# DIPLOMA THESIS

## Oliver S Kühnle

## Integration of multiple high-throughput data-types in cancer research

Department of Statistics
Ludwig Maximilians University of Munich
Field of study: Statistics
Supervision: Prof. Dr. Volker J Schmid

In cooperation with

Memorial Sloan Kettering Cancer Center, Bioinformatics Core, New York City
Supervision: PhD Raya Khanin and PhD Nicholas Socci

$30^{th}$ November 2011

# Abstract

This thesis deals with integrative analysis of two (different) multi-dimensional data types in the context of genomic cancer research. Firstly, simultaneous clustering of two high-throughput data sets optimizing the output of iCluster program. Secondly, developing a novel the Gene Set Score (GSS) that uses different data types to identify de-regulated pathways (or gene sets).

Regarding the clustering, the key concern is how to process and filter with the two different data types (Gene Expression and Copy Number data) and how consistent the results of iCluster are and how to determine the optimal (best) number of clusters. The output from different number of clusters is systematically studied using Wallace Index and Rand Index. A large sarcoma data set is analyzed, and the survival analysis of the samples of this cancer data with significant $p$-values is the resulting method of analysis. Altogether it can be said that iCluster is a good method to cluster samples into groups and this methodology is readily applicable to other data sets as well as to genomic areas where multi-dimensional data sets are produced (image analysis, psychological profiles).

The Gene Set Score is a novel score that uses Gene Expression and Copy Number data to identify significantly de-regulated gene sets/pathways/groups of genes. For the Gene Set Score, scores for single genes are calculated that merge together to a score of a given gene set/pathway. Events in the data then create the Gene Set Score. The resulting method identifies not only most of differentially regulated gene sets from a popular and widely used method (Gene Set Analysis, GSA) that is based on Gene Expression data only but additional pathways.

**Keywords:** iCluster, integrative clustering, Pathway analysis, Gene Set Score, Gene Expression data, Copy Number data, Gene Set Analysis

# Acknowledgement

First I wish to thank PhD Raya Khanin, who has been very generous in giving me the chance to come to New York again to write on this thesis with her and with her time and talents in discussing my questions with respect to this thesis and new research ideas.

I have been fortunate to be under the supervision of PhD Nicholas Socci and to have the support from all my colleagues of Bioinformatics Core as well as the labs providing me with the data.

My gratitude goes to Prof. Dr. Volker J Schmid for being the supervisor at the University of Munich.

I owe my girlfriend Deitra Myers a lot as she was not only a source of encouragement to me the whole time, but also graciously proofread my work.

A big "Thank you!" also to my parents, who always encouraged in all my projects with personal and financial support.

Last but not least, I would like to thank all my friends who visited me here in New York or supported me from home as well as all other people who helped me in a statistical or, more importantly, in a personal way to finish this thesis.

# Contents

# List of Figures

# List of Tables

# CHAPTER 1

# Introduction

I start with an introduction of the Memorial Sloan-Kettering Cancer Center and the Computational Biology Center including the research group cBio Core, where I wrote my thesis. Then I give a short overview about my thesis including the main goals.

## 1.1 Memorial Sloan-Kettering Cancer Center (MSKCC)

The Memorial Sloan-Kettering Cancer Center[1], founded in 1884, is the world's oldest and largest private cancer research center. More than 10000 employees work for the MSKCC in different fields. Up to 400 various cancer subtypes are treated by the 16 multidisciplinary cancer teams. In collaboration with Rockefeller University, Cornell University, and Weill Medical College of Cornell University, MSKCC is actively involved in the education and training of its employees. In addition, a significant amount of research is conducted within Sloan-Kettering in basic, translational, and clinical research. Important goals in research are to understand the biology of cancer through various programs, such as computational biology.

## 1.2 Computational Biology Center (cBio)

The Computational Biology Center was founded in 2002 by the MSKCC. cBio deals with research and service components. The Computational Biology Center is working on computational biology research projects as well as on developing tools in areas such as sequence-structure analysis, gene regulation, molecular pathways

[1] http://www.mskcc.org/

and networks, diagnostic and prognostic indicators in order to apply the theoretical methods and genome-scale data in everyday laboratory practice and use. For this computational methods are being created, that are available as open-source methods. The Bioinformatics Core is one of the four research groups of cBio. The Computational Biology Center is responsible for computational and bioinformatics services to the Tri-Institutions of the Upper East Side of Manhattan which include Memorial Sloan-Kettering Cancer Center, as well as Rockefeller University and Weill Cornell Medical College. To date, the Bioinformatics Core has developed analysis pipelines for next-gen and Sanger sequence data, CGH array normalization and segmentation pipeline and an array of other projects. Refer to cBio's webpage[2] for more detailed information. Another big topic is continuos work on the TCGA project.

## 1.3 Goals of the thesis

This thesis is dealing with one of the major current challenges in cancer research (and genomics in general). The most important point is the clustering of different types of multiple high-throughput data in one step using special methods including the validation of the results. Another big topic is the introduction of a new score, the so called Gene Set Score (GSS), that identifies gene sets that are significantly different between two conditions/groups of samples.

### 1.3.1 Clustering multiple data types with iCluster

There is a variety of different methods for clustering each data type on its own, and then applying a method that integrates the results manually. There are a lot of clustering methods that just cluster one data set at once (Hierarchical clustering, k-Means-Clustering). Besides the methodical issue (hierarchical, Partitional and spectral clustering can be used), it is also important to decide if one wants to cluster the samples, the genes or both together. Over the years a lot of those clustering methods were implemented and some provide good results for special cases. To manually integrate, let us say to combine the clustering results of more than one data type, special methods are needed that often lead to an increase in the quality of the samples as different methods for different data sets hardly result in the same clusters for all the samples. A lot of those methods also deal with a

[2] http://cbio.mskcc.org/

high dimension calculation that maybe reasonable for one data set at once, but obviously not for more than one data type at a time.
Knowing about all these problems there were some approaches to deal with them as Shen et al. [28] mentioned, but none of them was dealing with all the of the problems and that is why an integrative method called iCluster, which stands for 'Integrative clustering of multiple genomic data types using a joint latent variable model' was created. The iCluster method is capable of dealing with one data type as well as with more data types at once. This is important, since a lot of different data types have become increasingly available in cancer research within recent years. The two biggest challenges Shen et al. had to deal with were to 'capture both concordant and unique alterations across data types'. For this, one has to model the covariance between data types separate from the (co-)variance within each of the data types. While using integrative methods it is important to look at the concordant and unique alteration patterns, as it is possible that either way contains information about the subgroup of the cancer. The other challenge was to get good results in a reasonable time, what was solved by using a dimension reduction. Although iCluster still needs a lot of time to calculate the results for all the different tuning parameters to find the best results, such calculations would be impossible without a dimension reduction. Shen and their colleagues are doing this with the k-Means-Algorithm using the Principal Component Analysis as well as a latent Gaussian model. After getting the results one has to validate them. The Proportion Of Deviance is included in the iCluster analysis, but this does not test the stability of the clusters if one calculates them with different methods or different numbers of clusters $k$. This thesis deals with introducing different indices looking for the stability of the clusters and choses the best one. After that a survival analysis is done to validate the results using clinical data.

### 1.3.2 Gene Set Score (GSS)

Numerous methods exist that yield to identify significantly different gene sets/ pathways using Gene Expression data (Gene Set Analysis, Gene Set Enrichment Analysis, Signaling Pathway Impact Analysis - Section 7.1). The same methods are applicable to other quantitative data types such as protein abundances. The Gene Set Score deals with two different data types together and identifies differently expressed gene sets/pathways. One checks in both data types if there is an event in a gene set in Copy Number data as well as if there is an event in gene sets in

Gene Expression data. The combined score for both data types is then the Gene Set Score that finds differently expressed gene sets. A validation of this score is done by comparing the results with the results of the Gene Set Analysis. The GSS does not only find most of the pathways of the GSA, but it also finds other pathways that GSA does not find.

The methods include scores for each gene. Depending on the data type Fold Changes (Gene Expression data) or a more detailed method (Copy Number data) is used. With these Gene Scores it is possible to calculate the score for a given gene set/pathway.

# CHAPTER 2

# Methods

In this chapter I will give an overview of the methods used in my thesis. In the first sections I will introduce the methods contained in iCluster and methods to analyze the results of iCluster.

## 2.1 Expectation-Maximization-Algorithm (EM-Algorithm)

In 1977, Dempster et al. [6] created a new method to find the maximum likelihood estimates of parameters. The used statistical models depend on unobserved latent or missing variables. The algorithm works in two steps. One expectation (E) and one maximization (M) step.

**Algorithm 1:** Expectation-Maximization-Algorithm

```
Data: The goal is to calculate the parameter $\tau$ with complete data $c_i$ and
      known incomplete data $u_i$.
Starting value $\tau^{(0)}$
Result: $\tau$
repeat
    E-step:
    Calculate the conditional expectation over the latent/missing variable:
    $Q(\tau) = Q(\tau|\tau^{(i)}) = E[\ell(\tau, x)|y, \tau^{(i)}]$
    M-step (often analytical):
    Maximize $Q(\tau)$ respecting $\tau$ and get a new estimated value $\tau^{(i+1)}$:
    $\tau^{(i+1)} = \underset{\tau}{\text{argmax}} Q(\tau|\tau^{(i)})$
until convergence
```

## 2.2 Principal Component Analysis (PCA)

To simplify a data set Pearson [23] established a method called Principal Component Analysis (PCA) in 1901, that Hotelling [13] improved. PCA uses a dimension reduction method and is advantageous because it is implemented in most standard programs. The PCA projects data in a selected $d$-dimensional space. With each axis representing one of the original dimensions one can also create new axes with linear combinations of the original axis.

'The first principal component, is the axis through the data along which there is the greatest variation amongst the object' as Wit and McClure mention in their book [38]. Orthogonal to the first component is the second principal component that 'has the greatest variation in the data associated with it' [38]. The third component is then orthogonal to the first two principal components and so forth. This is done by first computing the covariance matrix for the complete data set. The next steps include calculating the eigenvectors and eigenvalues of that covariance matrix. The largest eigenvalue represents the first component, as it contains most of the variation.

The PCA maximizes the between-cluster variance, but the within-cluster variance is not minimized, what is a known problem.

## 2.3 $k$-Means-Algorithm

### 2.3.1 Standard $k$-Means

$k$-Means was first introduced in 1967 by MacQueen [17]. It is an algorithm that solves a clustering problem. This algorithm classifies a data set through a prior specified number of clusters $k$. A step-by-step tool[3] is used for the figures to illustrate this algorithm. The $k$-Means-Algorithm is working by the following steps:

[3] http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletKM.html

```
Algorithm 2: $k$-Means-Algorithm
  Data: Different data points. One has to determine how many clusters one
        wants to obtain.
  Result: Cluster membership of the data points.
  for numerous times (optional) do
      The starting points of the $k$ clusters are allocated randomly within the
      data set. One then calls this the $k$ group centroids, see Figure 2.1.
      repeat
          Assign each of the $n$ data points to the closest centroid using the
          euclidean metric, see Figure 2.2.
          For each of the $k$ centroids redetermine the center of the centroids, see
          Figure 2.3 and Equation 2.1.
      until convergence, see Figure 2.4
```

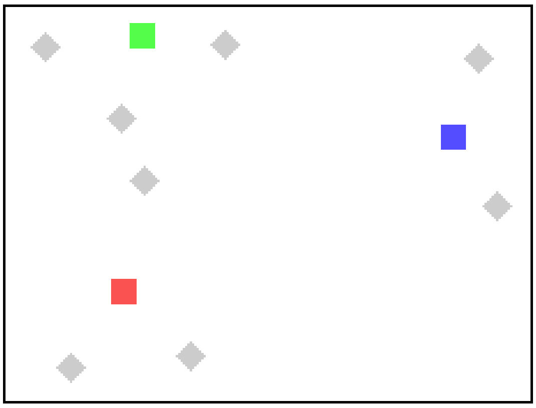


Figure 2.1: Data at the beginning - colored rectangles represent $k$ allocated starting points, grey circles represent the $n$ data points

The last two steps are repeated until convergence, such that the centroids are no longer moving. Though this algorithm always terminates, as proven in [17], it is obvious that this algorithm does not always find the global solution because of step one. A possible way to deal with this problem is to run this algorithm multiple times, which is done by iCluster.

The sum of within-cluster squared distances is minimized by $k$-Means, such that:

$$\min_k U_K = \min_k \sum_{i=1} \sum_{x_j \in K_j} ||x_j - \mu_i||^2 \tag{2.1}$$

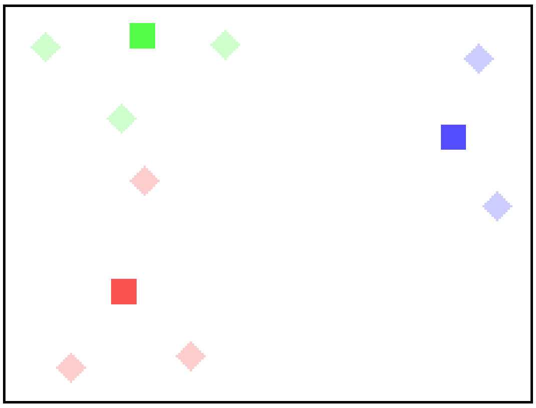


Figure 2.2: Data points assigned to the closest centroid

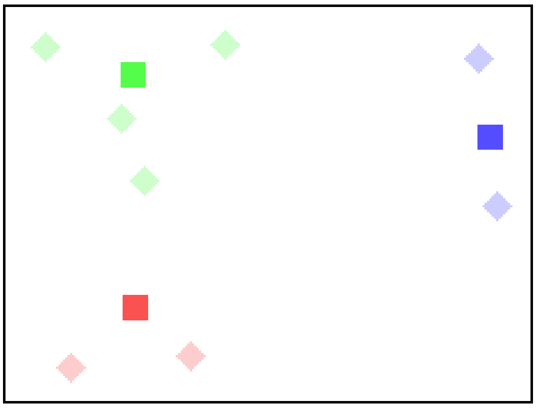


Figure 2.3: New centroids are calculated (rectangles) and data points assigned to the closest centroid

Where $X = (x_1, \ldots, x_j, \ldots, x_n)$ are the $n$ data points and $\mu_1, \ldots, \mu_i, \ldots, \mu_k$ are the mean vectors of all $k$ centroids.
As it was shown in the literature [7], [41] $k$-Means has better optimization, if one uses $k$-Means through Principal Component Analysis. A lack of good optimization occurs, because $k$-Means is sensitive to the starting points of the $k$ clusters. The $k$-Means-Algorithm often tends to find local solutions and not the global minimum.

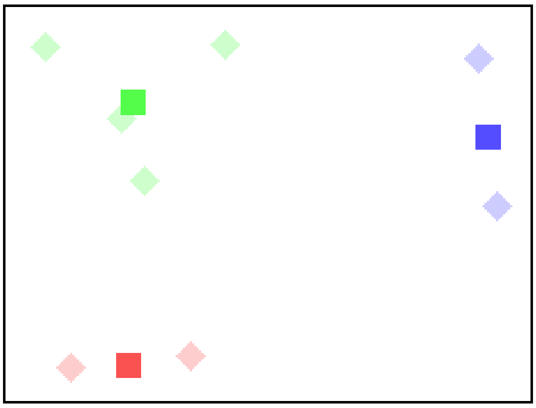


Figure 2.4: Steps continued until convergence

To represent the solution of the clustering, it is possible to show this with $k$ non-negative indicator vectors: $J_K = (j_1, \ldots, j_k)^T$. With $m_k$ being the number of points in Cluster $k$ and $\sum_{k=1}^{K} m_k = m$, $j_k$ is represented as:

$$j_k = (0, \ldots, 0, \underbrace{1, \ldots, 1}_{\mathrm{m_k}}, 0, \ldots, 0)^T / \sqrt{m_k} \qquad (2.2)$$

### 2.3.2 *k*-Means-Algorithm using the PCA

Equation 2.1 can then be rewritten as

$$U_K = tr(X^T X) - tr(J_K^T X^T X J_K). \qquad (2.3)$$

By calculating the minimum of $U_K$ the first part of Equation 2.3 is the total variance and the second part is the between-cluster variance. As the total variance is constant, since the data is always the same, it can be ignored. As a result the minimization of $U_K$, is the maximization of the between-cluster variance $tr(J_K^T X^T X J_K)$.

Under some assumptions this is equivalent to the eigenvalue decomposition. See Shen et. al [28] and Ding and He [41] for more details. The eigenvalue decomposition is a result in the Principal Component Analysis. The principal components are the indicator vectors $J_K$ of the $k$-Means-Algorithm. This is only possible for continuous $J_K$, otherwise there has to be an extra step to get the solution. PCA is a powerful method to reduce the dimension of a data set.

## 2.4 Gaussian variable model

There are two main reasons why it is useful to transcribe the $k$-Means into a Gaussian latent variable model. With a Gaussian latent variable model inference is possible as well as it naturally extends to multiple data types. To get this model we have to rewrite the $k$-Means-Algorithm using the PCA:

$$V = WJ + \epsilon \tag{2.4}$$

In this equation $V$ represents an $p \times n$ expression matrix with no intercept and it is mean-centered. The (cluster) indicator matrix, as defined in Equation 2.2, with $J = (j_1, \ldots, j_{K-1})^T$, has dimension $(K-1) \times n$. The $p \times (K-1)$ dimension matrix $W$ is defined as the coefficient matrix. $\epsilon = (\epsilon_1, \ldots, \epsilon_p)^T$ is the error-vector with mean 0 and a diagonal covariance matrix $Cov(\epsilon) = diag(\sigma_1, \ldots, \sigma_p) = \Sigma$. The true subtype in the data is the indicator matrix $J$, that is treated as latent variables. The discovery of the true subtypes is the goal of this Gaussian variable model. $W$ is used to get a dimension reduction, as it is used as a projection matrix. Under some assumptions it is possible to solve the $k$-Means problem with a likelihood-based approach. This is possible through the model 2.4, that is introduced above. For the inference one needs the data and the posterior mean of the indicator matrix with continuous values.

## 2.5 Introduction of iCluster

With iCluster it is possible to estimate the cluster indicator matrix $J$ for different types of data within one step. This is a big advantage in comparison to a lot of other methods, including hierarchical clustering, $k$-Means clustering and self-organizing maps. Most methods cluster a tumor subtype for each data type alone. In Figure 2.5 there is a stand alone tumor subtype $J_1$ for Gene Expression data and another tumor subtype $J_2$ for Copy Number data that need a manuel integration to get one tumor subtype $J$. With iCluster there is just one tumor subtype $J$ for all data, see Figure 2.6, because of the integrative clustering.

The data, each for the same set of $n$ samples, can be for example mRNA Gene Expression data, DNA Copy Number data, DNA Methylation data, et cetera. Let us denote the data with $V_1, V_2, \ldots, V_m$. The dimension of $U_l$ is $p_l \times n$. The iCluster

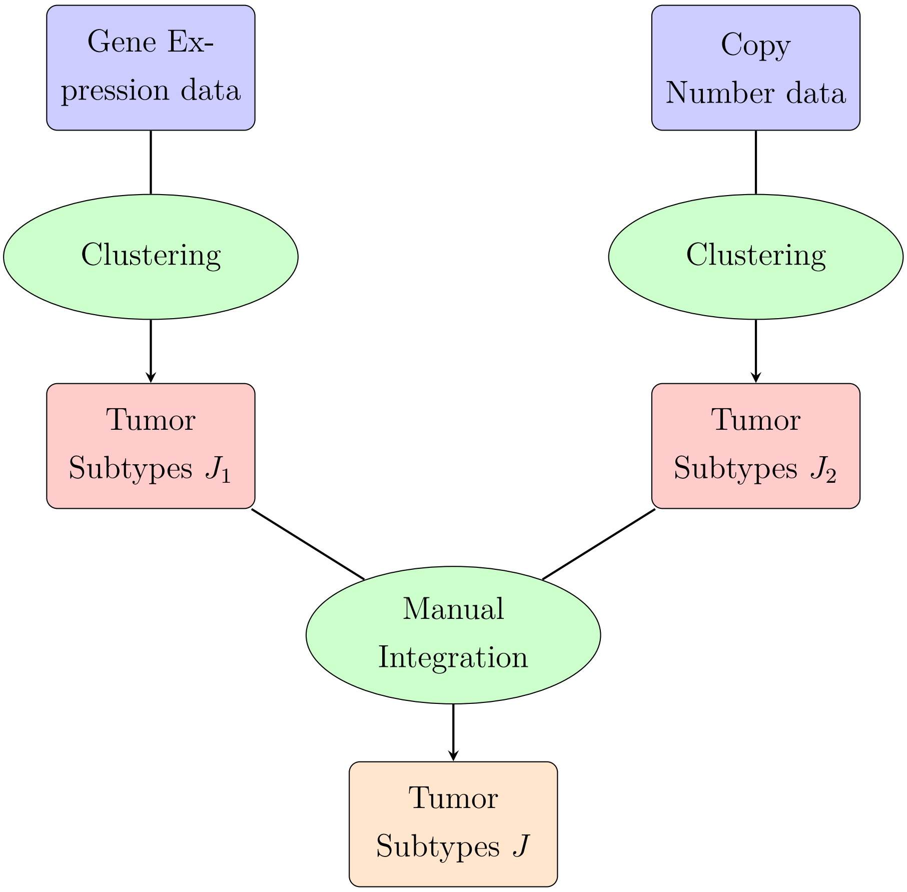


Figure 2.5: Clustering of different data types with multiple steps

model can then be written as:

$$\begin{aligned} V_1 &= W_1 J + \epsilon_1 \\ V_2 &= W_2 J + \epsilon_2 \\ &\vdots \\ V_m &= W_m J + \epsilon_m \end{aligned} \tag{2.5}$$

To estimate $J$ for all $m$ data sets at once, it is necessary that all data types are measured with the same samples. $J$ then becomes a latent component connecting all the $m$ sets of the data. In addition there are individual error terms $\epsilon_q$ for every data type with $q = 1, \ldots, m$. Each of these error terms has a mean of zero and a diagonal covariance matrix $\sigma_q$. This represents the variance that remains to each data type, after calculating the correlation across data types. The matrices $(W_1, \ldots, W_m)$ are called coefficient matrices.

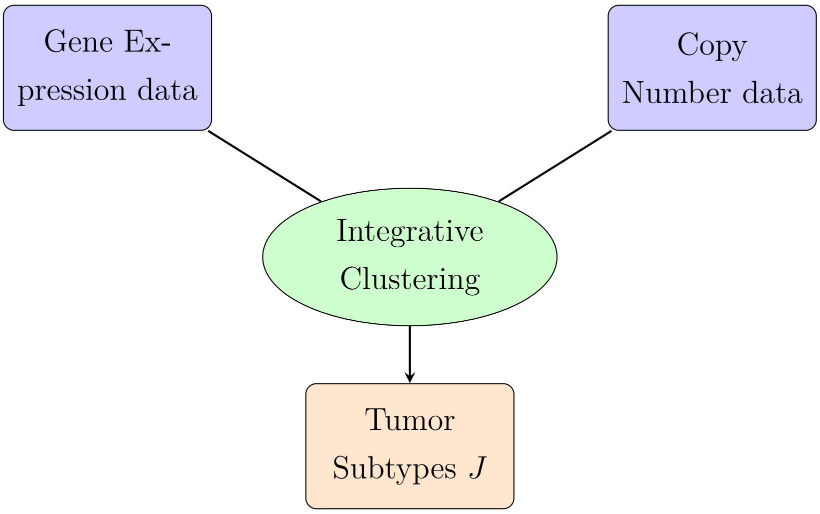


Figure 2.6: Integrative clustering of different data types with one step

A latent continuous parameterization with $J^* \sim \mathcal{N}(0, I)$ is used to get likelihood-bases solutions of Equation 2.5. The error term $\epsilon$ is also normally distributed with mean 0 and has a diagonal covariance matrix $Cov(\epsilon) = diag(\sigma_1, \ldots, \sigma_p) = \Sigma$ as in the Gaussian variable model.
The marginal multivariate normal distribution of the integrated data matrix $V = (V_1, \ldots, V_m)^T$ is the result. This distribution has mean 0 and covariance matrix $\Omega = WW^T = \Sigma$ with $W^T = (W_1, \ldots, W_m)$. The sample covariance matrix is:

$$G = \begin{pmatrix} G_{11} & G_{12} & \ldots & G_{1m} \\ G_{21} & G_{22} & \ldots & G_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ G_{m1} & G_{m2} & \ldots & G_{mm} \end{pmatrix} \tag{2.6}$$

With the distribution it is quite easy to write down the log-likelihood function:

$$\ell(W, \Omega) = -\frac{n}{2}\left(\sum_{i=1}^{m} p_i ln(2\pi) + ln(det(\Omega)) + tr(\Omega^{-1}G)\right) \tag{2.7}$$

To get the results of $W$ and $\Sigma$ of the maximum likelihood equation Shen et. al [28] suggest to use the Expectation-Maximization-Algorithm. They deal with the complete-data log-loglikelihood, that is the following:

$$\begin{aligned} \ell_c(W, \Sigma) = -\frac{n}{2}\left(\sum_{i=1}^{m} p_i ln(2\pi) + ln(det(\Sigma))\right) \\ -\frac{1}{2}\left(tr((V - WJ^*)^T \Sigma^{-1}(V - WJ^*)) + tr(J^{*T}J^*)\right). \end{aligned} \tag{2.8}$$

One advantage of the complete-data log-likelihood, Equation 2.8, is that one does not have to calculate the covariance matrix in Equation 2.6, which is computer intensive. This yields to a more efficient way of calculating the parameters than using the marginal data likelihood 2.7
Since there are many more data points than samples ($p >> n$) it is required to work with a sparse solution of the model to get good results. One possibility is to work with a sparse solution of the coefficient matrices. The idea is to work with a penalized log-likelihood of the complete-data log-likelihood, that is described in detail in Section 2.6.

## 2.6 A sparse version of iCluster

As described before, the complete data log-likelihood is penalized. $P_\lambda(W)$ is the penalty term of $W$ with the regularization parameter $\lambda > 0$. A lot of different penalty parameters can be used. Shen et. al [28] used a lasso type penalty. This $L_1-$norm penalty was introduced by Tibshirani [35] and can be written as:

$$P_\lambda(W) = \lambda \sum_{i=1}^{m} \sum_{k=1}^{K-1} \sum_{j=1}^{p_i} |w_{ikj}|. \tag{2.9}$$

The penalized complete-data log-likelihood is then:

$$\ell_{c,p}(W, \Sigma) = \ell_c(W, \Sigma) - P_\lambda(W). \tag{2.10}$$

With Equation 2.10 it is possible to run the EM-Algorithm. The E-step is responsible for a simultaneous dimension reduction. Updating of the parameters is done within the M-step. The EM-Algorithm runs until convergence. Therefore, a threshold must be set to determine when convergence is reached. Because the algorithm does not always converge, there has to be a maximum number of iterations for the EM-Algorithm. If the algorithm converges, the indicator matrix $J$ is calculated by running standard $k$-Means on the estimated expected value of the E-step until the convergence criterion is met.
As mentioned before, the sparse version of iCluster is achieved, since there is a sparse version of the coefficient matrices $W$. To achieve this, many coefficients of $W$ are shrunken toward zero. This leads to a reduction of the variance of this model and thereby to a better clustering performance. The better clustering performance can be shown with the bias-variance trade-off. By using this method

some of the coefficients of $W$ are shrunken exactly to zero by the Lasso penalty. One can now find genes that have non-zero loadings in the $k$-th cluster.

## 2.7 Proportion Of Deviance (POD)

It is possible to do a model selection based on a variety of different criterions such as $p$-values, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), et cetera. It is also possible to base it on the cluster separability. Shen et al. [28] are using the Proportion Of Deviance (POD) to show whether the cluster separability is weak or strong, with $0 \leq POD \leq 1$. Small values of POD indicate strong separability of the clusters. If there are non-overlapping subclasses ($POD = 0$), there is a perfect cluster separability and it is therefore an exact diagonal block matrix. One calculates a value of the sum of absolute differences between the 'perfect' diagonal block structure, a matrix of ones and zeros, and a matrix $\hat{B}^*$ with $\hat{B}^* = \hat{E}[Z^*|X]^T[Z^*|X]$.
For more details about $\hat{B}^*$ please refer to the following paper [28]. The best number of clusters $k$ and the best Lasso parameter $\lambda$ can be found with POD. For this one calculates the best POD for the data set. The lowest POD for all possible $k$ and all possible $\lambda$ is then the best POD.
As Shen et al. used three different thresholds for the tuning parameter $\lambda$ it is obvious that in a real data set one should not focus on the best global POD as the values for the POD do not differ a lot for the top ones for each $k$ and the different data sets as one can see in the Appendix A. For that reason the best $\lambda$ was the result one obtains. This case then was analyzed through all the subsets in Section 5.1 and was validated in extra steps like the stability of the different clusters and the clinical analysis in Sections 5.2 and 5.3.
As there are a lot of different subsets used to analyze the data it is necessary not only to look for the lowest POD in the analysis.

## 2.8 Comparison of the results for different $k$ and different data sets

There are two ways to compare different cluster analyses. One is how easy the clustering method is to use and how fast it produces results. This part is more computer-oriented than the second part. The most easiest and fast computed cluster method is useless if the results are worthless because they are just random results. One has to think about how to solve this problem and there were two methods used in the analysis of iCluster and looking at the results.

First of all, using different subsets of the data and showing that all yield to the same clustering is one part of it. In this analysis a broad range of different subsets are used to deal with this issue. The other part is how the clusters behave if one increases the number of clusters from, for example, $k = 2$ to $k = 3$. Are the samples still (for $k = 3$) behaving the same way as they did before ($k = 2$)? Or is this analysis random for different numbers of $k$, what one does not want to be the result. In the other part it is also of interest if one is comparing the clustering membership of the samples/patients for two (or more) different analyses. Here it should also not be a random result. In this section four different methods are introduced that are implemented in the R-package $profdpm$ of Shotwell [30]: The Rand Index [25], the Wallace Indices [37], the Fowlkes and Mallows Index [9] as well as the Jaccard Index [18].

For the indices one has to define some parameters:

To compare different clustering results one is looking at the triplet $(X, Y, m)$. Here $X$ stands for the $N$ samples that are clustered with $X = (X_1, X_2, \ldots, X_i, \ldots, X_N)$. $Y$ is a specific clustering of these samples into $K$ disjoint clusters:

$Y = (Y_1, Y_2, \ldots, Y_K)$. Each of these clusters contain one or multiple samples: $Y_k = (X_{k_1}, X_{k_2}, \ldots, X_{k_{n_k}})$ with $\sum n_k = N$ and $n_k \geq 1 \; \forall \; k = 1, 2, \ldots, K$. Finally, $m$ stands for the method used for this clustering.

In the following the pair $(X, Y)$ is used for $m = 1$ and $(X, \tilde{Y})$ for $m = 2$. Looking at a each pair of the $X_i$ and $X_j$, it can either be in the same cluster for $Y$ and $\tilde{Y}$ for both methods or within the same cluster for one of the methods or not in the same cluster at all. This leads us to define the following parameters that make the calculations of some indices more easy:

- $n_{11}$ represents the number of attributes where $Y$ and $\tilde{Y}$ are equal in both methods

- $n_{10}$ represents the number of attributes where $Y$ and $\tilde{Y}$ are equal in the first method, but different in the second method

- $n_{01}$ represents the number of attributes where $Y$ and $\tilde{Y}$ are different in the first method, but equal in the second method

- $n_{00}$ represents the number of attributes where $Y$ and $\tilde{Y}$ are different in both methods with $n_{11}+n_{10}+n_{01}+n_{00} = N$. An example for this would be the following. There are six samples with $Y = ([a,b,c],[d,e,f])$ being the clustering in two clusters with method one and $\tilde{Y} = ([a,b],[c,d,e],[f])$ that is similar to the example of Rand [25], but he does not split it the exact same way. Besides the four introduced methods there exist many more indices that are also worth thinking about. But for this analysis we use the four most common ones.

| Point-pair | ab | ac | ad | ae | af | bc | bd | be | bf | cd | ce | cf | de | df | ef | sum |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $n_{11}$ | $*$ | | | | | | | | | | | | $*$ | | | 2 |
| $n_{10}$ | | $*$ | | | | $*$ | | | | | | | | $*$ | $*$ | 4 |
| $n_{01}$ | | | | | | | | | | $*$ | $*$ | | | | | 2 |
| $n_{00}$ | | | $*$ | $*$ | $*$ | | $*$ | $*$ | $*$ | | | $*$ | | | | 7 |

Table 2.1: Example of the Comparison

### 2.8.1 Rand Index (RI)

The Rand Index was introduced in 1971 by Rand [25] with the left part of the following equation. The second part makes the formula more intuitive:

$$RI(Y,\tilde{Y}) = \frac{\sum_{i<j}^{N} \gamma_{ij}}{\binom{N}{2}} = \frac{n_{11}+n_{00}}{n_{11}+n_{00}+n_{10}+n_{01}} \tag{2.11}$$

$$\text{with } \gamma_{ij} = \begin{cases} 1 & \text{there is a } k \text{ and } \tilde{k} \text{ such that } X_i \text{ and } X_j \text{ are in } Y_k \text{ as well as in } \tilde{Y}_{\tilde{k}} \\ 1 & \text{there is a } k \text{ and } \tilde{k} \text{ such that } X_i \text{ is in } Y_k \text{ as well as in } \tilde{Y}_{\tilde{k}} \text{ while } X_j \\ & \text{is in none of them} \\ 0 & \text{for all the other cases} \end{cases}$$

with $\gamma_{ij} = 1$, if there is a $k$ and $\tilde{k}$ such that $X_i$ and $X_j$ are in $Y_k$ as well as in $\tilde{Y}_{\tilde{k}}$
with $\gamma_{ij} = 1$, if there is a $k$ and $\tilde{k}$ such that $X_i$ is in $Y_k$ as well as in $\tilde{Y}_{\tilde{k}}$ while $X_j$ is in none of them
and with $\gamma_{ij} = 0$, for all the other cases.
In the example $RI(Y,\tilde{Y}) = \frac{2+7}{2+7+4+2} = \frac{9}{15} = 0.60$.

