Empirical Report

Statistical Learning and Large Data

Matteo Orlandini[1] Riccardo Sommariva[1] Julian Tiedtke[2]

[1]PhD Economics, University of Côte d'Azur
[2]PhD Economics, Scuola Superiore Sant'Anna di Pisa

June 22, 2022

1 Introduction

Recently, there has been a lot discussions on importance of machine learning approaches in economics. For some this is one of the hottest topics. Others, argue that most of these methods are already familiar or are assumed to be of minor importance for the field. However, many economists are still unfamiliar with statistical learning and it finds only slowly it's way into the curriculum.

In this report, we discuss and apply some of the most common methods from unsupervised and supervised learning on data on the automobile market in Cyprus. In particular, we want to understand these methods and discuss its usefulness with respect to traditional methods from econometrics. The automobile market is a classical example for a product differentiated product market and therefore represents a classical topic in the economic literature.

The empirical report is organised as follows. Section 2 presents two of the most common unsupervised learning methods, i.e. Principal Component Analysis and Clustering. In section 3 we delve into supervised learning and we employ feature selection methods, like Lasso and Best Subset Selection, in order to reduce the dimensionality of the problem. Finally, section 5 introduces the issue of classification.

2 Data

We use data on the Cyprian automobile market between 1988 and 2000 that provides yearly information on units sold, retail and import prices and several product characteristics for different car models. The dataset stems from Clerides (2004), who combines several data sources: Information on every car registered within this period comes from the Cyprus Road Transport Department and information on prices and product characteristics were obtained from a local car magazine "Driver & Car". Prices on used cars are based on information provided by one of the biggest dealers of used imports.[1]

We make the following adjustments to the dataset: we use the Consumer Price Index to deflate all prices to 1995 Cyprus pound (CYP) value. Furthermore, we transform this panel dataset to cross sectional data. Additionally, following Berry et al. (1995), we assume that two observations in adjacent years represent the same car model if they have the same name, their engine size do not change by more than ten percent and are of the same import condition. Based on this definition, we have 313 distinct cars. We have a mixed data set, i.e. from our 35 variables, we have 26 continuous and 8 categorical variables.

Table (1) Descriptive statistics, 1989-2000

Year	Products	Sales	Ret. Price	Imp. Price	Japan	EU	Engine Size	SUV	Diesel	Used
1989	71	9202	14833.27	11228.87	0.56	0.37	1377.05	0.05	0.04	0
1990	71	9147	14219.82	11281.22	0.69	0.28	1406.77	0.04	0.07	0
1991	67	7503	14297.65	11946.21	0.72	0.26	1470.17	0.07	0.09	0
1992	70	8013	12838.41	11321.59	0.65	0.32	1462.87	0.05	0.06	0
1993	59	5049	12182.72	11305.78	0.54	0.40	1451.64	0.03	0.09	0
1994	65	5482	12597.34	12235.35	0.43	0.54	1545.86	0.05	0.17	0
1995	87	9678	11132.694	11083.13	0.45	0.50	1526.57	0.06	0.16	0.14
1996	96	11664	9728.49	10004.56	0.62	0.35	1585.13	0.08	0.17	0.44
1997	119	17221	8295.17	8864.42	0.76	0.23	1723.14	0.10	0.30	0.61
1998	131	21306	7788.55	8506.36	0.80	0.17	1797.93	0.16	0.32	0.69
1999	127	16647	8315.75	9213.54	0.74	0.23	1769.04	0.17	0.31	0.63
2000	131	15685	9602.12	11114.31	0.61	0.35	1755.53	0.11	0.28	0.53

Notes: The entry in each cell of the last eight columns is the sales weighted mean.

Table 1 and 2 provide some summary descriptive statistics for some selected variables. Table 1 illustrates several interesting trends. The number of products is constant in the first half of the period and nearly doubles in the second half. Sales decreased until 1994 and then rose strongly, peaking in 1998 with 21,306 sold cars. Both, import and retail prices fell from 1994 onward.

[1]Note that the sales data did not always identify the car model and that the number of models for which prices are available is much smaller than the number of models being sold. Clerides (2004) attempts to solve these issues by assigning cars to models on the basis of characteristics and by estimating a price equation with the available price data. The advantage of his approach is that it enables to include all sales but it's disadvantage is that these are not actual prices.

Evidently, these trends set in with the emergence of the used car market. Until 1995, the share of used cars was zero. Then, the share of sold used cars increased strongly to its high of 69% in 1998. This probably relates to a policy change in 1993, which relaxed the importation of used cars. Furthermore, a negative correlation between European and Japanese car imports is evident, with the Japanese cars being on a higher level. Additionally, a clear upward trend in the characteristics is observable. Apparently, this does not relate to an increase in prices, implying both: technological advances and more fierce competition.

Table (2) The range of continuous demand characteristics and associated models

	Percentile				
Variable	0	25	50	75	100
Retail price	Mitsubishi Minica	Renault Clio	Opel Astra	Rover 416	BMW 730
	1782.15	6936.69	10829.35	16617.00	100240.38
Import price	Mitsubishi Minica	Honda Accord	Renault Megane	Mazda MPV	BMW 730
	1872.14	6903.37	10485.88	15289.11	75751.66
Sales	*i.a.* Fiat Panda	*i.a.* Mazda MX6	IVECO unknown	Toyota Landcruiser	Mazda 323
	10	33	102	383	7367
Engine size	Mitsubishi Minica	*i.a.* Rover 214	*i.a.* Honda Accord	Toyota unkown	Jeep Cherokee
	657	1396	1598	1986	3960

Notes: The top entry for each cell gives the model name and the number directly below gives the value of the variable for this model. *i.a.* stands for *inter alia*. *Source*: similar to Berry et al. (1995).

Table 2 provides an indication of the range of continuous product attributes by presenting the quartiles of the distribution. Thereby, it associates some car models with the respective numbers. Evidently, there are stark differences between the car models. The BMW 730 is nearly 100 times as expensive as the Mitsubishi Minica. The Mazda 323 is the most popular product in the studied period. Engine capacity seems to be approximately normal distributed.

