IE 582 - STATISTICAL LEARNING FOR DATA MINING

Homework 1

15.11.2024

Ahmet Çeliker - 2020402126

This report presents my analysis of the relationships between antenna design parameters and their electromagnetic performance (S11 response) using dimensionality reduction and regression modeling techniques. The main objective of my work was to simplify the complexity of the design space through dimensionality reduction and develop predictive models to estimate the S11 response at specific frequency points.

1. Data and Analysis Process

1.1. Structure and Features of the Data

I used three main datasets for this analysis:

- **Input Data (input_data.csv)**: Contained geometric parameters of the antenna design.
- **Real Part Data (real_data.csv)**: Represented the real part of the S11 response across 201 frequency points.
- **Imaginary Part Data (img_data.csv)**: Represented the imaginary part of the S11 response across 201 frequency points.

The **input data** included parameters such as patch dimensions, substrate height, solder resist layer thickness, and dielectric constants, which are significant factors affecting the antenna's resonant behavior.

1.2. Data Preparation and Preprocessing

All design parameters are expressed in the same unit of measurement (mm) and cover similar ranges. Since all variables are in the same unit and lie within similar ranges, there are no scale differences that could cause one variable to become dominant over the others. When the data is within the same scale range, modeling algorithms can use this data directly without creating an imbalance in the importance of the variables. Consequently, the results of the model can be interpreted correctly.

As a result, there was no need for standardization in the data preprocessing step. I also checked for any missing values, and all data were complete and without any missing entries.

After confirming that all data fall within the given ranges, I proceeded to the next step.

1.3 Correlation Analysis

I calculated and visualized the correlation matrix of the design parameters to understand their relationships better. This process aimed to highlight which parameter pairs were most closely related, providing insight for feature selection and potential multicollinearity in regression modeling.

The pair with the highest correlation is **height of substrate and width of patch** with a correlation coefficient of **0.9237**. This extremely strong positive correlation indicates that these two parameters are highly related; when one increases, the other tends to increase proportionally, and vice versa. This could lead to potential multicollinearity issues if both variables are used simultaneously in a regression model, potentially inflating variance and affecting the model's interpretability and accuracy.

I took one of these two variables while setting up a linear regression to solve this problem. Additionally, when I used PCA, I had already managed redundancy, so there was no need for any extra processing in the case of PCA.

2. Dimensionality Reduction: Principal Component Analysis (PCA)

2.1. Implementation and Findings from PCA

To explore the underlying structure of the design parameters and reduce the dimensionality of the dataset, I conducted Principal Component Analysis (PCA) using two different methods for comparison.

First, I performed PCA using the prcomp() function. This function utilizes **Singular Value Decomposition (SVD)** to compute the principal components. I chose this method because SVD is known for its numerical stability, particularly when dealing with high-dimensional data or datasets with potential multicollinearity.

Next, for comparison, I also applied PCA using the princomp() function. The princomp() function uses **eigen decomposition** of the covariance or correlation matrix. This method, while useful for understanding the eigen decomposition of the dataset, can be less numerically stable compared to prcomp(), particularly in cases where the data is high-dimensional or highly correlated.

Both functions output the same standard deviations, proportion of variance, and cumulative proportion for each component. This shows that the overall variance explained by each principal component and their cumulative contributions are consistent across both methods.

For example, the first component explains approximately 20.7% of the total variance, and the cumulative variance explained by the first five components is about 60.6% in both cases.

At the same time, the eigenvalues were the same. I chose to use **princomp()**. This decision was guided by the fact that in our course, we specifically worked with the covariance matrix and used princomp() in class notes and exercises. Using princomp() aligns with the methodology emphasized in the course and ensures consistency with the approaches taught during our lessons.

2.2. Deciding Number of Principal Components

There was no specific decision metric, but typically, most data scientists were satisfied with at least 70% of the total variance explained or selected principal components with eigenvalues of at least 1 (Kaiser rule).

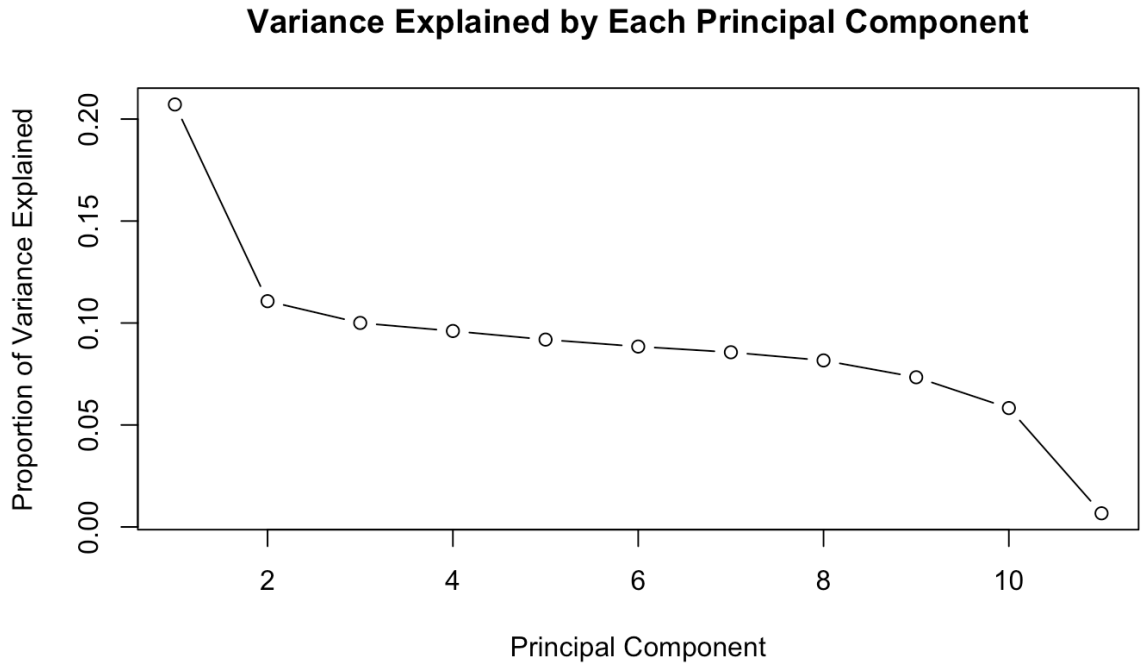
```
Importance of components:
                          Comp.1    Comp.2    Comp.3     Comp.4     Comp.5     Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
Standard deviation     1.5095325 1.1035356 1.0489863 1.02800587 1.00527544 0.98642834 0.97069535 0.94771083 0.89854601 0.80083052
Proportion of Variance 0.2071535 0.1107083 0.1000338 0.09607237 0.09187079 0.08845826 0.08565904 0.08165053 0.07339863 0.05830268
Cumulative Proportion  0.2071535 0.3178617 0.4178956 0.51396794 0.60583873 0.69429699 0.77995603 0.86160656 0.93500519 0.99330787
                          Comp.11
Standard deviation     0.271317885
Proportion of Variance 0.006692127
Cumulative Proportion  1.000000000
```

In our case, the first six principal components covered nearly 70% of the total variance.

Description: df [11 × 3]

	eigenvalue <dbl>	variance.percent <dbl>	cumulative.variance.percent <dbl>
Dim.1	2.27868828	20.7153480	20.71535
Dim.2	1.21779078	11.0708252	31.78617
Dim.3	1.10037220	10.0033837	41.78956
Dim.4	1.05679606	9.6072370	51.39679
Dim.5	1.01057871	9.1870791	60.58387
Dim.6	0.97304086	8.8458260	69.42970
Dim.7	0.94224946	8.5659042	77.99560
Dim.8	0.89815581	8.1650529	86.16066
Dim.9	0.80738493	7.3398630	93.50052
Dim.10	0.64132952	5.8302683	99.33079
Dim.11	0.07361339	0.6692127	100.00000

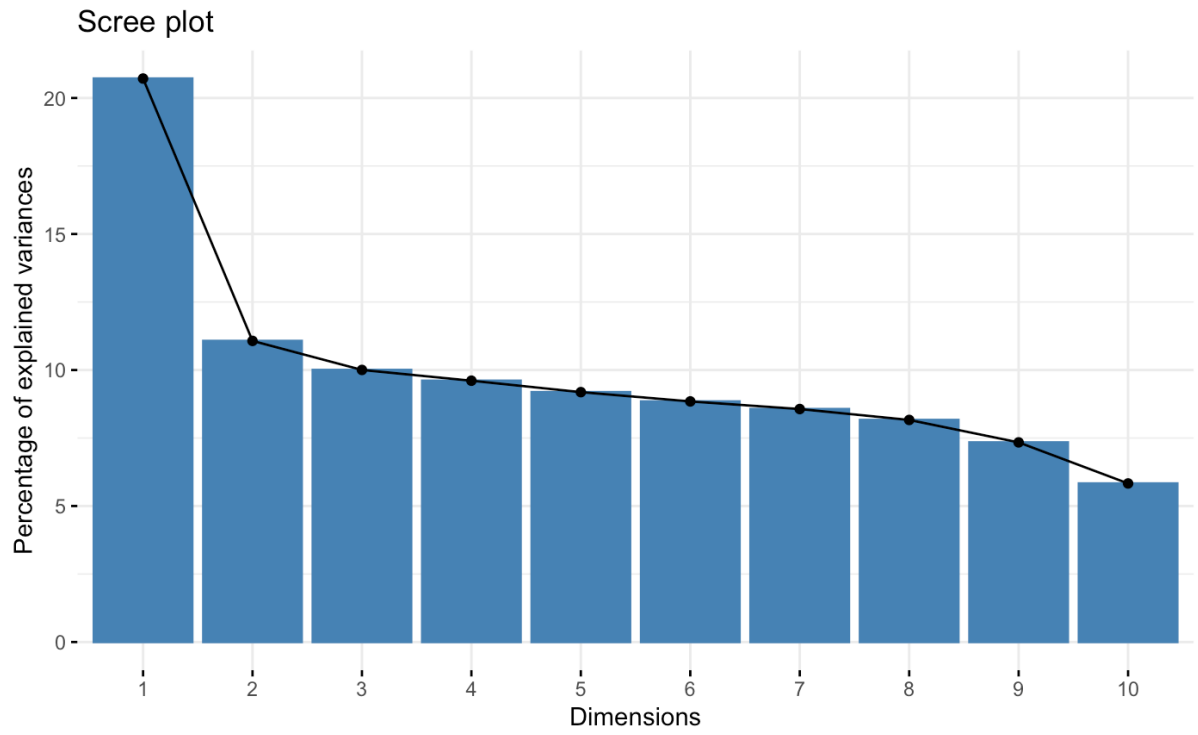
Additionally, I applied the "elbow method" to determine where the line stops curving.



The Y-axis represents the proportion of total variance explained by each principal component, while the X-axis shows the principal components (PC1 to PC11).

The first principal component (PC1) explains the largest portion of variance, approximately 20%. This indicates that PC1 captures the most significant variation in the dataset. There is a noticeable drop in the variance explained from PC1 to PC2, after which the rate of decrease becomes more gradual.By the time we reach the fifth or six component, the additional variance explained by each successive component is relatively small.

Notably, after the sixth component, the line straightened out and started to stabilize, indicating that we did not capture much additional information beyond this point.



