**Zoo animal dataset clustering and classification**

Rui Wang & Moran Wang

6/06/2018

**Data source and definitions explained:**

Zoo animal dataset is downloaded from Kaggle. There are two csv files in the zoo animal dataset. One is zoo.csv file. This file includes 18 columns: animal name, hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic, catsize, leg and class type. Hair, feathers, eggs, milk, airborne, aquatic, predator, toothed, backbone, breathes, venomous, fins, tail, domestic and catsize are binary variables that describe the features of animals. Leg is a numeric variable which describes the feature of animals. Another csv file is class.csv. This file includes class number, number of animal species, class type and animal names. In order to prediction, we merge this two csv files based on class type. After merging, the new dataset is shown below. There are 101 instances in our data.

A picture containing furniture

Description generated with very high confidence

**Main features of data set presented with appropriate graphics:**

* The response variable (animal class) is categorical. Then we explore how many animals in each animal class. We find that over 40 records are classified as mammal. And less than 10 records are classified as reptile and amphibian. This indicates that after the clustering and classification, the result should have more mammals and less amphibian and reptile.

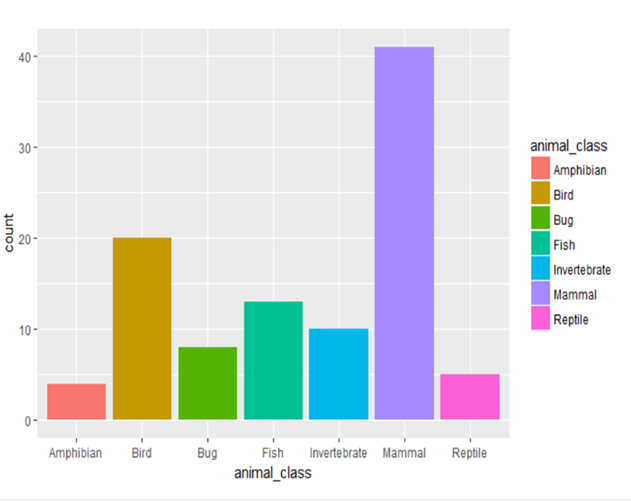


Fig1. count for distinct animal class

* we find some relationship between features and animal class. For example, we draw the graph between eggs and animal class as fig2. We find that Mammals almost don’t lay eggs. We also draw the legs VS type graph(fig3) and we find that fish don’t have legs and bugs have more legs than other type of animals. After we draw these relationship graphs between features and animal class, we find these features can precisely reflect the animal class.

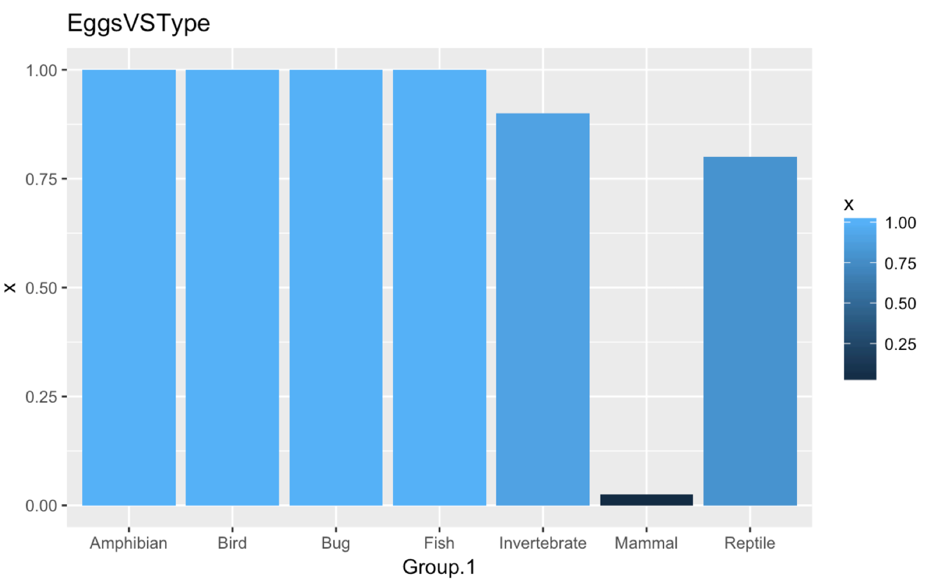
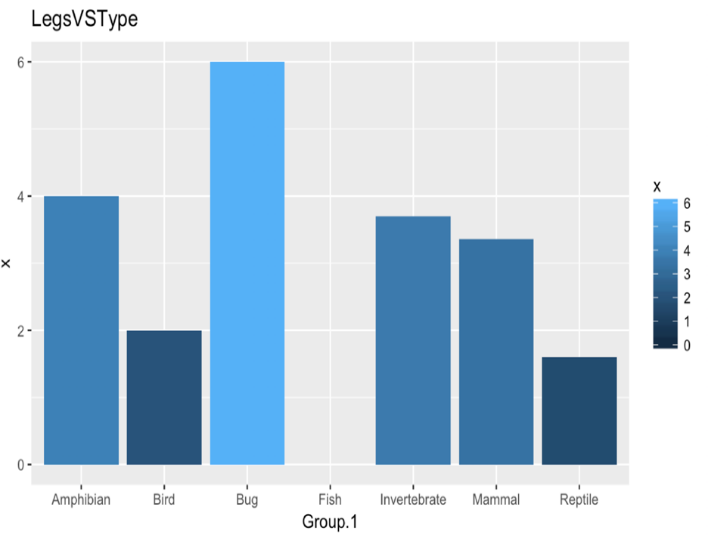
****

Fig2. eggs vs animal class Fig3.legs vs animal class

* Then we explore the relationship between attributes and we find that many of them are highly correlated. As the fig4 shows, darker blue and darker red mean high (positive and negative) correlation. For example, milk and hair have high positive relationship. Eggs and milk have high negative correlation.

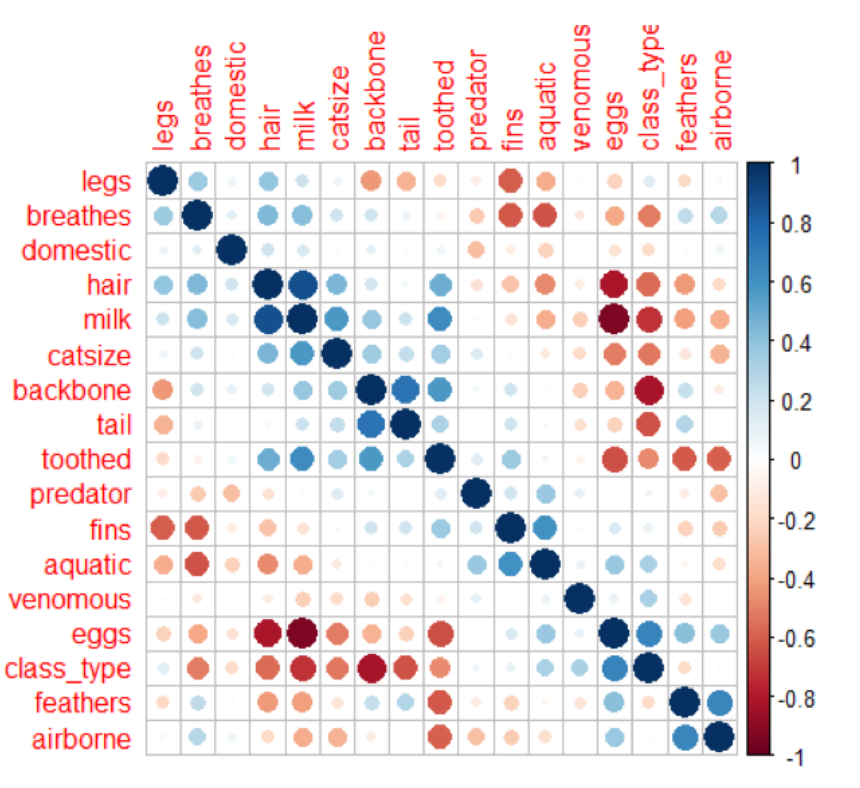


Fig4. correlation among attributes

**Research questions presented**

Q1: How could we reduce the dimension of variables and still contain as much as original information?

Q2: What are the clustering results by using Expectation Maximization method when assuming we do not know the actual animal types?

Q3: How could we predict the animals type?

