Classification of playing styles in football

The use of ball action data

C.J. Wensveen

Master of Science Thesis

TU Delft Delft University of Technology

ORTEC SPORTS

Classification of playing styles in football

The use of ball action data

Master of Science Thesis

by

Carolina Joanna Wensveen

to obtain the degree of Master of Science in Applied Mathematics
at the Delft University of Technology,
to be defended publicly on Wednesday October 5, 2016 at 10:00 AM.

Student number:	4063244	
Project duration:	January 4, 2016 – October 5, 2016	
Specialization:	Probability, Risk and Statistics	
Thesis committee:	Prof. dr. ir. G. Jongbloed,	TU Delft, supervisor
	Dr. L. Eveleens,	ORTEC Sports, supervisor
	Dr. D. Kurowicka,	TU Delft

Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS) - Delft University of Technology

TU Delft Delft University of Technology


TU Delft
Delft
University of
Technology


ORTEC SPORTS

Abstract

Coaches can benefit from objective information about playing styles applied in football matches. In this work, two methods for determining (characteristics of) the playing style of a football team in a certain match based on statistics of ball actions are constructed.

The first method assigns a match to one of four commonly applied playing styles in football based on a set of benchmark matches. Within this method, a relevant variable set with respect to the playing style of a team is selected based on the so-called minimum-redundancy-maximum-relevance algorithm. This algorithm makes use of mutual information as a measure of relevance. The mutual information between variables is estimated by the so-called Kraskov and adjusted Kraskov estimator. After a relevant variable set has been found, matches are assigned to one of the four playing styles by the use of a combination of a hierarchical scheme of K-means clustering and 1-Nearest Neighbors.

The second method focuses on general playing style characteristics of matches as opposed to labeling a match with a specific playing style. This way, details about the playing style in a match can be obtained without limiting to the four prior labeled playing styles. Using principal component analysis combined with domain knowledge, three characteristic variables are created which together give a general overview regarding the playing style applied in a match.

Application of both models on different sets of matches show satisfying results which agree with domain knowledge. These models can be used to provide football coaches with information regarding playing styles applied in matches. Coaches can use this information in order to both evaluate their own team as well as analyze their opponents.

Preface

This report describes the process of my graduation project as part of the Master Applied Mathematics at the Delft University of Technology. I carried this project out at ORTEC Sports which is part of the company ORTEC.

Within the MSc program Applied Mathematics I have chosen the specialization Probability, Risk and Statistics, mainly due to my interests in data analysis. Along with my passion for sports this lead me to the area of sports statistics. ORTEC Sports gave me the opportunity to carry out a project in this particular field, so that I was able to combine my two biggest interests optimally.

I would like to thank my supervisor Geurt Jongbloed for the useful advice and guidelines he provided me with throughout the process of my graduation project. Despite his busy schedule, he always made time for our weekly or biweekly meetings. Furthermore, my thanks go out to all my colleagues at ORTEC Sports who have helped me by sharing their domain knowledge about football and professional views with me.

Special thanks go to Laurenz Eveleens, my daily supervisor at ORTEC Sports, who always took the time to guide me in the right direction during the entire duration of my project. Apart from my thesis I have added value to ORTEC Sports by contributing to reports weekly and occasionally assisting in small analyses. Combining all of this with the daily table tennis sessions, the collaboration with ORTEC Sports has been very positive for me.

To conclude, I have learned a lot during the last nine months of my study - my knowledge of football has increased sincerely, I have learned many new things regarding data analysis and statistics and I got to see how important it is within a company to always keep in mind the goal of a project and the usefulness for potential clients. Despite some - I think inevitable - ups and downs throughout my project, I look back at the last nine months as a great conclusion of my time as a student at the Delft University of Technology.

Contents

Chapter 1

Objective playing style determination

Mathematics is usually not the first thing one thinks about in relation to sports. However, an increasing amount of scientific research in the area of sports shows that statistics along with data analyses and mathematical modeling are becoming more and more important in the field of sports. An example of the use of statistics in sports is shown in the well-known book and movie *Moneyball*. The book tells a story in which Billy Beane, manager of the baseball team Oakland Athletics, uses statistics such as stolen bases and batting averages to pick players for his team in order to be able to better compete against richer opponents. Another example is written about by Kjäll (2015), who talks about the Danish football team FC Midtjylland, which is an example of a team which nowadays operates in a similar way using sports statistics. These examples show that statistics are becoming increasingly important in sports.

ORTEC Sports responds to the increased need of statistics in sports by carrying out analyses of sports, in particular football. Based on registration of ball moments[1], information and insight about the performance, strength and weaknesses of players and teams is gained. ORTEC Sports is owner of a large number of statistics on the level of players and teams which give insight in specific qualities. Using these statistics, ORTEC Sports supports coaches by providing them with information which can increase their odds of winning.

In sports various aspects play an important role, such as physical, mental and technical strength, physical fitness and playing style. Especially in football, the playing style a team applies is relevant. A team will always apply the style of play with which they think their odds of winning are highest. The coach of a team is responsible for deciding on which style of play his team will apply during a match. A reason for deciding on a certain playing style can be that the coach believes the opponent of the team can best be defeated by using this specific playing style. Other reasons can be the form the team is in and the availability of the players. Coaches analyze teams and players to make statements about their playing styles. This way of working is quite subjective, since different coaches may have different opinions as to how different playing styles can be spotted. Coaches could benefit from information by an objective source about the playing styles of teams in order to have the best chance of defeating their opponents.

1.1 Problem description

As mentioned, it would be interesting if the playing styles of teams could be determined in an objective way, for example, by using statistics of ball moments. The question now is whether the large amount of statistics which ORTEC Sports owns, can be used for this purpose. The goal of this project is the following:

"Determine the playing style of a football team in a certain match based on statistics of ball moments."

In this study this problem will be subdivided into four different sub questions, which are all based on the usage of ball action statistics:

- Which possible styles of play exist and which properties do they have according to the literature?

[1]Ball moments are all moments during a match in which an action is performed on the ball. Registration of these ball moments can be in the form of, for example, a pass. From now on the words ball moment and ball action will be used interchangeably in this thesis.

- Can exploratory data analysis be used in order to get a first idea of the different playing styles in the data?
- Which statistics are especially important in determining differences in styles of play applied by teams in different matches?
- How can different matches of teams be subdivided into groups with similar styles of play?

The following section addresses the first sub question by giving a description of playing styles in general and of some common playing styles in football. Furthermore, research is done in this section as to how similar problems as the research goal in this study have been dealt with in past studies. Based on this research, the choice is made to use K-means clustering as classification method in order to assign matches of teams to different playing styles.

1.2 Literature study

This section gives a digression of playing styles in general as well as examples of common playing styles in football. Furthermore, an overview is given about the way playing styles-related problems have been dealt with in the past.

1.2.1 Styles of play

Author's note: this subsection is confidential.

1.2.2 Past studies

Much research has been carried out in the area of sports statistics. In particular the field of playing styles in sports has been studied quite thoroughly already. Examples of studies in this particular field are (Castellano et al., 2012), (Lago-Peñas et al., 2010), (Lorenzo Calvo et al., 2010), (Moura et al., 2014), (Wang, 2014), (Grunz et al., 2012), (Jäger and Schöllhorn, 2007), (Jäger and Schöllhorn, 2012), (Kempe et al., 2015), (Mooij, 2013), (Niu et al., 2012), (Pena and Touchette, 2012), (Pfeiffer and Perl, 2006), (Pollard et al., 1988), (Wang and Parameswaran, 2005) and (Wang et al., 2015). However, most of the found studies focus on a slightly different research question than the one in this study. Based on the found articles, a broad subdivision in the area of analysis of playing styles can be made. Part of the studies, up till (Wang, 2014) in the list given above, focus on the determination of match statistics which can discriminate between successful and unsuccessful teams. These research problems are of a supervised kind. Data about the outcome of a match, i.e. the score or winning/loss, are known and can be used in determination of a suitable model.

The second part of the studies, from (Grunz et al., 2012) till (Wang et al., 2015) in the list given above, focuses on the determination of the style of play a team or player applies during a match. Most studies approach this as an unsupervised problem, since experts are needed in order to provide data about the outcome variable, i.e. the playing style of a team or player. This does not only cost a lot of time and money, it also makes the problem less objective. These research problems are similar to the one in this study.

Most studies focusing on the latter type of research problems, however, make use of different types of data than the data in this study. In this research aggregated statistics about ball actions during (specific parts of) the match, such as the number of passes in a match, are used. The other studies mostly use data of consecutive ball actions and/or positional data of all the players. Due to these differences in data, not all methods applied in these studies can also be applied to the problem in this thesis.

Only Pollard et al. (1988) deal with the exact same research problem as in this study. They start by analyzing six statistics about the amount of ball actions during a match, namely *Number of long forward passes*[2], *Number of long goal clearances*[3], *Number of centers*[4], *Number of times possession is regained in attack*[5], *Number of defensive possession moments*[6] and *Number of multi-pass movements*[7]. All these statistics are expressed in

[2]Number of passes (excluding goal clearances) taking the ball at least 30 meters closer to the opponent's goal line.
[3]Number of long forward goal clearances made by a goalkeeper after picking up the ball.
[4]A center is defined as a cross, made at an angle of less than 45 degrees, taking the ball into the central 20 meters of the penalty area.
[5]Number of times possession is regained within 35 meters of the opponent's goal line.
[6]Number of possession moments of three or more completed passes that a team makes in its own half of the field.
[7]Average number of passes per match in all possession moments containing over three completed passes.

percentages. The study applies factor analysis in order to extract a smaller number of underlying features. The resulting factor scores show a clear distinction between two groups of matches corresponding to two different playing styles.

Some frequently called upon methods in the articles listed at the beginning of this section are *Principal Component Analysis (PCA)*, *Factor analysis (FA)*, *Linear Discriminant Analysis (LDA)*, *K-means clustering (K-means)*, *Artificial neural networks (ANN)* and *Self-Organizing Maps (SOM)*. There does not seem to be one specific study which produces the best quality of results, so the quality of results cannot be used as an indication of which method is best suited to solve the research problem in this study.

The research problem in this study demands clustering or classification of the objects into groups of similar playing styles. One of the most important aspects of the research question is the interpretability of the results. If a method is found by which the objects can be subdivided into groups of similar playing styles, but the characteristics describing the different groups are not known, i.e. which playing style corresponds to which group is not known, this adds no value. Neural networks often have the disadvantage that the results are hard to interpret. Therefore, these methods will not be called upon in this study.

A choice has to be made between supervised or unsupervised methods. The extra time and money it costs to obtain expert data about the playing styles of teams and the fact that using expert data makes the problem less objective, are disadvantages of treating this research problem as a supervised problem. That is why it would be optimal to treat the problem completely unsupervised. However, it will be shown in chapter 2 that doing so, does not lead to satisfying results. Also, from the results of (Mooij, 2013) it can be concluded that a completely unsupervised approach to the problem in this study will indeed not lead to satisfying results. Some expert knowledge with respect to which statistics are important, has to be used in order to improve the results. After this, the assignment of matches of teams to different playing styles will be done in an unsupervised way. Since K-means clustering is a frequently applied method in similar problems related to playing styles according to literature, as well as a popular unsupervised method in general, the choice is made to use K-means clustering as classification method.

A choice has now been made about which method to use in order to assign matches of teams to different playing styles. Before describing any analyses performed, it is important to give an overview of the data set which is used in this study. The next subsection gives a description of the process of gathering the data, provides an overview of what is in the data set at hand and describes some initial data analyses for preparation of the data for further analyses.

1.3 The ball actions data set

In this section the data set used in this study is described. First of all, the process of gathering the data is shown. Next, an overview is given as to what is contained in the data set used in this study. Lastly, initial data analyses are carried out in order to prepare the data for further analyses.

1.3.1 Data process

Author's note: this subsection is confidential.

1.3.2 Descriptive analysis

Author's note: this subsection is confidential.

1.3.3 Initial data analysis

Author's note: this subsection is confidential.

1.4 Outline

The first sub question in section 1.1 has been addressed in section 1.2.1. Chapter 2 is related to the second sub question. Exploratory data analysis is performed in this chapter in order to get a first idea of the structure and

the playing styles applied in the matches which are contained in the data set. This analysis shows that treating the problem completely unsupervised does not lead to satisfying results. Therefore, some expert knowledge has to be utilized in order to improve the results. This is done in chapter 3, which focuses on the third sub question of this study. The relevance of different statistics with respect to the detection of differences between various playing styles is researched in this chapter.

In chapter 4 the last sub question is addressed, i.e. matches of teams will be subdivided in different groups corresponding to different playing styles based on the chosen relevant statistics of ball actions. Also, some general playing style characteristics will be considered, by which properties of playing styles applied in matches can be analyzed. Finally, chapter 5 describes some results when applying the constructed models to different subsets of the data. The remaining chapter contains a conclusion and discussion.

Chapter 2

Exploratory analysis of the data set

Before starting real statistical analyses it is useful to perform some exploratory analysis in order to get an idea of the structure of the data. This way the presence of possible clusters[1] in the data can already be discovered. How many clusters there are exactly, what they represent and which observations lie in which clusters will be of later issue.

To explore the data it is useful to visualize them in some way, preferably in two dimensions for clarity. Ideally, visualizations of the data will show signs of clustering of the objects, preferably signs of five or more different clusters corresponding to the five different playing styles mentioned in subsection 1.2.1 and perhaps other ones. That would give an indication that the features in the ball actions data set are able of discriminating between different styles of play. Since the data set containing the features in *Feature set 1* consists of 106 features, some type of dimensionality reduction has to be performed in order to visualize the data. Principal component analysis (PCA) and multidimensional scaling (MDS) are considered for this goal. The next section gives a description of these methods.

2.1 Dimensionality reduction

In order to visualize high-dimensional data, such as the ball actions data set in this study, some type of dimensionality reduction has to be performed. For this goal, principal component analysis (PCA) is used. As comparison, also multidimensional scaling (MDS) is considered.

2.1.1 Principal component analysis

Principal component analysis is a method to transform a high-dimensional data set into a set of new, linearly uncorrelated variables, i.e. components, which explain a decreasing amount of variance in the original data set. Selecting only the first k components gives a lower-dimensional representation of the original data set which explains the most variance possible in k dimensions.

More formally, let $\mathbf{X} = (X_1, \ldots, X_p)$ be a random vector in $\mathbb{R}^p$. PCA aims to transform these p random variables into a set of p components, $Z_1, \ldots, Z_p$, with $Z_i = \alpha_{i,1}X_1 + \alpha_{i,2}X_2 + \cdots + \alpha_{i,p}X_p$, $i = 1, \ldots, p$, such that $\mathrm{Var}(Z_1) > \mathrm{Var}(Z_2) > \ldots > \mathrm{Var}(Z_p)$ and $\mathrm{Cov}(Z_i, Z_j) = 0$ for $i \neq j$. The elements $\alpha_{i,1}, \ldots, \alpha_{i,p}$, $i = 1, \ldots, p$, are called the *loadings* of the i^{th} principal component. For these loadings $\sum_{j=1}^{p} \alpha_{i,j}^2 = 1$ holds[2]. These loadings can be estimated based on n realizations of the random vector $\mathbf{X}$, denoted by $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$ with $\mathbf{x}_j$ the vector containing n realizations of the variable X_j, and using the corresponding covariance matrix $\mathbf{\Sigma}_n$. Algorithm 1 describes the process of estimating the principal components $Z_1, \ldots, Z_p$ based on these realizations. Denote by $z_{i,j}$, $i = 1, \ldots, n$ the value of the j^{th} principal component corresponding to the i^{th} realization of $\mathbf{X}$. These values are called the principal component *scores*.

[1] In this case clusters are considered groups of matches of teams with similar ball actions characteristics ,possibly corresponding to playing styles.

[2] If the sum of squares of $\alpha_{i,j}$, $j = 1, \ldots, p$ is not set to equal one, they can be set to equal an arbitrarily large value which can lead to an arbitrarily large variance.

Algorithm 1: Principal Component Analysis

Data: $n \times p$-dimensional matrix $\mathbf{X}_n$ corresponding to n realizations of the random vector $\mathbf{X} = (X_1, \ldots, X_p)$ in $\mathbb{R}^p$

Result: Estimations of the principal components $Z_1, \ldots, Z_p$ of the random vector $\mathbf{X}$

Calculate the covariance matrix $\mathbf{\Sigma}_n$ of the matrix $\mathbf{X}_n$;

Find the eigenvector $\boldsymbol{\alpha_k}$ of $\mathbf{\Sigma}_n$ corresponding to the k^{th} largest eigenvalue λ_k;

Estimate the k^{th} principal component by $Z_k = \boldsymbol{\alpha_k}^T \mathbf{X}$;

A derivation of algorithm 1 is given in appendix B.

Before performing PCA on a data set, it can be desirable to scale the data in order for variables not to have disproportional influence. This disproportionality can arise from differences in magnitude of variances of the variables. A variable with high variance will weigh higher in a principal component analyses than a variable with low variance, which might not be wishful. For example, in the ball actions data set the variable *Total possession time in milliseconds* will weigh much higher than the variable *Percentage of long passes*, due the the units of measurement , whilst it is wishful for them to have equally much influence. A way of overcoming this problem is by scaling the variables. There are various ways of doing this. The most common way of scaling is called *standard scaling*. Consider observation i with value $x_{1,i}$ for variable X_1. This value is scaled to $\frac{x_{1,i}-\mu_1}{\sigma_1}$ with μ_1 the sample mean of variable X_1 based on a set of realizations and σ_1 its sample standard deviation. The parameters μ_1 and σ_1 are called the scaling parameters. This way of scaling makes sure all observations are transformed to the signed number of standard deviations an observation lies above the mean. Disproportionally high variances and means will now not lead to disproportionally high influence in analyses anymore.

Note, however, that it might not be wishful for each variable to have equally much influence in analyses. Consider a variable, such as *Number of penalties*, which almost always takes on the value 0, but occasionally also the value 1 or 2. Originally, this variable has relatively low variance. It is probably not desirable for this variable to have equally much influence in analyses regarding playing style as, for example, the variable *Percentage of long passes*. However, when scaling this variable using the standard scaling procedure described above, the resulting variance and mean is equal to that of the other variables, which makes its influence equally high. To overcome issues like this, weighted scaling procedures could be used. This way, the variables are scaled using weigh factors which can increase or decrease the influence of important or less important variables. In this study, such weighted scaling is not utilized. It should be kept in mind, however, that this might improve the final results which is why it would be an interesting adjustment for further studies.

The ball actions data set is thus scaled using standard scaling. From now on, this scaled data set is used throughout the remaining part of this study. Data of new matches not contained in this data set yet, should be scaled using these same, fixed scaling parameters. After performing PCA, the first couple of principal components can be used to visualize the data. Mostly it suffices to only consider the first few components, since principal component analysis makes sure the principal components explain a decreasing amount of the total variance in the original data set. Therefore, in the case of the ball actions data set, the last principal component will account for a very small amount of the original variance and will therefore probably not add much information with respect to the playing style of a team. The first couple of components, however, have more discriminative value.

As an example, figure 2.1 shows a plot of the first two principal components resulting from performing PCA on the build-in data set *iris* in R[3]. This data set contains data about the sepal length, sepal width, petal length and petal width of 150 flowers from the species *setosa*, *virginica* and *versicolor*.

[3]R is used as programming language throughout this thesis.

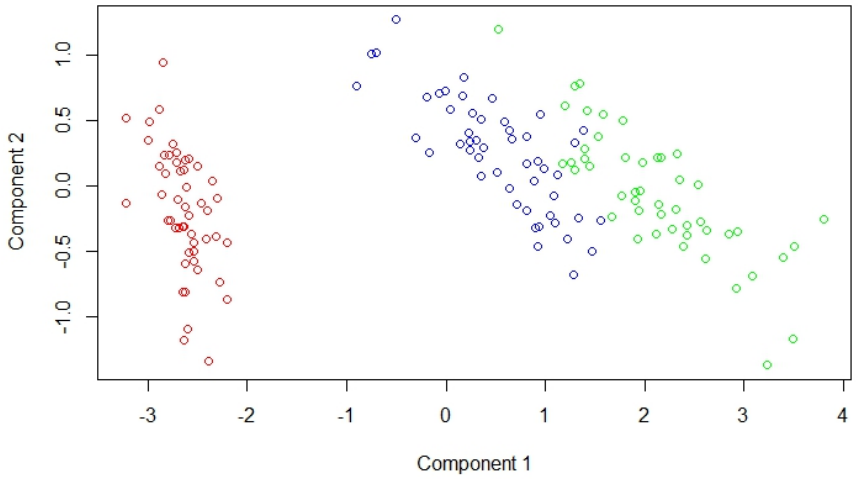

Figure 2.1: Plot of the first two principal components of the data set *iris*. Red dots correspond to flowers from the type *setosa*, green dots to the type *virginica* and blue dots to the type *versicolor*. The two components together account for 97.8% of the variance in the original data set.

