Application of PCA and KPCA to Ozone and CO2 Emissions Data

PSTAT 262 FE Final Project

Callum Weinberg*

*UCSB

PSTAT 262 FE

May 27th, 2022

Table of Contents

Table of Contents

Background

- CO2 and other greenhouse gas emissions are known to contribute to global warming and global climate change.
- The local effects of CO2 emissions are less clear: CO2 diffuses easily in the atmosphere, and therefore CO2 levels are generally similar throughout the globe.[6][5, p. 2497]
- However local levels can be higher near emission sources (particularly in urban and industrial areas.

Background

- In a 2010 paper, Mark Jacobson researches the effects of "domes" of CO2 over cities. The results of the paper suggest that CO2 levels may increase ozone and particulate matter and also may be associated with higher premature mortality.[5, p. 2497]
- Heightened ground-level ozone concentrations are associated with negative health outcomes.[2].
- Therefore it is important to understand any relationship between local CO2 emissions and local ozone concentration.

Background: Feature Extraction

- Since geospatial-temporal data can be high dimensional, feature extraction methods such as Principal Component Analysis (PCA) may be useful in dimensionality reduction for implementing statistical learning.
- Additionally, PCA is useful when covariates are highly correlated, as traditional regression techniques may fail/struggle (the matrix $X'X$ may be non-invertible or close to non-invertible). This applies to the CO2 data in this project.
- Kernel PCA (KPCA) allows for non-linear projections of the data by implementing and explicit or implicit feature map.

Background: Purpose

1. Explore the relationship between ozone concentration and CO2 emissions, by sector, at a local level
2. Do so by implementing PCA, KPCA, and PCR and KPCR (Principal Component Regression and Kernel Principal Component Regression). In doing so, explore Computational methods for PCR and KCPR and explore the implementation of different kernels.

Table of Contents

PCA

- The general idea in PCA is to solve the problem $X_{N\times p}W_{p\times p} = Z_{N\times d}$
- Where X represents the data, and Z is a projection of the data. The columns of Z are the principal components. The projection (defined by matrix W) projects the data onto a linear subspace that maximizes the variance of the data.[7, Part II Section 3.1]
- PCA can also be formulated as a case of the "eigenvalue problem", such that: $C_X W_X = \lambda W_X$, where C_X is the covariance Matrix of X: $X'X$.

Methods for Performing PCA

- Eigenvalue Decomposition
- Singular Value Decomposition of the Form: $X = U\Sigma V^T$, where $U\Sigma$ results in the principal components and V is the projection matrix.
- NIPALS algorithm. See R Markdown file.

Kernel PCA

- It may be that the principal components are better represented by a non-linear projection of the original data.
- Transform the data via a kernel (called the Gram matrix) into a higher dimensional space. In this higher dimensional space, a linear projection may be suitable and PCA can be performed.
- Kernel matrix is defined as $K = \phi(X)\phi^T(X)$. $\phi(X)$ is often unknown, but we can use the Kernel can be calculated from X, which provides point representation of the functions of the feature space.
- Matrix K is effectively treated as the Covariance Matrix, C_X in linear PCA.

Kernels

$$\text{RBF Kernel: } k(x, x') = \sigma^2 exp\left(\frac{-(x-x')^2}{2l^2}\right)$$

$$\text{RQF Kernel: } k(x, x') = \sigma^2\left(1 + \frac{-(x-x')^2}{2\alpha l^2}\right)^{-\alpha}.$$

Table of Contents

Data: Ozone Concentration and CO2 Data

- Ozone data from the U.S. Environmental Protection Agency (EPA)[1]
- CO2 emissions data from the Vulcan High-Resolution Hourly Fossil Fuel data set repository provided by the Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC). [4]
- CO2 emissions are estimated for the following sectors: residential, commercial, industrial, electricity production, onroad, nonroad, commercial marine vessel, airport, rail, and cement.

Ozone Data

- Ozone data available hourly. Specific file used: "hourly_44201_2014.zip"
- Measured as a concentration (parts per million in the atmosphere)
- 14 sites had ozone concentration data for the time frame. Lancaster and Santa Clarita locations were excluded from the analysis.

EPA Ozone Sensor Locations

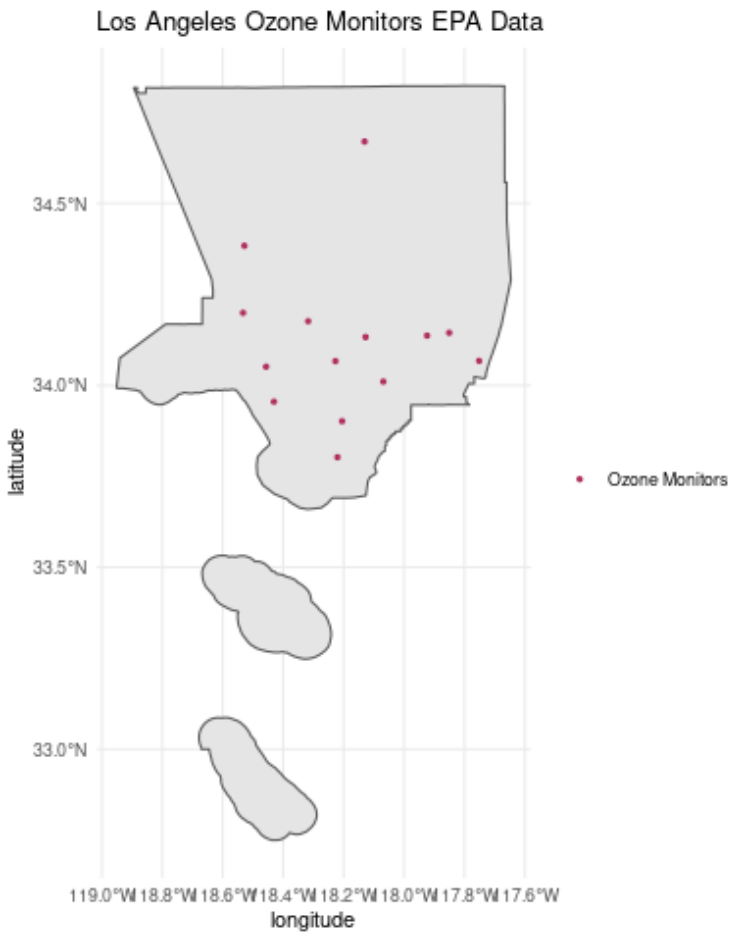

Figure: Ozone Monitors from EPA data, reporting ozone concentrations between PM on March 9th, 2014 until 3 PM on March 17th, 2014. L.A. County Site Numbers 9033 and 6012 are the two most northern points on the map.

Vulcan CO2 Data

- Provided hourly (each day and sector in its own file)for the continental U.S.
- .nc4 File type format
- Large amount of data (roughly 30 GB of raw data to get one weeks worth of processed data)

Limiting the Scope of the Data

- Data limited to 4PM on March 9th, 2014 until 3 PM on March 17th, 2014
- Limited to Los Angeles County area. Specifically limited to the square-bounds of the 12 ozone sensors.

Data Processing

R-Markdown File

EDA: Time Ozone Time Series

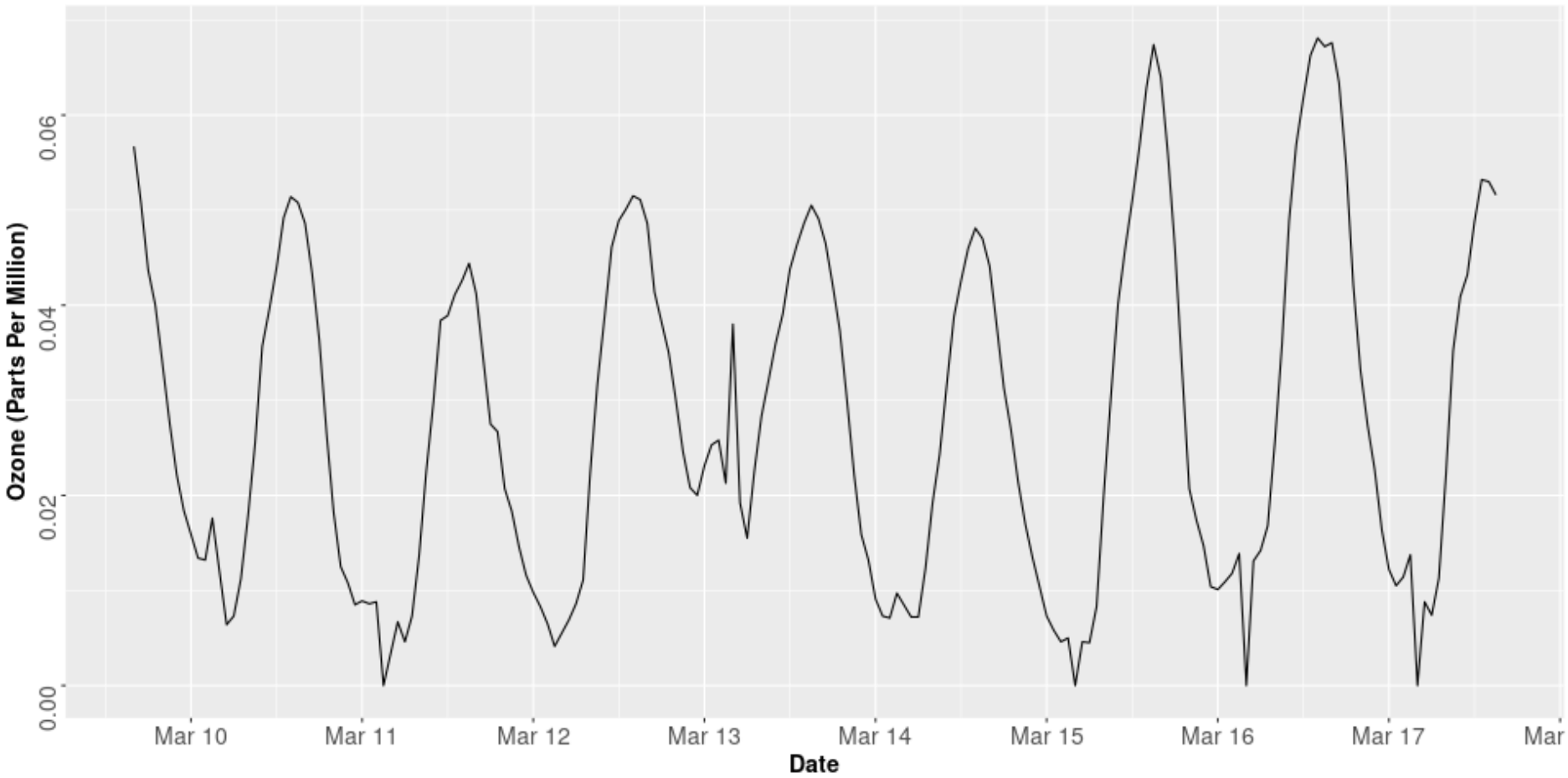

Figure: Average Ozone Concentration (Parts Per Million) Across Los Angeles County Sensors. Hourly from 4 PM May 9th, 2014 until 3 PM May 16th, 2014 (PST). Ozone sensor data excludes L.A. County Site Numbers 9033 and 6012 for the time period.