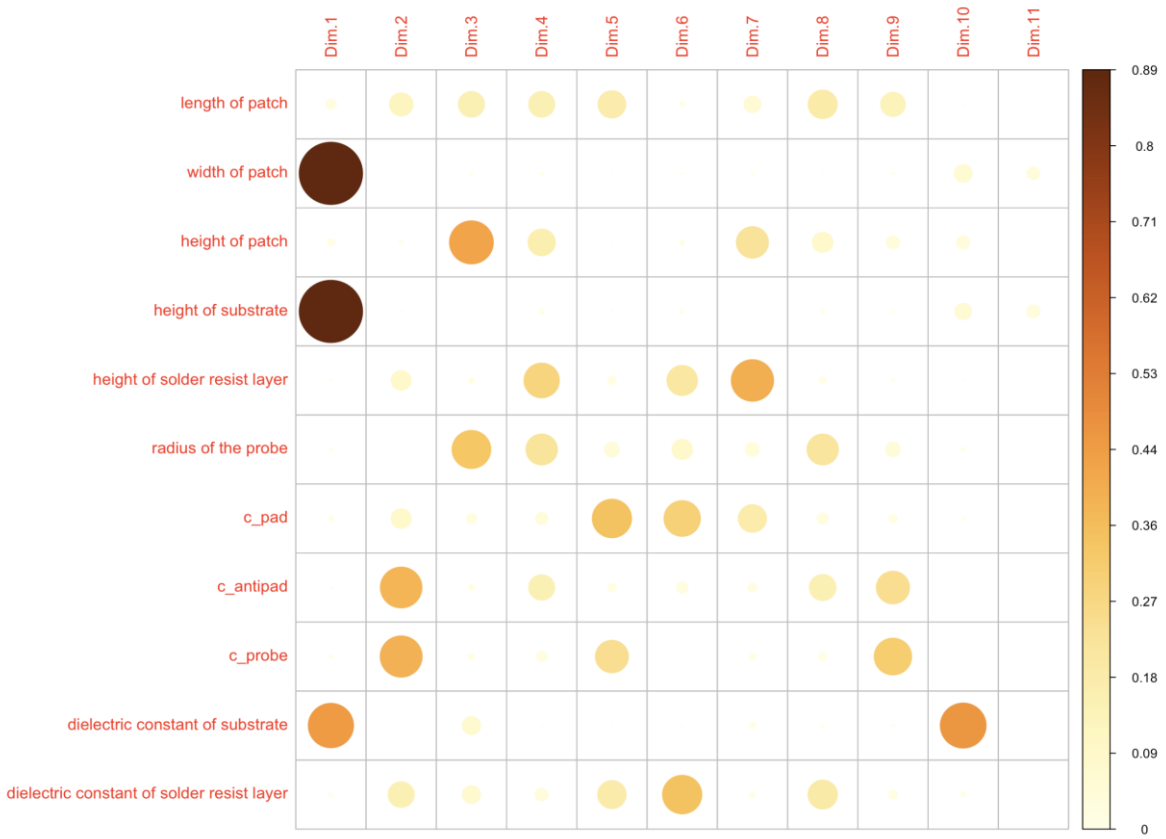
Based on the provided scree plot, it is evident that the first principal component (PC1) accounts for the highest proportion of variance, followed by the second principal component (PC2). This observation aligns with the measure of eigenvalues, where the initial PCs have larger values compared to subsequent PCs. This indicates that the first PCs correspond to the directions with the greatest variation within the dataset. Although there is no steep decline after the second principal component (PC2), selecting only the first two PCs would cover only about 32% of the total variance. Therefore, despite the lack of significant slope differences between PCs after the second component, I decided to include up to the sixth PC (PC6) to ensure that almost 70% of the total variance is captured. This approach strikes a balance between reducing dimensionality and retaining enough variance to maintain the dataset's explanatory power. By including the first six PCs, I preserved a meaningful portion of the variance for more accurate analysis and modeling.

As a result, The PCA results showed that the first six principal components explained 69.43% of the total variance. So, I chose first six principal component, which is a proper indicator for further analysis.

2.3 PCA Results for Variables

PCA results can be evaluated in terms of variables (e.g., design parameters) and rows (e.g.,geometry). I will start by extracting results for the variables.

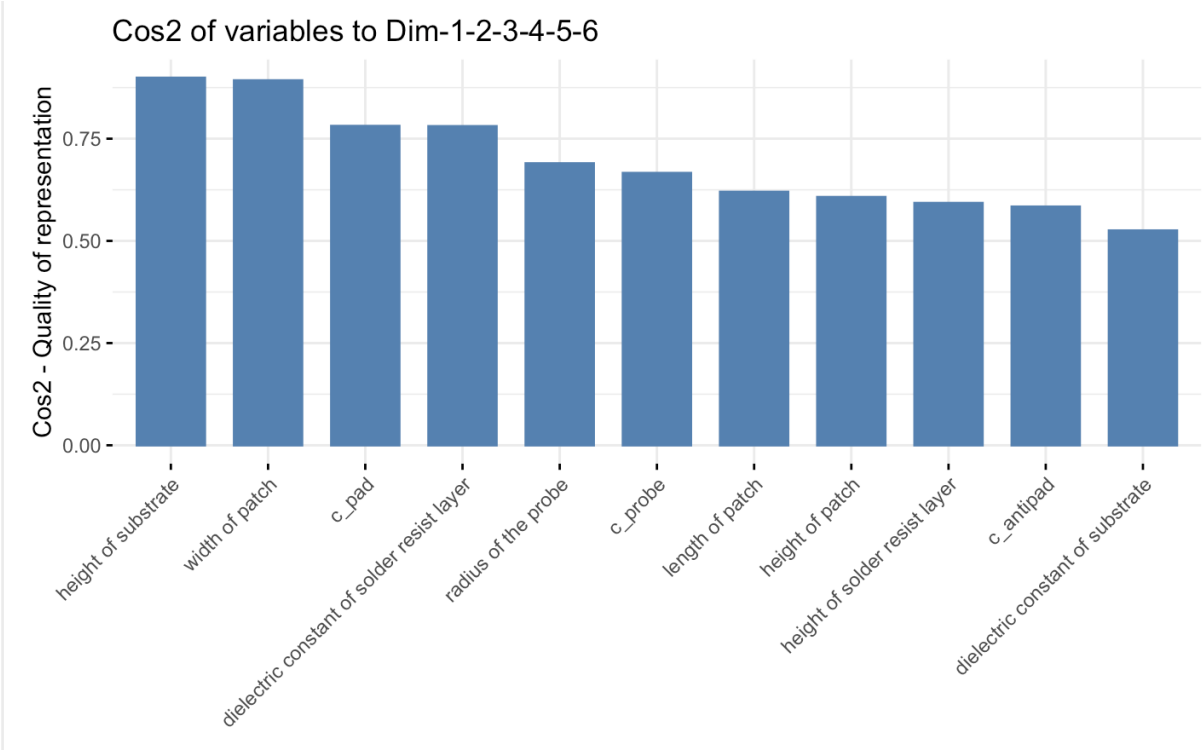
To achieve this, I used the get_pca_var() function, which generates a set of matrices containing detailed results for the active variables, including their coordinates, correlations with the principal axes, squared cosines (cos2), and contributions. The **cos2** value, or squared cosine, reflects how well a variable is represented on a given principal component. Higher values suggest that the variable is well represented by that component. In the plot, this is shown by larger and darker circles.



For instance, in this visualization:

- The **"width of patch"** and **"height of substrate"** have high cos2 values on the first dimension (Dim.1), indicating that these variables are well represented by PC1.
- **Dim.1 and Dim.2** appear to capture significant information from variables like **"width of patch," "height of substrate,"** and **"c_probe,"** which align with the earlier findings that these dimensions explain a large portion of the total variance.

- **Dim.3 to Dim.6** continue to show moderate cos2 values for various other parameters, supporting the decision to include up to six components to capture almost 70% of the total variance.



The provided graph illustrates the **cos² values (squared cosine)** of variables across the selected dimensions from the PCA analysis. In the chart, the **"height of substrate"** and **"width of patch"** show the highest cos² values, indicating that they are well-represented in the selected principal components (Dim-1 to Dim-6). I focused on variables with the highest cos² values to perform feature selection for **linear regression**, as these variables were most aligned with the principal components that explained a significant portion of the data's variance. I will explain it later in linear regression section detailly.

It is also important to point out that while PCR offers a straightforward approach to performing regression with fewer than p predictors (M<p), it does not function as a feature selection method. This is because each of the M principal components included in the regression is constructed as a linear combination of all p original features. Thus, feature selection cannot be performed for PCR and linear regression models where the predictors are principal components.

2.4 Comments about PCA Findings

PCA is a beneficial technique for this dataset, especially for reducing dimensions while capturing a high percentage of the variance, and is recommended for applications where interpretability of individual variables is less critical. However, if feature selection or interpretability is essential, PCA might need to be supplemented with other methods that allow a direct focus on specific variables.

There is a high correlation between the **width of patch** and **height of substrate** (correlation of approximately 0.92). Such a strong linear relationship indicates that these two features may contribute similar information to the dataset. This redundancy makes PCA suitable as it can capture the shared variance and reduce dimensionality by combining these variables into a single component.

Moderate correlation is observed between **width of patch** and **dielectric constant of substrate** (correlation of approximately 0.44), and between **height of substrate** and **dielectric constant of substrate** (around 0.46). These are not as strong as the previous pair, but they still indicate some degree of shared information, which PCA can exploit.

Most of the features exhibit low or negligible correlations with each other. For example, features like **height of solder resist layer**, **radius of the probe**, **c_pad**, **c_antipad**, and **c_probe** generally show weak correlations (close to zero) with other features. This means there is little linear dependence among these features, so PCA might not be as effective in reducing dimensionality for these uncorrelated features.

2.5 Assumptions of Principal Component Analysis (PCA)

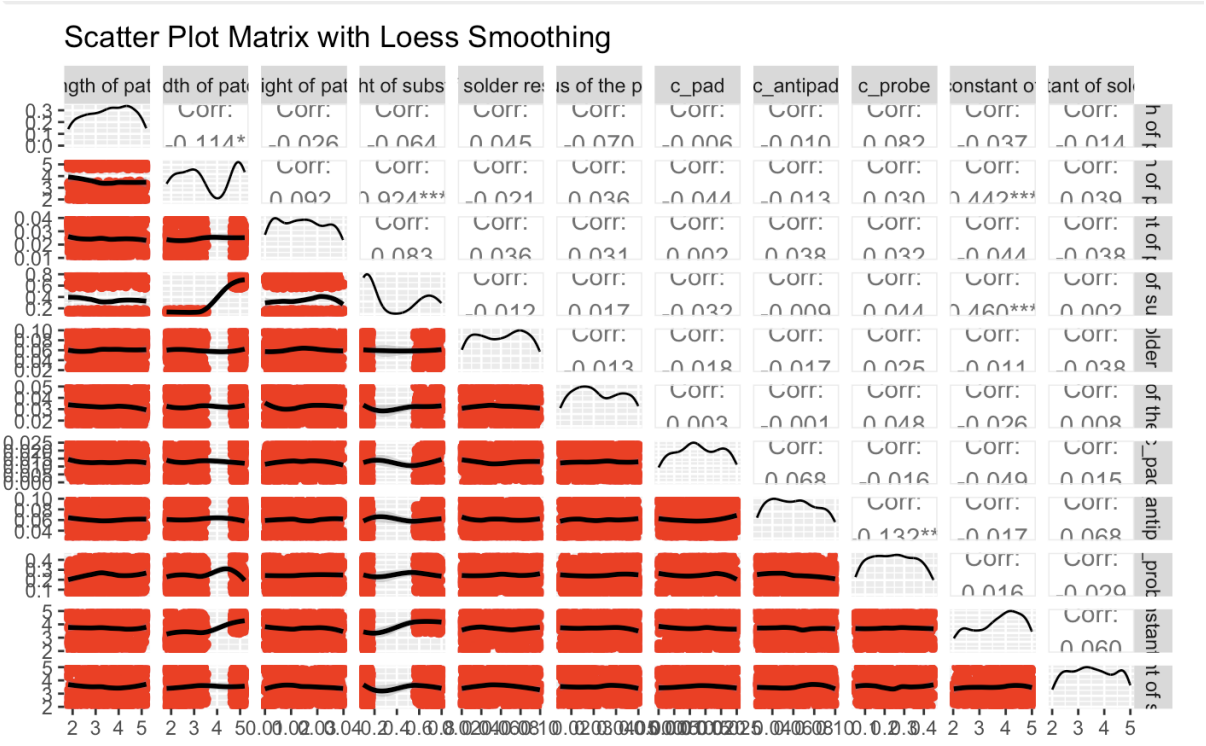
I performed the Shapiro-Wilk normality test on each variable in input_data to assess their normality.

