S-ACF: A selective estimator for the autocorrelation function of irregularly sampled time series

Lars T. Kreutzer,[1,2,3]★ Edward Gillen,[4,2]† Joshua T. Briegal,[2] Didier Queloz[2]
[1]Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, CB3 OWA, UK
[2]Astrophysics Group, Cavendish Laboratory, J.J. Thomson Avenue, Cambridge CB3 0HE, UK.
[3]Max Planck Institute for Gravitational Physics (Albert Einstein Institute), Am Mühlenberg 1, 14476 Golm, Germany
[4]Astronomy Unit, Queen Mary University of London, Mile End Road, London E1 4NS, UK

Accepted XXX. Received YYY; in original form ZZZ

ABSTRACT

We present a generalised estimator for the autocorrelation function, S-ACF, which is an extended version of the standard estimator of the autocorrelation function (ACF). S-ACF is a versatile definition that can robustly and efficiently extract periodicity and signal shape information from a time series, independent of the time sampling and with minimal assumptions about the underlying process. Calculating the autocorrelation of irregularly sampled time series becomes possible by generalising the lag of the standard estimator of the ACF to a real parameter and introducing the notion of selection and weight functions. We show that the S-ACF reduces to the standard ACF estimator for regularly sampled time series. Using a large number of synthetic time series we demonstrate that the performance of the S-ACF is as good or better than commonly used Gaussian and rectangular kernel estimators, and is comparable to a combination of interpolation and the standard estimator. We apply the S-ACF to astrophysical data by extracting rotation periods for the spotted star KIC 5110407, and compare our results to Gaussian process (GP) regression and Lomb-Scargle (LS) periodograms. We find that the S-ACF periods typically agree better with those from GP regression than from LS periodograms, especially in cases where there is evolution in the signal shape. The S-ACF has a wide range of potential applications and should be useful in quantitative science disciplines where irregularly sampled time series occur. A Python implementation of the S-ACF is available under the MIT license.

Key words: methods: analytical – methods: statistical – methods: data analysis – stars: rotation

1 INTRODUCTION

Time series are ubiquitous throughout the experimental sciences and give insight into the temporal evolution of systems and their underlying processes. Time series in astrophysics, for example, have been instrumental in our understanding of stellar and planetary systems: stellar light and radial velocity curves yield information about the temporal evolution of processes on the stellar surface, from the longitudinal inhomogeneity of starspot distributions and magnetic field mechanisms, to the presence of orbiting bodies and material.

Historically, detecting periodicity in time series has focused on either Fourier decomposition (for regularly sampled data) or fitting sinusoidal models (for irregularly sampled data). An example of the former is the Fast Fourier Transform (FFT; Cooley et al. 1969), and examples of the latter are the standard, modified and Bayesian Lomb-Scargle periodograms (Scargle 1982; Zechmeister & Kürster 2009; Mortier & Collier Cameron 2017). While the Lomb-Scargle method can be used for arbitrary samplings, the accuracy of the estimated periods can be limited for quasi-periodic processes and evolving periodic signals due to the inherent assumption that the process be well-described by a pure sine wave of fixed period. Similar issues affect methods based on phase folding and then minimising the variance or entropy of the data, as they also rely on strict periodicity and negligible phase evolution (e.g. Stellingwerf 2011; Graham et al. 2013a,b). More recently, flexible machine learning methods, applicable to both regular and irregular time series, have been used to describe quasi-periodic variations in stellar light curves (e.g. Angus et al. 2018).

The above approaches share the same basic principle: they all fit a model to the data to determine whether periodicity is present. The concept of autocorrelation, i.e. correlating the data with itself, is a distinct 'model-free' approach that uses only the time series data to extract periodicity (e.g. Shumway & Stoffer 2017). The autocorrelation function (ACF) is a powerful definition and a reliable method to obtain information from any regularly sampled time series, as it can capture both strictly periodic and quasi-periodic processes. It has been widely used on space-based photometric data given the regular sampling available (e.g. McQuillan et al. 2013, 2014), as well as on solar data (Morris et al. 2019) for the same reason. However, the requirement of the standard ACF estimator for regularly sampled data can be a limiting factor in its broader application, e.g. for ground-based photometric data.

Previous studies have attempted to address this problem by generalising the standard ACF estimator to irregularly sampled data. A number of these methods create an approximate regularly sampled

★ E-mail: lars.kreutzer@aei.mpg.de
† Winton Fellow

time series in order to apply the standard autocorrelation estimator enhanced with rules on which terms to discard in the series. The method proposed in Lukatskaia (1975) assumes that the irregular sampling arises from missing data points in a regularly sampled series, and further assumes that statistical properties of the missing data are the same as the observed data. It is therefore possible to calculate an autocorrelation estimator using only data points which fall on to this regular sampling, with the caveat that the time series must be much longer than the variability period of the signal of interest. The method from Andronov & Chinarova (2005) interpolates onto a regular sampling grid using a smoothing function. These methods can work well if the sampling is almost regular and only a subset of values are missing from a regularly sampled time series. The Discrete Autocorrelation Function, as proposed in Edelson & Krolik (1988), relies on binning values in time intervals to account for missing overlaps. A similar method was also proposed in Mayo et al. (1974).

As an extension of the binning proposed by Edelson & Krolik (1988), and drawing from the general kernel based methods proposed by Hall et al. (1994), Stoica & Sandgren (2006) and Bjørnstad & Falck (2001) both use a kernel to weight the product of observations according to the difference between the observation interval and the desired lag bin centre. This is also known as 'fuzzy slotting'. The kernels proposed are smooth density functions that tend to zero as lag increases or decreases from the desired lag subject to a characteristic width parameter. Stoica & Sandgren (2006) propose a sinc function, demonstrating the efficacy of this weighting on examples from over-the-internet temperature data, pulsar time-of-arrival measurements and ice core CO_2 measurements. Bjørnstad & Falck (2001) use a Gaussian kernel in the context of estimating a spatial autocorrelation for sparse ecological population data. A comparison of a number of correlation analysis techniques for irregularly sampled time series (linear interpolation, Lomb-Scargle periodogram, correlation slotting and several kernel based methods), in a geoscientific context, can be found in Rehfeld et al. (2011). These authors find that while all methods investigated lead to consistent results for time series with a relatively constant sampling density, the kernel based methods perform better for highly irregular time series.

Related to the problem of finding the autocorrelation function of irregularly sampled time series is the problem of finding the power spectral density (PSD), as the Fourier transform of the power spectrum of a (stochastic) time series is equivalent to the ACF (see e.g. Scargle 1989; Merrifield & McHardy 1994).

In this work we present a different generalisation of the standard estimator of the autocorrelation function, which we name the selective autocorrelation function estimator, or S-ACF. The S-ACF is an extended and generalised version of the standard estimator, which is applicable to both regularly and irregularly sampled time series, without making any assumptions about the time sampling or the statistical properties of the discrete time series and only assuming smoothness of the underlying process. In particular, there are no assumptions about regularity in the sampling of the time series.

The S-ACF method, presented in this publication, has previously been used under the name G-ACF in the publications Gillen et al. (2020) and Briegal et al. (2022).

A Python implementation of the S-ACF is available under the MIT license at github.com/joshbriegal/sacf/.

In Section 2 we define our mathematical notation and the standard estimator of the autocorrelation function. We then present the S-ACF and discuss its main properties in Section 3. In Section 4 we show that the S-ACF performs accurately on both synthetic and real data. We investigate the effect of different time samplings with both regularly and irregularly sampled time series and find that the S-ACF produces comparable estimates of the autocorrelation function and corresponding period estimates. We conclude in Section 5.

2 BASIC DEFINITIONS

We start with some elementary definitions concerning time series and the standard estimator of the autocorrelation function in order to clarify our vocabulary. We adopt the symbol := to indicate a definition. An overview of the notation can be found in Appendix A.

2.1 Time series

We define a *time series* $X_I(t)$ to be a finite ordered set

$$X_I(t) := \{(X_i, t_i) \in \mathbb{R} \times \mathbb{R}^+ | i \in I \subset \mathbb{N}, \quad (t_{i+1} - t_i) > 0 \, \forall i \in I\} \quad (1)$$

with $I \subset \mathbb{N}$ being a finite index set, which we can choose to be $I = \{0, 1, 2, \ldots, i_{\max}\}$. Furthermore we will refer to the set $T_I := \{t_i | i \in I\}$ as the set of *time labels* and to the set $X_I := \{X_i | i \in I\}$ as the set of *time series values*.

It can be useful to think of a time series $X_I(t)$ as a discrete sampling of a continuous process $X(t)$. The notation $X_i = X(t_i)$ will be used. Hence we define a time series to be *regularly sampled* if there exists a *sampling constant* $\Delta t > 0$, such that $t_k = t_0 + k \cdot \Delta t \quad \forall k \in I$, else we call the time series *irregularly sampled*.

2.2 Standard estimator of the autocorrelation function (ACF)

The *standard estimator of the autocorrelation function* (ACF; e.g. as described in Scargle 1989; Shumway & Stoffer 2017) relies on the self-similarity of the underlying process of the time series and can be applied to all regularly sampled time series to obtain information such as periodicity and signal shapes. We define the standard estimator of the autocorrelation function of a regularly sampled time series $X_I(t)$ as the function

$$\rho : \{0, 1, \ldots, i_{\max}\} \to [-1, 1] \quad (2)$$

$$\rho(k) := \frac{1}{N} \sum_{i=0}^{i_{\max}-k} (X_i - \langle X_I \rangle) \times (X_{i+k} - \langle X_I \rangle) \quad (3)$$

where $\langle X_I \rangle$ denotes the mean of the time series values and the normalisation N is the total sum of squares $N := \sum_{i \in I} (X_i - \langle X_I \rangle)^2$. The choice of this normalisation implies that $\rho(0) \equiv 1$, i.e. a time series is maximally similar to itself when there is no lag. The argument k is referred to as the *lag*.