EDA: CO2 Time Series

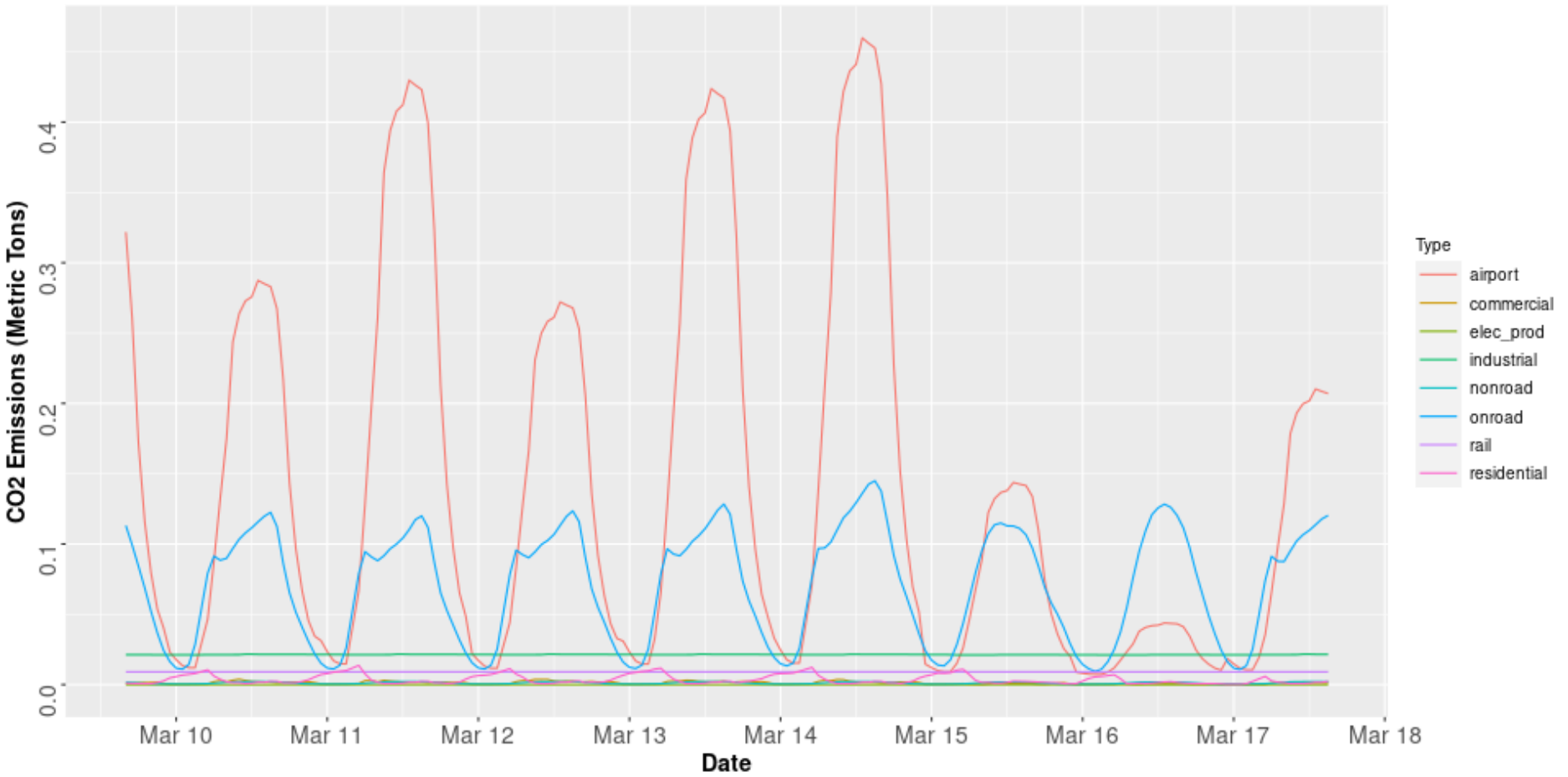

Figure: Average Metric Tons of CO2 Across Los Angeles County Sensors. Hourly from 4 PM May 9th, 2014 until 3 PM May 16th, 2014 (PST). Sensors Limited to bounds of Ozone Sensors, see Data Processing Section.

EDA: Scaled CO2 Time Series

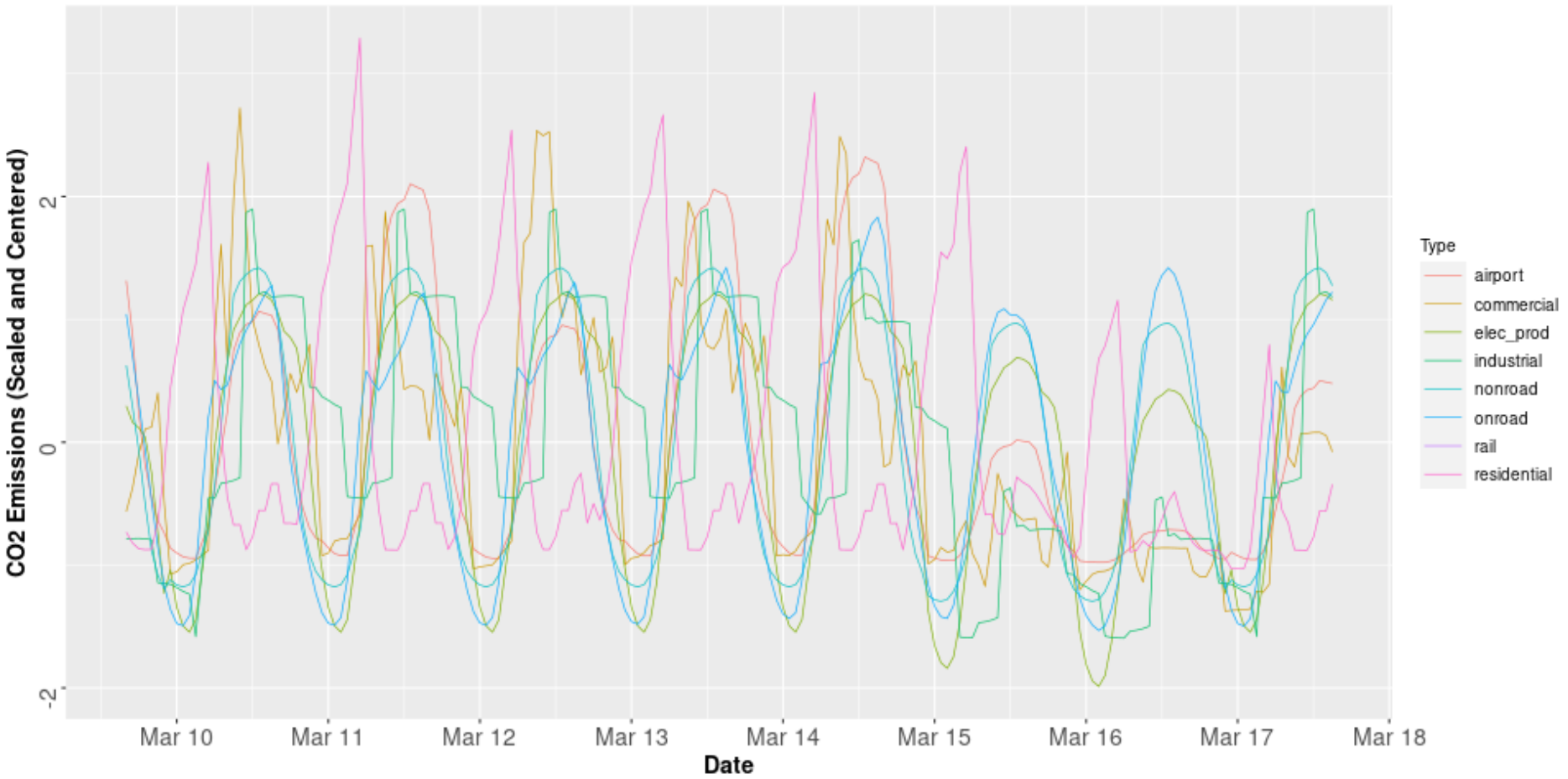

Figure: Average CO2 Scaled and Centered Across Los Angeles County Sensors. Hourly from 4 PM May 9th, 2014 until 3 PM May 16th, 2014 (PST). Rail emissions are not shown in this graph. Sensors Limited to bounds of Ozone Sensors, see Data Processing Section.