3 Unsupervised Learning

There is a substantial part of the literature in economics, e.g. Empirical Industrial Organizations, devoting much effort to understand the structure of demand in differentiated product markets. It is particularly interesting to understand the choice consumers make when they choose a product and which product types actually compete against each other. In other words, it is important to understand which products are similar to each other. To answer these kind of questions unsupervised learning approaches like clustering might be very useful. In clustering approaches an algorithm classifies objects into a number of groups wherein each group, objects are very similar to each other than those objects in other groups.

In this section we apply clustering in order to understand the demand structure of the automobile market in Cyprus. The automobile market is a classical example with its several product characteristics, e.g. brand, size, fuel source, etc. However, given our mixed data set we are confronted with two challenges: First, we need to pre-select the variables we include in our clustering algorithm. Some of the variables might not be meaningful since they do not contain any clustering information. Others might be multicollinear such that some characteristics are being over-weighted by the algorithm. Such superfluous variables can lead to identification problems and over-parametrization (Fop and Murphy, 2017).

Second, clustering mixed data is challenging because it is not possible to directly compute the distance between two categorical feature values. Therefore, it is challenging for such algorithms to deal with both continuous and categorical data at the same time.

We address these issues in the following way: First, we rely on Principal component analysis (PCA) in order to reduce the dimensionality and extract the meaningful informations. Second, since classical PCA cannot handle categorical variables too, we rely on Factor analysis of mixed data (FAMD) to handle our mixed dataset. Third and finally, we apply hierarchical clustering on our results from FAMD.

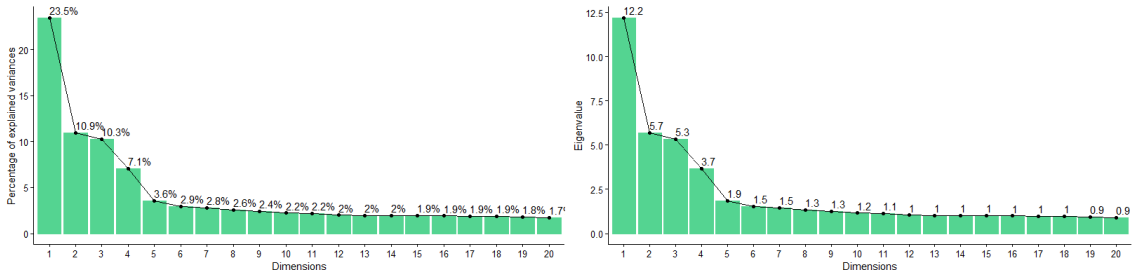
3.1 FAMD

FAMD allows to analyze the similarity between individuals by taking into account a mixed types of variables. The basic idea of the FAMD algorithm is to combine classic PCA and multiple correspondence analysis (MCA). Precisely, the continuous variables are scaled to unit variance and the categorical variables are transformed into a disjunctive data table and then scaled using the specific scaling of MCA.

The results of our FAMD algorithm are illustrated in figure 1). The scree plots show us that the first component explains around 23%, while the second and third explain around 10%. A common approach, often referred to as the elbow method, is to choose the number of principal components by eyeballing the scree plot, and looking for a point at which the proportion of variance explained by each subsequent principal component drops off. However, in our case this approach is not clear. The elbow might lie at the second or fourth component. Two alternative approaches are to collect all components that are able to explain in total 80% of the variance or all components with an eigenvalue higher than one. In both cases this holds true for 16 components.[2]

[2]Note that following this method instead of the elbow methods has in our case the disadvantage

Figure (1) Results of FAMD



(a) Scree plot: explained variance

(b) Scree plot: eigenvalues

3.2 Hierarchical Clustering

Next we apply hierarchical clustering on principal components (HCPC). Agglomerative hierarchical clustering creates a hierarchical tree called dendrogram. Here an algorithm compares iteratively each cluster, where at the beginning each observation is treated as its own cluster, based on a dissimilarity measure such that the two clusters that are most similar to each others are fused together. The algorithm continues in this fashion until only one cluster remains. (James et al., 2013, p.525).

Again, such algorithms have difficulties in comparing continuous and categorical variables. However, given our 16 principal components we do no longer need to worry about that.

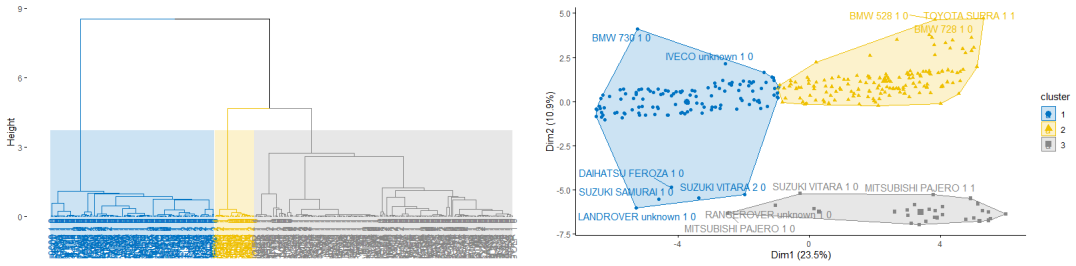
In our case we apply the HCPC algorithm in R, which is performed using the Ward's criterion on the selected principal components. Ward criterion is used in the hierarchical clustering because it is based on the multidimensional variance like principal component analysis (Kassambara, 2017).

The results of our hierarchical clustering algorithm are illustrated in figure 2). The dendogram depicts three main clusters. We can see that cluster one is completely separate from the others. This indicates that the distribution of product characteristics is substantially different with respect to the others.

As can be seen in the factor plot, this results from the explanatory power of dimension one, as group two and three can be found on the right hand side of the plot, while cluster is on the left hand side. Moreover, there is a clear shift along dimension two. This shift appears to be driven by categorical variables, particularly since the cars in the bottom of the figure are all SUVs.

that we include several components which have only minory explanatory power. Accordingly, it might be that we do not include relevant information but only white noise.