It can be seen that the different species are well separated in this plot, mainly by component 1. Note that the first and second principal component account for almost all of the variance in the original data set, namely 97.8%. Figure 2.2 shows a similar plot only now for the second and third component.

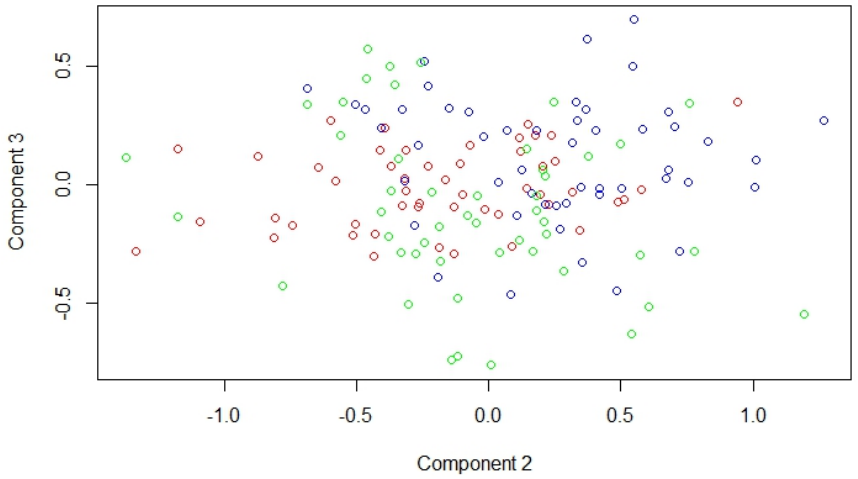

Figure 2.2: Plot of the second and third principal component of the data set *iris*. Red dots correspond to flowers from the type *setosa*, green dots to the type *virginica* and blue dots to the type *versicolor*. The two components together account for 7.0% of the variance in the original data set.

In this plot it can be seen that the species are not well separated. Note that the second and third principal component account for almost none of the variance in the original data set, namely 7.0%. This shows that it suffices to only use the first principal components, in this case mainly the first, in order to detect different groups in the data. In subsection 2.2 similar plots will be shown for the ball actions data set.

Even though relatively only little information is lost with respect to the original data set when considering only the first few principal components, for comparison also another dimensionality reduction technique will be considered. The next subsection describes the method multidimensional scaling.

2.1.2 Multidimensional scaling

When performing a dimension reduction technique on a high-dimensional data set, one is bound to lose information obtained in the original data set. This also holds for PCA, in the case of considering only the first couple of principal components. Therefore, it is desirable to also consider another dimensionality reduction technique in order to be more sure about the structure of the data, such as the presence of possible clusters. The technique multidimensional scaling (MDS) will therefore be considered as well. MDS is a method with which a lower-dimensional representation of the original data is found by trying to preserve the original pairwise distances as well as possible. So, if two data points in the original dimensional space are relatively close together, they should be represented by two points in a lower-dimensional space which are also relatively close together. Points which are far away from each other in the original space, should be represented by points which are far away from each other in the lower-dimensional space as well.

Multidimensional scaling can be seen as a class of techniques. That is to say, different variations of this technique exist, such as *classical* MDS, *metric* MDS and *non-metric* MDS. Furthermore, within these methods various distance measures can be used. The most important difference between PCA and MDS is the fact that PCA is based on variances of the data whereas MDS utilizes the distances or (dis)similarities of the data. Hastie et al. (2011, Ch. 14.8, pg. 571) note that when *classical* multidimensional scaling is used, i.e. when the similarities used in MDS are calculated by centered inner products, the coordinates resulting from MDS are exactly equal to the scores resulting from PCA. A proof of this statement is outside the scope of this thesis. Since MDS is considered in order to have a second lower-dimensional representation of the data apart from the representation obtained by PCA, it is not useful to consider classical MDS which gives the same results as PCA. The choice is made to consider metric MDS, which utilizes the actual distances of the data as opposed to non-metric MDS in which rankings of the distances are used.

As mentioned, so called metric multidimensional scaling is considered. In metric multidimensional scaling a lower-dimensional representation is found which preserves the original pairwise distances as well as possible[4]. Algorithm 2 describes how the coordinates of metric multidimensional scaling using Euclidean distances can be found.

Algorithm 2: Metric multidimensional scaling using Euclidean distances

Data: $n \times p$-dimensional matrix $\mathbf{X}_n$ representing n p-dimensional data points $x_1, \ldots, x_n$.
Result: k-dimensional representation of the original data $\mathbf{X}_n$ which preserves the original pairwise Euclidean distances as well as possible, with $k < p$.
Find a random mapping of n k-dimensional data points $z_1, \ldots, z_n$ by sampling from a normal distribution;
Calculate the pairwise distances $||z_i - z_j||_2$, $i, j = 1, \ldots, n$;
Calculate the stress function $S(z_1, \ldots, z_n) = \frac{\sum_{i \neq j} (||z_i - z_j||_2 - d_{i,j})^2}{\sum_{i \neq j} d_{i,j}}$ with $d_{i,j}$ the Euclidean distance between the original p-dimensional data points x_i and x_j;
while *Stress function is larger than some criterion* **do**
  Find a new mapping of the points $z_1, \ldots, z_n$;
  Recalculate the pairwise distances $||z_i - z_j||_2$, $i, j = 1, \ldots, n$;
  Recalculate the stress function $S(z_1, \ldots, z_n) = \frac{\sum_{i \neq j} (||z_i - z_j||_2 - d_{i,j})^2}{\sum_{i \neq j} d_{i,j}}$;
end

Note that in algorithm 2 Euclidean distance is used. This can be adjusted to any distance measure suitable to the problem at hand. Furthermore, note that algorithm 2 does not give any details about the procedure of finding a new low-dimensional mapping $z_1, \ldots, z_n$ leading to a smaller stress function. A common procedure for this is called *stress majorization*, but this is outside the scope of this thesis.

As an example of MDS, figure 2.3 shows a plot of the coordinates resulting from metric MDS performed on the *iris* data set using Euclidean distance.

[4]Note that just as with PCA scaled data should be used in order to overcome the problem of variables with disproportionally high variances to have disproportionally high influence in the multidimensional scaling process.

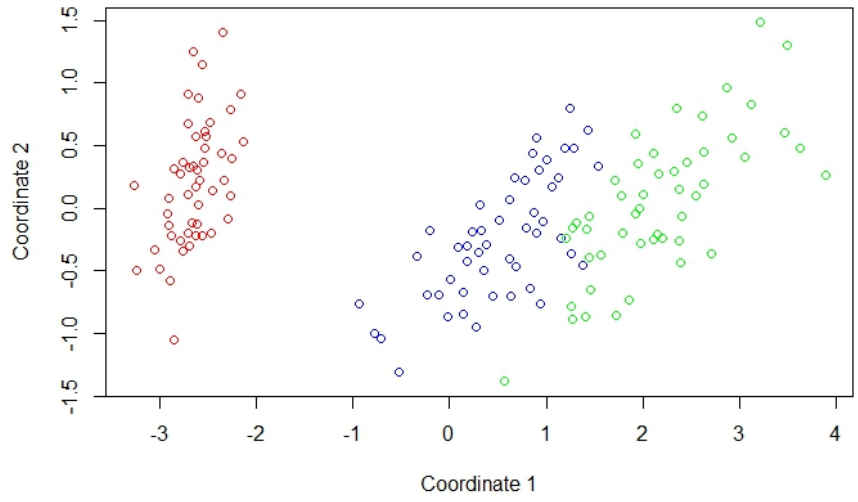

Figure 2.3: Plot of the two coordinates resulting from metric MDS performed on the data set *iris*. Red dots correspond to flowers from the type *setosa*, green dots to the type *virginica* and blue dots to the type *versicolor*.

It can be seen that the different species are again well separated in this plot. Note that reflection in the y-axis makes the plot look quite similar to the plot in figure 2.1, but investigating the plots closely shows that there are small differences.

Now that two dimensionality reduction methods have been considered, the next section applies these methods to subsets of matches in the ball actions data set in order to visualize and, thereby, explore the data. In the case of metric MDS, Euclidean distance is used.

2.2 Application on the ball actions data set

Author's note: this section is confidential.

Chapter 3

Information based feature selection

Chapter 2 showed that the features in *Feature set 1* are not well enough capable of discriminating between the five playing styles in subsection 1.2.1. In this chapter, the aim is to alter the set of features in order to obtain better discrimination. This new feature set can be formed both by adding new features as well as removing current features.

First of all, adding extra features to *Feature set 1* could improve the separation between different playing styles. The features mentioned in subsection 1.3.2 are therefore added to *Feature set 1*. These features were chosen based on domain knowledge, since they might be relevant with respect to the playing style of a team. After adding these new features the data set contains a total of 121 features, denoted by *Feature set 2*[1].

Similar visualizations of the set of matches of the benchmark teams as before are shown in order to check whether the separation between the different playing styles is improved by adding the new features. Again the standard scaled data are used. Table **??**, **??** and **??** in appendix D show the features which have the top 10 highest absolute loadings on the first three principal components, resulting from principal component analysis, along with an indication of a positive or negative loading of that feature. Given the top ten loadings on the principal components, the first component mainly says something about the ball possession of the team. The second component can be seen as a measure of offensive play and the third component as a measure of overall strength. Figure 3.1 shows a scatterplot of the first two principal components.

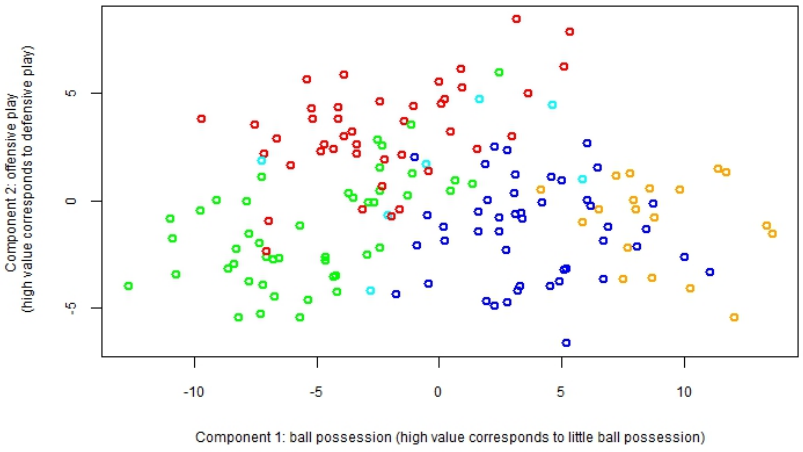

Figure 3.1: Scatterplot of the first two principal components when performing principal component analysis on the home matches of the benchmark teams using *Feature set 2* and standard scaling. These two components explain 32.7% of the variance in the original data set.

It can be seen that the separation between the matches of the benchmark teams is more clear in this plot

[1]Note that this set of new features does not contain any multicollinearity, so that no multicollinear features have to be removed.

than in the plot in figure **??**. The matches of FC Barcelona, Ajax, Atletico Madrid and Queens Park Rangers are all quite well separated now. Only the matches of Juventus do not seem to cluster together. This verifies the assumption that, as expected by domain knowledge, the new features are indeed relevant with respect to discriminating between different playing styles. Apparently, these new features were needed in order to be able to discriminate between the different playing styles more clearly.

Adding the new features made the separation between the matches of the five different teams more clear. However, this separation might still be improved. That is why the next section considers using only a selection of the features. If the reader is not interested in mathematical details, the remaining part of this chapter can be skipped.

3.1 Feature selection

Instead of adding new features, removing features from the ball actions data set could also improve the separation between different playing styles. Most of the studies considered in literature study about analysis of playing styles do not apply feature selection. Only the studies in which the goal was to find the most discriminative variables with respect to successful and unsuccessful teams deal with feature selection. The studies in which the goal was to determine playing styles, do not perform feature selection[2]. However, feature selection can improve final results drastically, since selecting a relevant subset of features reduces the effect of noisy features to the analysis.

There are many different feature selection methods. First of all, a choice should be made between supervised or unsupervised feature selection methods. Notice that the set of matches of the benchmark teams could be used in combination with a supervised method in order to select relevant features. However, the assumption has been made that the benchmark teams play according to their assumed playing style in *most* of their matches, so not necessarily in all. In other words, Ajax does not necessarily play according to the playing style *Hollandse School* in *all* of their matches. Therefore, labeling all of the matches of Ajax as *Hollandse School* could be wrong and using a supervised method could therefore lead to wrong results. However, if correct labels are known, supervised methods generally lead to more satisfying results than unsupervised methods, since in case of the former more prior information can be used. Thus, if the assumed labels of the benchmark teams are mostly correct, supervised methods could still lead to better results than unsupervised methods. However, if many of the assumed labels are incorrect, unsupervised methods probably lead to better results than supervised methods. Since it is not known how many of the assumed labels of the benchmark teams are correct, the choice is made to use a combination of supervised and unsupervised methods in order to determine the playing style applied in a match. For feature selection the choice is made to use a supervised method, since the assumption is made that in this interim step of the analysis, usage of wrong information causes less harm than in the final step of assigning matches of teams to playing styles. The actual assignment of matches to playing styles will, therefore, be done with an unsupervised method. In conclusion, the choice is made to make use of a supervised feature selection method by using the benchmark teams with their assumed playing styles as training data.

Within the variety of feature selection methods, a distinction can be made between *wrapper* and *filter* methods. Wrapper methods select features based on the results they give when used in a given classifier. The selected features are therefore dependent upon the choice of classifier. When using another classifier the chosen features might not be optimal anymore. Filter methods are, on the contrary, independent of any classifier. They select features based on certain statistical criteria, such as correlation or mutual information. Both of these type of methods have advantages and disadvantages. According to Alelyani et al. (2013) wrapper methods have been shown to outperform filter methods in terms of classification accuracy. However, in the case of a small amount of training data, wrapper methods have a high risk of overfitting. Filter methods, on the contrary, are robust to overfitting, as mentioned in Guyon and Elisseeff (2003). Also, if various classifiers should be tested because it is not known yet which classifier suits the problem at hand best, a filter method does not have to be performed over again each time a different classifier is used. A wrapper method, on the contrary, has to be applied again each time another classifier is considered due to the dependence on the classifier. Due to the advantages and disadvantages of both wrapper and filter methods, the choice is made to use a combination of both types of methods. First of all, a filter method will be used in order to find a ranking of the features such that the top k features are best capable of discriminating between the different playing styles of the benchmark teams. Next, a wrapper method will be applied in order to find which number of features k should best be selected.

[2]Some of these studies do perform feature extraction though, which is the process of transforming variables into a possibly smaller set of new features. Feature selection, on the other hand, does not alter the original variables. For interpretation purposes this study only focuses on feature selection.

It would be optimal to select a subset of features based on considering all possible subsets. However, this is extremely costly with respect to computational time. In the case of p original features in the data set, there are 2^p possible subsets. In the case of *Feature set 2*, this leads to a total number of possible subsets of $2^{121} \approx 2.7 \times 10^{36}$. Even if running the classifier method or calculating some statistic (wrapper versus filter methods) would only take 0.01 seconds for each feature subset, this would result in a total computational time of approximately 8.4×10^{26} years which is clearly not feasible.

Since considering all possible feature subsets is not computationally feasible, another way has to be found to search through the possible feature subsets. Two possible ways are to apply so-called forward selection or backward elimination. With forward selection one starts with an empty feature subset. A new feature, which results in the best subset of k features, with k the number of features in iteration k, is added to the feature subset in each iteration. This can be continued until a fixed number of features is achieved or until some stopping criterion has been reached. Backward elimination works the other way around, starting with a feature subset containing all features. One feature, which results in the best subset of k features, with k the number of features in iteration $p - k$, is removed from the feature subset in each iteration. This can be continued until a fixed number of features is achieved or until some stopping criterion has been reached.

Backward elimination and forward selection do not necessarily give the best subset. Considering forward selection, the best 1-feature model could contain feature x_1, whereas the best 2-feature model could contain features x_2 and x_3. Feature x_1 has already been added to the set then, so this optimal 2-feature subset can not be found using forward feature selection. The same holds for backward elimination. The best k-feature model could contain features $x_1, \ldots, x_k$, whereas the best $k-1$-feature model could contain features $x_1, \ldots, x_{k-2}, x_{k+1}$. Feature x_{k+1} has already been removed from the set then, so this optimal $k-1$-feature subset can not be found using backward feature elimination. Despite these disadvantages, forward selection and backward elimination are frequently used selection methods in the case that considering all possible subsets is not feasible.

Since considering all possible feature subsets is not computationally feasible, forward or backward feature selection will be used. The choice between these two depends on the number of desired features to select. If this number is small with respect to the total number of features, it makes more sense to use forward selection, since this method takes less time in this case. When the desired number of features to select is high with respect to the total number of features, backward elimination makes more sense for the same reason. For the ball actions data set it can be expected that the relevant number of features with respect to the applied playing style of a team is relatively small compared to the total number of features. Therefore, forward feature selection will be used.

In summary, a supervised forward, filter selection method will be used to find an appropriate ranking of the features. A wrapper method will then be used in order to choose the best number of features to select. According to Dash and Liu (1997) the statistical criteria used in filter methods can be subdivided in four categories, namely *distance*, *correlation*, *consistency* and *information*. Huang et al. (2007) note that the former three are all sensitive to noise and outliers, whereas for information measures this is not so much the case. Furthermore, correlation measures only take into account linear relations, whereas information measures are not limited to linearity. For this reason, the choice is made to apply a filter method based on an information measure. In order to measure the information that a feature gives about the output variable, i.e. the playing style of a team, *mutual information* will be evaluated. The higher this mutual information between a feature and the output variable, the more relevant the feature is with respect to the determination of the playing style of a team. The next section gives a more detailed description of mutual information.

3.2 Mutual information

The information that one variable gives about another variable shows how relevant one variable is with respect to predicting the other. For example, the information that the variable *Percentage of own goals* gives about the variable *Playing style* is probably quite low, whereas the information that *Total possession time in milliseconds* gives about *Playing style* is probably quite high. In order to measure the information between two random variables, the *mutual information* between these variables can be calculated. Mutual information between two variables is a quantification of the amount of information that one variable gives about the other (and vice versa). Similarly, mutual information between a set of variables and another variable is a quantification of the amount of information which that set of variables gives about the other variable. In other words, it is a measure

of predictability of a variable when the value of (an)other variable(s) is known.

Definition 1 and 2, as given in Yeung (2008), give formal definitions of the mutual information between two, respectively, k continuous variables. In the case of discrete or categorical random variables the integrals in these definitions are replaced by summation signs[3]. In the case of the use of log base 2, the mutual information is calculated in *bits*. Appendix E.1 gives some additional definitions and theorems with respect to mutual information which will be utilized later on.

Definition 1. *Let X and Y be two continuous random variables with possible outcome sets, respectively, Ω_X and Ω_Y. Let p_X and p_Y be the probability density functions of X and Y respectively and let p_{XY} be the joint probability density function of (X, Y). The mutual information of X and Y, $I(X;Y)$, is defined as*

$$I(X;Y) = \int_{\Omega_X} \int_{\Omega_Y} p_{XY}(x,y) \log \left(\frac{p_{XY}(x,y)}{p_X(x) p_Y(y)} \right) \mathrm{d}y \mathrm{d}x$$

Definition 2. *Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively and let $p_{X_1 \cdots X_k}$ be the joint probability density function of $(X_1, X_2, \ldots, X_k)$ and $p_{X_1 \cdots X_{k-1}}$ the joint probability density function of $(X_1, X_2, \ldots, X_{k-1})$. The multivariate mutual information of $(X_1, \ldots, X_{k-1})$ and X_k, $I(X_1, \ldots, X_{k-1}; X_k)$, is defined as*

$$I(X_1, \ldots, X_{k-1}; X_k) = \int_{\Omega_{X_1}} \cdots \int_{\Omega_{X_k}} p_{X_1 \cdots X_k}(x_1, \ldots, x_k) \log \left(\frac{p_{X_1 \cdots X_k}(x_1, \ldots, x_k)}{p_{X_1 \cdots X_{k-1}}(x_1, \ldots, x_{k-1}) p_{X_k}(x_k)} \right) \mathrm{d}x_k \cdots \mathrm{d}x_1$$

Mutual information can be used in order to find relevant variables with respect to discrimination between different playing styles. A relevant variable in the ball actions data set can be seen as a variable which gives a lot of information about the playing style. In other words, a relevant variable should have a high mutual information value with respect to the output variable *playing style*. Similarly, a relevant subset of features is a subset which has a high mutual information value with respect to the output variable.

The goal is to select a subset of k features $\{X_1, \ldots, X_k\}$ such that $I(X_1, \ldots, X_k; Z)$, with Z the output variable *playing style*, is highest among all other mutual information values between subsets of k features and the output variable Z. However, it is not known beforehand which number of features k should be selected. Therefore, subsets of all sizes $k = 1, \ldots, 121$ should be considered. As mentioned in the previous subsection, however, it is not computationally feasible to consider all possible sets of features. Therefore, forward feature selection will be used in order to select feature subsets of size $k = 1, \ldots, 121$.

