Reliable inference of exoplanet light-curve parameters using deterministic and stochastic systematics models

N. P. Gibson

European Southern Observatory, Karl-Schwarzschild-Str. 2, D-85748 Garching bei München, Germany

Accepted 2014 September 19. Received 2014 September 17; in original form 2014 July 30

ABSTRACT
Time series photometry and spectroscopy of transiting exoplanets allow us to study their atmospheres. Unfortunately, the required precision to extract atmospheric information surpasses the design specifications of most general purpose instrumentation, resulting in instrumental systematics in the light curves that are typically larger than the target precision. Systematics must therefore be modelled, leaving the inference of light-curve parameters conditioned on the subjective choice of systematics models and model selection criteria. This paper aims to test the reliability of the most commonly used deterministic systematics models and model selection criteria. As we are primarily interested in recovering light-curve parameters rather than the favoured systematics model, marginalization over systematics models is introduced as a more robust alternative than simple model selection. This can incorporate uncertainties in the choice of systematics model into the error budget as well as the model parameters. Its use is demonstrated using a series of simulated transit light curves. Stochastic models, specifically Gaussian processes, are also discussed in the context of marginalization over systematics models, and are found to reliably recover the transit parameters for a wide range of systematics functions. None of the tested model selection criteria – including the Bayesian information criterion – routinely recovered the correct model. This means that commonly used methods that are based on simple model selection may underestimate the uncertainties when extracting transmission and eclipse spectra from real data, and low significance claims using such techniques should be treated with caution. In general, no systematics modelling techniques are perfect; however, marginalization over many systematics models helps to mitigate poor model selection, and stochastic processes provide an even more flexible approach to modelling instrumental systematics.

Key words: methods: data analysis – methods: statistical – techniques: photometric – techniques: spectroscopic – planetary systems.

1 INTRODUCTION
Transiting exoplanets allow us to study their compositions and atmospheres via wavelength-dependent variations in their light curves, and significant progress has been made in this field since the first detection of an exoplanetary atmosphere (Charbonneau et al. 2002). During primary transit we can extract a transmission spectrum: a measurement of the transit depth or effective planet radius as a function of wavelength, which in turn depends on opacities in the planet’s upper atmosphere and therefore can reveal the presence of atomic and molecular species (Seager & Sasselov 2000; Brown 2001). During an eclipse or secondary transit, when the planet passes behind its host star, the eclipse depth is a measure of the star–planet flux ratio (e.g. Charbonneau et al. 2005; Deming et al. 2005), and measuring this at multiple wavelengths allows us to construct a spectrum of the planet’s dayside. Finally, phase curves can reveal the day/night contrast and map a planet’s emission spectrum as a function of longitude (e.g. Knutson et al. 2007, 2009; Cowan & Agol 2008).

All of these methods require exquisite temporal stability in the light curves in order to measure the tiny wavelength-dependent changes that the planetary atmosphere imprints on the light curves (typically $\sim 10^{-4}$ for hot Jupiters). In terms of photon noise, this is easily achievable for nearby, bright stellar systems. However, there are currently no dedicated instruments to study exoplanet atmospheres, and consequently we have been using common-user instruments for many years that were not designed with this level of precision in mind. This means that the vast majority of our observations are limited by our understanding
of the instrumental systematics, which manifest themselves as bumps and wiggles in the light curves, often of unknown origin. These are usually larger than the target precision of our observations, and must therefore be modelled when extracting measurements from exoplanet light curves in order to reach a useful precision.

Despite considerable efforts, systematics models remain a key factor in the analysis and interpretation of exoplanet spectra, and the community is yet to reach a consensus on the correct approaches to obtaining robust results in the presence of systematic noise. One prominent example is the Near Infrared Camera and Multi-Object Spectrometer (NICMOS) transmission spectrum of the hot Jupiter HD 189733b. The original analysis was performed using a simple, linear basis model to account for the systematics, where the bases consist of auxiliary parameters related to the orbital phase and telescope pointing. This resulted in the detection of molecular species (Swain, Vasisht & Tinetti 2008), but the choice of linear basis model was arbitrary rather than physically motivated. Gibson, Pont & Aigrain (2011) showed that this choice of linear basis model was arbitrary rather than physically motivated. These are usually larger than the target precision of our observations, and must therefore be modelled when extracting measurements from exoplanet light curves in order to reach a useful precision.

2 Inference of Light-Curve Parameters

To discuss the impact of systematics models on the inference of light-curve parameters, it is convenient to begin with a description of a light-curve model. The light curve consists of a series of $N$ flux measurements given by the column vector $f = (f_1, \ldots, f_N)^T$, observed at time $t = (t_1, \ldots, t_N)^T$, collectively referred to as the data, $D$. We fit a model light curve $T(t, \theta)$, a function of time with model parameters $\theta$, and wish to infer the joint probability distribution of $\theta$.

Under the assumption that the flux measurements are Gaussian distributed about the light-curve model with mean zero and standard deviation $\sigma = (\sigma_1, \ldots, \sigma_N)^T$, we proceed by constructing a likelihood function as the product of independent Gaussians:

$$p(D|\theta) = \prod_{i=1}^{N} \mathcal{N}(f_i|T(t_i, \theta), \sigma_i^2)$$

$$= \mathcal{N}(f|T(t, \theta), \Sigma).$$

Here, the light-curve model, $T(t, \theta)$, is the mean function of the Gaussian, and the $\Sigma$ is the covariance matrix, where $\Sigma = \text{diag}(\sigma)$. We can maximize this likelihood with respect to the light-curve parameters to obtain a point estimate of $\theta$, equivalent to minimizing the $\chi^2$ statistic if the uncertainties are held fixed. However, using the full likelihood means we can fit for a constant noise term, $\sigma$, or alternatively a scale factor if the flux uncertainties vary with time. This is conceptually similar to setting the reduced $\chi^2$ equal to 1, but allows the ‘uncertainty in the uncertainties’ to be taken into account.

To obtain the joint or marginal posterior probability distribution(s) for $\theta$, we turn to Bayes’ theorem:

$$p(\theta|D) = \frac{p(D|\theta) p(\theta)}{p(D)},$$

where $p(D|\theta)$ is the joint posterior probability distribution of $\theta$, $p(D|\theta)$ is the likelihood, $p(\theta)$ is our prior probability distribution for $\theta$, and $p(D)$ is the evidence under our model, independent of $\theta$. In general each of these probabilities should be conditioned on the context, which represents the explicit and implicit assumptions made about our data and models, such as the validity of the transit model and assumption of Gaussian noise, but we usually drop it for clarity. Indeed, for the remainder of this paper Gaussian noise is assumed, which may not always be the case in practice. The consistency of the residuals with a Gaussian distribution can be verified, and generally, rejecting a few outlying points when $N$ is large is a simple and uncontroversial procedure. However, a more detailed treatment of outliers is beyond the scope of this paper.

To infer the value of a single parameter in $\theta$, say $\theta_k$ we marginalize over the remaining parameters $\theta_{\neq k}$:

$$p(\theta_k|D) = \int p(\theta|D) d\theta_{\neq k}.$$
2.1 Inference using systematics models

Unfortunately the light curve is typically degraded by instrumental systematics, and our assumption that the flux values are independent and normally distributed around the light-curve model is no longer true. This must be accounted for when inferring $\theta$, a problem first discussed in detail by Pont, Zucker & Queloz (2006) in the context of transit surveys. Several methods have been suggested to account for additional systematic noise by rescaling the uncertainties to account for the loss of information (Pont et al. 2006; Winn et al. 2008), or shuffle the residuals from the best-fitting model to preserve the correlated noise in the light curves and evaluate the effect on the parameter uncertainties (residual permutation; Gillon et al. 2007). Alternatively we can use stochastic processes to fit for and account for time-correlated noise such as the wavelet approach of Carter & Winn (2009), or based on time-dependent GPs (Gibson et al. 2013a,b).