EDA: Ozone ACF, PACF

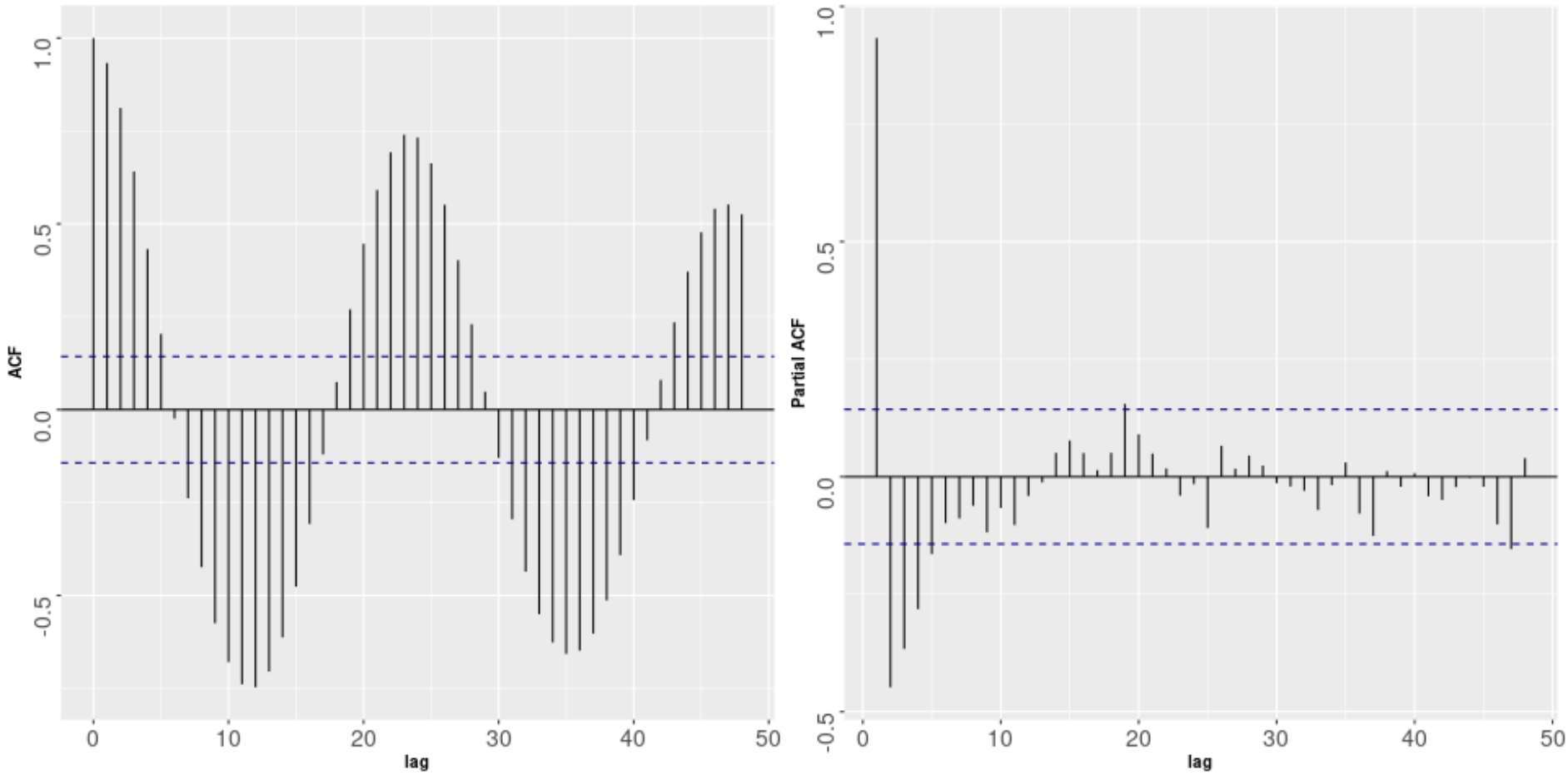

Figure: Autocorrelation function and partial autocorrelation function for ozone data over the time period, up to a lag of 48 hours. The blue lines indicate 95% confidence intervals for the ACF values.

EDA: Airport CO2 ACF, PACF

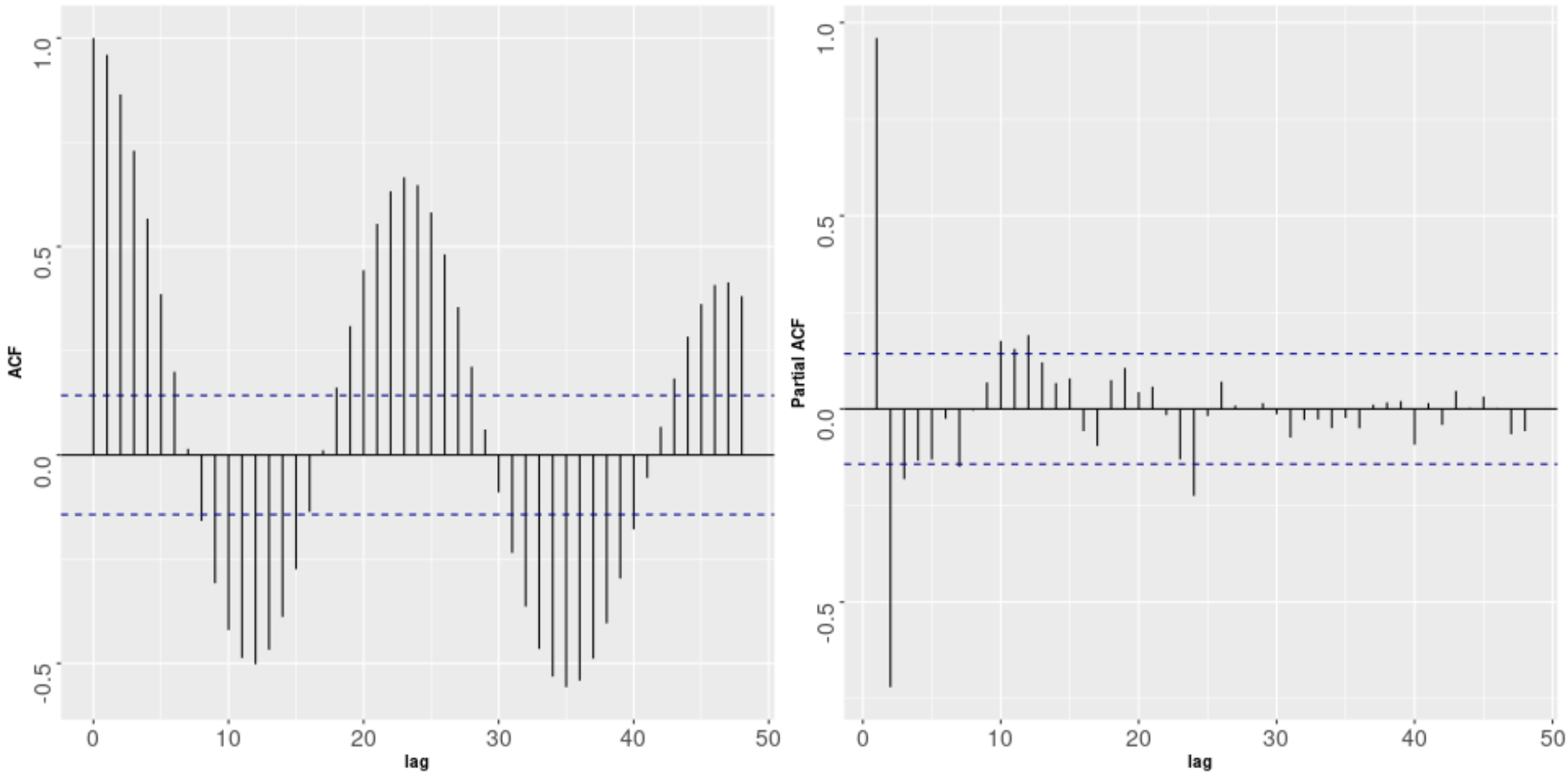

Figure: Autocorrelation function and partial autocorrelation function for Airport CO2 emissions data over the time period, up to a lag of 48 hours.

EDA: Cross-Correlation Functions

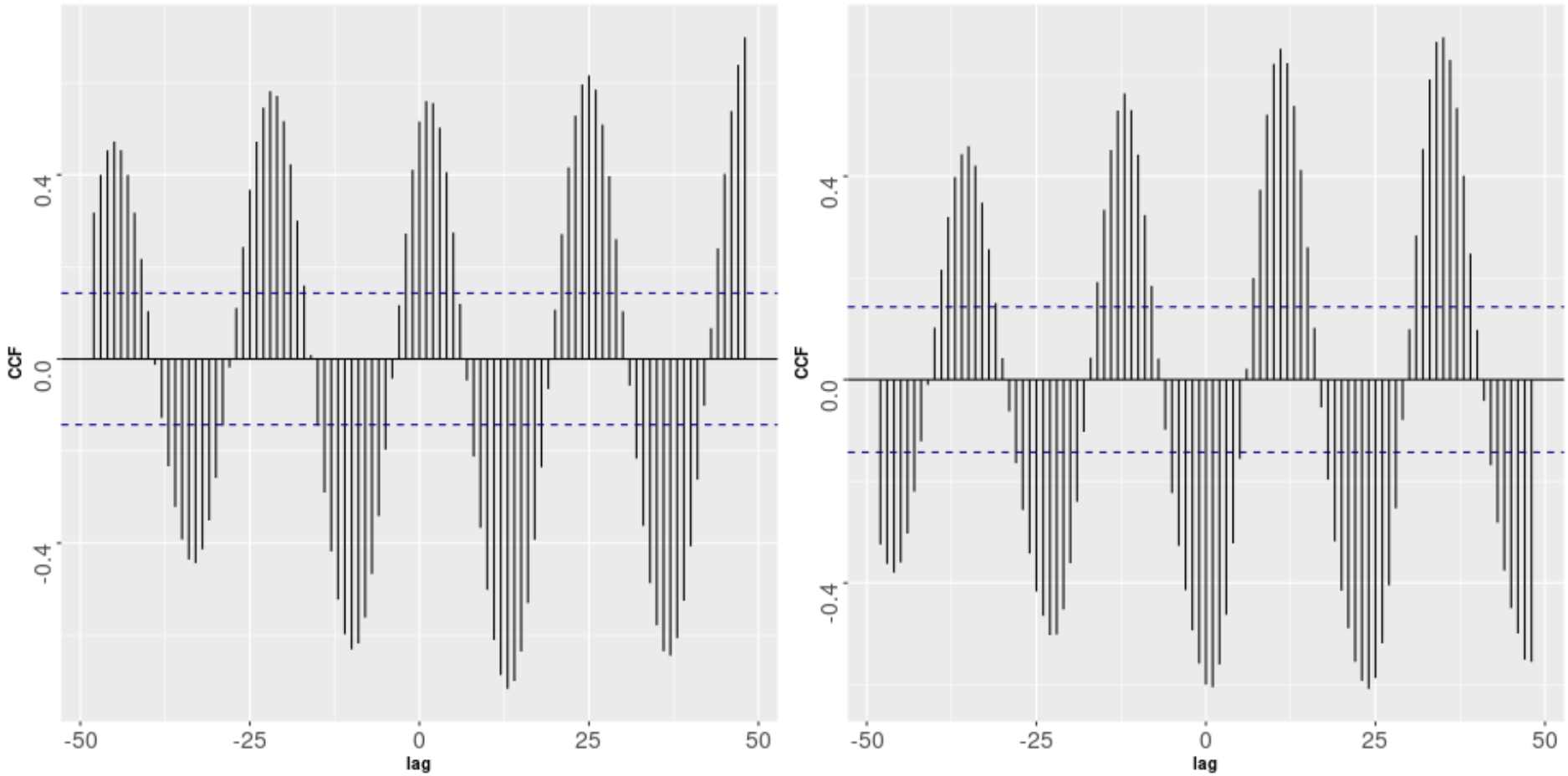

Figure: Cross-Correlation function functions Ozone vs. Airport CO2 (left) and Ozone vs. Residential CO2 (Right) over the time period, up to a lag of 48 hours.