Equation 3 is the standard estimator of the (true) autocorrelation function in the case of regularly sampled and finite time series. In the following we will use the abbreviation ACF to mean this standard estimator.

While this is a standard way to introduce the ACF, it can be useful to think of the ACF as a function with a time domain instead of an integer lag domain. We can make this domain modification explicit by multiplying the argument by the sampling constant Δt:

$$\rho(k\Delta t) : \{0, \Delta t, \ldots, \Delta t \cdot i_{\max}\} \to [-1, 1]. \quad (4)$$

These descriptions are equivalent, but the latter view is more useful for the forthcoming discussion. Specifically, it is clear that the standard estimator is only directly applicable to time series where the sampling is regular.

As discussed in the Introduction, various efforts have been made to interpolate or infer missing data to recover a regularly sampled time series and allow the use of the standard estimator. The accuracy of such efforts depends on the length of the gaps or irregularities in the time series sampling compared to the scale of structures in the signal. If the time series has large temporal gaps compared to the scale of the underlying process it is in general very difficult to restore the missing information using interpolation. Here, we seek to develop a new method to obtain information from arbitrary time series, regardless of their sampling, which is applicable to all time series. The approach presented in this work is a generalisation of the standard estimator, based on the fact that the standard estimator is already applicable to all regularly sampled time series irrespective of the underlying process. The key step is to formalise the time label dependence of the definition explicitly.

3 SELECTIVE AUTOCORRELATION FUNCTION ESTIMATOR (S-ACF)

We begin generalising the definition of the standard estimator for time series of arbitrary sampling by introducing two functions, the *selection function* $\hat{S}$ and the *weight function* $\hat{W}$, as well as generalising the notion of the lag to a *generalised lag* $\hat{k} \in [0\,, (\max(T_I) - \min(T_I))]$.

We define the selective autocorrelation function estimator, for a time series of any sampling, to be the function $\hat{\rho}(\hat{k};\, \hat{W}, \hat{S})$ which, restricted to the generalised lag $\hat{k}$, is a function of the form

$$\hat{\rho}(\hat{k}) : [0, (\max(T_I) - \min(T_I))] \to [-1, 1]. \tag{5}$$

A possible generalised definition is then given by

$$\hat{\rho}\left(\hat{k}; \hat{W}, \hat{S}\right) := \frac{1}{N} \sum_{\substack{i \in I \\ t_i + \hat{k} \leq \max(T_I)}} \Big[(X(t_i) - \langle X_I \rangle) \times \left(X(\hat{S}(t_i + \hat{k})) - \langle X_I \rangle \right) \times \hat{W}\left(\left| \hat{S}\left(t_i + \hat{k}\right) - \left(t_i + \hat{k}\right) \right| \right) \Big]. \tag{6}$$

Where $N := \sum_{i \in I} (X_i - \langle X_I \rangle)^2$ denotes the total sum of squares and $\langle X_I \rangle$ is the mean of the time series values set. The general form of the S-ACF is very similar to that of the standard estimator (Equation 3). The S-ACF differs from the standard estimator by the explicit inclusion of the selection function in the second factor, the restriction on the sum, the generalised lag and an additional third factor given by the weight function. In the following sub-sections we discuss the three new components: the generalised lag, the selection function and the weight function.

For regularly sampled time series, we want the S-ACF to reduce to the standard estimator when restricting the generalised lag to multiples of the sampling constant. A full proof and detailed explanation of this property is given in Appendix B. This reduction to the standard estimator is one of the core requirements of our generalisation and ensures that the S-ACF and the standard estimator are equivalent for regularly sampled time series. We note that this requirement motivates some of the restrictions that we impose on the selection and weight functions.

In the case of the ACF, the equation $\rho(0) = 1$ tells us that – trivially – if we do not shift the time series the correlation is perfect. The property $\hat{\rho}(0) = 1$ should also hold for the S-ACF, which we prove in Appendix C.

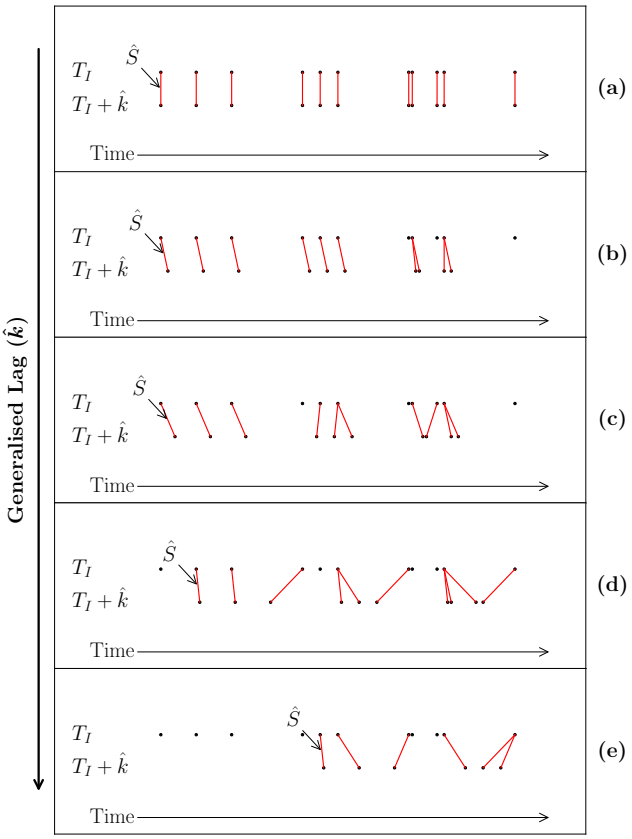

Figure 1. Each graphic (a) to (e) shows a set of time labels on the real axis and below them the same set of time labels shifted by a real generalised lag $\hat{k}$. The red lines indicate how the selection function $\hat{S}$ matches the shifted time labels to the original set of time labels above by choosing the closest time label from the set of time labels T_I. The generalised lag increases from panel (a) to (e), which corresponds to the lower set of labels 'shifting' to the right. A supplementary animated version of this figure is available on the journal website.

3.1 The generalised lag $\hat{k}$

As shown in Equation 6, the generalised lag $\hat{k} \in [0\,, (\max(T_I) - \min(T_I))]$ can now take any value within an interval in time instead of solely integer values. This is natural since in the most general case of irregular sampling there is no preferred time scale – i.e. there exists no sampling constant – and thus we can allow the lag to be a continuous variable.

Even though the S-ACF is a well defined function for any lag, it cannot contain meaningful information at a higher resolution than the time series itself, and we suggest setting the time-resolution of the generalised lag to values no smaller than the minimal difference between two neighbouring time labels $\delta\hat{k} \geq \min(t_{i+1} - t_i)$ for $\{t_i, t_{i+1}\} \in T_I$.

The condition $t_i + \hat{k} \leq \max(T_I)$ on the sum is the generalisation of the upper limit $i_{\max} - k$ of the sum in the ACF definition (Equation 3). This bound on the (generalised) lag enforces again that the maximum shifting of the process along itself is equal to the temporal length of the time series and thus when the first time label is matched up with the last time label.

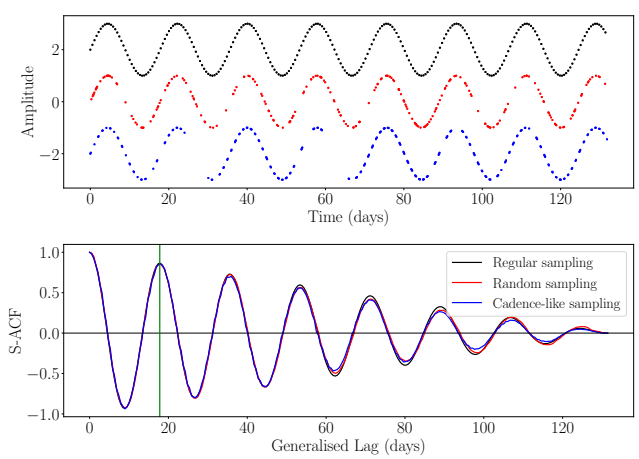

Figure 2. The top panel shows three time series with an underlying sine process (17.8 day period), sampled regularly (black), randomly (red) and with a cadence-like sampling that possesses additional larger gaps (blue). All time label sets have the same cardinality of $|T_I| = 250$. The bottom panel shows the selective autocorrelation functions (S-ACF) of the above time series. A vertical green line is plotted at the period of the signal (17.8 days) in generalised lag.

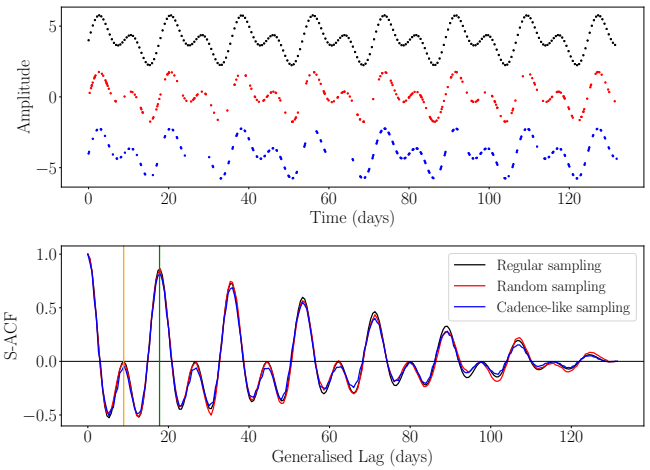

Figure 3. As Figure 2 but for a time series with an underlying process described by the sum of two sine functions with 8.9 and 17.8 day periods. Vertical orange and green lines are plotted at the periods of the signal (8.9 and 17.8 days respectively) in generalised lag.

