A multivariate analysis of automobile rating data

Sai Xiao, Cheng-Han Yu

1 Introduction

In this project, we are exploring a multivariate analysis of automobile data, which can be downloaded from `https://archive.ics.uci.edu/ml/datasets/Automobile`.

This data set contains 205 observations, each observation consists of three types of entities: (a) the specification of an auto in terms of various characteristics, including 10 categorical variables (e.g. the make, fuel type, number of doors, etc.) and 14 continuous variables (e.g. length, weight, height, price, etc.). (b) its normalized losses in use as compared to other cars. It is the relative average loss payment per insured vehicle year. This value is normalized for all autos within a particular size classification (two-door small, station wagons, sports/speciality, etc.), and represents the average loss per car per year. (c) the insurance risk rating, the dependent variable (response) of this data set. The rating is the degree to which the auto is more risky than its price indicates. It ranges from -3 to 3. A value of +3 indicates that the auto is risky, -3 that it is probably pretty safe.

Before doing analysis, several steps are conducted to clean up the data. The original data set contains 205 observations, including some missing values mainly in normalized-losses variable. After removing all observations including missing values, 159 observations are retained. In addition, we change all categorial variables to dummy (0/1) variables. In this report, we focus on building a binary classifier. So the rating are converted to binary variable. 0 means the original rating values are zero or negative. 1 means the original rating are positive. In total, we have 71 responses are 0 and 88 responses are 1.

The goal of this project is to apply several multivariate analysis methods, such as principal component analysis and factor analysis to do dimension reduction, and then use linear discriminant analysis and logistic regression to do classification.

2 Exploratory Data Analysis

To explore the data, we look at the categorial variables first and briefly summarize them as follows. For make of car, Toyota, Nissan and Honda are top three brands. For fuel type, there are only 15 cars using diesel and others uses gas. 27 cars have turbo aspirations. There are 64 two-doors and 95 four-door cars. Most of cars are sedan or hatchback, and have engine type `ohc`, with 12 `ohcf` and 8 for other three types. Also, most of cars have four cylinders (136), and 14 cars have six cylinders. The main fuel systems are `2bbl` and `mpfi`.

We further look at all continuous variables. Some variables are right or positively skewed, including `normalized_losses`, `wheel_base`, `width`, `curb_weight`, `engine_size`, `horse_power`, `city_mpg`, `highway_mpg` and `price`. `stroke` is left or negtively skewed. As for linear relationship, `price` is highly correlated with `width`, `curb_weight` and `engine_size`, with correlation of 0.84, 0.89 and 0.84 respectively. The `curb_weight` refers the weight of the car with all of the standard equipment and amenities, but without any passengers, cargo or any other separately loaded items in it. The `engine_size` usually contains small engines, medium size engines, and large engines. The reasons of different size of engine include the ability to fit into a particular size engine bay, a desired level of performance, and fuel economy. It is sensible that heavier vehicles have larger engine size, and is more expensive.

It is not surprising that `city_mpg` and `highway_mpg` have a strong correlation of 0.97 and `city_mpg` is usually lower than `highway_mpg` in a stop and go traffic. `horse_power` has correlations of -0.84 and -0.83 with `city_mpg` and `highway_mpg`. Another highly correlated group are `wheel_base`, `length`, `width`, and `curb_weight`. The wheelbase is the distance between the centers of the front and rear wheels. `wheel_base`, `length` and `width` are different measures of the size of vehicles, so it makes sense they are highly correlated.

Since many of these covariates are highly correlated to each other, it may not be a good idea to run a multiple linear regression using all the variables without considering multicollinearity problems, which may lead to problems such as insignificance and unreasonable $+/-$ sign of regression coefficients.

We also observe that some variables have week correlation with response, such as `stroke`, `compression_ratio`, etc. The stroke is the length of stroke of the cylinder in a piston engine. In a piston engine, the compression ratio is the ratio between the volume of the cylinder and combustion chamber when the piston is at the bottom of its stroke, and the volume of the combustion chamber when the piston is at the top of its stroke. A high compression ratio is desirable because it allows an engine to extract more mechanical energy from a given mass of air-fuel mixture due to its higher thermal efficiency. Both stroke and compression ratio might mainly depend on the manufacture, but are not good predictors to differentiate two groups.

3 Dimension reduction

Dimension reduction is critical for our data set. Due to the EDA, we observe that a lot of variables are highly correlated, so including all of them into the classifier might introduce more noises. In this section, we mainly use principal component analysis and factor analysis to extract features which represent the original variables and are independent as much as possible.

