Gaussian processes:
the next step in exoplanet data analysis

Suzanne Aigrain (University of Oxford)
Neale Gibson, Tom Evans, Amy McQuillan
Steve Roberts, Steve Reece, Mike Osborne

... let the data speak
Gaussian processes: the next step in exoplanet data analysis

Suzanne Aigrain (University of Oxford)
Neale Gibson, Tom Evans, Amy McQuillan
Steve Roberts, Steve Reece, Mike Osborne

Figure 1. The 'raw' HD u897ww NICMOS dataset used as an example for our Gaussian process model. Left: Raw light curves of HD u897ww for each of the u8 wavelength channels. Right: The optical step parameters extracted from the spectra plotted as a time series. These are used as the input parameters for our GP model. The red lines represent a GP regression on the input parameters used to remove the noise and test how this affects the GP model.

3.2 Type-II maximum likelihood
As mentioned in Sect., a useful approximation is Type-II maximum likelihood. First, the log posterior is maximised with respect to the hyperparameters and variable mean function parameters using a Nelder-Mead simplex algorithm (see e.g., Press et al. 1992). The transit model is the same as that used in GPAss, which uses Mandel & Agol (2002) models calculated assuming quadratic limb darkening and a circular orbit. All non-variables parameters were fixed to the values given in Pont et al. (2006), except for the limb darkening parameters, which were calculated for each wavelength channel in GPAss. Sing (2003). The only variable mean function variable parameters are the planet-to-star radius ratio and two parameters that govern a linear baseline model: an out-of-transit flux and time gradient.

An example of the predictive distributions found using type-II maximum likelihood is shown in Fig. for four of the wavelength channels. In this example, only orbits t and w are used to determine the parameters and hyperparameters of the GP for 'train' the GP and are shown by the red points. Orbit v (green points) was not used in the training set. Predictive distributions were calculated for orbits t–w and are shown by the grey regions, which plot the ±σ confidence intervals. The predictive distribution is a good fit to orbit v, showing that our GP model is effective at modelling the instrumental systematics. The systematics model will of course be even more constrained than in this example as we use orbits t–w to simultaneously infer parameters of the GP and transit function.

Now that all parameters and hyperparameters are optimised with respect to the posterior distribution, the hyperparameters are held fixed. This means the inverse covariance matrix and log determinant used to evaluate the log posterior are also fixed, and need calculated only once. An MCMC is used to marginalise over the remaining parameters of interest, in this case the planet-to-star radius ratio.
A Gaussian process in a nutshell

\[
P (t \mid \mathbf{X}, \theta, \phi) = \mathcal{N} [m(\mathbf{X}, \phi), \mathbf{K}]\]
A Gaussian process in a nutshell

The function $P(t | X, \theta, \phi)$ represents the likelihood distribution of the output vector (e.g. fluxes) given the input matrix (e.g. time, pointing, detector temperature...) and the (hyper-)parameters. It is a multivariate Gaussian distribution $\mathcal{N} [m(X, \phi), K]$ where $m(X, \phi)$ is the mean function and $K$ is the covariance matrix.
A Gaussian process in a nutshell

\[
P(t | X, \theta, \phi) = \mathcal{N}[m(X, \phi), K]
\]

where the covariance matrix \( K \)

\[
P \sim \mathcal{N}[\mathbf{0}, \mathbf{K}]
\]

output vector (e.g. fluxes)  \( \mathbb{X} \)

input matrix (e.g. time, pointing, detector temperature...)

likelihood

\( m(X, \phi) \)

mean function

\( \theta \)

(hyper-)parameters

\( \phi \)

(hyper-)parameters

\( \mathbf{K} \)

covariance matrix

\( \mathcal{N} \)

multivariate Gaussian
A Gaussian process in a nutshell

\[
P(t | X, \theta, \phi) = \mathcal{N}[m(x, \phi), K]
\]

