The Astronomical Journal, 160:240 (24pp), 2020 November https://doi.org/10.3847/1538-3881/abbc16

A Fast, Two-dimensional Gaussian Process Method Based on Celerite: Applications to Transiting Exoplanet Discovery and Characterization

Tyler A. Gordon[1], Eric Agol[1], and Daniel Foreman-Mackey[2]
[1] Department of Astronomy, University of Washington, Box 351580, U.W., Seattle, WA 98195-1580, USA; tagordon@uw.edu
[2] Center for Computational Astrophysics, Flatiron Institute, 162 5th Ave., New York, NY 10010, USA
Received 2020 July 10; revised 2020 September 17; accepted 2020 September 24; published 2020 November 3

Abstract

Gaussian processes (GPs) are commonly used as a model of stochastic variability in astrophysical time series. In particular, GPs are frequently employed to account for correlated stellar variability in planetary transit light curves. The efficient application of GPs to light curves containing thousands to tens of thousands of data points has been made possible by recent advances in GP methods, including the celerite method. Here we present an extension of the celerite method to two input dimensions where, typically, the second dimension is small. This method scales linearly with the total number of data points when the noise in each large dimension is proportional to the same celerite kernel and only the amplitude of the correlated noise varies in the second dimension. We demonstrate the application of this method to the problem of measuring precise transit parameters from multiwavelength light curves and show that it has the potential to improve transit parameters measurements by orders of magnitude. Applications of this method include transit spectroscopy and exomoon detection, as well a broader set of astronomical problems.

Unified Astronomy Thesaurus concepts: Exoplanets (498); Natural satellites (Extrasolar) (483); Transmission spectroscopy (2133); Transits (1711); Gaussian Processes regression (1930); Astrostatistics (1882); Bayesian statistics (1900)

1. Introduction

All exoplanet transit observations must contend with the presence of noise. Light curves can display both uncorrelated, or white, noise and correlated noise. While white noise often results from the the statistics of photon counting, and may only be ameliorated by collecting more photons, correlated noise can arise from a variety of sources. These can be broadly divided into two categories: astrophysical noise, which results from physical processes at the source of the observed photons such as stellar granulation and oscillations (Pereira et al. 2019; Barros et al. 2020; Morris et al. 2020; Sulis et al. 2020), and instrumental noise, which results from imperfections in detectors, errors in spacecraft pointing, or other processes taking place at the location of the observer rather than at the source.

Our ability to detect transits and infer their parameters depends on how well we can model both white and correlated noise. While white noise is straightforward to model as a Gaussian distributed random variable,[3] correlated noise can be more challenging to account for. Additionally, as more powerful telescopes yield more precise observations, photon-counting noise will decrease while astrophysical correlated noise (which does not depend on photon counts) will not. In fact, correlated noise will become more dominant as decreasing white noise amplitudes reveal previously undetectable variability.

A number of methods have been used to model, filter, or otherwise account for correlated noise in astrophysics, dating back to work by Rybicki & Press (1992, 1995). Among these techniques are wavelet filtering (Carter et al. 2008) and Kalman filtering (Kelly et al. 2014). A comprehensive study of various detrending methods is given in Hippke et al. (2019). These include various sliding filter methods (such as a sliding mean or median), sums of sines or cosines (Mazeh & Faigler 2010; Kipping et al. 2013), and others.

Our work focuses on the Gaussian process (GP) method of modeling correlated noise. In this paper we introduce an extension to the popular celerite code which can be used to model correlated noise in two-dimensions. We use this extension to simulate multiwavelength stellar variability in transit observations. We show that by accurately modeling correlation across wavelengths we can improve measurements of transit parameters by orders of magnitude in some common limits.

While this paper focuses on multiwavelength transit observations with a small number of bands, our method also naturally extends to transit spectroscopy as the number of bands becomes large. In this paper we consider a trapezoidal transit model that has no wavelength dependence, but a wavelength-dependent transit model can easily be incorporated. For transit spectroscopy, the transit depth and limb-darkening parameters should be allowed to vary between bands.

We further assume that the data have been preprocessed to remove instrumental systematics. Long-term trends in the observations may either be removed during preprocessing or incorporated into the mean of the GP model. If removed during preprocessing any uncertainties introduced should be carefully accounted for so that they can be incorporated into the GP. In the case studies presented in Section 4 we assume a zero mean, indicating that the observations have been normalized to zero in each band. We assume there are either no long-term trends present or that they have been removed in preprocessing without introducing any meaningful additional uncertainty.

The final assumption we make is that the correlated component of the variability is stationary, by which we mean that the parameters describing the correlated noise do not change with time. This is a fundamental limitation that is

[3] Which is the limit of a Poisson distribution at high photon count rates.

inherited from the 1D celerite method. Our method does, however,allow for a heteroscedastic white noise component. This means that each data point is allowed to have a unique measured uncertainty that varies from point to point in both time and wavelength.

1.1. A Short Introduction to GPs

While more general definitions of GPs may be formulated, it is most helpful for our purposes to view GPs as an ordered collection of random variables along one or more axes often representing time or space. In the case of an exoplanet transit the random variables model a series of observations of the star's flux taken at discrete times. The Gaussian aspect of a GP describes the relationship between random variables—we model $N_{\rm obs}$ observations with an $N_{\rm obs}$-dimensional Gaussian distribution. The covariance of the multidimensional Gaussian is described by a kernel function, which gives the covariance between any pair of observations as a function of their separation in time or space. The kernel function then defines the covariance matrix. For a kernel $k(x_i, x_j)$, we have

$$K_{ij} = k(x_i, x_j) + d_{ij}\sigma_i^2, \quad (1)$$

where d_{ij} is the Kronecker delta function and σ_i is the white noise component for the ith data point. In addition to the kernel function, a GP is characterized by its mean function, $m(t)$, which describes the deterministic component of the process. In the case of an exoplanet transit we use a transit model as the mean function. The GP likelihood function, $\mathcal{L}$, describes the likelihood that a set of observations, $\mathbf{y}$, is drawn from the GP. It is written as

$$\ln \mathcal{L} = -\frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^{\rm T}\mathbf{K}^{-1}(\mathbf{y}-\boldsymbol{\mu}) - \frac{1}{2}\ln\det(\mathbf{K}) - \frac{N_{\rm obs}}{2}\ln(2\pi) \quad (2)$$

where $\boldsymbol{\mu}$ is a vector where the entries are given by $\mu_i = m(x_i)$.

A typical and much simplified procedure for measuring exoplanet transit parameters using a GP noise model (as applied in Dawson et al. 2014; Barclay et al. 2015, and Chakrabarty & Sengupta 2019 among others) can be summarized as follows.

1. Choose a suitable kernel function to describe the correlated noise.
2. Choose a transit model to use as the GP mean function.
3. Optionally, maximize the GP likelihood with respect to the mean and kernel parameters or use another method to obtain a starting point for initializing Markov chain Monte Carlo (MCMC) chains.
4. Use an MC method to sample the posterior (defined by the GP likelihood and priors for each parameter) in order to estimate uncertainties for the transit and kernel parameters.

Choosing a suitable kernel function can be a complicated task. The choice of kernel might be motivated by prior knowledge of the characteristics of the data, or might be "learned" from the data, such spectral mixing kernels (Wilson & Adams 2013). In Section 2 we discuss and justify our choice of a kernel for modeling stellar variability which has the added benefit of being easily expressible in the celerite framework. For a more complete discussion of model selection in the general case we refer the reader to chapter 5 of Rasmussen & Williams (2006).

When searching for a previously undetected transit, the results of step 3 will suggest the most likely parameters of the transit. In a Bayesian framework the posterior estimates obtained from the MCMC analysis can then be used to estimate the evidence for a transit with respect to a flat mean.

In the case of a monochromatic light curve this procedure is effective at identifying transits when the depth or duration of the transit differs significantly from the amplitude and characteristic timescales of the noise. For instance, a transit that is much deeper than the noise amplitude is poorly described by the GP noise model and thus the likelihood will be sharply peaked at the location of the correct transit parameters. Similarly, a transit that occurs on a much shorter timescale than the characteristic timescale of the variability will be poorly described by the GP and hence easily detectable via the likelihood. Figures 5 and 6 below illustrate these instances.

A problem occurs when the transit depth and duration are comparable to the noise amplitude and timescale. In this case the GP covariance alone is able to fit the transit without the need for a mean model. The result is that the GP likelihood is not sharply peaked about the location of the correct transit parameters and the transit is thus difficult or impossible to detect. Gathering more photons with a larger telescope does not fix the problem as the correlated noise does not decrease with higher photon count rates as white noise does. One simply obtains a better measurement of the correlated noise, but the transit remains masked by the variability.

One solution to this problem is to gather light in multiple wave bands. With a multiband light curve we can leverage the difference in the spectral dependence of the transit as compared to the correlated variability to disentangle the transit from the noise, and thus detect shallower transits across a broader range in duration than is possible with monochromatic observations. This approach depends upon the assumption that the correlated noise has the same time dependence for each component of the power spectrum, but varies in amplitude with wavelength. If, on the other hand, the correlated noise is achromatic, multiple wave bands will not improve upon the monochromatic case. For the remainder of this paper we will assume that there is in fact a wavelength dependence to the correlated noise which shares a common time dependence. While in the general case a time delay may be included in these models, the common time dependence is a requirement of the fast GP methods derived here. We exploit the information contained in this time dependence to demonstrate an improvement on the inference of transiting planet parameters. We also ignore instrumental/systematic variations, and assume that the white noise is dominated by Poisson photon counting uncertainty.

In Section 2 we describe our wavelength-dependent stellar variability model. We then review the one-dimensional version of celerite before describing our extension to two-dimensions (Section 3). Next, we conduct an Information analysis to derive approximate, semi-analytic upper bounds on the precision that can be achieved when inferring transit parameters from multiband light curves with different noise properties, and compare the results of our Information analysis to a full MCMC treatment for select noise parameters (Section 4). In the discussion (Section 5) we outline additional applications of our method including exomoon detection and transmission

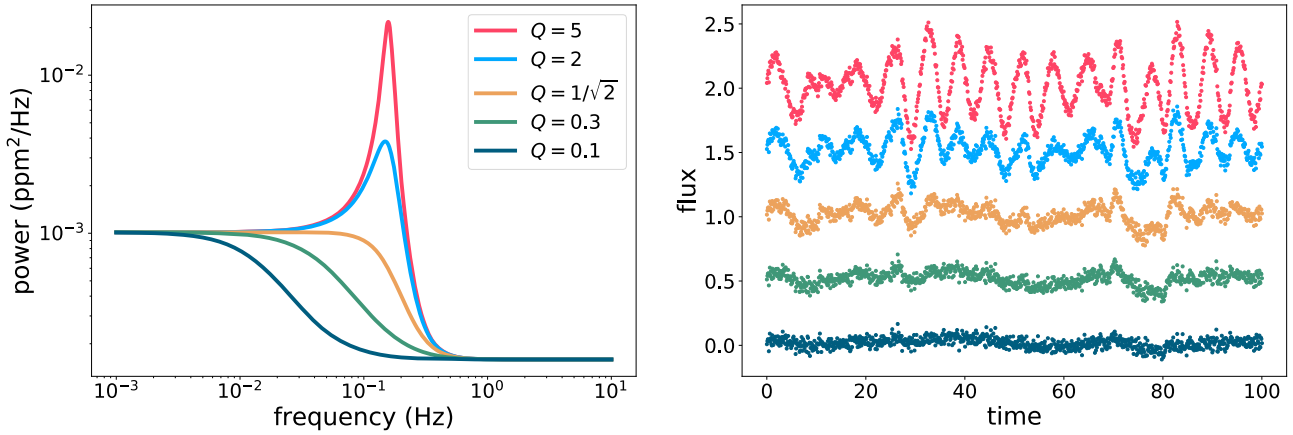

Figure 1. Left: power spectrum of the simple harmonic oscillator kernel for several values of the quality factor Q. For $Q < 1/\sqrt{2}$ the system is overdamped. For $Q > 1/\sqrt{2}$ the system is underdamped and the Gaussian process (GP) shows oscillations at the characteristic frequency. For our simulations we set $Q = 1/\sqrt{2}$, in which case the system is critically damped. Right: noise realizations for each power spectrum on the left. Note the decreasing coherency of the oscillations as we r from high to low values of Q. The decreasing noise amplitudes from top to bottom are a result of the fact that the GPs with larger Q values have more total power a constant S_0.

spectroscopy. We conclude in Section 6 with a discussion of the limitations of and potential improvements to our method.

2. Multiwavelength Noise Model

Here we describe our model for noise that is correlated across both time and wavelength. We start with a description of the time dependence of the noise.

2.1. Time-correlated Variability Model

Foreman-Mackey et al. (2017) describe how celerite can be used as a physically motivated model for stellar variability. The following discussion is closely based on the discussion in that paper.

We follow Anderson & Jefferies (1990) in modeling stellar oscillations as the result of stochastic excitations that are damped by convection and turbulent viscosity in the star. This process is described by the differential equation

$$\frac{1}{\omega_0^2}\frac{d^2}{dt^2}y(t) + \frac{1}{\omega_0 Q}\frac{d}{dt}y(t) + y(t) = \epsilon(t) \quad (3)$$

where ω_0 is the characteristic frequency of the oscillator, Q is the quality factor of the oscillator, $\epsilon(t)$ is a stochastic driving force, and y(t) is the amplitude of the oscillations. If $\epsilon(t)$ is Gaussian distributed then the solution to Equation (3) is a GP with the power spectral density

$$S(\omega) = \sqrt{\frac{2}{\pi}}\frac{S_0\omega_0^4}{(\omega^2 - \omega_0^2)^2 + \omega_0^2\omega^2/Q^2}. \quad (4)$$

Figure 1 shows this power spectrum for several values of Q. For our modeling we set $Q = 1/\sqrt{2}$, in which case the power spectral density simplifies to

$$S(\omega) = \sqrt{\frac{2}{\pi}}\frac{S_0}{(\omega/\omega_0)^4 + 1}. \quad (5)$$

This power spectrum has been used to describe granulation-driven stellar variability (Kallinger et al. 2014). We employ this kernel in our work both because stellar granulation is a significant source of noise on transit timescales and to simplify our discussion of relevent noise timescales in Section 4.1 which would be complicated by the presence of oscillations at the characteristic frequency. Choosing different values for Q will not affect the qualitative aspects of our results.

The corresponding kernel function to Equation (5) is

$$k(t) = S_0\omega_0 e^{-\omega_0 t/\sqrt{2}}\cos\left(\frac{\omega_0 t}{\sqrt{2}} - \frac{\pi}{4}\right), \quad (6)$$

where $t = |t_i - t_j|$.

2.2. Wavelength Dependence of Variability

We are now interested in constructing a simple model for the wavelength dependence of stellar variability based upon our time-dependent correlated variability model. To begin, we consider a two-component photosphere where each component has a unique spectrum and covering fraction. The star's variability is then a result of variations in the covering fraction of these components, and the covering fractions vary according to the stochastic process described in Section 2.1.

We label the two components "hot" and "cold." Their spectra are given by $S_h(\lambda)$ and $S_c(\lambda)$ and their covering fractions are given by x_h and $x_c = 1 - x_h$. In the absence of limb-darkening the flux observed in a band B_1 is given by

$$F_{B_1} = \frac{\pi R_*^2}{d^2}\int(x_c S_c(\lambda) + x_h S_h(\lambda))\,\tau_{B_1}(\lambda)\,d\lambda \quad (7)$$

where $\tau_{B_1}(\lambda)$ is the response curve for the filter and the integral is taken over all wavelengths, d is the distance from the observer to the star, and R_* is the stellar radius. Substituting $x_h = 1 - x_c$ allows us to rewrite this expression as

$$F_{B_1} = \frac{\pi R_*^2}{d^2}\left(\int S_h(\lambda)\,\tau_{B_1}(\lambda)\,d\lambda\right) - \frac{\pi R_*^2}{d^2}x_c\left(\int (S_h(\lambda) - S_c(\lambda))\,\tau_{B_1}(\lambda)\,d\lambda\right). \quad (8)$$

The first term of Equation (8) is the total flux for a photosphere completely covered by the hot component, and the second term is a correction dependent on the contrast between the hot and

cold components.For simplicity, we define

$$F_{\mathrm{B_1,hot}} = \frac{pR_*^2}{d^2}\int S_h(l)\,B_1(l)\,dl \quad (9)$$

and

$$a_1 = \frac{pR_*^2}{d^2}s_c\int (S_h(l) - S_c(l))\,B_1(l)\,dl, \quad (10)$$

where $s_c^2 = \mathrm{var}(x_c)$. With these definitions we have

$$F_{B_1} = F_{\mathrm{B_1,hot}} - \frac{x_c}{s_c}a_1. \quad (11)$$

We can do the same for a second hypothetical band, B_2, giving us

$$F_{B_2} = F_{\mathrm{B_2,hot}} - \frac{x_c}{s_c}a_2. \quad (12)$$

Since the only time-dependent quantity in Equations (11) and (12) is the covering fraction of the cold component, x_c, we see that the flux in each band will vary coherently with the same power spectral density and the amplitude of the variability will be set by the contrast between the hot and cold components of the photosphere in each band.