Understanding W and p-value:

- **W value**: The Shapiro-Wilk statistic (W) ranges between 0 and 1, where a value closer to 1 suggests a distribution that more closely approximates normality.
- **p-value**: If the p-value is below a chosen significance level (typically 0.05), I would reject the null hypothesis that the data is normally distributed.

While PCA does not strictly require the data to be normally distributed, the interpretation of the results can be more straightforward if the data follows a Gaussian distribution. This is because many statistical properties and the behavior of PCA are better understood under the assumption of normality.

For all variables, I found **very low p-values** (all significantly below 0.05), which led me to reject the null hypothesis of normality for each variable. These results suggested that there was a **significant deviation from normality** across all features, which could impact analyses that rely on the assumption of normality (such as PCA or linear regression). Recognizing this, I considered that non-normal data might not yield optimal results in these methods.



PCA is particularly effective when there are several highly correlated features, as it can capture the shared variance among them and reduce redundancy. Here, PCA might capture some of the variance from the high-correlation pairs, such as **width of patch** and **height of substrate**. However, because most pairs show low correlations, PCA may not lead to a substantial reduction in dimensionality for this dataset.

3. Determination of Key Frequencies

I determined the key frequencies by first calculating the mean S11 magnitude across all designs at each frequency index. This average profile allowed me to identify frequencies where significant resonance might occur. To pinpoint these potential resonance points, I looked for local minima in the mean S11 magnitude, as these points typically correspond to dips in magnitude, which are indicative of resonance behavior.

Using the diff() and sign() functions, I detected points where the slope changed from negative to positive, marking local minima in the profile. After identifying all local minima, I sorted them by their mean S11 magnitude values and selected the seven lowest points. This selection represented frequencies where resonance was likely most pronounced, and I planned to use these indices for further modeling.

4. Linear Regression Models

4.1 The Linear Regression Model Established by Taking into Account the Problem of Multicollinearity

When examining the correlation between features, I observed that **width of patch** and **height of substrate** are highly correlated, with a correlation coefficient of approximately 0.92. This high correlation indicates a strong linear relationship between these two variables, which can introduce several issues in the regression model. Multicollinearity complicates the interpretation of the model, as it blurs the unique effect of each predictor on the outcome.

To mitigate multicollinearity, one effective approach is to remove one of the highly correlated variables. In this case, the height of the substrate showed a stronger relationship with the target variable (S11 parameters) and greater explanatory power across various frequency indices.

To decide which parameter to include in the model, I created separate models for each key frequency—specifically 75, 94, 106, 113, 127, 146, and 168—using each feature individually and examined the outputs. By comparing the model results, I was able to assess the predictive power and stability of each feature in relation to the target variable.

Given the higher adjusted R^2 values, especially for the real part, and the strong significance of **height of substrate** across both parts (real and imaginary), it is preferable to retain **height of substrate** in the model over **width of patch**. Height of substrate not only captures more variance but also consistently contributes to a more statistically robust model, making it a stronger predictor for S11 behavior in this dataset. Specifically, with height of substrate included, adjusted R^2 values are around 0.78 to 0.80 for the real part of S11, which is higher than when width of patch is included. This indicates that height of substrate provides a better fit for predicting real S11 values.

As an example, the model created with the **width of patch** at the 75th frequency is as follows:

```
Regression Results for Real Part of S11

Frequency Index: 75

Call:
lm(formula = target ~ ., data = data_real)

Residuals:
     Min       1Q   Median       3Q      Max
-0.72839 -0.26113 -0.02432  0.20867  1.82594

Coefficients:
                                        Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -1.882298   0.178698 -10.533  < 2e-16 ***
`length of patch`                       0.011882   0.019482   0.610  0.54230
`width of patch`                        0.474161   0.017782  26.665  < 2e-16 ***
`height of patch`                      -0.479438   2.134898  -0.225  0.82243
`height of solder resist layer`        -0.354812   0.786649  -0.451  0.65222
`radius of the probe`                  -4.257738   1.803482  -2.361  0.01875 *
c_pad                                   1.215019   2.621359   0.464  0.64327
c_probe                                 0.444194   0.169449   2.621  0.00911 **
c_antipad                              -0.124375   0.875756  -0.142  0.88714
`dielectric constant of substrate`     -0.016573   0.024415  -0.679  0.49769
`dielectric constant of solder resist layer` -0.002522   0.021438  -0.118  0.90641
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3638 on 374 degrees of freedom
Multiple R-squared:  0.7061,    Adjusted R-squared:  0.6982
F-statistic: 89.85 on 10 and 374 DF,  p-value: < 2.2e-16
```

The model created with the **height of substrate** at the 75th frequency is as follows:

```
Regression Results for Real Part of S11

Frequency Index: 75

Call:
lm(formula = target ~ ., data = data_real)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7259 -0.1830 -0.0647  0.1165  1.4749

Coefficients:
                                       Estimate Std. Error t value Pr(>|t|)
(Intercept)                            -0.80478    0.15162  -5.308  1.9e-07 ***
`length of patch`                      -0.01390    0.01649  -0.843  0.39989
`height of substrate`                   2.20939    0.06578  33.587  < 2e-16 ***
`height of patch`                      -0.41158    1.81254  -0.227  0.82049
`height of solder resist layer`        -0.48179    0.66851  -0.721  0.47155
`radius of the probe`                  -3.57033    1.53200  -2.330  0.02031 *
c_pad                                   0.13698    2.22728   0.062  0.95099
c_probe                                 0.37544    0.14408   2.606  0.00953 **
c_antipad                              -0.33922    0.74429  -0.456  0.64882
`dielectric constant of substrate`     -0.05365    0.02098  -2.558  0.01093 *
`dielectric constant of solder resist layer`  0.02258    0.01822   1.239  0.21606
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3092 on 374 degrees of freedom
Multiple R-squared:  0.7877,    Adjusted R-squared:  0.782
F-statistic: 138.8 on 10 and 374 DF,  p-value: < 2.2e-16
```

As a result, the linear regression model built with height of substrate along with nine other features—totaling 10 independent variables—yields the following outcomes:

Regression Results for Real Part of S11

Frequency Index: 75

Adjusted R^2: 0.782

Frequency Index: 94

Adjusted R^2: 0.7715

Frequency Index: 106

Adjusted R^2: 0.7734

Frequency Index: 113

Adjusted R^2: 0.7673

Frequency Index: 127

Adjusted R^2: 0.7842

Frequency Index: 146

Adjusted R^2: 0.783

Frequency Index: 168

Adjusted R^2: 0.7934

Regression Results for Imaginary Part of S11

Frequency Index: 75

Adjusted R^2: 0.2442

Frequency Index: 94

Adjusted R^2: 0.2852

Frequency Index: 106

Adjusted R^2: 0.3113

Frequency Index: 113

Adjusted R^2: 0.3103

Frequency Index: 127

Adjusted R^2: 0.2905

Frequency Index: 146

Adjusted R^2: 0.2758

Frequency Index: 168

Adjusted R^2: 0.2767

Detailed output is shown in R code.

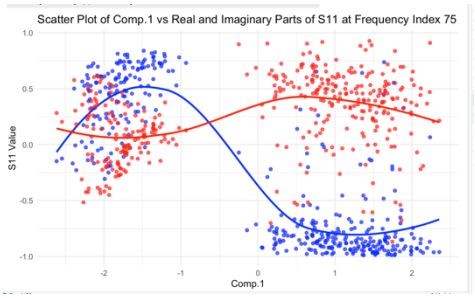
4.2 The Linear Regression Model Established Using PCA Components as Independent Variables

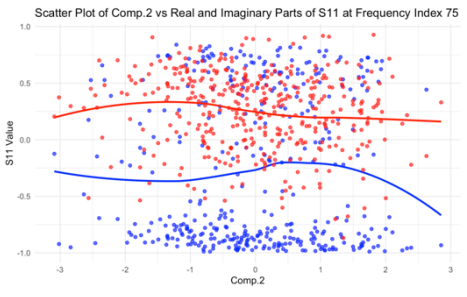
To build a Principal Component Regression (PCR) model, I first performed Principal Component Analysis (PCA) on the input data to reduce the dimensionality and then used the principal components as predictors in a regression model.

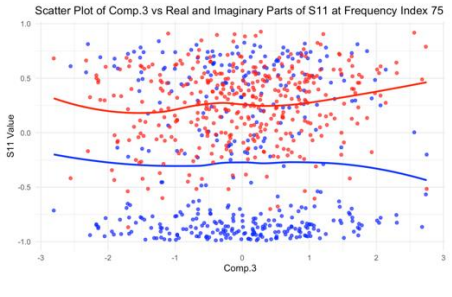
Although this method appears similar to Principal Component Regression (PCR), I actually implemented it by directly using the principal components as features in a standard linear regression model in R. Instead of following the full PCR procedure, where PCA and regression are performed as separate steps, I streamlined the process by inputting the principal components as independent variables directly into the linear regression model.

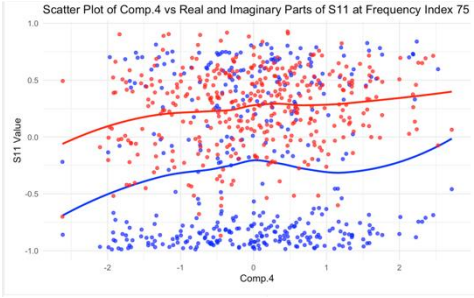
In this analysis, I used six principal components (PCs) as predictors in the regression model for the real part of S11 across various frequency indices. I selected six PCs because they capture approximately 69% of the total variance in the original data, which provides a good balance between reducing dimensionality and retaining the essential information from the design parameters.

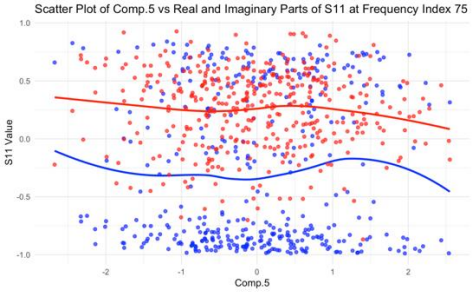
The scatter plots visually showed the relationship between each principal component (PC) and the target variable. This helped me understand which components had a greater impact on the target variable. If there was a strong relationship, those components became more important in the model. In the scatter plots below, the blue points represent the real part, while the red points represent the imaginary part.

