**Facioscapulohumeral muscular dystrophy**

**1. Introduction**

Facioscapulohumeral muscular dystrophy (FSHD) is a genetic muscle disorder in which the muscles of the face, shoulder blades and upper arms are among the most affected.

FSHD may be inherited through either the father or the mother, or it may occur without a family history. It is almost always associated with a genetic flaw (mutation) that leads to a shorter than usual segment of DNA on chromosome 4. The segment isn’t part of any particular gene, but it nevertheless seems to interfere with the correct processing of genetic material.

For the research will be used 14 samples of FMD and 18 samples of Normal states with 22,283 variables.

**2. Algorithms used for research**

The process of research will contain several steps. Firstly should be made the reduction of variables made by Fisher’s scoring and extracting first variables with highest weight. Then will be made comparing of classification methods using Naive Bayes, k-Nearest Neighbors and Discriminant functions of LDC and QDC to evaluate the minimum error while classifying the data. Comparing will contains two steps: classification with original data after scoring and data after PCA implementation.

Let us review the general description of mentioned algorithms.

**2.1 Fisher’s Score**

In mathematical statistics, the Fisher information (sometimes simply called information) is a way of measuring the amount of information that an observable random variable X carries about an unknown parameter θ of a distribution that models X. Formally, it is the variance of the score, or the expected value of the observed information. In Bayesian statistics, the asymptotic distribution of the posterior mode depends on the Fisher information and not on the prior (according to the Bernstein–von Mises theorem, which was anticipated by Laplace for exponential families). The role of the Fisher information in the asymptotic theory of maximum-likelihood estimation was emphasized by the statistician Ronald Fisher (following some initial results by Francis Ysidro Edgeworth). The Fisher information is also used in the calculation of the Jeffreys prior, which is used in Bayesian statistics.

The Fisher-information matrix is used to calculate the covariance matrices associated with maximum-likelihood estimates.

**2.2 PCA**

Principal component analysis (PCA) is a statistical procedure that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of distinct principal components is equal to the smaller of the number of original variables or the number of observations minus one. This transformation is defined in such a way that the first principal component has the largest possible variance (that is, accounts for as much of the variability in the data as possible), and each succeeding component in turn has the highest variance possible under the constraint that it is orthogonal to the preceding components. The resulting vectors are an uncorrelated orthogonal basis set. PCA is sensitive to the relative scaling of the original variables.

PCA Algorithm:

1. Calculate the average of the data

2. Subtract the mean of the data

3. Calculate the covariance matrix

4. Calculate the values and characteristic vectors of the covariance matrix

5. Sort the vectors according to the   characteristic values from highest to lowest

6. Project the data in the first L main components

**2.3 Classification algorithms**

***Naive Bayes***

Naive Bayes methods are a set of supervised learning algorithms based on applying Bayes’ theorem with the “naive” assumption of independence between every pair of features. Given a class variable y and a dependent feature vector x_1 through  x_n, Bayes’ theorem states the following relationship:

P(y \mid x_1, \dots, x_n) = \frac{P(y) P(x_1, \dots x_n \mid y)}
                                 {P(x_1, \dots, x_n)}

Using the naive independence assumption that

P(x_i | y, x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) = P(x_i | y),

for all i, this relationship is simplified to’

P(y \mid x_1, \dots, x_n) = \frac{P(y) \prod_{i=1}^{n} P(x_i \mid y)}
                                 {P(x_1, \dots, x_n)}

Since P(x_1, \dots, x_n) is constant given the input, we can use the following classification rule:

P(y \mid x_1, \dots, x_n) \propto P(y) \prod_{i=1}^{n} P(x_i \mid y)

\Downarrow

\hat{y} = \arg\max_y P(y) \prod_{i=1}^{n} P(x_i \mid y),

and we can use Maximum A Posteriori (MAP) estimation to estimate P(y) and P(x_i \mid y); the former is then the relative frequency of class y in the training set.

***k-Nearest Neighbors***

In pattern recognition, the k-nearest neighbors algorithm (k-NN) is a non-parametric method used for classification and regression. In both cases, the input consists of the k closest training examples in the feature space. The output depends on whether k-NN is used for classification or regression:

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

In k-NN regression, the output is the property value for the object. This value is the average of the values of its k nearest neighbors.

k-NN is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k-NN algorithm is among the simplest of all machine learning algorithms.

***Discriminant functions of LDC and QDC***

Both LDA and QDA can be derived from simple probabilistic models which model the class conditional distribution of the data P(X|y=k) for each class k. Predictions can then be obtained by using Bayes’ rule:

P(y=k | X) = \frac{P(X | y=k) P(y=k)}{P(X)} = \frac{P(X | y=k) P(y = k)}{ \sum_{l} P(X | y=l) \cdot P(y=l)}

and we select the class k which maximizes this conditional probability.