Figure (2) Results from Clustering



(a) Dendogram (b) Factor plot

Finally, table 3 illustrates some of the different product characteristic of the three proposed clusters. The average retail price is highest for cluster 1, but lowest with respect to unit sold, which is why the average sales are lower than in the other two clusters. With respect to the variables describing the power of the car, cluster 1 is the weakest cluster. However, the key difference to the other two clusters is that all cars are new cars. In contrast, almost 50% of the cars in cluster 2 are used. This probably is the reason why the retail price is lowest but unit sold is highest for cluster 2. Cluster 3 is mainly characterised by the fact that all cars are SUVs. Therefore the average retail price and the power of the cars are very high. Given these results, it appears that our algorithms is able to differentiate the most important product characteristics. Accordingly, unsupervised learning might be a powerful complementary tool for economists. In the following, we discuss one example where we believe that it could improve current econometric models.

Table (3) Summary Statistics of Clusters

	Cluster 1	Cluster 2	Cluster 3
Retail price	15593.35	10476.37	14703.30
Unit sold	82.39	105.5	102.07
Sales	6566.05	11032.24	10101.45
aloga	14.98	17.55	25.97
Engine Power	23.29	67.9	79.22
Engine Capacities	1491.761	1751.477	2580.375
Age	0	2.06	1.31
Used (in %)	0	47	34
SUV (in%)	4	0	100

3.3 Discussion

Since the seminal contributions of Bresnahan (1987), Berry (1994) and Berry et al. (1995) a large literature relies on the so-called "Random-Utility Nested Logit Model" in order to estimate demand systems in differentiated product markets. In these applications of these models to estimate own and cross-price elasticities it is fundamental to make assumptions about the nesting structure of the market. This means that substitution between products depends on the similarity of products. Accordingly, similar products are put together into groups/nests. Hereby, the model restricts correlation by allowing it within groups but not between. Hereby, we impose strong restrictions. Usually, the choice of nests is theory-driven. However, the underlying structure is often not clear and since the results of the estimation are sensitive to the choice of the nests this can be quite problematic.

Clearly, the structure presented by our dendogram relates to the nested structure of consumer's choices. Accordingly, clustering approaches might represent a data-driven alternative to the theory-driven implementation of the nesting structure. As our results show that our algorithm performs relatively well to create meaningful clusters. Additionally, such data-driven approach has the advantage that it is more sensible in constructing nests by merging several characteristics, instead of making hard discrete distinctions. Therefore, unsupervised learning might represent a useful technique in order to improve the choice of the nesting structure in discrete choice models.

4 Supervised Learning

In this section want to uncover the relationship between total sales of a given car model and its major determinants between the variable we have at our disposal.
The simple fact that we decide to impose a direction in the statistical relation between the variables in our model, shifts the analysis in the setting of supervised statistical learning.
Our main objective in this section is to find the best sparse model, or subset of variables, that predict our outcome variable *maketotalsales*.

4.1 Lasso and Ridge...

A subset of supervised dimension reduction techniques has developed in the direction of penalization algorithms that introduce a shrinkage factor in the estimation of those coefficients that contribute less to the prediction of the chosen outcome variable. Lasso and Ridge regressions belong to this family of methods.
The two methods differ in the type of penalization they impose in the estimation of the coefficients, namely, Lasso uses L1 Norm penalization and Ridge L2 Norm penalization:

$$\hat{\beta}_{Lasso} = \arg\min ||Y - X\beta||^2 + \lambda||\beta||_{(1)} \quad \wedge \quad \hat{\beta}_{Ridge} = \arg\min ||Y - X\beta||^2 + \lambda||\beta||_{(2)}$$

The result they produce is different: on one hand, Lasso drives the less significant coefficients to be exactly equal to 0, de facto reducing the dimension of the model, selecting the most relevant variables, on the other hand, Ridge only shrinks the values of those coefficient that offer little help in the prediction of Y, without setting them exactly equal to 0. The scalar λ determines the intensity of the penalization. We used both methods to determine whether we need a sparser model to predict total sales and here I report the plots that represent the evolution of the coefficients for various values of λ (Figure 3)

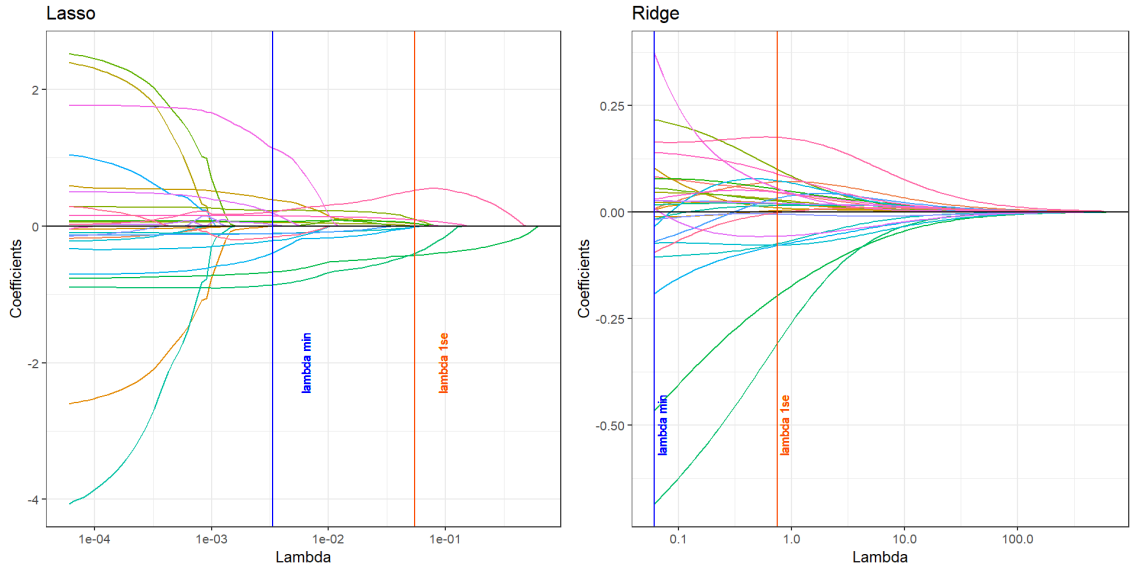


Figure (3) Lasso ($\alpha = 1$) vs Ridge ($\alpha = 0$). Notice that the scale of λ is different for the two plots.

As we can clearly see from the graphs, the Lasso coefficients exhibit a much more pronounced tendency to go towards zero, while Ridge coefficients have a much

smoother evolution towards lower and lower values.
The answer to the question: "which is the best model" is never definitive, but there is a third way, which exploits a mix between the two types of penalizations called Elastic Net.