In order to calculate the mutual information value $I(X_1, X_2, \ldots, X_k; Z)$, according to definition 2, the joint probability density functions $p_{X_1 \cdots X_k}$ and $p_{X_1 \cdots X_k Z}$ are needed. Since these density functions are not known, they need to be estimated. Various approaches in order to estimate these densities exist, such as binning approaches and kernel densities. Doquire et al. (2012) show, however, that mutual estimation estimates based on such density estimates can be quite inaccurate, especially in the case of many variables.

Instead of estimating the mutual information by using estimates of the probability density functions, other mutual information estimators have been constructed which avoid the use of probability density estimates. Kraskov et al. (2004) developed a mutual information estimator, from now on called the *Kraskov estimator*, which is primarily based on nearest neighbor statistics. This estimator builds on the idea that when the nearest neighbors of one variable (or a subset of variables) correspond to the nearest neighbors of another variable, those variables are related to each other. Doquire et al. (2012) show that this estimator is more accurate than estimators based on density estimates, no matter the number of dimensions. The next section gives more details about the Kraskov estimator.

3.3 Kraskov estimator

Kraskov et al. (2004) constructed a mutual information estimator for which no probability density estimator is needed. The estimator is based on nearest neighbor statistics. The estimator can only be used in the case of

[3] In the case of one/various discrete or categorical and one/various continuous random variable(s), naturally the integral(s) corresponding to the discrete or categorical random variable(s) changes into a summation sign(s) and the integral(s) corresponding to the continuous random variable(s) stays the same.

two (or multiple) continuous variables. Ross (2014) adjusted the Kraskov estimator in such a way that it can also be used in the case of the combination of continuous and categorical variables[4]. This estimator will be called *adjusted Kraskov estimator* from now on. Ross (2014) shows that this estimator is more accurate than the estimator based on binning-based density estimates.

The Kraskov estimator and the adjusted Kraskov estimator between two random variables is based on rewriting the mutual information between the variables as a function of their entropies and estimating these entropies based on nearest neighbor statistics. A mathematical derivation of the Kraskov estimator for the mutual information between two continuous random variables X and Y is given below. Appendix E.4 gives a similar derivation of the adjusted Kraskov estimator. Definitions 3 and 4 give the final expressions for the Kraskov estimator and the adjusted Kraskov estimator.

Let X and Y be two continuous random variables with possible outcome sets, respectively, Ω_X and Ω_Y, joint probability density function p_{XY} and marginal probability density functions p_X with $p_X(x) = \int_{\Omega_Y} p_{XY}(x,y)dy$ and p_Y with $p_Y(y) = \int_{\Omega_X} p_{XY}(x,y)dx$.

From theorem 1 in appendix E.1 it is known that

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= -\int_{\Omega_X} p_X(x)\log(p_X(x))dx - \int_{\Omega_Y} p_Y(y)\log(p_Y(y))dy + \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log(p_{XY}(x,y))dydx \\ &= -\mathbb{E}[\log(p_X(X))] - \mathbb{E}[\log(p_Y(Y))] + \mathbb{E}[\log(p_{XY}(X,Y))] \end{aligned} \tag{3.1}$$

Since p_X, p_Y and $p_{X,Y}$ are not known, equation 3.1 has to be rewritten. First consider $\mathbb{E}[\log(p_X(X))]$.

Consider drawing an independent sample of size n from the random variable X. Assume the i^{th} point in this sample, denoted by x_i, is given. There are $n-1$ remaining points. Choose a fixed, small ϵ. Denote by the random variable $D_X(i)$ the distance between x_i and its k^{th} nearest neighbor.

Let g be the probability density function of $D_X(i)$. Now consider the following lemma:

Lemma 1. *Consider an independent sample of size n drawn from a continuous random variable X. Assume the i^{th} point in this sample, denoted by x_i, is given. Let $D_X(i)$ be the random variable denoting the distance between x_i and its k^{th} nearest neighbor. In that case, the probability density function g of $D_X(i)$ approximately equals*

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} P_3$$

with

$$\begin{aligned} P1 &= \int_{x_i-\epsilon}^{x_i+\epsilon} p_X(x)dx \\ P2 &= 1 - \int_{x_i-\epsilon}^{x_i+\epsilon} p_X(x)dx \\ P3 &= p_X(x_i-\epsilon) + p_X(x_i+\epsilon) \end{aligned}$$

A proof of this lemma is given in appendix E.3.1.

Now, assume $p_X(x)$ is smooth in the interval $[x_i-\epsilon, x_i+\epsilon]$ and ϵ is small. Using lemma 1, the probability density function g of $D_X(i)$ now approximately equals

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} \left(2\epsilon p_X(x_i)\right)^{k-1} 2p_X(x_i)\left(1-2\epsilon p_X(x_i)\right)^{n-k-1} \tag{3.2}$$

It is known that for a random variable X with possible outcome set Ω_X and probability density function f, $\mathbb{E}[X] = \int_{\Omega_X} xf(x)dx$ holds. Using this and equation 3.2, it is found that

$$\begin{aligned} \mathbb{E}[\log(2D_X(i)p_X(x_i))] &= \int_0^\infty \log(2\epsilon p_X(x_i))g(\epsilon)d\epsilon \\ &= C\int_0^\infty \log(2\epsilon p_X(x_i))\left(2\epsilon p_X(x_i)\right)^{k-1} 2p_X(x_i)\left(1-2\epsilon p_X(x_i)\right)^{n-k-1} d\epsilon \end{aligned} \tag{3.3}$$

[4]In the case of the ball actions data set this estimator will be used for the mutual information between the output variable, which is a categorical variable, and the explanatory variables.

with $C = \dfrac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!}$
Setting $q = 2\epsilon p_X(x_i)$, equation 3.3 equals

$$\begin{aligned}
&\frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!}\int_0^1 \log(q)q^{k-1}(2p_X(x_i))(1-q)^{n-k-1}\frac{1}{2p_X(x_i)}dq \\
&= \frac{(n-1)!}{(k-1)!\,(n-k-1)!}\int_0^1 \log(q)q^{k-1}(1-q)^{n-k-1}dq \\
&= \frac{(n-1)!}{(k-1)!\,(n-k-1)!}\frac{\partial B(k,n-k)}{\partial k}
\end{aligned} \tag{3.4}$$

with $B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$ the beta function with $x, y > 0$.
Consider the following lemma:

Lemma 2. *Let $B(x,y)$ be the beta function with x and y positive integers. The following relation holds:*

$$\frac{1}{B(x,y)} = \frac{(x+y-1)!}{(x-1)!\,(y-1)!}$$

A proof of this lemma is added in appendix E.3.2.
Setting $x = k$ and $y = n-k$ in lemma 2, equation 3.4 equals

$$\frac{1}{B(k,n-k)}\frac{\partial B(k,n-k)}{\partial k} \tag{3.5}$$

Now consider the following lemma:

Lemma 3. *Let $B(x,y)$ be the beta function with $x, y > 0$. Let $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ be the digamma function with $\Gamma(x)$ the gamma function, $x > 0$. The following relation holds:*

$$\frac{\partial B(x,y)}{\partial x} = B(x,y)(\psi(x) - \psi(x+y))$$

A proof of this lemma is added in appendix E.3.3.
Setting $x = k$ and $y = n-k$ in lemma 3, equation 3.5 equals

$$\frac{1}{B(k,n-k)}B(k,n-k)(\psi(k) - \psi(n)) = \psi(k) - \psi(n)$$

In conclusion,

$$\mathbb{E}[\log(2D_X(i)p_X(x_i))] = \psi(k) - \psi(n) \tag{3.6}$$

Using equation 3.6, it is now found that

$$\begin{aligned}
\log(p_X(x_i)) &= \mathbb{E}[\log(p_X(x_i))] \\
&= \mathbb{E}[\log(2D_X(i)p_X(x_i)) - \log(2D_X(i))] \\
&= \mathbb{E}[\log(2D_X(i)p_X(x_i))] - \mathbb{E}[\log(2D_X(i))] \\
&= \psi(k) - \psi(n) - \mathbb{E}[\log(2D_X(i))]
\end{aligned} \tag{3.7}$$

Now, assume the points $x_1, \ldots, x_n$ in the sample are all known.
In that case, $\mathbb{E}[\log(2D_X(i))]$ can be estimated as follows:

$$\widehat{\mathbb{E}[\log(2D_X(i))]} = \frac{1}{n}\sum_{i=1}^{n}\log(2d_x(i)) \tag{3.8}$$

with $d_x(i)$ the realization of $D_X(i)$ for the sample $x_1, \ldots, x_n$.
Furthermore, $\mathbb{E}[\log(p_X(X))]$ can be estimated as follows:

$$\widehat{\mathbb{E}[\log(p_X(X))]} = \frac{1}{n}\sum_{i=1}^{n}\log(p_X(x_i)) \tag{3.9}$$

Using equations 3.7, 3.8 and 3.9, it can be concluded that

$$\mathbb{E}[\log(p_X(X))] \approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_x(i))\right) \tag{3.10}$$

Similarly, the following relation holds:

$$\mathbb{E}[\log(p_Y(Y))] \approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_y(i))\right) \tag{3.11}$$

with $d_y(i)$ the realization for the sample $y_1, \ldots, y_n$ of $D_Y(i) = |y_i - Y_{i,k}|$ with $Y_{i,k}$ the k^{th} nearest neighbor of y_i.

Now, only $\mathbb{E}[\log(p_{XY}(X,Y))]$ remains to be estimated.
Again, consider drawing an independent sample of size n from the random vector (X,Y). Assume the i^{th} point in this sample, denoted by (x_i, y_i), is given. Choose a fixed, small ϵ. Denote by $D_{XY}(i)$ the distance between (x_i, y_i) and its k^{th} nearest neighbor, where the distance is calculated using the maximum norm, i.e.

$$||(x_i, y_i) - (x_j, y_j)||_\infty = \max(|x_i - x_j|, |y_i - y_j|)$$

Let g be the probability density function of $D_{XY}(i)$. Consider lemma 4, which is an extension of lemma 1 in two dimensions:

Lemma 4. *Consider an independent sample of size n drawn from a continuous random vector (X,Y). Assume the i^{th} point in this sample, denoted by (x_i, y_i), is given. Let $D_{XY}(i)$ be the random variable denoting the distance between (x_i, y_i) and its k^{th} nearest neighbor, where the distance is calculated using the maximum norm, i.e.*

$$||(x_i, y_i) - (x_j, y_j)||_\infty = \max(|x_i - x_j|, |y_i - y_j|)$$

In that case, the probability density function g of $D_{XY}(i)$ approximately equals

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} P_3$$

with

$$\begin{aligned}
P1 &= \int_{x_i-\epsilon}^{x_i+\epsilon}\int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x,y)dydx \\
P2 &= 1 - \int_{x_i-\epsilon}^{x_i+\epsilon}\int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x,y)dydx \\
P3 &= \int_{x_i-\epsilon}^{x_i+\epsilon} p_{XY}(x, y_i-\epsilon)dx + \int_{x_i-\epsilon}^{x_i+\epsilon} p_{XY}(x, y_i+\epsilon)dx + \int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x_i-\epsilon, y)dy \\
&\quad + \int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x_i+\epsilon, y)dy
\end{aligned}$$

A proof of lemma 4 is given in appendix E.3.4.

Now, assume $p_{XY}(x,y)$ is smooth in the area $[x_i - \epsilon, x_i + \epsilon] \times [y_i - \epsilon, y_i + \epsilon]$ and ϵ is small. Using lemma 4, the probability density function g of $D_{XY}(i)$ now approximately equals

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!}\left((2\epsilon)^2 p_{XY}(x_i,y_i)\right)^{k-1}\left(8\epsilon p_{XY}(x_i,y_i)\right)\left(1-(2\epsilon)^2 p_{XY}(x_i,y_i)\right)^{n-k-1} \tag{3.12}$$

Using equation 3.12, it is now found that

$$\begin{aligned}
\mathbb{E}[\log((2D_{XY}(i))^2 p_{XY}(x_i,y_i))] &= \int_0^\infty \log((2\epsilon)^2 p_{XY}(x_i,y_i)) g(\epsilon)d\epsilon \\
&= C\int_0^\infty \log(q) q^{k-1}(1-q)^{n-k-1} 8\epsilon p_{XY}(x_i,y_i)d\epsilon
\end{aligned} \tag{3.13}$$

with $C = \dfrac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!}$ and $q = (2\epsilon)^2 p_{XY}(x_i, y_i)$.
Transforming to q, equation 3.13 equals

$$\begin{aligned}&\frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!}\int_0^1 \log(q) q^{k-1}(1-q)^{n-k-1} 8\epsilon p_{XY}(x_i,y_i)\frac{1}{8\epsilon p_{XY}(x_i,y_i)}dq \\ &= \frac{(n-1)!}{(k-1)!\,(n-k-1)!}\int_0^1 \log(q) q^{k-1}(1-q)^{n-k-1}dq\end{aligned}$$

In the same way as before, it is now found that

$$\mathbb{E}[\log((2D_{XY}(i))^2 p_{XY}(x_i,y_i))] = \psi(k) - \psi(n) \tag{3.14}$$

Using equation 3.14, it is now found that

$$\begin{aligned}\log(p_{XY}(x_i,y_i)) &= \mathbb{E}[\log(p_{XY}(x_i,y_i))] \\ &= \mathbb{E}[\log((2D_{XY}(i))^2 p_{XY}(x_i,y_i)) - \log((2D_{XY}(i))^2)] \\ &= \mathbb{E}[\log((2D_{XY}(i))^2 p_{XY}(x_i,y_i))] - \mathbb{E}[\log((2D_{XY}(i))^2)] \\ &= \psi(k) - \psi(n) - \mathbb{E}[\log((2D_{XY}(i))^2)]\end{aligned} \tag{3.15}$$

Now, assume the points $(x_1, y_1), \ldots, (x_n, y_n)$ in the sample are all known.
In that case, $\mathbb{E}[\log((2D_{XY}(i))^2)]$ can be estimated as follows:

$$\widehat{\mathbb{E}[\log((2D_{XY}(i))^2)]} = \frac{2}{n}\sum_{i=1}^{n}\log(2d_{xy}(i)) \tag{3.16}$$

with $d_{xy}(i)$ the realization of $D_{XY}(i)$ for the sample $(x_1, y_1), \ldots, (x_n, y_n)$.
Furthermore, $\mathbb{E}[\log(p_{XY}(X,Y))]$ can be estimated as follows:

$$\widehat{\mathbb{E}[\log(p_{XY}(X,Y))]} = \frac{1}{n}\sum_{i=1}^{n}\log(p_{XY}(x_i,y_i)) \tag{3.17}$$

Using equations 3.15, 3.16 and 3.17, it can be concluded that

$$\mathbb{E}[\log(p_{XY}(X,Y))] \approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n) - \frac{2}{n}\sum_{i=1}^{n}\log(2d_{xy}(i))\right) \tag{3.18}$$

The bias in equation 3.10 mainly comes from the assumption that $p_X(x)$ is smooth in the neighborhood of x_i. The effect of this bias depends on the realization of the distance $D_X(i)$; the larger the distance, the higher the bias. In order for the biases in equations 3.10, 3.11 and 3.18 to cancel out, the realizations of $D_X(i)$, $D_Y(i)$ and $D_{XY}(i)$ should therefore be the same. Note that equations 3.10, 3.11 and 3.18 hold for any value of k. Now, only fix the value of k in the derivation of $\widehat{\mathbb{E}[\log(p_{XY}(X,Y))]}$, i.e.

$$D_{XY}(i) = ||(x_i, y_i) - (X_i, Y_i)_k||_\infty$$

$$D_X(i) = |x_i - X_{i,v}|$$

$$D_Y(i) = |y_i - Y_{i,w}|$$

with v and w variable and k fixed.

Now, assume $d_x(i) = d_y(i) = d_{xy}(i)$. If $d_{xy}(i) = |x_i - (x_i)_k|$, then $d_x(i)$ is the distance to the $(n_x(i)+1)^{th}$ nearest neighbor of x_i with $n_x(i)$ the number of points x_j for which $|x_i - x_j| < d_x(i)$. In other words, the above value for v equals $n_x(i)+1$. This does not hold exactly if $d_{xy}(i) = |y_i - (y_i)_k|$. Similarly, if $d_{xy}(i) = |y_i - (y_i)_k|$, then $d_y(i)$ is the distance to the $(n_y(i)+1)^{th}$ nearest neighbor of y_i with $n_y(i)$ the number of points y_j for which $|y_i - y_j| < d_y(i)$. In other words, the above value for w equals $n_y(i)+1$. This does not hold exactly if $d_{xy}(i) = |x_i - (x_i)_k|$. Nevertheless, Kraskov et al. (2004) state that setting v and w equal to, respectively, $n_x(i)+1$ and $n_y(i)+1$ no matter the value of $d_{xy}(i)$, results in good estimations for both $\mathbb{E}[\log(p_X(X))]$ and $\mathbb{E}[\log(p_Y(Y))]$.

Using equations 3.10, 3.11 and 3.18, the following is now found:

$$\begin{aligned}
\mathbb{E}[\log(p_X(X))] &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_x(i)+1) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{xy}(i))\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(n_x(i)+1) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{xy}(i)) \\
\mathbb{E}[\log(p_Y(Y))] &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_y(i)+1) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{xy}(i))\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\psi(n_y(i)+1) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{xy}(i)) \\
\mathbb{E}[\log(p_{XY}(X,Y))] &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n) - \frac{2}{n}\sum_{i=1}^{n}\log(2d_{xy}(i))\right) \\
&= \psi(k) - \psi(n) - \frac{2}{n}\sum_{i=1}^{n}\log(2d_{xy}(i))
\end{aligned}$$

Using these equations and equation 3.1, it is now found that:

$$I(X;Y) \approx \psi(n) + \psi(k) - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_x(i)+1) + \psi(n_y(i)+1)\right) \tag{3.19}$$

□

The following estimator is thus found:

Definition 3. *The Kraskov estimator based on the k^{th} nearest neighbor for the mutual information between two continuous variables X and Y, based on a sample of size n is defined as*

$$\hat{I}_K(X;Y) = \psi(n) + \psi(k) - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_x(i)+1) + \psi(n_y(i)+1)\right) \tag{3.20}$$

with ψ the digamma function, i.e. $\psi(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ with Γ the gamma function, $n_x(i)$ the number of x_j, $j \neq i$, for which $|x_j - x_i| < \epsilon$ with ϵ the distance[5] from (x_i, y_i) to its k^{th} nearest neighbor and $n_y(i)$ the number of y_j, $j \neq i$, for which $|y_j - y_i| < \epsilon$.

Appendix E.4 gives a similar derivation of the adjusted Kraskov estimator, which leads to the following:

Definition 4. *The adjusted Kraskov estimator based on the k^{th} nearest neighbor for the mutual information between categorical variable X and continuous variable Y, based on a sample of size n is defined as*

$$\hat{I}_{AK}(X;Y) = \psi(n) + \psi(k) - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_y(i)) + \psi(n_d(i)+1)\right) \tag{3.21}$$

with ψ the digamma function, $n_d(i)$ the number of points whose value of X equals x_i and $n_y(i)$ the number of y_j, $j \neq i$, for which $|y_j - y_i| < \epsilon$ with ϵ the distance from point y_i to its k^{th} nearest neighbor among the $n_d(i)$ points whose value of X equals x_i.

Figure 3.2 and 3.3 show visual representations of the parameters used within the two estimators, n, k, $n_x(i)$, $n_y(i)$ and $n_d(i)$. In figure 3.2, $k = 1$ and $n = 15$. The resulting values for $n_x(i)$ and $n_y(i)$ are, respectively, 5 and 3. In figure 3.3 the red points correspond to points for which the value of x equals x_i. The parameter $n_d(i)$ equals 6. In this figure, $k = 3$ and $n = 12$. The resulting value for $n_y(i)$ is 5.

[5]The distance from (x_i, y_i) to (x_j, y_j) is calculated by using the maximum norm, i.e. $||(x_i, y_i) - (x_j, y_j)||_\infty = \max(|x_i - x_j|, |y_i - y_j|)$.

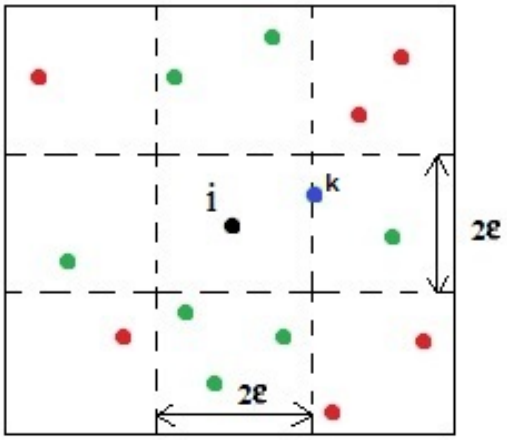

Figure 3.2: Visual representation of the parameters used in the Kraskov estimator. The point for which the k^{th} neighbor, based on the maximum norm, is found, is indicated by the letter i. The $k^{th} = 1^{st}$ nearest neighbor of this point is indicated by the letter k. The green points lie within distance ϵ from point i in the x or y direction, the red points do not.

