A partially marginalized likelihood for exoplanet inference

David W. Hogg (NYU CCPP, MPIA)
Paul Baines (UC Davis)
Rebekah Dawson (CfA, Berkeley)
Daniel Foreman-Mackey (NYU CCPP)
with help from everyone at SAMSI Kepler (exoSAMSI)

disclaimer: This document is a draft, and not yet ready for public consumption. In particular, any use of the content in this document might represent a serious lapse in judgement.

abstract: We present a likelihood function for the time-series photometry that makes up a lightcurve from the *Kepler* mission (or similar experiment). The function includes parameters that describe stochastic stellar variability and parameters that describe periodic exoplanet (or other companion) transits. The stellar variability model is a Gaussian Process in a wavelet basis that is capable of modeling non-trivial time correlations with a diagonal covariance matrix. These choices make it possible to *marginalize out* all stellar-variability parameters, leaving the user with a flexible likelihood function parameterized only by exoplanet parameters. We show something non-trivial.

1 Introduction

Transits are hard to find in general; harder in variable stars. All stars are variable.

Fitting transits and stellar variability simultaneously is The Right Thing To Do (tm).

Gaussian Processes are a Good Idea (tm).

Wavelets are a Good Idea (tm).

2 Likelihood generalities

There are N observations Y_n of a single star taken at times t_n. The model is

$$Y_n \;=\; [1 - Q(t_n \,|\, \omega)]\, F(t_n \,|\, \alpha) + e_n \quad , \tag{1}$$

where $Q(\cdot \,|\, \omega)$ is a function that describes the attenuation of the starlight caused by the transiting exoplanet, ω is a blob of parameters describing an exoplanet's size and orbit, $F(\cdot \,|\, \alpha)$ is a function that describes the apparent brightness of the star as a function of time, α is a blob of parameters describing the stellar mean flux and the variation of that flux with time, and e_n is the noise contribution to the nth datum (coming from photon and read noise, among other things). This description—plus a model for the noise—will lead to a justifiable likelihood function. If we model the noise contributions as being Gaussian and independent with known variances σ_n^2, the likelihood function becomes

$$p(\{Y_n\}_{n=1}^N \,|\, \omega, \alpha, \varphi) \;=\; \prod_{n=1}^N p(Y_n \,|\, \omega, \alpha, \varphi) \tag{2}$$

$$p(Y_n \,|\, \omega, \alpha, \varphi) \;=\; \mathcal{N}(Y_n \,|\, \mu_n, \sigma_n^2) \tag{3}$$

$$\mu_n \;\equiv\; [1 - Q(t_n \,|\, \omega)]\, F(t_n \,|\, \alpha) \quad , \tag{4}$$

where $\{Y_n\}_{n=1}^N$ is the set of all data, φ is an enormous blob of hyperparameters that includes all our decision-making and assumptions (and more, soon), the product in the likelihood encodes our "independent noise" assumption, and $\mathcal{N}(x \,|\, m, V)$ is the Gaussian for x with mean m and variance V,

By assumption, we care only about the exoplanet parameters ω and not in the least about the stellar variability parameters α. We marginalize by

$$p(\{Y_n\}_{n=1}^N \,|\, \omega, \varphi) \;=\; \int p(\{Y_n\}_{n=1}^N \,|\, \omega, \alpha, \varphi)\, p(\alpha \,|\, \varphi)\, \mathrm{d}\alpha \quad , \tag{5}$$

where we have had to introduce a prior PDF $p(\alpha \,|\, \varphi)$ for the variability parameters. This prior will in general depend on the hyperparameters φ (which is a very good thing (because learning can happen there). The object defined in (5) is the *partially marginalized likelihood* we seek.

3 Wavelet-based Gaussian Process

The considerations *(1)* that marginalization is paramount, and *(2)* that the variability of the star is stochastic, lead us naturally to think about Gaussian

Processes. In a Gaussian process in this context, there are hyperparameters controlling a non-trivial covariance matrix in the space of the data, and the star is modeled (in a prior sense) as a Gaussian draw from this covariance matrix. The brilliant idea generated at SAMSI is that we can model a non-trivial covariance matrix, which will be dense in the original data space, with a trivial, *diagonal* covariance matrix in a different basis. For all sorts of good reasons, we will think about wavelet bases, but what follows is pretty general.