... Elastic Net

In this case the penalization component is governed by a factor α:

$$(1-\alpha)/2||\beta||_2^2 + \alpha||\beta||_1$$

and the intuition is that it enforces an estimation procedure that at the same time shrinks the value of the coefficients while setting some of them to be exactly equal to 0. Using k-fold cross validation to find the best value of α we can recover the "best" model based on the ranking of the RMSE, which I will call *Optimal Elastic Net*:

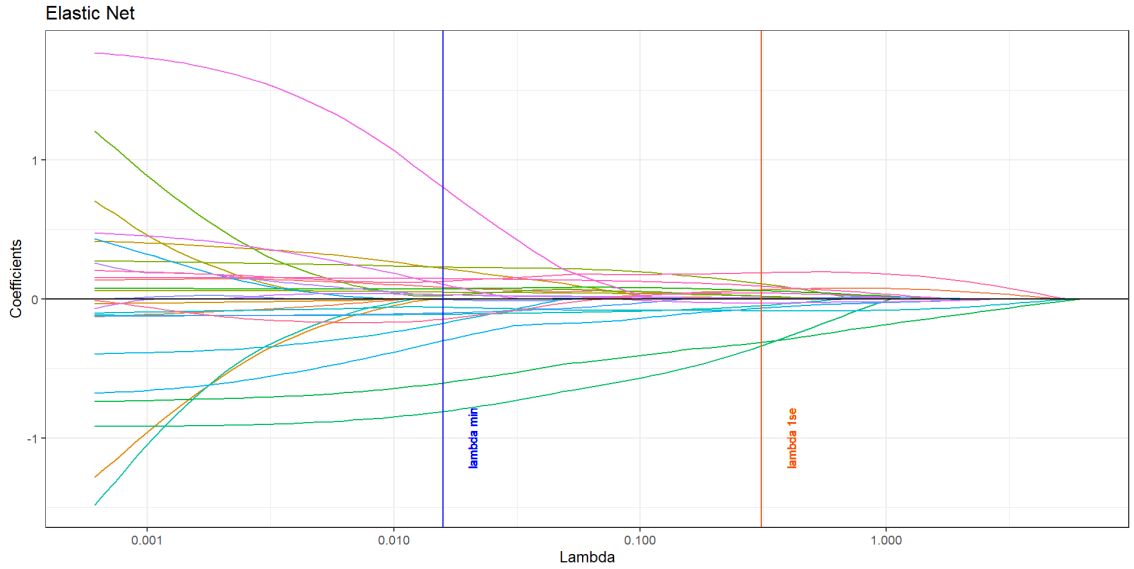


Figure (4) Optimal Elastic Net, $\alpha = 0.1$, using Cross Validation

Here I present also the crossvalidated MSE to determine to best value of λ for the three cases, i.e. the λ that minimizes the MSE:

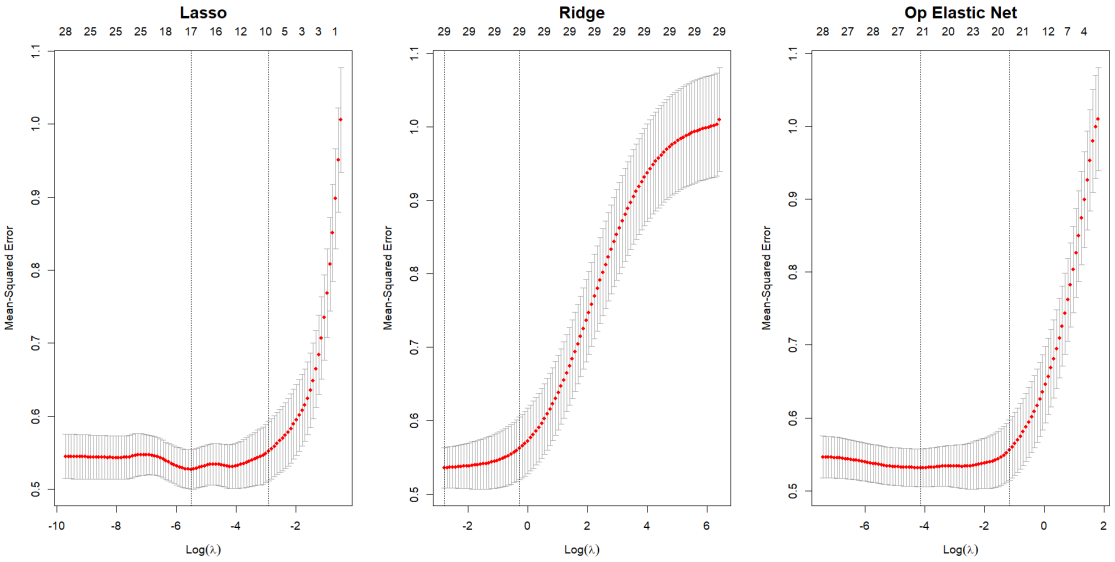


Figure (5) Lasso vs Ridge vs Optimal Elastic Net.

The result we obtain is that the "best" model is an elastic net with $\lambda = 0.1$, thus a more Ridge-like type of penalization that still retains some power to set some coefficients exactly equal to 0, operating feature selection.

Table (4) Models Performance Metrics

Statistic	Lasso	Ridge	Op. Elastic Net
RMSE	0.71	0.722	0.705
R Squared	0.509	0.493	0.512

4.2 Best subset selection

As an additional step in the pursuit of the best possible sparse model we employ another feature selection algorithm, slightly different from Lasso and Ridge regressions, that relies on the strength of brute force computation of the many possible alternative models to find the best one, Best Subset Selection. We use BSS on the model returned by the optimal elastic net.

The core idea of Best Subset Selection consists of testing all the possible combinations of the predictor variables and then selecting the best model according to some statistical criteria.

The most critical point in the evaluation statistics, however, is the fact that they are computed on training data that have been used to fit the model. This is a source

of possible overfitting and may cause poor performance when the selected model is used to predict new data. Thus we use k fold cross validation to get crossvalidated errors and we consider the subset that minimize such errors as an additional criterion in the determination of the best model.

