Gaussian Processes for Bayesian hypothesis tests on regression functions

Alessio Benavoli **Francesca Mangili**

Dalle Molle Institute for Artificial Intelligence (IDSIA)
SUPSI-USI, Lugano, Switzerland

Abstract

Gaussian processes have been used in different application domains such as classification, regression etc. In this paper we show that they can also be employed as a universal tool for developing a large variety of Bayesian statistical hypothesis tests for regression functions. In particular, we will use GPs for testing whether (i) two functions are equal; (ii) a function is monotone (even accounting for seasonality effects); (iii) a function is periodic; (iv) two functions are proportional. By simulation studies, we will show that, beside being more flexible, GP tests are also competitive in terms of performance with state-of-art algorithms.

1 Introduction

Gaussian processes (GPs) have found widespread use in machine learning, in different application domains such as classification, regression etc. [O'Hagan and Kingman, 1978, Neal, 1998, MacKay, 1998, Rasmussen and Williams, 2006, Rasmussen, 2011, Gelman et al., 2013]. The reason of such success can be attributed to the great modeling flexibility of GPs. The aim of this paper is to show that, because their flexibility, GPs can also be employed as a *universal tool* for developing a large variety of Bayesian statistical hypothesis tests. In particular, "as a proof of concept", we will show how GPs can be used for testing whether (i) two functions are *equal*; (ii) a function is *monotone* (even accounting for *seasonality* effects); (iii) a function is *periodic*; (iv) two functions are *proportional* (focusing on the proportionality of the intensity functions of two counting processes). To develop such universal tool, we follow a Bayesian estimation approach: any decision is based on the posterior distribution. This means that we place the GP as a prior distribution on the unknown f and we determine the posterior distribution of f given the observations. Once we have obtained the posterior distribution, we can perform different hypothesis tests about f by simply asking different questions to the posterior. Besides these advantages of GPs in terms of expressiveness and flexibility, we show that our Bayesian estimation based equality, monotonicity, periodicity and proportionality tests are competitive in terms of power when compared with the sate-of-art algorithms. After briefly introducing GPs, we illustrate how to theoretically devise these tests by exploiting the properties of GPs and Bayesian decision making. Then, we assess the performance (power) of these tests through simulation studies. For the equality test, we compare our GP test with two nonparametric frequentist methods [Neumeyer et al., 2003, Pardo-Fernández et al., 2007] and with a Bayesian test based on regression splines [Behseta and Kass, 2005], obtaining, on average, better accuracy. For the monotonicity test, we compare the GP based method with four non-Bayesian methods: [Zheng, 1996, Bowman et al., 1998, Baraud et al., 2005, Akakpo et al., 2014] and three Bayesian methods (based on Bayes factors) [Salomond, 2014, Dunson, 2005, Scott et al., 2013]. Our simulation study shows that our GP method achieves the same average accuracy of the best among these algorithms on standard benchmark functions. Moreover we will show that, while the aforementioned methods for monotonicity estimation have not been designed to account for the presence of seasonality (i.e., a periodic disturbance that affects the monotonic component), thanks to the flexibility GPs, our monotonicity test can be modified to account for seasonality disturbance. We develop a this method that removes the seasonality effects and test the monotonicity of the remaining non-seasonal component with the classical Mann-Kendall

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

test for monotonic trend with seasonality correction (KS), obtaining very similar performance, although KS requires the period of the seasonal component to be known, while our GP method estimates it from data. For the periodicity test, we compare the GP test for period detection with the classical Fisher's significance test for periodic components. Also in this case we prove by simulation that our method is competitive. Finally, we show that the GP test for the proportionality of intensity functions has much larger accuracy than the traditional test based on the Schoenfeld residuals [Grambsch and Therneau, 1994].

2 Gaussian Process

Consider the regression model

$$y = f(x) + v, \tag{1}$$

where $x \in \mathcal{X} \subseteq \mathbb{R}$, $f : \mathbb{R} \to \mathbb{R}$ and $v \sim N(0, \sigma^2)$, and assume that we observe the training data (x_i, y_i) for $i = 1, \ldots, n$. Our goal is to employ these observations to make inferences about the unknown function f. Following the Bayesian estimation approach, we place a prior distribution on the unknown f, and employ the observations to compute the posterior distribution of f; finally we use this posterior to make inferences about f. Since f is a function, the Gaussian process is a natural prior distribution for f [MacKay, 1998, Rasmussen and Williams, 2006]. Let $GP(0, k_{\boldsymbol{\theta}})$ denote a GP with zero mean function and covariance function $k_{\boldsymbol{\theta}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^+$, which depends on a vector of hyperparameters $\boldsymbol{\theta}$. If $f \sim GP(0, k_{\boldsymbol{\theta}})$, then, for any fixed m points $\mathbf{x}^* = [x_1^*, \ldots, x_m^*]^T$, the vector $\mathbf{f}^* = [f(x_1^*), \ldots, f(x_m^*)]^T$ is Gaussian distributed:

$$p(\mathbf{f}^*|\mathbf{x}^*, \boldsymbol{\theta}) = N(\mathbf{f}^*; \mathbf{0}, K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*)), \tag{2}$$

with zero mean and covariance matrix $K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*) = [k_{\boldsymbol{\theta}}(x_i^*, x_j^*)]_{ij}$ for each $i, j = 1, \ldots, m$. Consider a set of n inputs $\mathbf{x} = [x_1, \ldots, x_n]^T$ and a vector of noisy output data $\mathbf{y} = [y_1, \ldots, y_n]^T$. Based on the training data (x_i, y_i) for $i = 1, \ldots, n$, and given a test input $\mathbf{x}^*$, we wish to find the posterior distribution of $\mathbf{f}^* = [f(x_1^*), \ldots, f(x_m^*)]^T$. From (1) and the properties of the Gaussian distribution, it follows that [Rasmussen and Williams, 2006, Sec. 2.2]:

$$\begin{bmatrix} \mathbf{y} \\ \mathbf{f}^* \end{bmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{bmatrix} K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I} & K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}^*) \\ K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}) & K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*) \end{bmatrix}\right). \tag{3}$$

Hence, the posterior distribution of $\mathbf{f}^*$ is

$$p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}, \sigma^2) = N(\mathbf{f}^*; \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}^*|\mathbf{x}, \mathbf{y}), \hat{K}_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*|\mathbf{x})), \tag{4}$$

with mean and covariance given by:

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}^*|\mathbf{x}, \mathbf{y}) = K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x})(K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} \mathbf{y}, \tag{5}$$

$$\begin{aligned} \hat{K}_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*|\mathbf{x}) = K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*) \\ - K_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x})(K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}^*). \end{aligned} \tag{6}$$

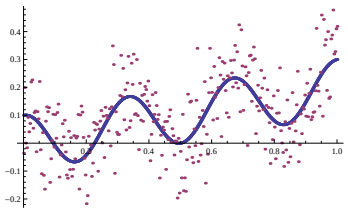

Figure 1: Function $f(x) = \frac{x}{5} + \frac{1}{10} cos(6\pi x)$ (blue) and its noisy observations $y_i = f(x_i) + v_i$ (red) with $x_i = \frac{i}{300}$, $v_i = N(0, 0.1^2)$, for $i = 1, \ldots, 300$.

