Analysis of horse's physical condition with multivariate methods

— STAT 456 final project

Cecily Liu
qliu273@wisc.edu

University of Wisconsin-Madison

Abstract

Applied statistical methods plays an essential role in many aspects, multivariate analysis is a effective can extract more information from data and visualize and analyze it. In this paper, I describe multiple multivariate ways and use them on a horse colic dataset to find out is any relationship between variables or observations and which aspects can have influence on horse's outcome(treated surgery or not, died or alive).

First I use some regression methods to deal with the missing values, and draw some plots to check the distribution of some continuous variables and visualize the relationship between variables.

Then, since there are many attributes in this dataset and there exists collinearity among attributes. I use PCA to reduce dimensionality and eliminate the effects of collinearity, which help us better analyze the relationship between variables and observations.

Next, instead of consider interrelationship within a set of variables, I'm interested in the relationships between two sets of variables. Thus I use canonical correlation analysis to analyze the connection between horse's dynamic physical variables and stable physical variables.

Furthermore, I want to more concentrate on the purpose that obtain a low di- mensional map of the data that can represent the relationship between observations. That brings in multidimensional scaling(MDS).This method has some advantages over PCA and CCA since it focus on the distance or dissimilarities among observa- tions, which has more clear representation according to our goal. We still use the five continuous variables on physical condition of horse and we can divide our whole dataset into sub-groups. We can see these groups have differences on other categori- cal data (such as age young or adult). Then in order to check whether it is correct to divide dataset into sub-groups according to these five continuous variables, I use K- means clustering and sample cluster into sub-groups based on 5 continuous variables, which verified the effectiveness of our previous methods.

Finally, I'm still interested other categorical variables and how they can affect a horse's outcome(surgery or not, died or alive). I use correspondence analysis to analyze the relationship between categorical variables, I draw mosaic plot and CA plot to get my conclusions.

keywords: visualization, principal component analysis, canonical correlation analysis, multidimensional scaling, K-means clustering, correspondence analysis

Contents

1 Introduction

With the development of the applied statistics, statistical and data-driven approaches are play an important role in multiple aspects. Multivariate analysis is a necessary field in applied statistics. In this paper, we will use several multivariate analysis methods to extract information of a horse physical condition and treatment dataset, visualizing the relationship and conclude results. This is a dataset about **Horse Colic** information, where I get from UC Irvine Machine Learning Repository data base**[1]**. It is a well documented data, this dataset is about three hundred horse and their information about their physical condition. This dataset allow us to explore whether these conditions has any connection with horse's surgeries and whether there euthanized. Next I will briefly introduce and depict this dataset.

This dataset is of 368 instances(300 training sample and 68 testing sample) with 28 attributes(which include continuous variables, discrete variables and nomial variables). There are about 30% missing values in this dataset. Since some variables are useless when we analyze the data(such as ID number), and some variables are of meaningless for our mul- tivariate analysis. I select 16 variables that depict the data. Some information about the 16 variables are shown below:

1. surgery.1 = Yes, it had surgery. 2 = It was treated without surgery.

2. Age .1 = Adult horse ,2 = Young (<= 6 months)

3. Rectal temperature . linear in degrees celsius. This parameter will usually change as the problem progresses, eg. may start out normal, then become elevated because of the lesion, passing back through the normal range as the horse goes into shock.

4. Pulse.The heart rate in beats per minute.It is a reflection of the heart condition: 30 -40 is normal for adults.It's rare to have a lower than normal rate although athletic horses may have a rate of 20-25.Animals with painful lesions or suffering from circulatory shock may have an elevated heart rate.

5. Respiratory rate.Normal rate is 8 to 10.Its usefulness is doubtful due to the great fluctuations.

6. Temperature of extremities.It's a subjective indication of peripheral circulation.Possible values: 1 = Normal, 2 = Warm, 3 = Cool , 4 = Cold .(Cool to cold extremities indi- cate possible shock, hot extremities should correlate with an elevated rectal temp.)

7. Peripheral pulse.1 = normal 2 = increased 3 = reduced 4 = absent .Normal or in- creased p.p. are indicative of adequate circulation while reduced or absent indicate poor perfusion.

8. Capillary refill time.A clinical judgement. The longer the refill, the poorer the circulation.1 : < 3 seconds .2 : >= 3 seconds.

9. Pain - a subjective judgement of the horse's pain level. The level range is 1 5, the pain's level increases when number increase. In general, the more painful, the more likely it is to require surgery. Prior treatment of pain may mask the pain level to some extent.

10. Peristalsis.An indication of the activity in the horse's gut. As the gut becomes more distended or the horse becomes more toxic, the activity decreases, the number increases. (level 1 $\sim$ 4).

11. Abdominal distension. The distension is more severe as number increases(range from 1 to 4).An animal with abdominal distension is likely to be painful and have reduced gut motility. A horse with severe abdominal distension is likely to require surgery just to relieve the pressure.

12. Packed cell volume. Linear.The amount of red cells by volume in the blood - normal range is 30 to 50. The level rises as the circulation becomes compromised or as the animal becomes dehydrated.

13. Total protein. Linear. Normal values lie in the 6-7.5 (gms/dL) range.The higher the value the greater the dehydration

14. Outcome.The final result of a horse.1 = lived, 2 = died ,3 = was euthanized.

15. Surgical lesion. Whether a horse has surgical harmful effect.1=yes.2=no.

16. Is pathology data present for this case? 1 = Yes 2 = No.

2 Goals

This dataset is not so large and with appropriate attributes that interest me how these attributes of a horse's physical condition can have effects on whether horse are sick or not(or have been operated surgery or not). And how the horse's different attributes are correlated to each other. To be more specifically, it can be expressed into some interesting questions.

1. How are these 16 measurements related? Does a large value for one variable tend to occur with large value for another variable? e.g: The rectal temperature and the temperature of extremities it is said to increase when the horse has the lesion according to background information. Will the data shows concordant with this information or contradict it?

2. Can I found some correlation between two variables that shows completely no common according to the background information.e.g: Such as refill time and pain level.

3. Do the survived horse and died horse(include them was euthanized) have statistically significant differences in their values of the variables?

4. Do the survived horse and died horse(include them was euthanized) show similar amount of variation for the variables.

5. Can we find some combinations of variables that maximize the variation of variables, and thus we can find a few of these components to depict the variation of whole dataset?

6. If the survived horse and died horse has difference in the distributions of variables, can we find some rules or construct some functions that separates these horse into different variables through the distribution of variables.

7. If we can get the distance or dissimilarities between the observations of these data. Can we find some way to depict these relationship with some 2-dimensional plots?

3 Main part

This is the main body of my report, I'll represent how I deal with the dataset and achieve my goals in detail. The concrete process will be shown step by step.

3.1 Preprocessing

Since I don't need to train the model with training sample and then testing it. I combine the train dataset and test dataset and get 368 observations. Then I found there are a bunch of missing values in some variable. Here is what I deal with missing values:

Since there are 368 observations.