- **Likelihood**: \(P(t | X, \theta, \phi)\)
- **Output vector (e.g. fluxes)**
- **Input matrix (e.g. time, pointing, detector temperature...)**
- **Mean function**: \(m(x, \phi)\)
- **Multivariate Gaussian**
- **Covariance matrix**: \(K_{nm} \equiv \text{cov} [x_n, x_m] = k(x_n, x_m, \theta)\)
- **Kernel function**: \(k(x_n, x_m, \theta)\)
- **(Hyper-)parameters**: \(\theta, \phi\)
A Gaussian process in a nutshell

\[ P(t \mid X, \theta, \phi) = \mathcal{N}[m(X, \phi), K] \]

- **Likelihood**: \( P(t \mid X, \theta, \phi) \)
- **Output vector (e.g. fluxes)**: \( t \)
- **Input matrix (e.g. time, pointing, detector temperature...)**: \( X \)
- **Mean function**: \( m(X, \phi) \)
- **Mean function parameters**: \( \phi \)
- **Covariance matrix**: \( K \)
- **Covariance function**: \( k(x_n, x_m, \theta) \)
- **Kernel function**: \( k \)
- **Kernel parameters**: \( \theta \)
Gaussian process regression
Gaussian process regression

a priori knowledge
Gaussian process regression

- a priori knowledge
  - chose kernel & mean functions
  - guess initial hyper-parameters
Gaussian process regression

a priori knowledge

chose kernel & mean functions

guess initial hyper-parameters

prior probability distribution over functions
Gaussian process regression

- a priori knowledge
  - chose kernel & mean functions
  - guess initial hyper-parameters
  - prior probability distribution over functions

get some data
Gaussian process regression

- Get some data
- Choose kernel & mean functions
- Guess initial hyper-parameters
- Prior probability distribution over functions
- Condition the GP

A priori knowledge
Gaussian process regression

- a priori knowledge
- chose kernel & mean functions
- guess initial hyper-parameters
- prior probability distribution over functions
- condition the GP
- predictive (posterior) distribution
- get some data
Gaussian process regression

- Get some data
- Choose kernel & mean functions
- Guess initial hyper-parameters
- Prior probability distribution over functions
- Condition the GP
- Predictive (posterior) distribution
- Marginal likelihood

Prior knowledge
Gaussian process regression

- a priori knowledge
  - chose kernel & mean functions
  - guess initial hyper-parameters
- get some data
  - condition the GP
  - prior probability distribution over functions
- posterior distribution over hyper-parameters
  - marginal likelihood
  - predictive (posterior) distribution
  - marginal likelihood
Gaussian process regression

- A priori knowledge
  - Chose kernel & mean functions
  - Guess initial hyper-parameters
- Prior probability distribution over functions
- Get some data
  - Condition the GP
- Marginal likelihood
  - Posterior distribution over hyper-parameters
- Predictive (posterior) distribution
  - Interpolate (predict)
A very simple example

$$k_{SE}(t, t') = A^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right)$$

from Gibson et al. (submitted)
A very simple example

\[ k_{SE}(t, t') = A^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right) \]

from Gibson et al. (submitted)
A very simple example

\[ k_{SE}(t, t') = A^2 \exp\left(-\frac{(t - t')^2}{2l^2}\right) \]

from Gibson et al. (submitted)
Example application 1:
instrumental systematics in transmission spectra

See Neale Gibson’s talk
Example application 2: modelling HD 189733b’s OOT light curve

5 Modelling the HD 30.955 light curve

5.1 Choice of model

We modelled the HD 30.955 data with all the individual kernels detailed in section 6 as well as a number of different combinations. We initially trained the GP on all three datasets but the MOST data had virtually no effect on the results mainly because they do not overlap with any of the transit observations of interest, so we did not use the MOST data for the final calculations. We used LOO-CV to compare different kernels, but we also performed a careful visual examination of the results, generative predictive distributions over the entire monitoring period and over individual seasons. Interestingly, we note that, in practice, LOO-CV systematically favours the simplest kernel which appears to give good results by eye.

We experimented with various combinations of SE and RQ kernels to form quasiperiodic models, and found that using an RQ kernel for the evolutionary term significantly improves the LOO-CV pseudo-likelihood relative to an SE-based evolutionary term. It makes very little difference to the mean of the predictive distribution when the observations are well-sampled, but it vastly increases the predictive power of the GP away from observations, as it allows for a small amount of covariance on long time-scales, even if the dominant evolution timescale is relatively short. On the other hand, we were not able to distinguish between SE and RQ kernels for the periodic term (the two give virtually identical best-fit marginal likelihoods and pseudo-likelihoods) and therefore opted for the simpler SE kernel.