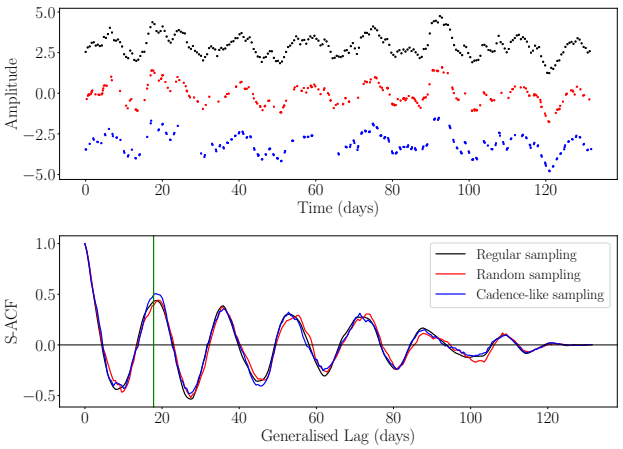

Figure 4. As Figure 2 but for a time series with an underlying process described by the sum of a comparable amplitude sine function and stochastic Gaussian process. A vertical green line is plotted at the period of the deterministic component of the signal (17.8 days) in generalised lag.

3.2 The selection function $\hat{S}$

The selection function is arguably the most important part of the S-ACF definition (Equation 6), as it deals with the irregular sampling issue that is at the core of the problem considered in this work. We define a selection function $\hat{S}$ to be a function $\hat{S} : \mathbb{R}^+ \rightarrow T_I$ that projects an arbitrary point in time onto the set of available time labels, thus selecting a specific time label for each point in time. There are many sensible functions that one could choose to accomplish this, however a natural selection function is the one that, for each point in time, selects the *closest allowed time label* (see Figure 1 for an illustration of this function). In the case that two time labels are equally close to the argument one can employ the convention of always choosing the smaller or larger value, or randomise the decision in any practical application of the S-ACF.

In order for this selection function to be justified we have to make the assumption that the process underlying the discrete time series has some degree of "smoothness" in between the time labels.

A key difference of this selection function, compared to the kernel based methods described in the Introduction, see e.g. Rehfeld et al. (2011), is that the selection function does not have a kernel with a fixed width. The distance in lag between a time label and the selected shifted time label can be as large as half the size of the largest gap in the time series sampling, as can be seen in Figure 1.

A possible alternative definition of the selection function would be to find the closest time label for the first shifted time label and then pair up all subsequent labels instead of finding the closest time label for each shifted label individually. While this definition reduces the computational complexity, it did not produce as accurate a reconstruction of the standard ACF as taking the closest time label for each individual time label when tested on synthetic time series.

3.3 The weight function $\hat{W}$

We define a weight function $\hat{W}$ to be a function $\hat{W} : [0, \infty) \rightarrow [0, 1]$ with $\hat{W}(0) \equiv 1$. We will interpret the weight function as a function that assigns time differences $\delta t \geq 0$ a weight within the interval $[0, 1]$. Specifically we see from Equation 6 that the weight function is used to assign a weight to the difference between the argument and the value of the selection function and is thereby a statement about the quality of the 'selection'. Every fixed point of the selection function $\hat{S}\left(t_i + \hat{k}\right) = t_i + \hat{k}$ will therefore lead to a term in the S-ACF with weight equal to one because of the requirement that $\hat{W}(0) \equiv 1$.

There are many choices for possible weight functions, but we have to make sure that the condition $\hat{W}(0) \equiv 1$ is observed since this is an important property (see Appendices B and C) and turns out to be a natural condition to ask for. Additionally, it would be natural for the weight function to be a monotonically decreasing function tending towards zero, since this reflects the interpretation that terms that involve time series values at similar points in time should be

preferred. There are infinitely many such functions, including an exponential function or one half of a Gaussian distribution – however a rational function is simpler and arguably more natural.

The Python implementation published with this work supports several different weight functions. Simple tests indicate that different weight functions, that fit the above criteria, do not lead to clearly significant changes in the accuracy of the method. Finding the optimal weight function is beyond the scope of this paper.

We propose the following weight function

$$\hat{W}(\delta t) = \frac{1}{1+\alpha\delta t}, \quad \alpha > 0,\ \delta t \geq 0 \tag{7}$$

where α is the inverse of the characteristic scale parameter of the time series labels, e.g. one may choose $\alpha = 1/\langle T_I \rangle$. The δt represents a generic time difference and does not have any interpretation as a sampling constant.

Naively one may expect that the S-ACF will depend on the scale of the time labels since we are free to re-scale time labels arbitrarily, but we know that the correlation between the different points in time of a process should not depend on the overall time scale. The use of the inverse scale parameter cancels out any re-scaling of the time labels since it will re-scale in the inverse fashion. This is the simplest continuous weight function that fits the above criteria and is also the most efficient for explicit calculations.

It is possible to consider a discrete weight function that satisfies $\hat{W}(0) = 1$ but is zero in all other cases, thus discarding all terms that do not have matching shifted time labels and hence eliminating the selection function from the definition. However in the case of irregularly sampled time series such a discrete weight function may eliminate the majority of the terms contributing to the S-ACF for a given lag and even almost matching terms would not be considered. This method is best suited to the case of an almost regular sampling where a small percentage of values are missing from an otherwise regularly sampled time series. In this case most terms in the autocorrelation function will have a match when the lag corresponds to an integer multiple of the 'regular' sampling constant and only a small number of terms without a matching time label need to be discarded.

4 S-ACF APPLIED TO SYNTHETIC AND REAL DATA

4.1 Synthetic data: simple sinusoidal time series

We created a periodic sinusoidal signal. To investigate the impact of the temporal sampling on the S-ACF we considered three different examples: (i) regularly sampled, (ii) randomly sampled, and (iii) cadence-like sampling with gaps. This third time series seeks to simulate the observing strategy of ground-based astronomical surveys, i.e. data during nighttime, gaps during daytime, and sporadic additional gaps due to bad weather. For ease of comparison, each time series contains the same number of data points ($|T_I| = 250$) and they differ only in the temporal distribution of the data points.

These three time series are shown in the top panel of Figure 2. The S-ACF of the regularly sampled time series is identical to the ACF, as expected due to their definitions. The S-ACF of the random and cadence-like samplings are similar to the ACF, but with small differences due to the data gaps and corresponding loss of information. These differences depend on the exact position and size of gaps within the data.

In Figure 3 we consider the sum of two sine functions with the same three temporal samplings described above. Similar behaviour is seen in both the two-sine and single sine function examples, i.e. both the random and cadence-like sampling cases display modest deviations from the regularly sampled case but overall comparable autocorrelation functions.

4.2 Synthetic data: more complex time series

In order to investigate the efficacy of the S-ACF on more realistic data sets, we generated a periodic signal with a large stochastic noise component. The periodic signal was again a sine function with a period of 17.8 days. The stochastic component was drawn from a Gaussian process (GP) using a simple harmonic oscillator (SHO) kernel (with quality factor $Q = 1/3$ and characteristic timescale $\rho = 5$ days), as implemented in the `celerite2` Python package (Foreman-Mackey et al. 2017; Foreman-Mackey 2018). The amplitude of the sinusoidal and stochastic components were comparable.

The same three temporal samplings were used as in Section 4.1, and the resulting time series and corresponding S-ACFs are shown in Figure 4. The S-ACF displays a prominent peak corresponding to the period of the sinusoidal component, although the exact position of this peak will be moderately affected by the large noise component, as expected. Despite the periodic and noise components having comparable amplitudes, the S-ACF is able to accurately recover a clear periodic signal in all three sampling cases.

4.3 Synthetic data, quantitative analysis: comparison to kernel estimators and interpolation

We want to quantitatively compare the performance of the S-ACF to that of the standard autocorrelation estimator, as well as to several other methods. To do this, we assume that the standard autocorrelation estimator of a regular sampling of a process is close to the true autocorrelation function and thus we measure the other estimators relative to this function. We focus on comparing the estimators directly without determining the periods of the process.

We consider the following methods: S-ACF, the rectangular and Gaussian kernel estimators (as implemented in Collenteur et al. (2019)), and the standard autocorrelation estimator following a linear interpolation of the irregularly sampled time series onto a regular sampling.

Throughout this section we consider the same kind of process as in section 4.2, i.e. processes that consist of a GP rotation signal with an additional GP noise component. The GP rotation component has a period between 0.1 - 50 days, standard deviation $\sigma = 1$, quality factor $Q_0 = 5$ (with $dQ = 1$) and fractional amplitude of the secondary mode relative to the primary mode $f = 0.5$. The noise is again drawn from a GP using a simple harmonic oscillator (SHO) kernel, with quality factor $Q = 1/3$ and characteristic timescale ρ between 0.1 - 50 days. The period of the periodic component of the process and the noise time scale are drawn uniformly at random from the interval 0.1 - 50. The signal-to-noise ratios considered are between 0.001 and 20. The overall length of the processes is 100 days.

From these processes we generate time series with sampling densities (or time label densities) varying between 0.1 - 20 time labels per day. The length of each time series is kept fixed at 100 days. The distributions of the time labels can be a uniform random distribution or a cadence-like distributions with regular sampling during nighttimes and additional larger gaps (e.g. simulating nightly observations with periods of bad weather).