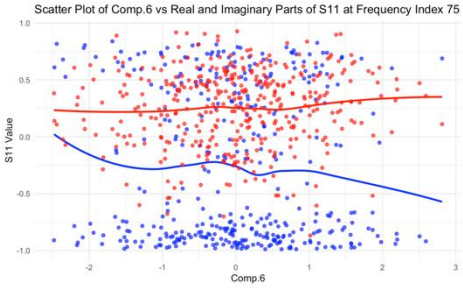












Upon examining the scatter plots, I observed that **PC1** showed a clear non-linear relationship with the real part of S11, indicating that it was likely an important component for the model. **PC2** displayed a weak association, while **PC3**, **PC5**, and **PC6** had almost no discernible relationship with the target variable, suggesting that they may not contribute significantly to the model. **PC4** showed a slight trend but with a limited effect.

Since it was clear that only PC1 had a linear relationship with the output, while the other components did not, I considered alternative approaches to improve the model. I explored potential options such as **non-linear models** (e.g., polynomial regression, support vector machines, neural networks, random forests), **transformations** on non-linear components, and **interaction terms** between components to better capture complex relationships. Additionally, I considered using a **Generalized Additive Model (GAM)** to allow for non-linear relationships in each component.

However, I did not implement these approaches in this assignment.

The outputs of the linear regression model I built using the first 6 principal components are provided below for both the real part and imaginary part, as an example for the 75th frequency.

```
Linear Regression Model Using the Principal Components for Real Part of S11

Frequency Index: 75

Call:
lm(formula = target ~ ., data = data_real_pcr)

Residuals:
     Min       1Q   Median       3Q      Max
-0.76908 -0.23840 -0.03737  0.16742  1.66024

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.28601    0.01880 -15.213   <2e-16 ***
Comp.1      -0.36183    0.01246 -29.051   <2e-16 ***
Comp.2       0.02472    0.01704   1.451   0.1477
Comp.3      -0.00197    0.01792  -0.110   0.9125
Comp.4       0.03959    0.01829   2.164   0.0311 *
Comp.5       0.03247    0.01870   1.736   0.0834 .
Comp.6      -0.04538    0.01906  -2.381   0.0178 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3689 on 378 degrees of freedom
Multiple R-squared:  0.6945,    Adjusted R-squared:  0.6897
F-statistic: 143.2 on 6 and 378 DF,  p-value: < 2.2e-16
```

```
Principal Component Regression for Imaginary Part of S11

Frequency Index: 75

Call:
lm(formula = target ~ ., data = data_img_pcr)

Residuals:
     Min       1Q   Median       3Q      Max
-1.11780 -0.18136  0.01306  0.19902  0.68974

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.25496    0.01539  16.568  < 2e-16 ***
Comp.1       0.08565    0.01019   8.402 9.02e-16 ***
Comp.2      -0.04089    0.01394  -2.932  0.00357 **
Comp.3       0.04419    0.01467   3.012  0.00277 **
Comp.4       0.04830    0.01497   3.226  0.00136 **
Comp.5      -0.02145    0.01531  -1.401  0.16191
Comp.6       0.02738    0.01560   1.755  0.08002 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3019 on 378 degrees of freedom
Multiple R-squared:  0.2153,    Adjusted R-squared:  0.2028
F-statistic: 17.29 on 6 and 378 DF,  p-value: < 2.2e-16
```

By using the principal components as predictors, I aimed to reduce dimensionality while preserving the most influential features of the original dataset. The PCR model for the real part of S11 consistently showed a strong fit, with R^2 values ranging from approximately 0.69 to 0.76 across key frequencies.

In contrast, the model for the imaginary part of S11 had lower R^2 values, typically ranging between 0.21 and 0.24. This indicated that the PCR model explained a limited portion of the variance for the imaginary part, suggesting that a linear approach might not fully capture its complexity. As seen in the scatter plots, there was no clear relationship between the principal components (PCs) and the imaginary part; data points were evenly distributed on the graph, and the smooth line was almost horizontal. This indicated that neither a linear nor a non-linear model would likely be strong enough for the imaginary part, as there was no evident correlation between the independent variables (PCs) and the imaginary part.

4.3 Principal Component Regression

In the model where I used the first 6 PCs as features, I applied a standard linear regression model without any evaluation process such as cross-validation. PCR, on the other hand, is an integrated method that combines PCA with linear regression. Cross-validation helped determine the model's generalizability and the optimal number of components to include. PCR is a dimensionality reduction technique combined with regression, which is particularly useful when predictor variables are highly collinear or when we want to reduce the complexity of the model.

I first created a model with all the PCs included, resulting in 11 independent variables. In this model, I aimed to identify the point at which the Cross-Validation Root Mean Square Error of Prediction (RMSEP) score reached its minimum and observe how many PCs were included up to that point.

As an example, the result of frequency 75 is given below. The results for the remaining key frequencies can be found in the HTML file.

```
Principal Component Regression for Real Part of S11

Frequency Index: 75
Data:   X dimension: 385 11
        Y dimension: 385 1
Fit method: svdpc
Number of components considered: 11

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
CV          0.6631   0.6488   0.6304   0.6267   0.6164   0.6032   0.5679   0.5494   0.5257   0.4578    0.4061    0.3117
adjCV       0.6631   0.6807   0.6904   0.6996   0.4411   0.4459   0.4455   0.4321   0.4292   0.3901    0.3610    0.3112

TRAINING: % variance explained
       1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X        9.091   18.182   27.273    36.36    45.45    54.55    63.64    72.73    81.82     90.91    100.00
target   1.852    2.865    3.017    71.22    71.52    71.52    74.89    74.89    77.77     78.90     79.07
NULL
Optimal number of components selected: 11
```

```
Principal Component Regression for Imaginary Part of S11

Frequency Index: 75
Data:   X dimension: 385 11
        Y dimension: 385 1
Fit method: svdpc
Number of components considered: 11

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
CV          0.3386   0.3376   0.3378   0.3357   0.3348   0.3307   0.3304   0.3254   0.3166   0.3131    0.3097    0.2980
adjCV       0.3386   0.3356   0.3349   0.3361   0.3114   0.3120   0.3134   0.3110   0.3049   0.3054    0.3039    0.2975

TRAINING: % variance explained
         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps  9 comps  10 comps  11 comps
X          9.091   18.182   27.273    36.36    45.45    54.55    63.64    72.73    81.82     90.91    100.00
target     3.150    5.923    7.309    21.96    22.89    22.99    24.44    26.12    26.23     26.84     26.96
NULL
Optimal number of components selected: 11
```

I used 10-fold cross-validation to measure Root Mean Square Error of Prediction (RMSEP) across different component levels. Using all 11 components yielded the best predictive accuracy, while the percentage of variance explained showed that the model was robust.

I created another PCR model using the 6 principal components I previously selected for the linear model. The results for this model at the 75th frequency are shown below:

```
Principal Component Regression for Real Part of S11

Frequency Index: 75
Data:   X dimension: 385 6
        Y dimension: 385 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.6631   0.6199   0.6049   0.5985   0.5469   0.5070   0.3716
adjCV       0.6631   0.6844   0.7039   0.4515   0.4450   0.4304   0.3712

TRAINING: % variance explained
         1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X       16.66667  33.3333    50.00    66.67    83.33   100.00
target   0.02529   0.4571    68.66    68.67    69.43    69.45
NULL
```

```
Principal Component Regression for Imaginary Part of S11

Frequency Index: 75
Data:   X dimension: 385 6
        Y dimension: 385 1
Fit method: svdpc
Number of components considered: 6

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV          0.3386   0.3360   0.3309   0.3307   0.3237   0.3107   0.3043
adjCV       0.3386   0.3322   0.3314   0.3096   0.3092   0.3046   0.3040

TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
X        16.667   33.333    50.00    66.67    83.33   100.00
target    6.372    6.389    21.04    21.40    21.51    21.53
NULL
```

The RMSEP levels off between the 5th and 6th components, reaching a similar range as the full 11-component model for some frequency indices. This suggests that the top 6 components can achieve comparable predictive accuracy while keeping the model simpler.

I found that using the 6-component model offered a streamlined approach with predictive performance comparable to the 11-component model. By limiting the number of components to 6, I was able to balance simplicity and predictive accuracy, capturing most of the variance while reducing the risk of overfitting. Although the 11-component model slightly improved RMSEP, this improvement was marginal beyond the 6th component and came at the cost of added complexity.

In summary, the 6-component PCR model offers nearly the same level of predictive accuracy as the 11-component model while maintaining simplicity and potentially better generalization. Therefore, it is a more practical choice for predicting the real part of S11 across the evaluated frequency indices.

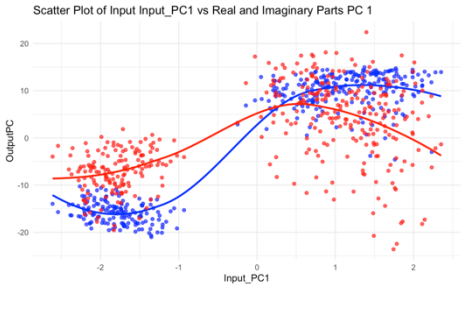
5. Visualization of Principal Component Relationships Between Input Data and Real/Imaginary Output Components

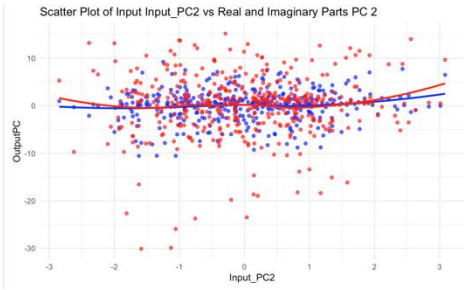
In this analysis, I extended the principal component comparison to six components to better understand the relationships between the input data and the real and imaginary parts of the output. My goal was to determine whether the patterns observed in the input data's primary six components aligned similarly with the structures in both the real and imaginary output parts.

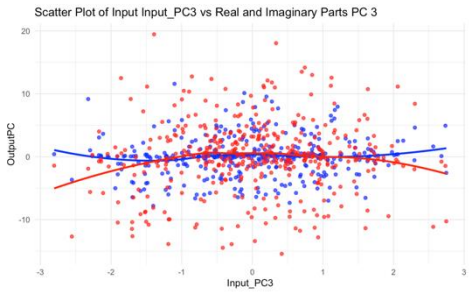
To achieve this, I first conducted PCA separately on the real and imaginary parts of the output, extracting the scores for the first six principal components in each case. By creating pairwise scatter plots and calculating correlations for each corresponding component, I could visually and quantitatively compare how closely the variance structures aligned across these datasets.

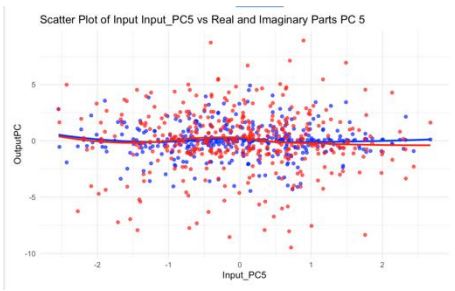
By creating pairwise scatter plots and calculating correlations for each corresponding component, I could visually and quantitatively compare how closely the variance structures aligned across these datasets.

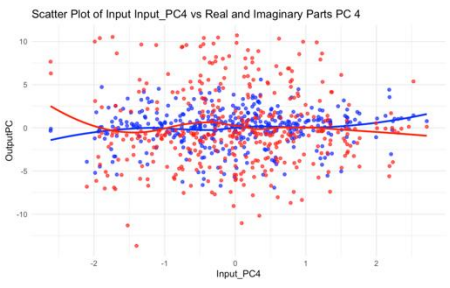
Plots are given below.

