A Gaussian process framework for modelling instrumental systematics: application to transmission spectroscopy

N. P. Gibson,[1]⋆ S. Aigrain,[1] S. Roberts,[2] T. M. Evans,[1] M. Osborne[2] and F. Pont[3]

[1]*Department of Physics, University of Oxford, Denys Wilkinson Building, Keble Road, Oxford OX1 3RH*
[2]*Robotics Research Group, Department of Engineering Science, University of Oxford, Oxford OX1 3PJ*
[3]*School of Physics, University of Exeter, Exeter EX4 4QL*

ABSTRACT

Transmission spectroscopy, which consists of measuring the wavelength-dependent absorption of starlight by a planet's atmosphere during a transit, is a powerful probe of atmospheric composition. However, the expected signal is typically orders of magnitude smaller than instrumental systematics and the results are crucially dependent on the treatment of the latter. In this paper, we propose a new method to infer transit parameters in the presence of systematic noise using Gaussian processes, a technique widely used in the machine learning community for Bayesian regression and classification problems. Our method makes use of auxiliary information about the state of the instrument, but does so in a non-parametric manner, without imposing a specific dependence of the systematics on the instrumental parameters, and naturally allows for the correlated nature of the noise. We give an example application of the method to archival NICMOS transmission spectroscopy of the hot Jupiter HD 189733, which goes some way towards reconciling the controversy surrounding this data set in the literature. Finally, we provide an appendix giving a general introduction to Gaussian processes for regression, in order to encourage their application to a wider range of problems.

Key words: methods: data analysis – techniques: spectroscopic – stars: individual: HD 189733 – planetary systems.

1 INTRODUCTION

Transiting exoplanetsΩoffer a unique opportunity to study the structures and atmospheres of planets outside our Solar system. When a planet passes in front of its parent star as we view it from Earth, it blocks a proportion of starlight which we can use to measure the planet's radius relative to its host. Transmission spectroscopy exploits the fact that the measured radius of a planet is wavelength dependent, as the effective size of a planet's occulting disc depends on the height in the atmosphere at which the planet becomes opaque to starlight, which in turn depends on the atomic and molecular species present in the atmosphere (e.g. Seager & Sasselov 2000; Brown 2001). Thus measuring planetary transits (i.e. the planet-to-star radius ratio) as a function of wavelength allows us to probe the makeup of a planet's atmosphere.

Observations of transmission spectra, in particular using the *Hubble Space Telescope* (*HST*) and *Spitzer Space Telescope* (*SST*) have provided some of the most detailed observations of exoplanet atmospheres to date (see e.g. Charbonneau et al. 2002, 2005;

Vidal-Madjar et al. 2003; Deming et al. 2005). However, extracting transmission spectra from these data sets is by no means trivial and has led to some discussion regarding the interpretation of these signals. Problems arise because the transmission signal is often dwarfed by instrumental systematics in the data sets. These are caused by imperfect observing conditions, such as changes in the pointing, detector temperature, optics and pixel-to-pixel sensitivity, and result in correlated noise in the light curves. This has led to the development of some novel techniques to correct and remove instrumental systematics, usually by modelling them as a deterministic function of auxiliary measurements from the data set (see e.g. Brown et al. 2001; Gilliland & Arribas 2003; Pont et al. 2007). Often, however, the choice of systematics model will significantly affect the transmission spectra, and therefore the detection of atmospheric species. Attempting to develop a more general and robust framework to infer transit parameters in the presence of instrumental systematics and correlated noise provides the motivation for this work.

The problem of correlated noise for observations of transiting exoplanets was first raised by Pont et al. (2006), who provided methods to estimate its effects on time-series analysis, and take it into account when inferring physical properties of exoplanets from

⋆ E-mail: Neale.Gibson@astro.ox.ac.uk

such data sets. Indeed, the problems associated with systematics appear in many different areas of exoplanet observations, including transit timing analysis (e.g. Gibson et al. 2009) and radial velocity measurements (e.g. Pont, Aigrain & Zucker 2011). Notably, Carter & Winn (2009) developed a fast and powerful wavelet-based method to model and remove time-correlated noise from transit light curves, without the need for additional inputs. Often, however, we have auxiliary information from the data, such as measurements of pointing drifts or changes in the detector and optical conditions, thought to be the cause of the instrumental systematics. This additional information should ideally be used to model systematic trends if available.

Instrumental systematics have proved to be particularly problematic for *HST*/NICMOS observations of transmission spectra (e.g. Swain, Vasisht & Tinetti 2008; Pont et al. 2009; Gibson, Pont & Aigrain 2011). Swain et al. (2008, hereafter SVT08) observed and analysed the transmission spectra of HD 189733b and interpreted the results as evidence for CH_4 and H_2O. Following previous *HST* analyses, the instrumental systematics in the light curves were modelled as a linear function of 'optical state parameters'. The optical state parameters are simply auxiliary measurements made directly from the spectra, such as the position, width and angle of the spectral trace, or other parameters reflecting the state of the detector and optics, such as the temperature and satellite orbital phase. Gibson et al. (2011, hereafter GPA11), reviewed the evidence for molecular species in this data set using similar methods to model the systematics. This study concluded that the transmission spectrum (and hence the detection of molecules) was too dependent not only on the functional form of the model, but also on the choice of which parameters to include. However, this was a rather unsatisfactory conclusion, as it did not provide an alternative method to determine the transmission spectrum, and we have since searched for more sophisticated techniques which can provide a robust interpretation of these data sets. This paper introduces the use of Gaussian process (GP) models to address this problem.

GP models are widely used in the machine learning community (Bishop 2006; Rasmussen & Williams 2006) for Bayesian regression and classification problems, and have recently been adopted in other areas of astrophysics (e.g. Way & Srivastava 2006; Mahabal et al. 2008; Way et al. 2009). Rather than imposing a deterministic model, GPs define a distribution over function space. This allows the instrumental systematics to be modelled in a *non-parametric* way. Using GPs, we can then marginalize out our ignorance of the systematics model and determine the posterior distribution of the parameters of interest, in this case the planet-to-star radius ratio. To put it another way, the form of the systematics model is inferred from the data themselves and does not have to be set a priori. As GPs are a Bayesian technique, they also automatically avoid the problem of overfitting. As we shall see later, through a sensible choice of prior distributions on our input variables, we may also determine which of the input variables are relevant to our data set.

This paper describes a GP model for inferring transit parameters in the presence of instrumental systematics and correlated noise, when additional parameters are measured from the data such as the optical state parameters mentioned above, but it can easily be generalized to time-correlated noise in the absence of such measurements. We provide a description of the GP model in Section 2, and its application to NICMOS transmission spectroscopy of HD 189733 in Section 3, followed by our conclusions in Section 4. We also provide an appendix giving a brief introduction to GPs with simple regression examples.

2 GAUSSIAN PROCESSES FOR TRANSMISSION SPECTROSCOPY

A GP is a non-parametric method for regression, used extensively for regression and classification problems in the machine learning community. A GP is defined as a collection of random variables, any finite number of which have a joint Gaussian distribution. A brief introduction to GPs is provided in the appendix, along with references for further reading.

Here, we use a GP to specify a non-parametric model of the instrumental systematics for a transit observation, along with a deterministic transit model to infer the transit parameters. We begin by defining the transmission spectroscopy data set, and briefly restating the 'standard' linear model used in previous analyses, before introducing our GP model.

2.1 Transmission spectroscopy data sets

Reduced transmission spectroscopy data sets consist of multiple transit light curves as a function of time. Here, we consider each wavelength channel independently and store the N flux measurements in the vector $\boldsymbol{f} = (f_1, \ldots, f_N)^T$ observed at time $\boldsymbol{t} = (t_1, \ldots, t_N)^T$. Often, additional parameters (or optical state parameters) are measured which describe the behaviour of the instrument as a function of time. These are given by the vector $\boldsymbol{x}_n = (x_{n,1}, \ldots, x_{n,K})^T$ at time n, where K is the number of additional parameters. These are collected in the $N \times K$ matrix $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)^T$, and hereafter are called the *input parameters*. As each light curve is treated independently, our model can easily be extended to single passband photometric transit observations, using additional measured inputs, or simply using time as the only input in order to specify time-correlated noise.

2.2 Linear baseline model

Often, the baseline flux, which is how the light curve would behave in the absence of a transit due to instrumental systematics, is modelled as a linear combination of the input parameters:

$$\boldsymbol{f} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ is a vector containing the coefficients of each input parameter and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_N)^T$ are the residuals, assumed to be independent Gaussian noise. The best-fitting coefficients are then found by linear least-squares for the out-of-transit data, given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \boldsymbol{f}.$$

The baseline flux is constructed for all data points using the best-fitting coefficients, and the transit light curve is divided through by the baseline model to remove the instrumental systematics. The 'decorrelated' light curve is then fitted with a transit model to determine the parameters of interest, using varying methods to propagate the uncertainties in the systematics model.

