To Higgs or not to Higgs

Project 1 - Statistical Methods in Data Mining

Bruno Parracho and Telmo Monteiro

8th November 2023

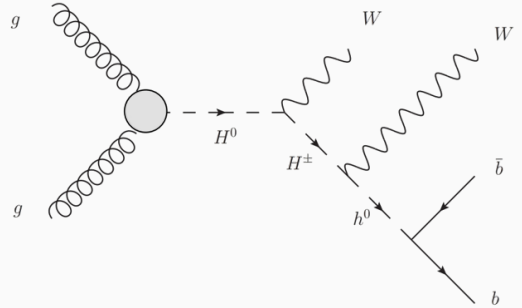
Mathematics Department, Faculty of Sciences of University of Porto

Introduction

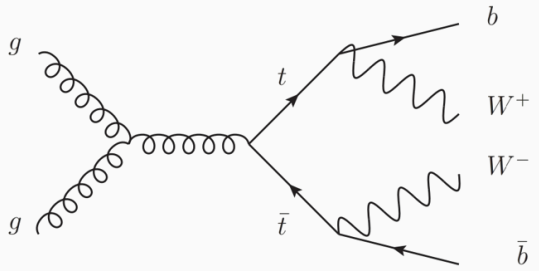
- Colliding particles, creates new particles
- Nobel Prize of Physics 2013: Higgs detection

Objective: Which collisions produce the Higgs

- Comparing the signal process and the background (null hypothesis)



(a) Signal Process.



(b) Background Process.

- Final products: 4 jets (quarks), 1 lepton (electron or muon) and 1 neutrino
- 21 Physical measurements: Transverse momentum (p_T), the pseudo-rapidity (η) and the angle (ϕ), and other quantities to fix conservation laws
- 7 Mass distributions

Data set

- Extracted 10 000 random points from the original data set.
- 21 features that are kinematic properties and 7 features that are functions of the first 21. Grand total of 28 features
- Target: 4639 for label 0 and 5361 for label 1.

Feature	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
lepton_pT	0.2747	0.5908	0.8655	1.0023	1.2509	5.7551
lepton_eta	-2.431080	-0.732722	0.006277	0.000139	0.725553	2.427076
lepton_phi	-1.74195	-0.82033	0.02844	0.02264	0.88376	1.74268
missing_energy_magnitude	0.0133	0.5662	0.8871	0.9958	1.2916	5.4989
missing_energy_phi	-1.74390	-0.88670	-0.01693	-0.01688	0.84569	1.74264
jet_1_pt	0.1945	0.6784	0.9020	0.9988	1.1796	5.7914
jet_1_eta	-2.941998	-0.673382	0.009382	0.009000	0.689175	2.943928
jet_1_phi	-1.741237	-0.853128	0.013466	0.001928	0.868451	1.741454
jet_1_b.tag	0.000	0.000	1.087	1.001	2.173	2.173
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
m_bb	0.07663	0.67552	0.87273	0.97943	1.14437	10.86245
m_wbb	0.3904	0.8232	0.9517	1.0407	1.1500	7.0254
m_wwbb	0.4206	0.7724	0.8752	0.9653	1.0682	5.4601

Table: Some basic statistics about the data used.

- Features have different scales

Principal Component Analysis (PCA)

Objectives

Reduce the high-dimensionality of the space

- Projection: $\mathbb{R}^N \rightarrow \mathbb{R}^K, K < N$
- Linear transformation $\rightarrow$ New efficient basis (more variance)
- This new basis vectors: $v'_i = \sum_j \lambda_j v_{ij}$
- Optimal number of Principal components (PCs): Kaiser-Guttman, Pearson's and the Cattell's criterions
- PCA: **prcomp** with scaling of the variables
- Biplot: **fviz_biplot**

PCA - Correlation matrix and PCs

- Highest eigenvalue → More influence on the PCs

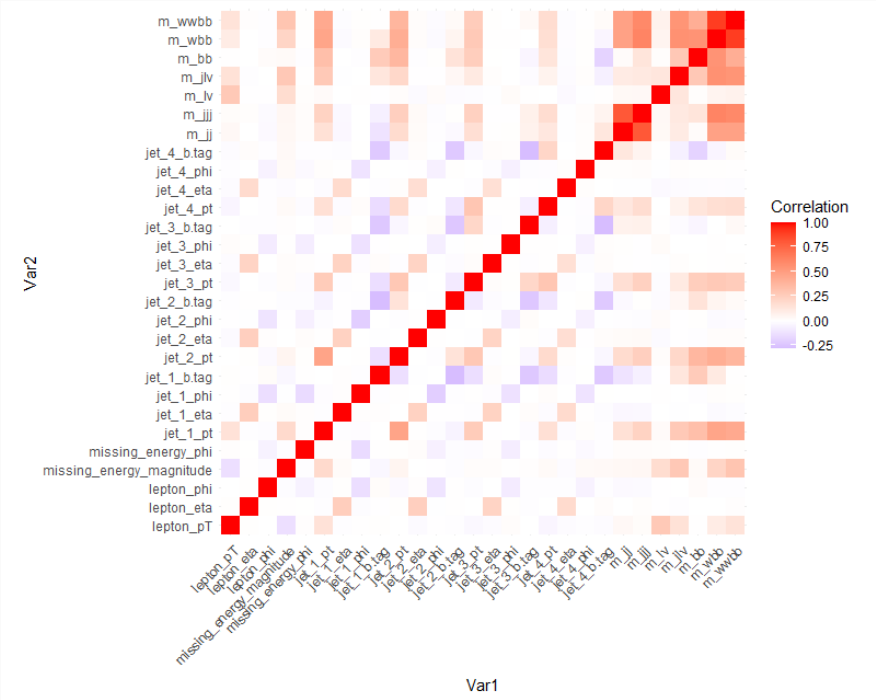


Figure: Correlation matrix of original variables using Pearson correlation coefficient.