3.1 Principle Component Analysis

Principle component analysis is used as a dimension reduction tool to find out the low dimensional representation of the original variables, which largely explains their total variation. The principal component is a linear combination of the original variables and principal components are mutually orthogonal.

Because the PCA is an analysis of eigenvectors of covariance matrix, so the covariance matrix based on dummy binary variables (converted from categorical variables) is incorrect and inaccurate. Here our PC analysis is only applied to continuous variables. The inclusion of categorical variables will be discussed in the future work.

The first step is to standardize the continuous variables. Because different variables have different units (price is in dollars and length is in centimeters) and variances of different variables differ a lot. Due to PCA is scale-invariant, the data standardization will let PCA to deal with correlation matrix rather the original covariance matrix.

The standard deviation explained by the first 5 principal components (PCs) are 2.8718, 1.4585, 1.2729, 0.94751 and 0.75617, respectively. The proportion of variances of the first 5 PCs are 0.5498, 0.1418, 0.1080, 0.05985 and 0.03812, respectively. The rest of PCs can only explain less than 3% of total variance. From the scree plot, we decide to retain the first three PCs, which can explain nearly 80% of the total variance.

The loadings of PC1, PC2 and PC3 are listed as follows,

Table 1: The loadings of first 3 principal components

	normalized losses	wheel base	length	width	height	curb weight	engine size	bore
PC1	-0.0554	-0.3157	-0.3251	-0.3141	-0.1741	-0.3412	-0.3129	-0.2584
PC2	-0.3544	0.1510	0.0602	0.0464	0.3592	0.0186	-0.0486	0.1044
PC3	-0.3644	-0.0316	0.0373	-0.1146	0.2672	-0.0499	-0.1545	0.3446
	stroke	compression ratio	horse power	peak rpm	city mpg	highway mpg	price	
PC1	-0.0957	-0.0818	-0.2898	0.1092	0.2864	0.2944	-0.3178	
PC2	0.0021	0.5008	-0.2924	-0.4786	0.2843	0.2387	0.0006	
PC3	-0.6209	-0.3691	-0.0142	-0.0630	-0.2051	-0.2079	-0.1548	

In PC1, the loadings of `peak_rpm`, `city_mpg`, `highway_mpg` are positive, while the loadings for other variables are negative. So the PC1 represents a contrast between `peak_rpm`, `city_mpg`, `highway_mpg` and other variables, especially `horse_power`, `bore`, `engine_size`, `curb_weight`, `wheel_base`, `length` and `width`. The `peak_rpm`, `city_mpg`, `highway_mpg` can be interpreted as fuel economy, while the `horse_power`, `bore`, `engine_size` can be interpreted as the engine power and `curb_weight`, `wheel_base`, `length` and `width` can be interpreted as the vehicle size.

Besides PC1, both PC2 and PC3 are hard to interpret. In PC2 and PC3, the weights of vehicle size related variables are very low, which should be mainly explained by PC1. However, some variables that have low weights in PC1 have high weights in PC2 or PC3, e.g. the weights of `normalized_losses`, `compression_ratio` and `stroke`. So these three PCs actually complement each other. Some sensible correlation can still be found in PC2 and PC3. The compression ratio and city mpg and highway mpg are positively correlated, because high compression ratio means high fueling efficiency.

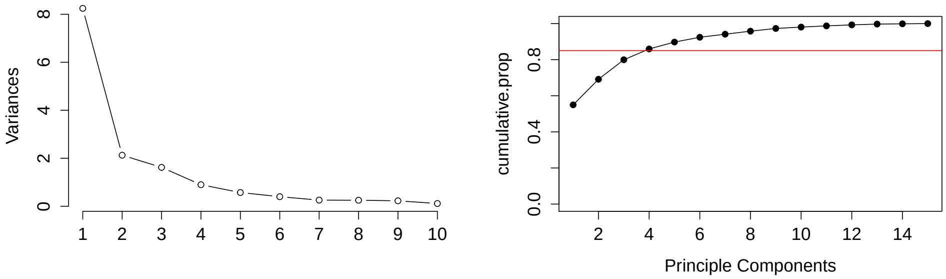

Figure 1: Left: scree plot. Right: Cumulative proportion of variances of principal components. The red line is 0.85.

