Exercises 5-4

Giovanni Zurlo

15/11/2021

Exercise 1

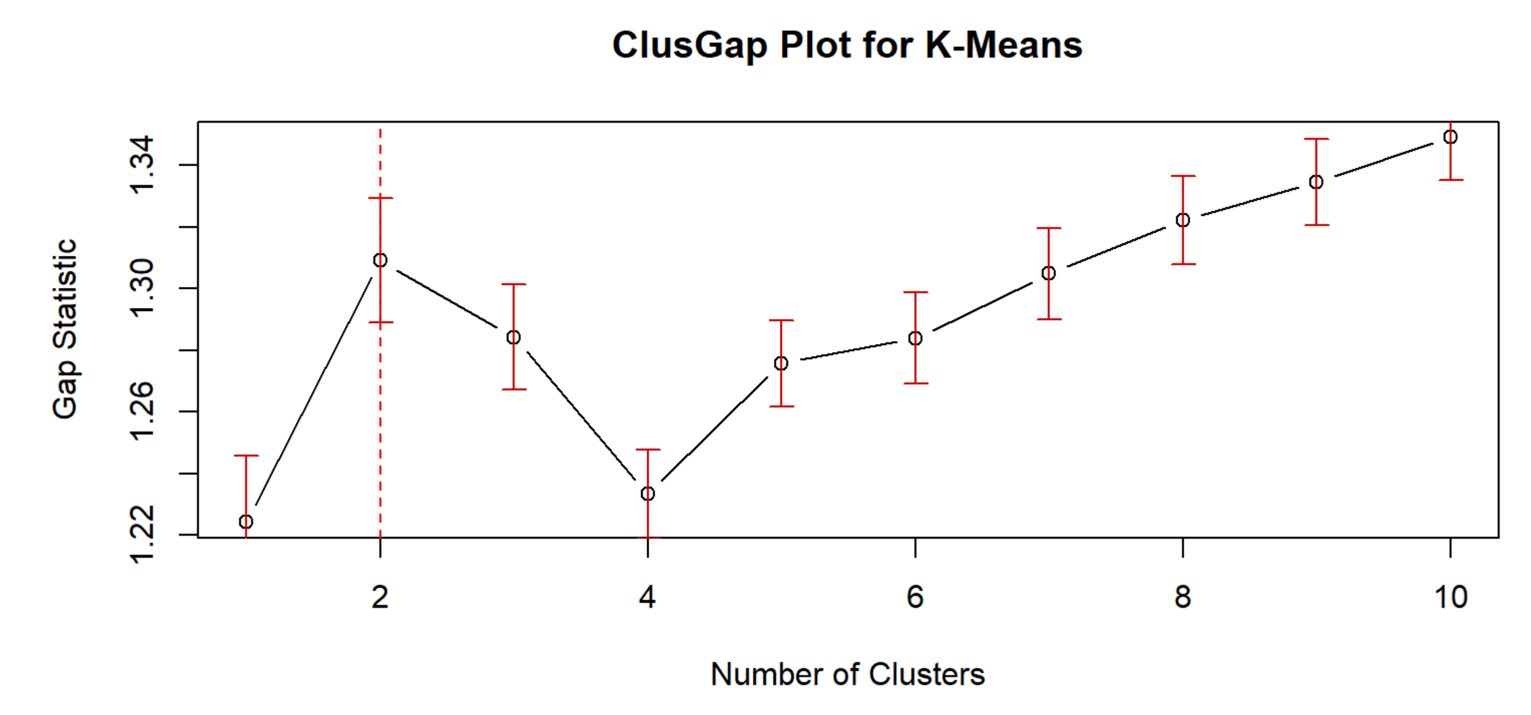
> *On Virtuale you'll find the data set wdbc.data. This data set is taken from the UCI Machine Learning Repository. Data are given about 569 breast cancer patients, and there are the two "true" classes of benign (357 cases) and malignant (212 cases) tumors. There are ten quantitative features in the dataset. Compute different clusterings of the data (use at least two different approaches including a Gaussian mixture model and try out numbers of clusters up to 10) and compare them first without using the information about benign vs. malign cancers in the diagnosis variable wdbcdiag. Which clustering do you think is best? Only after you have made a decision about your favourite clustering, use the ARI to compare all these clusterings to wdbcdiag. Discuss how it was possible, without using wdbcdiag, to recognise from the data whether a clustering would be similar or less similar to the "true" clustering in wdbcdiag.*

```
# Loading Data
wdbc <- read.csv("wdbc.dat",header=FALSE)
# Selecting variables 3-12
wdbcc <- wdbc[,3:12]
# Scaling data
swdbcc = scale(wdbcc)
# Coding true labels as integers
wdbcdiag <- as.integer(as.factor(wdbc[,2]))

# Computing Euclidean DistMat
dist=dist(swdbcc)
# Performing PCA for visualization
pca<- prcomp(swdbcc)
summary(pca)
```

```
## Importance of components:
##                           PC1    PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     2.3406 1.5870 0.93841 0.7064 0.61036 0.35234 0.28299
## Proportion of Variance 0.5479 0.2519 0.08806 0.0499 0.03725 0.01241 0.00801
## Cumulative Proportion  0.5479 0.7997 0.88779 0.9377 0.97495 0.98736 0.99537
##                            PC8     PC9    PC10
## Standard deviation     0.18679 0.10552 0.01680
## Proportion of Variance 0.00349 0.00111 0.00003
## Cumulative Proportion  0.99886 0.99997 1.00000
```