### 2.8.2 Wallace Indices (WI)

As the Wallace Index [37] is an asymmetric index there are two indices resulting in these circumstances. Again there are to possibilities two calculate each of the Wallace Indices with $n_{k\tilde{k}}$ representing the number of samples that are in cluster $Y_k$ in method one and in cluster $\tilde{Y}_{\tilde{k}}$ using the other method. These values are set in a $K \times \tilde{K}$ matrix like the following (filled with values from the example) with $a_i$ representing the sums of column in the $WI_{10}$ (2, 3, 1 in the example) case and $a_j$ the sums of rows in the $WI_{01}$ (3, 3 in the example) case:

$$\begin{array}{ll} \begin{pmatrix} 2 & 1 & 0 \\ 0 & 2 & 1 \end{pmatrix} & \begin{array}{c} 3 \\ 3 \end{array} \\ \begin{array}{ccc} \;\;2 & \;\;3 & \;\;1 \end{array} & \end{array}$$

With this information one is able to calculate the Wallace Indices as in the following equations:

$$WI_{10}(Y,\tilde{Y}) = \frac{\sum_i \sum_j m_{ij} \cdot (m_{ij} - 1)}{\sum_i a_i \cdot (a_i - 1)} = \frac{n_{11}}{n_{11} + n_{10}} \tag{2.12}$$

$$WI_{01}(Y,\tilde{Y}) = \frac{\sum_i \sum_j m_{ij} \cdot (m_{ij} - 1)}{\sum_j a_j \cdot (a_j - 1)} = \frac{n_{11}}{n_{11} + n_{01}} \tag{2.13}$$

In the example one can calculate $WI_{10}(Y,\tilde{Y}) = \frac{2\cdot1+1\cdot0+0\cdot(-1)+0\cdot(-1)+2\cdot1+1\cdot0}{3\cdot2+2\cdot1} = \frac{4}{8} = \frac{2}{2+2} = 0.50$ and $WI_{01}(Y,\tilde{Y}) = \frac{2\cdot1+1\cdot0+0\cdot(-1)+0\cdot(-1)+2\cdot1+1\cdot0}{3\cdot2+3\cdot2} = \frac{4}{12} = \frac{2}{2+4} \approx 0.33$.

### 2.8.3 Fowlkes and Mallows Index (FMI)

The third introduced index is the Fowlkes and Mallow Index [9] that one can calculate as the geometric mean of both of the Wallace Indices:

$$FMI(Y,\tilde{Y}) = \frac{\sum_i \sum_j m_{ij} \cdot (m_{ij} - 1)}{\sqrt{(\sum_i a_i \cdot (a_i - 1)) \cdot (\sum_j a_j \cdot (a_j - 1))}} = \frac{n_{11}}{\sqrt{(n_{11} + n_{10}) \cdot (n_{11} + n_{01})}} \tag{2.14}$$

As the result for the small example the index is
$FMI(Y,\tilde{Y}) = \frac{2\cdot1+1\cdot0+0\cdot(-1)+0\cdot(-1)+2\cdot1+1\cdot0}{\sqrt{(3\cdot2+3\cdot2)\cdot(3\cdot2+2\cdot1)}} = \frac{4}{\sqrt{96}} = \frac{2}{\sqrt{(2+2)\cdot(2+4)}} \approx 0.41$.

### 2.8.4 Jaccard Index (JI)

A well known way to compare two clustered data sets is by the Jaccard Index [18] that is the fraction of the amount of similar elements divided by the size of the union of the two sets:

$$JI(Y, \tilde{Y}) = \frac{|Y \cap \tilde{Y}|}{|Y \cup \tilde{Y}|} = \frac{n_{11}}{n_{11} + n_{10} + n_{01}} \tag{2.15}$$

To calculate this index it turned out that $JI(Y, \tilde{Y}) = \frac{2}{2+4+2} = 0.25$.

### 2.8.5 Choosing the best index

As it is shown in the literature all indices are within the interval $[0, 1]$ and values closer to one show a larger similarity between the two compared sets and a value of zero indicates that there is no similarity of the sets at all. It is always possible to compare just two sets with each other at a time.
For the comparison of the results of iCluster it is important to look at the similarity of two sets of results for the same data but with a different number of $k$ clusters within each set (case (i)) and to compare two sets from different data with the same number of clusters $k$ (case (ii)).

**Case (i)**
The best index for that case is the Wallace Index $WI_{10}$ from Section 2.8.2 as this index is looking especially at the sum of columns and not at the sum of rows. The sum of columns is a good way to include the value of different $k$ in the index. This is equal to exclude s$n_{01}$ that stands for the amount of attributes where the sets are different in the first method and equal in the second. This value ($n_{01}$) should be small or zero anyway as $k$ increases and new clusters are formed from existing clusters and no samples should move to another existing cluster. Therefore, $WI_{01}$ is not a useful index in our case as well as the Fowlkes and Mallows Index that is the geometric mean of the two Wallace Indices. With $WI_{10}$ the index just take $n_{11}$ and $n_{10}$ into account and $WI_{10} = 1$, even if a cluster splits into two new clusters, as long as there are no movements between the clusters. If one wants to add information about $n_{00}$ the Rand Index would also be a good choice. It is the second preferred index for (i). The Jaccard Index is also not that good for this case as it does not use the preferred values.

**Case (ii)**
For this case it is more important to use information about every of the four values. Therefore, one should use the Rand Index from Section 2.8.1 in the analysis of two sets created for the same $k$ with different data. The Wallace Indices are

also a good choice and as the $FMI$ is the geometric mean of both this one is even better and therefore choice number two after the Rand Index. The Jaccard Index is again not the best choice to compare clustering results as it was invented to compare the similarity of sets of objects.

## 2.9 Survival Analysis

Survival Analysis is an important method for checking the results of clustering in biological and medical problems. Before applying the methods for clusters of iCluster, one should be informed of the methods that will be used. In this section, the main methods will be introduced, such as the Kaplan-Meier-Estimator and the (Mantel-Haenszel) Log-Rank-Test. The most important parameter is the survival time, denoted with $S(t)$. This is the period from the beginning of the treatment until death or recurrence. The more general name for survival analysis is event analysis. More detailed information about this section can also be found in Schumacher and Schulgen [26][4].

### 2.9.1 Anomaly of event analysis

For event analysis, one has to deal with some special anomalies that are not subject in another way of looking at these criteria. The first one is that the time of an event is defined as the time from the starting point until the entering of an event. In randomized therapy studies, this is easy to determine: the starting point is just the point of randomization. For observation studies this has to be the date of the diagnostics. It is possible to determine the end point by observing the patient until the event of death or relapse/metastasis. Movement to the remission as the event makes it much more difficult to determine the event date. The second big problem for event analysis is that event dates are sometimes not observed completely. Incomplete means here, that the event does not occur until the end of observation, therefore so called (right) censored data results. Another problem that leads to incomplete data are so called drop-outs. Drop-outs are patients that leave the study for some reason. Those drop-outs, also censored data, are not easy to deal with because one would not be able to determine the survival time, even within an infinite time window. This is problematic, since one has no information about the reasons why this is happening. Types of drop-outs include the patient

[4] This literature is in German, but there are also several English books available on this topic.

feeling too sick to visit the doctor or feeling so good that he/she no longer goes to doctor. High standards in data quality are required to avoid biased results.

### 2.9.2 Kaplan-Meier-Estimator

As defined above the survival time is denoted as $S(t)$. This time is the probability to survive time $t$ without having any event until $t$ such that $S(t) = \mathbb{P}(T > t)$, where $T$ is the survival or event time. The estimator for $S(t)$ is the so called Kaplan-Meier-Estimator [19]. For this estimator, one has to order the event dates as in the following Table 2.2 for the $m$ different events:

| Date of the event | Number under risk | Number events |
|---|---|---|
| $t_1$ | $n_1$ | $d_1$ |
| $t_2$ | $n_2$ | $d_2$ |
| $\vdots$ | $\vdots$ | $\vdots$ |
| $t_m$ | $n_m$ | $d_m$ |

Table 2.2: Ordering of the event dates

The probability of surviving time $t$ for $t = t_1$ can be estimated by the number of patients surviving without any event until $t_1$ divided by the number of patients under risk just before $t_1$:

$$\hat{S}(t_1) = \frac{n_1 - d_1}{n_1} \tag{2.16}$$

Therefore, one can say that $\frac{n_2-d_2}{n_2}$ is the conditional probability to survive time $t_2$ without any event if there was already no event at time $t_1$ before:

$$\hat{S}(t_2) = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2} \tag{2.17}$$

The Kaplan-Meier-Estimator is hence the probability to survive a time $t$ after randomization without any event.

$$\hat{S}(t) = \frac{n_1 - d_1}{n_1} \cdot \frac{n_2 - d_2}{n_2} \cdot_{\dots} \cdot \frac{n_i - d_i}{n_i} = \prod_{i:t_i<t} \left(1 - \frac{d_i}{n_i}\right) \tag{2.18}$$

with time $t$ between $t_i$ and $t_{i+1}$.

This estimator enables one to show a curve for the survival analysis that is affected by random variations. The variations can be described with the Greenwood formula [10] that displays it as a standard error ($se$). For a $t$ between $t_i$ and $t_{i+1}$ this standard error is:

$$se(\hat{S}(t)) = \hat{S}(t) \cdot \sqrt{\sum_{i:t_i<t} \frac{d_i}{n_i \cdot (n_i - d_i)}} \tag{2.19}$$

With this standard error it is then possible to create a $100(1-\alpha)\%$ (pointwise) confidence interval. For this $u_{1-\frac{\alpha}{2}}$ is the $(1-\frac{\alpha}{2})$-quantile of the standard normal distribution and therefor the confidence interval for the event free survival probability is:

$$[\hat{S}(t) \pm u_{1-\frac{\alpha}{2}} \cdot se(\hat{S}(t)] \tag{2.20}$$

A simultaneous confidence interval, introduced by Hall and Wellner [12], is appropriate if one wants to regard the survival function as a function over the time. This interval is wider than the interval introduced before.

### 2.9.3 Mantel-Haenszel-Test/Log-Rank-Test

The Log-Rank-Test, which is also sometimes also called Mantel-Haenszel-Test, is a method to test if there is a difference between two or more survival curves. The survival curves are estimated with the Kaplan-Meier-Estimator in Section 2.9.2. Different survival curves occur if there is any difference in the survival of different groups of treatment or in our case in different clusters created by iCluster.

| | Number events at $t_i$ | Number event free samples at $t_i$ | Number samples under risk just before $t_i$ |
|---|---|---|---|
| Cluster 1 | $d_{1i}$ | $n_{1i} - d_{1i}$ | $n_{1i}$ |
| Cluster 2 | $d_{2i}$ | $n_{2i} - d_{2i}$ | $n_{2i}$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| Cluster q | $d_{qi}$ | $n_{qi} - d_{qi}$ | $n_{qi}$ |
| Together | $d_i$ | $n_i - d_i$ | $n_i$ |

Table 2.3: Table needed for construction of the test statistic

If the marginal totals are fixed one should find no difference under the null hypothesis between $d_{1i}$ and the expected number of events

$$\mathbb{E}(D_{1t_i}) = \frac{d_i \cdot n_{ji}}{n_i} \tag{2.21}$$

for time $t_i$ in this cluster. With a variance for the number of events at this time for two clusters:

$$\mathbb{V}(D_{jt_i}) = \frac{d_i \cdot (n_i - d_i) \cdot n_{1i} \cdot n_{2i}}{n_i^2 \cdot (n_i - 1)}. \tag{2.22}$$

This yields to the test statistic $T_{LR}$ to compare two ore more survival analysis distributions:

$$T_{LR} = \frac{\left(\sum_{t_i} w_{t_i} \cdot [d_{1t_i} - \mathbb{E}(D_{1t_i})]\right)^2}{\sum_{t_i} w_{t_i}^2 \cdot \mathbb{V}(D_{1t_i})}, \tag{2.23}$$

where $w_{t_i}$ is the weight for the $i-$th event point. To get the Mantel-Haenszel-Test, the weight is set to one: $w_{t_i} = 1$ for all $t_i$. This is also known as the Mantel-Haenszel version of the Log-Rank-Test. The normal Log-Rank-Test uses another variance and is more conservative than the Mantel-Haenszel version of the Log-Rank-Test. For more details see Schumacher and Schulgen [26].

The test statistic $T_{LR}$ is $\chi^2$ distributed with $k - 1$ degrees of freedom. With this information one is able to get $p$-values to test the null hypothesis, if there is a difference between two survival curves. Dealing with more than two survival curves is also possible.

In the R package *survival* [34] the variance function is calculated in two different ways. The default method is different and the $p$-values are more conservative. The test statistics and $p$-values calculated for this study in Section 5.3 are calculated as described in this section. Using the other method does not change the results significantly. The default variance function in this package is just the expected number for time $t_i$ in this cluster: $\mathbb{E}(D_{qt_i})$, instead of using the variance function 2.22.

# CHAPTER 3

# Preparation of the Sarcoma data set

## 3.1 About the data types

In this thesis Gene Expression (GE) and Copy Number (CN) data is used. To understand these data sets one has to understand the biological background.
The DeoxyriboNucleic Acid (DNA) is a double-helix molecular that exits in all living organisms and therefore in all humans. It contains all the genetic information about the creature. The DNA contains genes that contain all information needed for the creation of the single-helix RiboNucleic Acid (RNA). A special group of the RNAs, the messengerRNAs contain all the necessary information to create the proteins within the group of coding RNA. This is also known as the Central Dogma of molecular biology. Non-coding RNA (RNA that does not code for protein) is divided into two classes: long and short. The major class of short noncoding RNAs that has been intensively studied is microRNAs.
There are two copies for each gene in a human. Mutations on one or both copies can occur that may lead to cancer. In addition, one or both gene copies can become deleted (deletion event) or additional copies can be generated (amplification). In cancer one speaks about deleted or amplified (oncogenic) genes or regions. The Copy Number data is measured by Copy Number arrays or by DNA sequencing.

## 3.2 Raw Data

The data set from Dr Singer lab is used in this thesis and is clustered using iCluster. This data set consists of samples of sarcoma patients. There was an Excel file PMFH and MXF samples with CGH and U133A 06.07.11.xlsx provided with all the samples. This file has the following columns (first three rows included):

**Basename2:**

- PD_MXF2516_slide350_S01_CGH_107_Sep09
- PD_MFH2938_slide634_S01_CGH_107_Sep09
- PD_MFH0355_slide387_S01_CGH-v4_10_27Aug08

**CELLFILE:**

- PD_U133A_MFH2516
- PD_U133A_MFH2938
- PD_U133A_MFH0355

**SID:**

- MXF2516
- MXF2938
- MXF0355

**Type:**

- MXF
- MXF
- MXF

With this Excel file one is able to organize the Gene Expression and Copy Number data such that only samples, that occur in both data sets, are used in the final data matrices.

### 3.2.1 mRNA Gene Expression data

In the raw data lee_U133A_GCRMA.Rdata there were 22215 genes and 102 samples. The 22215 × 102 data matrix has the form:

$$\begin{array}{c|ccc} & MFH2516 & MFH2938 & MFH0355 \\ \hline 1007\_s\_at & 5.722348 & 6.321752 & 5.637633 \\ 1053\_at & 5.611733 & 6.493168 & 6.275949 \\ 117\_at & 4.831413 & 4.793718 & 5.363729 \end{array}$$

In this submatrix 1007_s_at, 1053_at and 117_at are the genes/probes and MFH2516, MFH2938 and MFH0355 (shorten for PD_U133A_MFH2516, PD_U133A_MFH2938 and PD_U133A_MFH0355) are the samples with genes/ probes in rows and sample names in columns.
With the Excel file it is possible to delete those samples that are not in that file and that are not in the Copy Number data.

### 3.2.2 DNA Copy Number data

A 18142 × 82 matrix of the row data leeA_MFH-lesions.Rdata is available for Copy Number data. There are 18142 probe levels for 82 samples and the matrix looks like the following subset, with shorten column names:

$$\begin{array}{r c} & \begin{array}{cc} PD\_MFH2938 & PD\_PMFH876 \end{array} \\ \begin{array}{r} 1q : 16367892 - 16451598 \\ 1q : 16453588 - 16469444 \\ 1q : 16470790 - 16490356 \end{array} & \begin{pmatrix} 0.93079973 & -0.00007654 \\ 0.15183170 & -0.00007654 \\ -0.04951276 & -1.00000000 \end{pmatrix} \end{array}$$

In this $1q\ldots$ stands for regions on chromosome arm $1q$ and PD_MFH2938 and PD_PMFH876 (shorten for PD_MFH2938_slide634_S01_CGH_107_Sep09 and PD_PMFH876_slide680_S01_CGH_107_Sep09) are the samples. One has also to deal with the above mentioned Excel sheet to exclude those samples that are not in the Gene Expression data.

### 3.2.3 Final data sets

After applying the Excel sheet there were 37 samples left that occur in both, Gene Expression and Copy Number, data sets. For iCluster one needs to transpose those matrices to run the package. The final data sets before creating smaller subsets for calculation the two data sets are:

- A 37 × 22215 Gene Expression data matrix.

- A 37 × 18142 Copy Number data matrix.

## 3.3 Creating subsets for Calculations

There are many reasons why one is interested in smaller subsets of the original data sets described in Section 3.2.3. There are for example genes that have a standard deviation of zero and therefore the data matrix does not have not a full rank, what is required in the $k$-Means-Algorithm. In other genes there is just noise within the characteristics of a gene. More detailed information about the Gene Expression subsets can be found in 3.3.1. Dealing with Copy Number data is different as those values are in the intervall $[-1, +1]$. There are a lot of possibilities how to deal with this data. Three of those are applied simultaneously with different thresholds/-values to get smaller subsets of the Copy Number data. All these methods are required to work with the data that contains the important information one needs for clustering the samples. It is also not possible to run iCluster in a reasonable time, even on big servers, with such big data sets. To run iCluster and choose the best subsets one has to run it several times for different clustering numbers $k$ (e.g. $k = 2, 3, \ldots, 6$) and different thresholds $\lambda$ (e.g. $\lambda = 0.00, 0.01, 0.02, \ldots, 0.20$). For each of these combinations it is an issue how the dimension of the data set is. In Tables 3.1 and 3.2 one sees the complete time for 37 samples and different subsets of genes and probes included. Phos and Ray are two different servers where it was calculated. Note: different numbers of the convergence number are not included to make this table easier. In this example the first columns of the sd_GE_11 Gene Expression data (see Section 3.3.1.2 for details) and the first columns of the SDK10 Copy Number were used as well as $k = 2, \lambda = 0.03$ and a maximum of $n = 100$ iterations for the convergence. In Table 3.3 the same data is used again by using the first 2000 genes/2000 probes of the data. $\lambda = 0.03$ is also used for all calculations for different $k$ as the convergence rate is different for different $k$. In this study the convergence number is included in the table, because it plays a big role for different $k$. It turns out, that for a higher $k$ the calculation within iCluster needs more time what is not surprising at all.

| Dimension of Gene Expression data | 3000 | 3000 | 3000 | 2500 | 2500 | 2500 |
|---|---|---|---|---|---|---|
| Dimension of Copy Number data | 3000 | 2500 | 2000 | 2500 | 2000 | 1500 |
| Time needed on Phos (in minutes) | 101 | 82 | 69 | 54 | 44 | 33 |
| Time needed on Ray (in minutes) | 107 | 86 | 70 | 58 | 48 | 37 |

Table 3.1: Calculation time for one subtype $k$ and one $\lambda$ as a tuning parameter - part 1

| Dimension of Gene Expression data | 2000 | 2000 | 2000 | 1500 | 1500 |
|---|---|---|---|---|---|
| Dimension of Copy Number data | 2000 | 1500 | 1000 | 1500 | 1000 |
| Time needed on Phos (in minutes) | 31 | 22 | 15 | 14 | 9 |
| Time needed on Ray (in minutes) | 34 | 24 | 16 | 15 | 10 |

Table 3.2: Calculation time for one subtype $k$ and one $\lambda$ as a tuning parameter - part 2

| Number of subtypes $k$ | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|
| Convergence number | 35 | 72 | 70 | 168 | 74 |
| Time needed on Phos (in minutes) | 31 | 79 | 113 | 270 | 142 |
| Time needed on Ray (in minutes) | 31 | 77 | 94 | 288 | 138 |

Table 3.3: Calculation time for $37 \times 2000$ data matrices of both data types for $\lambda = 0.03$ for different $k$

### 3.3.1 Gene Expression subsets

The histogram of the standard deviation of the genes, see Figure 3.1, shows that there are 5456 genes with standard deviation of smaller than 0.1 (red line). 2699 out of the 5456 genes have a standard deviation of 0. Around one third (7171 of 22215) of the genes have a standard deviation smaller than 0.3 (blue line) that is defined as the cutoff for noise only.

The higher the standard deviation of the genes the higher the information is in the genes.

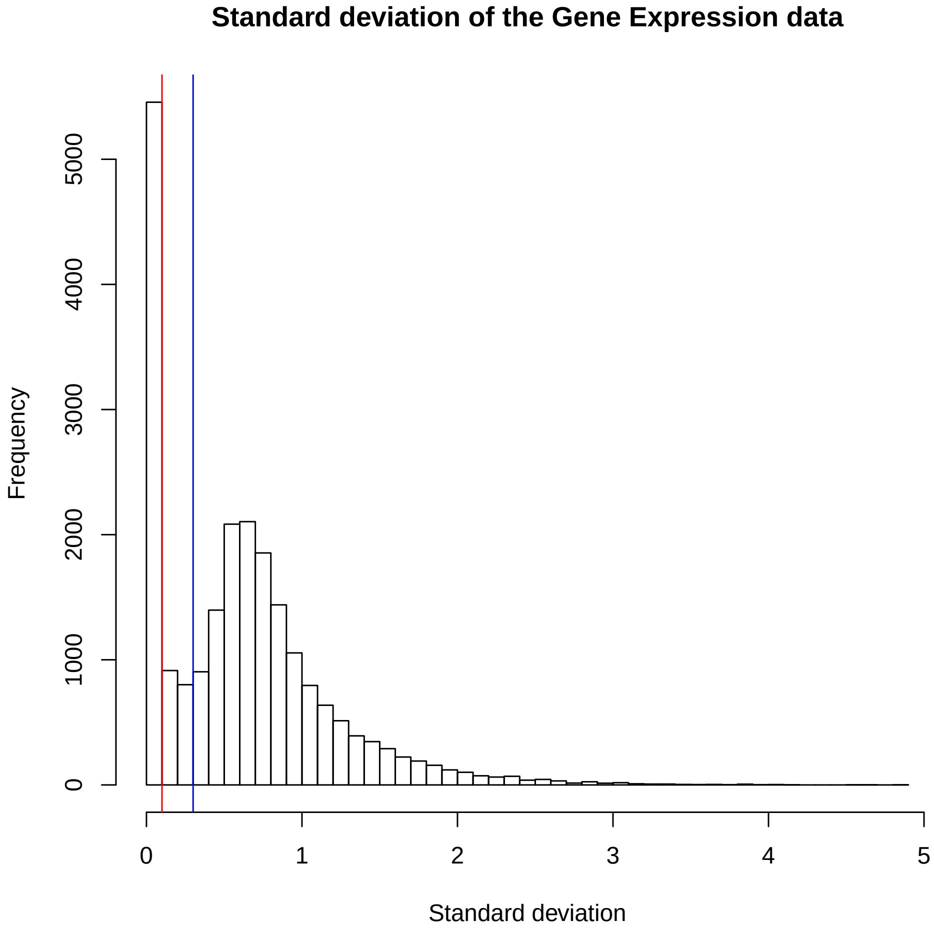


Figure 3.1: Standard deviation of the genes

The next histogram, Figure 3.2, again shows the standard deviation of the genes for Gene Expression data excluding those genes having a standard deviation of 0. It contains 19516 genes for the 37 samples. There are three cutoffs defined. The middle one (red line) is the cutoff for the main subset of the data for Gene Expression data. This cutoff (1.3) is one standard deviation above the noise level of 0.3. It is also around the mean plus the standard deviation of all genes, excluding genes with standard deviation equal to zero. This sum is equal to 1.242 and therefore close to the chosen threshold of 1.3.

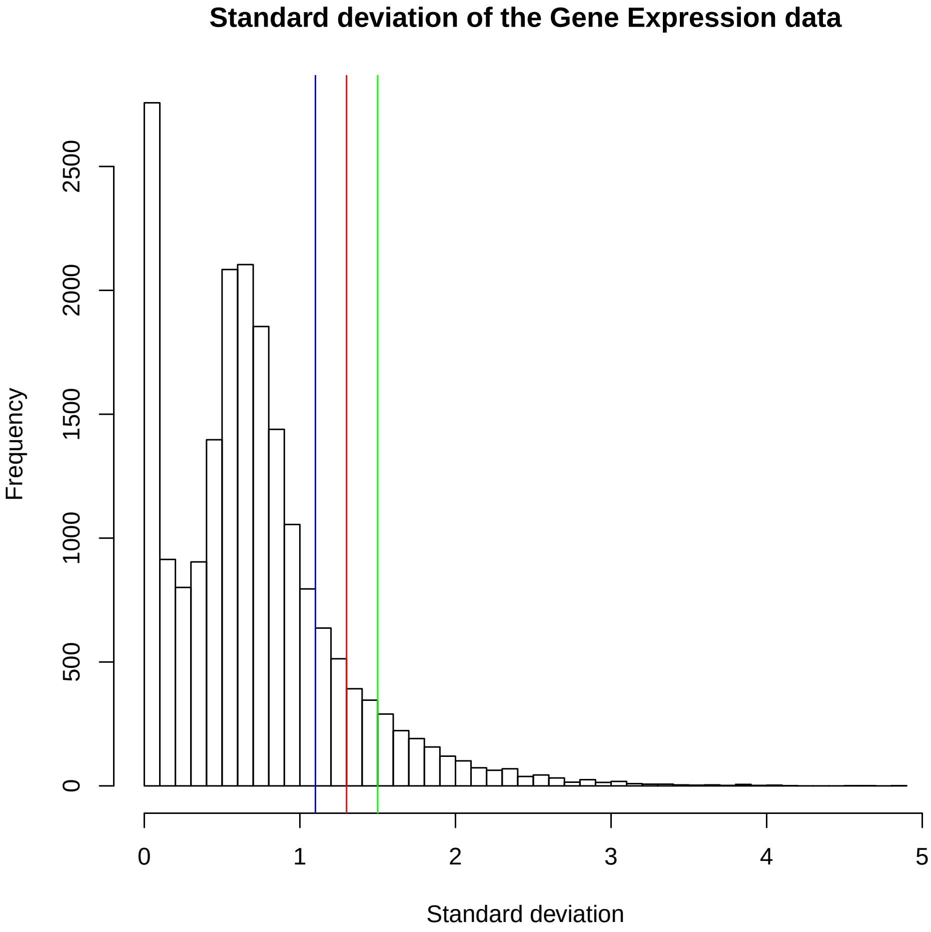


Figure 3.2: Standard deviation of the genes without genes with zero standard deviation

### 3.3.1.1 Subset 1 - sd_GE_13

The subset of Gene Expression data with standard deviation above 1.3 (red line) is then defined as *sd_GE_13*. This data set contains 2262 of the 22215 genes with high standard deviation.

### 3.3.1.2 Subset 2 - sd_GE_11

If one does not want to exclude that many genes it is possible to just exclude data with a standard deviation of below 1.1 (blue line). The subset then contains 3412 genes.

#### 3.3.1.3 Subset 3 - sd_GE_15

To exclude even more genes comparing to Subset 1 one may set the cutoff line to 1.5 (green line). In this case the data set is obviously smaller and contains 1524 genes.

#### 3.3.1.4 Subset 4 - Random

First, all genes with a standard deviation of zero are excluded. As the second and last step 10% of the genes were randomly picked and declared as the subset. A random subset in Gene Expression data is created as well as a random subset in the Copy Number data to study wether it is useful to create all the subsets with the rules or if it is also good to randomly pick genes/probes out of the original data to get a smaller subset of data.

#### 3.3.1.5 Summary of Gene Expression data subsets

Table 3.4 shows all the different subsets of the GE data.

| Subset | Name | Cutoff | Included genes | Ratio |
|---|---|---|---|---|
| 1 | sd_GE_13 | sd = 1.3 | 2262 of 22215 | 10.18% |
| 2 | sd_GE_11 | sd = 1.1 | 3412 of 22215 | 15.36% |
| 3 | sd_GE_15 | sd = 1.5 | 1524 of 22215 | 6.86% |
| 4 | Random | 10% of sd $>$ 0 | 1951 of 22215 | 8.78% |

Table 3.4: Summary of all different subsets of Gene Expression data

### 3.3.2 Copy Number subsets

For Copy Number data there are three different methods to do the shrinkage of the original data set to deal with. Each of them is applied to each subset of the data with different thresholds. Altogether there are four different tuning parameters in the function to get different subsets. For each parameter different values are used. There is a total of eight (nine) different subsets of Copy Number data.

#### 3.3.2.1 Subset 1 - ME537 as 'normal' data set

In Copy Number data it makes no sense if the absolute mean of the data is around zero. A zero mean occurs when around half of the data, that has non-zero samples, is $\sim -1$ and the other half is $\sim +1$. Genes with smaller absolute mean than a threshold are deleted. The default value was set to $\frac{5}{37} \approx 0.135$ as there are 37 samples.
For the 'normal' data set a threshold of $\frac{5}{37}$ is used. This is colored red in Figure 3.3. With this threshold for the parameter *mean_CN_exclude* in the function and default parameters (defined in the following pages) the subset contains 1560 genes of the Copy Number data. By applying only this reduction it would contain 7949 of the 18142 genes.

#### 3.3.2.2 Subset 2 - ME437

The next subset is created with all default parameters and a value of $\frac{4}{37} \approx 0.108$ for the parameter *mean_CN_exclude*. This is the blue line in Figure 3.3. The final data set contains 2162 genes. Before other tuning parameters are used the data set would be a $37 \times 9322$ matrix.

#### 3.3.2.3 Subset 3 - ME637

In Figure 3.3 the green line indicates subset 3. There are 6898 genes on the right of that line. 1509 of these genes are in the final data set after applying the other methods with their default parameters.

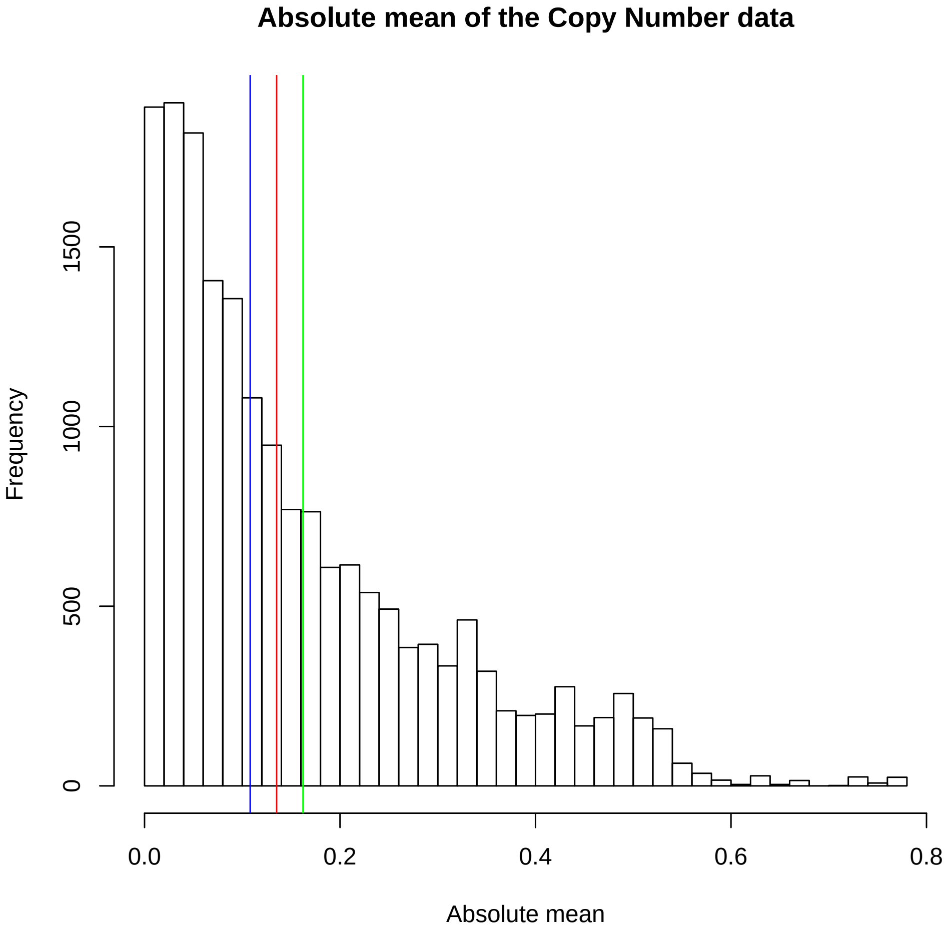


Figure 3.3: Mean of the genes

#### 3.3.2.4 Subset 4 - MSV8 and STD5 as ‘normal’ data set

With that method the largest of the absolute values for each gene are set to 1, the rest to 0. The default value of this is set to 0.8. This variable is named *cn_mat_set_val* in the algorithm to create subsets of the data. After applying this to the data matrix one has to calculate the sum for each gene. The lowest y% of the genes are then deleted. As a default parameter $st10\_CN\_delete0$ is set to 50%.

