A Time Series Decomposition Algorithm Based on Gaussian Processes

Massimo Bilancia[1(✉)], Fabio Manca[2], and Giovanni Sansaro[3]

[1] Ionian Department in Legal and Economic System of Mediterranean: Society, Environment, Culture (DSJGEM), Via Duomo 259, 74123 Taranto, Italy
massimo.bilancia@uniba.it
[2] Department of Education, Psychology, Communication (FORPSICOM), University of Bari Aldo Moro, Bari, Italy
[3] Graduate Program in Data Science, Department of Informatics (DIB), University of Bari Aldo Moro, Bari, Italy

Abstract. In this paper, we present an algorithm for decomposing time series based on Gaussian processes. Gaussian processes can be viewed as infinite-dimensional probability distributions over smooth functions and also provide a natural basis for additive decomposition of time series, since we can sum mutually independent Gaussian processes with a simple and elegant algebra involving covariance kernels. The component estimation algorithm we propose in this paper is general and does not depend on the number of components, nor on the correlation structure and interpretation of each component. Specifically, the proposed algorithm is based on nonparametric Bayesian Gaussian process regression, where the log-likelihood covariance is suitably structured to account for the presence of additive subcomponents. The numerical parameter estimation procedure finds MAP estimates by maximizing the unnormalized log-posterior density, with a great advantage in terms of computational cost and efficiency.

We apply our proposal to real data based on a time series of daily COVID-19 confirmed cases and daily swabs administered in Italy. The curve of daily new cases follows an oscillatory pattern, as the ratio between the number of new cases and the number of tested individuals is locally constant, while the number of daily new cases is systematically lower on some days of the week, as fewer swabs are performed. This wavelike pattern must be considered a nuisance that does not reflect the true dynamics of contagion, as it is the result of variability in human activities. We show how such a cyclic component can be successfully filtered out using the proposed algorithm.

Keywords: Time series decomposition · Gaussian processes · Bayesian hierarchical modelling · Computational methods for Bayesian statistics

1 Introduction

Real-world time series data exhibit complex patterns involving trend, oscillatory waveforms, and noise [1]. A robust decomposition of these components greatly

© Springer Nature Switzerland AG 2021
O. Gervasi et al. (Eds.): ICCSA 2021, LNCS 12952, pp. 577–592, 2021.
https://doi.org/10.1007/978-3-030-86973-1_41

facilitates typical time series tasks such as prediction and classification. Typical forms of fluctuations are: a) A smooth long-term tendency, the trend. b) A cyclical movement superimposed on the smooth trend. c) A less or more regular cyclical movement within a year, the seasonality. d) Residual or unexplained variation (noise). Whether or not a component is present in a particular time series depends on the human activities underlying the phenomenon and the frequency of measurement. In any case, it has often been assumed that the four components vary independently and interact on a linear additive scale. Additive time series decomposition models are an important tool in a data scientist's arsenal, and there are many ways to extract the trend and other components of a time series.

One of the first structured model-based approaches developed since the late 1970s was the LOESS (LOcal regrESSion) method for modeling trend and seasonal components using polynomial regression, which has several advantages over early decomposition methods. The STL algorithm based on LOESS is very versatile and robust, since the seasonal component is allowed to change over time and the smoothing of the trend cycle can also be controlled by the user [2]. Apart from some minor technical difficulties, the STL method is very easy to use and available on most data analysis-oriented programming platforms, such as R or Python.

However, a major drawback of classical methods based on additive decompositions is that they are rigidly bound to the subdivision into the three components trend/cycle/seasonality. Greater flexibility is only possible by introducing a new ad hoc decomposition model specific to the problem at hand. Another drawback common to all these approaches is that they are not based on stochastic processes and it is therefore impossible to determine the uncertainty associated with the subcomponent estimates. With this in mind, we have developed a decomposition algorithm based on Gaussian processes, a class of stochastic processes endowed with great modeling flexibility [3,4].

A Gaussian process can be viewed as an infinite-dimensional probability distribution over smooth functions [5]. In practice, we do not need to work with such infinite-dimensional function spaces, since it is sufficient to evaluate the probability distribution over the function at any finite collection of points. Gaussian processes also provide a natural basis for additive decomposition of time series, since we can sum mutually independent Gaussian processes with a simple and elegant algebra. The components of an additive Gaussian model can be structured very similarly to those of classical decompositions, with no restriction on their number and internal structure. Moreover, the component estimation algorithm we describe in this paper is general and does not depend on the number of components or on the correlation structure and interpretation of each component. Specifically, the proposed algorithm is based on the nonparametric Bayesian Gaussian process regression, where the log-likelihood covariance is suitably structured to account for the presence of additive subcomponents. The numerical parameter estimation procedure finds MAP estimates by maximiz-

ing the unnormalized log-posterior density, with a great advantage in terms of computational cost and efficiency.

We present a real-world experiment based on the time series of the daily total number of COVID-19 confirmed new cases and administered molecular swabs in Italy from February 21, 2020 to January 14, 2021. The curve of daily new cases follows an oscillating pattern, as the ratio of the number of new cases to that of tested subjects is locally constant, while the number of daily new cases is systematically lower on some days of the week, as fewer swabs are performed (a typical example of this pattern occurs on Monday, when most of the reported cases were tested on the previous Saturday and Sunday). This wave pattern must be considered a nuisance that does not reflect the true dynamics of contagion, as it is the result of variability in human activity, and its impact on subsequent analyses cannot be quantified. We show how the cyclic component can be successfully filtered out using the proposed algorithm.

The paper is organized as follows. Section 2 briefly recalls some alternative approaches to time series decomposition based on principal component analysis (PCA) and presents their advantages and disadvantages. Section 3 introduces the main definitions. Section 4 with the algebra of Gaussian processes and their covariance function and the way additive models can be structured. Section 5 introduces Bayesian Gaussian process regression and its approximate Bayesian inference, and provides the decomposition formula for computing the components of the additive decomposition. Section 6 presents the case study mentioned above. Section 7 draws final conclusions and suggests future research questions.