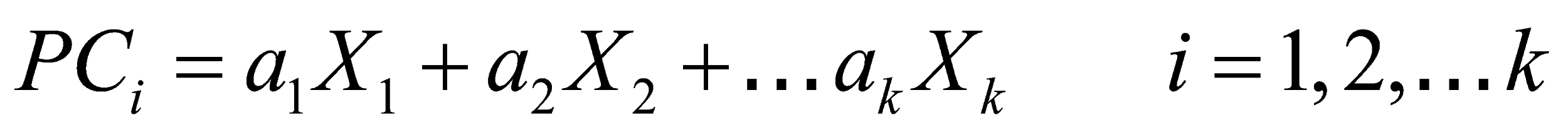
**For Q1:**

**Method and Explained:**

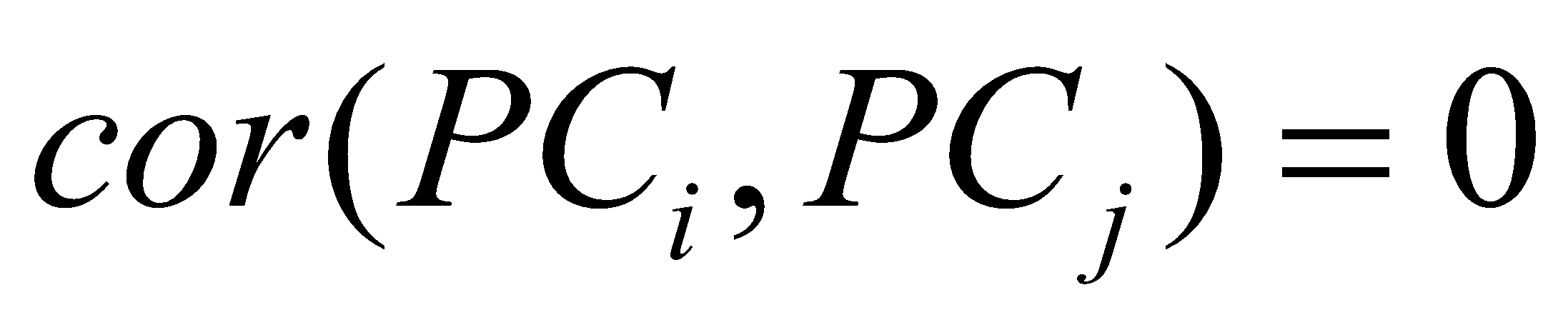
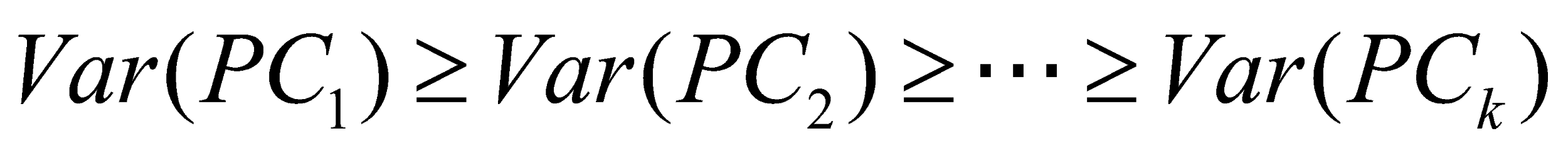
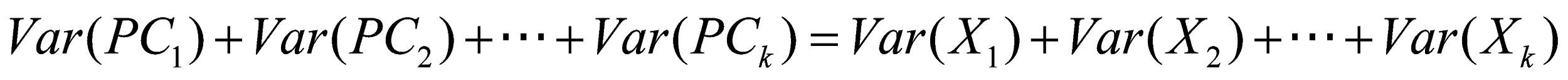
Principle component analysis(PCA) Method to solve 1st question.

PCA is a data reduction technique that transforms a larger number of correlated variables into a much smaller set of uncorrelated variables called principle components. (Kabacoff, 331) Principle

components are linear combination of original variables, the equation shown below.



Principal component properties:

1. ****
2. ****
3. 

**Data satisfaction of requirements of method:**

The goal of the PCA method is to use a smaller number of uncorrelated variables to instead of a large number of correlated variables. In our project, we use 16 correlated variables with 101 observations.

