
University of Cape Town

Department of Statistical Sciences

Honours in Statistics & Data Science

# Gaussian Processes for Time Series Analysis

*Author:*
Raphaela Azah
Sbonelo Gumede

*Student Number:*
AZRRAP001
GMDSBO006

September 17, 2025

# Contents

# 1 Introduction

In this paper we offer a basic introduction into time series analysis. The main focus is on using Gaussian process, for time series analysis. We then benchmark the peformance of Gaussian processes to other time series models. An area of application included is predicting the level of air pollution in the Table View station in Cape Town for the year of 2019.

# 2 Literature Review

## 2.1 Introduction to Time Series Analysis

[4] We have a training set of $\mathcal{D}$ of $n$ observations, $\mathcal{D} = \{(\mathbf{x}_i, y_i) \,|\, i = 1, \ldots, n\}$, where $\mathbf{x}$ denotes the input vector (covariates) of dimension $\mathcal{D}$ and y denotes a scalar output or target (dependent variable); the column vector of inputs for all n cases are arranged in the $\mathcal{D} \times n$ *design matrix* $X$, and the targets are collected in the vector $\mathbf{y}$, so we can write $\mathcal{D} = (X, \mathbf{y})$.

**Definition 2.1.1** [5] *A time series is a sequence of observations collected at regular equally spaced intervals over a period of time.*

Time series data are extremely common. They arize in virtually every application field, such as business (sales figures, production numbers, and customer frequencies), economics (stock prices, exchange rates, and interest rates), and official statistics (census data, personal expenditures, and road casualities).

**Definition 2.1.2** [5] *Time series analysis is a collection of statistical techniques that attempt to isolate and quantify the influence of events and changes in conditions in order to build a model that utilizes this information to forecast future values of the time series.*

Standard inferential techniques which assume independence of observations (e.g. regression) do not work well when data is collected at regular equally spaced time intervals because the observations are likely to be dependent. When this dependence occurs between observations $g$ time periods apart, it is called 'autocorrelation' at lag $g$. So we cannot assume that the data consitute a random sample.

The basic assumption underlying time series forecasting is that the factors that influence patterns of activity in past and present will continue to do so in more or less the same manner in the future. Thus the overall purpose of time series analysis is to identify and isolate these influencing factors from the past in order to better understand the process underlying the time series, for predictive purposes.

**Definition 2.1.3** [5] *A time series is said to be stationary if it's statistical properties are constant over time. This implies that the time series has a constant mean and variance over time.*

Most time series we encounter are non-stationary and we often need to transform them so that they exhibit stationarity. These transformations enable us to consider what other information exists in the data after we have removed the effect of a trend or seasonality and/or changing variance.

The first step in time series analysis is to plot the data and observe any patterns that have occurred

over time using a line graph. The time series plot enables us to detect and describe components of past behavior of the series. The identified components help in finding a suitable statistical model to decribe the data, which enables us to forecast future values of the time series.

Components of a non-stationary time series are:

- Trend
- Cyclical variation
- Seasonal variation
- Random variation

### 2.1.1 Trend

The long-term tendency of a time series. The pattern observed may move steadily in an upward or downward direction, or stay the same over time.

### 2.1.2 Cyclical variation

Irregular long-term wavelike movements through a time series. Due to extendend periods of prosperity followed by extended periods of recession.

### 2.1.3 Seasonal variation

Regular short-term repetitive wavelike movements through a time series, often when data is recorded houly, daily, weekly, monthly or quarterly. It repeats itself through the time series.

### 2.1.4 Random variation

Random variations in the data due to the combined effects of all unforeseen events such as war, strikes, natural disasters, power cuts, etc.

The simplest assumption about the relationship between the components in a time series is that they are additive and independent of each other. We write the additive model as:

$$Y_t = T_t + C_t + S_t + R_t,$$

where $t$ is the time period we are interested in, $Y_t$ is the observed value of the time series at time $t$, $T_t$ is the value of the trend component at time $t$, $C_t$ is the value of the cyclical component at time $t$, $S_t$ is the value of the seasonal component at time $t$, and $R_t$ is the value of the random component at time $t$.

Alternatively, we can assume that the four components of the time series are not necessarily independent and can affect one another. This is captured by the multiplicative model as:

$$Y_t = T_t \times C_t \times S_t \times R_t.$$

Note that this can be made additive by taking the logarithm of the series.

The additive model is most appropriate if the magnitude of the seasonal fluctuations or variation around the trend-cycle does not vary with the level of the time series. When the variation in seasonal

pattern, or the variation around the trend-cycle, appears to be proportional to the level of the time series, then a multiplicative model is more appropriate.

## 2.2 Simple forecasting models

[5] There are some forecasting methods that are simple, yet remarkably effective for some time series.

We fit simple forecasting methods to use as 'benchmarks' against which we compare other more complex forecasting methods i.e. if a more complicated forecasting method does not yield better forecasts than one of the simple methods here, there is no need to use it for the particular time series analysed.

There are four simple forecasting methods that we consider:

- The average method
- The naive method
- The seasonal naive method
- The drift method

### 2.2.1 Average method

The forecasts of all future values is the mean of the historical data:

$$\hat{y}_{T+h|T} = \frac{\sum_{t=1}^{T} y_t}{T}.$$

The notation $\hat{y}_{T+h|T}$ is a short-hand for the estimate of $y_{T+h|T}$ based on the data $\{y_1, \ldots, y_T\}$.