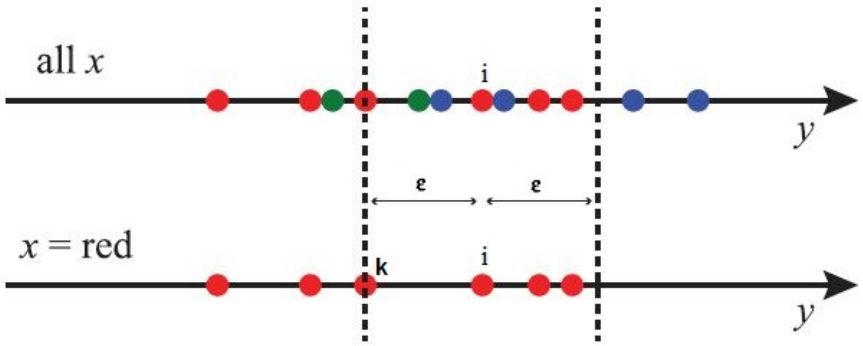

Figure 3.3: Visual representation of the parameters used in the adjusted Kraskov estimator. The colors correspond to different values of the categorical variable x. The point for which the k^{th} neighbor, among all points for which the value of x equals x_i, is found, is indicated by the letter i. The $k^{th} = 3^{rd}$ nearest neighbor of this point among all points for which the value of x equals x_i is indicated by the letter k.

Note that in the derivation of both estimators the non-categorical variables are assumed to be continuous. In the case of the ball actions data set, however, not all non-categorical variables are continuous. In fact, only the variables *Total possession time in milliseconds*, *Average possession time in milliseconds*, *Possession percentage*, *Average location of regaining the ball* and *Average time till the first duel after loss of possession* are continuous. All other variables, i.e. the variables which are expressed in percentages, are constructed from two discrete variables, such as *Number of long passes* and *Total number of passes*. Therefore, these variables are discrete variables themselves as well. In order to be able to apply the Kraskov estimator and the adjusted Kraskov estimator to these variables, very small noise[6] is added to each of these variables. In their experiments, Kraskov et al. (2004) proceed in the same way. As far as literature study shows, no other method for estimating mutual information of discrete variables based on the Kraskov estimator has been constructed. Note that in the case of discrete variables with relatively few different outcome states, the adjusted Kraskov estimator can be used after treating the discrete variables as categorical. In the case of relatively many different outcome states, however, treating the discrete variables as categorical and using the adjusted Kraskov estimator will not lead to accurate results. Unfortunately, most discrete variables in the ball actions data set have relatively many different outcome values.

An additional issue in the case of these estimators is the choice of k. Kraskov et al. (2004) suggest using

[6]Random number generated from a normal distribution with mean zero and standard deviation 10^{-15}.

a value for k in $\{2, 3, 4\}$[7]. Doquire et al. (2012) suggest to average the estimations obtained for all values of k within a reasonable range. In this study, therefore, the final estimates are obtained by averaging the estimates for $k = 2$, 3 and 4.

As mentioned, Doquire et al. (2012) and Ross (2014) show that, respectively, the Kraskov estimator and the adjusted Kraskov estimator are more accurate than probability density based estimators. However, Doquire et al. (2012) also show that the Kraskov estimator gets less accurate as the number of dimensions grows. In other words, the mutual information between a set of k variables and the output variable can not be estimated accurately for large k using the Kraskov estimator. Doquire et al. (2012) also state:

"Even though the task of estimating mutual information has been widely studied, it remains very challenging for high-dimensional vectors."

Fortunately, Peng et al. (2005) showed that instead of selecting a relevant feature subset based on these high-dimensional mutual information values, another algorithm which is similar in the case of forward feature selection can be used. The next section considers this substitutional algorithm.

3.4 Minimum-redundancy-maximum-relevance feature selection

As mentioned, forward feature selection can be used in order to select relevant feature subsets of size $k = 1, \ldots, 121$. For each k this can be done by finding the subset of size k with the maximum mutual information value $I(X_1, \ldots, X_k; Z)$ among all subsets of size k. The mutual information values are estimated by the use of the Kraskov and adjusted Kraskov estimators. Unfortunately, accurate estimation of high-dimensional mutual information values remains an issue, according to Doquire et al. (2012). Peng et al. (2005) proposed a substitutional algorithm, *minimum-redundancy-maximum-relevance feature selection (mRMR)*. This feature selection algorithm selects features based on forward feature selection by maximizing the mutual information of the individual features with respect to the output variable and at the same time minimizing the mutual information with respect to the already selected features. This way, the selected features are both informative with respect to the output variable as well as not too highly dependent upon each other.

If the features are selected based on only maximizing the mutual information of the individual features with respect to the output variable, the selected features can be dependent upon each other. If feature X_1 and X_2 both give a lot of information about the outcome variable but are highly correlated with each other, the increase in information when selecting X_2 after already having selected X_1 is quite low. In other words, feature X_2 is redundant when feature X_1 has already been selected. As opposed to irrelevant features, i.e. features which give no information about the outcome variable, redundant features do not harm the results. However, they also do not improve the results. Therefore, if a fixed number of features or a subset capable of discriminating between playing styles containing as little features as possible is to be selected[8], it is better to not select these redundant features. That way, only features which give much additional information about the outcome variable with respect to the already selected features, are taken into account.

Consider the variables $X_1, \ldots, X_p$ and outcome variable Z. Let S be the set of already selected features and F the set of non-selected features. Consider forward feature selection. In each iteration the mRMR algorithm maximizes the individual mutual information, $C_i = I(X_i; Z)$, with $X_i \in F$, and at the same time minimizes the average mutual information of the new feature and the set of selected features, $B_i = \frac{1}{|S|} \sum_{X_j \in S} I(X_i; X_j)$, with $X_i \in F$ and $X_j \in S$. This can be combined, for example, by maximizing $C_i - B_i$[9]. Note that a possible improvement here would be to place a regularization parameter λ in front of the quantity B_i in order to decrease or increase the penalty of selecting a feature which is partly redundant with respect to the already selected features. Peng et al. (2005) show that in the case of forward feature selection selecting features by the use of the mRMR algorithm is similar to selecting features based on maximization of high-dimensional mutual information values $I(X_1, \ldots, X_k; Z)$. A more detailed proof of this is given in appendix E.5. Algorithm 3 gives a description of the mRMR forward feature selection method. The mutual information values in this algorithm will be estimated by the Kraskov estimator and the adjusted Kraskov estimator in this thesis.

[7] According to Kraskov et al. (2004) high values of k lead to high systematic errors, whereas small values lead to high statistical errors. In order for the systematic errors not to outweigh the decrease of the statistical errors, k should be chosen within $\{2, 3, 4\}$.

[8] A feature set containing as little features as possible is desirable since this makes the interpretation of the resulting clusters easier.

[9] Another possibility would be to maximize $\frac{C_i}{B_i}$.

Algorithm 3: Minimum-redundancy-maximum-relevance forward feature selection

Data: $\mathbf{X} = \{X_1, \ldots, X_p\}$ a set of p features and output variable Z
Result: A subset S of $k < p$ features forming a maximum-relevance minimum-redundant feature set of k features with respect to the output variable Z
Set $S = \text{argmax}_{X_i \in \{X_1,\ldots,X_p\}}\{I(X_i; Z)\}$ with $I(X_i; Z)$ the mutual information between X_i and Z;
Set $F = \mathbf{X} \setminus S$;
while $|S| < k$ **do**
  $S = S \cup \text{argmax}_{X_i \in F}\{I(X_i; Z) - \frac{1}{|S|}\sum_{X_j \in S} I(X_i; X_j)\}$;
  $F = \mathbf{X} \setminus S$;
end

The mRMR algorithm can be performed on the ball actions data set using the matches of the benchmark teams in order to find a feature set of size k which is most relevant and least redundant with respect to the playing style. The mutual information values used within the mRMR algorithm are estimated using the Kraskov estimator and the adjusted Kraskov estimator. When performing the algorithm with k equal to 121, a ranking for all the features in the data set is found corresponding to a decreasing amount of additional information the features give about the playing style after higher ranked features have already been selected.

From now on, the reported results will be based on the matches of Ajax, FC Barcelona, Atletico Madrid and Queens Park Rangers, i.e. the matches of Juventus (the playing style *Catenaccio*) are left aside. All the steps have also been performed for the matches of the benchmark teams including Juventus, but the results for *Catenaccio* were not satisfying. There are two main issues when dealing with *Catenaccio*:

- There are very few *Catenaccio* matches in the set of matches of benchmark teams on which the model can be trained.
- The playing style *Catenaccio* is mostly characterized by defensive statistics of which not many are available[10].

These issues probably lead to the fact that *Catenaccio* can not be well discriminated. For that reason, the choice is made to drop *Catenaccio* as playing style in the analysis. Idea for further study is to gather more relevant statistics with respect to *Catenaccio* and find more *Catenaccio* benchmark matches. Appendix F.1 shows the ranking of the features in *Feature set 2* resulting from performance of the mRMR algorithm on the scaled data of the matches of the benchmark teams without Juventus.

A feature selection method has now been constructed with which a suitable ranking of the features in the ball actions data set can be obtained. Notice that the mRMR algorithm denoted in algorithm 3 selects k features based on a predefined number of features. As mentioned in section 3.1 the choice is made to choose an appropriate value for k based on the results feature subsets of different sizes k give when used in the final classification method. The next section deals with this issue and focuses on actually determining the playing style applied in a match.

[10] As mentioned in section 1.3.2, the data set only contains information about ball actions; information about actions not related to the ball is not available.

Chapter 4

Determining the playing style in a match

Chapter 3 described a feature selection method suited to the problem in this study. Applying this method to the matches of the benchmark teams gives a ranking of the features corresponding to a decreasing amount of additional information the features give about the playing style after higher ranked features have already been selected. Selecting the top k features of this ranking lists results in a feature set which is most likely better capable of discriminating between different playing styles than *Feature set 2*. The next step is to actually assign matches to different playing styles by the use of some classification or clustering method. In this chapter this classification issue is addressed and a final model with which matches can be assigned to a playing style is constructed. Instead of assigning matches to one of the playing styles *Hollandse School*, *Tiki Taka*, *Counterplay* and *Kick and Rush*, a method with which the scores of matches on different characteristics of playing styles in general can be determined, is also discussed.

The number of features to select from the ranking list should now be decided upon. As mentioned before, the choice is made to choose this number of features based on the results that feature subsets of different sizes, i.e. the top $k = 1, \ldots, 121$ features obtained from the ranking list using the mRMR method, give when using the final classification method in combination with these feature subsets. That way, the resulting feature subset is best capable of assigning matches to their corresponding playing styles using the final classification method. Before selecting the optimal number of features in the case of the ball actions data set, the next sections will now give details about the classification method which will be used.

4.1 Non-hierarchical clustering scheme

A choice has to be made as to which classification method will be used to assign matches to different playing styles. As mentioned in the beginning of section 3.1, an unsupervised method will be used in order to determine the playing style of a match. This is done, since the assumed playing styles in the matches of the benchmark teams might be incorrect. As mentioned in subsection 1.2.2, K-means clustering is a frequently applied unsupervised method in general as well as in similar problems as the one in this study. That is why the choice is made to apply K-means clustering in order to assign matches to playing styles[1]. The next subsection gives some details about the method K-means clustering.

4.1.1 K-means clustering

The choice is made to use K-means clustering in order to assign matches to playing styles. K-means clustering is a method with which unlabeled observations can be grouped into K clusters. In short, K-means clustering starts by choosing randomly (if not specified) K cluster centers. Within K-means clustering a cluster center is defined as the vector of the means of the feature values of the observations belonging to the specific cluster. Each observation then gets assigned to the cluster whose cluster center is closest with respect to the Euclidean distance. After having assigned each observation to a cluster, the cluster centers are calculated again based on the observations belonging to the clusters. Next, again the observations are assigned to the cluster with the

[1]Since other unsupervised methods might lead to better results, an idea for further studies is to also apply other methods and choose the one which produces the best results.

closest cluster center. This is repeated until the assignment of observations does not change anymore.

Note that within K-means clustering Euclidean distance is used as distance measure. Furthermore, the cluster centers are calculated as the means of the feature values of the observations belonging to the specific clusters. However, the same algorithm can be performed using different distance measures and different definitions of cluster centers. Subsection 4.1.3 discusses such a variation of K-means clustering in which Manhattan distance is used and cluster centers are calculated by medians instead of means. Also, weighted distances could be used such that important features have more influence on the cluster analysis.

Next, a more formal description of the K-means algorithm is given. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be n observations with $\mathbf{x}_i$, $i = 1, \ldots, n$, a p-dimensional vector. K-means clustering aims to divide these n observations into K sets $S_1, \ldots, S_K$ with $K \leq n$ by minimizing the within-cluster sum of squares. The within-cluster sum of squares measures the amount in which observations in a cluster differ from each other.

Definition 5. *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ be n observations with $\boldsymbol{x}_i$, $i = 1, \ldots, n$, a p-dimensional vector. Let $S = \{S_1, \ldots, S_K\}$ be the set of K clusters in total containing all of the n observations. The within-cluster sum of squares of S, $W(S)$, is defined as*

$$W(S) = \sum_{j=1}^{K} \sum_{\boldsymbol{x}_i \in S_j} (d_E(\boldsymbol{x}_i, \boldsymbol{c}_j))^2$$

with $d_E(\boldsymbol{x}_i, \mathbf{c}_j) = \sqrt{\sum_{l=1}^{p} (x_{i,l} - c_{j,l})^2}$ the Euclidean distance between $\boldsymbol{x}_i$ and $\mathbf{c}_j$ and $\mathbf{c}_j = (c_{1,j}, \ldots, c_{p,j}) = \frac{1}{N_j} \sum_{i:\boldsymbol{x}_i \in S_j} (x_{i,1}, \ldots, x_{i,p})$ the cluster center of cluster j.

Algorithm 4 describes the process of K-means clustering.

Algorithm 4: K-means clustering

Data: n observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$, with $\mathbf{x}_i$, $i = 1, \ldots, n$, a p-dimensional vector
Result: Subdivision of the n observations into K non-overlapping clusters $S = \{S_1, \ldots, S_K\}$
(Randomly) choose K initial cluster centers $\mathbf{c}_1, \ldots, \mathbf{c}_K$ with $\mathbf{c}_i$ a p-dimensional vector;
while *cluster assignments change* **do**
  Calculate the Euclidean distance $d_E(\mathbf{x}_i, \mathbf{c}_j)$ between observation i, $i = 1, \ldots, n$, and cluster center j, $j = 1, \ldots, K$;
  Assign observation i, $i = 1, \ldots, n$, to the cluster corresponding to the smallest distance between observation and cluster center;
  Recalculate the cluster centers $\mathbf{c}_j$, $j = 1, \ldots, K$, based on the observations belonging to the specific cluster;
end

Applying K-means clustering with K clusters on the set of matches in the scaled ball actions data set using *Feature set 2*, divides these matches in K different groups. In theory, these groups could correspond to all possible matters related to ball actions. Since exploratory analysis of the matches of the benchmark teams showed that the corresponding different playing styles were separated quite well using *Feature set 2*, see figure **??** in appendix F.2, it is expected that the resulting clusters will correspond to different playing styles. Assuming that this is indeed the case, these clusters can be labeled by either interpreting the clusters based on their centers or, in the case of clustering of matches which already have been labeled with a playing style, labeling the cluster as the playing style which is most frequently present in that cluster[2]. The question arises how the playing styles of new matches, which are not in the data set yet, should be determined. There are multiple ways in which this can be done in relation to K-means clustering:

- Perform K-means clustering again on the whole data set containing the new matches a well.
- Use the resulting clusters obtained by K-means clustering of the original data set and assign the new matches to the cluster whose cluster center is closest.
- Use the labels obtained by K-means clustering of the original data set to train a supervised classifier.

[2]In the case of the latter, the interpretation of the clusters based on their centers should agree with the assigned labels based on most present labels in the cluster.

In case of the first option, the interpretation of the clusters can change depending on which set of data points is being clustered. K-means clustering of the matches of the benchmark teams, with $K = 4$, (most likely) leads to four clusters corresponding to the playing styles *Hollandse School*, *Tiki Taka*, *Counterplay* and *Kick and Rush*. K-means clustering of the matches in the Eredivisie, however, might lead to four clusters with other interpretations, since the set of matches in the Eredivisie does not necessarily need to contain those four playing styles. For example, *Tiki Taka* might not get applied in the Eredivisie and/or some other playing style possibly does. Therefore, a particular match might get assigned to a cluster with a certain interpretation when it is being clustered together with a certain set of matches, whereas this same match might get assigned to a cluster with another interpretation when being clustered together with another set of matches. It is desirable to assign a match to a certain playing style independent of the set of matches within which it is being considered. Therefore, this option will not be utilized.

The second option is also known as 1-Nearest Neighbors trained on the cluster centers resulting from K-means clustering performed on the original data set. This option assigns matches to clusters independent of the set of matches within which it is being considered. For the third option the same holds. This study applies the second option for simplicity reasons. An idea for further studies is to consider the third option.

The choice is made to focus on the four playing styles *Hollandse School*, *Tiki Taka*, *Counterplay* and *Kick and Rush*, i.e. to apply K-means clustering on the set of matches of the benchmark teams only. This is done, because the feature selection method is based on only these four playing styles as well. When applying K-means to the set of all matches in the ball actions data set described in subsection 1.3.3, other playing styles than those four might be contained in this set, which makes the selected features less useful.

In conclusion, K-means clustering with $K = 4$ clusters will be applied to the matches of the benchmark teams after which 1-NN trained on the resulting cluster centers is applied to assign new matches to one of the four playing styles as well. Since K-means clustering chooses its initial cluster centers randomly, K-means is performed 100 times after which the best result is chosen. The best clustering result is assumed to be the result with the smallest within-cluster sum of squares $W(S)$ as defined in definition 5. Using the cluster centers corresponding to this best clustering result, other matches will be assigned to one of the four clusters. Before doing this, the number of features to select has to be chosen. The next subsection deals with this issue.

4.1.2 Number of features and misclassification error

Author's note: this subsection is confidential.

4.1.3 Clustering results

Author's note: this subsection is confidential.

4.2 Hierarchical clustering scheme

Author's note: this section is confidential.

4.3 Characteristic variables

Author's note: this section is confidential.

Chapter 5

Results

Author's note: this section is confidential.

Chapter 6

Conclusion and further directions

Conclusion

Coaches can benefit from objective information about playing styles applied in football matches. In this thesis, two methods for determining (characteristics of) the playing style of a football team in a certain match solely based on statistics of ball actions have been constructed. Exploratory analyses of the data set describing statistics of ball moments showed that treating the problem completely unsupervised does not lead to satisfying results. That is why domain knowledge is used to choose benchmark teams from which it is known that they play according to specific playing styles quite clearly in the majority of their matches. The playing styles *Hollandse School*, *Tiki Taka*, *Counterplay* and *Kick and Rush* with corresponding teams Ajax, FC Barcelona, Atletico Madrid and Queens Park Rangers are considered for this.

A subset of features from the data set is selected in order to improve the capability of discriminating between the four playing styles. This is done by using mutual information, estimated by the (adjusted) Kraskov estimator, as a measure of relevance of the features with respect to playing style. Features are then ranked based on the minimum-redundancy-maximum-relevance algorithm. Next, the matches of the benchmark teams have been divided into four groups by the use of K-means clustering based on a selected subset of features. New matches are then assigned to the cluster with the closest cluster center.

The optimal number of features to select is chosen based on the number corresponding to the smallest estimated expected misclassification error, which is found using 5-fold cross-validation. Splitting the matches of the benchmark teams in four clusters at once and assigning new matches to these four clusters leads to an estimated expected misclassification error of approximately 15%. A hierarchical structure is also built by first splitting the matches in two clusters and consecutively splitting the resulting two clusters in two new clusters each, by using K-means clustering based on a different selected subset of features in each clustering step. This *Hierarchical clustering model* leads to an estimated expected misclassification error of approximately 11% and is chosen as final model to assign matches to one of the four labeled playing styles.

A second method is constructed in order to obtain information about non-labeled playing styles as well. In this *Characteristic variables model*, characteristic variables *Dominance*, *Offensive play* and *Passing length* have been created by the use of principal component analysis. Based on these three variables important information about characteristics of playing styles in general can be obtained without the need for prior information about labeled playing styles.

Application of both models on subsets of the data show satisfying results which agree with domain knowledge. These models can be used to provide coaches with information regarding playing styles applied in matches. Coaches can use this information in order to both evaluate their own team as well as analyze their opponents. Based on this information, coaches can, for example, adjust their training procedures or change their game plan in order to increase their probability of winning. Even though satisfying results have been obtained in this study, improvements are still possible. The next section discusses such possible improvements as well as additional ideas for further studies.

Further directions

Two models have been built in this study by which the (characteristics of the) playing style of a football team in a match can be determined solely by the use of statistics about ball actions. Even though both models provide satisfying results, improvements and additions are still possible. Throughout this report some possible adjustments have already been mentioned. This section will give an overview of all these possibilities.