In many cases we instead attempt to model the systematics as a parametric function of a set of auxiliary parameters extracted from the data (often called optical state vectors or decorrelation parameters; see Gibson et al. 2011, for a more detailed discussion of linear basis models). These could be the position of the star on the detector, width of the stellar point spread function (PSF), temperature of the detector and so on, and the reader is referred to previously mentioned papers for more detail. This in principle allows us improve the precision of our parameter estimates. Using simple parametric models for systematics has long been a standard approach for studies of exoplanet atmospheres (e.g. Brown et al. 2001; Gilliland & Arribas 2003; Pont et al. 2007; Swain et al. 2008; Sing et al. 2011; Huitson et al. 2013; Nikolov et al. 2014). A common problem with this approach is that there is often no unique and obvious way of modelling the instrumental systematics, and our inference of $\theta$ is no longer dependent on simple assumptions like a transit model and Gaussian noise, but rather on the subjective choice of how to model the systematics. When we choose a systematics model, $S$, for our data, we are effectively conditioning our results on $S$, and we should rewrite Bayes’ theorem accordingly:

\[
p(\theta, \phi|D, S) = \frac{p(D|\theta, \phi, S) p(\theta, \phi|S)}{p(D|S)},
\]

where $\phi$ is an additional parameter vector containing parameters of the systematics model. We can similarly obtain a marginal distribution for $\theta_i$, by also marginalizing over the systematics parameters:

\[
p(\theta_i|D, S) = \int \int p(\theta, \phi|D, S) d\phi d\theta_i.
\]

Crucially, this takes into account the uncertainty in the parameters of the systematics model, not the choice of systematics model.

In this case, our knowledge of $\theta$ is explicitly dependent on the choice of $S$, whether it is an arbitrary choice or objectively selected from a group of models. If $S$ does not represent the true systematics model, then our inference of $\theta$ is unreliable. This is a crucial point to keep in mind for nearly all cases of transmission and emission spectroscopy. Naturally this will also affect any inference of atmospheric composition or structure, which will in turn be conditioned on the systematics model, $S$.

This is particularly problematic for seemingly arbitrary choices of systematics models. Gibson et al. (2011) discuss this in the context of Hubble Space Telescope (HST)/NICMOS transmission spectroscopy, where subtly changing the systematics models leads to significantly different transmission spectra and therefore to different physical interpretations of the planetary atmosphere. When an arbitrary choice of systematics model influences the exoplanet spectrum, and therefore the physical interpretation of the data, clearly we must test the assumptions that go into our systematics models in order to gain a more robust result from our data. Otherwise subjective choices will strongly influence our results, and different authors will come to different conclusions from the same data. More objective inference can be obtained by using model selection criteria to select the most appropriate model as discussed in the following section, and is the approach now followed by the majority of studies using parametric systematics models (e.g. Gibson et al. 2010; Sing et al. 2011; Crossfield et al. 2013; Huitson et al. 2013; Nikolov et al. 2014).

2.2 Model selection

Typically, systematics models are constructed from a set of auxiliary data extracted from the observations. The simplest model one can construct is a linear combination of these auxiliary data, and is given by the vector product $x_1^T \phi$, where $x_1$ is a vector containing $L$ auxiliary parameters for time $i$, and $\phi$ contains the $L$ weights for each parameter. The likelihood is now given by

\[
p(D|\theta, \phi, S) = \prod_{i=1}^N N\left(f_i|\langle T(t, \theta) \rangle \times X_1^T \phi, \sigma_i^2\right) = N(f_i|\langle T(t, \theta) \rangle \circ (X \phi), \Sigma),
\]

where $X = (x_1, \ldots, x_N)^T$ is the $N \times L$ design matrix, and $\circ$ denotes element wise multiplication. A vector of ones can be added to $X$ to account for an arbitrary offset in the flux, or alternatively absorbed into the light-curve model. An additive basis model could also be applied, but here only a multiplicative model is considered. Often a range of equally plausible basis models could be proposed to explain the systematics. For example we could include only a subset of the auxiliary parameters, or use higher order polynomials of the inputs to represent the systematics.

In the case of multiple systematics models $S_j$, it is not sufficient to select the model that gives the lowest $\chi^2$, or that produces the ‘whitest’ residuals. For mathematically principled model selection based on Bayesian inference, we must instead use the Bayesian evidence, $p(D|S)$. This is the denominator of Bayes’ theorem in Section 2.1. It is the normalization factor to ensure the posterior integrates to unity, and is interpreted as the probability of obtaining the data given a model $S$, after integrating over the parameters of the model. It is given by

\[
p(D|S_j) = \int \int p(D|\theta, \phi, S_j) p(\theta, \phi|S_j) d\theta d\phi.
\]

To select our best systematics model, we can calculate a probability for each systematics model and again call on Bayes’ theorem:

\[
p(S_i|D) = \frac{p(D|S_i) p(S_i)}{p(D)},
\]

where $p(S_i)$ is the prior belief of the model. Note that the evidence is often referred to as the marginal likelihood, given it now acts as the likelihood in Bayes’ theorem. In the absence of strong prior belief, the evidence $p(D|S)$ is the only term that contributes to the probability of $S$, and the ratio of these terms (known as the Bayes factor) can be used to assess the relative probabilities of systematics models and select the most appropriate one.

The model evidence naturally selects the model that best explains the data, but with the minimum required complexity; in other words it intrinsically applies Occam’s razor to the model selection. Whilst this term does not come naturally from MCMC methods, one can
use alternative numerical methods to evaluate it such as importance sampling (e.g. Bishop 2006; Gibson et al. 2013b) or nested sampling (Skilling 2004). One major difficulty is that for the Bayesian evidence to be well defined, we must place proper priors (i.e. prior probabilities that integrate to unity) on all the model parameters, or at least those that are not common to all models. In this case any selection of proper priors is also informative priors, and therefore the choice of prior can influence the model selection. An easy to implement approach to this based on Bayesian linear basis models is discussed in Section 3.2.

A commonly used albeit crude approximation to the evidence is the Bayesian information criterion (BIC; Schwarz 1978). Here the log evidence is approximated by

$$\ln p(D|S_\phi) \approx -\frac{1}{2} \text{BIC} = \ln p(D|\theta_*, \phi_*, S_\phi) - \frac{M}{2} \ln N,$$

where $\theta_*$ and $\phi_*$ are the best-fitting model parameters, $M$ is the number of variable parameters, and $N$ is the number of observations. Alternatively we can replace $-2\ln p(D|\theta_*, \phi_*, S_\phi)$ by $\chi^2$ after removing additive constants. The best-fitting model is that with the largest evidence, or lowest BIC. Note that the additive 'constants' depend on the uncertainties ($\sigma$), so BIC model comparison is invalid if calculated directly from the $\chi^2$ and the uncertainties are rescaled for different models; in this case the full likelihood is required.

An alternative information criterion, but less commonly used in the exoplanet literature is the Akaike information criterion\(^1\) (AIC; Akaike 1974). It is given by

$$\ln p(D|S_\phi) \approx -\frac{1}{2} \text{AIC} = \ln p(D|\theta_*, \phi_*, S_\phi) - M.$$

For large $N$, the AIC imposes a smaller penalty term for extra complexity as compared to the BIC.

### 3 MARGINALIZATION OVER SYSTEMATICS MODELS

One important caveat with model selection is that it assumes that one of the tested models is in fact the correct model. Given that the sources of instrumental systematics are poorly understood, we should not always assume that the systematics model will be well described by a linear basis model containing the auxiliary inputs and higher order terms. As we are generally only interested in the light-curve parameters rather than the specific systematics model, a more appropriate method is to marginalize over the various systematics models as well as the model parameters. This in principle can take into account the uncertainty associated with the choice of systematics model, provided that a suitable range of models is tested. As will be discussed in Section 3.3, this is one of the motivations for the GP framework introduced by Gibson et al. (2012a).

Again using Bayes’ theorem, the marginalized distribution of $\theta$ is given by a sum of the individual (model-specific) distributions of $\theta$, weighted on the evidence for each systematics model:

$$p(\theta|D) = \sum_q p(\theta|S_\phi, D) p(S_\phi|D),$$

where

$$p(\theta|S_\phi, D) = \int p(D|\theta, \phi_*, S_\phi) p(\theta, \phi|S_\phi) d\phi,$$

and is normalized so that $\int p(\theta|S_\phi, D) d\theta = 1$, and $p(S_\phi|D)$ is the marginal evidence as defined in Section 2.2 (normalized so that $\sum_q p(S_\phi|D) = 1$). If one of the models is heavily preferred over the others, then this result will be equivalent to model selection, but this is important to check during inference.

Even if the underlying distributions of $\theta$ conditioned on the individual systematics models were Gaussian, in general the resulting Gaussian mixture model is non-Gaussian. We could work with the full mixture distribution, but in the current work the mean and variance of the mixture distribution will suffice. For a single parameter, these are calculated assuming the conditional distributions are Gaussian with means and standard deviations $m_q$ and $s_q$. In this case the mean $m$ and standard deviation $s$ of the parameter with probability distribution given by the Gaussian mixture distribution are related by

$$m = \sum_q p(S_\phi|D) m_q,$$

and

$$s^2 = \sum_q p(S_\phi|D) \left[ (m_q - m)^2 + s_q^2 \right],$$

and are derived by summing the moments from the individual Gaussian components (e.g. Frühwirth-Schnatter 2006).