### 2.2.2 Naive method

The forecasts of all future values is the last observation:

$$\hat{y}_{T+h|T} = y_T.$$

### 2.2.3 Seasonal naive method

This is similar to the naive method, used when we have a time series with a strong seasonal component. The forecasted value for a particular 'season' is simply the value corresponding to the previous 'season':

$$\hat{y}_{T+h|T} = y_{T+h-m(k+1)},$$

Where $m$ is the seasonal period and $k$ is the integer part of $(h-1)/m$ i.e. the number of complete seasonal variations that have passed since the end of the original time series data. For example, if our time series is montly data and a seasonal variation lasts 12 months, then the future forecasts for all April values is simply the previous April value. If our time series is quartely data and a season lasts 4 quarters, then the future forecasts for all quarter 3 values is the previous quarter 3 value etc.

### 2.2.4 Drift method

An extension of the naive method to allow for the presence of a linear trend in the data. The amount of change over time (called the 'drift') is equal to the average change in the historical data:

$$\hat{y}_{T+h|T} = y_T + h\Big(\frac{y_T - y_1}{T-1}\Big).$$

## 2.3 ARIMA models

ARIMA models are very useful for modelling a wide variety of stationary and non-stationary time series process. ARIMA stands for 'Auto Regressive Integrated Moving Average' and refers to models that are mixtures of autoregressive and moving average models.

### 2.3.1 Autoregressive models

The observed value at time $t$ depends linearly on the last $p$ observed values, and the model looks like a regression model. It is denoted AR($p$).

A process $\{X_t\}$ is said to be an autoregressive process of order $p$ if

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t,$$

where $\epsilon_t$ is a white noise process with mean 0 and variance $\sigma^2$ that represents the error of the series, and $c$ is a constant related to the mean of the process.

### 2.3.2 Moving Average models

Rather than using past values of the forecast variable in a regression, a moving average model uses past forecast errors in a regression-like model. It is denoted MA($q$).

A process $\{X_t\}$ is said to be a moving average process of order $q$ if

$$X_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q},$$

where $\epsilon_t$ is a white noise process with mean 0 and variance $\sigma^2$ that represents the error of the series, and $c$ is a constant related to the mean of the process.

If a non-stationary time series is modelled by both an AR($p$) and an MA($q$) process, we say that it is modelled by an ARIMA($p$, $d$, $q$) process. It takes the form

$$X_t = c + \phi_1 X_{t-1} + \phi_2 X_{t-2} + \ldots + \phi_p X_{t-p} + \epsilon_t + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} + \ldots + \theta_q \epsilon_{t-q},$$

where $p$ is the number of autoregressive terms, $d$ is the order of differencing and $q$ is the number of moving average terms.

## 2.4 Gaussian Processes

**Definition 2.3.1** [4] *A Gaussian process is a collection of random variables, any finite number of which have a joint Gaussian distribution.*

A Gaussian process is completely specified by its mean function and co-variance function. We define mean mean function $m(x_i)$ and the covariance function $k(x_i, x_j)$ for a real process $f(x_i)$ as

$$m(x_i) = \mathbb{E}[f(x_i)],$$

$$k(x_i, x_j) = \mathbb{E}[(f(x_i) - m(x_i))(f(x_j) - m(x_j))],$$

and will write the Gaussian process as

$$f(x_i) \sim \mathrm{GP}(m(x_i),\, k(x_i,\, x_j)).$$

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}_n(\mathbf{m},\, \mathbf{K}),\ \text{where}\ \mathbf{m} = \begin{bmatrix} m(x_1) \\ m(x_2) \\ \vdots \\ m(x_n) \end{bmatrix}\ \text{and}\ \mathbf{K} = \begin{bmatrix} k(x_1,\, x_1) & k(x_1,\, x_2) & \cdots & k(x_1,\, x_n) \\ k(x_2,\, x_1) & k(x_2,\, x_2) & \cdots & k(x_2,\, x_n) \\ \vdots & \vdots & \ddots & \vdots \\ k(x_n,\, x_1) & k(x_n,\, x_2) & \cdots & k(x_n,\, x_n) \end{bmatrix}.$$

### 2.4.1 Mean functions

Usually in the literature people use a zero-mean function. [2] This is because every Gaussian process can be recovered by adding the mean function to a corresponding zero-mean Gaussian process

$$f \sim \mathrm{GP}(0,\, k)$$

$$m + f \sim \mathrm{GP}(m,\, k),$$

In this paper we explore this and a multiple linear regression mean function $m_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$.

### 2.4.2 Covariance functions

[1] In the following section, we briefly describe commonly used kernels. We start with a simple white noise, and then consider common *stationary* covariances, both uni- and multi-dimensional. We finish this section with periodic and quasi-periodic kernel functions. We note that sums (and products) of valid covariance kernels give valid covariance functions (i.e. the resultant covariance matrices are positive semi-definite) and so we may entertain with ease multiple explanatory hypothesis.

*White noise* with variance $\sigma^2$ is represented by

$$k(x_i,\, x_j) = \sigma^2\delta_{ij},\ \text{where}\ \delta_{ij} = 1\,\text{for}\,i = j\,\text{and}\,\delta_{ij} = 0\,\text{for}\,i \neq j.$$

This kernel allows us to entertain uncertainity in our observed data and is so typically found added to other kernels.

*The SE kernel* is given by

$$k(x_i,\, x_j) = h^2\mathrm{exp}\bigg( -\Big(\frac{x_i - x_j}{\lambda}\Big)^2\bigg),$$

where $h$ is an output-scale amplitude and $\lambda$ is an input (length, or time) scale. This gives rather smooth variations with typical time scale of $\lambda$ and admits functions drawn from the GP that are infinitely differentiable.