2.3 Gaussian process model

Rather than imposing a simple, deterministic function on the instrumental model, we model each wavelength channel as a GP with a transit mean function:

$$f(t, \boldsymbol{x}) \sim \mathcal{GP}\left(T(t, \boldsymbol{\phi}), \boldsymbol{\Sigma}(\boldsymbol{x}, \boldsymbol{\theta})\right).$$

Here, T is the transit function depending on t and the transit parameters $\boldsymbol{\phi}$, modelled using the analytic equations of Mandel & Agol (2002). $\boldsymbol{\Sigma}$ is the covariance matrix, which is a function of $\boldsymbol{x}$ and parameters $\boldsymbol{\theta}$, referred to as *hyperparameters* of the GP.[1] The definition of a GP means that the joint probability distribution for the finite set of observations $\boldsymbol{f}$ is a multivariate Gaussian distributed about a transit function T, with covariance $\boldsymbol{\Sigma}$, given by

$$p(\boldsymbol{f}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \mathcal{N}(T(\boldsymbol{t}, \boldsymbol{\phi}), \boldsymbol{\Sigma}(\mathbf{X}, \boldsymbol{\theta})). \tag{1}$$

The instrumental systematics and noise are specified fully by the covariance matrix, which in turn is specified by a covariance function or *kernel*. As the instrumental systematics are a function of the input parameters, we adopt the kernel:

$$\boldsymbol{\Sigma}_{nm} = k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \xi \exp\left[-\sum_{i=1}^{K} \eta_i (x_{n,i} - x_{m,i})^2\right] + \delta_{nm}\sigma^2.$$

This returns a scalar covariance for each pair of inputs, defining each element of the covariance matrix, $\boldsymbol{\Sigma}_{nm}$. ξ is a hyperparameter that specifies the maximum covariance and $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_K)^T$ are the inverse-scale parameters for each input vector. To incorporate white noise, we add a variance term σ^2 to the diagonal of the covariance matrix, where δ is the Kronecker delta function. We hereafter use the notation $\boldsymbol{\theta} = (\xi, \boldsymbol{\eta}^T, \sigma^2)^T$ to represent the covariance hyperparameter vector.

The interpretation of this kernel is straightforward; data points that are nearby each other in input space are highly correlated, and data points that are far from each other in input space are relatively uncorrelated. This kernel therefore describes a smooth function of the input parameters, with the addition of white noise. The lengthscales determine how important each vector in input space is to determine the 'closeness' of two data points. It therefore enables *Automatic Relevance Determination* (ARD; Neal 1996, also known as shrinkage or ridge regression), where the η_i associated with each input parameter becomes small when the parameter is of little relevance to explain the data set. When $\eta_i \to 0$ the parameter no longer influences the covariance matrix, and does not contribute to the model. In other words, the instrumental systematics model assumes a similar value when all *relevant* input parameters are close. This kernel is suitable for application to transmission spectroscopy data sets, as squared exponential kernels are suitable for processes with a dominant length-scale in each input dimension. In transmission spectroscopy we are interested in instrumental systematics within a narrow range of time-scales, commensurate with the duration of the transit, therefore input parameters that vary in these time-scales. Frequencies higher than the sampling can be treated as white noise and frequencies lower can be modelled using a simple baseline function incorporated into the mean function.

There are other kernel functions available that could be used to model such behaviour, in particular the Matérn class. This can be seen as a 'rougher' generalization of the squared exponential kernel, that results in less smooth functions of the inputs. However, this kernel would require an extra hyperparameter for each input parameter, and would likely prohibit marginalization over all the hyperparameters, and perhaps allow too much freedom in the instrument model. As the squared exponential kernel defines a prior distribution over the smooth function of the inputs, we can further justify the use of the this kernel by noting that this incorporates all common functions typically used in instrument models to account

for systematics; for example, models that are linear, quadratic and higher order functions of the inputs, and thus represent a generalization of previous work. Indeed, this kernel reflects our prior beliefs about the properties of the data. The systematics are dominated by pixel-to-pixel sensitivity variations, which we would expect to smoothly vary with the inputs, particularly when we consider that each wavelength channel is binned over many pixels. None the less, readers should be aware that there are many covariance kernels available with different properties. The choice of kernel is an important consideration in applying GP models, but the use of a non-parametric systematics model allows much greater flexibility than imposing a deterministic parametric form.

We have now defined a distribution over smooth functions of the input parameters to model the instrumental systematics. The functional form that describes the baseline function may now be inferred from the data themselves. As the joint probability of the observations defined by the GP is multivariate Gaussian (equation 1), the log marginal likelihood function is written as

$$\log \mathcal{L}(\boldsymbol{r}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) = -\frac{1}{2}\boldsymbol{r}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{N}{2}\log(2\pi), \tag{2}$$

where $\mathbf{r}$ is a vector containing the residuals from the mean function:

$$\boldsymbol{r} = \boldsymbol{f} - T(\boldsymbol{t}, \boldsymbol{\phi}).$$

This defines the likelihood as a function of the mean function parameters $\boldsymbol{\phi}$, which determine the underlying transit light curve and the hyperparameters $\boldsymbol{\theta}$, which determine the behaviour of the baseline function and the noise.

In practice, it is convenient to place explicit priors on the maximum covariance hyperparameter and the scalelength hyperparameters (or *hyperpriors*) to encourage their values towards zero if they are truly irrelevant to explain the data. In our model we use gamma hyperpriors (with shape parameter of unity) of the form

$$p(x) = \frac{1}{l}\exp(-x/l),$$

for $x \geq 0$ and l is the length-scale of the hyperprior. The log posterior distribution of the transit parameters and hyperparameters is proportional to the sum of the log marginal likelihood and log hyperpriors; hence, we add the log hyperpriors to equation (2) as follows:

$$\log \mathcal{P}(\boldsymbol{\theta}, \boldsymbol{\phi}|\boldsymbol{f}, \mathbf{X}) = \log \mathcal{L}(\boldsymbol{r}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) - \frac{\xi}{l_\xi} - \sum_{i=1}^{K}\left(\frac{\eta_i}{l_i}\right) + C, \tag{3}$$

for $\xi, \eta_i \geq 0$, l are the scalelengths of the hyperpriors, and C represents additional constant terms. Note that we do not explicitly add hyperpriors for the transit parameters or variance hyperparameter (implying uniform and improper hyperpriors). The hyperprior length-scales are set to a large value to ensure that the hyperpriors are non-informative. The challenge is now to infer the transit parameters and hyperparameters from the posterior probability distribution.

2.4 Inferring transit parameters

Now we have specified the log posterior as a function of the hyperparameters $\boldsymbol{\theta}$ and the transit parameters $\boldsymbol{\phi}$; it is straightforward (in theory) to infer the planet-to-star radius ratio ρ from the data. For example, to find the maximum posterior solution, the log posterior given by equation (3) is optimized with respect to the transit parameters and hyperparameters. In a fully Bayesian treatment, we

---

[1] Strictly speaking, the transit parameters are also hyperparameters of the GP, but here we refer to them as the transit parameters for simplicity.

should obtain the posterior distribution for each parameter of interest by marginalizing over all the other parameters of our model, i.e. all other mean-function parameters and hyperparameters. In practice, this can be done using Monte Carlo Markov Chains (MCMC) methods, which are widely used in the exoplanet community to explore the joint posterior probability distribution of multivariate models (see e.g. Holman et al. 2006; Collier Cameron et al. 2007; Winn et al. 2008). However, each time the hyperparameters are changed (at each step in a chain), we must re-calculate the covariance matrix and its inverse in order to evaluate the log posterior. This scales badly with the length of the input time series, requiring $\mathcal{O}(N^3)$ operations. This often makes a full marginalization over all hyperparameters intractable for large data sets, and at best slow to compute. It is therefore worth introducing an approximation known as type-II maximum likelihood.