EDA: Residential CO2 ACF, PACF

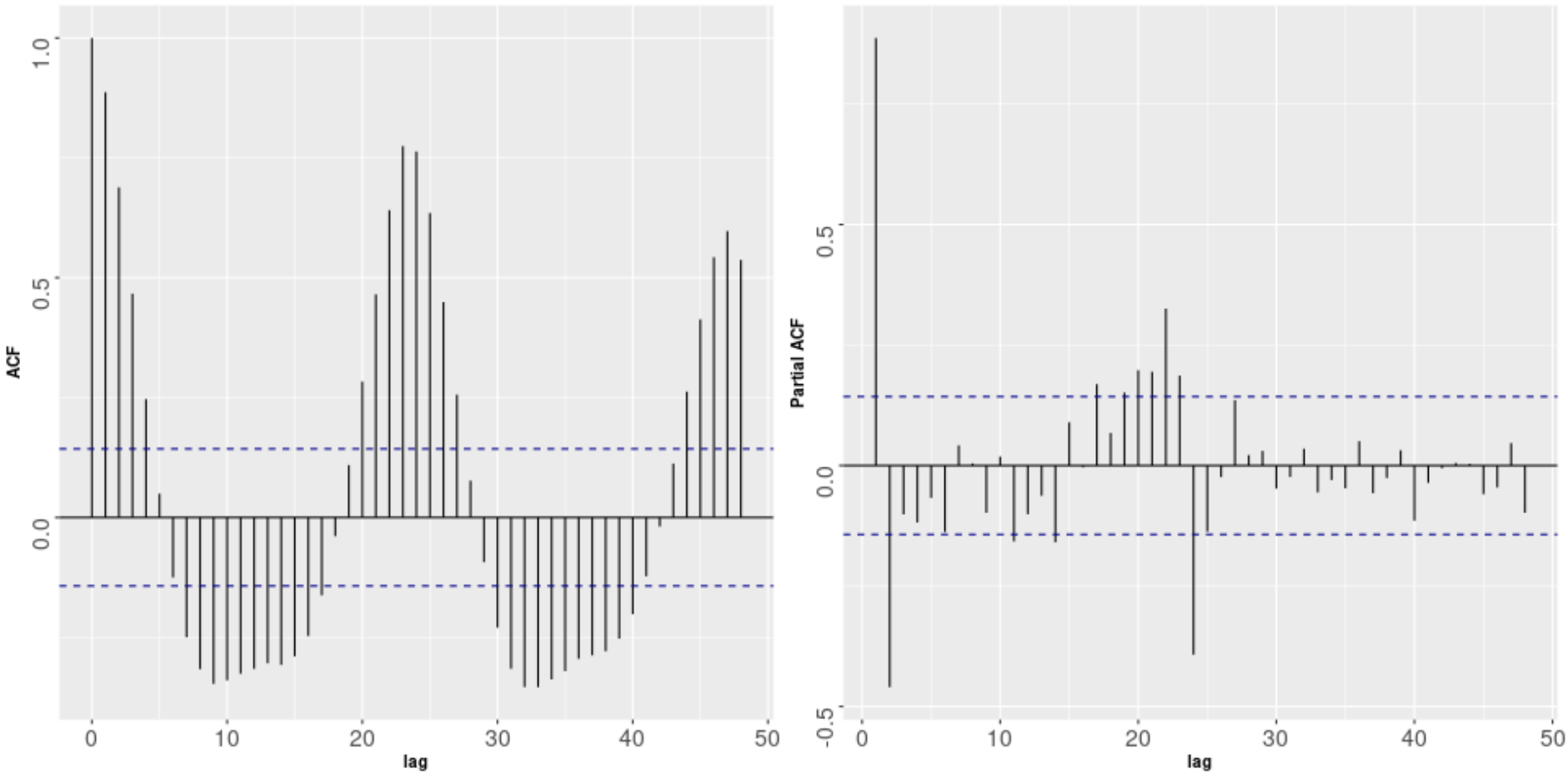

Figure: Autocorrelation function and partial autocorrelation function for Residential CO2 emissions data over the time period, up to a lag of 48 hours.

Table of Contents

PCA: Computation Time

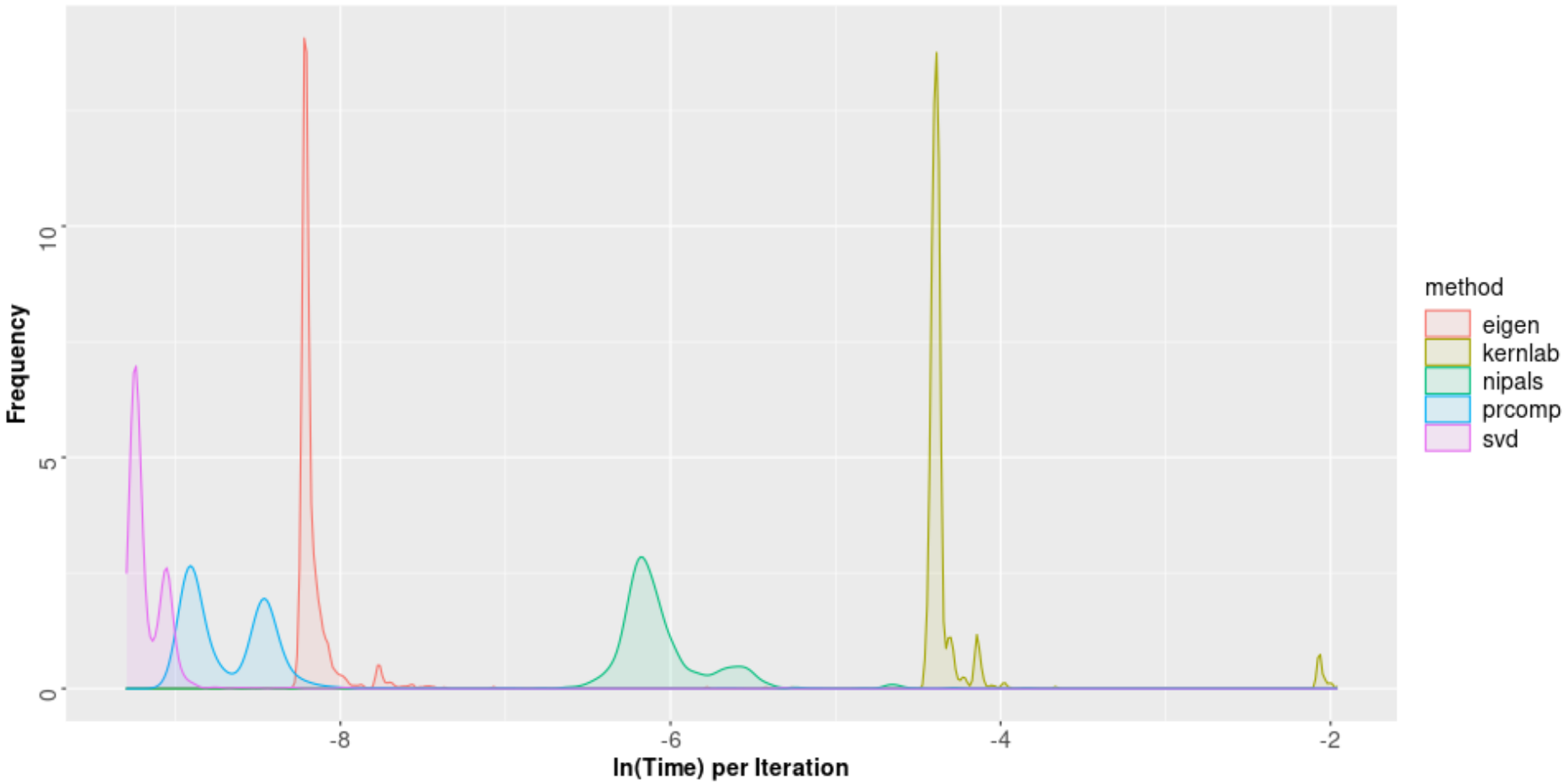

Figure: Log of time reported for visual clarity. 1000 iterations of performing PCA for five different methods. The data $X_{188\times 7}$ is the matrix of centered and scaled CO2 data.

PCA: Scree Plot

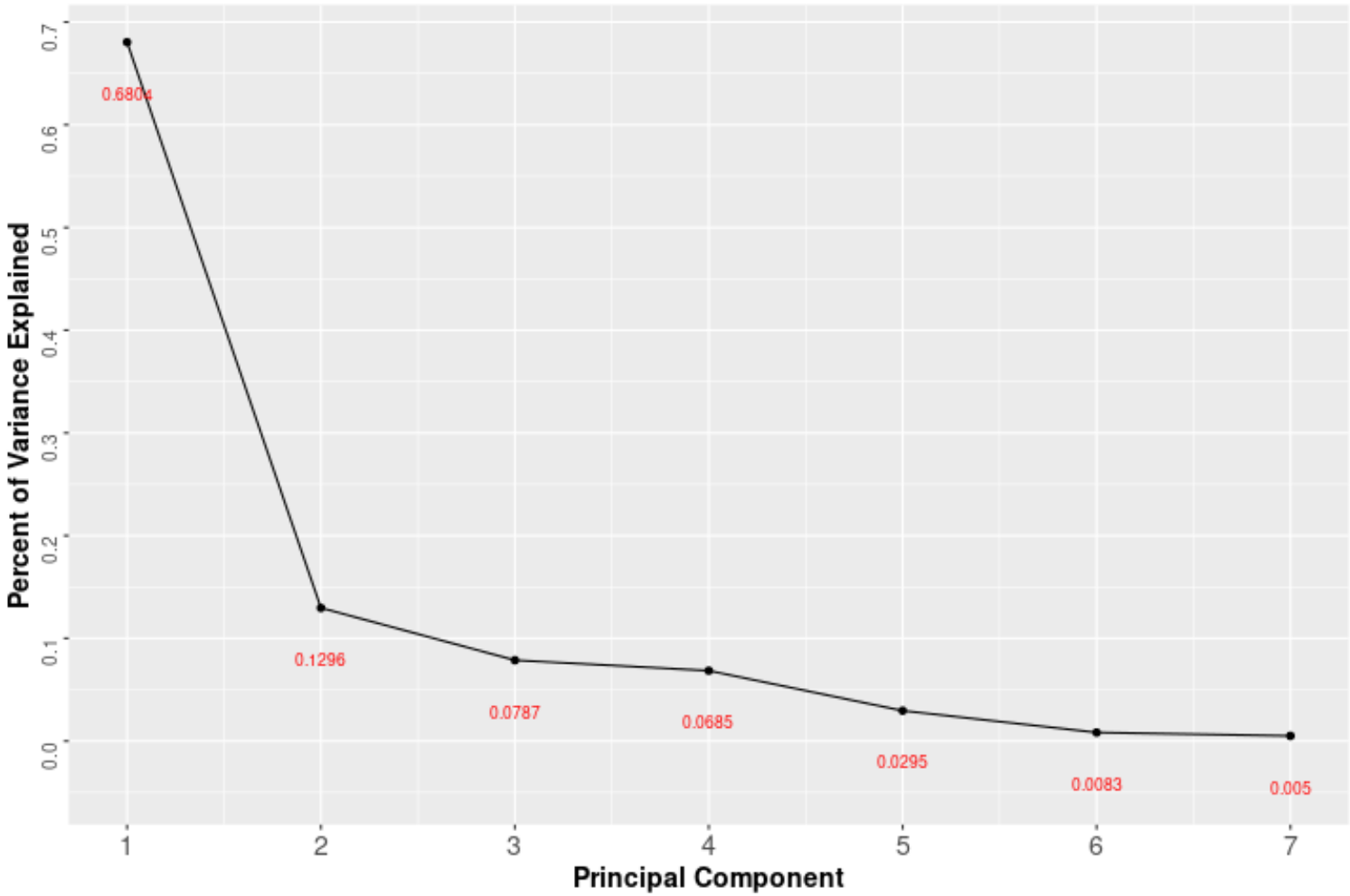

Figure: The percent of variation explained by each principal component for the CO2 data.