*The rational quadratic (RQ) kernel* is given by

$$k(x_i,\, x_j) = h^2\bigg(1 + \frac{(x_i - x_j)^2}{\alpha\lambda^2}\bigg)^{-\alpha},$$

where $\alpha$ is known as the index. [4] show that this is equivalent to a scale mixture of SE kernels with different length scales, the latter distributed according to a Beta distribution with parameters $\alpha$ and $\lambda^{-2}$. This gives variations with a range of time scales, the distribution peaking around $\lambda$ but extending significantly longer period (but remaining rather smooth). When $\alpha \to \infty$, the RQ kernel reduces to the SE kernel with length scale $\lambda$.

*The Matérn* class of covariance functions is defined by

$$k(x_i,\, x_j) = h^2 \frac{1}{\Gamma(v)2^{v-1}} \left(2\sqrt{v}\frac{|x_i - x_j|}{\lambda}\right)\mathbb{B}_v\left(2\sqrt{v}\frac{|x_i - x_j|}{\lambda}\right),$$

where $h$ is the output scale, $\lambda$ is the input scale, $\Gamma()$ is the standard Gamma function and $\mathbb{B}()$ is the modified Bessel function of second order. The additional hyperparameter $v$ controls the degree of differentiability of the resultant functions modelled by a GP with a Matérn covariance function, such that there are only $(v + \frac{1}{2})$ times differentiable. As $v \to \infty$, so the functions become infinitely differentiable and the Matern kernel becomes the SE one. Taking $v = \frac{1}{2}$ gives the exponential kernel

$$k(x_i,\, x_j) = h^2 \exp\left( - \left(\frac{x_i - x_j}{\lambda}\right)^2\right),$$

which results in functions that are only once differentiable, and corresponds to the Ornstein-Unlenbeck process, the continuous time equivalent of a first-order autoregressive model, AR(1). Time series models corresponding to AR($p$) processes are discrete time equivalents of GP models with Matérn covariance functions with $v = p - \frac{1}{2}$.

*Multiple inputs and outputs.* The simple distance metric, $|x_1 - x_2|$, used thus far clearly allows only for the simplest case of one-dimensional input $x$, which we have hitherto tacitly assummed to represent a time measure. In general, however, we assume our input space has finite dimension and write $x^{(e)}$ for the value of the $e$th element in $\mathbf{x}$ and denote $x_i^{(e)}$ as the value of the $e$th element at the $i$th index point. In such scenarios, we entertain multiple exogenous variables. Fortunately, it is not difficult to extend covariance functions to allow for these multiple input dimensions. Perhaps the simplest approach is to take a covariance function that is the product of one-dimensional covariances over each input (the *product correlation* rule),

$$k(x_i,\, x_j) = \prod_e k^{(e)}(x_i^{(e)},\, x_j^{(e)}),$$

where $k^{(e)}$ is a valid covariance function over the $e$th input. As the product of covariances is a covariance, so this defines a valid covariance over the multi-dimensional input space. We can also introduce distance functions appropriate for multiple inputs, such as the Mahalanobis distance,

$$d^{(M)}(\mathbf{x}_i,\, \mathbf{x}_j,\, \mathbf{\Sigma}) = \sqrt{(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}}\mathbf{\Sigma}^{-1}(\mathbf{x}_i - \mathbf{x}_j)},$$

where $\mathbf{\Sigma}$ is a covariance matrix over the input variable vector $\mathbf{x}$. Note that this is a matrix which represents hyperparameters of the model, and should not be confused with covariances formed from the covariance functions.

For multi-dimensional outputs, we consider a multi-dimensional space consisting of a set of time series along with a label $l$, which indexes the time series, and $x$ denoting time. Together, these hence form

the two-dimensional set of $[l,\, x]$. We will then exploit the fact that a product of covariance functions is a covariance function in its own right, and write,

$$k([l_m,\, x_i],\, [l_n,\, x_j]) = k_x(x_i,\, x_j)k_l(l_m,\, l_n),$$

taking covariance function terms over both time and time-series label.

*Periodic and quasi-periodic kernels.* Note that a valid covariance function under any arbitrary (smooth) map remains a valid covariance function [1]. For any function $u : x \to u(x)$, a covariance function $k()$ defined over the range of $x$ gives rise to a valid covariance $k'(x)$ over the domain of $u$. Hence, we can use simple, stationary covariances in order to construct more complex (possibly non-stationary) covariances. A particular relevent example of this,

$$u(x) = (u^{(a)}(x),\, u^{(b)}(x)) = \left(\cos\left(2\pi\frac{x}{T}\right),\, \sin\left(2\pi\frac{x}{T}\right)\right),$$

allows us to modify our simple covariance functions above to model periodic functions. We can now take this covariance over $u$ as a valid covariance over $x$. As a result, we have the covariance function, for example of the squared exponential,

$$k(x_i,\, x_j;\, h,\, w,\, T) = h^2\exp\left(-\frac{1}{2w^2}\sin^2\left(\pi\left|\frac{x_i - x_j}{T}\right|\right)\right).$$

In this case, the output scale $h$ serves as the amplitude and $T$ is the period. The hyperparameter $w$ is a 'roughness' parameter that serves a role similar to the input scale $\lambda$ in stationary covariances. With this formulation, we can peform inference about functions of arbitrary roughness and arbitrary period. Indeed a periodic covariance function can be constructed from any kernel involving the squared distance $(x_i - x_j)^2$ by replacing the latter with $\sin^2[\pi(x_i - x_j)/T]$, where $T$ is the period. The length scale $w$ is now relative to the period, and letting $w \to \infty$ gives sinusoidal variations, while increasingly small values of $w$ give periodic variations with increasingly complex harmonic content. Similar periodic functions could also be used, as long as they give rise to a symmetric positive definite covariance matrix - $\sin^2$ is merely the simplest.