**Method applied and interpreted:**

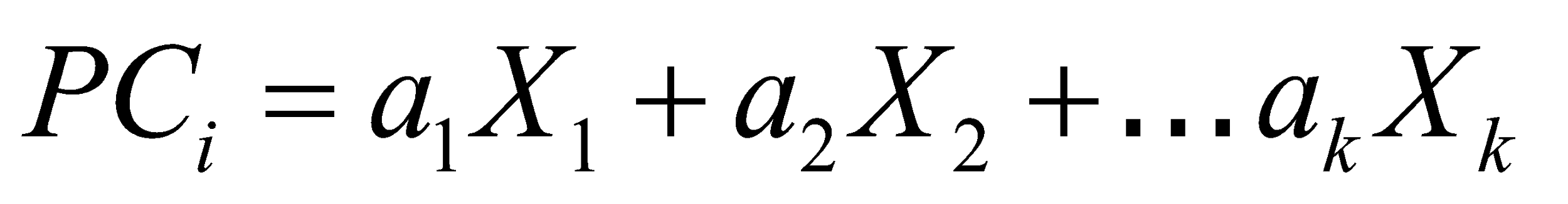
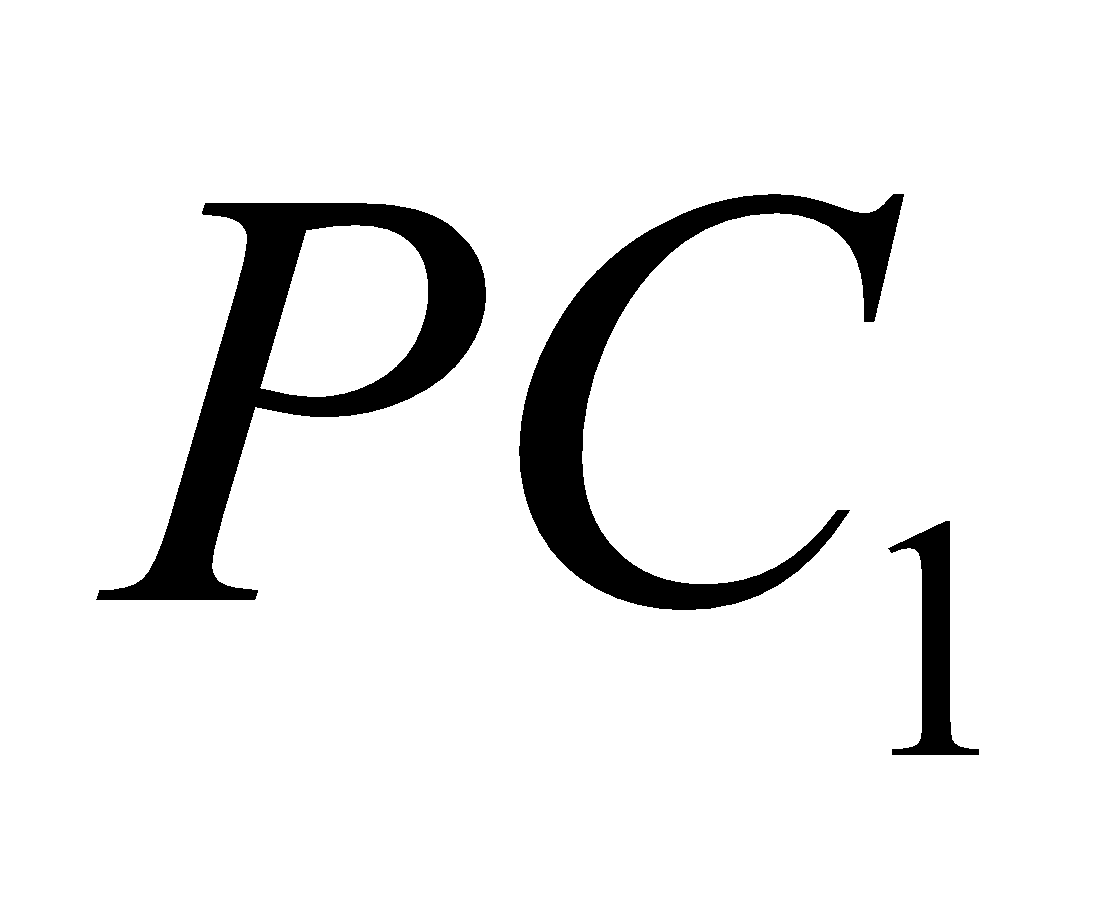
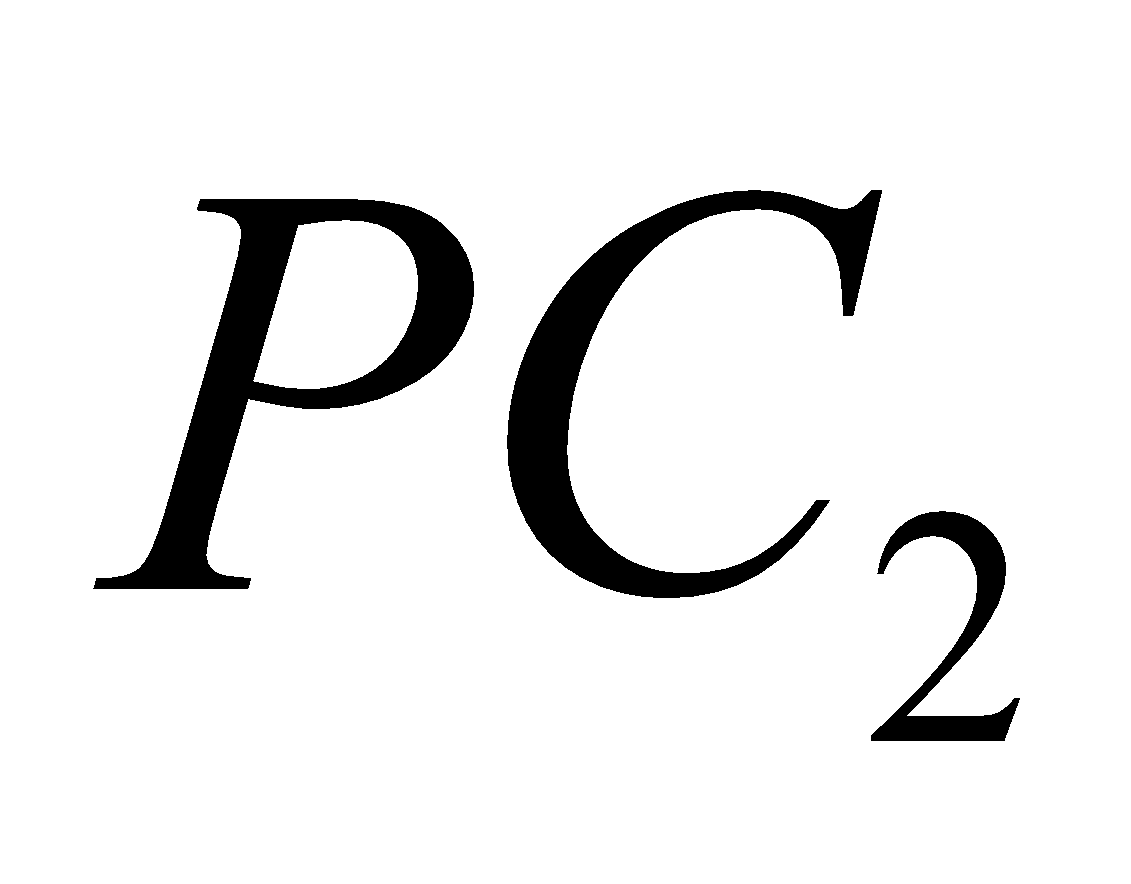
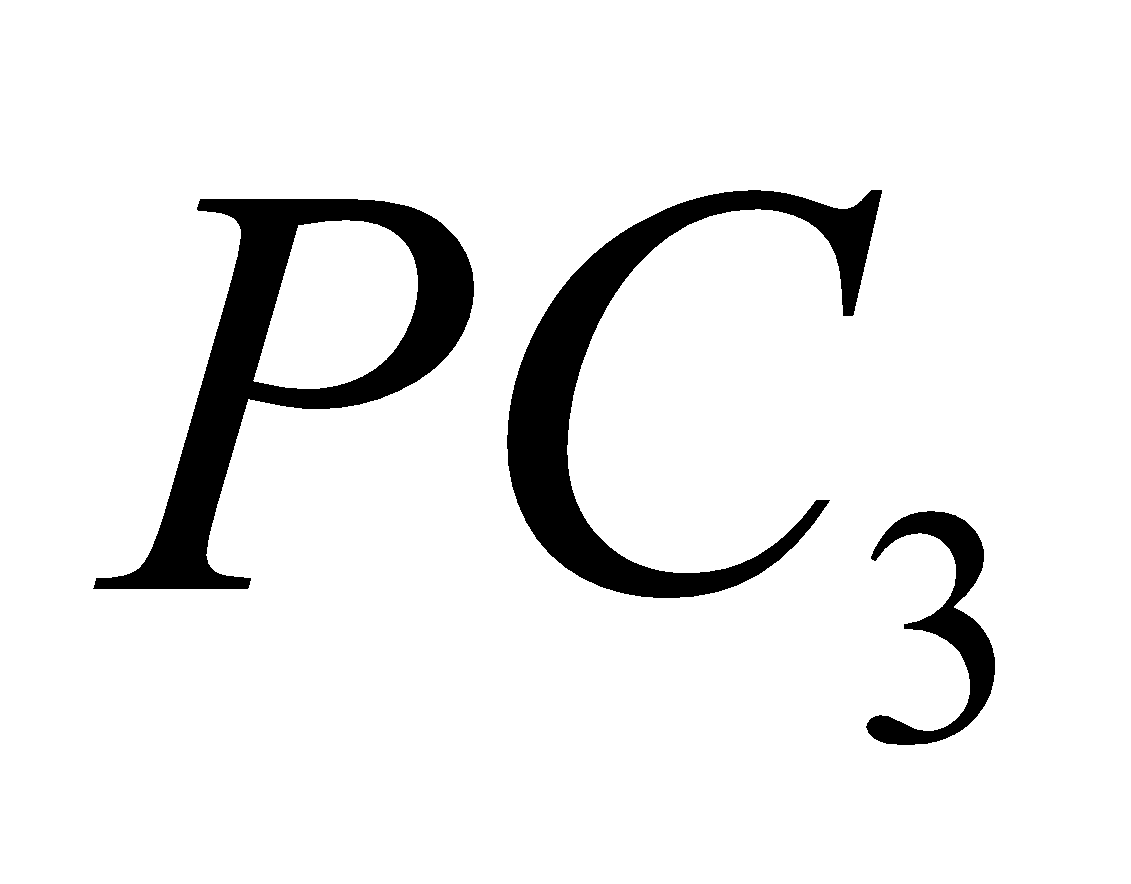
There are there parts in PCA method.

The first step is to find the number of components. The most common way is based on the eigenvalues. The largest eigenvalue is the PC1, the second-largest eigenvalue is the PC2, and so on. In R programming, there is a package called "psych". It contains a function "fa.parallel()", which can show three eigenvalue criteria at the same time.

Three eigenvalue criteria:

1. The Kaiser-Harris criterion: it suggests retaining components with eigenvalues greater than 1. Components with eigenvalues less than 1 explain less variance than contained in a single variable. (Kabacoff, 335)
2. The Cattell Scree test: the eigenvalues are plotted against their component numbers. Such plots typically demonstrate a bend or elbow, and the components above this sharp break are retained. (Kabacoff, 335)
3. The Parallel Analysis: you can run simulations, extracting eigenvalues from random data matrices of the same size as the original matrix. If an eigenvalue based on real data is larger than the average corresponding eigenvalues from a set of random data matrices, that component is retained. (Kabacoff, 335)

we use the "fa.parallel()" and abline( ) functions in R to draw fig5. The horizontal line is from the Kaiser-Harris criterion. The line with x-points is from the Cattell Scree test. The dashed line is from the Parallel Analysis. Finally, the plot suggests that three components are appropriate for summarizing our dataset.

The second step is to get loadings. In this step, we still use a built-in function in R called "principal()" from the "psych" package. we get the loading values as fig6. The third step is to convert original data to principle components data. We use the formula to convert data, where "a"s are component loadings from each principle component and "X"s are the attributes in original data. We got the values of ,  and , so the new data has three variables and they are independent as fig7. we can use this dataset for EM method to solve the 2nd question.