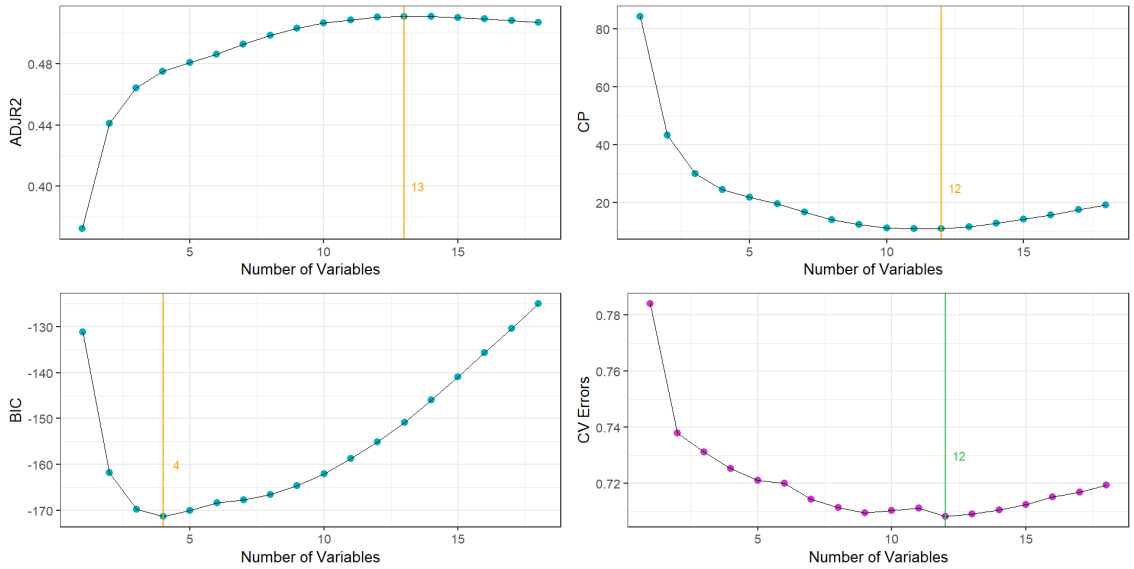


Figure (6) Selection Criteria

Figure (6) presents the 4 criteria we use to determine the best subset of predictors according to BSS algorithm. At a first look it seems that the best possible model is the one that uses a subset of 12 of the original variables that we fed in the algorithm, but, from a discussion that emerged during the presentation, one could argue that in the search for the optimal sparser model, a graphical rule of thumb of interpretation could be to select the best subset where the evolution of the statistical criterion shows a first pronounced flexion point, as a sort of "elbow" in evolution of the errors after which, the improvement of the metric related to an increase in the variables included, is marginally less significant. Following this rationale, one could argue in favor of the best 5 variables subset, as being the one where most criteria show a decise change in the steepness of the curve.
Here we plot another useful visualization tool to select the best subset depending on the statistical criterion of our choice:

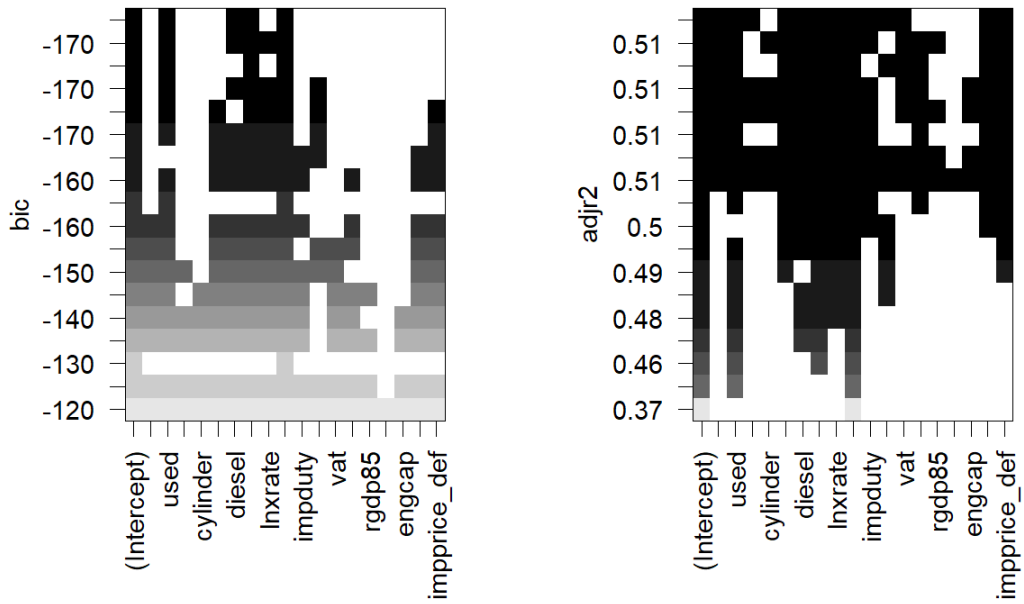


Figure (7) Selection Criteria

Finally we retrieve the best 5 variables subset and we run OLS[3] in order to get sensible estimates coefficients for the variables that should best predict the total sales in the market for cars in Cyprus. All the coefficients are statistically significant for a type I error $\alpha = 0.01$. Both the coefficients associated to the variables *used* and *diesel* are positive, meaning that used cars and diesel engines had a positive effect on the total number of sales. Being four wheel drive, an increase in the exchange rate and EU provenience of the car on the other hand, had a negative impact on the total number of sales, because the associated coefficients have a negative sign.

[3]Sort of following Belloni and Chernozhukov (2013)

Table (5)

	Dependent variable:
	makesalestotal
used	0.680^{***}
	(0.110)
diesel	0.296^{***}
	(0.104)
fourwd	-0.856^{***}
	(0.190)
lnxrate	-0.118^{**}
	(0.056)
EU	-0.564^{***}
	(0.064)
Constant	-0.204^{***}
	(0.054)
Observations	304
R^2	0.489
Adjusted R^2	0.481
Residual Std. Error	0.721 (df = 298)
F Statistic	57.128^{***} (df = 5; 298)

5 Classification

The goal of this paragraph is to come up with a model that is able to predict whether a car is brand new ("used"=0) or not ("used"=1). Since our response variable Y is a qualitative one (i.e. the fact that a car is/isn't a brand new one) we enter the realm of classification.