2 Related Work

Singular spectrum analysis (SSA) is a distinct technique related to additive time series decomposition, which is essentially PCA applied to lagged copies of the original data [6,7]. In the first stage, the time series is embedded in a lower-dimensional vector space of dimension M by considering M lagged copies of the data. In the second stage, the $M \times M$ lag covariance matrix estimator of the lag matrix is computed. Then, the lag-covariance matrix is diagonalized to obtain spectral information about the time series. Finally, temporal principal components are computed, and such components can be grouped to reconstruct a decomposition into a fixed number of subcomponent series. Extensions to multivariate series or 2D data (images or spatial data) are also available [8].

While SSA has significant advantages from a computational point of view, with many fast and stable programs available for extracting PCs, there are some drawbacks and differences with our approach that are worth some additional comments. From a purely algorithmic point of view, SSA is automatic and easy to use. However, the lag parameter M is a free and its choice is not obvious, and the obtained decomposition is strongly dependent on the choice of M. Finally, grouping the temporal principal components into a small number of subcomponents is not a simple task. The decomposition can be considered reasonable if each reconstructed subseries can be classified as part of either a trend or a periodic component or noise. To this end, a complete theory of SSA separability has

been introduced that attempts to answer questions such as which components can be separated by SSA and how to choose M and perform proper grouping for the extraction of interpretable components [9].

Although this theory is highly developed, it is often not easy to apply in practice. The main difference with our approach is that the interpretation of the components obtained by SSA is done after the decomposition of the original time series. In our case, on the other hand, the correlation structure and the interpretation of the additive components are determined before the execution of the algorithm. In some applications this fact can be a drawback, while in others it can be useful to accurately calibrate the decomposition model to the problem at hand.

3 Gaussian Processes

Denote f a real-valued continuous-time stochastic process and $\mathcal{X} \subseteq \mathbb{R}^D$ the input space. Sequences of input data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$ are embedded in a q-dimensional space, where they may be linearly ordered since the index reflects the temporal ordering of the trajectories. A stochastic process f is said to be Gaussian if $(f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))^\top$ is an n-dimensional Gaussian vector for all $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathcal{X}$ and all $n \geq 1$. In other words, a Gaussian stochastic process is an infinite collection of random variables such that any finite-dimensional distribution is multivariate Gaussian [10]. Each Gaussian process can be uniquely assigned by its mean function and its covariance kernel:

$$\mu(\mathbf{x}) = \mathsf{E}\left[f(\mathbf{x})\right], \quad \forall\, \mathbf{x} \in \mathcal{X}, \tag{1}$$

$$\kappa(\mathbf{x}, \mathbf{x}') = \mathsf{Cov}\left[f(\mathbf{x}), f(\mathbf{x}')\right], \quad \forall\, \mathbf{x}, \mathbf{x}' \in \mathcal{X}. \tag{2}$$

For a given mean function $\mu(\mathbf{x})$ and kernel function $\kappa(\mathbf{x}, \mathbf{x}')$, there exists a Gaussian process f whose finite-dimensional distributions are compatible with the two functions [11], which is given by:

$$f(\mathbf{x}) \sim \mathcal{GP}\left(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}')\right). \tag{3}$$

Some examples of trajectories simulated from a Gaussian process are shown in Fig. 1. On the left side we have a set of realizations from a Gaussian process with square-exponential covariance function, while on the right side the trajectories were simulated from a Gaussian process with a periodic covariance function of period $p = 7$ (see Sect. 4 for the relevant definitions).

4 The Algebra of Covariance Functions

A time series decomposition model is based on an appropriate decomposition of the underlying data generating process into additive subcomponents. To this

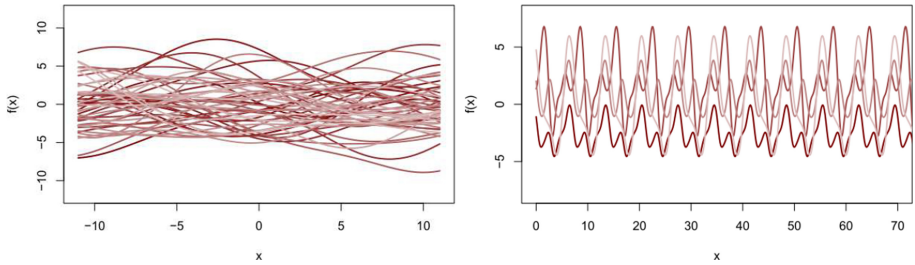


Fig. 1. Simulated trajectories from a Gaussian process.

end, let $\kappa_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)})$ and $\kappa_2(\mathbf{x}_{(2)}, \mathbf{x}'_{(2)})$ be valid covariance functions defined on the input spaces $\mathcal{X}_1$ and $\mathcal{X}_2$, respectively. Then:

$$\kappa(\mathbf{x}, \mathbf{x}') = \kappa_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)}) + \kappa_2(\mathbf{x}_{(2)}, \mathbf{x}'_{(2)}) \qquad \text{(direct sum)}, \tag{4}$$

$$\kappa(\mathbf{x}, \mathbf{x}') = \kappa_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)})\kappa_2(\mathbf{x}_{(2)}, \mathbf{x}'_{(2)}) \qquad \text{(tensor product)}, \tag{5}$$

are valid covariance functions defined on $\mathcal{X}_1 \times \mathcal{X}_2$.