The first thing which should be mentioned is that the applications of the methods used in this study are not limited to playing styles of football teams. The same methods can be used in order to determine the playing style of players instead of teams. The only difference is that the data should give information about specific players instead of an entire team. Furthermore, instead of benchmark teams, benchmark players should be chosen. Also, the hierarchical clustering scheme is not necessarily better than the non-hierarchical scheme in this case, so this should be checked. The number of clusters which ought to be constructed equals the desired number of different benchmark playing styles for players. Also note that the applications are not limited to football. The methods can be used for any ball sport under the condition that related data are available.

Regarding the data used in this study, some improvements and adjustments are also possible. The data in this study are aggregated data over an entire match. However, playing styles of teams are likely to change within a match as well. Teams can adjust their playing style after the first half of the match is over or after a goal has been made. If data about these shorter moments within a match would be used, the playing styles during these moments could be determined separately. Also note that the playing style *Catenaccio* is not considered in the final part of this study anymore, due to the lack of Catenaccio benchmark matches and possibly the lack of suitable features. Finding more of these benchmark matches and suitable features could lead to the fact that *Catenaccio* matches can also be recognized from the data. Regarding the benchmark matches in general, one of the most important improvements would be to find more and better benchmark matches such that the different playing styles can be discriminated between more clearly. Note that if enough benchmark matches are available from which it is known for sure which playing style is applied during these matches, there is no need to work (partly) unsupervised anymore. Supervised methods could be used in this case, which could lead to more satisfying results. Furthermore, in the case of the current unsupervised classification method, when using other benchmark matches it should be checked again whether the hierarchical clustering scheme still leads to better results than the non-hierarchical scheme.

There are a few more issues regarding the data which could possibly be improved. First of all, in this study standard scaling of the data has been used. This might not be the most suited scaling method for this specific data. For example, assigning different weight factors to the variables in the data set such that specific variables have more influence in analyses than others, could be desirable. More research as to what kind of scaling method suits best with this data could lead to improvements of the results. Also note that the variables in the data set have been transformed to percentage data based on domain knowledge. However, more suitable transformations might exist which could be found by using transformation methods for classification purposes such as the one introduced by Zhou et al. (2009). Also note that the multicollinear features have been removed from the data set before doing any analysis. In order to check whether these multicollinear features indeed give no additional information about playing styles, the methods could be performed on the data set including these multicollinear features.

Another part of this study for which improvements are possible, is the feature selection method. First of all, it should be noted that the applied feature selection method is of a supervised kind. This choice in combination with an unsupervised classification method was made due to the advantages and disadvantages of both supervised and unsupervised methods in the case of lack of training data for which the outcome values are known for sure. The assumption was made that usage of possibly wrong information in the feature selection step causes less harm than in the final step of assigning matches to playing styles, which is why a supervised feature selection method was used. Further studies could check whether this is indeed true or whether an unsupervised feature selection method might lead to better results after all. Furthermore, the number of features to select is now chosen based on the smallest estimated expected misclassification error which is calculated by 5-fold cross-validation. However, due to the variance in the results it might be more suitable to choose the number for which the upper bound of a confidence interval of the expected misclassification error is smallest.

Regarding the (adjusted) Kraskov estimator based on k nearest neighbors, an improvement would be to determine an optimal value for k. Unfortunately, during literature study no such methods have been encountered.

Using various values for k and choosing the one resulting in the most satisfying results would be an option. Furthermore, in this study the discrete variables have been adjusted by adding small noise in order for the variables to become continuous such that the (adjusted) Kraskov estimator can be used. Other mutual information estimators in the case of discrete data for which the data do not have to be manipulated might lead to more accurate estimates for the mutual information. Another possible improvement within the feature selection method is related to the mRMR criterion $C_i - B_i = I(X_i; Z) - \frac{1}{|S|} \sum_{X_j \in S} I(X_i; X_j)$ with S the set of already selected features and $X_i \notin S$. This criterion is maximized in order to select a feature based on high mutual information with respect to the outcome variable and low mutual information with respect to the already selected variables. This can also be achieved, however, by maximizing the criterion $\frac{C_i}{B_i}$. Furthermore, a regularization parameter can be introduced in order to control the importance of the regularization term $\frac{1}{|S|} \sum_{X_j \in S} I(X_i; X_j)$. Using this different criterion or making use of a regularization parameter could improve the results.

As mentioned, if a set of benchmark matches is available for which it is known for sure which playing styles are applied in them, a supervised classification method could be used instead of the constructed unsupervised method. However, if this is not the case, improvements could still be possible by using another unsupervised method than K-means clustering. Also, variations of K-means clustering, such as weighted K-means, could be used.

In the case of the *Characteristic variables model*, sparse PCA could be used in order to create the characteristic variables. This way, an optimal combination, with respect to maximizing the variance, of a subset of the variables can be found by taking into account all the variables instead of focusing on a subset. Furthermore, other characteristic variables could be created as well if information about other characteristics is desired.

All of these adjustments and additions could be investigated in further studies in order to check whether these would improve the results. The methods constructed in this study can also be used in order to be able to analyze new issues. For example, chapter 5 gave a short analysis of the probability of winning or losing when playing against an opponent with a certain playing style. More issues like this, which are interesting for coaches, can be analyzed by the use of the constructed models. Consider, for example, the issue of deciding which players to put on the field in order to be able to optimally perform a specific playing style as a team. Once the characteristics which are important for this specific playing style are known, players which have these specific characteristics as their qualities can be chosen. In the case of the *Characteristic variables model* the important qualities are known by considering the original variables which are used to create the characteristic variables. In the case of the *Hierarchical clustering model* the important qualities are known by considering the features which were selected by the mRMR features selection algorithm. More issues like this, which are important for coaches, can be analyzed now that models have been constructed by which the playing style of a football team applied in a match can be determined based on statistics of ball actions.

Appendices

Appendix A

Data

A.1 Features

Author's note: this section is confidential.

A.2 Irrelevant features

Author's note: this section is confidential.

A.3 Transformed features

Author's note: this section is confidential.

A.4 Multicollinear features

Author's note: this section is confidential.

Appendix B

Derivation principal components

Let $\mathbf{X} = (X_1, \ldots, X_p)$ be a random vector in $\mathbb{R}^p$. The principal components $Z_1, \ldots, Z_p$ of the random vector $\mathbf{X}$ are of the form $Z_i = \alpha_{i,1}X_1 + \cdots + \alpha_{i,p}X_p$ with $\sum_{j=1}^{p} \alpha_{i,j}^2 = 1$, $\mathrm{Var}(Z_1) > \mathrm{Var}(Z_2) > \cdots > \mathrm{Var}(Z_p)$ and $\mathrm{Cov}(Z_i, Z_j) = 0$ for $i \neq j$. In order to find values for $\alpha_{i,j}$, $i, j = 1, \ldots, p$, such that the above requirements hold, the covariance matrix of $\mathbf{X}$, $\mathbf{\Sigma}$, should be known. This covariance matrix $\mathbf{\Sigma}$ is not known, unfortunately, and will therefore be estimated by taking into account a sample of size n from the random vector $\mathbf{X}$.

Consider the $n \times p$-dimensional matrix $\mathbf{X}_n = (\mathbf{x}_1, \ldots, \mathbf{x}_p)$, corresponding to n realizations of the random vector $\mathbf{X}$. The principal components $Z_1, \ldots, Z_p$ of the random vector $\mathbf{X}$ will be estimated by using this set of realizations $\mathbf{X}_n$. The values of these principal components corresponding to the n realizations $\mathbf{X}_n$ are denoted by $\mathbf{z}_i = \alpha_{i,1}\mathbf{x}_1 + \alpha_{i,2}\mathbf{x}_2 + \cdots + \alpha_{i,p}\mathbf{x}_p$, $i = 1, \ldots, p$, with $\sum_{j=1}^{p} \alpha_{i,j}^2 = 1$.

By the definition of principal components, $\mathbf{z}_1$ should explain the maximum possible variance in the original data set, i.e. in $\mathbf{X}_n$. Therefore, $\mathrm{Var}(\mathbf{z}_1)$ should be maximized over all possible values of $\alpha_{1,j}$, $j = 1, \ldots, p$. Let $z_{i,j}$, $i = 1, \ldots, n$, $j = 1, \ldots, p$, denote the i^{th} score of the j^{th} principal component. The following should then be maximized

$$\begin{aligned}
\mathrm{Var}\left(\mathbf{z}_1\right) &= \frac{1}{n-1}\sum_{i=1}^{n}\left(z_{i,1} - \frac{1}{n}\sum_{i=1}^{n} z_{i,1}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\alpha_{1,j}x_{i,j} - \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p}\alpha_{1,j}x_{i,j}\right)^2 \\
&= \frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\alpha_{1,j}x_{i,j} - \sum_{j=1}^{p}\alpha_{1,j}\left(\frac{1}{n}\sum_{i=1}^{n}x_{i,j}\right)\right)^2
\end{aligned} \tag{B.1}$$

with $x_{i,j}$ the i^{th} element of $\mathbf{x}_j$.
For simplicity assume $\mathbf{x}_1, \ldots, \mathbf{x}_p$ all have mean zero[1]. Equation B.1 then equals

$$\frac{1}{n-1}\sum_{i=1}^{n}\left(\sum_{j=1}^{p}\alpha_{1,j}x_{i,j}\right)^2 \tag{B.2}$$

In matrix form this equals

$$\frac{1}{n-1}(\mathbf{X}_n\boldsymbol{\alpha_1})^T(\mathbf{X}_n\boldsymbol{\alpha_1}) \tag{B.3}$$

with $\boldsymbol{\alpha_1}$ the vector consisting of the elements $\alpha_{1,1}, \ldots, \alpha_{1,p}$. Equation B.3 can be rewritten as

$$\begin{aligned}
&\frac{1}{n-1}\boldsymbol{\alpha_1}^T\mathbf{X}_n^T\mathbf{X}_n\boldsymbol{\alpha_1} \\
&= \boldsymbol{\alpha_1}^T\frac{\mathbf{X}_n^T\mathbf{X}_n}{n-1}\boldsymbol{\alpha_1}
\end{aligned} \tag{B.4}$$

[1]This does not influence the results, since the observations $x_{i,j}$ can just be replaced by their centered versions $x_{i,j} - \hat{x}_j$ with $\hat{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}$.

Note that the constraint $\sum_{j=1}^{p} \alpha_{1,j}^2 = 1$, i.e. $\boldsymbol{\alpha_1}^T \boldsymbol{\alpha_1} = 1$, still holds. The problem of maximizing equation B.4 over all possible values of $\alpha_{1,j}$, $j = 1, \ldots, p$, subject to this constraint can be solved using the Lagrange multiplier. Consider the following equation:

$$u = \boldsymbol{\alpha_1}^T \frac{\mathbf{X}_n^T \mathbf{X}_n}{n-1} \boldsymbol{\alpha_1} - \lambda(\boldsymbol{\alpha_1}^T \boldsymbol{\alpha_1} - 1)$$

Differentiating to $\boldsymbol{\alpha_1}$ and setting equal to zero gives

$$\frac{\partial u}{\partial \boldsymbol{\alpha_1}} = 2\frac{\mathbf{X}_n^T \mathbf{X}_n}{n-1} \boldsymbol{\alpha_1} - 2\lambda \boldsymbol{\alpha_1} = 0$$

This equation leads to

$$\frac{\mathbf{X}_n^T \mathbf{X}_n}{n-1} \boldsymbol{\alpha_1} = \lambda \boldsymbol{\alpha_1}$$

Setting $\mathbf{V} = \frac{\mathbf{X}_n^T \mathbf{X}_n}{n-1}$ gives

$$\mathbf{V}\boldsymbol{\alpha_1} = \lambda \boldsymbol{\alpha_1} \tag{B.5}$$

This shows that $\boldsymbol{\alpha_1}$ is an eigenvector of matrix $\mathbf{V}$. Notice that the quantity to be maximized, i.e. equation B.4, equals $\boldsymbol{\alpha_1}^T \lambda \boldsymbol{\alpha_1} = \lambda \boldsymbol{\alpha_1}^T \boldsymbol{\alpha_1} = \lambda$ by substitution of equation B.5 and the constraint. Therefore, equation B.4 with constraint $\boldsymbol{\alpha_1}^T \boldsymbol{\alpha_1} = 1$ is maximized by choosing $\boldsymbol{\alpha_1}$ equal to the eigenvector of $\mathbf{V}$ corresponding to the largest eigenvalue λ.

The next principal components can be found in similar ways only adding a constraint which reassures that the components are uncorrelated. Derivation of the second principal component will now be shown. Parts of the steps are not shown, since they are exactly the same as in the derivation of the first principal component. By the definition of principal components $\mathbf{z}_1$ and $\mathbf{z}_2$ should be uncorrelated, i.e. $\mathrm{Cov}(\mathbf{z}_1, \mathbf{z}_2) = 0$. The goal is therefore to find values for $\alpha_{2,1}, \ldots, \alpha_{2,p}$ such that $\frac{1}{n-1} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \alpha_{2,j} x_{i,j} \right)^2$ is maximized under the constraints $\sum_{j=1}^{p} \alpha_{2,j}^2 = 1$ and $\mathrm{Cov}(\mathbf{z}_1, \mathbf{z}_2) = 0$.

In the same way as before it can be shown that

$$\mathrm{Cov}(\mathbf{z}_1, \mathbf{z}_2) = \frac{1}{n-1} \sum_{i=1}^{n} \left(\sum_{j=1}^{p} \alpha_{1,j} x_{i,j} \sum_{j=1}^{p} \alpha_{2,j} x_{i,j} \right)$$

In matrix form this equals

$$\begin{aligned}
&\frac{1}{n-1} (\mathbf{X}_n \boldsymbol{\alpha_1})^T (\mathbf{X}_n \boldsymbol{\alpha_2}) \\
&= \frac{1}{n-1} \boldsymbol{\alpha_1}^T \mathbf{X}_n^T \mathbf{X}_n \boldsymbol{\alpha_2} \\
&= \boldsymbol{\alpha_1}^T \frac{\mathbf{X}_n^T \mathbf{X}_n}{n-1} \boldsymbol{\alpha_2} \\
&= \boldsymbol{\alpha_1}^T \mathbf{V} \boldsymbol{\alpha_2} \qquad \text{(B.6)} \\
&= \left(\left(\boldsymbol{\alpha_1}^T \mathbf{V} \boldsymbol{\alpha_2} \right)^T \right)^T \\
&= \left(\boldsymbol{\alpha_2}^T (\boldsymbol{\alpha_1}^T \boldsymbol{V})^T \right)^T \\
&= \left(\boldsymbol{\alpha_2}^T \boldsymbol{V}^T \boldsymbol{\alpha_1} \right)^T \\
&= \left(\boldsymbol{\alpha_2}^T \boldsymbol{V} \boldsymbol{\alpha_1} \right)^T \\
&= \left(\boldsymbol{\alpha_2}^T \lambda \boldsymbol{\alpha_1} \right)^T \\
&= \lambda \boldsymbol{\alpha_1}^T \boldsymbol{\alpha_2} \qquad \text{(B.7)}
\end{aligned}$$

Using the method of the Lagrange multiplier again, the following equation is now formed:

$$u = \boldsymbol{\alpha_2}^T \mathbf{V} \boldsymbol{\alpha_2} - \lambda_2(\boldsymbol{\alpha_2}^T \boldsymbol{\alpha_2} - 1) - \lambda_3 \lambda \boldsymbol{\alpha_1}^T \boldsymbol{\alpha_2}$$

Differentiating to $\boldsymbol{\alpha_2}$ and setting equal to zero gives

$$\frac{\partial u}{\partial \boldsymbol{\alpha_2}} = 2\mathbf{V}\boldsymbol{\alpha_2} - 2\lambda_2\boldsymbol{\alpha_2} - \phi\boldsymbol{\alpha_1} = 0 \tag{B.8}$$

with $\phi = \lambda_3\lambda$.
Multiplying by $\boldsymbol{\alpha_1}^T$ gives

$$2\boldsymbol{\alpha_1}^T\mathbf{V}\boldsymbol{\alpha_2} - 2\lambda_2\boldsymbol{\alpha_1}^T\boldsymbol{\alpha_2} - \phi\boldsymbol{\alpha_1}^T\boldsymbol{\alpha_1}$$

Using equation B.6, B.7 and the constraint $\boldsymbol{\alpha_1}^T\boldsymbol{\alpha_1} = 1$ this equals

$$0 - 0 - \phi = 0$$

Therefore, $\phi = 0$ should hold. Equation B.8 now leads to

$$2\mathbf{V}\boldsymbol{\alpha_2} - 2\lambda_2\boldsymbol{\alpha_2} = 0$$

This is similar to equation B.5 in the derivation of the first principal component, so $\boldsymbol{\alpha_2}$ should again be set equal to the eigenvector of $\mathbf{V}$ corresponding to the largest eigenvalue λ_2. However, since in that case $\boldsymbol{\alpha}_2 = \boldsymbol{\alpha}_1$, i.e. $\mathbf{z}_2 = \mathbf{z}_1$, which violates $\text{Cov}(\mathbf{z}_1, \mathbf{z}_2) = 0$, the eigenvector corresponding to the second largest eigenvalue should be chosen as values for $\boldsymbol{\alpha_2}$. The remaining principal components can be derived in a similar way.

Appendix C

Results exploratory analysis

C.1 Eredivisie scattermatrix principal components

Author's note: this section is confidential.

C.2 Barcelona and Queens Park Rangers principal components

Author's note: this section is confidential.

C.3 Benchmark teams principal components

Author's note: this section is confidential.

Appendix D

Results additional features

Author's note: this section is confidential.

Appendix E

Feature selection

E.1 Mutual information

The mutual information of the random variables X and Y can also be calculated as a function of the *entropy* of the two variables. The entropy of a variable is a quantification of the uncertainty of the variable. All of the definitions and theorems in this appendix are given in Yeung (2008).

Definition 6. *Let X be a continuous random variable with possible outcome set Ω_X. Let p_X be the probability density function of X. The entropy of X, $H(X)$, is defined as*

$$H(X) = -\int_{\Omega_X} p_X(x) \log\left(p_X(x)\right) \mathrm{d}x$$

One can also consider the conditional entropy of two random variables. The conditional entropy of the random variables X and Y can be seen as a quantification of the uncertainty of variable X given the event Y.

Definition 7. *Let X and Y be continuous random variables with possible outcome sets, respectively, Ω_X and Ω_Y. Let p_X and p_Y be the probability density functions of X and Y respectively and let p_{XY} be the joint probability density function of (X, Y). The conditional entropy of X and Y, $H(X|Y)$, is defined as*

$$H(X|Y) = \int_{\Omega_X} \int_{\Omega_Y} p_{XY}(x, y) \log\left(\frac{p_Y(y)}{p_{XY}(x, y)}\right) \mathrm{d}y\mathrm{d}x$$

In the case of discrete or categorical random variables the integrals in definitions 6 and 7 are replaced by summation signs over all possible values of the random variables. In the case of the use of log base 2, the entropy is calculated in *shannon*.

Definitions 6 and 7 can also be extended to more than two random variables. One can consider the joint entropy of multiple random variables. The joint entropy of random variables $X_1, X_2, \ldots, X_k$ can be seen as a quantification of the uncertainty of the set of variables $\{X_1, X_2, \ldots, X_k\}$.

Definition 8. *Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively and let $p_{X_1 \cdots X_k}$ be the joint probability density function of $(X_1, X_2, \ldots, X_k)$. The joint entropy of $X_1, X_2, \ldots, X_k$, $H(X_1, \ldots, X_k)$, is defined as*

$$H(X_1, \ldots, X_k) = -\int_{\Omega_{X_1}} \cdots \int_{\Omega_{X_k}} p_{X_1 \cdots X_k}(x_1, \ldots, x_k) \log\left(p_{X_1 \cdots X_k}(x_1, \ldots, x_k)\right) \mathrm{d}x_k \cdots \mathrm{d}x_1$$

Also conditional joint entropy can be considered.

Definition 9. *Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively and let $p_{X_1 \cdots X_k}$ be the joint probability density function of $(X_1, X_2, \ldots, X_k)$. The conditional joint entropy of $X_1, X_2, \ldots, X_{k-1}$ given X_k, $H(X_1, \ldots, X_{k-1}|X_k)$, is defined as*

$$H(X_1, \ldots, X_{k-1}|X_k) = \int_{\Omega_{X_1}} \cdots \int_{\Omega_{X_k}} p_{X_1 \cdots X_k}(x_1, \ldots, x_k) \log\left(\frac{p_{X_k}(x_k)}{p_{X_1 \cdots X_k}(x_1, \ldots, x_k)}\right) \mathrm{d}x_k \cdots \mathrm{d}x_1$$

Definition 10. *Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively and let $p_{X_1 \cdots X_k}$ and $p_{X_2 \cdots X_k}$ be the joint probability density functions of, respectively, $(X_1, X_2, \ldots, X_k)$ and $(X_2, \ldots, X_k)$. The conditional joint entropy of X_1 given $X_2, \ldots, X_k$, $H(X_1|X_2, \ldots, X_k)$, is defined as*

$$H(X_1|X_2, \ldots, X_k) = \int_{\Omega_{X_1}} \int_{\Omega_{X_2}} \cdots \int_{\Omega_{X_k}} p_{X_1 \cdots X_k}(x_1, \ldots, x_k) \log \left(\frac{p_{X_2 \cdots X_k}(x_2, \ldots, x_k)}{p_{X_1 \cdots X_k}(x_1, x_2, \ldots, x_k)} \right) \mathrm{d}x_k \cdots \mathrm{d}x_2 \mathrm{d}x_1$$

Figure E.1 gives a visual interpretation of the mutual information between two correlated random variables X and Y.