The matrix looks like the following with z equal to the gene amount in the Copy Number data before that step:

$$\begin{array}{c|cccc|}
 & Gene_1 & Gene_2 & \dots & Gene_z \\
Sample_1 & 1 & 0 & \dots & 1 \\
Sample_2 & 1 & 1 & \dots & 1 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
Sample_{37} & 0 & 1 & \dots & 0
\end{array}$$

After excluding the genes with lowest sum of 1 the data matrix is again the matrix with Copy Number values and not the 1/0 matrix. As described above there are 1560 genes in the 'normal' data set for the default values. The genes (3121) kept after that step are the genes right of the red line in Figure 3.4.

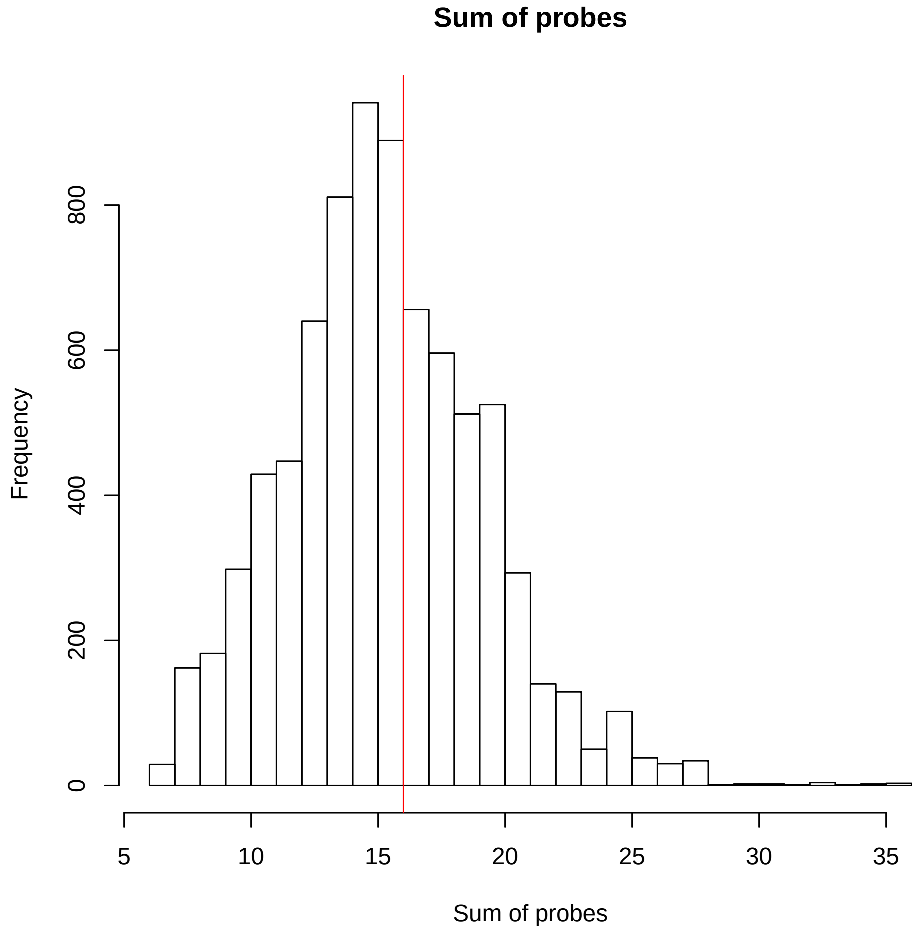


Figure 3.4: Sum of genes for $cn\_mat\_set\_val = 0.8$

#### 3.3.2.5 Subset 5 - MSV7

For this subset the variable *cn_mat_set_val* is set to 0.7. The 3497 genes kept after that step are right of the red line in Figure 3.5. After the third step it contains 1749 genes.

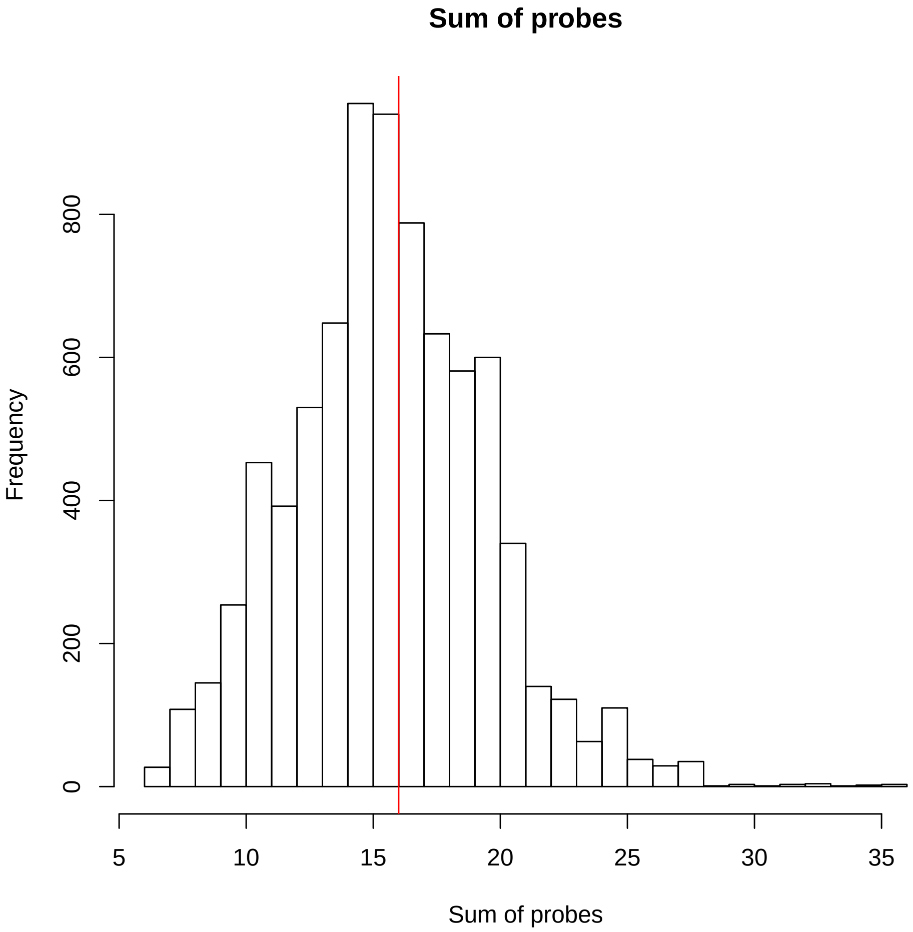


Figure 3.5: Sum of genes for $cn\_mat\_set\_val = 0.7$

#### 3.3.2.6 Subset 6 - MSV9

Here equal calculations are done. One sets $cn\_mat\_set\_val = 0.9$. The result can be found in Figure 3.6 right of the green line. There are 3683 genes left after that step and 1842 in the final subset of the data.

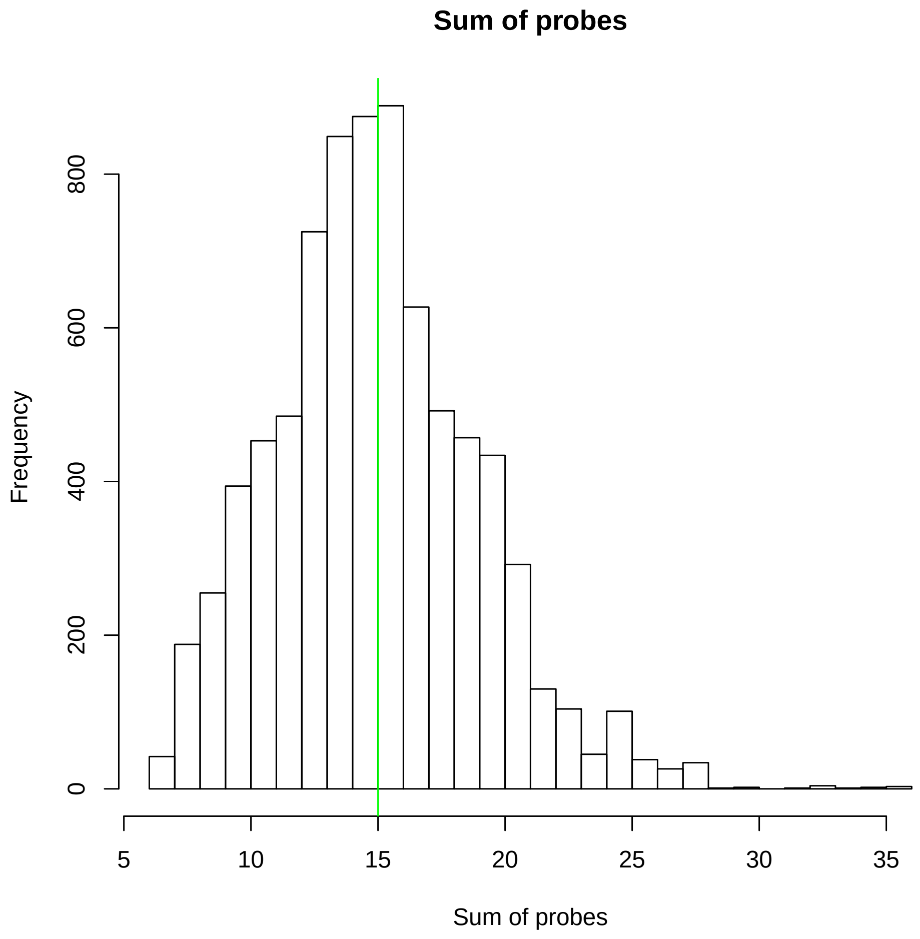


Figure 3.6: Sum of genes for $cn\_mat\_set\_val = 0.9$

#### 3.3.2.7 Subset 7 - STD4

Comparing to STD5 just 40% ($st10\_CN\_delete0 = 0.4$) instead of 50% of the genes with lowest sum of the genes is deleted so that there were 4010 genes left after that step. For this difference also see Figure 3.7, the red line. In the final data set for Copy Number data there are then 2010 genes included.

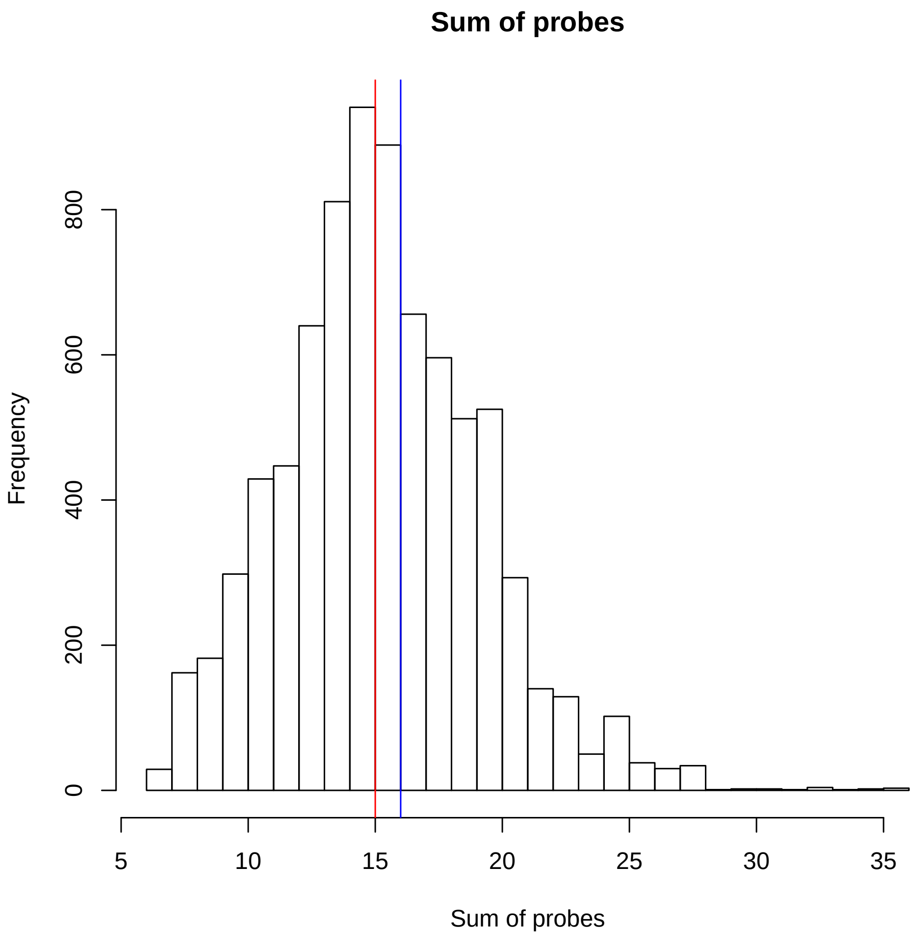


Figure 3.7: $cn\_mat\_set\_val = 0.8$ with different $st10\_CN\_delete0$

#### 3.3.2.8 Subset 8 - STD6

There was no further investigation with that subset as it is exactly the same data subset as STD5, as the sums are equal for 40% and 50% for that particular case.

#### 3.3.2.9 Subset 9 - SDK5 as 'normal' data set

As the last and third step one keeps x% of the genes with lowest standard deviation over the (absolute) mean (variable name $cn\_sd\_mean$). In the default scenario the variable is equal to 50% and produces finally the 'normal' data set with 1560 genes. One can see the histogram in Figure 3.8 with the red line equaling 1.500.

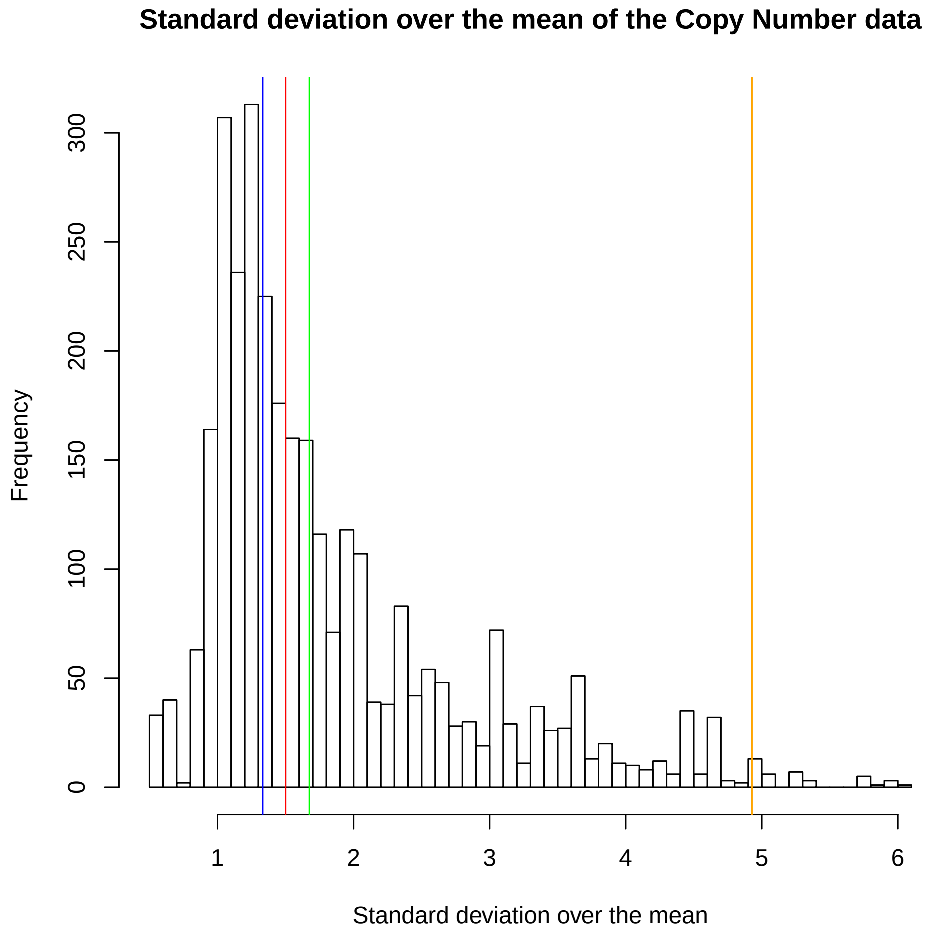


Figure 3.8: Standard deviation over the (absolute) mean for Copy Number data

### 3.3.2.10 Subset 10 - SDK4

The blue line in Figure 3.8 indicates the value for *cn_sd_mean* = 0.4. This value is equal to the standard deviation over the absolute mean of 1.332. This subset contains 1248 genes.

### 3.3.2.11 Subset 11 - SDK6

In this subset 60% of the data with lowest standard deviation over the (absolute) mean is kept. This is equal to a subset of 1872 genes and the green line in Figure 3.8. The value is equal to 1.675.

#### 3.3.2.12 Subset 12 - SDK10

This special case is indicated with the orange line in Figure 3.8 to get this bigger subset. One just deletes the highest 1% of the standard deviation above the (absolute) mean, that is equal to 4.928. 3094 genes are in this case in the subset.

#### 3.3.2.13 Subset 13 - Random

For this subset one just picks randomly 10% of the 18142 genes that is 1814 genes and declares this as the new subset.

#### 3.3.2.14 Summary of Copy Number data subsets

A summary of the CN data subset is Table 3.5.

| Subset | Name | mean_CN_exclude | cn_mat_set_val | st10_CN_delete0 | sd_mean_keep | Included genes | Ratio |
|---|---|---|---|---|---|---|---|
|  | Normal | $\frac{5}{37}$ | 0.8 | 0.5 | 0.5 | 1560 of 18142 | 8.60% |
| 1 | ME537 |  |  |  |  |  |  |
| 4 | MSV8 |  |  |  |  |  |  |
| 4 | STD5 |  |  |  |  |  |  |
| 9 | SDK5 |  |  |  |  |  |  |
| 2 | ME437 | $\frac{4}{37}$ | 0.8 | 0.5 | 0.5 | 2162 of 18142 | 11.92% |
| 3 | ME637 | $\frac{6}{37}$ | 0.8 | 0.5 | 0.5 | 1509 of 18142 | 8.32% |
| 5 | MSV7 | $\frac{5}{37}$ | 0.7 | 0.5 | 0.5 | 1749 of 18142 | 9.64% |
| 6 | MSV9 | $\frac{5}{37}$ | 0.9 | 0.5 | 0.5 | 1842 of 18142 | 10.15% |
| 7 | STD4 | $\frac{5}{37}$ | 0.8 | 0.4 | 0.5 | 2010 of 18142 | 11.08% |
| 8 | STD6 | $\frac{5}{37}$ | 0.8 | 0.6 | 0.5 | 1560 of 18142 | 8.60% |
| 10 | SDK4 | $\frac{5}{37}$ | 0.8 | 0.5 | 0.4 | 1248 of 18142 | 6.88% |
| 11 | SDK6 | $\frac{5}{37}$ | 0.8 | 0.5 | 0.6 | 1872 of 18142 | 10.32% |
| 12 | SDK10 | $\frac{5}{37}$ | 0.8 | 0.5 | 0.99 | 3094 of 18142 | 17.05% |
| 13 | Random | Randomly picking 10% of the genes |  |  |  | 1814 of 18142 | 10.00% |

Table 3.5: Summary of all different subsets of Copy Number data

# CHAPTER 4

# Run iCluster

After creating all the data sets described in Chapter 3 one is able to use the iCluster package in R. This chapter is structured as the following. First, iCluster is introduced as a R package with all the possible settings of the package. After this the settings used in the analysis are set.

## 4.1 iCluster as a R package

The software is publicly available within Bioconductor[5]. The additional R package of Shen is also downloadable[6]. For this studies version 1.2.0 of iCluster and version 2.13.0 of R [24] is used. R was used within RStudio[7] with version 0.94.92.
Three major functions of iCluster are described in the next three subsections. The first one is *iCluster* and that is used to compute the results, additional functions are *compute.pod* and *plotiCluster* to compare and plot the results.

### 4.1.1 iCluster

Shen described this function in [27] as the following. ‘Given multiple genomic data types (e.g., Copy Number, Gene Expression, DNA methylation) measured in the same set of samples, iCluster fits a regularized latent variable model based clustering that generates an integrated cluster assignment based on joint inference across data types’. The function call has to look like that:

[5] http://www.r-project.org/
[6] http://cran.r-project.org/web/packages/iCluster/
[7] http://rstudio.org/

```
fit <- iCluster(datasets, k, lambda, scalar = FALSE, max.iter = 50, epsilon = 1e-3)
```

The arguments of the function are the following:

- **datasets** - m different data sets as $samples \times genomicfeatures$ matrices. For example Copy Number, Gene Expression, DNA Methylation, . . . . Each of the m matrices must contain the same samples.
- **k** - number of subtypes in the data
- **lambda** - a vector of length m for the penalty terms (Lasso)
- **scalar** - default is false, if it is true one assumes a scalar covariance matrix
- **max.iter** - maximum of iterations (Expectation-Maximization-Algorithm)
- **epsilon** - the convergence criterion (Expectation-Maximization-Algorithm)

Those elements are returned:

- **expZ** - relaxed cluster indicator matrix
- **W** - coefficient matrix
- **clusters** - a vector that shows to the samples' cluster membership
- **conv.rate** - convergence history until convergence or max.iter

### 4.1.2 compute.pod

This is 'a function to compute the proportion of deviation from the perfect block diagonal matrix' [27].

```
pod <- compute.pod(fit)
```

- **fit** - the result of the iCluster function in Section 4.1.1

One gets the following results

- **pod** - as described in Section 2.7 - the proportion of deviation from the perfect block diagonal matrix

### 4.1.3 plotiCluster

‘A function to generate cluster separability matrix plot’ as Shen assumes [27].

```
pod <- compute.pod(fit)
```

- **fit** - the iCluster object calculated in Section 4.1.1
- **label** - sample names

There are no returned values, but a plot of the clusters.

## 4.2 Selected settings

To find the best result with iCluster one has to run iCluster several times with different values for the number of clusters $k$ and different values for the threshold for the penalty term $\lambda$. This threshold parameter is a vector of length $m$, with $m$ representing the number of different data sets used for the analysis. In every study in the paper Shen et al. [28] always used the same value for all $m$ value of $\lambda$. For this reason the values for all $\lambda$ are the same.

The values ranged from $\lambda = 0.00$ to $\lambda = 0.20$ in steps of 0.01, such that in total 21 different vectors for lambda were calculated for each data set and for each $k$.

As there were just 37 samples in this study, each cluster should not become to small and still contain enough samples. For this reason $k$ was set to $k = 2, 3, 4, 5, 6$. Note that $k = 5$ and $k = 6$ are calculated, even if the average size of each cluster would then just be 7.4 (6.167) samples and thus very small for survival analysis. The calculations were performed to validate the iCluster method and to show how iCluster works in general.

There are two more additional parameter settings within the iCluster package, as one can see in Section 4.1.1. The convergence criterion *epsilon* remained as default setting for all the analysis, such that $epsilon = 1e - 3$. The parameter *max.iter* was set to $max.iter = 100$ instead of using the default value of 50 to reach convergence more often, as the convergence rate is higher (for $k = 4, 5, 6$), see results in Appendix A. Altogether the settings were the following:

**datasets**

- 20110627_sd_GE_13
- 20110627_sd_GE_11
- 20110627_sd_GE_15
- 20110627_ME437
- 20110627_ME637
- 20110627_MSV7
- 20110627_MSV9
- 20110627_STD4
- 20110627_SDK4
- 20110627_SDK6
- 20110627_SDK10
- 20110707_Random

**k**

- 2
- 3
- 4
- 5
- 6

**lambda**

- 0.00
- 0.01
- ⋮
- 0.20

**max.iter**

- 100

**epsilon**

- 1e-3

# CHAPTER 5

# Results of iCluster

As shown in the previous Section 4.2 iCluster was run with many different settings for all the 11 (or 12 including the random subset) different data sets. To compare the results for the different values of lambda POD (Proportion Of Deviance from perfect block diagonal matrix) is used for each data set for each $k$. All the results for POD, the convergence rate and the convergence number can be found in the Appendix A.

## 5.1 Results for the data sets

Looking at all the different subsets for one specified number of clusters $k$ it turns out that there is a lambda for that the POD is small or even the smallest for all data sets used in that analysis. In the data set sd_GE_13 for $k = 4$ one can see in Figure 5.1 that for $\lambda = 0.13$ POD has its minimum and the values are small for neighboring thresholds ($\lambda = 0.14, 0.15$). Repeating the procedure for all the data set demonstrates that this $\lambda = 0.13$ is the best from all 21 possible $\lambda$ for that specific $k$.

After looking at all the different $k$ for all the data sets it turns out that there are thresholds for which the POD is the best in the way of being small as well as convergent in most of the cases. Sometimes it is not convergent for $conv.number = 100$, but if one looks at the $conv.rate$ it turns out that these should be convergent as well but did not reach the $conv.rate$ yet. There is again an example in Figure 5.2, when the $conv.rate = epsilon$ (R-package) $= 1e{-}3$ falls below the pre-definied (see Section 4.2) threshold for a $conv.number = 100$ smaller than the pre-defiend maximum steps. The epsilon is the default value, but the convergence number was

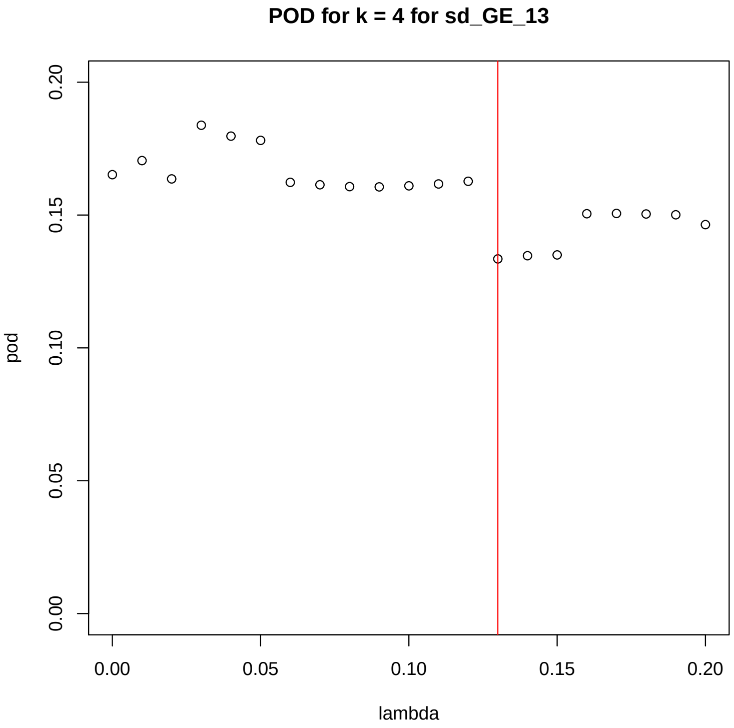


Figure 5.1: POD for different $\lambda$ for $k = 4$ within the sd_GE_13 data set

set to a value double than the default value, as there is convergence for this data sets between the pre-defined default value and the value chosen for this analysis. The result of the clusters with the POD is then looking like Figure 5.3.

The best thresholds for the different possibilities for $k$ are[8] for the different $k$ available in Figure 5.1.

| Number of clusters $k$ | best value for $\lambda$ |
|---|---|
| 2 | 0.03 |
| 3 | 0.20 |
| 4 | 0.13 |
| 5 | 0.05 |

Table 5.1: Best clusters for different $k$

[8] $\lambda = 0.00$ occurs more often as best POD for $k = 3$, but $\lambda = 0.20$ occurs more often as one of the smallest POD

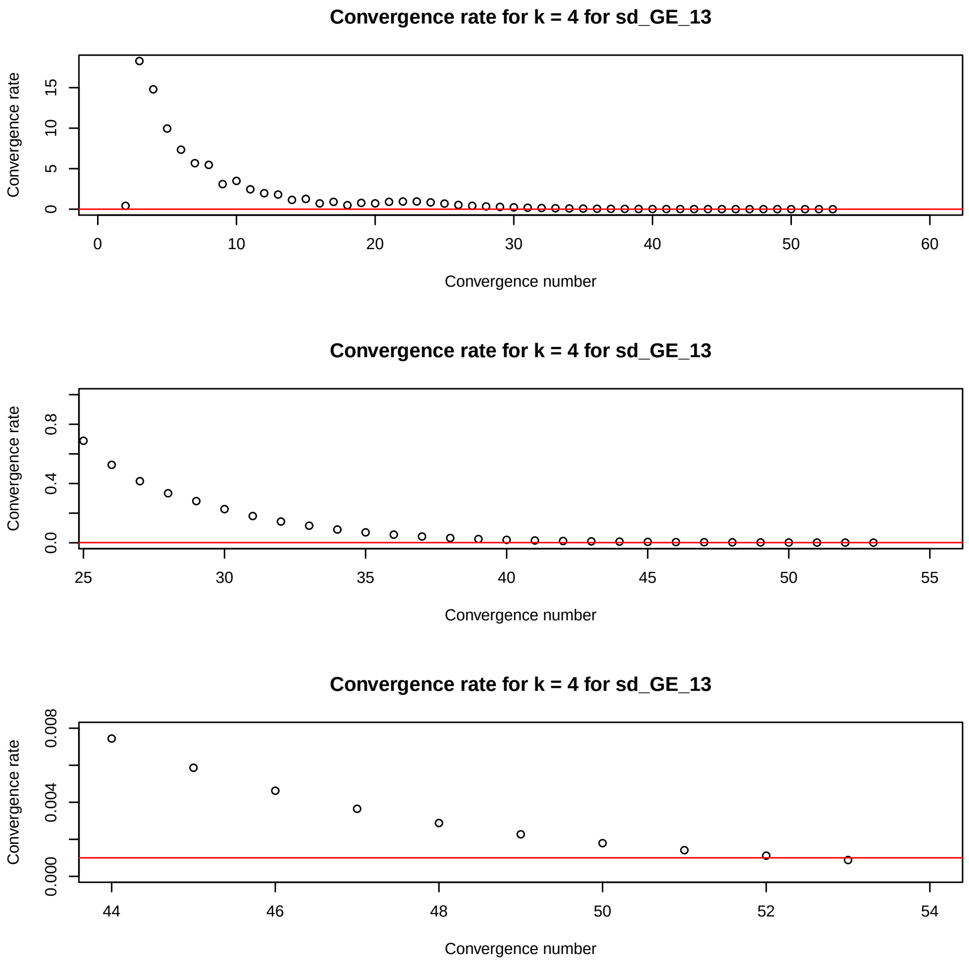


Figure 5.2: Convergence rate $\lambda = 0.13$ for $k = 4$ within the sd_GE_13 data set

For these best clusters it is possible to compare the data sets with the samples. It occurs that all the 11 data sets cluster the 37 samples in the same clusters. The minimum similarity number is that 9 out of 11 data sets cluster one sample in a particular cluster. This is a very good result as a lot of different data sets were used to run the algorithm. This result can be found in the Appendix A.13 and A.14 as well as an other comparison.

Cluster stability is also of interest to investigate: what happens when $k$ goes from $k = 2$ to $k = 3$ for example. Results where samples randomly move between the clusters are not useful regarding the clusters' stability.

