Gaussian Process Regression Applied to Climate Data

Sithila Premnath

McMaster University

STATS 4W03: Reading in Statistics

Abstract— **The dataset used to conduct this project consists of the responses collected from gas multi sensors deployed to measure emissions into the atmosphere by industrial waste. [4] Using Principle Component Regression (PCR) and Generalized Additive Models (GAMs), I have predicted how Benzene (C6H6) concentration is effected by the other gases present in the dataset. This was done using GAMs as well as Gaussian Process Regression conducted on a few principle components that were shown to have a strong relationship with Benzene. The results obtained show us that Titania, Tin Oxide and Carbon Monoxide has a direct relationship with Benzene concentrations. R code has been uploaded to Github at https://github.com/sith21/stats-4w03-code-project-sithila/tree/main**

Index Terms— **gaussian process regerssion – climate change– principle components**

I. INTRODUCTION

Linear regression (LR) provides a simplistic approach to analysis, where the slope and intercept of a line are used to predict whether two variables show a statistical relationship. Since we are unable to completely remove uncertainty from our data, probability distributions can offer a reliable solution. Gaussian Process Regression (GPR) is a non-parametric, probabilistic, Bayesian approach to making regression. The attractiveness of GPR lies in its ability to solve a variety of supervised learning problems for small datasets. We employ such GP to predict how Benzene - a pollutant that causes smog and pollutes water bodies - concentrations are effected by several gases recorded by a multi sensor device deployed in a climate measurement survey.

The dataset used [9] contains 9358 instances of hourly averaged responses of 5 metal oxides and ground truth hourly averaged concentrations of Carbon Monoxide (CO), Non-methane Hydrocarbons (NMHC), Benzene (C6H6), Nitrogen Oxides (NOx) and Nitrogen Dioxide (NO2). For this project, I will be going over the theory behind GP and GPR in section 2 and in section 3 I will discuss the predictions these measurements have on Benzene concentrations. Finally in section 4, I will conclude and explain how my results effect climate change due to air pollutants.).

II. THEORY

Before getting into GPR, I will talk about Gaussian Processes (GP). GP is class of stochastic processes that

define the realizations of random variables as functions of space and time. A GP is a collection of random variables, X, such that any subset of X is Gaussian [?].

$$\boldsymbol{X}_{i,...,}\boldsymbol{X}_j \sim \mathbf{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{1}$$

Based on a given input X, we add probabilities to a number of functions $f_n(X)$ for n $\in$ N and find which probability is more likely - GPR calculates the probability distribution over all functions that fit the data. This Bayesian approach infers a probability distribution over all possible values of $f_n(X)$. Assuming a linear function $y = wx + \epsilon$, the Bayesian approach specifies a prior distribution, $p(w)$ on the parameter w which is used to find an updated distribution, $p(w|y, X)$, known as the posterior distribution. Bayesian inference (BI) updates the probability, $p(w|y, X)$, distribution as more information becomes available and updates this probability distribution to fit the data. Given a random plot of points (page 3), if we implement linear regression there is uncertainty in the line used to fit the data. BI aims to fix this by producing a sample of smooth functions on the observed data, known as the prior sample. Using Bayes theorem, the model is updated to produce a probability distribution with an adjustment in our assumptions (posterior sample), where noise variance is added to get a predictive distribution, averaged over many intervals (page 3). A good example is predicting carbon monoxide concentrations with no precise data on its concentration within a certain time period (prior). However, we do know that gases such as methane and carbon dioxide are effected by carbon monoxide reactions in the air [14]. BI uses this data to update the probability distribution of carbon monoxide as the posterior.

In GPR, we assume a Gaussian process prior that takes the mean, $m(\boldsymbol{X})$ and covariance, $k(\boldsymbol{x}, \boldsymbol{x}')$ functions of the realization of the GP at some X, where X is a matrix of (n,d) training features and d is a collection of random variables that are jointly Gaussian.

