Draft version July 27, 2019
Typeset using LaTeX **preprint2** style in AASTeX62

Tyler A. Gordon[1] and Eric Agol[1]

[1]*Department of Astronomy, University of Washington, Box 351580, U.W., Seattle, WA 98195-1580, USA*

(Received; Revised; Accepted)

Submitted to

ABSTRACT

Gaussian processes are a model of stochastic variability frequently used in analyzing astrophysical time-series. They have been used to study variability in light curves of stars and AGN, as well as the spatial distribution of the extragalactic dust distribution. In the realm of stellar variability, GP regression is used to characterise stellar rotation periods, search for transit signals in noisy data, and analyze RV curves. One of the chief limitations of this method is the computational time incurred in computing a GP model for large or multi-dimensional data-sets. While approximate methods can make GP regression feasible for some of these applications, in others exact methods are preferable. Here we present a method based on the celerite GP implementation which enables fast, exact computation of two-dimensional GP models for data-sets with a small second dimension. This includes multi-bandpass photometry of transit and microlensing light curves. Our method also may have applications to RV and direct imaging methods.

1. INTRODUCTION

A Gaussian process (GP) is a generalization of a Gaussian distribution to function-space (**?**). GPs have been used extensively in astrophysics as a model of stochastic variability across both space and time, to study a diverse range of phenomena including variable stars, AGN(**?**), the extragalactic dust distribution(**?**), and gravitational wave signals(**?**). They have also been used to account for correlated noise in photometric and RV measurements of exoplanet hosts(**???**).

GP models offer advantages over other methods of modeling stochastic variability because they are able to account for correlated variations without the imposition of a parametric model. *Gordon*: more about advantages of GPs

The primary limitation of GP methods is that most GP operations are computationally expensive. For instance, the common tasks of computing likelihoods and sampling from the GP require $\mathcal{O}(N^3)$ operations for N data points. This is especially problematic for use cases which require repeated calls to the likelihood function, such as minimization and performing MCMC. This has prevented the adoption of GP methods for data-sets larger than $N \sim 10^4$ points(**?**).

The issue of computational expense can be partially overcome either by approximating the GP or by restricting the user to a subset GPs for which the computation can be sped up(**?**). In the latter category we find celerite (**?**), a fast one-dimensional GP method which increases

Corresponding author: Tyler A. Gordon
tagordon@uw.edu

the speed of most common GP computations to $\mathcal{O}(NJ^2)$ where J is typically small compared to N. celerite achieves this speedup by limiting the user to GPs with a specific functional form of the covariance. Another limitation of celerite is that the method is only applicable to one-dimensional GPs. Here we present a method to rectify this difficiency of the original celerite method. Our extension of celerite computes two-dimenional GPs on an N by M grid in $\mathcal{O}(NJ^2M^3)$ operations. This is an improvement over the general case in which the computation of the same GP would require $\mathcal{O}(N^3M^3)$ operations. Because the speed of our computation is still sharply dependent on M, it is most useful in cases where $M << N$. An example would be modeling variability in multi-bandpass photometry where M would represent the number of bands.

In this paper we first review the celerite method before introducing our extension of the original method. We then give a detailed example of GP regression on multi-band transit photometry to recover transit parameters in the presence of significant correlated noise. Finally, we discuss other potential applications to our method and conclude with an outline of future work in which we hope to further reduce the computational expense of a two-dimensional GP. This would enable the use of GP methods on even larger data-sets and potentially make feasible GP computations in higher dimensions.

2. GAUSSIAN PROCESSES AND THE CELERITE MODEL

A Gaussian process is a stochastic noise model which consists of a mean:

$$\boldsymbol{\mu_\theta}(\boldsymbol{x}) = \begin{pmatrix} \mu_{\boldsymbol{\theta}}(x_1) & \cdots & \mu_{\boldsymbol{\theta}}(x_N) \end{pmatrix} \quad (1)$$

and covariance defined the kernel function $k_{\boldsymbol{\alpha}}(x_n, x_m)$. The mean and kernel functions depend on the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ respectively, which are referred to as "hyperparameters" of the GP. The kernel function defines a covariance matrix $[K]_{n,m} = k_\alpha(x_n, x_m)$. This model has a likelihood function given by