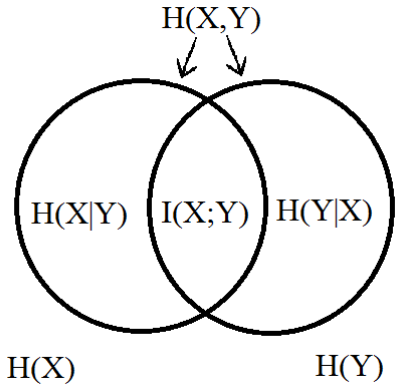

Figure E.1: Visual interpretation of the mutual information and entropies between two correlated random variables X and Y.

Also conditional mutual information can be considered.

Definition 11. *Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively and let $p_{X_1 \cdots X_k}$, $p_{X_2 \cdots X_k}$, $p_{X_1 \cdots X_{k-1}}$ and $p_{X_2 \cdots X_{k-1}}$ be the joint probability density functions of, respectively, $(X_1, X_2, \ldots, X_k)$, $(X_2, \ldots, X_k)$, $(X_1, X_2, \ldots, X_{k-1})$ and $(X_2, \ldots, X_{k-1})$. The conditional mutual information of X_1 and X_k given $X_2, \ldots, X_{k-1}$, $I(X_1; X_k|X_2, \ldots, X_{k-1})$, is defined as*

$$I(X_1; X_k|X_2, \ldots, X_{k-1}) = \int_{\Omega_{X_1}} \cdots \int_{\Omega_{X_k}} p_{X_1 \cdots X_k}(x_1, \ldots, x_k) \log \left(\frac{p_{X_2 \cdots X_{k-1}}(x_2, \ldots, x_{k-1}) p_{X_1 \cdots X_k}(x_1, \ldots, x_k)}{p_{X_1 \cdots X_{k-1}}(x_1, \ldots, x_{k-1}) p_{X_2 \cdots X_k}(x_2, \ldots, x_k)} \right) \mathrm{d}x_k \cdots \mathrm{d}x_1$$

The image in figure E.1 can also be extended to three random variables as seen in figure E.2.

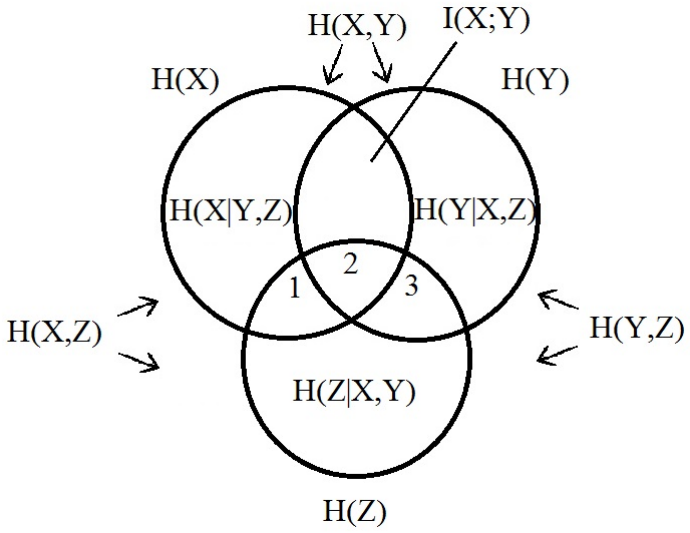

Figure E.2: Visual interpretation of the mutual information and entropies between three correlated random variables X, Y and Z.

The area included by the three circles can be interpreted as the joint entropy of the random variables X, Y and Z. Area 1 can be interpreted as the conditional mutual information $I(X;Z|Y)$, area 2 as $I(X;Y;Z)$ and area 3 as $I(Y;Z|X)$. The area included by area 1, 2 and 3 can be interpreted as the multivariate mutual information $I(X,Y;Z)$.

From figures E.1 and E.2 the following theorems can be deducted:

Theorem 1. *Let X and Y be two random variables with possible outcome sets, respectively, Ω_X and Ω_Y. Let p_X and p_Y be the probability density functions of X and Y respectively and let p_{XY} be the joint probability density function of (X,Y). The following equality holds:*

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Theorem 2. *Let X and Y be two random variables with possible outcome sets, respectively, Ω_X and Ω_Y. Let p_X and p_Y be the probability density functions of X and Y respectively and let p_{XY} be the joint probability density function of (X,Y). The following equality holds:*

$$I(X;Y) = H(X) - H(X|Y)$$

Appendix E.2.1 and E.2.2 give more formal proofs of the above theorems.

Furthermore, for multivariate mutual information the following theorem holds:

Theorem 3. *Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively and let $p_{X_1 \cdots X_k}$ be the joint probability density function of $(X_1, X_2, \ldots, X_k)$. For the multivariate mutual information of $(X_1, \ldots, X_{k-1})$ and X_k, $I(X_1, \ldots, X_{k-1}; X_k)$, the following equality holds:*

$$I(X_1, \ldots, X_{k-1}; X_k) = \sum_{i=1}^{k-1} I(X_i; X_k | X_1, \ldots, X_{i-1})$$

Using theorem 3, the following theorem can be conducted:

Theorem 4. *Let X, Y and Z be three random variables with possible outcome sets, respectively, Ω_X, Ω_Y and Ω_Z. Let p_X, p_Y and p_Z be the probability density functions of X, Y and Z respectively and let p_{XYZ} be the joint probability density function of (X,Y,Z). The following equality holds:*

$$I(X,Y;Z) = H(Z) + H(X,Y) - H(X,Y,Z)$$

Proofs of theorem 3 and 4 are given in appendix E.2.3 and E.2.4.

E.2 Proofs mutual information theorems

E.2.1 Proof theorem 1

To prove. Let X and Y be two random variables with possible outcome sets, respectively, Ω_X and Ω_Y. Let p_X and p_Y be the probability density functions of X and Y respectively and let p_{XY} be the joint probability density function of (X,Y). The following equality holds:

$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Proof. From definition 1 it is known that

$$\begin{aligned}
I(X;Y) &= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right)\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_{XY}(x,y)\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_X(x)\right)\mathrm{d}y\mathrm{d}x \\
&\quad - \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_Y(y)\right)\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_{XY}(x,y)\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\log(p_X(x))\left(\int_{\Omega_Y} p_{XY}(x,y)\mathrm{d}y\right)\mathrm{d}x \\
&\quad - \int_{\Omega_Y}\log(p_Y(y))\left(\int_{\Omega_X} p_{XY}(x,y)\mathrm{d}x\right)\mathrm{d}y \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_{XY}(x,y)\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\log(p_X(x))p_X(x)\mathrm{d}x \\
&\quad - \int_{\Omega_Y}\log(p_Y(y))p_Y(y)\mathrm{d}y
\end{aligned} \tag{E.1}$$

Using definitions 6 and 8 it can be concluded that equation E.1 equals

$$-H(X,Y) + H(X) + H(Y)$$

□

E.2.2 Proof theorem 2

To prove. Let X and Y be two random variables with possible outcome sets, respectively, Ω_X and Ω_Y. Let p_X and p_Y be the probability density functions of X and Y respectively and let p_{XY} be the joint probability density function of (X,Y). The following equality holds:

$$I(X;Y) = H(X) - H(X|Y)$$

Proof. From definition 1 it is known that

$$\begin{aligned}
I(X;Y) &= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p_X(x)p_Y(y)}\right)\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(\frac{p_{XY}(x,y)}{p_Y(y)}\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_X(x)\right)\mathrm{d}y\mathrm{d}x \\
&= -\int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(\frac{p_Y(y)}{p_{XY}(x,y)}\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_X(x)\right)\mathrm{d}y\mathrm{d}x \\
&= -\int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(\frac{p_Y(y)}{p_{XY}(x,y)}\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\log(p_X(x))\left(\int_{\Omega_Y} p_{XY}(x,y)\mathrm{d}y\right)\mathrm{d}x \\
&= -\int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(\frac{p_Y(y)}{p_{XY}(x,y)}\right)\mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\log(p_X(x))p_X(x)\mathrm{d}x
\end{aligned} \tag{E.2}$$

Using definitions 6 and 7 it can be concluded that equation E.2 equals

$$-H(X|Y) + H(X)$$

□

In the same way it can be shown that this theorem also holds for more variables:

$$I(X_1,\ldots,X_k;Y) = H(X_1,\ldots,X_k) - H(X_1,\ldots,X_k|Y)$$

E.2.3 Proof theorem 3

To prove. Let $X_1, X_2, \ldots, X_k$ be continuous random variables with possible outcome sets, respectively, $\Omega_{X_1}, \Omega_{X_2}, \ldots, \Omega_{X_k}$. Let $p_{X_1}, p_{X_2}, \ldots, p_{X_k}$ be the probability density functions of $X_1, X_2, \ldots, X_k$ respectively

and let $p_{X_1 \cdots X_k}$ be the joint probability density function of $(X_1, X_2, \ldots, X_k)$. For the multivariate mutual information of $(X_1, \ldots, X_{k-1})$ and X_k, $I(X_1, \ldots, X_{k-1}; X_k)$, the following equality holds:

$$I(X_1, \ldots, X_{k-1}; X_k) = \sum_{i=1}^{k-1} I(X_i; X_k | X_1, \ldots, X_{i-1})$$

Proof. The theorem will be proved for three variables, i.e. $I(X, Y; Z) = I(X; Z) + I(Y; Z|X)$. In this case, p_X, p_Y, p_Z denote the probability density functions of, respectively, X, Y and Z and p_{XYZ}, p_{XY}, p_{XZ} and p_{YZ} denote the joint probability density functions of, respectively, (X, Y, Z), (X, Y), (X, Z) and (Y, Z). Furthermore, the possible outcome sets of X, Y and Z are, respectively, Ω_X, Ω_Y and Ω_Z. The proof can easily be extended to show the version with more variables holds as well. From the extension of theorem 2 it is known that

$$\begin{aligned}
I(X, Y; Z) &= H(X, Y) - H(X, Y|Z) \\
&= -\int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x, y) \log\left(p_{XY}(x, y)\right) \mathrm{d}y\mathrm{d}x - \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{p_Z(z)}{p_{XYZ}(x, y, z)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x, y) \log\left(\frac{1}{p_{XY}(x, y)}\right) \mathrm{d}y\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{1}{p_Z(z)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(p_{XYZ}(x, y, z)\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x, y) \log\left(\frac{1}{p_{XY}(x, y)}\right) \mathrm{d}y\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Z} \log\left(\frac{1}{p_Z(z)}\right) \left(\int_{\Omega_Y} p_{XYZ}(x, y, z)\mathrm{d}y\right) \mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(p_{XYZ}(x, y, z)\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x, y) \log\left(\frac{1}{p_{XY}(x, y)}\right) \mathrm{d}y\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{1}{p_Z(z)}\right) \mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(p_{XYZ}(x, y, z)\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{1}{p_{XY}(x, y)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{1}{p_Z(z)}\right) \mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(p_{XYZ}(x, y, z)\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_X} p_{XZ}(x, z) \log\left(\frac{1}{p_Z(z)}\right) \mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{p_{XYZ}(x, y, z)}{p_{XY}(x, y)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{1}{p_Z(z)}\right) \mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{p_{XYZ}(x, y, z)}{p_{XY}(x, y)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{p_{XZ}(x, z)}{p_X(x)}\right) \mathrm{d}z\mathrm{d}x - \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{p_{XZ}(x, z)}{p_X(x)}\right) \mathrm{d}z\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{1}{p_Z(z)}\right) \mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{p_{XYZ}(x, y, z)}{p_{XY}(x, y)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{p_{XZ}(x, z)}{p_X(x)}\right) \mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{p_X(x)}{p_{XZ}(x, z)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x, z) \log\left(\frac{p_{XZ}(x, z)}{p_X(x)p_Z(z)}\right) \mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x, y, z) \log\left(\frac{p_X(x)p_{XYZ}(x, y, z)}{p_{XY}(x, y)p_{XZ}(x, z)}\right) \mathrm{d}z\mathrm{d}y\mathrm{d}x
\end{aligned} \tag{E.3}$$

According to definitions 1 and 11 equation E.3 equals

$$I(X; Z) + I(Y; Z|X)$$

□

E.2.4 Proof theorem 4

To prove. Let X, Y and Z be three random variables with possible outcome sets, respectively, Ω_X, Ω_Y and Ω_Z. Let p_X, p_Y and p_Z be the probability density functions of X, Y and Z respectively and let p_{XYZ} be the

joint probability density function of (X, Y, Z). Furthermore, denote by p_{XY}, p_{XZ} and p_{YZ} the joint probability density functions of, respectively, (X, Y), (X, Z) and (Y, Z). The following equality holds:

$$I(X, Y; Z) = H(Z) + H(X, Y) - H(X, Y, Z)$$

Proof. From theorem 3 it is known that

$$\begin{aligned}
I(X,Y;Z) &= I(X;Z) + I(Y;Z|X) \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{p_{XZ}(x,z)}{p_X(x)p_Z(z)}\right)\mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_X(x)p_{XYZ}(x,y,z)}{p_{XY}(x,y)p_{XZ}(x,z)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{p_{XZ}(x,z)}{p_X(x)}\right)\mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{1}{p_Z(z)}\right)\mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_{XYZ}(x,y,z)}{p_{XY}(x,y)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad - \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_{XZ}(x,z)}{p_X(x)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{p_{XZ}(x,z)}{p_X(x)}\right)\mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{1}{p_Z(z)}\right)\mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_{XYZ}(x,y,z)}{p_{XY}(x,y)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad - \int_{\Omega_X}\int_{\Omega_Z} \log\left(\frac{p_{XZ}(x,z)}{p_X(x)}\right)\left(\int_{\Omega_Y} p_{XYZ}(x,y,z)\mathrm{d}y\right)\mathrm{d}z\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{p_{XZ}(x,z)}{p_X(x)}\right)\mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{1}{p_Z(z)}\right)\mathrm{d}z\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_{XYZ}(x,y,z)}{p_{XY}(x,y)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad - \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{p_{XZ}(x,z)}{p_X(x)}\right)\mathrm{d}z\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Z} p_{XZ}(x,z)\log\left(\frac{1}{p_Z(z)}\right)\mathrm{d}z\mathrm{d}x + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_{XYZ}(x,y,z)}{p_{XY}(x,y)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{1}{p_Z(z)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad + \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(\frac{p_{XYZ}(x,y,z)}{p_{XY}(x,y)}\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(p_{XYZ}(x,y,z)\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad - \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(p_{XY}(x,y)p_Z(z)\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(p_{XYZ}(x,y,z)\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&\quad - \int_{\Omega_Z}\log\left(p_Z(z)\right)\left(\int_{\Omega_X}\int_{\Omega_Y} p_{XYZ}(x,y,z)\mathrm{d}y\mathrm{d}x\right)\mathrm{d}z \\
&\quad - \int_{\Omega_X}\int_{\Omega_Y}\log\left(p_{XY}(x,y)\right)\left(\int_{\Omega_Z} p_{XYZ}(x,y,z)\mathrm{d}z\right)\mathrm{d}y\mathrm{d}x \\
&= \int_{\Omega_X}\int_{\Omega_Y}\int_{\Omega_Z} p_{XYZ}(x,y,z)\log\left(p_{XYZ}(x,y,z)\right)\mathrm{d}z\mathrm{d}y\mathrm{d}x - \int_{\Omega_Z} p_Z(z)\log\left(p_Z(z)\right)\mathrm{d}z \\
&\quad - \int_{\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log\left(p_{XY}(x,y)\right)\mathrm{d}y\mathrm{d}x
\end{aligned} \tag{E.4}$$

According to definitions 6 and 8 equation E.4 equals

$$-H(X, Y, Z) + H(Z) + H(X, Y)$$

□

Note that this theorem can easily be extended to more variables, showing that

$$I(X_1, X_2, \ldots, X_k; Z) = H(Z) + H(X_1, X_2, \ldots, X_k) - H(X_1, X_2, \ldots, X_k, Z) \tag{E.5}$$

E.3 Proofs of lemmas used in Kraskov estimator

E.3.1 Proof of lemma 1

To prove. Consider an independent sample of size n drawn from a continuous random variable X. Assume the i^{th} point in this sample, denoted by x_i, is given. Let $D_X(i)$ be the random variable denoting the distance between x_i and its k^{th} nearest neighbor. In that case, the probability density function g of $D_X(i)$ approximately equals

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} P_3$$

with

$$\begin{aligned} P1 &= \int_{x_i-\epsilon}^{x_i+\epsilon} p_X(x)dx \\ P2 &= 1 - \int_{x_i-\epsilon}^{x_i+\epsilon} p_X(x)dx \\ P3 &= p_X(x_i-\epsilon) + p_X(x_i+\epsilon) \end{aligned}$$

Proof. Note that for a continuous random variable X with distribution function F and density function $F' = f$ the following holds

$$hf(x) = hF'(x) = \lim_{h\downarrow 0}\{F(x+h) - F(x)\} = \lim_{h\downarrow 0}\mathbb{P}(X \in [x, x+h]) \tag{E.6}$$

Now, let g be the probability density function of $D_X(i)$. Using equation E.6 it is found that

$$g(\epsilon) = \lim_{h\downarrow 0}\frac{\mathbb{P}(D_X(i) \in [\epsilon, \epsilon+h])}{h} = \lim_{h\downarrow 0}\frac{\mathbb{P}(|x_i - X_{i,k}| \in [\epsilon, \epsilon+h])}{h} \tag{E.7}$$

with $X_{i,k}$ the k^{th} nearest neighbor of x_i.

By the definition of $D_X(i)$ equation E.7 equals $\frac{1}{h}$ times the probability that

$$\begin{cases} |X_j - x_i| \in [\epsilon, \epsilon+h] & \text{for one } j,\ j \neq i \\ |X_j - x_i| < \epsilon & \text{for } k-1 \text{ } j,\ j \neq i \\ |X_j - x_i| > \epsilon + h & \text{for } n-k-1 \text{ } j,\ j \neq i \end{cases}$$

as $h \downarrow 0$.
Figure E.3 visualizes these three different regions in which a point could lie.

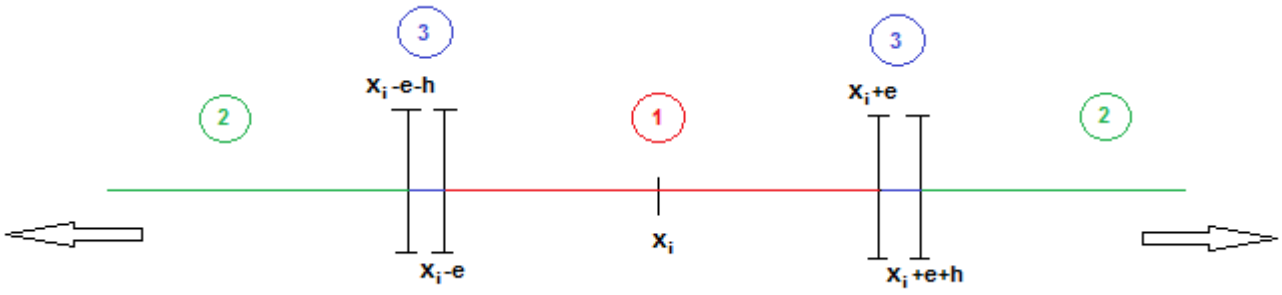

Figure E.3: Visualization of the three different regions in which a point x_j could lie.

