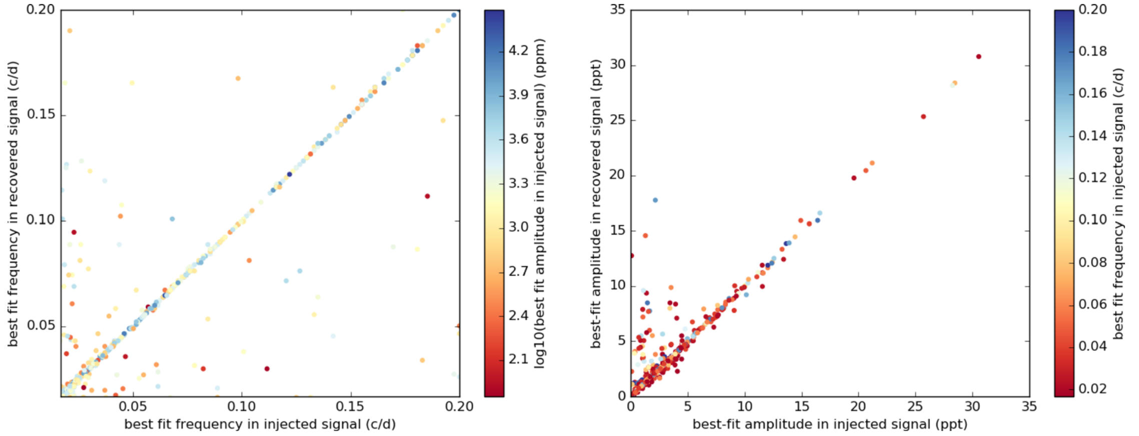

Figure 8. Sinusoid injection test results. The left-hand panel shows the frequency of the best-fitting sinusoids in the recovered versus the injected signal, colour coded according to the amplitude of the best-fitting sinusoid in the injected signal. The right-hand panel shows the amplitudes colour-coded according to frequency.

systematics as described in the previous section. The difference between the corrected light curves with and without injected signal, hereafter referred to as the *recovered* signal, is then compared to the injected signal itself: any discrepancies arise because the correction is affecting the injected signal. Fig. 7 illustrates this process for a few example light curves. Of course, the light curves into which the simulated signals are injected already contain astrophysical variability, which itself may have been affected by the correction, and this may contribute to the differences between the injected and recovered signals. However, we opted for this approach rather than attempting to generate light curves with simulated stellar signals *and* systematics, because we do not have a good generative model for the latter.

We quantify the effect of the correction on the simulated signals in several ways:

(i) by measuring the Pearson correlation coefficient P between the injected and recovered signals. We found that $P > 0.9$ for 83 per cent the time;

(ii) by measuring how much high-frequency noise the ARC2 correction adds to the recovered signal. We define the added noise as the quadrature difference between the point-to-point scatter in the recovered versus the injected signal (the latter is of course noise free but still has non-zero *measured* point-to-point scatter because of its discrete sampling). We found that the added scatter was <100 ppm 92 per cent of the time;

(iii) by comparing the dominant frequency and amplitude in the injected and recovered signals. We did this by performing a simple least squares fit of a single sinusoid over a grid of equally spaced trial frequencies and extracting the frequency and amplitude of the best-fitting sinusoid. The results are shown for a representative channel[7] and quarter in Fig. 8. Overall, the preservation of the signal is excellent: most points tightly cluster around the one-to-one line, and in over 80 per cent of the cases, the frequency error is <0.05 and the amplitude error <1 ppt. We visually examined a subset of the remainder and found that, for most of them, the sine-fitting process had identified a different component of the multisine signal in the injected and recovered versions, meaning that the relative amplitudes of the different sinusoids had been significantly altered, but the individual frequencies had not. In most of these cases, the

[7] The plots shown are for modout 17.2 in Q3, but the same tests were performed for all four of the test modouts in Q3–Q6 inclusive, and gave similar results.