$$\begin{aligned} \ln\ \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\alpha}) &= \ln\ p(\boldsymbol{y}|\boldsymbol{x}, \boldsymbol{\theta}, \boldsymbol{\alpha}) \quad (2) \\ &= -\frac{1}{2}\boldsymbol{r_\theta}^{\mathrm{T}} K_{\boldsymbol{\alpha}}^{-1} \boldsymbol{r_\theta} - \frac{1}{2}\ln\ \det(K_{\boldsymbol{\alpha}}) - \frac{N}{2}\ln(2\pi) \end{aligned}$$

where

$$\boldsymbol{y} = \begin{pmatrix} y_1 & \cdots & y_N \end{pmatrix} \quad (3)$$

are the data, N is the number of datapoints, and

$$\boldsymbol{r_\theta} = \boldsymbol{y} - \boldsymbol{\mu_\theta}(\boldsymbol{x}) \quad (4)$$

By maximizing the likelihood (or, in practice, minimizing the negative of the log-likelihood), one can obtain an estimate of the hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$. Markov Chain Monte Carlo (MCMC) methods can be used to sample the posterior distribution over the hyperparameters.

2.1. *The celerite model*

The primary limitation of Gaussian process noise models is that they are computationally expensive, with computation of the likelihood function scaling as $\mathcal{O}(N^3)$ due to the matrix inversion. This is especially problematic for use cases such as MCMC which require calling the likelihood function many times. This has prevented the adoption of Gaussian process noise models for datasets larger than $N \approx 10^3$, except in cases where approximate methods suffice. The celerite method overcomes this limitation by reducing the time to execute most GP computations to linear time for one class of kernel functions(**?**). We give a brief description of the celerite method and its limitations below.

Consider a one-dimensional gaussian process evaluated at the coordinates

$$\boldsymbol{x} = \begin{pmatrix} t_1 & \cdots & t_N \end{pmatrix} \quad (5)$$

The celerite kernel is given by

$$k_\alpha(t_n, t_m) = \sigma_n^2 \delta_{nm} + \sum_{j=1}^{J} a_j e^{-c_j \tau_{nm}} \quad (6)$$

where $\boldsymbol{\alpha} = (a_1...a_j, c_1...c_j)$, σ_n^2 is the variance of the gaussian-distributed white noise, and $\tau_{nm} = |t_n - t_m|$. The coefficients a and c may be complex, in which case we introduce b and d which are the imaginary parts of a and c respectively. This kernel defines a celerite model with J terms.

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

$$K = A + \mathrm{tril}(UV^\mathrm{T}) + \mathrm{triu}(VU^\mathrm{T}) \tag{7}$$

In the case of our covariance matrix, the generator matrices are specified by:

$$\begin{aligned} U_{n,2j-1} &= a_j e^{-c_j t_n} \cos(d_j t_n) + b_j e^{-c_j t_n} \sin(d_j t_n) \\ U_{n,2j} &= a_j e^{-c_j t_n} \sin(d_j t_n) - b_j e^{-c_j t_n} \cos(d_j t_n) \\ V_{m,2j-1} &= e^{c_j t_m} \cos(d_j t_m) \\ V_{m,2j} &= e^{c_j t_m} \sin(d_j t_m) \end{aligned} \tag{8}$$

and A is given by:

$$A_{n,n} = \sigma_n^2 + \sum_{j=1}^{J} a_j \tag{9}$$

Computing the Cholesky decomposition for this covariance matrix can be accomplished in $\mathcal{O}(NJ^2)$ operations, allowing for the fast computation of the GP likelihood function[1].

The main limitation of celerite is that the method only applies to one class of kernels. However, the celerite kernel overcomes this limitation by being both versatile in its ability to approximate other kernels, and also directly applicable to many problems. The versatility of this kernel is demonstrated in *Gordon*: figure reference which shows approximations to several popular kernels achieved by appropriate choices of the celerite coefficients a_j and c_j.

[1] The algorithm is described in appendex A

One model of stellar oscillations models the star as a stochastically-driven damped harmonic oscillator(**?**). The solution to the equations describing such an oscillator is a GP with a celerite kernel with coefficients

$$a_{j\pm} = \frac{1}{2} S_0 \omega_0 \left[1 \pm \frac{1}{\sqrt{1-4Q^2}}\right] \tag{10}$$

$$c_{j\pm} = \frac{\omega_0}{2Q}\left[1 \mp \sqrt{1-4Q^2}\right] \tag{11}$$

where S_0 is proportional to the power at the undamped frequency of the oscillator, ω_0, and Q is a parameter known as the quality factor of the oscillator. Variations in the stellar radius due to these oscillations gives rise to brightness variations in the star's light curve *Gordon*: reference?. This means that a celerite GP can be used as a model for stellar variability due to p-mode oscillations. Additionally, setting $Q = 1/\sqrt{2}$ results in a PSD which has been used to describe granulation-driven stellar variability(**?**).