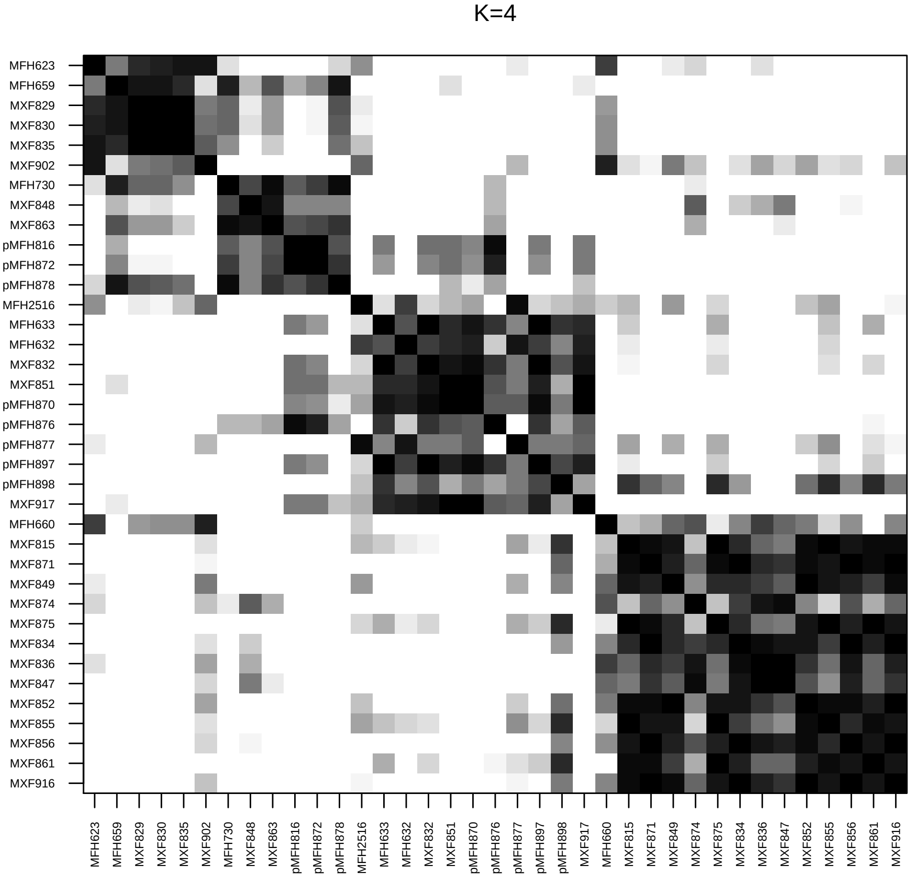


Figure 5.3: POD with Clusters for $\lambda = 0.13$ and $k = 4$ within the sd_GE_13 data set

## 5.2 Results of stability of $k$ different clusters

As mentioned in Section 2.8 it is of great interest to measure how the clustering results behave if one is increasing the number of clusters from $k = 2$ to $k = 3$ for a data set, for example. In this case the Wallace Index$_{10}$ is the most appropriate index to compare the cluster sets. In Table 5.2 numbers $1, 2, 3, 4, 5$ stand for the membership of one of the 37 samples to Cluster 1, Cluster 2, et cetera. The numbers of the clusters do not stand for anything particular and are not compulsory the same numbers then in the R output. Those numbers where chosen to make the understanding easier.

The stability of the cluster(s) is (very) good as samples of one cluster do not move to another cluster when number of clusters increases to $k = 3$ from $k = 2$. Most of the time by going to a higher $k$ one cluster splits into two new clusters. Also some of the samples of another cluster might join the new cluster.

In the following one can understand how the cluster(s) split into new cluster(s) and how the samples behave in that particular case by going from ... to .... Statistical tests for the quality can also be found on each page.

| No | Sample | k=2 | k=3 | k=4 | k=5 |
|---|---|---|---|---|---|
| 1 | PD_U133A_MFH2516 | 1 | 3 | 3 | 3 |
| 2 | PD_U133A_MFH633 | 2 | 3 | 3 | 3 |
| 3 | PD_U133A_MFH623 | 2 | 1 | 4 | 5 |
| 4 | PD_HG_U133A_MFH660 | 1 | 1 | 1 | 5 |
| 5 | PD_MXF815_HG-U133A | 1 | 1 | 1 | 1 |
| 6 | PD_U133A_MFH632 | 2 | 3 | 3 | 3 |
| 7 | PD_U133A_MFH659 | 2 | 2 | 4 | 4 |
| 8 | PD_MFH730_HG_U133A | 2 | 2 | 2 | 2 |
| 9 | PD_MXF871_HG-U133A | 1 | 1 | 1 | 1 |
| 10 | PD_MXF829_HG-U133A | 2 | 2 | 4 | 4 |
| 11 | PD_MXF830_HG-U133A | 2 | 2 | 4 | 4 |
| 12 | PD_MXF832_HG-U133A | 1 | 3 | 3 | 3 |
| 13 | PD_MXF849_HG-U133A | 1 | 1 | 1 | 1 |
| 14 | PD_MXF874_HG-U133A | 1 | 1 | 1 | 5 |
| 15 | PD_MXF875_HG-U133A | 1 | 1 | 1 | 5 |
| 16 | PD_MXF834_HG-U133A | 1 | 1 | 1 | 5 |
| 17 | PD_MXF835_HG-U133A | 2 | 2 | 4 | 4 |
| 18 | PD_MXF836_HG-U133A | 1 | 1 | 1 | 5 |
| 19 | PD_MXF847_HG-U133A | 1 | 1 | 1 | 1 |
| 20 | PD_MXF848_HG-U133A | 2 | 2 | 2 | 2 |
| 21 | PD_MXF851_HG-U133A | 2 | 3 | 3 | 3 |
| 22 | PD_MXF852_HG-U133A | 1 | 1 | 1 | 1 |
| 23 | PD_MXF855_HG-U133A | 1 | 1 | 1 | 1 |
| 24 | PD_MXF856_HG-U133A | 1 | 1 | 1 | 1 |
| 25 | PD_MXF861_HG-U133A | 1 | 1 | 1 | 1 |
| 26 | PD_MXF863_HG-U133A | 2 | 2 | 2 | 2 |
| 27 | PD_pMFH816_HG-U133A | 2 | 3 | 2 | 2 |
| 28 | PD_pMFH870_HG-U133A | 2 | 3 | 3 | 3 |
| 29 | PD_pMFH872_HG-U133A | 2 | 2 | 2 | 2 |
| 30 | PD_pMFH876_HG-U133A | 2 | 3 | 3 | 3 |
| 31 | PD_pMFH877_HG-U133A | 1 | 3 | 3 | 1 |
| 32 | PD_pMFH878_HG-U133A | 2 | 2 | 2 | 2 |
| 33 | PD_pMFH897_HG-U133A | 2 | 3 | 3 | 3 |
| 34 | PD_pMFH898_HG-U133A | 1 | 3 | 3 | 1 |
| 35 | PD_MXF902_HG-U133A | 1 | 1 | 4 | 5 |
| 36 | PD_MXF916_HG-U133A_2 | 1 | 1 | 1 | 5 |
| 37 | PD_MXF917_HG-U133A | 2 | 3 | 3 | 3 |

Table 5.2: Sample cluster membership for different $k$

### 5.2.1 Going from $k = 2$ to $k = 3$

| Cluster No for $k = 2$ | Cluster No for $k = 3$ | Number | Action |
|---|---|---|---|
| 1 | 1 | 15 | Stay |
| 1 | 3 | 4 | New |
| 2 | 1 | 1 | Move |
| 2 | 2 | 9 | Stay |
| 2 | 3 | 8 | New |

| Action | Number | Amount |
|---|---|---|
| Stay | 24 | 65 % |
| Move | 1 | 3 % |
| New | 12 | 32 % |

Table 5.3: Action of the samples for $k = 2, 3$

Cluster 1 stays more or less together, but four samples (definied: $S_1$) move to the new Cluster 3 that is mostly created by samples from Cluster 2. One can say: Cluster 1 stays, Cluster 2 splits into two clusters, the remaining Cluster 2 and the new Cluster 3. There is also one sample, that is moving between the two old clusters: it is moving from Cluster 2 to Cluster 1.
$WI_{10}(k = 2, \tilde{k} = 3) \approx 0.79$ shows statistically that the results for this analysis is good though the new cluster is created not only from one of the old clusters that would increase the index.

### 5.2.2 Going from $k = 3$ to $k = 4$

| Cluster No for $k = 3$ | Cluster No for $k = 4$ | Number | Action |
|---|---|---|---|
| 1 | 1 | 14 | Stay |
| 1 | 4 | 2 | New |
| 2 | 2 | 5 | Stay |
| 2 | 4 | 4 | New |
| 3 | 2 | 1 | Move |
| 3 | 3 | 11 | Stay |

| Action | Number | Amount |
|---|---|---|
| Stay | 30 | 81 % |
| Move | 1 | 3 % |
| New | 6 | 16 % |

Table 5.4: Action of the samples for $k = 3, 4$

Samples from Cluster 1 stays together in the same cluster. Only two samples (defined: $S_2$), from original Cluster 1, create the new cluster, Cluster 4. One of the two samples creating $S_3$ was former sample $S_1$. Cluster 3 also stays as it was before, with only on sample moving to Cluster 2, what was one of the samples that was previously in Cluster 2. Cluster 2 again splits into one cluster still named Cluster 2 and one new cluster, Cluster 4.

The value of Wallace Index $WI_{10}(k = 3, \tilde{k} = 4) \approx 0.93$ indicates that only a couple of samples move from/to their original clusters.

The wallace Index for $k = 2$, $k = 4$ is $\approx 0.81$ showing that the clustering is very stable.

### 5.2.3 Going from $k = 4$ to $k = 5$

| Cluster No for $k = 4$ | Cluster No for $k = 5$ | Number | Action |
|---|---|---|---|
| 1 | 1 | 8 | Stay |
| 1 | 5 | 6 | New |
| 2 | 2 | 6 | Stay |
| 3 | 1 | 2 | Move |
| 3 | 3 | 9 | Stay |
| 4 | 4 | 4 | Stay |
| 4 | 5 | 2 | New |

| Action | Number | Amount |
|---|---|---|
| Stay | 27 | 73 % |
| Move | 2 | 5 % |
| New | 8 | 22 % |

Table 5.5: Action of the samples for $k = 4, 5$

Going from four to five clusters leads to splitting of Cluster 1 into Cluster 1 and Cluster 5. There are two samples moving from Cluster 4 to the new cluster that we observed as $S_2$ before. There are no other samples joining the new cluster. Cluster 2 stays exactly the same. Cluster 3 stays together for most of the samples, but two samples move back to Cluster 1 where they came from as $S_1$ with four samples. There is also no other movement from Cluster 4 to other clusters than the movement to the new cluster already mentioned above.

As the new cluster is the result of another splitting and there are more or less no samples moving between the clusters the value $WI_{10}(k = 4, \tilde{k} = 5) \approx 0.78$ of Wallace Index shows that as well. $WI_{10}(k = 3, \tilde{k} = 5) \approx 0.84$ is also a good result if one is just interested in the index of increasing the clusters by two.

### 5.2.4 Going from $k = 2$ to $k = 5$

| Cluster No for $k = 2$ | Cluster No for $k = 5$ | Number | Action |
|---|---|---|---|
| 1 | 1 | 10 | Stay |
| 1 | 3 | 2 | New |
| 1 | 5 | 7 | New |
| 2 | 2 | 6 | Stay |
| 2 | 3 | 7 | New |
| 2 | 4 | 4 | New |
| 2 | 5 | 1 | New |

| Action | Number | Amount |
|---|---|---|
| Stay | 16 | 43 % |
| Move | 0 | 0 % |
| New | 21 | 57 % |

Table 5.6: Action of the samples for $k = 2, 5$

Altogether it is possible to say that Cluster 1 splits into Cluster 1 and Cluster 5 with only two samples going to Cluster 3. Cluster 2 splits into Cluster 2, Cluster 3 and Cluster 4 with only one sample going to Cluster 5. It is obvious that there is little movement between the clusters, but it turns out that the clusters are staying the same or creating new clusters as a subset of the old cluster. This is a good result in the way of stable results all over the samples and all over all 11 different data sets in this study.

The overall conclusion is that iCluster produces good stable results with the Wallace Index $WI_{10}(k = 2, \tilde{k} = 5) \approx 0.84$.

### 5.2.5 Overall stability

One can see in Figure 5.4 how clusters are formed from each other. Dotted lines represent the movements between one cluster and another, while dashed lines stand for samples going from cluster to a new cluster, but contain fewer samples then the other samples creating this new cluster. A good overlook about the results in the last pages is also possible by regarding Table 5.7. Another way to see this splitting of the clusters can be found in Appendix A.15, A.16 and A.17.

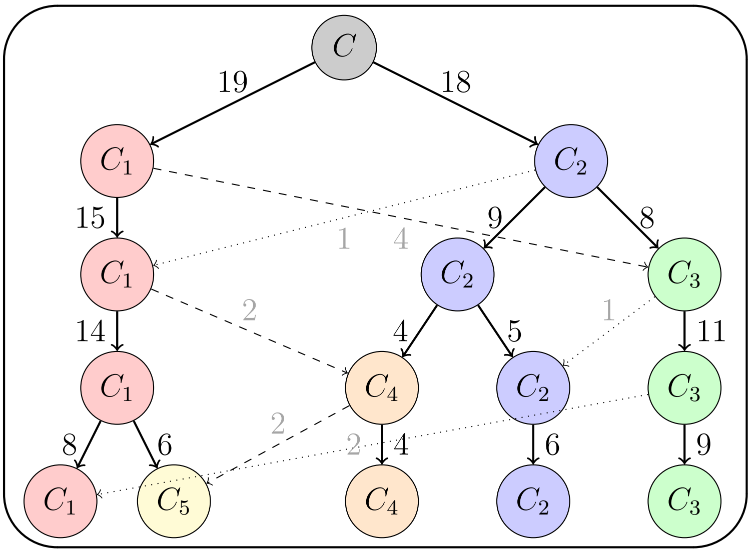


Figure 5.4: Tree diagram for all clusters

| Going from ... to ... | $WI_{10}(k, \tilde{k})$ |
|---|---|
| $k = 2 \ / \ \tilde{k} = 3$ | 0.79 |
| $k = 3 \ / \ \tilde{k} = 4$ | 0.93 |
| $k = 4 \ / \ \tilde{k} = 5$ | 0.78 |
| $k = 2 \ / \ \tilde{k} = 4$ | 0.81 |
| $k = 2 \ / \ \tilde{k} = 5$ | 0.84 |
| $k = 3 \ / \ \tilde{k} = 5$ | 0.84 |

Table 5.7: Increasing the number of clusters $k$

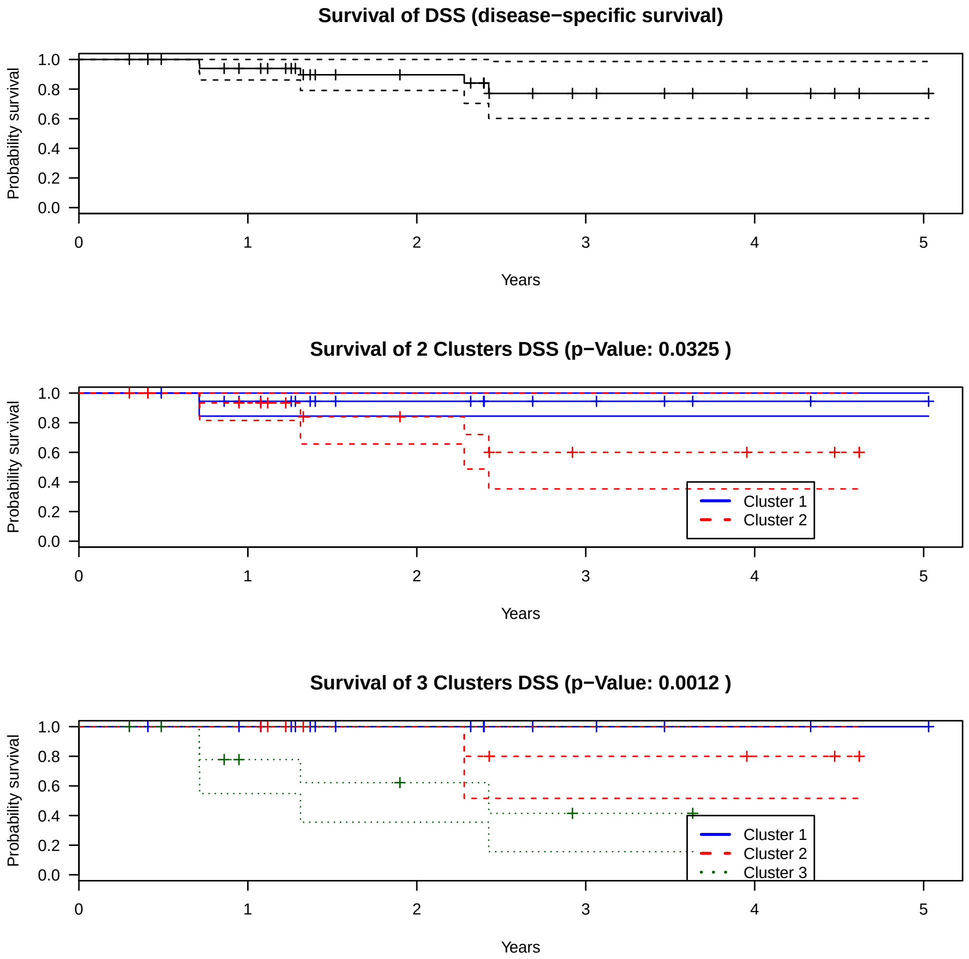
## 5.3 Clinical Analysis

As described in Section 2.9.1 it is important to perform survival analysis on obtained clusters to see if the results are meaningful from medical/biological point of view and to validate the quality of the clusters found with iCluster. Even the most stable results would be useless if one does not get any results from the clinical analysis.

For this data set there is survival data are available for 36 out of 37 samples used for the analysis with iCluster. No data is available for one sample. Two different types of survival analysis were carried out. The first one is DSS, a Disease-Specific Survival that is looking for the percentage of people survived since the diagnosis. The so called event that was regarded was death. Death from other causes was not counted as an event in DSS. The other analysis that was possible is a Distant Recurrence-Free Survival (DRFS). This is a measure of whether people have metastasis or not. So, if something is correlated with worse DRFS, that means that it correlates with an increase in metastasis.

First, a survival analysis for all the samples together is done. After this the survival curves for each cluster are plotted with the Kaplan-Meier-Estimator. The hypothesis if the survival curves are significantly different from each other are tested with the Log-Rank-Test with a significance level of $\alpha = 0.05$. A plot for the overall analysis as well as survival curves for two and three clusters can be found in Figure 5.5.

For all the 36 samples the times and events can be found in Table A.1. Figure 5.5 shows the output while Table 5.8 shows the results of the Log-Rank-Test with the $p$-value. Note that the number of degrees of freedom is $k-1$ for an analysis with $k$ different clusters. For the Disease-Specific Survival all $p$-values are significant for a significance level of $\alpha = 0.05$. This shows that the survival curves behave differently from each cluster comparing the survival curves. For $k = 2$ the Distant Recurrence-Free Survival analysis is not significant. However for all other $k$ it is also significant with this $p$-value cutoff. Note that for an increasing $k$ less samples are in each cluster and therefore the $p$-values have to be interpreted with caution. More detailed information about the values can be found in the RData-File.


Survival of DSS (disease−specific survival)

Survival of 2 Clusters DSS (p−Value: 0.0325 )

Survival of 3 Clusters DSS (p−Value: 0.0012 )

Figure 5.5: Survival curves for DSS for the overall (clusterfree), two cluster and three cluster case

|  | DSS |  | DRFS |  |
|---|---|---|---|---|
| Clusters $k$ | $T_{LR}$ | $p$-value | $T_{LR}$ | $p$-value |
| 2 | 4.58 | 0.0325 | 3.12 | 0.0774 |
| 3 | 13.41 | 0.0012 | 6.90 | 0.0318 |
| 4 | 17.53 | 0.0005 | 8.18 | 0.0424 |
| 5 | 17.90 | 0.0013 | 10.64 | 0.0310 |

Table 5.8: Log-Rank-Test with $p$-values for all the clusters for both types of clinical analysis

## 5.4 Differentially expressed (DE) genes and the DAVID tool

Many different methods exist to identify differentially expressed genes in Gene Expression data. The result is a list of genes that are significantly different between two conditions or tumor vs. normal or tumor subtypes. The gene list can be used to run, for example the DAVID tool [15], [16] to find out the over-represented/enriched genes/categories in a biological/medical way.

### 5.4.1 Differentially expressed genes

The goal is to find genes that behave different in two conditions. Two conditions can be two clusters. The analysis was done with a standard limma procedure [38] in R. There are 123 genes that are up-regulated and 265 genes that are down-regulated for all the genes in the Gene Expression data by comparing samples for two conditions. As condition one samples from Cluster 1 and as condition two samples from Cluster 2 were used from the $k = 2$ cluster case.

One gets these genes by testing a null hypothesis that a gene is not differentially expressed. As a result of that test one gets a $p$-value that is related to the false positive rate (FPR). As there is more than one gene one has to do a multivariate test. This is possible with the Benjamini and Hochberg FDR procedure [1].

The work to find these differentially expressed genes was mainly done for biological/medical and not statistical background. It was included in this thesis because one maybe study this topic further, in particular do enrichment analysis to find over-represented gene categories. See the following Section 5.4.2 for more details or do other biological analysis with the results.

### 5.4.2 DAVID tool

As described in Huang et. al [15], [16] the DAVID (Database for Annotation, Visualization and Integrated Discovery) tool is a web program[9]. This tool helps scientists to interpret their list of genes in a biological way. There are a lot of interesting results with this data that one may want to check out by using the two data files for the up- and down-regulated genes (provided in gene symbols) and trying some of the many options in the DAVID tool.

[9] http://david.abcc.ncifcrf.gov/

## 5.5 Accordance with other data

All methods with optimal parameters are useless if they get good and stable results in one data set, even if there were a lot of different sub sets. To show that iCluster provides good results with the methods and ways of data shrinkage explained before, a test run with another data set was done. In this data set altogether 64 samples occur in both, Copy Number and Gene Expression data with 10602 probes (CN) and 22215 genes (GE). The format of the data is the same as well the methods of creating subsets of the GE and CN data. Altogether iCluster was run with 13 different subsets including one random subset. To get the clusters one was looking at the POD again and then got the best clusters for different thresholds $\lambda$ as well as different amount of clusters $k$. Again best clusters, that are convergent, are found. It turned out that for an increasing $k$ the cluster behavior was fine until $k = 4$. For $k = 5$ the results of all the different subsets as well as the cluster membership was not straightforward anymore. Excluding this yields good results that can also be found in RData-Files on the DVD C. In addition, the survival analysis for this data returns significant $p$-values for the clusters when excluding the case DRFS for $k = 2$. In Table 5.12 one sees that 29 of the 37 samples of the Sarcoma data also occur in other data set. For this 29 of the 64 samples are common.

### 5.5.1 Comparison of the results for different $k$ and different data sets

Table 5.9 shows the just introduced results and with it that they are fine for $k = 2$, $k = 3$ and $k = 4$ but not for $k = 5$ as the value for the Wallace Index$_{10}$ (see Section 2.8.2 for details) drops a lot underneath the value of 0.80.

The same occurs if one is looking at the $p$-values for the survival analysis of this data:

As the tree diagram would not be readable for $k = 5$ anymore it is only for $k = 2, 3, 4$ in Figure 5.6. Note that cluster numbers $C_1, C_2, C_3$ and $C_4$ do not fit with the numbers of Table 5.12. The exact cluster membership of each sample can be found on the DVD.

| Going from $\ldots$ to $\ldots$ | $WI_{10}(k, \tilde{k})$ |
|---|---|
| $k = 2$ / $\tilde{k} = 3$ | 0.80 |
| $k = 3$ / $\tilde{k} = 4$ | 0.89 |
| $k = 4$ / $\tilde{k} = 5$ | 0.34 |
| $k = 2$ / $\tilde{k} = 4$ | 0.83 |
| $k = 2$ / $\tilde{k} = 5$ | 0.61 |
| $k = 3$ / $\tilde{k} = 5$ | 0.42 |

Table 5.9: Increasing the number of clusters $k$

|  | DSS | DRFS |
|---|---|---|
| Clusters $k$ | $p$-value | $p$-value |
| 2 | 0.0097 | 0.9070 |
| 3 | 0.0041 | 0.0646 |
| 4 | 0.0034 | 0.0513 |
| 5 | 0.0342 | 0.2264 |

Table 5.10: $p$-values for all the clusters for both types of clinical analysis for the new data

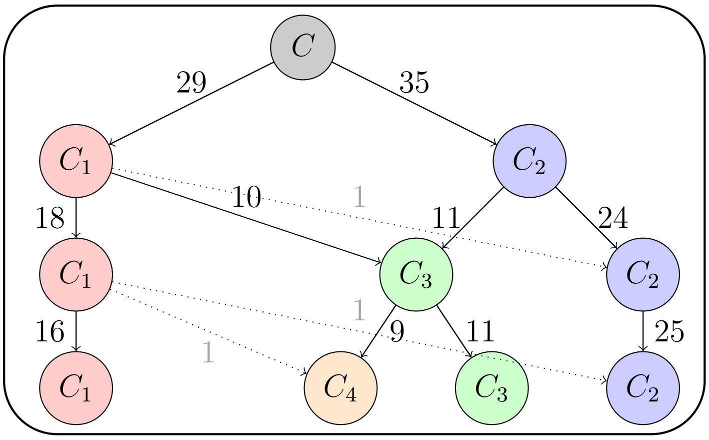


Figure 5.6: Tree diagram for all clusters

### 5.5.2 Comparison of the results for both data sets

To compare the clustering results that are created with two different data sets one should use the Rand Index 2.8.1 as it turned out that this index is the best one to compare the results in an appropriate way, see Section 2.8.5 for more details. As the results of the larger data set turned out to be not good for $k = 5$ they are not included in the following Table 5.11.

| Cluster $k$ | $RI(k)$ |
|---|---|
| $k = 2$ | 0.76 |
| $k = 3$ | 0.82 |
| $k = 4$ | 0.89 |

Table 5.11: Comparison of the two data sources

All values of the Rand Index are high enough to say that the results are similar in both approaches which demonstrates that results of iCluster are consistent and stable. A little bit surprising is that the value increases as $k$ increases but this can be explained as for a smaller $k$ some samples are not in the same cluster, as the clusters are different because one time there were 37 and the other time 64 samples in the data. This problem is getting smaller for an increased $k$ as more clusters are created and the missing 8 or 35 samples play a smaller role in the analysis.

| Sarcoma data | | 37 samples | | | 64 samples | | |
|---|---|---|---|---|---|---|---|
| No | Sample | k=2 | k=3 | k=4 | k=2 | k=3 | k=4 |
| 1 | PD_U133A_MFH2516 | 1 | 3 | 3 | 2 | 3 | 3 |
| 2 | PD_U133A_MFH633 | 2 | 3 | 3 | 2 | 3 | 3 |
| 3 | PD_U133A_MFH623 | 2 | 1 | 4 | 2 | 2 | 4 |
| 4 | PD_HG_U133A_MFH660 | 1 | 1 | 1 | 1 | 2 | 4 |
| 5 | PD_MXF815_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 6 | PD_U133A_MFH632 | 2 | 3 | 3 | 2 | 3 | 3 |
| 7 | PD_U133A_MFH659 | 2 | 2 | 4 | 2 | 2 | 2 |
| 8 | PD_MFH730_HG_U133A | 2 | 2 | 2 | 2 | 2 | 2 |
| 9 | PD_MXF871_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 10 | PD_MXF829_HG-U133A | 2 | 2 | 4 | 2 | 2 | 2 |
| 11 | PD_MXF830_HG-U133A | 2 | 2 | 4 | 2 | 2 | 2 |
| 12 | PD_MXF832_HG-U133A | 1 | 3 | 3 | 2 | 3 | 3 |
| 13 | PD_MXF849_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 14 | PD_MXF874_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 15 | PD_MXF875_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 16 | PD_MXF834_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 17 | PD_MXF835_HG-U133A | 2 | 2 | 4 | 2 | 2 | 4 |
| 18 | PD_MXF836_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 19 | PD_MXF847_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 20 | PD_MXF848_HG-U133A | 2 | 2 | 2 | 1 | 2 | 2 |
| 21 | PD_MXF851_HG-U133A | 2 | 3 | 3 | 2 | 3 | 3 |
| 22 | PD_MXF852_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 23 | PD_MXF855_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 24 | PD_MXF856_HG-U133A | 1 | 1 | 1 | 1 | 1 | 1 |
| 25 | PD_MXF861_HG-U133A | 1 | 1 | 1 | 2 | 1 | 1 |
| 26 | PD_MXF863_HG-U133A | 2 | 2 | 2 | 2 | 2 | 2 |
| 35 | PD_MXF902_HG-U133A | 1 | 1 | 4 | 1 | 2 | 4 |
| 36 | PD_MXF916_HG-U133A_2 | 1 | 1 | 1 | 1 | 1 | 1 |
| 37 | PD_MXF917_HG-U133A | 2 | 3 | 3 | 2 | 3 | 3 |

Table 5.12: Sample cluster membership for different $k$ for both data sources

# CHAPTER 6

# Summary and Conclusion of iCluster

Why is one interested in clustering samples? This is not only a statistical problems, it is also a medical issue as physicians are trying to identify similarities and subtle differences in patient's tumor samples in order to target specific therapies to groups of patients, for example. Clustering basically means grouping together samples that behave, in this analysis in a statistical way, similar to the other samples in the same cluster. 'Clustering is the most appropriate in typical clinical experiments' as Wit and McClure show in [38]. It is important to look at the samples, even if they come from the same phenotype as it is maybe possible through clustering methods to find new, so far unknown, subtypes of cancer. How can one find out that the samples are similar? Special clustering methods are the way to solve this question. Within this thesis the method from Shen et al., iCluster [28], is systematically studied and applied to an integrated data from Sarcoma patients. A short summary of the previous chapters as well as possible further studies and a conclusion is given in the following sections.

## 6.1 Summary of iCluster

Integrative clustering is an important topic as more and more genomic data sets of different types are becoming available iCluster is a powerful method to deal with multiple high-dimensional genomic data sets in cancer research at once. That an integrative approach is more powerful than $n$ stand-alone approaches with a followed manual integration is obvious.

During the analyses in this thesis one was looking into the details of iCluster and the methods behind the algorithm. Further indices as well as event analysis were introduced. The main example was done with the Sarcoma data set that includes

Gene Expression and Copy Number data. A lot of different subsets of the data were created with a set of four different thresholds for different parameters. It turned out that clusters, created by of all subsets, are equal for each $k$. Another good way to show the stability of the clusters is to look at how the samples behave if one increases the number of $k$, what is measured with the Wallace Index. The comparison with other data set, in that case with the Sarcoma data set for 64 samples, for the same $k$ also shows good results with the Rand Index. After showing all the methods and results in a statistical way it is also important to look at the survival analysis. This analysis also gives significantly different survival curves for most of the analysis of DSS and DRFS.

## 6.2 Further studies of iCluster

Further studies of the paper or of ongoing work with the existing methods are possible and would increase the quality of the results. For example, by creating a function or method that choses the best tuning parameter $\lambda$ on its own. This would lead to less manual work as well as may be saving computational time depending on how one is going to solve this problem. Closely related to this topic is the calculation of the POD that should be used to get the best threshold(s). Another important topic would be the whole analysis of the cluster membership, including the Wallace Index, as wells as the clinical analysis that should be implemented in the R package to give the analyst an easier way to use iCluster to solve the problem. POD yields result in a statistical way, meanwhile the other measurements provide an actual analysis of the real problems and answer biological and medical questions. Another field of study would be the integration of all the data such that one does not have to create subsets of the data manually. Depending on their importance, genes should not be deleted and others integrated to in the subset. It may be possible to integrate all the genes with a special weight function that shrinks genes without any information towards zero and just uses the most important ones.

## 6.3 Conclusion of iCluster

As this analysis demonstrates iCluster provides stable results that satisfy statistical criteria. In addition, an interesting from medical point of view insights is obtained from survival analysis of the data. The methods used in statistical analysis included the comparison of different subsets of the data, different numbers of clusters and also a comparison of the two main data sources. Some additional measurements make iCluster a perfect tool as an integrative clustering method.

# CHAPTER 7

# Methods of GSS

The idea of the Gene Set Score (GSS) came up after looking at differentially expressed (DE) genes (see Section 5.4.1) and the ACE (Analysis of Copy Number Alterations by Expression data) score [14] of a given region of genes. One can say that the GSS is something like a mixture of the two existing methods to deal with data. GSS integrates two different data types (Gene Expression and Copy Number data) and scores gene sets/pathways by taking into account events in both data types. Both other methods just deal with one type of data at once. The ACE uses Copy Number data of different genes connected by its loci. Meanwhile, Gene Expression data is used for the DE analysis. The new GSS uses both data types and it is open for further investigation concerning integration of the structure of a region/pathway/gene set. First, one should understand the DE and ACE better before looking at the GSS.

## 7.1 Analysis of gene sets

To identify gene sets, that are deregulated (activated, de-activated or differentially regulated) between different conditions, phenotypes and specifically cancer subtypes is a big challenge for genomics data analysis. The original idea was advanced by Bild and Febbo [2] and there are various approaches that do not take the structure of the pathway into account (GSEA, GSA, IGA, SPIA) as well as there are approaches that take the connections of genes in a pathway into account. The Gene Set (Enrichment) Analysis that do not take structure into account, are explained in the following. In particular the GSA method that is further compared with results of the GSS. We proceed by looking separately at Gene Expression data.

### 7.1.1 Individual Gene Analysis (IGA)

IGA as its name suggests performs analysis on the level of individual genes. In this analysis the following null-hypothesis is tested to compare two conditions with each other:

$H_0$: Gene i is not differentially expressed between condition 1 and condition 2

to compare two conditions with each other. Its simplest form is using a t-test to calculate a $p$-value for each of the genes. The resulting $p$-values need some adjustments for multiple hypothesis tests. The Bonferroni correction [3], [4] was used. As genes are intrinsically or biologically organized in pathways, it is a disadvantage of IGA that this analysis is just looking at each gene individually. Genes are assumed to be independent while they are connected and strictly speaking are not independent as well as different results occur if one uses different test statistics.

