Draft version November 28, 2020
Typeset using LaTeX **modern** style in AASTeX62

Normalized Gaussian Processes

Rodrigo Luger

1. INTRODUCTION

It is standard practice in astronomy to mean- or median- normalize datasets, since one is often more interested in deviations from some baseline than in the value of the baseline itself. This is the case, for example, in searches for transiting exoplanets or photometric studies of stellar variability, where the raw data consists of a timeseries of fluxes measured in counts on the detector. The absolute number of counts from a particular target is physically uninteresting, as it depends on a host of variables such as the distance to the target, the collecting area of the instrument, and the quantum efficiency of the detector. However, fractional deviations from the mean number of counts *are* physically meaningful, as they can encode information such as the size of the transiting planet or the size and contrast of star spots. Normalization by the mean (or median) allows one to analyze data in units of (say) parts per million rather than counts per second.

Another common practice in astronomy is to model one's data (or at least a component of one's data) as a Gaussian process (GP; e.g., Rasmussen & Williams 2005). GPs offer a convenient, flexible, and efficient way of modeling correlated data and have seen extensive use in both transiting exoplanet and stellar variability studies (?). In one dimension, a GP is fully described by a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$, the latter of which encodes information about correlations and periodicities in the data that are in some cases related to physical parameters of interest (such as the rotation period of a star or the lifetime of star spots).

In this note, we show that these two common practices are somewhat at odds with each other. Specifically, if a physical process that generates a dataset is distributed as a GP, the normalized process *cannot* be distributed as a GP. Provided certain conditions are met, the normalized process can be well *approximated* by a GP, albeit one with a different covariance matrix $\tilde{\boldsymbol{\Sigma}}$ that is not simply a scalar multiple of the original covariance matrix. Moreover, if the original process is described by a stationary kernel (i.e., one in which covariances are independent of phase), the normalized process is not guaranteed to be.

For many applications, the results of this note are not likely to make much of a difference, since GPs are often used to model nuisance signals; in this case, the optimal hyperparameters describing the GP covariance are physically uninteresting. However, in cases where one wishes to interpret the GP hyperparameters in a physical context (such as using a periodic GP kernel to infer stellar rotation rates), normalizing one's data to the mean or median value can impart (potentially significant) bias.

2. THE PROBLEM

Let $\mathbf{x} = (x_0 \ \cdots \ x_{K-1})^\top$ be a K-dimensional multivariate normal random variable distributed according to a Gaussian Process with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$:

$$\mathbf{x} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \ . \tag{1}$$

For simplicity, let us assume the mean is constant, i.e.,

$$\boldsymbol{\mu} = \mu\,\mathbf{j} \tag{2}$$

where $\mathbf{j}$ is the vector of K ones. Now suppose we cannot observe samples of $\mathbf{x}$ directly, but instead we can observe samples from the *normalized* process, which we will call $\tilde{\mathbf{x}}$:

$$\begin{aligned} \tilde{\mathbf{x}} &\equiv \frac{\mathbf{x}}{\bar{x}} \\ &= \frac{\mathbf{x}}{\frac{1}{K}\sum\limits_{k=0}^{K-1} x} \ . \end{aligned} \tag{3}$$

The random variable $\tilde{\mathbf{x}}$ is no longer normally distributed, but as long as the variance of $\bar{x}$ is small, we can attempt to approximate it as such. Let the mean and covariance of $\tilde{\mathbf{x}}$ be $\tilde{\boldsymbol{\mu}}$ and $\tilde{\boldsymbol{\Sigma}}$, respectively:

$$\tilde{\mathbf{x}} \mathrel{\dot\sim} \mathcal{N}\left(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\Sigma}}\right) \ . \tag{4}$$

By definition, the mean is unity, i.e,

$$\tilde{\boldsymbol{\mu}} = \mathbf{j}\,. \tag{5}$$

What is the expression for $\tilde{\boldsymbol{\Sigma}}$?

