Draft version November 30, 2017
Typeset using LaTeX **modern** style in AASTeX61

GAUSSIAN PROCESSES IN ASTRONOMY

Daniel Foreman-Mackey[1, 2, 3]

[1] *NASA Sagan Fellow*

[2] *Astronomy Department, University of Washington, Seattle, WA*

[3] *Center for Computational Astrophysics, Flatiron Institute, New York, NY*

ABSTRACT

This is an abstract.

Corresponding author: Daniel Foreman-Mackey
formeman.mackey@gmail.com

1. INTRODUCTION

2. NOTATION

3. PROBABILISTIC INFERENCE

To start, let's consider the general problem of fitting a model to data. Whether you are a frequentist or a Bayesian, the key quantity of interest in any fitting procedure is the likelihood function

$$\mathcal{L}(\boldsymbol{\theta}) \quad = \quad p(D\,|\,\boldsymbol{\theta}) \tag{1}$$

where D represents the data and $\boldsymbol{\theta}$ are the parameters of the model. The right-hand side of this equation can be read as "the probability of a dataset D given a specific set of model parameters $\boldsymbol{\theta}$." More formally, this is the probability density function (pdf) for D *conditioned* on $\boldsymbol{\theta}$. It is important to remember that this is a pdf *over datasets*. This means that

$$\int p(D\,|\,\boldsymbol{\theta})\,\mathrm{d}D \quad = \quad 1 \tag{2}$$

but

$$\int p(D\,|\,\boldsymbol{\theta})\,\mathrm{d}\boldsymbol{\theta} \quad \neq \quad 1 \quad . \tag{3}$$

It is our job as scientists to specify the model – and, hence, the likelihood function – that we want to fit. Once we have specified this pdf, there are two main ways to make inferences about the model parameters $\boldsymbol{\theta}$ based on a given dataset. The first method is to maximize the likelihood to find a point estimate of the "best-fit" parameters. This can be written as

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\mathrm{argmax}}\,\mathcal{L}(\boldsymbol{\theta}) \tag{4}$$

where $\boldsymbol{\theta}^*$ are the "maximum likelihood" parameters and the argmax operator indicates that we are finding the value of $\boldsymbol{\theta}$ that maximizes $\mathcal{L}(\boldsymbol{\theta})$. In practice, this maximization must usually be performed numerically using a non-linear optimization routine.

To quantify the uncertainties on our constraints on the model parameters or marginalize over our uncertainty in some nuisance parameters, we can quantify the posterior pdf for $\boldsymbol{\theta}$

$$p(\boldsymbol{\theta}\,|\,D) \quad = \quad \frac{p(\boldsymbol{\theta})\,p(D\,|\,\boldsymbol{\theta})}{p(D)} \quad . \tag{5}$$

As above, we can read the left-hand side of this equation as "the probability of the parameters conditioned on the data." In practice, this pdf is usually obtained by drawing samples from $p(\boldsymbol{\theta}\,|\,D)$ using a numerical method like Markov chain Monte Carlo (MCMC). In this case, we must specify the likelihood function and, additionally, the prior pdf $p(\boldsymbol{\theta})$.

The point here is that the likelihood function is the key ingredient in any model fitting context. In fact, some people (including this author) would argue that the likelihood function is as much part of the model as your code that computes the physics of your system. The following sections present some of the commonly made assumptions about the likelihood function for astronomical data analysis, and then derive the Gaussian process (GP) likelihood as a generalization of the standard equations. In Section 5, we derive the likelihood function for a GP model (Equation 17) and the key point is that this function is probably a *drop-in replacement* for the likelihood that you're currently using.

4. THE GAUSSIAN LIKELIHOOD

In this section, we derive the likelihood function that is most commonly used for astronomical data analysis and discuss this choice in the context of probabilistic data analysis. It is not uncommon for astronomers to compute and minimize something that we refer to as "chi-squared" χ^2, the sum of normalized squared residuals. This section demonstrates that this is equivalent to maximizing a likelihood function derived under a restrictive set of assumptions. This will then allow us, in the next section, to generalize this procedure.

It is commonly assumed that a set of data points with error bars represent independent measurements with Gaussian uncertainties of known variance. Under this assumption, for a model $f(\boldsymbol{x};\,\boldsymbol{\theta})$, the likelihood for a single data point y_n measured at coordinates $\boldsymbol{x}_n$ with error bar σ_n is

$$p(y_n\,|\,\boldsymbol{x}_n,\,\sigma_n,\,\boldsymbol{\theta}) \quad = \quad \frac{1}{\sqrt{2\,\pi\,{\sigma_n}^2}}\,\exp\left(-\frac{1}{2}\,\frac{[y_n - f(\boldsymbol{x};\,\boldsymbol{\theta})]^2}{{\sigma_n}^2}\right) \quad . \tag{6}$$

Therefore, the joint likelihood for a set of N data points $\{\boldsymbol{x}_n,\,y_n,\,\sigma_n\}_{n=1}^N$ is

$$p(\{y_n\}\,|\,\{\boldsymbol{x}_n,\,\sigma_n\},\,\boldsymbol{\theta}) \quad = \quad \prod_{n=1}^N \frac{1}{\sqrt{2\,\pi\,{\sigma_n}^2}}\,\exp\left(-\frac{1}{2}\,\frac{[y_n - f(\boldsymbol{x};\,\boldsymbol{\theta})]^2}{{\sigma_n}^2}\right) \quad . \tag{7}$$

It is common practice to work with the natural logarithm of this quantity instead of the likelihood directly. Taking the logarithm of Equation (7), we find

$$\log p(\{y_n\}\,|\,\{\boldsymbol{x}_n,\,\sigma_n\},\,\boldsymbol{\theta}) \quad = \quad -\frac{1}{2}\sum_{n=1}^N \left[\frac{[y_n - f(\boldsymbol{x};\,\boldsymbol{\theta})]^2}{{\sigma_n}^2} + \log\left(2\,\pi\,{\sigma_n}^2\right)\right] \quad . \tag{8}$$

It is worth qualitatively considering the roles of the two terms in Equation (8) because this discussion will come up repeatedly throughout this paper. The first term within the square brackets is what is commonly referred to as "χ^2" in the astronomy literature and it quantifies the "goodness-of-fit" of the model. The second term quantifies the specificity of the model and penalizes overly general models. For fixed uncertainties $\{\sigma_n\}$, this second term is a constant with respect to the parameters $\boldsymbol{\theta}$ and maximizing the log-likelihood in Equation (8) is equivalent to minimizing χ^2. In other words, calculating χ^2 requires assuming that the uncertainties are independent Gaussians with known variance.

This more general formulation of the Gaussian likelihood function will come in handy for our derivation of GP modeling shortly, but let's start with a concrete example where minimizing χ^2 is not sufficient. It is not uncommon for uncertainties on astronomical quantities to be underestimated or unknown. In this case, we must fit for a parametric representation of the uncertainties simultaneously with the model $f(\boldsymbol{x};\,\boldsymbol{\theta})$. To do this, we might include another parameter, we'll call it s, to quantify the amount by which the uncertainties are underestimated. In this case, the likelihood becomes

$$\log p(\{y_n\}\,|\,\{\boldsymbol{x}_n,\,\sigma_n\},\,\boldsymbol{\theta},\,s) \quad = \quad -\frac{1}{2}\sum_{n=1}^{N}\left[\frac{[y_n - f(\boldsymbol{x};\,\boldsymbol{\theta})]^2}{{\sigma_n}^2+s^2} + \log\left(2\,\pi\,[{\sigma_n}^2+s^2]\right)\right] \tag{9}$$

and we can now use this to fit for both $\boldsymbol{\theta}$ and s.

5. A GAUSSIAN PROCESS AS A MODEL OF CORRELATED NOISE

One way to motivate the definition of a GP is to think of it as a model of correlated noise. The derive this, we start by re-writing Equation (8) as a matrix equation

$$\log p(\{y_n\}\,|\,\{\boldsymbol{x}_n,\,\sigma_n\},\,\boldsymbol{\theta}) \quad = \quad -\frac{1}{2}\,{\boldsymbol{r}_{\boldsymbol{\theta}}}^{\mathrm{T}}\,K^{-1}\,\boldsymbol{r}_{\boldsymbol{\theta}} - \frac{1}{2}\log\det K - \frac{N}{2}\,\log(2\,\pi) \tag{10}$$

where $\boldsymbol{r}_{\boldsymbol{\theta}}$ is the residual vector

$${\boldsymbol{r}_{\boldsymbol{\theta}}}^{\mathrm{T}} = \begin{pmatrix} y_1 - f(\boldsymbol{x}_1;\,\boldsymbol{\theta}) & \cdots & y_N - f(\boldsymbol{x}_N;\,\boldsymbol{\theta}) \end{pmatrix} \tag{11}$$

and K is the "covariance matrix" that is, in this case, diagonal

$$K = \begin{pmatrix} {\sigma_1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & {\sigma_N}^2 \end{pmatrix} \quad . \tag{12}$$

Noting that, for a diagonal matrix like this K, the inverse is

$$K^{-1} = \begin{pmatrix} 1/{\sigma_1}^2 & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & 1/{\sigma_N}^2 \end{pmatrix} \quad . \tag{13}$$

and the log-determinant is

$$\begin{aligned} \log\det K \quad &= \quad \log\prod_{n=1}^{N}{\sigma_n}^2 && (14)\\ &= \quad \sum_{n=1}^{N}\log{\sigma_n}^2 \quad , && (15) \end{aligned}$$

we can see that Equation (8) and Equation (10) are equivalent.

You might remember that, in the previous section, we assumed that the data points are independent. This assumption is expressed by the fact that the covariance matrix in Equation (12) is diagonal – all the off-diagonal elements are zero. In order to take correlated noise into account, we introduce non-zero off-diagonal elements in the covariance K. The n, m-th entry in the $N \times N$ matrix K quantifies the covariance between data points n and m. If we have some method of *estimating* this covariance *a priori* – much like how we often assume that we can estimate the diagonal elements of this matrix – we can fill in this matrix K and evaluate Equation (10) with this new, dense K. However, it is often hard to estimate these covariances reliably and we will, instead, fit for them. In practice, we probably don't want to add the $N\,(N-3)/2 \sim N^2$ parameters that would be needed to fit for each entry in this matrix directly. Instead, we parameterize this covariance using a functional form where the n, m-th element of the matrix is given by

$$K_{n,m} \quad = \quad {\sigma_n}^2\,\delta_{n,m} + k(\boldsymbol{x}_n,\,\boldsymbol{x}_m;\,\boldsymbol{\alpha}) \tag{16}$$

where $\delta_{n,m}$ is the Kronecker delta, and $k(\boldsymbol{x}_n,\,\boldsymbol{x}_m;\,\boldsymbol{\alpha})$ is a function – parameterized by $\boldsymbol{\alpha}$ – that captures the covariance between the data points $\boldsymbol{x}_n$ and $\boldsymbol{x}_m$. This function $k(\boldsymbol{x}_n,\,\boldsymbol{x}_m;\,\boldsymbol{\alpha})$ goes by a few names in the literature, the most common of which are "covariance function" or "kernel function". Throughout this paper, we will refer to this function as the covariance function. To indicate that the covariance matrix is now parameterized by $\boldsymbol{\alpha}$, we will typeset it as $K_{\boldsymbol{\alpha}}$ and the log-likelihood function becomes

$$\log p(\{y_n\}\,|\,\{\boldsymbol{x}_n,\,\sigma_n\},\,\boldsymbol{\theta},\,\boldsymbol{\alpha}) \quad = \quad -\frac{1}{2}\,{\boldsymbol{r}_{\boldsymbol{\theta}}}^{\mathrm{T}}\,{K_{\boldsymbol{\alpha}}}^{-1}\,\boldsymbol{r}_{\boldsymbol{\theta}} - \frac{1}{2}\,\log\det K_{\boldsymbol{\alpha}} - \frac{N}{2}\,\log(2\,\pi) \tag{17}$$

It is worth noting that the case of independent data points (Equation 10) is a special (restrictive) case of Equation (17) where

$$k(\boldsymbol{x}_n,\,\boldsymbol{x}_m;\,\boldsymbol{\alpha}) \quad = \quad 0 \quad . \tag{18}$$

In the case of GPs, we relax this assumption and choose a more flexible covariance function that approximates the real covariance structure in the data generation process. If we fit this GP model to a dataset where the data are actually independent and the diagonal variances ${\sigma_n}^2$ are correctly estimated, the second term in Equation (17) will drive the covariance function to zero and correctly capture the fact that the data are independent.

One of the touchiest subjects in GP modeling is the choice of covariance function and we will return to a detailed discussion of this in *DFM*: SOME SECTION, but first, let's consider a concrete example.

6. KERNEL FUNCTIONS

Any kernel function used for GP modeling must be positive semi-definite. If this is not satisfied, the covariance matrix will not be invertible and the determinant will be zero or negative, giving a likelihood of zero. This requirement can also be written informally as

$$\int k(\boldsymbol{x},\,\boldsymbol{x}')\,f(\boldsymbol{x})\,f(\boldsymbol{x}')\,p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,p(\boldsymbol{x}')\,\mathrm{d}\boldsymbol{x}' \ge 0 \tag{19}$$

for all finite functions $f(\boldsymbol{x})$. This property is generally hard to prove for a new kernel function so the standard practice is to use sums and products of common functions that have been proved to be valid. Most common functions are discussed in Chapter 4 of Rasmussen & Williams (2006). Just like any model selection problem, there are few different methods of choosing the kernel and some of these might be better than others depending on the goals of the user. Sometimes, the choice can be motivated using physics and the parameters of the model can then be interpreted in this same framework. In other cases, the GP model might just be acting as an effective model. For these situations, it is generally useful to demonstrate quantitatively that the results of your inference are not sensitive to your choice of kernel. It is also possible to compare different kernel functions using standard model comparison techniques like Bayes factors, cross validation, or information criteria.

- an example: linear fit - prediction: interpolation and extrapolation - kernel choice - outliers - examples: exoplanet transit fitting, radial velocity fitting, emulation, - practical considerations: scaling,

6.1. *Code availability*

Alongside this paper, we have released a well-tested and documented open source software package that implements the method and all of the examples discussed in these pages. This software is available on GitHub `https://github.com/dfm/george`[1] and Zenodo *DFM*: ADD ZENODO, and it is made available under the MIT license.

It is a pleasure to thank for helpful discussions informing the ideas and code presented here.

This work was performed in part under contract with the Jet Propulsion Laboratory (JPL) funded by NASA through the Sagan Fellowship Program executed by the NASA Exoplanet Science Institute.

Facility: Kepler

Software:

REFERENCES

[1] This version of the paper was generated with git commit `9120d69` (2017-11-30).

Rasmussen, C. E., & Williams, K. I. 2006, Gaussian Processes for Machine Learning (MIT Press)