PCA: Bi-Plot PC1 and PC2

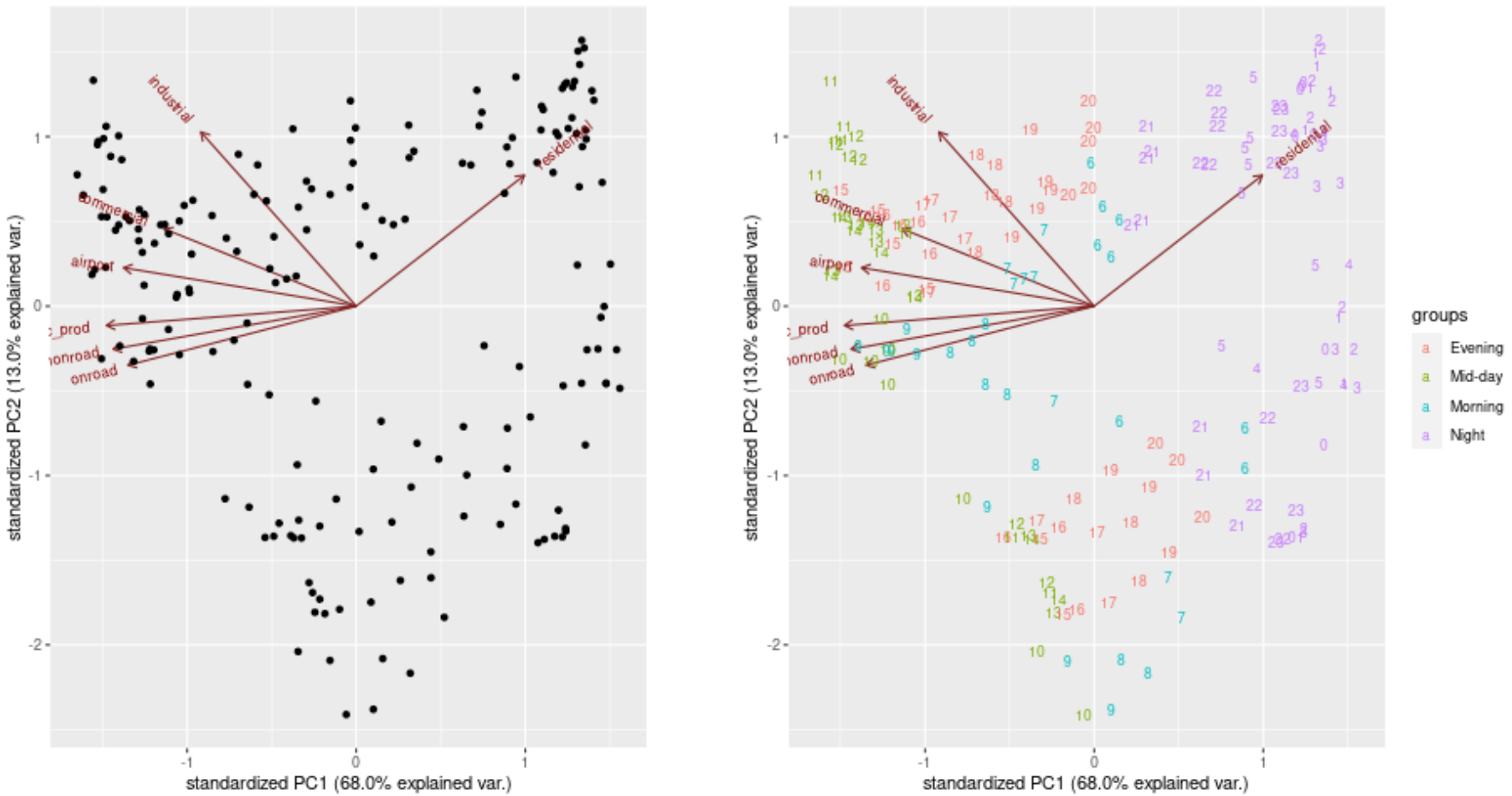

Figure: First and Second Principal Components Compared.

PCA: Potential Data Groupings

- Hours 21, 22, 23, 0, 1, 2, 3, 4, 5 grouped as "Night"
- Hours 15, 16, 17, 18, 19, 20 grouped as "Evening"
- Hours 6, 7, 8, 9 grouped as "Morning"
- Hours 10, 11, 12, 13, 14 grouped as "Mid-day"