### 3.1 Feature space expansion

Before model selection or marginalization over multiple models, we first must decide on a set of systematics models to test. Typically, a combination of the basis inputs and higher order terms is explored. For the remainder of this paper, such a feature space is used to define the set of models $S_\phi$. Given a set of measured auxiliary parameters $X$, we might wish to test a variety of models using simple polynomial expansions of them, and models that leave out some of the inputs. Assuming that we never use a higher order term without all corresponding lower order terms for a given input, the number of possible models is given by $(1 + p)^L$, where $p$ is the highest order of the polynomial expansion and $L$ are the number of auxiliary inputs. For example, if we have two auxiliary inputs $a$ and $b$, then a systematics model of third order in $a$ and second order in $b$ is given by

$$\mathbf{x}^T \phi \triangleq \phi_1 a^1 + \phi_2 a^2 + \phi_3 a^3 + \phi_4 b_1 + \phi_5 b^2,$$

where higher order terms of $a$ and $b$ are incorporated into $\mathbf{x}$ and therefore $X$.

For large numbers of auxiliary inputs and/or high expansion orders this can get prohibitively expensive, particularly if using numerical methods to calculate the evidence. Such expansions are commonly used as systematics models (e.g. Sing et al. 2011; Huitson et al. 2013; Nikolov et al. 2014), but are used for model selection rather than marginalization. We might also consider using cross-terms in the basis expansion, which are used in some systematics models, but this is not considered in the current work.

### 3.2 Laplacian approximation of the evidence and shrinkage priors

In Section 2.2 we discussed the Bayesian evidence and the BIC and AIC as crude approximations to it, often used for model

\(^1\) In this paper a modified form of the AIC called AIC\(_C\) is used to correct for finite sample sizes (Hurvich & Tsai 1989). In this case $\ln p(D|S_\phi) \approx \ln p(D|\theta_*, \phi_*, S_\phi) - MN/(M - N - 1)$, and is equivalent to the AIC for large $N$. 

selection. The difficulty in using the full Bayesian evidence is partly that it is relatively slow to calculate using numerical methods, but more importantly that we need to assume proper priors on the parameters that are not common to the models we are comparing. We can improve our evidence approximation by using a Laplacian approximation to the posterior integral (e.g. Bishop 2006).

The Laplacian approximation uses a second-order Taylor expansion of the log posterior around the maximum posterior value. The Jacobian is zero at a stationary point so we need only consider the Hessian matrix, which is the negative inverse covariance matrix. We therefore only need an estimate of the maximum (unnormalized) posterior value, and the covariance matrix of our variables \( \mathbf{K} \), both of which are easily obtained from an MCMC or a Levenberg–Marquart optimization. In this case we are approximating the posterior by a Gaussian distribution, and the integral can be estimated from the ratio of the maximum posterior value to that of an equivalent normalized Gaussian distribution. The marginal likelihood in this case is estimated as

\[
    p(D | \mathbf{S}_q) \approx \frac{q(\theta^*, \phi^* | D, \mathbf{S}_q)}{N(\theta^*, \theta^* | \phi^*, \mathbf{K})} = \sqrt{2\pi|\mathbf{K}|} \cdot q(\theta^*, \phi^* | D, \mathbf{S}_q),
\]

where \( q(\theta^*, \phi^* | D, \mathbf{S}_q) = p(D | \theta^*, \phi^*, \mathbf{S}_q) p(\theta^*, \phi^* | \mathbf{S}_q) \) is the un-normalized posterior. In practice it is better to calculate the log evidence:

\[
    \ln p(D | \mathbf{S}_q) \approx \ln q(\theta^*, \phi^* | D, \mathbf{S}_q) + \frac{M}{2} \ln 2\pi + \frac{1}{2} \ln |\mathbf{K}|
    \approx \ln q(\theta^*, \phi^* | D, \mathbf{S}_q) + \frac{1}{2} \ln |2\pi|\mathbf{K}|.
\]

This is in many ways resembles the BIC in that increasing the number of data points or the model complexity typically increases a penalty term (indeed the BIC and AIC can be derived following these arguments using a Gaussian approximation and vague priors; e.g. Bishop 2006), but uses the covariance matrix to obtain a more accurate estimate of the log evidence, as verified from importance sampling using the technique described in Gibson et al. (2013b).

In order to apply priors to the non-common systematics parameters, we follow a Bayesian treatment of linear regression, and introduce zero mean priors to each of the basis model weights (Bishop 2006). In this case we need only assign proper priors to the non-common inputs, and this is where the Laplace approximation shows its value. Instead of using simple model selection (or marginalization) to choose the systematics models, we can use more complex models providing that we have shrinkage priors assigned to each additional basis vector. This is the general principle behind Bayesian linear basis models. This allows linear models with large numbers of basis functions to be fit to relatively few data points, whilst mitigating overfitting. This is because the shrinkage priors restrict the complexity of the model, by avoiding the model being able to fine-tune itself using large basis weights. The width of the priors cannot be set by optimizing the likelihood or posterior, but can be set through optimizing the marginal likelihood. The reader is referred to Bishop (2006) for more details.

To implement this for our systematics models, for each of the systematics parameters \( \phi \), we can assign a zero mean Gaussian prior with variance \( \alpha^2 \):

\[
p(\phi | \mathbf{S}_q) = \mathcal{N}(0, \alpha^2 I).
\]

We first use an MCMC to determine the mean and covariance for the likelihood function (i.e. with posterior uniform priors), and approximate it as a Gaussian. We can then analytically multiply the likelihood and prior as they are both Gaussian; and use the Laplace approximation to evaluate the log evidence. Thus by using the Laplace approximation we can easily maximize the evidence with respect to \( \alpha \) to find suitable shrinkage priors for our inference. We can subsequently use the posterior distribution with the maximized value for \( \alpha \) to obtain our desired distribution for \( \theta \), perform another MCMC analysis using priors on the systematics parameters fixed to \( \alpha \), or again use the values for the log evidence for each tested model to find the best-fitting model. Alternatively, we can marginalize over the feature space to obtain a marginalized distribution for \( \theta \).

3.3 Gaussian processes as marginalization over systematics models

A GP can be seen as a convenient shortcut to the above marginalization over many systematics models. Formally, a GP is an infinite collection of random variables, any finite number of which have a joint Gaussian distribution (Bishop 2006; Rasmussen & Williams 2006). They were introduced for modelling time series light curves by Gibson et al. (2012a) and have been subsequently used in various forms in Gibson et al. (2012b, 2013a,b), Aigrain, Pont & Zucker (2012), Evans et al. (2013), which the reader is referred to for further details.

A GP may be derived from a probabilistic linear basis model, similar to those discussed in the previous section, by considering the mean and covariance of a random draw from the model. To make the connection between GPs and feature space models discussed earlier, we follow the derivation of Rasmussen & Williams (2006) and Bishop (2006), starting with a basis model as described in Section 2.2:

\[
f = \mathbf{X}\phi + \epsilon,
\]

where \( \epsilon \) is the white noise term distributed according to \( \sigma \), i.e. \( p(\epsilon) = \mathcal{N}(0, \sigma^2 I) \), and \( \sigma^2 \) is the variance of the white noise. To generate a probability distribution over the resulting functions, we must place a prior distribution over the basis weights, \( \phi \). Similarly to the Laplace approximation, we place a zero mean Gaussian with prior variance \( \alpha^2 \):

\[
p(\phi) = \mathcal{N}(0, \alpha^2 I).
\]

Following the derivation of Rasmussen & Williams (2006) and Bishop (2006), we then proceed to find the joint probability distribution of \( f \), which is also Gaussian, and therefore we only need to calculate the mean, \( \mu \), and covariance, \( \Sigma \). The mean function is given by the expectation of \( f \):

\[
\mu = \mathbb{E}[f] = \mathbb{E}[\mathbf{X}\phi] + \mathbb{E}[\epsilon] = 0,
\]

and the covariance matrix is given by the expectation of \( (f - \mu)(f - \mu)^T \):

\[
\Sigma = \mathbb{E}[(f - \mu)(f - \mu)^T] = \mathbb{E}[(\mathbf{X}\phi + \epsilon)(\mathbf{X}\phi + \epsilon)^T] = \alpha^2 \mathbf{X}\mathbf{X}^T + \sigma^2 I.
\]

\(2\) For convenience a very broad Gaussian distribution for each of the variable light-curve parameters is taken as the prior. These do not affect the mean and variance of the posterior, and is equivalent to adding an arbitrary constant to the log evidence of each model.
Our GP is fully defined by mean and covariance, and is now given by
\[ p(f) = \mathcal{N}(0, \sigma_f^2 \mathbf{X} \mathbf{X}^T + \sigma_n^2 I). \]

Using this procedure, we can write any Bayesian linear basis model as a GP. Similarly, we can directly define the kernel function of a GP, and convert it back into feature space. This is where GPs become extremely powerful for non-parametric data analysis, as we can define kernels that have infinite dimensions in feature space. A GP can therefore be used to place a probability distribution over a huge class of functions, and therefore automatically marginalize over many systematics models simultaneously. This was the motivation behind the GP systematics framework introduced by Gibson et al. (2012a). For example, the squared exponential kernel used by Gibson et al. (2012a) can be represented by an infinite number of basis functions (Rasmussen & Williams 2006).