The covariance between two bands can now be computed:

$$\begin{aligned}\mathrm{cov}(F_{B_1}, F_{B_2}) &= s_c^{-2}\mathrm{cov}(x_c a_1, x_c a_2)\\ &= a_1 a_2 \mathrm{corr}(x_c, x_c).\end{aligned} \quad (13)$$

Now we let x_c be a function of time, $x_c(t)$, and assert that it is drawn from a one-dimensional GP evaluated at times t_i for $i = 1, K, N$ (i.e., a correlated time series) with a kernel which can be described with the celerite formalism. Then the full covariance matrix for the time and wavelength dimensions is given by the block matrix

$$K = \begin{bmatrix} S_1 + T_{1,1}R & T_{1,2}R & \dots & T_{1,N}R \\ T_{2,1}R & & & \\ \vdots & & & \\ T_{N,1}R & & & S_N + T_{N,N}R \end{bmatrix}, \quad (14)$$

where

$$S_i = \begin{pmatrix} s_{i,1}^2 & 0 \\ 0 & s_{i,2}^2 \end{pmatrix} \quad (15)$$

is a diagonal matrix containing the white noise components for each band at time i; $T_{ij} = \mathrm{corr}(x_c(t_i), x_c(t_j))$ is the time covariance matrix for the process described in Section 2.1 normalized by the variance of x_c, and R is the covariance matrix across bands, defined as

$$R = \begin{pmatrix} a_1^2 & a_1 a_2 \\ a_2 a_1 & a_2^2 \end{pmatrix}. \quad (16)$$

For M bands $B_1, B_2, \dots B_M$ with amplitudes given by $a_1, a_2, \dots a_M$, R becomes

$$\begin{aligned} R &= \begin{pmatrix} a_1^2 & a_1 a_2 & \dots & a_1 a_M \\ a_2 a_1 & & & a_2 a_M \\ \vdots & & & \vdots \\ a_M a_1 & a_M a_2 & \dots & a_M^2 \end{pmatrix} \\ &= \boldsymbol{a}\boldsymbol{a}^T \end{aligned} \quad (17)$$

where $\boldsymbol{a} = (a_1, a_2, \dots, a_M)^T$. The covariance matrix can now be written

$$K = S + T \otimes R, \quad (18)$$

where Σ is the block matrix

$$S = \begin{bmatrix} S_1 & \dots & 0 \\ 0 & & \\ \vdots & & S_N \end{bmatrix}, \quad (19)$$

and where ⊗ denotes the Kronecker product. The Kronecker product is defined for two matrices A and B with dimensions $N \times M$ and $P \times Q$, respectively, as the $NP \times MQ$ block matrix

$$A \otimes B = \begin{bmatrix} a_{1,1}\boldsymbol{B} & a_{2,1}\boldsymbol{B} & \dots & a_{1,N}\boldsymbol{B} \\ a_{2,1}\boldsymbol{B} & a_{2,2}\boldsymbol{B} & & \\ \vdots & & & \\ a_{N,1}\boldsymbol{B} & & & a_{N,N}\boldsymbol{B} \end{bmatrix}. \quad (20)$$

An important consideration for constructing new GP covariance matrices is that a valid covariance matrix must be positive-definite for all inputs (Rasmussen & Williams 2006). For a detailed discussion of the postive-definiteness of the 1D celerite kernel we refer the reader to Appendix A of Foreman-Mackey et al. (2017). Assuming that the covariance matrix T is positive-definite, the positive-definiteness of the full covariance $K = T \otimes R + S$ can be ascertained by considering the eigenvalues of the Kronecker product plus diagonal covariance matrix, which are uniformly positive if and only if the matrix is positive-definite. For now we state the conclusion that K is positive-definite if R is positive-definite, or if R is positive-semidefinite and Σ has all nonnegative entries along its diagonal. A proof is given in Appendix E. In the case that R is the outer product $\boldsymbol{a}\boldsymbol{a}^T$, R is positive-semidefinite and positive-definiteness is thus ensured as long as a nonzero white noise component is given for each data point.

In practice, providing too small a white noise component when R is positive-semidefinite may result in numerical instabilities. For this reason we recommend that care be taken when applying this method to extremely high-precision data. For arbitrary definitions of R positive-definiteness should be ensured on a case-by-case basis. In general it is sufficient to show that R is positive-definite, or that R is positive-semidefinite with a nonzero white noise amplitude provided for each data point.

When the number of bands, M, is small, this covariance matrix can be used to model multiband observations. We can also allow M to become arbitrarily large, in which case the resultant covariance matrix can be used to model spectral observations. Here each entry in $\boldsymbol{a}$ would represent the amplitude of the correlated variability in one wavelength bin

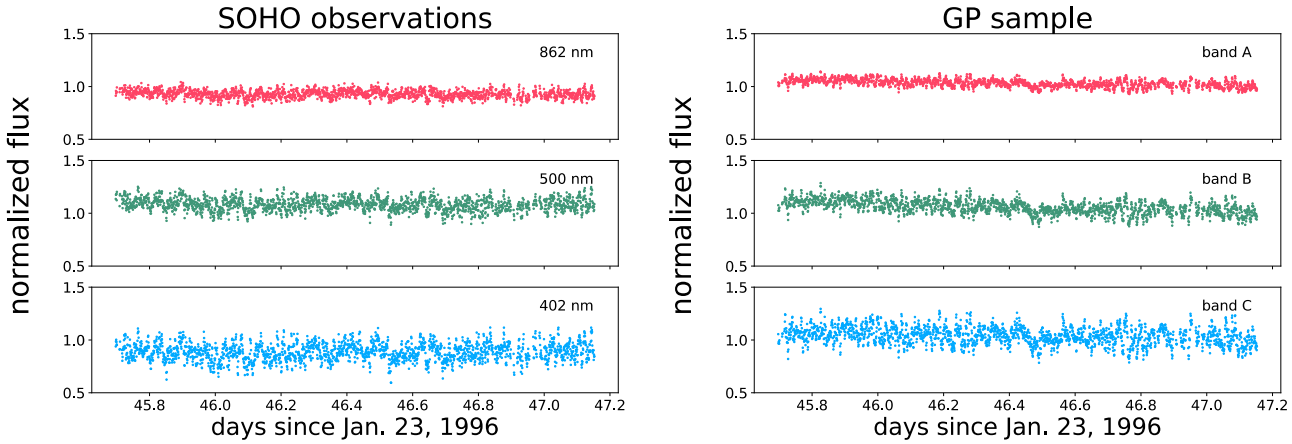

Figure 2. Left: SOHO three-channel Sun photometer time series of the Sun. Right: a three-band light curve simulated from a GP with a kernel consisting of three Kronecker-product terms (see Equation (59)), each term having the covariance described by Equation (17). The GP hyperparameters were obtained by optimizing GP likelihood with respect to the data in the left panel.

of the spectrum. The linear scaling of our method with respect to both the time and wavelength dimension makes it feasible to model high spectral resolution time series this way. We include additional discussion on the subject of modeling transmission spectra in Section 5.

To validate this model of multiwavelength stellar variability, we compare with observed solar variability in Figure 2. This figure shows a time series from the SOHO VIRGO three-channel Sun photometer (SPM; Frohlich et al. 1995). The SPM monitors the Sun's variability in three visible light wave bands at one minute cadence, and each of these bands exhibits a power spectrum which has the same shape, but with amplitude which increases from red (862 nm) to blue (402 nm) as shown in Sulis et al. (2020). Alongside the SOHO SPM data we show a GP drawn from our two-dimensional celerite algorithm in which the amplitudes in each band have been scaled to match the SOHO SPM multiband data. The qualitative agreement between the observed and simulated data is remarkable, and indicates that our model contains the necessary properties to capture high-precision multiwavelength stellar variability.

The algorithm used to simulate multiwavelength stellar variability and to compute the likelihood model is described in Section 3.2. Our implementation of the multiband GP, which is based on the celerite GP method, achieves $\mathcal{O}(NMJ^2)$ scaling where N is the size of T corresponding to the length of the vector $\boldsymbol{x}$, and M is the number of bands and corresponds to the size of the vector $\boldsymbol{a}$. Appendix A introduces a more general form of the two-dimensional GP which scales as $\mathcal{O}(NJ^2M^3)$ for arbitrary covariance in the second dimension. The remaining component of the likelihood function is the mean model, which for this paper we take to be a transit model described next.

Similarly, Loper et al. (2020) recently derived a multivariate generalization of celerite with linear scaling for a class of covariance functions called latent exponentially generated (LEG) kernels. These LEG kernel functions are presented for multivariate outputs instead of multivariate inputs as described here, but it should be possible to express the kernels described here as members of the LEG family. However, for our restricted application, the computational cost and scaling of our method is better, since LEG GPs will scale as $\mathcal{O}(NJ^2M^3)$ in the notation above.

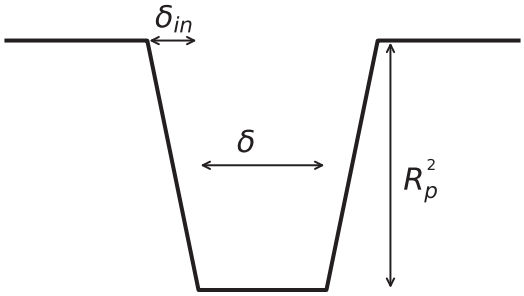

Figure 3. Schematic of the trapezoidal transit model. The center of transit t_0 is the midpoint of the transit.

2.3. Transit Model

To simplify and sharpen our simulated light curves, we use a trapezoidal transit model (Carter et al. 2008); this is the mean model whose parameters we wish to infer. For all our simulations the out-of-transit flux is normalized to unity in order to reduce the number of parameters to be inferred, though we note that this would represent an additional free parameter when modeling real observations. A schematic of this transit model is shown in Figure 3. For the purposes of this paper, we ignore limb-darkening (which can have a wavelength dependence), and we ignore the slight curvature which occurs during ingress and egress. We also assume that the radius of the transiting planet is constant with respect to wavelength. This requirement can be relaxed to accommodate transmission spectroscopy, which we discuss in Section 5.

The model is described by the function $m_{\rm trap}(t, \boldsymbol{q})$ with $\boldsymbol{q} = (R_p, t_0, d, d_{\rm in})$ where R_p is the planet's radius in units of the star's radius, t_0 is the time at center of transit, δ is the transit duration, and $d_{\rm in}$ is the ingress/egress duration. Note that we set the normalization of this model to one under the assumption that the out-of-transit data will be sufficiently lengthy to constrain the unocculted stellar flux.

With the noise and mean models specified, we next describe our simulated data.

2.4. Simulations

We simulate a suite of multiband light curves and construct a parallel set of monochromatic light curves by summing the flux

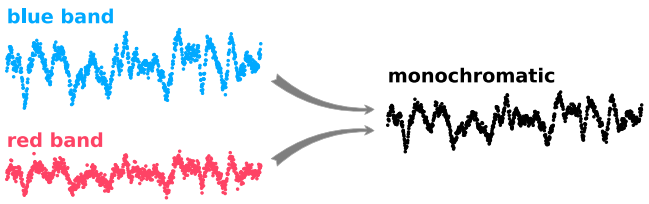

Figure 4. Two bands from a multiband simulation combined to simulate a monochromatic light curve with the same noise realization. Note that the white noise amplitude is smaller in the monochromatic light curve than for either individual band, while the amplitude of the correlated noise is the photon-weighted mean of the amplitude in the two bands. Here the blue band has a correlated noise amplitude twice that of the red band.

between the bands of our multiband light curves. Figure 4 shows schematically how we produce a monochromatic light curve from the simulated multiband light curve. We compute the Information matrix (see Section 2.5) and run MCMC analysis on each light curve using our multiband GP model. The Information matrix tells us the theoretical lower limit for the uncertainty of each parameter, while the MCMC analysis gives us an estimate of the uncertainty on the parameters.

We split our simulations into three noise regimes based on the ratio between the characteristic variability timescale and the ingress/egress and total duration of the transit. The characteristic variability timescale is given by $2\pi\omega_0^{-1}$ where ω_0 is the characteristic frequency of the variability appearing in Equation (4). We define the three regimes as follows:

1. regime I: $1/f_0 > d$
2. regime II: $d_{\rm in} < 1/f_0 < d$
3. regime III: $1/f_0 < d_{\rm in}$

where $f_0 = \omega_0/(2\pi)$ is the characteristic frequency of the variability. Figure 5 contains representative light curves from each regime, chosen where the white and correlated noise amplitudes are comparable. In regime I the transit signal is distinguishable from the noise by its duration—all of the power in the correlated variability is on longer timescales than the transit duration. In regime II the characteristic timescale of the noise is smaller than the transit duration, but longer than the ingress/egress timescale. The transit still stands out from the noise because the transition into and out of transit is sharper than is characteristic for the simple harmonic oscillator (SHO) variability. In regime III the variability timescale is shorter than all of the relevant transit durations. We can see from Figures 1 and 6 that the SHO power spectrum allocates equal power to all oscillations on timescales longer than the characteristic timescale. The transit durations are thus swamped by correlated noise. As a result in the monochromatic case it is difficult to differentiate between the transit signal and noise, both by eye and with the GP. Fortunately the multiband GP is able to make use of additional information in the correlation between bands to disentangle the transit signal from the variability.

Among all of our simulations we hold constant the total noise, $\bar{a}^2 + s^2$ where $\bar{a}^2$ is the weighted variance of the correlated noise over all bands and s^2 is the variance of the white noise summed over all bands. We then vary the ratio between the noise amplitudes in order to analyze the simulations as a function of $\bar{a}/s$. For the multiband simulations, $\bar{a}$ is the weighted mean of the amplitudes of variability in the individual bands, given by a_i. For all of our simulations, unless otherwise specified, we use a two-band model with $a_2 = 2a_1$ to represent the multiband case.

We hold the transit duration and ingress/egress duration constant so that the value of ω_0 changes to determine which noise regime we fall under.

Into all of our simulations we inject a transit signal with a fractional depth of 1% of the star's flux. We use a transit duration of 12 hr in the middle of a 10 day baseline. The ingress/egress duration is set to 1.2 hr.

2.5. Information Matrix Analysis

The Information matrix encodes the amount of information about a signal that can be determined from observations taken in the presence of noise with a given covariance. For a model made up of a mean function $\boldsymbol{m}_{\boldsymbol{q}}$ with N_θ parameters $q_1, q_2, \ldots q_{N_q}$ obscured by noise drawn from a multivariate Gaussian with covariance K, the Information matrix is the $N_q \times N_q$ matrix with entries given by

$$[\mathcal{I}_{\boldsymbol{q}}]_{ij} = \left(\frac{d\boldsymbol{m}}{dq}\right)^{\rm T} K^{-1} \left(\frac{d\boldsymbol{m}}{dq}\right). \quad (21)$$

The covariance between parameters of the mean are then approximated by

$$[\mathcal{I}_{\boldsymbol{q}}^{-1}]_{ij} \gtrsim \mathrm{cov}(q, q). \quad (22)$$

This approximation represents a lower limit on the covariance that can be estimated in practice via methods such as MCMC simulation. It is valid in the limit that the posterior probability is a multidimensional Gaussian distribution near the maximum likelihood solution. This corresponds to the limit in which a signal may be approximated as linear with respect to its parameters, known as the linear signal approximation (LSA). Vallisneri (2008) shows that in order for the LSA to apply we must be in the high signal-to-noise ratio (S/N) limit. Accordingly, the following analysis should be taken to apply only to a transit with a depth much larger than both the correlated and white noise components of the noise. While the approximation may continue to be accurate for smaller signal-to-noise, a full quantification of the uncertainty in the low-S/N limit should rely on sampling the posterior directly via MCMC analysis.

We compute the Information matrix for the transit parameters assuming that the hyperparameters of the GP are known exactly. In practice the GP hyperparameters will be unknown, and should be fit simultaneously with the transit parameters. Our results thus represent a scenario in which there are sufficient out-of-transit observations to determine the covariance of the noise to arbitrary precision.

We adopt a semi-analytic approach to computing the Information matrix by using exact derivatives of the trapezoidal transit model and using celerite to compute products of the inverted covariance matrix with the transit model's derivatives. This approach is necessary because the covariance matrix for our GP model cannot be inverted analytically except in special cases.

2.6. Analytical Estimates for Parameter Uncertainties

The Information matrix approach can yield analytic results for the depth uncertainty in the limit that limb-darkening is ignored, the ingress/egress duration is short $d_{\rm in} \approx 0$, and all other parameters are assumed to have no uncertainty. In

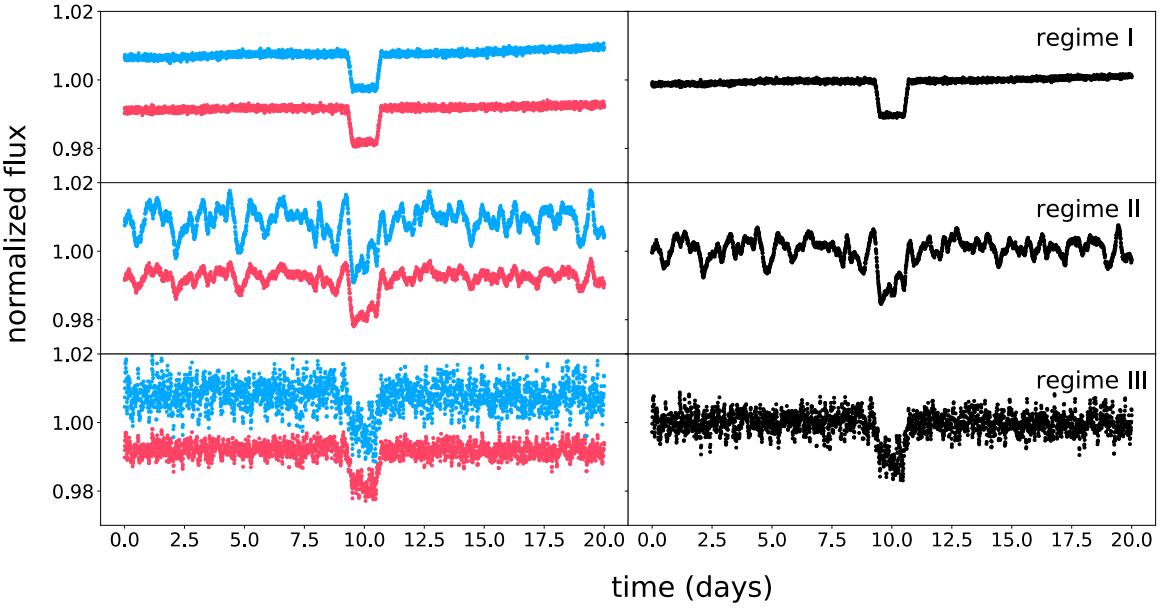

Figure 5. Representative light curves for the three noise regimes. The left panels show the two bands separately and the right panels show the monochromatic light curve resulting from the summation of the two bands. Top: in regime I the variability timescale is much longer than the transit duration. Middle: in regime II the variability timescale is between the transit duration and ingress/egress duration. Bottom: in regime II the variability timescale is shorter than the ingress/egress duration. Figure 6 shows power spectra corresponding to each of these regimes (but not to the light curves pictured here).