Type-II maximum likelihood involves fixing all the hyperparameters at their maximum likelihood values and marginalizing over the remaining parameters of interest. In this case we will find the maximum posterior solution, and marginalize over the remaining transit parameters, although we still refer to this procedure as type-II maximum likelihood as in the literature. This is a valid approximation when the posterior distribution is sharply peaked at its maximum around the covariance hyperparameters. First, the log posterior function is optimized with respect to the hyperparameters θ *and* transit parameters ϕ, in order to determine the maximum posterior solution. This procedure still requires the covariance matrix is inverted, but much fewer times than for a full marginalization over all parameters. The hyperparameters θ are then held fixed at their maximum posterior values, and to obtain the (approximate) posterior distributions for each parameter of interest, one then marginalizes over the remaining transit parameters, which does not require a matrix inversion at each step. One caveat is that we do not know if the posterior distribution is sharply peaked until we marginalize over all parameters. Consequently, it is always preferable to marginalize over all hyperparameters of the GP when possible, and indeed, this is the final approach used in this work. However, for many sets of data using the same kernel function, we might confirm the effectiveness of type-II maximum likelihood for a particular subset and apply it to the remainder. Nevertheless, in many cases fixing the covariance hyperparameters might be a better approach than imposing a deterministic model of the systematics, but such approximations must be used with caution. In the following section, we apply our GP model to NICMOS transmission spectroscopy of HD 189733, and use both type-II maximum likelihood and a full marginalization over all transit parameters and hyperparameters to obtain the transmission spectrum. Of course, results from the full marginalization are preferred, but the comparison to type-II maximum likelihood is a worthwhile exercise.

3 APPLICATION TO NICMOS OBSERVATIONS OF HD 189733

3.1 The data set

Here we provide a summary of the *HST*/NICMOS observations of HD 189733, used as a test case for our GP model. These were originally analysed by SVT08 and re-analysed by GPA11, and the data set used in this paper is the same as that used in GPA11. For a more detailed account of the observations and data reduction methods, we refer the readers to the aforementioned papers.

A transit of HD 189733 was monitored on 2007 May 25, using the G206 grism covering the wavelength range 1.4–2.5 μm. As

HD 189733 is not in the continuous viewing zone of the *HST*, the light curve was observed over 5 half-orbits, although orbit 1 is excluded as it shows much larger systematic effects attributed to spacecraft 'settling'. The light curves for 18 wavelength channels were extracted for 638 spectra, along with optical state parameters. These are the positions of the spectral trace along the dispersion axis (ΔX), the average position of the spectral trace along the cross-dispersion axis (ΔY), the angle the spectral trace makes with the *x*-axis (ψ) and the average width of the spectral trace (W). The temperature[2] (T) and orbital phase (ϕ_H) were also determined for each image.

In total, our data set consists of a time series of 519 flux measurements (for orbits 2–5 only) in 18 wavelength channels and time series of six optical state parameters. The 18 light curves and the optical state parameters are shown in Fig. 1. The light curves show clear instrumental systematics, and these must be accounted for when measuring the transit depth at each wavelength. We have strong reasons to suspect that the systematics in the light curves are related to the optical state parameters in some way. This is a reasonable assumption, as changes in the position, angle and width of the spectrum can obviously have an effect on the flux collected in each wavelength bin, as can the temperature of the detector. However, it is not possible to construct a reliable deterministic physical model which relates these parameters to the baseline flux, as it depends on each individual pixel's sensitivity and response to temperature, as well as other complex changes in the optics.

For each wavelength channel, the GP model outlined in Section 2 was applied, where f contains the 519 flux measurements and $\mathbf{X}$ contain the input parameters, where $x_n = (\Delta X_n, \Delta Y_n, W_n, \psi_n, T_n, \phi_n)^T$. To test that the model was not influenced by the noise of the input parameters, a GP regression model with zero mean function was fitted to each input parameter using a squared exponential kernel, with time as the only input. The hyperparameters were optimized with respect to the likelihood function, and the regression is shown by the red lines in Fig. 1. The inference described in the following two sections was repeated using these de-noised input parameters.[3] We found that this had little influence on the results, and report the results using the noisy input parameters. We also checked that the choice of hyperprior length-scale had little effect on the results, by setting them to large values, and repeating the procedure with varying length-scales ensuring that the transmission spectra were not significantly altered.

3.2 Type-II maximum likelihood

As mentioned in Section 2.4, a useful approximation is Type-II maximum likelihood. First, the log posterior is maximized 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 GPA11, which uses Mandel & Agol (2002) models calculated assuming quadratic limb darkening and a circular orbit. All non-variable parameters were fixed to the values given in Pont et al. (2008), except for the limb darkening parameters, which were calculated for each wavelength channel (GPA11; Sing 2010). The only variable mean function

---

[2] No direct measurement of the temperature of the detector exists, rather a the temperature of the NIC1 mounting cup is used as a proxy measurement, see GPA11 for details.

[3] We did not use the GP smoothed parameters for T, as we were not confident that the fit reliably represented the underlying structure.
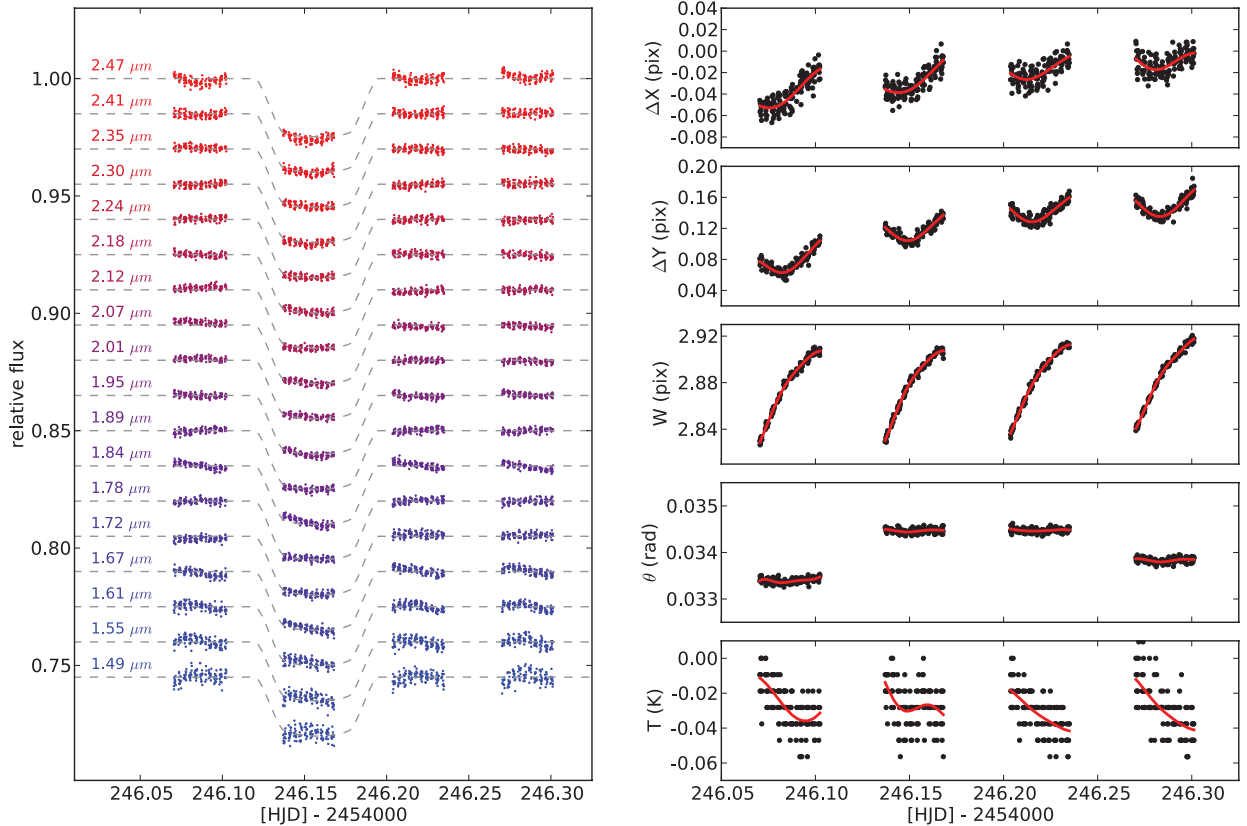
Figure 1. The 'raw' HD 189733 NICMOS data set used as an example for our Gaussian process model. Left: raw light curves of HD 189733 for each of the 18 wavelength channels, from 2.50 μm to 1.48 μm top to bottom. Right: the optical state parameters extracted from the spectra plotted as a time series. 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.

variable parameters are the planet-to-star radius ratio, and two parameters that govern a linear baseline model; an out-of-transit flux f_{oot} and (time) gradient T_{grad}.

An example of the predictive distributions found using type-II maximum likelihood is shown in Fig. 2 for four of the wavelength channels. In this example, only orbits 2, 3 and 5 are used to determine the parameters and hyperparameters of the GP (or 'train' the GP) and are shown by the red points.[4] Orbit 4 (green points) was not used in the training set. Predictive distributions were calculated for orbits 2–5, and are shown by the grey regions, which plot the 1σ and 2σ confidence intervals. The predictive distribution is a good fit to orbit 4, showing that our GP model is effective at modelling the instrumental systematics. The final systematics model will of course be even more constrained than in this example, as we use orbits 2–5 to simultaneously infer parameters of the GP and transit function, in particular since orbit 4 has the most similar input parameters to the in-transit orbit (Fig. 1).

