Monthly Notices
of the
ROYAL ASTRONOMICAL SOCIETY

MNRAS **504,** 5829–5839 (2021)
Advance Access publication 2021 April 26

doi:10.1093/mnras/stab1178

Matched filtering with non-Gaussian noise for planet transit detections

Jakob Robnik [1]★ and Uroš Seljak[2,3]

[1]*Department of Physics, ETH-Hönggerberg, CH-8093 Zürich, Switzerland*
[2]*Department of Physics, University of California, Berkeley, CA 94720, USA*
[3]*Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Berkeley, CA 93720, USA*

Accepted 2021 April 21. Received 2021 March 24; in original form 2020 October 4

ABSTRACT

We develop a method for planet detection in transit data, which is based on the matched filter technique, combined with the Gaussianization of the noise outliers. The method is based on Fourier transforms and is as fast as the existing methods for planet searches. The Gaussianized matched filter (GMF) method significantly outperforms the standard baseline methods in terms of the false positive rate, enabling planet detections at up to 30 per cent lower transit amplitudes. Moreover, the method extracts all the main planet transit parameters, amplitude, period, phase, and duration. By comparison to the state-of-the-art Gaussian process methods on both simulations and real data, we show that all the transit parameters are determined with an optimal accuracy (no bias and minimum variance), meaning that the GMF method can be used for both the initial planet detection and the follow-up planet parameter analysis.

Key words: methods: data analysis – methods: statistical – planets and satellites: detection.

1 INTRODUCTION

Exoplanet detection using transits has become the leading method to detect new planets and determine their demographics. *Kepler Space Telescope* (Koch et al. 2010) photometrically measured flux from about 200 000 stars, with over 4000 confirmed planets (Akeson et al. 2013). One of the primary goals of such studies is to determine the planet demographics, which is the planet occurrence rate as a function of its parameters, such as the radius of the planet and the distance from the star. This relies heavily on our ability to estimate how reliable these candidates are (Steffen & Coughlin 2016). There are several origins of a false positive (Thompson et al. 2018): A planet can be mimicked by an eclipsing binary star, a single or multiple outlier noise event, a fluctuation of host star's brightness, a sudden instrumental drop, an event in an off-set star (a star in the same field of view that has no physical association with a given star; Bryson et al. 2017), etc.

For small planets far from the star, distinguishing them from false signals is difficult, as such planets are close to or below the detection limit of the *Kepler Space Telescope*. One prominent such group are habitable-zone Earth-like planets. Estimates for their occurrence rate vary wildly, with 95 per cent confidence interval covering more than one order of magnitude (Kopparapu 2013). This has implications for the prospects of follow-up missions such as the proposed Luvoir (France et al. 2017), where the predictions for the number of spectroscopically detectable habitable-zone planets also vary by a similar amount.

A traditional Kepler approach towards false positives is to perform a series of tests, each designed to target a specific group of false positives, and eliminate them if a candidate does not pass these individual tests. Those tests are, however, binary, meaning that a candidate is either rejected or not, which is a rather crude approach when on the detection limit. A likelihood analysis or a Bayesian evidence analysis would be more informative for the subsequent hierarchical analysis. Also, some of the cuts are rather heuristic, such as checking for the proper shape of the transit by calculating a metric distance (LPP metric) from the known planet shapes (Thompson et al. 2018).

The goal of this paper is to develop a new and independent planet detection pipeline, which is near-optimal, fast, and provides sufficient statistical information for downstream tasks such as planet demographics. Our goal is to develop a rigorous analysis of the stellar variability and outlier false positives. Various Gaussian process (GP)-based methods have been developed that can be used to model stellar variability (Foreman-Mackey et al. 2017; Robnik & Seljak 2020) and determine how likely it is that a given candidate is caused by stellar variability. These methods are, however, computationally expensive, so that they can only be used once a good candidate has been found and an initial estimate on its parameters is known. They must be combined with a simplified analysis of an actual planet search, where a simplified model for the stellar variability and planet transits is assumed. We will show that this is a crude assumption and results in higher significance of false positives, and therefore a loss of many real planet candidates, already at this initial stage of analysis.

We will present an alternative approach that is as fast as the simplified analysis currently used, but with the near-optimal performance of the full GP analysis. It is therefore applicable to the complete *Kepler* data set with no need for the secondary GP analysis. The general idea behind our method is to use the Fourier-based GP (Robnik & Seljak 2020), which describes the stellar variability as a frequency-dependent noise. This naturally connects to the matched filtering technique, which Fourier transforms the planet transit signal template(s) and performs signal-to-noise weighting in Fourier space

★ E-mail: jakob.robnik@gmail.com

first. Afterwards, one looks for the highest peaks in its inverse Fourier transform. Matched filtering can be shown to be optimal under the assumption of Gaussian noise, and is the method of choice in many statistical analyses, such as in Laser Interferometer Gravitational-wave Observatory (LIGO) signal detection (Abbott et al. 2016).

Kepler data noise is not Gaussian, and noise outliers must be dealt with; otherwise, they can lead to a significant increase in the false positive rate. Here, we develop a Gaussianization transformation approach, which maps a signal with non-Gaussian power-law tails in the *Kepler* data to a Gaussian. Specifically, we take advantage of the uncorrelated nature of noise to develop a method where this procedure does not change the planetary transit signal component. We will show that this approach eliminates the outlier false positives, and gives superior results to the alternatives such as robust statistics or outlier elimination (Tenenbaum et al. 2013).

2 NOISE GAUSSIANIZATION TRANSFORMATION

Here, we first review the key results of Robnik & Seljak (2020). In general, we write the data model $d(t)$ as a sum of a transit signal $s(t)$, noise $n(t)$, and stellar variability $y(t)$. Here, we first discuss the noise and the stellar variability, which can be viewed as a correlated noise, so we assume there is no transit signal. Stellar variability is correlated, and assumed to be Gaussian, which has been shown to be a good assumption (Robnik & Seljak 2020). Noise is uncorrelated but non-Gaussian, distributed according to the noise probability distribution $p(n)$. In the absence of planet signal, we assume to have a stationary, time-ordered and equally spaced data $d_i = d(n\Delta t)$ for $n = 0, 1, 2, \ldots, N-1$.

If we assume stationarity of the signal, the correlations depend only on the time difference between the points, and the GP kernel also depends only on their relative separation. Stellar variations could also be described with a non-stationary form, but this would require a significant increase in the complexity of the kernel, which we want to avoid. Later, we will, however, describe the non-stationary generalization due to the gaps in the data. To describe a stationary kernel of a uniform time series, the most general approach is to use the Fourier basis and describe the GP kernel using the power spectrum (Robnik & Seljak 2020). A Fourier transform

$$\tilde{y}_k = \mathcal{F}\{y\}_k = \frac{1}{\sqrt{N}} \sum_{n=1}^{N} y_n e^{i\omega_k n \Delta t}$$
$$y_n = \mathcal{F}^{-1}\{\tilde{y}\}_n = \frac{1}{\sqrt{N}} \sum_{k=1}^{N} \tilde{y}_k e^{-i\omega_n k \Delta t} \quad (1)$$

introduces a new basis in which the covariance matrix is diagonal and can be described with the power spectrum, and the Fourier modes $\mathcal{F}\{y\}_k$ are uncorrelated. We denoted $\omega_k = 2\pi k/N$.

If, on the other hand, the data are uncorrelated, but non-Gaussian, then we can Gaussianize it as a simple 1D point-wise non-linear transformation:

$$\psi_i = \psi^{(1D)}(n_i). \quad (2)$$

This transformation can be obtained by mapping the cumulative distributions of the data to a Gaussian (Robnik & Seljak 2020).

Here, we want to apply to the *Kepler* data, which we assume are composed of correlated and nearly Gaussian stellar variability, added to an uncorrelated noise, containing non-Gaussian outliers. Gaussianizing thus requires identifying the stellar component of the data, subtracting it from the data, Gaussianizing the remaining uncorrelated part, adding the stellar component back, and transforming to the Fourier basis:

$$\mathcal{G}(y) = \mathcal{F}\{\Psi(d - y) + y\}, \quad (3)$$

in the absence of the correlated structures, such as planet transits, $\Psi = \psi^{(1D)}$. In general, it is an invertible non-linear scalar function that is local in the sense that it depends only on d_i and possibly on its neighbours. We will use this generalization when dealing with the correlated structures in the data, such as planet transits. The correlated Gaussian component $y(t)$ can be extracted from the data with the Fourier GP (Robnik & Seljak 2020). Note that this step does not add much to the cost of our pipeline because it is done only once and because GP is never used for planet search and parameter inference, where it would have to be iterated with the optimization of the planet parameters.

We choose a function $\Psi(n)$ to Gaussianize the noise probability distribution $p(d)$, which then becomes

$$p(d) = \exp\left\{-\frac{1}{2}\sum_{k=1}^{N}\left(\frac{|\mathcal{G}_k(d)|^2}{P_k} + \ln 2\pi + \ln P_k\right)\right\}, \quad (4)$$

where P_k is the kth component of the power spectrum of y. Here, $||^2$ is a product of a complex mode with its complex conjugate, which equals adding the squares of its real and imaginary components.

2.1 Uncorrelated non-Gaussian noise

We need to determine the non-linear local $\psi(n)$ in equation (3). Assuming uncorrelated noise, a given realization n_i is distributed according to some probability density function q, independent of realizations at different times. Gaussianization $\Psi_i(n_i) = \psi^{(1d)}(n_i)$ will then also act point-wise and will satisfy

$$P(\psi^{(1D)}(n_i) > X_N) = P(n_i > X_q), \quad (5)$$

where X_N and X_q are random variables distributed according to the normal and q distributions. A unique function with the required property is

$$\psi^{(1D)} = \mathrm{CDF}_N^{-1} \circ \mathrm{CDF}_q, \quad (6)$$

where CDF_N^{-1} is an inverse of the cumulative density function of the normal distribution and CDF_q is cumulative density function corresponding to q.

In our application, noise is distributed normally except for the outliers. Outliers have been shown to be uncorrelated in the *Kepler* data (Robnik & Seljak 2020), so we can write for one data point n_i

$$q(n_i) = (1 - a)\, N(n_i) + a\, NCT(n_i), \quad (7)$$

where N is a Gaussian distribution that is defined by zero mean and variance that by Parseval's theorem is the sum over all power spectrum components, *NCT* is a distribution modelling outliers and a is a probability that a given realization is an outlier. As shown in Robnik & Seljak (2020), the outlier probability density function can be modelled well with a non-central t-distribution, and we determine its parameters and a by a fit to the data PDF. For small amplitude y the Gaussian contribution dominates, as a is typically a very small number of the order of 10^{-3}. For large y the *NCT* contribution dominates as it decays only as a power law, in contrast to the Gaussian that decays very rapidly. Correlated stellar variability never produces a large outlier signal, so it is not affected by the Gaussianization.

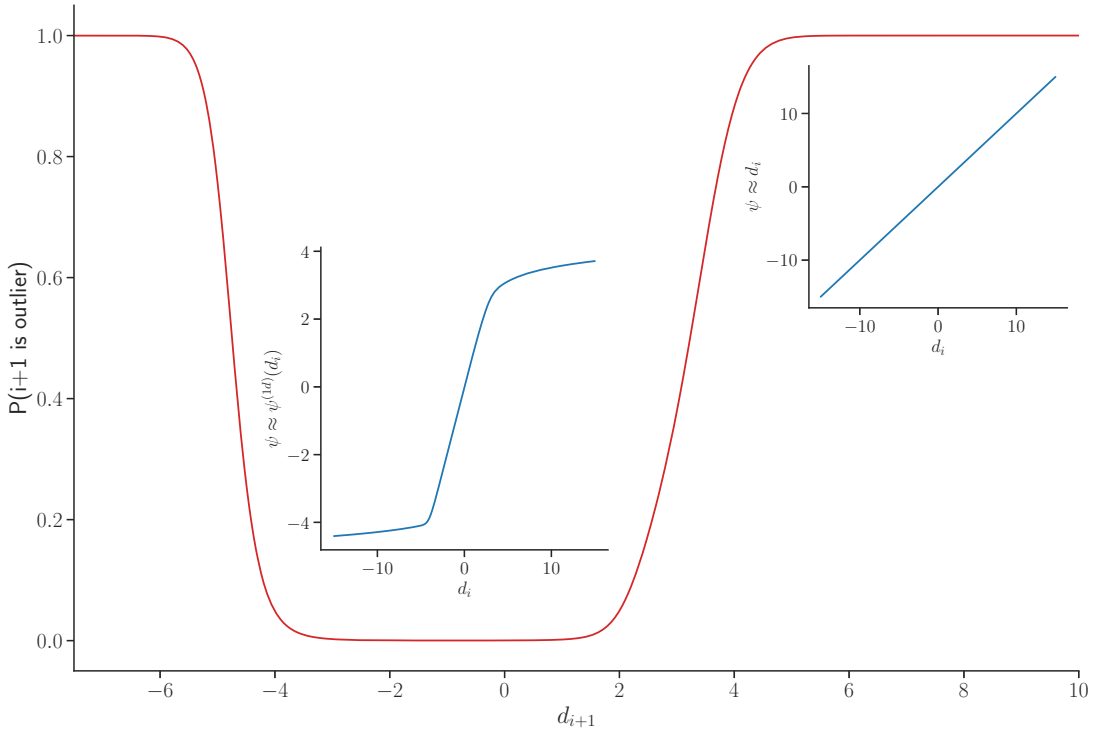

Figure 1. Gaussianization transformation at a point i (blue line) depends on a flux d_i, as well as on its neighbours d_{i+1} and d_{i-1}. If neighbours are not outliers (central part of the red figure), a point i is Gaussianized according to a one-dimensional Gaussianization; i.e. outlier part of the distribution is mapped to the central Gaussian part. If, however, neighbours are in the non-Gaussian part of the distribution (left and right parts of the red plot), then it is by far more likely that they are all a part of the correlated structure generated by a real transit and the Gaussianization acts as an identity.

2.2 Filtering of correlated structures

So far, we assumed $d(t)$ to be a sum of noise and stellar variability, but in reality it also contains correlated structures such as planet transits, which may depart significantly from the average of the flux. We do not want this signal to be affected by the Gaussianization, i.e. we do not want the deep transit signatures to be mistaken for the noise outliers and have their depth reduced. Since outlier noise is uncorrelated while real transit is not, we thus require that Gaussianization $\psi^{1D}(d_i)$ to be identity if it acts on a correlated structure, and acts as a 1D Gaussianization $\psi^{1D}(d_i)$ on an uncorrelated outlier. We write a full Gaussianization at a point i as a mixture

$$\Psi_i = \mathscr{P} d_i + (1 - \mathscr{P})\, \psi^{1D}(d_i). \tag{8}$$

$\mathscr{P}$ can be arbitrary if $|d_i|$ is small as $\psi^{1D}(d_i) = d_i$ then, and $\mathscr{P}$ cancels out. On the other hand, if $|d_i| \gg 0$, $\mathscr{P}$ should be 1 if point i is a part of the correlated structure and 0 if it is not. A simple but effective choice is

$$\mathscr{P}(d_i|d_{i-1}, d_{i+1}) = P(i+1 \text{ or } i-1 \text{ is an outlier}|\, d_{i-1},\ d_{i+1}), \tag{9}$$

where P is probability for an outlier that is calculated under the assumption that there are no correlated structures. We show that such $\mathscr{P}$ satisfies the required properties:

(i) A priori probability of point i being an outlier is a. If i is not a part of a correlated structure, then its neighbours are also outliers with probability a. Probabilities are independent, since we assume noise outliers are not correlated. A probability that i and one of its neighbours are both outliers is then $2a^2$, which can be well approximated to be zero for a typical value of a of 10^{-3}: We assume that the probability of two neighbours both being an outlier is zero. Therefore, $|d_i| \gg 0 \Rightarrow i \pm 1$ are not outliers $\Rightarrow \mathscr{P} = 0$.

(ii) On the other hand, if $|d_i| \gg 0$ and point i is a part of the correlated structure then also $|d_{i+1}| \gg 0$ or $|y_{i-1}| \gg 0$ and $\mathscr{P}$ is close to 1, as required.

$\mathscr{P}$ is straightforward to compute:

$$\begin{aligned}\mathscr{P} &= 1 - P(i+1 \text{ and } i-1 \text{ not outliers}|\, d_{i-1},\ d_{i+1}) \\ &= 1 - P(i+1 \text{ not outlier}|\, d_{i+1}) P(i-1 \text{ not outlier}|\, d_{i-1}),\end{aligned} \tag{10}$$

where $P(\text{not outlier}\,|\,d)$ is computed by a simple application of the Bayes theorem:

$$P(\text{not outlier}\,|\,d) = \frac{(1-a)N(d)}{(1-a)N(d) + aNCT(d)}. \tag{11}$$

The Gaussianization of equation (8) preserves the correlated structures, such as planet transits or binary star transits, and reduces the noise outliers by mapping them to the Gaussian distribution, thus reducing their impact on the outlier false positives in the search for the planets. It is visualized in Fig. 1. We will show that this procedure is more effective than outlier rejection, and is easy to implement.

3 MATCHED FILTER

Next, we look for a signal s in the data d in the presence of the correlated non-Gaussian noise $y + n$, such as considered in Section 2. The result will be a frequency-dependent filter, which is a generalization of the Gaussian matched filter because of the Gaussianization we perform on the noise. In Section 3.1, we will specialize to the localized periodic templates that can be used for the planet search, which will be further addressed in the next section.

The signal has a template form of an event with a time profile

$$s(t) = A\, s_0(t), \tag{12}$$

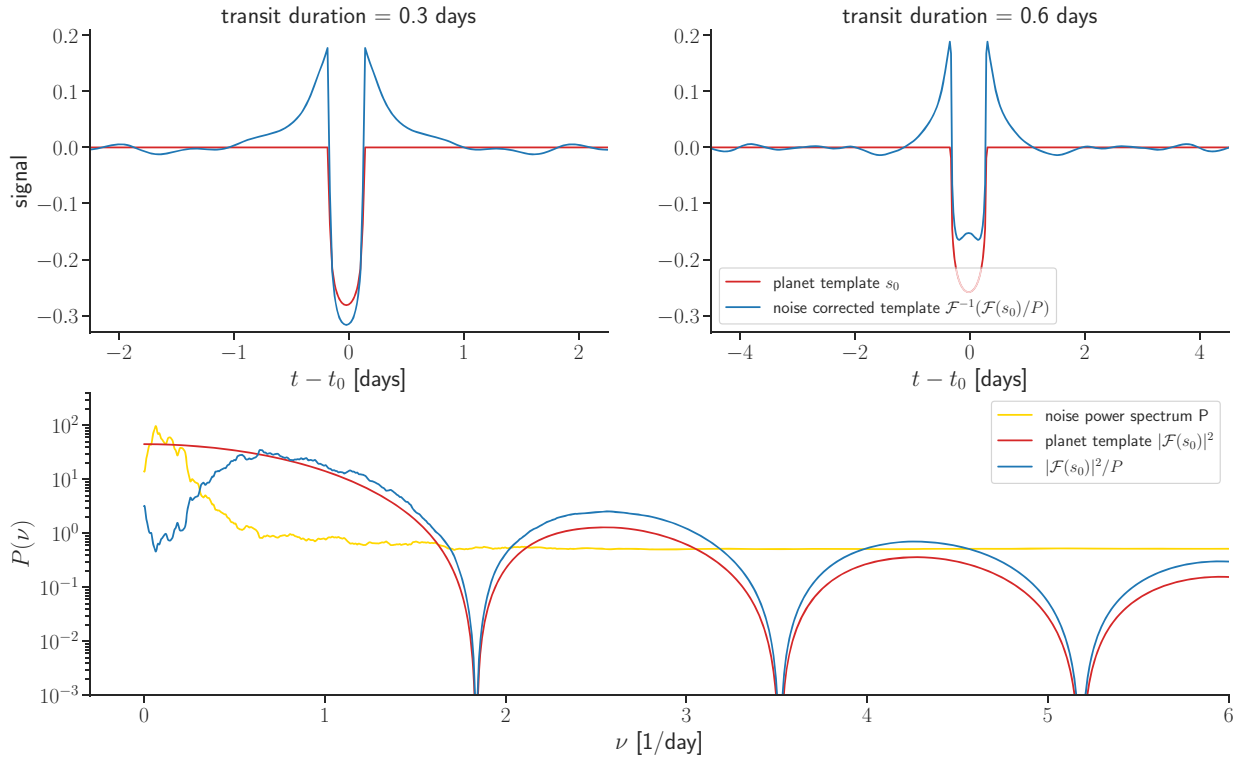

Figure 2. Top panels: a planet's transit signal s_0 (red) for the transit duration of 0.3 d in the left-hand panel and 0.6 d in the right-hand panel. Match filtering with the frequency-dependent noise as in equation (23) is equivalent to a convolution of a Gaussianized signal with the template, if we replace the original template $s_0(t)$ with $\mathcal{F}^{-1}\{\mathcal{F}\{s_0\}/P_k\}(t)$ and normalize it. The matched filter template (blue) shows a partially compensated profile characteristic of a red noise power spectrum $P(\nu)$ with high power at low frequencies. The effect is more significant for longer transits, which can be more easily contaminated by the stellar variability. Bottom panel: noise power spectrum is taken from a realistic stellar variability analysis of star Kepler 90 (yellow), together with the 0.6 d planet transit template in the Fourier domain (red). Inverse noise weighting of the matched filter suppresses the low-frequency components (blue) and leads to a compensated profile in time domain (top panels).

where A is an amplitude of the signal and s_0 is a template. Template depends on additional parameters such as the time of transit t_0 or transit duration. The shape of the template is for now assumed to be known; we will determine its parameters by a template bank search method. An example of a signal is shown in Fig. 2. It shows the effect of a typical red power spectrum taken from a realistic stellar variability analysis of star Kepler 90 on the matched filter. If the power spectrum were white, there would have been no effect, but for the red power spectrum the effect shows up as a partially compensated profile, which suppresses the low frequencies that are contaminated by the correlated noise and hence need to be filtered out. The effect is more significant for longer transits, because the stellar variability has more power on longer time-scales.

We can write the data $d(t_i)$ as

$$d(t_i) = y(t_i) + n(t_i) + As_0(t_i). \tag{13}$$

By Gaussianizing the residuals under the assumptions of Section 2.2, we obtain

$$\Psi_i(d - s) = \Psi_i(d) - s_i, \tag{14}$$

$$\mathcal{G}_k(d - s) = \mathcal{G}_k(d) - \mathcal{F}\{s\}_k. \tag{15}$$

Equation (4) becomes

$$-2\ln p(y) = \sum_{k=1}^{N} \frac{|\mathcal{G}_k(d) - A\mathcal{F}\{s_0\}_k|^2}{P_k} + c, \tag{16}$$

where $c = \ln 2\pi + \ln P_k - 2\ln J(d - s)$. The main advantage of the matched filter technique is that it can analyse the data for every possible value of t_0, by performing Fast Fourier Transform (FFT)-based convolution of the data with the signal. This is complicated in the non-linear case by the presence of the Jacobian term in c. In a typical Gaussianization application, the Jacobian term is negligible, so we will drop this term.

At a given template, there is a unique extremal value of the amplitude $\widehat{A}$, which can be obtained from the maximum likelihood, for which the derivative

$$\left.\frac{\partial \ln p}{\partial A}\right|_{\widehat{A}} = 0. \tag{17}$$

Taylor expanding the log-likelihood function around the optimal amplitude, we obtain the variance of A

$$\sigma_A^{-2}(t_0) = -\left.\frac{\partial^2 \ln p}{\partial A^2}\right|_{\widehat{A}}. \tag{18}$$

Equating first derivative to zero gives

$$\widehat{A}(t_0, s_0) = \frac{\sum_{k=1}^{N} \mathcal{G}_k^* \mathcal{F}\{s_0\}_k / P_k}{\sum_{k=1}^{N} |\mathcal{F}\{s_0\}_k|^2 / P_k}. \tag{19}$$

Equation (18) gives

$$\sigma_A^{-2}(t_0, s_0) = \sum_{k=1}^{N} |\mathcal{F}\{s_0\}_k|^2 / P_k. \tag{20}$$

A signal to noise for the event s_0 happening at time t_0 is then defined as the ratio of the signal to variance

$$\mathrm{SNR}(t_0) = \frac{\widehat{A}(t_0)}{\sigma_A(t_0)} = \tag{21}$$

$$= \frac{\sum_{k=1}^{N} \mathcal{G}_k^* \mathcal{F}\{s_0\}_k / P_k}{\left(\sum_{k=1}^{N} |\mathcal{F}\{s_0\}_k|^2 / P_k\right)^{1/2}} \tag{22}$$

$$= \frac{\mathcal{F}^{-1}\left\{\mathcal{G}_k^* \mathcal{F}\{s_0\}_k / P_k\right\}}{\left(\sum_{k=1}^{N} |\mathcal{F}\{s_0\}_k|^2 / P_k\right)^{1/2}}, \tag{23}$$

where the last step is valid under the assumption that the template is stationary. As expected from a matched filter, the SNR is proportional to a convolution of the Gaussianized signal $\Psi(d)$ with the filter profile s_0, which is a multiplication of their corresponding Fourier transforms, followed by an inverse Fourier transform. Noise power modulates this convolution with the inverse noise weighting: The larger the noise power P_k, the less weight a given Fourier component contributes. The denominator term properly normalizes the SNR by the expected signal of the matched filter profile, again inverse weighted by the noise power.

The computational complexity of evaluating equation (23) for some value of t_0 is $O(N \log N)$. More importantly, evaluating it for all t_0 on a time lattice with a lattice spacing Δt is still $O(N \log N)$ thanks to the fast Fourier transform. This is very useful for a search over the whole parameter space, which is required in the initial planet search where we do not know planet period, phase, or transit amplitude.

Note, however, that the template may not be stationary. For example, if there are gaps in the data, the template $s_0(t, t_0)$ has zeros where the data are missing and those zeros cannot be shifted. In this case, the matched filter cannot be calculated by an inverse Fourier transform and the cost of evaluating equation (23) for all t_0 on a time lattice is $O(N^2 \log N)$. We will show in the next section how to simplify and avoid incurring this cost in the planet search.

3.1 Periodic template

A special case of interest is a template containing multiple events that repeat with a period P (not to be confused with the power spectrum P_k). The template has the form

$$S_0(t, P, \phi) = \sum_{m \in I} s_0(t - mP - \phi), \tag{24}$$

where ϕ is a phase, I is a set of all integers m for which the data at $mP + \phi$ are available, and s_0 is a template of each individual event. An example of S would be multiple transits of the planet.

In principle, one would have to add P as a parameter of the template and find it using a template bank approach. However, for the purpose of finding good planet candidates in the *Kepler* data a few simplifications can be made to make this faster.

We will assume the following:

(i) Events s_0 are localized: Overlap sums containing different events $\mathcal{F}\{s_0(t)\}_k \mathcal{F}\{s_0(t - P)\}_k^*$ are zero. This can be used to write the $\mathrm{SNR}(P, \phi, S_0)$ in terms of the $\mathrm{SNR}(t_0, s_0)$. This assumption is in principle problematic for the planets with short periods, but we verified that even for planets with a 3 d period this would result in only a 3 per cent bias of the SNR.

(ii) For each m, we assume that for all t close to $\phi + mP$ either almost all data are available or almost all data are missing. By close we mean those t that contribute most to the SNR of the event. Then, the template is approximately stationary and equation (23) can be solved by the inverse FFT. This assumption fails only for the transits with partially missing data. We justify this assumption by noting that most of the missing data are collected in the time gaps that are long compared to the transit time.

Inserting equation (24) into equation (23), using the linearity of the Fourier transform, and the above stated assumptions, we derive

$$\mathrm{SNR}(P, \phi, S_0) = \frac{1}{\sqrt{|I|}} \sum_m \mathrm{SNR}(mP + \phi, s_0), \tag{25}$$

where $\mathrm{SNR}(P, \phi, S_0)$ is a joint signal-to-noise ratio of all periodic events with the period P and phase ϕ found with the template $S_0(t, P, \phi)$. $\mathrm{SNR}(t_0, s_0)$ is an SNR of one individual event centred at t_0, found with the template $s_0(t, t_0)$. Thus, one can look for periodic events by first applying convolution 23 to get $\mathrm{SNR}(t_0, s_0)$, and then fold it at different periods using equation (25).

4 PLANET SEARCH

We now apply the Gaussianization and matched filter (GMF) formalism to an example planet search in the *Kepler* data. We use the *Kepler* data processed through the pre-search data conditioning module (Jenkins et al. 2017), which eliminates systematic instrumental errors. Specifically, we use PDCSAP flux, where long-term trends have been eliminated. We normalize flux in different quarters as described in Robnik & Seljak (2020), to get an evenly spaced time series (except for gaps in the data), with unit variance of the Gaussian part of the distribution and zero average. We Gaussianize the data with equation (8) to obtain a normally distributed, correlated noise, which may contain correlated structures, such as planet transits.

Kepler time streams have gaps where the data are not available. Filling this point with zeros is not advisable if the stellar flux is highly correlated on short time-scales (e.g. Kepler 1517 in Fig. 3), as this is introducing a jump in the data that may trigger a false positive planet event. Instead, we use the Fourier GP (Robnik & Seljak 2020) to determine the correlated stellar component of the data and insert it in places where there are gaps. An example of gap filling is shown in Fig. 3.

We will first discuss the template form of the signal from equation (12). It is modelled by the dimensionless template form U and the transit duration τ as described in Robnik & Seljak (2020) and Parviainen (2015):

$$s_0(t - t_0) = U((t - t_0)/\tau, u_1, u_2). \tag{26}$$

Limb darkening parameters u_1 and u_2 are a property of the star; their impact will be discussed in Section 4.1. U depends weakly on the radius of the planet, which can be accounted for in an iterative manner, but we verified that its impact on the SNR is less than one in a thousand, so we will ignore it. In the planet search, we will first adopt a template bank approach and search over the entire parameter space to find the best planet candidates, and then optimize with respect to the planet's parameters once we are close to the peak.

For a matched filter analysis, we first need the noise power spectrum P_k, whose determination will be described in Section 4.2. Next, we look for the planets that are significant enough to be detectable without the period folding (Section 4.3). In this approach, we do not lose information on the potential transit timing deviations (TTVs) and transit duration deviations, which can be used for planetary dynamics studies, e.g. planet–planet or planet–moon gravitational

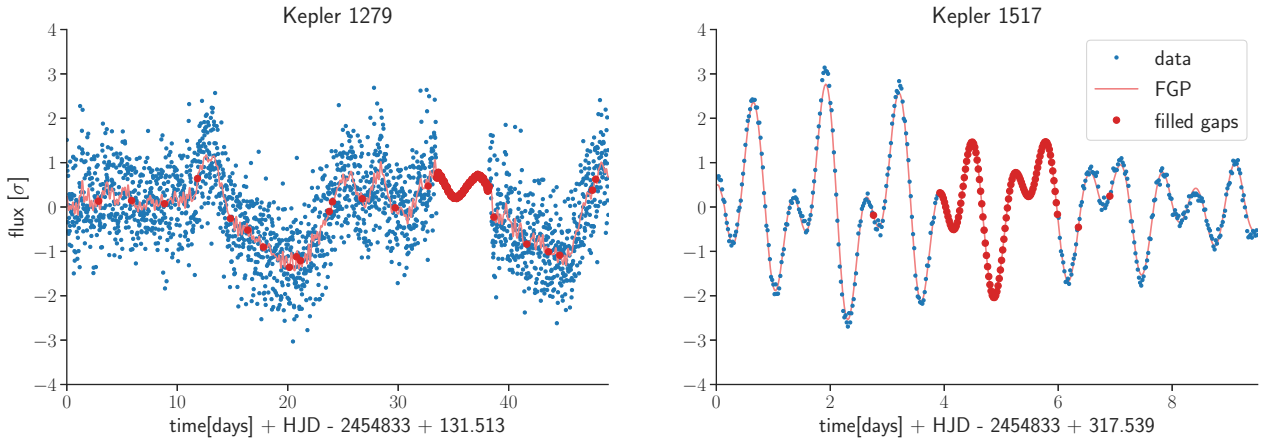

Figure 3. The *Kepler* data has gaps where no data are available. Data are missing in the form of either isolated data points or as larger gaps than can be as long as 3 months. We fill the gaps by fitting the Fourier GP to the data. We show a segment of the data for two stars, namely Kepler 1279 and Kepler 1517. Red dots are added to the data to fill the gaps.

interactions. Also, if TTVs are large, a folded analysis would leave residuals, which we want to avoid. In the last stage (Section 4.4), we look for small planets, where period is added as a parameter, and a folded analysis is required to reach a sufficient signal to noise.

4.1 Limb darkening parameters

Stellar flux density over the stellar disc is modelled by a quadratic law in cosine of the angle spanned by the observer, centre of the star, and the point on the surface of the star, as proposed in Kipping (2013). Coefficients of the polynomial u_1 and u_2 can only take values in the triangular region of the $\mathbb{R}^2$ set by the requirement that physical profiles must have positive flux density and that flux density is decreasing as the angle is increasing. In Kipping (2013), they give a convenient reparametrization $(u_1, u_2) \mapsto (q_1, q_2)$, such that physical region corresponds to $(q_1, q_2) \in [0, 1]^2$. Reparametrization also decreases correlations between the limb darkening coefficients, so we use these coordinates.

Limb darkening parameters can be calculated from the properties of the star: effective surface temperature, surface gravitational acceleration, and metallicity (Claret & Bloemen 2011). As a result of the experimental error in the stellar parameters and systematic errors in the predictions, limb darkening coefficients are still only constrained to a small subspace of the $[0, 1]^2$ plane. Limb darkening parameters do not have a strong impact on the SNR of the planet candidate as compared to period P, phase ϕ, and transit duration τ: Theoretical predictions from the stellar properties can be used as an initial guess to find all of the planet candidates. Joint SNR of all planet candidates can then be maximized with respect to the limb darkening parameters if needed.

4.2 Noise power spectrum

Noise is composed of a white detector noise and a correlated component due to the stellar variations. Power spectrum of the stellar variations approaches zero at high frequencies, making the white noise component dominant. We assume that the true noise power spectrum is a smooth function of frequency. After Fourier transforming the data and multiplying with the complex conjugate, we perform a band power averaging to reduce the variance of individual bandpowers. At high frequencies, we use wider bands, as power spectrum is roughly constant, and variations are mostly the standard fluctuations of a GP.

The presence of the data gaps requires the iterative estimation of the power spectrum (Robnik & Seljak 2020) where in the n-th step the just estimated power spectrum $P^{(n)}$ is used to simulate a flux series with gaps whose power spectrum $P^{(n)}_{\rm sim}$ is used to correct the estimate of the power spectrum in the next step $P^{(n+1)} = P + P^{(n)} - P^{(n)}_{\rm sim}$, where P is the estimate of the power spectrum obtained from the data. In practice, a few steps are sufficient for the convergence.

If a large planet signal is present in the data, it affects the power spectrum at high frequencies due to its U-shape. Thus, we first find the large planets using a first rough estimation of the power spectrum, eliminate these large planets from the flux, and recalculate the power spectrum. We only do this once, as we find that an additional iteration on this process is not necessary.

4.3 Search for the large planets

By large planets we mean those planets whose individual events are significant on their own. First, using equation (23) we convolve the data with the template forms of different transit durations. We maximize over the transit duration first and then find peaks over the time of the transit. For the found candidates, we then optimize the exact SNR from equation (22) that properly accounts for the gaps. These candidates are not forced to have equal transit time, and a subsequent analysis can be used to determine their TTVs.

4.4 Search for the small planets

Finding small planets is more challenging, especially those on the edge of detectability. It requires a search over their period, phase, and transit duration. First, using equation (23) we convolve the data with template forms of different transit durations. Then, using equation (25) we search over the planet's parameter space. If the planet's orbit was circular and perfectly aligned with the line of sight the transit duration would be completely fixed by the period: $\tau_K = qP^{1/3}$, by the Kepler's third law. The proportionality constant is given by the stellar radius R_* and the stellar mass M_*: $q = R_*(4/\pi GM_*)^{1/3}$. Inclined and elliptical orbits allow for a different transit duration, but it is sufficient to first fix the transit duration to $\tau_K(P)$, maximizing over ϕ, and then only consider different τ for the highest SNR candidates. This yields a top candidate at a

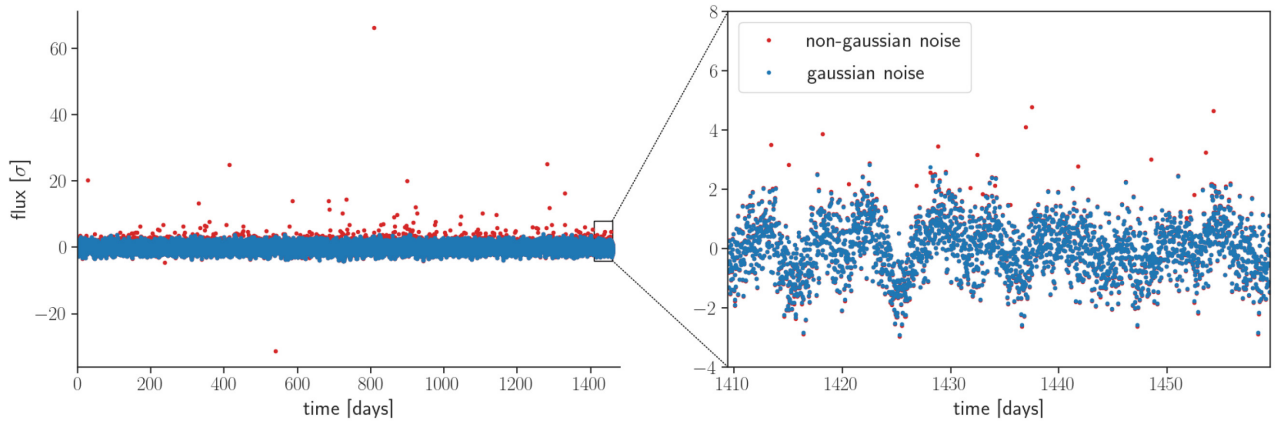

Figure 4. Simulation of stellar background time series (blue) and added non-Gaussian outliers (red). Noise is correlated with a red power spectrum. In this example, the power spectrum is extracted from the Kepler 90 data (shown in Fig. 2). The right-hand panel is zoomed part of the left-hand panel, showing the correlations. This is one realization of the stellar background time series that will be used to test the false positive rate in Fig. 5.

given period, after neglecting all the other candidates at the same period, knowing that there cannot be two planets at exactly the same period. Peaks in the thus found $\mathrm{SNR}(P) = \max_{\phi,\tau}\mathrm{SNR}(P, \phi, \tau)$ correspond to the planet candidates. This procedure assures no real planets have been eliminated, but some of the candidates may be false positives. We eliminate next those candidates that are caused by some confirmed planet with higher SNR, e.g. its higher harmonics. Such false positives appear because it would be unnecessarily expensive to terminate the search every time a new candidate is found, since it would require to eliminate it from the data and start over, which can be expensive especially for the systems with many planets. Instead, we find all the candidates first, and then sweep over the candidates, starting at the highest SNR, removing them from the data one by one, and identifying candidates lower on the list that still have SNR above the threshold. This procedure ensures that the found planet candidates are independent and not higher harmonics of higher SNR candidates.

5 TESTING GMF ON SIMULATIONS AND REAL DATA

We now compare our GMF method with baseline methods used in the literature. We test the ability to extract the correct amplitude of the injected planet, provide the smallest errors and accurately estimate it, and to produce as few false positives as possible. We show that GMF significantly outperforms the baselines in terms of the false positives, and at the same time is as accurate in reproducing the SNR as the most sophisticated GP methods (which cannot even be used as a planet search engine due to their excessive computational cost).

5.1 False positive test

As our initial test, we simulate a background-only time series. Stellar background is a Gaussian correlated noise, with the power spectrum resembling the power spectrum of Kepler 90 (Robnik & Seljak 2020). A given realization is a random sample from a Gaussian distribution. Later, we also add the noise outliers, such that each point is an outlier with some small probability a, independently of the other points, and is thus drawn from the non-Gaussian distribution (Section 2). A realization of both processes is shown in Fig. 4.

We search for the planet-shaped transits in the background over periods in range between $T_{\mathrm{min}} = 3$ d and $T_{\mathrm{max}} = 300$ d, over all phases and fix a transit duration to the Kepler value τ_K. We determine the maximal SNR as a function of period T cumulative from T_{min} to T. We report the median over different realizations, so that 50 per cent of realizations have higher SNR. In Fig. 5, we show the median of the maximal SNR as a function of the period T, so this is the highest SNR between 3 d and T range over which the search is performed.

First, we show the different methods of the planet search using the stellar background only. One common method is to approximate the stellar variability by spline fitting with spacing of nodes on a time-scale that is larger than the transit duration (Vanderburg & Johnson 2014), and then subtract it from the time series. A matched filter with a flat noise power spectrum is then performed to find the planet candidates. Another approximation for the stellar variability is to take a moving median across bins that are longer than a typical transit duration (Foreman-Mackey et al. 2016). Wavelet-based methods for the planet search (Tenenbaum et al. 2012) similarly assume a separation of the time-scale between the planet duration and the star variability. In Fig. 5, we show that our method has significantly lower false positive rate than the spline fitting and the moving median.

Next, we add outliers to the background. Here, we use our best performed matched filter method to address the impact of noise outliers only. A standard practice (Jenkins et al. 2017) is to introduce a cut-off on the outliers such that all points with a flux deviation from the mean larger than a cut-off are discarded from the series, here chosen to be 7σ. This is problematic for the negative outliers because the excluded points may be from a planet signal. In contrast, Gaussianization fully preserves the planet signatures. However, even in the absence of this problem, such as in Fig. 5, cut-off method is not as good as the Gaussianization, because it must be made in the region where the Gaussian distribution is negligible, whereas Gaussianization also operates in the region below 7σ, where both the outlier distribution and the Gaussian distribution are important. Gaussianization procedure is optimal noise outlier suppression method by construction, and this is reflected in the reduced false positive rate.

We note that positive outliers can produce false positive events in the presence of the stellar variation because the effective template is positive in some regions; see Fig. 2. Positive outliers in the *Kepler* data are more prominent than the negative outliers, which makes this effect comparable to the false positives from the negative outliers. This is not an artefact of match filtering with the stellar variability,

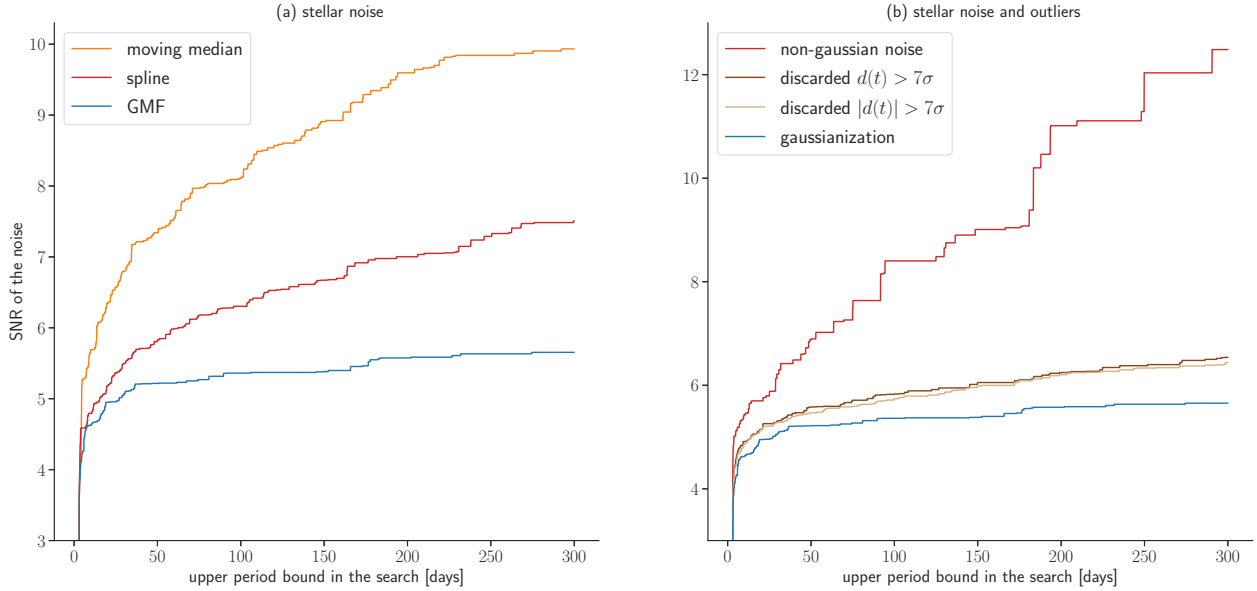

Figure 5. Expected maximal SNR event in the background-only simulation as a function of the maximal period T, scanning over a period from 3 d to T. Maximum SNR event in the search at each upper limit T is computed for 30 noise realizations. We show the median over the realizations (50 per cent have higher false positive SNR). Left-hand panel (a): Here, the background is a Gaussian, correlated noise with a power spectrum taken from the Kepler 90 (example of a realization of such noise is shown in blue in Fig. 4). We show different methods to the planet search: Moving median (orange) eliminates star flux by a moving median, and spline (red) does this with a spline interpolation. They both eliminate the estimated stellar flux and proceed by using a matched filter with a flat noise power spectrum. GMF (blue) proposed in this paper does not eliminate the stellar flux but treats it as a correlated component of the noise. GMF has significantly lower median SNR of false events. Right-hand panel (b): Non-Gaussian outliers are added to the time series (example of a realization is shown in red in Fig. 4). We show the case where outliers are not accounted for (red), a cut-off of positive outliers (dark brown), a cut-off of all outliers (light brown), and the Gaussianization (blue). In all cases, we use our frequency-dependent match filter that performed best in the case (a). Gaussianization has the lowest SNR of the false positive events, so it has the lowest false positive rate at a given SNR.

and standard Kepler pipeline also encounters this problem (Jenkins et al. 2017).

The baseline methods in the literature use outlier removal at 12.3σ (Jenkins et al. 2017), which would fall between the 7σ rejection and no rejection in the right-hand side of Fig. 5.

As a result, we expect the median of a false positive to be around 6–7 for the lower periods below 50 d, increasing to 7–12 for the longer periods. Kepler threshold of 7.1 (Jenkins et al. 2017) may give a low false positive rate for the shorter periods if spline or wavelets are used, but for the longer periods it may not be sufficiently conservative, as we expect many false positives with SNR > 7. In contrast, our GMF method achieves a median false positive SNR of 5.3 even at long periods: This can lead to a dramatic difference in the efficiency of planet detections, especially for Earth-like planets in the habitable zone with periods longer than 200 d. At the equal false positive rate, we expect GMF to be sensitive to up to 30 per cent lower transit amplitudes.

5.2 Planet parameters

To explore the accuracy of the planet parameter estimates, we inject a planet transit signature with a known SNR_0 into the stellar background and test GMF capability to reproduce the injected parameters of the planet: period, phase, transit duration, and amplitude. We test our method against the more sophisticated methods Fourier GP (Robnik & Seljak 2020) and Celerite GP (Foreman-Mackey et al. 2017), which are computationally too expensive for a planet search, as they need a good initial guess of the planets' parameters. The Fourier GP method is argued to be optimal under the assumption of Gaussian stellar variability, so comparing GMF against these methods is informative on the (sub)optimality of GMF.

We show in Fig. 6 that GMF is as good in reproducing the parameters of a planet as the Fourier GP, and is significantly better than the spline fitting, both in terms of bias and variance. The errors equal those of Fourier GP, which is near-optimal. We also show that the matched filter correctly predicts the errors on the planet's parameters. Covariance matrix of the planet's parameters is calculated as an inverse of the Hessian of the $-\log p$, and this analytical method matches the variance obtained from the simulations well. Note the ease of estimating the covariance matrix compared to the GP methods where a marginalization over the stellar parameters is required.

There is good agreement between GMF and Fourier GP. This shows that GMF can be used for the full analysis, not only for the initial detection analysis, as it gives optimal results on all the parameters of interest.

We also test GMF's performance in different stellar backgrounds, to show that it is a general method that can be applied to any star. We take a selection of *Kepler*'s targets and perform the planet search as described in Section 4. We retrieve the known planets and subtract them from the data. We then inject a planet in the data and test GMF's ability to extract its parameters. Results are shown in Fig. 7. GMF performs well on all tested stars, despite their qualitatively different behaviour.

5.3 Analysing real planets

We have applied the GMF to the planet search in several *Kepler*'s targets and have retrieved all officially confirmed planets. We show the results for the Kepler 90 system in Table 1 for the large planets

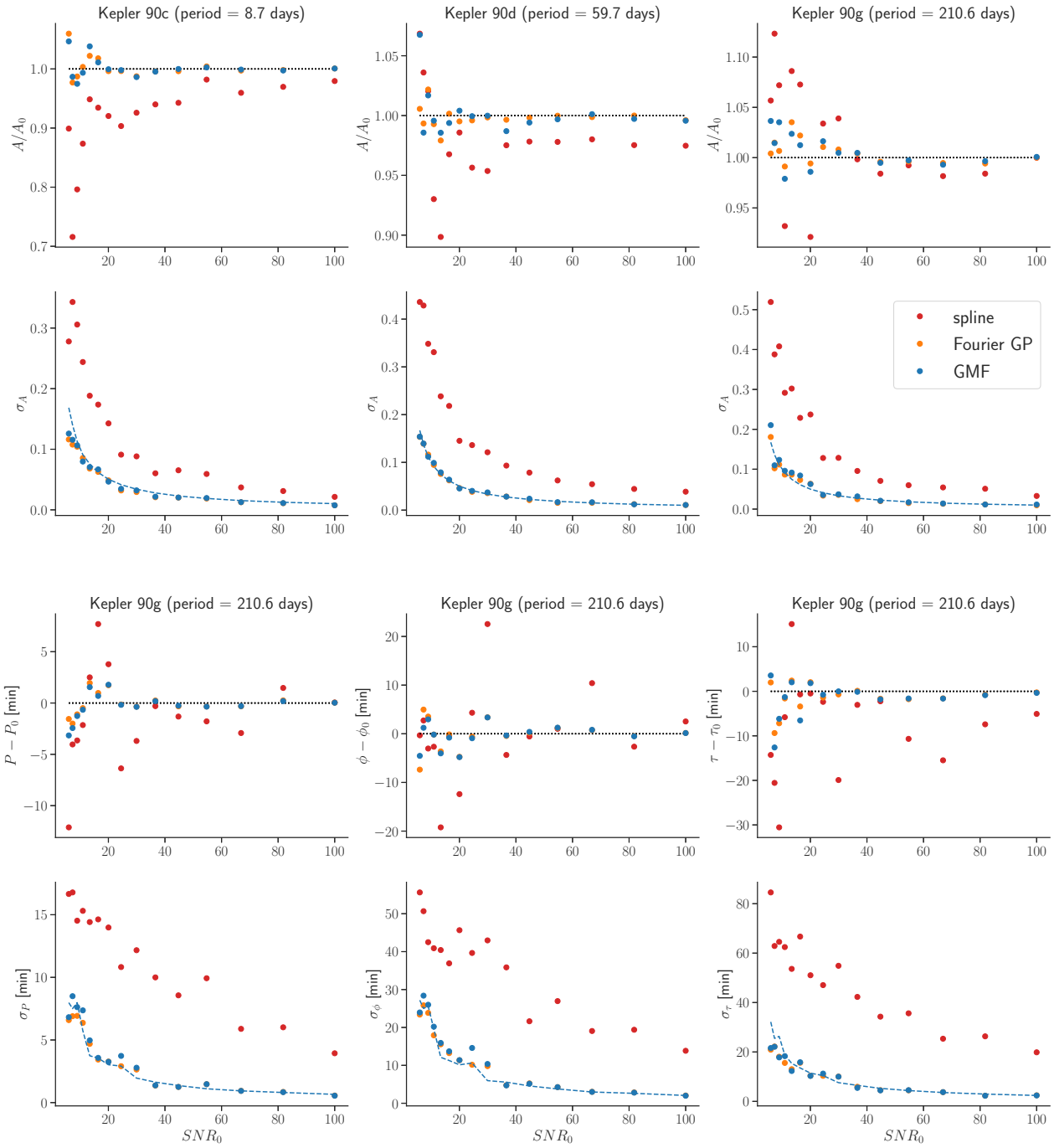

Figure 6. Parameters obtained by the matched filter are compared to the known value of the injected synthetic planets' parameters, as a function of the injected SNR_0 in the range of 6–100. Upper row shows the expected value of the ratio A/A_0 averaged over 30 randomly chosen planet phases to reveal possible bias. Second row shows the variance of A. Three choices for the period of the injected planet are chosen resembling the known planets Kepler 90c, Kepler 90d, and Kepler 90g. The last two rows show the bias and variance of the other three parameters: period, phase, and transit duration. In all figures, a dotted envelope is a GMF prediction of the error as computed from the Hessian at the SNR peak. It matches well the variances from the simulations. Matched filter is compared to the Fourier GP and to the spline fitting, showing that is as accurate as FGP and significantly better than the spline fitting in both bias and variance.

where each transit is identified individually, and in Table 2 for the smaller planets that are folded over their period. In this paper, we do not consider the proposed Kepler 90i (Shallue & Vanderburg 2018), which is close to the detection threshold, and requires a more careful analysis of the look-elsewhere effect (Bayer & Seljak 2020), which we will pursue elsewhere. Transit time is defined as the time at the centre of the transit measured relative to the beginning of the *Kepler*'s measurements in Kepler 90, which is HJD - 2454833 + 131.5124 d. Phase is defined as the transit time of the first observed transit.

We compare these results to those of Fourier GP. We again confirm a perfect agreement between the GMF and the Fourier GP on estimated SNR; see Table 2. GMF can be used not only as a planet search algorithm, but can also replace expensive GP-based analysis methods such as Fourier GP or Celerite.

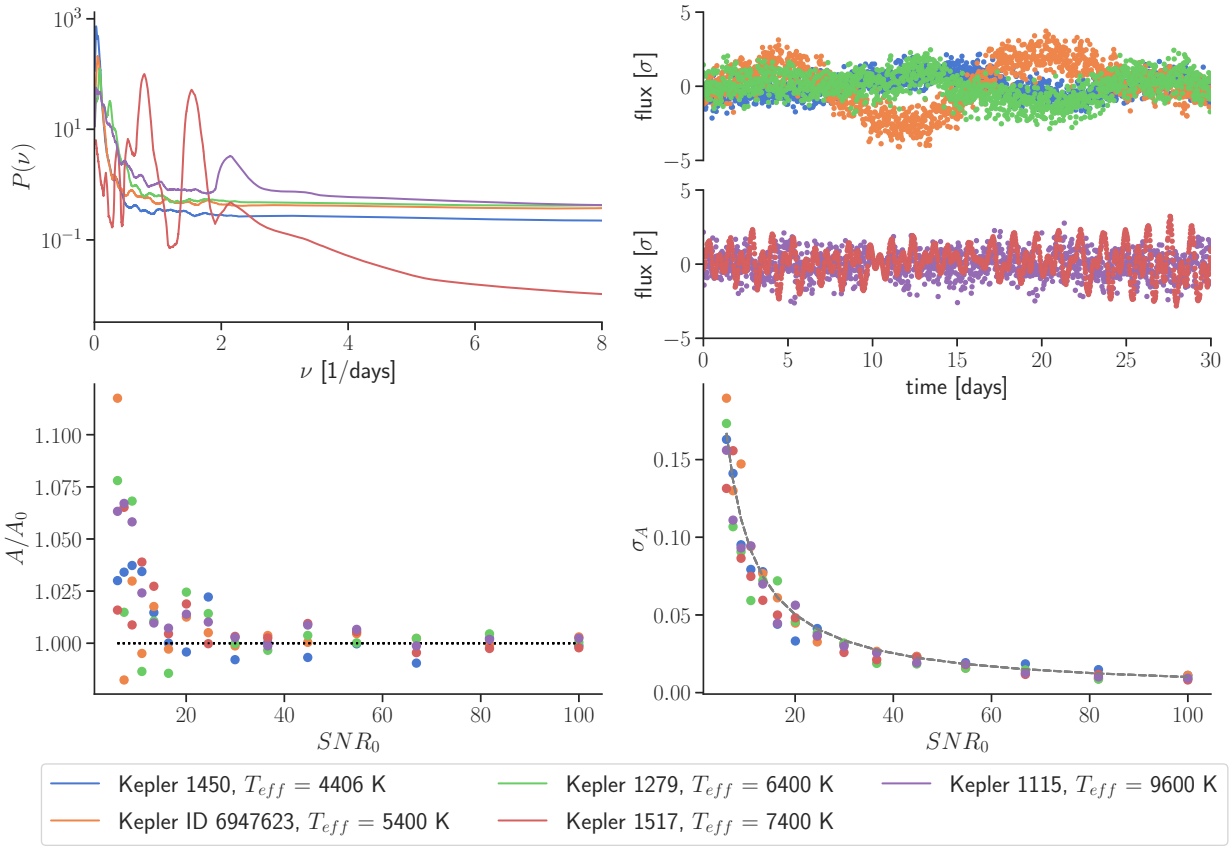

Figure 7. We test GMF on a variety of stellar backgrounds. The upper left plot shows the power spectra of the analysed stars. Note the large differences in the stellar power spectra; for example, Kepler 7515679 (green) has a power spectrum amplitude that ranges over four orders of magnitude across the range of frequencies shown, while the power spectrum of Kepler 5022828 (red) spans only two orders of magnitude. Upper right plot shows a segment of *Kepler*'s measurements for those stars. Time is measured relative to the beginning of *Kepler*'s measurement for that star. Very red spectra (large power at low frequencies compared to the high frequencies), for example Kepler 7515679, are seen as a strongly correlated stellar flux on the time-scales shown. We take *Kepler*'s measurements, find planets, and eliminate them. We then inject a synthetic planet with a period 30 d in the stellar flux and test GMF's ability to extract the injected planet's amplitude A_0. Lower left plot shows the expected value of the ratio A/A_0 averaged over 30 randomly chosen planet phases as a function of the injected planet's SNR_0 in the range of 6–100. Lower right plot shows the variance of A. A dotted envelope is a GMF prediction of the error as computed from the Hessian at the SNR peak. It matches well the variances from the simulations. GMF is performing well, regardless of the stellar background.

Table 1. Individual transits of the large planets of Kepler 90 (90g and 90h) for GMF, both transit time, transit duration, and the estimated error. Results from the Fourier GP (Robnik & Seljak 2020) are also shown relative to GMF, but the two are not completely comparable, because FGP results also account for the non-uniformity of *Kepler* data point measurements, which is an effect of the order of the error of the parameter. As a consequence, for both transit time and transit duration the agreement between GMF and FGP can be a factor of 2 larger than the estimated GMF errors. Note that for the transit time a typical difference between the two is of the order of 3 min, while the data are given in 29.4 min intervals.

Planet	SNR	Transit time (d)	Transit duration (h)	FGP transit time (d)	FGP transit duration (h)
h	108.1	8.964 22 ± 0.0006	13.341 ± 0.029	0.0018	0.06
g	53.9	15.5869 ± 0.0011	11.392 ± 0.055	0.0018	0.07
g	53.9	226.0428 ± 0.0011	11.349 ± 0.054	−0.002	0.02
h	109.7	340.607 74 ± 0.0007	13.462 ± 0.034	0.0005	−0.11
g	53.6	436.7707 ± 0.0011	11.432 ± 0.055	0.0025	0.04
g	53.5	857.96 ± 0.0012	11.621 ± 0.059	0.0051	−0.0
g	55.4	1068.5496 ± 0.0025	11.5 ± 0.12	0.0045	0.01
g	52.8	1280.221 ± 0.0012	11.676 ± 0.06	0.0062	−0.04
h	105.9	1335.3784 ± 0.000 58	13.297 ± 0.028	0.0026	0.06

6 CONCLUSIONS

In this paper, we propose a method for planet detection in transit data, which is based on matched filter technique, combined with the Gaussianization method for the noise outliers, which we call Gaussianized matched filter (GMF). We show that GMF significantly outperforms standard baselines in terms of reducing the false positive rate: While standard methods give median false positive signal to noise as high as 8, the corresponding number for GMF is 5.3. Since the number of false positives explodes exponentially at lower SNR values, this could enable a significantly lower false positive rate of

Table 2. Joint transits of the small Kepler 90 planets from GMF method. We also show SNR of Fourier GP, which agrees well with the SNR of GMF.

Planet	SNR	Period (d)	Phase (d)	Transit duration (h)	FGP SNR
d	27.0	59.736 88 ± 0.000 28	27.4461 ± 0.003	8.003 ± 0.097	29.6
e	23.1	91.9401 ± 0.000 42	2.7822 ± 0.0037	8.96 ± 0.12	23.5
f	17.9	124.915 ± 0.0013	123.1868 ± 0.0075	10.08 ± 0.18	16.6
c	16.2	8.719 734 ± 2.9e-05	8.0097 ± 0.002	3.831 ± 0.092	15.8
b	15.4	7.008 546 ± 3.2e-05	6.1096 ± 0.0033	4.32 ± 0.13	15.5

faint planets, especially for Earth-like planets in the habitable zone. Alternatively, at the equal false positive rate we expect GMF to detect planets with up to 30 per cent lower transit amplitudes.

A procedure for GMF planet search can be summarized as follows:

(i) Gaussianize the data by remapping the outliers.

(ii) Calculate the noise power spectrum and use it in inverse noise weighting in the convolution of the data with the transit profile Fourier transform.

(iii) Search over the time of transit and transit duration to find the individual transits of the big planets.

(iv) Search over the period, phase, and transit duration for the small planets.

The method eliminates outlier false positives and stellar variability positives that in the standard Kepler pipeline need to be eliminated in the post-processing phase using RoboVetter (Thompson et al. 2018). Further false positives that need to be considered are binary stars (Foreman-Mackey et al. 2016) and off-target false positives, and we plan to address these elsewhere.

A remarkable feature of the GMF method is that it can be used not only for the initial planet detection but also for the final planet parameter analysis. By comparison against the state-of-the-art GP methods on both simulations and real data, we observe that GMF achieves near-optimal results on amplitude, period, phase, and transit time. Moreover, a simple analytic Laplace approximation of the Hessian gives reliable error estimates on these parameters. Thus, GMF may be not only the fastest method to detect planets, with the lowest false positive rate, but also the most accurate method to extract their parameters.

GMF provides parameter estimates and their errors as a compressed summary statistic, from which one can form a likelihood that one can use for the more involved inverse problems, where optimization or Markov chain Monte Carlo analysis is needed to find the solution. One such example is a transit timing variation (TTV) analysis (Liang, Robnik & Seljak 2020), where transit times and transit durations and their errors from the GMF analysis of individual transits in Table 1 have been used to form a data likelihood. This enabled a subsequent inverse problem analysis that identified the models that can explain TTVs. In this example, this led to a determination of all of the orbital parameters and the masses of Kepler 90g and h, and the discovery of Kepler 90g as a superpuff.

ACKNOWLEDGEMENTS

We acknowledge Ad futura Slovenia for supporting JR MSc study at Eidgenössische Technische Hochschule (ETH) Zürich. This material is based on work supported by the National Science Foundation under grant numbers 1814370 and NSF 1839217, and by National Aeronautics and Space Administration (NASA) under grant number 80NSSC18K1274. This paper includes data collected by the *Kepler* mission. Funding for the *Kepler* mission is provided by the NASA Science Mission directorate.

DATA AVAILABILITY

The data underlying this article are available in NASA Exoplanet Archive, at https://exoplanetarchive.ipac.caltech.edu/bulk_data_download/.

REFERENCES

Abbott B. P. et al., 2016, Phys. Rev. D, 93, 122003
Akeson R. et al., 2013, PASP, 125, 989
Bayer A. E., Seljak U., 2020, J. Cosmol. Astropart. Phys., 2020, 009
Bryson S. T. et al., 2017, Technical Report, Kepler Certified False Positive Table. NASA, Washington, DC
Claret A., Bloemen S., 2011, A&A, 529, A75
Foreman-Mackey D., Morton T. D., Hogg D. W., Agol E., Schölkopf B., 2016, AJ, 152, 206
Foreman-Mackey D., Agol E., Ambikasaran S., Angus R., 2017, AJ, 154, 220
France K. et al., 2017, in Oswald H. S., ed., Proc. SPIE Conf. Ser. Vol. 10397, UV, X-ray, and Gamma-Ray Space Instrumentation for Astronomy XX. SPIE, Bellingham, p. 1039713
Jenkins J. M., Tenenbaum P., Seader S., Burke C. J., McCauliff S. D., Smith J. C., Twicken J. D., Chandrasekaran H., 2017, Kepler Data Processing Handbook: Transiting Planet Search (Kepler Science Document). NASA Ames Res. Cent., Moffett Field, CA
Kipping D. M., 2013, MNRAS, 435, 2152
Koch D. G. et al., 2010, ApJ, 713, L79
Kopparapu R. K., 2013, ApJ, 767, L8
Liang Y., Robnik J., Seljak U., 2020, AJ, 161, 202
Parviainen H., 2015, MNRAS, 450, 3233
Robnik J., Seljak U., 2020, AJ, 159, 224
Shallue C. J., Vanderburg A., 2018, AJ, 155, 94
Steffen J. H., Coughlin J. L., 2016, Proc. Natl. Acad. Sci., 113, 12023
Tenenbaum P. et al., 2012, ApJS, 199, 24
Tenenbaum P. et al., 2013, ApJS, 206, 5
Thompson S. E. et al., 2018, ApJS, 235, 38
Vanderburg A., Johnson J. A., 2014, PASP, 126, 948

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