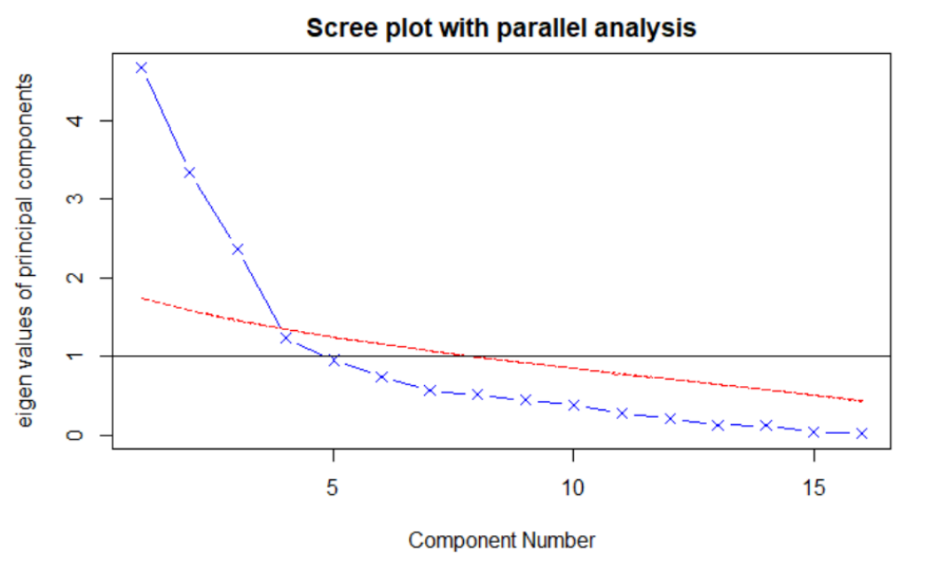
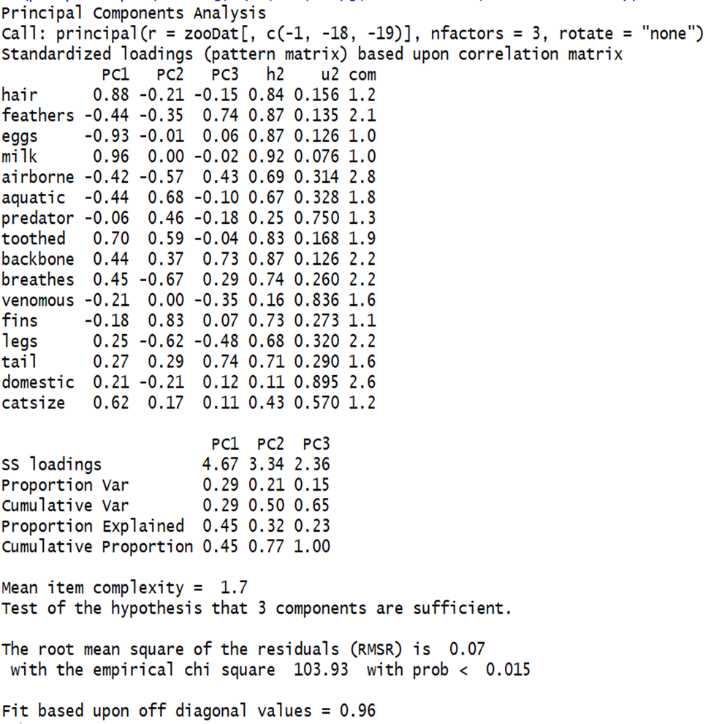


