Colon Cancer Subtypes from Gene Expression Data

OxWaSP Module 3: Applied Statistics

Emmanuelle Dankwa Natalia Garcia Martin William Thomas

16 November 2018

Overview

Background and Data

The Data

- Gene expression data for 54,675 genes from 90 patients
- Patients have stage II colon cancer and underwent curative surgery
- Samples taken in Amsterdam, The Netherlands

Aims

- Use clustering to identify colon cancer subtypes
- Assess the robustness of the clustering procedures
- Build a classification rule for new patients

Data Processing

We want to:

- Reduce the dimensionality of the data.
- Identify the most informative genes.

Processing Steps

- Normalise and summarise the raw data using fRMA.
- Adjust for any batch effects.
- Apply the barcode algorithm to determine which genes are expressed.
- Select the most informative genes using median absolute deviance.

k-means Clustering

Algorithm 1 k-means Clustering

1. Initialise a set of k means.
2. Using Euclidean distance, assign each observation to the cluster with the nearest mean.
3. Recompute cluster means as the centroids of the clusters.
4. Repeat the allocation procedure.
5. End algorithm when no observations are re-allocated.

k-medoids Clustering

Definition

A *medoid* is a point whose average dissimilarity to all other points in a cluster is minimal i.e. the central-most point in a cluster.

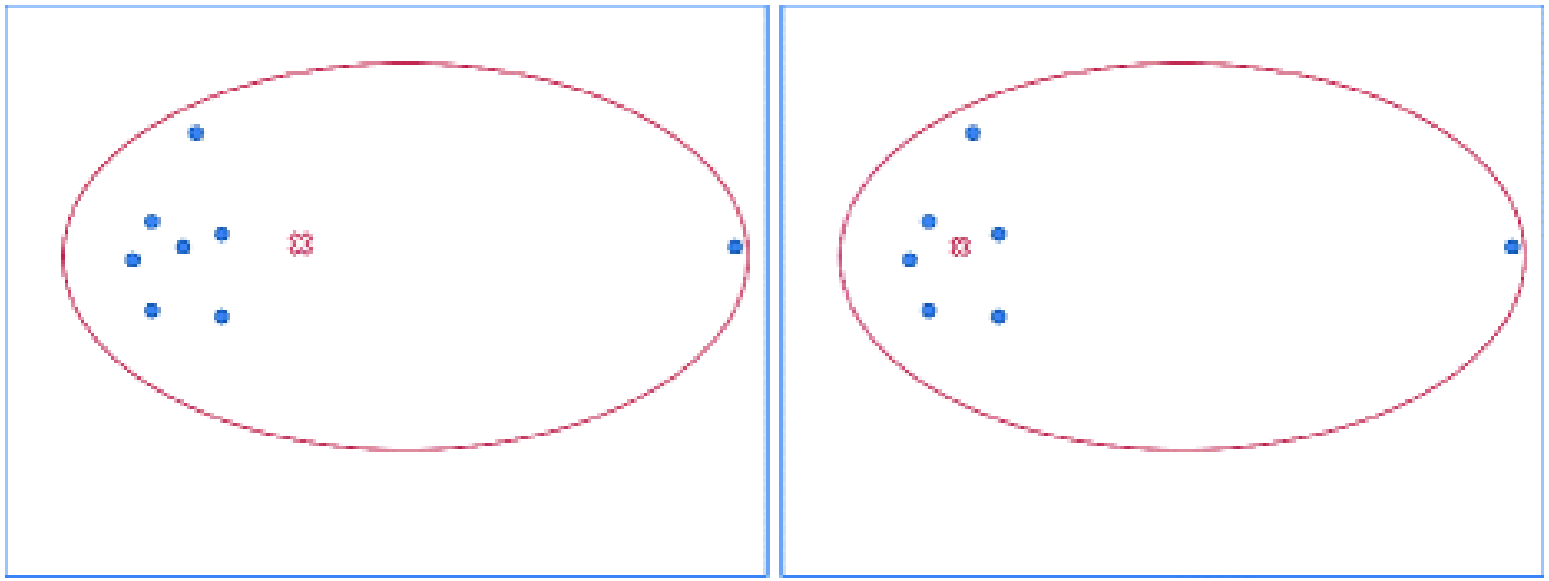

Figure: Sensitivity of the mean to outliers. Original image from Jin and Han (2010).

k-medoids Clustering

Algorithm 2 k-medoids Clustering

1. Initialise a set of k medoids.
2. Assign each observation to the nearest medoid.
3. Within each cluster, make each observation in turn the medoid and evaluate some cost function.
4. If the cost decreases, take the corresponding observation to be the new medoid. Else retain the previous medoid.
5. Reassign based on the new medoids and repeat until the cost function cannot be reduced further.