To apply GPs to exoplanet light curves, a mean function representing the light-curve model can be easily included in the GP, and the kernel function (the stochastic component) represents the systematics model. In this case the likelihood function for a GP is identical to that given in Section 2:
\[ p(D | \theta, \phi) = \mathcal{N}(f | T(t, \theta), \Sigma). \]

The difference is that the covariance matrix is no longer a diagonal matrix, but is given by the kernel function which defines each element \( \Sigma_{ij} \) in \( \Sigma \). The squared exponential kernel used in Gibson et al. (2012a) is given by
\[ \Sigma_{ij} = k(x_i, x_j) = \xi^2 \exp \left( -\sum_{k=1}^{K} \eta_k (x_{i,k} - x_{j,k})^2 \right) + \delta_{ij} \sigma^2. \]

Here \( \eta = (\eta_1, \ldots, \eta_K) \) are the inverse length scale parameters, one for each input in \( \mathbf{X} \), \( \xi \) is the height scale, and \( \delta_{ij} \) is the Kronecker delta. Note that the parameters of our mean function and covariance matrix are often referred to as hyperparameters in the context of GPs. Two inputs are highly correlated when all inputs in \( \mathbf{x} \) are similar (with similarity governed by \( \eta \)), and uncorrelated otherwise. This therefore represents a joint systems model, i.e. the systematics are not independent functions of the components of \( \mathbf{X} \) added together.

We can make the systematics components independent by modifying the kernel so that each input has its own, independent height scale:
\[ \Sigma_{ij} = k(x_i, x_j) = \sum_{k=1}^{K} \xi_k^2 \exp \left( -\eta_k (x_{i,k} - x_{j,k})^2 \right) + \delta_{ij} \sigma^2. \]

This kernel now represents a systems model made up of independent functions of the inputs in \( \mathbf{X} \). The first kernel is hereafter referred to as the squared exponential (SE) kernel, and the second as the summed squared exponential (SSE) kernel.

Inference of our light-curve parameters proceeds identically to before, where we numerically sample from the joint posterior probability distribution, and parameters in the covariance kernel are treated in just the same way as our light-curve parameters, or the systematics parameters of the deterministic models. For this work, a logarithmic prior is imposed on the length scales \( \eta \) of the form
\[ p(x) = \begin{cases} 1/x & \text{if } x > 0, \\ 0 & \text{if } x \leq 0. \end{cases} \]

These are improper priors, and are equivalent to re-parametrizing \( \eta \) as \( \log \eta \), and is the natural parametrization for such a scale parameter. This also acts as a shrinkage prior, as it encourages the inverse length scale to be small unless required to explain the data. However, even without priors of this type, GPs are intrinsically Bayesian and try to find the simplest explanation for the data, as a complexity penalty is introduced in the first \( \Sigma \) in the Gaussian likelihood.

### Table 1. Transit and basis function parameter parameter distributions used to produce simulated light curves.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Symbol</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Central transit time (d)</td>
<td>( T_0 )</td>
<td>0</td>
</tr>
<tr>
<td>Period (d)</td>
<td>( P )</td>
<td>(2.75, 3.25)</td>
</tr>
<tr>
<td>Scaled semimajor axis</td>
<td>( a/R_\star )</td>
<td>(9, 11)</td>
</tr>
<tr>
<td>Impact parameter</td>
<td>( b )</td>
<td>(0.05, 0.25)</td>
</tr>
<tr>
<td>Planet-to-star radius ratio</td>
<td>( \rho = R_p/R_\star )</td>
<td>(0.08, 0.12)</td>
</tr>
<tr>
<td>Limb darkening parameters</td>
<td>( c_1, c_2 )</td>
<td>(0.1, 0.3)</td>
</tr>
<tr>
<td>White noise</td>
<td>( \sigma_w )</td>
<td>(0.0001, 0.0004)</td>
</tr>
<tr>
<td>Red noise</td>
<td>( \sigma_r )</td>
<td>(0.0001, 0.0004)</td>
</tr>
</tbody>
</table>

### 4 SIMULATIONS OF TRANSIT LIGHT CURVES WITH INSTRUMENTAL SYSTEMATICS

In order to investigate the best practices for inferring light-curve parameters in the presence of systematics, a series of transit light curves was simulated and degraded with simulated systematics, and their parameters were recovered using a variety of systematics modelling techniques, based on those discussed in Section 2. The transit models were simulated assuming a circular orbit, with parameters given in Table 1. They were generated using the analytic models of Mandel & Agol (2002), using a quadratic limb darkening law with parameters \( c_1 \) and \( c_2 \). The simulated light curves were of typical hot Jupiter-like transits, with the transit and noise parameters for each model picked from a uniform distribution from the (arbitrary) ranges outlined in Table 1.

Next, a set of stochastic basis models \( \mathbf{X} \) was generated by taking draws from a GP with a rational quadratic kernel (Rasmussen & Williams 2006):
\[ k(t_i, t_j) = \exp \left( 1 + \frac{(t_i - t_j)^2}{2\beta l^2} \right)^{-\beta}, \]

where \( t_i \) and \( t_j \) are the times of the two inputs, \( l \) is the (temporal) length scale, and \( \beta \in (0, \infty) \). This can be interpreted as a sum of squared exponential kernels of varying length scales, and becomes the squared exponential kernel when \( \beta \to \infty \). The kernel parameters were fixed at \( l = 0.07 \) and \( \beta = 0.1 \); these were chosen to generate ‘realistic’ looking, smooth input bases with typically 2–4 turning points over the transit duration. Examples of generated basis functions are shown in Fig. 1.

Each basis was first scaled to have mean 0 and variance 1. In some cases discussed below, some inputs from the generated basis set were excluded from the instrument model, or higher order terms were included to generate a feature space (i.e. the basis inputs were

3 In practice the matrix equation \( \mathbf{Ax} = \mathbf{b} \) is solved using Cholesky decomposition, where \( \mathbf{A} \) is a positive semidefinite matrix.
Figure 1. Examples of simulated light curves generated to test the fitting techniques. The red line is the model light curve, the dashed lines are basis inputs used to generate the systematics, and the green lines are the resulting systematics function. The left-, centre, and right-hand panels are light curves generated with the linear systematics model (Section 4.1), the feature space systematics model (Section 4.2), and the stochastic systematics model (Section 4.3), respectively. For each set of simulations, 10 000 light curves were generated and fitted to recover the planet-to-star radius ratio.
raised to integer powers and added to the design matrix \( \mathbf{X} \). For each basis set, the weight parameters \( \mathbf{\phi} \) were randomly generated from a uniform distribution with range \((-1,1)\), and the systematics model was generated as \( \mathbf{X} \mathbf{\phi} \). The resulting systematic function was scaled and shifted to have mean 1 and standard deviation (or ‘red noise’) \( \sigma_{\epsilon} \). The systematic function and transit function were then multiplied together, and white noise was added with standard deviation \( \sigma_{\text{w}} \) to generate the final simulated light curve.

For each simulated light curve, the planet-to-star radius ratio \( \rho \) was recovered using a variety of systematics models outlined below, with all other transit parameters fixed except for \( f_{\text{occ}} \), the out-of-transit flux, which was allowed to vary freely. All parameters of the systematics model were allowed to vary, as was the white noise. For each systematics model, the fitted planet-to-star radius ratio \( \hat{\rho} \), its uncertainty \( \sigma_{\hat{\rho}} \), and also the true value \( \rho \) were recorded for each transit fit. \( \hat{\rho} \) and \( \sigma_{\hat{\rho}} \) were found using an MCMC (see Gibson et al. 2012a, and references therein for more details of the implementation used here). For each model fit, two MCMC chains were run in parallel. After an initial burn-in period (typically of length 5000, but increased for more complex models), the chains were continued until the Gelman–Rubin statistic (Gelman & Rubin 1992) was below 1.01 for all variable parameters, or a maximum chain length was reached. In almost all cases the chains converged, but this was not confirmed for every single run.