1. First I select the variables whose missing value is less than 52 (1/7 of the whole sample). These missing values has less than 1/7 of the total observations, thus I thought they can be filled through the information of the non-missing values for these variables. Then I got 1:surgery,4:pulse, 8:capillary refill time, 10:peristalsis. 12:packed cell volume, 13: total protein, 14:outcome. Then I divided them into discrete variable and continuous variable. For each of the discrete variables, I found the most frequent occurring non-missing value of this variable among the sample and use them to fill the missing values of this variable. For each of the continuous variables, I replace the missing value with the mean of non-missing value of the variable.

2. Then I select the variables whose missing value is larger than 52 (1/7 of the whole sample). These missing values has larger than 1/7 of the total observations, thus I thought they need to be filled through the information of the non-missing values of other variables, which can be meet through a regression by treating missing values as response variable. I got 3:rectal temperature,5 Respiratory rate, 6: temperature of extremities, 7: peripheral pulse.9: pain - a subjective judgement of the horses pain level.11: abdominal distension. Then I divided them into discrete variable and continuous variable. For each of the discrete variables, I treat them as response

variable and fit logistic regression with other variable as explanatory variables, and predict the missing value.For each of the continue variables, I treat them as response variable and fit multiple linear regression with other variable as explanatory variables, and predict the missing value.

3.2 Exploratory data analysis and visualization

As an old saying goes according to Chambers, Cleveland, Kleiner, and Tukey (1983) **[2]**, there is no statistical tool that is as powerful as a well-chosen graph. Certainly graphical presentation has a number of advantages over tabular displays of numerical results, not least in creating interest and attracting the attention of the viewer. The graph can more directly shows the relationship between the variables or the data, it deliver massage more quickly can numbers or equations. So visualization plays an important role in our data analysis.

3.2.1 Normal quantile plot

Since we do not know the distribution of this data, it is important to find its distribution(or asymptotic distribution in most cases), a wise choice is normal quantile plots. I randomly select 4 continuous variables: Rectal temperature, pulse, respiratory rate and packed cell volume. Here is the quantile plots at figure 1:

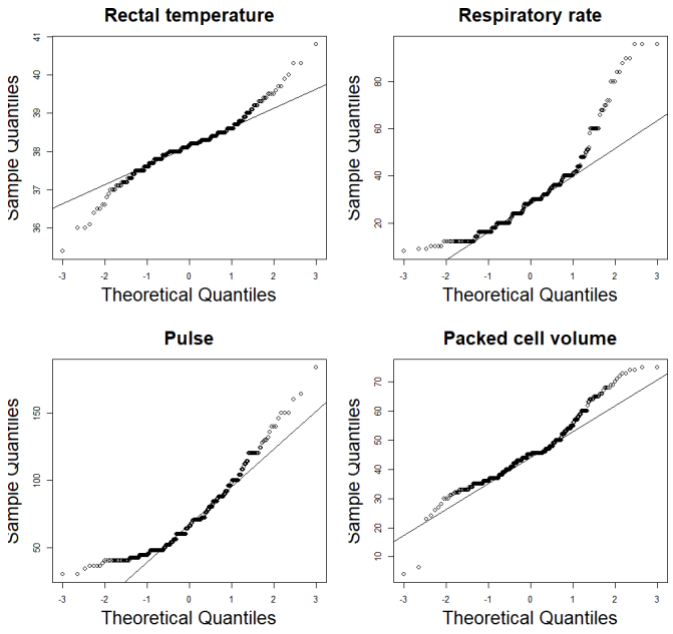

Figure 1: QQplots

As we can see, the data of continuous variables can be treated as normal diostribution,

this is kind of important in our following analysis(such as principle components analysis or the calculation of Mahalanobis distance between observations).

3.2.2 bivariate boxplot

Consider overall disribution of data, the first plot I come up with is scatter plot, however, in many actual cases, it might be helpful to have a more formal and objective method for labelling observations as outliers, and such a method is provided by the bivariate box- plot, which is a two-dimensional analogue of the boxplot for univariate data proposed by Goldberg and Iglewicz. I select two continuous variables pulse(The heart rate in beats per minute) and respiratory rate of every horse. According to the background information, animals with painful lesions or suffering from circulatory shock may have an elevated heart rate, thus I thought it might be helpful to identify the outliers(higher pulse and respiratory horse) for the convenience of the analysis.

Since the data is asymptotically normal, the non-robust bivariate boxplot is reliable to provide some trustworthy information. As for the outliers, I select 8 horses with the highest pulse and 8 horses with the highest respiratory rate and combine them, usually there are a few horses has the highest pulse meanwhile the highest respiratory rate, so the total number of outliers might be lower than 8 + 8 = 16. The figure is shown below at figure 2:

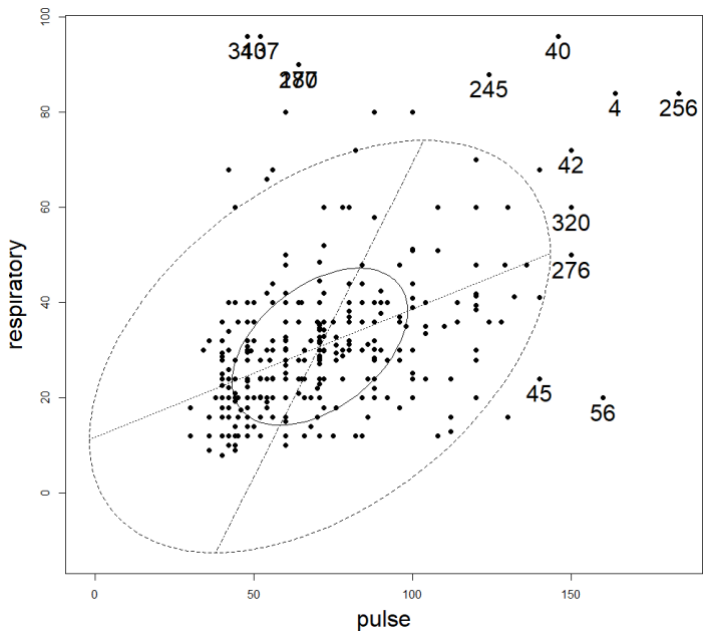

Figure 2: bvbox plot

As we can see from the picture above, there are aproximately 50% of data are in the

inner solid ellipse hinge. And the points out of outer dashed ellipse fence are potentially outliers. But not all the points out of ”fence” are labeled points (the points with the indexes). We can simply got the horse 4, 40 42 45 56, 107 ,187, 245,256,270, 276, 320, 343 are potential outliers with high pulse and high respiratory rate. Also, we can see from the figure that the pulse and respiratory of horses has positive correlation. We will detailedly focus on it in the following analysis.

3.2.3 scatterplot matrix 1:boxplot and linear regerssion line

Since there are several variables in the horse colic data, which between them generate several plots if we want to compare multivariate variables. But just making the graphs without any coordination will often result in a confusing collection of graphs that are hard to integrate visually. Consequently, it is very important that the separate plots be presented in the best way to aid overall comprehension of the data.

The scatterplot matrix is intended to accomplish this objective. I select four continuous variables that mentioned before and want to check whether they have somecorrelations.