Hierarchical clustering

- Agglomerative vs. divisive
- Average linkage $\text{dist}_{AL}(C_r, C_s) = \frac{1}{n_r n_s} \sum_{i=1}^{n_r} \sum_{j=1}^{n_s} d(x_{ri}, x_{sj})$

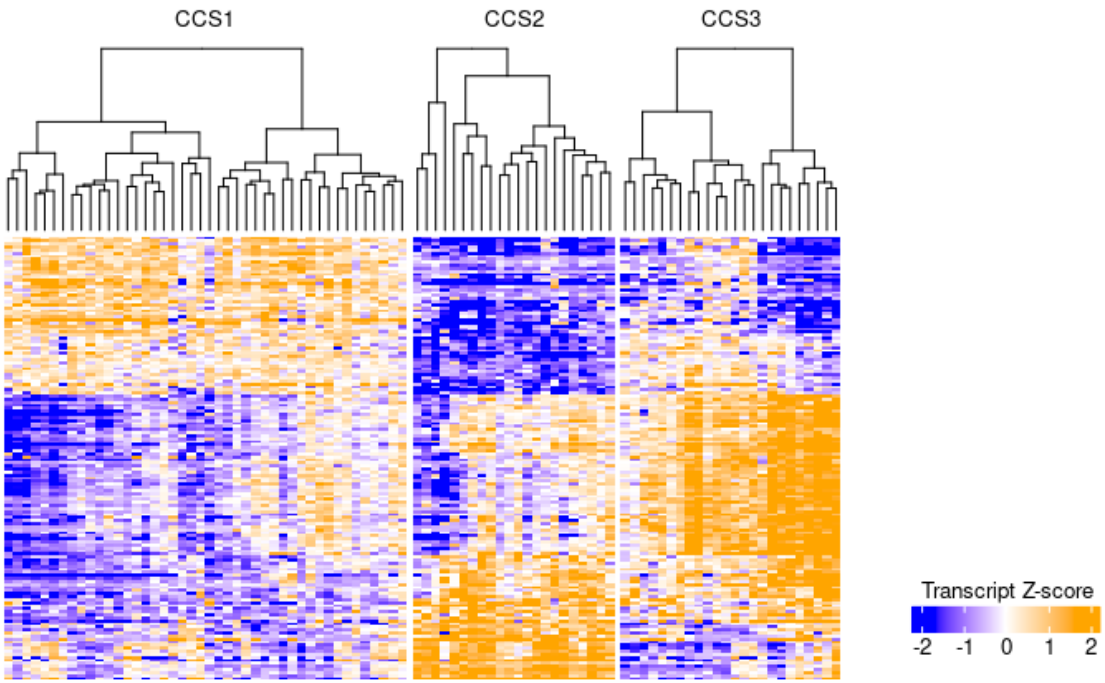

Figure: Heatmap of hierarchical clustering of the 90 patients (columns) using the 146 classifying genes (rows). Points correspond to median-centered $\log_2$ gene expression values (orange, high expression; blue, low expression).

Consensus clustering

Algorithm 3 Consensus clustering with hierarchical clustering

For $K \in \mathcal{K}$, for $h \in 1, \ldots, H$:

1. Bootstrap original data D to obtain perturbed dataset $D^{(h)}$
2. Compute indicator matrix: $I^{(h)}(i,j) = 1$ if patients i and j are both in $D^{(h)}$
3. Cluster $D^{(h)}$ into K clusters using hierarchical clustering
4. Construct connectivity matrix: $M^{(h)}(i,j) = 1$ if patients i and j are in the same cluster
5. At iteration H, construct the consensus matrix given by $\mathcal{M}^{(K)}(i,j) = \frac{\sum_h M^{(h)}(i,j)}{\sum_h I^{(h)}(i,j)}$

Choose best K based on the $\mathcal{M}^{(K)\prime}s$ and partition D into $\hat{K}$ clusters