Given two independent Gaussian processes f_1 and f_2 whose covariance function is given respectively by $\kappa_1(\mathbf{x}_{(1)}, \mathbf{x}'_{(1)})$ and $\kappa_2(\mathbf{x}_{(2)}, \mathbf{x}'_{(2)})$, the additive sum:

$$f(\mathbf{x}) = f_1(\mathbf{x}_{(1)}) + f_2(\mathbf{x}_{(2)}) \tag{6}$$

is itself a Gaussian process whose covariance function is given by the direct sum (4).

We can extend this approach to define an additive model using D mutually independent Gaussian processes $f_1, \ldots, f_D$:

$$f(\mathbf{x}) = f_1(x_{(1)}) + f_2(x_{(2)}) + \cdots + f_D(x_{(D)}), \tag{7}$$

where D is the size of the input space. Each Gaussian process f_i can be defined on its own one-dimensional input space. The covariance function of the additive model is the direct sum of D covariance functions [12].

A covariance function can be further classified into the following categories:

- stationary: $\kappa(\mathbf{x}, \mathbf{x}')$ depends on the input variables only by the step size $\mathbf{x} - \mathbf{x}'$
- isotropic: $\kappa(\mathbf{x}, \mathbf{x}')$ depends on the input variables only by the Euclidean distance $r = \|\mathbf{x} - \mathbf{x}'\|$

If a covariance function does not satisfy at least the stationary condition, it is said to be nonstationary. The class of these covariance functions is invariant to translations of the coordinate system. Isotropic covariance functions are invariant to translations, rotations, and reflections (rigid motions) of the coordinate system.

4.1 Kernel Matrix

Consider T sample observations $y_i = f(\mathbf{x}_i)$, $i = 1, 2, ..., n$, from a Gaussian process $f(\mathbf{x}) \sim \mathcal{GP}(m(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$. The kernel matrix $\mathbf{K}$ is the Gram matrix containing the covariances between all possible pairs of observations:

$$\mathbf{K} = \begin{pmatrix} \kappa(\mathbf{x}_1, \mathbf{x}_1) & \kappa(\mathbf{x}_1, \mathbf{x}_2) & \dots & \kappa(\mathbf{x}_1, \mathbf{x}_n) \\ \kappa(\mathbf{x}_2, \mathbf{x}_1) & \kappa(\mathbf{x}_2, \mathbf{x}_2) & \dots & \kappa(\mathbf{x}_2, \mathbf{x}_n) \\ \vdots & \vdots & \ddots & \vdots \\ \kappa(\mathbf{x}_n, \mathbf{x}_1) & \kappa(\mathbf{x}_n, \mathbf{x}_2) & \dots & \kappa(\mathbf{x}_n, \mathbf{x}_n) \end{pmatrix}. \tag{8}$$

To be a valid covariance function, the resulting kernel matrix must be symmetric definite positive. Conversely, any symmetric definite positive matrix is the kernel matrix of a Gaussian process. Classes of valid covariance functions used to model the correlation structure of additive subcomponents are:

Squared Exponential Kernel. Squared exponential kernel (or Radial Basis Function) is an isotropic kernel having the following expression:

$$\kappa_{\mathsf{SE}}(x, x') = \sigma^2 \exp\left(-\frac{(x - x')^2}{2\ell^2}\right). \tag{9}$$

The characteristic length scale ℓ controls the smoothing of trajectories, since it can be shown that the expected value of the number of times a trajectory crosses the x-axis in the unit interval is equal to $(2\pi\ell)^{-1}$. This means that as ℓ increases, the function becomes less irregular and smoother as it wobbles less quickly above and below the x-axis. The amplitude σ determines the maximum excursion of the trajectories.

Periodic Kernel. Mapping a one-dimensional input variable $x \in \mathcal{X}$ onto the two-dimensional u-space $u(x) = (\cos(x), \sin(x))$, allows us to define a periodic stationary kernel [13,14]. Using the squared exponential kernel in u-space, we can define a periodic stationary covariance function of period 2π:

$$\kappa_{\mathsf{P}}(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2(\frac{x - x'}{2})}{\ell^2}\right), \tag{10}$$

as:

$$(\cos(x_i) - \cos(x_j))^2 + (\sin(x_i) - \sin(x_j))^2 = 4\sin^2\left(\frac{x_i - x_j}{2}\right).$$

We can easily rescale the covariance function defined above to obtain the following covariance function of period p:

$$\kappa_{\mathsf{P}}(x, x') = \sigma^2 \exp\left(-\frac{2\sin^2(\frac{\pi(x - x')}{p})}{\ell^2}\right). \tag{11}$$

The parameter σ controls the amplitude of the trajectories, while ℓ controls the regularity of the waveform.

Locally Periodic Kernel. Multiplying a quadratic exponential kernel by a periodic kernel of period p can increase flexibility:

$$\kappa_{\mathrm{LP}}(x, x') = \sigma^2 \exp\left(-\frac{(x-x')^2}{2\ell_1^2}\right) \exp\left(-\frac{2\sin^2(\frac{\pi(x-x')}{p})}{\ell_2^2}\right). \tag{12}$$

The trajectories of the resulting Gaussian process are periodic time series of period p where the waveform is not stationary since its amplitude may change over time (see Fig. 2 where some synthetic trajectories are shown).