3. THE SOLUTION

The covariance is given by

$$\begin{aligned} \tilde{\boldsymbol{\Sigma}} &= \mathbb{E}\big[(\tilde{\mathbf{x}} - \tilde{\boldsymbol{\mu}})(\tilde{\mathbf{x}} - \tilde{\boldsymbol{\mu}})^\top\big] \\ &= \mathbb{E}\left[\left(\frac{\mathbf{x}}{\bar{x}} - \mathbf{j}\right)\left(\frac{\mathbf{x}}{\bar{x}} - \mathbf{j}\right)^\top\right] \\ &= \mathbb{E}\left[\frac{\mathbf{x}\mathbf{x}^\top}{\bar{x}^2} - \frac{\mathbf{x}\,\mathbf{j}^\top}{\bar{x}} - \frac{\mathbf{j}\,\mathbf{x}^\top}{\bar{x}} + \mathbf{j}\,\mathbf{j}^\top\right] \\ &= \mathbb{E}\left[\frac{\mathbf{x}\mathbf{x}^\top}{\bar{x}^2}\right] - \mathbb{E}\left[\frac{\mathbf{x}\,\mathbf{j}^\top}{\bar{x}}\right] - \mathbb{E}\left[\frac{\mathbf{j}\,\mathbf{x}^\top}{\bar{x}}\right] + \mathbf{j}\,\mathbf{j}^\top . \end{aligned} \tag{6}$$

To evaluate this, it is convenient to write

$$\mathbf{x} = \boldsymbol{\mu} + \mathbf{L}\mathbf{u}\,, \tag{7}$$

where $\mathbf{L}$ is the lower Cholesky decomposition of $\boldsymbol{\Sigma}$, i.e.,

$$\boldsymbol{\Sigma} = \mathbf{L}\,\mathbf{L}^{\top} \tag{8}$$

and $\mathbf{u}$ is a standard multivariate normal random variable,

$$\mathbf{u} \sim \mathcal{N}\left(0, \mathbf{I}\right)\,, \tag{9}$$

where $\mathbf{I}$ is the $K \times K$ identity matrix. The expression for the covariance $\tilde{\boldsymbol{\Sigma}}$ may now be written

$$\tilde{\boldsymbol{\Sigma}} = \mathbb{E}\left[\frac{(\boldsymbol{\mu} + \mathbf{L}\mathbf{u})(\boldsymbol{\mu} + \mathbf{L}\mathbf{u})^{\top}}{\bar{x}^2}\right] - \mathbb{E}\left[\frac{(\boldsymbol{\mu} + \mathbf{L}\mathbf{u})\mathbf{j}^{\top}}{\bar{x}}\right] - \mathbb{E}\left[\frac{\mathbf{j}(\boldsymbol{\mu} + \mathbf{L}\mathbf{u})^{\top}}{\bar{x}}\right] + \mathbf{j}\,\mathbf{j}^{\top} \tag{10}$$

The mean $\bar{x}$ is

$$\begin{aligned} \bar{x} &= \sum_{j=0}^{K-1} (\mu_j + L_{i,j} u_j) \\ &= \mu + \frac{1}{K}\mathbf{j}^{\top}\mathbf{L}\,\mathbf{u} \\ &= \mu(1 + \epsilon)\,, \end{aligned} \tag{11}$$

where we define the quantity

$$\begin{aligned} \epsilon &\equiv \frac{\bar{x}}{\mu} - 1 \\ &= \frac{1}{\mu K}\mathbf{j}^{\top}\mathbf{L}\,\mathbf{u}\,. \end{aligned} \tag{12}$$