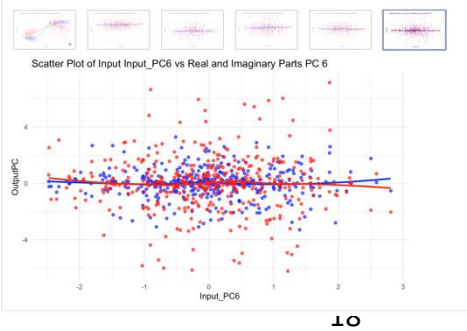










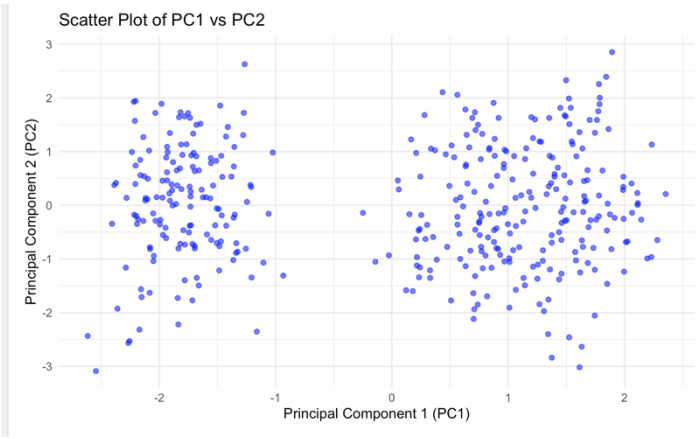


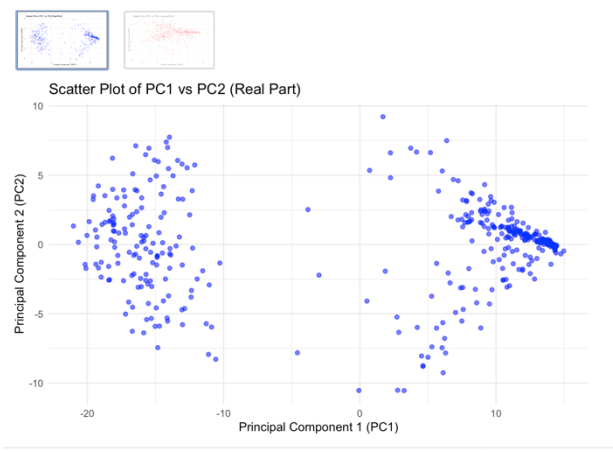
Ahmet Çeliker
2020402126

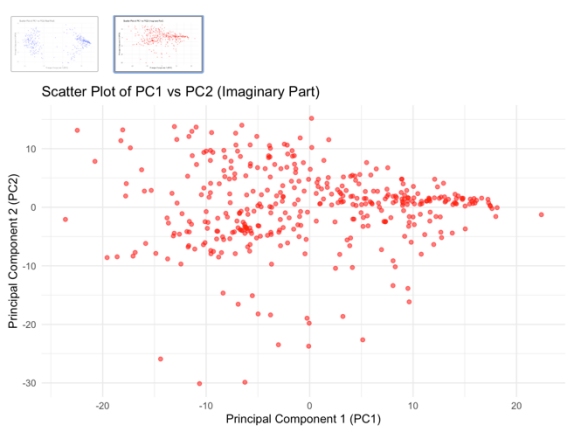
In my analysis, I found that only the first principal component (PC1) displayed a meaningful, non-linear relationship with both the real and imaginary parts of the output data. For the remaining principal components (PC2 and beyond), I noticed a nearly horizontal line in their plots with the output. This flat line indicated that these higher-order components did not contribute significantly to changes in the output. This suggested to me that while these components captured additional variance in the input data, they did not hold relevant information for predicting the real or imaginary parts of the output. In other words, the variance represented by these components might consist of noise or patterns unrelated to the output's structure.

I saw a similar pattern when the scatter plot of the first 2 principal components of the input data is compared with the PC of the output part.

Below are these two plots.







In the input data's first two principal components (PC1 and PC2), the data points are primarily concentrated in two distinct regions, with higher densities on the left and right sides of the plot. This pattern suggests that certain features or parameter values in the input data are creating natural groupings or tendencies to fall within these specific areas along the first two principal components. When I examined the output data's real part in terms of its first two principal components, I noticed a similar tendency, where data points also concentrated in particular regions.

In contrast, when I examined the imaginary part of the output data in terms of its first two principal components, I observed a completely different distribution. Unlike the real part, which showed similar density patterns to the input data, the imaginary part displayed a more dispersed, random spread across the plot, without any clear areas of concentration.

From this observation, I concluded that while the real part of the output data retained some underlying structure from the input features, the imaginary part did not exhibit this relationship.

However, since I used the first six principal components in my models, this analysis did not provide a definitive direction for my approach. My takeaway was as follows: if I had only used the first two principal components, the real part would have shown a fairly similar relationship with the input data, while the imaginary part would have fallen short in explaining the variance effectively. This insight reinforced the importance of including more components to capture the underlying complexity in both the real and imaginary parts of the output.

6. Cross Validation of Linear Models

Cross-validation was crucial for my linear regression model because it allowed me to evaluate how well the model could generalize to unseen data. By dividing the data into multiple segments and validating the model on each, I could assess the model's performance more robustly, reducing the risk of overfitting. This method provided an unbiased estimate of how the model would perform in practice, rather than just on the training dataset.

6.1 Cross Validation of Linear Regression Model

The cross-validation (CV) results for the linear regression model using the 10 predictors that I decided upon in part 4.1 (input data) for both the real and imaginary parts of S11 offer

insights into the model's performance in terms of RMSE (Root Mean Squared Error), R-squared, and MAE (Mean Absolute Error) across various key frequency indices.

For the real part, the R-squared values range from approximately 0.68 to 0.75 across the frequency indices, indicating that the model explains between 68% and 75% of the variance in the data. The RMSE values range from about 0.32 to 0.37, which indicates the average error in prediction. The MAE, which measures the average absolute differences between predicted and actual values, ranges from around 0.26 to 0.29.

In contrast, for the imaginary part, the R-squared values are notably lower, ranging from approximately 0.20 to 0.24 across frequency indices, suggesting that the model explains only about 20% to 24% of the variance in the data. The RMSE values for the imaginary part are in the range of 0.30 to 0.32, and the MAE is around 0.23 to 0.24.

Below are the results of CV for some key frequencies.

Real part:

```
Cross-Validation Results for Linear Regression Model (Real Part of S11)

Frequency Index: 75
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 347, 347, 346, 345, 347, 348, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3631303  0.7082465  0.2856734

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 94
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 345, 348, 347, 345, 348, 346, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3661306  0.6836447  0.283499

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 106
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 346, 346, 345, 347, 348, 346, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3516909  0.6997634  0.2732546

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 113
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 345, 347, 347, 347, 346, 346, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3458379  0.7090632  0.2710509

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 127
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 348, 346, 346, 346, 348, 346, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3322016  0.7221621  0.2590718

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 146
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 346, 346, 348, 346, 346, 347, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3213795  0.7381832  0.2575878

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 168
Linear Regression

385 samples
 10 predictor

No pre-processing
```

Imaginary Part:

```
Cross-Validation Results for Linear Regression Model (Imaginary Part of S11)

Frequency Index: 75
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 348, 346, 348, 347, 346, 347, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3075202  0.1951476  0.2360048

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 94
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 348, 346, 346, 347, 348, 347, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.3010571  0.204205  0.2388721

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 106
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 346, 346, 346, 346, 347, 347, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3027145  0.2327399  0.2374258

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 113
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 347, 347, 348, 347, 346, 346, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3052654  0.2158634  0.2381139

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 127
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 346, 346, 345, 346, 346, 348, ...
Resampling results:

  RMSE       Rsquared   MAE
  0.3156169  0.2135843  0.2436342

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 146
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
Summary of sample sizes: 345, 346, 346, 346, 346, 348, ...
Resampling results:

  RMSE       Rsquared  MAE
  0.3070245  0.227906  0.238302

Tuning parameter 'intercept' was held constant at a value of TRUE

Frequency Index: 168
Linear Regression

385 samples
 10 predictor

No pre-processing
Resampling: Cross-Validated (10 fold)
```

6.2 Cross Validation of PCR

The cross-validation (CV) results for the linear regression models using the first six principal components (PCs) as predictors, applied separately to the real and imaginary parts of the S11 parameter, offer insights into the model's performance and effectiveness.

For the real part, R2 values range from 0.69 to 0.76, meaning the model with six principal components captures 69% to 76% of the variance. The RMSE, indicating prediction error,

spans 0.31 to 0.37, while the MAE lies between 0.24 and 0.28, showing a close match between predictions and actual values.

For the imaginary part, the R2R^2R2 values remain low, between 0.20 and 0.24, showing that the model explains only 20% to 24% of the variance using six principal components. RMSE ranges from 0.30 to 0.32, with MAE around 0.23 to 0.24, indicating moderate accuracy but limited predictive reliability.

6.3 Comments about Cross Validation Results

When I compared the linear regression model using principal components as predictors with the regular regression model (which used the original features directly), I found that the regular regression model had higher RMSE values and lower R-squared values. This suggests that, in terms of both predictive accuracy and explanatory power, the regular regression model underperformed compared to the model based on principal components.

This outcome implies that, for this dataset, the PCR model was more effective at capturing the key patterns in the data. By transforming the original features into principal components, the PCR model may have reduced redundancy and focused on the most important information, which enhanced its predictive and explanatory power.

7. Research Questions

Although I addressed these questions indirectly throughout the report, I wanted to provide a clear and direct answer for the sake of completeness and clarity.

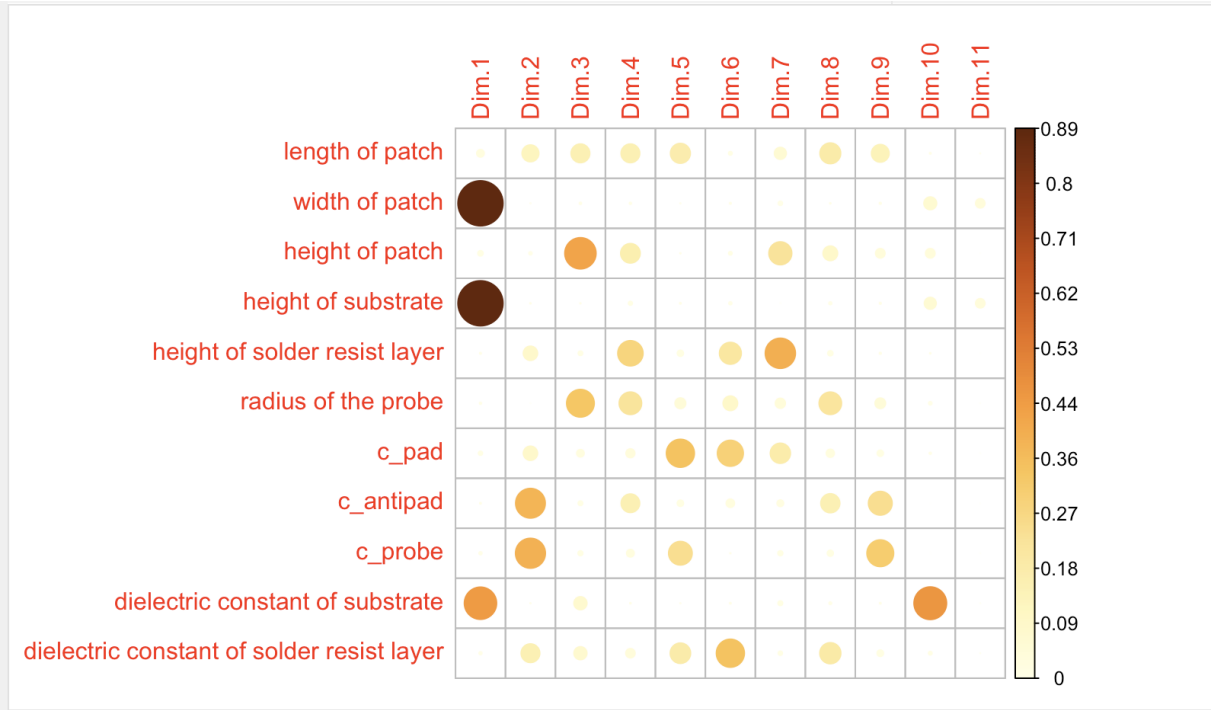
7.1 Dimensionality Reduction with PCA:

7.1.1 Can we reduce the complexity of the design space by using Principal Component Analysis (PCA) to identify key parameters that most influence the S11 response of the antenna?