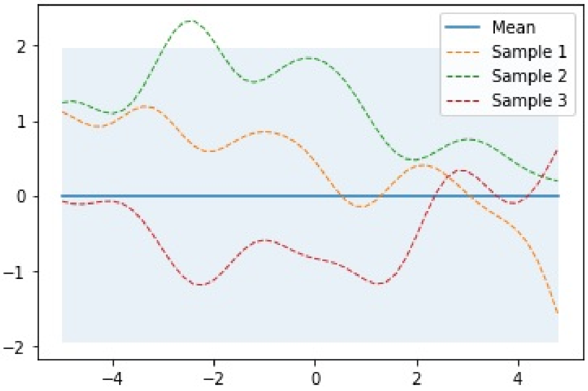
$$m(\boldsymbol{X}) = \mathbf{E}\left[f(\boldsymbol{X})\right]$$

$$k(\boldsymbol{x}, \boldsymbol{x}') = \mathbf{E}\left[f(\boldsymbol{X}) - m(\boldsymbol{X})(f(\boldsymbol{X}' - m(\boldsymbol{X}'))\right]$$

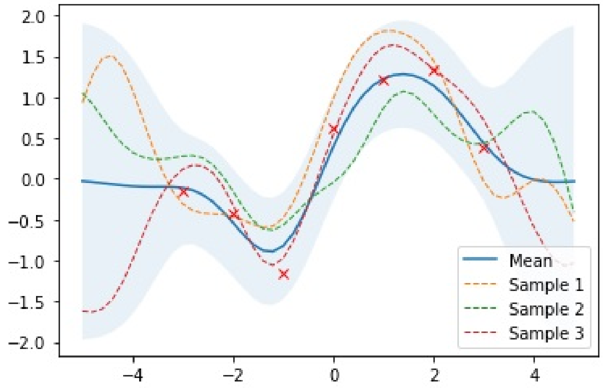
for an $f(\boldsymbol{X}) \sim GP(m(\boldsymbol{X}), k(\boldsymbol{X}, \boldsymbol{X}'))$ where E denotes the expected value of a function. Estimates of the mean of $f(X)$ are produced as linear combinations of an observed target value y, where y is an (n,1) vector of training points; using GP with our dataset leaves y as measurements of benzene.

III. Results

Spline functions displayed in figure 5 show us how using an array of splines to predict Benzene can have a distinct effect on predictions. An accurate model has a residuals vs linear predictor graph centered around 0 and a linear correlation of response vs fitted values. The results suggest that spline functions on Titania and Carbon Monoxide and Tin Oxide have a noticeable relationship, indicating that predictions of Benzene concentration are effected by these measurements. However Titania, or more commonly known as Titanium Dioxide, is a photo catalyst used to decompose Benzene under ultraviolet lab conditions and cannot have any effect in gaseous emissions in the atmosphere. Tin Oxide is another photo catalyst used to degrade Benzene into Carbon Dioxide and water, leaving Carbon Monoxide to



Random sample of functions [8]



Posterior sample incorporating noise to the Gaussian fit [8]