To compare results, the reduced \( \chi^2 \) (\( \chi^2 \)) for each systematics modelling technique was calculated, which should be distributed according to a normal distribution with mean 1 and standard deviation \( \sqrt{2/\mathcal{N}_{\epsilon}} \), for large samples, where \( \mathcal{N}_{\epsilon} \) are the number of transit light curves fitted. The ‘number of sigma’ statistic (Carter & Winn 2009) was also calculated: \( \mathcal{N}_{\epsilon} = (\hat{\rho} - \rho) / \sigma_{\hat{\rho}} \). This should be distributed with a mean of zero and variance of unity if the distribution of \( \hat{\rho} \) is normally distributed around its true value with the derived uncertainty. These are both measures of the reliability of the inferred probability distribution of \( \rho \). Models were also compared using the mean uncertainty \( \sigma_{\rho} \) and the mean accuracy (i.e. mean of \( |\hat{\rho} - \rho| \)). The favoured systematics model is the one with the lowest mean uncertainty and accuracy but also with reliable uncertainties. For the remainder of this paper, an acceptable fit is roughly defined as one with a mean uncertainty within \( \approx 10 \) per cent of the true systematics model, and reliable uncertainties when \( \chi^2 \) is less than \( \approx 1.1 \).

Unless stated otherwise, 100 data points were generated over a 4.8 h duration. This is at the low end of the number of data points in a typical light curve, but was chosen to speed up inference over thousands of light curves. However, the resulting cadence of \( 3.8 \) h is similar to typical HST/Space Telescope Imaging Spectrograph (STIS) light curves, and in principle systematics modelling should become easier with higher cadence. The results of these simulations are only strictly valid for the range of transit and systematics parameters chosen here, but nonetheless such simulations should be able to establish some sensible guidelines for inferring light-curve parameters in the presence of instrumental systematics. Examples of the resulting basis functions and light curves are shown in Fig. 1 for a variety of the light-curve simulation methods outlined below.

### 4.1 Linear systematics model

The first set of simulated light curves was generated using a linear combination of known basis functions for the systematics. For each simulated transit, five basis functions were generated, but only the first three were used to create the systematics models as described in the previous section. This is a realistic scenario where many auxiliary inputs are extracted, but it is unclear which ones are the primary cause of the systematics (although the assumption of linearity is typically unrealistic). 10 000 simulated light curves were generated in total, and the transit depth was recovered using a series of systematics models.

These included simple linear basis models (LBM), using the correct, known basis expansion (LBM: correct exp), using only two of the correct bases (LBM: 2 inputs), and finally all five of the bases (LBM: 5 inputs). A series of first-order feature space (FS) expansions of the first four linear bases in \( \mathbf{X} \) was also used to fit each light curve, in this case resulting in 16 combinations of basis models. The distributions for the light-curves parameters were then obtained by marginalizing over the various feature space models using equation (1). The evidence was estimated using the BIC (FS: marginalized BIC), the AICc (FS: marginalized AIC), and the Laplace approximation to the evidence combined with shrinkage priors (FS: marginalized LP) as described in Section 3.2. The statistics were also recorded for the best-fitting systematics models found using each evidence approximation. Finally, a Laplacian approximation was used to compute a model with all five input bases, and after the value of \( \alpha \) was determined by maximizing the evidence, it was fixed and the posterior was resampled using MCMC (LP: 5 inputs).

The results of the using various systematics models are shown in Table 2, including \( \chi^2 \), the mean of the \( \mathcal{N}_{\epsilon} \) distribution, the mean uncertainty, the mean accuracy, and the rescaling of uncertainties required to set the \( \chi^2 \) equal to 1. Note the standard deviation of \( \mathcal{N}_{\epsilon} \) is given by \( \sqrt{2/\mathcal{N}_{\epsilon}} \), if the mean of \( \mathcal{N}_{\epsilon} \) is exactly zero. Clearly the best model to use for recovery of the planet-to-star radius ratio is the correct, known basis model, which generates ideal statistics and results in reliable uncertainties. However, in general the systematics model is not known a priori. Using more complex systematics models which contain the correct model as a subset (e.g. LBM: 5 inputs) also results in reliable uncertainties, but at the cost of inflated uncertainties. Therefore in order to get the most from our data, model selection or marginalization is required.

From the feature space models, it is clear that marginalizing over the feature space models is better than selecting one in all cases. This is because marginalization will help reduce the effects of poor model selection, and in particular is important when the evidence from multiple feature space models does not clearly select one over the others. Fig. 2 shows \( \mathcal{N}_{\epsilon} \) distributions for the best fit and marginalization over feature space using the Laplace approximation. It is clear that marginalization over feature space gives \( \mathcal{N}_{\epsilon} \) distributions that are closer to a uniform distribution than model selection. It is also clear that the AICc is better than the BIC. This is due to the AICc preferring more complex models, which results in more reliable yet larger uncertainties. This is a desirable quality for selecting or marginalizing over systematics models, as we are not particularly interested in which model is most likely to accurately describe the systematics models, but prefer to pick the least complex one that gives reliable uncertainties. The Laplace approximation to the evidence provides the best model selection and marginalization criterion, but is only slightly better than the BIC here, and with significantly more computation required. Finally, the Laplace approximation using shrinkage priors and a more complex model gives reasonable uncertainties along with a modest increase in the uncertainties. This is a good balance between allowing model complexity and preventing overfitting, but it is still dependent on the specific basis inputs used, so methods using model selection or marginalization are generally preferred.

Perhaps the most noteworthy result is that no systematics model selection procedure results in reliable statistics, even when we know
that the correct systematics model is within the tested models. Most strikingly, the commonly used method of BIC model selection (at least in this case) typically underestimates the uncertainties by a factor of ~100 per cent, although this is likely not always the case in practice (Section 5).

### 4.2 Feature space systematics

An additional two sets of simulated light curves were generated using a feature space expansion of basis inputs. Two linear basis functions were generated as before, and higher order terms were added to the design matrix \( \mathbf{X} \). The order for expansion was selected for each basis at random from the range 1–3 for each simulated light curve. The first set of simulations used \( N = 100 \) data points as in the previous section, and the second set was generated using \( N = 200 \) data points to test the systematics models with higher cadence.

For both sets of simulations, 10 000 light curves were generated, and \( \rho \) was recovered using a similar set of techniques as before. These methods included using only a linear basis expansion (LBM: linear exp), plus the known, correct basis expansion for each light curve (LBM: correct exp). A feature space expansion for up to fourth order in both bases was used, resulting in 25 combinations of basis models. The BIC (FS: marginalized BIC), AIC \(_C\) (FS: marginalized AIC), and Laplace approximations to the evidence (FS: marginalized LP) were then used as before to recover the fitted parameters, as well as the best-fitting model in each case. Laplace approximations were also recorded for up to second-order and third-order expansions (LP: 2nd/3rd-order exp), including refits after fixing the priors. Results for the same expansions were also recorded without the shrinkage priors on the weight parameters (LBM: 2nd/3rd-order exp). The results for both sets of simulations are shown in Tables 3 and 4.

In both cases we find similar results to the linear systematics simulations. The best model is unsurprisingly using the correct, known basis expansion. Marginalization over many models is again preferable to model selection, and the \( \mathcal{N}_\chi \) distributions for the Laplacian approximation are shown in Fig. 2, which again show that the marginalization results in a \( \mathcal{N}_\chi \) distributions closer to the unit normal. Similarly, using the BIC for model selection typically underestimates the uncertainty on average by ~100 per cent. The AIC \(_C\) is perhaps the best model selection criterion here, as the Laplace approximation slightly overestimates the uncertainties, although only by a small factor. Both methods provide reliable recovery of the planet-to-star radius ratio, although the AIC \(_C\) is significantly easier to calculate. Qualitatively, the simulations for \( N = 200 \) produce the same results as for \( N = 100 \) (i.e. the same model selection/marginalization techniques produce the best results), but quantitatively the model selection techniques are more reliable than before. This is not surprising, as model selection should become easier as the number of data points and data quality increases.