For ease of notation, hereafter we will omit $\boldsymbol{\theta}, \sigma^2$ in $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}, \boldsymbol{\theta}, \sigma^2)$. Once we have computed $p(\mathbf{f}^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y})$ we can make any inference about $\mathbf{f}^*$.

2.1 Kernels, composition and marginalization

GP models use a kernel to define the covariance between any two function values: $Cov(f(x), f(x^*)) = k_{\boldsymbol{\theta}}(x, x^*)$. The kernel specifies which functions are likely under the GP prior. Commonly used kernels families include the squared exponential (SE), periodic (PE), constant-linear-quadratic (QD):
QD: $k_{\boldsymbol{\theta}}(x_1, x_2) = s_0 + s_1 x_1 x_2 + s_2 x_1^2 x_2^2$,
SE: $k_{\boldsymbol{\theta}}(x_1, x_2) = \sigma_s^2 \exp(-0.5(x_1 - x_2)^2/\ell_s^2)$,
PE: $k_{\boldsymbol{\theta}}(x_1, x_2) = \sigma_p^2 \exp(-2\sin(\pi(x_1 - x_2)/p_e)^2/\ell_p^2)$,
with hyperparameters $s_i > 0$, $\sigma_s, \ell_s > 0$ and, respectively, $\sigma_p, \ell_p > 0$ with period p_e. Positive semidefinite kernels (i.e. those which define valid covariance functions) are closed under addition and multiplication. This allows one to create richly structured and interpretable kernels by kernel composition. In this paper, we will focus on kernel summation. By summing kernels, we can model the data as a superposition of independent functions. For instance, in time series models, sums of kernels can express superposition of different processes, possibly operating at different scales. A typical example is a monotonic increasing time series with seasonal variation, in which a monotonic (linear) increasing function f_a is superposed to a periodic function f_b, as shown in the example in Fig. 1. To model a superposition of two (or more) functions, we can assume two (or more) independent GP priors for f_a, f_b, i.e., $f_a \sim GP(0, k_{\boldsymbol{\theta}_a}^a)$, $f_b \sim GP(0, k_{\boldsymbol{\theta}_b}^b)$, then $f = f_a + f_b \sim GP(0, k_{\boldsymbol{\theta}_c}^c) = GP(0, k_{\boldsymbol{\theta}_a}^a + k_{\boldsymbol{\theta}_b}^b)$ with $\boldsymbol{\theta}_c = (\boldsymbol{\theta}_a, \boldsymbol{\theta}_b)$. Moreover, assume we have determined the posterior distribution of f but that we are interested on only, say, phenomenon f_a, we can consider the f_b component as a disturbance and evaluate the predictive distribution for f_a only, which is:

$$p(\mathbf{f}_a^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = N(\mathbf{f}_a^*; \hat{\boldsymbol{\mu}}_a(\mathbf{x}^*|\mathbf{x}, \mathbf{y}), \hat{K}_a(\mathbf{x}^*, \mathbf{x}^*|\mathbf{x})), \tag{7}$$

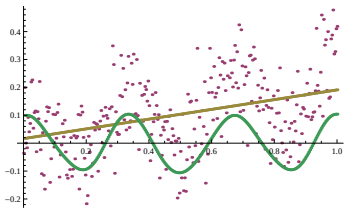

Figure 2: Estimate of the quadratic or periodic component for the example in Fig. 1.

with mean and covariance given by:

$$\hat{\boldsymbol{\mu}}_a(\mathbf{x}^*|\mathbf{x},\mathbf{y}) = K^a_{\boldsymbol{\theta}_a}(\mathbf{x}^*,\mathbf{x})(K^c_{\boldsymbol{\theta}_c}(\mathbf{x},\mathbf{x}) + \sigma^2\mathbf{I})^{-1}\mathbf{y}, \quad (8)$$

$$\hat{K}_a(\mathbf{x}^*,\mathbf{x}^*|\mathbf{x}) = K^a_{\boldsymbol{\theta}_a}(\mathbf{x}^*,\mathbf{x}^*) - K^a_{\boldsymbol{\theta}_a}(\mathbf{x}^*,\mathbf{x})(K^c_{\boldsymbol{\theta}_c}(\mathbf{x},\mathbf{x}) + \sigma^2\mathbf{I})^{-1}K^a_{\boldsymbol{\theta}_a}(\mathbf{x},\mathbf{x}^*). \quad (9)$$

Fig. 2 shows the mean $\hat{\boldsymbol{\mu}}_a(\mathbf{x}^*|\mathbf{x},\mathbf{y})$ ($\hat{\boldsymbol{\mu}}_b(\mathbf{x}^*|\mathbf{x},\mathbf{y})$ if instead we focus on the periodic component) for the example of Fig. 1, when $k^a_{\boldsymbol{\theta}_a}$ is the quadratic kernel and $k^b_{\boldsymbol{\theta}_b}$ the periodic kernel.

2.2 Determining the hyperparameters

Once we have selected a kernel or a particular kernel composition, we must determine the values of the hyperparameters $\boldsymbol{\theta}$ and the variance σ^2. The proper Bayesian procedure is to choose a prior for $\boldsymbol{\theta}$ and σ^2 and then determine the posterior distribution of the quantities of interest. For instance, inferences on $\mathbf{f}^*$ can be carried out by marginalizing out $\boldsymbol{\theta}$ and σ^2:

$$p(\mathbf{f}^*|\mathbf{x}^*,\mathbf{x},\mathbf{y}) = \int p(\mathbf{f}^*|\mathbf{x}^*,\mathbf{x},\mathbf{y},\boldsymbol{\theta},\sigma^2)dP(\boldsymbol{\theta},\sigma^2).$$