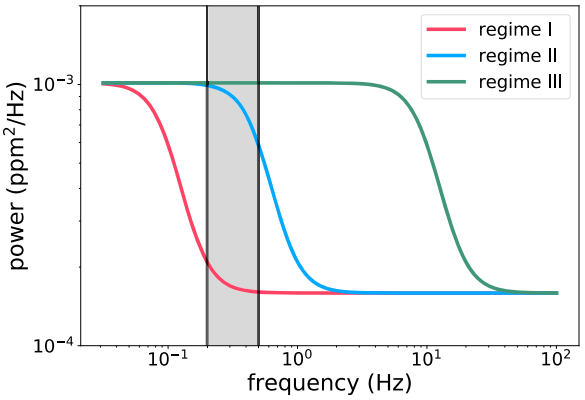

Figure 6. Power spectral densities for the three regimes. The shaded region spans from the inverse transit duration on the left to the inverse ingress/egress duration on the right. Note that the densities plotted here are only meant to be illustrative, and do not correspond to the power spectra of the light curves in Figure 5.

particular we make the approximation that the out-of-transit flux is measured to high precision from extensive monitoring. In this limit the transit model has a derivative of

$$\frac{\partial m_{\rm trap}}{\partial R_p^2} = \begin{cases} 0 & \text{out-of-transit} \\ -1 & \text{in-transit} \end{cases}, \tag{23}$$

where R_p^2 is the depth of the transit. If we assume that the transit duration matches exactly a single observation cadence then the covariance matrix may be written in the two-band case as

$$K = \begin{pmatrix} s_1^2 + a_1^2 & a_1 a_2 \\ a_1 a_2 & s_2^2 + a_2^2 \end{pmatrix}, \tag{24}$$

where $s_{1,2}$ are the white noise components on the timescale of the transit and $a_{1,2}$ are the correlated noise amplitudes on the timescale of the transit in the two bands.

For this covariance matrix, the Information matrix gives an uncertainty on the depth of the transit, $s^2_{R_p^2}$, of

$$s^2_{R_p^2,\rm poly} = \left(\frac{1}{s_1^2} + \frac{1}{s_2^2}\right)^{-1} \left(1 + \frac{\left(\frac{a_1}{s_1}\right)^2 + \left(\frac{a_2}{s_2}\right)^2}{1 + \frac{(a_1 - a_2)^2}{s_1^2 + s_2^2}}\right). \tag{25}$$

Note that the prefactor equals the noise in the limit of no correlated noise components ($a_1 = a_2 = 0$).

In the monochromatic case we can compute the uncertainty assuming that the noise is Poisson, in which case the mean amplitude of correlated noise is given by

$$s^2_{R_p^2,\rm mono} = \left(\frac{1}{s_1^2} + \frac{1}{s_2^2}\right)^{-1} + \bar{a}^2, \tag{26}$$

where we have assumed the noise to be Poisson and $\bar{a}$ is defined to be the weighted mean amplitude of the correlated noise in both bands, given by

$$\bar{a} = \left(\frac{1}{s_1^2} + \frac{1}{s_2^2}\right)^{-1} \left(\frac{a_1}{s_1^2} + \frac{a_2}{s_2^2}\right). \tag{27}$$

The relations for the polychromatic and monochromatic cases are plotted in Figure 7 in the case $a_2 = 2a_1$ and $s_1 = s_2 = s$ in which the sum of the white noise and correlated noise is held fixed. Compare with Figure 12 to see the similarity of this analytic approximation with the Information matrix results for the full trapezoidal model.

We can generalize these expressions for the depth uncertainty in the monochromatic and two-band case to an arbitrary number of bands when the white noise is identical for each band (i.e., $s_i = s$ for M bands indexed by i). In this case the

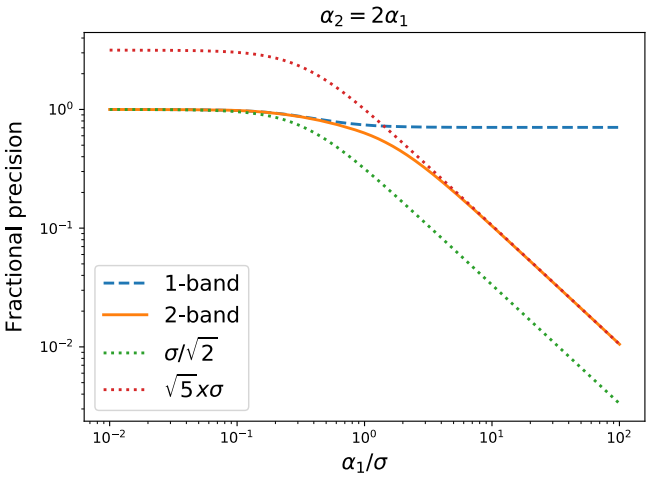

Figure 7. Analytic approximation for the fractional uncertainty on depth for two bands vs. the ratio of the correlated noise to white noise in first band, in the limit of a constant amplitude of the sum of correlated and white noise (so that the white noise declines as the correlated noise increases). The ratio of the correlated noise in the two bands is two, i.e., $a_2 = 2a_1$. Plotted are the single-band case (blue dashed), two-band case (orange solid), and the white noise in each band, σ, times $1/\sqrt{2}$ and $\sqrt{5}$ (dotted). The fractional precision is normalized to the case $a_1 = 0$.

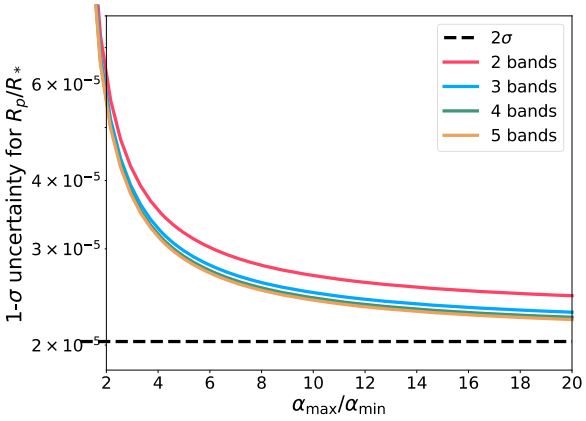

Figure 8. Information uncertainty curves for the planet–star radius ratio as a function of contrast ratio for a spectrum that increases linearly with photon flux from a_{min} to a_{max}. We plot the Information uncertainty for different values of M, the number of bands into which the spectrum is binned for modeling. The dashed line is the minimum uncertainty achievable as the contrast ratio becomes infinite which for the two-band case is equal to 2σ where σ is the information uncertainty in the absence of correlated noise.

uncertainties are given by

$$\frac{s^2_{R_p^2 M,\mathrm{poly}}}{s^2} = \frac{1 + s^{-2}\sum_{i=1}^{M} a_i^2}{M\left(1 + s^{-2}\sum_{i=1}^{M} a_i^2\right) - \left(s^{-1}\sum_{i=1}^{M} a_i\right)^2}, \quad (28)$$

for the M-band case, and

$$\frac{s^2_{R_p^2 M,\mathrm{mono}}}{s^2} = \frac{1}{M} + \left(\frac{1}{M s}\sum_{i=1}^{M} a_i\right)^2. \quad (29)$$

for the corresponding monochromatic case. Similar expressions may likely be found for the other transit parameters as well as for non-uniform noise in M bands, which we leave to future work.

While the uncertainties predicted by these equations differ from those found by a full Information matrix analysis of the trapezoidal transit, we find that they correctly predict the relationship between the monochromatic and multiband uncertainties in the limits $a \ll s$ and $a \gg s$ not only for the depth, but for the other parameters of the trapezoidal transit as well.

This is illustrated by Figure 12 which shows the Information uncertainties for each parameter of the trapezoidal transit model in the presence of correlated noise. We use a two-band noise model with $a_2 = 2 * a_1$. When the white noise dominates over the correlated noise ($s \gg a_{1,2}$), the Information uncertainties for the model with correlated noise are identical to those for a white noise-only model with the same white noise component, as we expect given that the correlated noise components is insignificant in this limit. We can use Equation (28) to predict the Information matrix for the two-band model in the limit that the correlated noise component dominates over the white noise component ($s \ll a_{1,2}$). Taking this limit Equation (28) becomes

$$\lim_{s \ll a_{1,2}} \frac{s^2_{R_p^2 M,\mathrm{poly}}}{s^2} = \frac{\sum_{i=1}^{M} a_i^2}{M\sum_{i=1}^{M} a_i^2 - \left(\sum_{i=1}^{M} a_i\right)^2}. \quad (30)$$

Setting $a_1 = 1$ and $a_2 = 2$, we find $s_{R_p^2 M,\mathrm{poly}}/s = \sqrt{10}$ which explains the scaling of the Information uncertainty at large α/σ in Figure 12.

We also examine the Information uncertainties as a function of number of bands. We consider a photon spectrum for which the variability increases from a value of a_{min} to a_{max}. We assume that the photon spectrum variability is split into M bands with an equal photon count rate in each band to give equivalent Poisson noise across all bands. In addition, we assume that α varies linearly with the photon count rate across all bands, so that the ith band has a correlated noise amplitude of $a_i = a_{min} + (a_{max} - a_{min})(i - 1/2)/M$. For example, in the case of two bands with $a_{max}/a_{min} = 5$, we have $a_2 = 2a_1$, as in Figure 7. Figure 8 shows the uncertainty for the planet–star radius ratio as a function of the ratio between the minimum and maximum variability a_{max}/a_{min}, for several values of M. The minimum achievable uncertainty as M approaches infinity and $a_{max} \gg a_{min}$, which can be arrived at by taking the appropriate limits of Equation (28) and transforming the sums into integrals as M approaches infinity. In these limits the minimum achievable uncertainty is twice that for the white noise-only case, which is represented by the dashed line in Figure 8.

The same calculation may be performed for alternative spectra. For a blackbody spectrum we arrive at a limit of 2.2 times the white noise-only case when the number of bands and the contrast ratio is large. For arbitrary spectra the integrals can be computed numerically to yield the minimum achievable uncertainties for realistic stellar spectra and spot models.

The Information matrix and analytic approaches describe approximations to the parameter uncertainties. We next summarize our MCMC analysis to check and validate these approximations.

2.7. MCMC Analysis

We use the exoplanet package (Foreman-Mackey et al. 2019) which interfaces with PyMC3 to conduct our MCMC simulations. Each simulation is initialized with the true parameters. During MCMC we hold the GP hyperparameters constant as we did for the Information matrix analysis, and vary only the parameters of the trapezoidal transit model. We use PyMC3's implementation of No U-Turn sampling (NUTS; Hoffman & Gelman 2014), which requires the derivatives of the log likelihood to carry out the Hamiltonian markov chain integration. The NUTS sampler is initialized by tuning each simulation for 2000 steps. Subsequently, the simulation is run another 2000 steps to sample the posterior. This procedure results in about 1300 effective samples for each parameter of the model for each simulation as the autocorrelation length of the chains is extremely short (one of the advantages of using the NUTS sampler).

The final ingredient needed for our Information matrix and MCMC simulations involves our novel two-dimensional version of celerite, which we describe next.

3. Implementation of the Multiwavelength Variability Model

We implement our multiwavelength variability model as an extension of the celerite GP method to two dimensions. The celerite algorithm (Foreman-Mackey et al. 2017) is a method for computing GPs in one dimension that scales as $\mathcal{O}(NJ^2)$ where N is the number of data points being modeled and J is the number of terms used to represent the covariance matrix. While one-dimensional GPs are suitable for a wide range of applications, there are many problems for which we need to model covariance between data points in two or more dimensions. Here we describe a method for computing a two-dimensional GP when the covariance in the second dimension can be written as the outer product of a vector with itself. This covariance matrix is relevant to the common task of modeling time-variable spectra, as in our multiband transit model application. Our method is scalable, with computational time increasing linearly with the number of data points. In this section we introduce the method and revisit the celerite algorithm for Cholesky decomposition of the covariance matrix as it applies to a two-dimensional data set. In Appendices B, C, and D we discuss the algorithms for computing the likelihood, predicting or extrapolating from the GP, and sampling from the GP.

For problems where the covariance cannot be modeled as an outer product we offer a more general extension of celerite where the covariance matrix for the second dimension can be arbitrary. We discuss our implementation of the arbitrary covariance method in Appendix A.

3.1. The One-dimensional celerite Method

Until the past few decades, the adoption of GP methods was limited by computational expense. As a reminder, the log-likelihood function for a GP model for a series of N flux measurements $\boldsymbol{y} = (y_1, y_2, \ldots y_N)$, so that the total number of data points $N_{\rm dp} = N$, taken at times $\boldsymbol{t} = (t_1, t_2, \ldots t_N)$ is given by

$$\ln \mathcal{L} = -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{m})^{\mathrm{T}} K^{-1} (\boldsymbol{y} - \boldsymbol{m}) - \frac{1}{2}\ln\det(K) - \frac{N}{2}\ln(2\pi) \quad (31)$$

where $\boldsymbol{m} = (m(t_1), m(t_2), \ldots m(t_N))$ and K is the covariance matrix of the GP. This equation involves the inverse and determinant of the N × N matrix K. In general, computing the inverse and determinant of an N × N matrix requires $\mathcal{O}(N^3)$ operations. Thus computing the likelihood for a GP by directly inverting K becomes prohibitively expensive for data sets larger than about 10^4 observations (Deisentroth & Ng 2015). This is especially true for applications that require repeated calls to the likelihood function as is the case for minimization and MCMC.

Because of this, much work has been done to reduce the complexity of GP computations. This can be accomplished primarily in two nonexclusive ways. The first is by employing inexact methods in which the full GP covariance matrix is approximated by a matrix for which the relevant matrix operations (primarily inversion and computation of the determinant) can be computed more efficiently than $\mathcal{O}(N^3)$ (Rasmussen & Williams 2006). Members of this class of methods include the HODLR factorization method of Ambikasaran et al. (2015) which achieves $\mathcal{O}(n \log^2 n)$ scaling as well as various sparse GP methods (Csató & Opper 2002; Snelson & Ghahramani 2006; Almosallam et al. 2016).

The second is by restricting the user to covariance matrices of a specific form. These methods are often known as structure-exploiting methods since they take advantage of the properties of specially structure matrices (e.g., low-rank matrices, Toeplitz matrices, Kronecker-product matrices) to speed up Gaussian process operations (Zhang et al. 2005; Nickson et al. 2015; Wilson & Nickisch 2015).

The celerite algorithm is a fast, one-dimensional GP method which exploits the properties of semiseparable plus diagonal matrices to accelerate GP computations, achieving $\mathcal{O}(NJ^2)$ scaling where J is the number of celerite terms that make up the kernel function and N is the number of data points. For commonly used kernel models the number of terms will be very small compared to N.

celerite works by representing the GP covariance matrix as the sum of a diagonal matrix and J semi-separable matrices. The Cholesky factorization of J semi-separable matrices plus a diagonal matrix can be computed in $\mathcal{O}(NJ^2)$ rather than the $\mathcal{O}(N^3/3)$ required for an ordinary matrix. Once the Cholesky factors are in hand, the inverse and determinant of the covariance matrix can be computed in $\mathcal{O}(NJ)$ and $\mathcal{O}(N)$ respectively. Here we briefly describe the celerite algorithm, referring the reader to Foreman-Mackey et al. (2017) for a more detailed exposition of the method.

Consider a one-dimensional Gaussian process evaluated at the coordinates

$$\boldsymbol{x} = (t_1 \ldots t_N) \quad (32)$$

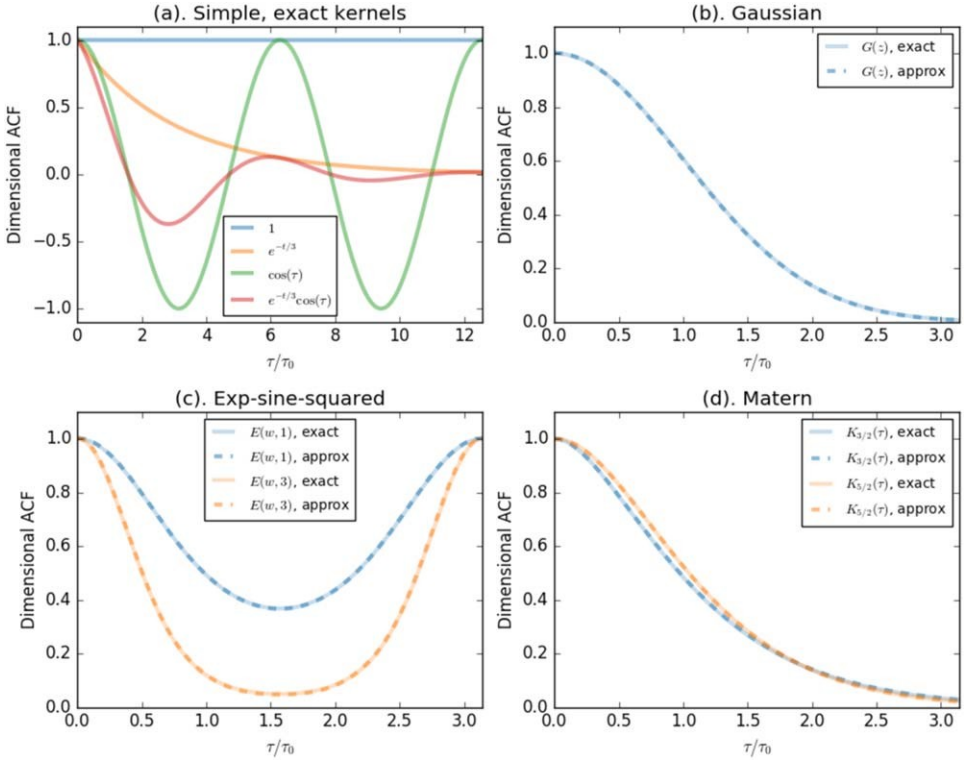

Figure 9. Approximation to various commonly used GP kernels. (a) Simple kernels with an exact celerite representation: cosine, exponential times cosine. (b) Approximation of (a) referred to as "exponential-squared" to distinguish it from sine-squared kernels discussed in (c). (d) Matern kernels.