We can measure the influence of key parameters on the principal components and observe how each original feature contributes to these new components. However, by applying PCA,

we cannot eliminate any features directly from the model; instead, we combine the original features into new components based on their variance contributions. This allows us to understand the importance or impact of key parameters on the model without discarding specific features. Thus, we can only use PCA to observe the influence of parameters, not to select or remove features directly from the model.

In the following plot, we can observe how each variable contributes to the principal components (PCs).



Principal Component Regression (PCR) is useful for reducing the dimensionality of the design space by using a smaller number of principal components ($M < p$) to capture the most significant variation in the data. However, PCR is not a feature selection method, as each principal component is a combination of all original features. This means that PCR does not identify specific individual parameters that influence the S11 response but rather captures overall patterns and variability in the data through a transformed set of components.

In the context of the design space for the antenna's S11 response, while PCA can reduce complexity, it does not isolate key parameters. Instead, it compresses the influence of all parameters into a few principal components, similar to how ridge regression distributes influence across features rather than selecting specific ones. Therefore, while PCA may simplify the analysis by reducing the number of predictors, it does not directly indicate which

original parameters most affect the S11 response, and the results may still involve contributions from all features.

However, for the imaginary part of S11, the linear relationship captured by the PCs is weaker, as evidenced by lower R-squared values in regression models. This indicates that while PCA can reduce dimensionality, it may not capture all complex relationships, especially non-linear interactions. Overall, PCA proves useful in simplifying the design space, but further refinement may be needed for complex patterns in the imaginary part.

7.1.2 How much of the total variance in the design parameter space can be explained by the principal components?

Based on the PCA results, the first principal component (Comp.1) explains approximately 20.7% of the variance.

The cumulative explained variance by the first two components (Comp.1 and Comp.2) is around 31.8%.

By the time we include the first six components, about 69.4% of the variance is explained.

Including all eleven components explains nearly 100% of the variance.

This indicates that while the first few components capture a significant portion of the variance, to fully capture all variability in the design space, all components are needed. However, focusing on the first 6-8 components would still retain a high level of the design space variance, allowing for a reduction in complexity while preserving most of the information.

7.1.3 What insights can we draw from the PCA regarding the relationship between geometry and electromagnetic behavior?

From the PCA results, we can derive several insights into the relationship between the geometry of the design parameters and the antenna's electromagnetic (EM) behavior.

By examining the loadings of the first few principal components, we see which geometric features contribute most to each PC. For example, features like the "width of patch" and "height of substrate" have high loadings in the first component, indicating these dimensions play a critical role in influencing the S11 response.

The lower R-squared values for the imaginary part in the regression analysis with PCs suggest that some aspects of the S11 response, especially the imaginary component, may involve more complex or non-linear relationships with the geometry.

Additionally, when examining the scatter plot of PC1 and PC2, we observe a strong resemblance to the scatter plot of the real part of S11. This similarity in pattern suggests that the design parameters which heavily influence PC1 and PC2 also significantly impact the S11 response. Consequently, parameters with high loadings on these principal components are likely to play a key role in determining the electromagnetic behavior represented by S11. This insight is valuable for design optimization, as it indicates that focusing on the parameters contributing most to PC1 and PC2 could yield meaningful adjustments to the S11 response.

Also, the fact that a large portion of the variance (around 70%) is explained only by the first six principal components, rather than just a few, suggests that there is not a high linear correlation among the parameters. If strong linear relationships existed between the parameters, fewer components would be able to explain a larger portion of the variance. This distribution of variance across multiple components indicates that each parameter contributes unique information to the model, with limited redundancy among them.

As a result, parameters such as **width of patch**, **height of substrate**, and **c_probe** strongly influence the first two components, suggesting that these design parameters play a significant role in shaping the S11 response.

The **height of patch** and **radius of the probe** are particularly impactful in PC3, indicating these parameters also contribute to the antenna's performance but in a secondary manner compared to PC1 and PC2.

Parameters that are highly correlated, like **width of patch** and **height of substrate** (correlation of 0.92), contribute similarly to specific components, particularly PC1. This reinforces the idea that these parameters interact closely in determining the antenna's characteristics.

The high variance captured by the first few components suggests that designers can focus on optimizing a subset of the parameters (such as width, height, and certain capacitive elements like c_probe and c_antipad) for effective S11 response.

7.2 Regression Modeling for S11

7.2.1 Given that S11 parameters are evaluated at 201 frequency points, predicting them simultaneously through multitarget regression may be computationally intensive and beyond the scope of the content covered so far. Instead, can we simplify the regression task by focusing on predicting S11 at a few key frequency points?

Yes, focusing on predicting S11 at a few key frequency points is a practical approach to simplify the regression task. Given that evaluating S11 at all 201 frequency points would require significant computational resources and may add unnecessary complexity, selecting representative frequency points can make the task more manageable.

I chose the local minima in the S11 magnitude plot because these points represented frequencies where the S11 parameter reached its lowest values. In antenna theory, these low points were critical since they corresponded to resonance frequencies, which held particular significance for the antenna's performance.

At these resonance frequencies, maximum power transfer occurred. This was because the antenna's impedance matched the source impedance at these points, allowing maximal power transfer from the source to the antenna.

To identify the key frequency indices, I calculated the S11 magnitude by taking the square root of the sum of squares of the real and imaginary parts. Afterward, I examined the S11 magnitude profiles for the first six designs to understand their behavior across frequencies. By averaging the S11 magnitudes across all designs, I derived the mean S11 magnitude across

frequencies. I then used a method to detect local minima in this mean S11 magnitude curve, focusing on frequencies where S11 reached its lowest points. These local minima were marked as key frequency indices because they indicated points of resonance where the antenna's efficiency was at its peak, characterized by minimal reflected power.

7.2.3 How effective are linear regression models in predicting the real and imaginary components of S11 at these selected frequency points, based on the geometric parameters of the antenna design?

In my analysis, I compared the predictive performance of two regression models: a standard linear regression model using the original features and a principal component regression (PCR) model that utilized principal components as predictors. The goal was to determine which model provided greater accuracy and explanatory power when predicting the S11 parameter's real and imaginary components at selected frequency points.

When I examined the real component predictions, I found that the PCR model performed better overall. Specifically, the PCR model had lower root mean square error (RMSE) values and higher R2 values compared to the regular linear regression model. This indicated that PCR provided both better predictive accuracy and a stronger fit to the data. By transforming the original features into principal components, the PCR model likely reduced redundancy among highly correlated variables, allowing it to focus on the most significant patterns within the dataset. As a result, the PCR model captured the essential relationships more effectively, enhancing both its predictive accuracy and explanatory power.

For the imaginary component, however, neither the PCR model nor the regular linear regression model produced satisfactory results. Both models exhibited low R2 values and moderate RMSE scores, indicating limited predictive accuracy for this component.

This outcome suggested that the features in the dataset might be insufficient for explaining the imaginary component.

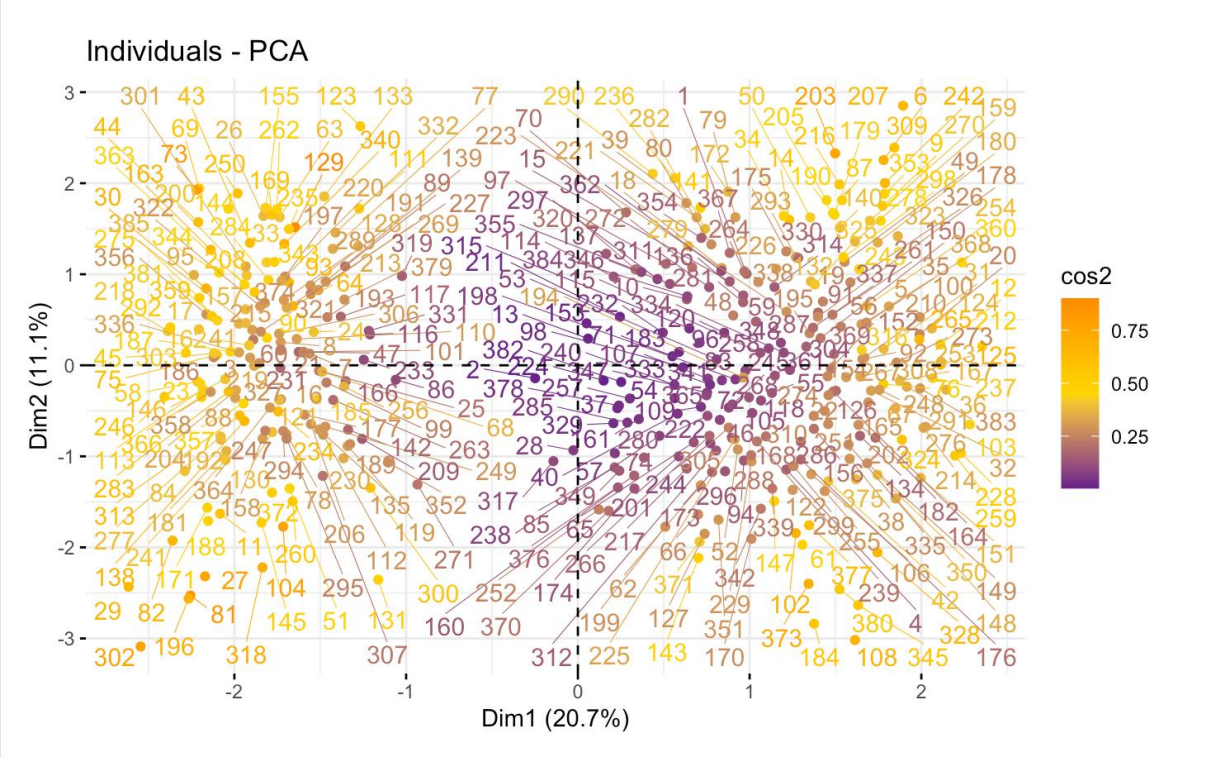
7.2.4 What patterns emerge when linear regression is applied to individual frequency points, and do these patterns suggest any broader trends in the design space?

