A partially marginalized likelihood for exoplanet inference

David W. Hogg (NYU)
with help from Paul Baines (Davis) and everyone else at SAMSI Kepler

disclaimer: `This document is a draft, and not yet ready for public consumption. In particular, any use of the content in this document would 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 generalities

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

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

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 variability, 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

$$
\begin{aligned}
p(\{Y_n\}_{n=1}^N \,|\, \omega, \alpha, \varphi) &= \prod_{n=1}^{N} p(Y_n \,|\, \omega, \alpha, \varphi) && (2) \\
p(Y_n \,|\, \omega, \alpha, \varphi) &= N(Y_n \,|\, \mu_n, \sigma_n^2) && (3) \\
\mu_n &\equiv [1 - q(t_n \,|\, \omega)]\, F(t_n \,|\, \alpha) \quad , && (4)
\end{aligned}
$$

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 $N(x \,|\, m, V)$ 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 star parameters α. We marginalize by

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

where we have had to introduce a prior PDF $p(\alpha \,|\, \varphi)$ for the star 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.

The considerations *(1)* that marginalization is of 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