### 4.3 Gaussian process fitting

In order to establish the reliability of GPs for recovering light-curve parameters when the systematics can be well described by simple basis models, a GP model was used to recover the planet-to-star radius ratio of similar systematics simulations. These fits were performed separately from the linear basis and feature space fits described earlier, but used the same sets of parameters to generate the light curves. In order to recover the planet-to-star radius ratio, a squared exponential kernel was used as described in Section 3.3, using both the SE and SSE forms as outlined below. A uniform prior was used for the height scale, and a logarithmic prior was used for the inverse length scales, in other words we fitted for \( \log \eta \) instead of \( \eta \) as discussed in Section 3.3. A logarithmic prior was also tried for the height scale, but this was found to be too aggressive. For the majority of cases this did not affect the fits; however, it was found to produce more outliers in the distribution for \( N_\gamma \), presumably when the prior suppressed modelling of the systematics. Strictly speaking, a logarithmic prior should also be imposed on the height scale, but it proved difficult to check the details of individual fits over thousands of light curves, and we proceeded using uniform priors on \( \xi \). As the posterior distribution is no longer well approximated by a Gaussian distribution, an affine-invariant MCMC was used instead of the Metropolis–Hastings algorithm (Goodman & Weare 2010; Foreman-Mackey et al. 2013), with two chains computed for each fit and the same convergence tests were applied as before.

Table 2. Results from fitting 10 000 light curves simulated using the linear systematics model described in Section 4.1, i.e. a linear combination of three basis functions, and fitting using a combination of those three basis inputs and a further two inputs. The light curves were fitted using linear basis models (LBM), feature space (FS) expansions, and Laplacian approximations using shrinkage priors on the basis weights (LP). The feature space evidence was calculated using the BIC, AIC, and a Laplacian expansion with shrinkage. \( \chi^2_\nu \), the mean of \( \mathcal{N}_\chi \), the mean uncertainty (\( \sigma_\rho \)), mean accuracy (mean acc), and the rescaling of the uncertainties required to set \( \chi^2_\nu = 1 \) are reported for the various fitting methods.

<table>
<thead>
<tr>
<th>Fitting method</th>
<th>( \chi^2_\nu )</th>
<th>( \mathcal{N}_\chi )</th>
<th>( \sigma_\rho )</th>
<th>Mean acc</th>
<th>Rescaling (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBM: 2 inputs</td>
<td>8.76</td>
<td>-0.004</td>
<td>0.00044</td>
<td>0.00080</td>
<td>196.0</td>
</tr>
<tr>
<td>LBM: correct exp</td>
<td>1.05</td>
<td>-0.016</td>
<td>0.00054</td>
<td>0.00044</td>
<td>2.5</td>
</tr>
<tr>
<td>LBM: 5 inputs</td>
<td>1.02</td>
<td>-0.022</td>
<td>0.00082</td>
<td>0.00065</td>
<td>1.2</td>
</tr>
<tr>
<td>FS: best-fitting BIC</td>
<td>4.73</td>
<td>-0.022</td>
<td>0.00036</td>
<td>0.00054</td>
<td>117.6</td>
</tr>
<tr>
<td>FS: best-fitting AIC</td>
<td>3.38</td>
<td>-0.020</td>
<td>0.00041</td>
<td>0.00052</td>
<td>83.9</td>
</tr>
<tr>
<td>FS: best-fitting LP</td>
<td>3.29</td>
<td>-0.026</td>
<td>0.00039</td>
<td>0.00048</td>
<td>81.3</td>
</tr>
<tr>
<td>FS: marginalized BIC</td>
<td>1.40</td>
<td>-0.014</td>
<td>0.00055</td>
<td>0.00048</td>
<td>18.3</td>
</tr>
<tr>
<td>FS: marginalized AIC</td>
<td>1.06</td>
<td>-0.014</td>
<td>0.00059</td>
<td>0.00047</td>
<td>3.1</td>
</tr>
<tr>
<td>FS: marginalized LP</td>
<td>1.00</td>
<td>-0.015</td>
<td>0.00060</td>
<td>0.00046</td>
<td>-0.1</td>
</tr>
<tr>
<td>LP: 5 inputs</td>
<td>1.38</td>
<td>-0.016</td>
<td>0.00056</td>
<td>0.00050</td>
<td>17.5</td>
</tr>
</tbody>
</table>
Figure 2. Histograms of the number-of-sigma ($N_s$) statistic for a selection of the simulated light curves. The systematics model and the fitting methods are marked on the plots. The left-hand panels show the difference between using the best-fitting feature space expansion and marginalizing over the full feature space, showing that in all cases marginalization provides a more reliable recovery of the planet-to-star radius ratio. The right-hand panel shows the results for the GP fits to various systematics models, showing that in all cases the GP models provide a good model for the systematics.

In general the GP models all produce acceptable fits to the light curves, and provide reliable uncertainties. This is despite the GP models not always being ideal models for the data. First, the GP kernel (GP: SE kernel) as well as for the correct basis expansion (LBM: correct exp) to recover $\hat{\rho}$ and $\sigma_\rho$. The second set of simulated light curves was identical to the linear basis systematics described in Section 4.1. Here, a GP with SE kernels was used with both 3 and 5 of the basis inputs (GP: SE kernel 3/5 inputs). The final set of simulated light curves was identical to that described in Section 4.2 (with $N = 100$), using 2 basis inputs and an expansion for each selected at random from 1 to 3. These were fitted using a GP with both SE and SSE kernels (GP: SE/SSE kernel). The results are shown in Table 5, and a selection of $N_s$ distributions is shown in Fig. 2.
Table 3. Results from 10 000 simulated using the first set of simulations ($N = 100$) from the feature space systematics model, where two basis inputs were used. The basis expansion for each input was selected randomly between first and third order. These were fitted using similar models to those in Table 2.

<table>
<thead>
<tr>
<th>Fitting method</th>
<th>$\chi^2$</th>
<th>$\bar{N}_i$</th>
<th>$\bar{\sigma}_p$</th>
<th>Mean acc</th>
<th>Rescaling (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBM: linear exp</td>
<td>5.99</td>
<td>0.036</td>
<td>0.00046</td>
<td>0.00076</td>
<td>144.7</td>
</tr>
<tr>
<td>LBM: correct exp</td>
<td>1.04</td>
<td>0.004</td>
<td>0.00058</td>
<td>0.00046</td>
<td>2.2</td>
</tr>
<tr>
<td>LBM: 2nd-order exp</td>
<td>1.74</td>
<td>0.008</td>
<td>0.00060</td>
<td>0.00058</td>
<td>31.9</td>
</tr>
<tr>
<td>LBM: 3rd-order exp</td>
<td>1.05</td>
<td>0.007</td>
<td>0.00076</td>
<td>0.00062</td>
<td>2.7</td>
</tr>
<tr>
<td>FS: best-fitting BIC</td>
<td>3.64</td>
<td>0.008</td>
<td>0.00039</td>
<td>0.00053</td>
<td>90.7</td>
</tr>
<tr>
<td>FS: best-fitting AIC</td>
<td>2.65</td>
<td>0.006</td>
<td>0.00045</td>
<td>0.00055</td>
<td>62.7</td>
</tr>
<tr>
<td>FS: best-fitting LP</td>
<td>2.31</td>
<td>0.008</td>
<td>0.00040</td>
<td>0.00044</td>
<td>51.9</td>
</tr>
<tr>
<td>FS: marginalized BIC</td>
<td>1.36</td>
<td>0.002</td>
<td>0.00054</td>
<td>0.00048</td>
<td>16.5</td>
</tr>
<tr>
<td>FS: marginalized AIC</td>
<td>0.94</td>
<td>0.009</td>
<td>0.00065</td>
<td>0.00049</td>
<td>3.1</td>
</tr>
<tr>
<td>FS: marginalized LP</td>
<td>0.77</td>
<td>0.011</td>
<td>0.00071</td>
<td>0.00046</td>
<td>−12.1</td>
</tr>
<tr>
<td>LP: 2nd-order exp</td>
<td>1.93</td>
<td>0.001</td>
<td>0.00048</td>
<td>0.00049</td>
<td>38.9</td>
</tr>
<tr>
<td>LP: 3rd-order exp</td>
<td>1.55</td>
<td>0.002</td>
<td>0.00046</td>
<td>0.00044</td>
<td>24.3</td>
</tr>
</tbody>
</table>

Table 4. Same as Table 3 but for the second set of simulations using feature space systematics with $N = 200$.