K-Means

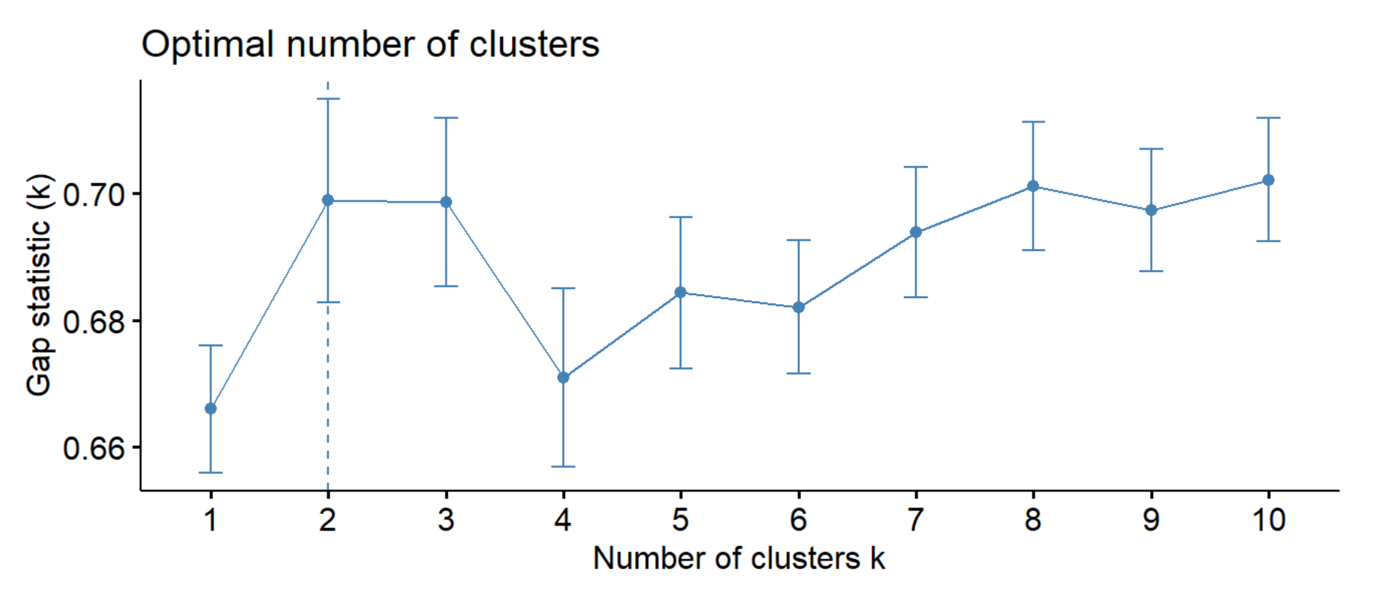


K-Means (2)

Cluster	Freq
1	169
2	400

K = 2 solution was preferred over a higher number of clusters also according to the ASW distribution (Max ASW for K-Means (2) = **0.395**)