No closed form solution exists for $p(\mathbf{f}^*|\mathbf{x}^*,\mathbf{x},\mathbf{y})$ or for the posterior of the hyperparameters and, therefore, inferences must be computed numerically by Markov Chain Monte Carlo methods (MCMC). The convergence of MCMC methods can be quite slow when the dimension of $\boldsymbol{\theta}$ is high and, therefore, when we are not interested in the posterior distribution of $\boldsymbol{\theta},\sigma^2$, we can approximate the marginal of $\mathbf{f}^*$ with (4) by using the maximum a-posteriori (MAP) estimate for the values of $\boldsymbol{\theta},\sigma^2$. This means that instead of performing MCMC we maximize w.r.t. $\boldsymbol{\theta}$ and σ^2 the joint marginal probability of $\mathbf{y},\boldsymbol{\theta},\sigma^2$, whose logarithm can be computed analytically [Rasmussen and Williams, 2006, Ch.2]:

$$\begin{aligned} L(\mathbf{y},\boldsymbol{\theta},\sigma^2|\mathbf{x}) = \; & -\tfrac{1}{2}\mathbf{y}^T(K_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}) + \sigma^2\boldsymbol{I})^{-1}\mathbf{y} \\ & -\tfrac{1}{2}\log|K_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}) + \sigma^2\boldsymbol{I}| \\ & -\tfrac{n}{2}\log 2\pi + \log p(\boldsymbol{\theta},\sigma^2). \end{aligned} \quad (10)$$

The values of $\boldsymbol{\theta},\sigma^2$ can then be determined by maximizing this score. Unfortunately, optimizing over parameters is not a convex optimization problem, and the space can have many local optima. To tackle this problem, we have used a global search algorithm based on the algorithm developed by Ugray et al. [2007] and implemented in MATLAB [2013] by the function "GlobalSearch".

3 Equality test

An equality test is used to decide whether two regression functions are equal. In particular, our aim is to compare two regression functions f_1 and f_2 given the two independent samples $(\mathbf{x}^{(1)},\mathbf{y}^{(1)})$ and $(\mathbf{x}^{(2)},\mathbf{y}^{(2)})$ of, respectively, n_1 and n_2 observations. Nonparametric frequentist tests for the equality of regression curves are described in [Neumeyer et al., 2003, Pardo-Fernández et al., 2007], while a Bayesian test based on regression splines is presented in [Behseta and Kass, 2005]. We assume that the covariates $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ have the same support $\mathcal{X}$. Our aim is to develop an equality test using GPs. A way to devise such test is to assume the same GP prior $GP(0,k_{\boldsymbol{\theta}})$ for the two functions f_1 and f_2 and compute the posterior marginal GPs $p(\mathbf{f}^*_1|\mathbf{x}^*,\mathbf{x}^{(1)},\mathbf{y}^{(1)})$ and $p(\mathbf{f}^*_2|\mathbf{x}^*,\mathbf{x}^{(2)},\mathbf{y}^{(2)})$ at the $n = n_1 + n_2$ test inputs $\mathbf{x}^* = [\mathbf{x}^{(1)},\mathbf{x}^{(2)}]$, which are Gaussian and given by (4). In this way, the equality of the two functions is tested at the covariates of the observations, that is, where we have the experimental evidence. Note that the two posteriors share the same hyperparameters and test inputs. The hyperparameters are determined by the MAP approach described in Section 2.2. Assuming that f_1 and f_2 are independent Gaussian processes, we have that $p(\mathbf{y}^{(1)},\mathbf{y}^{(2)}|f_1,f_2) = p(\mathbf{y}^{(1)}|f_1)p(\mathbf{y}^{(2)}|f_2)$ and thus the logarithm of the joint marginal of $\mathbf{y}^{(1)},\mathbf{y}^{(2)},\boldsymbol{\theta},\sigma^2$ is equal to $L(\mathbf{y}^{(1)},\boldsymbol{\theta},\sigma^2|\mathbf{x}^{(1)}) + L(\mathbf{y}^{(2)},\boldsymbol{\theta},\sigma^2|\mathbf{x}^{(2)})$, where $L(\mathbf{y},\boldsymbol{\theta},\sigma^2|\mathbf{x})$ is given in (10). Let us denote the means of the posterior distributions of $\mathbf{f}^*_1$ and $\mathbf{f}^*_2$ as $\hat{\boldsymbol{\mu}}^{*(1)}$, $\hat{\boldsymbol{\mu}}^{*(2)}$ and their covariance matrices as $\hat{K}^{*(1)}$, $\hat{K}^{*(2)}$. Since the difference of two Gaussian variables are Gaussian, it follows that the posterior of $\Delta\mathbf{f}^* = \mathbf{f}^*_1 - \mathbf{f}^*_2$ is also Gaussian with mean $\Delta\hat{\boldsymbol{\mu}}^* = \hat{\boldsymbol{\mu}}^{*(1)} - \hat{\boldsymbol{\mu}}^{*(2)}$ and covariance matrix $\hat{K}^*_\Delta = \hat{K}^{*(1)} + \hat{K}^{*(2)}$. Then, we say that the two functions are equal with posterior probability $1-\alpha$ if the credible region for $\Delta\mathbf{f}^*$ includes the zero vector or, in other words, if:

$$(\Delta\hat{\boldsymbol{\mu}}^*)^T(\hat{K}^*_\Delta)^{-1}\Delta\hat{\boldsymbol{\mu}}^* \leq \chi^2_\nu(1-\alpha), \quad (11)$$

where $\chi^2_\nu(1-\alpha)$ is the $(1-\alpha)$-quantile of a Chi-squared distribution with ν degrees of freedoms and ν is the number of positive eigenvalues of $\hat{K}^*_\Delta$. Indeed, as the number n of test inputs is likely to be considerably larger than the dimensionality of the covariance function, the matrix $\hat{K}^*_\Delta$ is not full rank. Thus, we decompose it as PDP^T, where D is the diagonal matrix

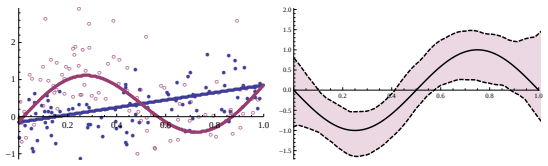

Figure 3: Left: functions f_1 (blue) and f_2 (red) and corresponding noisy observations. Right: estimated credible region for $f_1 - f_2$ (dashed lines) and its true value (continuous line).

of the eigenvalues $\lambda_1, \ldots, \lambda_n$ (sorted in descending order), and retain only the sub-matrices $P_\nu D_\nu P_\nu^T$ corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_\nu$ which verify the condition $\lambda_{\nu+1} / \sum_{i=1}^{n} \lambda_i < \epsilon$, where ϵ is a small, positive constant. For the experiments of this paper we have used $\epsilon = 0.01$.

Fig. 3 shows the estimated credible region of $f_1 - f_2$ when $f_1(x) = x$ and $f_2(x) = x + sin(2\pi x)$ and $n_1 = n_2 = 100$. As the region does not include the zero function, we can conclude that f_1 and f_2 are different.