The random sampling is generated by selecting time labels uniformly at random until the given average time label density is reached. The cadence-like sampling is generated by placing regularly sampled time labels during the night time (considered to last 8 hours) around

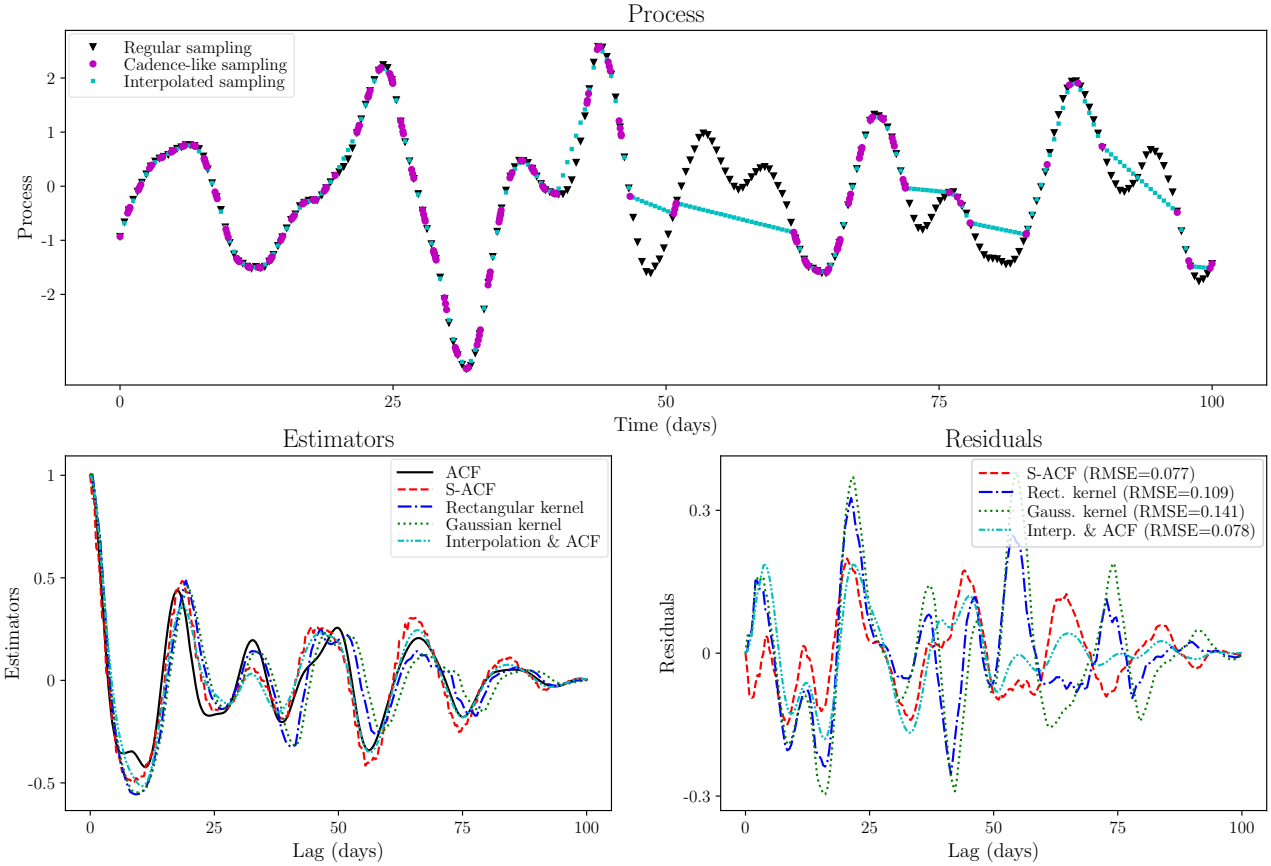

Figure 5. Top: three versions of a process comprising both a periodic component (P = 17.8 days) and a correlated noise component (characteristic timescale l = 5 days), which has an average sampling density of 2.5 time labels per day and a signal-to-noise ratio of 5.0. The three versions are a regular sampling of the process (black triangles), a cadence-like sampling (magenta circles) and an interpolation of the cadence-like process onto the regular sampling (cyan squares). Bottom left: comparison of the different estimators of the autocorrelation function for this process. The standard estimator (black line) is based on the regularly sampled time series. The S-ACF estimator (red dashed), rectangular kernel estimator (blue dash-dotted) and Gaussian kernel estimator (green dotted) are based on the cadence-like sampling of the process. Finally, we use the interpolated regular sampling of the process to again apply the standard estimator (cyan dash-dot-dotted). Bottom right: residuals of the different estimators (S-ACF, rectangular and Gaussian kernels, and interpolation) relative to the standard estimator (i.e. the ACF). Root-mean-square errors (RMSEs) are indicated in the legend.

10 (larger) gaps, where the edges of the gaps are chosen uniformly at random. A small number of time labels are then added, at random, around the gaps until the given average time label density is reached.

Figures 5 and 6 illustrate how we compare the different estimators in the case of cadence-like sampling (the process of comparing the different estimators is equivalent for the random sampling). Figure 5 shows an 'easy' time series with an average sampling density of 2.5 time labels per day and a signal-to-noise ratio (S/N) of 5. We then compute the estimators derived from these methods based on the cadence-like sampling of the process. Finally, we compute the residuals and root-mean-square errors (RMSE) of each estimator relative to the standard autocorrelation estimator of the regularly sampled time series (i.e. no gaps). Figure 6 repeats this for a 'hard' time series with an average sampling density of 0.5 time labels per day and a S/N of 0.1. It is worth noting that the discrete time series shown in Figure 6 (top) appear "smoother" than the underlying continuous process, due to their low sampling density.

Using the RMSEs computed in this way we can quantitatively compare the S-ACF to the standard estimator for time series with random and cadence-like sampling. In the case of regular sampling the S-ACF reduces to the standard estimator (see Appendix B), which makes this comparison trivial by design. Figure 7 shows the average RMSEs of the S-ACF for a wide range of parameters. The RMSEs are generally much lower in the case of random sampling. Moreover we can see that the signal-to-noise ratio has essentially no effect when comparing the S-ACF directly to the standard estimator of the regularly sampled time series, without applying further methods to detect periodicity. For sampling densities below 0.2 the RMSEs increase in both sampling cases.

In order to compare the S-ACF to other estimators of irregularly sampled time series, we compare the RMSEs of the S-ACF, relative to the standard estimator of the regularly sampled time series, to the equivalent RMSEs of the other methods for the same processes. This is equivalent to running many analyses as in Figure 5 and taking the differences between the RMSEs of each method and the RMSE of the S-ACF. The differences between the RMSEs are always computed for the same time series and are averaged over a large number of different time series. In Figures 8 and 9 we show the results for time series with random and cadence-like sampling, respectively.

The S-ACF performs better than the two kernel methods in all areas of the parameter space, but especially when the time label density is very low ($\leqslant 0.2$ day^{-1}), and this effect is more pronounced in the case of random sampling. The performance of the S-ACF is very similar to the interpolation method in all areas of the parameter

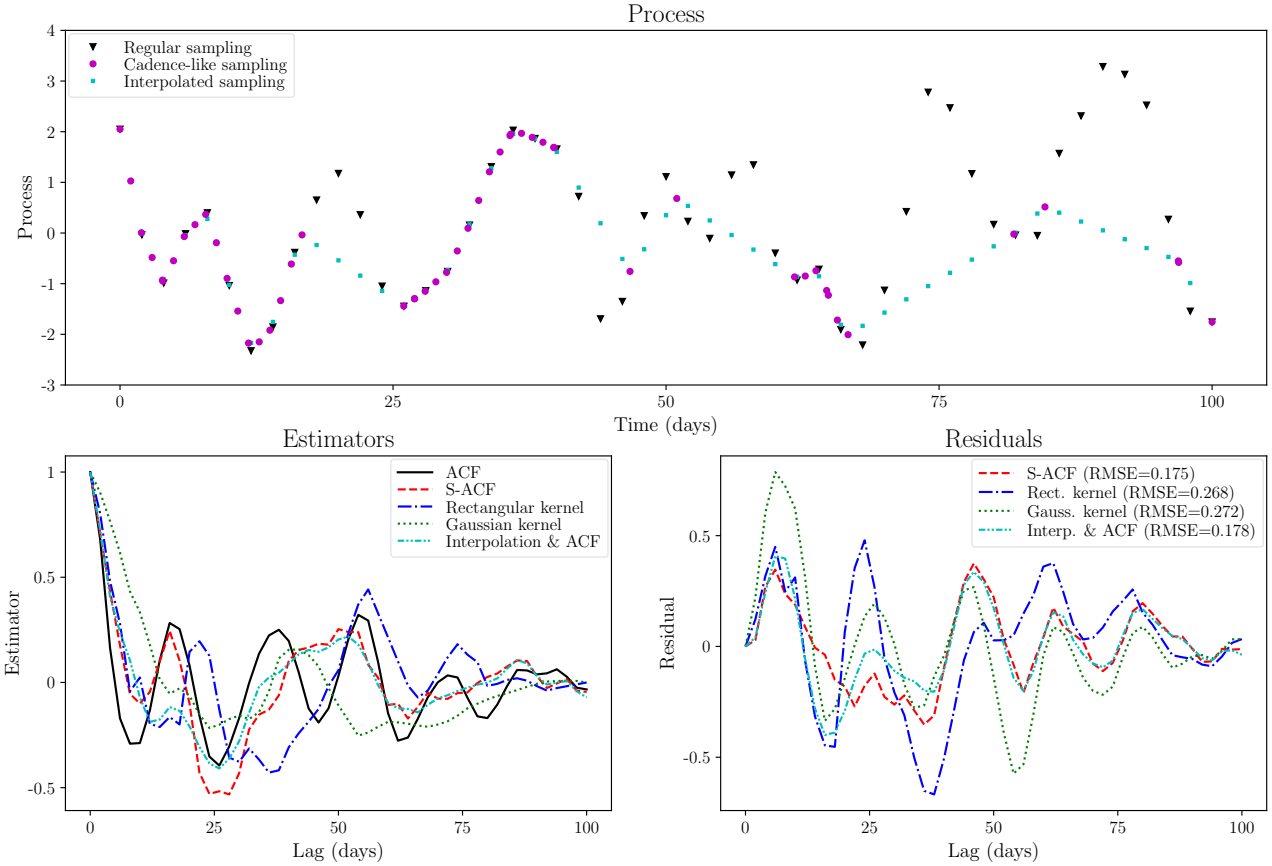

Figure 6. Same as Figure 5 for a process with a periodic component (P = 17.8 days) and a correlated noise component (characteristic timescale l = 15 days), an average sampling density of 0.5 time labels per day and a signal-to-noise ratio of 0.1.