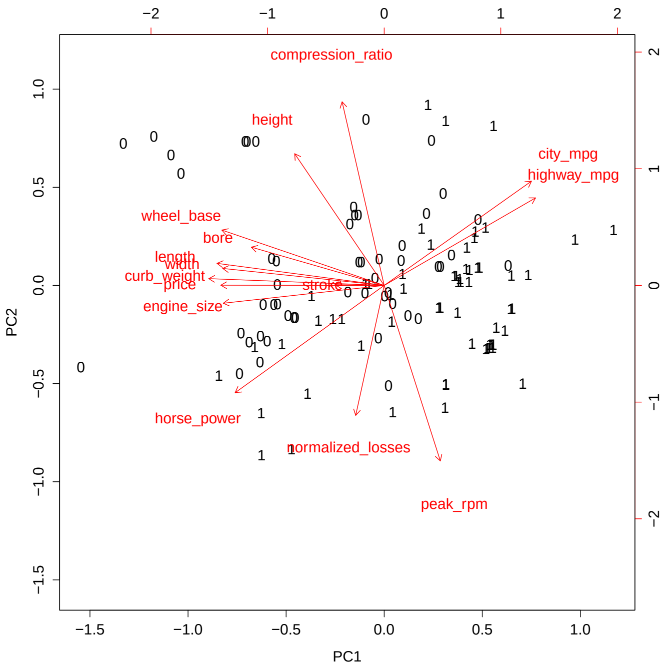

Figure 2: Biplot

From the biplot, the directions of arrows can show the relationships between the variables and the dots represent the observations. We can observe that the there are two main clusters of variables: one includes `city_mpg` and `highway_mpg`; the other cluster includes size and engine related variables, e.g. `wheel_base`, `length`, `engine_size`, etc. These two sets of variables are negative correlated with each other. Moreover, the `hight`, `compression_ratio`, `horse_power`, `normalized_losses` and `peak_rpm` are relative independent from other variables. The `horse_power` is almost perfect negatively correlated with `city_mpg` and `highway_mpg`, because usually powerful vehicle consumes more fuel. For the observations (0 refers to safe cars and 1 refers to risky cars), group 0 mainly stays around the negative values of PC1 and positive values of PC2, which means the bigger/heavier vehicles with large engine are safer. The majority of cars with high `city_mpg` and `highway_mpg` are risky. Most of cars with higher `normalized_losses` are risky, which make sense. Interestingly, the observations with higher `compression_ratio` and lower `peak_rpm` are safer, which is hard to explain.

3.2 PCA only for variables of significant predictability

As we know, PCA is an unsupervised learning method and our final goal is to combine PCA with a classification method. Before carrying out PC analysis, we want to double check if all variables have significant predictability for the response . If some variables are not significant for predicting response, it is not necessary to apply PCA based on those unrelated variables. In particular, if those insignificant variables have large covariances with other variables, then the chosen principal components might have large weight on these insignificant variables, thus using the resulting principal components as covariates would suffer the predictability of the classifier. So we want to try PCA only for those variables of significant predictability. We perform several logistic regressions on the dependent variable by one independent variable at a time. It turns out that `stroke` and `compression_ratio` have insignificant relationships with the response. So we remove these two variables and apply the PCA again. We denote the results as PCA2 and denote the results in last part as PCA in what follows. In PCA2, its PC1 explains 0.6259 of total variance, PC1-2 explains 0.7634 and PC1-3 explains 0.83731 cumulatively. We will keep these three principal components. The other details will not be discussed here. In the next section, we will incorporate the PCs to different classifiers, and compare the performance with two versions of PCs, respectively.

3.3 Factor Analysis

We are also considering (Exploratory) Factor Analysis (FA) to extract low dimensional latent variables to explain continuous variables in our data. Since `normalized_losses`, the relative average loss payment per insured vehicle year, is not directly related to the features or characteristics of a car, like `width` or `height`, it is not in the FA. The covariance matrix of the remaining 14 variables has three eigenvalues that are greater than one (8.225, 1.992, 1.458), so three latent factors are considered because the contribution on total communality K of these factors is greater than the variation of one covariate.[1]

After doing FA, we found that variable `height`, `width`, `bore`, `stroke`, `compression_ratio` and `peak_rpm` have relatively high uniqness value (`u2` ≥ 0.16), indicating that they may not be well explained and replaced by the chosen three factors. To improve the explanatory power of latent factors and the fit of factor models, we decide to remove them and do the FA again on the remaining nine variables.