4 Monotonicity test

A continuously differentiable function $f : \mathcal{X} \to \mathbb{R}$ on a closed interval $\mathcal{X}$ is said to be monotonically increasing (or decreasing) in $\mathcal{X}$ if $f_d(x') = \frac{df}{dx}(x') > 0$ (or $\frac{df}{dx}(x') < 0$) for each $x' \in \mathcal{X}$. Without loss of generality, we will focus on monotonically increasing functions. Our goal is to employ the training data to test the positive monotonicity of f based on its first derivative. Monotonicity tests based on the derivative have been proposed by [Hall and Heckman, 2000, Ghosal et al., 2000]. Assuming as prior on f the Gaussian Process $GP(0, k_{\boldsymbol{\theta}})$, we compute the posterior of $\frac{df}{dx}$ given the training data and test inputs. Since differentiation is a linear operator, the derivative of a GP is another GP, whose mean and covariance functions can be computed analytically [Rasmussen, 2003, Solak et al., 2003, Riihimäki and Vehtari, 2010]:

Theorem 1. *Assume that $f \sim GP(0, k_{\boldsymbol{\theta}})$ and that $k_{\boldsymbol{\theta}}$ is differentiable, then it follows that*

$$p(\mathbf{f}_d^*|\mathbf{x}^*, \mathbf{x}, \mathbf{y}) = N\Big(\mathbf{f}_d^*; \hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}^*|\mathbf{x}, \mathbf{y}), \hat{K}_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*|\mathbf{x})\Big) \quad (12)$$

where $\mathbf{f}_d^* = [\frac{df}{dx}(x_1^*), \ldots, \frac{df}{dx}(x_m^*)]^T$,

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}^*|\mathbf{x}, \mathbf{y}) = K_{\boldsymbol{\theta}}^d(\mathbf{x}^*, \mathbf{x})(K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1}\mathbf{y}, \quad (13)$$

$$\hat{K}_{\boldsymbol{\theta}}(\mathbf{x}^*, \mathbf{x}^*|\mathbf{x}) = K_{\boldsymbol{\theta}}^d(\mathbf{x}^*, \mathbf{x}^*) - K_{\boldsymbol{\theta}}^d(\mathbf{x}^*, \mathbf{x})(K_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{x}) + \sigma^2 \mathbf{I})^{-1} K_{\boldsymbol{\theta}}^d(\mathbf{x}^*, \mathbf{x})^{\mathrm{T}}, \quad (14)$$

with $K_{\boldsymbol{\theta}}^d(\mathbf{x}^*, \mathbf{x}) = [\frac{\partial}{\partial x_a} k_{\boldsymbol{\theta}}(x_a, x_j)\big|_{x_a = x_i^*}]_{ij}$, *and* $K_{\boldsymbol{\theta}}^d(\mathbf{x}^*, \mathbf{x}^*) = [\frac{\partial^2}{\partial x_a \partial x_b} k_{\boldsymbol{\theta}}(x_a, x_b)\big|_{x_a = x_i^*, x_b = x_l^*}]_{il}$ *for* $i, l = 1, \ldots, m$ *and* $j = 1, \ldots, n$. ■

Thus, we can use GPs to make inferences about derivatives and in particular test the monotonicity of f on $\mathbf{x}^* = \mathbf{x}$ (again, the test is performed at the observations covariates). First, we define a loss function for each decision:

$$L(\mathbf{f}_d^*, a) = \begin{cases} C_0 I_{\{\mathbf{f}_d^* > 0\}} & \text{if } a = 0, \\ C_1 I_{\{\mathbf{f}_d^* \not> 0\}} & \text{if } a = 1. \end{cases} \quad (15)$$

where the notation $\mathbf{f}_d^* > 0$ ($\mathbf{f}_d^* \not> 0$) indicates that all (not all) values in $\mathbf{f}_d^*$ are larger than 0, and where C_0 and C_1 are the losses we incur, respectively, by wrongly taking action $a = 0$ (i.e., declaring that $\mathbf{f}_d^* \not> 0$ when actually $\mathbf{f}_d^* > 0$), and by wrongly taking action $a = 1$ (i.e., declaring that $\mathbf{f}_d^* > 0$ when actually $\mathbf{f}_d^* \not> 0$).

Second, we compute the expected value of this loss given the training data and the test inputs $\mathbf{x}^*$. The expected loss is given by:

$$E[L(\mathbf{f}_d^*, a)] = \begin{cases} C_0 P[\mathbf{f}_d^* > 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}] & \text{if } a = 0, \\ C_1 P[\mathbf{f}_d^* \not> 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}] & \text{if } a = 1, \end{cases}$$

where we have exploited the fact that $E[I_{\{A\}}] = P[A]$. Thus, we choose $a = 1$ if

$$\begin{aligned} &C_0 P[\mathbf{f}_d^* > 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}] \leq C_1 P[\mathbf{f}_d^* \not> 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}] \\ &\text{equiv. } P[\mathbf{f}_d^* > 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}] > \tfrac{C_1}{C_1 + C_0}, \end{aligned} \quad (16)$$

or $a = 0$ otherwise. In the above derivation, we have exploited the fact that $P[\mathbf{f}_d^* \not> 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}] = 1 - P[\mathbf{f}_d^* > 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}]$. When the last inequality in (16) is satisfied, we can declare that $f_d > 0$ with probability $\frac{C_1}{C_1 + C_0}$. For comparison with the traditional test we may take $\frac{C_1}{C_0 + C_1} = 1 - \alpha$ with $\alpha = 0.05$; notice however that, while in the traditional tests a principled way of choosing α is lacking, in this GP based test the use of a Bayesian approach allows setting the decision rule in a more informed way based on the losses C_0 and C_1 expected for Type I and II errors. The probability $P[\mathbf{f}_d^* > 0|\mathbf{x}^*, \mathbf{x}, \mathbf{y}]$ can be computed by Monte Carlo (MC) sampling many vectors $\mathbf{f}_d^*$ from (12) and computing the proportion of runs in which the condition $\mathbf{f}_d^* > 0$ is satisfied.

4.1 Accounting for Seasonality

In time series analysis, we must often deal with seasonality effects, i.e., the function of interest may be the superposition of a monotonic and a periodic function, (e.g., $x/5 + (1/10)\cos(6\pi x)$). This composition is clearly non-monotonic. However, we may interpret the periodic function as a seasonal component, i.e., a periodic disturbance that affects the non-seasonal component (in the example $x/5$). In these cases, it is of

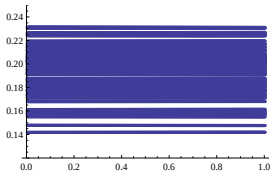

Figure 4: Samples of $\mathbf{f}_d^*$ for the non-periodic component $(x/5)$ of the example of Fig. 1.