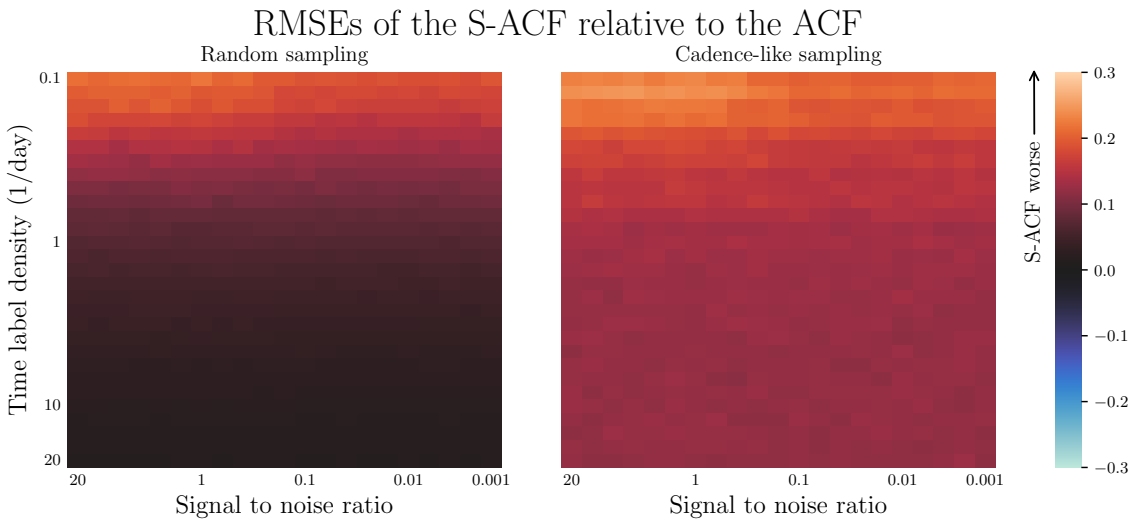

Figure 7. Average RMSEs of the S-ACF estimator for randomly sampled time series (left) and for time series with cadence-like sampling (right) relative to the standard estimator of the same process with regular sampling. Each bin shows the average of many evaluations (totalling 140700 processes per plot). Negative RMSEs are not possible and the colour scale is chosen to be consistent with Figures 8 and 9, where differences between RMSEs are shown.

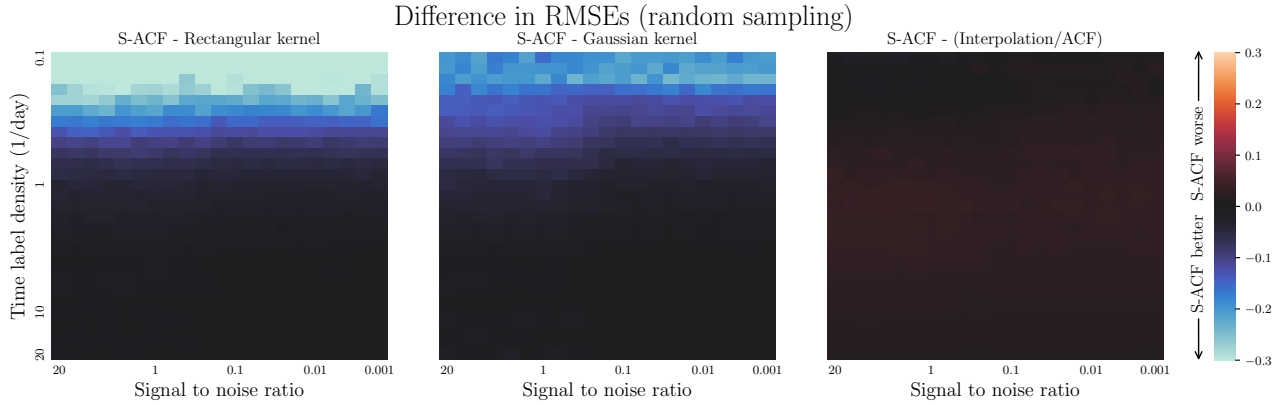

Figure 8. Differences between RMSEs (relative to the standard estimator of the regularly sampled time series) using randomly sampled time series for the S-ACF estimator vs. the rectangular kernel estimator (left), the Gaussian kernel estimator (centre) and the standard estimator on the interpolation of the randomly sampled time series (right). These are comparable to the RMSEs of the left panel of Figure 7. Red indicates a larger RMSE of the S-ACF estimator, blue indicates a larger RMSE of the respective other estimator, and black indicates comparable RMSEs.

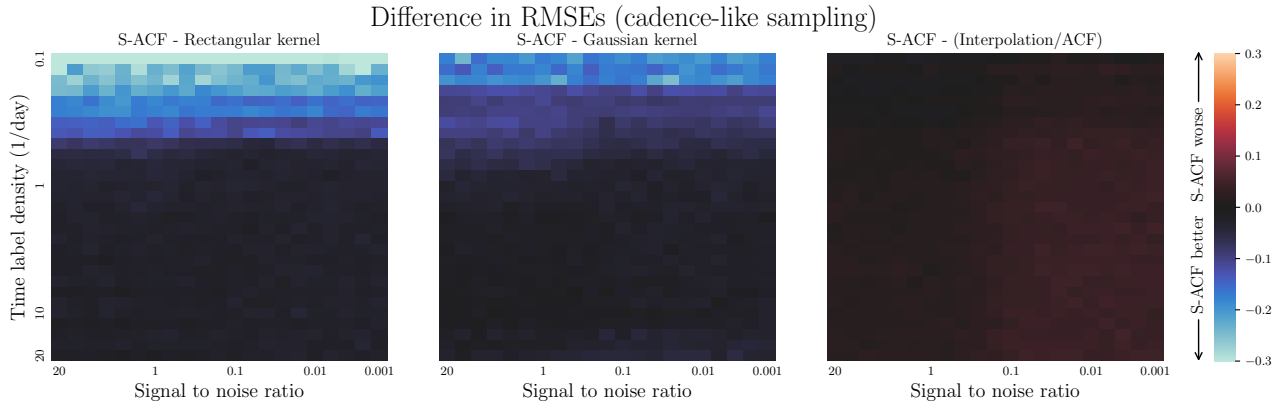

Figure 9. Same as Figure 8 but for time series with a cadence-like sampling.

space for the processes considered here. In the regime of very low S/N ratios ($\leqslant 0.1$) the interpolation method slightly outperforms the S-ACF, at least in terms of the RMSE value. We note, however, that while the RMSE is a useful indicator of an accurate estimator, we are often concerned about extracting periods from these estimator functions, e.g. from the location of the first peak (or weighted average of the first few peaks). Considering Figure 6, as an illustrative example, the bottom left panel shows that only the S-ACF has a first peak in general agreement with the ACF, so any period extracted from these estimator functions would likely favour S-ACF over the other methods, at least in this particular example. Further tests, using a robust period finding algorithm across the full parameter space, would be needed to ascertain whether S-ACF or interpolation produces more accurate period estimates. This is beyond the scope of this work.

4.4 Real data: the *Kepler* light curve of the spotted star KIC 5110407

We wish to test the efficacy of the S-ACF on real time series data and explore how the periods estimated from the S-ACF compare to other commonly used period estimation techniques. We estimate periods from the S-ACF by calculating a Fast Fourier Transform (FFT) of the first 3 peaks of the S-ACF. Restricting the lag time used in the period estimation reduces the effect of signal shape evolution on the auto-correlation function at long lag times and correspondingly improves the accuracy of the period estimated. Using a FFT to calculate the periodicity of the S-ACF is possible as the S-ACF is a continuous function by definition. We note that another method of extracting periodicity from the S-ACF would be to calculate the position of the first peak in the S-ACF, or calculating the positions of subsequent peaks in addition in order to refine this period estimate, such as the technique used in McQuillan et al. (2013).

While the S-ACF is not restricted to astronomy, the standard ACF has been widely used to estimate the rotation periods of stars from time series photometry. Therefore, as an illustrative example, we selected a spotted star observed by *Kepler*, KIC 5110407 (e.g. Roettenbacher et al. 2013), and compare the period predictions of S-ACF to two other techniques for rotation period estimation: Gaussian process (GP) regression and Lomb-Scargle (LS) periodograms. Using a rotationally variable star for this comparison allows us to probe the efficacy of S-ACF on time series that display evolution in the signal (phase) shape. While we focus on the period here, we note that other useful information, such as the evolution timescale, can also be extracted from the S-ACF. Our approach to comparing S-ACF, GP regression and LS periodograms follows Gillen et al. (2020) and we refer the reader to Section 3 of that paper for further details, but give a brief overview below of the GP and LS models used here.