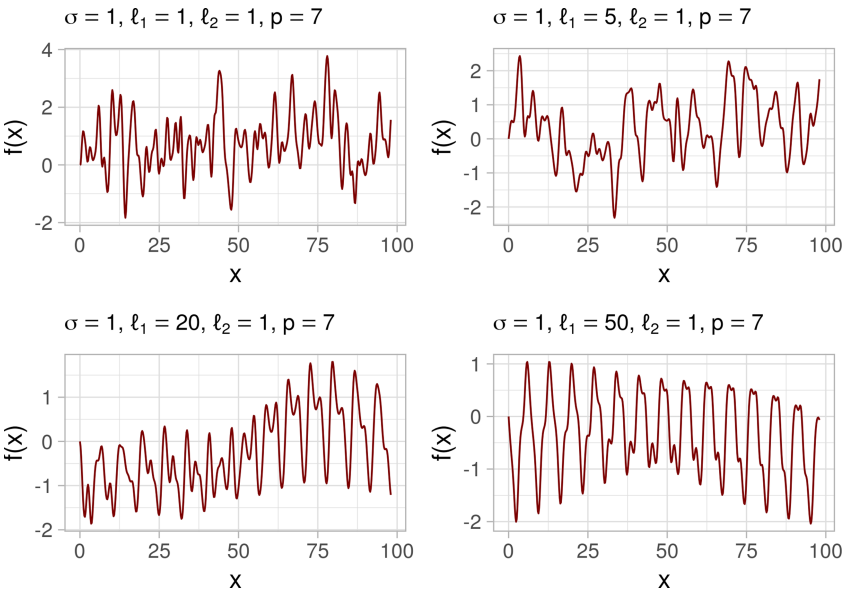


Fig. 2. Four trajectories drawn (discretized over $n = 500$ points) from a GP with locally periodic covariance function of period $p = 7$, for different values of σ, ℓ_1 and ℓ_2.

5 Gaussian Process Regression

The basic framework of our decomposition model is a noisy Gaussian process regression model, which can be expressed as $y_i = f(\mathbf{x}_i) + \epsilon_i$, $i = 1, 2, ..., n$, with $f(\mathbf{x}) \sim \mathcal{GP}(\mu(\mathbf{x}), \kappa(\mathbf{x}, \mathbf{x}'))$ and $\epsilon(\mathbf{x}) \sim \mathcal{GP}(0, \sigma_\epsilon^2 \mathbb{1}(\mathbf{x} = \mathbf{x}'))$. The Gaussian processes f and ϵ are assumed to be independent, where $\mathbb{1}(\cdot)$ denotes the indicator function. We first assume that the mean, the covariance function of f and the noise variance σ_ϵ^2 are known.

Given the design matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_n)^\top \in \mathbb{R}^{n \times D}$, we define:

$$\mathbf{y} = (y_1, y_2, ..., y_n)^\top, \tag{13}$$

$$\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), ..., f(\mathbf{x}_n))^\top \tag{14}$$

$$\boldsymbol{\mu}(\mathbf{X}) = (\mu(\mathbf{x}_1), \mu(\mathbf{x}_2), ..., \mu(\mathbf{x}_n))^\top . \tag{15}$$

A nonparametric Gaussian process regression model can be formulated as the following Bayesian hierarchical model with a Gaussian likelihood and a latent Gaussian process prior [15]:

$$\begin{aligned} \mathbf{y}|\mathbf{f} &\sim \mathcal{N}_n(\mathbf{f}, \sigma_\epsilon^2 \mathbb{I}_n) \quad \text{(likelihood)}, \\ \mathbf{f}|\mathbf{X} &\sim \mathcal{N}_n(\boldsymbol{\mu}(\mathbf{X}), \mathbf{K}) \quad \text{(prior)}, \end{aligned}$$

where $\mathbf{K}$ is the Kernel matrix of $\mathbf{f}$ and $\mathbb{I}_n$ denotes the $n \times n$ identity matrix.

The latent Gaussian process $\mathbf{f}$ can also be viewed as an infinite-dimensional prior over the space of regression functions. For a Gaussian likelihood we can simply marginalize over $\mathbf{f}$ and obtain [16]:

$$p(\mathbf{y}|\mathbf{X}) = \int p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{X}) \; d\mathbf{f} \sim \mathcal{N}_n(\boldsymbol{\mu}(\mathbf{X}), \mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_N) \tag{16}$$

The joint distribution of the observed data points and the latent Gaussian process has expression:

$$\begin{pmatrix} \mathbf{y} \\ \mathbf{f} \end{pmatrix} \sim \mathcal{N}_{2n} \left(\begin{pmatrix} \boldsymbol{\mu}(\mathbf{X}) \\ \boldsymbol{\mu}(\mathbf{X}) \end{pmatrix}, \begin{pmatrix} \mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_n & \mathbf{K} \\ \mathbf{K} & \mathbf{K} \end{pmatrix} \right). \tag{17}$$

Exploiting the standard properties of multivariate Gaussian vectors [17], we have that the posterior finite-dimensional distribution of the process at the observed data points is:

$$\mathbf{f}|\mathbf{y}, \mathbf{X} \sim \mathcal{N}_n(\boldsymbol{\mu}_{\mathsf{post}}, \mathbf{K}_{\mathsf{post}}), \tag{18}$$

with:

$$\boldsymbol{\mu}_{\mathsf{post}} = \boldsymbol{\mu}(\mathbf{X}) + \mathbf{K}(\mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_n)^{-1}(\mathbf{y} - \boldsymbol{\mu}(\mathbf{X})), \tag{19}$$

$$\mathbf{K}_{\mathsf{post}} = \mathbf{K} - \mathbf{K}(\mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_n)^{-1}\mathbf{K}. \tag{20}$$

Since a Gaussian process is flexible enough to model the regression function even in the case where the mean function is identically zero, i.e. $\boldsymbol{\mu}(\mathbf{X}) \equiv \mathbf{0}$, in this case the posterior mean obviously reduces to:

$$\boldsymbol{\mu}_{\mathsf{post}} = \mathbf{K}(\mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_n)^{-1}\mathbf{y}, \tag{21}$$

while the expression of the posterior covariance matrix is left unchanged. This marginal likelihood model is computationally more efficient since it uses a lower-dimensional parameter space. The latent Gaussian process formulation is useful when the dependent variable is non-Gaussian and is not pursued further.