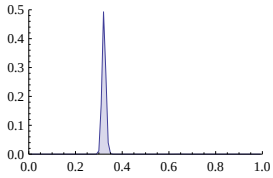

Figure 5: Posterior distribution of p_e computed on a grid of 100 elements.

interest to develop a statistical method that removes the seasonality effects and test the monotonicity of the remaining non-seasonal component. The development of such test is immediate with GPs. We can simply include a periodic kernel in the GP, i.e., $f = f_a + f_b$ with $f_a \sim GP(0, k^a_{\boldsymbol{\theta}_a})$ and $f_b \sim GP(0, k^b_{\boldsymbol{\theta}_b})$ (where $k^b_{\boldsymbol{\theta}_b}$ is the periodic kernel). Then, we determine the posterior distribution of the function f, remove the periodic component f_b as a disturbance and evaluate the posterior distribution of the non periodic components only, i.e., f_a, as discussed at the end of Section 2.1. Finally, we use this posterior to perform the monotonicity test on the non seasonal component. Fig. 4 shows samples of $\mathbf{f}_d^*$ (with $\mathbf{x}^* = \mathbf{x}$) for the example of Fig. 1, when $k^a_{\boldsymbol{\theta}_a}$ is the quadratic kernel, $k^b_{\boldsymbol{\theta}_b}$ the periodic kernel. The periodic component, considered a disturbance, has been removed, and thus all the derivatives of $\mathbf{f}_d^*$ are constant and distributed around $\frac{1}{5}$, that is the actual derivative of the non seasonal component $f_{ns}(x) = \frac{x}{5}$. As all derivatives are positives, we can declare with probability ≈ 1 that $f_{ns}(x)$ is monotone increasing in $[0, 1]$.

5 Periodicity test

In this case our goal is to test if a function $f : \mathcal{X} \to \mathbb{R}$ on a closed interval $\mathcal{X}$ (w.l.o.g. we can take $\mathcal{X} = [0, 1]$) is periodic based on noisy observations of f. We can detect the periodicity of the function only if its period is less than half of the range of x, that is 0.5 as $\mathcal{X} = [0, 1]$. To perform this test, we use a GP with only the periodic kernel. By defining a loss function similar to that in Section 4, we declare that the function is periodic if

$$P\left[p_e < 0.5|\mathbf{x}, \mathbf{y}\right] \geq \tfrac{C_1}{C_1+C_0},$$

that is, if the posterior probability that the period hyperparameter p_e of the periodic kernel is less than 0.5 is greater than $\frac{C_1}{C_1+C_0}$. The posterior of the period p_e can be obtained by MCMC sampling from the posterior obtained from the joint $\exp[L(\mathbf{y}, \boldsymbol{\theta}, \sigma^2|\mathbf{x})]$, where $L(\mathbf{y}, \boldsymbol{\theta}, \sigma^2|\mathbf{x})$ is given in (10). Fig. 5 shows the posterior distribution of p_e for the case in which the observations are generated according to $y_i = (1/10)cos(6\pi x_i) + v_i$ with $x_i = i/100$, $v_i = N(0, 0.2^2)$, for $i = 1, \ldots, 100$, and we use a uniform prior in $[0, 1]$ for p_e and weak priors for the other parameters of the periodic kernel. The maximum of p_e is on $1/3$ which is the true period of the function.

6 Proportional intensity test

In this case we are interested in testing the proportionality of the intensity function of counting processes based on counts data generated by them. Let us assume that the data are generated by a Poisson process whose intensity $\lambda(t, x)$ is a function of time t and of a covariate x. Then, the number of counts y in the time interval $[t, t + \Delta t]$ has a Poisson distribution with parameter $\Lambda(t, x) = \int_t^{t+\Delta t} \lambda(\tau, x) d\tau$. The proportionality assumption, which is widely used to model $\lambda(t, x)$, states that x has a multiplicative effect on the intensity and implies that

$$\Lambda(t, x) = \Lambda_0(t) e^{f(x)} \tag{17}$$

where $\Lambda_0(t)$ is a baseline function representing the time dependence, whereas $f(x)$ represents the dependence on the covariate x. Proportional intensity is a strong assumption which is not always necessarily reasonable and needs to be checked. Popular proportional intensity tests are based on the Schoenfeld residuals [Grambsch and Therneau, 1994]. Here we focus on the case where x is a categorical variable with two possible values x_1 and x_2 and we test whether the intensity functions $\lambda_1(t) = \lambda(t, x_1)$ and $\lambda_2(t) = \lambda(t, x_2)$ are proportional. This assumption implies the equality of the derivatives of $f_1(t) = \log[\Lambda_1(t)]$ and $f_2(t) = \log[\Lambda_2(t)]$, since from (17) we have:

$$\frac{d[\log \Lambda_1(t) - \log \Lambda_2(t)]}{dt} = \frac{d[f(x_1) - f(x_2)]}{dt} = 0.$$

Then, to test the proportionality assumption, we assume the same GP prior for $f_1(t)$ and $f_2(t)$ and compute the posteriors given the associated counts, respectively, $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$, observed for each time bin $[t_i, t_i + \Delta t]$, $i = 1, \ldots, n$. As the likelihood of a count data y_i is not Gaussian but Poisson with parameter $\exp(f(t_i))$, conjugacy is lost and the posterior distribution has to be obtained either by approximation (Laplace method or Expectation propagation) or

MCMC methods [Rasmussen and Williams, 2006]. We focus on the Laplace method that uses a Gaussian approximation of the posterior $p(\mathbf{f}|\mathbf{x},\mathbf{y})$ around its maximum $\hat{\mathbf{f}}^*$, thus recovering conjugacy as an approximation. The posterior mean and covariance matrix obtained using the Laplace method are

$$\hat{\boldsymbol{\mu}}_{\boldsymbol{\theta}}(\mathbf{x}^*|\mathbf{x},\mathbf{y}) = K_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x})K_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x})^{-1}\hat{\mathbf{f}}^*, \quad (18)$$

$$\hat{K}_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x}^*|\mathbf{x}) = K_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x}^*) - K_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x})(K_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}) + W^{-1})^{-1}K_{\boldsymbol{\theta}}(\mathbf{x},\mathbf{x}^*),$$

where $W = -\nabla\nabla \log p(\mathbf{y}|\mathbf{f})$. As shown in Section 4, the posterior distribution of the derivative of $f_1(t)$ and $f_2(t)$ is obtained by using $K^d_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x}^*)$ and $K^d_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x})$ instead of $K_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x}^*)$ and $K_{\boldsymbol{\theta}}(\mathbf{x}^*,\mathbf{x})$ in (18). Finally, the equality test for the derivatives is performed as described in Section 3. The procedure can be extended to deal with continuous covariates, based on the fact that the proportionality assumption implied that $\frac{\partial^2[\log \Lambda(t,x)]}{\partial t \partial x} = 0$, as follows from (17).

