Gaussian Processes for Time Series

Raphaela Azar and Sbonelo Gumede

August 7, 2025

Introduction

In this paper we hope to define what is a Gaussian process. How can you use a Gaussian process for time series analysis? Compare the peformance of a Gaussian process to other time series analysis models such as state-space 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. Another area of application is predicting the gold price for South Africa for the year of 2020. This is done in a Bayesian framework in order to make probabilistic statements and obtain a predictive distribution for future forecasts.

A Gaussian process is a stochastic process $\{X_t\}_{t \in \mathcal{I}}$, where the parameter space $\mathcal{I}$ and the state space $\mathcal{U}$ are real numbers. At each point $t \in \mathcal{I}$, the state $X_t \sim \mathcal{N}(f(t), g(t))$, where $f(t)$ is a mean function and $g(t)$ is the variance function. Since $g(t)$ denotes the variance then it must be that $g(t) \geq 0 \, \forall \, t \in \mathcal{I}$. Consider a simple example shown in figure 1, where $f(t) = t + e$, where $e \sim \mathcal{N}(0, 1)$ and $g(t) = \sin^2(t) + 1$.
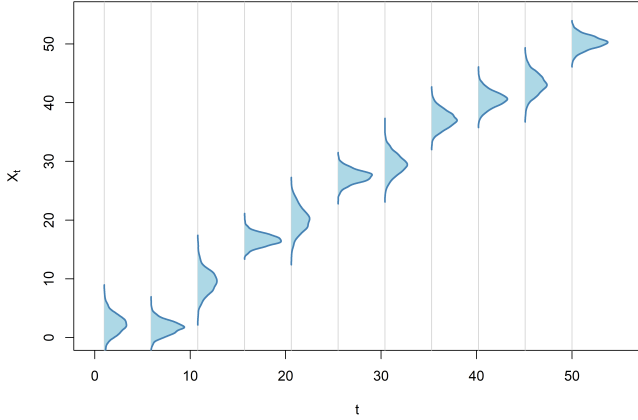

Figure 1: An example of a Gaussian process.

We defined $g(t)$ as being a univariate non-negative function of the variance over time. We can generalize this to a function of the covariance over time. Simply we consider two points t_i and t_j, for i and $j \in \{1, \ldots, p\}$ such that $g(t_i, t_j)$ gives us the covariance between the two points. Consider a countable finite set of points of the parameter space $\{t_1, \ldots, t_p\}$. Each $X_{t_k} \sim \mathcal{N}(f(t_k), g(t_k))$, for $k \in \{1, \ldots, p\}$.

Thus $\mathbf{X} = \begin{bmatrix} X_{t_1} \\ \vdots \\ X_{t_p} \end{bmatrix} \sim \mathcal{N}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu} = \begin{bmatrix} f(t_1) \\ f(t_2) \\ \vdots \\ f(t_p) \end{bmatrix}$ and $\boldsymbol{\Sigma} = \begin{bmatrix} g(t_1, t_1) & g(t_1, t_2) & \cdots & g(t_1, t_p) \\ g(t_2, t_1) & g(t_2, t_2) & \cdots & g(t_2, t_p) \\ \vdots & & \ddots & \vdots \\ g(t_p, t_1) & g(t_p, t_2) & \cdots & g(t_p, t_p) \end{bmatrix}$.

Often in the literature people use $f(t) = 0$ and $g(t_i, t_j) = \alpha^2 \exp(-\frac{1}{2}(\frac{|t_i - t_j|}{\rho})^2)$, where α and ρ are hyperparameters. The latter is known as the squared-exponential kernel. Some kernel functions common in the literature are the following:

There are two ways of modeling time series data. The first one is functional mapping y = f(x) + e, where y is the response variable, x is the explanatory variable, and e is typically white additive noise. The noise is assumed to be uncorrelated with one another. The second one is curve fitting, where x is often time and y is the response. Often, we work with random effects models $y_i = X_i^T \boldsymbol{\beta_i} + a_i + e_i$, where $X_i^T \boldsymbol{\beta_i}$ is the regression term, a_i is the random effect term, and e_i is the random variation.