Because we are about to get all linear-algebra-y, let's make some notation changes: Instead of the "set" of data $\{Y_n\}_{n=1}^N$ we will move to thinking about the N data points as components of a column vector y. Instead of the "function" $Q(\cdot \,|\, \omega)$, we will think of the column vector q. Instead of individual scalar noise variances, we will think of variance tensors.

We are going to transform to another basis; that is, we are going to "rotate" between the time basis in which vector y lives to a basis-function-amplitude vector a wavelet-amplitude by a transformation like

$$y \;\leftarrow\; W^\mathsf{T} \cdot a \quad , \tag{6}$$

where W is a matrix of "weights" or basis functions, and a is a vector of basis-function amplitudes. If this transformation is unitary, the inverse is

$$a \;\leftarrow\; W \cdot y \quad . \tag{7}$$

In the real world, to make the transformation unitary, the user either must have evenly spaced, homoskedastic data (all σ_n^2 identical), or else build a unique, inhomogeneous wavelet basis customized to every non-uniform, heteroskedastic data set or time series. In what follows, we won't need this unitarity to be true.

In this framework, the star variability parameters α—the parameters that set the particular shape of the star's flux as a function of time—are contained in the vector a that comprises the amplitudes of the wavelet coefficients. This model is equivalent to assuming that the star's flux as a function of time can be written as a superposition of basis functions

$$F(t \,|\, \alpha) \;=\; \sum_{k=1}^{K} \sum_{p=1}^{P_k} a_{kp}\, g_{kp}(t) \tag{8}$$

where the a_{kp} are the elements of the vector a (and therefore the contents of the parameter blob α), and the g_{kp} are the wavelet basis functions. The

basis functions are indexed by two indices, one (k) going over the frequency domain, and the other (p) going over the location or time domain.

The two key ideas of the variability model we propose here—the model for $p(\alpha \,|\, \varphi)$ or equivalently $p(a \,|\, \varphi)$—are *(1)* that in the wavelet basis the Gaussian from which the vector a is drawn can have a diagonal covariance matrix, and *(2)* that the variances on the diagonal of that matrix will depend only on the frequency index k and not at all on the location index p. Under these assumptions (really only the diagonality assumption), we can write the partially marginalized likelihood as:

$$p(y \,|\, \omega, \varphi) = \mathcal{N}(y \,|\, m, V) \quad (9)$$

$$m \equiv \bar{f}\,[O - q] \quad (10)$$

$$V \equiv \Psi + W^{\mathsf{T}} \cdot \Phi \cdot W \quad , \quad (11)$$

where m and V are the mean vector and variance tensor of the Gaussian Process, $\bar{f}$ is the mean stellar flux, O is the N-dimensional vector of ones (unities), q is the column vector made up of the $Q(t_n \,|\, \omega)$ values describing the attenuation caused by the transiting planet, Ψ is the diagonal noise variance tensor with σ_n^2 values down the diagonal, W is the linear operator that transforms points from the (natural) time basis to the wavelet basis, and Φ is the diagonal tensor of variances appropriate to the wavelet components. That is, down the diagonal of Φ are the variances from which, in a prior sense, the amplitudes a of the wavelet components are expected to be drawn, in the absence of (or prior to) data. Another way to put it is that the variance tensor V is dense, but it is made from the two diagonal tensors Ψ and Φ, each of which is diagonal (but in different bases). Some notes:

(a) Although we don't explicitly show the integral over the vector a of amplitudes, it is included implicitly in the marginalized likelihood $p(y \,|\, \omega, \varphi)$; the integral is a convolution of Gaussians, which produces a Gaussian with broader variance. In this way of thinking about it, the wavelet-amplitude diagonal variance tensor Φ is part of the hyperparameter blob φ. With additional hyperparameters in this blob, and choices about hyperpriors on the hyperparameters, the variability model itself can be marginalized out. This may indeed become one of our goals below.