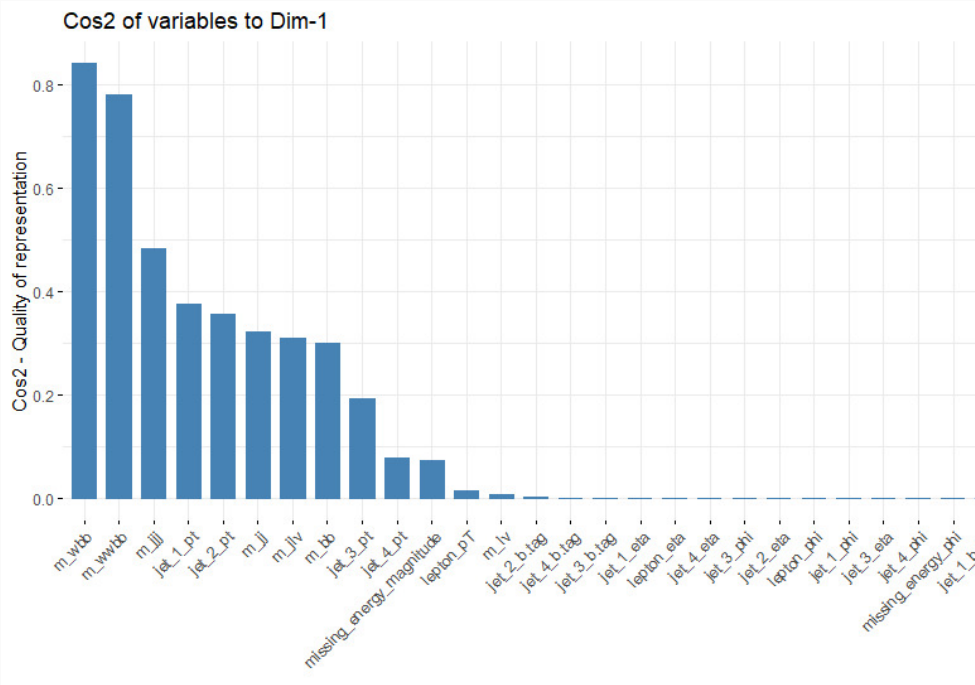


Figure: Dependence of the First PC with the original variables

PCA-Biplot

- Perpendicular vectors ⇒ Uncorrelated
- Opposite ⇒ Negatively correlated
- Co-linear ⇒ Positively correlated

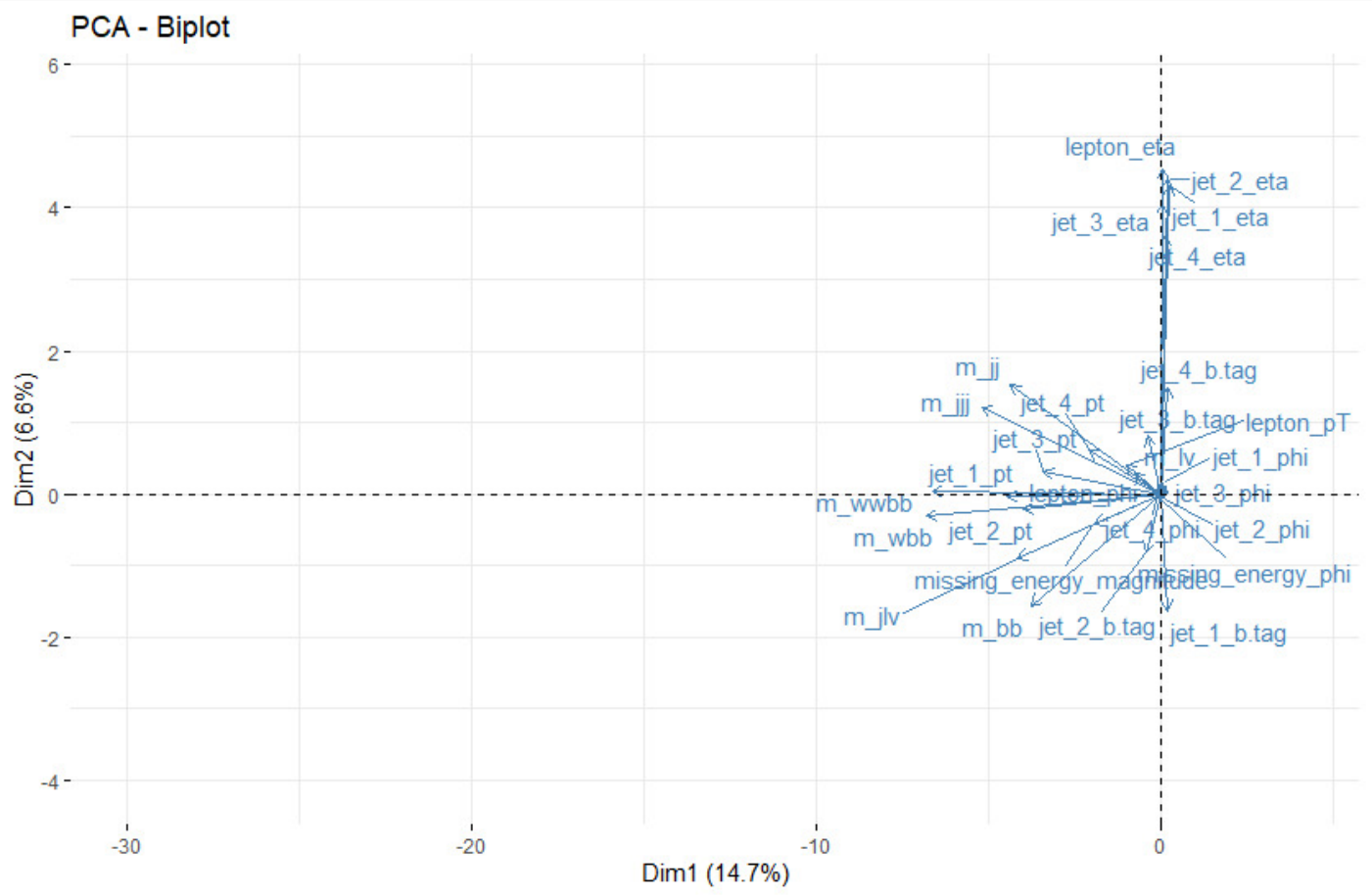
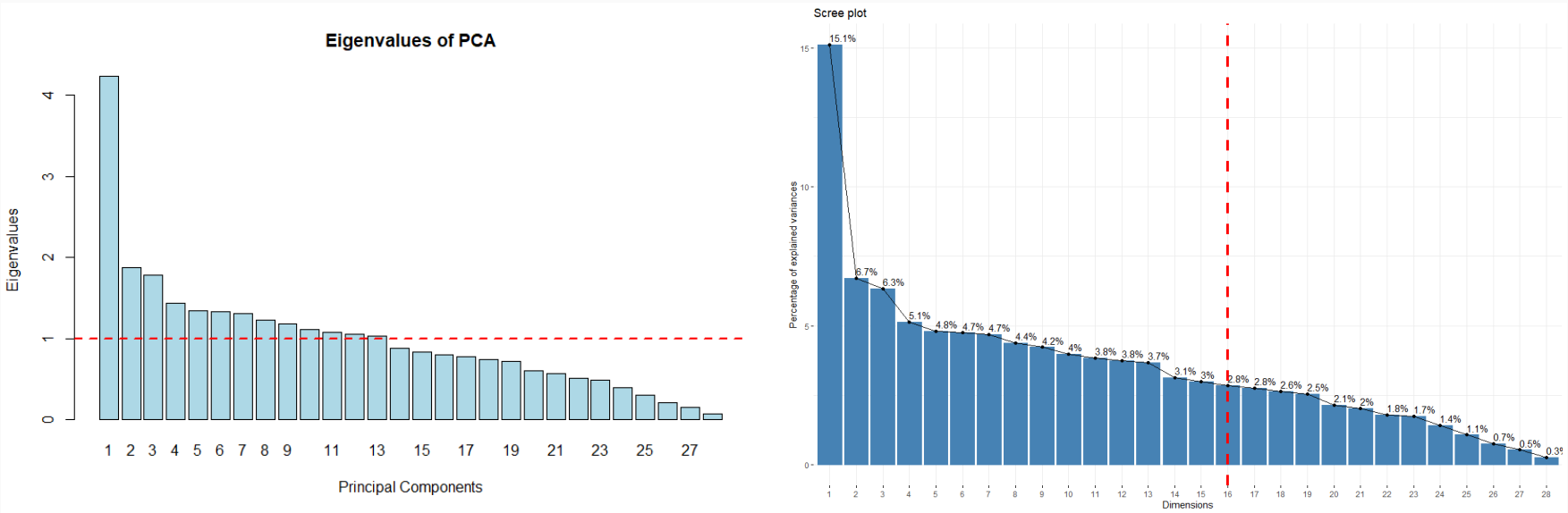