I set diagonal panels of the plot matrix as boxplot, thus we can clearly see the quantiles of each variables and whether they have outliers. The other panels I draw the scatter plot of the each pair of variables and add a linear regression line on it. The plot is shown below at figure 3:

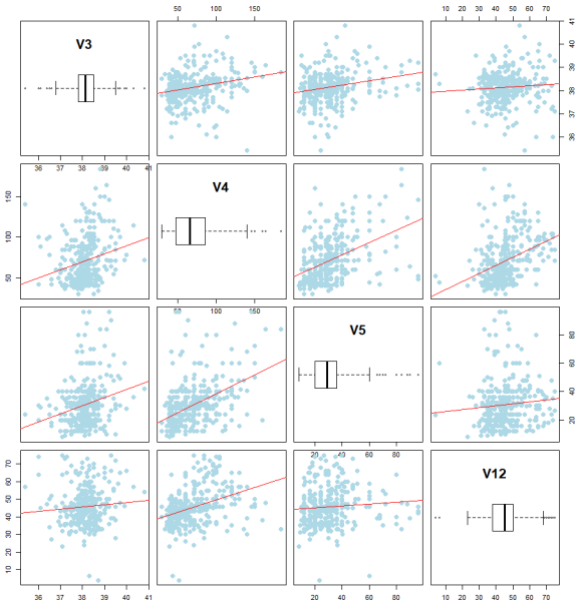

Figure 3: boxplot and lm

As shown in the plot, the V3,V4, V5, V12 represents rectal temperature, pulse, respi- ratory rate and packed cell volume. We can conclude that:

- The median of rectal temperature is around 38.3, there is a few outliers horses whose temperature is higher than 40 or lower than 36.5.
- The median of pulse is around 66, there is a few outliers horses whose pulse is higher than 140.
- The median of respiratory rate is around 30, there is a few outliers horses whose respiratory is higher than 60.
- The median of packed cell volume is around 45, there is a few outliers horses whose packed cell volume is higher than 65 or lower than 23.
- There is positive correlation between each pair of variable, which will be explored more precisely later.

3.2.4 scatterplot matrix 2:histogram and scatterplot with lowess regression line

Considering the boxplot may not represent the overall distribution of each variables, I change the diagonal panel's plot to histograms, and set the regression method on the scatter to lowess regression to check whether we can find some more information. Since the scatter plot and smooth line of each pair of variables are the same in the upper and lower panels, I put the correlation coecient in the lower panels to see it more intuitively. As we can see below at figure 4:

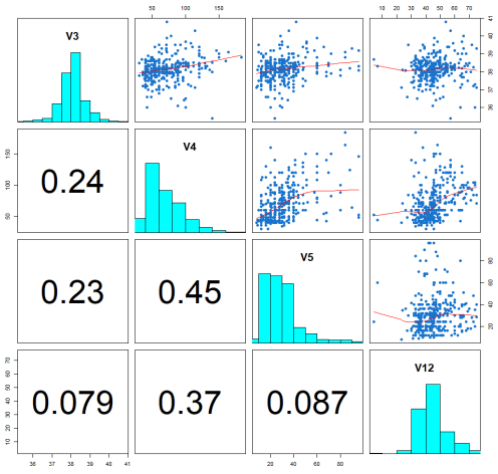

Figure 4: histogram and correlation coefficient

As mentioned before, the V3,V4, V5, V12 represents rectal temperature, pulse, respi- ratory rate and packed cell volume. We can conclude that:

- The rectal temperature and packed cell volume are cluster in the median, while the respiratory rate and pulse are right skewed.
- There are relatively strong positive correlation between pulse and respiratory rate, and there are come linear correlation between rectal temperature and pulse.

3.3 Principal Component Analysis

Since visualization can intuitively shows the relationship between the variables or the data, there is still some problems we encountered.

We still want to do some further analysis to explore the information contained in this dataset, considering this dataset. One of the them is with a lot of sets of multivariate data, it becomes harder to make the application of the graphical techniques described previously successful in providing an informative initial assessment of the data. The other is there exists multicollinearity among variables, which may cause trouble when we want to extract some information. This condition can be improved a lot if we can find some objects(or conponents) that are uncorrelated and can also represent most of the variation (which also known as information) contained in original variables.

A very popular and effective way is principal component analysis, which is a multi- variate technique with the aim of reducing dimensionality of a multivariate dataset while accounting for as much of the original variation as possible in the dataset. PCA can trans- form original variables into a principal component, the linear combinations of the original variables, which are uncorrelated and are ordered. These principal components are gener- ated with goal of maximizing variance so that the first few of them account for most of the variation in the original variables. PCA can be achieved from the following steps.

3.3.1 Standardization

As we learned in the lecture, the principal components are derived from covariance matrix of the variable, which may lead to a question: When we operate PCA on data, we only know the sample covariance matrix, should we do standardization before the analysis? The answer is yes.

One character of the PCA is it is not scale-invariant, this can be clearly explained by holding this dataset as example: Consider the rectal temperature has the unit Celsius, the pulse is the heart rate in beats per minute, the total protein has the unit gms/dL. These value of variables may change when we choose the different unit, thus the structure of the principal components derived from the covariance matrix will depend upon the essentially arbitrary choice of units of measurement. Since we want principal component to represent the most of the variation of variable itself instead of the large value which is the result of randomly chosen unit. In other words, if there are large dierences between the variances of the original variables, then those whose variances are largest will tend to dominate the early components. Therefore, principal components should only be extracted from the sample covariance matrix when all the original variables have roughly the same scale. But this is

rare in practise and consequently, in practise, principal components are extracted from the correlation matrix of the variables.

3.3.2 Got PCA from correlation matrix

I select all the continuous variables. After extraction from the correlation matrix, we got the R summary table including loadings, as we can see blow at figure 5:

```
Importance of components:
                          Comp.1    Comp.2    Comp.3    Comp.4    Comp.5
Standard deviation     1.3372055 1.0123622 0.9683891 0.8926162 0.6726537
Proportion of Variance 0.3576237 0.2049754 0.1875555 0.1593527 0.0904926
Cumulative Proportion  0.3576237 0.5625992 0.7501547 0.9095074 1.0000000
```

Figure 5: R summary table

As shown in the summary table: V3 represents rectal temperature, V4 represents pulse, V5 represents respiratory, V12 represents packed cell volume, V13 represents total protein. As we can see clearly, each principal components accounts for part of the variation of the original data. According to cumulative proportion, we know that we should at least picked 4 principal components to that can account for at least 90 % variations. We let the abbreviation of the five variables denotes them, we have the loading matrix is shown below in figure 6:

```
> load
            pca1       pca2       pca3       pca4
rctt   0.3831497  0.4576714  0.2701764  0.7547547
puls   0.6226358 -0.0177487 -0.1698775 -0.2125224
rspr   0.5263454  0.2467412  0.2900720 -0.5438190
pccc   0.4065647 -0.4522388 -0.6166527  0.2702387
ttlp -0.1522717  0.7244451 -0.6585925 -0.1281072
>
```

Figure 6: loading matrix

As we can see clearly, the first principal component

$$z_1 = 0.383\ rctt + 0.622\ puls + 0.526\ rspr + 0.407\ pccc - 0.152\ ttlp$$

shows the positive character in rectal temperature, pulse, respiratory rate, packed cell volume and negative total protein. The variation explained by this components is 35.8%.