The celerite kernel is given by

$$k_{\boldsymbol{b}}(t_n, t_m) = s_n^2 \delta_{nm} + \sum_{j=1}^{J} \frac{1}{2}\left[(a_j + i b_j) e^{-(c_j - i d_j) t_{nm}} + (a_j - i b_j) e^{-(c_j + i d_j) t_{nm}}\right] \quad (33)$$

where $\boldsymbol{b} = (a_1 \ldots a_J, b_1 \ldots b_J, c_1 \ldots c_J, d_1 \ldots d_J)$, s_n^2 is the variance of the Gaussian-distributed white noise, and $t_{nm} = |t_n - t_m|$ with $n, m \in 1, \ldots, N$. This kernel defines a celerite model with J terms.

For a kernel function of this form, the covariance matrix is a symmetric, semiseparable matrix with semiseparability rank $P = 2J$. A matrix of this type can be written in terms of two generator matrices U and V, both of size $(N \times P)$, along with a diagonal matrix A:

$$K = A + \mathrm{tril}(UV^{\mathrm{T}}) + \mathrm{triu}(VU^{\mathrm{T}}), \quad (34)$$

where tril is the lower-triangular operator which, when applied to a square matrix, preserves the entries below the diagonal and replaces all entries on and above the diagonal with zeros. The triu operator does the same for the upper-triangular entries in the matrix. In the case of our covariance matrix, the generator matrices are specified by

$$\begin{aligned} U_{n,2j-1} &= a_j e^{-c_j t_n} \cos(d_j t_n) + b_j e^{-c_j t_n} \sin(d_j t_n), \\ U_{n,2j} &= a_j e^{-c_j t_n} \sin(d_j t_n) - b_j e^{-c_j t_n} \cos(d_j t_n), \\ V_{m,2j-1} &= e^{c_j t_m} \cos(d_j t_m), \\ V_{m,2j} &= e^{c_j t_m} \sin(d_j t_m), \end{aligned} \quad (35)$$

and A is given by

$$A_{n,n} = s_n^2 + \sum_{j=1}^{J} a_j. \quad (36)$$

We will soon see that the Cholesky decomposition for this covariance matrix can be computed in $\mathcal{O}(NJ^2)$ operations, allowing for the fast evaluation of the GP likelihood function.

The kernel function implemented by celerite is versatile in that by choosing appropriate coefficients it can be made to approximate a wide range of other kernel functions. Furthermore, Loper et al. (2020) demonstrated that celerite kernels provide a complete basis for one-dimensional stationary covariance functions, meaning that these methods can, in principle, be used to approximate any stationary kernel, though there might be issues with numerical precision and computational cost when a large number of terms are required for accuracy. This versatility is demonstrated qualitatively in Figure 9 which shows approximations of several popular kernels achieved by carefully choosing $\{a_j\}$, $\{b_j\}$, $\{c_j\}$, and $\{d_j\}$. Since each of these kernels may be approximated well by a celerite kernel, the products and sums of these component kernels are also celerite kernels, meaning that complex kernels can still be approximated within the celerite kernel formalism.

The Cholesky factorization of the covariance matrix K is given by

$$K = LDL^{\mathrm{T}} \quad (37)$$

where L is the lower-triangular Cholesky factor and D is a diagonal matrix. Foreman-Mackey et al. (2017) begin their derivation of the Cholesky factorization algorithm with the ansatz that L can be represented in terms of U and a new (at this

point unknown) matrix W with the same dimensions as U,

$$L = I + \mathrm{tril}(UW^{\mathrm{T}}). \quad (38)$$

Then W and D can be found via the recursion relations

$$\begin{aligned} S_{n,j,k} &= S_{n-1,j,k} + D_{n-1,n-1} W_{n-1,j} W_{n-1,k} \\ D_{n,n} &= A_{n,n} - \sum_{j=1}^{P} \sum_{k=1}^{P} U_{n,j} S_{n,j,k} U_{n,k} \\ W_{n,j} &= \frac{1}{D_{n,n}} \left[V_{n,j} - \sum_{k=1}^{P} U_{n,k} S_{n,j,k} \right], \end{aligned} \quad (39)$$

where $S_{1,j,k}$ is a matrix of zeros and P is both the rank of the semiseparable covariance matrix and the number of columns in U and V, here equal to $2J$. In the original celerite paper it was found that, in order to avoid numerical stability issues caused by the exponential factors in Equation (35), it was necessary to redefine the generator matrices U and V and to define an additional matrix f of the same dimensions as U and V. The generators become

$$\begin{aligned} \tilde{U}_{n,2j-1} &= a_j \cos(d_j t_n) + b_j \sin(d_j t_n) \\ \tilde{U}_{n,2j} &= a_j \sin(d_j t_n) - b_j \cos(d_j t_n) \\ \tilde{V}_{m,2j-1} &= \cos(d_j t_m) \\ \tilde{V}_{m,2j} &= \sin(d_j t_m). \end{aligned} \quad (40)$$

The unknown matrix W becomes

$$\begin{aligned} \tilde{W}_{n,2j-1} &= e^{-c_j t_n} W_{n,2j-1} \\ \tilde{W}_{n,2j} &= e^{-c_j t_n} W_{n,2j} \end{aligned} \quad (41)$$

And the new matrix f is defined

$$f_{n,2j-1} = f_{n,2j} = e^{-c_j (t_n - t_{n-1})}. \quad (42)$$

The algorithm for decomposing the covariance matrix becomes

$$S_{n,j,k} = f_{n-1} f_{n-1} [S_{n-1,j,k} + D_{n-1,n-1} \tilde{W}_{n-1,j} \tilde{W}_{n-1,k}] \quad (43)$$

$$D_{n,n} = A_{n,n} - \sum_{j=1}^{P} \sum_{k=1}^{P} \tilde{U}_{n,j} S_{n,j,k} \tilde{U}_{n,k} \quad (44)$$

$$\tilde{W}_{n,j} = \frac{1}{D_{n,n}} \left[\tilde{V}_{n,j} - \sum_{k=1}^{P} \tilde{U}_{n,k} S_{n,j,k} \right]. \quad (45)$$

This completes our recap of the one-dimensional version of celerite; next we describe our novel two-dimensional version.

3.2. Computing the Two-dimensional GP

We now consider the Cholesky decomposition of the covariance matrix for a two-dimensional GP when the covariance in the second dimension can be written as the outer product of a vector with itself. This form of the covariance applies when the correlated component of the noise has the same shape along the first large dimension (of size N) and varies proportionally in amplitude along the second small dimension (of size M), as is the case for the multiwavelength stellar variability problem discussed above.

This covariance matrix is given by Equation (18), reproduced here:

$$K = S + T \otimes R, \quad (46)$$

which has size $N' \times N' = NM \times NM$. Here Σ is a diagonal matrix containing the white noise components for each data point, which may be heteroscedastic, T is the covariance matrix in the first dimension, which must be defined by a celerite kernel, and R is the covariance matrix for the second dimension which must be an outer product of the form

$$R = \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}, \quad (47)$$

where $\boldsymbol{a} = (a_1, a_2, ..., a_M)^{\mathrm{T}}$ is a vector of length M.

Writing K in terms of the celerite generator matrices from Equation (34):

$$\begin{aligned} K &= S + [A_0 + \mathrm{tril}(UV^{\mathrm{T}}) + \mathrm{triu}(VU^{\mathrm{T}})] \otimes R \\ &= S + \mathrm{diag}(A_0 \otimes R) \\ &\quad + \mathrm{tril}(UV^{\mathrm{T}} \otimes R) + \mathrm{triu}(VU^{\mathrm{T}} \otimes R), \end{aligned} \quad (48)$$

where A_0 is the diagonal component of T obtained by setting $s_n = 0$ for all $n \in 1, \ldots, N$ in Equation (36). Substituting the outer product $\boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}$ for R inside the upper and lower triangular operators we have

$$\begin{aligned} K = S &+ \mathrm{diag}(A_0 \otimes R) \\ &+ \mathrm{tril}(UV^{\mathrm{T}} \otimes \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}) \\ &+ \mathrm{triu}(VU^{\mathrm{T}} \otimes \boldsymbol{a}\boldsymbol{a}^{\mathrm{T}}). \end{aligned} \quad (49)$$

Applying the formula for mixed Kronecker and matrix products,

$$(AB) \otimes (CD) = (A \otimes C)(B \otimes D), \quad (50)$$

we can rewrite the covariance matrix as

$$\begin{aligned} K = S &+ \mathrm{diag}(A_0 \otimes R) \\ &+ \mathrm{tril}((U \otimes \boldsymbol{a})(V \otimes \boldsymbol{a})^{\mathrm{T}}) \\ &+ \mathrm{triu}((V \otimes \boldsymbol{a})(U \otimes \boldsymbol{a})^{\mathrm{T}}). \end{aligned} \quad (51)$$

We now see that the two-dimensional covariance matrix has exactly the same semi-separable structure as the one-dimensional covariance matrix with new definitions of the generator matrices in terms of their Kronecker products with a:

$$\begin{aligned} A' &= S + \mathrm{diag}(A_0 \otimes R) \\ U' &= U \otimes \boldsymbol{a} \\ V' &= V \otimes \boldsymbol{a} \end{aligned} \quad (52)$$

The components of the refactored generator matrices, corresponding to Equation (40), are now given by

$$\begin{aligned} \tilde{U}'_{M(n-1)+p,2j-1} &= a_p (a_j \cos(d_j t_n) + b_j \sin(d_j t_n)) \\ \tilde{U}'_{M(n-1)+p,2j} &= a_p (a_j \sin(d_j t_n) - b_j \cos(d_j t_n)) \\ \tilde{V}'_{M(m-1)+p,2j-1} &= a_p \cos(d_j t_m) \\ \tilde{V}'_{M(m-1)+p,2j} &= a_p \sin(d_j t_m), \end{aligned} \quad (53)$$

and f' is given by

$$f'_{M(n-1)+p,:} = \begin{cases} e^{-c_j (t_n - t_{n-1})} & p = 1 \\ 1 & p > 1, \end{cases} \quad (54)$$

with $n, m \in 1, \ldots, N$, $p \in 1, \ldots, M$, and the colon indicating that the element is identical for every entry of that row. For these definitions of the generator matrices the recursive

Cholesky decomposition algorithm becomes

$$S_{n,j,k} = f_{n,j} f_{n,k} [S_{n-1,j,k} + D_{n-1,n-1} W_{n-1,j} W_{n-1,k}],$$
$$D_{n,n} = A_{n,n} - \sum_{j=1}^{P} \sum_{k=1}^{P} U_{n,j} S_{n,j,k} U_{n,k},$$
$$W_{n,j} = \frac{1}{D_{n,n}} \left[V_{n,j} - \sum_{k=1}^{P} U_{n,k} S_{n,j,k} \right], \quad (55)$$

where again P is the number of columns in $U_{n,}$ and $V_{n,}$.

The recursive algorithm defined above requires one pass through each of the $N_{tot} = NM$ rows of $U_{n,}$ and $V_{n,}$. At each step we compute a double sum over the P columns of these matrices. The resultant scaling is thus $O(NMP^2)$. For the outer-product definition of R we have $P = 2J$ and the method scales as $O(NMJ^2)$ (see Appendix B for benchmarks).

As shown in Appendix A, we can come up with similar definitions of $U_{n,}$, $V_{n,}$, and $f_{n,}$ for arbitrarily defined R which yield $P = 2JM$, allowing us to compute the Cholesky decomposition in $O(NJ^2M^3)$. Algorithms for computing the likelihood function, computing predictions or extrapolations from the GP, and sampling the GP are given in Appendices B, C, and D respectively for both outer-product and arbitrary definitions of R.

For this two-dimensional GP the set of observations used to compute the GP likelihood is also two-dimensional. Actual computation of the likelihood, however, requires that the input be reduced to one dimension. The Kronecker structure of the covariance matrix determines the form of the vector of observations. For input defined on a grid of size $t \times r$ where t represents the dimension along which the covariance is described by a celerite kernel and r represents the second small dimension, we have a two-dimensional matrix of observations:

$$Y_{ij} = y(r_i, t_j). \quad (56)$$

We define the observation vector to be

$$\boldsymbol{y} = \mathrm{vec}(Y), \quad (57)$$

where $\mathrm{vec}(Y)$ is the concatenation of the rows of Y. In other words,

$$\boldsymbol{y} = (Y_{:,1}, Y_{:,2}, \ldots Y_{:,N}). \quad (58)$$

With the description of our computational methods completed, we now turn to the results of transit simulations.

4. Results

We have carried out an analysis of simulated transit light curves with a wide range of noise amplitudes, timescales, and ratios of correlated to white noise, which we summarize the results of here. We start with a discussion of the results from a case study of seven examples with different ratios of correlated to white noise (Section 4.1), and then expand the discussion to a wider range of simulations for which we compare the Information matrix, analytic, and MCMC error analyses (Section 4.2).

4.1. Case Studies

To start with, Figure 10 shows seven examples of our simulations for two bands with correlated noise amplitudes which differ by a ratio of two. These were made with moderate S/N and with $w_0 d = 100$, which corresponds to a characteristic timescale of the correlated noise which is shorter than the transit duration and the ingress/egress timescales (regime III). In this case we held the white noise in the two bands to be identical in amplitude (corresponding to an identical photon count rate in both bands), and we compared a joint analysis of the two bands (we refer to this as "polychromatic") with an analysis of a single band consisting of the sum of the same simulated light curves from the two bands (this analysis we refer to as "monochromatic"). Across these simulations we have varied the ratio of the total correlated noise to the white noise, α/σ, over seven values, {0.02, 0.55, 1, 2, 4, 20, 143}, to examine the precision of the two-band analysis compared with a monochromatic analysis.

For the first two simulations, α = 0.02σ and α = 0.55σ, variance of the correlated noise is smaller than that of the white noise. At this low ratio of α/σ we find that the measurement of the transit depth and timing parameters is about the same in the two-band case as in the monochromatic case (top panel, Figure 10). In the third panel where the white and correlated noise amplitudes are equal, we see a slight improvement in the measurement of the transit time and depth. In the remaining panels (bottom four panels of 10), we find an increasing degree of improvement in all the measured parameters as α increases relative to σ. As we approach the small white noise limit the improvement in all parameters between the single-band and two-band analyses is dramatic, with the transit depth improving by a factor of 18 at α = 20σ and by a factor of 143 at α = 143σ. The transit time measurement improves by a factor of 21 and 65 respectively for these simulations. This improvement results from the ability to distinguish correlated noise variations from the transit signal when two bands are utilized, thanks to the different amplitudes of the correlated noise in the two bands; the correlated noise variations are measured to high precision in this case due to the small photon noise. Even so, the precision of the transit parameters is worse than it would be if there were no correlated noise by a factor of $\sqrt{10}$. This is an astrophysical limitation, and yet it still demonstrates a dramatic improvement in the analysis which splits the the photons into two bands versus a single summed band.

The intermediate values of α/σ = {1, 2, 4} shown in Figure 10 have a behavior which is intermediate between the high white noise and low white noise limits that we discuss above: a monotonic improvement in all of the measurements with the increase in α/σ.

The general trends of these simulations hold over a broader range of parameters. To examine a larger number of cases, we summarize the uncertainties of the monochromatic cases and polychromatic cases based on the measurement precision as a function of the noise parameters, which amounts to measuring the breadth of the posterior distributions inferred for each parameter (left-hand panels of Figure 10). We also compare these to the uncertainty estimates using the Information matrix approach and the analytic estimates given in Sections 2.5 and 2.6, which we discuss next.