Air pollution example

Exploratory data analysis

Variable	Name	Description	Unit
NO$_2$	Nitrogen dioxide	A harmful gas from vehicles and industry.	μg/m^3
PM$_{10}$	Particulate matter 10	Small inhalable dust particles.	μg/m^3
SO$_2$	Sulphur dioxide	Mainly from burning fossil fuels.	μg/m^3
Direction	Wind direction	Indicates where the wind is coming from.	Degrees (0–360˚)
Speed	Wind speed	How fast the wind is moving.	m/s

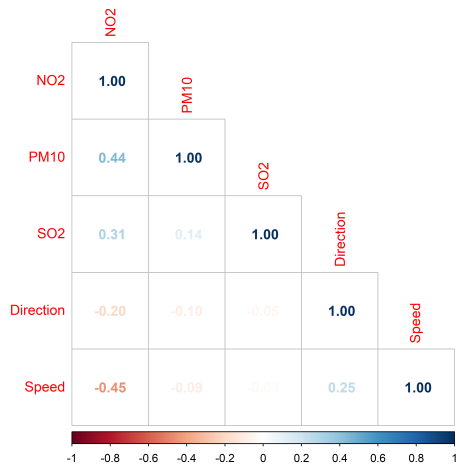Table 1: Description of variables used in the analysis.

Figure 2: Correlation plot of the variables.

Our response variable NO$_2$ appears to be moderately positively correlated with PM$_{10}$ and SO$_2$, weakly negatively correlated with Direction 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.

Before looking at the data. The proposed model for the mean function is $f(x) = x^T\beta + \alpha$ and $g(x) = \alpha^2\exp(-\frac{1}{2}(\frac{|x_i - x_j|}{\rho})^2)$ for the covariance function. This allows us to capture the trend (if it exists) in the time series through regression. This also allows us to model a changing variance since it is likely that the levels of nitrogen in the atmosphere differ thoughout the seasons of the year.
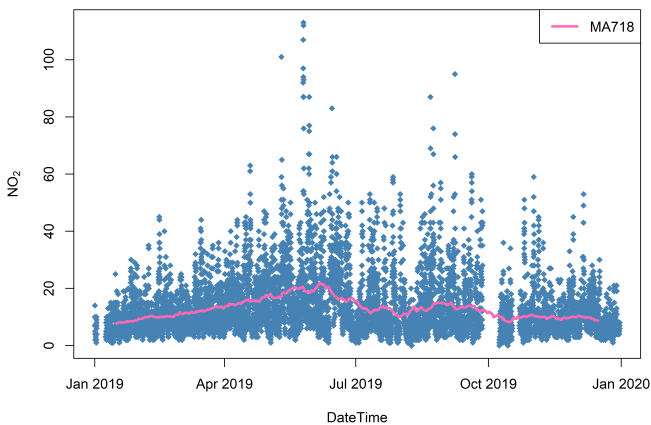
(a) Scatter plot of NO$_2$.
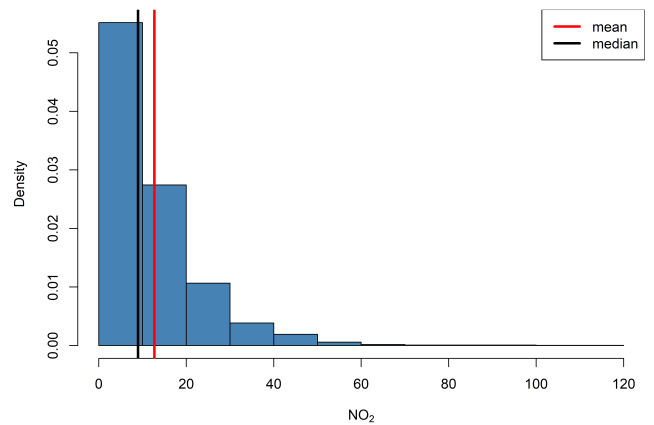
(b) Histogram of NO$_2$.

The mean appears to be constant, but the variance seems to be changing with time. Thus it is better if $f(x) = 0$ and $g(x) = \alpha^2 \exp(-\frac{1}{2}(\frac{|x_i - x_j|}{\rho})^2)$. The prior distributions of the exponantiated quadratic kernel are the following: $\alpha \sim \text{half-normal}(0, 1)$, $\rho \sim \mathcal{IG}(20, 10)$, and $\sigma \sim \text{half-normal}(0, 1)$. There is seasonality since regular patterns repeat for the seasons of the year. Random variation is present. However, there is no cyclical component present.

References

1. https://www.lung.org/clean-air/outdoors/what-makes-air-unhealthy/nitrogen-dioxide