The second principal component

$$z_2 = 0.458\ rctt - 0.018\ puls + 0.247\ rspr - 0.452\ pccc + 0.724\ ttlp$$

shows the positive character in rectal temperature, respiratory rate and total protein and negative pulse and packed cell volume. The variation explained by this components is 20.5%.

The third principal component

$$z_3 = 0.270\ rctt - 0.170\ puls + 0.290\ rspr - 0.617\ pccc - 0.659\ ttlp$$

shows the positive character in rectal temperature, respiratory rate and negative pulse and packed cell volume and total protein. The variation explained by this components is 18.7%.

The fourth principal component

$$z_4 = 0.755\ rctt - 0.213\ puls - 0.544\ rspr + 0.270\ pccc - 0.128\ ttlp$$

shows the positive character in rectal temperature and packed cell volume, and negative pulse and respiratory rate and total protein. The variation explained by this components is 16.0%.

3.3.3 pairwise plot of PCAcoefficient

Since there are 4 PCs, we cannot show them into one 2 dimensional plots, thus we use pairwise plots to exhibit, as we can see in the figure 7:

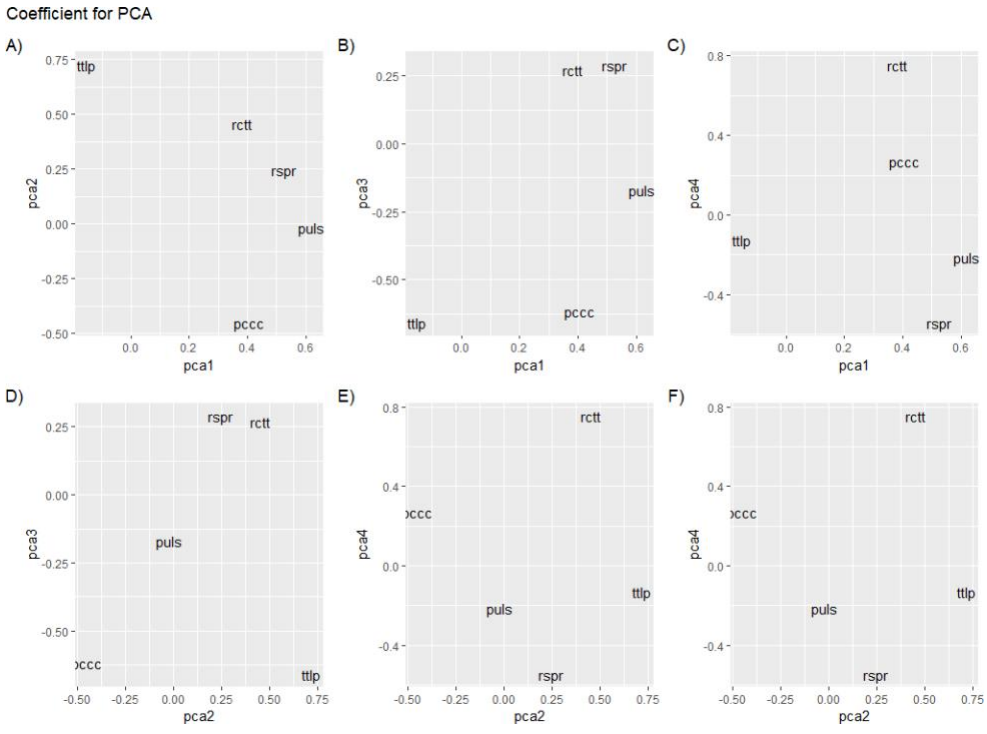

Figure 7: loading matrix

As we can see in the figure, hold PC1 as instance, the total temperature is low in PC1 while other 4 variables are high, which is concordant with what we has concluded above.

3.3.4 pairwise plot of PCA scores

Besides the PC's coefficient, we are still interested in the scores of PCs, which can show how the data distribute on these principal components, which can help us to analysis the differences between the samples and the effects other variables may cause to them. In case for the analytic complexity, I only draw the sample scatter plot on the two selected component.

Here is the plot on the first two components and I distinguish the horses based on 7-th variable Peripheral pulse through colors and points characters, the variable peripheral pulse is a categorical variable with 4 levels:1 = normal;2 = increased; 3 = reduced; 4 = absent. The plot is shown below :

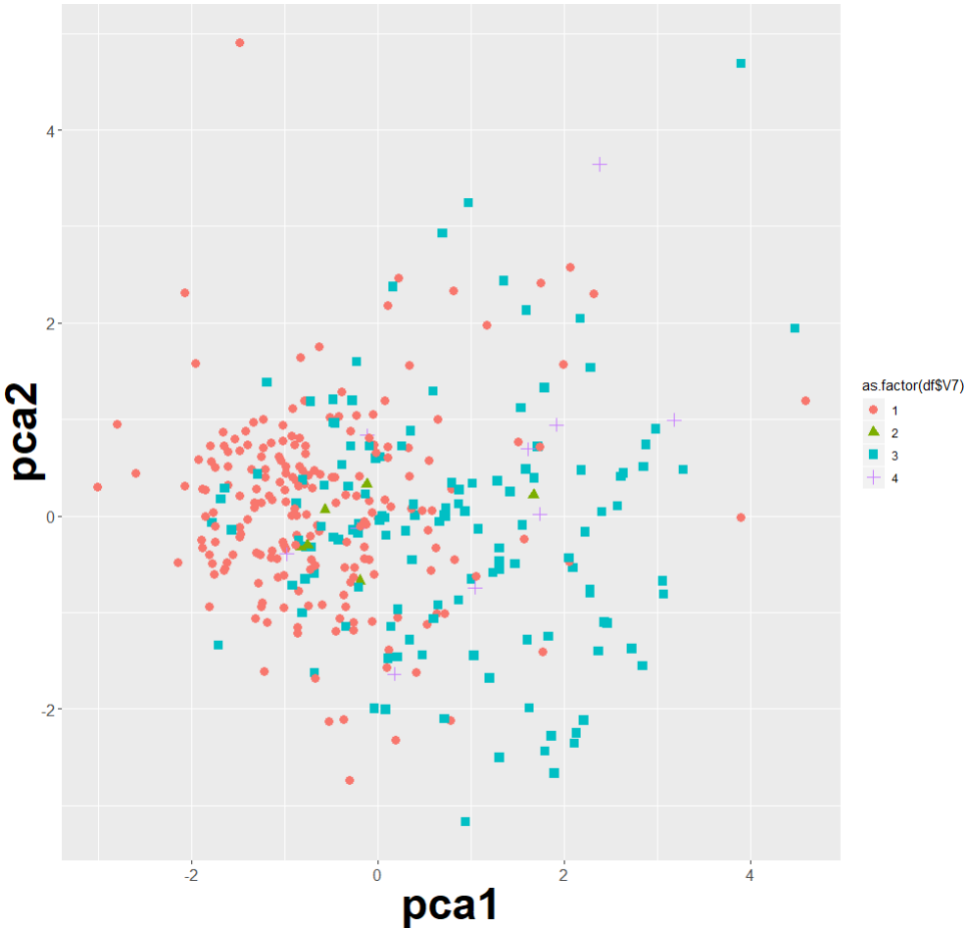

Figure 8: PC's scores

As we can see in figure 8:

- The horses with normal or increasing peripheral pulse have relatively lower PC1 value and are more cluster in lower PC1 and medium PC2, which indicates they are

of normal level in rectal temperature, pulse, respiratory rate, packed cell volume, and has adequate protein in their body. According to what has been illustrated in the introduction, Normal or increased peripheral pulse are indicative of adequate circulation, which is concordant with the information shown in the figure.