PCA: Bi-Plot PC1 and PC2, Detailed

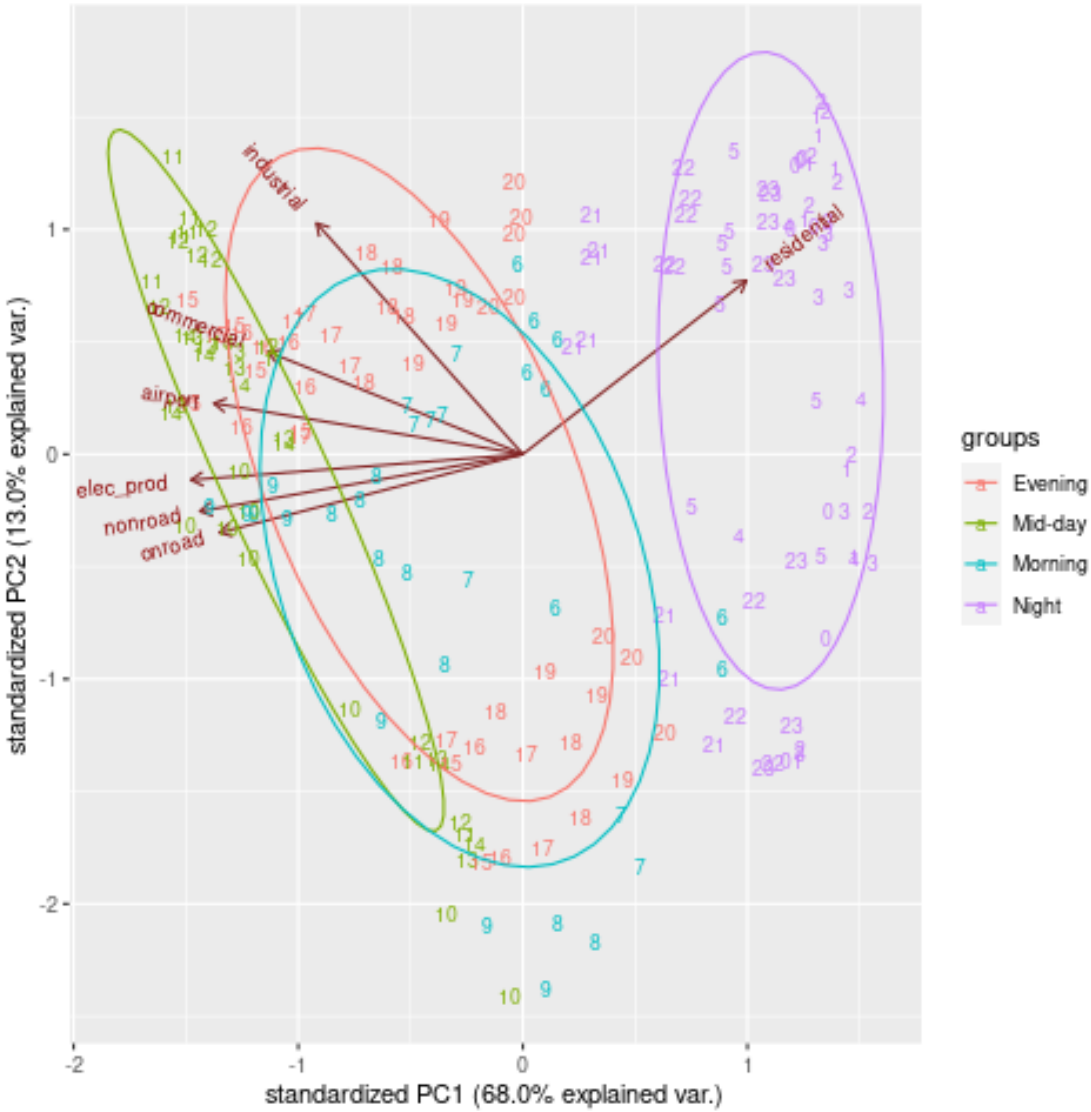

Figure: First and Second Principal Components Compared, with normal ellipse

PCA: Bi-Plot PC3 and PC4, Detailed

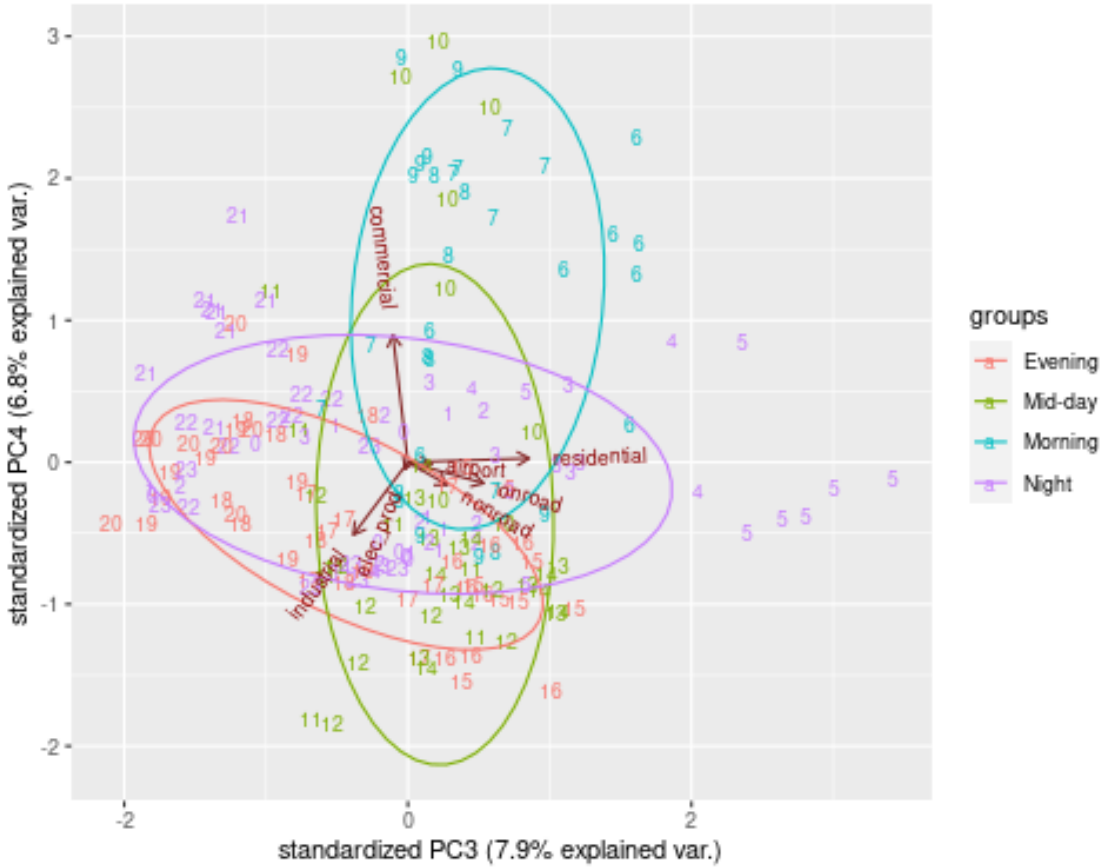

Figure: Third and fourth Principal Components Compared, with normal ellipses

PCA: Regression

- A simple linear model is not feasible in this setting due to high multicollinearity between the CO2 variables (covariates).
- Instead, use the principal components to model Ozone and CO2 from different emissions sources.
- Not all of the principal components are necessary to get (effectively) the same model. With only seven PC's, it is easy to see that the model does not benefit from having more than the first four PC's as covariates.

PCR Regression

PC	Coefficient	Standard Error	P-Value
PC1	-0.0055698	0.0010155	$1.35x10^{-}7$
PC2	-0.0057490	0.0023270	0.01440
PC3	0.0001153	0.0029865	0.96924
PC4	-0.0097947	0.0032011	0.00255

Table: Principal Component Regression between Ozone and the Principal Components of CO2 Data. 184 df, Adjusted $R^2 = .181$

KPCA Methods

- Practically for KPCA, the hyperparameters need to be selected to represent a feature map that captures the variation of the data in a useful manner
- In this section the RBF kernel and the Polynomial kernel will be considered for hyperparameter selection

KPCA: RBF, l = .01

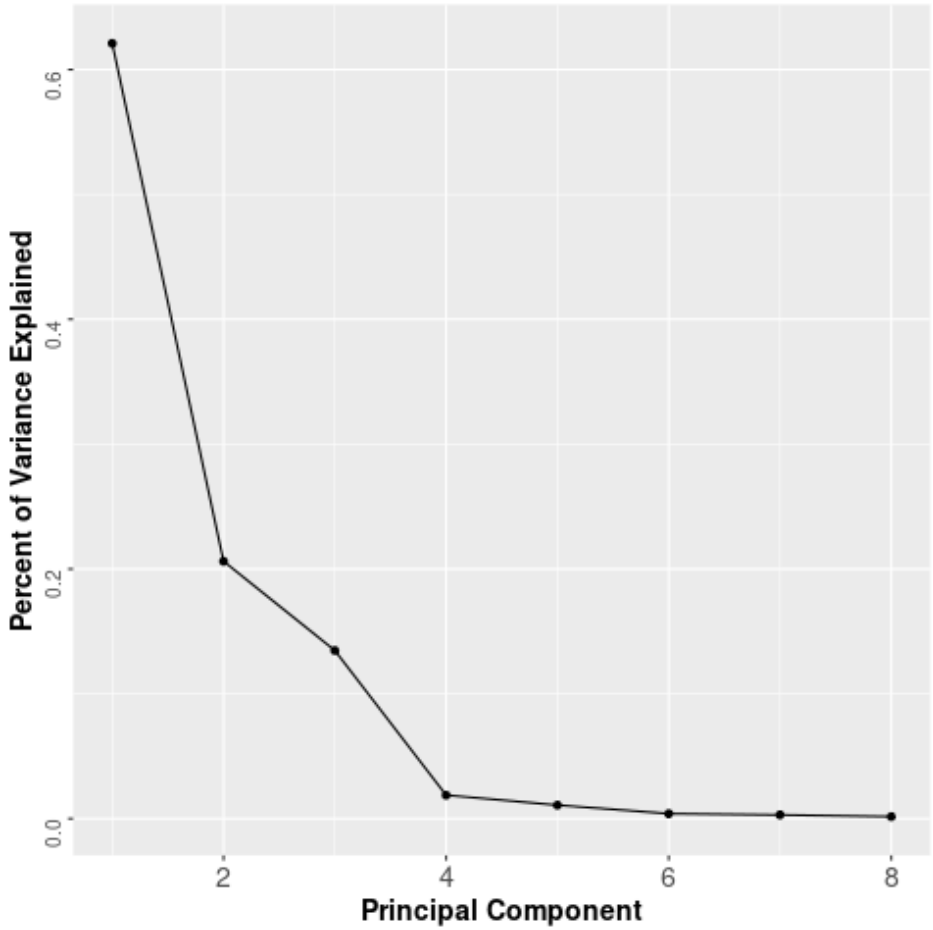

Figure: l = inverse parameter. kernlab package used.

KPCA: RBF, l = 1

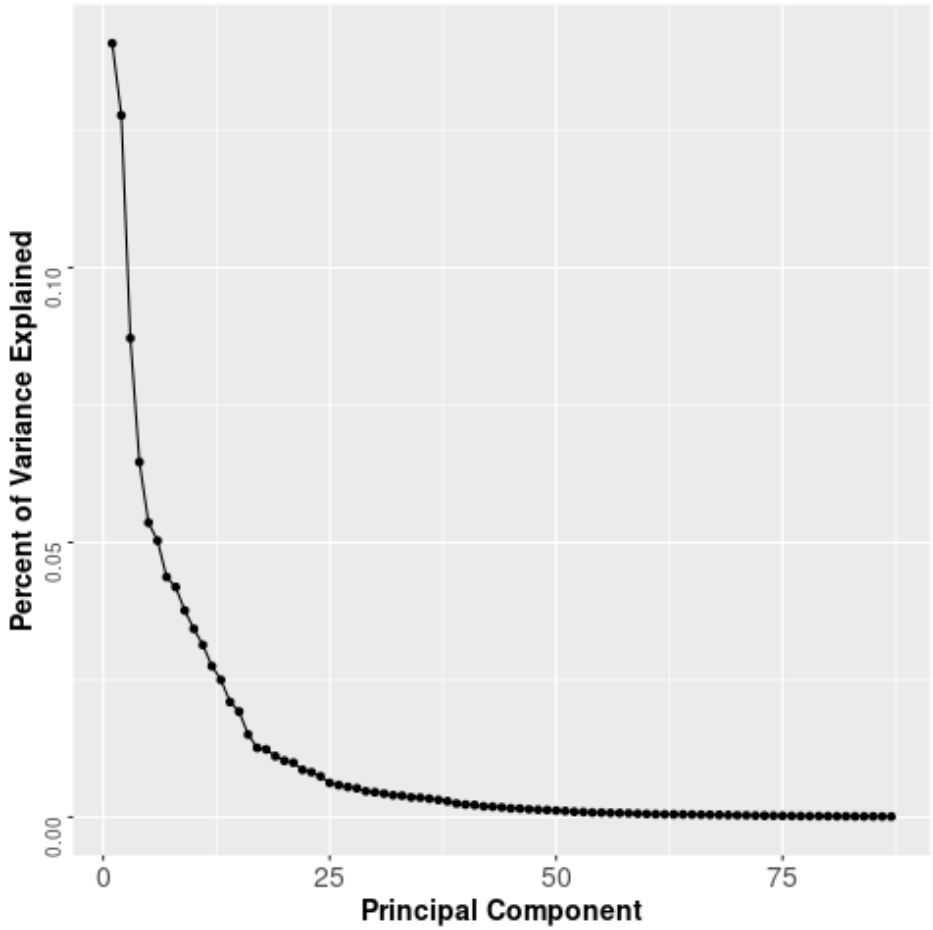

Figure: l = inverse parameter. kernlab package used.

KPCA: RBF, l = 2

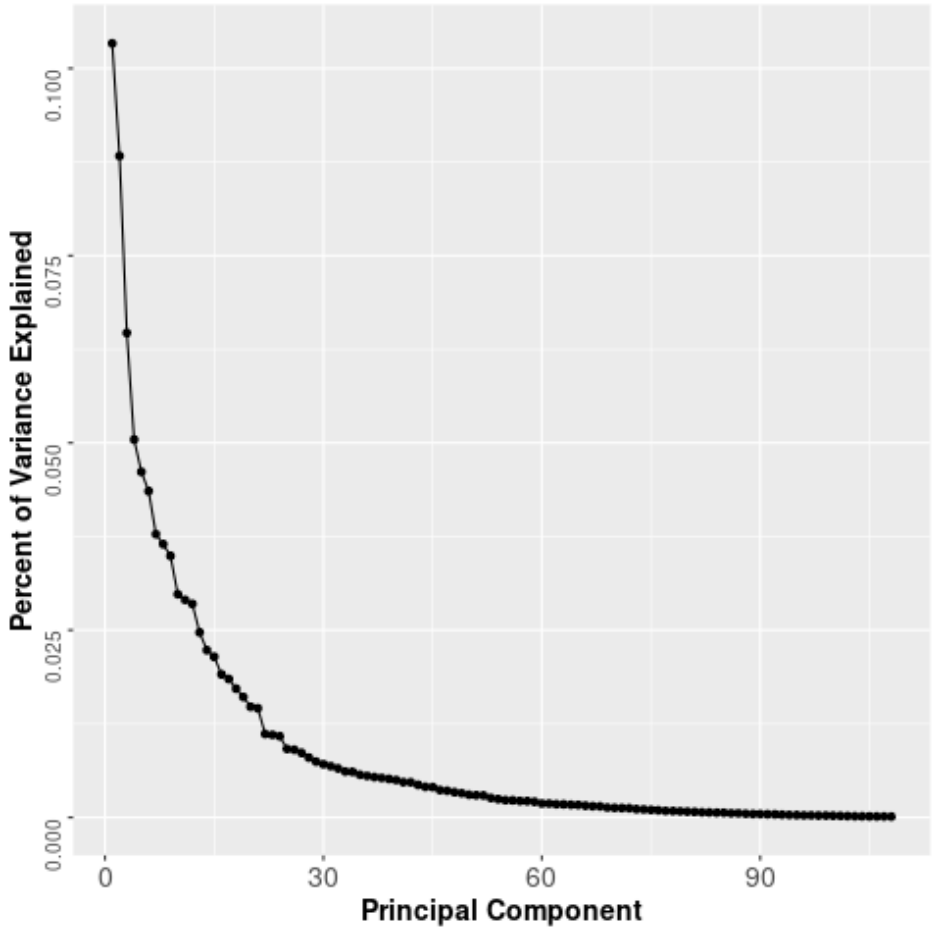

Figure: l = inverse parameter. kernlab package used.