5.1 Logistic Regression

The first thing we try to use is a simple logistic built on 20 predictors:

```
Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -1.124e+02  7.127e+02  -0.158   0.8747
generation1     1.997e+00  9.470e-01   2.109   0.0350 *
tottaxrate      1.076e+02  6.452e+02   0.167   0.8676
cpi             1.270e+00  2.093e+00   0.606   0.5442
gdp            -2.003e-08  1.408e-08  -1.422   0.1550
unitsold        1.273e-02  6.194e-03   2.055   0.0399 *
refugee         3.090e+07  4.230e+07   0.731   0.4650
vat             9.131e+01  1.510e+03   0.060   0.9518
aloga           2.309e-01  1.939e-01   1.191   0.2338
engpow          7.051e-03  2.554e-02   0.276   0.7825
cylinder       -1.097e+00  1.198e+00  -0.916   0.3596
diesel1        -4.373e-01  1.433e+00  -0.305   0.7602
lnxrate        -6.121e+00  3.055e+01  -0.200   0.8412
EU1            -6.339e+01  1.481e+03  -0.043   0.9659
makesalestotal -2.034e-05  4.984e-05  -0.408   0.6832
impduty         3.090e+07  4.230e+07   0.731   0.4650
constax         3.090e+07  4.230e+07   0.731   0.4650
captariff       4.254e-03  2.388e-03   1.782   0.0748 .
fourwd1        -4.452e-01  2.057e+00  -0.216   0.8287
advaltax       -3.090e+07  4.230e+07  -0.731   0.4650
suv1           -1.087e+01  8.278e+00  -1.313   0.1893
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure (8) simple logistic regression

As we can see, apart from the generation of the car, the number of unit sold and "captariff" (which represents a particular tax on emissions) no variable seems to be significant. The out-of-sample confusion matrix relative to this regression is:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  1
         1  6 15

               Accuracy : 0.8833
                 95% CI : (0.7743, 0.9518)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.004021

                  Kappa : 0.7287

 Mcnemar's Test P-Value : 0.130570

            Sensitivity : 0.8636
            Specificity : 0.9375
         Pos Pred Value : 0.9744
         Neg Pred Value : 0.7143
             Prevalence : 0.7333
         Detection Rate : 0.6333
   Detection Prevalence : 0.6500
      Balanced Accuracy : 0.9006

       'Positive' Class : 0
```

Figure (9) confusion matrix of the simple logistic

Even though the number of significant regressors is small we still have a quite good result: our model is able to predict the condition of a car in the 88% of cases circa. Now we try another approach to see whether we can improve this result. Now we are going to use the so called stepwise logistic in order to try to improve the above result. Regression stepwise techniques are useful tools that allow us to come up with a restricted number of predictors. The most common approaches are 4:

1. forward iteration: we start from a model without any predictors and we choose among the possible ones (20 in our case) the one that predicts our response variable better than all the remaining one and we keep it fixed. At this point we have the first model with 1 predictor. Then, among those predictors who are left outside we choose the one that, combined with the first one we picked, gives us the best prediction of Y, we add it to the model and we keep it fixed. Now, we have our second model with two regressors. Iterating forward we replicate this procedure up to the point in where no predictor is left outside (i.e. the final model we come up with is identical to the "plain vanilla" logistic regression). Once we have our p models (where p equals the total number of available predictors) we pick the one that performs better according to some

metrics. It should be clear that this methodology suffers from the problem of multicollinearity among the predictors.

2. backward iteration: its functioning is completely symmetric to the one described above with the distinction that the model we start with is the "plain vanilla" logistic regression with all the predictors and we keep iterate backward by removing one predictor per iteration up to the point where we have just one predictor. Also this technique suffers from multicollinearity issues.

3. "brute force" best subset selection: this method is the one that is statistically more correct among the four but at the same time when the number of regressors is large it is extremely expensive in terms of both time and computation power. The alghoritm simply evaluates as many models as the number of the possible combinations of predictors available. Here multicollinearity do not matter anymore since all the possible combinations of regressors are exploited.

4. mixed stepwise logistic: this technique is a mixture of the first two methodologies. It partially takes into account multicollinearity but not as much as the "brute force" best subset selection does.

The approach we follow is the fourth:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.659e+01  2.759e+01  -0.964   0.3352
lnxrate     -2.175e+00  9.608e-01  -2.264   0.0236 *
EU1         -3.670e+01  1.859e+03  -0.020   0.9842
impduty     -2.163e+01  1.151e+01  -1.880   0.0602 .
captariff    1.757e-03  8.688e-04   2.023   0.0431 *
tottaxrate   3.716e+01  2.444e+01   1.520   0.1284
advaltax    -3.916e+01  2.635e+01  -1.486   0.1373
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure (10) stepwise logistic regression

Unfortunately, still only 3 predictors are statistically significant: "lnxrate", which is the exchange rate; "impduty" a tax on imports; "captariff" the same as before. Let's check how the model predicts by means of the confusion matrix:

```
Confusion Matrix and Statistics

          Reference
Prediction  0  1
         0 38  2
         1  6 14

               Accuracy : 0.8667
                 95% CI : (0.7541, 0.9406)
    No Information Rate : 0.7333
    P-Value [Acc > NIR] : 0.0105

                  Kappa : 0.6842

 Mcnemar's Test P-Value : 0.2888

            Sensitivity : 0.8636
            Specificity : 0.8750
         Pos Pred Value : 0.9500
         Neg Pred Value : 0.7000
             Prevalence : 0.7333
         Detection Rate : 0.6333
   Detection Prevalence : 0.6667
      Balanced Accuracy : 0.8693

       'Positive' Class : 0
```

Figure (11) confusion matrix of the stepwise logistic regression

Also the performance is worse: 86% circa of the predictions are accurate (before it was 88%). We should then stick to the classical logistic regression.