<table>
<thead>
<tr>
<th>Fitting method</th>
<th>$\chi^2$</th>
<th>$\bar{N}_i$</th>
<th>$\bar{\sigma}_p$</th>
<th>Mean acc</th>
<th>Rescaling (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBM: linear exp</td>
<td>8.08</td>
<td>0.006</td>
<td>0.00030</td>
<td>0.00058</td>
<td>184.2</td>
</tr>
<tr>
<td>LBM: correct exp</td>
<td>1.01</td>
<td>0.007</td>
<td>0.00033</td>
<td>0.00026</td>
<td>0.6</td>
</tr>
<tr>
<td>LBM: 2nd-order exp</td>
<td>2.17</td>
<td>0.016</td>
<td>0.00034</td>
<td>0.00037</td>
<td>47.2</td>
</tr>
<tr>
<td>LBM: 3rd-order exp</td>
<td>1.01</td>
<td>0.001</td>
<td>0.00040</td>
<td>0.00032</td>
<td>0.7</td>
</tr>
<tr>
<td>FS: best-fitting BIC</td>
<td>2.67</td>
<td>0.007</td>
<td>0.00027</td>
<td>0.00031</td>
<td>63.3</td>
</tr>
<tr>
<td>FS: best-fitting AIC</td>
<td>1.80</td>
<td>0.005</td>
<td>0.00030</td>
<td>0.00031</td>
<td>34.0</td>
</tr>
<tr>
<td>FS: best-fitting LP</td>
<td>1.79</td>
<td>0.006</td>
<td>0.00027</td>
<td>0.00027</td>
<td>33.7</td>
</tr>
<tr>
<td>FS: marginalized BIC</td>
<td>1.36</td>
<td>0.004</td>
<td>0.00033</td>
<td>0.00029</td>
<td>16.5</td>
</tr>
<tr>
<td>FS: marginalized AIC</td>
<td>0.95</td>
<td>0.003</td>
<td>0.00037</td>
<td>0.00029</td>
<td>−2.5</td>
</tr>
<tr>
<td>FS: marginalized LP</td>
<td>0.91</td>
<td>0.001</td>
<td>0.00038</td>
<td>0.00027</td>
<td>−4.6</td>
</tr>
<tr>
<td>LP: 2nd-order exp</td>
<td>2.21</td>
<td>0.014</td>
<td>0.00031</td>
<td>0.00034</td>
<td>48.8</td>
</tr>
<tr>
<td>LP: 3rd-order exp</td>
<td>1.37</td>
<td>0.008</td>
<td>0.00031</td>
<td>0.00028</td>
<td>17.0</td>
</tr>
</tbody>
</table>

Table 5. Results for 10 000 light curves simulated using the range of models described in Section 4.3, and recovered using various GP systematics models.

<table>
<thead>
<tr>
<th>Fitting method</th>
<th>$\chi^2$</th>
<th>$\bar{N}_i$</th>
<th>$\bar{\sigma}_p$</th>
<th>Mean acc</th>
<th>Rescaling (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single input systematics model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LBM: correct exp</td>
<td>1.02</td>
<td>0.002</td>
<td>0.00037</td>
<td>0.00029</td>
<td>1.1</td>
</tr>
<tr>
<td>GP: SE kernel</td>
<td>0.84</td>
<td>0.004</td>
<td>0.00045</td>
<td>0.00033</td>
<td>−8.4</td>
</tr>
<tr>
<td>Linear systematics model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP: SE kernel (3 inputs)</td>
<td>1.21</td>
<td>0.000</td>
<td>0.00070</td>
<td>0.00057</td>
<td>10.2</td>
</tr>
<tr>
<td>GP: SE kernel (5 inputs)</td>
<td>0.87</td>
<td>0.000</td>
<td>0.00093</td>
<td>0.00064</td>
<td>−7.0</td>
</tr>
<tr>
<td>Feature space systematics model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GP: SE kernel</td>
<td>1.28</td>
<td>0.008</td>
<td>0.00063</td>
<td>0.00051</td>
<td>12.9</td>
</tr>
<tr>
<td>GP: SSE kernel</td>
<td>0.79</td>
<td>0.000</td>
<td>0.00072</td>
<td>0.00047</td>
<td>−11.3</td>
</tr>
</tbody>
</table>

used here is additive, and the true systematics models are multiplicative. Secondly, the SE kernel is not the correct choice for the models presented here, as we know a priori that the systematics models for each input basis are independent, whereas the SE kernel assumes the systematics are a joint function of the inputs, i.e. that all (relevant) inputs must be similar for the covariance to be high. This results in larger uncertainties than the corresponding marginalization over feature space, as is the case for the linear systematics models. The SSE kernel is the more appropriate choice for independent systematics, but it was not tried for the linear systematics model as the number of hyperparameters increased significantly and made inference over thousands of light curves infeasible. For the feature space systematics, the SSE kernel performs better, and provides almost identical results to the feature space marginalization with Laplacian approximation. This is despite it being far more flexible than even marginalization over feature space modes, and required only one posterior distribution to be computed. Forcing the systematics model to be a simple deterministic function of the basis inputs is perhaps a better method for fitting systematics when the true systematics model is contained in the tested model space. In general however, this is quite a rigid assumption, as the true systematics model is very unlikely to be precisely described by a simple polynomial expansion of the inputs, suggesting that GPs are the better choice as they provide greater flexibility and are easier...
to implement. The following section tests the use of deterministic models on a more general class of systematics models.

### 4.4 Stochastic systematics model

Now we have established the reliability of marginalization over feature space and stochastic processes (GPs) to recover transit parameters in the presence of linear basis models and higher order expansions of them, we might ask whether linear basis models and feature space expansions can reliably recover light-curve parameters when the systematics are not exactly described by a simple expansion of basis inputs. This is a more realistic scenario, as systematics are unlikely to be exactly described by such simple functions.

To answer this question, a series of 10 000 light curves was again generated from a basis set \( \mathbf{X} \), only now generating the systematics using a stochastic function of the inputs. For each simulated light curve, two basis functions were generated as before, and a GP with squared exponential kernel was used to generate independent systematics. This was so cross-terms in the feature space expansion did not need to be considered. For each of the two bases in \( \mathbf{X} \), a random draw was taken from a GP with a mean of 1 and squared exponential kernel function:

\[
k(x_i, x_j) = \exp \left( \frac{(x_i - x_j)^2}{2l^2} \right),
\]

with \( l \) equal to 2. This typically resulted in 1–2 turning points per draw, so the generated systematics should be fit reasonably well with low-order polynomial expansions of \( \mathbf{X} \). The two draws for each input were then multiplied together, scaled according to \( \sigma_r \), and multiplied to the transit model to generate the simulated light curve.

The planet-to-star radius ratio was then recovered as before, using a similar set of linear basis models, marginalization over feature space expansions of the inputs (to fourth-order expansion), Laplacian approximations with shrinkage priors (to fourth-order expansion), and GPs using both the SE and SSE kernels. The results are shown in Table 6, and example \( \mathcal{N}_r \) distributions is shown in the bottom panels of Fig. 2.

The results indicate that marginalization over feature space is yet again better than model selection, and delivers more reliable results. The marginalization using the AIC and Laplace approximations provides reliable inference of the transit depth, and the BIC underestimates the uncertainties as for the other systematics models. The most reliable method is now the GP with SSE kernel, which provides almost ideal statistics. The SE kernel also does a good job in recovering the transit depth, despite imposing that the systematics are a joint model of the inputs rather than an independent model. Ideally these tests would be extended to higher order feature spaces and more basis inputs, along with joint systematics models, but this would require too much computation over thousands of light curves.

### 5 Practical Implementation

These results show that marginalization over feature space models is the best way to recover light-curve parameters in the presence of instrumental systematics described by (independent) low-order polynomial expansions of basis inputs. This method can also recover reliable transit parameters when the systematics are not exactly given by a low-order polynomial of the basis inputs, but well approximated by it. However, these results were based on simulated light curves, and in practice several other considerations should be taken into account.

To compute the evidence required for marginalizing over feature space, both AIC\(_C\) and Laplacian approximations to the evidence are better than the more commonly used BIC. Technically, the Laplacian approximation should give a better estimate of the evidence, especially if the evidence is recomputed from the full distribution once the priors on the weight parameters are fixed, but the AIC\(_C\) is more convenient to calculate. The methods used here also have the added bonus of being calculable directly from a Levenberg–Marquardt optimization, which could significantly speed up computation time.

These results also show that methods employing model selection with BIC can commonly underestimate the uncertainties by ~100 per cent. However, this does not mean we should revise the uncertainties of all data extracted using such techniques by this amount, as in practice most authors take more care to select the systematics models and ensure that the model provides an acceptable fit to the data. For example, one could use the residuals from a simple white light-curve fit in order to inform the model selection and the required complexity. Furthermore, if the model selection criteria show a clear preference for one model over the others, then model selection and marginalization should produce identical results, and the choice of model selection criteria will become less important. As the quality of the data increases, model selection will also become easier, as was shown in Section 4.2. This means that for high-quality data such as STIS light curves (e.g. Sing et al. 2011), the best-fitting systematics model might be well constrained, and marginalization over many models is unnecessary and again might produce the same results.