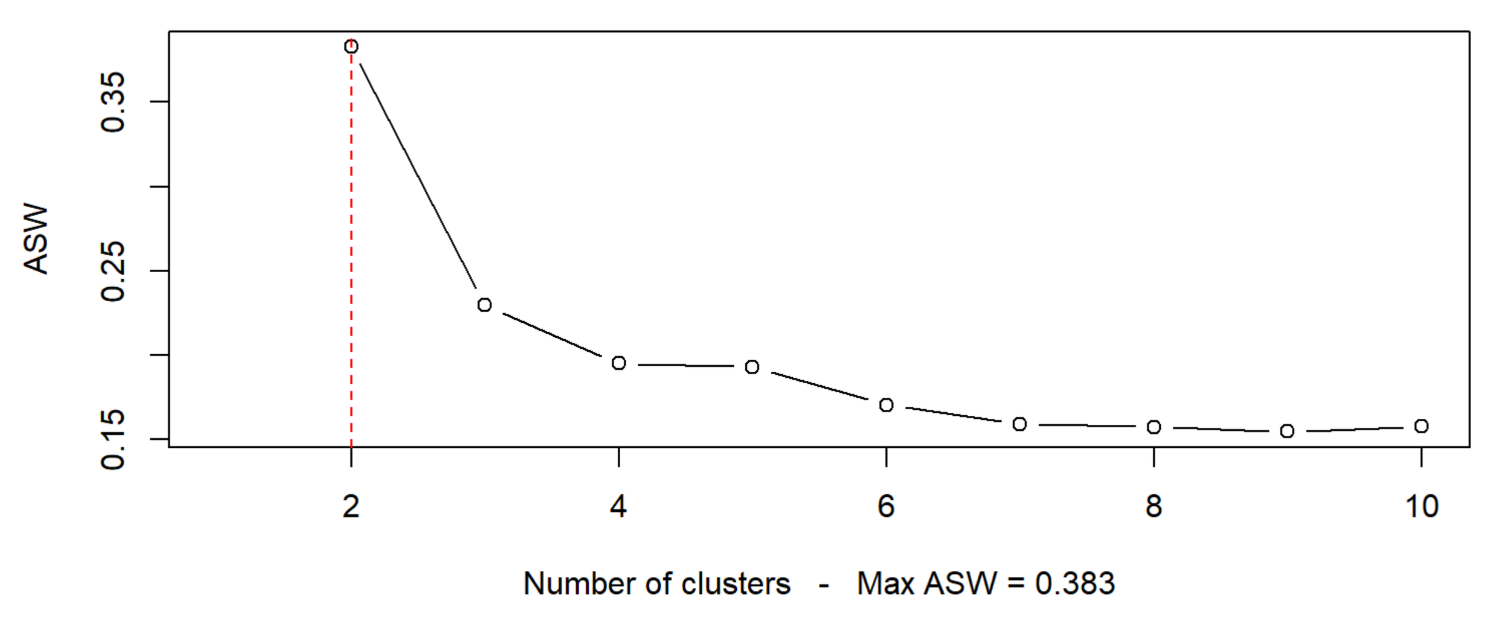
Ward Method



Ward Method (2)

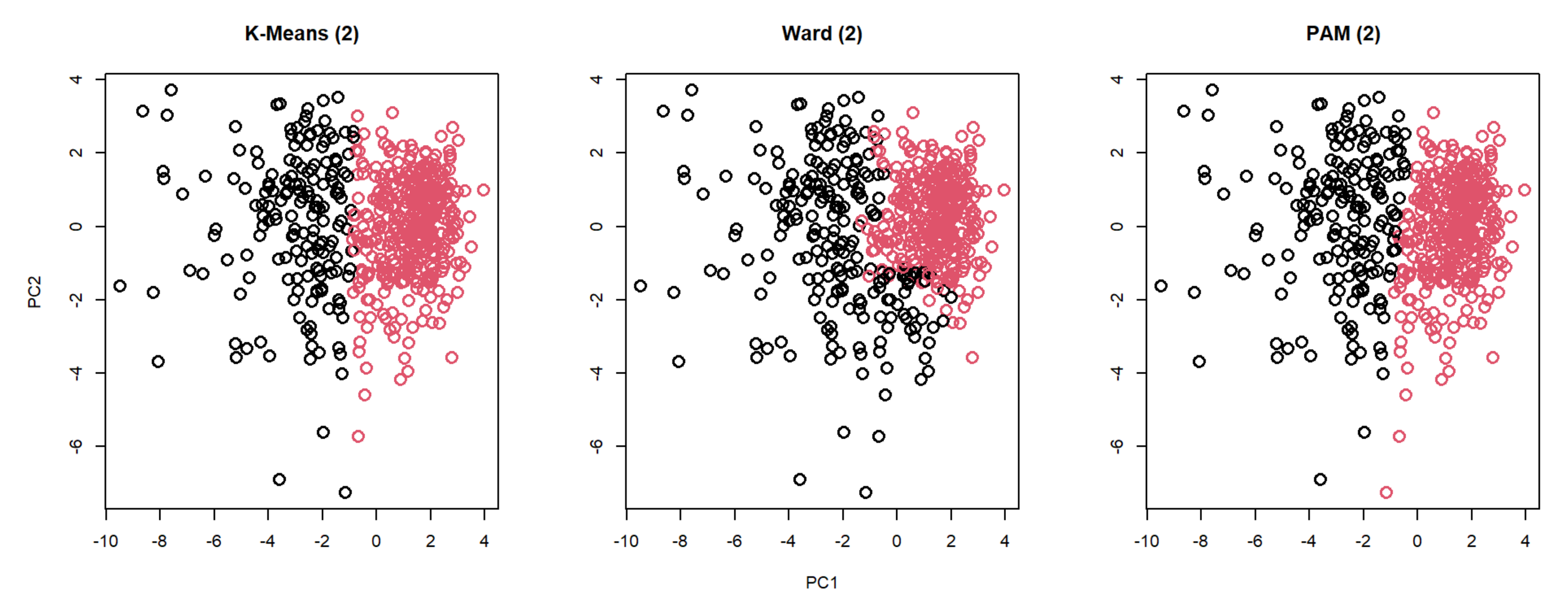
Cluster	Freq
1	217
2	352

PAM



PAM (2)

Cluster	Freq
1	184
2	385



Solutions have been reported on a 2D PCA plot summarizing about 80% of the original data variability. Cluster 1, in *coral*, can be recognized as a density cluster whose correct identification will lead to better correspondence with the true classes. Graphically, Ward clustering seems to lack a bit in isolating it while PAM and K-Means solutions are pretty similar and convincing.