Fig5. Scree plot

A picture containing wall

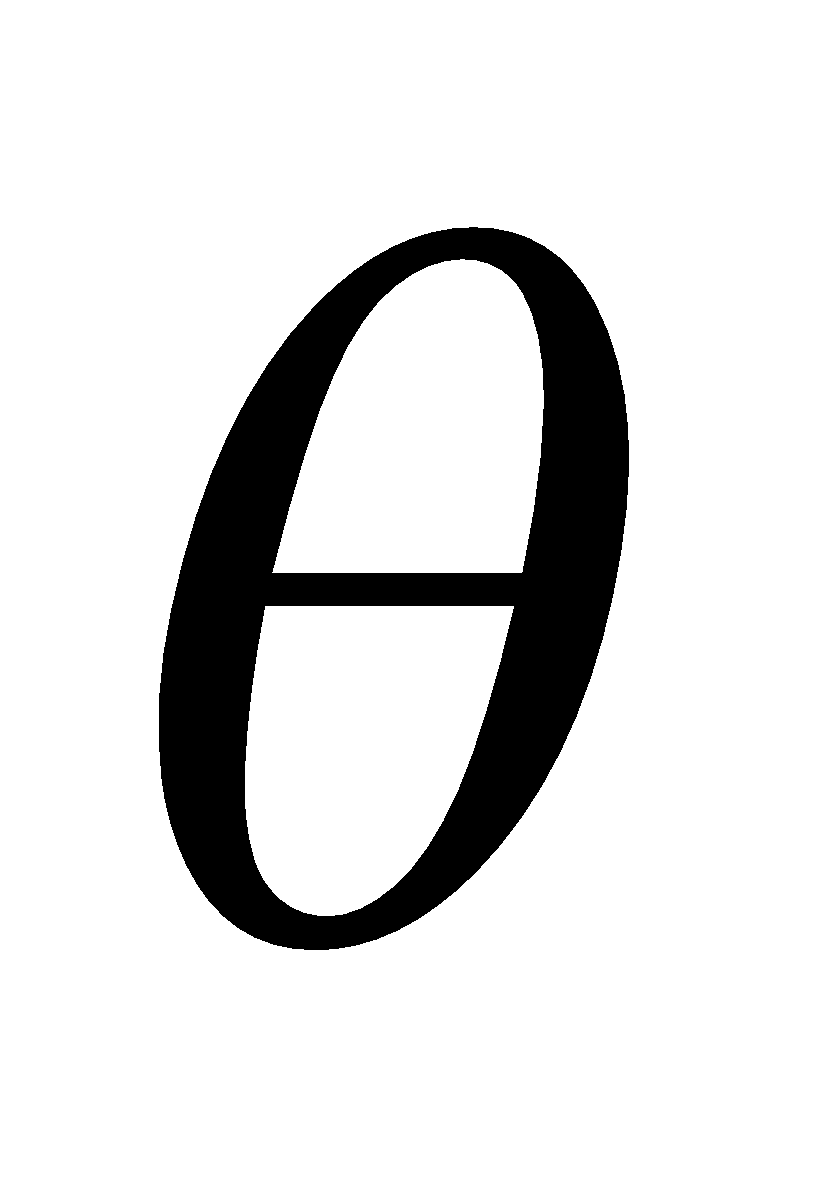
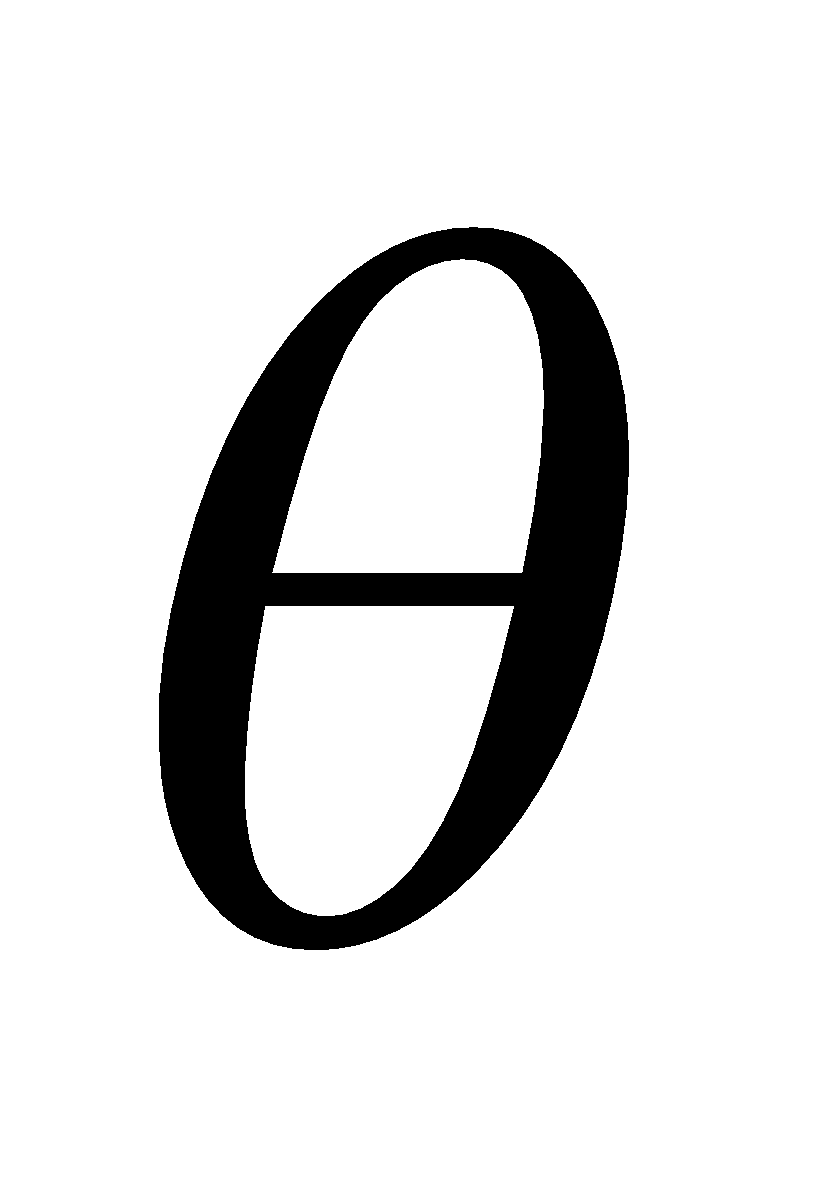
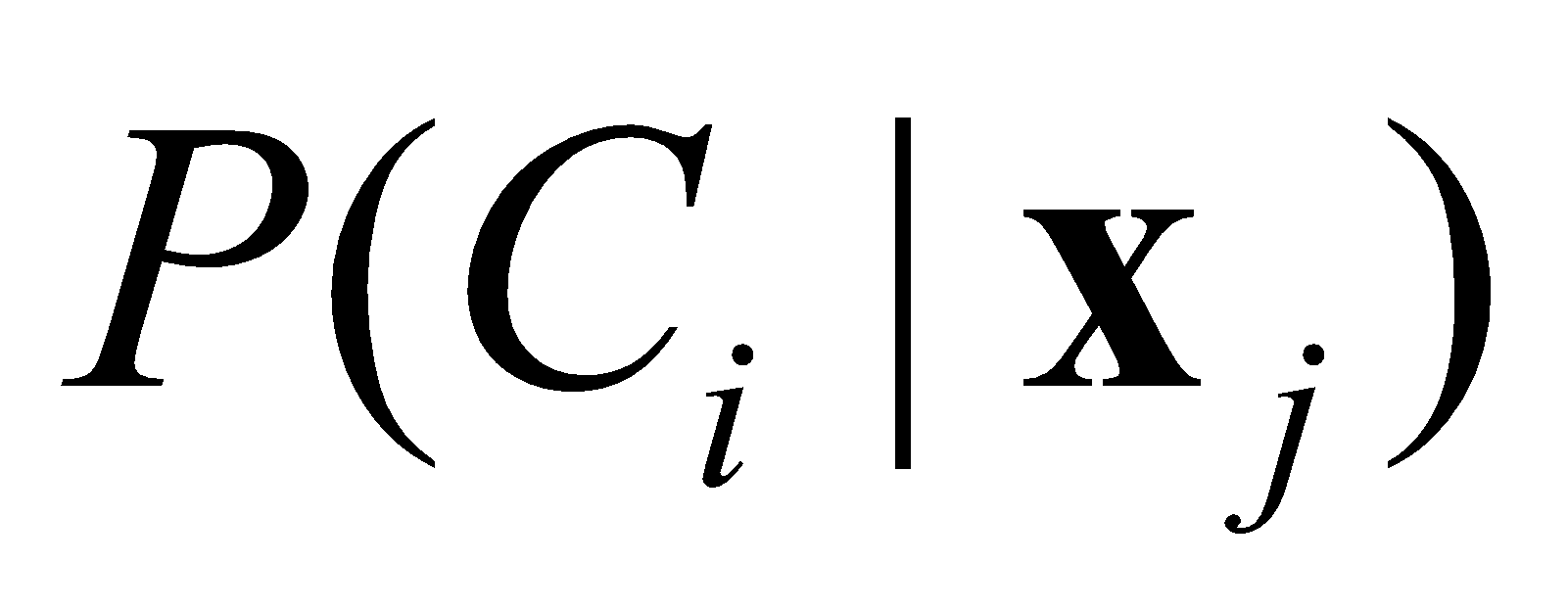
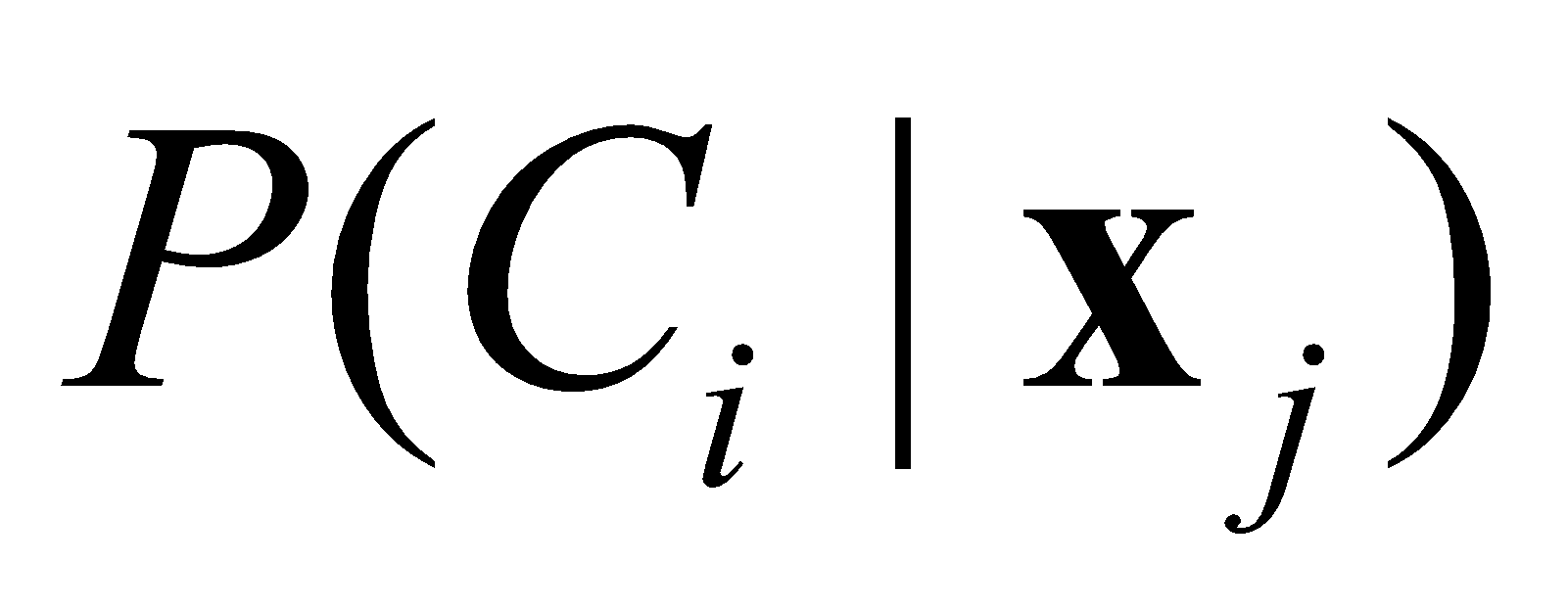
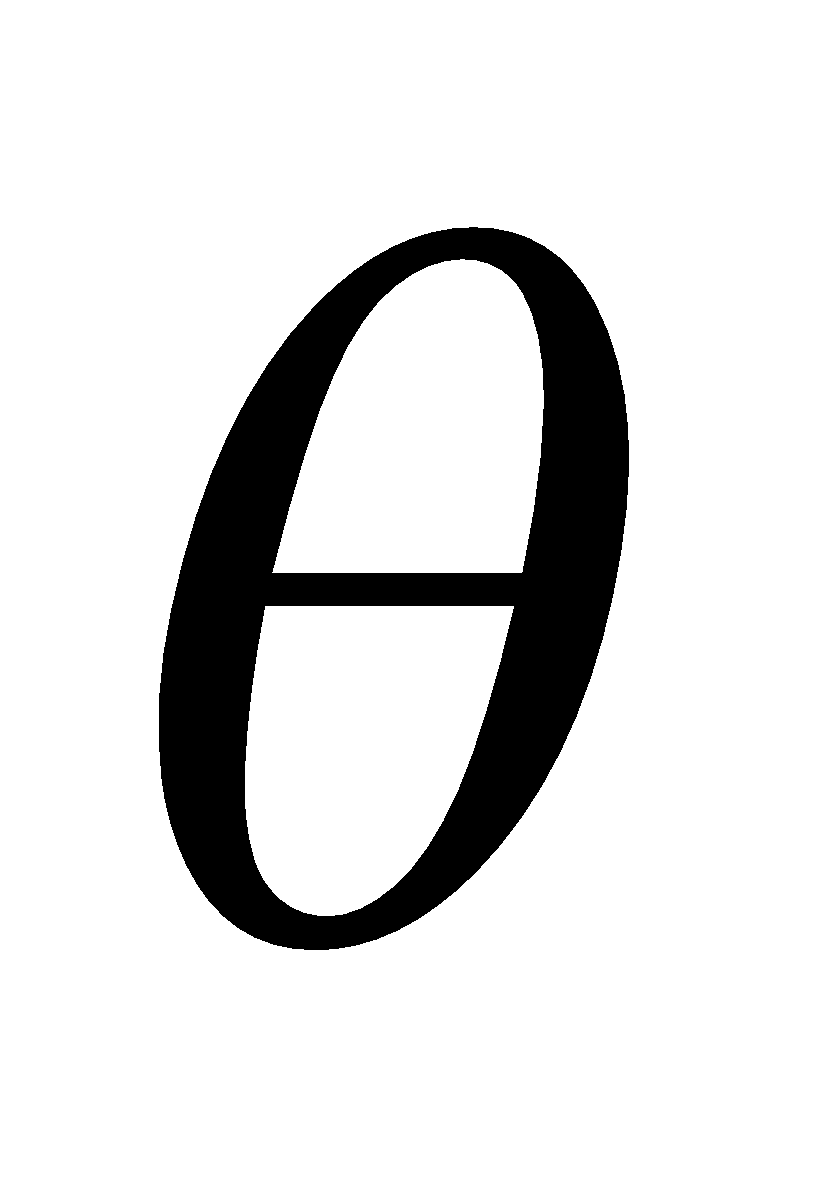
Description generated with high confidence