KPCA: RBF, # PCs Required to Explain 75% Variation

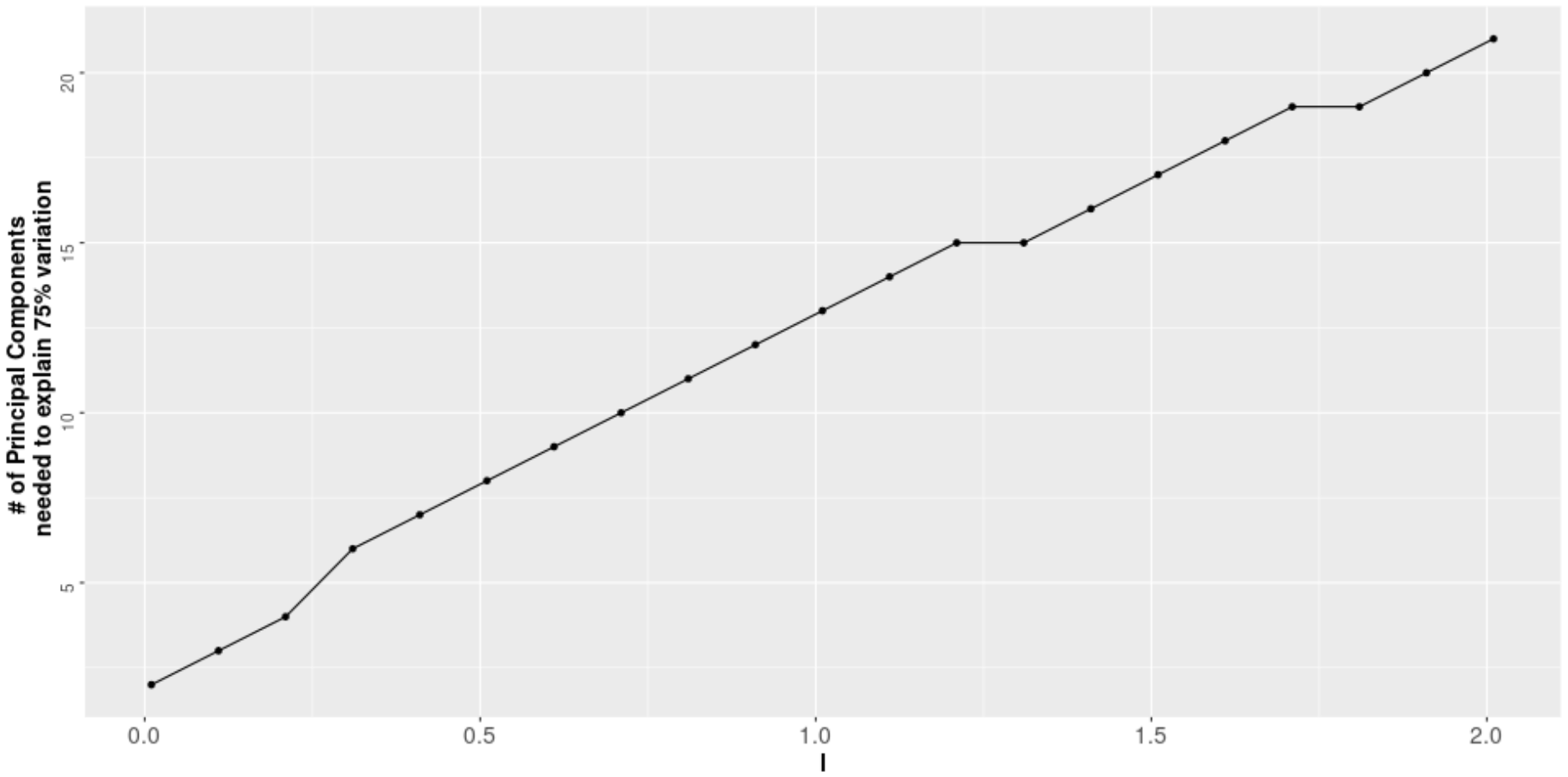

Figure: l = inverse parameter. kernlab package used.

KPCA: Polynomial, d = 1

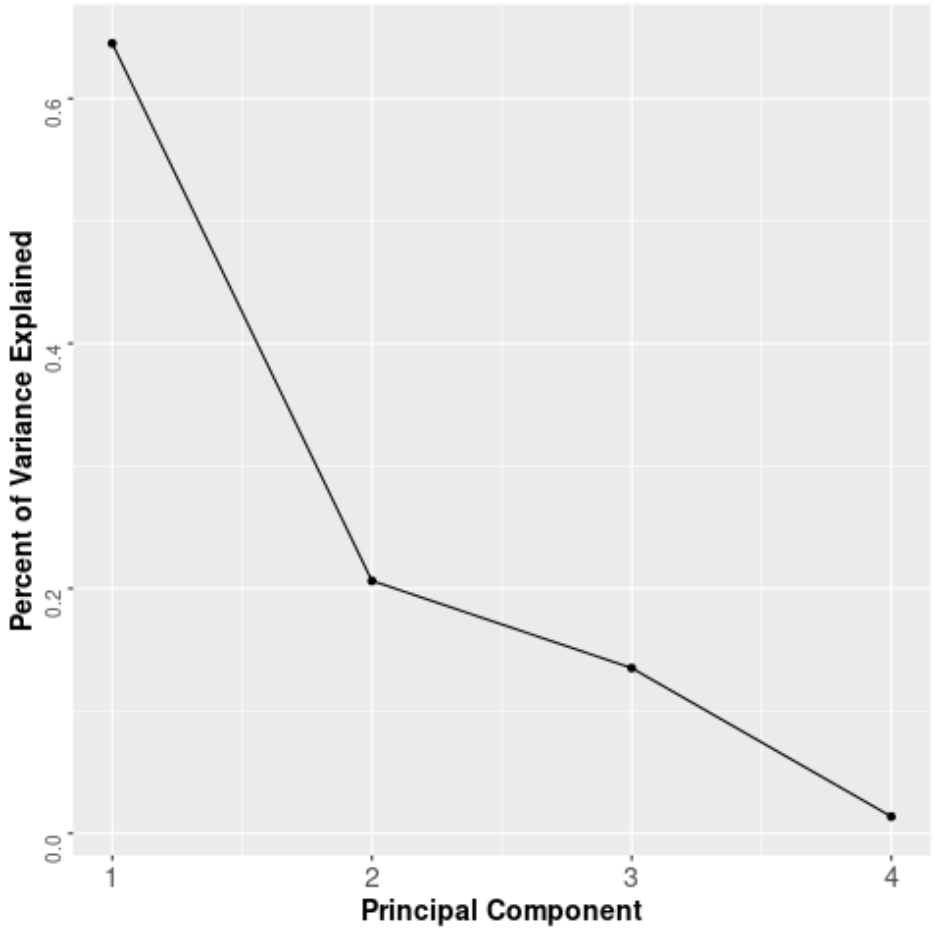

Figure: d = degree parameter. kernlab package used.

KPCA: Polynomial, d = 2

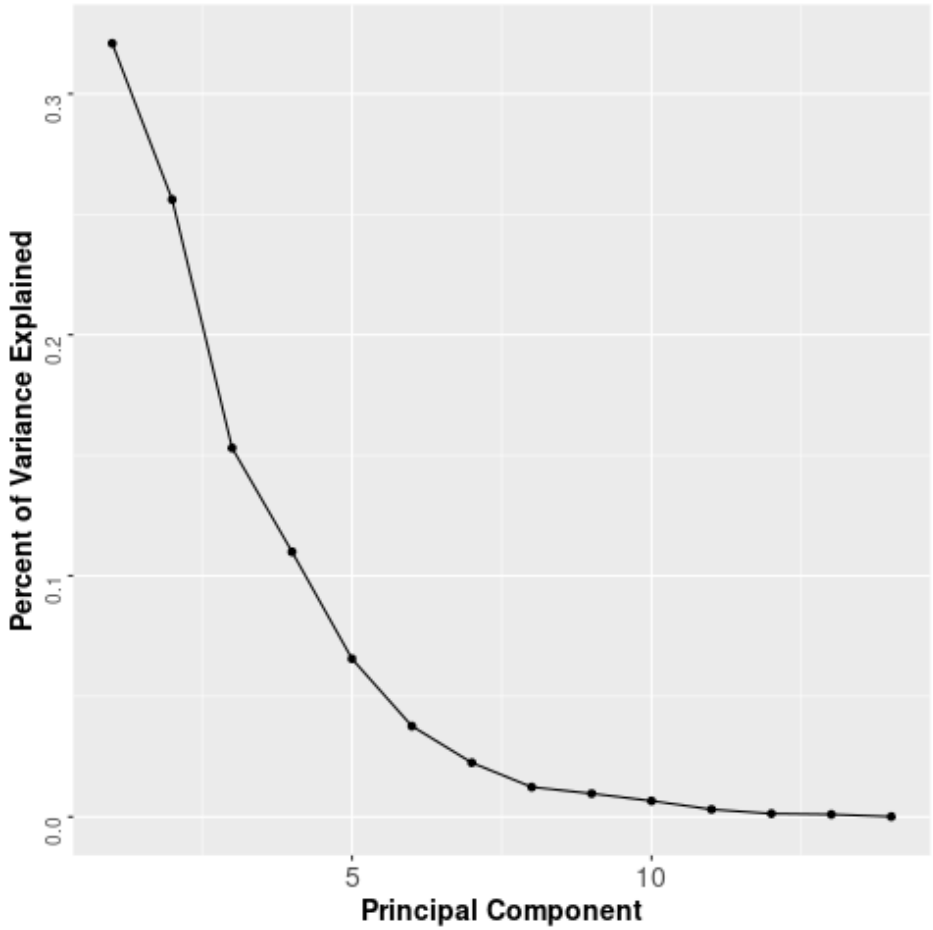

Figure: d = degree parameter. kernlab package used.