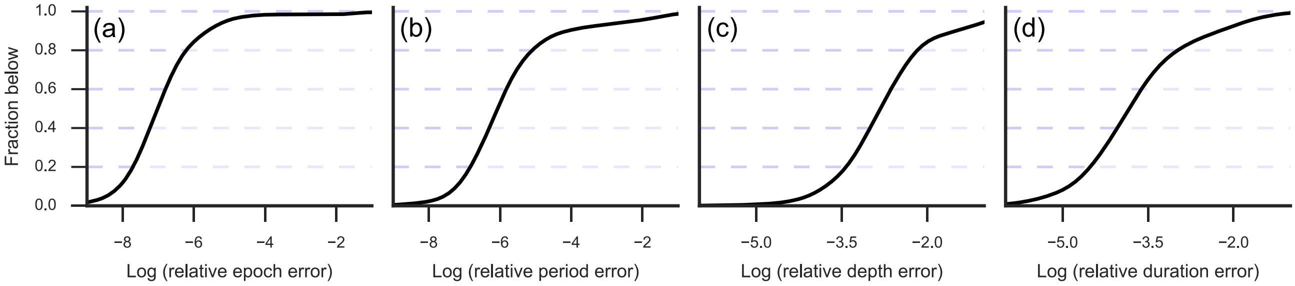

Figure 9. Transit injection test results as cumulative distributions of the relative errors (the absolute difference between the true value and the recovered value divided by the true value) for the recovered transit epoch (a), orbital period (b), transit depth (c) and transit duration (d).

dominant sinusoid in the injected signal had relatively low frequency and/or amplitude, as one would expect.

Transit signals

We carry out the transit-signal injection tests likewise to the rotation-signal injection tests. The orbital periods are drawn randomly from a log uniform distribution ranging from 5 to 60 d, the transit depths are drawn from a log normal distribution with mean 10^{-3} and standard deviation 0.5 dex, as with the rotation-signal injection test, and the transit epoch is uniformly distributed. The orbital inclination is fixed to $\frac{\pi}{2}$, semimajor axis is calculated assuming stellar density of 1.5 g cm^{-3}, and the transit duration is derived from the semimajor axis and period assuming a circular orbit. The transit signals are generated using PYTRANSIT[8] (Parviainen 2015) with six subsamples per long-cadence exposure.

We add the transit signals to a sample of 500 light curves, carry out the detrending and fit a transit model to the recovered transit signals (the difference between the detrended light curves with and without the injected signal), similarly to the rotation-like signal injection tests. The transit model is fitted using a Nelder–Mead optimizer with the injected signal parameters as starting values.

The results from the transit injection test are shown in Fig. 9 as cumulative distributions of the relative errors (absolute difference of the true and recovered values divided by the true value) for the transit epoch, orbital period, transit depth and transit duration. As with the rotation-signal injection test, we see that the detrending has only a minimal effect on the recovered parameters.

4.1.1 *Testing the jump correction*

We also used the transit injection tests to check whether the jump correction sometimes unintentionally removes planetary transits. To do this, we check what fraction of the time one of the injected transit was detected as a discontinuity, and misclassified as a jump. This happened on average in 10 per cent of the light curves, but in each light curve the jump correction never removed more than of the injected transit events. Overall, about 2 per cent of the injected transits were removed.

We also performed an additional set of injection tests using simple step function discontinuities, to test how effective the jump correction was at detecting and correcting jumps. The discontinuities were injected at random locations, with (positive or negative) amplitudes ranging from 0 to 10σ (where σ is the point-to-point scatter of the light curve). The jump correction was then applied and we checked which of the injected jumps were correctly detected, classified and corrected. The results are shown in Fig. 10. The performance improves with jump amplitude, as one would expect, but the overall sensitivity is lower than expected, and it is puzzling that even at the largest amplitude a significant fraction (10 per cent) of the injected jumps are still missed or misclassified. Most of these cases occur where a simulated jump has been inserted close to a pre-existing, real discontinuity. Our approach relies on the assumption that there are relatively few discontinuities per light curve, so that the likelihood of encountering two of them close together is very low. This assumption is appropriate for real *Kepler* light curves, but is not always satisfied in our injection tests. Therefore, the results of these tests should be taken as a lower limit to the performance of our discontinuity detection and classification method.