7 Experiments

7.1 Equality test

In this section we study the behaviour of the GP-based equality test by means of a simulation study taken from Neumeyer et al. [2003]. Here $n_1 = n_2 = 50$ data are sampled from the models

$$\mathbf{y}^{(1)} = f_1(\mathbf{x}^{(1)}) + \boldsymbol{v}_1; \quad \mathbf{y}^{(2)} = f_2(\mathbf{x}^{(2)}) + \boldsymbol{v}_2,$$

where $\mathbf{x}^{(1)}$ and $\mathbf{x}^{(2)}$ are uniformly distributed in $[0,1]$, $\boldsymbol{v}_1, \boldsymbol{v}_2$ are Gaussian noises with variances $\sigma_1^2 = 0.25$ and $\sigma_2^2 = 0.5$, respectively, and for f_1 and f_2 nine benchmark cases are considered:

i	$f_1(x) = f_2(x) = 1$,	
ii	$f_1(x) = f_2(x) = e^x$,	
iii	$f_1(x) = f_2(x) = \sin(2\pi x)$,	
iv	$f_1(x) = 1$,	$f_2(x) = 1 + x$,
v	$f_1(x) = e^x$,	$f_2(x) = e^x + x$,
vi	$f_1(x) = \sin(2\pi x)$,	$f_2(x) = \sin(2\pi x) + x$,
vii	$f_1(x) = 1$,	$f_2(x) = 1 + \sin(2\pi x)$,
viii	$f_1(x) = e^x$,	$f_2(x) = e^x + \sin(2\pi x)$,
ix	$f_1(x) = \sin(2\pi x)$,	$f_2(x) = 2\sin(2\pi x)$.

To limit the number of hyper-parameters to be estimated in the GP regression, as we are not interested in accurate estimates of the functions f_1 and f_2, we have used the square exponential kernel alone. The hyper-parameters have been estimated using the MAP approach with the weak prior $G(\sigma_s^2; 2, 1/2)G(\ell_s; 2, 1/2)$, where $G(x; \alpha, \beta)$ is the pdf of a Gamma distribution with mean α/β and variance α/β^2. The simulations results for the GP test are shown in Table 1 (averaged over 1000 MC runs) and compared against those of three frequentist tests: the test $K_N^{(2)}$ in Neumeyer et al. [2003], and the tests T^1_{CM} and T^2_{CM} in Pardo-Fernández et al. [2007]. We have directly implemented the T^1_{CM} and T^2_{CM} methods, and used, instead, the results in Neumeyer et al. [2003] for the $K_N^{(2)}$ method. Results show that the GP method is calibrated under the null hypothesis (cases i,ii,iii) and, on average, is the most accurate. The GP test always outperforms the T^2_{CM} method. In cases iv, v and vi it has power less than or similar to the $K_N^{(2)}$ and T^1_{CM} methods; however these two methods, perform rather poorly in situations vii, viii and ix, where they are largely outperformed by the GP test.

	$K_N^{(2)}$	T^1_{CM}	T^2_{CM}	**GP**
i	94.8	94.8	95.6	**98.8**
ii	94.9	96.0	96.9	**97.0**
iii	95.1	96.5	95.8	**98.1**
iv	95.0	96.3	85.8	**95.4**
v	94.8	95.4	85.2	**95.0**
vi	93.2	93.3	72.6	**89.0**
vii	57.4	11.7	81.4	**95.2**
viii	61.9	14.1	84.1	**97.5**
ix	51.3	7.0	51.0	**98.6**
av	81.04	67.23	83.15	**96.1**

Table 1: Percentage of MC runs in which the functions were correctly classified by the $K_N^{(2)}$, T^1_{CM}, T^2_{CM} and GP equality tests with $\alpha = 0.05$. Last row reports the average correct classification rate across all 9 cases.

To provide comparison with a Bayesian approach, we consider the simulation study carried out in Behseta and Kass [2005] using a test based on Bayesian regression adaptive splines (BARS) and compare the results with those obtained by the GP test in the same situations. The data $\mathbf{y}^{(1)}$ and $\mathbf{y}^{(2)}$ are in the form of counts on 10ms time bins of events generated by a Poisson processes with intensity functions

x	$\lambda_1(t) = rN(t; 47, 7^2)$;	$\lambda_2(t) = rN(t; 47, 7^2)$,
xi	$\lambda_1(t) = rN(t; 47, 7^2)$;	$\lambda_2(t) = rN(t; 57, 7^2)$,
xii	$\lambda_1(t) = rp_{\chi^2}(t; 40)$;	$\lambda_2(t) = rN(t; 57, 7^2)$.

where r is a positive constant, $p_{\chi^2}(t;\kappa)$ is the pdf of a Chi-squared distribution with κ degrees of freedom and where the intensities, means and variances are given, respectively, in events/s, ms, and ms^2. We assume a GP prior with square exponential kernel and a weak prior $G(\sigma_s^2; 2, 1/2)G(\ell_s; 2, 1/2)$ on the hyper-parameters, and use the Laplace approximation to obtain the posterior distribution of the functions $f_1(t) = \log \lambda_1(t)$ and $f_2(t) = \log \lambda_2(t)$. Table 2 compares the rejection probabilities evaluated over 1000 MC runs for BARS and GP test in situations x, xi and xii for two values of r. We have not directly implemented the BARS test, but reported the results in Behseta and Kass [2005]. The results for situation x are not presented in the original paper, but as the test is adjusted to have size $\alpha = 0.05$, we expect the rejec-

tion probabilities in this case to be very close to the nominal value. Results show that the GP test is very conservative under the null hypothesis (case x), as its type I error is much smaller than α, and that under the alternative hypothesis (cases xi and xii) it has equal or better power than the BARS test.

	x	xi	xii
BARS	-	82|91	98|99
GP	**100**|**100**	**82**|**100**	**100**|**100**

Table 2: Percentage of MC runs in which the functions were correctly classified BARS and GP equality tests with $\alpha = 0.05$ for $r = 30|50$.

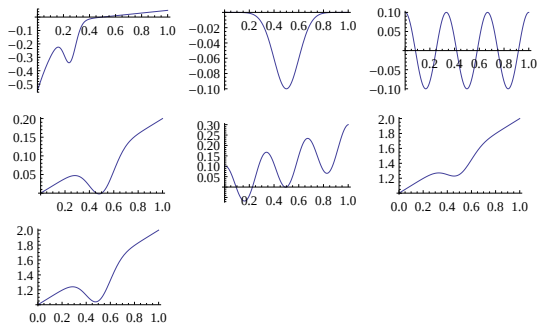

Figure 6: From left to right $g_1, g_3, g_4, g_5, g_6, g_7, g_{11}$.