Figure: Biplot with the correlation between variables

PCA - Criterion

Kaiser-Guttman Criterion: remove the PC for which the eigenvalues are below 1.

Pearson's Criterion: maintain the q components that explain at least 80% of the variance.



Cattell's Criterion: eigenvalues of the components to be kept need to have a small difference with respect to previous component, such that $\lambda_\alpha - \lambda_{\alpha-1} < \epsilon$.

There is no possible ϵ that keeps the first principal component and eliminates some of the others, so we opt to not use this criterion.

PCA - Correlation with the target

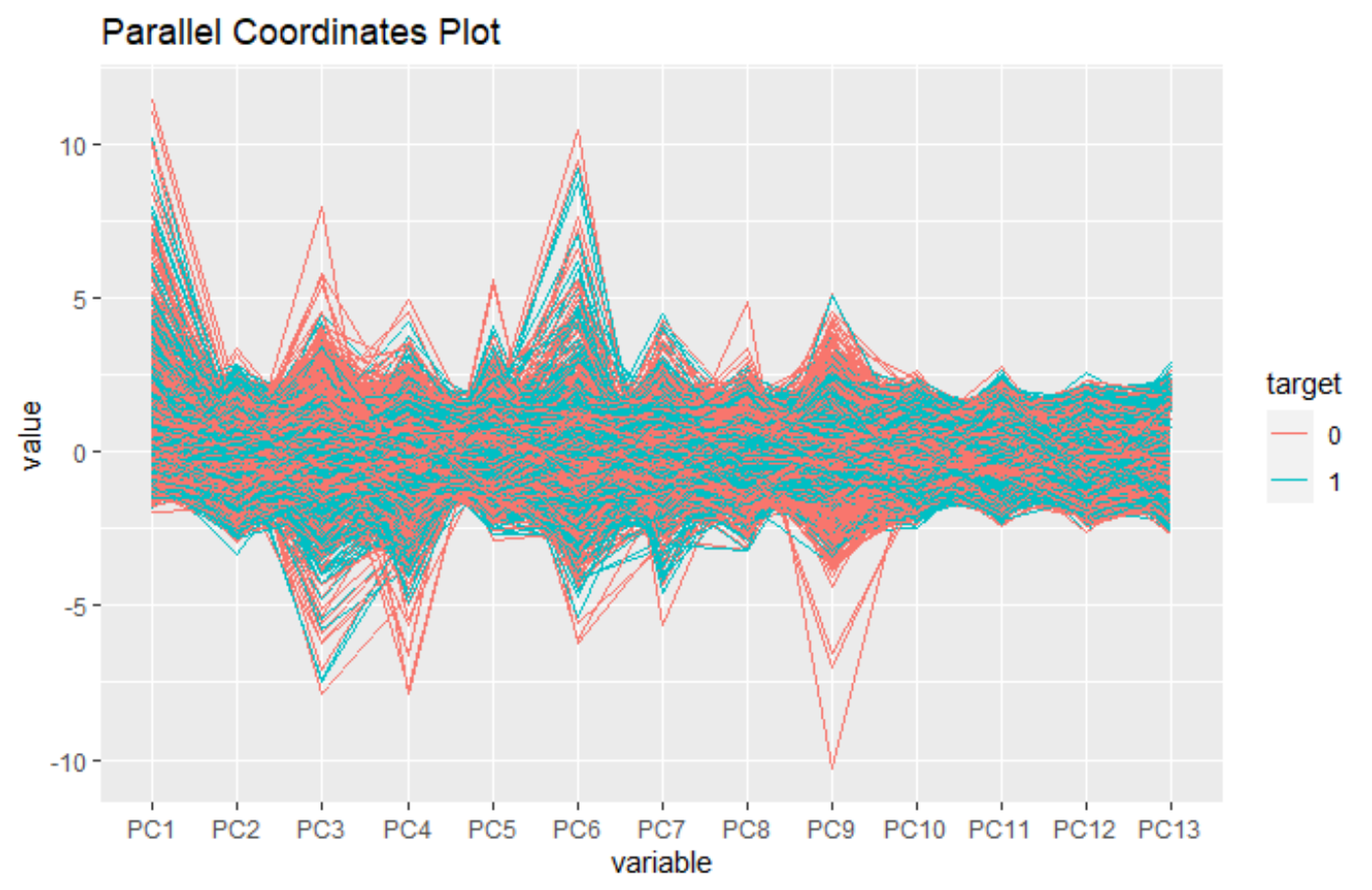


Figure: Parallel coordinates plot for the 13 PC that were kept, colored according to label.