As described in Rasmussen & Williams [4], valid covariance functions can be constructed by adding or multiplying simpler covariance functions. Thus, we can obtain a quasi-periodic kernel simply by multiplying a periodic kernel with one of the basic stationary kernels described earlier. The latter then specifies the rate of evolution of the periodic signal. For example, we can multiply the previous equation with a squared exponential kernel,

$$k_{\text{QP, SE}}(x_i,\, x_j) = h^2\exp\left(-\frac{\sin^2[\pi(x_i - x_j)/T]}{2w^2} - \frac{(x_i - x_j)^2}{\lambda^2}\right),$$

to model a quasi-periodic signal with a single evolutionary time scale $\lambda$.

### 2.4.3 Prediction

[2] To simulate sampling a function from a Gaussian process and then evaluating the sampled function at the grid of covariate values we can just directly sample from a multivariate normal random number generator.

We can also consider *predictions* by taking advantage of the conditional structure of a multivariate normal density function. Consider a grid of observed covariate values

$$\{x_1^{\text{obs}},\, \ldots,\, x_{n_{\text{obs}}}^{\text{obs}}\}$$

where we know the function values $f(x_i^{\text{obs}})$, and a grid of unobserved covariate values where we want to predict the functional values,

$$\{x_1^{\text{pred}}, \ldots, x_{n_{\text{pred}}}^{\text{pred}}\}.$$

The parameters of the multivariate normal density function of the combined covariate values decomposes into the parameters of the multivariate normal density function for each component grid plus mixed covariate function evaluations.

$$\mathbf{m} = \begin{bmatrix} m(x_1^{\text{obs}}) \\ \vdots \\ m(x_{n_{\text{obs}}}^{\text{obs}}) \\ m(x_1^{\text{pred}}) \\ \vdots \\ m(x_{n_{\text{pred}}}^{\text{pred}}) \end{bmatrix} \text{and} \mathbf{K} = \begin{bmatrix} k(x_1^{\text{obs}}, x_1^{\text{obs}}) & \cdots & k(x_1^{\text{obs}}, x_{n_{\text{obs}}}^{\text{obs}}) & k(x_1^{\text{obs}}, x_1^{\text{pred}}) & \cdots & k(x_1^{\text{obs}}, x_{n_{\text{pred}}}^{\text{pred}}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ k(x_{n_{\text{obs}}}^{\text{obs}}, x_1^{\text{obs}}) & \cdots & k(x_{n_{\text{obs}}}^{\text{obs}}, x_{n_{\text{obs}}}^{\text{obs}}) & k(x_{n_{\text{obs}}}^{\text{obs}}, x_1^{\text{pred}}) & \cdots & k(x_{n_{\text{obs}}}^{\text{obs}}, x_{n_{\text{pred}}}^{\text{pred}}) \\ k(x_1^{\text{pred}}, x_1^{\text{obs}}) & \cdots & k(x_1^{\text{pred}}, x_{n_{\text{obs}}}^{\text{obs}}) & k(x_1^{\text{pred}}, x_1^{\text{pred}}) & \cdots & k(x_1^{\text{pred}}, x_{n_{\text{pred}}}^{\text{pred}}) \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ k(x_{n_{\text{pred}}}^{\text{pred}}, x_1^{\text{obs}}) & \cdots & k(x_{n_{\text{pred}}}^{\text{pred}}, x_{n_{\text{obs}}}^{\text{obs}}) & k(x_{n_{\text{pred}}}^{\text{pred}}, x_1^{\text{pred}}) & \cdots & k(x_{n_{\text{pred}}}^{\text{pred}}, x_{n_{\text{pred}}}^{\text{pred}}) \end{bmatrix}.$$

More compactly

$$\mathbf{m} = \begin{bmatrix} \mathbf{m}_{\text{obs}} \\ \\ \mathbf{m}_{\text{pred}} \end{bmatrix} \text{ and } \mathbf{K} = \begin{bmatrix} \mathbf{K}_{\text{obs}} & \mathbf{K}_{\text{mix}} \\ \\ (\mathbf{K}_{\text{mix}})^{\text{T}} & \mathbf{K}_{\text{pred}} \end{bmatrix}.$$

The conditional probability density function for the unobserved function values $(f_{\text{pred}})_i = f(x_i^{\text{pred}})$ given the observed function values $(f_{\text{obs}})_i = f(x_i^{\text{obs}})$ falls into the multivariate normal family,

$$\pi(\mathbf{f}_{\text{pred}}|\mathbf{f}_{\text{obs}}) \sim \mathcal{N}_n(\boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

with the location parameters

$$\boldsymbol{\mu} = \mathbf{m}_{\text{pred}} + (\mathbf{K}_{\text{mix}})^{\text{T}} \cdot (\mathbf{K}_{\text{obs}})^{-1} \cdot (\mathbf{f}_{\text{obs}} - \mathbf{m}_{\text{obs}}),$$

and the covariance matrix parameters

$$\boldsymbol{\Sigma} = \mathbf{K}_{\text{pred}} - (\mathbf{K}_{\text{mix}})^{\text{T}} \cdot (\mathbf{K}_{\text{obs}})^{-1} \cdot \mathbf{K}_{\text{mix}}.$$

Inference with Gaussian process priors proceeds similarly. Once we've identified the covariate values at which we have observations or will want to make predictions we can specify the marginalized prior model with the corresponding multivariate normal density function. If the observational model is normal then we can generate predictions analytically with the conditioning operations for the posterior covariance function.