Now that all parameters and hyperparameters are optimized 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 to be calculated only once. An MCMC is used to marginalize over the remaining parameters of interest, in this case the planet-to-star

radius ratio and the linear baseline coefficients. Our MCMC implementation is adapted from Gibson et al. (2010), and uses the log posterior function given by equation (3) as the merit function.

For each of the 18 wavelength channels, four separate chains of length 20 000 were computed, with the starting parameter set initialized to the maximum posterior distribution parameters with small perturbations applied to the variable parameters. During the burn-in phase (first 20 per cent of each chain), the global and relative stepsizes are adapted so that ∼25 per cent of proposal parameters sets are accepted. After burn-in is complete the chains have converged and now sample from the posterior probability distribution. These samples are used to estimate the marginalized posterior distributions for the planet-to-star radius ratio. For each wavelength channel, convergence was confirmed for the four chains by calculating the Gelman–Rubin (GR) statistic (Gelman & Rubin 1992), and in all cases it was within 1 per cent of unity, a good sign of mixing and convergence.

The resulting transmission spectrum consists of the planet-to-star radius ratio for each wavelength channel and is shown in the left-hand plot of Fig. 3. The dashed grey line shows the value of ρ for the 'white' light curve (i.e. the light curve given by the sum of all wavelength channels) reported in GPA11.

3.3 Marginalization over the hyperparameters

Despite the need to invert a matrix at every calculation of the log likelihood function, it is still feasible to marginalize over all the parameters and hyperparameters of the GP model for the HD 189733

---

[4] We used the smoothed hyperparameters here for aesthetic purposes. Using the noisy hyperparameters would simply result in noisier 1 and 2σ limits, but with similar structure.
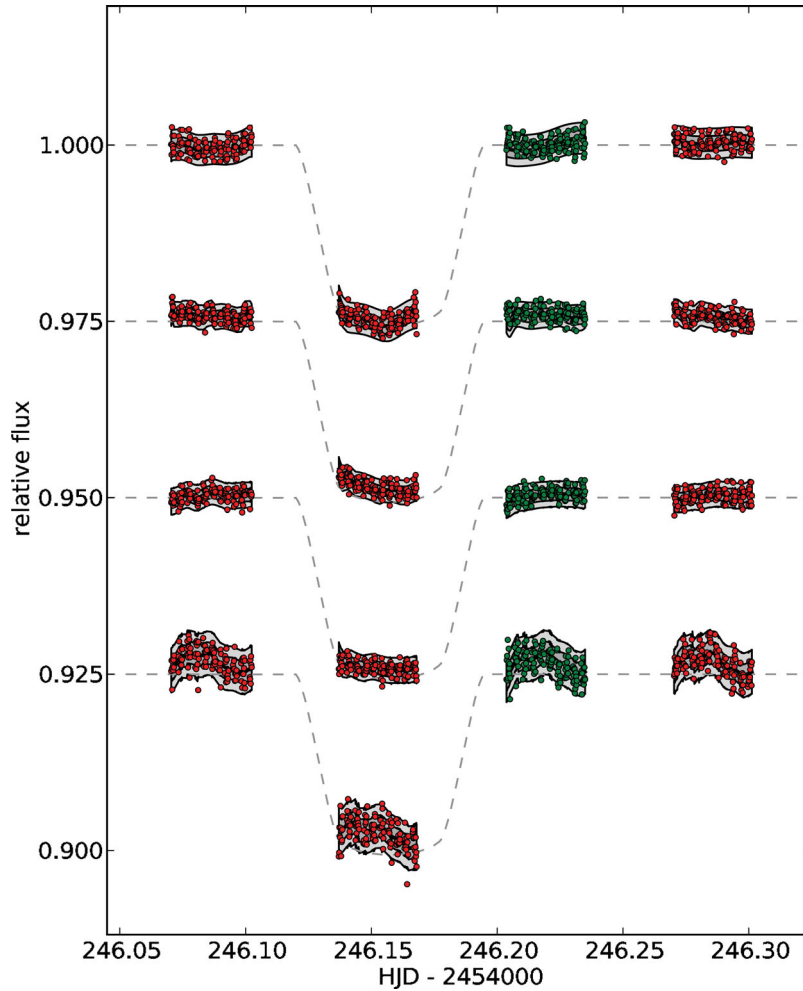
Figure 2. Examples of type-II maximum likelihood regression for four light curves of HD 189733. The red points show those data used in the training process (orbits 2, 3 and 5), and the green points show the remaining data (orbit 4). The predictive distributions are shown by the grey regions, which mark the 1σ and 2σ confidence intervals. The GP model predictions are consistent with the measured data, indicating that our GP model is an effective model of the instrumental systematics. In practice all four orbits are used to model the light curve and systematics, and infer the planet-to-star radius ratio, hence the systematics model will be even better constrained than shown here.
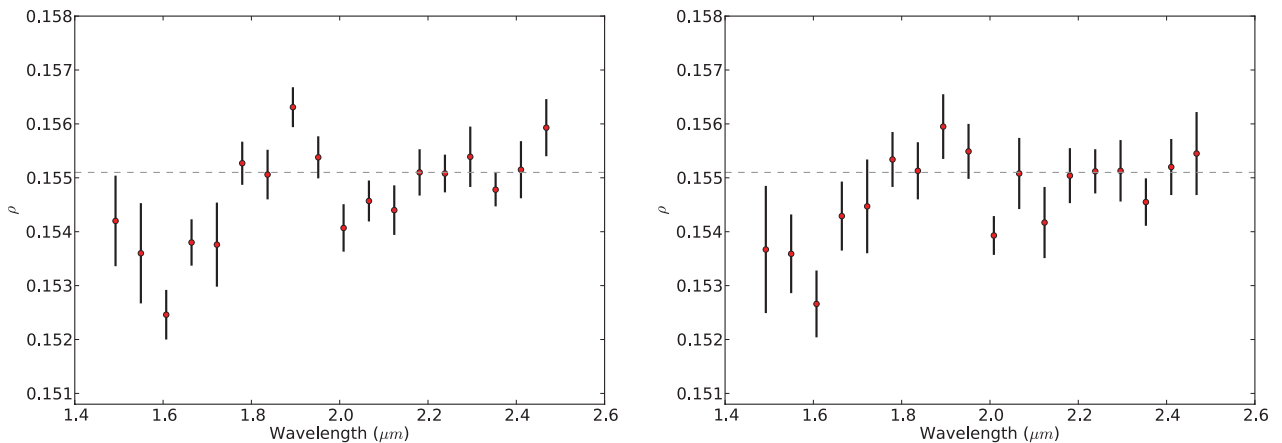

Figure 3. NICMOS transmission spectra of HD 189733 produced from the GP model. The horizontal dashed line represents the planet-to-star radius ratio determined from the white light curve. Left: using the maximum likelihood type-II approximation (see Section 3.2). Right: obtained by marginalizing over all other parameters and hyperparameters (see Section 3.3).

data set using MCMC. The same MCMC routine as described in the previous section is used, this time allowing all the hyperparameters of the GP to vary along with the mean function parameters. The chains were initialized from the maximum posterior values, again with a small perturbation applied to each variable parameter.

For each wavelength channel, four chains of length 150 000 were calculated. As each evaluation of the posterior probability required a matrix inversion, each chain took about ∼2.5 h to run on a standard desktop computer. In order to run tests in a reasonable time frame, all 72 chains were run in parallel using the Oxford e-Research Centre supercomputing cluster. The chains were again checked for convergence using the GR statistic. For the majority of parameters, convergence within 1 per cent of unity was achieved. However, some parameters did not converge as efficiently and the GR statistic

was as much as ∼5 per cent from unity. This can be understood in terms of the degeneracy between input parameters. Given the inputs show similar structure, the same overall systematics model may be described using different combinations of scalelength parameters. Also, input parameters that have small η values are not strongly constrained by the data (or priors) and effectively exhibit random walk behaviour. However, importantly the main parameters of interest ρ did converge for all wavelength channels. Correlation plots for two of the wavelength channels are shown in Fig. 4.