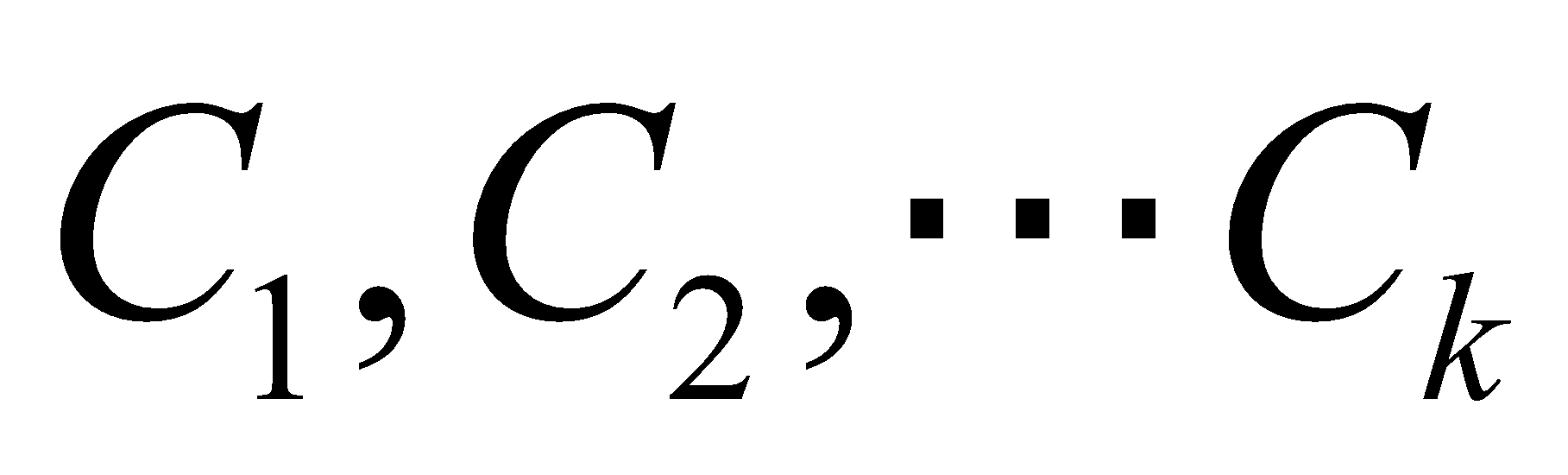
Fig6. Loadings of principle components Fig7. principlecomponents data

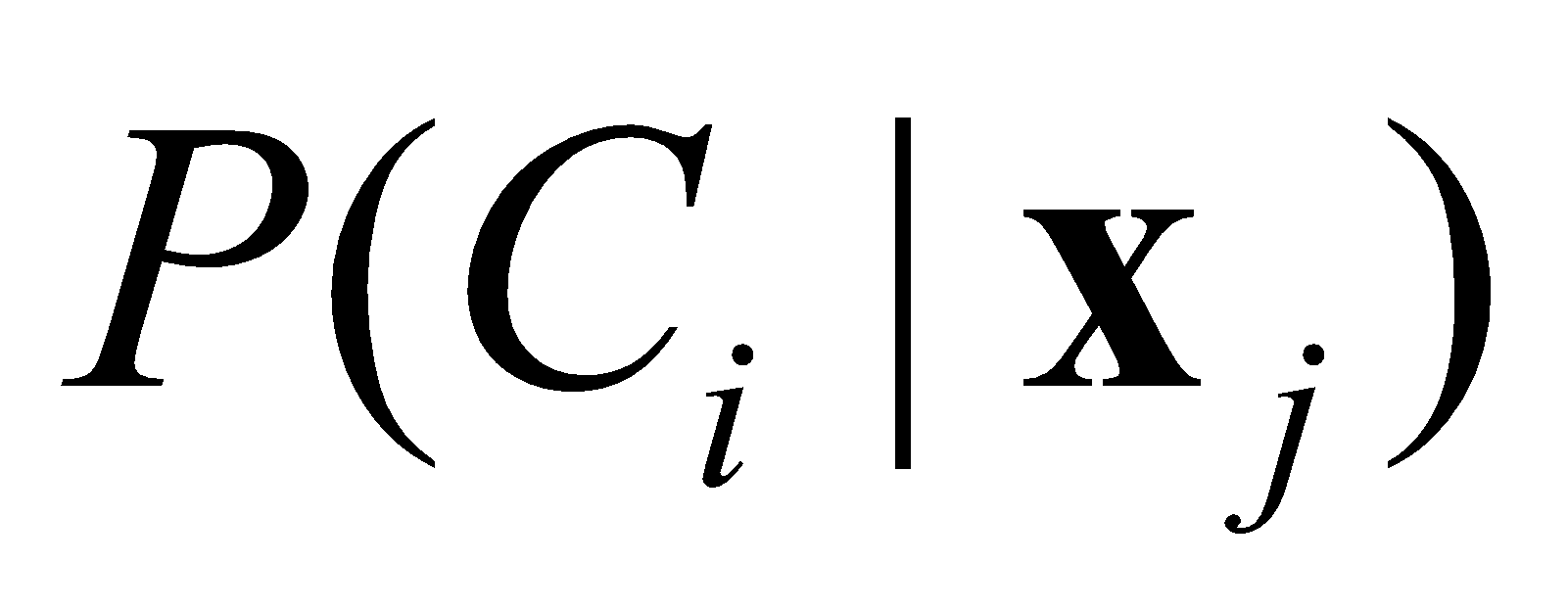
**For Q2:**

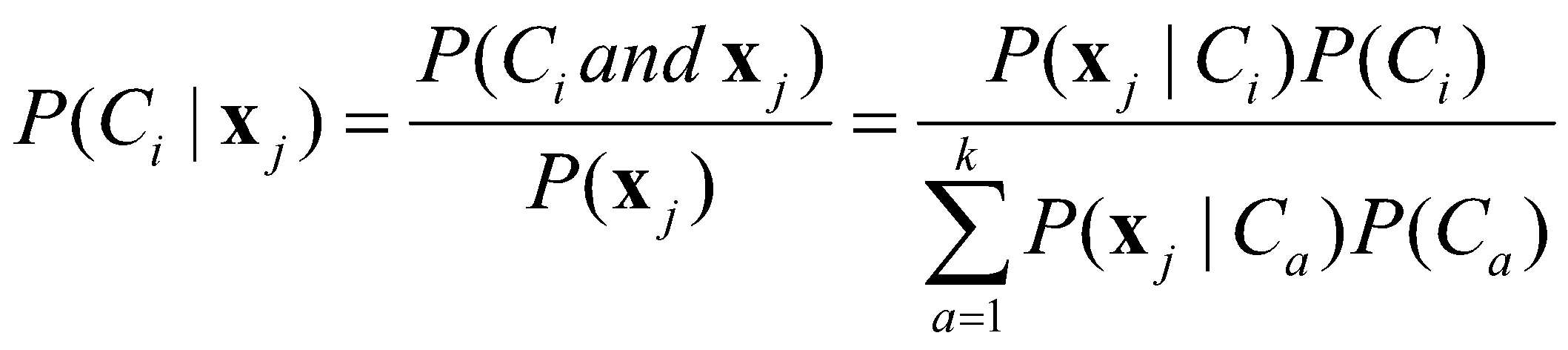
**Method and Explained:**

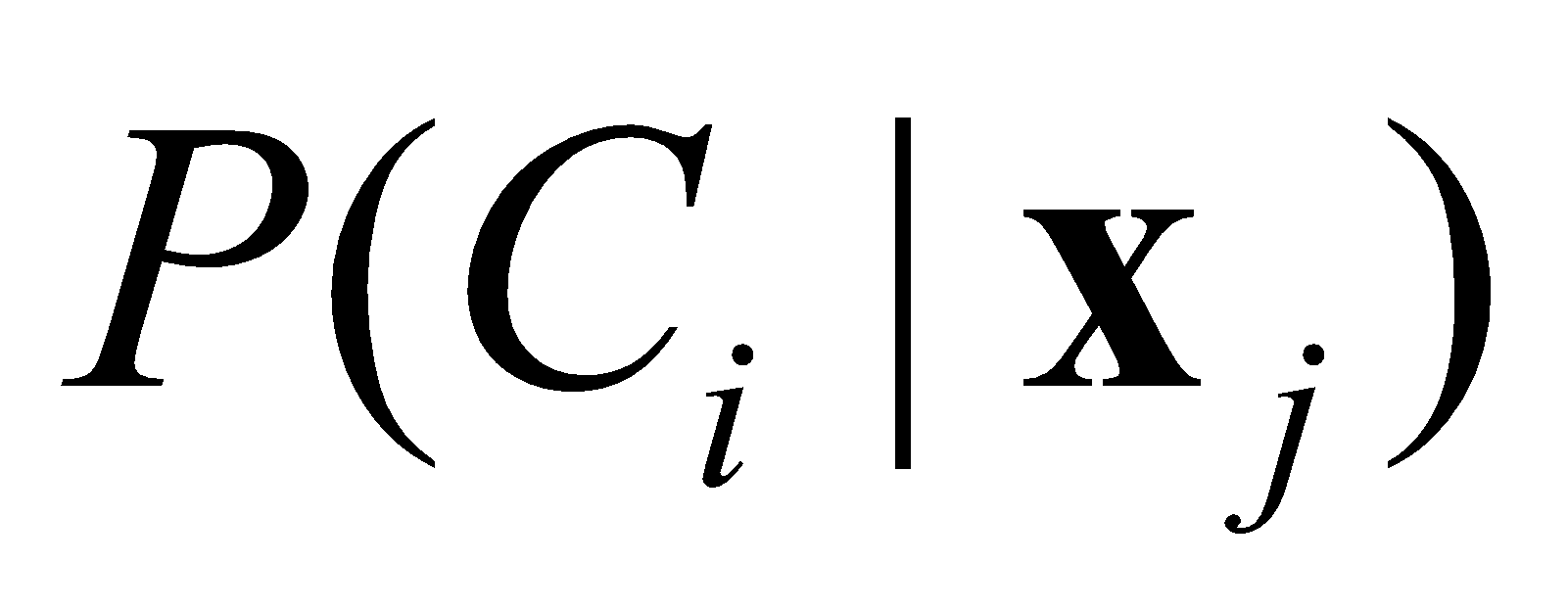
Expectation maximization (EM) algorithm to solve the question 2.