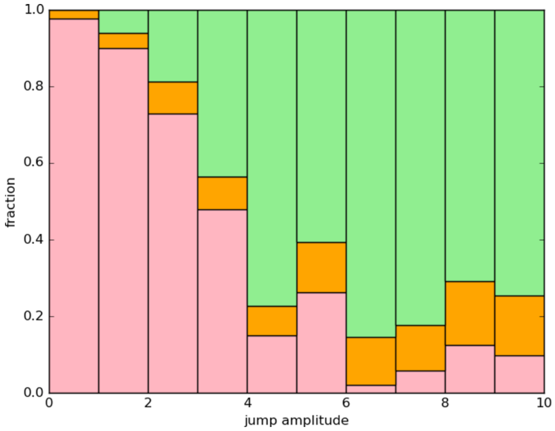

Figure 10. Jump injection test results. The x-axis shows the amplitude of the injected jumps relative to the high-frequency scatter in the light curves. In each bin, the pink, orange and green bars represent the fraction of injected jumps that were respectively missed (not detected), misclassified (and hence not corrected) and corrected.

4.2 Comparison to PDC-MAP

Figs 1 and 2 were introduced in Section 1.1 when discussing the PDC pipeline. We now return to them for the purposes of comparing the PDC to our new ARC2 pipeline. The most important remark, whether comparing individual examples as in Fig. 1, or statistics over a large ensemble of light curves, as in Fig. 2, is how similar the PDC and ARC2 corrections are. This is not altogether surprising, since both corrections are based on the same set of basis vectors, but it is reassuring: it indicates that the different choices of priors do not usually have too strong an effect on the results. It also means

[8] https://github.com/hpparvi/PyTransit

that we have succeeded in one of our initial aims, namely to match the overall performance of the PDC pipeline.

Both corrections do a good job of preserving, or even reducing, the high-frequency noise level. However, for a subset of the stars, the PDC correction (a) reduces the range R significantly more and (b) injects significantly more noise on transit time-scales than the ARC2 one, as illustrated by the bottom two rows of Fig. 1, and by the middle and right panels of the bottom row of Fig. 2. In quantitative terms, for the four test channels in Q3, the median value of $R_{\rm PDC}/R_{\rm ARC2}$ was 0.94, and the PDC pipeline increased the 6.5-h CDPP by >1.1 in 5.7 per cent of the stars, compared to only 0.3 per cent of the stars for the ARC2 pipeline. This suggests that the ARC2 correction is marginally more robust than the PDC one, though the differences are minimal.

As an aside, we note that comparing σ and the 6.5-h CDPP before and after correction can be used as a useful diagnostic of problematic light curves. For example, when the ARC2 correction increases σ by more than 10 per cent, the visual examination of the light curves often reveals abnormal behaviour, apparently associated with image artefacts, stars located near the edge of a CCD, or very high proper motion stars (where the fraction of the flux collected within the photometric aperture might change significantly during a quarter).

5 SUMMARY AND CONCLUSIONS

The injection tests and comparison to the PDC-MAP data detailed in the previous section indicate that the approach we took succeeded with respect to our key objectives: developing an open-source alternative to the PDC-MAP pipeline, which matches (or even slightly improves on) the latter's performance in terms of systematics suppression, while providing demonstrably excellent preservation of astrophysical signals such as rotational modulation of star-spots and planetary transits.