5.2 LDA and QDA

Now we move to another technique of classification to see whether it is possible to improve our preditions. We are going to use Linear Discriminant Analysis (LDA) and Quadratic Discriminant Analysis (QDA). Both LDA and QDA, instead of modelling "directly" $P_k(x) = P(Y = k|X = x)$, they try to come up with $P_k(x)$ in an "indirect" way by passing through tha Bayes' rule. Our priors are simply the fraction of observation in the training sample that belongs to a determined class. What is more difficult to handle is $f_k(x) = P(X|Y = k)$. In order to retrieve it we impose some assumptions: in LDA we assume that all the variables are sampled from a multivariate Normal with class-specific mean and common variance-covariance matrix; same assumptions hold in QDA except from the fact that here the variance-covariance matrix is class-specific. For sake of exposition (the reason will be clear in the next paragraph) we will only use as predictors those variables that were significant in the final stepwise logistic. Let's start with the confusion matrix of the LDA:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 159  57
         1  18  10

               Accuracy : 0.6926
                 95% CI : (0.6306, 0.7499)
    No Information Rate : 0.7254
    P-Value [Acc > NIR] : 0.8877

                  Kappa : 0.0581

 Mcnemar's Test P-Value : 1.145e-05

            Sensitivity : 0.8983
            Specificity : 0.1493
         Pos Pred Value : 0.7361
         Neg Pred Value : 0.3571
             Prevalence : 0.7254
         Detection Rate : 0.6516
   Detection Prevalence : 0.8852
      Balanced Accuracy : 0.5238

       'Positive' Class : 0
```

Figure (12) confusion matrix of the linear discriminant analysis

As we can notice, LDA performs very poorly compared to the previous two techniques: only 69% of the predictions where correct. This might be due to the fact that some of the assumptions involved in LDA do not hold. Let's move to QDA:

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 158   2
         1  19  65

               Accuracy : 0.9139
                 95% CI : (0.8714, 0.9459)
    No Information Rate : 0.7254
    P-Value [Acc > NIR] : 1.931e-13

                  Kappa : 0.7997

 Mcnemar's Test P-Value : 0.0004803

            Sensitivity : 0.8927
            Specificity : 0.9701
         Pos Pred Value : 0.9875
         Neg Pred Value : 0.7738
             Prevalence : 0.7254
         Detection Rate : 0.6475
   Detection Prevalence : 0.6557
      Balanced Accuracy : 0.9314

       'Positive' Class : 0
```

Figure (13) confusion matrix of the quadratic discriminant analysis

QDA tremendously overperform LDA: 91% circa of the in-sample predictions are accurate.

5.3 Sample Rebalancing

In our dataset brand new cars ("used"=0) represent the 73% circa of the cases and the remaining ones are used ones.

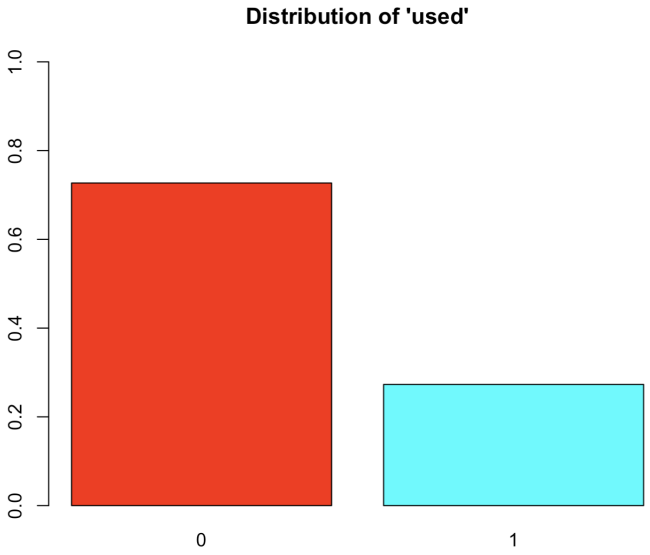


Figure (14) frequency of new ("used"=0) vs used ("used"=1) cars

Hence, when we develop a prediction model this would be affected by the issue of having a class which is over-represented and another one which is under-represented. In the statistical jargon, problems as this one goes under the name of unbalanced sample. Put it simple, no matter how we build a model, if we start from such a sample our predictions will be accurate if we are interested in recognising when a car is new whereas it will perform worse when we are interested in predicting whether a car is used. To visualize the problem consider the following picture representing a confusion matrix:

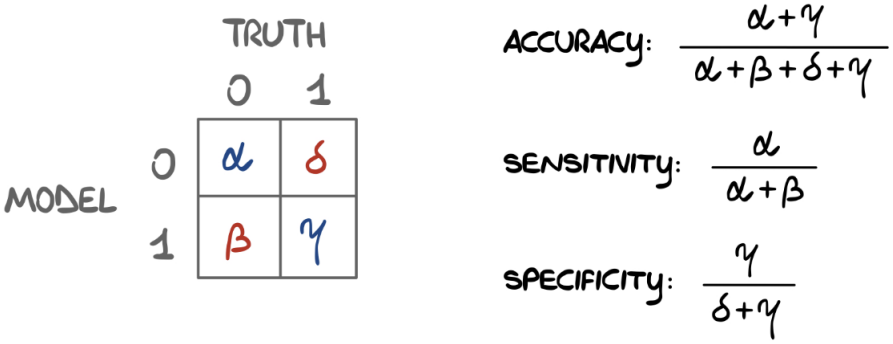


Figure (15) meaning of the confusion matrix

In our case, we face the problem of low specificity: the relative high number of new car w.r.t. used ones makes the model "well-trained" for predictions of "used"=0, and poorly trained for predictions of "used=1". In order to see the issue let's represent our dataset into a xyz-plane where green points correspond to a new car and pink

points corresponds to a used car (at this point it should be clear that we only choose 3 predictors from the previous paragraph for reasons of viewability).