7.2 Proportional intensity test

In this section we evaluate the behavior of the proportional intensity test when data are generated based on the models x, xi and xii in Section 7.1. In Table 3 the results of the GP test are compared with those obtained with the Schoenfeld residuals test [Grambsch and Therneau, 1994] implemented by the `cox.zph` function in the R package `survival` (hereafter denoted as ZPH). Like in the equality test with count data, the GP test is very conservative under the null hypothesis (case x). On the other side, the GP test outperforms the ZPH test when the alternative hypothesis is true (cases xi and xii).

	x	xi	xii
ZPH	94|96	17|36	44|73
GP	**100**|**100**	**73**|**99**	**100**|**100**

Table 3: Percentage of MC runs in which the functions were correctly classified by ZPH and GP test with $\alpha = 0.05$ for $r = 30|50$.

7.3 Monotonicity test

This section reports the results of a simulation study on eleven test functions that were used in previous works as benchmarks:

$$
\begin{aligned}
&g_1 = 4(x-\tfrac{1}{2})^3 I_{\{x\leq 0.5\}} + 0.1(x-\tfrac{1}{2}) - \tfrac{1}{4}\exp(-250(x-\tfrac{1}{4})^2),\\
&g_2 = \tfrac{x}{10}, && g_3 = -\tfrac{1}{10}\exp(-50(x-\tfrac{1}{2})^2),\\
&g_4 = \tfrac{1}{10}\cos(6\pi x), && g_5 = \tfrac{x}{5} + g_3(x),\\
&g_6 = \tfrac{x}{5} + g_4(x), && g_7 = x + 1 - \tfrac{1}{4}\exp(-50(x-\tfrac{1}{2})^2),\\
&g_8 = \tfrac{x^2}{2}, && g_9 = 0,\\
&g_{10} = x + 1, && g_{11} = x + 1 - \tfrac{9}{20}\exp(-50(x-\tfrac{1}{2})^2).
\end{aligned}
$$

Functions g_2, g_8, g_{10} are monotone increasing, while all the other functions are non-monotone, see Fig. 6. Following [Scott et al., 2013, Akakpo et al., 2014, Salomond, 2014], the observations have been generated according to the model $y_i = g(x_i) + v_i$, with $n = 100$ equally spaced x values on $(0, 1]$ and the v_i are i.i.d. Normal variables with zero-mean and standard deviation 0.1. We have generated 100 datasets for each of the eleven test functions and compared our GP monotonicity test (with $\alpha = 0.05$) against the results of seven alternative methods that previously appeared in literature. These methods are four Bayesian algorithms denoted with S_1: the method from Salomond [2014], S_2: smoothing spline test [Scott et al., 2013], G: Gaussian regression spline-based test [Scott et al., 2013], M: regression-spline test with method-of-moments priors [Scott et al., 2013] and three frequentist methods, U: U-test [Zheng, 1996], B: the test from Baraud et al. [2005], A: the test from Akakpo et al. [2014] We have not directly implemented these methods but compared the performance of our test with the results reported in [Scott et al., 2013, Tab.1] for the same simulation setting. For the frequentist tests, Scott et al. [2013] calculated a p-value under the null hypothesis of monotonicity, and rejected the null whenever $p \leq p^o$. For the Bayesian tests, they rejected the null hypothesis of monotonicity whenever the Bayes factor in favor of a non-monotone function exceeded a critical value b^o. The thresholds p^o, b^o were selected so that the frequentist and Bayesian tests are calibrated when $g = g_9$ (the zero function). Our GP method has been implemented using the quadratic and square exponential kernels, i.e., $f = f_1 + f_2$ with $f_1 \sim GP(0, k_{QD})$ and $f_2 \sim GP(0, k_{SE})$. The hyperparameters have been estimated using the MAP approach with the following weak prior $\prod_i TN(s_i; 1, 5^2)TN(\sigma_s^2; 7, 5^2)TN(\ell_s; 7, 5^2)$, where $TN(x; \mu, \sigma^2)$ is the pdf of a truncated Gaussian distribution on $\mathbf{R}^+$. This prior penalizes the complexity of the model as it assumes a length-scale for the SE kernel much larger than the range of $x \in (0, 1]$. Thus, a-priori the contribution of the SE kernel component is reduced to an approximately constant term. The simulation results are shown in Table 4. The results

of our method are in the last column (GP). Looking at the results for g_9, it is evident that our GP test with a weak prior and performed with $\alpha = 0.05$ is automatically calibrated for the zero function. Our GP test performs much better than the other tests on the difficult function g_3 (whose single oscillation is masked by the noise). Conversely, it is not very powerful on g_8, as it cannot efficiently detect the monotonicity of g_8 close to zero at this level of noise. However, overall, our GP based test obtains the same average accuracy of the best method (the Gaussian regression spline-based test).

	S_1	S_2	G	M	U	B	A	**GP**
g_1	19	99	100	99	59	6	9	**100**
g_2	83	72	74	63	59	64	33	**73**
g_3	51	34	35	49	59	53	43	**94**
g_4	73	80	91	98	0	92	92	**99**
g_5	56	95	85	90	99	24	25	**79**
g_6	86	96	99	100	34	77	75	**92**
g_7	13	92	91	47	16	1	4	**85**
g_8	98	80	93	93	41	100	100	**41**
g_9	96	98	95	95	99	97	94	**96**
g_{10}	99	99	97	99	28	100	99	**99**
g_{11}	100	100	99	99	100	71	8	**100**
Av	70	86	87	85	54	62	60	**87**

Table 4: Percentage of MC runs in which the function was correctly classified by each monotonicity test. Last row reports the average correct classification rate across all 11 test functions.

7.3.1 Seasonality

The function $g_6 = x/5 + (1/10)\cos(6\pi x)$ is clearly non-monotone, being the superposition of a linear and a periodic function. To test for the monotonicity of its non-periodic component we add a periodic Kernel to the quadratic one used in the previous section, i.e., $f = f_1 + f_2$ with $f_2 \sim GP(0, k_{PE})$; we assume the prior $TN(\sigma_p^2; 7, 5^2)TN(\ell_p; 7, 5^2)Unif(p_e; 0, 1)$ for the hyperparameters of the periodic kernel. Then, we determine the posterior distribution of the function f and retain non periodic components only, i.e., f_1, the periodic component f_2 being a disturbance. Finally, we employ this posterior to perform our monotonicity test. Table 5 reports the accuracy of the monotonic test for the functions g_6 (monotonic without the seasonal component) and $g_9 + g_4$ (non monotonic without the seasonal component) and compares it with that of the Mann-Kendall trend test (KS) with seasonality adjustment [Mann, 1945]. For fair comparison with KS, which is a trend test, we have included only the quadratic component in the GP kernel and not the SE component which models local oscillations (such as those of g_1 or g_5) which contradict monotonicity but not the assumption that the function has a trend. It can be verified that GP has similar performance (sometimes better) as KS (which assumes that the period is known).