### 7.1.2 Gene Set Enrichment Analysis (GSEA)

GSEA is a method for comparing two conditions that uses gene sets instead of looking at single genes and it was first introduced by Subramanian et al. [31]. 'The idea behind this method is that even if there is only a weak expression change in the individual genes it can occur that in association with large gene sets there will be a significant pattern. While the IGA method cannot discover this, the GSEA is a powerful, analytical tool which gets its power by focusing on pathways', as it is written in the summary of Kühnle and Pfundstein [20]. The GSEA compares two conditions by first ordering all genes in a ranked list with absolute values due to their differential expression between two conditions with a Signal2Noise-Statistic. For a given pathway GSEA then calculates whether the genes of that pathway are randomly distributed throughout this list or found at the top or bottom of that list. For this one uses an Enrichment Score (ES) in the Gene Set Enrichment Analysis. The ES shows how many genes in a given gene set are found at the extreme values (bottom or top) of the ranked list. The correlation of the gene expression with a certain phenotype or condition is also included in the score. To identify significant pathways, one is doing a permutation test by permuting the conditions many times.

### 7.1.3 Gene Set Analysis (GSA)

Efron and Tibshirani [8] introduced GSA as an improvement of GESA [31]. One advantage of GSA in dealing with the data is that it is implemented in R instead of using special software for that so that analysts can easily continue working with that. Another advantage, that is not needed in the analysis of GSS yet, is that one is able to compare multiple phenotypes in one step instead of just comparing two at a time. The biggest analytical advantage of the Gene Set Analysis is that GSA uses the Maxmean-Statistic instead of the Signal2Noise-Statistic. In the paper of Efron and Tibshirani [8] it was shown that this is often more powerful than the statistic used in GSEA which is a modified Kolmogorov-Smirnov-Statistic. GSA also includes the permutation of the genes and not only the permutation of samples.

### 7.1.4 Signaling Pathway Impact Analysis (SPIA)

SPIA by Tarca et al [32] goes one step further than GSA and GSEA taking into account the structure of a pathway. The Signaling Pathway Impact Analysis requires pathway representation as directed graphs, availability and accuracy of which, in particular in cancer, is still an issue. SPIA calculates an Impact Factor for each pathway as a sum of two terms. One of the terms uses the information from the $p$-value from IGA in Section 7.1.1 with a variety of possibilities of how to use it (maximal observed significant one, mean of all significant $p$-values, etc.). The other term is a functional term that depends on the identity of the differentially expressed genes from a gene set and the interactions described by that set of genes.

### 7.1.5 Choosing one method for comparison with the GSS

As the Gene Set Score that is being developed in this thesis is not yet dealing with the structure of the pathways, it is therefore clear that one is not interested in using SPIA. Using IGA does not make sense as this method is only looking at single genes and not at pathways or gene sets. GSA contains all the features of GSEA plus additional advantages that it uses more robust statistics and is implemented in R. GSA is therefore the best choice for the analysis of pathways or gene sets and the method that will be used for comparison to the GSS.

## 7.2 ACE - Analysis of CNAs by Expression data

Alterations in Copy Number data (CNAs) are detected in a lot of human cancers and therefore amplification/deletions in the genes are often mediators of malignancy in the genomes. Various techniques are available to deal with these events. Detection of CNAs from expression data is technically difficult as Hu et al. [14] say in their paper because the 'expression data reflect multiple layers of gene regulation beyond genomic alterations.' They introduced the ACE, a computational algorithm, to identify regions of gain and loss.

Before each step of the ACE is introduced in more detail it is useful to look at it in an algorithm:

```
Algorithm 3: ACE Algorithm
  Data: Copy Number data by expression data
  Result: The goal is to define locations/regions of gain and loss
repeat
    As the indicator of CNA likelihood one calculates the Neighboring Score
    (NS) for each of the chromosomal locus (i)
    The significance of the NS is estimated (ii)
    Regions of loss and gain are defined (iii)
until it is done for all the genes
```

Now it is useful to look at the algorithm in more detail.

**(i)**

The Neighboring Score (NS) is calculated for each of the chromosomal locus. In their paper Hu et al. [14] used paired t-statistics for ovarian cancer cell lines as well as independent t-statistics for other data sets to calculate the Expression Score (ES) for each gene. This is done by 'the correlation of its expression with the phenotypes in comparison' [14]. An extension of this method is that one can use other metrics than t-statistics to calculate the ES. For all $j = 1, 2, \ldots, N$ genes ordered by their physical position on genome one can define the NS at the locus $i$:

$$NS_i = \sum_{j=1}^{N} w_{ji} \cdot ES_j \ , \tag{7.1}$$

where weight $w_{ji}$ is the weight of gene $j$ at locus $i$. It depends on the distance between the two loci. Obviously the weight is getting weaker when the distance increases. The weight is calculated by using a Gaussian function:

$$w_{ij} = c \cdot exp\left(-\frac{(j-1)^2}{2 \cdot \sigma^2}\right) \ , \tag{7.2}$$

where $c$ is a rescaling/nomalizing constant to get a NS, within the interval $[-1, +1]$. In the analysis of ACE the variation controls the weight decay rate. $\sigma$ was set to values between 3 and 10 and a default of around 7. Hu et al. [14] claim that, as written in the paper, similar values can be used in other studies. With the weight function given by Equation 7.2 physically close genes have a big influence on the NS compared to genes that are far away. Positive and negative values for the NS are possible and stand for gain or loss.

**(ii)**

To calculate $p$-values and the significance of the NS various permutation schemes can be used. For example, one can permute the gene positions and, if there are enough samples available, the samples. This permutations are done 1000 times and one can calculate the Neighboring Score for each permutation. Hu et al. [14] then mention that 'the $p$ values of observed NS are then computed using the distribution of permuted NS and adjusted to false discovery rate (FDR) $q$ values by the Benjamini-Hochberg procedure'.

**(iii)**

The final step is to define regions of gain or loss. If there are 20 or more continuous NSs that are positive, and a $q \leq 0.01$ (DFR), one speaks about a gain in the region. For 20 ore more negative NS one speaks of a loss.

# CHAPTER 8

## Data preparation of GSS

To calculate the GSS and determine its significance the data has to be organized such that the calculations are possible. Some preprocessing steps are needed and are described in detail in the next sections.

### 8.1 Gene Expression and Copy Number data

The data is again a set of two data sets, one containing Gene Expression data and one with Copy Number data. This set is publicly available in the internet[10] and it will be referred as the ‘Prostate Dataset’.
The raw data for Gene Expression (MSKCC_PCa_mRNA_data) looks like the following subset:

$$\begin{array}{lll}
GeneID & GeneSymbol & \begin{array}{ccc} PCA0001 & PCA0002 & PCA0003 \end{array} \\
\begin{array}{l} 1 \\ 9974 \\ 54715 \end{array} & \begin{array}{c} A1BG \\ A1CF \\ A2BP1 \end{array} & \begin{pmatrix} 8.056938 & 7.955632 & 8.287924 \\ 5.387556 & 5.573194 & 5.781781 \\ 6.876291 & 6.614546 & 6.874874 \end{pmatrix}
\end{array}$$

This is a $3 \times 3$ subset out of a $26447 \times 185$ matrix. This means that there are 26447 genes/probes and 185 samples in the data set. The following subset (again a $3 \times 3$ matrix), is from the $18202 \times 207$ data matrix (MSKCC_PCa_RAE_gene_calls) of Copy Number data:

[10] http://cbio.mskcc.org/cancergenomics/prostate/data/

$$\begin{array}{llc} GeneID & GeneSymbol & \begin{array}{ccc} PCA0001 & PCA0002 & PCA0003 \end{array} \\ \begin{array}{l} 1 \\ 9974 \\ 54715 \end{array} & \begin{array}{l} A1BG \\ A1CF \\ A2BP1 \end{array} & \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} \end{array}$$

In this subset just zero entries occur. In the whole matrix the following five values occur:

$$CN\ data\ parameter\ values = \{-2, -1, 0, +1, +2\}$$

Comparing to Section 3.2.2 the values of the Copy Number are not continuos within an interval $[-1, +1]$ and are now discrete while the Gene Expression data is continuous.

## 8.2 Similar samples and conditions

The next step is to extract similar samples profiled on both data sets as well as organize them in one of the two conditions. It turned out that there are 133 samples which are in both data types. 128 out of the 133 had some information about the condition. The two conditions are ‘PRIMARY’ and ‘MET’ which is also provided. This data is for prostate tumors. ‘PRIMARY’ is for initial or primary tumor and ‘MET’ for metastasis. 109 samples have condition ‘PRIMARY’ and the other 19 ‘MET’. As the data set of Copy Number contained similar genes, these double genes are deleted as entries are the same. This occurred for 23 genes, such that 18179 genes are left. Altogether a vector with the conditions as well as the following two matrices are used:

- A 128 × 26447 Gene Expression data matrix.
- A 128 × 18179 Copy Number data matrix.

## 8.3 Pathways as gene sets

Genes interact with each other on different levels. Interaction between genes are a complex biological process. Such interactions or regulations between genes can be viewed as being organized in different pathways. In other words groups of genes perform a specific function in the cell or participate in the same biological process.

Most of the genes are not just part of one pathway and some genes occur more than once in a certain pathway. The exact structure of many pathways and interactions within a pathway are not yet determined. For the GSS it is enough to consider pathways as sets of genes. Altogether seven different pathway collections are used for GSS. The first six of them can be found at the GSA website[11] and the seventh (KEGG) collection of pathways can also be found in the internet[12]. In Table 8.1 one can have a look at the sources of the pathways.
The number of (unique) genes in each pathway varies. The mean (blue line) of (unique) genes per pathway is around 46 genes while the median (green line) is 21 genes. The smallest pathway just included one gene and the largest pathway included 2728 genes. Altogether there were 625 pathways with 10 and less (orange line) genes and 124 pathways with more than 200 (red line) genes. The largest are not displayed in Figure 8.1 as it would make the histogram not as good readable as with excluding them. Note that later in the GSS just pathways with at least 10 genes in CN and GE data are included.

[11] http://www-stat.stanford.edu/~tibs/GSA/

[12] http://www.genome.jp/kegg/

| Number | Short name | Long name | # pathways |
|---|---|---|---|
| 1 | arms | Chromosome Arms from Stanford Microarray Database | 46 |
| 2 | cyto | Cytobands from Stanford Microarray Database | 797 |
| 3 | proc | Cellular processes gene sets from Stanford Microarray Database | 24 |
| 4 | sega | Cancer module gene sets from Eran Segal's lab | 456 |
| 5 | tile | 5MbChromosomalTiles from Stanford Microarray Database | 1192 |
| 6 | tiss | Tissues gene sets from Stanford Microarray Database | 80 |
| 7 | kegg | KEGG: Kyoto Encyclopedia of Genes and Genomes | 206 |
|  | TOTAL | All pathways together | 2801 |

Table 8.1: Name of the pathways

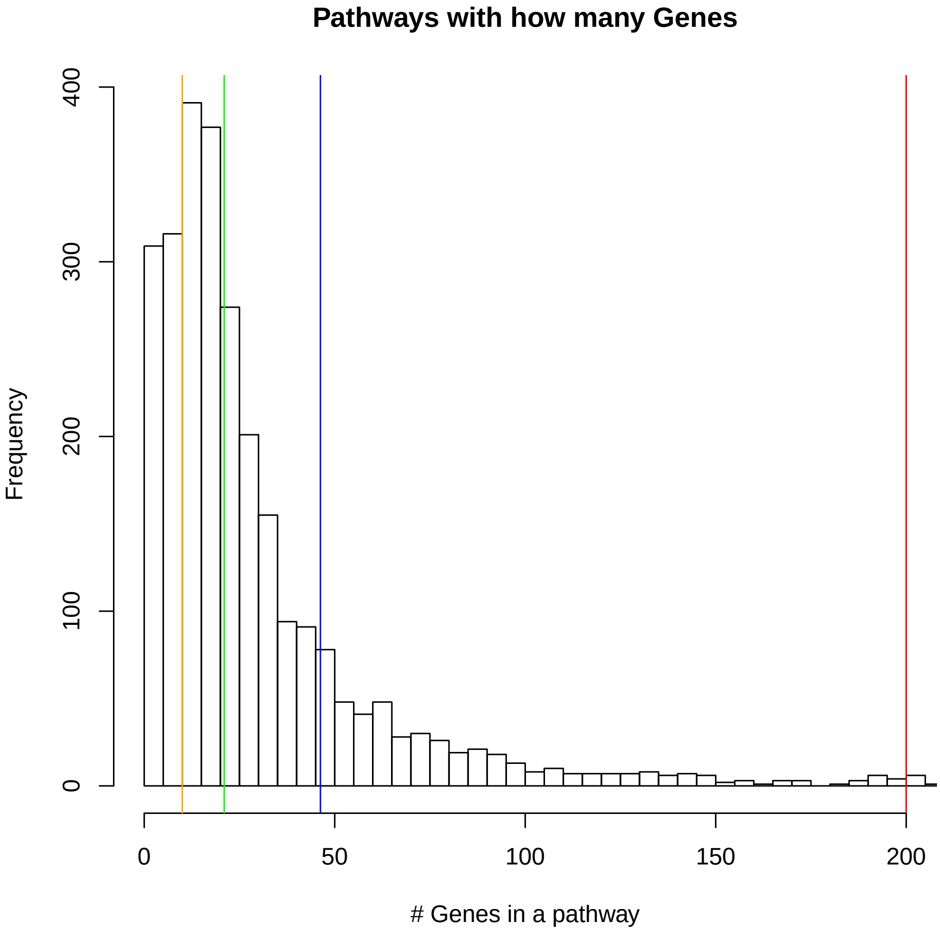


Figure 8.1: Histogram of the number of genes in each pathway

# CHAPTER 9

# Gene Score $GS$

In order to calculate the score for a gene set, a score for each gene in the set is required. Since Gene Expression and Copy Number data are of different types (one is discrete, the other one is continuous), the challenge is to combine two scores into one. In this chapter, first the Gene Score for Gene Expression data is explained (in Section 9.1) followed by the description of Gene Score for Copy Number data (Section 9.2). Having computed the Gene Scores for single genes, it is then possible to calculate Gene Score for a gene set.

## 9.1 Gene Score for Gene Expression data $GS_{GE}$

To calculate the Gene Score for each gene it is first required to calculate the Fold Change for each gene.

### 9.1.1 Fold Changes (FCs) of Gene Expression data

Various methods for identifying statistically significant differentially expressed genes are shown in section 5.4.1. The methods yield $p$-values as a result of the analysis of the difference of the two conditions. Another possibility is to look at the Fold Changes of the genes rather than the $p$-values from the t-statistics, mainly because the results are more reproducible as Witten and Tibshirani showed [39]. The first attempt of creating a FC for GE data was done by Tusher et al. [36]. They used the standard definition of a Fold Change that is calculated by the following equation:

$$FC_i = \frac{\overline{x_{i\cdot}}}{\overline{y_{i\cdot}}} \tag{9.1}$$

Here $\overline{x_{i\cdot}}$ and $\overline{y_{i\cdot}}$ represent the mean on raw expression level for gene $i$ in condition one ($\overline{x_{i\cdot}}$) or condition two ($\overline{y_{i\cdot}}$).
Another possibility to calculate the FC is the following method explained by Choe et al. [5] and Guo et al. [11]:

$$FC_i = \overline{x_{i\cdot}} - \overline{y_{i\cdot}} \tag{9.2}$$

For the GSS the second method is used, because the Gene Expression data is on $\log_2$-scale.

### 9.1.2 Gene Score

We would like the Gene Score for all genes in Gene Expression data to be within the interval $[0, 1]$. The closer to 1 the $GS_{GE}$ is the bigger the difference between the Fold Changes for the two conditions. After calculating the FC it is now possible to get the Gene Score. At least three possibilities can be considered:

**First possibility:**
For gene $i$, one gets the Gene Score for Gene Expression data as the ratio between the absolute Fold Change value of that gene divided by the maximum of all absolute Fold Change values across all genes.

$$GS_{GE}(i) = \frac{abs(FC_i)}{max(abs(FC))} \tag{9.3}$$

For the Prostate Dataset the mean of the GS of GE data is around 0.0418.

**Second possibility:**
As a second method one can think of setting a threshold to get rid of large outliers. All absolute FC values larger than the pre-defined threshold are set to the threshold value. For our data, as we consider some large data sets with many thousands genes (e.g. 20000 genes) it does not matter if fold changes are outliers, as they do not change the mean that much. However, if this happens one should definitely use a threshold. After that the Gene Score for each gene is calculated as in Equation 9.3. Results of both methods do not vary much as in the test data set enough genes were included. Method one is more conservative and is therefore chosen in this study.

**Third possibility:**

Another way is to order the values of the absolute values of the FC and then calculate the score using the rank:

$$GS_{GE}(i) = \frac{rank(abs(FC_i)) - 1}{\#genes - 1} \tag{9.4}$$

This Gene Score for Gene Expression data is also within the interval [0, 1]. As one can see the mean of this score is always 0.5. The results in the simulation were not as good as for the first two possibilities. Therefore, one should not use this score for the Gene Set Score.

There are also other methods for calculating a Gene Score that one can think about. Those can be subject of further studies.

An event for a gene in Gene Expression level occurs if its Gene Score is larger than the mean of all Gene Scores for all the genes.

## 9.2 Gene Score for Copy Number data $GS_{CN}$

Unlike the Gene Score for Gene Expression data the Gene Score for Copy Number data is more difficult. The values of CN data are mapped from continuous $log_2$ ratios to constant values between 0 and ±1 [33] and being discretized. An event represent an amplification or a deletion and is often set to values $< -0.9$ (defined as $s$) for deletion and $> +0.9$ (defined as $t$) for amplification. Let us assume that the two conditions in that example are 0 and 1 defined as $c_0$ and $c_1$. Then, the following four parameters can be defined for each gene $i$ as the number of samples that have one condition larger than the threshold divided by the number of samples of that condition:

$$\begin{aligned} a &= \frac{\#(c_0 \cap t)}{\#c_0} \\ b &= \frac{\#(c_0 \cap s)}{\#c_0} \\ c &= \frac{\#(c_1 \cap t)}{\#c_1} \\ d &= \frac{\#(c_1 \cap s)}{\#c_1} \end{aligned} \tag{9.5}$$

As one can see it is required to look at various steps to find the most common event, as long as this event does not occur in the other condition as well. This threshold is pre-defined to 0.3 and it can also be changed. This threshold of 30%

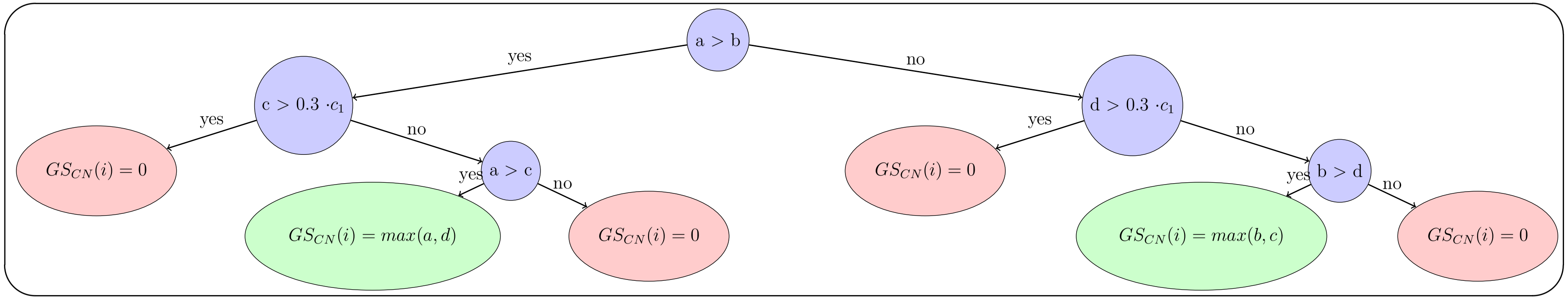


Figure 9.1: Tree diagram for the calculation of the Gene Score for Copy Number data

is set, because it makes no sense to define an event if it occurs in more than 30% of the other condition as well. In other words, an event is declared in $c_1$ if it occurs in at least 30% of samples in this condition but in less than 30% in another condition plus some more strict additional checks.
The Copy Number Gene Score in that data set, by comparing $c_0$ with $c_1$, has a mean of 0.1941. It is also possible to compare $c_1$ with $c_0$ where different results are the output, but are similar as one can see in Section 10.3.
The Gene Score for Copy Number data, $GS_{CN}$, is defined as:

$$GS_{CN}(i) = max(a, d) \tag{9.6}$$

for $a > b$, $c \leq 0.3 \cdot c_1$ and $a > c$. And for $a < b$, $d \leq 0.3 \cdot c_1$ and $b > d$:

$$GS_{CN}(i) = max(b, c) \tag{9.7}$$

otherwise

$$GS_{CN}(i) = 0 \tag{9.8}$$

## 9.3 Gene Score for a Gene Set

In the previous sections the Gene Score was introduced for one gene only. Normally one is not interested in the behavior of a single gene but in the behavior of a set of genes. Gene Sets are usually biological pathways that include various number of genes. The detailed structure of the pathways is not yet included in the GSS but is a possible topic for further studies. This could add more information for a pathway. For all it is important to determine a Gene Score for a gene set (for a given pathway) for Copy Number and Gene Expression data:
The Copy Number Score (CNS) and the Gene Expression Score (GES) is the sum of all available Genes Scores for data divided by the number of genes that are present in both GE and CN data, and their corresponding scores are in the interval $[0, 1]$. For a given pathway $j$ with $i = 1, \ldots, n$ genes the scores can be defined as follows:
For the Gene Expression Score for a pathway $j$

$$GES(j) = \frac{\sum_{i=1}^{n} GS_{GE}(i)}{\sum_{i=1}^{n} I_{GS_{GE}(i) \geq 0}} \tag{9.9}$$

and for the Copy Number Score for a pathway $j$.

$$CNS(j) = \frac{\sum_{i=1}^{n} GS_{CN}(i)}{\sum_{i=1}^{n} I_{GS_{CN}(i) \geq 0}} \tag{9.10}$$

For the overall analysis of the data set one uses all the pathways as described from Section 8.3 and calculates the GES and CNS.

The threshold of 10 genes is not a fixed threshold and thresholds of 20, or 40 genes are used in the analysis as well. But the threshold should be at least 10 genes because otherwise one or two genes that have a large Gene Score can substantially affect the GES and/or CNS.

The subject of the next chapter is how the Gene Set Score is used for identifying deregulated pathways.

# CHAPTER 10

# Null hypothesis and results of combining two data types

Before the null hypothesis for the behavior of gene sets/pathways is formulated it is necessary to determine what one understands as an event in GE and CN data regarding the Gene Expression Score and the Copy Number Score.

## 10.1 Events in CNS and GES

There are several ways to determine an event that occurs/is observed/measured in both data types. One can say that an event occurs if the score is larger than the mean score for all genes. The following two definitions are the result of it.

### 10.1.1 Gene Expression Event (GEE)

A Gene Expression Event (GEE) occurs if the Gene Expression Score (GES) for a given gene set/pathway $j$ is larger than the mean (on expression level) of the $GS_{GE}$ for all N genes.

$$GEE = \begin{cases} yes & GES(j) > \frac{1}{N}\sum_{i=1}^{N} GS_{GE}(i) \\ no & GES(j) \leq \frac{1}{N}\sum_{i=1}^{N} GS_{GE}(i) \end{cases} \tag{10.1}$$

### 10.1.2 Copy Number Event (CNE)

There is an event in Copy Number data (CNE) if the Copy Number Score (CNS) for a pathway/gene set $j$ is larger than the mean of the $GS_{CN}$ of all M genes, that does not compulsory have to be the same number of genes as for GEE.

$$CNE = \begin{cases} yes & CNS(j) > \frac{1}{M}\sum_{i=1}^{M} GS_{CN}(i) \\ no & CNS(j) \leq \frac{1}{M}\sum_{i=1}^{M} GS_{CN}(i) \end{cases} \tag{10.2}$$

## 10.2 Null Hypothesis $H_0$

If one is working with pathways different analyses are possible. There are (at least) four different ways on how the analysis can be done:

1. Working with Gene Expression data.

2. Working with Gene Expression data including structure.

3. Working with Gene Expression data and Copy Number data.

4. Working with Gene Expression data and Copy Number data including structure.

A lot of research is done for the first method including GSEA [31], GSA [8], SPIA [32]. For the second method also various work was done including the work of Kühnle and Pfundstein [20] as well as these methods are described in Section 7.1. Here the GSS is computed using the third method as not much work has so far been done on integrating GE and CN data. We formulate the null hypothesis as one of the most important questions in analyzing high-throuhput data sets/pathways in cancer and other complex diseases is to identify de-regulated gene sets/pathways.

**$\mathbf{H_0}$**:
(1) $H_0$: Pathways/Gene sets that have at least one CNE are as likely to have a GEE as all other gene sets.
(2) $H_0$: The relationships of genes in the pathways do not change between two conditions (or phenotypes).

## 10.3 Results of combining CNS and GES

To test the null hypothesis

$H_0$: Pathways/Gene sets that have a CNE are as likely to have a GEE as all other gene sets.

one can proceed as follows:
Out of 2801 pathways from Section 8.3, 1902 have at least 10 genes in both of the data types (GE and CN) available to calculate the CNS and GES. As one can see in Section 8.2 it is possible to compare 'PRIMARY' with 'MET' as well as 'MET' with 'PRIMARY'. As it turns out that the rates for a CNE accompanied by a GEE are similar (both ratios are equal to 0.74) it is not necessary to check out the different ways each time. The results are shown in Table 10.1 for all the pathways (PWs).
There is more than one possibility to calculate the Gene Score for GE data and therefore the ratios are also calculated with the 'Interval Method' with Equation 9.4. The total ratios (0.60, 0.67) are still better than a random picking, but not as good as the other ones. The exact results are available in Appendix B.1. The next step is to come up with a score and show whether the pathways with a Copy Number Event and a Gene Expression Event are significant different to other pathways. For this one introduces the Gene Set Score for a given pathway and compares this GSS with a GSS for a random set of genes.

|  |  |  | 'PRIMARY' with 'MET' |  |  | 'MET' with 'PRIMARY' |  |  |
|---|---|---|---|---|---|---|---|---|
| Name | # PW | # PW $\geq 10$ | # CNE | # (CNE $\cap$ GEE) | Ratio | # CNE | # (CNE $\cap$ GEE) | Ratio |
| arms | 46 | 39 | 18 | 16 | 0.89 | 14 | 12 | 0.86 |
| cyto | 797 | 399 | 211 | 135 | 0.64 | 115 | 63 | 0.55 |
| proc | 24 | 23 | 16 | 14 | 0.88 | 12 | 12 | 1.00 |
| sega | 456 | 406 | 225 | 190 | 0.84 | 180 | 171 | 0.95 |
| tile | 1192 | 820 | 438 | 302 | 0.69 | 251 | 160 | 0.64 |
| tiss | 80 | 73 | 50 | 47 | 0.94 | 43 | 39 | 0.91 |
| kegg | 206 | 142 | 92 | 68 | 0.74 | 58 | 41 | 0.71 |
| TOTAL | 2801 | 1902 | 1050 | 772 | 0.74 | 673 | 498 | 0.74 |

Table 10.1: Ratio of a CNE is accompanied by a GEE

# CHAPTER 11

# Gene Set Score (GSS)

## 11.1 Gene Set Score

In the last chapters the CNS and GES were introduced and the results presented. This chapter is combining these two scores into one score. The name for this combined score is Gene Set Score (GSS). For this score various possibilities are worth thinking about as this is a novel score and the best one was chosen.
In the GSS it should play a role how the CNS behaves compared to a random CNS as well as how the GES behaves compared to a random GES. Therefore, ratios of the CNS against a random set are used as well as ratios of the GES against a random set. The random sets are equal to the mean over all genes by the Gene Score of GE or CN times the number of genes within the regarded pathway. The final GSS comes up with a product of two ratios including indicator functions of each ratio. Using a sum instead of the product also provides nearly the same results as including indicator functions, but using indicator functions slightly increases the quality of the results and is therefore also included. The indicator function is not just included because of getting better results. It is included, because it shows that there is a Copy Number Event accompanied by a Gene Expression Event which one was investigating in the previous sections. Because comparing 'MET' with 'PRIMARY' and vice versa is the same one just focuses in on one direction for the GSS. So the total Gene Set Score for a given pathway/gene set $j$ with $n$ genes is the following as there are $i = 1, \ldots, N$ genes in Gene Expression data and $i = 1, \ldots, M$ genes in Copy Number data:

$$GSS(j) = \underbrace{\frac{CNS(j)}{n \cdot \frac{1}{M} \cdot \sum_{i=1}^{M} GS_{CN}(i)}}_{y} \cdot I_{y>1}(y) \cdot \underbrace{\frac{GES(j)}{n \cdot \frac{1}{N} \cdot \sum_{i=1}^{N} GS_{GE}(i)}}_{z} \cdot I_{z>1}(z) \quad (11.1)$$

## 11.2 GSS Random

As the sample size is often not big enough for a randomization one just focuses on the randomization of the genes. A combined randomization of genes and samples would also be possible. For the random Gene Set Score $GSS_{Random}(j)$ for a pathway $j$ with $n$ genes one samples randomly picks $n$ genes from the available data set of all genes $K = 1000$ times and calculates the $GSS_{Random(k)}(j)$ each time:

$$GSS_{Random}(j) = (GSS_{Random(1)}(j), \ldots, GSS_{Random(1000)}(j))^T \tag{11.2}$$

The $GSS_{Random(k)}(j)$ is calculated with the same equation as the 'normal' GSS. With that vector of $GSS_{Random}(j)$ it is then possible to calculate significant pathways.

## 11.3 Significant different pathways with the GSS

As a reminder the null hypothesis one wants to test is:

$H_0$: The relationships of genes in the pathways do not change between two conditions (or phenotypes).

To calculate the $p$-values from the test of the null hypothesis it is required to calculate a random Gene Set Score as well. These $p$-values result from those pathways which have a CNS and GES which are both larger than for a random pathway, because indicator functions are added. The result of this is only pathways that have a Copy Number Event and a Gene Expression Event that are maybe significant.

The $p$-values are calculated as in the following equation for one pathway $j$:

$$p - value(j) = \frac{\sum_{k=1}^{K} I_{GSS_{Random(k)}(j) > GSS(j)}}{K} \tag{11.3}$$

As one can see the $p$-value is the fraction of the random GSS of all permutations being larger than the GSS and the number of permutations.

Different thresholds for the $p$-values are possible by comparing those gene sets/ pathways with the pathways which are differently expressed with GSA. Therefore three different thresholds for the $p$-values are included in the results in Table 11.1. This analysis is done with pathways with at least 10 genes. As it is also interesting how the results change if one is looking at pathways that are larger, Table B.2 and Table B.3 in the appendix show how the results look for pathways with at least 20 or 40 genes.

The ratio of significant pathways of GSS (see the definition above) is increasing if the *p*-values are getting larger which is no surprise. With a significance level of 0.05 already 74% of the pathways are significantly different than random pathways if they have a CNE and a GEE. By increasing this level to a threshold of 0.10 87% of all pathways with a Copy Number Event accompanied by a Gene Expression Event are significant.

Increasing the minimum number of genes in the pathways also increases the significance rate. This may be the result of the increased power of the GSS could also be the result of the decreasing affect of the randomization. The randomization, as a reminder, randomly picks as many genes as in the pathway to compare with. The smaller the pathway the larger the possible affect of one gene (or more genes) affecting the GSS distinctly. This can be shown as the affect getting smaller if the *p*-value is getting larger. This needs some further analysis with other data sets. Including a parameter which adjusts the number of genes within the randomization is a possibility to handle this. For the different results one can compare Table 11.1 to Tables B.2 and B.3 in the appendix.

|  | # (CNE | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| Name | ∩ GEE) | $p < 0.01$ |  | $p < 0.05$ |  | $p < 0.10$ |  |
| arms | 16 | 16 | 1.00 | 16 | 1.00 | 16 | 1.00 |
| cyto | 135 | 61 | 0.45 | 100 | 0.74 | 124 | 0.92 |
| proc | 14 | 11 | 0.79 | 12 | 0.86 | 12 | 0.86 |
| sega | 190 | 110 | 0.58 | 147 | 0.77 | 169 | 0.89 |
| tile | 302 | 157 | 0.52 | 224 | 0.74 | 259 | 0.86 |
| tiss | 47 | 29 | 0.62 | 39 | 0.83 | 44 | 0.94 |
| kegg | 68 | 12 | 0.18 | 33 | 0.49 | 49 | 0.72 |
| TOTAL | 772 | 396 | 0.51 | 571 | 0.74 | 673 | 0.87 |

Table 11.1: Significant pathways for different thresholds for the $p$-value with at least 10 genes per pathway

# CHAPTER 12

# Comparing the Gene Set Score with the Gene Set Analysis

## 12.1 Introduction

A comparison of significant pathways of the GSS with the significant pathways found by the GSA is an important step to validate the results and show the quality of the Gene Set Score. As introduced in Section 7.1.3 GSA is an important and well known method to detect significantly different pathways. The goal is that the novel Gene Set Score detects most of the significant pathways that are detected by the Gene Set Analysis plus additional pathways that GSA does not detect. There are two thresholds, one is the *p*-value of GSS and one is the *p*-value of GSA, that one has to set to a value before the analysis to get a ratio of how many significant pathways of the GSA method are also found with the GSS method.