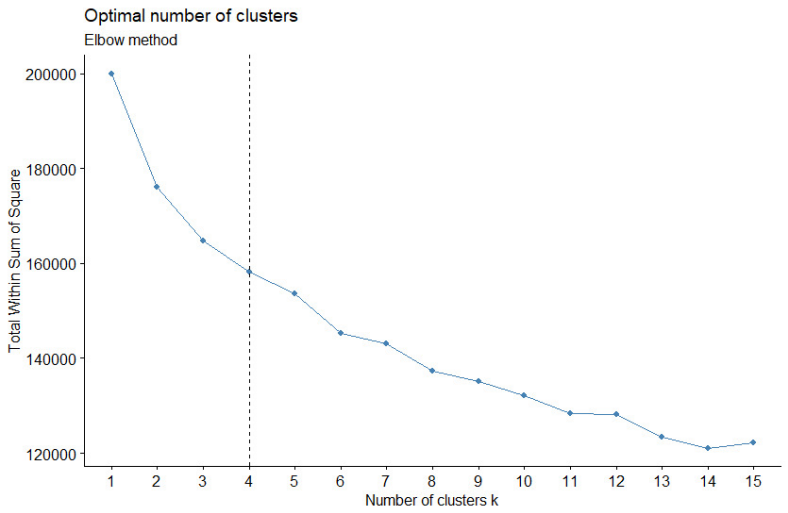
Objectives

- Pattern recognition
- Group the data into K clusters
- Validity methods $\rightarrow$ Optimal number of clusters, K
- Loss function:

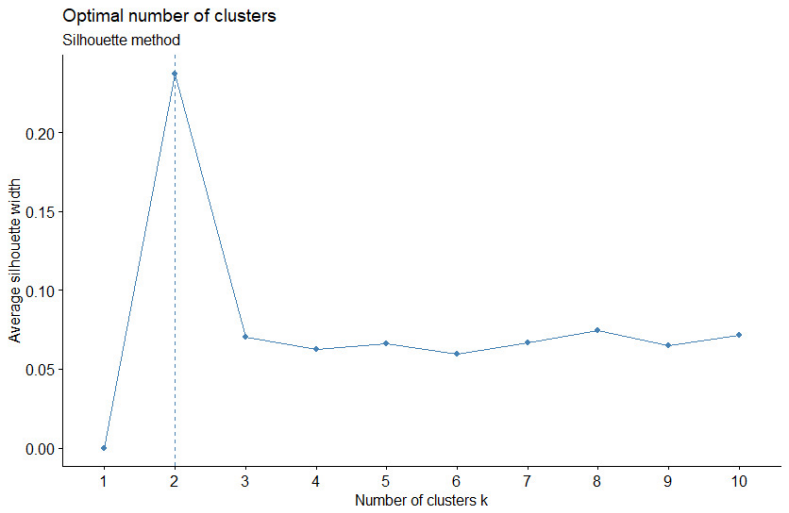
$$W(C) = \sum_{k=1}^{K} \sum_{C(i)=k} ||X_i - \overline{X}_k||^2 \text{ (Euclidean distance)} \quad (1)$$

- Minimize this function: Naive Algorithm
- K-means clustering: **kmeans**, and to visualise **splom**

K-means clustering - Validity methods



(a) Elbow method: Total loss function in terms of K



(b) Silhouette method: Average silhouette width in terms of K

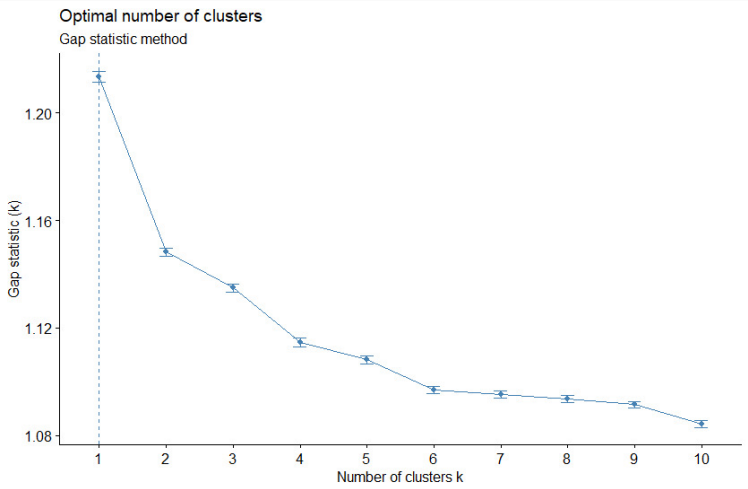


Figure: Gap statistic: Variation of total loss function with the expected values of the data in terms of K

- The initial problem consists of binary classification $\rightarrow$ 2 clusters;
- $K = 1$ (gap statistic) is redundant, as the objective in clustering is to divide the data in more than 1 group. The second best K is $K = 2$.

2 methods that support $K = 2$ and 1 that supports $K = 4$. By majority and adding the first factor, we opted to use $K = 2$.