3. 2D GAUSSIAN PROCESSES

In addition to being limited to one class of kernels celerite , as formulated above, is inherently one-dimensional. For many applications, such as modeling noise in single-bandpass light curves or RV time series GP computations in one dimension are all that is necessary. However in other cases such as modeling noise in simultaneous multi-bandpass light curves or disentangling noise in RV data using multiple activity indicators, two-dimensional GP methods are necessary. We outline a method to compute a GP in two dimensions utilizing the celerite method for the larger of the two dimensions to speedup most GP operations by a factor of $\sim \mathcal{O}(N^2)$ over ordinary 2D GPs.

In two-dimensions the independent variable is

$$\boldsymbol{x} = \begin{pmatrix} (t_1, u_1) & \cdots & (t_N, u_N) \end{pmatrix} \tag{12}$$

If the covariance between two points satisfies

$$k_\alpha(x_n, x_m) = q_\alpha(t_n, t_m) r_\alpha(u_n, u_m) \tag{13}$$

then the covariance matrix has the structure of a Hadamard matrix product (element-wise multiplication)

$$[K_\alpha]_{nm} = [Q_\alpha]_{nm} [R_\alpha]_{nm} \tag{14}$$

or

$$K = Q \circ R \tag{15}$$

Now consider a GP evaluated on a $(N \times M)$ grid. We begin by defining

$$\boldsymbol{t'} = \begin{pmatrix} t'_1 \cdots t'_N \end{pmatrix} \tag{16}$$

$$\boldsymbol{u'} = \begin{pmatrix} u'_1 \cdots u'_M \end{pmatrix} \tag{17}$$

to be vectors containing only the unique t and u values in increasing order. We also define the corresponding covariance matrices for each dimension:

$$[Q'_\alpha]_{nm} = q_\alpha(t'_n, t'_m) \tag{18a}$$

$$[R'_\alpha]_{pq} = r_\alpha(u'_p, u'_q) \tag{18b}$$

In this case, assuming (13), the full covariance matrix can be written as a Kronecker product between these two one-dimensional covariances[2]:

$$K_\alpha = Q'_\alpha \otimes R'_\alpha \tag{19}$$

Since the rest of this discussion will focus on the case of gridded data, we will drop the prime notation and use Q_α and R_α to refer to the matrices in (18). In the following discussion we assume that the covariance in t is described by a celerite kernel, and that the covariance in u is arbitrary. This means that Q_α can be expressed in terms of the generator matrices U and V. Following equation (7), in the absence of a white noise term, we can write the covariance matrix as follows[3]:

$$\begin{aligned} K &= \left(A + \mathrm{tril}(UV^\mathrm{T}) + \mathrm{triu}(VU^\mathrm{T})\right) \otimes R \\ &= A' + \mathrm{tril}(U'V'^\mathrm{T}) + \mathrm{triu}(V'U'^T) \end{aligned} \tag{20}$$

where

$$A' = A \otimes R \tag{21}$$

$$U' = U \otimes R \tag{22}$$

$$V' = V \otimes I_M \tag{23}$$

To include white noise we add a white noise term to each element of A:

$$A'_{n,n} \to A'_{n,n} + \sigma_n^2 \tag{24}$$

From the form of equation 20 we can see that the covariance is, as before, semiseparable. However Kronecker multiplication by R and I_M has increased the size of the generator matrices from $(N \times 2J)$ to $(NM \times 2JM)$. As a result, the number of operations necessary to compute the Cholesky decomposition to increase from $\mathcal{O}(NJ^2)$ to $\mathcal{O}(NJ^2M^3)$. The number of operations for sampling from and extrapolating/interpolating with the two-dimensional GP scales the same. Some improvement can be made in the scaling for extrapolation and interpolation if the covariance in the second dimension can be expressed via a celerite kernel. In this case the algorithm requires $\mathcal{O}(nNP + mMP)$ operations where P is the number of points at which the prediction is evaluated and n, m are constants. The Cholesky composition algorithm as well as algorithms for interpolation/extrapolation and sampling from the GP are described in appendix D.