# 3 Methodology

The three approaches that we consider are:

- Likelihood approach

- Bayesian approach
- Latent variable approach

Note that for all of these approaches we use the same likelihood function.

Since

$$f(x_i) \sim \text{GP}(m(x_i),\, k(x_i,\, x_j)).$$

Then

$$\mathbf{x} = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix} \sim \mathcal{N}_n(\mathbf{m},\, \mathbf{K}).$$

Therefore, the likelihood function is:

$$\begin{aligned} \mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x}|\, \alpha,\, \rho,\, \sigma) &= (2\pi)^{-\frac{n}{2}} \det(\mathbf{K})^{-\frac{1}{2}} \exp\Big( -\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}} \mathbf{K}^{-1} (\mathbf{x}-\mathbf{m}) \Big), \text{ for } \mathbf{x} \in \mathbb{R}^n, \\ &\propto \det(\mathbf{K})^{-\frac{1}{2}} \exp\Big( -\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}} \mathbf{K}^{-1} (\mathbf{x}-\mathbf{m}) \Big). \end{aligned}$$

## 3.1 Likelihood approach

### 3.1.1 Zero mean function and white noise kernel

Let

$$f(x_i) = 0$$
$$k(x_i,\, x_j) = \sigma^2 \delta_{ij}, \text{ where } \delta_{ij} = 1 \text{ for } i = j \text{ and } \delta_{ij} = 0 \text{ for } i \neq j.$$

Then

$$\mathbf{m} = \mathbf{0} \text{ and } \mathbf{K} = \sigma^2 \mathbf{I}_n$$

Therefore

$$\begin{aligned} \big[\hat{\sigma}\big] &= \arg\max_{\sigma} \quad \mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x}|\, \sigma) \\ &= \arg\min_{\sigma} \quad -\ln(\mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x}|\, \sigma)) \\ &= \arg\min_{\sigma} \quad -\ln\bigg( \det(\sigma^2 \mathbf{I}_n)^{-\frac{1}{2}} \exp\Big( -\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}} (\sigma^2 \mathbf{I}_n)^{-1} (\mathbf{x}-\mathbf{m}) \Big) \bigg) \\ &= \arg\min_{\sigma} \quad -\ln\bigg( (\sigma^2)^{-\frac{n}{2}} \exp\Big( -\frac{1}{2\sigma^2} \mathbf{x}^{\mathrm{T}} \mathbf{x} \Big) \bigg) \end{aligned}$$

One can use the 'optim' funtion in 'R' to obtain an estimate of $\sigma$.

### 3.1.2 Zero mean function and squared exponential kernel

Let

$$f(x_i) = 0$$
$$k(x_i,\, x_j) = \alpha^2 \exp\bigg( -\Big( \frac{x_i - x_j}{\rho} \Big)^2 \bigg)$$

Then

$$\mathbf{m} = \mathbf{0}$$

Therefore

$$\begin{aligned}
\begin{bmatrix}\hat{\alpha}\\ \hat{\rho}\end{bmatrix} &= \underset{\alpha,\,\rho}{\arg\max} \quad \mathcal{L}(\mathbf{m},\,\mathbf{K},\,\mathbf{x}|\,\alpha,\,\rho)\\
&= \underset{\alpha,\,\rho}{\arg\min} \quad -\ln(\mathcal{L}(\mathbf{m},\,\mathbf{K},\,\mathbf{x}|\,\alpha,\,\rho))\\
&= \underset{\alpha,\,\rho}{\arg\min} \quad -\ln\Big(\det(\mathbf{K})^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{m})\Big)\Big)\\
&= \underset{\alpha,\,\rho}{\arg\min} \quad -\ln\Big(\det(\mathbf{K})^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{x}\Big)\Big)
\end{aligned}$$

One can use the 'optim' funtion in 'R' to obtain estimates of $\alpha$ and $\rho$.

### 3.1.3 Multiple linear regression mean function and squared exponential kernel

$$f(x_i) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}.$$

$$k(x_i,\,x_j) = \alpha^2\exp\Big(-\Big(\frac{x_i - x_j}{\rho}\Big)^2\Big).$$

Then

$$\mathbf{m} = \mathbf{X}\boldsymbol{\beta}$$

Therefore

$$\begin{aligned}
\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\alpha}\\ \hat{\rho}\end{bmatrix} &= \underset{\boldsymbol{\beta},\,\alpha,\,\rho}{\arg\max} \quad \mathcal{L}(\mathbf{m},\,\mathbf{K},\,\mathbf{x}|\,\boldsymbol{\beta},\,\alpha,\,\rho)\\
&= \underset{\boldsymbol{\beta},\,\alpha,\,\rho}{\arg\min} \quad -\ln(\mathcal{L}(\mathbf{m},\,\mathbf{K},\,\mathbf{x}|\,\boldsymbol{\beta},\,\alpha,\,\rho))\\
&= \underset{\boldsymbol{\beta},\,\alpha,\,\rho}{\arg\min} \quad -\ln\Big(\det(\mathbf{K})^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{X}\boldsymbol{\beta})\Big)\Big)\\
&= \underset{\boldsymbol{\beta},\,\alpha,\,\rho}{\arg\min} \quad -\ln\Big(\det(\mathbf{K})^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{X}\boldsymbol{\beta})\Big)\Big)
\end{aligned}$$

One can use the 'optim' funtion in 'R' to obtain estimates of $\boldsymbol{\beta}$, $\alpha$ and $\rho$.