- As for horses with reduced peripheral pulse or such pulse is absent(sometimes undetected) in some horses, these horses' PC scores in the first PCs are more like a erratic and scatter pattern. This pattern shows these horse have a more unsteady rectal temperature, pulse respiratory rate and packed cell volume.

Here is the plot on thefirst two components and I distinguish the horses based on 14-th variable outcome through colors and points characters, the variable peripheral pulse is a categorical variable with 3 levels:1 = lived;2 = died; 3 = was euthanized. The plot is shown below :

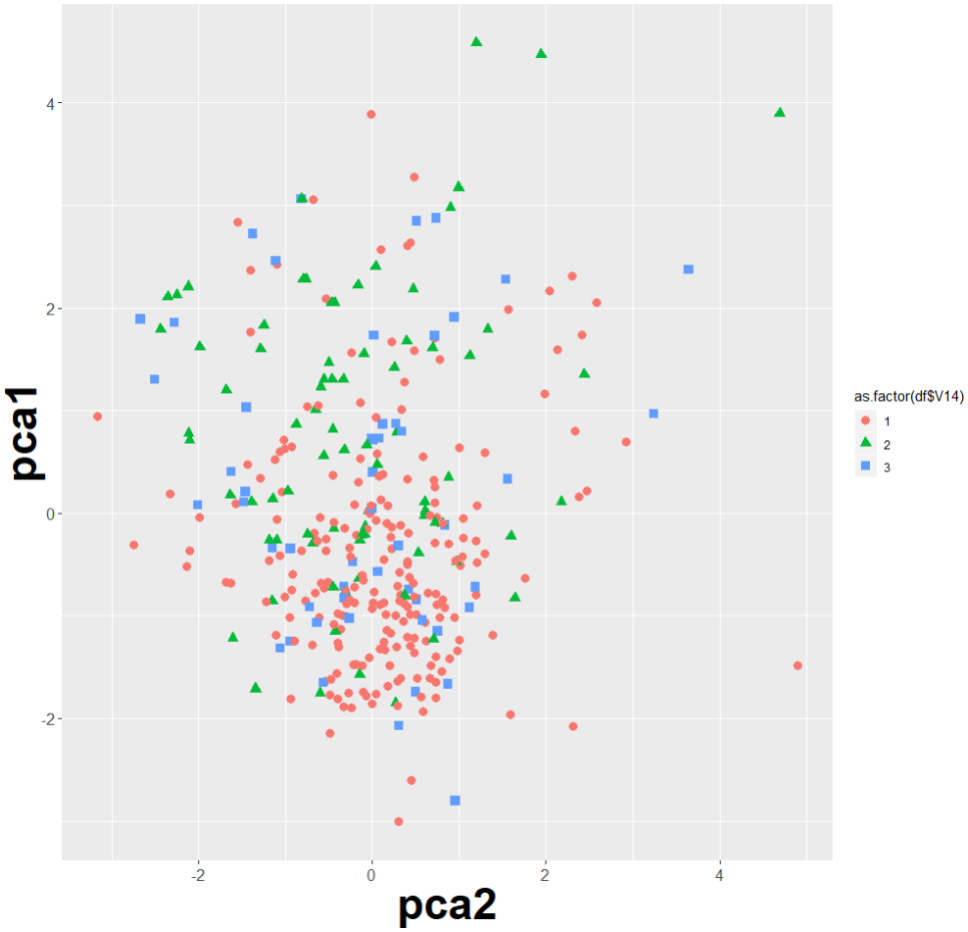

Figure 9: PC's scores

As we can see in the figure 9 :

- The lived horse has a lower PC1 and relatively lower PC2, and are more clustered, which indicates their physical condition such as rectal temperature, pulse and packed cell volume is stable, this result is consistent with previous conclusion.

- The died horse(include one which has been euthanized) are higher in PC1 and PC2 than the lived horse, and are more scatter in the plots, which indicates their physical condition is unstable.

3.4 Canonical correlation analysis(CCA)

Principal components analysis considers interrelationships within a set of variables. But there are situations we may be interested in assessing the relationships between two sets of variables.

After analysis the relationship of all variables(considered to be in one sets), we still want to analysis the relationship between the horse's dynamic physical situations (which can be fluctuated during sleep, or physical motions such as running) and stable situations (which will not easily fluctuate on one horse). Our goal is to detect: Is dynamic physical situations related to stable physical situations?

In this case, we need to measure a number of dynamic variables meanwhile a number of stable variables and analysis the association of them. One technique for addressing such questions is canonical correlation analysis.

Suppose we have two sets of variables, say sets x and sets y, the approach adopted in CCA is to take the association between x and y to be the largest correlation between two single variables, u1 and v1, derived from x and y, with u1 being a linear combination of variables in sets x and v1 being a linear combination of variables in sets y. But often a single pair of variables (u1,v1) is not sucient to quantify the association between the x and y variables, and we may need to consider some or all of s pairs where s is the minimum number of variables in both sets.

3.4.1 CCA approach

I select pulse and respiratory rate as dynamic situation variables and rectal temperature, packed cell volume and total protein as stable situation variables. Then I use CCA proce- dure on it. As I illustrated in PCA part, we still need to standardize the data, Then I use CCA procedure to get canonical correlations and the corresponding u and v. This is the R result we got:

```
> eigen(e1)
eigen() decomposition
$`values`
[1] 0.18737813 0.03573377

$vectors
             [,1]       [,2]
[1,] 0.999955446 -0.4175152
[2,] 0.009439618  0.9086699

> eigen(e2)
eigen() decomposition
$`values`
[1]  1.873781e-01  3.573377e-02 -2.217482e-18

$vectors
            [,1]       [,2]       [,3]
[1,]  0.4839294  0.7813950  0.3811774
[2,]  0.8586942 -0.4814121 -0.2702414
[3,] -0.1686910 -0.3970697  0.8841229
```

Figure 10: cca result

As we can see in figure 10, the canonical correlations, being the square root of eigen-values, are 0.4329 and 0.1890 (I take 4 decimals). The canonical variables are:

$$
\begin{aligned}
u_1 &= puls + 0.01\ rspr \\
v_1 &= 0.484\ rctt + 0.859\ pccc - 0.169\ ttlp \\
u_2 &= -0.42\ puls + 0.91\ rspr \\
v_2 &= 0.781\ rctt - 0.481\ pccc - 0.397\ ttlp
\end{aligned}
\tag{1}
$$

As we can see in equation (1):

- u1 mainly represent the pulse and v1 is mainly a function of rectal temperature and packed cell volume. Thus the higher rectal temperature and packed cell volume tend to be in the horse with higher pulse rate.
- u2 is mainly the contrast between respiratory rate and rectal temperature(where respiratory rate makes positive effect), and v2 is mainly the contrast between rectal temperature and other two stable vatiables, where rectal temperature makes positive effect. Thus we can conclude that horse with higher rectal temperature and lower packed cell volume tends to have lower pulse and higher respiratory rate.