4.2. Noise Comparison

We have carried out a much broader parameter study, varying the ratio of α/σ over a wide range of values for three values of the timescale $w_0 d = 0.1$, 10, and 100. We compare

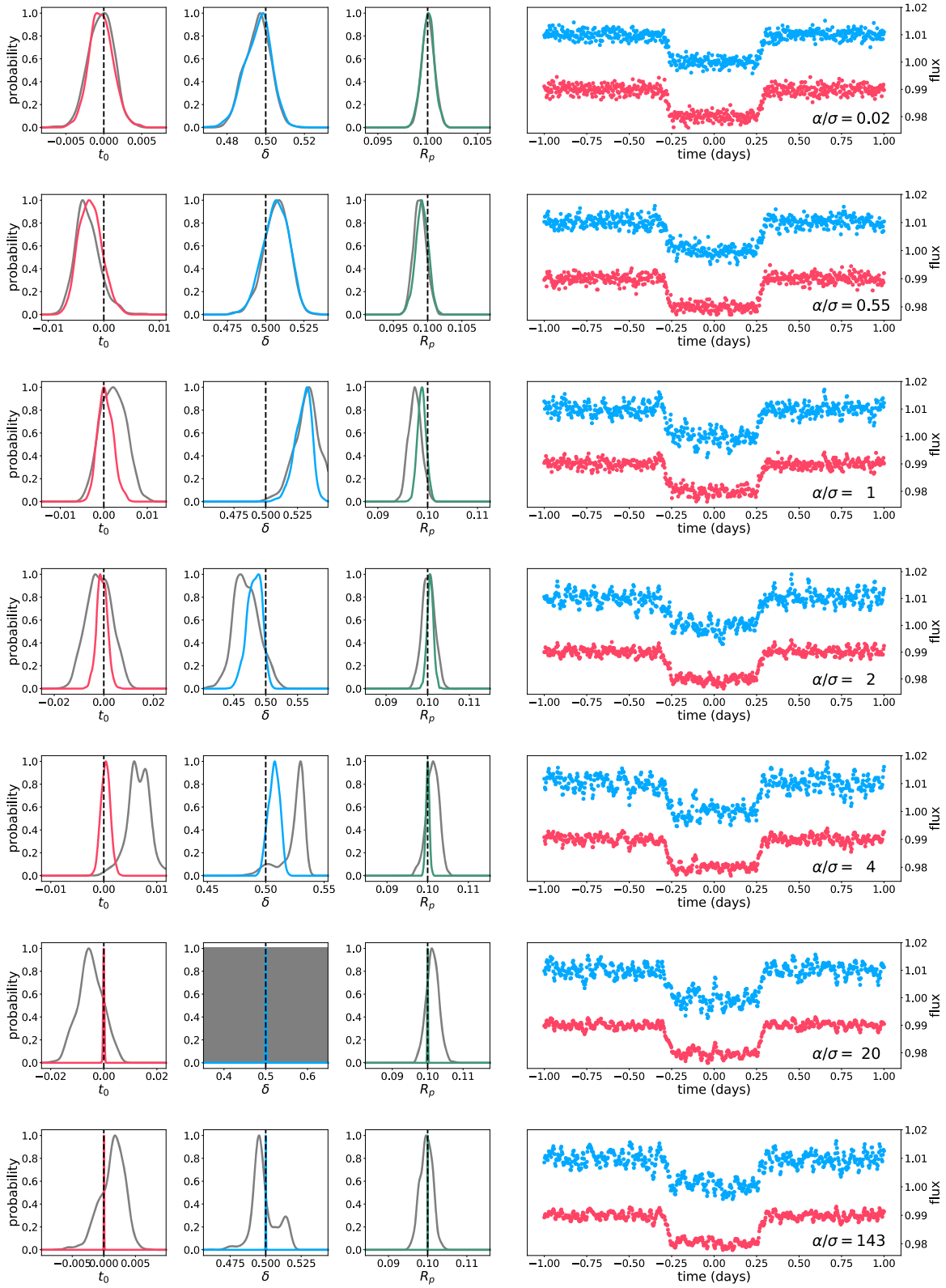

Figure 10. Left: posteriors for three transit parameters estimated by MCMC analysis on the two-band (colored) and single-band (gray) data. Posteriors are smoothed using Gaussian kernel density estimation with $d = 100$ (corresponding to the final panel of Figure 11). From left to right: the center of transit t_0, transit duration δ, and radius ratio R_p/R_*. For $\alpha/\sigma = 20$ and $\alpha/\sigma = 143$ the posterior distributions for the two-band case are too sharply peaked to be visible. Right: representative light curves for each value of the noise amplitude ratio α/σ zoomed in on the transit signal (the input light curves have a duration of 10 days).

the Information matrix analysis against the MCMC anlysis in the monochromatic case with the two-band case also with $a_2 = 2a_1$, in Figure 11. The MCMC uncertainty estimates agree closely with the Information uncertainty curves for almost all of our simulations, as demonstrated by Figure 11 for moderate S/N.

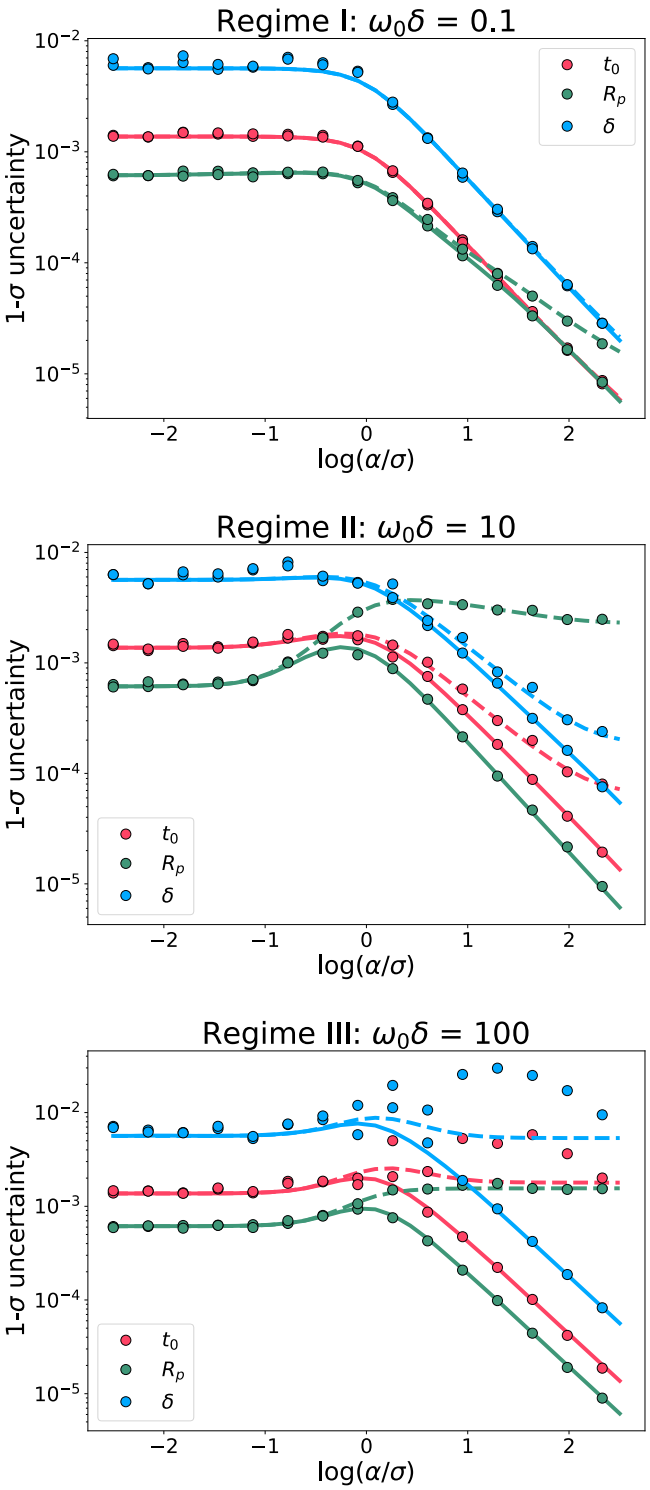

Figure 11. Information uncertainty curves overlaid with MCMC uncertainty estimates for trapezoidal transit parameters. Dashed lines show results for the monochromatic noise model and solid lines show results for the two-band noise model. Circles represent the MCMC uncertainty for distinct realizations of the noise and transit.

In regime I, $w_0\,d = 0.1$, in which the characteristic variability timescale is longer than the transit duration, the uncertainties on the transit parameters are nearly identical between the monochromatic and multiband simulations up to $a/s \approx 10$, where the multiband uncertainties begin to diverge slightly from the monochromatic uncertainties. Since the transit signal is distinguishable from the noise on the basis of its duration alone, the amount of additional information contained in the inter-band correlation is insignificant and both models perform similarly well.

We now skip to regime III, with $w_0\,d = 100$ (the same as the case studies in the previous subsection), in which the characteristic variability timescale is smaller than the transit duration. Because the SHO power spectrum allocates equal power to all oscillations on timescales longer than $1/w_0$, the transit signal is not distinguishable from the variability on the basis of its duration. In this case the inter-band correlation contains the additional information necessary to correctly infer transit parameters. Both models perform similarly when the correlated noise amplitude is small compared to the white noise, but when the correlated noise amplitude α begins to dominate over the white noise σ the monochromatic model does a poor job of inferring parameters (as evidenced by the large uncertainties) while the multiband model infers more and more precise values as the white noise decreases relative to the correlated noise.

The results for regime II, here represented by $w_0\,d = 10$, fall intermediately between regimes I and III. In regime II, the characteristic timescale of the variability falls between the transit duration and the ingress/egress timescale so that measurements of the transit duration must contend with correlated noise on the same timescale, whereas measurements of the ingress and egress are affected primarily by white noise rather than correlated noise. Since the transit time is constrained by the ingress and egress times rather than by the transit duration, measurements of t_0 are also primarily affected by white noise. This is why we see significant improvement in the measurement of the transit depth at high α and low σ between the single-band and two-band simulations, while the timing parameters show much less improvement until we reach the low white noise limit. At this point the white noise amplitude is small enough compared to the correlated noise amplitude that the relatively low correlated noise on the timescale of the ingress/egress duration does begin to interfere with timing measurements in the single-band case.

In Figure 12 we plot Information uncertainty curves in regime III for the multiband model against those for the monochromatic model having the same transit parameters but with only a white noise component—the correlated noise amplitude is set to zero. The colored curves representing the Information uncertainties for the full noise model (white and correlated noise) match the white noise-only uncertainty in the limit that the correlated noise component is very small, as expected. As we increase the relative amplitude of the correlated noise component the uncertainty for the full model jumps from the white noise-only curve with the same white noise amplitude to the white noise-only curve with $\sqrt{10}$ times greater amplitude. As the correlated noise amplitude further increases, the Information uncertainty for the full model behaves as though we are doing inference on an equivalent model with the correlated noise component exchanged for a larger white noise amplitude.

The behavior seen here is explained by the analytical model outlined in Section 2.6. In particular, Equation (30) explains why the uncertainty scales as the white noise-only uncertainty with $\sqrt{10}\,s$ in the large correlated noise limit for two bands with amplitudes related by $a_2 = 2a_1$.

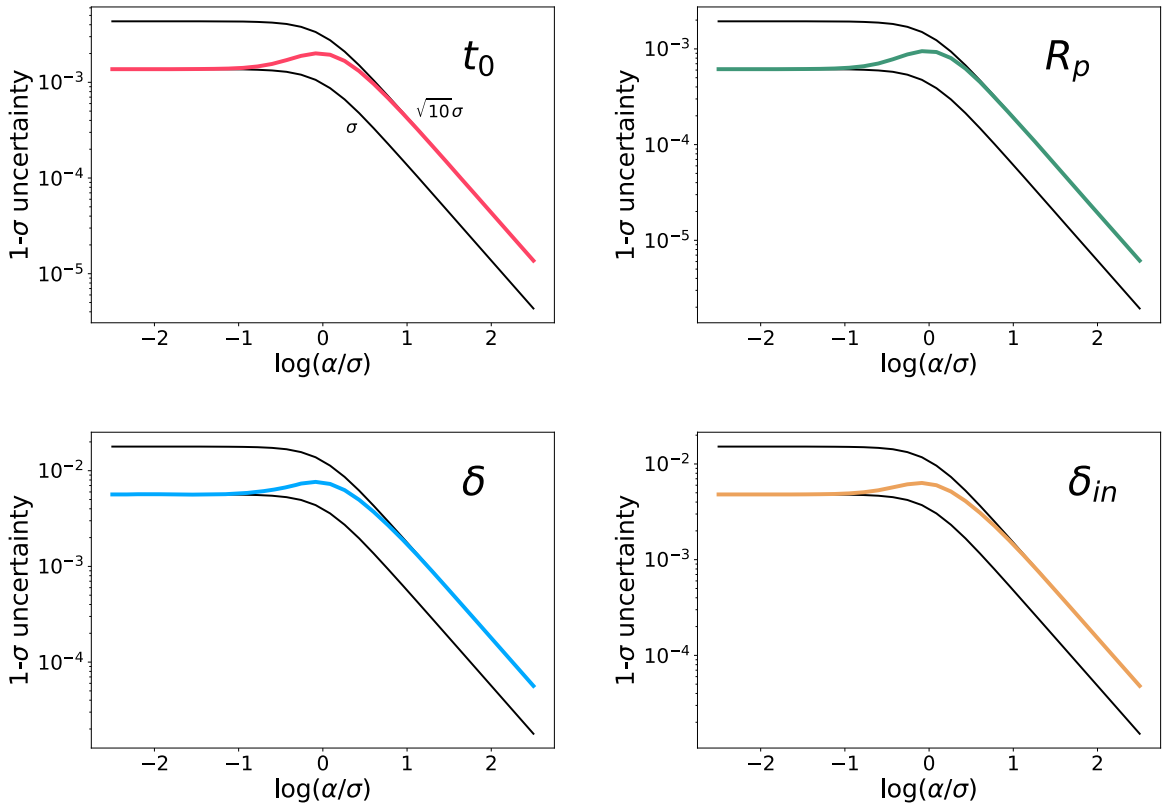

Figure 12. Information uncertainty curves (colored lines) for the two-band model compared to the white noise-only versions of the corresponding monochromatic noise model (black lines) in regime III. For the white noise-only models we set the correlated noise amplitude to zero and leave all other parameters the same as th monochromatic model. As we transition from the white noise-dominated to the correlated noise-dominated regimes the Information uncertainty curves for the two-band model transition from following the white noise model with $s_c = s$ to the white noise model with $s_c = \sqrt{10}s$. In effect perfect knowledge of the two-band correlated noise hyperparameters allows us to recover transit parameters at the same precision as if the correlated noise were simply white noise with a $\sqrt{10}$ larger amplitude.

This completes our description of the simulated light curves and the results from these simulations. We next discuss the implications of these results.

5. Discussion

We have demonstrated the application of our method to the problem of fitting a transit observed in multiple bands in the presence of correlated noise. We now revisit and summarize the results of that demonstration before outlining some other potential applications of our method.

Monochromatic transit observations are ill-equipped to deal with correlated noise, as the wavelength-integrated flux does not provide enough information to distinguish between transits and noise features except when the correlated noise amplitude is low on the timescale of the transit duration. When transits occur on timescales similar to or longer than the variability timescale we must rely on the spectral dimension to provide the information necessary to distinguish between the two.

We use the Information matrix to explore the difference between inference on a monochromatic noise model and a multiband model with wavelength-dependent variability. We construct sets of monochromatic and multiband models with identical noise properties by splitting a given number of photons per wavelength into different spectral bins. We find that our results depend strongly on the timescale of the noise with respect to the transit duration. When the timescale of the correlated variability is much longer than the transit duration the monochromatic and multiband models perform similarly, though the multiband model still allows us to infer slightly more precise parameters in the limit that the correlated noise amplitude is much larger than the white noise amplitude (see Figure 11).

For the noise regime in which the correlated variability timescale is similar to or shorter than the transit duration we summarize our results as follows.

1. As the white noise amplitude decreases and the correlated noise amplitude increases, the precision inferred by the monochromatic noise model stays approximately constant, getting slightly worse for the radius ratio but improving slightly for the timing parameters δ and t_0. In contrast, the precision inferred by the multiband noise model improves as the white noise amplitude decreases even with increasing correlated noise amplitude. The increase in precision scales the same as if the correlated noise were held constant. The presence of correlated noise simply decreases the precision of the parameters by a constant factor which is related to the form of the variability as a function of wavelength.
2. Most of the benefits of the multiband noise model can be realized by splitting the monochromatic variability into just two bands, but more bands achieve slightly better precision (see Figure 8).
3. In the limit that we approach an infinitely high-resolution spectrum we can derive the factor by which the precision of the transit parameters is worse than the case where there is no correlated variability. Using Equation (28) we

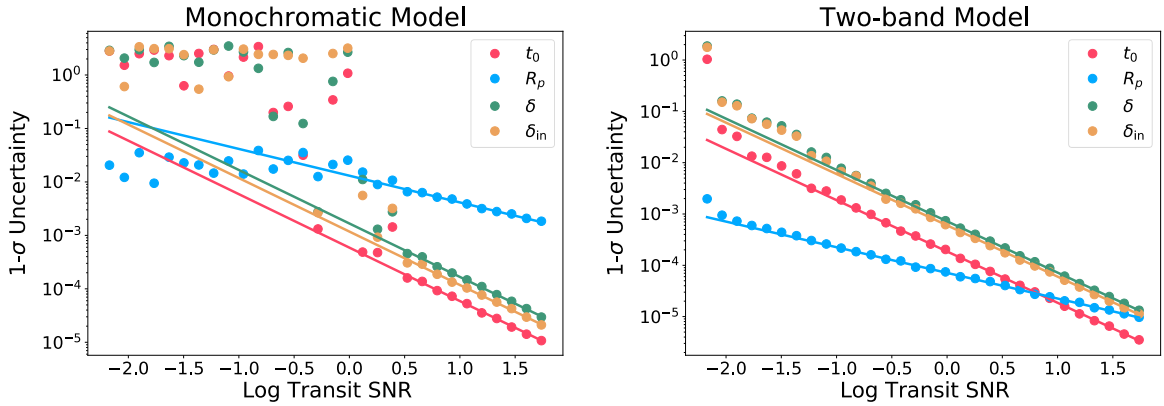

Figure 13. MCMC uncertainties (dots) and Information matrix uncertainties (lines) for monochromatic and two-band noise models as a function of the transit signal-to-noise ratio (S/N) with $a_2 = 2a_1$ for the two-band simulations. For these simulations the correlated noise is held constant at 150 times the amplitude of the white noise component and the total noise, defined to be the sum in quadrature of the white noise and correlated noise amplitudes, is conserved. The variability timescal $1/w_0 = d/10$, placing these simulations in regime II. For the monochromatic model, the Information and MCMC uncertainties correspond down to an S/N of about 10, which is the point at which the MCMC simulations no longer converge to the correct transit solution, as evidenced by the scatter in MCMC uncertainties at lowe S/N. For the two-band simulations the Information and MCMC uncertainties correspond down to an S/N of 1/100.