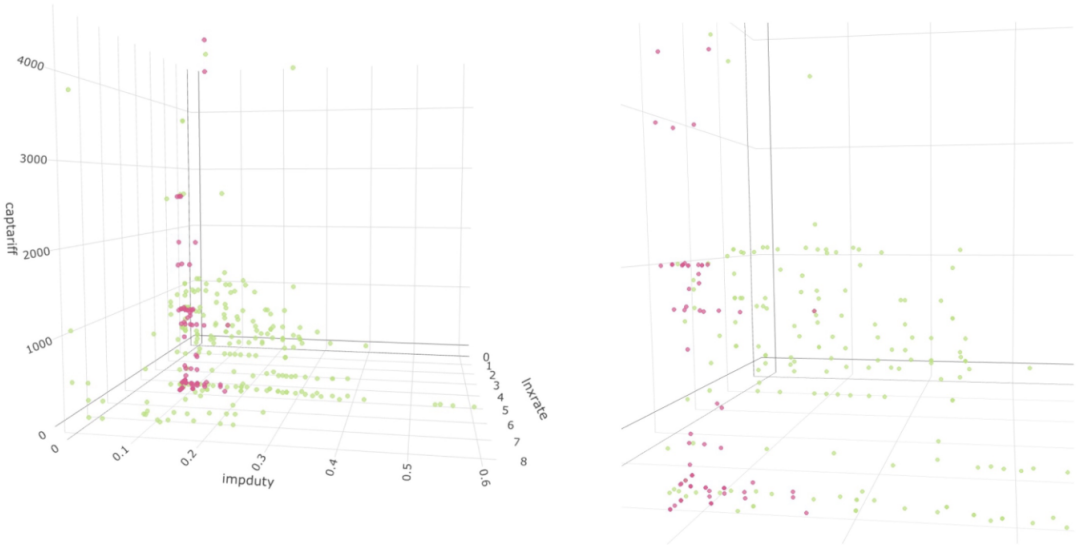


Figure (16) data distribution (left) and zoom-in on pink points (right)

We then tackle the problem in 4 different ways:

- Oversampling: we over-sample (in a bootstrap fashion) the under-represented class up to the point in where we get an equal number of variables in the two classes.
- Undersampling: we under-sample the over-represented class up to the point in where we get an equal number of variables in the two classes.
- Convex rebalancing: we artificially create new data points which are convex combinations of points labelled as "used"=1 up to the point in where we get an equal number of variables in the two classes. This procedure makes the resulting cloud of points more dense.
- Gaussian rebalancing: we artificially create new data points in the gaussian neighbourhood of each original point labelled as "used=1". This procedure makes the resulting cloud of points more sparse.

In order to visualize how the last two techniques work, consider the following picture:

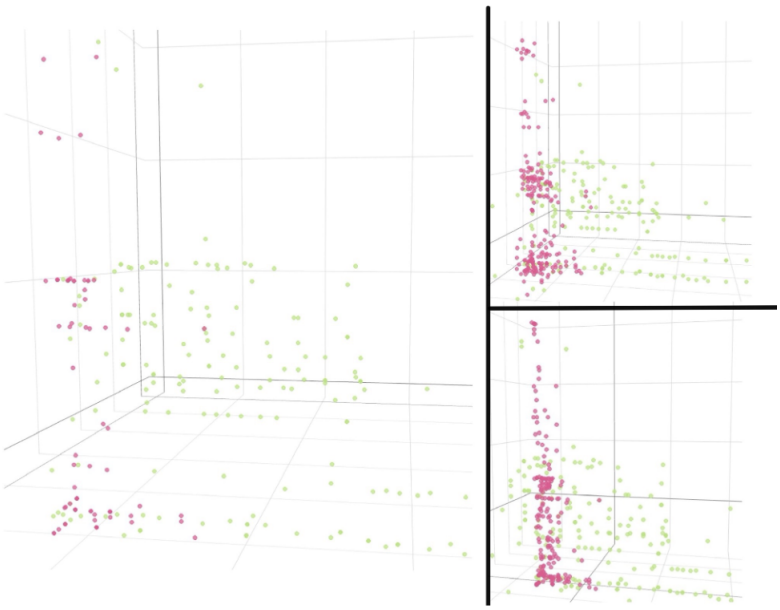


Figure (17) zoom of the original plot (left), zoom of the gaussian rebalancing (up-right), zoom of the convex rebalancing (down-right)

As we can notice, the cloud of pink points becomes more sparse and more dense respectively in the case three and case four. Finally we present a comparison of confusion matrices relative to LDA:

	Standard	Oversampling	Undersampling	Gauss. Rebal.	Conv. Rebal.
Accuracy	0.6926	0.8418	0.8284	0.8371	0.8371
Sensitivity	0.8983	0.6893	0.6866	0.7014	0.6923
Specificity	0.1493	0.9944	0.9701	0.9729	0.9819

Figure (18) comparison of the various techniques of sample rebalancing

Each method increments the capability of the model to predict correctly the class: this is due to the fact that specificity hugely increases.

References

Belloni, Alexandre and Victor Chernozhukov (2013) "Least squares after model selection in high-dimensional sparse models," *Bernoulli*, 19 (2), 521 – 547, 10.3150/11-BEJ410.

Berry, Steven, James Levinsohn, and Ariel Pakes (1995) "Automobile Prices in Market Equilibrium," *Econometrica*, 63 (4), 841–890, http://www.jstor.org/stable/2171802.

Berry, Steven T. (1994) "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25 (2), 242, https://doi.org/10.2307/2555829.

Bresnahan, Timothy F. (1987) "Competition and Collusion in the American Automobile Industry: The 1955 Price War," *The Journal of Industrial Economics*, 35 (4), 457, https://doi.org/10.2307/2098583.

Clerides, Sofronis (2004) "Gains from Trade in Used Goods: Evidence from the Global Market for Automobiles," University of Cyprus Working Papers in Economics 6-2004, https://ideas.repec.org/p/ucy/cypeua/6-2004.html.

Fop, Michael and Thomas Murphy (2017) "Variable Selection Methods for Model-based Clustering," *Statistics Surveys*, 12, 10.1214/18-SS119.

Hollander, M., D.A. Wolfe, and E. Chicken (2013) *Nonparametric Statistical Methods*, Wiley Series in Probability and Statistics: Wiley, https://books.google.it/books?id=-V7jAQAAQBAJ.

James, Gareth, Daniela Witten, Trevor Hastie, and Robert Tibshirani (2013) *An Introduction to Statistical Learning: with Applications in R*: Springer, https://faculty.marshall.usc.edu/gareth-james/ISL/.

Kassambara, Alboukadel (2017) *Practical Guide To Principal Component Methods in R: PCA, M (CA), FAMD, MFA, HCPC, factoextra*, 2: STHDA.