More specifically, for linear and quadratic discriminant analysis, P(X|y) is modelled as a multivariate Gaussian distribution with density:

p(X | y=k) = \frac{1}{(2\pi)^n |\Sigma_k|^{1/2}}\exp\left(-\frac{1}{2} (X-\mu_k)^t \Sigma_k^{-1} (X-\mu_k)\right)

To use this model as a classifier, we just need to estimate from the training data the class priors P(y=k) (by the proportion of instances of class k), the class means \mu_k (by the empirical sample class means) and the covariance matrices (either by the empirical sample class covariance matrices, or by a regularized estimator: see the section on shrinkage below).

In the case of LDA, the Gaussians for each class are assumed to share the same covariance matrix: \Sigma_k = \Sigma for all k. This leads to linear decision surfaces between, as can be seen by comparing the log-probability ratios \log[P(y=k | X) / P(y=l | X)]:

\log\left(\frac{P(y=k|X)}{P(y=l | X)}\right) = 0 \Leftrightarrow (\mu_k-\mu_l)\Sigma^{-1} X = \frac{1}{2} (\mu_k^t \Sigma^{-1} \mu_k - \mu_l^t \Sigma^{-1} \mu_l)

In the case of QDA, there are no assumptions on the covariance matrices \Sigma_k of the Gaussians, leading to quadratic decision surfaces.

**3. Analysis**

**3.1 Fisher’s Score**

Implementing Fisher’s score we will get the next results represented on Fig. 1.



Fig. 1 – Fishers Score

That is why we can describe firs three gens that have more influence than others.

**3.1.1 HOXC10 (218959\_at)**

***Gene Information and Sequence***

* HOXC10 spans 5215 bps of chromosome 12 from 53985065 to 53990279.
* HOXC10 has 5 transcripts containing a total of 10 exons on the forward strand.
* Annotation for this gene includes both automatic annotation from Ensembl and [Havana](https://vega.sanger.ac.uk/index.html) manual curation, see [article](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/info/genome/genebuild/genome_annotation.html).
* [View the gene sequence in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000180818;r=12:53985065-53990279)
* [View the chromosome region for this gene in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000180818;r=12:53985065-53990279)

***Variations***

* HOXC10 has 790 SNPs.
* [View sequence variations such as polymorphisms, along with genotypes and disease associations in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Image?db=core;g=ENSG00000180818;r=12:53985065-53990279)

***Orthologues***

* HOXC10 has 77 orthologues in Ensembl
* [View homology between species inferred from a gene tree in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Compara_Ortholog?db=core;g=ENSG00000180818;r=12:53985065-53990279)

***Paralogues***

* HOXC10 has 13 paralogues in Ensembl
* [View homology arising from a duplication event, inferred from a gene tree in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Compara_Paralog?db=core;g=ENSG00000180818;r=12:53985065-53990279)

***Regulation***

* There are 3 regulatory elements located in the region of HOXC10.
* [View the gene regulatory elements, such as promoters, transcription binding sites, and enhancers in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXIvZ2VuZT90ZXJtPUhPWEMxMCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TTVVTRzAwMDAwMDIyNDg0&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Regulation?db=core;g=ENSG00000180818;r=12:53985065-53990279)

**3.1.2 Fasciculation and elongation protein zeta 2 (215000\_s\_at)**

***Gene Information and Sequence***

* FEZ2 spans 114283 bps of chromosome 2 from 36531805 to 36646087.
* FEZ2 has 16 transcripts containing a total of 49 exons on the reverse strand.
* Annotation for this gene includes both automatic annotation from Ensembl and [Havana](https://vega.sanger.ac.uk/index.html) manual curation, see [article](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/info/genome/genebuild/genome_annotation.html).
* [View the gene sequence in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000171055;r=2:36531805-36646087)
* [View the chromosome region for this gene in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000171055;r=2:36531805-36646087)

***Variations***

* FEZ2 has 16315 SNPs.
* [View sequence variations such as polymorphisms, along with genotypes and disease associations in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Image?db=core;g=ENSG00000171055;r=2:36531805-36646087)

***Orthologues***

* FEZ2 has 92 orthologues in Ensembl
* [View homology between species inferred from a gene tree in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Compara_Ortholog?db=core;g=ENSG00000171055;r=2:36531805-36646087)

***Paralogues***

* FEZ2 has 1 paralogue in Ensembl
* [View homology arising from a duplication event, inferred from a gene tree in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Compara_Paralog?db=core;g=ENSG00000171055;r=2:36531805-36646087)

***Regulation***

* There are 17 regulatory elements located in the region of FEZ2.
* [View the gene regulatory elements, such as promoters, transcription binding sites, and enhancers in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1GRVoyJmNsYXNzaWZpY2F0aW9uPTk2MDYmdGlkPW5hbWVPcmdFTlNHMDAwMDAxNzEwNTU=&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Regulation?db=core;g=ENSG00000171055;r=2:36531805-36646087)