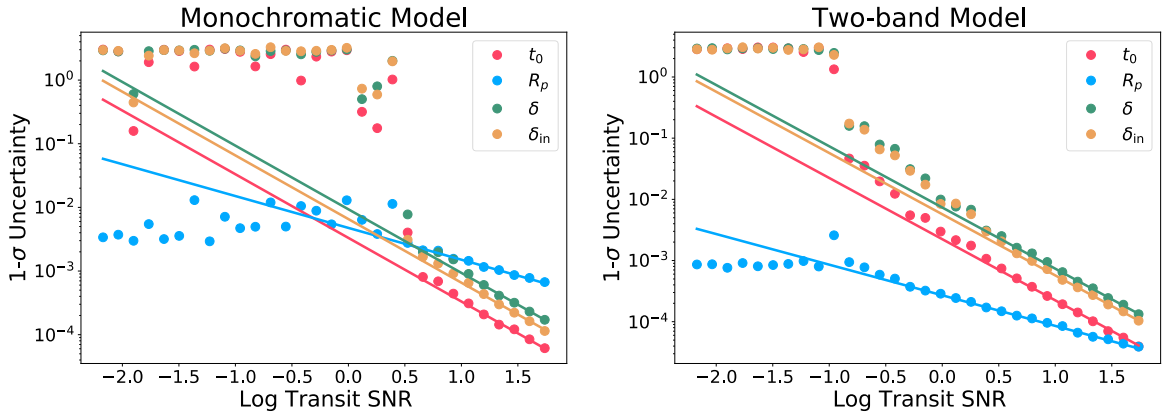

Figure 14. MCMC uncertainties (dots) and Information uncertainties (lines) for monochromatic and two-band noise models as a function of the transit S/N with wavelength dependence specifie $a_2 = 2a_1$ for the two-band simulations. For these simulations the correlated noise is held constant at 10 times the amplitude of the white noise componen and the totalnoise, defined to be the sum in quadrature of the white noise and correlated noise amplitudes, is conserved. The variability timescale $1/w = d/10$, placing these simulations in regime II. The larger white noise component compared to Figure 13 pushes the S/N limit below which the MCMC and Information uncertainties diverge to higher S/N. As before, there is an abrupt transition at this limiting S/N where the MCMC suddenly fails to converge to the correct transit solution.

find that the precision inferred in the presence of correlated noise is worse than in the white noise-only case by a factor of 2 when the variability amplitude scales linearly with cumulative photon counts with wavelength and 2.2 when the variability amplitude is distributed according the the blackbody distribution. In other words, in the presence of linearly scaling correlated variability amplitudes, we need four times as many photons to achieve the same precision in the presence of correlated noise as can be achieved when there is only white noise, provided we use a multiband noise model to do our inference.

5.1. Low Transit S/N Limit

The limit where the transit depth is small compared to the correlated noise amplitude is important if we are interested in detecting planets with small radii, or rocky planets around Sun-like stars. The Information matrix analysis above was done in the high-S/N limit, because that is the limit in which the Information matrix can be shown to approximate the uncertainty on model parameters. We now include results on the correspondence between the Information matrix and MCMC uncertainties in the low-S/N limit. Since we are primarily interested in the correlated noise component, we use S/N to refer to the ratio of the transit depth to the correlated noise amplitude.

Figure 13 shows the MCMC-derived uncertainties and the Information uncertainties for our four trapezoidal transit parameters in both the monochromatic and two-band cases. We use a correlated noise to white noise amplitude ratio of 150 for this portion of the analysis.

When we use a monochromatic model the Information uncertanties diverge from the MCMC uncertainties at an S/N of about 10. This corresponds to the point at which the MCMC uncertainties jump to very high values for the timing parameters, indicating that the MCMC fails to converge to the correct solution.

This contrasts strongly with the two-band model. Using two bands the Information analysis finds the same uncertainty as the MCMC analysis down to an S/N of about 1/100, for which the ratio of the transit depth to the white noise is near unity.

In Figure 14 we repeat the analysis for $\alpha/\sigma = 10$. With a larger white noise component the MCMC uncertainties diverge

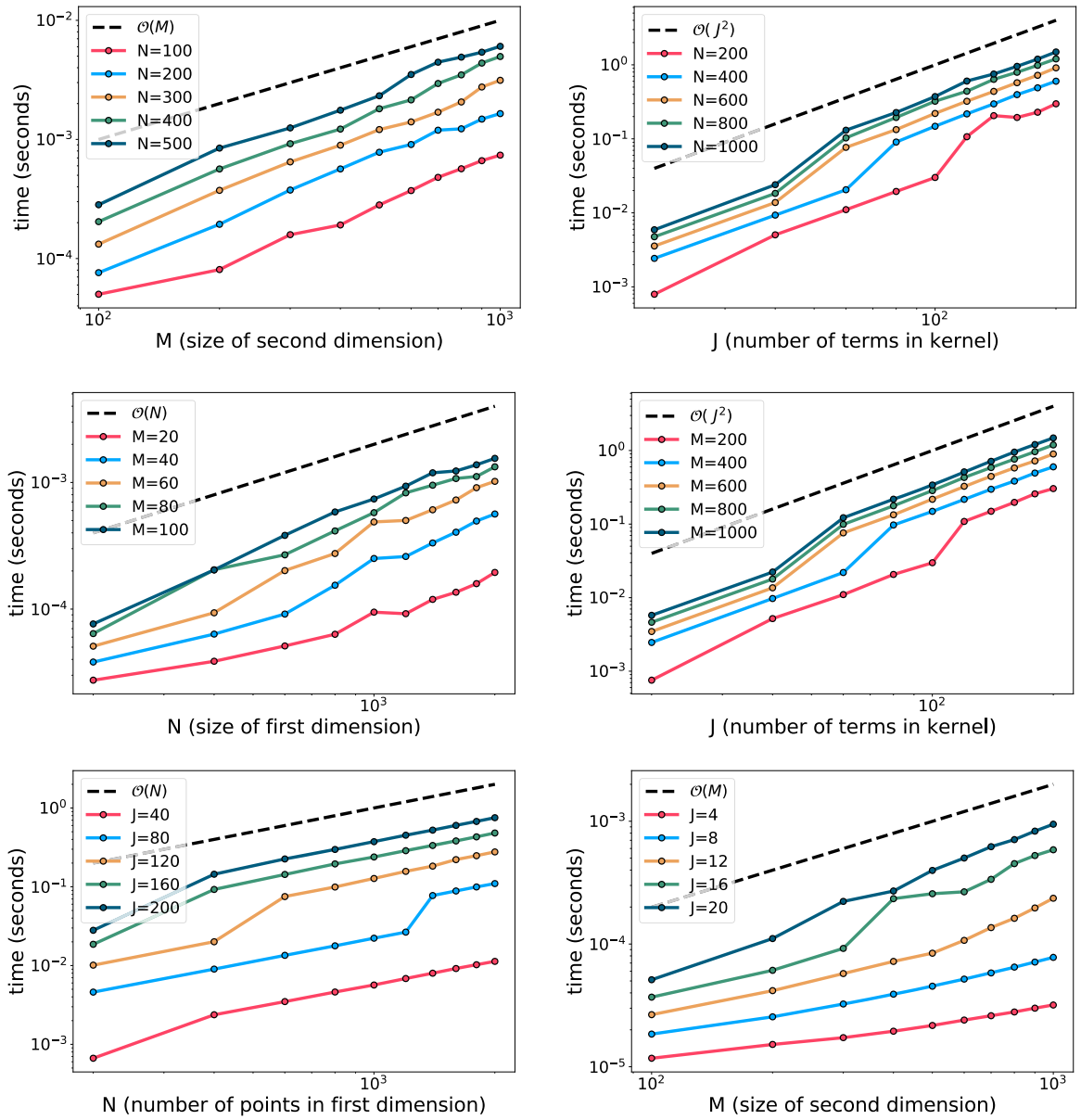

Figure 15. Benchmarks for the two-dimensional celerite implementation with outer-product covariance in the second dimension. We recover the anticipated linear scaling with respect to both N and M, and the quadratic scaling with respect to J.

from the Information uncertainties at a higher S/N. However, the two-band model still outperforms the monochromatic model with the MCMC corresponding to the Information and converging to the correct solution down to an S/N of about 1/10, where again the transit depth is comparable to the white noise.

These results simply that the improvement resulting from multiple bands applies only when the signal is larger than the white noise, and in this limit, the Information matrix provides an adequate estimate of the uncertainties on the model parameters, assuming that the GP parameters are well constrained as these were not varied in our analysis. This approach may be used to estimate sensitivity and detection of transiting bodies, such as exomoons, discussed next.

5.2. Other Applications

Exomoons, or moons of exoplanets, are oft-theorized but thus far undetected objects of interest both for their ability to inform understanding of planetary formation and for their potential habitability. While one candidate exomoon, Kepler-1625b-i (Teachey & Kipping 2018), has been identified it remains unconfirmed (Kreidberg et al. 2019; Teachey et al. 2020). The saga of Kepler-1625b i illustrates one of the primary barriers to observing exomoons their small size and correspondingly shallow transits. An additional complication is that exomoon transits will not be strictly periodic, due to orbital motion about their planets. This means that folding the light curve on the planet's orbital period to increase the S/N for a detection will not be effective.

Observations designed for detecting transiting exomoons may likely need to consist of very high-S/N photometry of more than one transit of a known exoplanet. In the near future the James Webb Space Telescope (JWST) will be the observatory best suited to these observations (Beichman et al. 2014). It has the ability to observe time series spectra of bright objects via the NIRSpec instrument (Bagnasco et al. 2007). Our

method is well-suited to model these observations, and we believe it may end up being the optimal method of identifying an exomoon transit signal. Simulating JWST observations of transiting planet systems with realistic noise (Sarkar et al. 2019), while applying our multiwavelength GP model to the results, would reveal what sensitivity JWST would have to shallow transiting bodies such as exomoons.

Transit transmission spectroscopy aims to measure the transmission spectrum of an exoplanet by measuring the effective radius of the planet as a function of wavelength. This is typically accomplished by varying the transit depth in the fit to the time series photometry at each wavelength as in (Berta et al. 2012) and Mandell et al. (2013). In studies like these the effects of stellar variability have been minimal and largely ignored. However in the future high-precision observations of bright stars at optical and near-IR wavelengths will likely have to contend with variability resulting from stellar granulation and/or pulsations (Sarkar et al. 2018).

Our method offers an elegant means of measuring the transmission spectrum. Given a sufficiently long time baseline, the wavelength dependence of a star's variability can be arbitrarily well-determined. In this case any "leftover" variability—variations in transit depth that are not explained by the wavelength dependence of the star's variability—can be attributed to the planet's transmission spectrum. By allowing the GP mean function to vary in transit depth across wavelength during MCMC analysis we can recover an estimate of the transmission spectrum with uncertainties in the presence of stellar variability. As such, this is a straightforward extension of our model as the only change involves varying the depth and limb-darkening as a function of wavelength, while the covariance remains the same as in the examples we have already shown.

Transit timing variations occur when the gravitational interaction between planets in a multi-planet system perturbs a transiting planet away from a Keplerian orbit (Agol et al. 2005; Holman 2005). The perturbed planet will transit earlier or later than the Keplerian solution would dictate based on the relative position of the transiting planet and perturbing planet. Observations of these transit timing variations over the course of many orbits help to constrain the orbital parameters of the perturber as well as the masses of both the perturber and the transiting planet. A notable application of this technique is to the seven-planet TRAPPIST-1 system (Gillon et al. 2017; Grimm et al. 2018). Correlated noise on timescales similar to the ingress/egress time of a transit can substantially affect measurements of the transit time (Agol & Fabrycky 2018).

At present correlated noise is observable for transiting planets around evolved stars. A notable example is Kepler-91b (Barclay et al. 2015), a hot Jupiter orbiting a red giant. Individual transits of Kepler-91b are nearly undetectable due to correlated noise on similar timescales and amplitudes to the transit signal. While most main-sequence Kepler targets do not show significant correlated variability, we expect that this variability will become observable in the near future with the advent of larger space-based telescopes such as the JWST. This means that accurate transit timing measurements for small planets transiting main-sequence stars will require the use of methods like ours to overcome the effects of correlated noise.

Variable phenomena. While we are primarily interested in the transiting planet problem, our multiwavelength GP implementation is likely to be useful for studies of other astronomical objects displaying time-correlated, stochastic variations. Many subfields in astronomy make use of GP variability models, or stochastic models that are equivalent to a Gaussian process, including the study of eclipsing binaries (e.g., Mahadevan et al. 2019), pulsating binaries (e.g., Hey et al. 2020), X-ray variability of the logarithm of the flux of X-ray binaries and active galactic nuclei (e.g., Uttley et al. 2005; Kelly et al. 2014), the study of transient phenomena such as supernovae (e.g., Kim et al. 2013), quasar variability (Kelly et al. 2009; MacLeod et al. 2010), reverberation mapping (Zu et al. 2011; Pancoast et al. 2014), and gravitational lensing time delays (Press et al. 1992; Hojjati et al. 2013; Hojjati & Linder 2014). Multiwavelength data may be exploited to better characterize these systems. For example, Boone (2019) found much better characterization of transients with multiwavelength Gaussian process modeling, while Peters et al. (2015) use the color dependence of the time-correlation of quasar variability to better characterize their physical properties. It is our hope that some of these fields may benefit from applying our new multiwavelength GP implementation to study the wavelength dependence of these various phenomena.

5.3. Limitations of the Method

When the second dimension's covariance matrix can be represented in terms of an outer product between a vector and itself, our method has a fast scaling with the number of data points. If the second dimension cannot be described as an outer product, then we obtain a poor scaling with the size of this dimension cubed. For the method to be computationally efficient in this case, the non-celerite dimension should be small compared to the size of the dimension along which the covariance is specified by a celerite kernel function. For problems where the second dimension is comparable in size to the first and where R must be arbitrarily defined, approximate methods such as the HODLR method (Ambikasaran et al. 2015), KISS-GP (Wilson & Nickisch 2015), or the black box methods implemented in GPyTorch (Gardner et al. 2018) may be more efficient.

The celerite method is a stationary GP method, meaning that the covariance kernel is constant in the one-dimensional coordinate. In other words, $k(x_i, x_j) = k(|x_i - x_j|)$. Our two-dimensional method inherits this limitation. What this means is that our method is not suited to modeling variability which changes substantially in amplitude or timescale over the time period in question. For instance, while our method is well-suited to modeling stellar variability over relatively short time periods, it would not be able to model solar variability across an entire solar cycle because the changing amplitude of the Sun's variability could not be captured by our stationary kernel.

If non-stationarity is required for a particular application, we refer users to other methods such as the sparse GP method of Almosallam et al. (2016) or the tree-structured GP of Bu & Turner (2014), both of which are approximate methods. It also may be possible to extend the celerite algorithm to non-stationary kernels by, for example, allowing the kernel coefficients to vary explicitly as a function of time, but we have yet to implement this.

Another limitation is fundamental to the Gaussian process framework: our method, like all GP methods, does a poor job of modeling outliers. When analyzing observational data, outliers are often dealt with by discarding them prior to analysis. However, in some cases outliers may represent useful

information and should be included in a model. One method of dealing with outliers without discarding them is to adopt a Student-t likelihood (Vanhatalo et al. 2009; Jylänki et al. 2011; Shah et al. 2014; Tang et al. 2017). The wider Student-t likelihood better accommodates outliers than the Gaussian likelihood, decreasing their influence on the regression. Similarly, a Gaussian mixture likelihood may be adopted, again increasing the robustness of the method to outliers (Daemi et al. 2019). We consider that a Student-t process (TP) may be an even better model for data sets containing outliers as well as having other advantages, especially with regard to TP prediction (Tracey & Wolpert 2018). We leave to future investigation the prospects for implementing a TP or TP likelihood version of celerite and evaluating the performance of these models on transit photometry. Unfortunately, Gaussian mixture likelihoods appear not to be compatable with the celerite formalism.

5.4. Limitations of the Multiband Photometric Noise Model

We make several assumptions in the construction of our multiband noise model, which likely do not hold in all cases. First and foremost, a Gaussian process assumes that the noise is stationary and Gaussian. This does not apply to some sources of noise, such as stellar flares, or sources that undergo outbursts in which the amplitude and/or shape of the power spectrum change dramatically. Likewise our method does not apply if there is a significant time delay between the bands, if one band involves a time convolution of the other, or if the correlated components of the bands have no correlation with one another.

Second, the specific form we have chosen for the wavelength covariance assumes that the wavelength dependence of the flux is due to varying covering fractions of a hot and cold component in a two-component photosphere. We expect that this model will work under different assumptions; for instance, small-amplitude temperature variations should have a similar behavior as area fluctuations. However, different sources of variability will result in different forms for the covariance in the wavelength dimension.