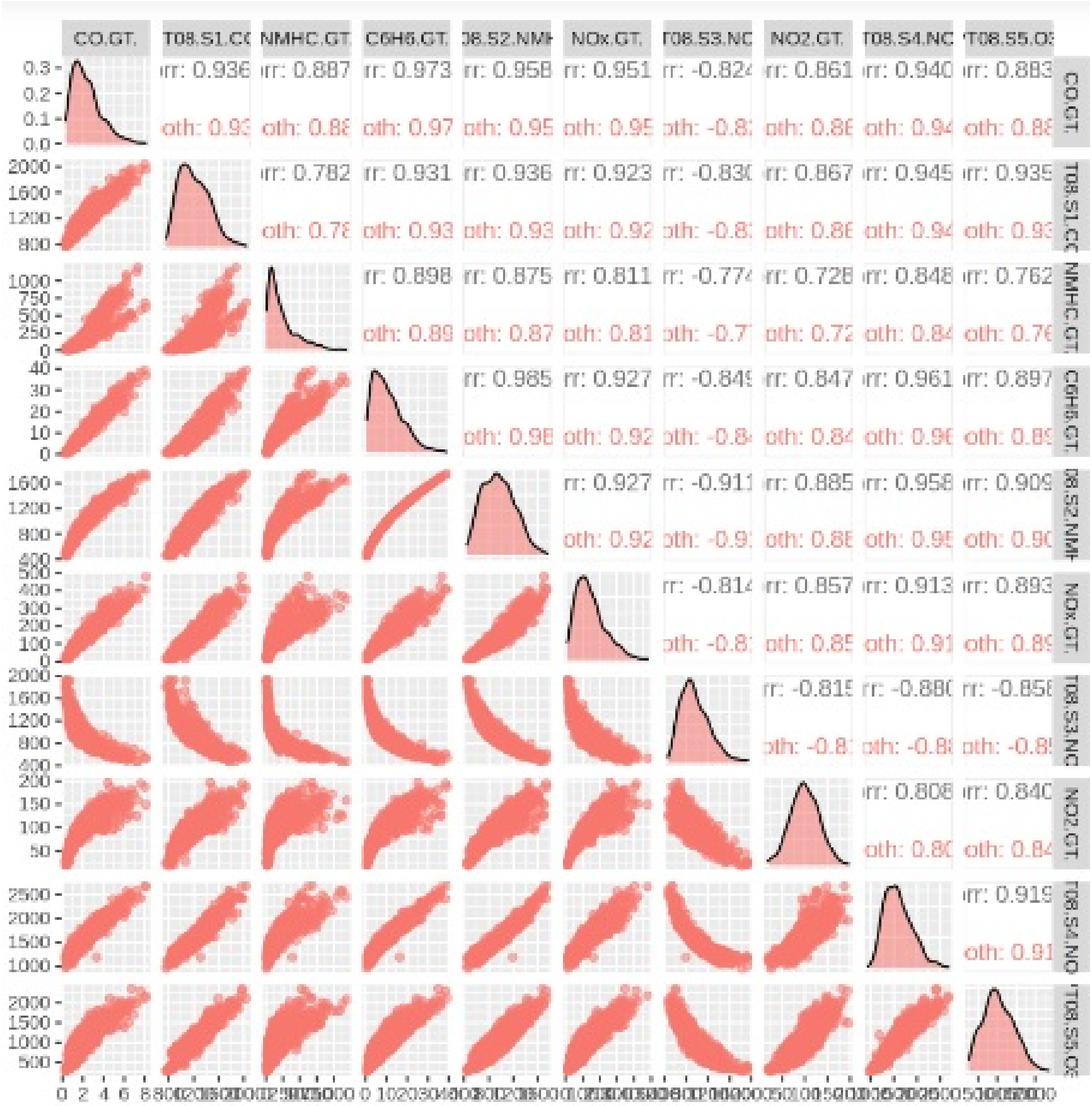
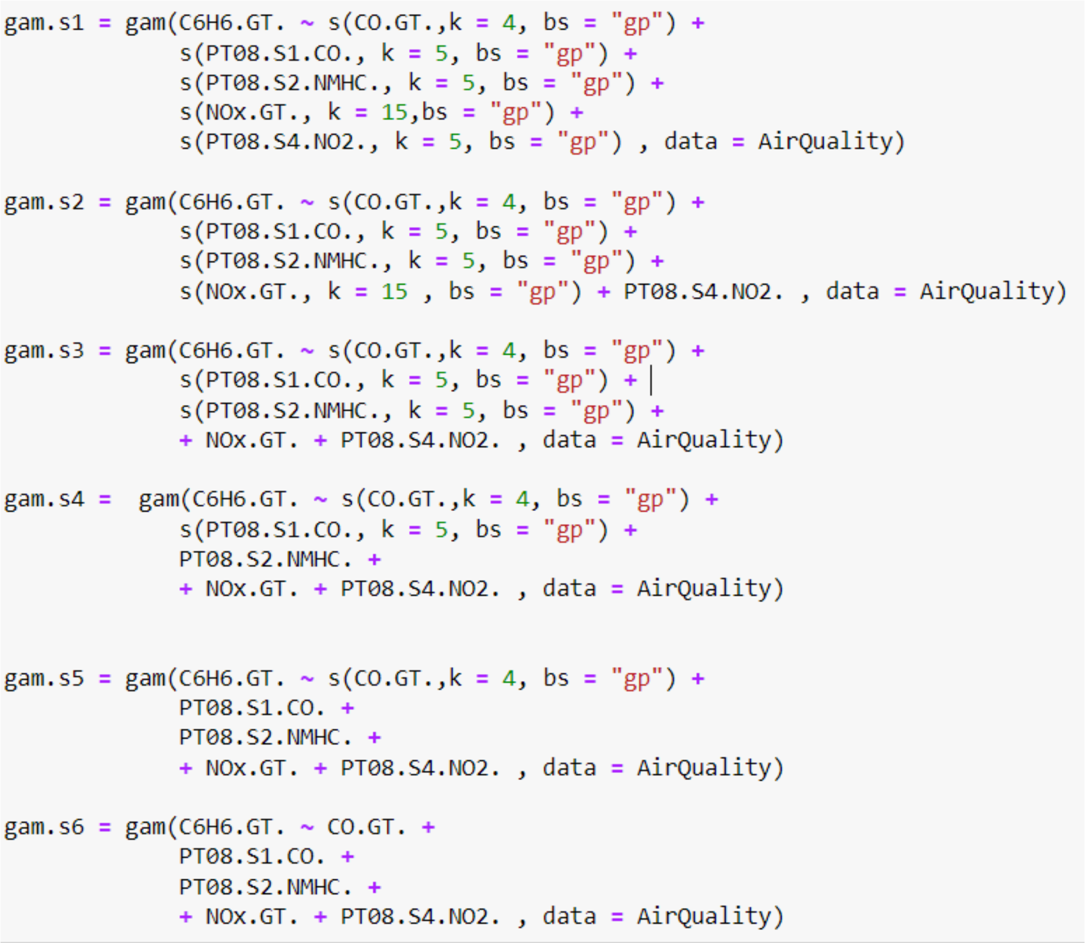