## 12.2 Results

The *p*-value of the GSA was set to 0.01 as this yields to the most significant pathways that are found with GSA. The smaller the *p*-value for GSS the less accordance of pathways found with both methods. Though this effect is not that large and even for a *p*-value of 0.01 for the GSS there is still a ratio of 0.66 of detected pathways by the GSS of the significant pathways of the GSA. For a *p*-value of 0.10 this ratio increases to 74%. Increasing the numbers of genes for a pathway increases the ratio most of the time but less pathways fulfill the selection as the number of pathways is getting lower by increasing the minimum number of genes in each pathway. Results with a *p*-value of 0.01 for GSA and 10 genes can be found in Table 12.1. The results of 20 or 40 genes can be found in the appendix in Tables B.4 and B.5.

| Name | # PW<br>$p_{GSA} < 0.01$ | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p_{GSS} < 0.01$ | | $p_{GSS} < 0.05$ | | $p_{GSS} < 0.10$ | |
| arms | 0 | 0 | - | 0 | - | 0 | - |
| cyto | 6 | 3 | 0.50 | 3 | 0.50 | 4 | 0.67 |
| proc | 3 | 3 | 1.00 | 3 | 1.00 | 3 | 1.00 |
| sega | 33 | 24 | 0.73 | 25 | 0.76 | 25 | 0.76 |
| tile | 16 | 8 | 0.50 | 9 | 0.56 | 10 | 0.62 |
| tiss | 5 | 4 | 0.80 | 4 | 0.80 | 4 | 0.80 |
| kegg | 5 | 3 | 0.60 | 3 | 0.60 | 4 | 0.80 |
| TOTAL | 68 | 45 | 0.66 | 47 | 0.69 | 50 | 0.74 |

Table 12.1: Detected significant pathways of the GSA ($p_{GSA} < 0.01$) by significant GSS with at least 10 genes

In Tables B.6 and B.7 in the appendix one can also compare the results with results of a Gene Set Score with a significance level of 0.05 or 0.10. The ratios are lower than the ratios for a significance level of 0.01 what shows that GSS detects primarily pathways that are highly significant with the Gene Set Analysis and does not focus at the low significant pathways plus additional other pathways that one does not get by using the GSA.

It is also interesting to see that there are zero significant pathways with the Gene Set Analysis for the 'arms' pathways. Looking at this source shows that these pathways contain couple of hundreds up to some thousands (!) of genes. For some reason GSA does not detect those large pathways as significant.

Depending on the thresholds for the *p*-value and minimum number of genes the Gene Set Score detects up to 81% of the pathways that the Gene Set Analysis also detect. As the primary version with a *p*-values of 0.01 for the GSA and a *p*-value of 0.10 for the GSS with a minimum of 10 genes per pathway still finds 74% of the pathways and shows that the GSS is a good method using two data types to find differentially expressed pathways.

## 12.3 Further studies

As mentioned above further studies should include a parameter that is dealing with the influence of the randomization. The randomization, as far as the results seem to indicate and one thinks about, plays a small role in the way that for pathways with fewer genes one gets not as many significant pathways as for pathways with more genes.

Another way to continue working on the Gene Set Score is to take the structure of pathways into account. There are a lot of possibilities of how one may add this into the Gene Set Score. It should obviously just be included if the power is increased, such that the ratio of detected significant pathways of the GSA by the GSS is increased.

Comparing the GSS with the ACE it is not necessary to deal with neighboring genes to calculate a gene score based on the neighboring genes. But this idea can be somehow also implemented in the way of how pathways interact with each other as far as data is available for this.

Not available data for a gene is maybe also not treated in a way of how it should be treated and loss in the power is the result. If there is no data available for one gene of a pathway in either the Gene Expression data, the Copy Number data or in both of the data types one just ignores this gene for the Gene Set Score. In the version of the Gene Scores for a gene set the sum of all available Gene Scores divided by the number of available genes is calculated. But there is maybe a reason why data is not available for some genes. This is particular a problem if the data sets are not as big as the data sets used in this analysis.

## 12.4 Summary and Conclusion

The new developed Gene Set Score is a powerful score to detect most of the already known pathways that are significantly different between two conditions or phenotypes in the Gene Set Analysis as well as the GSS finds more pathways that are not significant with GSA but with the GSS.

First one is looking at Copy Number data and whether there is an event for a given gene set. If there is an event one checks if there is also an event in the Gene Expression data. This accompanied event occurs in three quarters of the CNEs. The results of both data types are then combined with the novel Gene Set Score. With this score one is able to detect significantly different (between two conditions or phenotypes) pathways using two data types instead of just one as the Gene Set Analysis does.

As a conclusion it is possible to say that looking at two data types at once to solve the problem of finding significantly different pathways adds more information about the data and therefore increases the power of the quality of these pathways. This yields to detecting the most relevant pathways that are also found with just one data type as well as additional pathways that are not detected as significantly different with just one data type.

# APPENDIX A

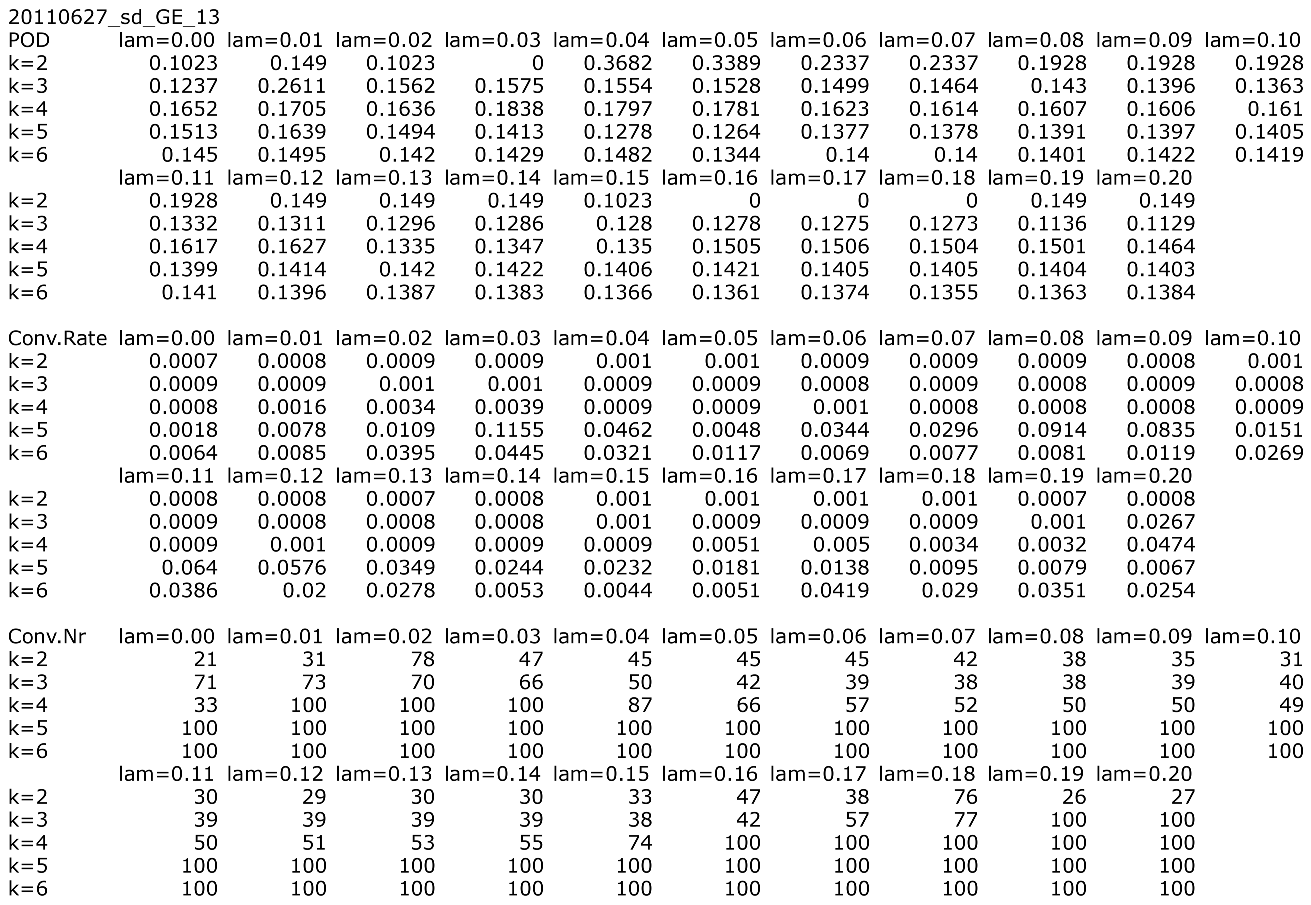
# Results iCluster

## A.1 POD, Conv.Rate and Conv.Nr

In this section, all the results for using iCluster with the Sarcoma data set including 37 samples are shown. Results include the POD, Conv.Rate and Conv.Nr for the following subsets of the original data:

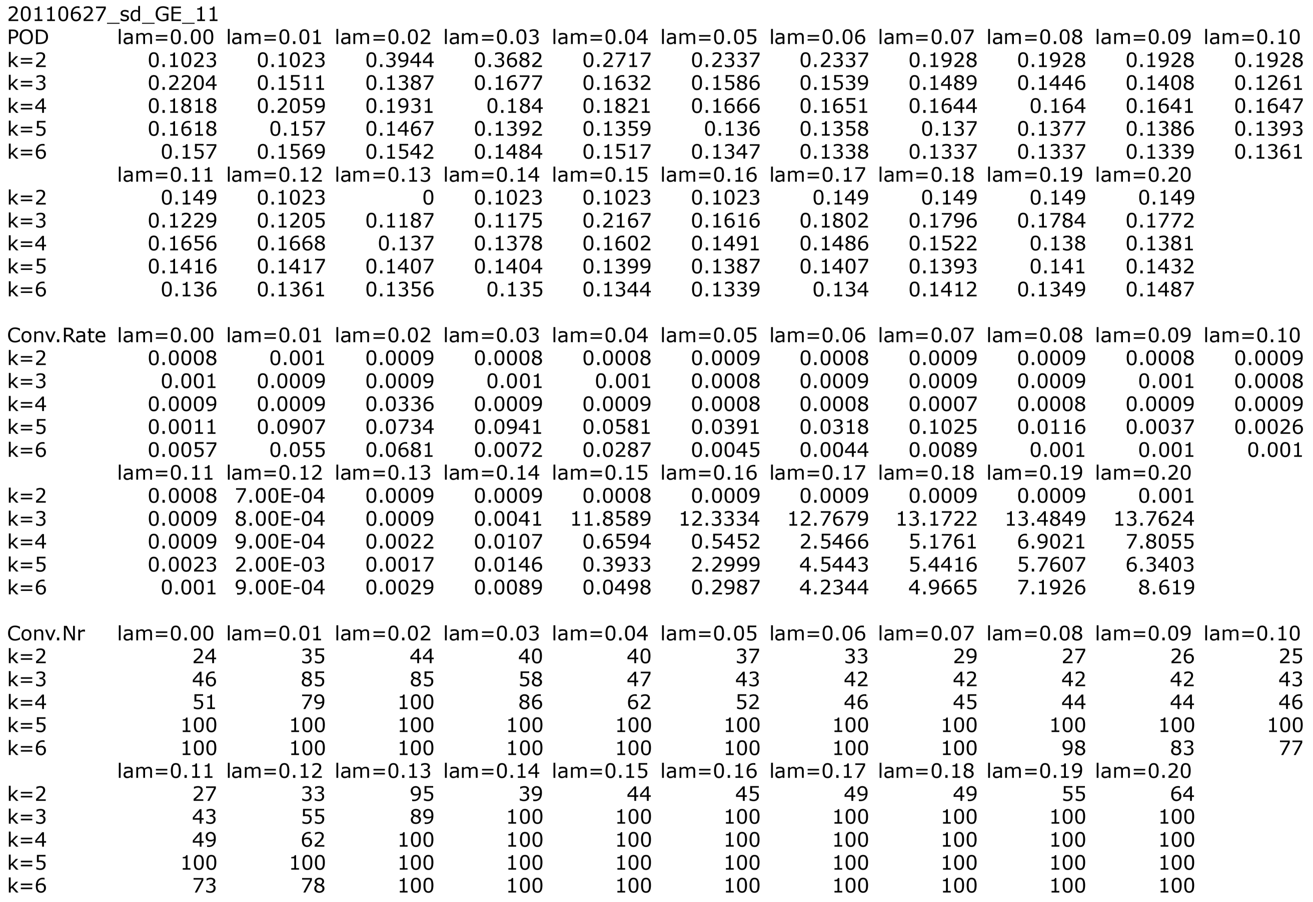
- sd_GE_13
- sd_GE_11
- sd_GE_15
- ME437
- ME637
- MSV7
- MSV9
- STD4
- SDK4
- SDK6
- SDK10
- Random

With the represented results one is able to figure out which of the 21 lambda is the best for each of the $k = 2, \ldots, 6$. Conv.Nr equal to 100 occurs, if the Conv.Rate was above the pre-defined threshold. One should look at these specific plots to figure out if it is still convergent or not (for another threshold or a larger Conv.Nr.). NAs occur for the reason: 'more cluster centers than distinct data points'. This is an error from the R-package iCluster from Shen [27].



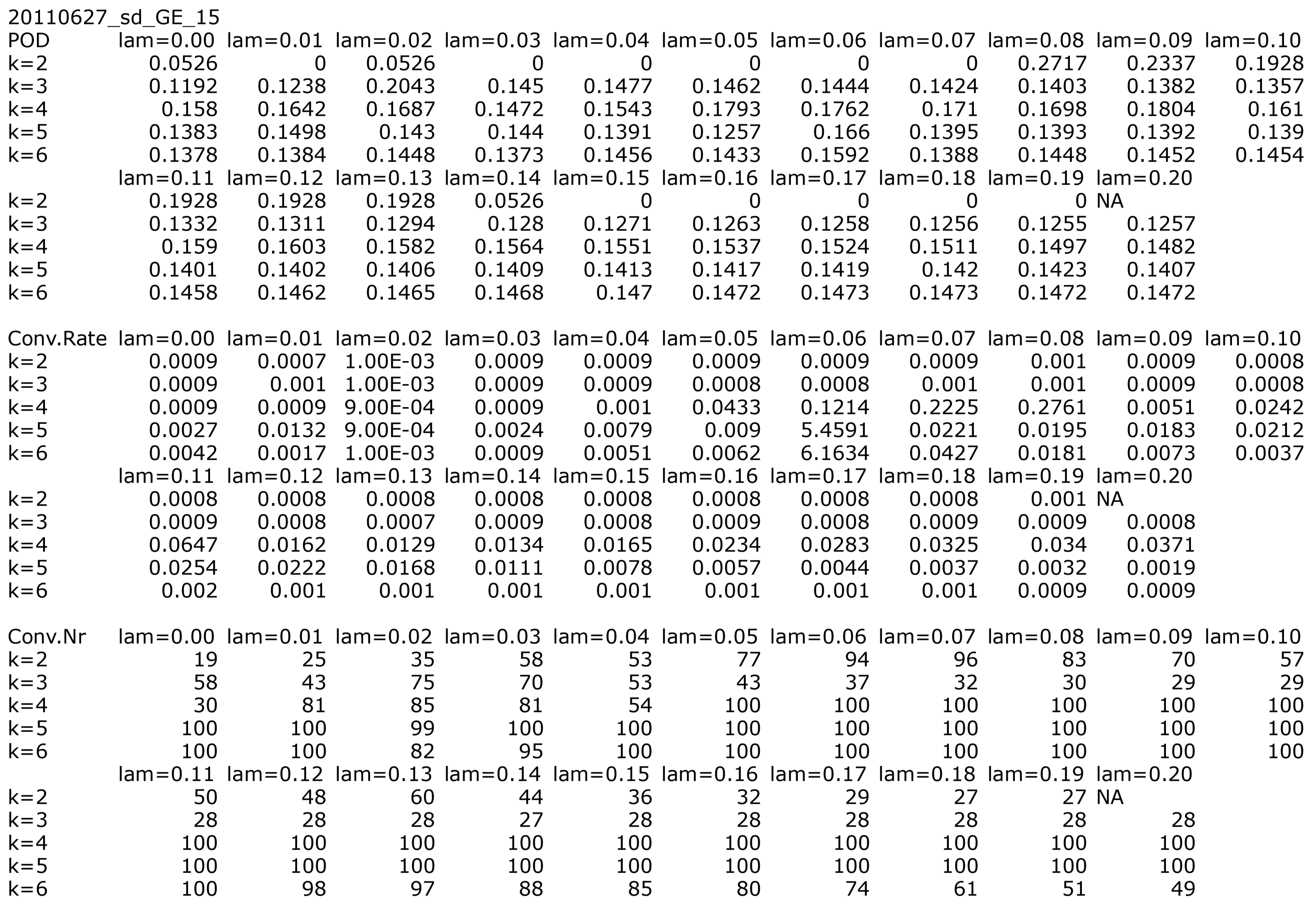
| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 20110627_sd_GE_13 | | | | | | | | | | | |
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.149 | 0.1023 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1237 | 0.2611 | 0.1562 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1652 | 0.1705 | 0.1636 | 0.1838 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1513 | 0.1639 | 0.1494 | 0.1413 | 0.1278 | 0.1264 | 0.1377 | 0.1378 | 0.1391 | 0.1397 | 0.1405 |
| k=6 | 0.145 | 0.1495 | 0.142 | 0.1429 | 0.1482 | 0.1344 | 0.14 | 0.14 | 0.1401 | 0.1422 | 0.1419 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | 0 | 0 | 0 | 0.149 | 0.149 | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1505 | 0.1506 | 0.1504 | 0.1501 | 0.1464 | |
| k=5 | 0.1399 | 0.1414 | 0.142 | 0.1422 | 0.1406 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.141 | 0.1396 | 0.1387 | 0.1383 | 0.1366 | 0.1361 | 0.1374 | 0.1355 | 0.1363 | 0.1384 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0007 | 0.0008 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.001 |
| k=3 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.0008 | 0.0016 | 0.0034 | 0.0039 | 0.0009 | 0.0009 | 0.001 | 0.0008 | 0.0008 | 0.0008 | 0.0009 |
| k=5 | 0.0018 | 0.0078 | 0.0109 | 0.1155 | 0.0462 | 0.0048 | 0.0344 | 0.0296 | 0.0914 | 0.0835 | 0.0151 |
| k=6 | 0.0064 | 0.0085 | 0.0395 | 0.0445 | 0.0321 | 0.0117 | 0.0069 | 0.0077 | 0.0081 | 0.0119 | 0.0269 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0008 | 0.0008 | 0.0007 | 0.0008 | 0.001 | 0.001 | 0.001 | 0.001 | 0.0007 | 0.0008 | |
| k=3 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0267 | |
| k=4 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0051 | 0.005 | 0.0034 | 0.0032 | 0.0474 | |
| k=5 | 0.064 | 0.0576 | 0.0349 | 0.0244 | 0.0232 | 0.0181 | 0.0138 | 0.0095 | 0.0079 | 0.0067 | |
| k=6 | 0.0386 | 0.02 | 0.0278 | 0.0053 | 0.0044 | 0.0051 | 0.0419 | 0.029 | 0.0351 | 0.0254 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 21 | 31 | 78 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 71 | 73 | 70 | 66 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 33 | 100 | 100 | 100 | 87 | 66 | 57 | 52 | 50 | 50 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 30 | 29 | 30 | 30 | 33 | 47 | 38 | 76 | 26 | 27 | |
| k=3 | 39 | 39 | 39 | 39 | 38 | 42 | 57 | 77 | 100 | 100 | |
| k=4 | 50 | 51 | 53 | 55 | 74 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |

Figure A.1: POD, Conv.Rate and Conv.Nr of sd_GE_13



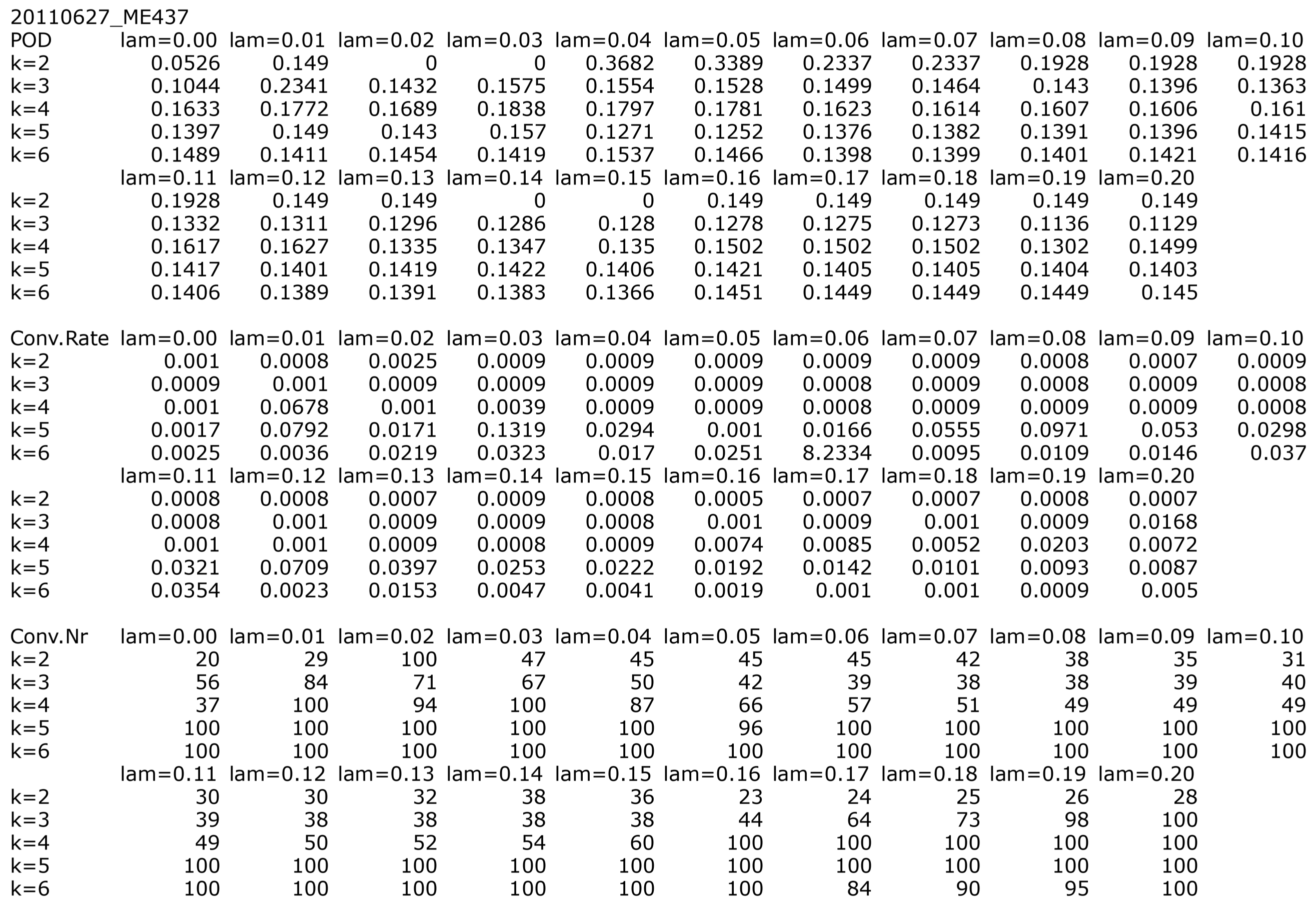
| 20110627_sd_GE_11 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.1023 | 0.3944 | 0.3682 | 0.2717 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.2204 | 0.1511 | 0.1387 | 0.1677 | 0.1632 | 0.1586 | 0.1539 | 0.1489 | 0.1446 | 0.1408 | 0.1261 |
| k=4 | 0.1818 | 0.2059 | 0.1931 | 0.184 | 0.1821 | 0.1666 | 0.1651 | 0.1644 | 0.164 | 0.1641 | 0.1647 |
| k=5 | 0.1618 | 0.157 | 0.1467 | 0.1392 | 0.1359 | 0.136 | 0.1358 | 0.137 | 0.1377 | 0.1386 | 0.1393 |
| k=6 | 0.157 | 0.1569 | 0.1542 | 0.1484 | 0.1517 | 0.1347 | 0.1338 | 0.1337 | 0.1337 | 0.1339 | 0.1361 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.149 | 0.1023 | 0 | 0.1023 | 0.1023 | 0.1023 | 0.149 | 0.149 | 0.149 | 0.149 | |
| k=3 | 0.1229 | 0.1205 | 0.1187 | 0.1175 | 0.2167 | 0.1616 | 0.1802 | 0.1796 | 0.1784 | 0.1772 | |
| k=4 | 0.1656 | 0.1668 | 0.137 | 0.1378 | 0.1602 | 0.1491 | 0.1486 | 0.1522 | 0.138 | 0.1381 | |
| k=5 | 0.1416 | 0.1417 | 0.1407 | 0.1404 | 0.1399 | 0.1387 | 0.1407 | 0.1393 | 0.141 | 0.1432 | |
| k=6 | 0.136 | 0.1361 | 0.1356 | 0.135 | 0.1344 | 0.1339 | 0.134 | 0.1412 | 0.1349 | 0.1487 | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0008 | 0.001 | 0.0009 | 0.0008 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0009 | 0.0008 | 0.0009 |
| k=3 | 0.001 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0008 |
| k=4 | 0.0009 | 0.0009 | 0.0336 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0007 | 0.0008 | 0.0009 | 0.0009 |
| k=5 | 0.0011 | 0.0907 | 0.0734 | 0.0941 | 0.0581 | 0.0391 | 0.0318 | 0.1025 | 0.0116 | 0.0037 | 0.0026 |
| k=6 | 0.0057 | 0.055 | 0.0681 | 0.0072 | 0.0287 | 0.0045 | 0.0044 | 0.0089 | 0.001 | 0.001 | 0.001 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0008 | 7.00E-04 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.001 | |
| k=3 | 0.0009 | 8.00E-04 | 0.0009 | 0.0041 | 11.8589 | 12.3334 | 12.7679 | 13.1722 | 13.4849 | 13.7624 | |
| k=4 | 0.0009 | 9.00E-04 | 0.0022 | 0.0107 | 0.6594 | 0.5452 | 2.5466 | 5.1761 | 6.9021 | 7.8055 | |
| k=5 | 0.0023 | 2.00E-03 | 0.0017 | 0.0146 | 0.3933 | 2.2999 | 4.5443 | 5.4416 | 5.7607 | 6.3403 | |
| k=6 | 0.001 | 9.00E-04 | 0.0029 | 0.0089 | 0.0498 | 0.2987 | 4.2344 | 4.9665 | 7.1926 | 8.619 | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 24 | 35 | 44 | 40 | 40 | 37 | 33 | 29 | 27 | 26 | 25 |
| k=3 | 46 | 85 | 85 | 58 | 47 | 43 | 42 | 42 | 42 | 42 | 43 |
| k=4 | 51 | 79 | 100 | 86 | 62 | 52 | 46 | 45 | 44 | 44 | 46 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 98 | 83 | 77 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 27 | 33 | 95 | 39 | 44 | 45 | 49 | 49 | 55 | 64 | |
| k=3 | 43 | 55 | 89 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=4 | 49 | 62 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 73 | 78 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |

Figure A.2: POD, Conv.Rate and Conv.Nr of sd_GE_11



| 20110627_sd_GE_15 | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0526 | 0 | 0.0526 | 0 | 0 | 0 | 0 | 0 | 0.2717 | 0.2337 | 0.1928 |
| k=3 | 0.1192 | 0.1238 | 0.2043 | 0.145 | 0.1477 | 0.1462 | 0.1444 | 0.1424 | 0.1403 | 0.1382 | 0.1357 |
| k=4 | 0.158 | 0.1642 | 0.1687 | 0.1472 | 0.1543 | 0.1793 | 0.1762 | 0.171 | 0.1698 | 0.1804 | 0.161 |
| k=5 | 0.1383 | 0.1498 | 0.143 | 0.144 | 0.1391 | 0.1257 | 0.166 | 0.1395 | 0.1393 | 0.1392 | 0.139 |
| k=6 | 0.1378 | 0.1384 | 0.1448 | 0.1373 | 0.1456 | 0.1433 | 0.1592 | 0.1388 | 0.1448 | 0.1452 | 0.1454 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.1928 | 0.1928 | 0.0526 | 0 | 0 | 0 | 0 | 0 | NA | |
| k=3 | 0.1332 | 0.1311 | 0.1294 | 0.128 | 0.1271 | 0.1263 | 0.1258 | 0.1256 | 0.1255 | 0.1257 | |
| k=4 | 0.159 | 0.1603 | 0.1582 | 0.1564 | 0.1551 | 0.1537 | 0.1524 | 0.1511 | 0.1497 | 0.1482 | |
| k=5 | 0.1401 | 0.1402 | 0.1406 | 0.1409 | 0.1413 | 0.1417 | 0.1419 | 0.142 | 0.1423 | 0.1407 | |
| k=6 | 0.1458 | 0.1462 | 0.1465 | 0.1468 | 0.147 | 0.1472 | 0.1473 | 0.1473 | 0.1472 | 0.1472 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0009 | 0.0007 | 1.00E-03 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0009 | 0.0008 |
| k=3 | 0.0009 | 0.001 | 1.00E-03 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.001 | 0.001 | 0.0009 | 0.0008 |
| k=4 | 0.0009 | 0.0009 | 9.00E-04 | 0.0009 | 0.001 | 0.0433 | 0.1214 | 0.2225 | 0.2761 | 0.0051 | 0.0242 |
| k=5 | 0.0027 | 0.0132 | 9.00E-04 | 0.0024 | 0.0079 | 0.009 | 5.4591 | 0.0221 | 0.0195 | 0.0183 | 0.0212 |
| k=6 | 0.0042 | 0.0017 | 1.00E-03 | 0.0009 | 0.0051 | 0.0062 | 6.1634 | 0.0427 | 0.0181 | 0.0073 | 0.0037 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.001 | NA | |
| k=3 | 0.0009 | 0.0008 | 0.0007 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0009 | 0.0008 | |
| k=4 | 0.0647 | 0.0162 | 0.0129 | 0.0134 | 0.0165 | 0.0234 | 0.0283 | 0.0325 | 0.034 | 0.0371 | |
| k=5 | 0.0254 | 0.0222 | 0.0168 | 0.0111 | 0.0078 | 0.0057 | 0.0044 | 0.0037 | 0.0032 | 0.0019 | |
| k=6 | 0.002 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.001 | 0.0009 | 0.0009 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 19 | 25 | 35 | 58 | 53 | 77 | 94 | 96 | 83 | 70 | 57 |
| k=3 | 58 | 43 | 75 | 70 | 53 | 43 | 37 | 32 | 30 | 29 | 29 |
| k=4 | 30 | 81 | 85 | 81 | 54 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=5 | 100 | 100 | 99 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 82 | 95 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 50 | 48 | 60 | 44 | 36 | 32 | 29 | 27 | 27 | NA | |
| k=3 | 28 | 28 | 28 | 27 | 28 | 28 | 28 | 28 | 28 | 28 | |
| k=4 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 98 | 97 | 88 | 85 | 80 | 74 | 61 | 51 | 49 | |

Figure A.3: POD, Conv.Rate and Conv.Nr of sd_GE_15



| 20110627_ME437 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0526 | 0.149 | 0 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1044 | 0.2341 | 0.1432 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1633 | 0.1772 | 0.1689 | 0.1838 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1397 | 0.149 | 0.143 | 0.157 | 0.1271 | 0.1252 | 0.1376 | 0.1382 | 0.1391 | 0.1396 | 0.1415 |
| k=6 | 0.1489 | 0.1411 | 0.1454 | 0.1419 | 0.1537 | 0.1466 | 0.1398 | 0.1399 | 0.1401 | 0.1421 | 0.1416 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0 | 0 | 0.149 | 0.149 | 0.149 | 0.149 | 0.149 | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1502 | 0.1502 | 0.1502 | 0.1302 | 0.1499 | |
| k=5 | 0.1417 | 0.1401 | 0.1419 | 0.1422 | 0.1406 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.1406 | 0.1389 | 0.1391 | 0.1383 | 0.1366 | 0.1451 | 0.1449 | 0.1449 | 0.1449 | 0.145 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.001 | 0.0008 | 0.0025 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0007 | 0.0009 |
| k=3 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.001 | 0.0678 | 0.001 | 0.0039 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0008 |
| k=5 | 0.0017 | 0.0792 | 0.0171 | 0.1319 | 0.0294 | 0.001 | 0.0166 | 0.0555 | 0.0971 | 0.053 | 0.0298 |
| k=6 | 0.0025 | 0.0036 | 0.0219 | 0.0323 | 0.017 | 0.0251 | 8.2334 | 0.0095 | 0.0109 | 0.0146 | 0.037 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0008 | 0.0008 | 0.0007 | 0.0009 | 0.0008 | 0.0005 | 0.0007 | 0.0007 | 0.0008 | 0.0007 | |
| k=3 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.0008 | 0.001 | 0.0009 | 0.001 | 0.0009 | 0.0168 | |
| k=4 | 0.001 | 0.001 | 0.0009 | 0.0008 | 0.0009 | 0.0074 | 0.0085 | 0.0052 | 0.0203 | 0.0072 | |
| k=5 | 0.0321 | 0.0709 | 0.0397 | 0.0253 | 0.0222 | 0.0192 | 0.0142 | 0.0101 | 0.0093 | 0.0087 | |
| k=6 | 0.0354 | 0.0023 | 0.0153 | 0.0047 | 0.0041 | 0.0019 | 0.001 | 0.001 | 0.0009 | 0.005 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 20 | 29 | 100 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 56 | 84 | 71 | 67 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 37 | 100 | 94 | 100 | 87 | 66 | 57 | 51 | 49 | 49 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 96 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 30 | 30 | 32 | 38 | 36 | 23 | 24 | 25 | 26 | 28 | |
| k=3 | 39 | 38 | 38 | 38 | 38 | 44 | 64 | 73 | 98 | 100 | |
| k=4 | 49 | 50 | 52 | 54 | 60 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 84 | 90 | 95 | 100 | |