3.5 Multidimensional scaling

As we mentioned in last section, one of the main purposes of the principal components analysis and canonical correlation analysis is to find some components to represent the original variables and analysis the correlation between variables.

Another main purpose is to obtain a low dimensional ”map” of the data that can represent the relationship between the observations. There is a other method that can achieve (or more concentrate on) this goal, called multidimensional scaling. This method

aims to produce similar maps of data but do not operate directly on the usual multivariate data matrix X. Instead they are applied to some kind of distance matrices which are derived from the matrix X, and also are applied to so-called dissimilarity or similarity matrices(A more generalized name is proximity matrix, a proximity matrix may measure similarity or dissimilarity). This character of this method is also a advantage since it doesn't need the original data matrix, and can depict the relationship(distance or dissimilarity) among the samples on lower dimensional plots just from observation's distance matrix or dissimilarity matrix. If the original dataset's dimension is 2, thus we can reconstruct the map of observations without their original data just through proximity matrix.

3.5.1 Classical multidimensional scaling

First, like all MDS techniques, classical scaling seeks to represent a proximity matrix by a simple geometrical model or map. I used the original dataset with 5 variables: rectal temperature, pulse, respiratory rate, packed cell volume and total protein. Thus I got 368 observations with 5 variables. Since these variables are normal distribution, I calculate the Mahalanobis distance between the observations.

Since the sum of first 3 largest eigen values has the overall 0.8 proportion of sum of all the eigen values, we choose k=3 is a good fit, where k represent is best-fitting k-dimensional representation. Then I use a 2-dimensional figure to show the relationship of the total observations. Here is the 2D classical MDS solution for all the horses besed on their five physical variables mentioned previously, I distinguish them based on their age:1= adult horse,2=young horse(<= 6 months). Here is shown below:

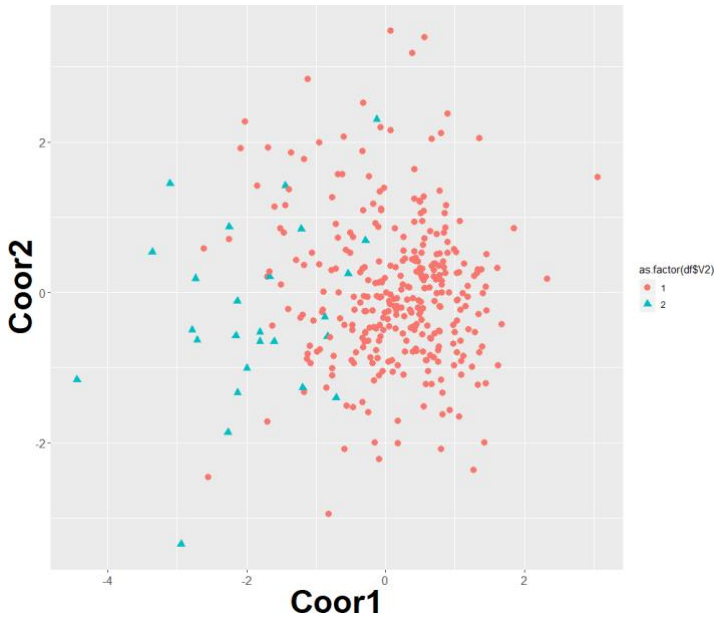

Figure 11: mds result

As we can see in figure 11, the MDS successfully distinguish the horse by the age, thus we can conclude the younger horse do have different rectal temperature, pulse, respiratory rate, packed cell volume and total protein with the adult horse.

Here is another 2D classical MDS solution for all the horses and I distinguish them based on their capillary refill time:1:<= 3 seconds,2: >= 3 seconds.

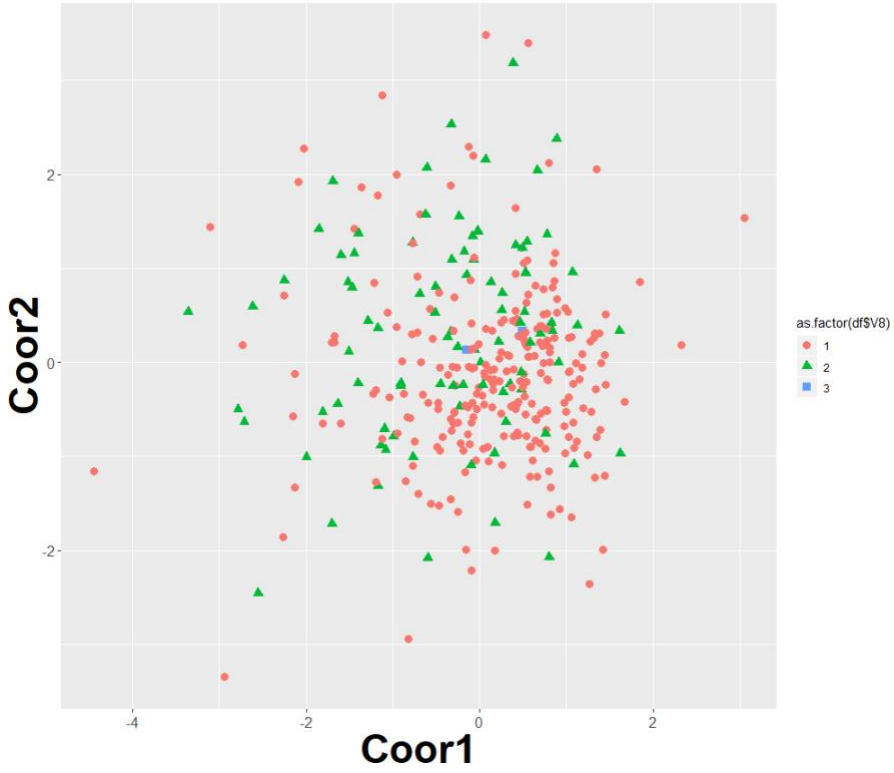

Figure 12: mds result

As we can see in figure 12, the MDS successfully distinguish the horse by thecapillary refill time, according what we have introduced in the introduction, the longer the capillary refill time, the poorer the circulation. We can conclude that the horses with with longer refill time (also the poorer circulation) have different rectal temperature, pulse, respiratory rate, packed cell volume and total protein with the adult horse, which is concordant to the previous conclusion.

3.5.2 Non-metric multidimensional scaling

In some cases, we may want to focus on the order of the distance or dissimilarity of samples due to the irregular magnitude of the sample. This horse data is related to biostatistic thus it may be unreasonable to calculate the numerical distance in the data as the difference of the horses. Such considerations led to the search for a method of multidimensional

scaling that uses only the rank order of the proximities to produce a spatial representation of them. In other words, a method was sought that would be invariant under monotonic transformations of the observed proximity matrix, in which case, the derived coordinates will remain the same if the numerical values of the observed proximities are changed but their rank order is not.

Here is the 2D classical MDS solution for all the horses besed on their five physical vari- ables mentioned previously, I distinguish them based on their age:1= adult horse,2=young horse(<= 6 months).