Transmission and eclipse spectroscopy measurements also have the added bonus of multiple, simultaneous light curves from the

<table>
<thead>
<tr>
<th>Fitting method</th>
<th>( \chi^2 )</th>
<th>( \mathcal{N}_\rho )</th>
<th>( \sigma_r )</th>
<th>Mean acc</th>
<th>Rescaling (per cent)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LBM: linear exp</td>
<td>6.33</td>
<td>-0.000</td>
<td>0.00045</td>
<td>0.00076</td>
<td>151.7</td>
</tr>
<tr>
<td>LBM: 4th-order exp</td>
<td>1.36</td>
<td>-0.001</td>
<td>0.00090</td>
<td>0.00078</td>
<td>16.6</td>
</tr>
<tr>
<td>FS: best-fitting BIC</td>
<td>4.60</td>
<td>-0.009</td>
<td>0.00038</td>
<td>0.00057</td>
<td>114.4</td>
</tr>
<tr>
<td>FS: best-fitting AIC</td>
<td>3.31</td>
<td>-0.008</td>
<td>0.00045</td>
<td>0.00060</td>
<td>82.0</td>
</tr>
<tr>
<td>FS: best-fitting LP</td>
<td>3.60</td>
<td>-0.014</td>
<td>0.00039</td>
<td>0.00053</td>
<td>89.8</td>
</tr>
<tr>
<td>FS: marginalized BIC</td>
<td>1.78</td>
<td>-0.013</td>
<td>0.00054</td>
<td>0.00052</td>
<td>33.4</td>
</tr>
<tr>
<td>FS: marginalized AIC</td>
<td>1.23</td>
<td>-0.010</td>
<td>0.00065</td>
<td>0.00054</td>
<td>11.1</td>
</tr>
<tr>
<td>FS: marginalized LP</td>
<td>1.18</td>
<td>-0.012</td>
<td>0.00067</td>
<td>0.00051</td>
<td>8.7</td>
</tr>
<tr>
<td>LP: 4th-order exp</td>
<td>2.49</td>
<td>0.004</td>
<td>0.00049</td>
<td>0.00056</td>
<td>57.8</td>
</tr>
<tr>
<td>GP: sqexp/ARD kernel</td>
<td>1.29</td>
<td>0.038</td>
<td>0.00061</td>
<td>0.00049</td>
<td>13.8</td>
</tr>
<tr>
<td>GP: sqexp/SUM kernel</td>
<td>0.98</td>
<td>0.000</td>
<td>0.00071</td>
<td>0.00051</td>
<td>-1.2</td>
</tr>
</tbody>
</table>
same instrument. This should in principle make model selection easier. The current results show that more complex models are a more reliable way to model systematics, suggesting that authors should select the most complex systematics model required for any of the wavelength-dependent light curves to achieve more robust inference. Marginalization over multiple models is still probably a better approach, as we should not expect systematics to behave the same way in all wavelength channels, particularly if the dominant effects are related to pixel-to-pixel sensitivity variations, i.e. the systematics change spatially over the detector, which is the case for many instruments.

Simultaneous observations of multiple light curves also have the advantage of being able to remove common mode systematics. These are systematics that are common to all wavelength channels, and can be determined from analysis of the white light curve. In this case the results are conditioned on the common mode correction, but this is generally not problematic as long as additional systematics that are inserted into the light curves are taken into account during inference. Marginalization over a common mode correction could be achieved by simultaneously modelling all the light curves, although this is difficult in practice.

When the systematics do not follow a simple polynomial expansion of the basis inputs, then GPs become the method of choice for modelling instrumental systematics. This is unsurprising, seeing that the feature space expansion does not contain the correct systematics model, and therefore our inference is not always so reliable. In the case of GPs using a SE kernel (or SSE), we are in effect using an infinite number of basis functions, and can be sure that the true systematics model, or something very close, is contained within the model space.

The model selection criteria used in this paper also rely on the posterior distribution being well approximated by a Gaussian distribution. This may not always be the case, for example if we fit for all of the transit parameters, or when using other light-curve models. One could use more sophisticated methods to calculate the evidence, but care must be taken to ensure that the priors are selected in an objective manner. When the posterior is no longer well approximated by a Gaussian, the Gaussian mixture model used to marginalize over the feature space models may also be inappropriate. This is certainly an advantage for GPs, which do not rely on calculation of the model evidence, and should be applicable to highly non-Gaussian posterior distributions. Arguably, kernels still need to be selected for GPs, which could be performed with model evidence and even marginalization over many kernels, but the flexibility in even the simplest GP kernels makes this less critical.

Nevertheless, marginalization over a high enough order feature space can still provide reliable results. This could become prohibitively expensive as the order of expansion and the number of basis inputs increases, particularly if one wants to include cross-terms in the feature space expansion. GP models are therefore a better choice of systematics model if \( N \) is small enough for GPs to remain computationally feasible. This is typically for \( N \lesssim 1000 \).

Unfortunately for GPs, the computational complexity scales with \( O(N^3) \). This leaves feature space models, with complexity scaling as \( O(N) \), as a more feasible approach for large \( N \). They are therefore applicable to larger data sets such as Spitzer light curves without the necessity to bin the light curves, but going to high-order feature space will still be challenging.

Another important consideration in using either feature space or GP models is whether the systematics are well described by a function of the available inputs. For example, the dominant systematic could be a function of a detector property that we have no measurement of (or too noisy a measurement). In this case, the additional systematics could be modelled as a time-dependent function, and could be taken into account by using a stochastic process. For GPs this would require adding a time-dependent component to the kernel, and for feature space models we could use the wavelet method of Carter & Winn (2009) to compute the likelihood for each model.

Finally, it is worth noting that no systematics modelling technique will provide perfect, reliable statistics in all cases. Rather than picking a single technique to extract transit parameters, the best approach is always to use multiple techniques and test whether the inference is dependent on subjective choices made. Where multiple epochs of data exist for the same planetary system, either with the same instrumental set-up or with overlapping wavelength regions, common techniques should be tested to validate their use on each instrument.

### 6 Conclusions

This study has investigated the use of common model selection criteria used in the analysis of exoplanet light curves in the presence of instrumental noise, which remains of central importance in our understanding of exoplanet atmospheres. The use of marginalization over multiple, parametric systematics models was introduced, and demonstrated on a series of transit light curves with known systematics functions. This is shown to be a much more reliable method to recover light-curve parameters than simple model selection. Of particular note, the commonly used BIC is shown to be the worst model selection criteria out of those tested, and can result in underestimated uncertainties if care is not taken during inference, therefore, low significance claims using such techniques should be treated with caution.

GP models were also discussed in the context of model selection, and their use was demonstrated on similar sets of simulated transit light curves. They were shown to be an extremely flexible and reliable method to recover transit parameters, even when the kernel used is not the most appropriate. Given the flexibility of GP models, these are perhaps the best method to used when the number of data points is relatively small (\( N \lesssim 1000 \)). Feature space models however can also provide reliable inference when the systematics are well approximated by low-order polynomials of the input basis set, and are therefore more easily applied to large data sets such as Spitzer light curves.

The technique of marginalizing over systematics models contributes to the growing number of statistical tools that we can use to infer parameters from light curves. The choice of systematics model and model selection methods are clearly critical to performing reliable inference. We should therefore take a critical look at our choices of systematics models, and assess how sensitive our results are to subjective decisions. If the results change with subjective decisions, then we should not trust the light-curve parameters, or indeed any extraction of atmospheric data based on them. Furthermore, the choice of systematics model as well as the parameters of the systematics model will always contribute to the error budget; therefore, pushing to the photon limit is not feasible in the presence of significant systematics. When in doubt, an extra ‘layer’ of Bayesian inference can be used to marginalize over our ignorance of the systematics to reflect our lack of understanding, and allows us to perform more objective inference. Re-analysing and re-observing targets, in particular using different instruments, are invaluable in assessing the validity of the most commonly used systematics models, and therefore in re-evaluating what we truly know about exoplanet atmospheres.
ACKNOWLEDGEMENTS

I am extremely grateful to Suzanne Aigrain and the anonymous referee for a careful reading of the manuscript, and for insightful comments that significantly improved the clarity of the paper.

REFERENCES

Hurvich C. M., Tsai C.-L., 1989, Biometrika, 76, 297

This paper has been typeset from a \TeX/\LaTeX\ file prepared by the author.