For each of the wavelength channels, different input parameters proved to be relevant to describe the instrument model as inferred from the values of the η parameters. The phase is a consistently relevant input parameter for the majority of wavelength channels, and the temperature is not particularly relevant for any channels,
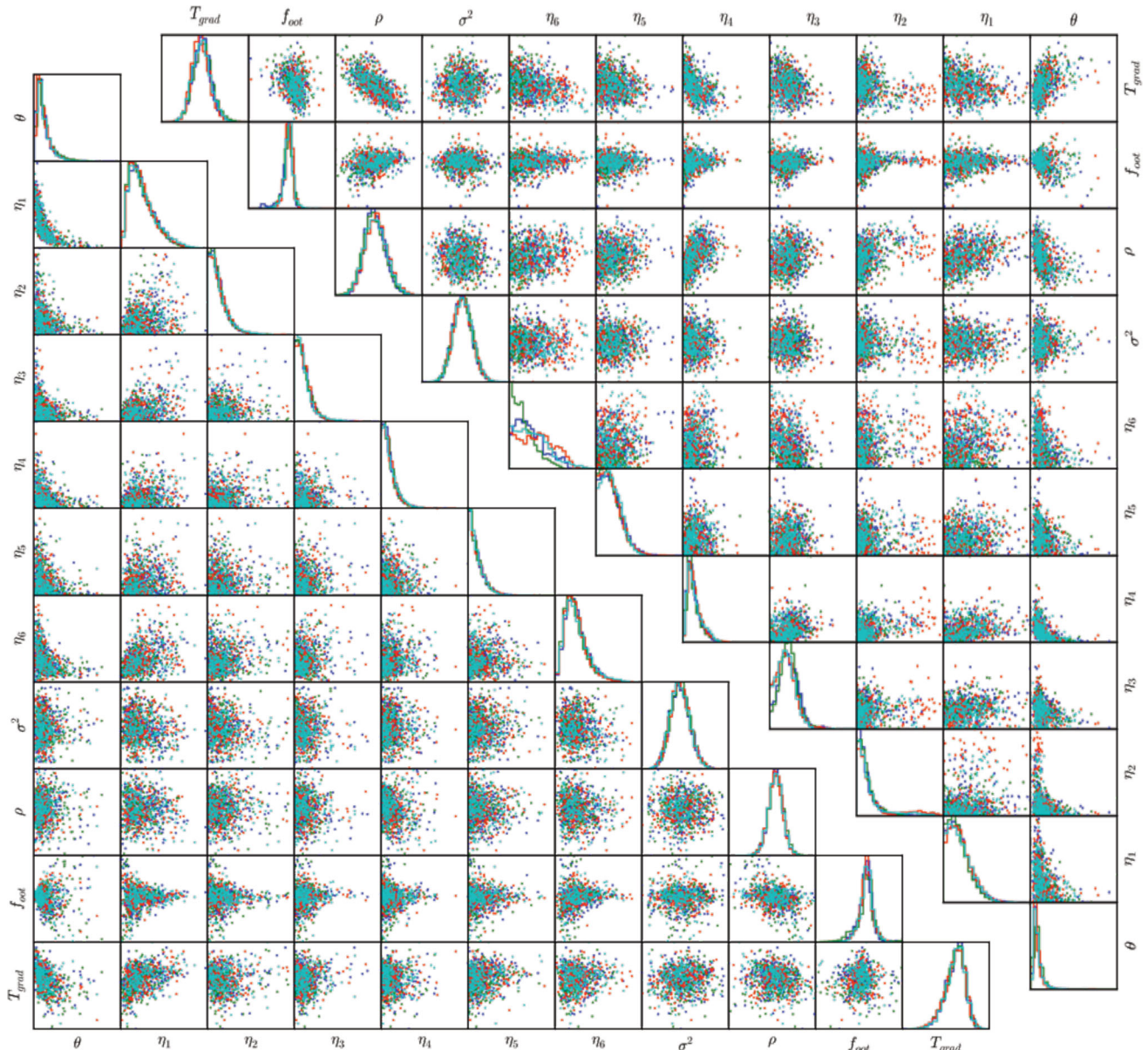
Figure 4. Posterior distributions of the variable transit parameters and hyperparameters from the MCMC chains for two of the wavelength channels, shown in the lower and upper triangles. The scatter plots show all pairs of parameters plotted after marginalization over all other parameters, and the histograms show the marginalized posterior distribution of each individual parameter. The different colours represent the four separate MCMC chains. The lower triangle shows one of the better converged wavelength channels, and the upper one shows one with some poorly converged hyperparameters.

likely due to the measurement being particularly noisy and only a proxy measurement. Of course the *true* temperature of the detector may still have a significant influence on the light curves, but unfortunately this is not available. The other parameters tend to vary in their influence, in part due to the degeneracy discussed earlier. SVT08 and GPA11 recognized the angle as the most important input parameter. This at first might seem contradictory to our results, but in this case importance was judged on the overall (visual) change in features of the outputted transmission spectrum. Here we judge it on the hyperparameter values from the posterior distribution. Indeed, parameters that are only relevant for a subset of the wavelength channels may in fact have the greatest impact on any features in the transmission spectrum.

The resulting transmission spectrum is shown in the right-hand plot of Fig. 3. Again the dashed line shows the white light curve planet-to-star radius ratio from GPA11. The spectrum is consistent with the type-II maximum likelihood results, the main difference being that the uncertainties are larger. Therefore, in this case it was important to marginalize over all the hyperparameters to obtain good estimates of the uncertainties. Given the convergence was not perfect, we ran the same MCMC procedure several more times. The transmission spectrum and uncertainties produced were almost indistinguishable.

This spectrum is also consistent with the results of SVT08 and GPA11, when using the simple linear basis models. However, the uncertainties are much larger and do not provide strong evidence for the molecular features reported in SVT08. The central 'feature' in the spectrum at ~2.1 μm is no longer significant. Bluewards of ~1.7 μm the same dip appears, but with lower significance. We emphasize that if the simple linear basis functions provided a good explanation for the systematics, our GP model would have reproduced the same transmission spectrum (and uncertainties), as the GP marginal likelihood function favours long length-scales (i.e. slowly varying functions) if they are sufficient to explain the instrumental systematics. Therefore, we strongly encourage anyone wishing to carry out comparisons with theoretical models to use the more conservative, but robust transmission spectrum determined in the present work using the GP model. Not doing so would entail the risk of overinterpreting features in the spectrum. These results are provided in Table 1.

The dip on the left-hand side of the spectrum could indeed be the result of a drop in molecular absorption in the atmosphere. However, we warn the reader that this region of the spectrum is where the flux is lowest, and therefore is susceptible to systematics from poor background subtraction as discussed in GPA11. This has the effect of stretching or compressing the transit light curve in the flux axis, and therefore the dip could be the result of systematics remaining from the data reduction. This raises an important point; GP models, or indeed any sophisticated models, can still suffer from systematic effects not removed from the data sets at the initial reduction stage. A more detailed interpretation of the HD 189733 spectrum will form the subject of a later paper (Gibson et al., in preparation), combined with WFC3 near-infrared observations that overlap the blue region of the NICMOS spectrum, and will shed some light on the interpretation of this feature.

4 SUMMARY AND DISCUSSION

We have introduced GPs as an alternative to deterministic models for the modelling and removal of systematics in the presence of auxiliary input parameters. GPs provide a powerful Bayesian approach to place distributions over functions. Rather than imposing

Table 1. NICMOS transmission spectrum of HD 189733 from our GP model, giving the wavelength, planet-to-star radius ratio ρ and its uncertainty $\Delta\rho$. The wavelengths are displayed red to blue, so that they correspond to the plots in Fig. 1.

Wavelength (μm)	ρ	$\Delta\rho$
2.468	0.155 45	0.000 77
2.411	0.155 20	0.000 52
2.353	0.154 55	0.000 44
2.296	0.155 13	0.000 57
2.238	0.155 12	0.000 41
2.181	0.155 04	0.000 51
2.124	0.154 17	0.000 66
2.066	0.155 08	0.000 66
2.009	0.153 93	0.000 36
1.951	0.155 49	0.000 51
1.894	0.155 95	0.000 60
1.837	0.155 13	0.000 53
1.779	0.155 34	0.000 51
1.722	0.154 47	0.000 87
1.665	0.154 29	0.000 64
1.607	0.152 66	0.000 62
1.550	0.153 59	0.000 73
1.492	0.153 67	0.001 18

a parametric model of instrumental systematics, by using GPs, one can marginalize out their ignorance of the functional form of the systematics model. GPs allow an arbitrary number of additional input parameters and a framework to determine which parameters are important to the analysis.

We demonstrated the application of our GP model to NICMOS transmission spectroscopy of HD 189733. The transmission spectrum is consistent with those found in SVT08 and GPA11 using simple linear basis functions, although with larger uncertainties. This is because the linear basis functions are not sufficient to explain the instrumental systematics, and therefore do not provide realistic treatment of the uncertainties. Detailed interpretation of these results will be included in a later paper.