Fig. 1. Scatterplots of each pair of measurements are drawn on the left part of the figure. Pearson correlation is displayed on the right and the variable distribution is available on the diagonal.



```
gam.s1 = gam(C6H6.GT. ~ s(CO.GT.,k = 4, bs = "gp") +
             s(PT08.S1.CO., k = 5, bs = "gp") +
             s(PT08.S2.NMHC., k = 5, bs = "gp") +
             s(NOx.GT., k = 15,bs = "gp") +
             s(PT08.S4.NO2., k = 5, bs = "gp") , data = AirQuality)

gam.s2 = gam(C6H6.GT. ~ s(CO.GT.,k = 4, bs = "gp") +
             s(PT08.S1.CO., k = 5, bs = "gp") +
             s(PT08.S2.NMHC., k = 5, bs = "gp") +
             s(NOx.GT., k = 15 , bs = "gp") + PT08.S4.NO2. , data = AirQuality)

gam.s3 = gam(C6H6.GT. ~ s(CO.GT.,k = 4, bs = "gp") +
             s(PT08.S1.CO., k = 5, bs = "gp") +
             s(PT08.S2.NMHC., k = 5, bs = "gp") +
             + NOx.GT. + PT08.S4.NO2. , data = AirQuality)

gam.s4 =  gam(C6H6.GT. ~ s(CO.GT.,k = 4, bs = "gp") +
             s(PT08.S1.CO., k = 5, bs = "gp") +
             PT08.S2.NMHC. +
             + NOx.GT. + PT08.S4.NO2. , data = AirQuality)


gam.s5 = gam(C6H6.GT. ~ s(CO.GT.,k = 4, bs = "gp") +
             PT08.S1.CO. +
             PT08.S2.NMHC. +
             + NOx.GT. + PT08.S4.NO2. , data = AirQuality)

gam.s6 = gam(C6H6.GT. ~ CO.GT. +
             PT08.S1.CO. +
             PT08.S2.NMHC. +
             + NOx.GT. + PT08.S4.NO2. , data = AirQuality)
```

Fig. 2. 6 GAM functions used in predicting Benzene concentrations

be the only measurement that directly effects Benzene composition in the atmosphere.

IV. Conclusions

The results obtained indicate that Gaussian Process Regression infers accurate Carbon Monoxide dependencies on Benzene concentrations in the atmosphere. The observations on Titania and Tin Oxide can be attributed to inertness of reaction due to the lack of requirements needed for chemical decomposition of Benzene, thus the relationship observed by both chemicals shown by the GPR.

Benzene is a pollutant that can react with other chemicals to create smog which can attach to rain and leads to contaminated water and soil. Furthermore, Benzene has been attributed to adverse health affects such as bone marrow damage leading to anemia or leukemia [13], immune system depression, irregular menstrual periods and produces carbon dioxide when reacted with oxygen in the atmosphere. The partial plots (figure 1) tell us that benzene can be found where there is a high concentration of carbon monoxide. Both chemicals are known to adversely effect the atmosphere and efforts to reduce their concentrations in the air have been taken. Most notably from my results, Titania has been used as a catalyst to reduce benzene into both carbon dioxide and water. The carbon dioxide is then reacted with methane to produce a mixture of hydrogen