5.1 Decomposition Formulas

If $f(\mathbf{x})$ is an additive a-Gaussian process, see Eq. (7), then each f_i is a Gaussian process with finite-dimensional posterior mean and posterior covariance matrix, given respectively by:

$$\mu_{\mathsf{post}}(f_i) = K_i(\mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_n)^{-1}\mathbf{y}, \tag{22}$$

$$K_{\mathsf{post}}(f_i) = K_i - K_i(\mathbf{K} + \sigma_\epsilon^2 \mathbb{I}_n)^{-1}K_i, \tag{23}$$

where, for $i = 1, 2, ..., D$:

$$f_i(x_{(i)}) \overset{\mathsf{ind.}}{\sim} \mathcal{GP}(0, \kappa_i(x_{(i)}, x'_{(i)})), \tag{24}$$

and:

$$\mathbf{K} = K_1 + K_2 + \cdots + K_D.$$

The expressions (22) and (23) fully determine a Bayesian estimate of the trajectory of the i-th component entering the additive model.

5.2 Approximate Bayesian Inference

When the covariance function depends on a vector of unknown parameters $\boldsymbol{\theta}$, we have the following extended hierarchical model:

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\theta}, \sigma_\epsilon^2 \sim \mathcal{N}_n(\mathbf{0}, \mathbf{K}(\boldsymbol{\theta}) + \sigma_\epsilon^2 \mathbb{I}_n), \tag{25}$$

$$\boldsymbol{\psi} \sim \pi(\boldsymbol{\phi}). \tag{26}$$

where $\boldsymbol{\psi} = (\boldsymbol{\theta}, \sigma_\epsilon^2)^\top$ and $\pi(\boldsymbol{\phi})$ is a joint prior distribution for both kernel parameters $\boldsymbol{\theta}$ and the noise variance σ_ϵ^2, with fixed hyperparameters $\boldsymbol{\phi}$.

Combining the marginal likelihood with the prior, we get an unnormalized posterior, and inference could be done using standard Markov Chain Monte Carlo (MCMC) methods [18,19]. In principle, we could sample from the joint posterior of the parameters and the finite-dimensional distribution of the process using the factorization:

$$p(\mathbf{f}, \boldsymbol{\psi}|\mathbf{y}, \mathbf{X}) = p(\mathbf{f}|\boldsymbol{\psi}, \mathbf{X}, \mathbf{y})p(\boldsymbol{\psi}|\mathbf{y}), \tag{27}$$

and sequentially simulate $\boldsymbol{\psi}$ from $p(\boldsymbol{\psi}|\mathbf{y})$, using MCMC, and f from $p(\mathbf{f}|\boldsymbol{\psi}, \mathbf{X}, \mathbf{y})$, which is in closed form. A major problem with this approach is that its basic complexity is $\mathcal{O}(n^3)$, due to the inversion of a $n \times n$ matrix at each MCMC step, which is prohibitive for large data sets. There are quite a number of proposals to solve this problem [20–22].

In practice, we use the posterior modes $\widehat{\boldsymbol{\psi}}_{\mathsf{MAP}}$, obtained by maximizing the unnormalized log-posterior density:

$$\widehat{\boldsymbol{\psi}}_{\mathsf{MAP}} = \arg\max_{\boldsymbol{\psi}} \log p(\mathbf{y}|\mathbf{X}, \boldsymbol{\psi}) + \log \pi(\boldsymbol{\phi}). \tag{28}$$

Having found the MAP estimate for the hyperparameters, we can reconstruct the trajectories of each process component using (22) and (23). This solution ignores the uncertainty in posterior estimates of $\boldsymbol{\psi}$ but it can be handled appropriately using feasible iterative maximization algorithms. For example, Stan (a high-level language in which the user specifies a Bayesian hierarchical model and the posterior computations are implemented automatically) is particularly efficient at computing the gradients of the marginal likelihood as well as estimating the marginal posterior modes using a second-order iterative quasi-Netwon method, known as the Broyden-Fletcher-Goldfarb-Shanno (BFGS) algorithm, which forms an approximation of the second-order partial derivatives and the Hessian matrix by using only gradient information [23].

6 Example

The data used in this experiment are the time series of the daily total number of COVID-19 new cases and molecular swabs administered in Italy from February 21, 2020, when the first cases were officially reported, to January 14, 2021. Scraping of COVID-19 national data was performed directly from Wikipedia, using the template *COVID-19 pandemic data/Italy medical cases*, which aggregates daily data from regions and autonomous provinces (PP/AA) collected by the Ministry of Health and Istituto Superiore di Sanità (ISS) [24].

The curve of daily new cases exhibits a characteristic sawtooth pattern (see Fig. 3), as the ratio of the number of new cases to the number of subjects tested is locally constant (in the sense that it is approximately constant over short time windows), while the number of new cases is systematically lower on some days of the week, as fewer swabs are performed (a typical example of this pattern occurs on Monday, when most of the reported cases were tested on the previous Saturday and Sunday).

Such a wave pattern observed in the outcome data takes the form of a weekly cyclic oscillation with time-dependent amplitudes. This oscillatory wave exhibits phase coupling with the number of daily swabs administered, in the sense that the phase difference between the series of daily new cases and that of daily swabs remains constant for obvious reasons. However, this wave pattern must be considered a nuisance that does not reflect the actual dynamics of contagion. Rather, it is the result of variability in human activity, and its impact on subsequent analyses cannot be quantified unless different scenarios are considered for data analysis. It is therefore reasonable to construct an alternative dataset in which the periodic component is filtered out by the proposed algorithm based on Gaussian processes.