Additionally, if there are more than two components to the photosphere, then we must consider the possibility that each component's covering fraction varies with a different characteristic timescale. In this case rather than pairing a single wavelength covariance matrix T with a single time covariance matrix R to form the full covariance matrix $K = T \otimes R$, we should pair multiple wavelength covariance matrices with corresponding time covariance matrices, each having different characteristic timescales:

$$K = \sum_{i=1}^{N} T_i \otimes R_i . \quad (59)$$

Our code accepts multiple kernel components, each with a unique T matrix. While we have limited ourselves to the case of a single kernel component in this paper for the sake of clarity and simplicity, we plan to introduce this extension in more detail in a future paper.

In the examples in this paper we chose to fix the kernel parameters. In practice the kernel parameters will need to be measured alongside the parameters of the mean model. This brings up the question of how long of a time series is required to produce a sufficiently strong constraint on the kernel parameters that the inference of a transit is unambiguous. We also defer this question to future work.

Finally, our formulation assumes that the observations are complete; i.e., in the multiband times series example, every time of observation contains data in every band. In principle this assumption could be relaxed, and in Equation (52) the Kronecker products with a could be replaced with an a (and corresponding R matrix) which varies with time stamp, and only contains the amplitudes of the bands observed at each time stamp. This would also require modifying the indexing in Equations (53) and (54), but the rest of the method would remain the same.

6. Conclusions

We have extended the celerite method for fast one-dimensional GP computations to two dimensions. Our method inherits the $\mathcal{O}(N)$ scaling of celerite in one of the two dimensions while incurring a computational cost of $\mathcal{O}(M)$ for a grid with size M in the second dimension. Computing the two-dimensional GP on an N × M grid thus costs $\mathcal{O}(NM)$ using our method, compared to $\mathcal{O}(N^3M^3)$ for the direct solution (i.e., inverting the full $NM \times NM$ covariance matrix). This scaling applies only when the amplitude of correlated noise varies across the bands; a more general dependence on the second dimension has a poorer scaling, yet still improves upon direct solution.

This extension may have many possible applications, among them simultaneous modeling of stellar variability across wavelength. This application is of particular interest to us, as we would like to mitigate the effects of stellar variability on detecting transiting exoplanets and measuring their properties. We demonstrate that we can improve the precision of transit depth, time, and duration measurements by modeling the transit in multiple wavelengths when compared to the monochromatic case.

When the S/N is high, we have shown that a precision which is proportional to the photon noise limit is achievable. For instance, in the two-band case in which the correlated noise in one band is twice that in the second band, one can achieve $\sqrt{10}$ of the photon-noise limit. This means that to reach the same precision as the no correlated noise case requires 10 times as many photons, or a telescope which has a collecting area 10 times larger. In the limit of a blackbody which is photon-noise dominated, with a large number of bands, one can reach 2.2 times the photon-noise limit in which the correlated noise is absent. Hence, one needs to use a telescope with $2.2^2 = 4.8$ times the collecting area. Thus, in general one can achieve a precision of measurement which is comparable to the pure photon-noise limit, but this requires about an order of magnitude more photons to do so.

In future work, we plan to extend our variability model to model more realistic stellar variability by including terms in the covariance kernel function that capture variability on different timescales with different wavelength dependencies. We suggest that the SOHO spacecraft's three-channel SPM data may be a useful starting point for exploring the wavelength dependence of variability in Sun-like stars. This data set consists of measurements of the Sun's irradiance in three visible-light bands at one minute cadence (Frohlich et al. 1995).

We are additionally interested in applying our method to radial velocity observations of exoplanet host stars, following the method demonstrated by Rajpaul et al. (2015). This

requires us to compute linear combinations of the GP and its time derivatives,which in principle should be feasible.

Our code is available in the form of a pip installable python package called specgp. specgp extends exoplanet [4] to enable two-dimensional GP computations. Interested users can find instructions and tutorials at https://github.com/tagordon/specgp.

We acknowledge support from NSF grant AST-1907342. We thank Jackson Loper for useful conversations about CELERITE GPs. E.A. was supported by a Guggenheim Fellowship and NSF grant AST-1615315. We also acknowledge support from NASA's NExSS Virtual Planetary Laboratory, funded under NASA Astrobiology Institute Cooperative Agreement Number NNA13AA93A, and the NASA Astrobiology Program grant 80NSSC18K0829. This research was partially conducted during the Exostar19 program at the Kavli Institute for Theoretical Physics at UC Santa Barbara, which was supported in part by the National Science Foundation under grant No. NSF PHY-1748958.

This work was facilitated though the use of the advanced computational, storage, and networking infrastructure provided by the Hyak supercomputer system at the University of Washington.

Appendix A
celerite Algorithm for the Arbitrary Covariance Matrix in the Second Dimension

In this section we assume that the covariance in the second dimension, defined by the covariance matrix R, is arbitrary, subject to the constraint that the full covariance matrix K must be positive-definite.

We start by rewriting T in terms of the celerite generator matrices A, U, and V from Equation (34):

$$\begin{aligned} K &= S + [A_0 + \mathrm{tril}(UV^{\mathrm{T}}) + \mathrm{triu}(VU^{\mathrm{T}})] \otimes R \\ &= S + \mathrm{diag}(A_0 \otimes R) \\ &\quad + \mathrm{tril}(UV^{\mathrm{T}} \otimes R) + \mathrm{triu}(VU^{\mathrm{T}} \otimes R). \end{aligned} \tag{A1}$$

We rewrite R as RI_M where I_M is the $M \times M$ identity matrix, which allows us to write K as

$$\begin{aligned} K &= S + \mathrm{diag}(A_0 \otimes R) \\ &\quad + \mathrm{tril}((U \otimes R)(V \otimes I_M)^{\mathrm{T}}) \\ &\quad + \mathrm{triu}((V \otimes I_M)(U \otimes R)^{\mathrm{T}}) \end{aligned} \tag{A2}$$

where we have again applied Equation (50). As for the outer product case, we now have a semi-separable matrix defined by a new set of generators:

$$\begin{aligned} A' &= S + \mathrm{diag}(A_0 \otimes T) \\ U' &= U \otimes T \\ V' &= V \otimes I_M. \end{aligned} \tag{A3}$$

In terms of the celerite coefficients the refactored generator matrices are defined element-wise as follows:

$$\begin{aligned} A'_{(n-1)M+p,(n-1)M+p} &= s^2_{(n-1)M+1} + R_{p,p}\sum_{j=1}^{J} a_j \\ \tilde{U}'_{(n-1)M+p,(2j-1)M+q} &= R_{p,q}\tilde{U}_{n,2j-1} \\ \tilde{U}'_{(n-1)M+p,2jM+q} &= R_{p,q}\tilde{U}_{n,2j} \\ \tilde{V}'_{(n-1)M+p,(2j-1)M+q} &= \delta_{p,q}\tilde{V}_{n,2j-1} \\ \tilde{V}'_{(n-1)M+p,2jM+q} &= \delta_{p,q}\tilde{V}_{n,2j}, \end{aligned} \tag{A4}$$

where $\tilde{U}$ and $\tilde{V}$ are the refactored generator matrices defined in Equation (40), n ranges over $(1, N)$, p and q range over $(1, M)$, and $\delta_{p,q}$ is the Kronecker delta function:

$$\delta_{p,q} = \begin{cases} 1 & p = q \\ 0 & p \neq q. \end{cases} \tag{A5}$$

The recursive algorithm for carrying out the Cholesky decomposition is identical to the outer-product case. Starting with $D_{1,1} = A'_{1,1}$ and $W_{1,j} = \tilde{V}_{1,j}/D_{1,1}$, we then recursively define:

$$\begin{aligned} S_{n,j,k} &= f'_{n,j} f'_{n,k}[S_{n-1,j,k} + D_{n-1,n-1} W_{n-1,j} W_{n-1,k}], \\ D_{n,n} &= A'_{n,n} - \sum_{j=1}^{P}\sum_{k=1}^{P} \tilde{U}'_{n,j} S_{n,j,k} \tilde{U}'_{n,k}, \\ W_{n,j} &= \frac{1}{D_{n,n}}\left[\tilde{V}'_{n,j} - \sum_{k=1}^{P} \tilde{U}'_{n,k} S_{n,j,k}\right], \end{aligned} \tag{A6}$$

for $n = 2, \ldots, N$, $N' = NM$, with $P = 2JM$ the number of rows in $\tilde{U}'$ and $\tilde{V}'$. This additional factor of M accounts for the relatively poorer scaling of the method for arbitrary R over the outer-product case. For arbitrary definitions of R, $P = 2JM$ and the Cholesky decomposition thus scales as $\mathcal{O}(NJ^2M^3)$.

Appendix B
Computing the Log-likelihood

The log-likelihood is given by

$$\begin{aligned} \ln \mathcal{L} &= -\frac{1}{2}(\boldsymbol{y} - \boldsymbol{\mu})^{\mathrm{T}} K^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) \\ &\quad - \frac{1}{2}\ln\det(K) - \frac{N'}{2}\ln(2\pi), \end{aligned} \tag{B1}$$

which incorporates both the inverse and log-determinant of the covariance matrix, K. We therefore begin by describing the algorithms for each of these computations separately. The following algorithm comes directly from the original celerite paper, but with our modified definitions of the semi-separable matrix components, $\tilde{U}'$, $\tilde{V}'$, and W, and $f'_{n,j}$ rather than $f_{n,j}$ (see Section 3.2).

The product of the inverse covariance matrix with a vector, $\boldsymbol{z} = K^{-1}\boldsymbol{y}$, is computed with a two-part algorithm. We first compute the intermediate $\boldsymbol{z}'$, setting $z'_1 = y_1$, and then using the recursion relation

$$f_{n,j} = f'_{n,j}[f_{n-1,j} + W_{n-1,j} z'_{n-1}] \tag{B2}$$

[4] https://github.com/exoplanet-dev/exoplanet

$$z_n = y_n - \sum_{j=1}^{P} U_{nj} f_{n,j}, \quad \text{(B3)}$$

for $n = 2, \ldots, N$, where $N = NM$ and $f_{0,j} = 0$ for all j. We then use z to compute z in the second step of the algorithm, first setting $z_N = z_N / D_{N,N}$ and then using downward recursion

$$g_{n,j} = f_{n+1,j}[g_{n+1,j} + U_{n+1,j} z_{n+1}] \quad \text{(B4)}$$

$$z_n = \frac{z_n}{D_{n,n}} - \sum_{j=1}^{P} W_{n,j} g_{n,j} \quad \text{(B5)}$$

for $n = N - 1, \ldots, 1$, where $g_{N,j} = 0$ for all j and P is the number of columns in U, V, and W.

The log-determinant of K is given by

$$\ln(\det K) = \sum_{n=1}^{N} \ln(D_{n,n}). \quad \text{(B6)}$$

Putting these two steps together we can compute the log-likelihood. Because the algorithm for taking products of the inverse requires $O(NMP)$ operations, whereas the log-determinant can be computed in only $O(NM)$ operations, the log-likelihood computation as a whole scales as $O(NMP)$. In practice, the bottleneck for applications such as maximizing the likelihood or MCMC is computing the Cholesky factor rather than computing the log-likelihood, since the log-likelihood computation itself is faster by $O(P)$. Again we have $P = 2J$ when R is an outer product and $P = 2JM$ when R is any arbitrary covariance matrix. Figure 15 shows benchmarks for the log-likelihood computation demonstrating that the predicted scalings hold.

Appendix C
Prediction Algorithm

A GP prediction is an interpolation or extrapolation of the observed data using with the GP model. A prediction evaluated at each data point can also be thought of as a smoothing operation as it yields an estimate of the function with white noise removed.

The predictive distribution of a GP is a multivariate normal with a mean m^* and covariance K^* evaluated at the input coordinates x^*. For a GP with no white noise component the mean is constrained to pass directly through each observation of the data points y. For a GP with a non-zero white noise component the GP will act as a filter such that when the mean is subtracted from the data the residuals will be distributed according to a Gaussian distribution whose width is given by the GP white noise.

The predictive mean and covariance are computed as follows:

$$m^* = m_\theta(x^*) + K(x^*, x) K(x, x)^{-1} [y - m_\theta(x)] \quad \text{(C1)}$$

$$K^* = K(x^*, x^*) - K(x^*, x) K(x, x)^{-1} K(x, x^*) \quad \text{(C2)}$$

where $K(x^*, x)$ and $K(x, x^*)$ are the covariance kernel evaluated between the input coordinates and the data coordinates. If the input coordinates consist of N^* points in the first dimension and M^* points in the second then these matrices have dimensions $(M^* N^* \times NM)$ and $(NM \times N^* M^*)$ respectively.

For the two-dimensional Kronecker-structured covariance matrix $K = T \otimes R$, we can rewrite Equation (C1) as

$$m^* = m_\theta(x^*) + [T(x^*, x) \otimes R(x^*, x)] K(x, x)^{-1} \times [y - m_\theta(x)] \quad \text{(C3)}$$

$$= m_\theta(x^*) + [T(x^*, x) \otimes R(x^*, x)] z \quad \text{(C4)}$$

where $z = K(x, x)^{-1} [y - m_\theta(x)]$. Writing the second term of Equation (C3) in terms of the vectorization operator we have

$$[T(x^*, x) \otimes R(x^*, x)] z = [T(x^*, x) \otimes R(x^*, x)] \mathrm{vec}(Z) \quad \text{(C5)}$$

where $Z = Y - m_\theta(X)$ with X and Y matrices of size $N \times M$ defined by $x = \mathrm{vec}(X)$ and $y = \mathrm{vec}(Y)$ respectively. For matrices A, B, and C of sizes $(n \times m)$, $(m \times p)$, and $(p \times q)$ respectively there is an identity that states the following:

$$\mathrm{vec}(ABC) = (A \otimes C^{\mathrm{T}}) \mathrm{vec}(B). \quad \text{(C6)}$$

Applying this to Equation (C5) gives

$$[T(x^*, x) \otimes R(x^*, x)] z = \mathrm{vec}(TZR). \quad \text{(C7)}$$

The full expression for the predictive mean is now

$$m^* = m_\theta(x^*) + \mathrm{vec}(TZR). \quad \text{(C8)}$$

The matrix product TZR can be computed via a modified version of the celerite prediction algorithm presented in Foreman-Mackey et al. (2017).

First, we compute the product ZR at a computational cost of $O(NM)$ when R is outer product and $O(NM^2)$ for arbitrary R. We then compute

$$m^*_{p,m} = \sum_{n=1}^{N} \sum_{j=1}^{J} e^{-c_j |t^*_p - t_n|} [a_j \cos(d_j |t^*_p - t_n|) + b_j \sin(d_j |t^*_p - t_n|)] [ZR]_{n,m}. \quad \text{(C9)}$$

in two parts. Here p and m index the elements of the predicted mean matrix. The first part consists of a forward pass through $n_0 = 1, \ldots, N$ where we define:

$$G^-_{n,m,k} = [G^-_{n-1,p,k} + [ZR]_{n,m} V^*_{n,k}] e^{-c_{k//2}(t_{n+1} - t_n)} \quad \text{(C10)}$$

$$H^-_{p,n,k} = e^{-c_{k//2}(t^*_p - t_{n+1})} U^*_{p,k}, \quad \text{(C11)}$$

and[5] the second consisting of a backward pass through $n_0 = N, \ldots, 1$ where we define

$$G^+_{n,m,k} = [G^+_{n+1,p,k} + [ZR]_{n,m} U^*_{n,k}] e^{-c_{k//2}(t_n - t_{n-1})} \quad \text{(C12)}$$

$$H^+_{p,n,k} = e^{-c_{k//2}(t^*_{n-1} - t_p)} V^*_{p,k}, \quad \text{(C13)}$$

where $t_0 = t_1$, $t_{N+1} = t_N$, $G^-_{0,m,k} = 0$, and $G^+_{N+1,m,k} = 0$ for $k = 1, \ldots, 2J$ and for all m. The expressions for $U^*_{p,i}$ and $V^*_{p,i}$ are evaluated at t^*_p. For each value of p, G are evaluated recursively from n to n_0 and then the prediction $m^*_{p,m}$ is computed from

$$m^*_{p,m} = \sum_{k=1}^{P} [G^-_{n_0,m,k} H^-_{p,n_0,k} + G^+_{n_0+1,p,k} H^+_{p,n_0+1,k}]. \quad \text{(C14)}$$

[5] $k//2$ denotes integer division of k by 2. In other words, $k//2 = \mathrm{floor}(k/2)$.

This two-part computation scales as $\mathcal{O}(nN + n^*N^*)$ where n and n^* are constants. The overall scaling is therefore determined by the cost of the matrix multiplication step.

Appendix D
Sampling from the GP

A sample $\boldsymbol{y}$ can be drawn from a Gaussian process by computing

$$\boldsymbol{y} = \boldsymbol{m} + L\boldsymbol{n} \tag{D1}$$

where $\boldsymbol{m}$ is the mean function and $\boldsymbol{n}$ is a vector of draws from a normal distribution

$$n_i \sim \mathcal{N}(0, D_{i,i}^{1/2}) \tag{D2}$$

for each entry n_i in $\boldsymbol{n}$. The ordering of entries in $\boldsymbol{m}$ and consequently $\boldsymbol{y}$ is determined by the structure of K. For the Kronecker structured covariance matrix given in Equation (18), $\boldsymbol{m}$ is the concatenation of the N length-M vectors containing the mean function evaluated at each point in the second dimension at a given point in the first. In other words,

$$\boldsymbol{m} = (\boldsymbol{m}_1, \boldsymbol{m}_2, \ldots \boldsymbol{m}_N) \tag{D3}$$

where $\boldsymbol{m}_i = (m_{i,1}, m_{i,2}, \ldots m_{i,M})$ is the mean function evaluated at the ith point in the first dimension.

Thus $\boldsymbol{m}$ is a one-dimensional vector of length $N_\cdot = NM$ where N is the size of the first dimension and M the size of the second. The sample vector $\boldsymbol{y}$ then has the same structure. Most users will wish to either unpack the sample into M separate vectors obtained by taking every Mth entry in $\boldsymbol{y}$ or reshape it into an N × M array before displaying or examining the sample.