For better interpretation, this time we chose two factors accoording to the eigenvalues of the covariance of the eight variables. Both PCA and MLE methods and both varimax

[1] Kaiser (1960)

and promax methods are implemented. We found that there is an interpretable factor under promax method. The results of the method of MLE with promax are shown as below.

```
Loadings:
            Factor1 Factor2
wheel_base   0.94
length       0.78
curb_weight  0.96 <----- high performance
engine_size  0.88        (big car with high quality engine)
price        0.92
horse_power  0.40   -0.53
city_mpg             1.04 <-- green
highway_mpg          0.93     (economical/fuel efficiency)
```

The first factor is mainly constructed by the variable `wheel_base`, `length`, `curb_weight`, `engine_size`, `price`, and `horse_power`. The second factor is formed by `city_mpg` and `highway_mpg`. Obviously, the first factor describes how big a car is and how powerful its engine is; the second factor focuses on if a car is energy efficient. Hence we could call the first factor **high performance** (big car with high quality engine) and the second one could be called **green** (economical and fuel efficiency). Now we can use these two factors **high performance** and **green** and corresponding factor scores to do our further analysis, especially LDA.

4 Classification

In this section, we will use linear discriminant analysis to do classification by using original variables, and compare with other classifier such as logistic regression with regularization.

4.1 LDA

The purpose of linear discriminant analysis (LDA) is to find the linear combinations of the original variables (the 15 continuous variables here) that gives the best possible separation between the groups in our data set. Here, we use Gaussian LDA, which assumes that variables follow a multivariate Gaussian distribution and covariance matrix is common for two groups. Based on this assumption, we still only introduce continuous variables to LDA model.

First of all, the data preprocessing include data standardization. Although, standardization is not necessary for LDA, however, the interpretation of coefficient is much easier after data standardization. Then we conduct the normality test of all continuous variables. As a result, we decide to remove `compression_ratio` because of its bimodal distribution. Also, we did log transformation for `normalized_loss`, `curb_weight`, `engine_size`, `horse_power`, `city_mpg`, `highway_mpg` and `price` because of their skewed distribution shapes, and experiments have shown that this transformation improves the performance in classification.

Table 2: The coefficients of variables in linear discriminant function

normalized losses	wheel base	length	width	height	curb weight	engine size
0.7476	-0.9192	-0.2474	1.0546	0.3083	-1.0628	0.0661
bore	stroke	horse power	peak rpm	city mpg	highway mpg	price
-0.0837	0.2868	0.2406	0.0935	-0.1991	0.0477	-0.4882

The coefficients of variables in the linear discriminant function are in table 2, where the weights of `normalized_losses`, `wheel_base`, `width`, `height`, `curb_weight` and `price` are relatively high. The loadings for `normalized_losses`, `width` and `hight` are positive, while those for `wheel_base`, `curb_weight` and `price` are negative. Therefore, the discriminant function represents a contrast between values of `normalized_losses`, `width` and `hight` and the values of `wheel_base`, `curb_weight` and `price`. The following plots show the distribution of discriminant function's values in two groups in training set and test test set respectively. The separation achieved by linear discriminant function is not very good because there is still some overlap between group 0 and 1, even worse in test set.

Considering the size of the whole data set is small (159 observations in total), we report the classification performance based on 20 pairs of training sets and test sets. the average accuracy rate of LDA in training sets is 84.82% and in test sets is 78.72%.

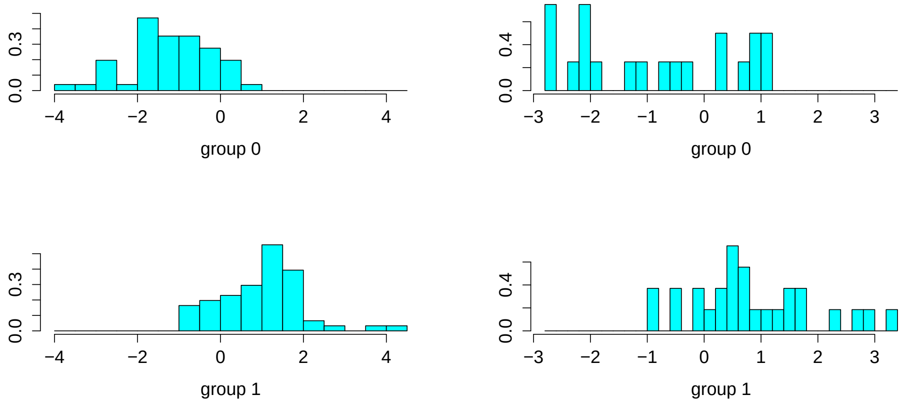

Figure 3: The stacked histogram of the discriminant function's values in training (left panel) and test (right panel) data sets.