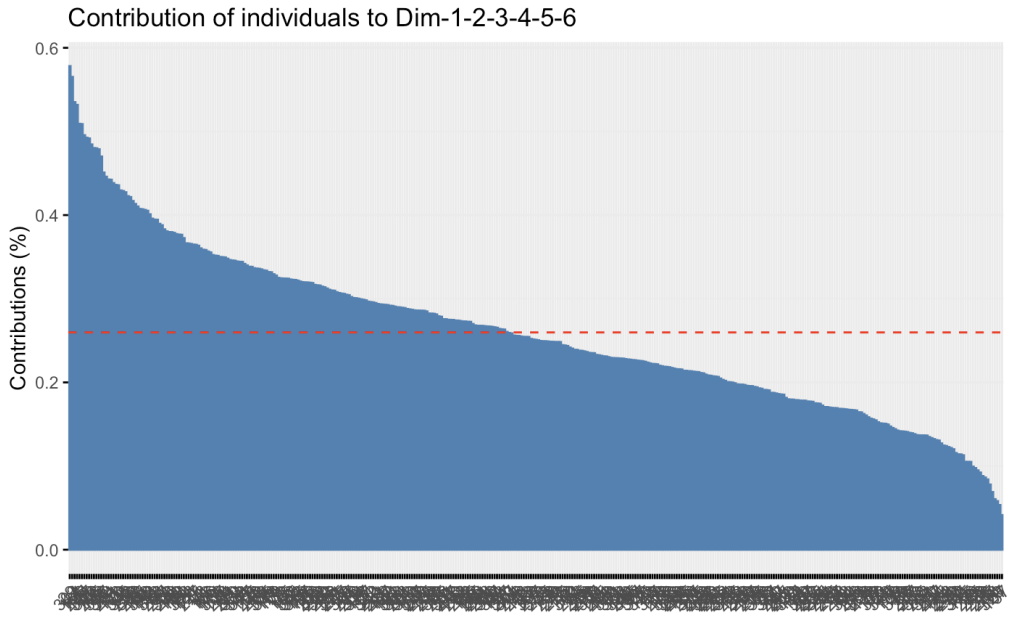
Across different frequency points, some design parameters consistently show significance in the linear regression model. For example, parameters such as the width and length of the patch and radius of the probe often emerge as influential predictors for the real component of

S11. This consistency suggests that these dimensions play a central role in shaping the antenna's performance at various frequencies, highlighting their importance in the overall design space.

Multicollinearity is observed between certain features, such as the height of substrate and width of the patch. This collinearity can inflate the variance of the regression coefficients if not managed properly. This pattern suggests that, in the broader design space, some features contribute redundant information, which may not add much additional insight but can complicate model interpretation and accuracy.

The influence of certain features on S11 may vary with frequency. For instance, at lower frequencies, the effects of specific geometric parameters might be more pronounced, while at higher frequencies, other parameters may become more influential.





Based on these two graphs, I observed that each frequency had a unique impact on the principal components (PCs). In the first graph, the distribution of contribution percentages indicated that certain frequencies contributed more significantly than others. This suggested that specific frequencies added more variance to certain dimensions, making them more informative in those directions.

In the second graph, which displayed the contribution of each individual frequency to PC1 and PC2, I noted that certain frequencies had stronger or weaker influences on these two principal components, as represented by the color gradient. Frequencies with higher cos2 values, shown in darker colors, indicated a stronger contribution to these PCs. The clustering patterns in different areas implied that not all frequencies behaved similarly, with some frequencies adding more variance or significance to the analysis.

These observations led me to conclude that the relationship between frequencies and design parameters was multidimensional, with each frequency contributing differently to the principal components. This result suggested that certain geometric features of the antenna were more influential at specific frequencies, impacting the electromagnetic response in unique ways across the frequency spectrum.

7.3 Model Performance and Interpretability:

7.3.1 How do PCA and regression models compare in terms of their ability to simplify and predict the antenna's performance?

I used PCA and regression models to simplify and predict the antenna's performance, each serving a distinct but complementary role. With PCA, I aimed to reduce the dimensionality of the data by transforming it into a smaller set of principal components that captured most of the variance in the original design parameters. In the context of antenna design, PCA helped me identify which combinations of geometric parameters had the largest influence on overall performance. However, I found that PCA alone did not predict performance metrics like the S11 parameter; rather, it provided a more streamlined representation of the design space that could be used in further analyses.

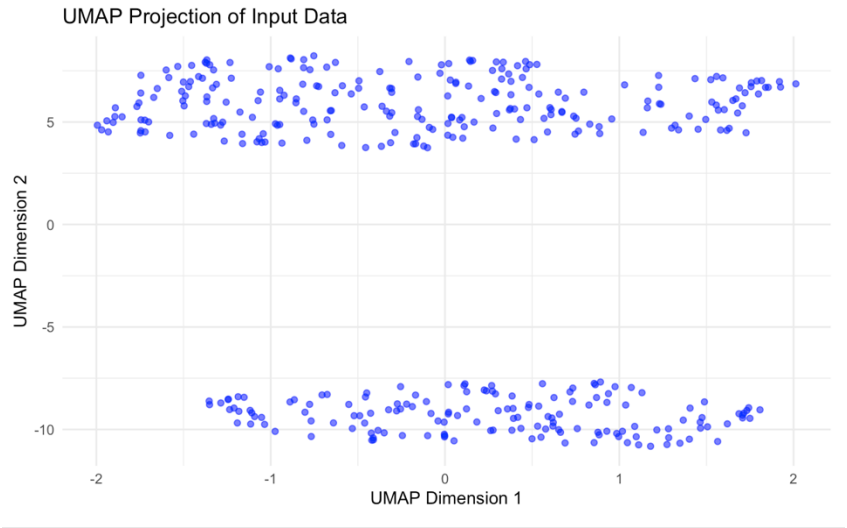
In contrast, regression models allowed me to directly predict the antenna's performance. By using PCA-derived principal components as inputs for the regression model, I was able to reduce overfitting and focus the predictive model on the most impactful components. This approach also enhanced the computational efficiency of the model, as I was no longer using the full set of original features. Although this combination of PCA and regression often worked well in capturing the antenna’s behavior, I understood that some interpretability could be lost, given that the principal components are linear combinations of the original features rather than direct parameters.

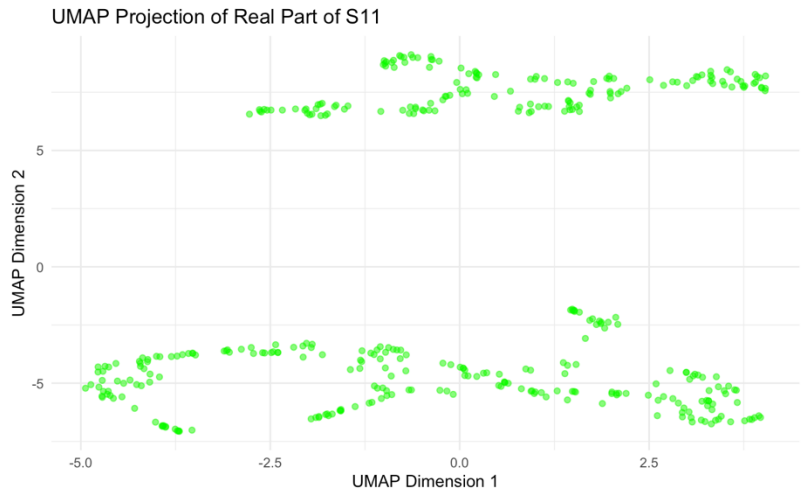
7.3.2 What are the potential limitations of these models, and how could they be improved to more accurately represent complex, nonlinear electromagnetic behavior?

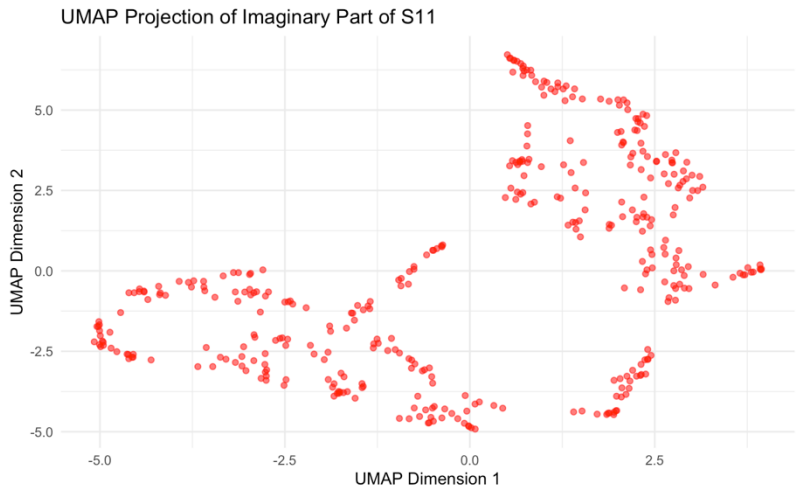
The models I employed have several potential limitations, primarily in capturing complex, nonlinear electromagnetic behavior. Both PCA and linear regression techniques assume linear relationships, which may not fully represent the intricate interactions within the antenna's geometry and its electromagnetic response. While PCA helps reduce dimensionality, it may overlook nonlinear dependencies that influence performance at different frequency ranges. Linear regression models, even when enhanced with principal components, struggle to capture these nonlinearities effectively, as they are limited to explaining relationships that can be modeled linearly.

To improve accuracy, I could incorporate advanced nonlinear modeling techniques, such as kernel-based methods or machine learning models like neural networks and support vector machines (SVMs), which are better suited for capturing nonlinear dependencies. Additionally, using polynomial regression or transforming the input features could help introduce nonlinearity into the model. Exploring these approaches could enhance the model's ability to represent the antenna's complex behaviors more effectively across various operating frequencies.

I used the UMAP (Uniform Manifold Approximation and Projection) method to analyze and visualize complex relationships within the dataset in a simplified, lower-dimensional space. UMAP is a non-linear dimensionality reduction technique, particularly effective for capturing and preserving the global structure of high-dimensional data in a compact form. Unlike linear approaches like PCA, UMAP leverages a graph-based algorithm to approximate the underlying manifold structure, which allows it to retain both local and global data relationships effectively. This approach provided a clearer visualization of patterns, clusters, and potential separations within the data, helping to reveal relationships that might be harder to detect in the original high-dimensional space. By projecting the data into two dimensions, I could visually inspect how the input parameters and S11 responses were distributed, which was valuable for understanding any patterns or similarities across different subsets of the data.