Figure A.4: POD, Conv.Rate and Conv.Nr of ME437

20110627_ME637

| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 0.1023 | 0.149 | 0.1023 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1138 | 0.2647 | 0.1546 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1661 | 0.1677 | 0.1631 | 0.1837 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1519 | 0.1647 | 0.1505 | 0.1405 | 0.1277 | 0.1263 | 0.1377 | 0.1378 | 0.139 | 0.1398 | 0.1404 |
| k=6 | 0.1459 | 0.152 | 0.1426 | 0.1436 | 0.1485 | 0.1345 | 0.14 | 0.14 | 0.1401 | 0.1422 | 0.1419 |
|  | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 |  |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | 0 | 0 | 0 | 0.149 | 0.149 |  |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 |  |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1505 | 0.1506 | 0.1504 | 0.1501 | 0.1494 |  |
| k=5 | 0.1399 | 0.1414 | 0.142 | 0.1422 | 0.1406 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 |  |
| k=6 | 0.1411 | 0.1397 | 0.1387 | 0.1383 | 0.1366 | 0.1361 | 0.1428 | 0.1356 | 0.1364 | 0.1384 |  |

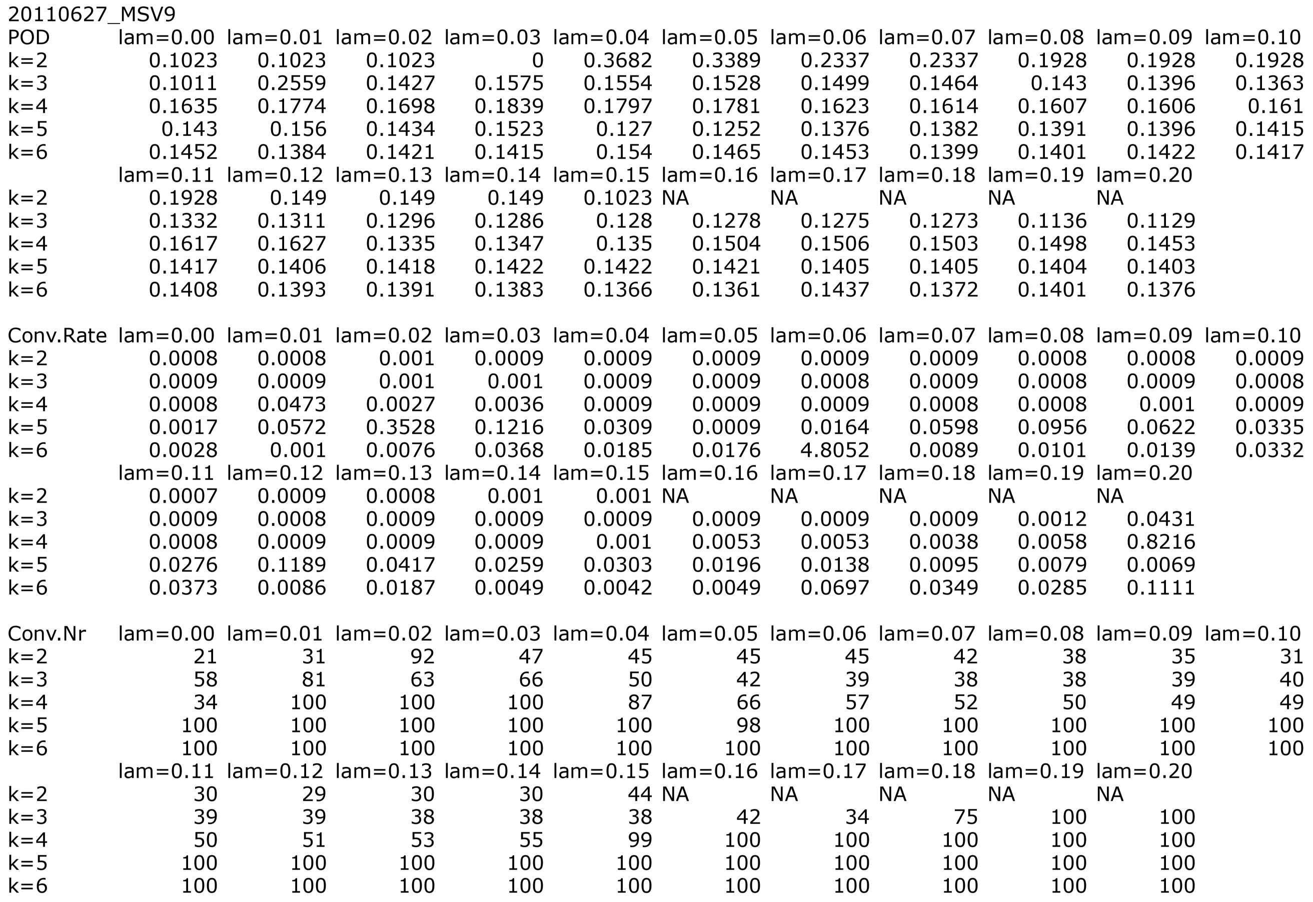
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 0.0008 | 0.0009 | 0.001 | 0.0009 | 0.001 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0008 | 0.001 |
| k=3 | 0.0009 | 0.001 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.0009 | 0.0012 | 0.0028 | 0.0042 | 0.0009 | 0.001 | 0.001 | 0.0009 | 0.0008 | 0.0008 | 0.0009 |
| k=5 | 0.0018 | 0.0075 | 0.0265 | 0.1118 | 0.0467 | 0.0057 | 0.0366 | 0.0193 | 0.0802 | 0.0927 | 0.0859 |
| k=6 | 0.003 | 0.0063 | 0.0403 | 0.0446 | 0.0325 | 0.0096 | 0.0072 | 0.0078 | 0.0082 | 0.0119 | 0.0266 |
|  | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 |  |
| k=2 | 0.0008 | 0.0008 | 0.0007 | 0.0008 | 0.001 | 0.0008 | 0.0009 | 0.001 | 0.0008 | 0.0007 |  |
| k=3 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.001 | 0.0011 | 0.0259 |  |
| k=4 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.001 | 0.005 | 0.0049 | 0.0033 | 0.0031 | 0.0105 |  |
| k=5 | 0.0643 | 0.0564 | 0.0347 | 0.0244 | 0.023 | 0.0181 | 0.0139 | 0.0096 | 0.0079 | 0.0068 |  |
| k=6 | 0.0386 | 0.0212 | 0.0294 | 0.0054 | 0.0044 | 0.0051 | 0.1139 | 0.031 | 0.0351 | 0.0254 |  |

| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 21 | 31 | 73 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 75 | 70 | 74 | 66 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 33 | 100 | 100 | 100 | 87 | 66 | 57 | 52 | 50 | 50 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
|  | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 |  |
| k=2 | 30 | 29 | 30 | 30 | 33 | 49 | 57 | 76 | 26 | 27 |  |
| k=3 | 39 | 39 | 39 | 39 | 38 | 42 | 57 | 70 | 100 | 100 |  |
| k=4 | 50 | 51 | 53 | 55 | 82 | 100 | 100 | 100 | 100 | 100 |  |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |  |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |  |

Figure A.5: POD, Conv.Rate and Conv.Nr of ME637

| 20110627_MSV7 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.0526 | 0.1023 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1303 | 0.2555 | 0.143 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1617 | 0.1717 | 0.165 | 0.1839 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.148 | 0.1575 | 0.1444 | 0.1404 | 0.1269 | 0.1245 | 0.1381 | 0.1381 | 0.1391 | 0.1397 | 0.1421 |
| k=6 | 0.1445 | 0.1492 | 0.1406 | 0.1416 | 0.1547 | 0.146 | 0.14 | 0.1399 | 0.1401 | 0.1422 | 0.1418 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | 0.149 | 0 | 0.149 | 0.149 | 0.149 | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.113 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1504 | 0.1506 | 0.1504 | 0.1501 | 0.1297 | |
| k=5 | 0.1399 | 0.141 | 0.142 | 0.1422 | 0.1406 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.1409 | 0.1394 | 0.1392 | 0.1383 | 0.1366 | 0.1361 | 0.1367 | 0.1358 | 0.1371 | 0.14 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0006 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0009 |
| k=3 | 0.001 | 0.0009 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.0008 | 0.0105 | 0.0026 | 0.0035 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.001 | 0.0009 |
| k=5 | 0.0018 | 0.0235 | 0.0285 | 0.1177 | 0.0312 | 0.0009 | 0.0039 | 0.0499 | 0.097 | 0.0857 | 0.1165 |
| k=6 | 0.011 | 0.0035 | 0.0238 | 0.045 | 0.019 | 0.028 | 0.0076 | 0.0081 | 0.009 | 0.0129 | 0.0301 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0007 | 0.0008 | 0.0008 | 0.0009 | 0.0007 | 0.0006 | 0.0009 | 0.0006 | 0.0007 | 0.0008 | |
| k=3 | 0.0009 | 0.0008 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0011 | 0.0868 | |
| k=4 | 0.0008 | 0.0009 | 0.0008 | 0.0008 | 0.001 | 0.0053 | 0.0052 | 0.0036 | 0.0037 | 0.0269 | |
| k=5 | 0.0627 | 0.0744 | 0.0348 | 0.0244 | 0.024 | 0.018 | 0.0137 | 0.0094 | 0.0077 | 0.0067 | |
| k=6 | 0.0378 | 0.0119 | 0.0211 | 0.005 | 0.0043 | 0.0049 | 0.042 | 0.0339 | 0.0313 | 0.0332 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 21 | 30 | 87 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 59 | 78 | 63 | 66 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 33 | 100 | 100 | 100 | 87 | 66 | 57 | 52 | 50 | 49 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 30 | 29 | 30 | 30 | 36 | 28 | 56 | 26 | 26 | 27 | |
| k=3 | 39 | 39 | 38 | 38 | 38 | 42 | 56 | 76 | 100 | 100 | |
| k=4 | 50 | 51 | 53 | 55 | 69 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |

Figure A.6: POD, Conv.Rate and Conv.Nr of MSV7



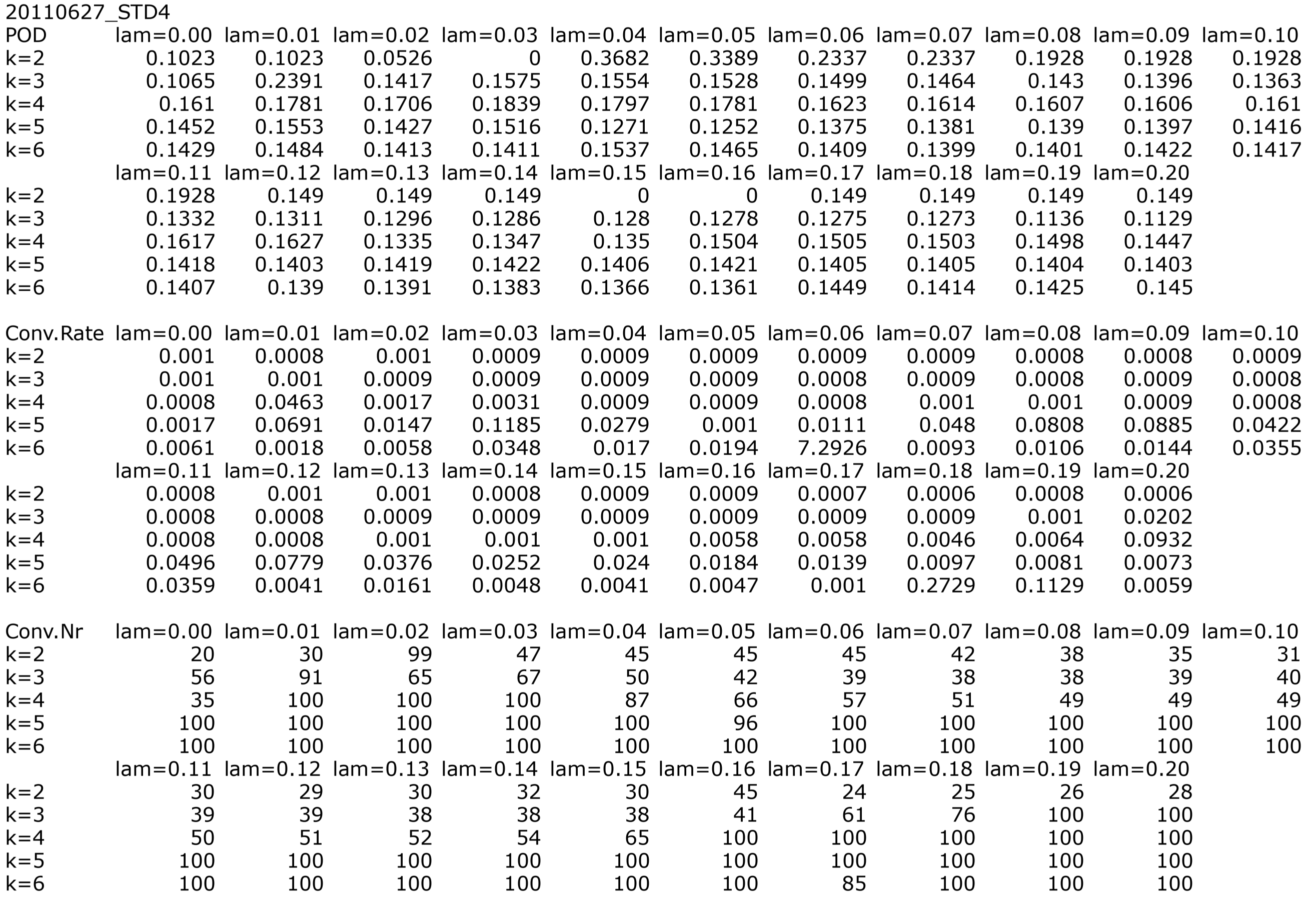
20110627_MSV9

| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 0.1023 | 0.1023 | 0.1023 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1011 | 0.2559 | 0.1427 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1635 | 0.1774 | 0.1698 | 0.1839 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.143 | 0.156 | 0.1434 | 0.1523 | 0.127 | 0.1252 | 0.1376 | 0.1382 | 0.1391 | 0.1396 | 0.1415 |
| k=6 | 0.1452 | 0.1384 | 0.1421 | 0.1415 | 0.154 | 0.1465 | 0.1453 | 0.1399 | 0.1401 | 0.1422 | 0.1417 |
|  | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 |  |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | NA | NA | NA | NA | NA |  |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 |  |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1504 | 0.1506 | 0.1503 | 0.1498 | 0.1453 |  |
| k=5 | 0.1417 | 0.1406 | 0.1418 | 0.1422 | 0.1422 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 |  |
| k=6 | 0.1408 | 0.1393 | 0.1391 | 0.1383 | 0.1366 | 0.1361 | 0.1437 | 0.1372 | 0.1401 | 0.1376 |  |

| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 0.0008 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0009 |
| k=3 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.0008 | 0.0473 | 0.0027 | 0.0036 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.001 | 0.0009 |
| k=5 | 0.0017 | 0.0572 | 0.3528 | 0.1216 | 0.0309 | 0.0009 | 0.0164 | 0.0598 | 0.0956 | 0.0622 | 0.0335 |
| k=6 | 0.0028 | 0.001 | 0.0076 | 0.0368 | 0.0185 | 0.0176 | 4.8052 | 0.0089 | 0.0101 | 0.0139 | 0.0332 |
|  | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 |  |
| k=2 | 0.0007 | 0.0009 | 0.0008 | 0.001 | 0.001 | NA | NA | NA | NA | NA |  |
| k=3 | 0.0009 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0012 | 0.0431 |  |
| k=4 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0053 | 0.0053 | 0.0038 | 0.0058 | 0.8216 |  |
| k=5 | 0.0276 | 0.1189 | 0.0417 | 0.0259 | 0.0303 | 0.0196 | 0.0138 | 0.0095 | 0.0079 | 0.0069 |  |
| k=6 | 0.0373 | 0.0086 | 0.0187 | 0.0049 | 0.0042 | 0.0049 | 0.0697 | 0.0349 | 0.0285 | 0.1111 |  |

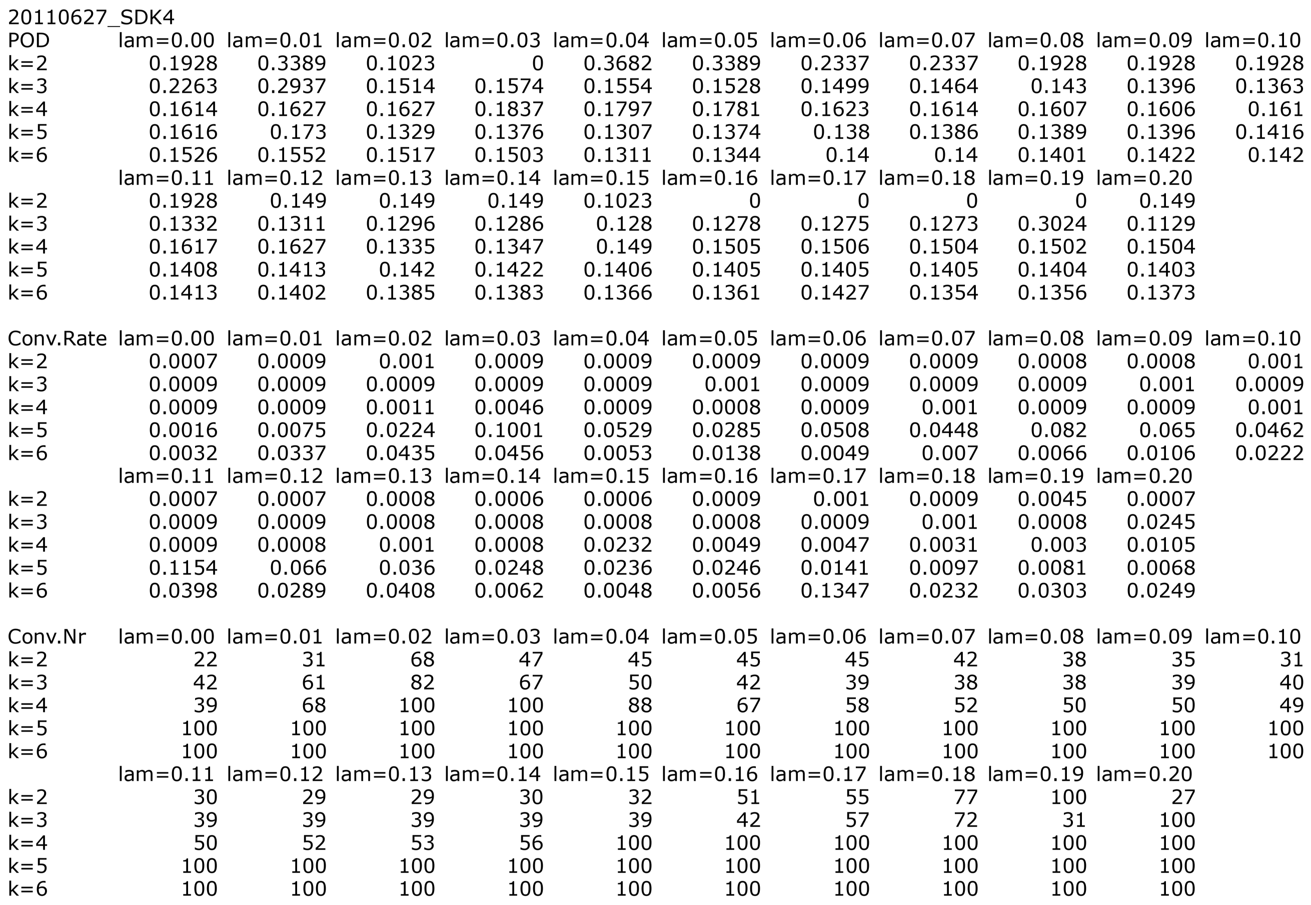
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 21 | 31 | 92 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 58 | 81 | 63 | 66 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 34 | 100 | 100 | 100 | 87 | 66 | 57 | 52 | 50 | 49 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 98 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
|  | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 |  |
| k=2 | 30 | 29 | 30 | 30 | 44 | NA | NA | NA | NA | NA |  |
| k=3 | 39 | 39 | 38 | 38 | 38 | 42 | 34 | 75 | 100 | 100 |  |
| k=4 | 50 | 51 | 53 | 55 | 99 | 100 | 100 | 100 | 100 | 100 |  |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |  |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |  |

Figure A.7: POD, Conv.Rate and Conv.Nr of MSV9



| 20110627_STD4 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.1023 | 0.0526 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1065 | 0.2391 | 0.1417 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.161 | 0.1781 | 0.1706 | 0.1839 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1452 | 0.1553 | 0.1427 | 0.1516 | 0.1271 | 0.1252 | 0.1375 | 0.1381 | 0.139 | 0.1397 | 0.1416 |
| k=6 | 0.1429 | 0.1484 | 0.1413 | 0.1411 | 0.1537 | 0.1465 | 0.1409 | 0.1399 | 0.1401 | 0.1422 | 0.1417 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0 | 0 | 0.149 | 0.149 | 0.149 | 0.149 | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1504 | 0.1505 | 0.1503 | 0.1498 | 0.1447 | |
| k=5 | 0.1418 | 0.1403 | 0.1419 | 0.1422 | 0.1406 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.1407 | 0.139 | 0.1391 | 0.1383 | 0.1366 | 0.1361 | 0.1449 | 0.1414 | 0.1425 | 0.145 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.001 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0009 |
| k=3 | 0.001 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.0008 | 0.0463 | 0.0017 | 0.0031 | 0.0009 | 0.0009 | 0.0008 | 0.001 | 0.001 | 0.0009 | 0.0008 |
| k=5 | 0.0017 | 0.0691 | 0.0147 | 0.1185 | 0.0279 | 0.001 | 0.0111 | 0.048 | 0.0808 | 0.0885 | 0.0422 |
| k=6 | 0.0061 | 0.0018 | 0.0058 | 0.0348 | 0.017 | 0.0194 | 7.2926 | 0.0093 | 0.0106 | 0.0144 | 0.0355 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0008 | 0.001 | 0.001 | 0.0008 | 0.0009 | 0.0009 | 0.0007 | 0.0006 | 0.0008 | 0.0006 | |
| k=3 | 0.0008 | 0.0008 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0202 | |
| k=4 | 0.0008 | 0.0008 | 0.001 | 0.001 | 0.001 | 0.0058 | 0.0058 | 0.0046 | 0.0064 | 0.0932 | |
| k=5 | 0.0496 | 0.0779 | 0.0376 | 0.0252 | 0.024 | 0.0184 | 0.0139 | 0.0097 | 0.0081 | 0.0073 | |
| k=6 | 0.0359 | 0.0041 | 0.0161 | 0.0048 | 0.0041 | 0.0047 | 0.001 | 0.2729 | 0.1129 | 0.0059 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 20 | 30 | 99 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 56 | 91 | 65 | 67 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 35 | 100 | 100 | 100 | 87 | 66 | 57 | 51 | 49 | 49 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 96 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 30 | 29 | 30 | 32 | 30 | 45 | 24 | 25 | 26 | 28 | |
| k=3 | 39 | 39 | 38 | 38 | 38 | 41 | 61 | 76 | 100 | 100 | |
| k=4 | 50 | 51 | 52 | 54 | 65 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 85 | 100 | 100 | 100 | |

Figure A.8: POD, Conv.Rate and Conv.Nr of STD4



20110627_SDK4

| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 0.1928 | 0.3389 | 0.1023 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.2263 | 0.2937 | 0.1514 | 0.1574 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1614 | 0.1627 | 0.1627 | 0.1837 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1616 | 0.173 | 0.1329 | 0.1376 | 0.1307 | 0.1374 | 0.138 | 0.1386 | 0.1389 | 0.1396 | 0.1416 |
| k=6 | 0.1526 | 0.1552 | 0.1517 | 0.1503 | 0.1311 | 0.1344 | 0.14 | 0.14 | 0.1401 | 0.1422 | 0.142 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | 0 | 0 | 0 | 0 | 0.149 | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.3024 | 0.1129 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.149 | 0.1505 | 0.1506 | 0.1504 | 0.1502 | 0.1504 | |
| k=5 | 0.1408 | 0.1413 | 0.142 | 0.1422 | 0.1406 | 0.1405 | 0.1405 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.1413 | 0.1402 | 0.1385 | 0.1383 | 0.1366 | 0.1361 | 0.1427 | 0.1354 | 0.1356 | 0.1373 | |

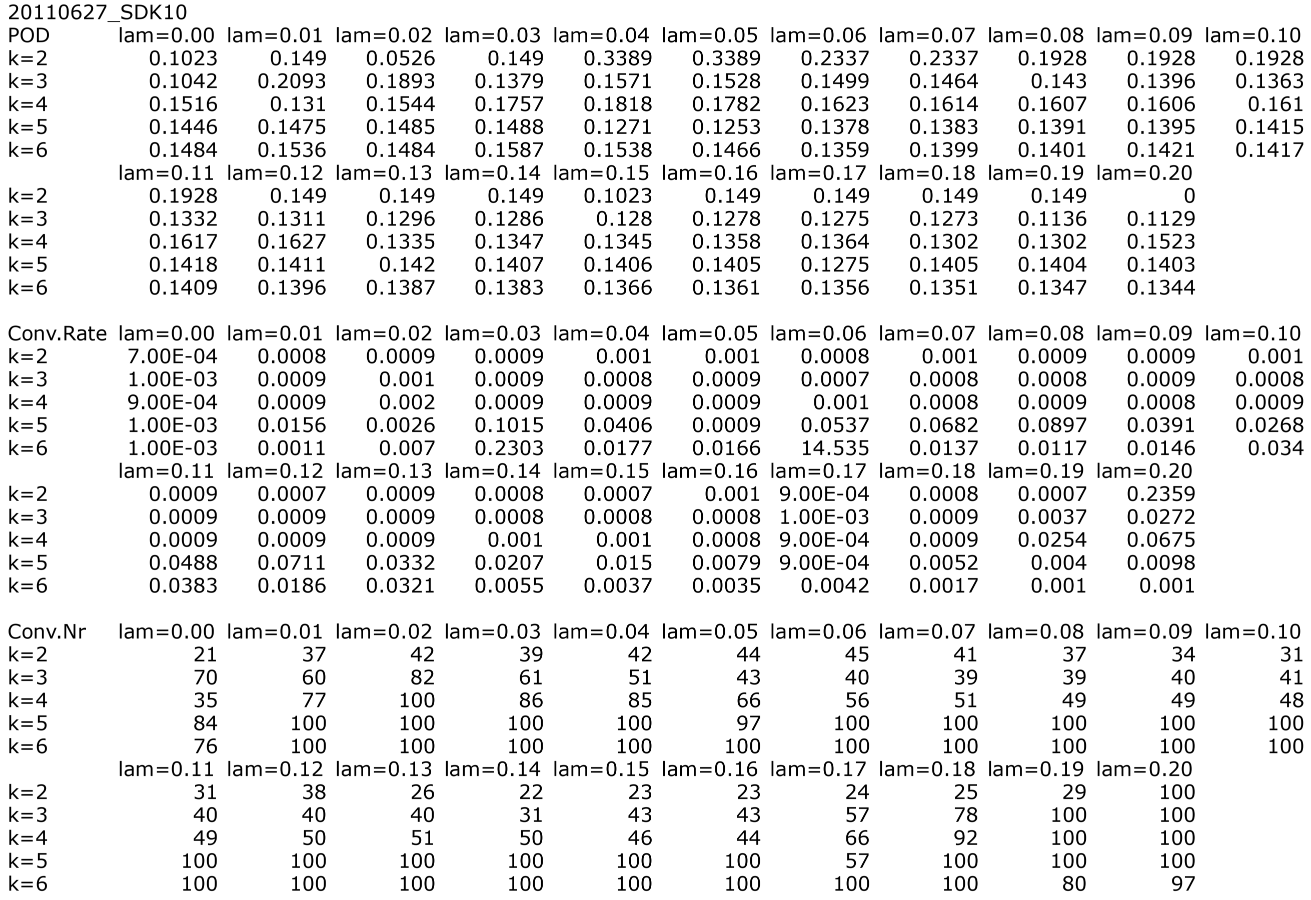
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 0.0007 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.001 |
| k=3 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0009 |
| k=4 | 0.0009 | 0.0009 | 0.0011 | 0.0046 | 0.0009 | 0.0008 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.001 |
| k=5 | 0.0016 | 0.0075 | 0.0224 | 0.1001 | 0.0529 | 0.0285 | 0.0508 | 0.0448 | 0.082 | 0.065 | 0.0462 |
| k=6 | 0.0032 | 0.0337 | 0.0435 | 0.0456 | 0.0053 | 0.0138 | 0.0049 | 0.007 | 0.0066 | 0.0106 | 0.0222 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0007 | 0.0007 | 0.0008 | 0.0006 | 0.0006 | 0.0009 | 0.001 | 0.0009 | 0.0045 | 0.0007 | |
| k=3 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 0.0008 | 0.0009 | 0.001 | 0.0008 | 0.0245 | |
| k=4 | 0.0009 | 0.0008 | 0.001 | 0.0008 | 0.0232 | 0.0049 | 0.0047 | 0.0031 | 0.003 | 0.0105 | |
| k=5 | 0.1154 | 0.066 | 0.036 | 0.0248 | 0.0236 | 0.0246 | 0.0141 | 0.0097 | 0.0081 | 0.0068 | |
| k=6 | 0.0398 | 0.0289 | 0.0408 | 0.0062 | 0.0048 | 0.0056 | 0.1347 | 0.0232 | 0.0303 | 0.0249 | |

| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2 | 22 | 31 | 68 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 42 | 61 | 82 | 67 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 39 | 68 | 100 | 100 | 88 | 67 | 58 | 52 | 50 | 50 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 30 | 29 | 29 | 30 | 32 | 51 | 55 | 77 | 100 | 27 | |
| k=3 | 39 | 39 | 39 | 39 | 39 | 42 | 57 | 72 | 31 | 100 | |
| k=4 | 50 | 52 | 53 | 56 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |

Figure A.9: POD, Conv.Rate and Conv.Nr of SDK4

| 20110627_SDK6 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.0526 | 0.1023 | 0 | 0.3682 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.102 | 0.2549 | 0.143 | 0.1575 | 0.1554 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1607 | 0.1755 | 0.1694 | 0.1839 | 0.1797 | 0.1781 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1477 | 0.1533 | 0.144 | 0.1461 | 0.1271 | 0.1251 | 0.1375 | 0.1382 | 0.1391 | 0.1397 | 0.1418 |
| k=6 | 0.1496 | 0.1474 | 0.1423 | 0.1413 | 0.1539 | 0.1345 | 0.14 | 0.1399 | 0.1401 | 0.1422 | 0.1418 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | NA | NA | NA | NA | NA | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.135 | 0.1504 | 0.1506 | 0.1503 | 0.1498 | 0.1499 | |
| k=5 | 0.1424 | 0.1405 | 0.142 | 0.1422 | 0.1422 | 0.1421 | 0.1405 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.1408 | 0.1393 | 0.1391 | 0.1383 | 0.1366 | 0.145 | 0.1358 | 0.1357 | 0.1372 | 0.1407 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0006 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0009 |
| k=3 | 0.0009 | 0.0013 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 0.0009 | 0.0571 | 0.0031 | 0.003 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.001 | 0.0009 |
| k=5 | 0.0017 | 0.063 | 0.1068 | 0.119 | 0.0286 | 0.0009 | 0.0099 | 0.051 | 0.0923 | 0.0887 | 0.0701 |
| k=6 | 0.0024 | 0.0029 | 0.0101 | 0.0364 | 0.0182 | 0.0092 | 0.008 | 0.0084 | 0.0094 | 0.0133 | 0.0321 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0008 | 0.0009 | 0.0009 | 0.0007 | 0.0007 | NA | NA | NA | NA | NA | |
| k=3 | 0.0009 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0009 | 0.0012 | 0.0403 | |
| k=4 | 0.0008 | 0.0009 | 0.0008 | 0.0009 | 0.001 | 0.0053 | 0.0054 | 0.0038 | 0.0064 | 0.0074 | |
| k=5 | 0.1129 | 0.0795 | 0.036 | 0.0252 | 0.0333 | 0.0181 | 0.0136 | 0.0094 | 0.0077 | 0.0066 | |
| k=6 | 0.0378 | 0.0091 | 0.0185 | 0.0049 | 0.0043 | 0.001 | 0.0262 | 0.0332 | 0.0305 | 0.0375 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 21 | 30 | 89 | 47 | 45 | 45 | 45 | 42 | 38 | 35 | 31 |
| k=3 | 53 | 100 | 63 | 67 | 50 | 42 | 39 | 38 | 38 | 39 | 40 |
| k=4 | 34 | 100 | 100 | 100 | 87 | 66 | 57 | 51 | 49 | 49 | 49 |
| k=5 | 100 | 100 | 100 | 100 | 100 | 97 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 30 | 29 | 30 | 31 | 44 | NA | NA | NA | NA | NA | |
| k=3 | 39 | 39 | 38 | 38 | 38 | 42 | 34 | 75 | 100 | 100 | |
| k=4 | 50 | 51 | 53 | 55 | 78 | 100 | 100 | 100 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 90 | 100 | 100 | 100 | 100 | |