Since we have not demonstrated any significant improvement in performance over the PDC-MAP pipeline, and the PDC-MAP light curves are publicly available for the entire *Kepler* data set, it might be reasonable to ask whether publishing an alternative pipeline is worthwhile. The key difference is that our pipeline code is publicly available, in contrast to the *Kepler* pipeline. This makes it straightforward for users to test the effect of the systematics correction on whichever signals they are most interested in, using, for example, injection tests such as those described in Section 4.1. This ability is key for any study that aims to draw statistical inferences, for example measuring the distribution of rotation periods and amplitudes for stellar signals, or period and radius ratios for planetary transits.

Furthermore, the detailed comparison of the two methods, which we have presented in this paper, helps validate both of them. On the whole, we have shown that both are performing well, and if they have limitations, these are mostly common to the two methods. Having two sets of publicly available light curves (or the ability to generate them easily, in the case of the ARC2) is also valuable. If the two versions of a given light curves differ significantly, or if subsequent analysis of the two light-curve versions yields different results, this should act as a warning sign to users that something may be amiss. By flagging and examining these cases, the community may eventually help to identify the source of the discrepancy and to improve the performance of both methods.

ACKNOWLEDGEMENTS

SA was supported by grant ST/K00106X/1 from the UK Science and Technology Facilities Council, and HP by grant RPG-2012-661 from the Leverhulme Trust. The authors are grateful to the anonymous referee for their careful reading of the manuscript and helpful suggestions. Some of the data presented in this paper were obtained from the Mikulski Archive for Space Telescopes (MAST). STScI is operated by the Association of Universities for Research in Astronomy, Inc., under NASA contract NAS5-26555. Support for MAST for non-HST data is provided by the NASA Office of Space Science via grant NNX09AF08G and by other grants and contracts.

REFERENCES

Aigrain S., Parviainen H., Pope B. J. S., 2016, MNRAS, 459, 2408
Bakos G. Á. et al., 2007, ApJ, 670, 826
Basri G. et al., 2010, ApJ, 713, L155
Basri G. et al., 2011, AJ, 141, 20
Fanelli M., Jenkins J., Haas M., Gautier T., 2011, Kepler Data Processing Handbook. KSI-19081-001
Huang N. E. et al., 1998, Proc. R. Soc. A, 454, 903
Jenkins J. M. et al., 2010, ApJ, 713, L87
Kovács G., Bakos G., Noyes R. W., 2005, MNRAS, 356, 557
McQuillan A., Aigrain S., Mazeh T., 2013a, MNRAS, 432, 1203
McQuillan A., Mazeh T., Aigrain S., 2013b, ApJ, 775, L11
McQuillan A., Mazeh T., Aigrain S., 2014, ApJS, 211, 24
Nielsen M. B., Gizon L., Schunker H., Karoff C., 2013, A&A, 557, L10
Parviainen H., 2015, MNRAS, 450, 3233
Reinhold T., Reiners A., Basri G., 2013, A&A, 560, A4
Roberts S., McQuillan A., Reece S., Aigrain S., 2013, MNRAS, 435, 3639 (Paper I)
Smith J. C. et al., 2012, PASP, 124, 1000
Stumpe M. C. et al., 2012a, PASP, 124, 985
Stumpe M. C., Smith J. C., Catanzarite J. H., Van Cleve J. E., Jenkins J. M., Twicken J. D., Girouard F. R., 2014, PASP, 126, 100
Tamuz O., Mazeh T., Zucker S., 2005, MNRAS, 356, 1466
Thompson S., Fraquelli D., Van Cleve J. E., Caldwell D. A., 2016, Kepler Archive Manual. KDMC-10008-005
Twicken J. D., Chandrasekaran H., Jenkins J. M., Gunter J. P., Girouard F., Klaus T. C., 2010a, Proc. SPIE, 7740, 77401U
Twicken J. D., Clarke B. D., Bryson S. T., Tenenbaum P., Wu H., Jenkins J. M., Girouard F., Klaus T. C., 2010b, Proc. SPIE, 7740, 774023
van Cleve J., Caldwell D., Thompson R., Haas M., Koch D., Borucki W., 2009, Kepler Instrument Handbook. KSCI-19033-002

This paper has been typeset from a TEX/LATEX file prepared by the author.