Gap statistic

Definition

$$\text{Gap}(k) = E^*\{\log(W_k)\} - \log(W_k)$$

$$\hat{K} = \min\{K \in 1, \ldots K_{max} : Gap(k) \geq Gap(k+1) - s.e._{k+1}\}$$

Notation

- Sum of pairwise distances $D_r = \sum_{i,i' \in C_r} d_{ii'}$
- Pooled within-cluster sum of squares $W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r$

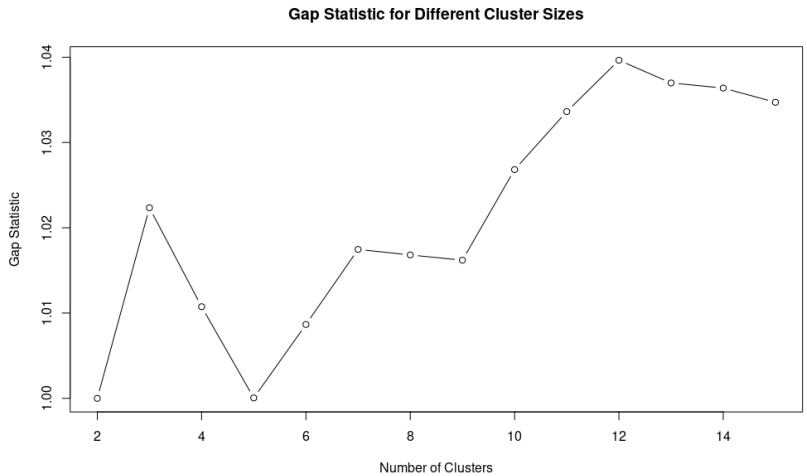

Figure: Values of the gap statistic for different numbers of clusters with 100 Bootstrap iterations.

Results

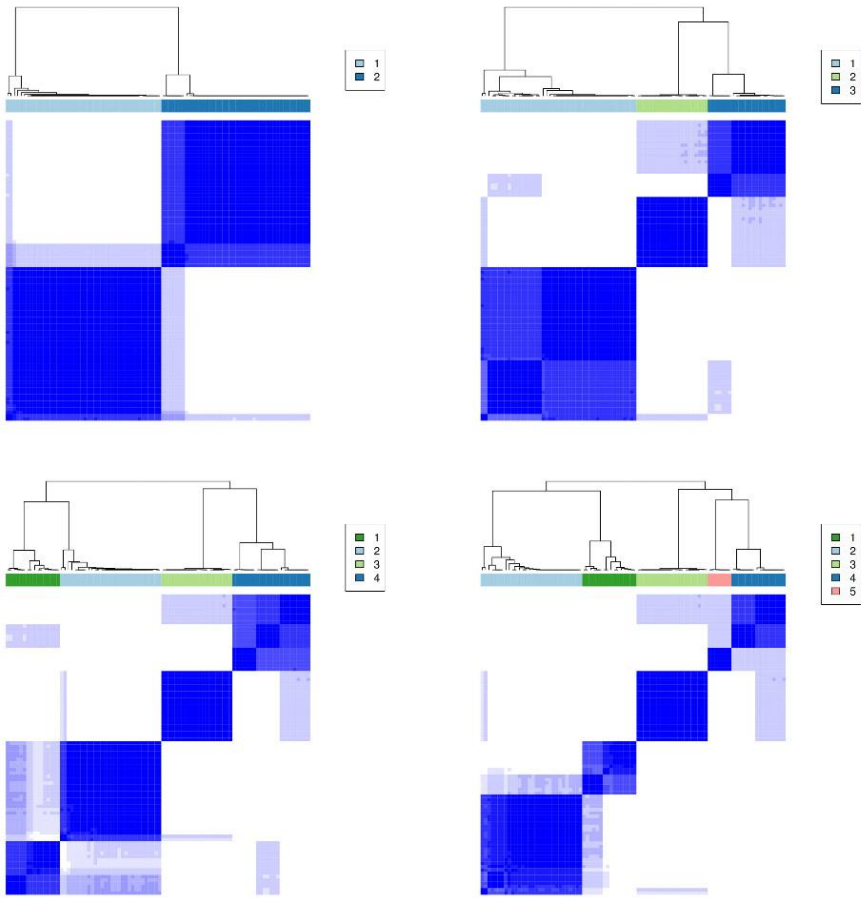

Figure: Consensus matrices for $K = 2$ to $K = 5$ produced with hierarchical clustering with 0.98 subsampling ratio and 1000 iterations using the selected 7,747 genes.