To describe the observational noise, we experimented with a separate, additive SE kernel as well as a white noise term. However, we found that this did not significantly improve the marginal or pseudo-likelihood, and the best-fit length-scale was comparable to the typical interval between consecutive data points. We therefore reverted to a white noise term only. The final kernel was thus:

\[
k_{\text{QP,mixed}}(t, t') = A^2 \exp \left( -\frac{\sin^2[\pi(t - t')/P]}{2L^2} \right) \times \left( 1 + \frac{(t - t')^2}{2\alpha l^2} \right)^{-\alpha} + \sigma^2 I.
\]

The best-fit hyperparameters were \(A = 8.80\) mmag, \(P = 33.08\) days, \(L = 2.3\) days, \(\alpha = 2.45\), \(l = 39.02\) days, and \(\sigma = 4.33\) mmag. Our period estimate is in excellent agreement with Henry & Winn (2020). We note that very similar best-fit hyperparameters were obtained with the other kernels we tried (where those kernel shared equivalent hyperparameters). The relatively long periodic timescale \(L\) indicates that the variations are dominated by fairly large active regions. The evolutionary term has a relatively short timescale \(l\) (about 3.7 times the period) but a relatively shallow index \(\alpha\) which is consistent with the notion that individual active regions evolve relatively fast, but that there are preferentially active longitudes where active regions tend to be located as inferred from better-sampled long-duration CoRoT light curves for similarly active stars by Lanza et al.

Note that we should really treat the noise term separately for each dataset and the constant relative offsets or linear trends where applicable as parameters of the mean function. However, we doubt this would make a big difference to the results.
Example application 2: modelling HD 189733b’s OOT light curve

\[ k_{\text{QP,mixed}}(t, t') = A^2 \exp \left( -\frac{\sin^2[\pi(t - t')/P]}{2L^2} \right) \times \left( 1 + \frac{(t - t')^2}{2\alpha l^2} \right)^{-\alpha} + \sigma^2 \mathbf{I}. \]

predict brightness at times of transit observations
correct measured radius ratio for effect of un-occulted spots (Frederic Pont’s talk)
Example application 2: modelling HD 189733b’s OOT light curve

\[ k_{\text{QP,mixed}}(t, t') = A^2 \exp \left( -\frac{\sin^2[\pi(t - t')/P]}{2L^2} \right) \times \left( 1 + \frac{(t - t')^2}{2\alpha l^2} \right)^{-\alpha} + \sigma^2 I. \]

hyper-parameters contain information about rotation rate, spot distribution and evolution...

predict brightness at times of transit observations
correct measured radius ratio for effect of un-occulted spots (Frederic Pont’s talk)
Pros ...

- Rigorous error propagation
- Extremely versatile
- Built-in Ockam’s razor
- Joint modelling of arbitrary number of inputs (and outputs)
- Easy to combine with other techniques e.g. MCMC

... and cons

- Computationally intensive: $O(N^3)$
- ok up to $N \sim 1000$
- alternative: Variational Bayes (see Tom Evans’ poster)

Pros ...                           ... and cons
<table>
<thead>
<tr>
<th>Pros ...</th>
<th>... and cons</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Rigorous error propagation</td>
<td>• Computationally intensive: $O(N^3)$</td>
</tr>
<tr>
<td>• Extremely versatile</td>
<td>• ok up to $N \sim 1000$</td>
</tr>
<tr>
<td>• Built-in Ockam’s razor</td>
<td>• alternative: Variational Bayes (see Tom Evans’ poster)</td>
</tr>
<tr>
<td>• Joint modelling of arbitrary number of inputs (and outputs)</td>
<td></td>
</tr>
<tr>
<td>• Easy to combine with other techniques e.g. MCMC</td>
<td></td>
</tr>
</tbody>
</table>

Want to try?
Python GP module under development