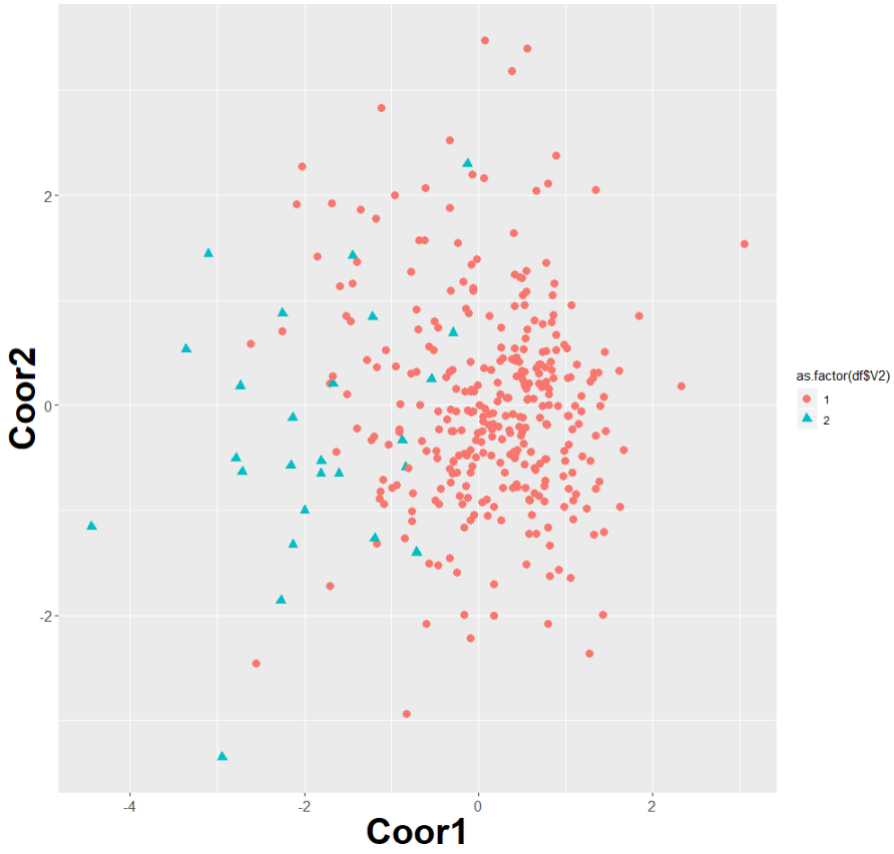

Figure 13: nonmetric mds result

As we can see in the figure 13, we get a better result than the classical MDS. The nonmetric MDS better distinguish the horses based on their ages, which verify the reliability of this method.

As we can see, as for looking for the relationship between the observations and con- structing a 2-D map to show this relationship, the MDS methods have advantages over PCA since it has a more clear representation.

3.5.3 Clustering

I wonder whether it is correct we use this five physical condition variables(rectal temperature, pulse, respiratory rate, packed cell volume and total protein) to construct the plot and distinguish them based on the difference of observations based on this 5 variables. So, I want to check whether this five variables has dramatically differences and whether we can divide into sub-groups only based on this 5 variables data.

Thus, clutering come into my mind, if we can cluster the observations into sub-groups then the conclusions listed previously are concordant. I use K-means to achieve this goal since it will update the center of each cluster when clustering. Here can we see below:

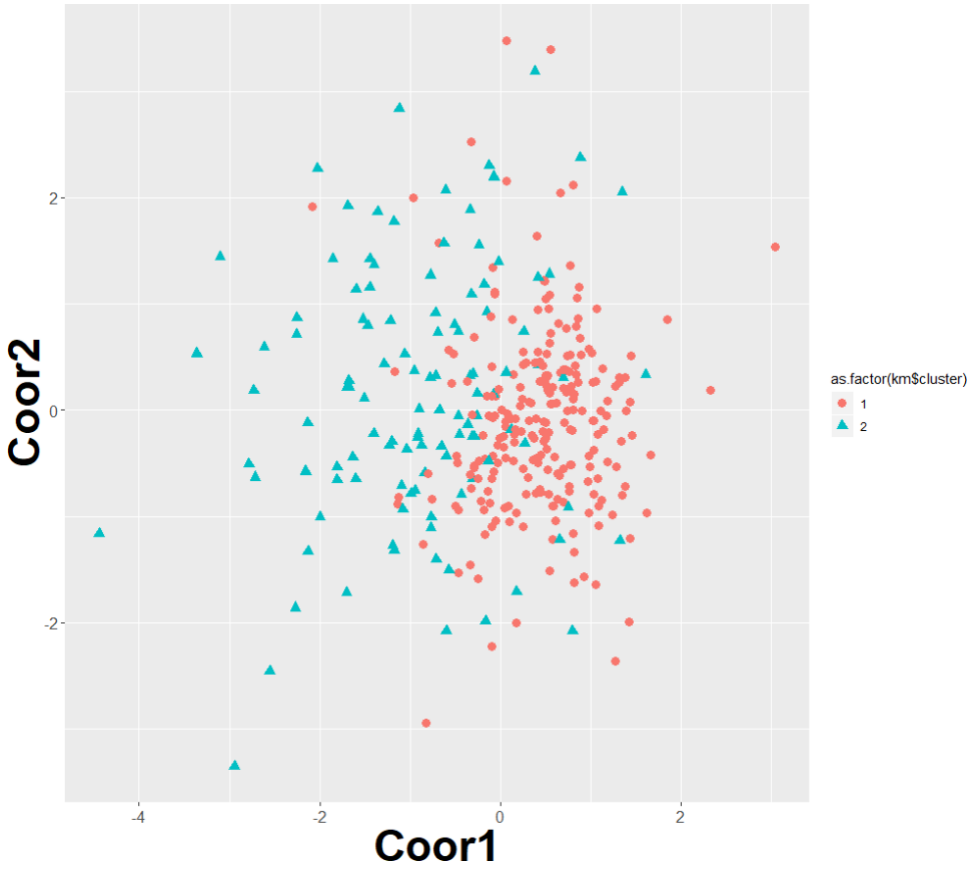

Figure 14: kmeans

As we can see in figure 14, the observations cluster into 2 sub-groups, which is consistent with the previous results. This result conduct the effectiveness of the previous 5 physical condition variables.

3.6 Correspondence Analysis

In the previous sections, we have fully demonstrate and analysis the relationship between the continuous variables and the observations based on these variables.

But considering the attributes of the horse dataset, there are a bunch of categorical variables play a essential role in horse's physical condition, and also interest me how this conditions interact and have some influence on each horse.

That brings in the correspondence analysis, which is a technique for displaying multivariate (in most cases bivariate) categorical data graphically by deriving coordinates to represent the categories of both the row and column variables, which may then be plotted so as to display the pattern of association between the variables graphically.

As what has been introduced in introduction, the horses temperatures of extremities have strong influence on horse's physical condition, cool to cold extremities indicate possible shock, hot extremities should correlate with an elevated rectal temp. What interest me is whether this temperature may have connection with the horse's pain level, since in general a horse who is more painful is likely to treated surgery, thus the temperature of extremities might be the indicator of whether this horse need surgery.