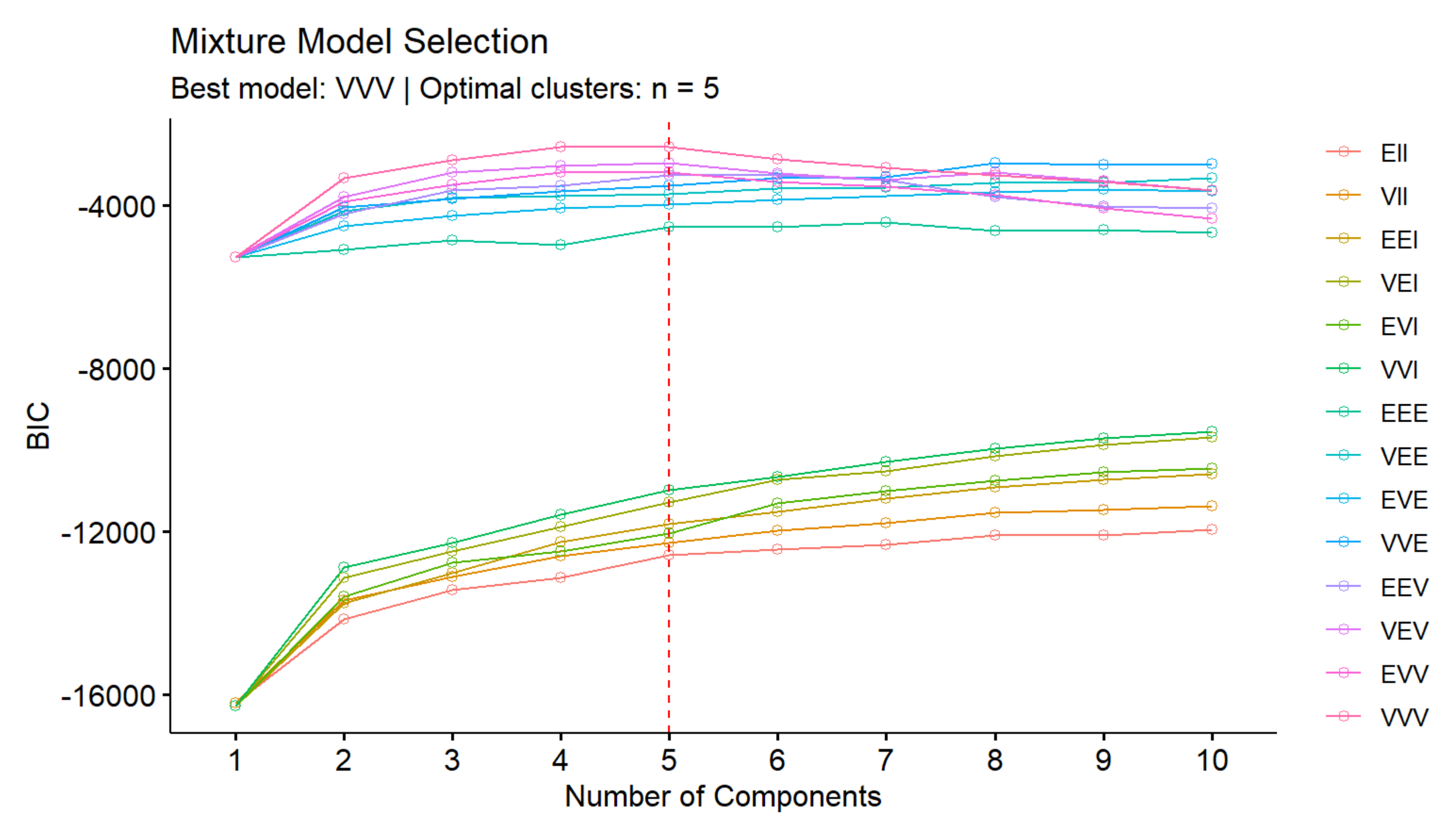
Mixture Models

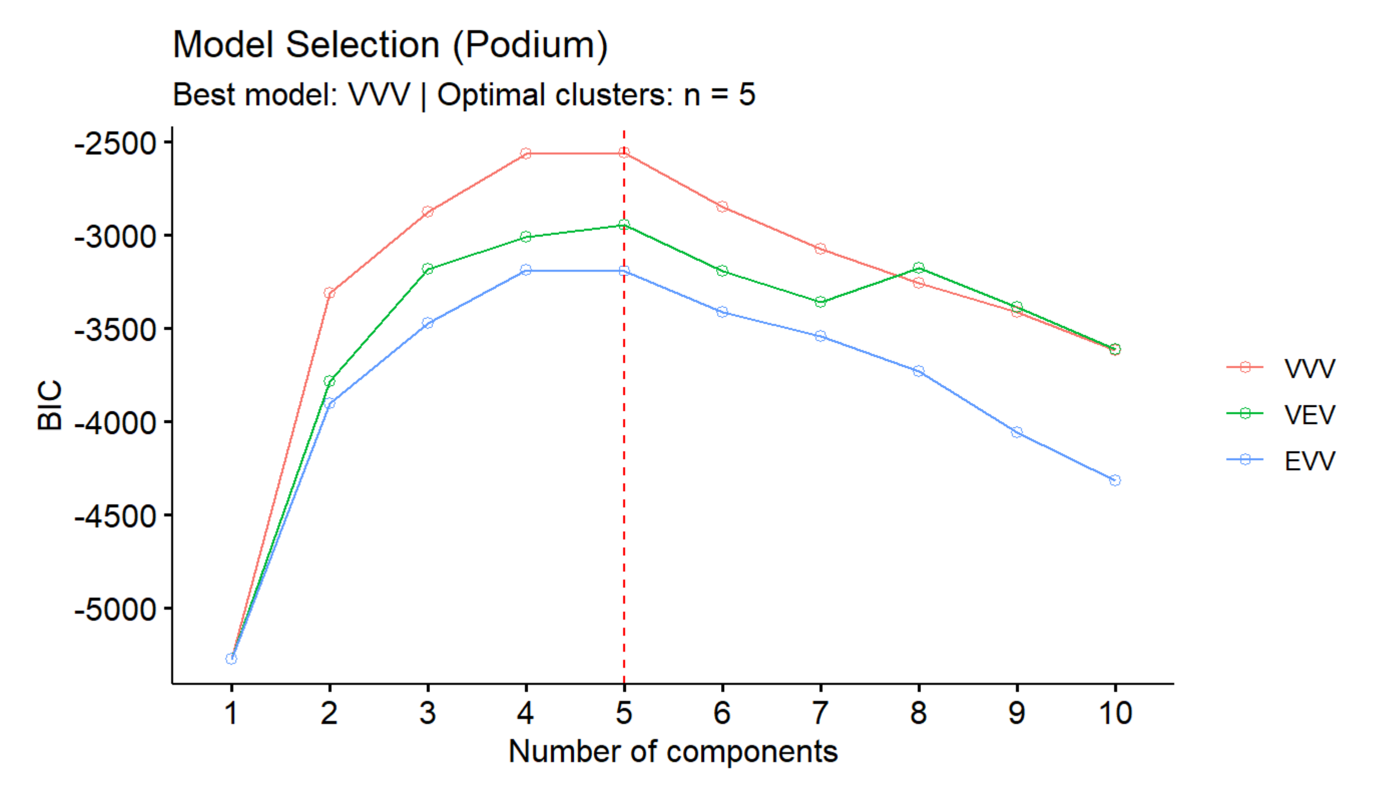
```
# Fitting different Gaussian Mixtures up to 10 components
mixtures=Mclust(swdbcc, G=1:10)
summary(mixtures)
```

```
## ----------------------------------------------------
## Gaussian finite mixture model fitted by EM algorithm
## ----------------------------------------------------
##
## Mclust VVV (ellipsoidal, varying volume, shape, and orientation) model with 5
## components:
##
##  log-likelihood   n  df       BIC       ICL
##       -235.8921 569 329 -2558.921 -2606.322
##
## Clustering table:
##   1   2   3   4   5
##  57 130 132 108 142
```

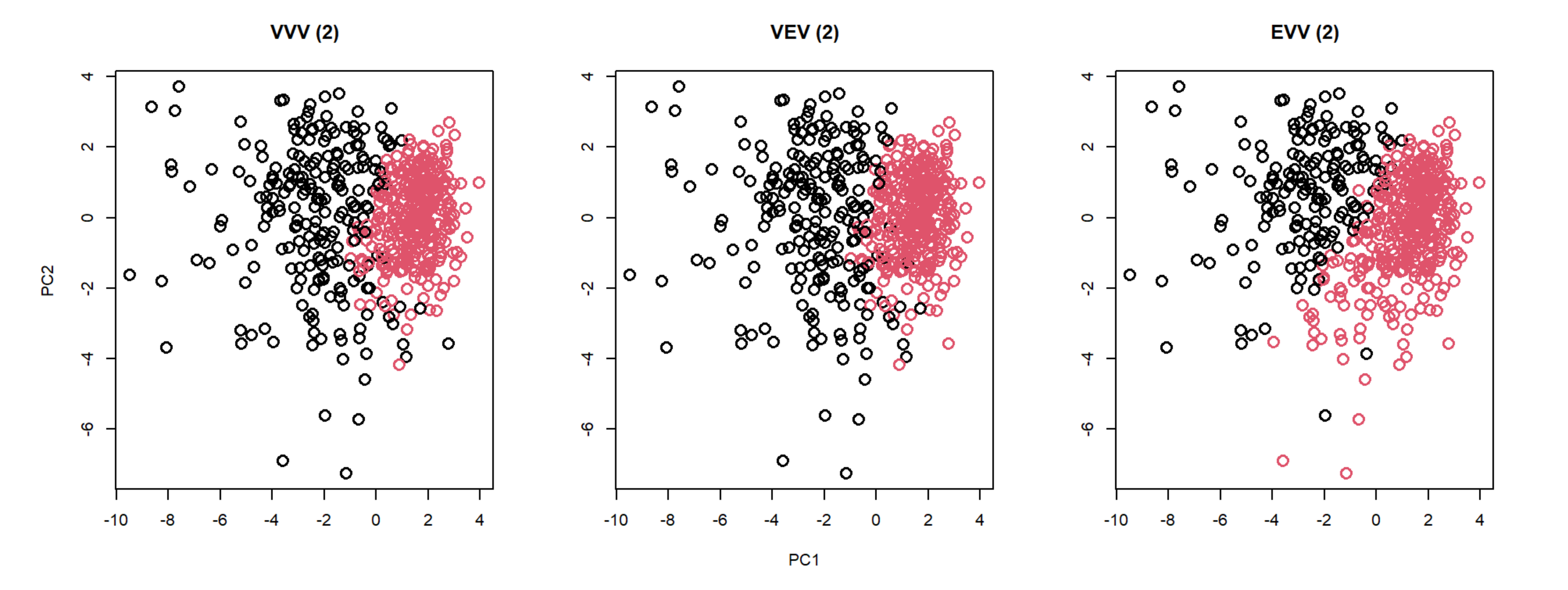
```
# Storing the selected best models
VVV=Mclust(swdbcc, G=2, 'VVV')
VEV=Mclust(swdbcc, G=2, 'VEV')
EVV=Mclust(swdbcc, G=2, 'EVV')

# Plotting BIC for each fitted model
library(factoextra)
fviz_mclust_bic(mixtures, model.names = NULL, shape = 1, lwd=3,
                color = "model", palette = NULL, legend = NULL,cex=2,
                main = "Mixture Model Selection", xlab = "Number of Components",
                ylab = "BIC")
```