## 3.2 Bayesian approach

### 3.2.1 Zero mean function and squared exponential kernel

$$f(x_i) = 0.$$

$$k(x_i,\,x_j) = \sigma^2\delta_{ij},\ \text{where}\ \delta_{ij} = 1\ \text{for}\ i = j\ \text{and}\ \delta_{ij} = 0\ \text{for}\ i \neq j.$$

This is the simplest model we can fit. It has the implication that there is no trend in the data and observations are independent. These assumptions are not true for most time series.

$$\mathbf{m} = \mathbf{0} \text{ and } \mathbf{K} = \sigma^2 \mathbf{I}_n$$

The likelihood function becomes

$$\begin{aligned}\mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x}|\, \alpha,\, \rho,\, \sigma) &\propto \det(\mathbf{K})^{-\frac{1}{2}} \exp\Big( -\frac{1}{2}(\mathbf{x}-\mathbf{m})^{\mathrm{T}} \mathbf{K}^{-1} (\mathbf{x}-\mathbf{m})\Big) \\ &\propto \det(\sigma^2 \mathbf{I}_n)^{-\frac{1}{2}} \exp\Big( -\frac{1}{2}(\mathbf{x}-\mathbf{0})^{\mathrm{T}} (\sigma^2 \mathbf{I}_n)^{-1} (\mathbf{x}-\mathbf{0})\Big) \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp\Big( -\frac{1}{2\sigma^2} \mathbf{x}^{\mathrm{T}} \mathbf{x}\Big)\end{aligned}$$

Prior for $\sigma$ for a known $\phi$

$$\sigma \sim \text{half-normal}(0,\, \phi^2).$$

$$\begin{aligned}\pi(\sigma) &= \frac{2}{\sqrt{2\pi\phi^2}} \exp\Big( -\frac{\sigma^2}{2\phi^2}\Big), \text{ for } \sigma \geq 0, \\ &\propto \exp\Big( -\frac{\sigma^2}{2\phi^2}\Big).\end{aligned}$$

Therefore, the posterior distribution is

$$\begin{aligned}\pi(\sigma|\, \mathbf{x}) &\propto \mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x}|\sigma) \cdot \pi(\sigma) \text{ for } \sigma \geq 0, \\ &\propto (\sigma^2)^{-\frac{n}{2}} \exp\Big( -\frac{1}{2\sigma^2} \mathbf{x}^{\mathrm{T}} \mathbf{x}\Big) \cdot \exp\Big( -\frac{\sigma^2}{2\phi^2}\Big).\end{aligned}$$

### 3.2.2 Zero mean function and squared exponential kernel

$$f(x_i) = 0.$$

$$k(x_i,\, x_j) = \alpha^2 \exp\Big( -\Big(\frac{x_i - x_j}{\rho}\Big)^2\Big).$$

We could improve from the previous model by accounting for autocorrelation present in the data. We do this via the squared exponential kernel. It is a natural way of describing the autocorrelation decay.

Likelihood function

$$\mathcal{L}(\mathbf{m},\, \mathbf{K},\, \mathbf{x}|\, \alpha,\, \rho) \propto \det(\mathbf{K})^{-\frac{1}{2}} \exp\Big( -\frac{1}{2} \mathbf{x}^{\mathrm{T}} \mathbf{K}^{-1} \mathbf{x}\Big).$$

Prior for $\alpha$ for a known $\tau$

$$\alpha \sim \text{half-normal}(0,\, \tau^2).$$

$$\begin{aligned}\pi(\alpha) &= \frac{2}{\sqrt{2\pi\tau^2}} \exp\Big( -\frac{\alpha^2}{2\tau^2}\Big), \text{ for } \alpha \geq 0, \\ &\propto \exp\Big( -\frac{\alpha^2}{2\tau^2}\Big).\end{aligned}$$

Prior for $\rho$ for a known $\lambda$

$$\rho \sim \text{Inverse-Gamma}(\lambda,\, \beta).$$

$$\pi(\rho) = \frac{\beta^{\lambda}}{\Gamma(\lambda)}\rho^{-\lambda-1}\exp\Big(-\frac{\beta}{\rho}\Big), \text{ for } \rho > 0,$$
$$\propto \rho^{-\lambda-1}\exp\Big(-\frac{\beta}{\rho}\Big).$$

Posterior density

$$\pi(\alpha, \rho|\,\mathbf{x}) \propto \mathcal{L}(\mathbf{m}, \mathbf{K}, \mathbf{x}|\,\alpha, \rho)\cdot\pi(\alpha)\cdot\pi(\rho) \text{ for } \alpha \geq 0,\ \rho > 0$$
$$\propto \det(\mathbf{K})^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}\mathbf{x}^{\mathrm{T}}\mathbf{K}^{-1}\mathbf{x}\Big)\cdot\exp\Big(-\frac{\alpha^2}{2\tau^2}\Big)\cdot\rho^{-\lambda-1}\exp\Big(-\frac{\beta}{\rho}\Big).$$

### 3.2.3 Multiple linear regression mean function and squared exponential kernel

$$f(x_i) = \beta_0 + \sum_{j=1}^{p}\beta_j x_{ij}.$$
$$k(x_i, x_j) = \alpha^2\exp\bigg(-\Big(\frac{x_i - x_j}{\rho}\Big)^2\bigg).$$

A futher improvement on the previous models that allows accounting for a linear-trend and autocorrelation in the data. This is a reasonable assumption to make for some time series data. Then

$$\mathbf{m} = \mathbf{X}\boldsymbol{\beta}$$

The prior distributions for the beta coefficients are as follows

$$\boldsymbol{\beta} \sim \mathcal{N}_p(\mathbf{0}, \mathrm{I}_p).$$

$$\pi(\boldsymbol{\beta}) = (2\pi)^{-\frac{p}{2}}\exp\Big(-\frac{1}{2}(\boldsymbol{\beta}-\mathbf{0})^{\mathrm{T}}(\mathrm{I}_p)^{-1}(\boldsymbol{\beta}-\mathbf{0})\Big), \text{ for } \boldsymbol{\beta}\in\mathbb{R}^p,$$
$$\propto \exp\Big(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\Big).$$