Results

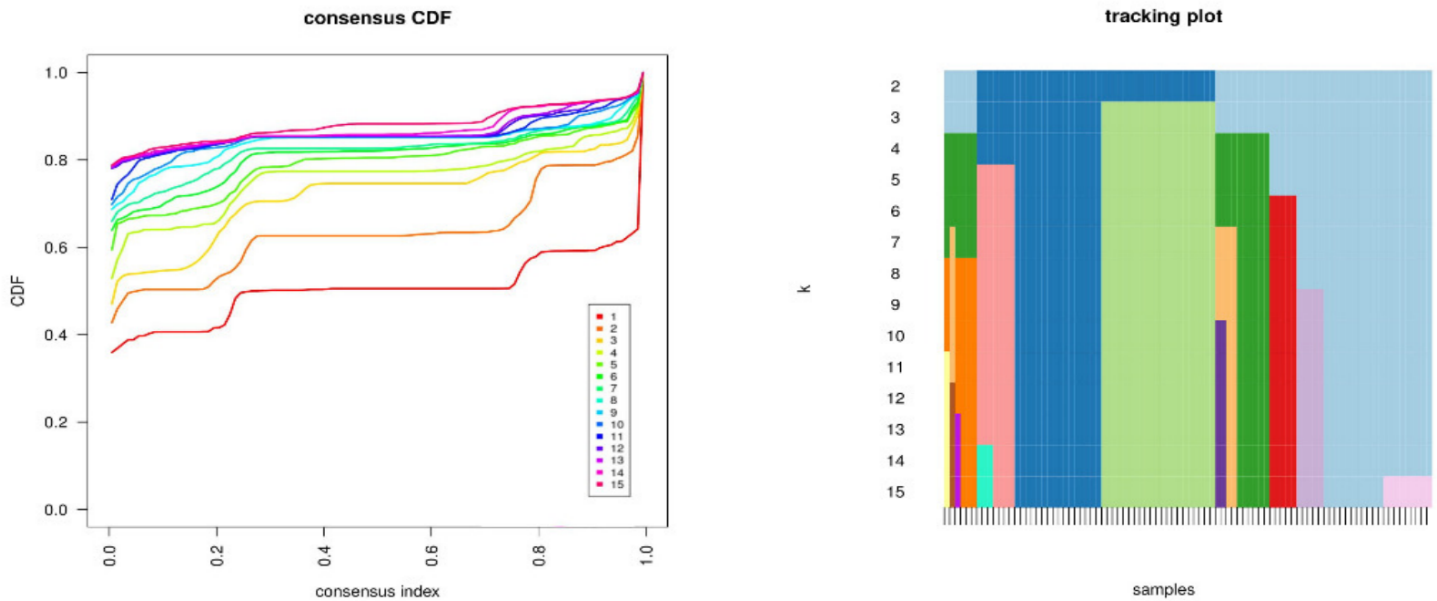

Figure: (Left) Consensus CDF for $K_{max} = 15$. (Right) Consensus tracking plot for $K_{max} = 15$.

Model-based Clustering

Concept

- Entire dataset $\mathbf{x} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots \mathbf{x}_i, ..., \mathbf{x}_n\}$ as coming from a mixture distribution with each component modelling a particular cluster
- Mixture Density: $f(\boldsymbol{x}_i; \boldsymbol{\Psi}) = \sum_{k=1}^{G} \pi_k f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$
  - $\boldsymbol{\Psi}$ denotes the set of parameters for the mixture
  - π_k represents a mixing weight for the k^{th} component where $\sum_{k=1}^{G} \pi_k = 1$
  - $f_k(\boldsymbol{x}_i; \boldsymbol{\theta}_k)$ is the density function for the kth cluster
- Each cluster is modelled by a component of the mixture distribution.