6.1 Model Setting

For time series data, we have a single common input variable $\mathbf{x} \equiv t \in \mathbb{R}$, where t is the time index. We consider the following additive model:

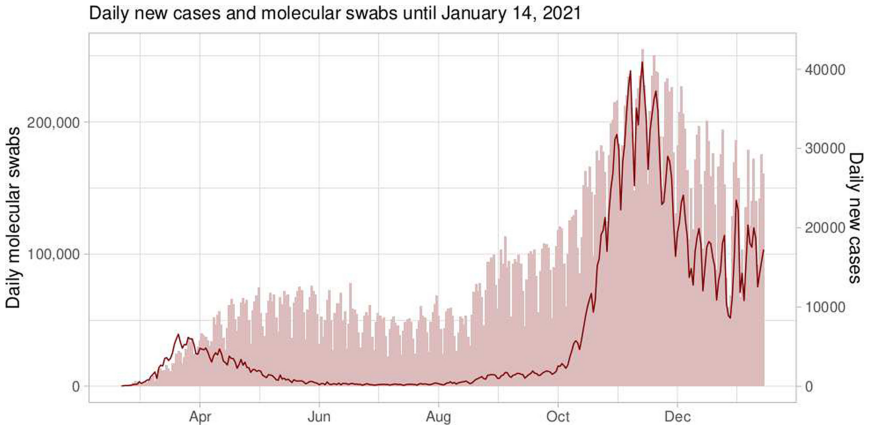


Fig. 3. Number of daily newly diagnosed cases and molecular swabs in Italy, between 21 February 2020 and 14 January 2021. Daily new cases are represented by the solid line; the bar graph shows the number of daily molecular swabs administered.

$$f(t) = f_1(t) + f_2(t) + f_3(t) + \epsilon_t, \quad t = 1, 2, \ldots, n, \tag{29}$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma_\epsilon^2)$ models the Gaussian white noise. The covariance function of this Gaussian process is:

$$\kappa(t, t') = \kappa_1(t, t') + \kappa_2(t, t') + \kappa_3(t, t') + \sigma_\epsilon^2 \mathbb{1}(t = t'), \tag{30}$$

with different terms representing different scales and periodicities. In particular:

- $f_1(t)$ models the smooth long-term variation, using a Squared Exponential kernel:

$$\kappa_1(t, t') = \sigma_1^2 \exp\left(-\frac{(t - t')^2}{2\ell_1^2}\right). \tag{31}$$

- $f_2(t)$ models short-term variation, using a Squared Exponential kernel (with different parameters)

$$\kappa_2(t, t') = \sigma_2^2 \exp\left(-\frac{(t - t')^2}{2\ell_2^2}\right). \tag{32}$$

- f_3 represents the weekly periodic pattern with time-dependent amplitudes due to daily variations in the number of tests administered. In this case, the correlation structure can be conveniently described by a locally periodic kernel with period $p = 7$:

$$\kappa_3(t, t') = \sigma_3^2 \exp\left(-\frac{(t - t')^2}{2\ell_{3.1}^2}\right) \exp\left(-\frac{2\sin^2\frac{\pi(t-t')}{7}}{\ell_{3.2}^2}\right). \tag{33}$$

Our approach is quite informal, in the sense that the proposed specification is based on the structural knowledge of the domain of study, rather than insisting on a formal Bayesian view to select a model from a set of candidate models [25,26]. The specification of prior distributions over model parameters is:

$$\sigma_1, \sigma_2, \sigma_3 \overset{\text{ind.}}{\sim} \text{Half-}\mathcal{N}(0,1), \tag{34}$$

$$\ell_1, \ell_2, \ell_{3.1}, \ell_{3.2} \overset{\text{ind.}}{\sim} \text{Inv-Gamma}(5,5), \tag{35}$$

which is a weekly informative setting that attempts to introduce weak prior assumptions about the true location of the unknown parameters [27,28] into our analysis. The Half-$\mathcal{N}(0,1)$ distribution gives the distribution of the absolute value of a $\mathcal{N}(0,1)$ random variable. The Inv-Gamma $(5,5)$ with shape parameter $\alpha > 0$ and scale parameter $\beta > 0$ is the distribution of the inverse of a Gamma$(5,5)$ random variable. With the chosen setting, the latter distribution is approximately centered around 1 (the expected value is $5/4 = 1.25$) and quite diffuse, since the range between 0.025 and 0.975 percentiles is $(0.49, 3.08)$. The optimization of the unnormalized posterior (28) was performed using R 4.0.4 + rstan 2.21.2 [29].

6.2 Identifiability Issues

It is important to note that there is a fundamental ambiguity in which parameters are associated with each Gaussian process component, since the probability of the sum is invariant to each permutation of the indices. With a weakly informative prior, the posterior distribution inherits this permutational invariance. One way to resolve this ambiguity is to restrict the exploration of posterior modes to a single ordering of parameters, a trick that can be interpreted as a method of making an exchangeable prior non-exchangeable. It can be proved that the resulting inferences are correct after imposing an artificial identifiability constraint. Moreover, this approach is often able to separate the symmetric posterior modes [30,31]. The order we used for our problem can be expressed by the following restricted prior, which naturally reflects the relative weight of the additive components of the Gaussian process:

$$\pi(\sigma_1, \sigma_2, \sigma_3)\mathbb{1}(\sigma_2 < \sigma_3 < \sigma_1). \tag{36}$$

A complete decomposition of the trajectory into its component trajectories can be estimated by computing the expected value and standard deviation of the approximate posterior distribution of $f_1(t)$, $f_2(t)$, $f_3(t)$ using the expression derived for $\boldsymbol{\mu}_{\text{post}}$ and $\mathbf{K}_{\text{post}}$.

6.3 Results

Preliminarily, the data were standardized because in Gaussian processes it is usually assumed that the expected value function is a function of time and is identically zero. This assumption reduces the size of the parameter space, but at the same time is not a real constraint, since a Gaussian process is flexible enough to model any trajectory, even if one assumes an identically-zero expected value. The estimated components of the resulting decomposition are shown in Fig. 4. Approximate posterior confidence bands, based on the diagonal elements of the estimated $\mathbf{K}_{\mathsf{post}}(f_i)$ matrices, have been superimposed.

