Bank Transaction Analysis

Josh Meyers and Jackson Curtis

Abstract

This paper demonstrates identifying bank fraud through iterative outlier detection on a dataset of customer transactions. This paper also explores the multivariate normality of this.

1 Introduction

Financial institutions are interested in reducing costs by identifying and reducing fraud. This report analyzes a dataset containing 250,000 financial transactions with the goal of detecting fraudulent claims and assessing multivariate normality. To respect confidentiality the 20 variables in this dataset are the first 20 principle components of the original data. It is believed that fraudulent records account for less than 0.2% of the data. This dataset can then be thought of being sampled from two populations, the majority of the data from the authentic transaction population and the small minority from the fraudulent transaction population. This report describes the methods used to identify 250 fraudulent records, the process of assessing multivariate normality, and the transformations used to improve non-normality.

2 Strategy for Identifying Fraudulent Cases

Because the data has no identified examples of fraudulent transactions a few assumptions must be made about the characteristics of fraudulent transactions before attempting to identify them. We assume when looking at all 20 variables that the fraudulent transactions look different from the rest of the data, i.e., they are the outliers in the dataset. This assumption gives us a clear way to classify fraudulent transactions.

2.1 Methods

The Mahalanobis Distance ($\mathcal{D}$) can be thought as a measure similarity between transactions. The Mahalanobis Distance is a measure of how far away each transaction is from the average transaction, thus the transactions with small $\mathcal{D}'s$ are most similar to the mean and transaction with a large $\mathcal{D}'s$ are most different from the mean (see Figure 1 for a visual demonstration of the Mahalanobis Distance).The green line in Figure 2 shows the ordered distances calculated on the original data.

The problem with simply classifying the 250 transactions with the largest $\mathcal{D}$ as fraudulent is that the Mahalanobis Distance is calculated using *all* the data. The fraudulent transactions within the dataset may be skewing the calculation of $\mathcal{D}$ and consequently the 250 transactions with the largest $\mathcal{D}$ may not be the most different from the true mean of the authentic transactions. We use iterative methods to determine which transactions are farthest from the authentic mean as opposed to the mean of the mixed data set.

2.1.1 Backward Search

The first method is a type of backwards search. The premise of this method is the assumption that the transaction with the largest $\mathcal{D}$ is most likely to be fraudulent. The algorithm is as follows:

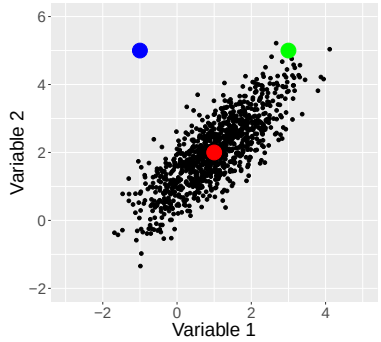

Figure 1: *Using the traditional Euclidean Distance the blue and green dots are the exact same distance (or equally similar) to the red dot. The Mahalanobis Distance, on the other hand, incorporates the correlation within the data so that the the blue dot's $\mathcal{D}$ is much larger than the green dot's $\mathcal{D}$.*

Backwards Search Algorithm:

1. Calculate μ and $\mathbf{S}$ from the entire dataset
2. Repeat the following 250 times

    a. Calculate the Mahalonobis Distance using μ (sample mean) and $\mathbf{S}$ (sample covariance matrix)
    b. Mark the transaction with the largest $\mathcal{D}$ as fraudulent and remove it from the dataset
    c. Calculate μ and $\mathbf{S}$ from the new (now smaller) dataset

Notice that each time the Mahalonobis Distance is calculated in step 2.a the inputs $(\mu, \mathbf{S})$ are different because a transaction is being removed each iteration. This allows $\mu, \mathbf{S}$ to get closer to the $\mu_{authentic}, \mathbf{S}_{authentic}$ with each iteration. This method had an overlap of 144 transactions with the 250 transactions with the highest $\mathcal{D}$ from the entire dataset.

2.1.2 Forward Search

The second method we used is a type of forward search. The premise of this method is the assumption that the transactions with the smallest $\mathcal{D}$ are most likely to be authentic. The algorithm is as follows:

Forward Search Algorithm:

1. Calculate μ and $\mathbf{S}$ from the entire dataset
2. Repeat the following until there are only 250 records remaining

    a. Calculate the Mahalonobis Distance using μ and $\mathbf{S}$
    b. Mark the $\mathbf{N}*$(loop number) transactions with the smallest $\mathcal{D}$ as authentic
    c. Calculate μ and $\mathbf{S}$ using only the authentic records

To ensure that 250 records remain that in the last iteration less than $\mathbf{N}$ records were added to the 'authentic' dataset. This method slowly builds an 'authentic' dataset by only adding transactions that are *most* like the transactions already in the dataset. This leaves the 250 last transactions as fraudulent. This method had an overlap of 123 transactions with the 250 transactions with the highest $\mathcal{D}$ from the entire dataset.

2.1.3 Clustering Search

The last method we used implemented clustering techniques. The premise of this method is the assumption that there is more than one type of authentic transaction. That is, different types of people may have different types of transactional behavior. This methods seeks to find outliers from these different groups. The algorithm is as follows:

Custering Search Algorithm: Repeat until 250 fraudulent transactions are found

1. Use K-means (with a random seed) to cluster the data into 10 clusters
2. Calculate the Mahalonobis Distance within each cluster
3. Within each cluster mark the transaction with the largest $\mathcal{D}$ as potentially fraudulent
4. When a transaction has been marked potentially fraudulent in 50 different clusters mark it as fraudulent and remove it from the dataset

We admit that some of our parameter choices are arbitrary. This method ensures that a transaction was an outlier in many different clusters before marked as fraudulent. This method only has an overlap of 34 transactions with the 250 transactions with the highest $\mathcal{D}$ from the entire dataset.