EM algorithm is a soft assignment of points to clusters, so that each point has a probability of belonging to each cluster. EM is a two-step iterative approach. It has Expectation (E) step and Maximization (M) step (Zaki and Meira, 380). EM algorithm is usually used for the maximum likelihood estimation in the latent class analysis. It starts from an initial guess for the parameters in model. Given the current estimates for , E-step is to compute the cluster posterior probabilities via Bayes theorem. Next, M-step uses the  to re-estimates the parameters  for each cluster (Zaki and Meira, 381).

Assume there are k clusters:  and the specific steps are as following.

1. Initialize parameters including mean, covariance matrix and prior probability for each cluster. (set random values for means, set 1 to the diagonal 0 for others in covariance matrix, set uniform prior probability)
2. E-step: compute the cluster posterior probabilities  by using the below formula:



1. M-step: use  to re-estimate the parameters for each cluster.
2. Repeat the E-step and M-step until algorithm convergence.

**Data satisfaction of requirements of method:**

The clustering problem could be solved by getting the maximum likelihood of parameters or getting the maximum log likelihood of parameters, but directly to obtain the maximum likelihood is very hard. In this case, we could use EM algorithm to solve this problem. There are unknown parameters and known data observations.

**Method applied and interpreted correctly:**

Apply the EM method to our data. We assume there are 7 clusters because in our data there are 7 actual class labels. In this way, we could test the performance of EM algorithm easily. Test performance will be written later. Begin with the assumption that the first 15 instances are in cluster 1, next 15 instances in cluster 2, etc. Accordingly initialize the means, covariance matrixes and prior probabilities. The stop condition is the Euclidean distance between the old means and new means less than 0.001. We do not use the package in R and we write code by ourselves. The code will be shown in appendix. The clustering result against PC1 and PC2 is shown as fig8. The distinct colors mean different clusters.

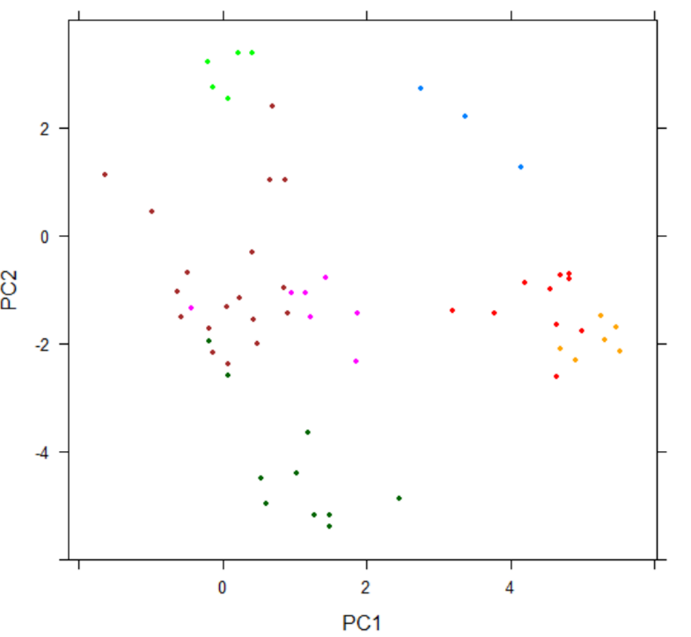


Fig8. Clustering result against PC1 and PC2

Next, test the performance of EM algorithm. Fig9 is a contingency table for actual partitions in data and clusters by using EM algorithm.

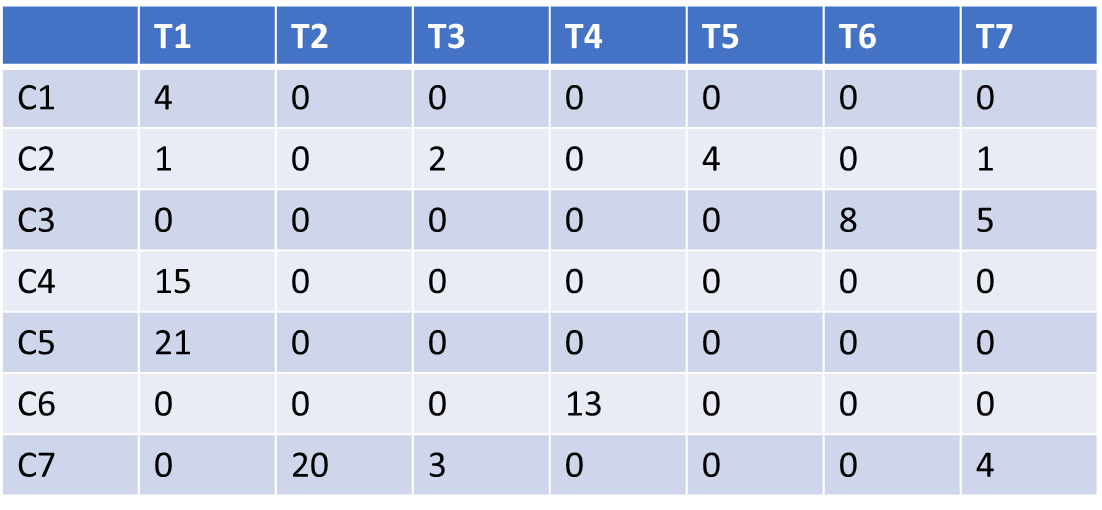
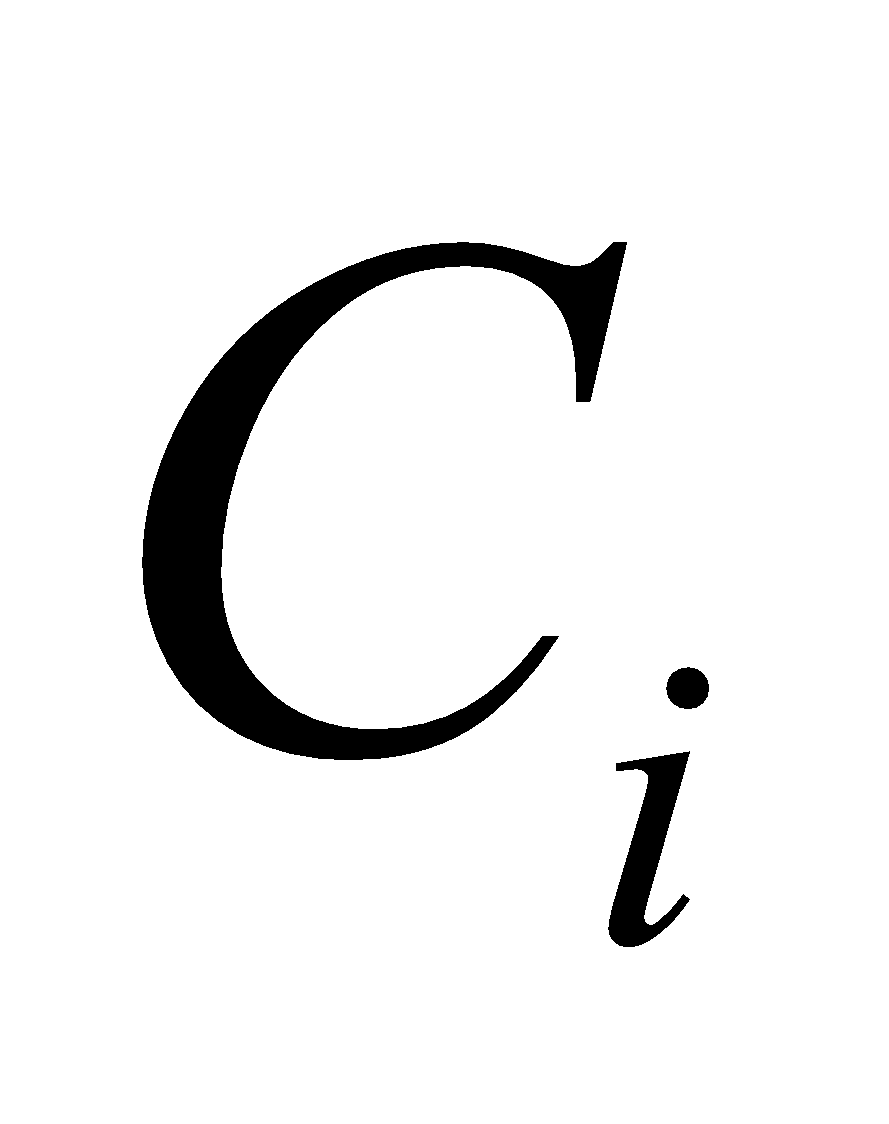