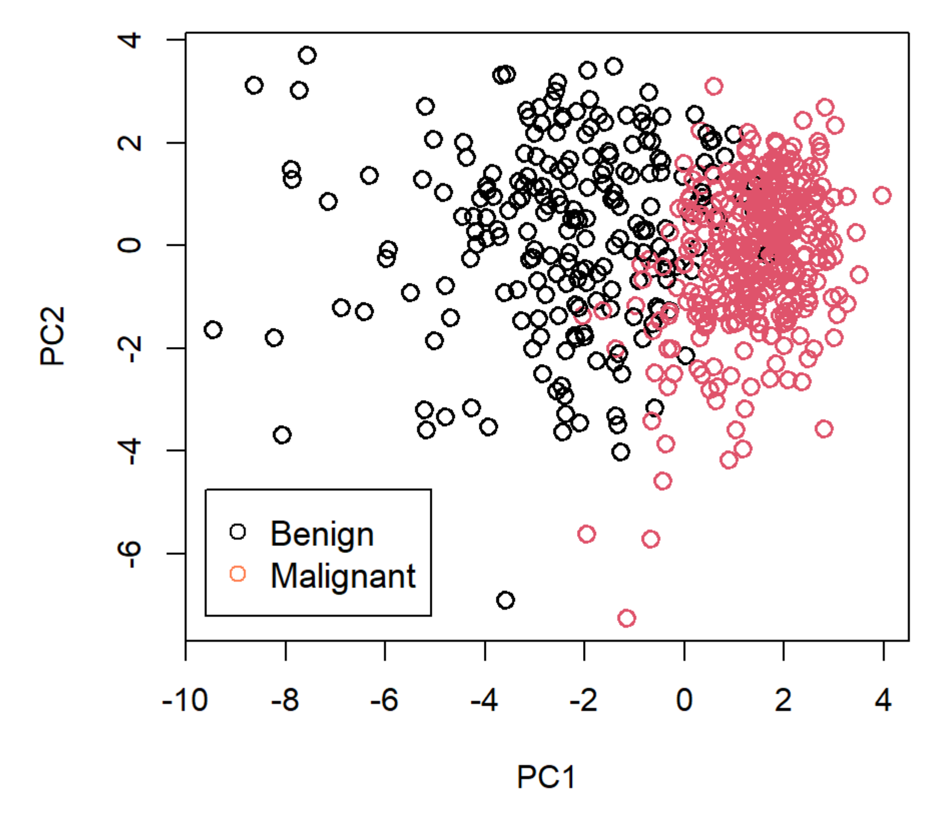
Two clusters solutions were preferred over *K* = 5 to which higher BIC values correspond. ASW index also suggests this choice but because of a well known bias it has. EVV covariance matrix mixture includes in the coral cluster some more observations from the lower part of the plot but it is also worse in isolating the density cluster we have (it assumes equal volume). VVV and VEV models are hard to distinguish graphically (remember we're exploiting a PC approximation of the original sample) and also analytically they're very similar (*ARI = 0.9*). I think that these two are the best together with PAM and K-Means, but I cannot decide which is the most appropriate volume/shape for the coral cluster.




```
adjustedRandIndex(VVV$classification,VEV$classification)
```

```
## [1] 0.89692
```