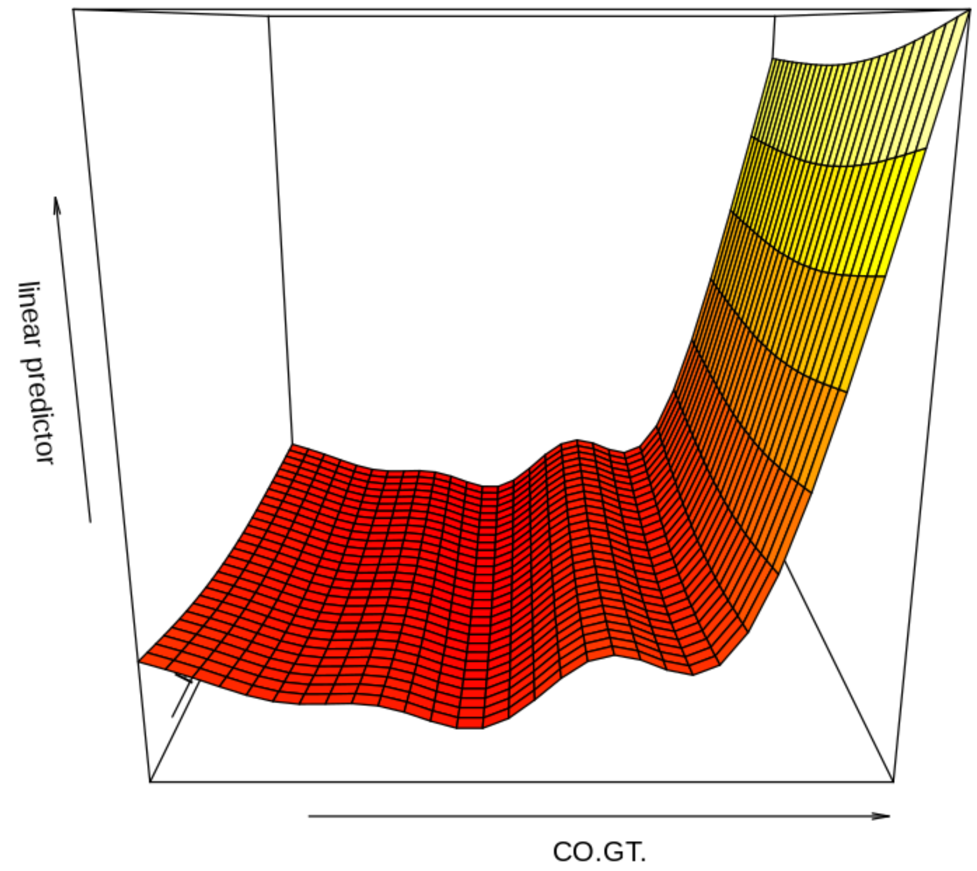


Fig. 3. A contour plot of the PCA conducted on Carbon Monoxide and Tin Oxide.

and carbon monoxide that is converted into ammonia. [10]

REFERENCES

[1] **Hilarie Sitt, Towards Data Science, Quick Start To Gaussian Process Regression**

[2] **Gaussian Process Regression From First Principles, Ryan Sander, MIT**

[3] **1.7 Gaussian Process, SciKit Learn**

[4] **Debin Fang, Xiaoling Zhang, Qian Yu, Trenton Chen Jin, Luan Tian, A novel method for carbon dioxide emission forecasting based on improved Gaussian processes regression, Journal of Cleaner Production, Volume 173, 2018, Pages 143-150, ISSN 0959-6526, https://doi.org/10.1016/j.jclepro.2017.05.102.**

[5] **How does climate change affect migration?, Stanford Earth Matters**

[6] **Gaussian Process Regression with tfprobability, Sigrid Keydana**

[7] **Noamross, Visualizing GAMs https://noamross.github.io/gams-in-r-course/chapter2**

[8] **Krasserm, Gaussian Processes for Classification http://krasserm.github.io/2020/11/04/gaussian-processes-classification/**

[9] **UCI Machine Learning Repository https://archive.ics.uci.edu/ml/datasets/air+quality**

[10] **Nanocomposite Titania–Carbon Spheres as CO2 and CH4 Sorbents https://pubs.acs.org/doi/10.1021/acsomega.9b03806**

[11] **https://noamross.github.io/gams-in-r-course/chapter2**

[12] **http://gaussianprocess.org/gpml/chapters/RW2.pdf**

[13] **https://emergency.cdc.gov/agent/benzene/basics/facts.asp**

[14] **https://www.epa.gov/ghgemissions/overview-greenhouse-gases**