Our GP model may also be used in the absence of auxiliary information, by using a time dependant covariance kernel to model time-correlated noise in exoplanet transits. However, due to the considerable computation time required, existing methods, in particular the wavelet method of Carter & Winn (2009), are considerably faster where applicable for large transit data sets, and are preferable in the absence of additional inputs, although limited to regularly sampled data. Comparing results of both methods for time-correlated noise will provide useful tests and may form the subject of future work.

Recently, Waldmann (2011) proposed another non-parametric method for removing systematics from transmission spectroscopy data sets, based on independent component analysis. This uses a completely different approach, and searches for common signals in multiple transit light curves, used to remove the systematics from the data. These new methods based on non-parametric models should provide effective new tools to robustly extract signals in the presence of unknown systematic noise, and the development of such methods are important to future exoplanet research. It would be interesting to compare results from our GP model to those using the approach proposed by Waldmann (2011), as the two methods may

prove complementary, given that they make different assumptions about the systematics.

GP models provide a general framework that can be applied to many different problems in regression, interpolation and prediction where a deterministic function is not available. The major limitation to applying GP models is that each evaluation of the posterior probability requires a matrix inversion, which scales badly with the size of the data set. This is unavoidable for data sets which require marginalization over function space. This difficulty may be somewhat eased in the near future by using sparse GP methods (e.g. Quiñonero-Candela & Rasmussen 2005; Walder, Kim & Schölkopf 2008), which approximate the covariance matrix as sparse, or using utilizing rank-reduction methods for large matrices where appropriate (e.g. Foster et al. 2009; Way et al. 2009), and/or by using GPUs to vastly speed up matrix inversion (e.g. Volkov & Demmel 2008). Indeed, ongoing research in this area should open up the application of GPs to many more astronomical data sets.

ACKNOWLEDGMENTS

REFERENCES

Bishop C. M., 2006, Pattern Recognition and Machine Learning. Springer, Berlin
Brown T. M., 2001, ApJ, 553, 1006
Brown T. M., Charbonneau D., Gilliland R. L., Noyes R. W., Burrows A., 2001, ApJ, 552, 699
Carter J. A., Winn J. N., 2009, ApJ, 704, 51
Charbonneau D., Brown T. M., Noyes R. W., Gilliland R. L., 2002, ApJ, 568, 377
Charbonneau D. et al., 2005, ApJ, 626, 523
Collier Cameron A. et al., 2007, MNRAS, 380, 1230
Deming D., Seager S., Richardson L. J., Harrington J., 2005, Nat, 434, 740
Foster L. et al., 2009, J. Machine Learning Res., 10, 857
Garnett R., Osborne M., Reece S., Rogers A., Roberts S., 2010, Comput. J., 53, 1430
Gelman A., Rubin D. B., 1992, Stat. Sci., 7, 457
Gibson N. P. et al., 2009, ApJ, 700, 1078
Gibson N. P. et al., 2010, MNRAS, 401, 1917
Gibson N. P., Pont F., Aigrain S., 2011, MNRAS, 411, 2199 (GPA11)
Gilliland R. L., Arribas S., 2003, Instrum. Sci. Rep. NICMOS 2003-001
Holman M. J. et al., 2006, ApJ, 652, 1715
Mahabal A. et al., 2008, in Bailer-Jones C. A. L., ed., AIP Conf. Ser. Vol. 1082, Classification and Discovery in Large Astronomical Surveys. Am. Inst. Phys., New York, p. 287
Mandel K., Agol E., 2002, ApJ, 580, L171
Neal R. M., 1996, Bayesian Learning for Neural Networks. Springer, Berlin
Pont F., Aigrain S., Zucker S., 2011, MNRAS, 411, 1953
Pont F., Gilliland R. L., Knutson H., Holman M., Charbonneau D., 2009, MNRAS, 393, L6
Pont F. et al., 2007, A&A, 476, 1347
Pont F., Knutson H., Gilliland R. L., Moutou C., Charbonneau D., 2008, MNRAS, 385, 109
Pont F., Zucker S., Queloz D., 2006, MNRAS, 373, 231
Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., 1992, in Press W. H., Teukolsky S. A., Vetterling W. T., Flannery B. P., eds, Numerical Recipes in C. MIT Press, Cambridge, MA
Quiñonero-Candela J., Rasmussen C., 2005, J. Machine Learning Res., 6, 1939
Rasmussen C. E., Williams K. I., 2006, Gaussian Processes for Machine Learning. MIT Press, Cambridge, MA
Seager S., Sasselov D. D., 2000, ApJ, 537, 916
Sing D. K., 2010, A&A, 510, A21
Swain M. R., Vasisht G., Tinetti G., 2008, Nat, 452, 329 (SVT08)
Vidal-Madjar A., Lecavelier des Etangs A., Désert J.-M., Ballester G. E., Ferlet R., Hébrard G., Mayor M., 2003, Nat, 422, 143
Volkov V., Demmel J., 2008, Tech. Rep. UCB/EECS-2008-49, EECS Department, University of California, Berkeley
Walder C., Kim K., Schölkopf B., 2008, in Proc. 25th Int. Conf., Machine Learning. ACM, New York, p. 1112
Waldmann I. P., 2011, preprint (arXiv:1106.1989)
Way M. J., Foster L. V., Gazis P. R., Srivastava A. N., 2009, ApJ, 706, 623
Way M. J., Srivastava A. N., 2006, ApJ, 647, 102
Winn J. N. et al., 2008, ApJ, 683, 1076

APPENDIX A: GAUSSIAN PROCESSES FOR REGRESSION

Gaussian process (GP) models are extensively used in the machine learning community for Bayesian inference in non-parametric regression and classification problems. They can be seen as an extension of kernel regression to probabilistic models. This appendix aims to give a brief introduction to GPs used for regression problems. Our explanations are based on the textbooks by Rasmussen & Williams (2006) and Bishop (2006), where the interested reader will find more complete and detailed information.

A1 Introducing Gaussian processes

We first define a collection of N observations, consisting of independent variables x_n and observed values y_n. Here $x_n = (x_{1,n}, \ldots, x_{D,n})^T$ is a D-dimensional input vector and y_n is the corresponding scalar output.[5] We arrange the N inputs into an $N \times D$ matrix $\mathbf{X} = (x_1, \ldots, x_N)^T$ and the N observed outputs into a vector $y = (y_1, \ldots, y_N)^T$.

A standard approach is to model data as

$$y = m(\mathbf{X}, \phi) + \epsilon,$$

where m is the *mean function* with parameter vector ϕ, and ϵ represents independent and identically distributed (i.i.d.) Gaussian noise:

$$p(\epsilon) = \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I}).$$

Here, $\mathcal{N}$ represents a multivariate Gaussian distribution with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2 \mathbf{I}$, where σ^2 is the variance of a single observation and $\mathbf{I}$ is the $N \times N$ identity matrix. The joint probability distribution for a set of outputs y is therefore

$$p(y \mid \mathbf{X}, \phi, \sigma) = \mathcal{N}(m(\mathbf{X}, \phi), \sigma^2 \mathbf{I}).$$

For example to represent photometric observations of a planetary transit, y would be the measured flux, typically the input matrix $\mathbf{X}$ contains only time and ϕ are the transit parameters. Thus, the above expression is the standard likelihood function often used for transit

---

[5] It is straightforward to extend this model to multidimensional outputs, but for simplicity we consider only scalars.

light-curve fitting. This could be extended to include additional inputs in the matrix **X**, such as auxiliary data pertaining to the state of the instrument (e.g. detector temperature) and observing conditions (e.g. airmass).

Formally, Rasmussen & Williams (2006) define a GP as a collection of random variables, any finite number of which have a joint Gaussian distribution. GPs are often written as

$$y(\boldsymbol{x}) \sim \mathcal{GP}\left(m(\boldsymbol{x}), \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix, and we have dropped the mean function parameters from the notation for simplicity. A GP has only two parameters: the mean and covariance. The parameters of the mean function and covariance function, $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$, respectively, are the *hyperparameters* of the GP.

As any finite collection of observations has a joint Gaussian distribution, the joint probability distribution of $\boldsymbol{y}$ is given by

$$p(\boldsymbol{y} \mid \mathbf{X}, \boldsymbol{\phi}, \boldsymbol{\theta}) = \mathcal{N}(m(\mathbf{X}), \boldsymbol{\Sigma}). \qquad (A1)$$

We can see that assuming an i.i.d. noise model is in fact a special case of a GP with a diagonal covariance matrix. In a GP model, each element in the covariance matrix is defined by the *kernel* or *covariance function* which has hyperparameters $\boldsymbol{\theta}$ and is given by

$$\boldsymbol{\Sigma}_{nm} = k(\boldsymbol{x}_n, \boldsymbol{x}_m),$$

where the kernel maps the input vectors $\boldsymbol{x}$ to a scalar covariance. The specification of a covariance function implies a probability distribution over functions; thus, a GP can be viewed as a distribution over functions. To illustrate this we will adopt a commonly used covariance function, the squared exponential, given by

$$k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \theta_0 \exp\left(-\theta_1 \sum_{i=1}^{D} (x_{i,n} - x_{i,m})^2\right),$$

where θ_0 is the maximum covariance and θ_1 is an inverse length-scale parameter (this becomes more obvious if we write $\theta_1 = 1/2l^2$ to get an unnormalized Gaussian). As there is only one hyperparameter governing the length-scale for all dimensions, this is called an *isotropic* kernel.