Comparison with True Classes



Clustering Similarity to Diagnosis Variable

	K-Means (2)	Ward (2)	PAM (2)	VVV (2)	VEV (2)	EVV (2)
ARI	0.6464599	0.4694395	0.7239038	0.6265249	0.5987271	0.6182784

Average Silhouette Width

	K-Means (2)	Ward (2)	PAM (2)	VVV (2)	VEV (2)	EVV (2)
ASW	0.395	0.333	0.383	0.336	0.346	0.347

Finally, the classifications were compared to the actual labels by means of the ARI. PAM clustering matches best the diagnoses we observed by a consistent margin. It is followed by K-Means clustering (graphically *wider shaped*) and the mixtures are behind them. A fully parametric, gaussian approach seems not to be the optimal one, while maximum flexibility is at least required to match the lables. I mainly followed a visual validation focusing on the idea of a higher density cluster (on the right side of the PCA plot) in a heap of noise (benign cases); in reality the malignant cases cluster has rather a skewed normal shape. The choice of *K = 2* for the mixture models has been confirmed from the analysis of the Adjusted Rand Index which decreases a lot as we add more and more components. ASW ranking seems to suggest which partitions perform better in terms of ARI, but this cannot be generalized and its usage may also be not recommended in this setting.

I would say that these unsupervised techniques are not completely appropriate to perform classification, since these do not take into account true lables. Like in this situation, decent results can be obtained if clear structures are present in the feature space, but this is not always the case. Supervised learning methods like SVMs are way more favored in this kind of tasks. As I observed in the principal components plot for the PAM clustering, a separating hyperplane could have done a great job with less effort.

Exercise 3

> *On Moodle you can find the article "Regularized k-means clustering of high-dimensional data and its asymptotic consistency" by Wei Sun and Junhui Wang, EJS 6 (2012).*

Explain and motivate in your own words how regularized k-means differs from k-means, and how regularized model-based clustering differs from model-based clustering.

In the regularized k-means algorithm, an adaptive group lasso penalty term is introduced in the classical objective function. Basically, we sum to the usual within-cluster distances a regularization term depending on p different lambda parameters (one for each dimension) which allows for a flexible variable selection. In this way, as for other regularized regression techniques, we balance model fitting and sparsity straight in the objective function optimization and we obtain some "shrinked" new centers. Fitting procedure does not change since we still adopt an iterative scheme to approximate the global minimum of the objective (sensitive to initialization). This idea can also be extended to model-based techniques such as mixture models by regularizing the log-likelihood function in the same fashion, and then employing the EM algorithm.

The authors state that the X-variables should be "centralized". Why is this important? Do you think it would also be useful to standardize them to unit variance? Why, or why not, or under what circumstances?

I guess that lambda parameters and clustering results may depend on the different scale of both variables and centers. Since we have the L2 norm of each center in the penalty term, I expect larger dimensions to "dominate" over smaller ones, as the euclidean norm is not scale-invariant. Whenever variables do not share a common unit of measure and/or scale, I think it could be useful to standardize them first in order to take into account each variable's true relative importance. As an example, common regularized regression techinques are not scale invariant and standardization is highly recommended.

As opposed to the use of the Lasso in regression, the tuning constant λ here cannot be chosen by optimizing a prediction error estimated by cross-validation. Why not? What do the authors propose instead to choose the λ? Do you think this could also be done involving the Rand or adjusted Rand index?

We cannot use cross-validation as we do not have "labels" at disposal and classification isn't our goal. The authors suggest employing bootstrap resampling to investigate cluster stability over different parameters choice; the key idea is that a good clustering algorithm (with optimal parameters'values) should produce clusterings that are the more similar from one sample to another. In the paper, instability of the regularized k-means clustering is measured by a generic expected distance between any two clustering assignments; the authors define such distance as a measure of the probability of their disagreement, so that ARI could not be assimilated to such measure (it takes negative value). Rand Index, however, does range between zero and one and since simulations are performed holding K constant, RI could also be involved in such tuning process.

What are advantages and good properties of the method according to the authors? Do you think this is convincing? Can you think of any potential disadvantages of the method?

According to the authors, this new clustering method overcomes some of k-means issues with high-dimensional data (where many dimensions may contain no information about the clustering structure, i.e. may be redundant) such as deteriorated performances and clusters interpretability. Regularized k-means simultaneously performs cluster analysis and variables selection and has the advantage of cutting down the "curse of dimensionality" suffered by the classical algorithm; as an example, with diverging data dimension, this last shows the tendency of grouping together neighboring clusters. Assuming a pre-specified true number of clusters K, two important properties are proved: estimation consistency and selection consistency. Given a properly selected λ, the regularized algorithm assures the a.s. convergence of the estimated cluster centers to the true ones; by the property of selection consistency, the algorithm can eliminate the uninformative variables in the centers with probability tending to one. However, further investigation of these traits is required in the case of unknown K; in general, one disadvantage of this procedure is the difficulty of tuning K and λ simultaneously. This means that, in several settings, regularized k-means clustering can lead to results that are biased and sub-optimal with respect to those offered by the unpenalized algorithm, and so it should be employed with great care.