K-means clustering - Splom

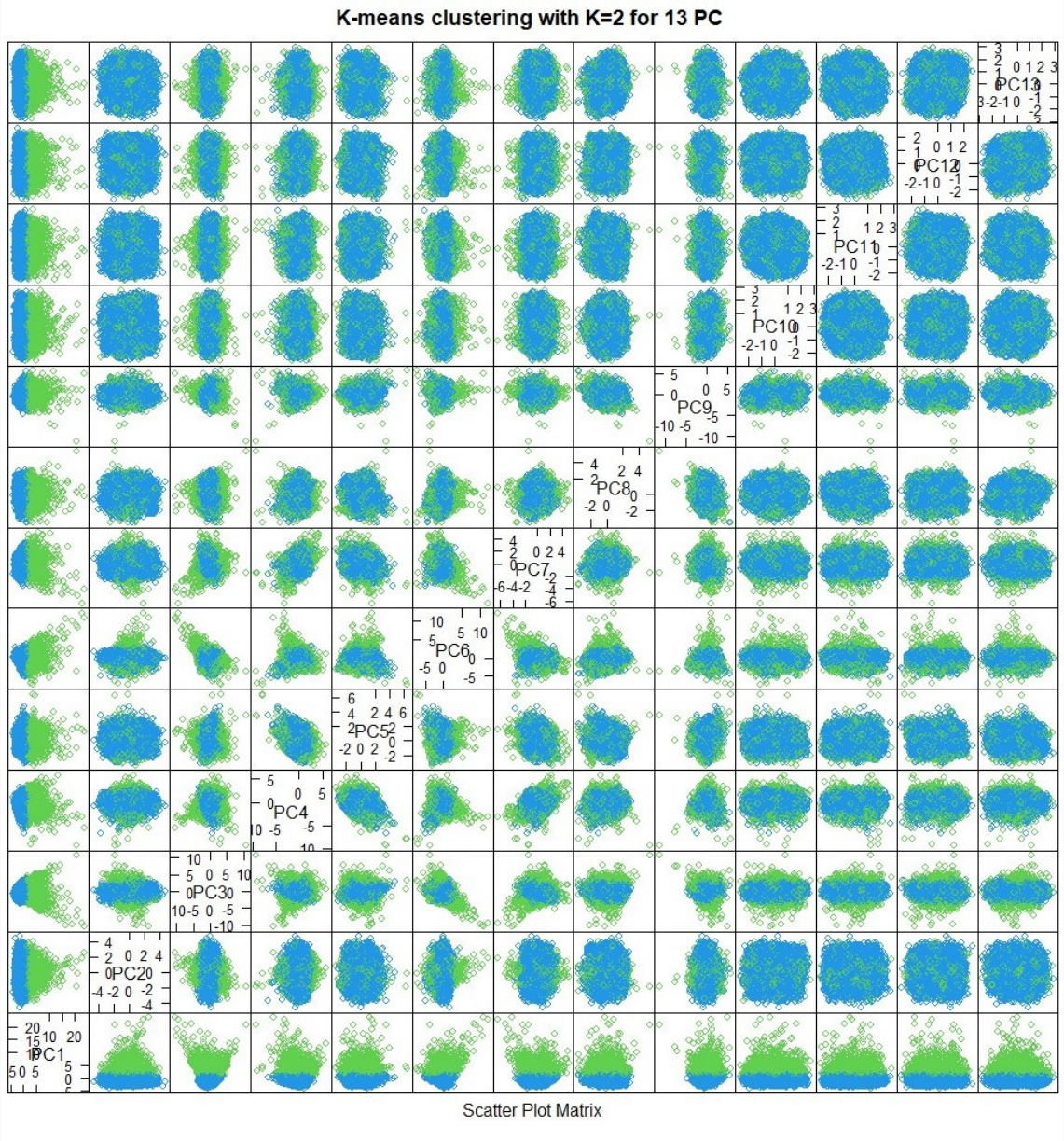


Figure: Graphical representation of the K-means clustering for every combination of principal components.

K-means clustering - PC1 vs PC2

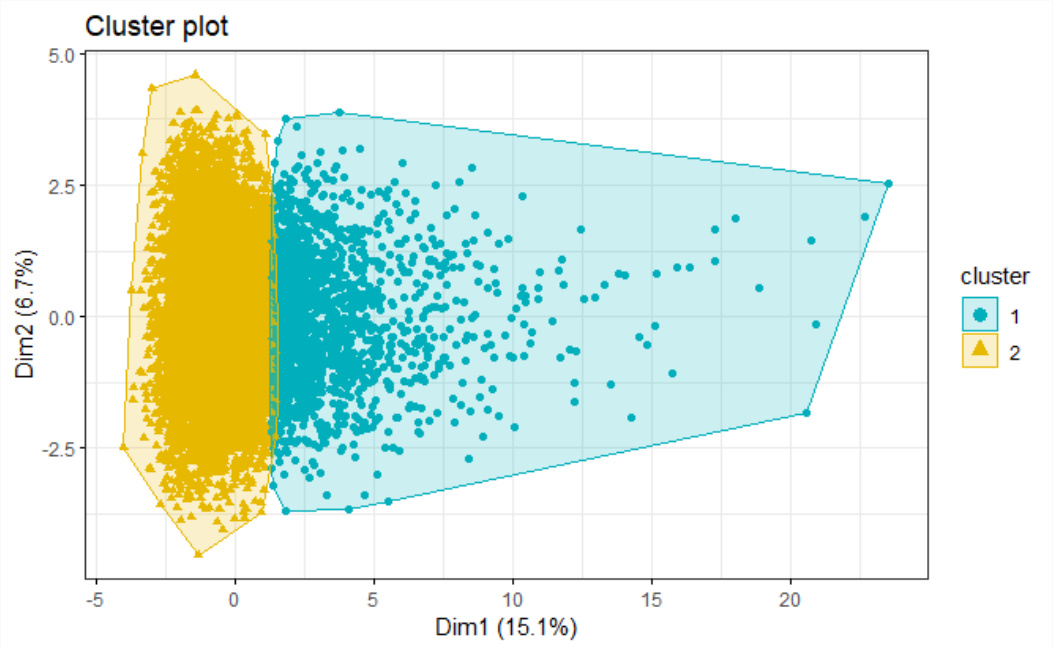


Figure: Cluster plot for the first and second principal components.

- Within-cluster sum of squares: 56 265.79 (cluster 1) and 119 201.10 (cluster 2)
- Between-cluster sum of squares / Total sum of squares: 11.9%

	Target	
Cluster	0	1
1	1034	898
2	3653	4415

Table: Label of the original points and the cluster they were assigned to.