The GP now has a simple interpretation; data points that lie near each other in input space are highly correlated, and data points which are distant in input space are relatively uncorrelated. We can draw samples from the GP by choosing a number of inputs, calculating the corresponding covariance matrix, and generating a random Gaussian vector from the multivariate distribution. The top two plots in Fig. A1 show examples of three vectors drawn at random from a GP with a one-dimensional input space, with a squared exponential kernel function and zero mean function. The dark and light grey shaded areas represent the 1σ and 2σ boundaries of the prior distribution, respectively. The left-hand plot has hyperparameters $\{\theta_0, \theta_1\} = \{25, 1/8\}$ ($l = 2$), and the right-hand plot has hyperparameters $\{\theta_0, \theta_1\} = \{25, 2\}$ ($l = 0.5$). These sample vectors represent random smooth functions drawn from the prior (specified by the hyperparameters). The longer inverse length-scale means that the functions on the right-hand plot vary more quickly.

A GP with a squared exponential kernel function therefore defines a distribution over smoothly varying functions $y(\boldsymbol{x})$. The θ_0 hyperparameter determines the maximum variation of the function, and the inverse length-scale θ_1 determines how quickly varying the function is; in other words, how far input values need to be
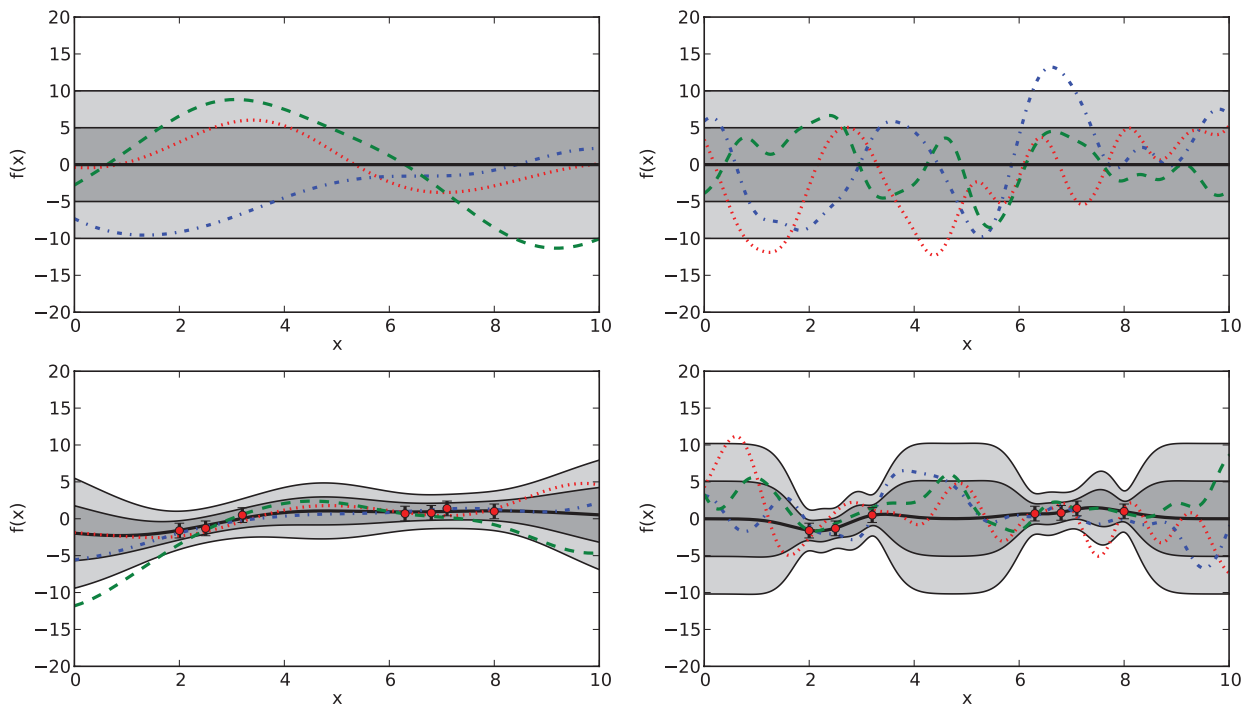

Figure A1. Top: examples of random functions drawn from a GP with a squared exponential covariance kernel. The left- and right-hand plots have hyperparameters $\{\theta_0, \theta_1\} = \{25, 2\}$ and $\{\theta_0, \theta_1\} = \{25, 0.5\}$, respectively. The dark and light grey shaded areas represent the 1σ and 2σ boundaries of the distribution. Bottom: the predictive distributions generated after adding some training data and conditioning using the same hyperparameters as the corresponding plots above, with the addition of a white noise term ($\sigma^2 = 1$). The coloured lines are now random vectors drawn from the posterior distribution.

apart in order to become uncorrelated. It can be shown that GP regression with the squared exponential kernel function is equivalent to Bayesian linear regression with an infinite number of basis functions.[6]

The mean function of the GP controls the deterministic component of the model, whilst the covariance function controls its stochastic component. Multiple covariance functions can be added or multiplied together to model multiple stochastic processes (as they are still valid covariance functions). For example, most GP models include an i.i.d. component. To incorporate this, we simply add a term to the diagonal of the covariance matrix. The covariance function is then given by

$$k(\boldsymbol{x}_n, \boldsymbol{x}_m) = \theta_0 \exp\left(-\theta_1 \sum_{i=1}^{D}(x_{i,n} - x_{i,m})^2\right) + \delta_{nm}\sigma^2,$$

where δ is the Kronecker delta function, and σ^2 is a new hyperparameter representing the variance of the white noise. Therefore, a GP with this covariance function now defines a distribution over smoothly varying functions with the addition of white noise.

The joint probability distribution of the GP given by equation (A1) is the likelihood of observing $\boldsymbol{y}$, given the inputs $\mathbf{X}$, mean function parameters $\boldsymbol{\phi}$ and hyperparameters $\boldsymbol{\theta}$. This likelihood is already marginalized over all possible realizations of the stochastic component that satisfy the covariance matrix (or all possible functions in the distribution specified by the GP) and hence is commonly referred to as the marginal likelihood. In this sense, GPs are intrinsically Bayesian and help to mitigate against the problem of over-fitting associated with non-Bayesian methods.

A2 Gaussian process regression

Given a set of N observed data points $\boldsymbol{y}$ with corresponding inputs $\mathbf{X}$ (in the machine learning literature known as the training data), we can use a GP to make predictions for additional input values given by $\boldsymbol{x}_\star$ (known as the test data). We want to obtain the predictive distribution for the output $y_\star$ conditioned on the training data. According to the prior, the joint probability distribution of the training outputs $\boldsymbol{y}$ and the test output $y_\star$ is Gaussian and given by

$$p\left(\begin{bmatrix} \boldsymbol{y} \\ y_\star \end{bmatrix}\right) = \mathcal{N}\left(\begin{bmatrix} m(\mathbf{X}) \\ m(\boldsymbol{x}_\star) \end{bmatrix}, \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{k}_\star \\ \boldsymbol{k}_\star^T & c \end{bmatrix}\right),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix for the training data points, $\boldsymbol{k}_\star$ is the column vector with elements $k(\boldsymbol{x}_n, \boldsymbol{x}_\star)$ for $n = (1, \ldots, N)$, and c is the scalar given by $k(\boldsymbol{x}_\star, \boldsymbol{x}_\star) + \sigma^2$.

The predictive distribution of $y_\star$ is then obtained by *conditioning* on the observed data points. Using standard results the conditional joint posterior distribution is another Gaussian:

$$p\left(y_\star | \boldsymbol{x}_\star, \boldsymbol{y}, \mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}\right) = \mathcal{N}\left(\hat{y}_\star, \hat{\sigma}_\star^2\right),$$

where $\hat{y}_\star$ and $\hat{\sigma}_\star^2$ are the mean and variance of the distribution, respectively, given by

$$\hat{y}_\star = m(\boldsymbol{x}_\star) + \boldsymbol{k}_\star^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r},$$

and

$$\hat{\sigma}_\star^2 = c - \boldsymbol{k}_\star^T \boldsymbol{\Sigma}^{-1} \boldsymbol{k}_\star,$$

and $\boldsymbol{r} = \boldsymbol{y} - m(\mathbf{X})$ is the vector containing the residuals from the mean function.