Plugging this in and rearranging, we obtain

$$\tilde{\boldsymbol{\Sigma}} = \mathbf{P} - \frac{1}{\mu}\left(\mathbf{Q} + \mathbf{Q}^{\top}\right) + \frac{1}{\mu^2}\mathbf{R}\,, \tag{13}$$

where we define the matrices

$$\mathbf{P} \equiv \mathbb{E}\Big[\frac{\epsilon^2\,\mathbf{j}\,\mathbf{j}^{\top}}{(1+\epsilon)^2}\Big] \tag{14}$$

$$\mathbf{Q} \equiv \mathbb{E}\left[\frac{\epsilon\,\mathbf{L}\,\mathbf{u}\,\mathbf{j}^{\top}}{(1+\epsilon)^2}\right] \tag{15}$$

$$\mathbf{R} \equiv \mathbb{E}\left[\frac{\mathbf{L}\,\mathbf{u}\,\mathbf{u}^{\top}\,\mathbf{L}^{\top}}{(1+\epsilon)^2}\right] \tag{16}$$

The denominators in the expressions above make direct evaluation of the expectations intractable. Provided $\left|\epsilon\right| < 1$ (an assumption we'll revisit below), we can Taylor expand the matrices as

$$\mathbf{P} = \sum_{n=0}^{\infty}(-1)^n(n+1)\,\mathbb{E}\left[\epsilon^{n+2}\,\mathbf{j}\,\mathbf{j}^\top\right] \tag{17}$$

$$\mathbf{Q} = \sum_{n=0}^{\infty}(-1)^n(n+1)\,\mathbb{E}\left[\epsilon^{n+1}\,\mathbf{L}\,\mathbf{u}\,\mathbf{j}^\top\right] \tag{18}$$

$$\mathbf{R} = \sum_{n=0}^{\infty}(-1)^n(n+1)\,\mathbb{E}\left[\epsilon^{n}\,\mathbf{L}\,\mathbf{u}\,\mathbf{u}^\top\mathbf{L}^\top\right] \tag{19}$$

In the Appendix, we show that the expecations in the expressions above may be computed from

$$\mathbb{E}\left[\epsilon^{n+2}\,\mathbf{j}\,\mathbf{j}^\top\right] = \frac{(n+1)g_n m^{\frac{n+2}{2}}}{\mu^{n+2}}\mathbf{j}\,\mathbf{j}^\top \tag{20}$$

$$\mathbb{E}\left[\epsilon^{n+1}\,\mathbf{L}\,\mathbf{u}\,\mathbf{j}^\top\right] = \frac{(n+1)K\,g_n m^{\frac{n}{2}}}{\mu^{n+1}}\,\mathbf{m}\,\mathbf{j}^\top \tag{21}$$

$$\mathbb{E}\left[\epsilon^{n}\,\mathbf{L}\,\mathbf{u}\,\mathbf{u}^\top\mathbf{L}^\top\right] = \frac{g_n}{\mu^n}\left(n\,m^{\frac{n-2}{2}}\mathbf{m}\,\mathbf{m}^\top + m^{\frac{n}{2}}\mathbf{\Sigma}\right) \tag{22}$$

where

$$g_n \equiv \begin{cases} \dfrac{n!}{2^{\frac{n}{2}}\left(\frac{n}{2}\right)!} & n\,\text{even} \\ 0 & n\,\text{odd}\,, \end{cases} \tag{23}$$

is the expression for the n^{th} moment of the standard normal distribution,

$$\mathbf{m} \equiv \frac{1}{K}\mathbf{\Sigma}\,\mathbf{j} \tag{24}$$

is the average of each row in $\mathbf{\Sigma}$, and

$$m \equiv \frac{1}{K^2}\mathbf{j}^\top\mathbf{\Sigma}\,\mathbf{j} \tag{25}$$

is the average of all elements in $\mathbf{\Sigma}$. Inserting these expressions into Equation (13) and rearranging, we obtain the final expression for the normalized covariance,

$$\tilde{\mathbf{\Sigma}} = \frac{\alpha\,\mathbf{\Sigma}}{\mu^2} + \frac{\alpha\,\left(\mathbf{s}\,\mathbf{s}^\top - \mathbf{m}\,\mathbf{m}^\top\right)}{\mu^2 m} + \frac{\beta\,\left(\mathbf{s}\,\mathbf{s}^\top\right)}{\mu^2 m}\,, \tag{26}$$

where we define

$$\mathbf{s} \equiv m\mathbf{j} - \mathbf{m} \tag{27}$$

$$\alpha \equiv \sum_{n=0}^{\infty} \frac{(2n+1)!}{2^n \, n!} z^n \tag{28}$$

$$\beta \equiv \sum_{n=0}^{\infty} \frac{2n(2n+1)!}{2^n \, n!} z^n \tag{29}$$

as well as the dimensionless parameter

$$z \equiv \frac{m}{\mu^2} \,, \tag{30}$$

equal to the average element of the covariance matrix divided by the square of the GP mean.