Thus, the two categorical variables we want to analysis is :

1. Temperature of extremities:1 = normal. 2 = warm. 3 = cool. 4 = cold.
2. Pain level: the level range is $1 \sim 5$,the pain increases with level increases. This

is the table we get from data:

	level 1	**level 2**	**level 3**	**level 4**	**level 5**
norm	32	20	47	9	8
warm	12	7	14	4	2
cool	13	53	42	34	35
cold	2	13	6	5	10

Table 1: Table caption

We will analysis table 1 by following two procedures(plots).

3.6.1 Mosaic plot

Here is the mosaic plot which can directly show the relationship between these two variables:

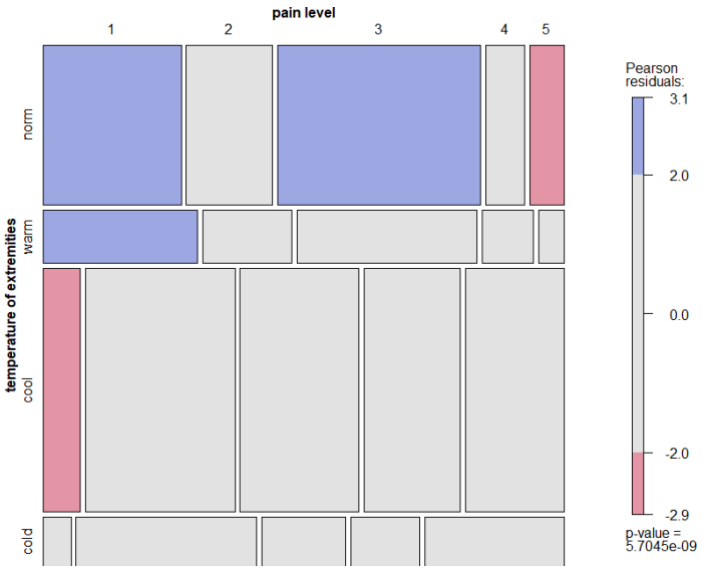

Figure 15: Mosaic plot

As we can see in the figure 15, the warm temperature are easily occurring in a horse accompanying the lowest pain level, the cool temperature are less likely to be in the horse with lowest pain level, which verify the background previously.

3.6.2 CA plot

A CA plot can help us better understand the relationship between two variables as well as the specific level of each variables, such as row profiles and column profiles. Here is the CA plot:

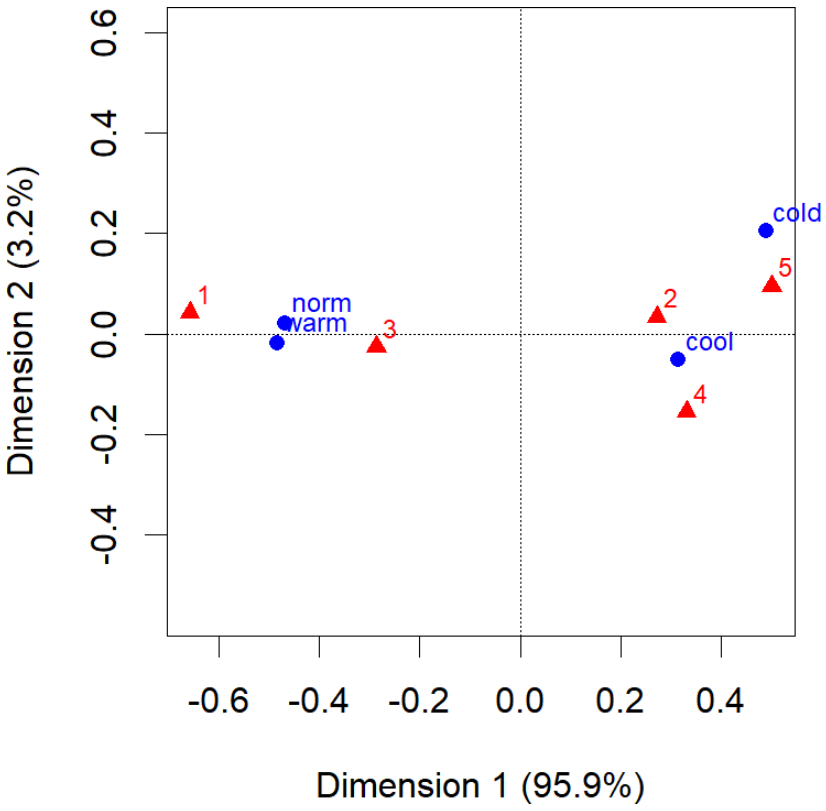

Figure 16: CA plot

As we can see in figure 16:

- As for temperature, the norm level of the horse and warm level of the horse has the same row profiles.
- We can get similar result in mosaic plot. The norm level and warm level of temper- ature combined with pain level 1 occurs more frequently in the table than would be expected if the two variables were independent.
- The cool level of temperature combined with pain level 1 occurs less frequently in the table than would be expected if the two variables were independent.

Through correspondence analysis, we can safely conclude that the temperature of extrem- ities is not independent with the horse's pain level.

4 Conclusion

We have use multiple ways to analysis this horse colic data, extract information about it and conclude useful results which is concordant and also supplement the background knowledge according to biostatistic. Each methods has its advantages and disadvantages. The reliable way is to combine them and conclude at a overall situation.

1. In the PCA section, from figure 8 and figure 9, we learned the horses with normal or increasing peripheral pulse are of normal level in rectal temperature, pulse, respira- tory rate, packed cell volume,and has adequate protein in their body. As for horses with reduced peripheral pulse or such pulse is absent(sometimes undetected) in some horses, these horse have a more unsteady rectal temperature, pulse respiratory rate and packed cell volume.
2. In the CCA section, from figure 10, we learned the higher rectal temperature and packed cell volume tend to be in the horse with higher pulse rate. And we can conclude that horse with higher rectal temperature and lower packed cell volume tends to have lower pulse and higher respiratory rate.
3. In the MDS section, we can conclude the younger horse do have different rectal temperature, pulse, respiratory rate, packed cell volume and total protein with the adult horse. We also learned that the horses with with longer refill time (also the poorer circulation) have different rectal temperature, pulse, respiratory rate, packed cell volume and total protein with the adult horse, which is concordant to the previous conclusion.
4. In the Clustering section, we learned that five continuous physical condition vari- ables(rectal temperature, pulse, respiratory rate, packed cell volume and total pro- tein) has dramatically differences and can divide observations into sub-groups.

5. In the CA section, we learned the temperature of extremities of a horse has relatively strong correlations with a horse's pain level, which may result in whether such horse receive a surgery. We have concluded that the warm temperature are easily occurring in a horse accompanying the lowest pain level, the cool temperature are less likely to be in a horse with the lowest pain level.

Thus, in multivariate analysis, there are several methods to analysis a dataset and extract information. Not only should we conduct these methods and draw such plots, but more importantly to know the advantages and disadvantages of these methods and choose the most suitable method on different dataset, and never overlook any method that can shed more light on the data.

References

[1] Machine Learning Repository:
https://archive.ics.uci.edu/ml/datasets/Horse+Colic

[2] Chambers, J. M., Cleveland, W. S., Kleiner, B., and Tukey, P. A. (1983), Graphical Methods for Data Analysis, London, UK: Chapman & Hall/CRC. Cited on p. 25

[3] Everitt, Brian, and Torsten Hothorn. An Introduction to Applied Multivariate Analysis with R. Vol. 4, Springer New York, 2011.