These results may easily be extended to make predictions for $N_\star$ test inputs given by $\mathbf{X}_\star$. If we define $\mathbf{K}_\star$ as the $N \times N_\star$ matrix containing the covariance for all pairs of training inputs and test inputs, and $\mathbf{K}_{\star\star}$ as the $N_\star \times N_\star$ covariance matrix of the test inputs, the distribution is again Gaussian with mean $\hat{\boldsymbol{y}}$ and covariance matrix $\mathbf{C}$, given by

$$\hat{\boldsymbol{y}}_\star = m(\mathbf{X}_\star) + \mathbf{K}_\star^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r},$$

and

$$\mathbf{C} = \mathbf{K}_{\star\star} - \mathbf{K}_\star^T \boldsymbol{\Sigma}^{-1} \mathbf{K}_\star.$$

Fig. A1 gives simple examples of GP regression conditioned on some artificial training data, shown by the red points. The top plots show the prior probability distribution and random vectors drawn from it, as described in previous section. The lower plots show the distribution conditioned on the training data, with the mean indicated by the thick black line, surrounded by the 1σ and 2σ predictive distributions. These use the same hyperparameters as the corresponding plots above, with an additional white noise hyperparameter equal to 1. We can think of this process as a weighted average of the functions generated by the prior, weighted on their likelihood function, or mathematically, marginalizing over function space. The predictive distribution therefore becomes narrower as more of function space is restricted. Note that with the longer lengthscale the distribution is narrowed along most of the x range plotted, whereas with the shorter length-scale the distribution quickly resorts to the prior distribution far from any observed data. The plotted distributions include white noise for the training data, and the three coloured lines are random functions drawn from the posterior distribution.[7] Clearly, the value of the hyperparameters has a large impact on the regression; luckily it is possible to infer these from the training data.

A3 Learning the hyperparameters

GPs are extremely effective for interpolation and prediction, and indeed this is one of their main applications in the machine learning community. However, GPs can also be used to infer distributions of the mean function parameters and/or hyperparameters, either of which may be the quantities of interest. We consider the marginal likelihood function from the previous sections:

$$\mathcal{L}(\boldsymbol{r}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{r}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r}\right),$$

where we have now explicitly written the Gaussian function. It is convenient to use the log likelihood:

$$\log\mathcal{L}(\boldsymbol{r}|\mathbf{X}, \boldsymbol{\theta}, \boldsymbol{\phi}) = -\frac{1}{2}\boldsymbol{r}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{r} - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{n}{2}\log(2\pi).$$

Ideally, in a Bayesian framework one should obtain the posterior probability distribution of each hyperparameter of interest by marginalizing over all the others. This can be done, for example, using a sampling technique such as Markov Chain Monte Carlo. However, this is expensive to compute as each evaluation of the likelihood function we must invert the $N \times N$ covariance matrix, requiring $\mathcal{O}(n^3)$ operations.

---

[6] In practice, we need only to consider the function values for a finite set of inputs making it possible to work in the infinite space of basis functions. This is impossible using standard basis function models.

[7] When plotting the distribution we need only to consider the diagonal of the matrix $\mathbf{K}_{\star\star}$, whereas when generating random vectors we must consider the full covariance matrix of the test inputs.
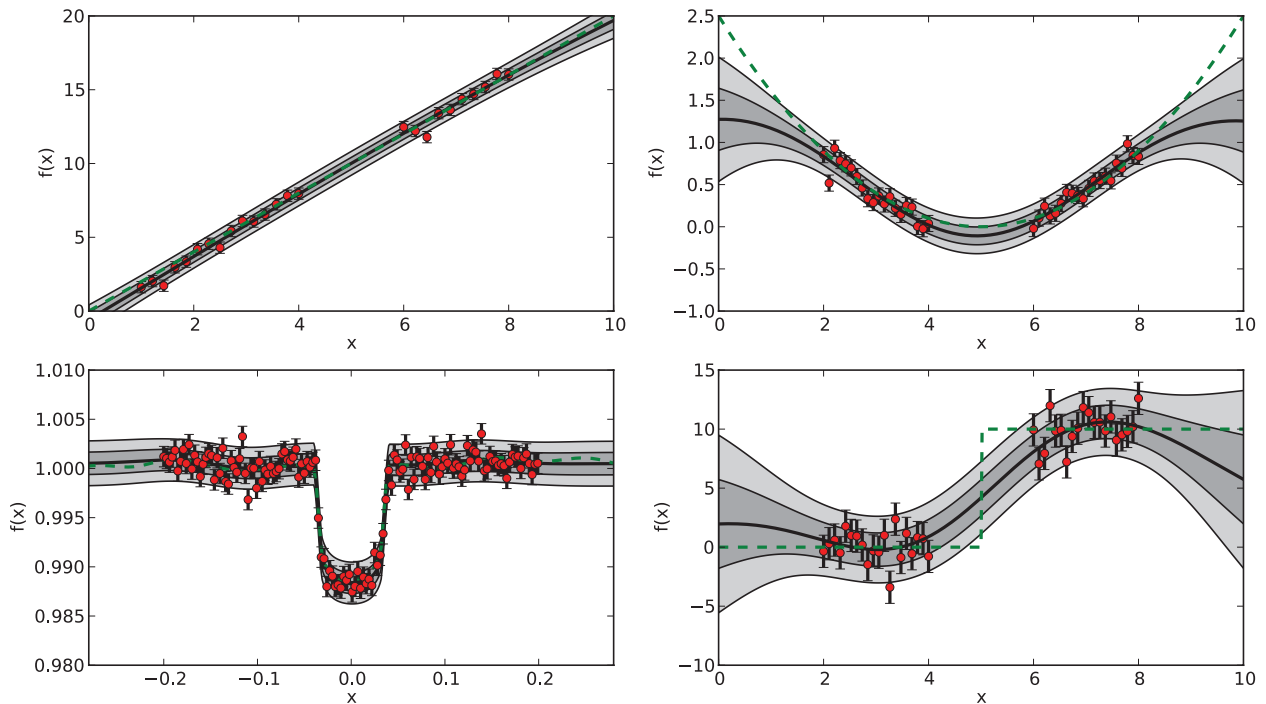
Figure A2. Examples of GP regression using type-II maximum likelihood. The dashed green lines represent the functions that the training data were drawn from, before adding Gaussian i.i.d. noise. All data were fitted using a GP with squared exponential kernel, and all used a zero mean function with the exception of the bottom left-hand plot, which used a transit mean function. The solid grey and black regions represent the mean, and 1σ and 2σ confidence intervals. Clockwise from top left, the function used to generate the data is: a straight line, a quadratic function, a step-function and a planetary transit model with the addition of correlated 'noise'. See the text for discussion.

One alternative is to make an approximation known as type-II maximum likelihood. This involves fixing the covariance hyperparameters at their maximum likelihood values and marginalizing over the remaining mean function hyperparameters of interest. This is a valid approximation when the posterior distribution is sharply peaked at its maximum with respect to the covariance hyperparameters and should give similar results to the full marginalization.

Fig. A2 gives examples of GP regression using a squared exponential covariance function, after optimizing the hyperparameters via a Nelder–Mead simplex algorithm (see e.g. Press et al. 1992). The training data (red points) were generated from various functions (shown by the green dashed line) with the addition of i.i.d. noise. All were modelled using a zero-mean function, with the exception of the transit function at the bottom left, which used a transit mean function. In the top-left the data were generated from a straight line function. The marginal likelihood strongly favours functions with a longer length-scale, and the GP is able to predict the underlying function both when interpolating and extrapolating away from the data. The top right-hand data were generated from a quadratic function and again the GP reliably interpolates the underlying function.

However, as the length-scale is now shorter, the predictive distribution veers towards the prior more quickly outside the training data. These two examples show that GPs with squared exponential kernels are very good at interpolating data sets providing the data are generated from a smooth function. The bottom left-hand plot shows data drawn from a transit mean function with sinusoidal and Gaussian time-domain 'systematics' added. The GP model has a transit mean function and the additional 'systematics' are modelled by the GP. Finally, the bottom right-hand plot shows data drawn from a step function. In this case the GP interpolation is less reliable. This is because the squared exponential function implies smooth functions. A more appropriate choice of covariance kernel is required for such functions (e.g. Garnett et al. 2010). Indeed, a wide choice of covariance kernels exist which effectively model periodic, quasi-periodic and step functions (see Rasmussen & Williams 2006), enabling GPs to be a very flexible and powerful framework for modelling data in the absence of a parametric model.

This paper has been typeset from a TEX/LATEX file prepared by the author.