Model-based Clustering

Finite Gaussian mixture models

- Assumes a multivariate Gaussian distribution for each component (cluster) ; $f_k(\boldsymbol{x}; \boldsymbol{\theta}_k) \sim N(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$
  - $\boldsymbol{\mu}_k$ - center of cluster
  - $\boldsymbol{\Sigma}_k$ - covariance matrix
- Aim is to estimate parameters such that the likelihood is maximized.
  - Issues with non-uniqueness of maxima (Neal, 2011)
  - EM algorithm offsets this

Model-based Clustering

Finite Gaussian Mixture Model applied in the clustering of 7,747 genes from AMC-AJCCII-90 series dataset: R -output using 'mclust' package

```
> summary(FGM_Classif)
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust VEI (diagonal, equal shape) model with 4 components:

 log.likelihood  n    df      BIC      ICL
      -672251.4 90 38741 -1518830 -1518830
```

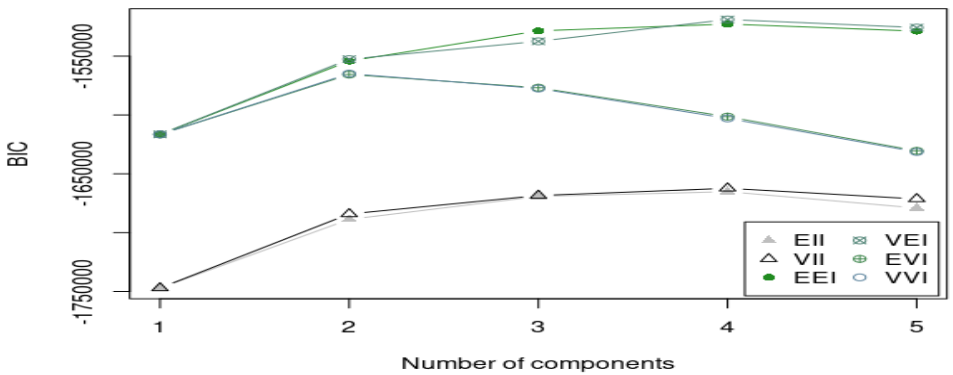

Figure: Plot indicating Bayesian Information Criterion (BIC) values for varying numbers of mixture components

Rand Index

- Computes a measure of agreement between two clustering assignments
- Given n observations $x_1, \ldots, x_n$ the rand index (RI) is given by

$$R = \frac{a+b}{\binom{n}{2}},$$

where a = number of pairs of observations which are clustered together by both methods;

b = denotes the number of pairs of observations which are not clustered together by both methods.

- Assumes values from 0 to 1 - no agreement to perfect agreement.
- Downside: Non-constant expected value since 'by-chance agreements' are unaccounted for.

Rand Index

- Computes a measure of agreement between two clustering assignments
- Given n observations $x_1, \ldots, x_n$ the rand index (RI) is given by

$$R = \frac{a+b}{\binom{n}{2}},$$

where a = number of pairs of observations which are clustered together by both methods;

b = denotes the number of pairs of observations which are not clustered together by both methods.

- Assumes values from 0 to 1 - no agreement to perfect agreement.
- Downside: Non-constant expected value since 'by-chance agreements' are unaccounted for.
- Solution: **Adjusted Rand Index (ARI)**!

Comparison of Clustering Methods

Classification methods	RI (Rand index)	ARI (adjusted RI)
HC with and without consensus	0.829	0.646
K means with and without consensus	1	1
K-medoids clustering with and without consensus	1	1
Model-based clustering and HC with consensus	0.7583021	0.4546874

Table: Measures of agreement between different clustering methods.

Prediction Analysis for Microarrays (PAM) (Tibshirani et al., 2002)

- **Nearest shrunken centroid method**:
- Define $\bar{x}_{ik} = \sum_{j \epsilon C_k} x_{ij}/n_k$ as the i^{th} component of the centroid for class k and $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$ as the i^{th} component of the overall centroid
- Class centroids are shrunk toward the overall centroid using a shrinkage parameter (threshold), Δ, which controls number of genes used in classifier.
- Optimal Δ is chosen by cross-validation and chosen to have minimal cross-validation errors.

Prediction Analysis for Microarrays (PAM) (Tibshirani et al., 2002)

- **Nearest shrunken centroid method**:
- Define $\bar{x}_{ik} = \sum_{j \epsilon C_k} x_{ij}/n_k$ as the i^{th} component of the centroid for class k and $\bar{x}_i = \sum_{j=1}^{n} x_{ij}/n$ as the i^{th} component of the overall centroid
- Class centroids are shrunk toward the overall centroid using a shrinkage parameter (threshold), Δ, which controls number of genes used in classifier.
- Optimal Δ is chosen by cross-validation and chosen to have minimal cross-validation errors.