(b) The form (11) for the variance tensor assumes diagonality but does not yet capitalize on the frequency and location of the wavelets. The idea that the variances will depend on the freqency index but *not* the location index. That means that diagonal tensor Φ will have repeated elements; the

number of free parameters in Φ will be far smaller than the number of wavelet amplitudes in a.

(c) For many wavelet bases, the number K of distinct frequencies k is only of order $\log_2(N)$, where N is the number of data points in the lightcurve. This means that even a non-parametric form for Φ will only grow logarithmically with the size of the data set.

(d) The wavelet amplitude prior variance Φ does *not* need to be diagonal in principle. It could be sparse, or tri-diagonal, or dense. Of course if it is fully dense, the advantage of going to the wavelet basis is reduced. We won't consider that case further here.

(e) If we only use the weight matrix W to transform *covariance matrices*—and not transform and transform back the data or residuals—we can in fact make the W non-square, both by removing rows that correspond to wavelet scales at which we expect no power (more on this below), and by removing columns that correspond to data locations at which there are no observations (vanishing $1/\sigma_n^2$). Finally, If we drop columns in blocks (corresponding to contiguous intervals of data locations), there will in general be some "unsupported" wavelet components. These orphaned rows ought to also be dropped to keep relevant matrices full-rank. These changes break orthonormality and completeness of the basis, but they do not change the symbolic form (11) of the total variance tensor.

In what follows, we are going to treat the diagonal data noise variance tensor Ψ as known, the wavelet basis W as fixed, but parameterize and permit to be fit the Gaussian Process variance tensor Φ, parameterized in a form that makes the variances sensitive to scale but not location.

4 *Kepler* lightcurve specifics

My proposal—it can't be described as "our proposal" just yet—is to do the following violence to the model to keep the core ideas, but make it tractable, at the cost of strengthening or distorting some assumptions.

`Going to work on the SAP data. Explain why`

`Turning Untrendy knobs to 11. We don't want any discontinuities or artifacts.`

For the purposes of computing the wavelet transformation matrices W, treat the data as coming homoskedastic on a perfectly equally-spaced grid. Then we will drop rows and columns where there is no data, or no support

for particular wavelet basis functions.

We will also punk (slightly) the wavelet transform in the following sense: We will treat the data as uniform in cadence for the purposes of the W, for the purposes of computing the *exoplanet transits*, treat the data as coming from the true, non-uniform grid in BJD. That is, we will be able to use "out of the box" code to generate the wavelet transform matrix W.

Our non-parametric form for Φ is just to give one variance σ_k^2 to every wavelet frequency scale k. That's the most flexible possible model; it only has of order $\log_2(N)$ parameters.

Let's show here that the wavelet coefficients do have low autocorrelation amplitude.

5 Search

We have to say something about the horrible, horrible things we do to get started.

6 Experiments

Experiment 1: Run on a single injection into a single good G star. Show that we recover correctly.

Experiment 2: Run on a representative sample of G stars with a representative sample of injections. Show that we rock it.

Experiment 3: Run on 10,000 G stars. Find all ELPaSLS.

7 Discussion

Most astrophysical data are generated by processes that are a mixture of stochastic and deterministic signals. We need to learn more about stochastic models. This project is a baby step towards this.

We only considered transit and lightcurve data. Does this have anything to do with radial velocities? Of course it does. Ideas about modeling low-Q oscillations.

We just *made up* a wavelet basis and the data looked pretty uncorrelated in this basis. That's just lucky. In principle a more rigid GP concept could be explicitly diagonalized. More speculatively, we could *learn the basis*. This is related to sparse coding and dictionary methods.

We only looked at a tiny fraction of the available data. We only looked at a tiny fraction of the possible exoplanets in those data. Let's scale up!

Acknowledgements: It is a pleasure to thank Eric Ford for organizing the workshop that got this started.

Yann LeCun

It is a pleasure to thank SAMSI...

Bibliography