The GP model is based on the `celerite2` package (Foreman-

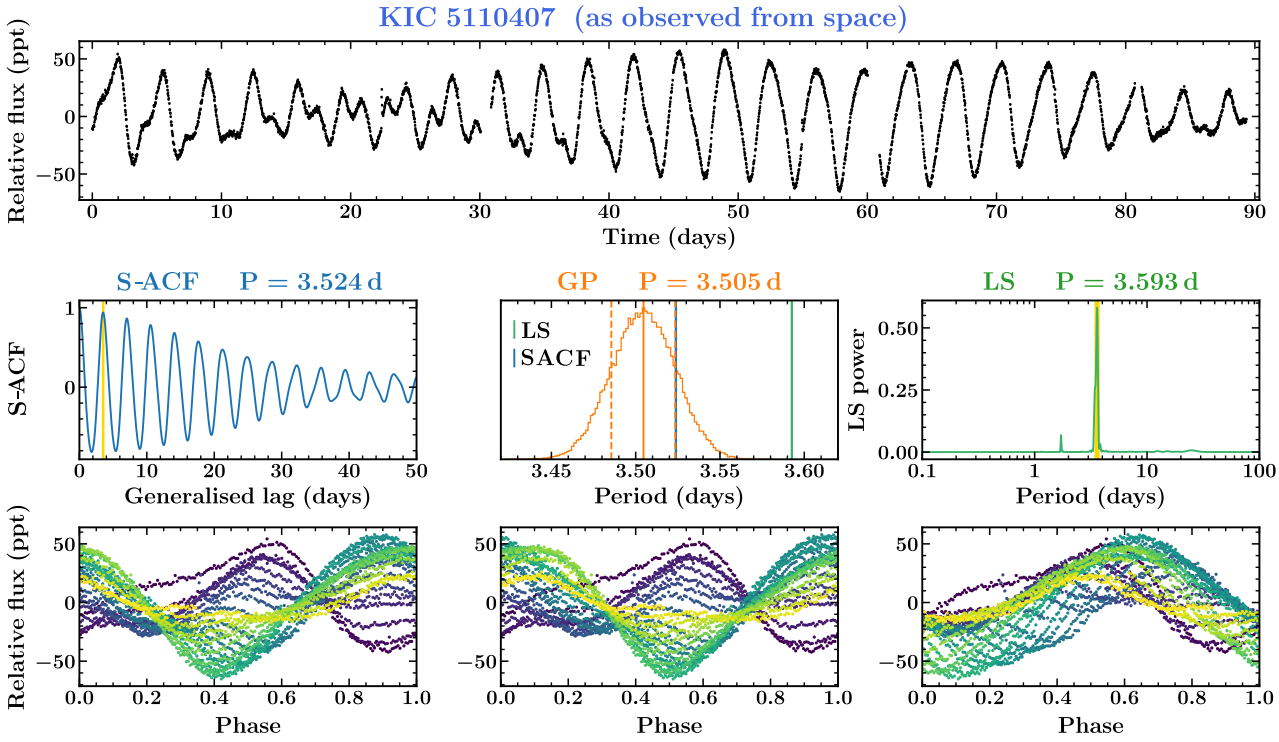

Figure 10. Rotation period estimates for the spotted star KIC 5110407 from S-ACF, Gaussian process (GP) regression, and Lomb-Scargle (LS) periodogram. *Top panel*: the system's quarter 7 *Kepler* light curve. *Middle left*: Selective autocorrelation function (blue) with the identified period highlighted (yellow). *Middle centre*: GP posterior period distribution (orange) with the median and 1 σ uncertainties highlighted (solid and dashed orange lines). For comparison, the S-ACF and LS periods are also shown (blue and green solid lines, respectively). *Middle right*: LS periodogram (green) with the identified period highlighted (yellow). *Bottom row*: The *Kepler* light curve phase-folded on the corresponding method's period (S-ACF, GP and LS; left-to-right) and coloured from the beginning (blue) to the end (yellow) of the observations.

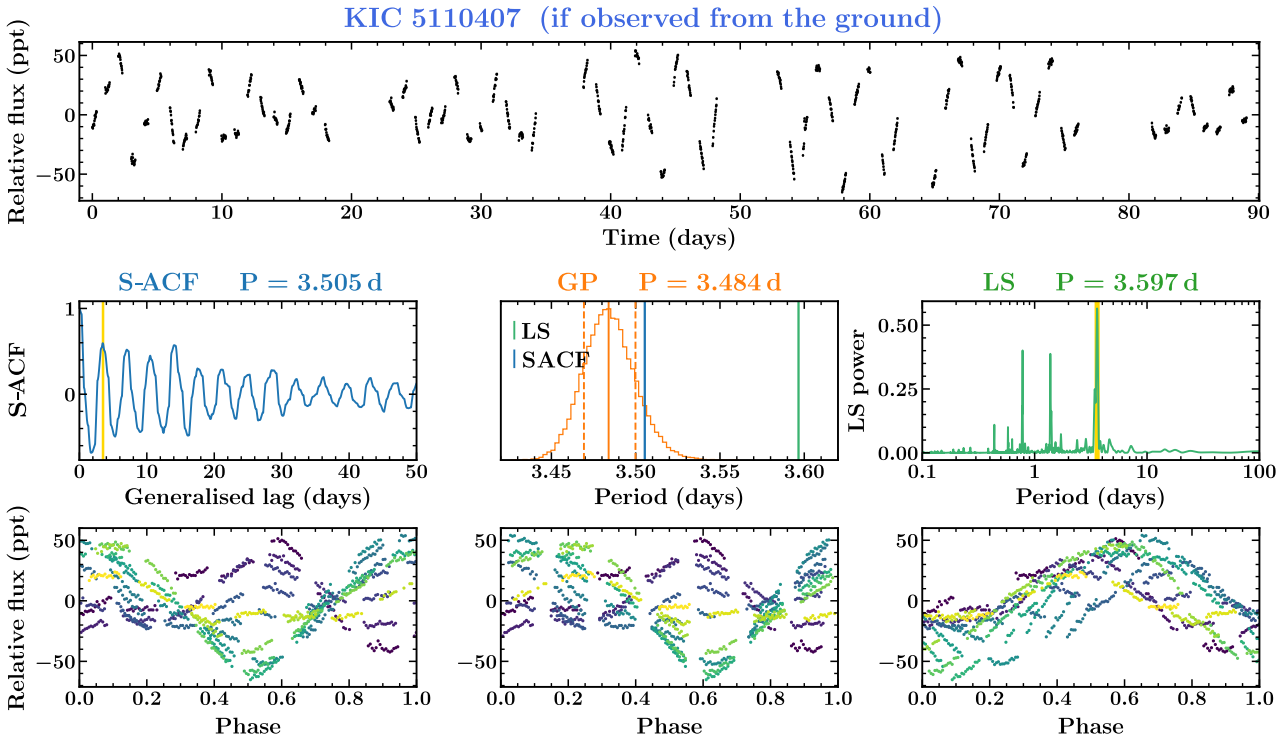

Figure 11. Same as Figure 10 but simulating KIC 5110407 being observed from the ground (i.e. observations during night time only with additional gaps from bad weather).

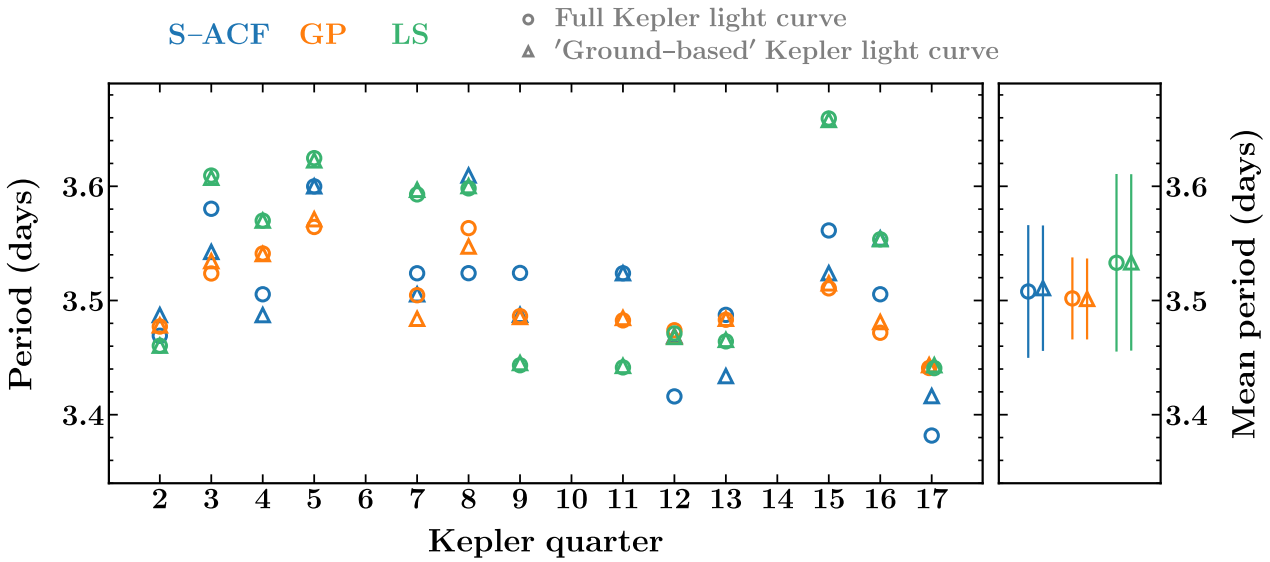

Figure 12. Rotation period estimates for KIC 5110407 from S-ACF (blue), GP (orange) and LS (green) for all quarters with *Kepler* data. Circles show the period estimates from the full *Kepler* light curve and triangles from the 'ground-based' version of the light curve. The S-ACF periods typically agree better with the GP periods than with LS, especially in cases where there is evolution in the signal shape (e.g. quarters 7 and 15). The right hand panel shows the mean and standard deviation of the period estimates across all quarters. The mean S-ACF periods from both the full and 'ground-based' versions of the light curves agree with each other, and with the values from both the GP and LS methods. The scatter in the S-ACF periods across quarters is smaller than the scatter in LS periods but larger than the scatter in GP periods.