Therefore, the posterior distribution is

$$\pi(\alpha, \rho|\,\mathbf{x}) \propto \mathcal{L}(\mathbf{m}, \mathbf{K}, \mathbf{x}|\,\alpha, \rho)\cdot\phi(\beta)\cdot\pi(\alpha)\cdot\pi(\rho) \text{ for } \alpha \geq 0,\ \rho > 0$$
$$\propto \det(\mathbf{K})^{-\frac{1}{2}}\exp\Big(-\frac{1}{2}(\mathbf{x}-\mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{x}-\mathbf{X}\boldsymbol{\beta})\Big)\cdot\exp\Big(-\frac{1}{2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\Big)\cdot\exp\Big(-\frac{\alpha^2}{2\tau^2}\Big)\cdot\rho^{-\lambda-1}\exp\Big(-\frac{\beta}{\rho}\Big)\cdot\exp\Big(-\frac{\sigma^2}{2\phi^2}$$

# 4 Application to air pollution forecasting

## 4.1 Description of the data

### 4.1.1 Exploratory data analysis

| Variable | Name | Description | Unit |
|---|---|---|---|
| $NO_2$ | Nitrogen dioxide | A harmful gas from vehicles and industry. | $\mu g/m^3$ |
| $PM_{10}$ | Particulate matter 10 | Small inhalable dust particles. | $\mu g/m^3$ |
| $SO_2$ | Sulphur dioxide | Mainly from burning fossil fuels. | $\mu g/m^3$ |
| Speed | Wind speed | How fast the wind is moving. | m/s |

Table 1: *Description of variables of the air pollution dataset.*

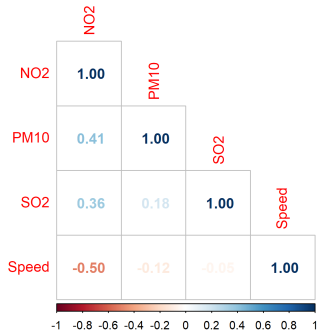
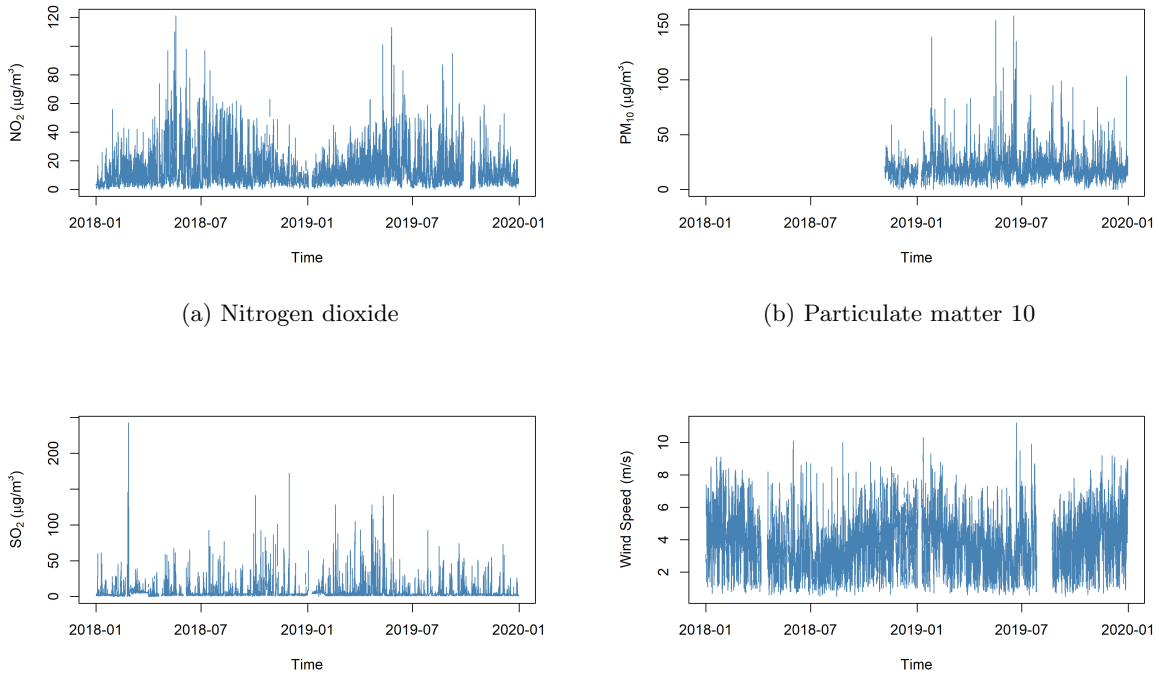



Figure 1: *Correlation plot of the air pollution dataset.*

Our response variable $NO_2$ appears to be moderately positively correlated with $PM_{10}$ and $SO_2$, and moderately negatively correlated with Speed. These are not ideal explanatory variables since we typically would like them to be strongly correlated with the response variable. The explanatory variables are weekly correlated with one another, whether it be positive or negative correlation. This is ideal since some models do not work well with correlated explanatory variables, often leading to unstable point estimates and inflated standard errors.