Appendix E
Proof of the Positive-definiteness of the Two-dimensional Kernel

To begin, we give the following equivalent definitions of positive-definite and positive-semidefinite matrices.

1. A square matrix $A \in \mathbb{R}^{n \times n}$ is positive-definite if and only if all of its eigenvalues are positive.
2. Equivalently, a square matrix $A \in \mathbb{R}^{n \times n}$ is positive-definite if and only if the scalar $\boldsymbol{x}^T A \boldsymbol{x}$ is positive for all real-valued vectors $\boldsymbol{x} \in \mathbb{R}^n$.
3. A square matrix $A \in \mathbb{R}^{n \times n}$ is positive-semidefinite if and only if all of its eigenvalues are nonnegative (they may be zero).
4. Equivalently, a square matrix $A \in \mathbb{R}^{n \times n}$ is positive-semidefinite if and only if the scalar $\boldsymbol{x}^T A \boldsymbol{x}$ is nonnegative for all real-valued vectors $\boldsymbol{x} \in \mathbb{R}^n$.

For a square matrix $A \in \mathbb{R}^{n \times n}$ with eigenvalues given by l_i for i = 1, K, n and $B \in \mathbb{R}^{m \times m}$ with eigenvalues m_j for $j = 1, \ldots, m$, the eigenvalues of the Kronecker product $A \otimes B$ are given by $l_i m_j$ for all values of i and j. To see this, note that for eigenvectors u and v corresponding to eigenvalues λ and μ of A and B respectively, $A\boldsymbol{u} \otimes B\boldsymbol{v} = (A \otimes B)(\boldsymbol{u} \otimes \boldsymbol{v}) = lm(\boldsymbol{u} \otimes \boldsymbol{v})$.

We now consider a positive-definite covariance matrix $T \in \mathbb{R}^{N \times N}$ and a positive-semidefinite square matrix $R \in \mathbb{R}^{M \times M}$. We consider the Kronecker product $T \otimes R$: since the eigenvalues of T are positive and the eigenvalues of R are nonnegative, the products of their eigenvalues that make up the eigenvalues of $T \otimes R$ will also be nonnegative. Therefore $T \otimes R$ is a positive-semidefinite matrix.

Similarly, if R is positive-definite (rather than positive-semidefinite), the eigenvalues of $T \otimes R$ will be uniformly positive and $T \otimes R$ will be a positive-definite matrix.

We now consider the effect of adding a real, positive-valued diagonal matrix $S \in \mathbb{R}^{NM \times NM}$ to the Kronecker product $T \otimes R$.

First consider the case that R is positive-definite. In this case $T \otimes R$ is positive-definite. Using the definition of positive-definiteness that states that a matrix A is positive definite if and only if $\boldsymbol{x}^T A \boldsymbol{x}$ is a positive scalar for all $\boldsymbol{x} \in \mathbb{R}^n$, we compute

$$\boldsymbol{x}^T (T \otimes R + S)\boldsymbol{x} = \boldsymbol{x}^T (T \otimes R)\boldsymbol{x} + \boldsymbol{x}^T S \boldsymbol{x} \tag{E1}$$

Assuming that R is positive-semidefinite, we have already established that $T \otimes R$ is positive-semidefinite as well. The left term above is therefore nonnegative. Since the matrix Σ is a diagonal matrix with positive entries, its eigenvalues, which are given by the diagonal entries, are positive, and therefore Σ is positive-definite. Consequently $\boldsymbol{x}^T S \boldsymbol{x}$ is a positive scalar and the matrix $T \otimes R + S$ is proven to be positive-definite. This completes the proof that the covariance matrix $K = T \otimes R + S$ is positive-definite in the case that T is a positive-definite kernel function, R is positive-definite, and Σ is a diagonal matrix with positive entries.

If we assume that R is positive-definite rather than positive-semidefinite then, as shown previously, $T \otimes R$ is itself a positive-definite matrix. In this case the covariance matrix $K = T \otimes R$ is positive-definite even without the addition of Σ. By the same logic as above, the addition of Σ will preserve the positive-definiteness of the covariance and $K = T \otimes R + S$ will be positive-definite as well.

Appendix F
Notation

Notation and symbols, in order of appearance:

K: covariance matrix.
k: kernel function corresponding to K.
x: general independent variable for the GP.
$\mathcal{L}$: GP likelihood.
$\boldsymbol{m}$: GP mean vector.
$\boldsymbol{y}$: vector of observations.
$N_\cdot$: number of observations corresponding to the length of vector $\boldsymbol{y}$.
ω_0: characteristic frequency of simple harmonic oscillator (SHO) term.
$\mathcal{F}$: stochastic force term, driving force of SHO.
t: an independent variable used to represent time.
Q: quality factor of SHO.
ω: an independent variable used to represent frequency in expressions for the power spectral density of a process.
S_0: Amplitude of the SHO.
τ: an independent variable used to represent time lag, as in $\tau_{i,j} = |t_i - t_j|$.
n, m, i, j, p, q: integers used to index independent variables and matrices.
x_h, x_c: covering fractions of a hot and cold component of the stellar photosphere.
R_*: stellar radius.
d: distance from star to observer.
F: flux.
B_n: nth spectral band.

l: independent variable representing wavelength.
$S_c(l)$, $S_h(l)$: spectra of the hotand cold components of a stellar photosphere.
$\phi_{B_i}(l)$: Response curve for band$_i$.B
a_i: variability amplitude integrated over band$_i$.B
s_c, s_h: var$(x_c)^{1/2}$, var$(x_h)^{1/2}$ respectively; the rms of the cold and hot covering fraction.
S_i: diagonal matrix containing. the white noise variances for each wavelength at the ith time index.
T: covariance matrix representing the first dimension or time dimension of the two-dimensionalGP. T will always be described by a celerite kernel function.
R: covariance matrix representingthe second dimension or wavelength dimension of the two-dimensional GP. R may be an arbitrary covariance matrix or an outer-product.
N: length of the first dimension, equal to the number of times in our example application of multiband time series.
M: length of second dimension, equal to the number of bands in our example application of multiband time series.
J: number of celerite terms in kernel function.
P: rank of celerite generator matrices.
$\boldsymbol{a}$: vector of correlated noise amplitudes in the second dimension.
S: diagonal matrix containing the white noise variances for each observation;the white noise component of the GP covariance matrix.
s_i^2: white noise variance for ith data point.
R_p: planetary radius.
t_0: time of center of transit.
d_n: duration of transit ingress/egress
d: transit duration (mid-ingress to mid-egress).
$\boldsymbol{q}$: vector of transit parameters.
$m_{\rm trap}$: transit mean model.
f_0: characteristic frequency of the correlated noise model.
$\bar{a}$: weighted mean of$\boldsymbol{a}$ used to represent the total amplitude of the correlated variability component of the GP summed over all bands (“monochromatic”).
s: mean of $\boldsymbol{s}$, the vector of white noise terms; used to represent the total amplitude of the uncorrelated variability component of the GP.
Φ: Information matrix.
N_θ: number of mean-model parameters,equal to the length of $\boldsymbol{q}$.
$s_{R_p^2}$: uncertainty on the transit depth (with “poly” and “mono” to indicate the polychromatic and monochromatic values).
$\boldsymbol{b}$: vector of coefficients used in defining the celerite kernel (Foreman-Mackey et al2017 use α).
a, b, c, d: celerite coefficients.
A: diagonal component of full kernel function; $K = A + \mathrm{tril}(UV^{\mathrm{T}}) + \mathrm{triu}(VU^{\mathrm{T}})$.
U,V : celerite generator matrices.
L: lower triangular matrix used in LDLT Cholesky decomposition.
tril, triu: lower and upper triangular matrix operators.
D: diagonal matrix used in decomposition.
W: matrix used in semi-separableLDLT Cholesky decomposition.
I: identity matrix.
S: intermediary matrix used in the celerite decomposition algorithm.
D: diagonal matrix in the Cholesky decomposition of K.
A_0: diagonal component of K with white noise amplitude set to zero; $A_0 = A - S$.
U_ϕ, V_ϕ, A_ϕ: Kronecker products of$U$,V , and A taken withΦ or R and I_M.
$\tilde{U}$,$\tilde{V}$,$\tilde{W}$: refactored celerite matrices corresponding to U, V, and W.
$\tilde{U}_\phi$, $\tilde{V}_\phi$, $\tilde{W}_\phi$: refactored celerite matricescorresponding to U_ϕ, V_ϕ, and A_ϕ.
f, f_ϕ : matrices used in the refactored version of celerite.
F, G : intermediary matrices for prediction algorithm.
$f_{n,j}$, $g_{n,j}$: intermediary vectorsused to compute the likelihood of the GP model.
$\boldsymbol{m}^*$: predictive mean model.
K^*: predictive covariance.
$\boldsymbol{x}^*$: independentvariable used to represent the points at which the predictive mean and covariance ofthe GP are evaluated.
$\boldsymbol{z}$: product between the observed vector$\boldsymbol{y}$and the inverse of the covariance matrix K^{-1} used to compute the GP likelihood.
$\boldsymbol{z}_\phi$: intermediary vector used to compute$\boldsymbol{z}$
X,Y,Z : $\boldsymbol{x} = \mathrm{vec}(X)$, $\boldsymbol{y} = \mathrm{vec}(Y)$, and $\boldsymbol{z} = \mathrm{vec}(Z)$ respectively; the matrix versions of $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ for the two-dimensional GP.
t^*: independent variable used to represent the points at which the predictive mean and covariance of the GP are evaluated; same as x^* when the independent variable is time.
H : intermediary matrix used to compute the GP prediction (Q in Foreman-Mackey et al2017).
N^*: the number of points at which the prediction is evaluated in the first dimensions.
n^*: constanton which the computationalscaling of the prediction algorithm depends.
$\boldsymbol{n}$: vector of random draws from a standard normal distribution used to draw a sample from the GP.

ORCID iDs

Tyler A. Gordon https://orcid.org/0000-0001-5253-1987
Eric Agol https://orcid.org/0000-0002-0802-9145
Daniel Foreman-Mackey https://orcid.org/0000-0002-9328-5652

References

Agol, E., & Fabrycky, D. C. 2018, in Handbook of Exoplanets, ed. H. Deeg & J. Belmonte (Cham: Springer), 797
Agol, E., Steffen,J., Sari, R., & Clarkson, W. 2005, MNRAS, 359, 567
Almosallam, I. A., Lindsay, S. N., Jarvis, M. J., & Roberts, S. J. 2016, MNRAS, 455, 2387
Ambikasaran,S., Foreman-Mackey,D., Greengard,L., Hogg, D. W., & O'Neil, M. 2015, ITPAM, 38, 252
Anderson,E. R., & Jefferies,S. M. 1990, ApJ, 364, 699
Bagnasco,G., Kolm, M., Ferruit, P., et al. 2007, Proc. SPIE, 6692, 66920M
Barclay,T., Endl, M., Huber,D., et al. 2015, ApJ, 800, 46
Barros,S. C. C., Demangeon,O., Díaz, R. F., et al. 2020, A&A, 634, A75
Beichman,C., Benneke,B., Knutson,H., et al. 2014, PASP, 126, 1134
Berta,Z. K., Charbonneau,D., Désert,J.-M., et al. 2012, ApJ, 747, 35
Boone,K. 2019, AJ, 158, 257
Bui, T. D., & Turner, R. E. 2014, in Advances in Neural Information Processing Systems 27, ed. Z. Ghahramani et al. (Red Hook, NY: Curran Associates), 2213
Carter,J. A., Yee, J. C., Eastman, J., Gaudi,B. S., & Winn, J. N. 2008, ApJ, 689, 499
Chakrabarty,A., & Sengupta,S. 2019, AJ, 158, 39
Csató,L., & Opper, M. 2002, Neural Comput., 14, 641

Daemi, A., Kodamana, H., & Huang, B. 2019, J. Process Control, 81, 209
Dawson, R. I., Johnson, J. A., Fabrycky, D. C., et al. 2014, ApJ, 791, 89
Deisenroth, M. P., & Ng, J. W. 2015, PMLR, 37, 1481
Foreman-Mackey, D., Agol, E., Ambikasaran, S., & Angus, R. 2017, AJ, 154, 220
Foreman-Mackey, D., Czekala, I., Luger, R., et al. 2019, dfm/exoplanet v0.2.3, Zenodo, doi:10.5281/zenodo.1998447
Frohlich, C., Romero, J., Roth, H., et al. 1995, SoPh, 162, 101
Gardner, J. R., Pleiss, G., Bindel, D., Weinberger, K. Q., & Wilson, A. G. 2018, arXiv:1809.11165
Gillon, M., Triaud, A. H. M. J., Demory, B.-O., et al. 2017, Natur, 542, 456
Grimm, S. L., Demory, B.-O., Gillon, M., et al. 2018, A&A, 613, A68
Hey, D. R., Murphy, S. J., Foreman-Mackey, D., et al. 2020, AJ, 159, 202
Hippke, M., David, T. J., Mulders, G. D., & Heller, R. 2019, AJ, 158, 143
Hoffman, M. D., & Gelman, A. 2014, JMLR, 15, 1593
Hojjati, A., Kim, A. G., & Linder, E. V. 2013, PhRvD, 87, 123512
Hojjati, A., & Linder, E. V. 2014, PhRvD, 90, 123501
Holman, M. J. 2005, Sci, 307, 1288
Jylänki, P., Vanhatalo, J., & Vehtari, A. 2011, JMLR, 12, 3227
Kallinger, T., De Ridder, J., Hekker, S., et al. 2014, A&A, 570, A41
Kelly, B. C., Bechtold, J., & Siemiginowska, A. 2009, ApJ, 698, 895
Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A., & Uttley, P. 2014, ApJ, 788, 33
Kim, A. G., Thomas, R. C., Aldering, G., et al. 2013, ApJ, 766, 84
Kipping, D. M., Hartman, J., Buchhave, L. A., et al. 2013, ApJ, 770, 101
Kreidberg, L., Luger, R., & Bedell, M. 2019, ApJL, 877, L15
Loper, J., Blei, D., Cunningham, J. P., & Paninski, L. 2020, arXiv:2003.05554
MacLeod, C. L., Ivezić, Ž, Kochanek, C. S., et al. 2010, ApJ, 721, 1014
Mahadevan, S., Bender, C. F., Hambleton, K., et al. 2019, ApJ, 884, 126
Mandell, A. M., Haynes, K., Sinukoff, E., et al. 2013, ApJ, 779, 128
Mazeh, T., & Faigler, S. 2010, A&A, 521, L59
Morris, B. M., Bobra, M. G., Agol, E., Lee, Y. J., & Hawley, S. L. 2020, MNRAS, 493, 5489
Nickson, T., Gunter, T., Lloyd, C., Osborne, M. A., & Roberts, S. 2015, arXiv:1510.07965
Pancoast, A., Brewer, B. J., Treu, T., et al. 2014, MNRAS, 445, 3073
Pereira, F., Campante, T. L., Cunha, M. S., et al. 2019, MNRAS, 489, 5764
Peters, C. M., Richards, G. T., Myers, A. D., et al. 2015, ApJ, 811, 95
Press, W. H., Rybicki, G. B., & Hewitt, J. N. 1992, ApJ, 385, 404
Rajpaul, V., Aigrain, S., Osborne, M. A., Reece, S., & Roberts, S. 2015, MNRAS, 452, 2269
Rasmussen, C. E., & Williams, C. K. I. 2006, Gaussian Processes for Machine Learning (Cambridge, MA: MIT Press)
Rybicki, G. B., & Press, W. H. 1992, ApJ, 398, 169
Rybicki, G. B., & Press, W. H. 1995, PhRvL, 74, 1060
Sarkar, S., Argyriou, I., Vandenbussche, B., Papageorgiou, A., & Pascale, E. 2018, MNRAS, 481, 2871
Sarkar, S., Madhusudhan, N., & Papageorgiou, A. 2019, MNRAS, 491, 378
Shah, A., Wilson, A., & Ghahramani, Z. 2014, JMLR, 33, 877
Snelson, E., & Ghahramani, Z. 2006, in Advances in Neural Information Processing Systems 18, ed. Y. Weiss, B. Schölkopf, & J. C. Platt (Cambridge, MA: MIT Press), 1257
Sulis, S., Lendl, M., Hofmeister, S., et al. 2020, A&A, 636, A70
Tang, Q., Niu, L., Wang, Y., et al. 2017, IJCAI, 17, 2822
Teachey, A., & Kipping, D. M. 2018, SciA, 4, eaav1784
Teachey, A., Kipping, D., Burke, C. J., Angus, R., & Howard, A. W. 2020, AJ, 159, 142
Tracey, B. D., & Wolpert, D. H. 2018, arXiv:1801.06147
Uttley, P., McHardy, I. M., & Vaughan, S. 2005, MNRAS, 359, 345
Vallisneri, M. 2008, PhRvD, 77, 042001
Vanhatalo, J., Jylänki, P., & Vehtari, A. 2009, Advances in Neural Information Processing Systems, 1910
Wilson, A., & Adams, R. 2013, in Int. Conf. on Machine Learning, ed. S. Dasgupta & D. McAllester (Brookline, MA: Microtome Publishing), 1067
Wilson, A. G., & Nickisch, H. 2015, arXiv:1503.01057
Zhang, Y., Leithead, W. E., & Leith, D. J. 2005, in Proc. 44th IEEE Conf. on Decision and Control, ed. F. Allgöwer et al. (Piscataway, NJ: IEEE), 3711
Zu, Y., Kochanek, C. S., & Peterson, B. M. 2011, ApJ, 735, 80