	σ	KS	**GP**
g_6	0.1|0.2	100|70	**100|81**
g_4	0.1	98	**96**

Table 5: Percentage of MC runs in which the function was correctly classified by each trend test.

7.4 Periodicity test

In this case the goal is to test whether the function of interest is periodic. We compare our GP method described in Section 5 with the well known Fisher's significance test for periodic components [Fisher, 1929, Percival, 1993, Wichert et al., 2004]. The simulations results are shown in Table 6 for a periodic and non periodic function under different levels of noise. Also in this case, the performance of GP is high.

	σ	Fisher	**GP**
g_9	0.1	96	**100**
g_4	0.1|0.2	100|39	**96|35**

Table 6: Percentage of MC runs in which the function was correctly classified by each periodicity test.

8 Conclusions

We have proposed a novel Bayesian method based on Gaussian Processes (GP) for performing hypothesis tests on regression functions. The advantage of the Bayesian approach is that, once we have obtained the posterior distribution of the regression function f, we can perform different hypothesis tests about f by simply asking different questions to the posterior. This has allowed us to develop tests for equality, monotonicity (which can also take into account seasonality), periodicity and proportionality of regression functions. We have evaluated the performance of our GP method against state-of-art algorithms and shown that it is very competitive. We plan to use this approach to implement other tests for regression functions, time series and spatial statistics.

Acknowledgments

This work was partly supported by the Swiss NSF grants nos. 200021_146606 / 1 and 200020_137680 / 1.

References

Nathalie Akakpo, Fadoua Balabdaoui, Cécile Durot, et al. Testing monotonicity via local least concave majorants. *Bernoulli*, 20(2):514–544, 2014.

Yannick Baraud, Sylvie Huet, and Béatrice Laurent. Testing convex hypotheses on the mean of a Gaussian vector. Application to testing qualitative hypotheses on a regression function. *Annals of statistics*, pages 214–257, 2005.

Sam Behseta and Robert E Kass. Testing equality of two functions using bars. *Statistics in medicine*, 24 (22):3523–3534, 2005.

AW Bowman, MC Jones, and Irène Gijbels. Testing monotonicity of regression. *Journal of computational and Graphical Statistics*, 7(4):489–500, 1998.

David B Dunson. Bayesian semiparametric isotonic regression for count data. *Journal of the American Statistical Association*, 100(470):618–627, 2005.

Ronald Aylmer Fisher. Tests of significance in harmonic analysis. *Proceedings of the Royal Society of London. Series A*, 125(796):54–59, 1929.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press, 2013.

Subhashis Ghosal, Arusharka Sen, and Aad W. van der Vaart. Testing monotonicity of regression. *The Annals of Statistics*, 28(4):1054–1082, 08 2000.

Patricia M Grambsch and Terry M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

Peter Hall and Nancy E. Heckman. Testing for monotonicity of a regression mean by calibrating for linear functions. *The Annals of Statistics*, 28(1):pp. 20–39, 2000.

David JC MacKay. Introduction to Gaussian processes. *In Bishop, C. M., editor, Neural Networks and Machine Learning*, pages 133–166, 1998.

Henry B Mann. Nonparametric tests against trend. *Econometrica: Journal of the Econometric Society*, pages 245–259, 1945.

MATLAB. *version 8.1.0 (R2013a)*. The MathWorks Inc., Natick, Massachusetts, 2013.

R. M. Neal. Regression and classification using gaussian process priors. In *in Bernardo, JM and Berger, JO and Dawid, AP and Smith, AFM editors, Bayesian Statistics 6: Proceedings of the sixth Valencia international meeting*, volume 6, page 475, 1998.

Natalie Neumeyer, Holger Dette, et al. Nonparametric comparison of regression curves: an empirical process approach. *The Annals of Statistics*, 31(3): 880–920, 2003.

A. O'Hagan and J. F. C. Kingman. Curve fitting and optimal design for prediction. *Journal of the Royal Statistical Society. Series B (Methodological)*, 40(1): pp. 1–42, 1978.

Juan Carlos Pardo-Fernández, Ingrid Van Keilegom, Wenceslao González-Manteiga, et al. Testing for the equality of k regression curves. *Statistica Sinica*, 17 (3):1115, 2007.

Donald B Percival. *Spectral analysis for physical applications*. Cambridge University Press, 1993.

Carl Edward Rasmussen. Gaussian processes to speed up hybrid monte carlo for expensive bayesian integrals. In *Bayesian Statistics 7: Proceedings of the 7th Valencia International Meeting*, pages 651–659. Oxford University Press, 2003.

Carl Edward Rasmussen. The gaussian processes web site, February 2011. URL `http://www.gaussianprocess.org/`.

Carl Edward Rasmussen and CKI Williams. Gaussian processes for machine learning. 2006. *The MIT Press, Cambridge, MA, USA*, 38:715–719, 2006.

Jaakko Riihimäki and Aki Vehtari. Gaussian processes with monotonicity information. In *International Conference on Artificial Intelligence and Statistics*, pages 645–652, 2010.

Jean-Bernard Salomond. Adaptive Bayes test for monotonicity. In *The Contribution of Young Researchers to Bayesian Statistics*, pages 29–33. Springer, 2014.

James G Scott, Thomas S Shively, and Stephen G Walker. Nonparametric bayesian testing for monotonicity. *arXiv preprint arXiv:1304.3378*, 2013.

Ercan Solak, Roderick Murray-Smith, William E Leithead, Douglas J Leith, and Carl Edward Rasmussen. Derivative observations in Gaussian process models of dynamic systems. 2003.

Zsolt Ugray, Leon Lasdon, John Plummer, Fred Glover, James Kelly, and Rafael Martí. Scatter search and local nlp solvers: A multistart framework for global optimization. *INFORMS Journal on Computing*, 19(3):328–340, 2007.

Sofia Wichert, Konstantinos Fokianos, and Korbinian Strimmer. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics*, 20(1):5–20, 2004.

John Xu Zheng. A consistent test of functional form via nonparametric estimation techniques. *Journal of Econometrics*, 75(2):263–289, 1996.