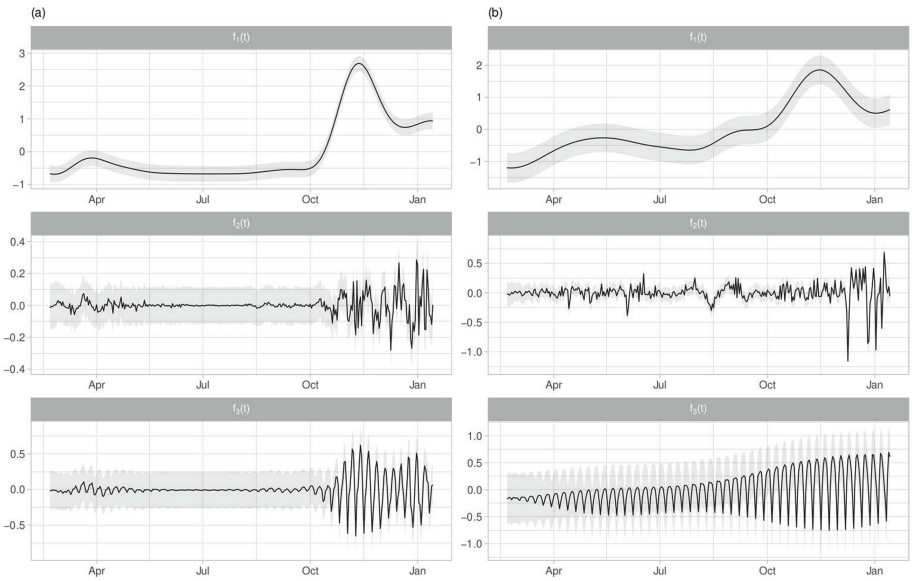


Fig. 4. Decomposition of the daily number of new cases (a) and the daily number of molecular swabs administered (b) into three linearly superimposed components, estimated using a sum of Gaussian processes. The proposed decomposition model includes: 1) Long-term trend f_1. 2) Shorter-term smooth fluctuation f_2. 3) Weekly periodic pattern with time-dependent amplitudes f_3. 4) Noise (not shown here). Gray shaded areas represent 95% posterior confidence intervals. Prior to decomposition, the original data were standardized to ensure that the empirical mean was close to zero.

The smoothed data are shown in Fig. 5. These smooth versions were obtained by subtracting the estimated periodic component from the original standardized data and then transforming back to the original scale.

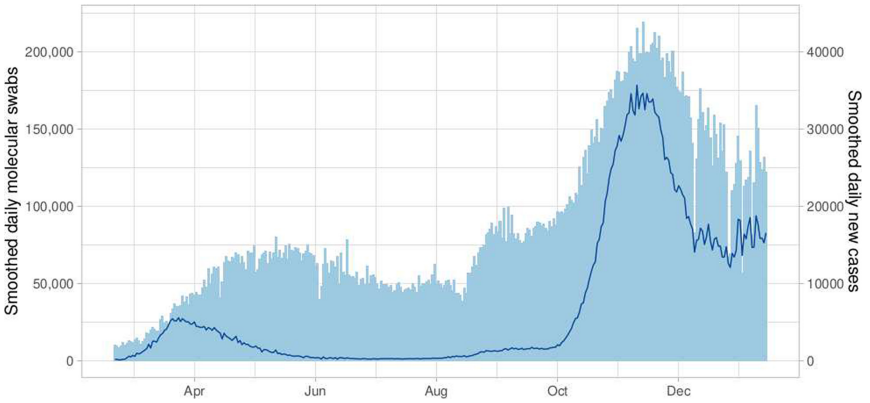


Fig. 5. Smoothed series of daily new cases and molecular swabs. These series were estimated by subtracting from the original data the irregular periodic weekly component, based on the sum of suitably specified Gaussian processes. The smoothed data were transformed back to the original scale.

7 Discussion and Conclusions

In this paper, we have presented a model for decomposing a time series into a set of additive components. Our approach is based on Gaussian processes, and the decomposition is achieved by exploiting the additivity of the mean and the covariance function of the sum of independent Gaussian processes. The proposed algorithm is different from those currently available in the literature. Indeed, the latter are generally based on a suitable rotation of the coordinate system in which the data are expressed, causing the new axes to lie according to the directions of maximum variability. Our approach, on the other hand, is less mechanistic, requiring ad hoc knowledge about the domain of study. On the one hand, this can be an advantage since the decomposition model can be calibrated accurately. From this point of view, it is obvious that the decomposition (29) is specific to the example we are dealing with and must be modified accordingly for other problems. In other contexts, the lack of an automatic procedure for determining the components of the additive decomposition may be a drawback. Whether or not the increased flexibility of our approach can be a disadvantage depends on the problem at hand.

The estimation algorithm we use does not provide a complete exploration of the posterior surface, but is limited to determining the MAP estimates of the parameters of interest. However, this approach requires caution since the target is not convex and local minima may be a problem. Sensitivity analysis to the initial conditions need to be further explored. Similarly, it is necessary to further investigate the sensitivity of the obtained results when the joint prior distribution of the parameters varies. All these problems will be addressed in future work.