KPCA: Polynomial, d = 10

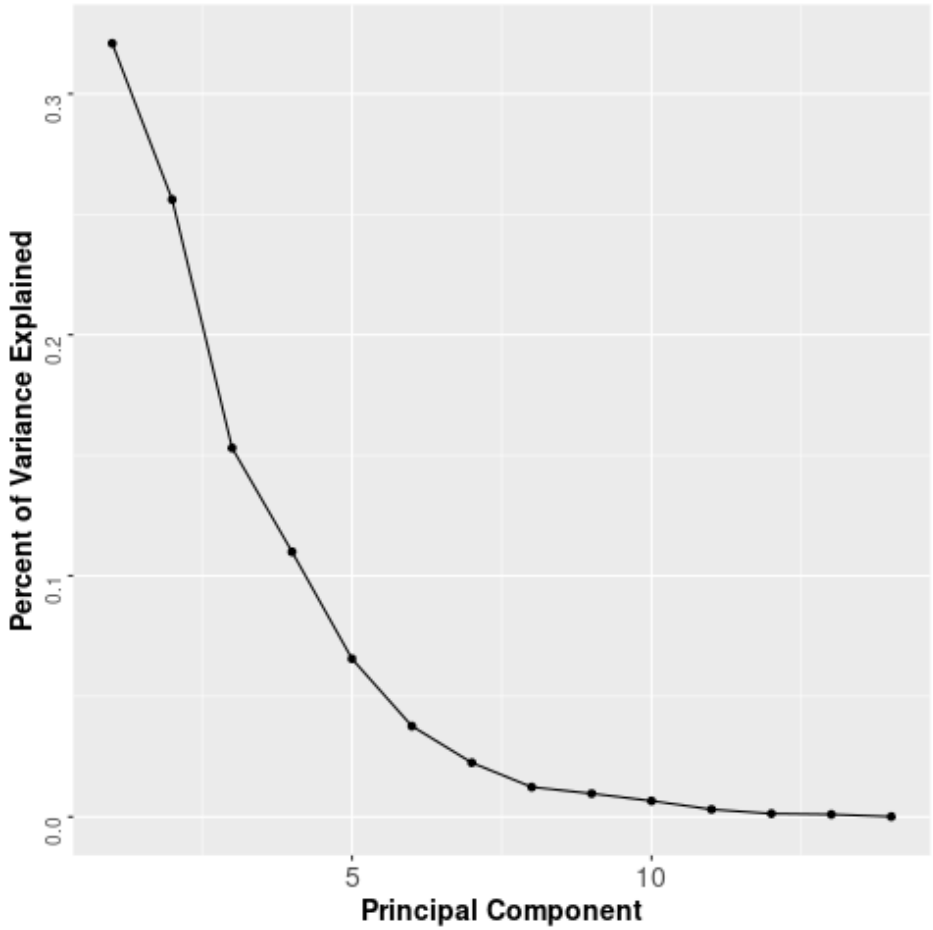

Figure: l = degree parameter. kernlab package used.

KPCA: Poly, # PCs Required to Explain 75% Variation

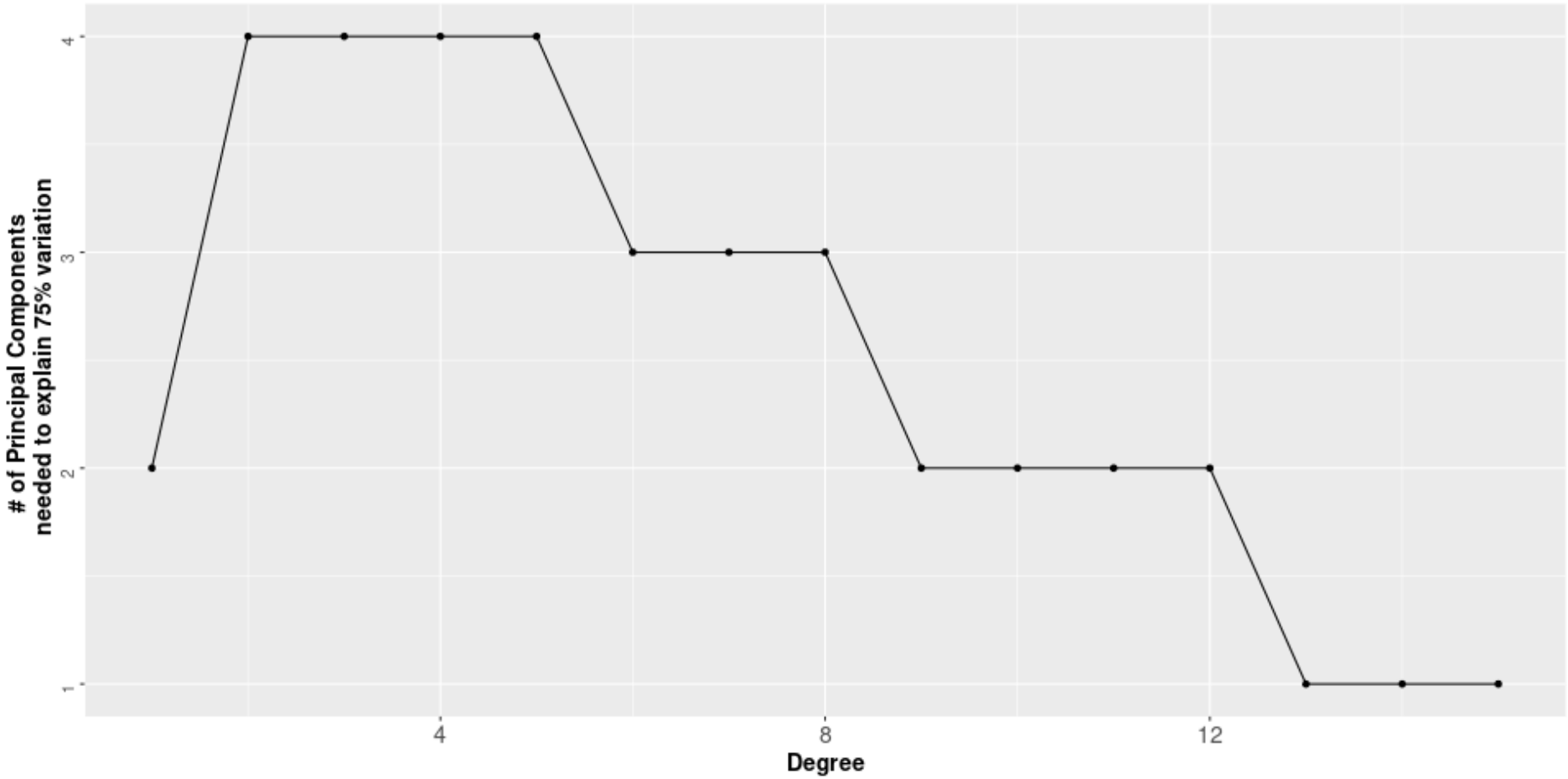

Figure: d = degree parameter. kernlab package used.

Table of Contents

Summary of Results

- EDA and PCA indicated a clear relationship between Ozone and CO2 (but not causal!)
- PCA proved useful in explaining the sources of variation, and PCR a viable method for this data
- Choice of hyperparameters has significant impact on KPCA results

Future Work

- Issues implementing user defined kernel function
- Kernel Principal Component Analysis Regression
- Ideally would download a bit more data for a test set to get test error for regression, in an effort to compare methods for statistical learning purposes.

Acknowledgements

Thank you to Professor Gareth Peters who taught the class on this material, held multiple meetings advising me on the project, and suggested the Vulcan CO2 Emissions data set for use in this report. Thank you to Laurel Abowd who helped me with the geospatial processing of the Ozone and CO2 data. In particular, she advised me how to process the .nc4 formatted data files and how work with the Lambert Conformal Conic projection in R.

Table of Contents

Statistical Code and Software

All of the statistical programming in this report is completed in R. R version 4.2.0 was used at the time this report was written [[3]]. The R-Markdown file related to this report can be found on the author's Github page (see Introduction). The following packages were implemented in addition to the functions available in 'base' R: ncdf4, stringr, ggplot2, dplyr, tidyr, lubridate, tmap, sf, rematch, pls, kernlab, cowplot, knitr, ggbiplot (see report paper for citations for these packages).

Table of Contents

Bibliography I

[1] EPA Air Data. *Air Data: Pre-Generated Data Files*. Available at: https://aqs.epa.gov/aqsweb/airdata/download_files.html#Raw. File in Question referred to as "hourly_44201_2014.zip". Nov. 2021.

[2] EPA. *Health Effects of Ozone Pollution*. Tech. rep. Available at: https://www.epa.gov/ground-level-ozone-pollution/health-effects-ozone-pollution. U.S. Environmental Protection Agency, 2021.

[3] The R Foundation. *The R Project for Statistical Computing*. Version 4.2.0: Vigorous Calisthenics. 2022. URL: https://www.r-project.org/.

[4] K.R. Gurney et al. *Vulcan: High-Resolution Hourly Fossil Fuel CO2 Emissions in USA, 2010-2015, Version 3*. Available at: https://daac.ornl.gov/cgi-bin/dsviewer.pl?ds_id=1810. Oct. 2020.

Bibliography II

[5] Mark Jacobson. "Enhancement of Local Air Pollution by Urban CO 2 Domes".
In: *Environmental Science & Technology* 44.7 (Mar. 2010). Available at https://web.stanford.edu/group/efmh/jacobson/Articles/V/es903018m.pp. 2497–2502.

[6] Charlotte Munson. *You Asked: How Does Carbon Dioxide Get So High Up Into the Atmosphere?* Tech. rep. Available at https://news.climate.columbia.edu/2020/09/23/carbon-dioxide-distribution-atmosphere/. Columbia Climate School, Sept. 2020.

[7] Gareth Peters. *PSTAT262FE Kernel Machines and Feature Extraction Methods*. Specific version provided as part of PSTAT 262 FE course. A 2017 version of these notes can be accessed here https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3050592. Apr. 2022.