4. THE BAD NEWS

In the previous section we derived a compact expression for the covariance of a normalized Gaussian process (Equation 26), which can be computed as the sum of a term proportional to the original covariance $\mathbf{\Sigma}$ and a low-rank correction. Unfortunately, the expressions for the scaling constants α and β (Equations 28 and 29) involve series expansions that *do not converge*. This is demonstrated in Figure 1, which shows the terms in the summations in Equations 28 and 29 for different values of z. While the terms initially decrease in magnitude (left panel), asymptotically approaching a finite value for both α and β (right panels), as the expansion order increases the terms eventually diverge, leading to infinite values for both coefficients. Even though the series take longer to diverge for smaller values of z, the divergence occurs for all $z \neq 0$: i.e., the radius of convergence for both series is zero.

Why is this? Recall the assumption we made when expanding the expectation integrals in Equations (17), (18), and (19): the Taylor expansion converges only for $|\epsilon| < 1$. The quantity ϵ (Equation 12) is indeed small provided the sample mean $\bar{x}$ is close to the GP mean μ. However, ϵ is a random variable, equal to the weighted sum of K standard normal random variables. Because a normal distribution has infinite support, ϵ is guaranteed to take on values greater than unity in the limit that an infinite number of samples are drawn from the process. In particular, it is guaranteed to take on values arbitrarily close to unity, in which case the sample mean approaches zero and the values in the normalized sample diverge. This reveals a fundamental flaw in our premise: it is simply not correct to normalize a sample from a Gaussian process by its mean value, since the resulting covariance is formally infinite.

Nevertheless, if one were to draw a very large number of samples from a normal distribution with unit mean and small standard deviation, the probability of drawing