Mackey et al. 2017; Foreman-Mackey 2018), as implemented through the `exoplanet` framework (Foreman-Mackey et al. 2021b; Foreman-Mackey et al. 2021a), and uses the standard rotation kernel with an additional simple harmonic oscillator (SHO) kernel (with quality factor $Q = 1/3$) to capture any non-periodic structure in the light curves. The posterior parameter space was explored via gradient-based Markov-chain Monte Carlo (MCMC) using the No U-Turn Sampler (NUTS), as available through `exoplanet`, which in turn uses `PyMC3` and `theano` (Hoffman & Gelman 2014; Kumar et al. 2019; Salvatier et al. 2016; Theano Development Team 2016). For each quarter, we ran 5 independent chains of 5,000 tuning steps followed by 10,000 sampling steps. GP periods were taken as the median of the posterior period distribution. It is worth noting that the GP model requires an initial period guess, in contrast to both S-ACF and LS, for which we give the average of the S-ACF and LS period estimates. The GP model is also sensitive to data not well captured by the chosen rotation kernel, such as stellar flares, which we account for by performing an initial maximum a posteriori fit, masking $3\,\sigma$ outliers, and refitting. For the LS model, we use the version available through the `astropy` project (Astropy Collaboration et al. 2013; Price-Whelan et al. 2018). LS periods are estimated from the largest peak in the periodogram. Both the LS and S-ACF models were run on the data without further processing such as flare masking.

Kepler observed KIC 5110407 for almost four years spanning 13 of the 17 quarters. *Kepler* quarters typically last ~90 days and have essentially continuous observations with a cadence of ~30 mins. The ACF has been successfully applied to such *Kepler* data (e.g. McQuillan et al. 2013, 2014) but, as noted, the ACF is not applicable to non-continuous data that cannot be accurately interpolated onto a regularly spaced time series grid, i.e. time series with large data gaps, such as ground-based photometry. We therefore estimated the stellar rotation period of KIC 5110407 from two versions of its *Kepler* light curve: (i) the full *Kepler* light curve and (ii) the *Kepler* light curve as though it had been observed from the ground (i.e. with gaps during daytime and simulated 'bad weather' events[1]). Figure 10 shows the results for the full *Kepler* light curve observed during quarter 7 and Figure 11 shows the results for 'ground-based' version of the light curve. The *Kepler* data from this quarter shows moderate evolution throughout and displays both 'double-dip' patterns (e.g. at ~20 days) and sinusoidal modulation (e.g. at ~40–80 days). For this quarter, therefore, the S-ACF and GP periods agree best whereas the LS period prediction is slightly larger. This is the case for both the full and 'ground-based' light curves. The better agreement between S-ACF and GP is because they are more flexible than LS (i.e. they do not assume a rigid sinusoidal model), and hence are more applicable to such evolving time series. The periods can be best compared in the middle centre panel of Figures 10 and 11 and by comparing the phase-folded light curves.

We performed the same analysis on each available quarter of *Kepler* data and compare the period predictions for S-ACF, GP and LS across quarters in Figure 12. Across quarters, and for both the full and 'ground-based' light curves, the S-ACF and GP periods agree best overall. The LS predictions agree well for some quarters, mainly those which show sinusoidal modulation, but less well for those that show evolving modulation patterns, which results in a larger scatter and correspondingly larger uncertainties on the mean rotation period prediction compared to S-ACF or GP. The mean periods and standard deviations across quarters are: S-ACF = 3.51 ± 0.06 and 3.51 ± 0.06 days for the full and 'ground-based' light curves, respectively; GP = 3.50 ± 0.04 and 3.50 ± 0.04 days; and LS = 3.53 ± 0.08 and 3.53 ± 0.08 days. We note that Roettenbacher et al. (2013) estimate

[1] All quarters had the same relative times masked. Nighttime was considered to last 8 hours of each 24 hour period and bad weather was simulated between the following times: 18.5–22.5, 34.5–37.5, 48.5–52.5, 62.5–64.5 and 76.5–81.5 days (relative to the start of each quarter).

a rotation period for KIC 5110407 through light-curve inversion of 3.4693 days, which agrees to within 1σ for all three methods.

This comparison between the S-ACF and the GP and LS methods, for both continuous and irregularly sampled time series, illustrates the validity of the S-ACF for such applications. Furthermore, as the S-ACF is a very general approach with minimal assumptions about the process, it can be applied to time series data of essentially any form, without the need to adapt the kind of model chosen (in the case of GP) or assume a rigid sinusoidal model (in the case of LS). The S-ACF method took approximately 0.6 seconds on the quarter 7 KIC 5110407 light curve (4,117 data points) using a single laptop core[2]. The GP regression took ~8.3 seconds for the maximum a posteriori fit (and ~7 minutes for the MCMC), while the LS periodogram took approximately 0.01 seconds to run on the same laptop core. Each method is performing different calculations, and the GP MCMC method additionally provides a period uncertainty, so their respective times are simply included here for completeness and general interest. The S-ACF is a powerful and efficient approach to extract periodicity, quasi-periodicity and short-term self-similarity from time series data in general, and especially data for which the true functional form is unknown.

Furthermore, we note that the S-ACF has been successfully applied to ground-based data from the Next-Generation Transit Survey (NGTS; Wheatley et al. 2018) to extract various kinds of stellar variability, including rotation, pulsations and eclipsing binaries (Gillen et al. 2020; Briegal et al. 2022).

4.5 A note on aliasing

In the case of cadence-like sampling (e.g. ground-based astronomical surveys), which possess a (mostly) fixed periodicity of sampling gaps, alias signals can appear in the S-ACF due to the missing information in between well-sampled clusters of data. These aliases can be easily identified since their periodicity will be equal to the periodicity of the data clusters and their amplitude will be proportional to the relative size of the gaps between clusters (see e.g. Briegal et al. 2022). This effect will not be relevant for most applications unless the structure (or period) of interest is comparable to the structure of the sampling clusters (i.e. sidereal day for ground-based astronomical surveys). We suspect that it may be possible to reduce this modest effect by generalising the normalisation to a function that depends on both the lag and time label density. However, a full removal of these aliases will likely not be possible since gaps imply missing information that cannot be restored without additional information or assumptions. We note, however, that these effects are small if there are sufficient data points per period of the process.

5 CONCLUSIONS

The S-ACF, or selective autocorrelation function estimator, is a new and versatile definition that can reliably and efficiently extract - amongst others - periodicity and signal shape information from any time series, virtually independent of the time series sampling and only assuming the smoothness of the underlying process. We show that the standard estimator of the autocorrelation function can be generalised and applied to irregularly sampled time series by generalising the lag to a real variable and introducing both selection and weight functions. We show that the S-ACF reduces to the standard estimator for regularly sampled time series and possesses the property of maximal correlation at zero lag.

The S-ACFs derived from both simple and more complex synthetic time series with different samplings (regular, random and 'cadence-like') agree well, however there are small deviations due to the data gaps and corresponding loss of information. We calculate the root-mean-square error (RMSE) of the S-ACF of irregularly sampled synthetic processes, relative to the standard estimator of the ACF of the same process with a regular sampling. The RMSEs are calculated for a large number of processes spanning a wide range of time label densities and signal-to-noise ratios. The RMSEs of the S-ACF are then compared to the equivalent RMSEs of other methods that aim to estimate the true ACF, including Gaussian and rectangular kernel estimators and a combination of linear interpolation and the standard estimator. The RMSEs of the S-ACF increase significantly at low time label densities (< 1 day^{-1}) and the same effect can be seen with the other methods. The RMSEs of the S-ACF have essentially no dependency on the signal-to-noise ratio. For the processes considered here, the S-ACF performs better than the two kernel methods (most notably at low time label densities) and comparable to the interpolation method (although we note that the interpolation method performs slightly better at low signal-to-noise ratios, which may be due to properties of the GP noise that dominates the process in this regime). At high time label densities and modest-to-high S/N all considered methods perform well and are very close to the standard estimator of the ACF.

We compare the period predictions of S-ACF to those from GP regression and LS periodograms by extracting rotation periods for the spotted star, KIC 5110407. The S-ACF and GP periods typically agree best across the different *Kepler* quarters, with LS periods being comparable in quarters with mainly sinusoidal modulation but more discrepant for quarters displaying more complex or evolving patterns. All three methods achieve consistent mean periods and uncertainties.

There are a wide range of potential applications for the S-ACF, not only within astronomy but also in other quantitative sciences where irregularly sampled time series occur, such as economics, finance, climatology, geology, biology and others.

The Python implementation used in this work is available open-source under the MIT license at github.com/joshbriegal/sacf, and additionally can be installed through PyPI using the command: `pip install sacf`.

ACKNOWLEDGEMENTS

LTK would like to thank the Bridgwater bursary of the Faculty of Mathematics at the University of Cambridge for their financial support and the Battcock Centre, Trinity College Cambridge and Sidney-Sussex College Cambridge for their hospitality. EG gratefully acknowledges support from the David and Claudia Harding Foundation in the form of a Winton Exoplanet Fellowship. JTB acknowledges support from the Science and Technologies Facilities Council (STFC) as part of the Centre for Doctoral Training in Data Intensive Science. The authors thank Dan Foreman-Mackey for his insightful comments during the refereeing, helping them to improve the paper.

[2] The run time of the S-ACF is dependent on both the number of data points and the number of lag time steps.

DATA AVAILABILITY

The *Kepler* data used in this article are available through the Mikulski Archive for Space Telescopes (MAST) portal system at the URL https://mast.stsci.edu.