References

1. Shumway, R.H., Stoffer, D.S.: Time Series Analysis and Its Applications. Springer, Heidelberg (2017). https://doi.org/10.1007/978-3-319-52452-8
2. Cleveland, R.B., Cleveland, W.S., McRae, J.E., Terpenning, I.: STL: a seasonal-trend decomposition procedure based on loess (with Discussion). J. Off. Stat. **6**, 3–73 (1990)
3. Zhang, X., et al.: Gaussian process. In: Encyclopedia of Machine Learning, pp. 428–439. Springer, US, Boston (2011). https://doi.org/10.1007/978-0-387-30164-8_324
4. Gonzalvez, J., Lezmi, E., Roncalli, T., Xu, J.: Financial applications of gaussian processes and bayesian optimization. SSRN Electron. J. (2019). https://doi.org/10.2139/ssrn.3344332
5. Roberts, S., Osborne, M., Ebden, M., Reece, S., Gibson, N., Aigrain, S.: Gaussian processes for time-series modelling. Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. **371**, 20110550 (2013). https://doi.org/10.1098/rsta.2011.0550
6. Golyandina, N., Zhigljavsky, A.: Singular Spectrum Analysis for Time Series. Springer, Heidelberg (2013). https://doi.org/10.1007/978-3-642-34913-3
7. Golyandina, N., Korobeynikov, A.: Basic Singular Spectrum Analysis and forecasting with R. Comput. Stat. Data Anal. **71**, 934–954 (2014). https://doi.org/10.1016/j.csda.2013.04.009
8. Golyandina, N., Korobeynikov, A., Shlemov, A., Usevich, K.: Multivariate and 2D extensions of singular spectrum analysis with the rssa package. J. Stat. Softw. **67**, 1–78 (2015). https://doi.org/10.18637/jss.v067.i02
9. Hassani, H., Mahmoudvand, R., Zokaei, M.: Separability and window length in singular spectrum analysis. Comptes Rendus Math. **349**, 987–990 (2011). https://doi.org/10.1016/j.crma.2011.07.012
10. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press (2005)
11. Stein, M.L.: Interpolation of Spatial Data. Springer, New York (1999). https://doi.org/10.1007/978-1-4612-1494-6
12. Duvenaud, D.K., Nickisch, H., Rasmussen, C.: Additive gaussian processes. In: Shawe-Taylor, J., Zemel, R., Bartlett, P., Pereira, F.,Weinberger, K.Q. (eds.) Advances in Neural Information Processing Systems, pp. 226–234 (2011)
13. MacKay, D.J.C.: Introduction to gaussian processes. In: Bishop, C.M. (ed.) Neural Networks and Machine Learning, pp. 133–166. Kluwer Academic Press (1998)
14. Durrande, N., Hensman, J., Rattray, M., Lawrence, N.D.: Detecting periodicities with Gaussian processes. PeerJ Comput. Sci. **2**, e50 (2016). https://doi.org/10.7717/peerj-cs.50
15. Cheng, L., Ramchandran, S., Vatanen, T., Lietzén, N., Lahesmaa, R., Vehtari, A., Lähdesmäki, H.: An additive Gaussian process regression model for interpretable non-parametric analysis of longitudinal data. Nat. Commun. **10**, 1798 (2019). https://doi.org/10.1038/s41467-019-09785-8
16. Vanhatalo, J., Riihimäki, J., Hartikainen, J., Jylänki, P., Tolvanen, V., Vehtari, A.: GPstuff: Bayesian modeling with gaussian processes. J. Mach. Learn. Res. **14**, 1175–1179 (2013)
17. Ravishanker, N., Dey, D.K.: A First Course in Linear Model Theory. Chapman & Hall/CRC, New York (2001)
18. Robert, C.P., Casella, G.: Monte Carlo Statistical Methods. Springer, New York (2004). https://doi.org/10.1007/978-1-4757-4145-2

19. Robert, C., Casella, G.: Introducing Monte Carlo Methods with R. Springer, New York (2010). https://doi.org/10.1007/978-1-4419-1576-4
20. Neal, R.: MCMC Using Hamiltonian Dynamics. In: Brooks, S., Gelman, A., Jones, G., Meng, X.-L. (eds.) Handbook of Markov Chain Monte Carlo, pp. 116–62. Chapman and Hall/CRC (2011). https://doi.org/10.1201/b10905
21. Hoffman, M.D., Gelman, A.: The No-U-Turn sampler: adaptively setting path lengths in hamiltonian Monte Carlo. J. Mach. Learn. Res. **15**, 1593–1623 (2014)
22. Betancourt, M., Girolami, M.: Hamiltonian Monte Carlo for hierarchical models. In: Current Trends in Bayesian Methodology with Applications, pp. 79–101. Chapman and Hall/CRC (2015). https://doi.org/10.1201/b18502-5
23. Carpenter, B., et al.: Stan: a probabilistic programming language. J. Stat. Softw. **76**, 1–32 (2017). https://doi.org/10.18637/jss.v076.i01
24. Template: COVID-19 pandemic data/Italy medical cases - Wikipedia. https://tinyurl.com/njz8y89p. Accessed 04 Feb 2021
25. Vehtari, A., Ojanen, J.: A survey of Bayesian predictive methods for model assessment, selection and comparison. Stat. Surv. **6**, 142–228 (2012). https://doi.org/10.1214/12-ss102
26. Gelman, A., Hwang, J., Vehtari, A.: Understanding predictive information criteria for Bayesian models. Stat. Comput. **24**(6), 997–1016 (2013). https://doi.org/10.1007/s11222-013-9416-2
27. Gelman, A.: Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). Bayesian Anal. **1**, 515–534 (2006). https://doi.org/10.1214/06-BA117A
28. Gelman, A.: Prior Choice Recommendations. https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations. Accessed 11 Feb 2020
29. R Core Team: R: A Language and Environment for Statistical Computing (2020). https://www.r-project.org/
30. Papastamoulis, P.: label.switching: an R package for dealing with the label switching problem in MCMC outputs. J. Stat. Softw. **69** (2016). https://doi.org/10.18637/jss.v069.c01
31. Betancourt, M.: Identifying Bayesian Mixture Models. https://mc-stan.org/users/documentation/case-studies/identifying_mixture_models.html. Accessed 22 Dec 2020