It should be noted that the scaling of our method is still relatively poor as we add points to our grid in the second dimension. This is why the method is most suitable for cases where the

[2] See appendix B for a proof of (19) and notes on the Hadamard and Kronecker products.

[3] See appendix C for proof

second dimension is much smaller than the first, i.e. where $M << N$. This condition applies to the important cases of modeling noise in multi-bandpass observations.

4. EXAMPLE: SOLAR SYSTEM TRANSITS

One application of fast two-dimensional Gaussian Processes is modeling noise in multi-bandpass light curves to, for example, search for transits and infer their parameters. One-dimensional GPs are already in use as noise models for transit light curves. By setting the mean of the GP equal to a transit model and minimizing the GP likelihood, transit parameters can be more accurately inferred in the presence of correlated stellar noise. However, for cases where a shallow transit signal is obscured by correlated noise, a one-dimensional GP noise model is not sufficient to disentangle signal from noise.

This problem can be mitigated by examining the wavelength-dependence of the star's variability. The correlated noise component of the light curve will be subject to this wavelength-dependence whereas the mean (the transit signal) will not share the same dependence. Thus by including multiple bandpasses and simultaneously modeling the covariance between bands using a two-dimensional GP.

In what follows we aim to demonstrate this method using transits of Solar System planets as an example. We use publicly available light curves from the SOHO mission's three-channel sunphotometer (SPH) (**???**) and add Gaussian noise to simulate observations taken at varying distances from the Sun. We've injected transits of the Earth and Venus into that data and we demonstrate the recovery of transit parameters for both objects.

4.1. *modeling the wavelength-dependence of stellar variability*

Our first task is to construct a model of the Sun's variability that allows us to determine the form of the R matrix. The sun's p-mode oscillations peak at between 5 and 6 minutes (**?**) compared to transit durations for Venus and Earth of 11 and 13 hours respectively, so we ignore this source of variability in our model. Of more relevance are brightness variations modulated by changing patterns of hotter and cooler regions of the photosphere. We model the solar photosphere as a mix of two regions of temperatures T_h and T_c with $T_h > T_c$. The flux in a filter A is

$$F_A = \int_{\lambda_1}^{\lambda_2} \phi_A(\lambda) \left[x I_\lambda(T_h) + (1-x) I_\lambda(T_c) \right] d\lambda \tag{25}$$

where $\phi_A(\lambda)$ is the profile of the filter and x is the covering fraction of the region at T_h. As the covering fraction varies, the flux changes according to

$$\frac{dF_A}{dx} = \int_{\lambda_1}^{\lambda_2} \phi_A(\lambda) \left[I_\lambda(T_h) - I_\lambda(T_c) \right] d\lambda \tag{26}$$

From this we see that the rate at which the flux varies with variations in x depends on the contrast in brightness between the two regions within the filter. We therefore expect the flux to be more variable where the contrast between the two regions is higher than where it is lower. For instance, in the infrared, we should see higher amplitude variability features in filters centered at shorter wavelengths than longer ones.

We can expand the flux as a function of covering fraction in a Taylor series to first order:

$$F_A(x) = F_A(\bar{x}) + \frac{dF_A}{dx}(x - \bar{x}) \tag{27}$$

where $\bar{x}$ is the mean covering fraction. We can then compute the covariance between the flux in filter A at time t_n and the flux in some filter B at time t_m/ Filter B is assumed to be centered at a longer wavelength than A:

$$\begin{aligned} \mathrm{cov}(F_A(t_n), F_B(t_m)) &= \mathrm{E}\left[(F_A(t_n) - \bar{F_A})(F_B(t_m) - \bar{F_B}) \right] \\ &= \frac{dF_A}{dx}\frac{dF_B}{dx} \mathrm{E}\left[(x_n - \bar{x})(x_m - \bar{x}) \right] \\ &= C_A C_B \mathrm{cov}(x_n, x_m) \end{aligned} \tag{28}$$

where C_A and C_B are the derivatives of F_A and F_B with respect to covering fraction and we've used the fact that $\mathrm{cov}(x_n, x_m) = \mathrm{E}\left[(x_n - \bar{x})(x_m - \bar{x})\right]$. This result satisfies equation (13) with $r(u_n, u_m) = C_n C_m$. The corresponding covariance matrix for the second dimension is

$$[R]_{nm} = C_n C_m \tag{29}$$