REFERENCES

Andronov I. L., Chinarova L. L., 2005, in Koester D., Moehler S., eds, Astronomical Society of the Pacific Conference Series Vol. 334, 14th European Workshop on White Dwarfs. p. 659

Angus R., Morton T., Aigrain S., Foreman-Mackey D., Rajpaul V., 2018, MNRAS, 474, 2094

Astropy Collaboration et al., 2013, A&A, 558, A33

Bjørnstad O. N., Falck W., 2001, Environmental and Ecological Statistics, 8, 53

Briegal J. T., et al., 2022, MNRAS, 513, 420

Collenteur R., Bakker M., Caljé R., Klop S., Schaars F., 2019, Groundwater, 57, 877

Cooley J. W., Lewis P. A. W., Welch P. D., 1969, IEEE Transactions on Education, 12, 27

Edelson R. A., Krolik J. H., 1988, ApJ, 333, 646

Foreman-Mackey D., 2018, Research Notes of the American Astronomical Society, 2, 31

Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017, AJ, 154, 220

Foreman-Mackey D., et al., 2021a, exoplanet-dev/exoplanet v0.5.1, doi:10.5281/zenodo.1998447, https://doi.org/10.5281/zenodo.1998447

Foreman-Mackey D., et al., 2021b, arXiv e-prints, p. arXiv:2105.01994

Gillen E., et al., 2020, MNRAS, 492, 1008

Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., 2013a, MNRAS, 434, 2629

Graham M. J., Drake A. J., Djorgovski S. G., Mahabal A. A., Donalek C., Duan V., Maker A., 2013b, MNRAS, 434, 3423

Hall P., Fisher N. I., Hoffmann B., 1994, The Annals of Statistics, 22, 2115

Hoffman M. D., Gelman A., 2014, Journal of Machine Learning Research, 15, 1593

Kumar R., Carroll C., Hartikainen A., Martin O. A., 2019, The Journal of Open Source Software

Lukatskaia F. I., 1975, in Sherwood V. E., Plaut L., eds, IAU Symposium Vol. 67, Variable Stars and Stellar Evolution. p. 179

Mayo W. T., Shay M. T., Riter S., 1974, Technical Report AEDC-TR-74-53, The Development of New Digital Data Processing Techniques for Turbulence Measurements with a Laser Velocimeter. Department of the Air Force, Arnold Engineering Development Center

McQuillan A., Aigrain S., Mazeh T., 2013, MNRAS, 432, 1203

McQuillan A., Mazeh T., Aigrain S., 2014, ApJS, 211, 24

Merrifield M. R., McHardy I. M., 1994, MNRAS, 271, 899

Morris B. M., Davenport J. R. A., Giles H. A. C., Hebb L., Hawley S. L., Angus R., Gilman P. A., Agol E., 2019, MNRAS, 484, 3244

Mortier A., Collier Cameron A., 2017, A&A, 601, A110

Price-Whelan A. M., et al., 2018, AJ, 156, 123

Rehfeld K., Marwan N., Heitzig J., Kurths J., 2011, Nonlinear Processes in Geophysics, 18, 389

Roettenbacher R. M., Monnier J. D., Harmon R. O., Barclay T., Still M., 2013, ApJ, 767, 60

Salvatier J., Wiecki T. V., Fonnesbeck C., 2016, PeerJ Computer Science, 2, e55

Scargle J. D., 1982, ApJ, 263, 835

Scargle J. D., 1989, ApJ, 343, 874

Shumway R. H., Stoffer D. S., 2017, Time Series Analysis and Its Applications, 4 edn. Springer Texts in Statistics, Springer International Publishing, doi:10.1007/978-3-319-52452-8

Stellingwerf R. F., 2011, in McWilliam A., ed., Carnegie Observatories, Astrophysics Vol. 5, RR Lyrae Stars, Metal-Poor Stars, and the Galaxy. p. 47 (arXiv:1108.4984)

Stoica P., Sandgren N., 2006, Digital Signal Processing, 16, 712

Theano Development Team 2016, arXiv e-prints, abs/1605.02688

Wheatley P. J., et al., 2018, MNRAS, 475, 4476

Zechmeister M., Kürster M., 2009, A&A, 496, 577

APPENDIX A: NOTATION

Table A1 lists the notation used in this work.

APPENDIX B: PROOF OF THE REDUCTION OF THE S-ACF TO THE STANDARD ESTIMATOR FOR REGULARLY SAMPLED TIME SERIES

From the definition of the S-ACF (Equation 6), the selection function (Section 3.2) and the property $\hat{W}(0) \equiv 1$ of the weight function, we can derive a consistency property of the S-ACF for the case of regularly sampled time series, which is that the S-ACF reduces to the standard estimator in this case.

If a time series is regularly sampled, then there exists a sampling constant Δt that describes the time difference between any two neighbouring time labels t_i and t_{i+1} by $t_{i+1} - t_i = \Delta t$ with $i, i+1 \in I$. In this case we can compute the standard estimator as well as the S-ACF. If we want to compare the two functions for the same time series then we can only compare their values on their common domain, which means that the generalised lag $\hat{k}$ of the S-ACF has to satisfy the relation $\hat{k} = k \cdot \Delta t$ with respect to the lag k of the standard estimator. We thus have to restrict the real generalised lag to the domain of the standard estimator, given by integer multiples of the sampling constant.

Using the restriction on the generalised lag, all points in time of the form $t_i + \hat{k}$ will satisfy

$$t_i + \hat{k} = t_i + k \cdot \Delta t, \tag{B1}$$

but, since the time series is regularly sampled, adding multiples of the sampling constant will give another time label

$$t_i + \hat{k} = t_{i+k} \in T_I\,. \tag{B2}$$

If we now apply the selection function to this equation we obtain that

$$\hat{S}\left(t_i + \hat{k}\right) = \hat{S}\left(t_{i+k}\right), \tag{B3}$$

but since the selection function – by construction – maps its argument to the closest time label, we can use the fact that time labels are fixed points of the selection function, meaning that

$$\hat{S}\left(t_{i+k}\right) = t_{i+k} \tag{B4}$$

to arrive at the equation

$$\hat{S}\left(t_i + \hat{k}\right) = t_{i+k}\,. \tag{B5}$$

This result will be central to reducing the S-ACF to the definition of the standard estimator for regularly sampled time series.

If we look at the definition of the S-ACF (Equation 6), we see that it only differs from the standard estimator by the factor of the weight function and the insertions of the selection function. If we focus only on the last factor in Equation 6, we can directly apply Equations B2 and B5 to give

$$\hat{W}\left(\left|\hat{S}\left(t_i + \hat{k}\right) - \left(t_i + \hat{k}\right)\right|\right) = \hat{W}\left(\left|t_{i+k} - t_{i+k}\right|\right) = \hat{W}(0). \tag{B6}$$

Table A1. A brief summary of the notation of the sets, functions and parameters used.

Symbol	Description
I	Finite index set of natural numbers
$i_{\max} := \max(I)$	The maximum of the index set
$X(t)$	A continuous real process
$X_I(t)$	A time series with index set I
$X_i = X(t_i)$	A time series value corresponding to the label t_i
X_I	The set of time series values
T_I	The set of time labels
Δt	A positive sampling constant
$\rho(k)$	The standard estimator of the autocorrelation function (ACF)
k	The integer lag of the standard estimator
$\langle T_I \rangle$	The mean value of the time label set
$\langle X_I \rangle$	The mean value of the time series value set
$N := \sum_{i \in I} (X_i - \langle X_I \rangle)^2$	The normalisation of the ACF/S-ACF
$\hat{\rho}(\hat{k})$	The selective autocorrelation function estimator (S-ACF)
$\hat{k}$	The real generalised lag of the S-ACF
$\hat{W}(\delta t)$	The weight function of the S-ACF
δt	A generic positive time difference
$\hat{S}(t)$	The selection function of the S-ACF
α	A positive constant

But we defined the weight function to satisfy the property $\hat{W}(0) \equiv 1$ and thus we obtain

$$\hat{W}\left(|\hat{S}\left(t_i + \hat{k}\right) - \left(t_i + \hat{k}\right)|\right) = 1 \tag{B7}$$

removing the weight factor from the definition of the S-ACF.

The second factor in the definition of the S-ACF (Equation 6)

$$X(\hat{S}(t_i + \hat{k})) - \langle X_I \rangle \tag{B8}$$

is also modified by the selection function. However we can apply Equation B5 and obtain

$$X(\hat{S}(t_i + \hat{k})) - \langle X_I \rangle = X_{i+k} - \langle X_I \rangle \tag{B9}$$

and thus we again recover the factor from the definition of the standard estimator.

Since the first factor of the S-ACF (Equation 6) is the same as in the standard estimator, the only further modification is the restriction

$$t_i + \hat{k} \leq \max(T_I) \tag{B10}$$

on the sum. We can again use Equation B2 to obtain

$$t_{i+k} \leq \max(T_I). \tag{B11}$$

If we write this equation in terms of the indices we arrive at

$$i + k \leq i_{\max} \tag{B12}$$

which is equivalent to writing the upper limit of the sum on i as $i_{\max} - k$, as is the case in the definition of the standard estimator.

The above proof shows that the S-ACF reduces to the definition of the standard estimator if the time series is regularly sampled and we restrict the generalised lag to the domain of the standard estimator.

APPENDIX C: THE PROOF OF $\hat{\rho}(0) = 1$

We prove that $\hat{\rho}(0) = 1$ by considering the S-ACF definition (Equation 6) at zero lag. For $\hat{k} = 0$ we find

$$\hat{\rho}(0) = \frac{1}{N} \sum_{\substack{i \in I \\ t_i \leq \max(T_I)}} \Bigg[(X(t_i) - \langle X_I \rangle) \times \left(X(\hat{S}(t_i)) - \langle X_I \rangle\right) \times \hat{W}\left(|\hat{S}(t_i) - (t_i)|\right) \Bigg], \tag{C1}$$

but since $\hat{S}(t_i) = t_i$ and $\hat{W}(0) = 1$ we have

$$\hat{\rho}(0) = \frac{1}{N} \sum_{\substack{i \in I \\ i \leq i_{\max}}} \Bigg[(X(t_i) - \langle X_I \rangle) \times (X(t_i) - \langle X_I \rangle) \Bigg]. \tag{C2}$$

Using the definition of the normalisation we arrive at the desired equation

$$\hat{\rho}(0) = \frac{N}{N} = 1. \tag{C3}$$

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