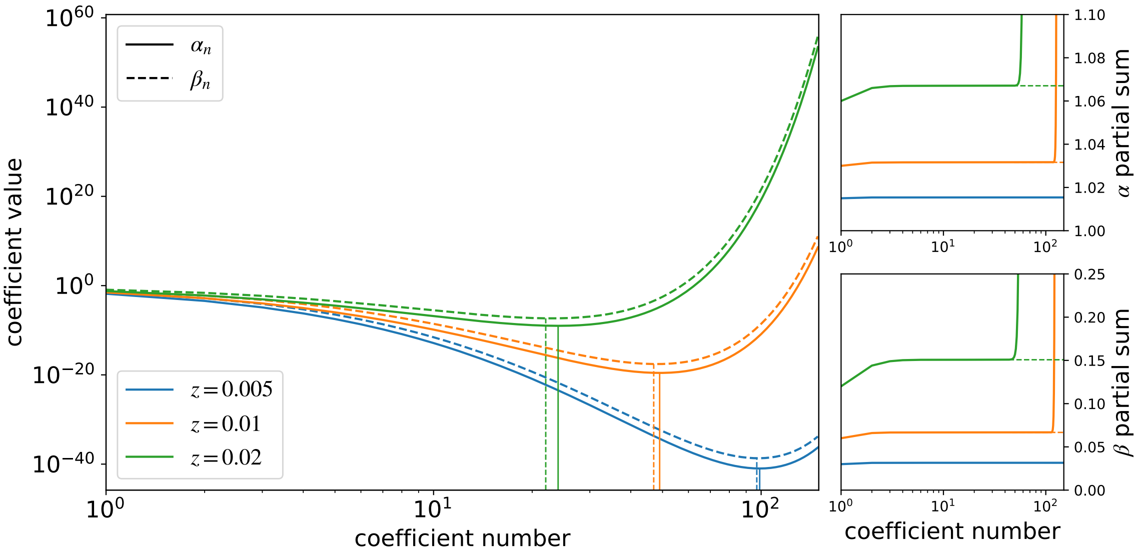

Figure 1. Asymptotic series expansions for α and β (Equations 28 and 29). The left panel shows the value of each coefficient (α: solid, β: dashed) as a function of the index n for three different values of z. In all cases, the value initially decreases with n but eventually diverges to $+\infty$ as $n \rightarrow \infty$. The vertical lines indicate the optimal truncation order N, for which the partial sums yield the best approximation to the asymptotic values of the cofficients. The right panels show the partial sums evaluated at each order; the value at the optimal truncation order is marked with the horizontal dashed lines.

a sample with mean close to or smaller than zero is vanishingly small (for example, for $\sigma \lesssim 0.1$, this probability is $\lesssim 10^{-22}$). Any numerical (sampling) estimate of the covariance matrix of the normalized Gaussian process will therefore yield a result that asymptotically approaches a finite, consistent value as the number of samples increases. If it were practical to draw $\sim 10^{22}$ samples, however, the estimate would eventually diverge, since some of the normalized samples would have divergent values.

This is the exact same behavior we see in the series solution above. While the series formally diverges, the finite asymptotic value obtained by truncating the expansion early can be understood as an estimate of the covariance of the normalized process *ignoring the divergent tails of the distribution.*

In the following sections, we empirically show that this interpretation is correct, and that the expression for the normalized covariance (Equation 26) is accurate provided $z \ll 1$.

5. NON-STATIONARITY

REFERENCES

Rasmussen, C. E., & Williams, C. K. I.
2005, Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning) (The MIT Press)

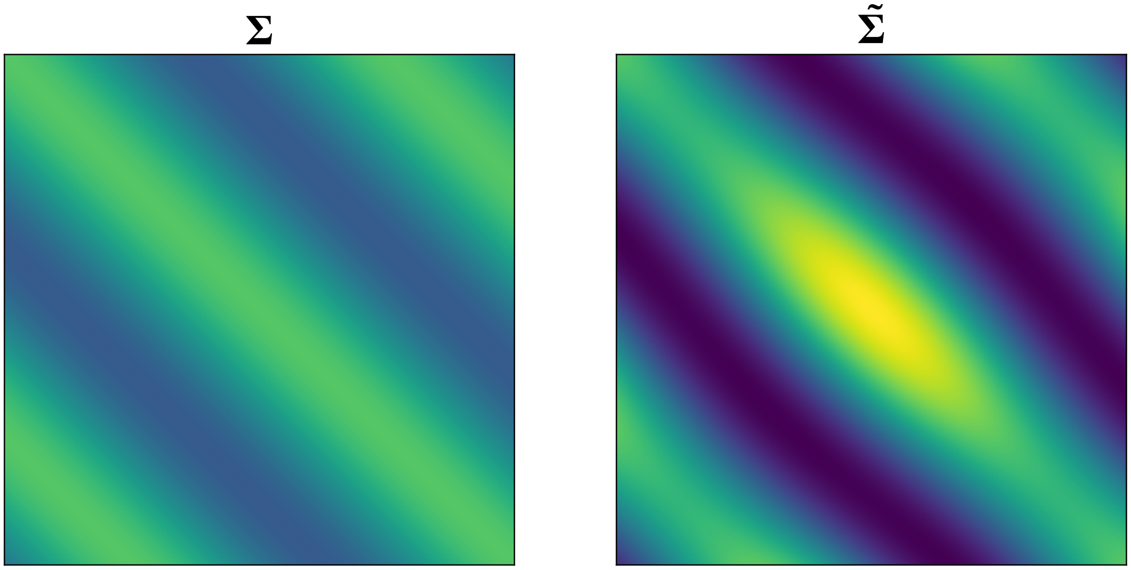

Figure 2. The covariance matrix corresponding to a periodic kernel (left) and the covariance matrix of the corresponding normalized process (right).