The probability of a point falling in region 1 equals

$$P_1 = \int_{x_i-\epsilon}^{x_i+\epsilon} p_X(x)dx \tag{E.8}$$

The probability of a point falling in region 2 equals

$$P_2 = 1 - \int_{x_i-\epsilon}^{x_i+\epsilon} p_X(x)dx - \mathcal{O}(h) \text{ as } h \downarrow 0 \tag{E.9}$$

The probability of a point falling in region 3 equals

$$h \cdot P_3 = h \cdot (p_X(x_i - \epsilon) + p_X(x_i + \epsilon)) + \mathcal{O}(h^2) \text{ as } h \downarrow 0 \tag{E.10}$$

In other words, the random variable $Y(i) = |X_j - x_i|$ can take on values in three mutually exclusive regions with corresponding probabilities denoted in equations E.8, E.9 and E.10. Consider the independent sample of size n drawn from X with given point x_i, i.e. $n-1$ unknown trials remain. Let N_i denote the number of times $Y(i)$ takes on a value in region i based on this sample. Since the possible outcomes are mutual exclusive and the $n-1$ trials are independent, the random vector $N = (N_1, N_2, N_3)$ follows a multinomial distribution with parameters $n-1$ and the probabilities in equations E.8, E.9 and E.10.
Consider the following definition of the probability mass function of a random variable following a multinomial distribution:

Definition 12. *Let $X = (X_1, \ldots, X_k)$ be a random vector following a multinomial distribution with parameters n and $p_1, \ldots, p_k$. Let $x_1, \ldots, x_k$ denote realizations of this random vector with $\sum_{i=1}^{k} x_i = n$. The probability mass function f of this multinomial distribution is*

$$f(x_1, \ldots, x_k; n, p_1, \ldots, p_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

Now, setting $X = N$, $n = n-1$, $p_1 = P_1$, $p_2 = P_2$, $p_3 = h \cdot P_3$, $x_1 = k-1$, $x_2 = n-k-1$ and $x_3 = 1$ in definition 12 leads to the following probability mass function for the random vector N:

$$\frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} \cdot h \cdot P_3$$

Therefore, equation E.7, i.e. the probability density function of $D_X(i)$, approximately equals

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} P_3$$

□

E.3.2 Proof of lemma 2

To prove. Let $B(x, y)$ be the beta function with x and y positive integers. The following relation holds:

$$\frac{1}{B(x,y)} = \frac{(x+y-1)!}{(x-1)!\,(y-1)!}$$

Proof. It is known that

$$B(x,y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \tag{E.11}$$

with Γ the gamma function.
For x a positive integer, the following holds

$$\Gamma(x) = (x-1)!$$

Therefore, it follows that equation E.11 equals

$$\frac{(x-1)!\,(y-1)!}{(x+y-1)!}$$

Taking the inverse gives

$$\frac{1}{B(x,y)} = \frac{(x+y-1)!}{(x-1)!\,(y-1)!}$$

□

E.3.3 Proof of lemma 3

To prove. Let $B(x,y)$ be the beta function with $x,y>0$. Let $\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}$ be the digamma function with $\Gamma(x)$ the gamma function, $x>0$. The following relation holds:

$$\frac{\partial B(x,y)}{\partial x}=B(x,y)(\psi(x)-\psi(x+y))$$

Proof. It is known that

$$\begin{aligned}\frac{\partial B(x,y)}{\partial x}&=\frac{\partial}{\partial x}\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\\&=\frac{\Gamma(x+y)\Gamma'(x)\Gamma(y)-\Gamma(x)\Gamma(y)\Gamma'(x+y)}{(\Gamma(x+y))^2}\\&=\frac{\Gamma'(x)\Gamma(y)}{\Gamma(x+y)}-\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\frac{\Gamma'(x+y)}{\Gamma(x+y)}\\&=\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\frac{\Gamma'(x)}{\Gamma(x)}-\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}\frac{\Gamma'(x+y)}{\Gamma(x+y)}\end{aligned}$$

Using $\psi(x)=\frac{\Gamma'(x)}{\Gamma(x)}$ and $\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$, it now follows that

$$\frac{\partial B(x,y)}{\partial x}=B(x,y)(\psi(x)-\psi(x+y))$$

□

E.3.4 Proof of lemma 4

To prove. Consider an independent sample of size n drawn from a continuous random vector (X,Y). Assume the i^{th} point in this sample, denoted by (x_i,y_i), is given. Let $D_{XY}(i)$ be the random variable denoting the distance between (x_i,y_i) and its k^{th} nearest neighbor, where the distance is calculated using the maximum norm, i.e.

$$||(x_i,y_i)-(x_j,y_j)||_\infty=\max(|x_i-x_j|,|y_i-y_j|)$$

In that case, the probability density function g of $D_{XY}(i)$ approximately equals

$$g(\epsilon)=\frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!}P_1^{k-1}P_2^{n-k-1}P_3$$

with

$$\begin{aligned}P1&=\int_{x_i-\epsilon}^{x_i+\epsilon}\int_{y_i-\epsilon}^{y_i+\epsilon}p_{XY}(x,y)dydx\\P2&=1-\int_{x_i-\epsilon}^{x_i+\epsilon}\int_{y_i-\epsilon}^{y_i+\epsilon}p_{XY}(x,y)dydx\\P3&=\int_{x_i-\epsilon}^{x_i+\epsilon}p_{XY}(x,y_i-\epsilon)dx+\int_{x_i-\epsilon}^{x_i+\epsilon}p_{XY}(x,y_i+\epsilon)dx+\int_{y_i-\epsilon}^{y_i+\epsilon}p_{XY}(x_i-\epsilon,y)dy\\&\quad+\int_{y_i-\epsilon}^{y_i+\epsilon}p_{XY}(x_i+\epsilon,y)dy\end{aligned}$$

Proof. Let g be the probability density function of $D_{XY}(i)$. Similarly as in appendix E.3.1, it is found that

$$g(\epsilon)=\lim_{h\downarrow 0}\frac{\mathbb{P}(D_{XY}(i)\in[\epsilon,\epsilon+h])}{h}=\lim_{h\downarrow 0}\frac{\mathbb{P}(\max(|x_i-(X_i)_k|,|y_i-(Y_i)_k|)\in[\epsilon,\epsilon+h])}{h}\tag{E.12}$$

with $(X_i)_k$ and $(Y_i)_k$, respectively, the $x-$ and $y-$coordinate of the k^{th} nearest neighbor of (x_i,y_i).

By the definition of $D_{XY}(i)$ equation E.12 equals $\frac{1}{h}$ times the probability that

$$\begin{cases}||(X_j,Y_j)-(x_i,y_i)||_\infty\in[\epsilon,\epsilon+h] & \text{for one } j,\ j\neq i\\ ||(X_j,Y_j)-(x_i,y_i)||_\infty<\epsilon & \text{for } k-1\ j,\ j\neq i\\ ||(X_j,Y_j)-(x_i,y_i)||_\infty>\epsilon+h & \text{for } n-k-1\ j,\ j\neq i\end{cases}$$

as $h \downarrow 0$. Figure E.4 visualizes these three different regions in which a point could lie.

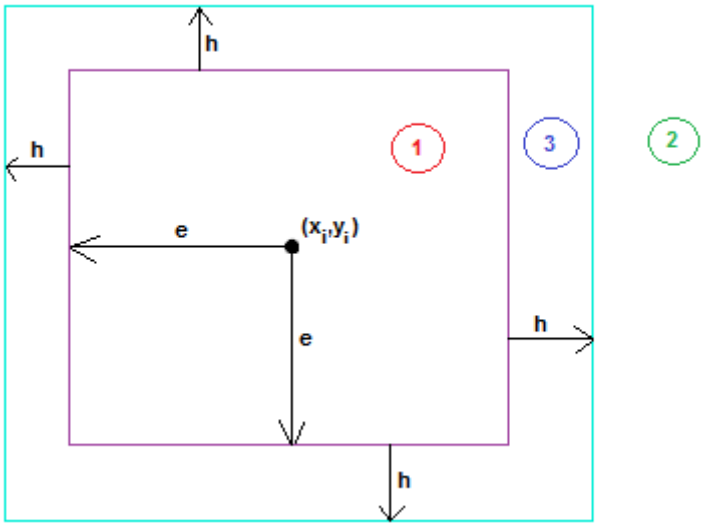

Figure E.4: Visualization of the three different regions in which a point (x_j, y_j) could lie.

The probability of a point falling in region 1 equals

$$P_1 = \int_{x_i-\epsilon}^{x_i+\epsilon} \int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x,y)dydx \tag{E.13}$$

The probability of a point falling in region 2 approximately equals

$$P_2 = 1 - \int_{x_i-\epsilon}^{x_i+\epsilon} \int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x,y)dydx \text{ as } h \downarrow 0 \tag{E.14}$$

The probability of a point falling in region 3 approximately equals

$$\begin{aligned} h \cdot P_3 = h \cdot \Bigg(&\int_{x_i-\epsilon}^{x_i+\epsilon} p_{XY}(x, y_i-\epsilon)dx + \int_{x_i-\epsilon}^{x_i+\epsilon} p_{XY}(x, y_i+\epsilon)dx \\ &+ \int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x_i-\epsilon, y)dy + \int_{y_i-\epsilon}^{y_i+\epsilon} p_{XY}(x_i+\epsilon, y)dy \Bigg) \text{ as } h \downarrow 0 \end{aligned} \tag{E.15}$$

In other words, the random variable $Z(i) = ||(X_j, Y_j) - (x_i, y_i)||_\infty$ can take on values in three mutually exclusive regions with corresponding probabilities denoted in equations E.13, E.14 and E.15. Consider the independent sample of size n drawn from (X, Y) with given point (x_i, y_i), i.e. $n-1$ unknown trials remain. Let N_i denote the number of times $Z(i)$ takes on a value in region i based on this sample of size $n-1$ and point (x_i, y_i). Since the possible outcomes are mutually exclusive and the $n-1$ trials are independent, the random vector $N = (N_1, N_2, N_3)$ follows a multinomial distribution with parameters $n-1$ and the probabilities in equations E.13, E.14 and E.15.

Now, setting $X = N$, $n = n-1$, $p_1 = P_1$, $p_2 = P_2$, $p_3 = h \cdot P_3$, $x_1 = k-1$, $x_2 = n-k-1$ and $x_3 = 1$ in definition 12 in appendix E.3.1 leads to the following probability mass function for the random vector N:

$$\frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} \cdot h \cdot P_3$$

Therefore, equation E.12, i.e. the probability density function of $D_{XY}(i)$, approximately equals

$$g(\epsilon) = \frac{(n-1)!}{1!\,(k-1)!\,(n-k-1)!} P_1^{k-1} P_2^{n-k-1} P_3$$

□

E.4 Derivation adjusted Kraskov estimator

Let X and Y be, respectively, a categorical and a continuous random variable with possible outcome sets, respectively, Ω_X and Ω_Y, joint probability density function p_{XY}, marginal probability (density) functions q_X with $q_X(x) = \int_{\Omega_Y} p_{XY}(x,y)dy$ and p_Y with $p_Y(y) = \sum_{x\in\Omega_X} p_{XY}(x,y)$ and conditional probability density function $p_{Y|X}$ with $p_{Y|X}(y|x) = \frac{p_{XY}(x,y)}{q_X(x)}$. For a categorical variable X and continuous variable Y, define the k^{th} nearest neighbor of (x,y), with $x \in \Omega_X$, $y \in \Omega_Y$, as the k^{th} nearest neighbor of y among all points for which the value of X equals x.

From theorem 2 it is known that

$$\begin{aligned} I(X;Y) &= H(Y) - H(Y|X) \\ &= -\int_{\Omega_Y} p_Y(y)\log(p_Y(y))dy + \sum_{x\in\Omega_X}\int_{\Omega_Y} p_{XY}(x,y)\log(p_{Y|X}(y|x))dy \\ &= -\mathbb{E}[\log(p_Y(Y))] + \mathbb{E}[\log(p_{Y|X}(Y|X))] \end{aligned} \tag{E.16}$$

Since p_Y and $p_{Y|X}$ are not known, equation E.16 has to be rewritten. First consider $\mathbb{E}[\log(p_Y(Y))]$. Similarly as in the derivation of the Kraskov estimator, it is found that

$$\mathbb{E}[\log(p_Y(Y))] \approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_y(i))\right) \tag{E.17}$$

with $d_y(i)$ the realization based on the sample $y_1,\ldots,y_n$ of $D_Y(i) = |y_i - Y_{i,k}|$ with $Y_{i,k}$ the k^{th} nearest neighbor of y_i.

Now, consider $\mathbb{E}[\log(p_{Y|X}(Y|X))]$.
Consider again drawing a sample of size n from the random vector (X,Y). Assume the i^{th} point in this sample, denoted by (x_i,y_i), is given. Choose a fixed, small ϵ. Denote by the random variable $D_{Y|X}(i)$ the distance between y_i and its k^{th} nearest neighbor among all points for which the value of X equals x_i. Furthermore, denote by $n_d(i)$ the number of points in the sample for which the value of X equals x_i.

Let g be the probability density function of $D_{Y|X}(i)$. Now consider the following lemma:

Lemma 5. *Consider an independent sample of size n drawn from a random vector (X,Y) with X a categorical random variable and Y a continuous random variable. Assume the i^{th} point in this sample, denoted by (x_i,y_i), is given. Let $D_{Y|X}(i)$ be the random variable denoting the distance between y_i and its k^{th} nearest neighbor among all points for which the value of X equals x_i. Furthermore, denote by $n_d(i)$ the number of points in the sample for which the value of X equals x_i. In that case, the probability density function g of $D_{Y|X}(i)$ approximately equals*

$$g(\epsilon) = \frac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!}P_1^{k-1}P_2^{n_d(i)-k-1}P_3$$

with

$$\begin{aligned} P_1 &= \int_{y_i-\epsilon}^{y_i+\epsilon} p_{Y|X}(y|x_i)dy \\ P_2 &= 1 - \int_{y_i-\epsilon}^{y_i+\epsilon} p_{Y|X}(y|x_i)dy \\ P_3 &= p_{Y|X}(y_i-\epsilon|x_i) + p_{Y|X}(y_i+\epsilon|x_i) \end{aligned}$$

A proof of lemma 5 is given in appendix E.4.1.

Now, assume $p_{Y|X}(y|x)$ is smooth in the interval $[y_i-\epsilon, y_i+\epsilon]$ and ϵ is small. Using lemma 5, the probability density function g of $D_{Y|X}(i)$ now approximately equals

$$g(\epsilon) = \frac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!}\left(2\epsilon p_{Y|X}(y_i|x_i)\right)^{k-1}\left(2p_{Y|X}(y_i|x_i)\right)\left(1-2\epsilon p_{Y|X}(y_i|x_i)\right)^{n_d(i)-k-1} \tag{E.18}$$

Using equation E.18, it is now found that

$$\begin{aligned}\mathbb{E}[\log(2D_{Y|X}(i)p_{Y|X}(y_i|x_i))] &= \int_0^\infty \log(2\epsilon p_{Y|X}(y_i|x_i))g(\epsilon)d\epsilon \\ &= C\int_0^\infty \log(q)q^{k-1}(1-q)^{n_d(i)-k-1}2p_{Y|X}(y_i|x_i)d\epsilon(i) \end{aligned} \tag{E.19}$$

with $C = \dfrac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!}$ and $q = 2\epsilon p_{Y|X}(y_i|x_i)$.
Transforming to q, equation E.19 equals

$$\begin{aligned}&\frac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!}\int_0^1 \log(q)q^{k-1}(1-q)^{n_d(i)-k-1}2p_{y|X}(y_i|x_i)\frac{1}{2p_{y|X}(y_i|x_i)}dq \\ =\ &\frac{(n_d(i)-1)!}{(k-1)!\,(n_d(i)-k-1)!}\int_0^1 \log(q)q^{k-1}(1-q)^{n_d(i)-k-1}dq\end{aligned}$$

In the same way as in the derivation of the Kraskov estimator, it is now found that

$$\mathbb{E}[\log(2D_{Y|X}(i)p_{Y|X}(y_i|x_i))] = \psi(k) - \psi(n_d(i)) \tag{E.20}$$

Using equation E.20, it follows that

$$\begin{aligned}\log(p_{Y|X}(y_i|x_i)) &= \mathbb{E}[\log(p_{Y|X}(y_i|x_i))] \\ &= \mathbb{E}[\log(2D_{Y|X}(i)p_{Y|X}(y_i|x_i)) - \log(2D_{Y|X}(i))] \\ &= \mathbb{E}[\log(2D_{Y|X}(i)p_{Y|X}(y_i|x_i))] - \mathbb{E}[\log(2D_{Y|X}(i))] \\ &= \psi(k) - \psi(n_d(i)) - \mathbb{E}[\log(2D_{Y|X}(i))]\end{aligned} \tag{E.21}$$

Now, assume the points $(x_1, y_1), \ldots, (x_n, y_n)$ in the sample are all known.
In that case, $\mathbb{E}[\log(2D_{Y|X}(i))]$ can be estimated as follows:

$$\widehat{\mathbb{E}[\log(2D_{Y|X}(i))]} = \frac{1}{n}\sum_{i=1}^{n}\log(2d_{y|x}(i)) \tag{E.22}$$

with $d_{y|x}(i)$ the realization of $D_{Y|X}(i)$ for the sample $(x_1, y_1), \ldots, (x_n, y_n)$.
Furthermore, $\mathbb{E}[\log(p_{Y|X}(Y|X))]$ can be estimated as follows:

$$\widehat{\mathbb{E}[\log(p_{Y|X}(Y|X))]} = \frac{1}{n}\sum_{i=1}^{n}\log(p_{Y|X}(y_i|x_i)) \tag{E.23}$$

Using equations E.21, E.22 and E.23, it can be concluded that

$$\mathbb{E}[\log(p_{Y|X}(Y|X))] \approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n_d(i)) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{y|x}(i))\right) \tag{E.24}$$

For the same reason as in the derivation of the Kraskov estimator, set $d_y(i) = d_{y|x}(i) = |y_i - y_{i,k,x_i}|$. That way, $d_y(i)$ is the distance to the $(n_y(i)+1)^{th}$ nearest neighbor of y_i with $n_y(i)$ the number of points y_j for which $|y_i - y_j| < d_y(i)$.

Using equations E.17 and E.24, the following is now found:

$$\begin{aligned}\mathbb{E}[\log(p_Y(Y))] &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_y(i)+1) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{y|x}(i))\right) \\ &= \frac{1}{n}\sum_{i=1}^{n}\psi(n_y(i)+1) - \psi(n) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{y|x}(i)) \\ \mathbb{E}[\log(p_{Y|X}(Y|X))] &\approx \frac{1}{n}\sum_{i=1}^{n}\left(\psi(k) - \psi(n_d(i)) - \frac{1}{n}\sum_{i=1}^{n}\log(2d_{y|x}(i))\right) \\ &= \psi(k) - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_d(i)) + \log(2d_{y|x}(i))\right)\end{aligned}$$

Using these equations and equation E.16, it is now found that:

$$I(X;Y) \approx \psi(n) + \psi(k) - \frac{1}{n}\sum_{i=1}^{n}\left(\psi(n_y(i)+1) - \psi(n_d(i))\right)$$

E.4.1 Proof of lemma 5

To prove. Consider an independent sample of size n drawn from a random vector (X, Y) with X a categorical random variable and Y a continuous random variable. Assume the i^{th} point in this sample, denoted by (x_i, y_i), is given. Let $D_{Y|X}(i)$ be the random variable denoting the distance between y_i and its k^{th} nearest neighbor among all points for which the value of X equals x_i. Furthermore, denote by $n_d(i)$ the number of points in the sample for which the value of X equals x_i. In that case, the probability density function g of $D_{Y|X}(i)$ approximately equals

$$g(\epsilon) = \frac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!} P_1^{k-1} P_2^{n_d(i)-k-1} P_3$$

with

$$\begin{aligned} P_1 &= \int_{y_i-\epsilon}^{y_i+\epsilon} p_{Y|X}(y|x_i)dy \\ P_2 &= 1 - \int_{y_i-\epsilon}^{y_i+\epsilon} p_{Y|X}(y|x_i)dy \\ P_3 &= p_{Y|X}(y_i-\epsilon|x_i) + p_{Y|X}(y_i+\epsilon|x_i) \end{aligned}$$

Proof. Let g be the probability density function of $D_{Y|X}(i)$. Similarly as in appendix E.3.1, it is found that

$$g(\epsilon) = \lim_{h\downarrow 0} \frac{\mathbb{P}(D_{Y|X}(i) \in [\epsilon, \epsilon+h])}{h} = \lim_{h\downarrow 0} \frac{\mathbb{P}(|y_i - Y_{i,k,x_i}| \in [\epsilon, \epsilon+h])}{h} \tag{E.25}$$

with Y_{i,k,x_i} the k^{th} nearest neighbor of y_i among all points for which the value of X equals x_i.

By the definition of $D_{Y|X}(i)$ equation E.25 equals $\frac{1}{h}$ times the probability that

$$\begin{cases} |y_i - Y_j| \in [\epsilon, \epsilon+h] & \text{for one } j \in A(i),\ j \neq i \\ |y_i - Y_j| < \epsilon & \text{for } k-1\ j \in A(i),\ j \neq i \\ |y_i - Y_j| > \epsilon + h & \text{for } n_d(i)-k-1\ j \in A(i),\ j \neq i \end{cases}$$

as $h \downarrow 0$ with $A(i) \subset \{1, \ldots, n\}$ such that $X_j = x_i\ \forall j \in A(i)$ and $n_d(i) = |A(i)|$. Figure E.5 visualizes these three different regions in which a point could lie.