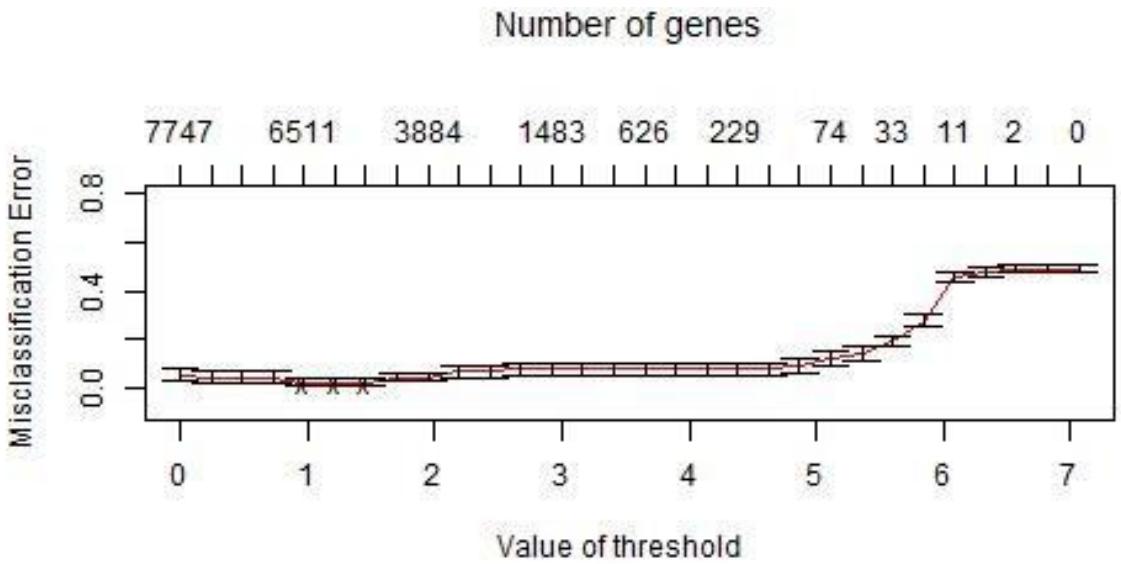

Figure: Plot showing cross-validation errors for different threshold levels (Δ). Desired Δ = 4.625 with 164 genes

Discussion

- Three subgroups of colon cancer were identified.
- Clustering procedures were robust and supported by other methods.
- We are able to classify patients into each cluster.
- There is a difference between the survival rates in each cluster.

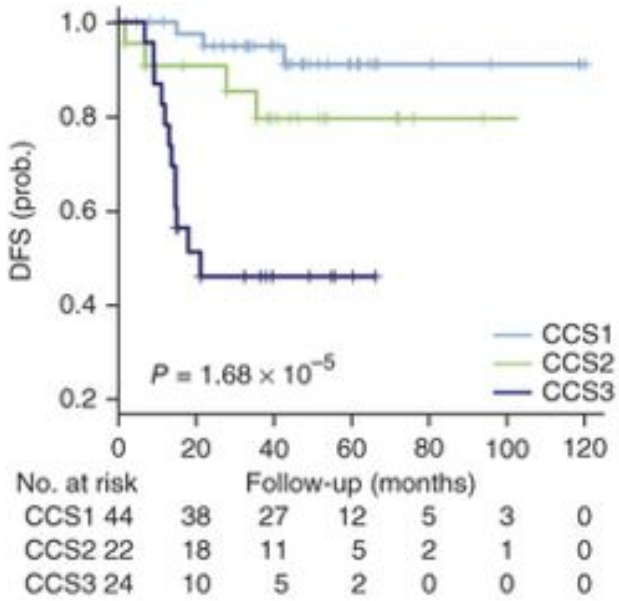

Figure: Kaplan-Meier survival plots for each of the identified subgroups. Original image is from De Sousa E Melo et al. (2013).

References

Felipe De Sousa E Melo et al. Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. **Nature medicine**, 19 (5):614, 2013.

Xin Jin and Jiawei Han. **K-Medoids Clustering**, pages 564–565. Springer US, Boston, MA, 2010. ISBN 978-0-387-30164-8. doi: 10.1007/978-0-387-30164-8_426. URL https://doi.org/10.1007/978-0-387-30164-8_426.

R. Neal. Clustering and mixture models. http://www.utstat.utoronto.ca/~radford/sta414.S11/week10a.pdf, 2011. [Online: Accessed 15-Nov-2018].

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. **Proceedings of the National Academy of Sciences**, 99(10):6567–6572, 2002.