EXERCISE 2

We would like to maximize the expected complete log lik

$$E_\eta = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\log \pi_k + \log \varphi_{a_k,\sigma_k^2}(x_i)\right)$$

with respect to the mixing parameters $\pi_1, \dots \pi_K$ and the densities parameters $a_1, \dots, a_K$ (MEANS) and $\sigma_1^2, \dots, \sigma_K^2$ (VARIANCES) from K one-dimensional gaussian distributions.
the optimization takes place under the constraint $\sum_{k=1}^{K} \pi_k = 1$

Q)

$$\operatorname{argmax}\ Q(\pi_1, .. \pi_K, \lambda) =$$

$$\operatorname*{argmax}_{\pi_1, \dots \pi_K, \lambda} \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\log \pi_k + \log \varphi_{a_k,\sigma_k^2}(x_i)\right) - \lambda\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

WLOG, I show the maximization wrt an arbitrary param π_e but the steps and result are the same for any π_k $k = 1, .. K$

$$\frac{\partial}{\partial \pi_e} Q(\pi_1, .. \pi_K, \lambda) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\frac{\partial}{\partial \pi_e}\log \pi_k\right) - \lambda \frac{\partial}{\partial \pi_e}\left(\sum_{k=1}^{K} \pi_k - 1\right)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\frac{1}{\pi_e}\mathbb{1}(k=e)\right) - \lambda\,\mathbb{1}(k=e)$$

$$= \frac{1}{\pi_e}\sum_{i=1}^{n} p_{ie} - \lambda = 0 \implies \pi_e^* = \frac{1}{\lambda}\sum_{i=1}^{n} p_{ie}$$

BUT SINCE $\frac{\partial}{\partial \lambda} Q(\pi_1, .. \pi_K, \lambda) = 1 - \sum_{k=1}^{K} \pi_k = 0$

$$\iff \frac{1}{\lambda}\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik} = 1 \qquad \text{REPLACING FOR } \pi_k^*$$

$$\iff \lambda = n \qquad \text{SINCE } p_{ik}^{(t-1)} = P(Z_i = k \mid \eta^{t-1}, x_i)$$
$$\text{AND } \sum_{k=1}^{K} p_{ik} = 1$$

$$\implies \pi_k^* = \frac{1}{n}\sum_{i=1}^{n} p_{ik} \qquad k = 1, \dots e \dots, K$$

b)

$$\operatorname*{argmax}\ Q(\alpha_1,\dots,\alpha_K;\sigma_1^2,\dots,\sigma_K^2) =$$

$$\operatorname*{argmax}_{\alpha_c,\,\sigma_c^2} \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\log \pi_k + \log \varphi_{\alpha_k,\sigma_k^2}(x_i)\right) - \lambda\left(\sum_{k=1}^{K}\pi_k - 1\right)$$

$$\frac{\partial}{\partial \alpha_c} Q(\alpha_1,\dots,\alpha_K;\sigma_1^2,\dots,\sigma_K^2) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\frac{\partial}{\partial \alpha_c}\left(-\frac{1}{2}\frac{(x_i-\alpha_k)^2}{\sigma_k^2}\right)\right) =$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(-\frac{1}{2}\right)\left(-2\frac{(x_i-\alpha_k)}{\sigma_k^2}\right)\mathbb{1}(k=c) = \sum_{i=1}^{n} p_{ic}\frac{(x_i-\alpha_c)}{\sigma_c^2} = 0$$

$$\iff \sum_{i=1}^{n} p_{ic}\,x_i = \sum_{i=1}^{n} p_{ic}\,\alpha_c \implies \alpha_c^* = \frac{\sum_{i=1}^{n} p_{ic}\,x_i}{\sum_{i=1}^{n} p_{ic}}$$

HOLDS FOR ANY $k = 1,\dots c\dots,K$

$$\frac{\partial}{\partial \sigma_c} Q(\alpha,\sigma) = \sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(\frac{\partial}{\partial \sigma_c}\left(-\frac{1}{2}\ln\sigma_k^2 - \frac{1}{2}\frac{(x_i-\alpha_k)^2}{\sigma_k^2}\right)\right) =$$

$$\sum_{i=1}^{n}\sum_{k=1}^{K} p_{ik}\left(-\frac{1}{\sigma_k} + \frac{(x_i-\alpha_k)^2}{\sigma_k^3}\right)\mathbb{1}(k=c) = \sum_{i=1}^{n} p_{ic}\left(-\frac{1}{\sigma_c} + \frac{(x_i-\alpha_c)^2}{\sigma_c^3}\right) = 0$$

$$\iff \frac{1}{\sigma_c^3}\sum_{i=1}^{n} p_{ic}(x_i-\alpha_c)^2 = \frac{1}{\sigma_c}\sum_{i=1}^{n} p_{ic} \implies {}^{*}\sigma_c^2 = \frac{\sum_{i=1}^{n} p_{ic}(x_i-\alpha_c)^2}{\sum_{i=1}^{n} p_{ic}}$$

HOLDS FOR ANY $k = 1,\dots c\dots,K$