2.2 Results of Identifying Fraudulent cases

To create the final list of 250 fraudulent records, we used the combined results from the methods above. We found that the forward search and backwards search had an overlap of 229 records. Another 14 records had an overlap between either the forward search and the clustering search or the backwards search and the clustering search. The remaining 7 records were taken from transactions in the backwards search that had no overlap with the other methods. Thus we gave priority to the transactions that were marked fraudulent in more than one method.

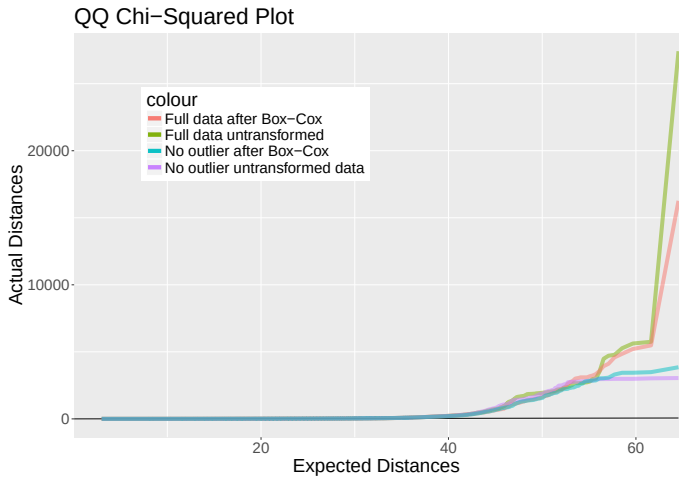

Figure 2: *If data followed a multivariate normal distribution, it would look similar to the black line at the bottom of the plot.*

3 Assessing Multivariate Normality

The outliers (presumably caused by fraudulent transactions) clearly depart from normality. For example, the seventh principle components has an observation over 97 standard deviations above the mean. Instead, we will consider the fraudulent transactions and the common transactions as coming from two different distributions, and assess whether the common transactions are distributed multivariate normal.

The properties of the multivariate normal (MVN) distribution make it easy to dismiss the raw data as clearly not distributed MVN. The raw data still has far too many extreme values, and far too much skewness, even after the 250 fraudulent cases were removed. Instead, we will explore whether transformations to the data can help the data meet normality assumptions.

The transformation we will use is the Box-Cox transformation. This transformation (know as the "power law" transformation) can be useful when data is either left-skewed or right-skewed. By raising each variable to a power (and dividing by that power), the transformation brings far away points closer to the rest. A maximum-likelihood method is used to find which power the data

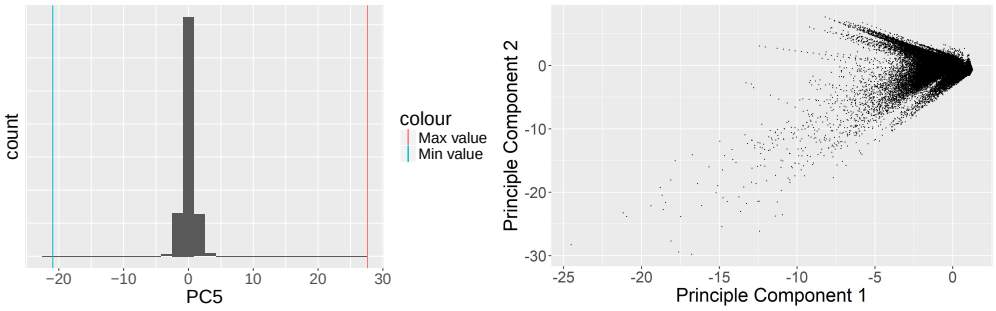

Figure 3: Extreme values and non-linear relations in transformed data

should be raised to in order to make it the most normal. This method only works for positive points, so we shift the data so all points are positive and then scale it so they are reasonably small.

Figure 2 provides an overall assessment of normality. The theory behind this plot is that the Mahalonobis Distance for a point from an MVN distribution follows a χ^2_p distribution. We can order our data by $\mathcal{D}$ and plot each of our distances against their quantile of the χ^2_p distribution. If they are MVN we expect most points to fall near the line with intercept 0 and slope 1.

From the graph, it becomes clear that the data does not follow the MVN distribution. Removing fraudulent transactions got rid of the worst offenders, but the distances are still orders of magnitude larger than we would expect under normality. The Box-Cox transformation seems to have helped very little in normalizing the data.

Digging into the data, two reasons stand out as explanations for why the Box-Cox transformation had so little effect on normality. First, Box-Cox is effective at dealing with right- or left-skew of individual variables. However, when there are extreme points on both sides (such as the histogram in Figure 3), the transformation cannot reduce the distance on one side without increasing it on the other. Thus many of our variables could not be made univariate normality.

Second, and more fundamentally problematic, is that our data exhibits non-linear relationships, something that cannot be represented by an MVN distribution. Figure 3 shows that two variables, despite having almost zero correlation, have a strong dependence on each other. In contrast, when MVN data has no correlation the data must be independent. No one-to-one transformation is going to remove the dependence between these two variables.

4 Conclusion

In conclusion, we were able to successfully identify transactions that looked most dissimilar from regular transactions. Without having any labeled cases of fraud, we cannot be sure if they were fraud, only that they failed to exhibit typical patterns in the data. If we could obtain a sample of verified cases of fraud, we could test our theory that fraud will exhibit this extreme behavior.

Upon assessing multivariate normality, we found problems with univariate normality within individual variables and with multivariate normality among variables. Thus we cannot recommend this dataset be used in any statistical analysis that requires strict normality. Strong non-normal trends in the data recommend that more robust methods be used on any future analysis.