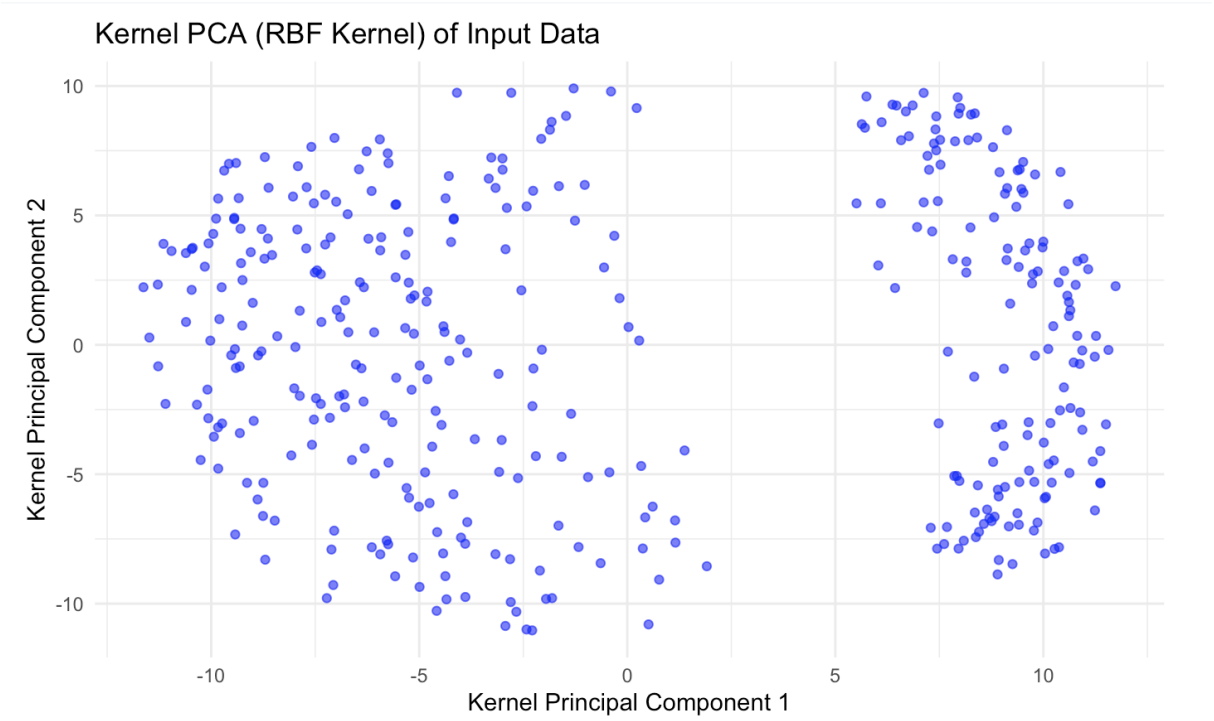


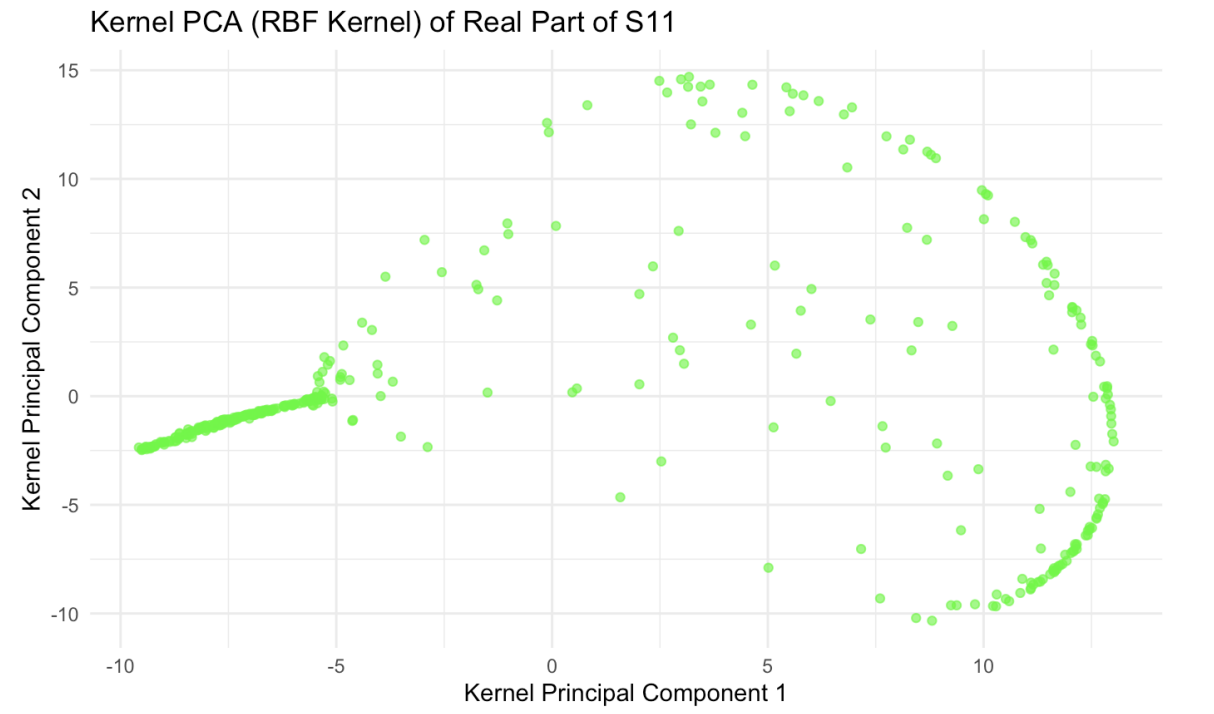


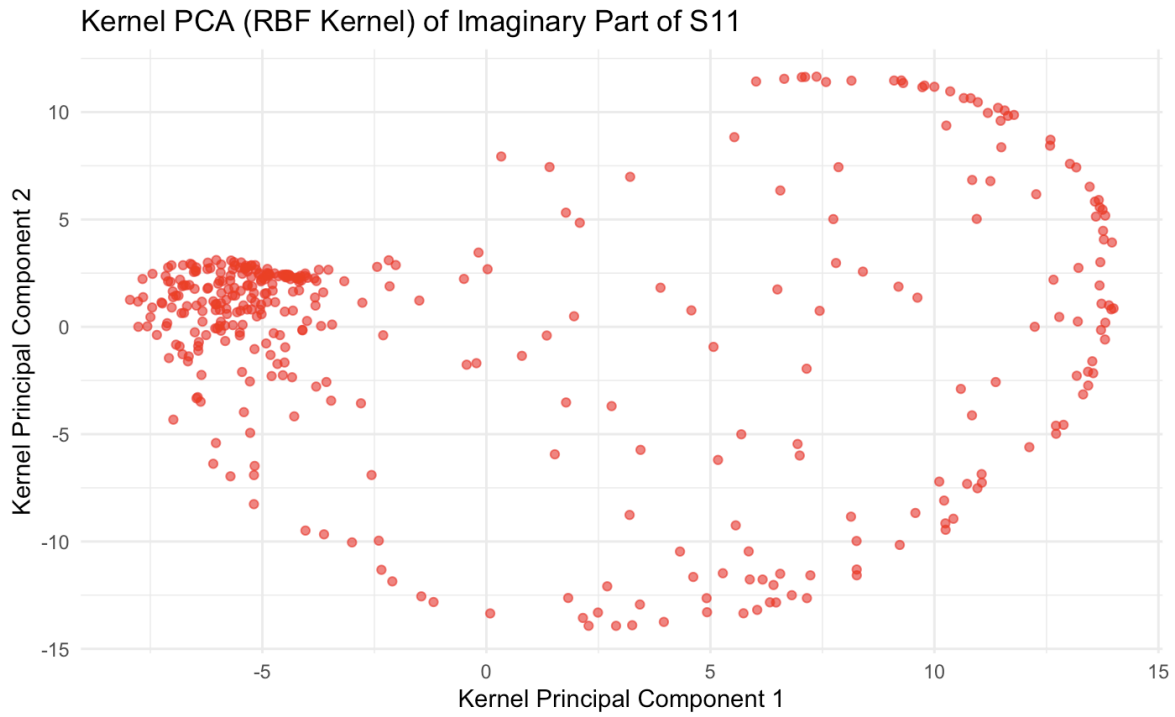


After applying UMAP to the data, I observed that if there had been a strong non-linear relationship between the input parameters and the imaginary part of S11, UMAP should have been able to capture it in the low-dimensional projection. However, the UMAP projection for the imaginary part showed a structure distinctly different from that of the input data. This led me to consider two possibilities. First, the relationship between the input parameters and the imaginary part might be non-linear but complex enough that UMAP could not fully capture it. Second, the non-linear relationship might be weak or dependent on other factors, making it less prominent in the UMAP representation. This observation suggested that more advanced modeling techniques might be needed to accurately capture the behavior of the imaginary part. For the real part of S11, the UMAP projection showed a structure that was more closely aligned with the input data's projection, suggesting a stronger and potentially more straightforward relationship between the input parameters and the real component of S11.

Secondly, after applying Kernel PCA, I used the resulting principal components as predictors in the regression model and examined the outcomes.







When I compared the Kernel PCA regression results with those from Principal Component Regression (PCR), I noticed some interesting patterns. For the real part of S11, the Kernel PCA model showed good predictive performance, with R2 values ranging from approximately 0.72 to 0.79 and RMSE values between 0.30 and 0.35. These results aligned closely with, and in some cases slightly improved upon, the PCR results. This suggested to me that Kernel PCA, with its ability to capture nonlinear relationships, had a slight advantage over PCR when dealing with the real part of S11.

However, for the imaginary part of S11, the Kernel PCA regression performed similarly to PCR, yielding low R2 values (around 0.13 to 0.22) and relatively high RMSE values (about 0.30 to 0.33). This told me that neither Kernel PCA nor PCR could effectively explain much of the variance in the imaginary component. I interpreted this as an indication that the imaginary part of S11 likely does not have strong linear or simple nonlinear relationships with the input parameters, possibly due to more complex dependencies that neither model was able to fully capture.

8. Conclusion and Inferences

In this analysis, I aimed to simplify the complexity of antenna design parameters using dimensionality reduction techniques and to develop predictive models to estimate the S11 response at specific frequency points. I applied Principal Component Analysis (PCA) to reduce the dimensionality of the dataset, identifying the main components that capture the variance in the design space. Despite PCA's general suitability for data with highly correlated features, I found that it was not as efficient for this specific dataset.

The results indicated that the first six principal components explained approximately 69.43% of the total variance, which initially suggested that PCA could be beneficial in capturing a substantial portion of the dataset's variability. However, my correlation analysis showed that most pairs of features had low or negligible correlations, with a few exceptions like the high correlation between the width of the patch and the height of the substrate. These weak correlations among most features imply limited linear dependence, which reduces PCA's effectiveness in significantly lowering dimensionality. In cases where variables are mostly uncorrelated, PCA may not provide substantial gains in variance capture with fewer components.

Additionally, the Shapiro-Wilk normality tests revealed that none of the features followed a normal distribution, which may affect the interpretability of PCA results. Although PCA does not strictly require normality, non-normal data can hinder the statistical assumptions underlying PCA, making the results less robust.

Nevertheless, The PCA approach not only made the data more manageable but also revealed underlying relationships between certain features.

I developed three models to evaluate predictive performance: a linear regression model using the 10 predictors, a Principal Component Regression (PCR) model, and a PCA-based model that used the first six principal components as predictors. I then performed cross-validation for each model to assess their generalization capabilities. When I compared models, I found that the linear regression model with these 10 predictors had a lower R2 and a higher RMSE than the Principal Component Regression (PCR) model. This result highlighted the effectiveness of dimensionality reduction in enhancing model performance.

For the real part of the S-parameters, the results were satisfactory, with the models achieving a reasonable level of accuracy. However, the results for the imaginary part were significantly

worse, indicating a potential gap in the model's ability to capture complex relationships. To address this, I applied advanced techniques like UMAP and Kernel PCA to explore potential non-linear patterns. Unfortunately, even these methods did not yield improved results for the imaginary part.

Through these analyses, I concluded that the relationship between the design parameters and the imaginary part of the S-parameters was highly complex and non-linear. Despite my efforts with advanced techniques, I was unable to develop an effective model for the imaginary part. This outcome suggested that further exploration of non-linear modeling approaches or additional feature engineering would be necessary to capture this intricate relationship more accurately.

References

1) https://www.researchgate.net/publication/371196641_Neural_Network_Modeling_of_Antennas_on_Package_for_5G_Applications
2) https://en.wikipedia.org/wiki/Principal_component_analysis
3) **Shlens, J. (3 Apr 2014). A tutorial on principal component analysis. *Google Research*. (https://doi.org/10.48550/arXiv.1404.1100)**
4) https://www.geeksforgeeks.org/principal-component-analysis-pca/
5) https://en.wikipedia.org/wiki/Kernel_principal_component_analysis
6) **Cadima, J. ve Jolliffe, İ. T. (2016 Apr 13). Principal component analysis: a review and recent developments. *Pmc Pubmed Central*.10.1098/rsta.2015.0202**
7) https://developer.ibm.com/tutorials/awb-how-to-implement-pca-in-r/
8) **An Introduction to Statistical Learning with Applications in R, Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, Springer, NY, 2013**
9) **https://builtin.com/data-science/step-step-explanation-principal-component-analysis**