Fig9. Contingence table

We use external validation measures because we know actual class labels for each point. We compute the purity and F-measure. Purity quantifies the extent to which a cluster  contains entities from only one partition. In other words, it measures how “pure” each cluster is. F-measure tries to balance the precision and recall values across all the clusters. For a perfect clustering, when r = k, the maximum value of the F-measure is one ((Zaki and Meira, 474). We got the purity from values in fig8 is 0.84 and the F-measure is 0.67. It could say we get a good clustering by using EM algorithm.

**For Q3:**

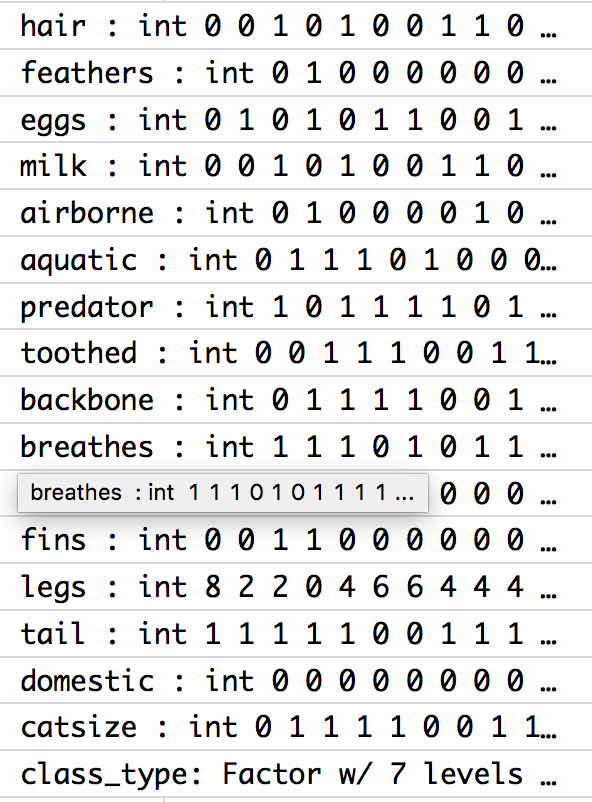
**Method and Explained:**

Naive Bayes algorithm to solve 3rd question based on the features of animal.

Naive Bayes algorithms is to predict the categorical variable. Naive Bayes algorithm can be applied for binary, numeric and categorical response variables.

**Data satisfaction of requirements of method:**

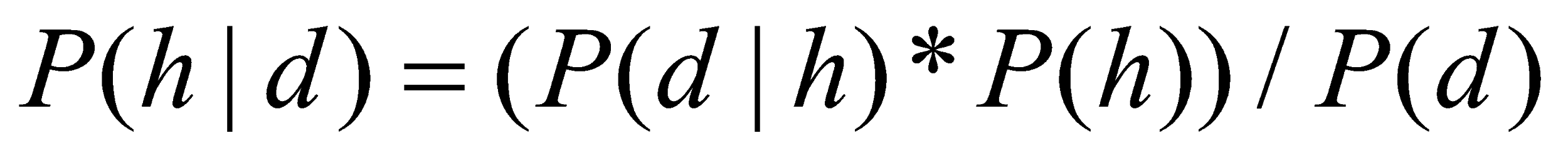
The dataset includes 1 categorical response variable (Animal Type). The dataset also includes 15 binary variables and 1 numeric variable. The dataset is shown below. So, Naive Bayes algorithm is adaptive.



**Method applied and interpreted correctly:**

The assumption of Naive Bayes algorithm is that each response variable is independent and they contribute equally to the animal type.

The formula to Naive Bayes algorithm is:



Apply Naive Bayes algorithms to our data:

* Divide the dataset into training and testing data. There are 60 instances in the training data and 41 instances in the testing data.
* Calculating the prior probabilities (using training data):
  + For binary response variables: using “P(class=1) = count(class=1) / (count(class=0) + count(class=1))” (Jason, Naive Bayes for Machine Learning)
  + For numeric response variables: using number of instance with label c1/ number of instances
* Calculating conditional probabilities (using training data):
  + For binary response variables: “The conditional probabilities are the frequency of each attribute value for a given class value divided by the frequency of instances with that class value.” (Jason, Naive Bayes for Machine Learning) For example: P(hair = 1| class = 1) = count (instances with hair = 1 and class = 1)/ count (instances with class = 1)
  + For numeric response variables: calculate the mean and variance for each class and use density function to estimate the conditional probabilities.
* Doing prediction (using testing data):
  + Calculating class probability based on the instance in the testing data by using Naive Bayes theorem.
  + Then we choose the class that has the largest probability as the instance’s class

The contingency table of Naive Bayes algorithm shows as fig10:

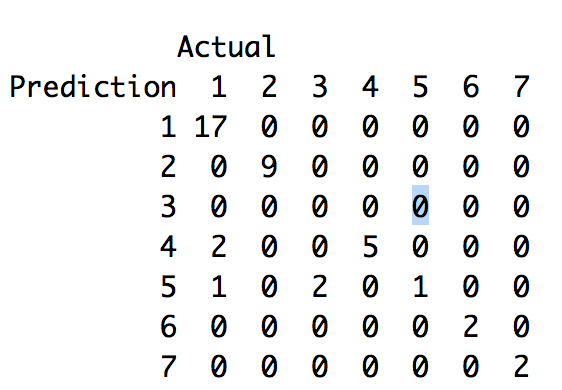


Fig10. The contingency table of Naive Bayes algorithms

The overall accuracy is the count of instance which prediction value equals to actual value divided by the total number of points in test data. The overall accuracy is 0.878. We successfully predict animal type which is class type 1 (mammal), 6 (bug) and 7 (fish). We think Naive Bayes algorithm is really good for this data because the result accuracy is high.

**Conclusion:**

We use Principle Component Analysis method to reduce the 16 feature variables of animals to 3 principle components and still contain as much as original information. We use EM algorithm to cluster animals based on components data and get high purity and F-measure which verify that it is a good clustering. Finally, we successfully predict animal types which are class type 1 (mammal), 6 (bug) and 7 (fish). We think Naive Bayes algorithm is really good for this data because the result accuracy is high.

**Citation**

**Kabacoff, Robert.** *R in action: data analysis and graphics with R*. Manning, 2015

**Zaki, Mohammed J, Meira Jr., Wagner.** *Data Mining and Analysis:Fundamental Concepts and Algorithms.*

**Brownlee, Jason.** “Naive Bayes for Machine Learning.” *Machine Learning Mastery*, 11 Apr.2016. machinelearningmastery.com/naive-bayes-for-machine-learning/.

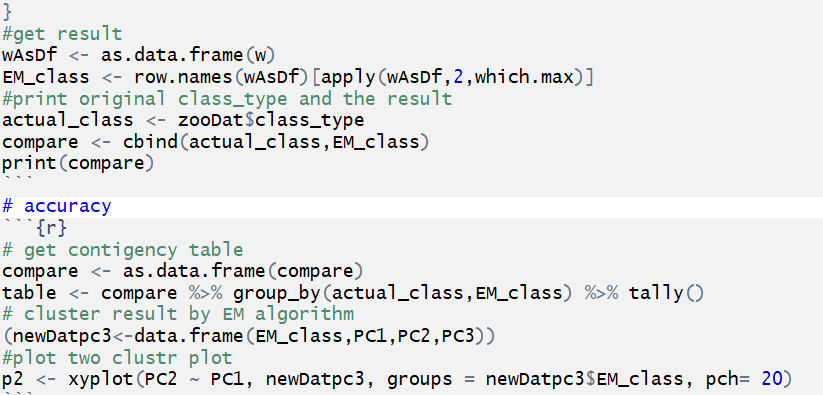
**Appendix**

Main Feature



PCA and EM





Naive Bayes