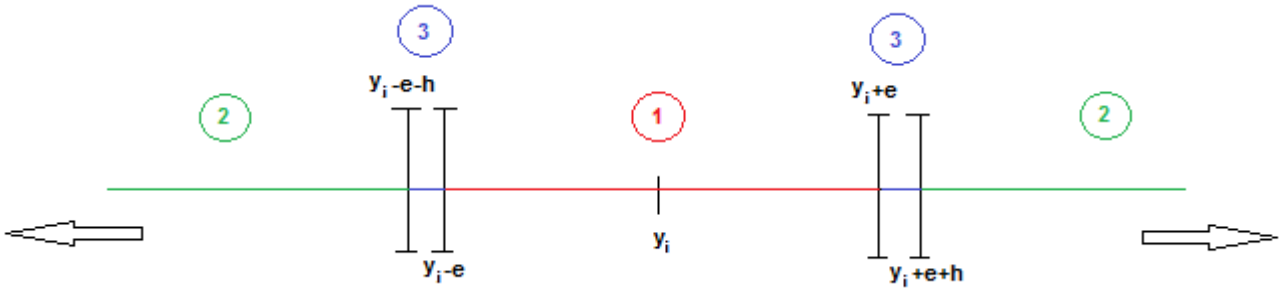

Figure E.5: Visualization of the three different regions in which a point y_j could lie.

The probability of a point falling in region 1 equals

$$P_1 = \int_{y_i-\epsilon}^{y_i+\epsilon} p_{Y|X}(y|x_i)dy \tag{E.26}$$

The probability of a point falling in region 2 approximately equals

$$P_2 = 1 - \int_{y_i-\epsilon}^{y_i+\epsilon} p_{Y|X}(y|x_i)dy \text{ as } h \downarrow 0 \tag{E.27}$$

The probability of a point falling in region 3 approximately equals

$$h \cdot P_3 = h \cdot (p_{Y|X}(y_i-\epsilon|x_i) + p_{Y|X}(y_i+\epsilon|x_i)) \text{ as } h \downarrow 0 \tag{E.28}$$

In other words, the random variable $Z(i) = |Y_j - y_i|$ with $j \in A(i)$,$A(i) \subset \{1, \ldots, n\}$ such that $X_j = x_i$ $\forall j \in A(i)$, can take on values in three mutually exclusive regions with corresponding probabilities denoted in equations E.26, E.27 and E.28. Consider the independent sample of size n drawn from (X, Y) with given point (x_i, y_i), i.e. $n-1$ unknown trials remain. Let N_i denote the number of times $Z(i)$ takes on a value in region i based on this sample of size $n-1$ and point (x_i, y_i). Since the possible outcomes are mutually exclusive and the $n-1$ trials are independent, the random vector $N = (N_1, N_2, N_3)$ follows a multinomial distribution with parameters $n-1$ and the probabilities in equations E.26, E.27 and E.28.

Now, setting $X = N$, $n = n-1$, $p_1 = P_1$, $p_2 = P_2$, $p_3 = h \cdot P_3$, $x_1 = k-1$, $x_2 = n-k-1$ and $x_3 = 1$ in definition 12 in appendix E.3.1 leads to the following probability mass function for the random vector N:

$$\frac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!} P_1^{k-1} P_2^{n_d(i)-k-1} \cdot h \cdot P_3$$

Therefore, equation E.25, i.e. the probability density function of $D_{Y|X}(i)$, approximately equals

$$g(\epsilon) = \frac{(n_d(i)-1)!}{1!\,(k-1)!\,(n_d(i)-k-1)!} P_1^{k-1} P_2^{n_d(i)-k-1} P_3$$

□

E.5 Proof mRMR feature selection

To prove. In the case of forward feature selection, selecting features by the use of the mRMR algorithm is similar to selecting features based on maximization of high-dimensional mutual information values $I(X_1, \ldots, X_k; Z)$.

Proof. Let $\mathbf{X}$ be the set of all features and S_m the set of already selected features after iteration m. In iteration $m+1$ an optimal feature then needs to be selected from the set $F = \mathbf{X} \backslash S_m$. Note that if $m = 0$, S_m is empty.

The optimal feature is now found by maximizing $I(S_m, X_{m+1}; Z)$ with $X_{m+1} \in F$. From the extension of theorem 4, it is known that the following equation holds:

$$I(S_m, X_{m+1}; Z) = H(Z) + H(S_m, X_{m+1}) - H(S_m, X_{m+1}, Z) \tag{E.29}$$

Let $p_{X_1}, \ldots, p_{X_{m+1}}, p_Z$ be the probability density functions of, respectively, $X_1, \ldots, X_{m+1}$ and Z. Furthermore, let $p_{X_1 \cdots Z}$ and $p_{X_1 \cdots X_{m+1}}$ be the joint probability density functions of, respectively, $(X_1, \ldots, X_{m+1}, Z)$ and $(X_1, \ldots, X_{m+1})$.
Define

$$J(S_m, X_{m+1}) = \int_{x_1} \cdots \int_{x_{m+1}} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \log \left(\frac{p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1})}{p_{X_1}(x_1) \cdots p_{X_{m+1}}(x_{m+1})} \right) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \tag{E.30}$$

and

$$J(S_m, X_{m+1}, Z) = \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, z) \log \left(\frac{p_{X_1 \cdots Z}(x_1, \ldots, z)}{p_{X_1}(x_1) \cdots p_{X_{m+1}}(x_{m+1}) p_Z(z)} \right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \tag{E.31}$$

Equation E.30 equals

$$\begin{aligned}
&\int_{x_1} \cdots \int_{x_{m+1}} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \log \left(p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \right) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&- \int_{x_1} \cdots \int_{x_{m+1}} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \log \left(p_{X_1}(x_1) \cdots p_{X_{m+1}}(x_{m+1}) \right) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
= &\int_{x_1} \cdots \int_{x_{m+1}} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \log \left(p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \right) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&- \int_{x_1} \log \left(p_{X_1}(x_1) \right) \left(\int_{x_2} \cdots \int_{x_{m+1}} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_2 \right) \mathrm{d}x_1 \\
&- \ldots - \int_{x_{m+1}} \log \left(p_{X_{m+1}}(x_{m+1}) \right) \left(\int_{x_1} \cdots \int_{x_m} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \mathrm{d}x_m \cdots \mathrm{d}x_1 \right) \mathrm{d}x_{m+1}
\end{aligned}$$

$$
\begin{aligned}
&= \int_{x_1} \cdots \int_{x_{m+1}} p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1}) \log\left(p_{X_1 \cdots X_{m+1}}(x_1, \ldots, x_{m+1})\right) \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&\quad - \int_{x_1} \log\left(p_{X_1}(x_1)\right) p_{X_1}(x_1) \mathrm{d}x_1 - \ldots - \int_{x_{m+1}} \log\left(p_{X_{m+1}}(x_{m+1})\right) p_{X_{m+1}}(x_{m+1}) \mathrm{d}x_{m+1} \\
&= -H(S_m, X_{m+1}) + H(X_1) + \cdots + H(X_{m+1})
\end{aligned}
$$

In the same way it can be shown that equation E.31 equals

$$-H(S_m, X_{m+1}, Z) + H(X_1) + \cdots + H(X_{m+1}) + H(Z)$$

The following equations are now found

$$H(S_m, X_{m+1}) = \sum_{i=1}^{m+1} H(X_i) - J(S_m, X_{m+1})$$

$$H(S_m, X_{m+1}, Z) = \sum_{i=1}^{m+1} H(X_i) + H(Z) - J(S_m, X_{m+1}, Z)$$

Substituting these equations into E.29 leads to

$$
\begin{aligned}
I(S_m, X_{m+1}; Z) &= H(Z) + \sum_{i=1}^{m+1} H(X_i) - J(S_m, X_{m+1}) - \left(\sum_{i=1}^{m+1} H(X_i) + H(Z) - J(S_m, X_{m+1}, Z)\right) \\
&= J(S_m, X_{m+1}, Z) - J(S_m, X_{m+1})
\end{aligned}
$$

In summary,

$$I(S_m, X_{m+1}; Z) = J(S_m, X_{m+1}, Z) - J(S_m, X_{m+1}),$$

so maximizing $I(S_m, X_{m+1}; Z)$ is similar to maximizing $J(S_m, X_{m+1}, Z) - J(S_m, X_{m+1})$, i.e. maximizing $J(S_m, X_{m+1}, Z)$ and minimizing $J(S_m, X_{m+1})$ at the same time.

Now it will be shown that maximizing $J(S_m, X_{m+1}, Z)$ is equal to using the maximum relevance criterion to find an optimal feature X_{m+1} and minimizing $J(S_m, X_{m+1})$ is equal to using the minimum redundancy criterion to find an optimal feature X_{m+1}. Combination of these two leads to the mRMR algorithm with forward feature selection.

Denote by $p_{X_1|X_2 \cdots Z}$ and $p_{X_{m+1}|Z}$ the conditional probability density functions of, respectively, $(X_1|X_2, \ldots, X_{m+1}, Z)$ and $(X_{m+1}|Z)$.
Consider $J(S_m, X_{m+1}, Z)$.

$$
\begin{aligned}
J(S_m, X_{m+1}, Z) &= \int_{x_1} \cdots \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(\frac{p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z)}{p_{X_1}(x_1) \cdots p_{X_{m+1}}(x_{m+1}) p_Z(z)}\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&= \int_{x_1} \cdots \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, z) \log\left(\frac{p_{X_1|X_2 \cdots Z}(x_1|x_2, \ldots, z) \cdots p_{X_{m+1}|Z}(x_{m+1}|z) p_Z(z)}{p_{X_1}(x_1) \cdots p_{X_{m+1}}(x_{m+1}) p_Z(z)}\right) \mathrm{d}z \cdots \mathrm{d}x_1 \\
&= \int_{x_1} \cdots \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_{X_1|X_2 \cdots X_{m+1}Z}(x_1|x_2, \ldots, x_{m+1}, z)\right) \mathrm{d}z \cdots \mathrm{d}x_1 \\
&\quad + \ldots + \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_{X_{m+1}|Z}(x_{m+1}|z)\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&\quad + \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_Z(z)\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&\quad - \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_{X_1}(x_1)\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&\quad - \ldots - \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_{X_{m+1}}(x_{m+1})\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&\quad - \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_Z(z)\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1 \\
&= \int_{x_1} \cdots \int_{x_{m+1}} \int_{z} p_{X_1 \cdots Z}(x_1, \ldots, x_{m+1}, z) \log\left(p_{X_1|X_2 \cdots Z}(x_1|x_2, \ldots, x_{m+1}, z)\right) \mathrm{d}z \mathrm{d}x_{m+1} \cdots \mathrm{d}x_1
\end{aligned}
$$

$$
\begin{aligned}
&+\ldots+\int_{x_1}\cdots\int_{x_{m+1}}\int_{z} p_{X_1\cdots Z}(x_1,\ldots,x_{m+1},z)\log\left(p_{X_{m+1}|Z}(x_{m+1}|z)\right)\mathrm{d}z\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
&-\int_{x_1}\cdots\int_{x_{m+1}}\int_{z} p_{X_1\cdots Z}(x_1,\ldots,x_{m+1},z)\log\left(p_{X_1}(x_1)\right)\mathrm{d}z\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
&-\ldots-\int_{x_1}\cdots\int_{x_{m+1}}\int_{z} p_{X_1\cdots Z}(x_1,\ldots,x_{m+1},z)\log\left(p_{X_{m+1}}(x_{m+1})\right)\mathrm{d}z\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
=&\int_{x_1}\cdots\int_{x_{m+1}}\int_{z} p_{X_1\cdots Z}(x_1,\ldots,x_{m+1},z)\log\left(\frac{p_{X_1\cdots Z}(x_1,x_2,\ldots,x_{m+1},z)}{p_{X_2\cdots Z}(x_2,\ldots,x_{m+1},z)}\right)\mathrm{d}z\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
&+\ldots+\int_{x_1}\cdots\int_{z} p_{X_1\cdots Z}(x_1,\ldots,z)\log\left(\frac{p_{X_{m+1}Z}(x_{m+1},z)}{p_Z(z)}\right)\mathrm{d}z\cdots\mathrm{d}x_1+\sum_{i=1}^{m+1}H(X_i)\\
=&-H(X_1|X_2,\ldots,X_{m+1},Z)-H(X_2|X_3,\ldots,X_{m+1},Z)-\ldots-H(X_{m+1}|Z)+\sum_{i=1}^{m+1}H(X_i)\\
\leq&\sum_{i=1}^{m+1}H(X_i)
\end{aligned}
\tag{E.32}
$$

The inequality sign in equation E.32 changes into an equality sign if all the conditional entropies are zero. This is the case if the variables $X_1,\ldots,X_{m+1},Z$ are maximally dependent. Since the first m features are already selected, this dependency criterion means that feature X_{m+1} and Z should be maximally dependent, which is exactly the maximum relevance criterion.

Now consider $J(S_m,X_{m+1})$. From Jensen's inequality it is known that

$$
\phi\left(\mathbb{E}\left[X\right]\right)\leq\mathbb{E}\left[\phi(X)\right] \tag{E.33}
$$

for X a random variable and ϕ a convex function. Equality holds if X is a degenerate random variable.
Now consider the random variable $Y(X_1,\ldots,X_{m+1})=\frac{p_{X_1}(X_1)\ldots p_{X_{m+1}}(X_{m+1})}{p_{X_1\cdots X_{m+1}}(X_1,\ldots,X_{m+1})}$ and $\phi(x)=-\log(x)$, which is a convex function.
Using equation E.33 the following holds:

$$
\begin{aligned}
J(S_m,X_{m+1})&=\int_{x_1}\cdots\int_{x_{m+1}}p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})\log\left(\frac{p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})}{p_{X_1}(x_1)\cdots p_{X_{m+1}}(x_{m+1})}\right)\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
&=-\int_{x_1}\cdots\int_{x_{m+1}}p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})\log\left(\frac{p_{X_1}(x_1)\cdots p_{X_{m+1}}(x_{m+1})}{p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})}\right)\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
&=\int_{x_1}\cdots\int_{x_{m+1}}p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})\left(-\log\left(\frac{p_{X_1}(x_1)\cdots p_{X_{m+1}}(x_{m+1})}{p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})}\right)\right)\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\\
&=E\left[\phi\left(Y(X_1,\ldots,X_{m+1})\right)\right]\\
&\geq\phi\left(E\left[Y(X_1,\ldots,X_{m+1})\right]\right)\\
&=-\log\left(\int_{x_1}\cdots\int_{x_{m+1}}p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})\frac{p_{X_1}(x_1)\cdots p_{x_{m+1}}(x_{m+1})}{p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})}\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\right)\\
&=-\log\left(\int_{x_1}\cdots\int_{x_{m+1}}p_{X_1}(x_1)\cdots p_{X_{m+1}}(x_{m+1})\mathrm{d}x_{m+1}\cdots\mathrm{d}x_1\right)\\
&=-\log\left(1\right)=0
\end{aligned}
$$

In summary, $J(S_m,X_{m+1})$ is bounded from below by zero. The minimum is reached if $Y(X_1,\ldots,X_{m+1})$ is a degenerate random variable, i.e. if $p_{X_1}(x_1)\cdots p_{X_{m+1}}(x_{m+1})$ equals $p_{X_1\cdots X_{m+1}}(x_1,\ldots,x_{m+1})$ almost everywhere. This is the case if the random variables $X_1,\ldots,X_m$ and X_{m+1} are all independent of each other. Since the first m features are already selected, this independence criterion means that the mutual information between the variable X_{m+1} and each of the already selected variables is minimized, which is exactly the minimum redundancy criterion.

In summary, selecting a feature X_{m+1} by maximizing $J(S_m,X_{m+1},Z)$ and minimizing $J(S_m,X_{m+1})$ with S_m the already selected features is the same as selecting a feature X_{m+1} by using the mRMR algorithm with forward feature selection. Since the former is also equal to maximizing $I(S_m,X_{m+1};Z)$ with forward feature

selection, the equality between mRMR and maximizing $I(S_m, X_{m+1}; Z)$ with forward feature selection has now been shown. □

Appendix F

Results feature selection

F.1 Ranking list features

Author's note: this section is confidential.

F.2 Principal components selected features

Author's note: this section is confidential.

Appendix G

Results hierarchical clustering scheme

G.1 Results accurate versus non-accurate playing styles

Author's note: this section is confidential.

G.2 Results *Hollandse School* versus *Tiki Taka* matches

Author's note: this section is confidential.

G.3 Results *Counterplay* versus *Kick and Rush* matches

Author's note: this section is confidential.

Appendix H

Results Eredivisie and chi-squared tests

H.1 Eredivisie hierarchical clustering results

Author's note: this section is confidential.

H.2 Results Chi-squared tests

Author's note: this section is confidential.

Bibliography

Alelyani, S., Tang, J., and Liu, H. (2013). Feature selection for clustering: A review. *Data Clustering: Algorithms and Applications*, 29.

Castellano, J., Casamichana, D., and Lago, C. (2012). The use of match statistics that discriminate between successful and unsuccessful soccer teams. *Journal of human kinetics*, 31:137–147.

Dash, M. and Liu, H. (1997). Feature selection for classification. *Intelligent data analysis*, 1(3):131–156.

Doquire, G., Verleysen, M., et al. (2012). A comparison of multivariate mutual information estimators for feature selection. In *ICPRAM (1)*, pages 176–185.

Grunz, A., Memmert, D., and Perl, J. (2012). Tactical pattern recognition in soccer games by means of special self-organizing maps. *Human movement science*, 31(2):334–343.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, 3(Mar):1157–1182.

Hastie, T. J., Tibshirani, R. J., and Friedman, J. H. (2011). *The elements of statistical learning: data mining, inference, and prediction*. Springer.

Huang, J., Cai, Y., and Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters*, 28(13):1825–1844.

Jäger, J. M. and Schöllhorn, W. I. (2007). Situation-orientated recognition of tactical patterns in volleyball. *Journal of Sports Sciences*, 25(12):1345–1353.

Jäger, J. M. and Schöllhorn, W. I. (2012). Identifying individuality and variability in team tactics by means of statistical shape analysis and multilayer perceptrons. *Human movement science*, 31(2):303–317.

Kempe, M., Grunz, A., and Memmert, D. (2015). Detecting tactical patterns in basketball: Comparison of merge self-organising maps and dynamic controlled neural networks. *European journal of sport science*, 15(4):249–255.

Kjäll (2015). Midtjylland: Meet the men behind moneyball fc.

Kraskov, A., Stögbauer, H., and Grassberger, P. (2004). Estimating mutual information. *Physical review E*, 69(6):066138.

Lago-Peñas, C., Lago-Ballesteros, J., Dellal, A., and Gómez, M. (2010). Game-related statistics that discriminated winning, drawing and losing teams from the spanish soccer league. *Journal of Sports Science and Medicine*, 9(2):288–293.

Lorenzo Calvo, A., Gómez Ruano, M. Á., Ortega Toro, E., Ibañez Godoy, S. J., and Sampaio, J. (2010). Game related statistics which discriminate between winning and losing under-16 male basketball games. *Journal of Sports Science and Medicine*, 9(4):664–668.

Mooij (2013). Ball possession classification using clustering algorithms. Master thesis, Erasmus University, Rotterdam, Netherlands.

Moura, F. A., Martins, L. E. B., and Cunha, S. A. (2014). Analysis of football game-related statistics using multivariate techniques. *Journal of sports sciences*, 32(20):1881–1887.

Niu, Z., Gao, X., and Tian, Q. (2012). Tactic analysis based on real-world ball trajectory in soccer video. *Pattern Recognition*, 45(5):1937–1947.

Pena, J. L. and Touchette, H. (2012). A network theory analysis of football strategies. *arXiv preprint arXiv:1206.6904*.

Peng, H., Long, F., and Ding, C. (2005). Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on pattern analysis and machine intelligence*, 27(8):1226–1238.

Pfeiffer, M. and Perl, J. (2006). Analysis of tactical structures in team handball by means of artificial neural networks. *International Journal of Computer Science in Sport*, 5(1):4–14.

Pollard, R., Reep, C., and Hartley, S. (1988). The quantitative comparison of playing styles in soccer. *Science and football*, pages 309–315.

Ross, B. C. (2014). Mutual information between discrete and continuous data sets. *PloS one*, 9(2):e87357.

Wang, J. R. and Parameswaran, N. (2005). Analyzing tennis tactics from broadcasting tennis video clips. In *11th International Multimedia Modelling Conference*, pages 102–106. IEEE.

Wang, M. (2014). Evaluating technical and tactical abilities of football teams in euro 2012 based on improved information entropy model and som neural networks. *International Journal of Multimedia and Ubiquitous Engineering*, 9(11):293–302.

Wang, Q., Zhu, H., Hu, W., Shen, Z., and Yao, Y. (2015). Discerning tactical patterns for professional soccer teams: An enhanced topic model with applications. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 2197–2206. ACM.

Yeung, R. W. (2008). *Information theory and network coding*. Springer Science & Business Media.

Zhou, J., Fu, Z., and Robles-Kelly, A. (2009). Learning the optimal transformation of salient features for image classification. In *Digital Image Computing: Techniques and Applications, 2009. DICTA'09.*, pages 125–131. IEEE.