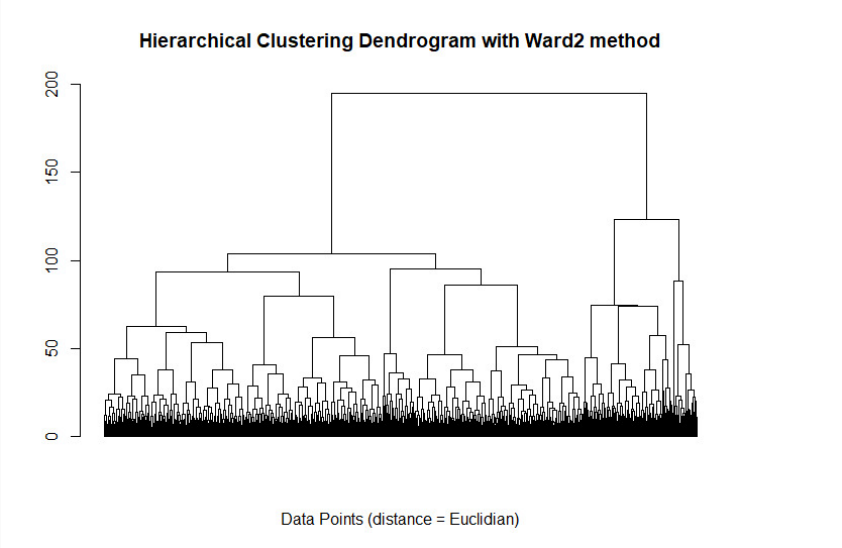
Hierarchical Clustering

Objective

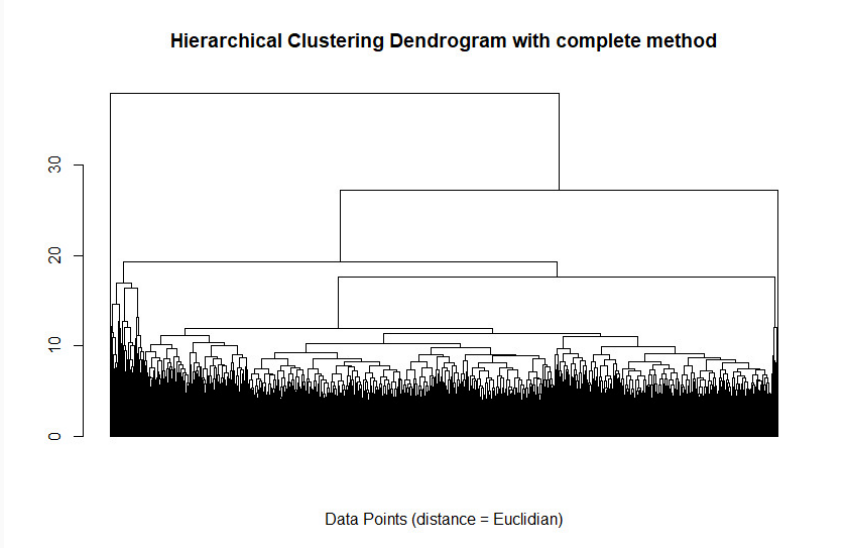
Group the data into clusters

- Doesn’t require a pre-selected number of clusters
- Requires distance measures
- Agglomerate method: Iterative increasing the levels
- Methods: complete, average and Ward
- Distances: Euclidean, Manhattan and Chebychev
- Hierarchical clustering: **hclust**

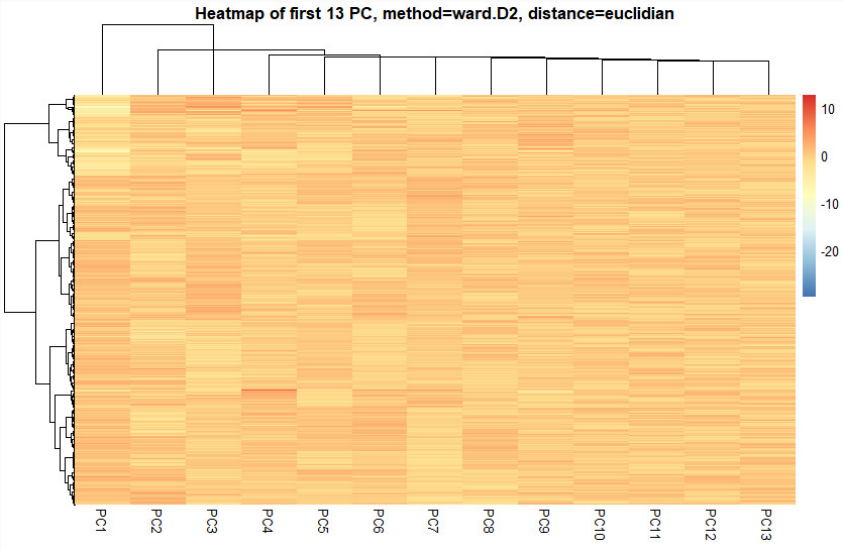
Hierarchical Clustering - Some results



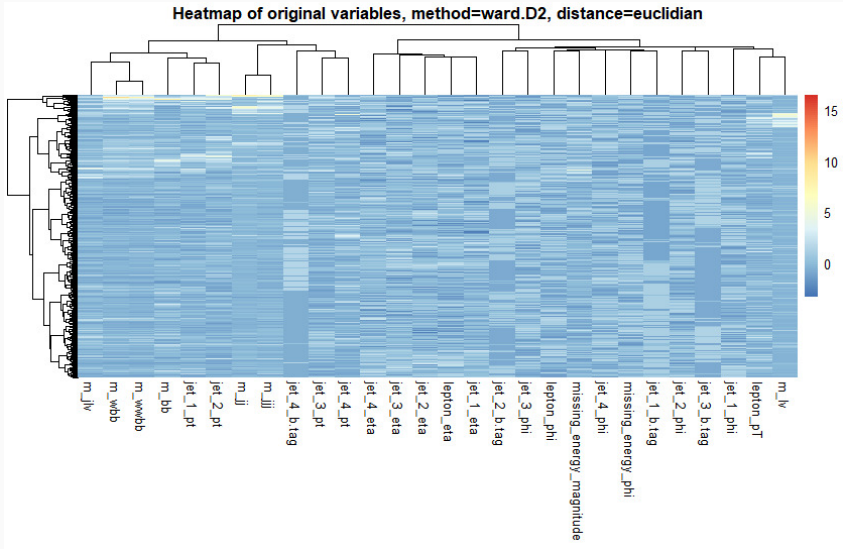
(a) Dendrogram with Ward2 method and Euclidean distance.



(b) Dendrogram with complete method and Euclidean distance.



(c) Heatmap with method Ward D2 and Euclidean distance.



(d) Same heatmap for original variables.

Normal Mixture Models for Clustering

Objective

Model the data into a multi-modal gaussian function

- Cluster: Uni-modal gaussian function, $p_k(X) = N_p(\mu_k, \Sigma_k)$
- More precise and adaptable method
- Overlaying clusters
- Multi-modal function: $p(X) = \sum_{k=1}^{K} \pi_k p_k(X)$
- Method: Maximum likelihood estimation
- System of equations:

$$\begin{cases} \pi_i = \frac{1}{N} \sum_{j=1}^{N} q_i(X_j) \\ \mu_i = \frac{1}{N\pi_i} \sum_{j=1}^{N} q_i(X_j) X_j \\ \Sigma_i = \frac{1}{N\pi_i} \sum_{j=1}^{N} q_i(X_j)(X_j - \mu_i)(X_j - \mu_i)^T \end{cases}$$

- Method to solve: Expectation-Maximization Algorithm
- Normal mixture clustering with EM algorithm: **mclust**
- Various models to fit: Bayesian Information Criterion (BIC)

Normal Mixture Models for Clustering

- The best fit model: "VVV", ellipsoidal, varying volume, shape and orientation.