(a) Nitrogen dioxide

(b) Particulate matter 10

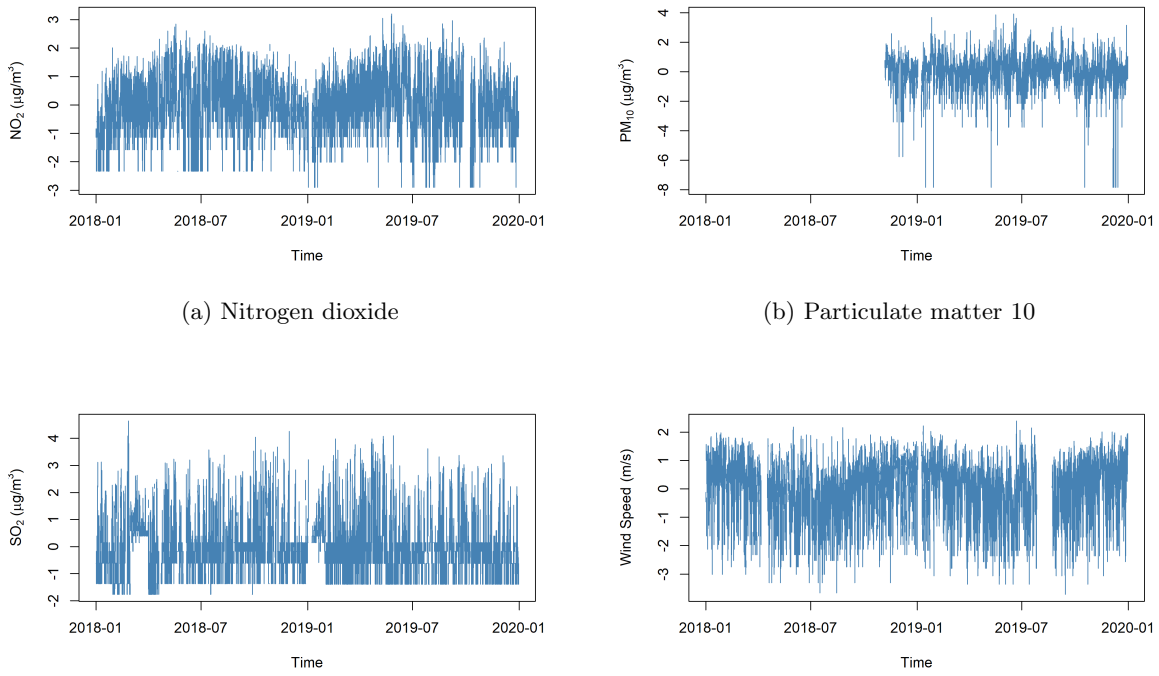
(c) Sulphur dioxide

(d) Wind speed

Figure 2: *Time-series plots of nitrogen dioxide and mereorological processes from 01/01/2018 to 31/12/2019 measured hourly. Similar cyclic patterns can be observed in the meteorological processes, and a weaker seasonality component can be noted in the yearly nitrogen dioxide processes.*

### 4.1.2 Data pre-processing

The response variable time-series $x_t$ is not statationary because the mean and variance are changing with time. To achieve stationarity there is a need for detrending and a variance stabilizing transformation. In order to stabilize the variance we use Box-Cox transformations in the training set, $y_t = (x_t^\lambda - 1)/\lambda$. The Box-Cox transformations allow us to experiment with a wide variety of $\lambda$ values. A good value of $\lambda$ is one the makes the variation in the data constant through time [5]. The `R` package `forecast` was used to perform the Box-Cox transformation which yielded an optimal value of $\hat{\lambda} \approx 0$, suggesting a logarithmic transformation. Then, conducting first-order differencing at lag one to remove the trend $z_t = y_t - y_{t-1}$ for $t \in \{2, 3, \ldots, p\}$. The time series $z_t$ is now stationary. Thus, we only need to account for the mean and variance in our models.




(a) Nitrogen dioxide

(b) Particulate matter 10

(c) Sulphur dioxide

(d) Wind speed

Figure 3: *Post-processed daily curves of the nitrogen dioxide and meteorological data, both standardized by their overall mean and standard deviation.*

## 4.2 Results

| Model | RMSE | | | MAE | | |
|---|---|---|---|---|---|---|
| | Forecasts ($h$ day time horizon) | | | Forecasts ($h$ day time horizon) | | |
| | 24 | 168 | 744 | 24 | 168 | 744 |
| Average | 7.731 | **5.397** | 7.439 | **6.042** | **4.254** | 5.385 |
| Naive | 10.553 | 7.864 | 10.040 | 7.708 | 6.071 | 7.308 |
| Drift | 10.606 | 8.128 | 11.358 | 7.764 | 6.391 | 8.809 |
| AR(1) | 8.662 | 5.912 | 8.044 | 6.029 | 4.422 | 5.616 |
| GP-0-WN | 13.756 | 11.141 | 13.169 | 11.719 | 9.786 | 11.081 |
| GP-0-SE | | | | | | |
| GP-MLR-WN | **7.692** | 6.134 | **7.225** | 6.119 | 4.490 | **5.136** |
| GP-MLR-SE | | | | | | |

Table 2: *RMSE and MAE of forecasting models across horizons.*

# References

[1] Roberts et al. "Gaussian Processes for Time-Series Modelling". In: *Philosophical Transactions of the Royal Society* (2013). URL: https://royalsocietypublishing.org/.
[2] Michael Betancourt. *Robust Gaussian Process Modeling*. https://betanalpha.github.io/assets/case_studies/gaussian_processes.html. 2020.
[3] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2019. URL: https://www.R-project.org/.
[4] Rasmussen and Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2006. URL: https://gaussianprocess.org/gpml/.
[5] Neil Watson. *Introduction to Time Series Analysis*. Unknown Publisher, 2025.

# 5 Appendix

## 5.1 Code: