IE 582

STATISTICAL LEARNING FOR DATA MINING

Homework 1

by

Hatice DOĞAN

Course Instructor: Assoc. Prof. Mustafa Gökçe BAYDOĞAN

Submitted to: Abdullah KAYACAN

Date of Submission: 15.11.2024

Department of Industrial Engineering
Boğaziçi University
Bebek, Istanbul

1. PROBLEM STATEMENT

As listed in Table 1, there are 11 parameters that are considered while designing an antenna. Each of these parameters are expected to have an effect on the S11 parameter, i.e. return loss. S11 is a measure for the power reflected back to the source when a signal is transmitted to an antenna, often measured in dB. The ideal S11 value is usually below -10 dB for an antenna, implying more efficient energy transmission (Pozar, 2012). This parameter is usually evaluated over a range of frequencies to determine the bandwidth of the antenna. Frequencies minimizing S11 indicate the frequencies at which the antenna performs optimally (Pozar, 2012).

Table 1 Design parameters and their ranges (Saçın & Durgun, 2023).

Parameter	**Range**
Patch Length (L)	1.8–5.2 mm
Patch Width (W)	1.8–5.2 mm
Probe radius	0.015–0.05 mm
Metal thickness (h_{cond})	0.01–0.04 mm
Substrate thickness (h_{sub})	0.1–0.8 mm
Solder resist thickness (h_{sr})	0.02–0.1 mm
Pad-probe radius difference	0–0.025 mm
Anti-pad-pad radius difference	0.025–0.1 mm
Probe position/patch length ratio	0.05–0.45
Substrate dielectric constant	2–5
Solder resist dielectric constant	2–5

In this homework, there are 385 simulation runs, i.e. 385 different antenna designs, in accordance with the given design parameters as the input, and the corresponding S11 values over 201 frequency levels as the output. The aim is to understand the relevance of electromagnetic behavior to geometry, and to evaluate the suitability of PCA (Principal Component Analysis) and linear regression models on reducing the complexity while analyzing such dataset.

2. SOLUTION APPROACH

2.1. Dimensionality Reduction with PCA

Given data is normalized, and principal component analysis is performed on the original space. Number of principal components (PC) is determined by plotting the explained percentage of the data by the PCs. Number of PCs is selected to be 9, as they explain approximately 94% of the data. Then, the original data is projected onto this new, *slightly* lower-dimensional space. Since there were already 11 parameters, PCA is not very successful at reducing the complexity of the design space.

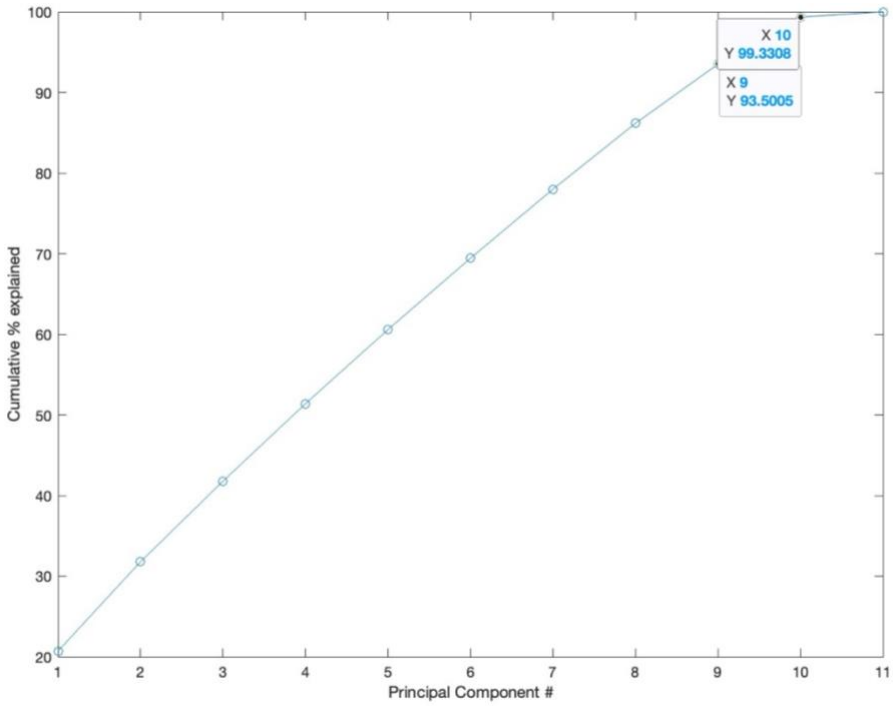

Figure 1 Cumulative % explained by the principal components.

Since there are 201 different frequencies that S11 is measured on, an elimination process is required. If at a specific frequency level, S11 values deviate significantly, it implies that the changes in the antenna's geometry significantly affect S11s. Standard deviations are calculated for both the real and imaginary parts of S11at each frequency level and represented with bar charts. Threshold values are determined with the following:

$$Threshold = k \cdot std + mean \tag{2.1}$$

where k is specified as 1.

It is observed that the imaginary parts do not experience a meaningful deviation among different frequencies. On the contrary, as examined in Figure 2, standard deviations of the real parts exceed the threshold value for frequency levels of 0-42.

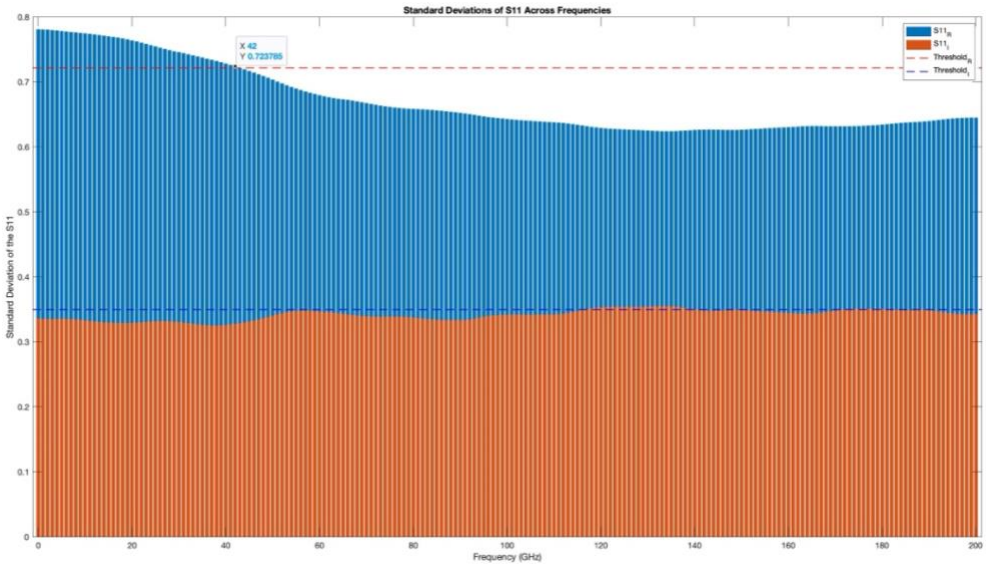

Figure 2 Standard deviation of real and imaginary parts of S11 across 201 frequencies.

The relationship between S11 values at the first 43 frequencies and the design parameters is examined by checking their correlations.

Figure 3 Correlation of S11 with the design parameters.

It is observed that S11 is highly positively correlated with design parameters 2, 4, and 10, which are the *width of patch*, *height of substrate*, and *dielectric constant of substrate*, respectively. These characteristics of the antenna design seem to have the most significant effect on the S11 parameter.

PCA was not a very useful tool throughout this process, as mentioned before. Nevertheless, a relationship between the geometric features and the 9 principal components can be drawn by examining the PC matrix.

Table 2 Principal component matrix.

Feature/PC	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
length of patch	-0.1013	0.3154	0.3696	-0.3769	0.4098	-0.0733	0.2533	0.4543	0.4062
width of patch	0.6242	-0.0097	-0.0408	-0.0396	0.0122	0.0323	-0.0004	0.0221	0.0402
height of patch	0.0711	0.0557	-0.6205	-0.3930	0.0177	-0.0697	0.4901	-0.3179	0.2166
height of substrate	0.6241	0.0188	-0.0164	-0.0765	0.0280	0.0606	-0.0028	0.0595	0.0382
height of solder resist layer	-0.0210	0.2663	-0.0864	-0.5107	-0.1246	-0.4598	-0.6472	-0.1104	-0.0402
radius of the probe	0.0260	0.0028	-0.5484	0.4544	0.2190	-0.3049	-0.2128	0.4915	0.2403
c_pad	-0.0542	-0.2660	-0.1448	-0.1775	0.5801	0.5476	-0.4300	-0.1744	0.1403
c_antipad	-0.0187	-0.5595	-0.0883	-0.3739	0.1278	-0.1721	0.1381	0.4159	-0.5486
c_probe	0.0392	0.5648	-0.0931	0.1510	0.4875	-0.0163	0.1025	-0.1293	-0.6187
dielectric constant of substrate	0.4451	-0.0128	0.2562	0.0225	0.0123	-0.0278	-0.0973	0.0532	-0.0319
dielectric constant of solder resist layer	0.0384	-0.3509	0.2553	0.1850	0.4203	-0.5917	0.0868	-0.4603	0.1456

A large positive or negative value indicates that the component strongly correlates with the feature, whereas a value close to zero indicates no correlation. Here, it is seen that none of the numbers are extremely close to 1 or -1, so the limit is set to ±0.5 to be able to make an interpretation. PC1 is positively related to the width of the patch and height of the substrate, meaning that an increase in the direction of the first principal component would indicate an increase in these two features. PC3 is negatively related to the height of the patch and radius of the probe, indicating an inverse relationship. Similar kinds of comments can be made for the other principal components.

2.2. Linear Regression Model

Again, high-deviation frequencies (0-42) were selected as they represent points of greatest sensitivity to design parameters, making them critical for understanding the relationship between x and y. The linear regression model is computed in MATLAB as:

$$y \sim 1 + x_1 + x_2 + x_3 + x_4 + x_5 + x_6 + x_7 + x_8 + x_9 + x_{10} + x_{11}$$

This model considers the effects of each design parameter on S_{11} to be equal, as all the coefficients are equal to 1. The graphs of actual and predicted values of the S_{11} parameters are plotted, both for the real and imaginary parts. It is observed from Figure 4, that the predicted values (red lines) are quite consistent with the actual values (blue lines) for $S_{11,\,Re.}$ R^2 of this model is calculated as ~ 0.85, which confirms the validity of the model. For the imaginary part however, the model is not as successful, with an R^2 value of 0.18. From Figure 5, it seems that tit predicts the directions mostly correctly, but the magnitudes seem quite different. This discrepancy likely stems from the more complex, nonlinear nature of the imaginary component or its limited impact on the overall variance.

Figure 4 Prediction of the real parts of S11 parameter for frequency levels 0-42.

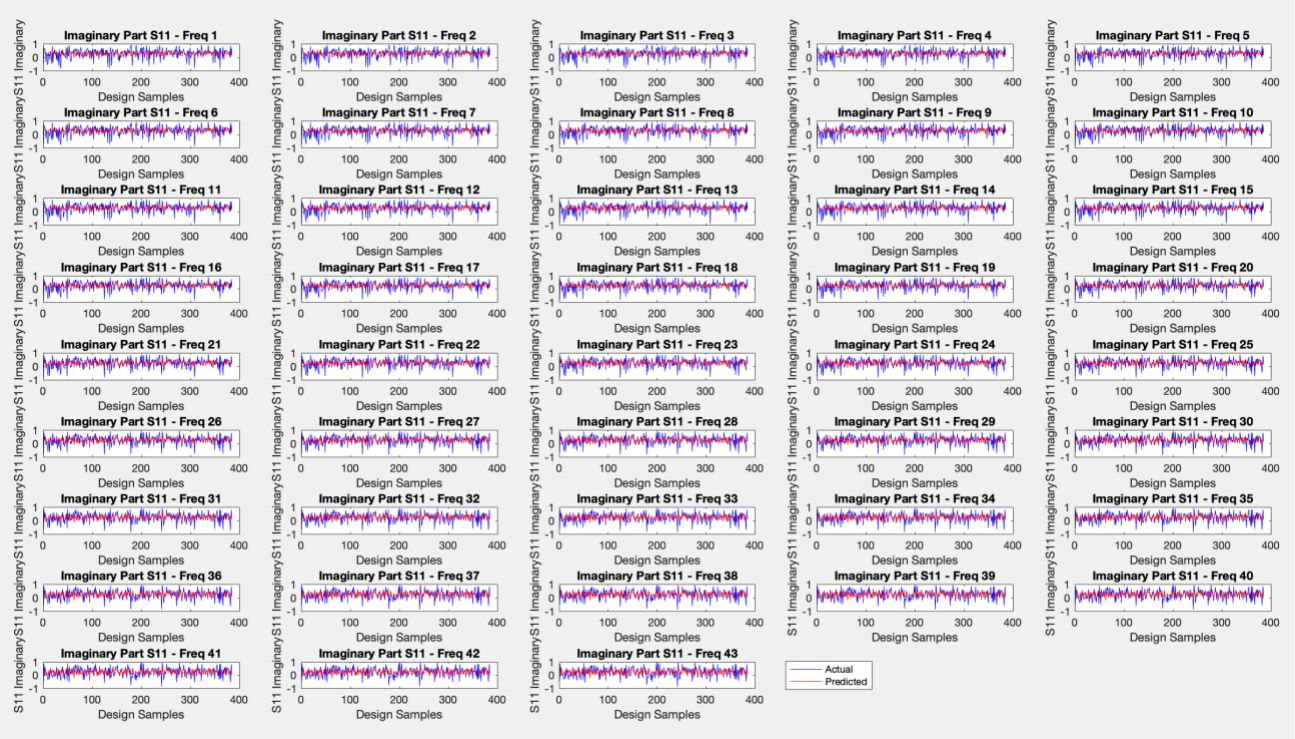

Figure 5 Prediction of the imaginary parts of S11 parameter for frequency levels 0-42.

2.3. Model Performance and Interpretability

When comparing PCA and linear regression, each has its strengths and limitations. PCA is a powerful tool for dimensionality reduction, but in this case, it didn't significantly simplify the problem due to the relatively small number of original parameters. While reducing 11 parameters to 9 principal components preserved 94% of the data's variance, it didn't notably ease the complexity of the design space. In higher-dimensional datasets, PCA might have had a greater impact. Linear regression, on the other hand, directly connects the geometric parameters to S11, making it more interpretable and actionable. PCA could still add value when combined with regression, for instance, by using the principal components as inputs to the model. Exploring nonlinear dimensionality reduction methods, like Kernel PCA, might reveal hidden patterns that linear techniques miss.

There are clear limitations with the current approach, especially with linear regression, which assumes a strictly linear relationship between inputs and outputs—an assumption that may not always capture the complex interactions inherent in electromagnetic behavior. This is particularly evident in predictions for the imaginary component of S11. Advanced models might uncover these

nonlinear relationships and improve predictive accuracy. While PCA does simplify data, its utility here was limited, and this could be addressed by incorporating domain knowledge into feature selection or by using alternative techniques that better capture nonlinear relationships and provide more meaningful dimensionality reduction.

As for the interpretability, linear regression is highly interpretable because its coefficients directly show how each geometric parameter affects S11. For example, a positive coefficient indicates that increasing the corresponding parameter increases S11, while a negative coefficient shows the opposite. This clarity makes it easy to understand the impact of parameters like patch width or substrate height on antenna performance. However, the model struggles with nonlinear relationships, especially for the imaginary component of S11. PCA, while useful for reducing dimensionality, sacrifices interpretability by transforming the original parameters into principal components, which are combinations of multiple features. Although we can identify the most influential parameters for each principal component, the results are less intuitive. Combining PCA with regression or using tools like SHAP can make these relationships clearer, even for more complex models.

Lastly, it's important to reflect on the physical context of these findings. The difficulty in modeling the imaginary components of S11 may be tied to the specific electromagnetic properties they represent, which can involve highly nonlinear and frequency-dependent behavior. Incorporating insights from electromagnetic theory into the model design could make future analyses both more grounded and more effective.

REFERENCES

1. ChatGPT. (2024, November 15). *PCA S11 Veri Analizi.* OpenAI.
2. Pozar, D. M. (2012). Microwave engineering, 4-th edition. John Wiley&Sons.
3. Saçın, E. S., & Durgun, A. C. (2023, March). Neural network modeling of antennas on package for 5G applications. In *2023 17th European Conference on Antennas and Propagation (EuCAP)* (pp. 1-5). IEEE.