### 3.1.3 Cytoskeleton associated protein 4 (200999\_s\_at)

***Gene Information and Sequence***

* CKAP4 spans 66403 bps of chromosome 12 from 106237877 to 106304279.
* CKAP4 has 3 transcripts containing a total of 8 exons on the reverse strand.
* Annotation for this gene includes both automatic annotation from Ensembl and [Havana](https://vega.sanger.ac.uk/index.html) manual curation, see [article](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1DS0FQNCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TRzAwMDAwMTM2MDI2&hash=hash&url=https://www.ensembl.org/info/genome/genebuild/genome_annotation.html).
* [View the gene sequence in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1DS0FQNCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TRzAwMDAwMTM2MDI2&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Sequence?db=core;g=ENSG00000136026;r=12:106237877-106304279)
* [View the chromosome region for this gene in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1DS0FQNCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TRzAwMDAwMTM2MDI2&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG00000136026;r=12:106237877-106304279)

***Variations***

* CKAP4 has 7778 SNPs.
* [View sequence variations such as polymorphisms, along with genotypes and disease associations in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1DS0FQNCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TRzAwMDAwMTM2MDI2&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Variation_Gene/Image?db=core;g=ENSG00000136026;r=12:106237877-106304279)

***Orthologues***

* CKAP4 has 68 orthologues in Ensembl
* [View homology between species inferred from a gene tree in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1DS0FQNCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TRzAwMDAwMTM2MDI2&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Compara_Ortholog?db=core;g=ENSG00000136026;r=12:106237877-106304279)

***Paralogues***

* CKAP4 has no paralogues in Ensembl

***Regulation***

* There are 15 regulatory elements located in the region of CKAP4.
* [View the gene regulatory elements, such as promoters, transcription binding sites, and enhancers in Ensembl.](https://www.ebi.ac.uk/s4/jump?from=aHR0cDovL3d3dy5lYmkuYWMudWsvczQvc3VtbWFyeS9tb2xlY3VsYXI/dGVybT1DS0FQNCZjbGFzc2lmaWNhdGlvbj05NjA2JnRpZD1uYW1lT3JnRU5TRzAwMDAwMTM2MDI2&hash=hash&url=https://www.ensembl.org/Homo_sapiens/Gene/Regulation?db=core;g=ENSG00000136026;r=12:106237877-106304279)

**3.2 PCA**

After implementing PCA to scored data first three variables cover the 95.26% of variations. All weights are represented on Fig. 2.

As can be seen from Fig. 3 and Fig. 4 represented data can be clearly separated in to two classes.



Fig. 2 – Part of variation captured by firs 31 components



Fig. 3 – PCA two components



Fig. 4 – PCA three components

After implementing mentioned methods of classification all obtained errors can be seen on Fig. 5 - Fig. 8 for data after PCA and Fig. 9 - Fig. 12 for raw data after scoring.

For error accumulating were used methods of cross validation for 20 times for each iteration.



Fig. 5 – Naive Bayes errors   
obtained using data after PCA



Fig. 6 – kNN errors obtained using data after PCA



Fig. 7 – LDC errors obtained using data after PCA



Fig. 8 – QDC errors obtained using data after PCA



Fig. 9 – Naive Bayes errors   
obtained using original data



Fig. 10 – kNN errors obtained using original data



Fig. 11 – LDC errors obtained using original data



Fig. 12 – QDC errors obtained using original data

More shortly results are presented in   
Tab 1 – Tab 3.

**Table 1**

Average Errors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NB** | **kNN** | **LDC** | **QDC** |
| **Original data** | 0.533854 | 0.059375 | 0.486632 | 0.565799 |
| **PCA data** | 0.529514 | 0.057118 | 0.471007 | 0.57691 |

**Table 2**

Minimal Errors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NB** | **kNN** | **LDC** | **QDC** |
| **Original data** | 0.4278 | 0.02778 | 0.4056 | 0.06667 |
| **PCA data** | 0.3889 | 0.0222 | 0.3889 | 0.3778 |

**Table 3**

Amount of components used to have a minimal errors

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **NB** | **kNN** | **LDC** | **QDC** |
| **Original data** | 18 | 20 | 18 | 1 |
| **PCA data** | 22 | 20 | 20 | 11 |

**4. Conclusions**

For this case kNN methods gives the smallest error 2.22% using the first 20 variables or 5.71% of average error using the PCA data and 2.8% with 5.93% with raw data.

**References**

1. Larose, Daniel T. Discovering knowledge in data: an introduction to data mining. John Wiley & Sons, 2014.

2. Abdi, Hervé, and Lynne J. Williams. "Principal component analysis." Wiley interdisciplinary reviews: computational statistics 2.4 (2010): 433-459.

3. Lachenbruch, Peter A., Cheryl Sneeringer, and Lawrence T. Revo. "Robustness of the linear and quadratic discriminant function to certain types of non‐normality." Communications in Statistics 1.1 (1973): 39-56.

4. Murphy, Kevin P. "Naive bayes classifiers." University of British Columbia (2006).