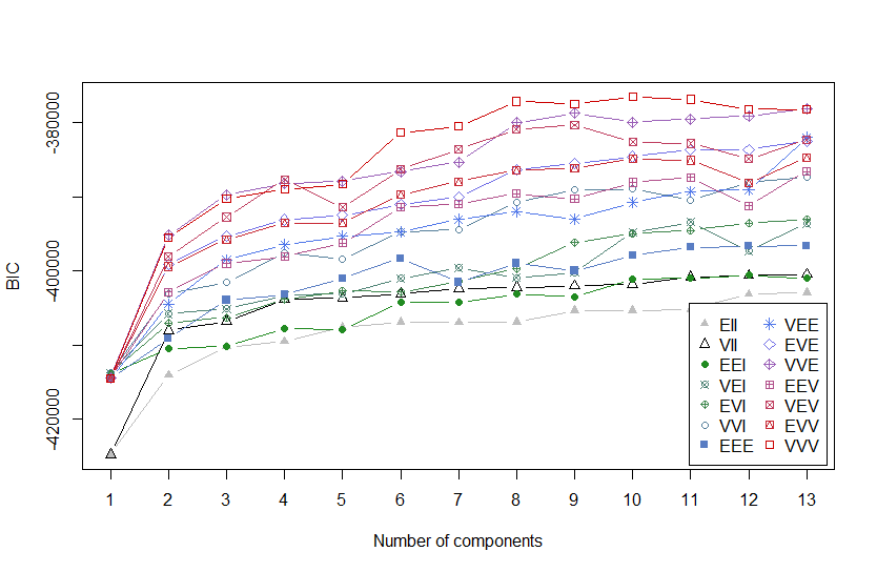


Figure: Bayesian Information Criterion for various models as a function of the PCs.

log-likelihood		n		df		BIC		ICL	
-183450.1		10000		1049		-376561.9		-378118.6	
Clustering table:									
1	2	3	4	5	6	7	8	9	10
1003	1543	1578	631	1665	1167	638	239	651	885
Mixing probabilities:									
1	2	3	4	5	6	7	8	9	10
0.1034	0.1527	0.1554	0.0626	0.1622	0.1135	0.0736	0.0256	0.0636	0.0869

Table: Some statistics for Gaussian finite mixture model fitted by EM algorithm, as well as the mixing probabilities and the number of points for each cluster.

Normal Mixture Models for Clustering

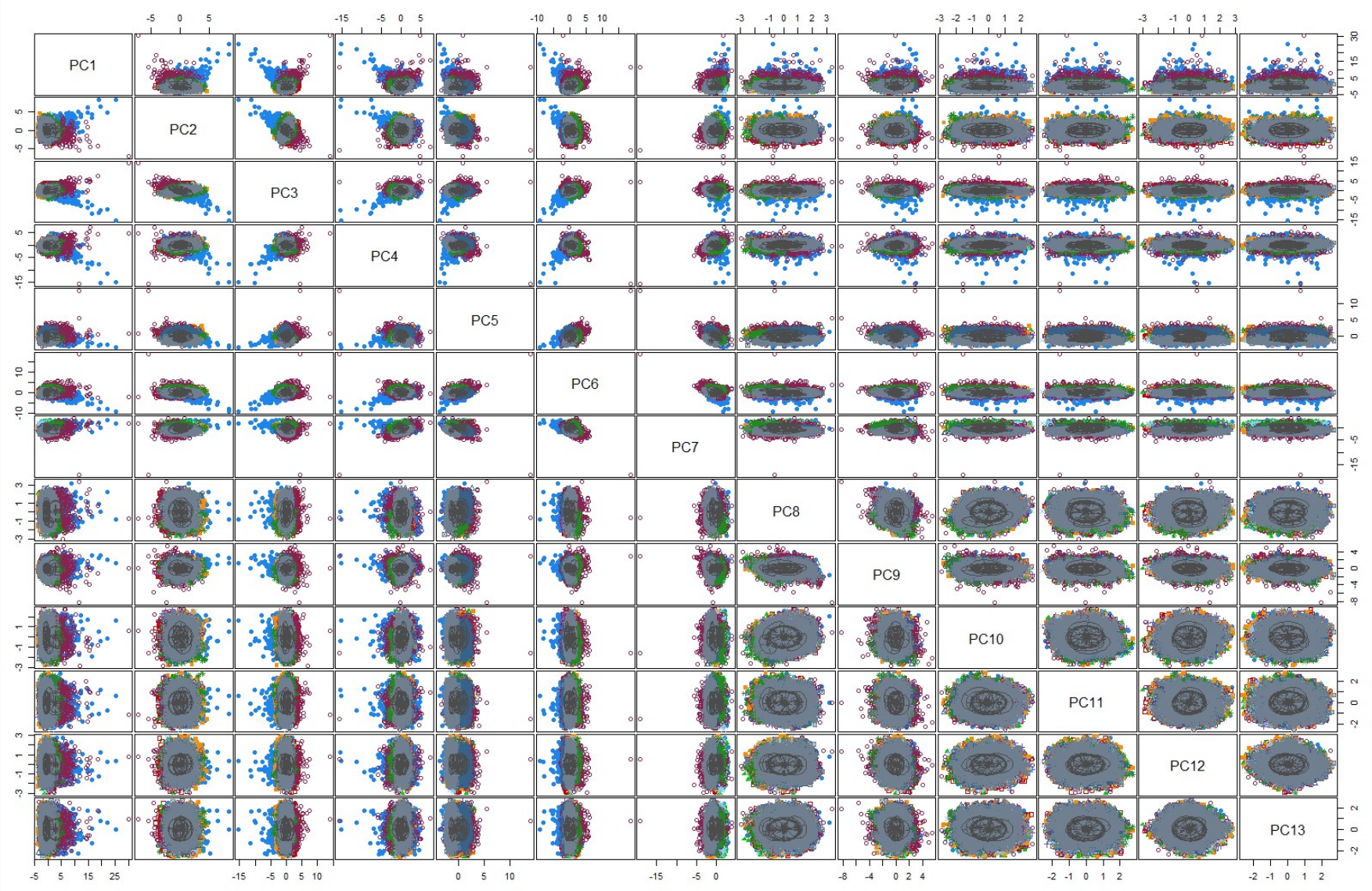
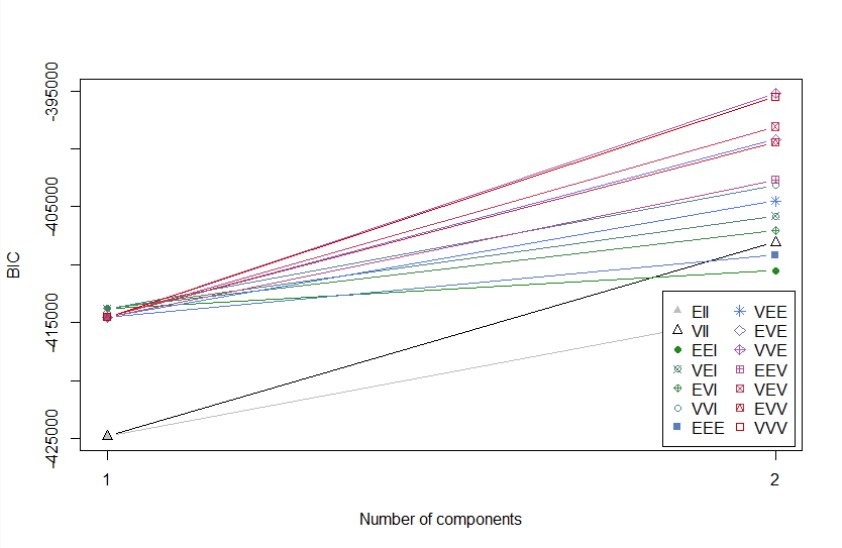
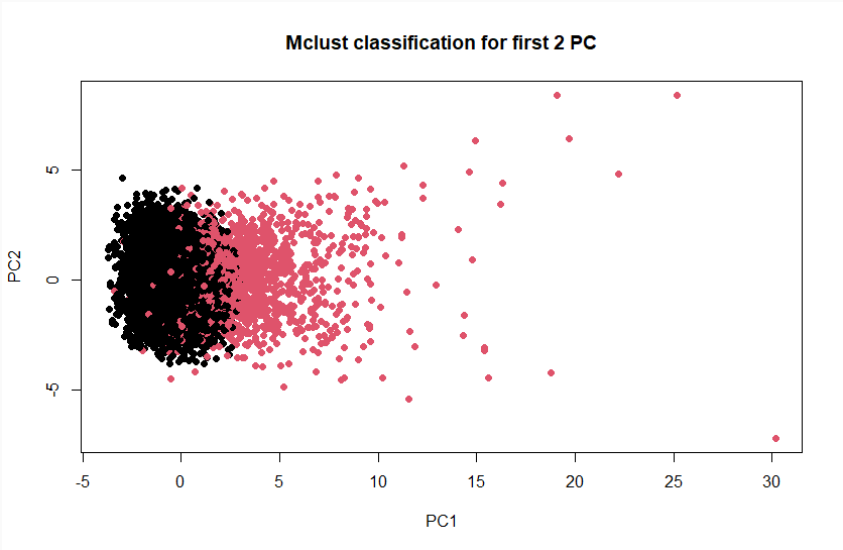