4.2 QDA

The overlap of values of linear discriminant functions in two groups motived us to try nonlinear discriminant analysis. Quadratic Discriminant Analysis (QDA) relaxes the assumption that the covariance matrix in two groups is common, instead it allows the covariance matrix in each group to be estimated separately. In this case, the discriminant function is quadratic (nonlinear). Based on 20 pairs of training sets and test sets, the average accuracy rate of LDA in training sets is 91.03% and in test sets is 81.70%. The performance in both training and test sets are largely improved, indicating the quadratic discriminant function achieves the separation much better than linear discriminant function.

4.3 Logistic Regression with Lasso

We use the logistic regression as the classification method and apply the lasso (L1-norm) regularization to force the coefficient of some variables to be zero. This classification will automatically select a subset of variables to the logistic regression model. It can handle both continuous variables and dummy variable converted from categorical variables, however, in order to compare with LDA and QDA model, we only introduce continuous variables in this model too. After 10-fold cross validation in the training set, we choose the regularization parameter λ which produces the minimum mean cross-validated error.

Based on 20 pairs of training sets and test sets, the average accuracy rate in training sets is 88.89% and in test sets is 82.98%. Although the performance of this model in training set is less than that of QDA, however, the it performs better in test set.

The logistic regression classifier based on all continuous variables and dummy variables can achieve good accuracy rate of 99.11% in the training set, and decent accuracy rate of 89.36% in the test set.

4.4 Incorporate dimension reduction with classifiers

In this part, we will introduce the principal components/factors as predictable variables into different classifiers and compare their performances.

Our previous PCA and FA analysis is based on one training set, the PC scores and Factor scores for the test set is calculated by using the loading of PCs and Factors estimated in the training set. To keep consistent with section 3, we will only compare the classification result in the same pair of training and test set, which we used to estimate the PCs and Factors.

Firstly, we introduce three PCs and two factors into the logistic regression, LDA, QDA classifiers. The performance results are in table 3, we can see the accuracy rates in training set do not differ much. LDA with FA performs a little better than other models. The accuracy rates in test set are same for all the models, even if the false positive rates and false negative rates differ a little which are not shown here.

Table 3: Accuracy rate of different classifiers by using three PCs from section 3.1

	LR with PCA	LDA with PCA	QDA with PCA	LDA with FA
Train	83.04%	82.14%	82.14%	85.71%
Test	72.34%	72.34%	72.34%	72.34%

Then we introduce three PCs from PCA2 which are yielded after removing `stroke` and `compression_ratio`, into different classifiers. The results are in table 4. Compared with table 3, for logistic regression classifiers, the accuracy rate in training set decreases from 83.04% to 81.25%, while when generalizing the model to test set the accuracy rate largely increases from 72.34% to 78.72%. This means that as we expect, much less noises are introduced after removing `stroke` and `compression_ratio`. However, LDA with PCA2 performs even worse than LDA with PCA. QDA with PCA2 has lower accuracy rate but retain the same accuracy rate (72.34%) in test set. It seems PCA2 only improves the performance in logistic regression classifier, which might be because the selection of variables are conducted by using logistic regression classification method.

Table 4: Accuracy rate of different classifiers by using three PCs from section 3.2

	LR with PCA2	LDA with PCA2	QDA with PCA2
Train	81.25%	77.67%	81.25%
Test	78.72%	70.21%	72.34%

5 Future work

The future work includes inclusion of categorical variables in the dimension reduction part. Most of real world data including our data set are in the form of mixed data, containing both continuous and categorical variables. So far, we only applied PCA and FA to continuous variables because the main ideas of PCA and FA are both based on the eigenvectors of the covariance matrix. Here, we searched some dimension reduction methods for non-continuous variables and will try them on the vehicle rating data in the future.

A simple and quick method is to assign an order to categories if possible and treat the ordinal variables as continuous. Another approach is to use polychoric correlation matrix for the categorical and ordinal variables. A more reasonable approach is to use multiple correspondence analysis (MCA), which is usually applied to contingency table. The idea is to find scores for the row and column categories on a subset of dimensions which account for the greatest proportion of the χ^2 statistics for association between the row and column categories. That is similar with PCA, which is to maximize the variance. A more comprehensive method called Multiple Factor Analysis tailed to handle mixed data. It actually combines PCA and MCA and the tool has been implemented in the FactoMineR R package.