Figure: Graphical representation of the normal mixture clustering using the "VVV" model between every combination of PCs.

Normal Mixture Models for Clustering



(a) Bayesian Information Estimator (BIC) as a function of the PCs for various models for the first two principal components.



(b) Mclust classification for the first two principal components.

log-likelihood	n	df	BIC	ICL
-196996.3	10000	131	-395199.2	-396060
Clustering table:				
1	2			
8267	1733			
Mixing probabilities:				
1	2			
0.8154968	0.1845032			

Table: Some statistics for Gaussian finite mixture model fitted by EM algorithm, as well as the mixing probabilities and the number of points for each cluster.

Normal Mixture Models for Clustering

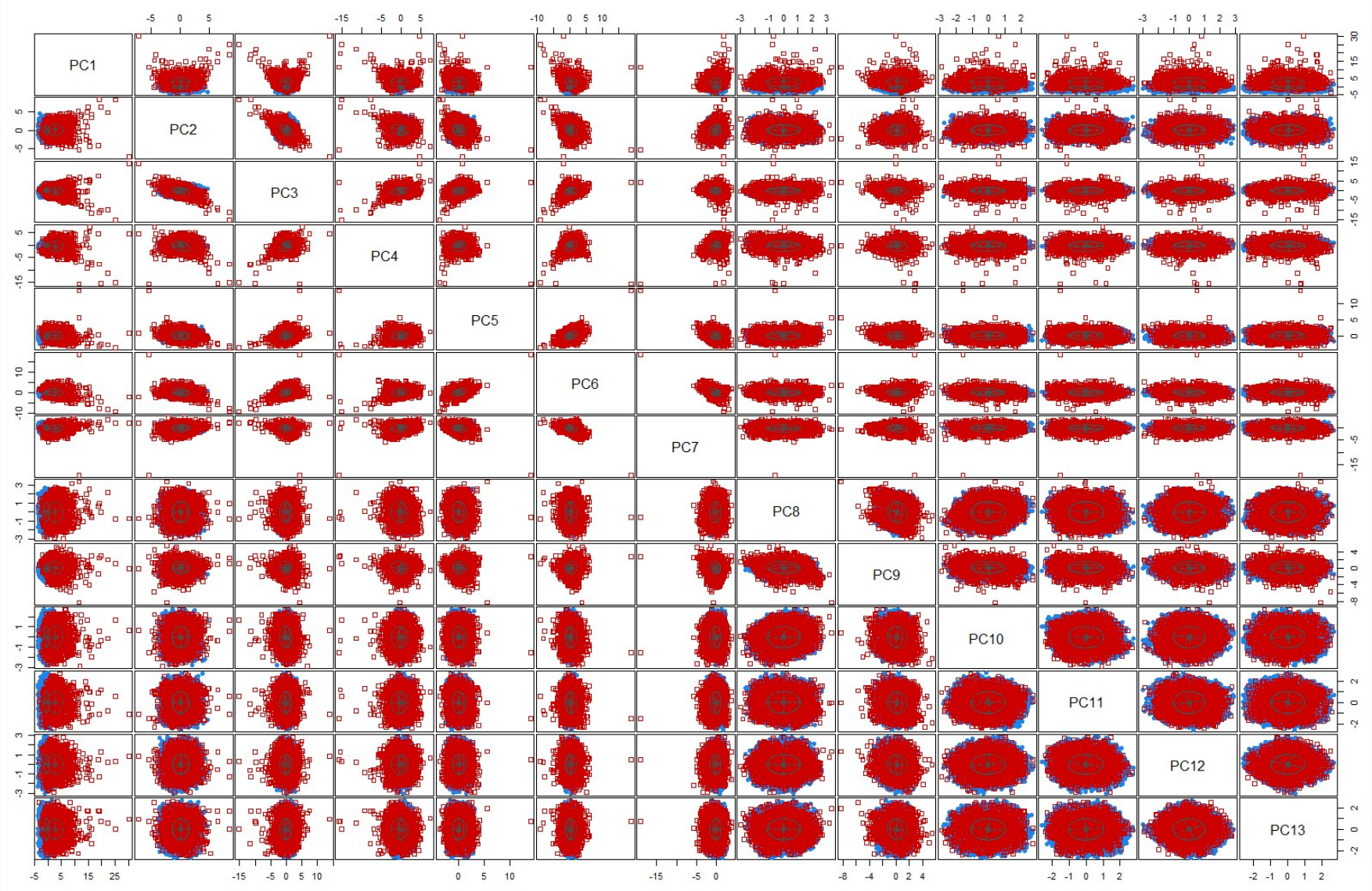


Figure: Graphical representation of the normal mixture clustering using the "VVE" model between two first PCs.

Thank you!