Figure A.10: POD, Conv.Rate and Conv.Nr of SDK6



| 20110627_SDK10 | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.149 | 0.0526 | 0.149 | 0.3389 | 0.3389 | 0.2337 | 0.2337 | 0.1928 | 0.1928 | 0.1928 |
| k=3 | 0.1042 | 0.2093 | 0.1893 | 0.1379 | 0.1571 | 0.1528 | 0.1499 | 0.1464 | 0.143 | 0.1396 | 0.1363 |
| k=4 | 0.1516 | 0.131 | 0.1544 | 0.1757 | 0.1818 | 0.1782 | 0.1623 | 0.1614 | 0.1607 | 0.1606 | 0.161 |
| k=5 | 0.1446 | 0.1475 | 0.1485 | 0.1488 | 0.1271 | 0.1253 | 0.1378 | 0.1383 | 0.1391 | 0.1395 | 0.1415 |
| k=6 | 0.1484 | 0.1536 | 0.1484 | 0.1587 | 0.1538 | 0.1466 | 0.1359 | 0.1399 | 0.1401 | 0.1421 | 0.1417 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1928 | 0.149 | 0.149 | 0.149 | 0.1023 | 0.149 | 0.149 | 0.149 | 0.149 | 0 | |
| k=3 | 0.1332 | 0.1311 | 0.1296 | 0.1286 | 0.128 | 0.1278 | 0.1275 | 0.1273 | 0.1136 | 0.1129 | |
| k=4 | 0.1617 | 0.1627 | 0.1335 | 0.1347 | 0.1345 | 0.1358 | 0.1364 | 0.1302 | 0.1302 | 0.1523 | |
| k=5 | 0.1418 | 0.1411 | 0.142 | 0.1407 | 0.1406 | 0.1405 | 0.1275 | 0.1405 | 0.1404 | 0.1403 | |
| k=6 | 0.1409 | 0.1396 | 0.1387 | 0.1383 | 0.1366 | 0.1361 | 0.1356 | 0.1351 | 0.1347 | 0.1344 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 7.00E-04 | 0.0008 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.0008 | 0.001 | 0.0009 | 0.0009 | 0.001 |
| k=3 | 1.00E-03 | 0.0009 | 0.001 | 0.0009 | 0.0008 | 0.0009 | 0.0007 | 0.0008 | 0.0008 | 0.0009 | 0.0008 |
| k=4 | 9.00E-04 | 0.0009 | 0.002 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.0008 | 0.0009 | 0.0008 | 0.0009 |
| k=5 | 1.00E-03 | 0.0156 | 0.0026 | 0.1015 | 0.0406 | 0.0009 | 0.0537 | 0.0682 | 0.0897 | 0.0391 | 0.0268 |
| k=6 | 1.00E-03 | 0.0011 | 0.007 | 0.2303 | 0.0177 | 0.0166 | 14.535 | 0.0137 | 0.0117 | 0.0146 | 0.034 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0009 | 0.0007 | 0.0009 | 0.0008 | 0.0007 | 0.001 | 9.00E-04 | 0.0008 | 0.0007 | 0.2359 | |
| k=3 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 0.0008 | 1.00E-03 | 0.0009 | 0.0037 | 0.0272 | |
| k=4 | 0.0009 | 0.0009 | 0.0009 | 0.001 | 0.001 | 0.0008 | 9.00E-04 | 0.0009 | 0.0254 | 0.0675 | |
| k=5 | 0.0488 | 0.0711 | 0.0332 | 0.0207 | 0.015 | 0.0079 | 9.00E-04 | 0.0052 | 0.004 | 0.0098 | |
| k=6 | 0.0383 | 0.0186 | 0.0321 | 0.0055 | 0.0037 | 0.0035 | 0.0042 | 0.0017 | 0.001 | 0.001 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 21 | 37 | 42 | 39 | 42 | 44 | 45 | 41 | 37 | 34 | 31 |
| k=3 | 70 | 60 | 82 | 61 | 51 | 43 | 40 | 39 | 39 | 40 | 41 |
| k=4 | 35 | 77 | 100 | 86 | 85 | 66 | 56 | 51 | 49 | 49 | 48 |
| k=5 | 84 | 100 | 100 | 100 | 100 | 97 | 100 | 100 | 100 | 100 | 100 |
| k=6 | 76 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 31 | 38 | 26 | 22 | 23 | 23 | 24 | 25 | 29 | 100 | |
| k=3 | 40 | 40 | 40 | 31 | 43 | 43 | 57 | 78 | 100 | 100 | |
| k=4 | 49 | 50 | 51 | 50 | 46 | 44 | 66 | 92 | 100 | 100 | |
| k=5 | 100 | 100 | 100 | 100 | 100 | 100 | 57 | 100 | 100 | 100 | |
| k=6 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | 80 | 97 | |

Figure A.11: POD, Conv.Rate and Conv.Nr of SDK10

| 20110707_Random | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| POD | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.1023 | 0.0526 | 0 | 0.1023 | 0.1023 | 0.1928 | 0.1928 | 0.1928 | 0.149 | 0.149 | 0.149 |
| k=3 | 0.1838 | 0.216 | 0.2159 | 0.1716 | 0.1403 | 0.1287 | 0.1268 | 0.1241 | 0.1409 | 0.1262 | 0.124 |
| k=4 | 0.2027 | 0.1938 | 0.1754 | 0.273 | 0.2773 | 0.1463 | 0.152 | 0.1292 | 0.1293 | 0.1502 | 0.1475 |
| k=5 | 0.1593 | 0.1659 | 0.1718 | 0.1942 | 0.1896 | 0.1867 | 0.1582 | 0.1414 | 0.1483 | 0.1531 | 0.1227 |
| k=6 | 0.1399 | 0.1449 | 0.1702 | 0.1903 | 0.1841 | 0.1842 | 0.1409 | 0.1628 | 0.1504 | 0.13 | 0.1269 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.1023 | 0.149 | 0.149 | 0.149 | 0.149 | 0.149 | 0.149 | 0.149 | 0.149 | 0.149 | |
| k=3 | 0.1223 | 0.1406 | 0.1381 | 0.1362 | 0.1348 | 0.1337 | 0.1327 | 0.1318 | 0.131 | 0.3097 | |
| k=4 | 0.1397 | 0.1394 | 0.1394 | 0.1395 | 0.1387 | 0.1378 | 0.1613 | 0.1657 | 0.1605 | 0.1645 | |
| k=5 | 0.1477 | 0.1464 | 0.1597 | 0.1223 | 0.1422 | 0.1397 | 0.1616 | 0.1457 | 0.142 | 0.1671 | |
| k=6 | 0.1254 | 0.1629 | 0.1372 | 0.1415 | 0.1658 | 0.1664 | 0.1371 | 0.1446 | 0.1436 | 0.1566 | |
| | | | | | | | | | | | |
| Conv.Rate | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 0.0008 | 0.0009 | 0.001 | 0.0009 | 0.0009 | 0.001 | 0.0008 | 0.0008 | 0.0009 | 0.0008 | 8.00E-04 |
| k=3 | 0.0104 | 0.001 | 0.001 | 0.0047 | 0.0291 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0008 | 6.00E-04 |
| k=4 | 0.0008 | 0.001 | 0.001 | 0.001 | 0.006 | 0.2274 | 0.001 | 0.0009 | 0.0009 | 0.0027 | 5.00E-04 |
| k=5 | 0.001 | 0.003 | 0.0089 | 0.001 | 0.001 | 0.001 | 0.0219 | 0.4276 | 0.0083 | 0.0049 | 8.00E-04 |
| k=6 | 0.0921 | 0.084 | 0.0071 | 0.001 | 0.001 | 0.0114 | 0.0033 | 0.0829 | 0.0009 | 0.001 | 8.00E-04 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 0.0007 | 0.0008 | 8.00E-04 | 0.0007 | 0.0009 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0005 | |
| k=3 | 0.0006 | 0.0007 | 8.00E-04 | 0.0009 | 0.0009 | 0.0008 | 0.0009 | 0.0008 | 0.001 | 0.0009 | |
| k=4 | 0.0629 | 0.001 | 8.00E-04 | 0.0009 | 0.0009 | 0.001 | 0.3141 | 0.0009 | 0.7809 | 0.0024 | |
| k=5 | 0.001 | 0.001 | 9.00E-04 | 0.0335 | 5.71 | 5.3763 | 0.0008 | 5.2366 | 0.0057 | 0.0008 | |
| k=6 | 0.0009 | 5.2223 | 9.00E-04 | 0.1867 | 9.5631 | 8.0256 | 7.428 | 7.5207 | 7.3143 | 4.6154 | |
| | | | | | | | | | | | |
| Conv.Nr | lam=0.00 | lam=0.01 | lam=0.02 | lam=0.03 | lam=0.04 | lam=0.05 | lam=0.06 | lam=0.07 | lam=0.08 | lam=0.09 | lam=0.10 |
| k=2 | 31 | 31 | 41 | 65 | 41 | 34 | 32 | 27 | 25 | 23 | 22 |
| k=3 | 100 | 73 | 69 | 100 | 100 | 87 | 59 | 47 | 39 | 28 | 31 |
| k=4 | 45 | 63 | 87 | 89 | 100 | 100 | 57 | 51 | 43 | 100 | 34 |
| k=5 | 52 | 100 | 100 | 91 | 100 | 68 | 100 | 100 | 100 | 100 | 99 |
| k=6 | 100 | 100 | 100 | 85 | 81 | 100 | 100 | 100 | 53 | 41 | 37 |
| | lam=0.11 | lam=0.12 | lam=0.13 | lam=0.14 | lam=0.15 | lam=0.16 | lam=0.17 | lam=0.18 | lam=0.19 | lam=0.20 | |
| k=2 | 22 | 22 | 23 | 27 | 21 | 20 | 22 | 33 | 29 | 28 | |
| k=3 | 33 | 35 | 36 | 41 | 46 | 40 | 45 | 40 | 32 | 32 | |
| k=4 | 100 | 65 | 51 | 47 | 45 | 49 | 100 | 35 | 100 | 100 | |
| k=5 | 66 | 37 | 49 | 100 | 100 | 100 | 89 | 100 | 100 | 97 | |
| k=6 | 44 | 100 | 45 | 100 | 100 | 100 | 100 | 100 | 100 | 100 | |

Figure A.12: POD, Conv.Rate and Conv.Nr of Random

## A.2 Clusters for different $k$ and lambdas

In this section one the membership of all the samples in the different clusters which one specified as best clusters before are shown. This is done for all the 12 (including the random subset) subsets of the Sarcoma data set. A sum of the number how often that sample (in the 11 subsets) is within the main cluster is also provided as well as the number of samples in each cluster. The going from . . . to . . . is also provided in a more detailed version in color. The event analysis data is also shown for 36 samples, where data was available.

Clusters for different k and lambdas

| k=2, lam=0.03 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20110627_sd_GE_13 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_sd_GE_11 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| 20110627_sd_GE_15 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_ME437 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_ME637 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_MSV7 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_MSV9 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_STD4 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_SDK4 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_SDK6 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| 20110627_SDK10 | 1 | 1 | 2 | 1 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 2 |
| 20110707_Random | 1 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 1 | 1 | 1 | 1 | 1 |
| Most common Cluster | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| Sum without Random | 11 | 9 | 9 | 11 | 11 | 9 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 9 | 11 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 10 | 11 | 11 | 9 | 11 | 11 | 11 | 10 |

1 19    2 18    ▨ 6

| k=3, lam=0.20 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20110627_sd_GE_13 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_sd_GE_11 | 2 | 3 | 1 | 1 | 1 | 2 | 2 | 2 | 1 | 1 | 1 | 3 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 1 | 1 | 2 | 2 | 3 | 3 | 3 | 2 | 2 | 3 | 3 | 2 | 1 | 1 | 3 |
| 20110627_sd_GE_15 | 1 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 3 | 3 | 3 | 2 | 3 | 1 | 1 | 1 | 3 |
| 20110627_ME437 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_ME637 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_MSV7 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_MSV9 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_STD4 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_SDK4 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_SDK6 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_SDK10 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110707_Random | 1 | 3 | 3 | 3 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 3 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 3 | 1 | 3 |
| Most common Cluster | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| Sum without Random | 9 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 11 | 10 | 10 | 11 | 11 | 11 | 10 | 11 | 10 | 11 | 11 | 10 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 11 | 9 | 10 | 10 | 10 | 11 | 9 | 11 | 11 | 11 |

1 16    2 9    3 12    ▨ 13

Figure A.13: Cluster membership for $k = 2, 3$

Clusters for different k and lambdas

| k=4, lam=0.13 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20110627_sd_GE_13 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_sd_GE_11 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_sd_GE_15 | 3 | 3 | 1 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 4 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| 20110627_ME437 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_ME637 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_MSV7 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_MSV9 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_STD4 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_SDK4 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_SDK6 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110627_SDK10 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| 20110707_Random | 3 | 3 | 3 | 1 | 1 | 3 | 4 | 4 | 1 | 4 | 4 | 3 | 1 | 1 | 3 | 1 | 4 | 1 | 1 | 4 | 3 | 1 | 1 | 1 | 3 | 4 | 3 | 3 | 3 | 3 | 2 | 4 | 3 | 1 | 1 | 1 | 3 |
| Most common Cluster | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| Sum without Random | 11 | 11 | 10 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 10 | 11 | 11 |

1 14   2 6   3 11   4 6   ▨ 12

| k=5, lam=0.05 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 20110627_sd_GE_13 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_sd_GE_11 | 5 | 3 | 5 | 5 | 1 | 5 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 5 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 5 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 5 | 1 | 3 |
| 20110627_sd_GE_15 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 1 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 2 | 1 | 2 | 2 | 1 | 5 | 5 | 2 |
| 20110627_ME437 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_ME637 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_MSV7 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_MSV9 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_STD4 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_SDK4 | 5 | 3 | 5 | 5 | 1 | 5 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 5 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 5 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 5 | 1 | 3 |
| 20110627_SDK6 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110627_SDK10 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| 20110707_Random | 4 | 5 | 5 | 2 | 3 | 4 | 4 | 4 | 1 | 2 | 4 | 5 | 1 | 1 | 5 | 5 | 2 | 5 | 5 | 4 | 4 | 1 | 5 | 1 | 5 | 4 | 2 | 5 | 4 | 2 | 3 | 2 | 3 | 3 | 1 | 5 | 4 |
| Most common Cluster | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |
| Sum without Random | 9 | 11 | 11 | 11 | 11 | 9 | 11 | 11 | 11 | 11 | 11 | 11 | 11 | 9 | 11 | 9 | 11 | 9 | 11 | 10 | 11 | 11 | 11 | 11 | 9 | 11 | 11 | 11 | 11 | 10 | 9 | 11 | 10 | 9 | 11 | 9 | 10 |

1 10   2 6   3 9   4 4   5 8   ▨ 21

Figure A.14: Cluster membership for $k = 4, 5$

Splitting of the clusters

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2, lam=0.03 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| k=3, lam=0.20 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |

| | |
|---|---|
| 1-1 | 15 |
| 1-3 | 4 |
| | |
| 2-2 | 9 |
| 2-3 | 8 |
| 2-1 | 1 |

| | | |
|---|---|---|
| stay | 24 | 0.65 |
| new | 12 | 0.32 |
| other | 1 | 0.03 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=3, lam=0.20 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| k=4, lam=0.13 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |

| | |
|---|---|
| 1-1 | 14 |
| 1-4 | 2 |
| | |
| 2-2 | 5 |
| 2-4 | 4 |
| | |
| 3-3 | 11 |
| 3-2 | 1 |

| | | |
|---|---|---|
| stay | 30 | 0.81 |
| new | 6 | 0.16 |
| other | 1 | 0.03 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=4, lam=0.13 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| k=5, lam=0.05 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |

| | |
|---|---|
| 1-1 | 8 |
| 1-5 | 6 |
| | |
| 2-2 | 6 |
| | |
| 3-3 | 9 |
| 3-1 | 2 |
| | |
| 4-4 | 4 |
| 4-5 | 2 |

| | | |
|---|---|---|
| stay | 27 | 0.73 |
| new | 8 | 0.22 |
| other | 2 | 0.05 |

Figure A.15: Splitting of the clusters, part 1

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2, lam=0.03 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| k=4, lam=0.13 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |

| | |
|---|---|
| 1-1 | 14 |
| 1-3 | 4 |
| 1-4 | 1 |
| 2-2 | 6 |
| 2-3 | 7 |
| 2-4 | 5 |

| | | |
|---|---|---|
| stay | 20 | 0.54 |
| new | 17 | 0.46 |
| other | 0 | 0.00 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2, lam=0.03 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| k=5, lam=0.05 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |

| | |
|---|---|
| 1-1 | 10 |
| 1-3 | 2 |
| 1-5 | 7 |
| 2-2 | 6 |
| 2-3 | 7 |
| 2-4 | 4 |
| 2-5 | 1 |

| | | |
|---|---|---|
| stay | 16 | 0.43 |
| new | 21 | 0.57 |
| other | 0 | 0.00 |

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=3, lam=0.20 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| k=5, lam=0.05 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |

| | |
|---|---|
| 1-1 | 8 |
| 1-5 | 8 |
| 2-2 | 5 |
| 2-4 | 4 |
| 3-3 | 9 |
| 3-1 | 2 |
| 3-2 | 1 |

| | | |
|---|---|---|
| stay | 22 | 0.59 |
| new | 12 | 0.32 |
| other | 3 | 0.08 |

Figure A.16: Splitting of the clusters, part 2

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| k=2, lam=0.03 | 1 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 1 | 2 |
| k=3, lam=0.20 | 3 | 3 | 1 | 1 | 1 | 3 | 2 | 2 | 1 | 2 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 3 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 1 | 1 | 3 |
| k=4, lam=0.13 | 3 | 3 | 4 | 1 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 1 | 1 | 1 | 4 | 1 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 3 | 2 | 3 | 3 | 4 | 1 | 3 |
| k=5, lam=0.05 | 3 | 3 | 5 | 5 | 1 | 3 | 4 | 2 | 1 | 4 | 4 | 3 | 1 | 5 | 5 | 5 | 4 | 5 | 1 | 2 | 3 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 2 | 3 | 1 | 2 | 3 | 1 | 5 | 5 | 3 |

| | |
|---|---|
| 1-1 | 10 |
| 1-3 | 2 |
| 1-5 | 7 |
| | |
| 2-2 | 6 |
| 2-3 | 7 |
| 2-4 | 4 |
| 2-5 | 1 |

| | | |
|---|---|---|
| stay | 16 | 0.43 |
| new | 21 | 0.57 |
| other | 0 | 0.00 |

Figure A.17: Splitting of the clusters, part 3

| No | Sample Name | Follow up Years | Die/ Censor | Time to recur/Censor | DR |
|---|---|---|---|---|---|
| 1 | PD_U133A_MFH2516 | 3.633 | 0 | 3.633 | 0 |
| 2 | PD_U133A_MFH633 | 1.900 | 0 | 1.900 | 0 |
| 3 | PD_U133A_MFH623 | 0.408 | 0 | 0.408 | 0 |
| 4 | PD_HG_U133A_MFH660 | 5.029 | 0 | 5.029 | 0 |
| 5 | PD_MXF815_HG-U133A | 4.331 | 0 | 4.331 | 0 |
| 6 | PD_U133A_MFH632 | 1.311 | 1 | 0.244 | 1 |
| 7 | PD_U133A_MFH659 | 3.953 | 0 | 3.953 | 0 |
| 8 | PD_MFH730_HG_U133A | 2.281 | 1 | 0.164 | 1 |
| 9 | PD_MXF871_HG-U133A | 1.257 | 0 | 0.526 | 1 |
| 10 | PD_MXF829_HG-U133A | 4.619 | 0 | 4.619 | 0 |
| 11 | PD_MXF830_HG-U133A | 4.474 | 0 | 4.474 | 0 |
| 12 | PD_MXF832_HG-U133A | 0.712 | 1 | 0.386 | 1 |
| 13 | PD_MXF849_HG-U133A | 2.398 | 0 | 1.366 | 1 |
| 14 | PD_MXF874_HG-U133A | 0.947 | 0 | 0.947 | 0 |
| 15 | PD_MXF875_HG-U133A | 1.076 | 0 | 1.076 | 0 |
| 16 | PD_MXF834_HG-U133A | 3.064 | 0 | 3.064 | 0 |
| 17 | PD_MXF835_HG-U133A | 1.224 | 0 | 1.224 | 0 |
| 18 | PD_MXF836_HG-U133A | 3.466 | 0 | 3.466 | 0 |
| 19 | PD_MXF847_HG-U133A | 2.319 | 0 | 2.319 | 0 |
| 20 | PD_MXF848_HG-U133A | 2.428 | 0 | 2.428 | 0 |
| 21 | PD_MXF851_HG-U133A | 2.426 | 1 | 0.621 | 1 |
| 22 | PD_MXF852_HG-U133A | 2.396 | 0 | 2.396 | 0 |
| 23 | PD_MXF855_HG-U133A | 1.399 | 0 | 1.333 | 1 |
| 24 | PD_MXF856_HG-U133A | 1.369 | 0 | 1.369 | 0 |
| 25 | PD_MXF861_HG-U133A | 1.520 | 0 | 1.520 | 0 |
| 26 | PD_MXF863_HG-U133A | 1.328 | 0 | 1.328 | 0 |
| 27 | PD_pMFH816_HG-U133A | 2.921 | 0 | 0.830 | 1 |
| 28 | PD_pMFH870_HG-U133A | 0.715 | 1 | 0.285 | 1 |
| 29 | PD_pMFH872_HG-U133A | 1.076 | 0 | 0.496 | 1 |
| 30 | PD_pMFH876_HG-U133A | 0.947 | 0 | 0.947 | 0 |
| 31 | PD_pMFH877_HG-U133A | 0.860 | 0 | 0.860 | 0 |
| 32 | PD_pMFH878_HG-U133A | 1.117 | 0 | 1.117 | 0 |
| 33 | PD_pMFH897_HG-U133A | 0.298 | 0 | 0.238 | 1 |
| 34 | PD_pMFH898_HG-U133A | 0.487 | 0 | 0.487 | 0 |
| 35 | PD_MXF902_HG-U133A | 2.686 | 0 | 2.686 | 0 |
| 36 | PD_MXF916_HG-U133A_2 | 1.281 | 0 | 0.331 | 1 |

Table A.1: Sample cluster membership for different $k$

# APPENDIX B

## Results GSS

In this chapter additional results of the Gene Set Score are provided. With the Gene Set Score various possibilities are worth to think about. But some of them do not provide as good results as others that were chosen as best and were represented in the main chapters.

|  |  |  | 'PRIMARY' with 'MET' |  |  | 'MET' with 'PRIMARY' |  |  |
|---|---|---|---|---|---|---|---|---|
| Name | # PW | # PW<br>≥ 10 | # CNE | # (CNE<br>∩ GEE) | Ratio | # CNE | # (CNE<br>∩ GEE) | Ratio |
| arms | 46 | 39 | 18 | 10 | 0.56 | 14 | 7 | 0.50 |
| cyto | 797 | 399 | 211 | 115 | 0.55 | 115 | 72 | 0.63 |
| proc | 24 | 23 | 16 | 13 | 0.81 | 12 | 10 | 0.83 |
| sega | 456 | 406 | 225 | 162 | 0.72 | 180 | 144 | 0.80 |
| tile | 1192 | 820 | 438 | 232 | 0.53 | 251 | 148 | 0.59 |
| tiss | 80 | 73 | 50 | 34 | 0.68 | 43 | 30 | 0.70 |
| kegg | 206 | 142 | 92 | 64 | 0.70 | 58 | 43 | 0.74 |
| TOTAL | 2801 | 1902 | 1050 | 630 | 0.60 | 673 | 454 | 0.67 |

Table B.1: Ratio of a CNE is accompanied by a GEE with the 'Interval Method'

| Name | # (CNE $\cap$ GEE) | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p < 0.01$ | | $p < 0.05$ | | $p < 0.10$ | |
| arms | 16 | 16 | 1.00 | 16 | 1.00 | 16 | 1.00 |
| cyto | 63 | 28 | 0.44 | 46 | 0.73 | 60 | 0.95 |
| proc | 14 | 11 | 0.79 | 12 | 0.86 | 12 | 0.86 |
| sega | 145 | 93 | 0.64 | 115 | 0.79 | 132 | 0.91 |
| tile | 129 | 58 | 0.45 | 99 | 0.77 | 117 | 0.91 |
| tiss | 41 | 28 | 0.68 | 35 | 0.85 | 39 | 0.95 |
| kegg | 38 | 9 | 0.24 | 22 | 0.58 | 30 | 0.79 |
| TOTAL | 446 | 243 | 0.54 | 345 | 0.77 | 406 | 0.91 |

Table B.2: Significant pathways for different thresholds for the *p*-value with at least 20 genes per pathway

| Name | # (CNE $\cap$ GEE) | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p < 0.01$ | | $p < 0.05$ | | $p < 0.10$ | |
| arms | 16 | 16 | 1.00 | 16 | 1.00 | 16 | 1.00 |
| cyto | 14 | 7 | 0.50 | 10 | 0.71 | 13 | 0.93 |
| proc | 10 | 9 | 0.90 | 10 | 1.00 | 10 | 1.00 |
| sega | 105 | 72 | 0.69 | 87 | 0.83 | 97 | 0.92 |
| tile | 21 | 10 | 0.48 | 17 | 0.81 | 19 | 0.90 |
| tiss | 30 | 22 | 0.73 | 27 | 0.90 | 29 | 0.97 |
| kegg | 15 | 5 | 0.33 | 9 | 0.60 | 12 | 0.80 |
| TOTAL | 211 | 141 | 0.67 | 176 | 0.83 | 196 | 0.93 |

Table B.3: Significant pathways for different thresholds for the *p*-value with at least 40 genes per pathway

| Name | # PW<br>$p_{GSA} < 0.01$ | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p_{GSS} < 0.01$ | | $p_{GSS} < 0.05$ | | $p_{GSS} < 0.10$ | |
| arms | 0 | 0 | - | 0 | - | 0 | - |
| cyto | 3 | 1 | 0.33 | 1 | 0.33 | 2 | 0.67 |
| proc | 3 | 3 | 1.00 | 3 | 1.00 | 3 | 1.00 |
| sega | 30 | 22 | 0.73 | 23 | 0.77 | 23 | 0.77 |
| tile | 6 | 0 | 0.00 | 2 | 0.33 | 2 | 0.33 |
| tiss | 4 | 4 | 1.00 | 4 | 1.00 | 4 | 1.00 |
| kegg | 5 | 1 | 0.20 | 3 | 0.60 | 4 | 0.80 |
| TOTAL | 51 | 31 | 0.61 | 36 | 0.71 | 38 | 0.75 |

Table B.4: Detected significant pathways of the GSA ($p_{GSA} < 0.01$) by significant GSS with at least 20 genes

| Name | # PW<br>$p_{GSA} < 0.01$ | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p_{GSS} < 0.01$ | | $p_{GSS} < 0.05$ | | $p_{GSS} < 0.10$ | |
| arms | 0 | 0 | - | 0 | - | 0 | - |
| cyto | 0 | 0 | - | 0 | - | 0 | - |
| proc | 2 | 2 | 1.00 | 2 | 1.00 | 2 | 1.00 |
| sega | 19 | 15 | 0.79 | 15 | 0.79 | 15 | 0.79 |
| tile | 3 | 0 | 0.00 | 1 | 0.33 | 2 | 0.67 |
| tiss | 1 | 1 | 1.00 | 1 | 1.00 | 1 | 1.00 |
| kegg | 1 | 1 | 1.00 | 1 | 1.00 | 1 | 1.00 |
| TOTAL | 26 | 19 | 0.73 | 20 | 0.77 | 21 | 0.81 |

Table B.5: Detected significant pathways of the GSA ($p_{GSA} < 0.01$) by significant GSS with at least 40 genes

| Name | # PW<br>$p_{GSA} < 0.05$ | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p_{GSS} < 0.01$ | | $p_{GSS} < 0.05$ | | $p_{GSS} < 0.10$ | |
| arms | 0 | 0 | - | 0 | - | 0 | - |
| cyto | 35 | 11 | 0.31 | 14 | 0.40 | 16 | 0.46 |
| proc | 5 | 4 | 0.80 | 4 | 0.80 | 4 | 0.80 |
| sega | 57 | 34 | 0.60 | 37 | 0.65 | 37 | 0.65 |
| tile | 101 | 39 | 0.39 | 48 | 0.48 | 55 | 0.54 |
| tiss | 10 | 9 | 0.90 | 9 | 0.90 | 9 | 0.90 |
| kegg | 13 | 5 | 0.38 | 6 | 0.46 | 7 | 0.54 |
| TOTAL | 221 | 102 | 0.46 | 118 | 0.53 | 128 | 0.58 |

Table B.6: Detected significant pathways of the GSA ($p_{GSA} < 0.05$) by significant GSS with at least 10 genes

| Name | # PW<br>$p_{GSA} < 0.10$ | # PW | Ratio | # PW | Ratio | # PW | Ratio |
|---|---|---|---|---|---|---|---|
| | | $p_{GSS} < 0.01$ | | $p_{GSS} < 0.05$ | | $p_{GSS} < 0.10$ | |
| arms | 1 | 0 | 0.00 | 0 | 0.00 | 0 | 0.00 |
| cyto | 81 | 25 | 0.31 | 40 | 0.49 | 43 | 0.53 |
| proc | 8 | 5 | 0.63 | 5 | 0.63 | 5 | 0.63 |
| sega | 77 | 44 | 0.57 | 48 | 0.62 | 48 | 0.62 |
| tile | 183 | 71 | 0.39 | 88 | 0.48 | 97 | 0.53 |
| tiss | 12 | 11 | 0.92 | 11 | 0.92 | 11 | 0.92 |
| kegg | 22 | 8 | 0.36 | 13 | 0.59 | 14 | 0.64 |
| TOTAL | 384 | 164 | 0.43 | 205 | 0.53 | 218 | 0.57 |

Table B.7: Detected significant pathways of the GSA ($p_{GSA} < 0.10$) by significant GSS with at least 10 genes

# APPENDIX C

## DVD Content

A digital version of this thesis, all the program code (R-Files) and most of the workspace files (RData-Files), the original data files (txt-Files and xls-Files), various plots (PDF-Files) as well as most of the cited papers are attached in a DVD. For iCluster there are always two folders including both Sarcoma data set (37 or 64 samples). As the case with 64 was done after the other one the R-Code is better commented.

# Bibliography

[1] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[2] Andrea Bild and Phillip G. Febbo. Application of a priori established gene sets to discover biologically important differential expression in microarray data. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15278–15279, October 2005.

[3] Carlo E. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. In *Studi in Onore del Professore Salvatore Ortu Carboni*, pages 13–60. Rome, 1935.

[4] Carlo E. Bonferroni. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze*, 8:3–62, 1936.

[5] Sung Choe, Michael Boutros, Alan Michelson, George Church, and Marc Halfon. Preferred analysis methods for Affymetrix GeneChips revealed by a wholly defined control dataset. *Genome Biology*, 6(2):R16+, 2005.

[6] Arthur Dempster, Nan Laird, and Donald Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39, 1977.

[7] Chris Ding and Xiaofeng He. K-means clustering via principal component analysis. pages 225–232. ACM Press, 2004.

[8] Bradley Efron and Robert J. Tibshirani. On testing the significance of sets of genes. *Annals of Applied Statistics*, pages 107–129, 2006.

[9] Edward B. Fowlkes and Colin L. Mallows. A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78:553–569, 1983.

[10] Major Greenwood. The natural duration of cancer. *The Series Report on Public Health and Medical Subjects*, 33:1–26, 1926.

[11] Lei Guo, Edward K. Lobenhofer, Charles Wang, Richard Shippy, Stephen C. Harris, Lu Zhang, Nan Mei, Tao Chen, Damir Herman, Federico M. Goodsaid, Patrick Hurban, Kenneth L. Phillips, Jun Xu, Xutao Deng, Yongming Sun Andrew, Weida Tong, Yvonne P. Dragan, and Leming Shi. Rat toxicogenomic study reveals analytical consistency across microarray platforms. *Nat Biotechnol*, 24(9):1162–9, 2006.

[12] Wendy J. Hall and Jon A. Wellner. Confidence bands for a survival curve from censored data. *Biometrika*, 67:133–143, 1980.

[13] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *J. Educ. Psych.*, 24, 1933.

[14] Guohong Hu, Robert A. Chong, Qifeng Yang, Yong Wei, Mario A. Blanco, Feng Li, Michael Reiss, Jessie L.-S. Au, Bruce G. Haffty, and Yibin Kang. Mtdh activation by 8q22 genomic gain promotes chemoresistance and metastasis of poor-prognosis breast cancer. *Cancer Cell*, 15(1):9 – 20, 2009.

[15] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1):1–13, 2009.

[16] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nature Protocols*, 4(1):44–57, 2009.

[17] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. Le Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[18] Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et des jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37:547–579, 1901.

[19] Edward L. Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53:457–481, 1958.

[20] Oliver Kühnle and Georg Pfundstein. *Gene Expression and Pathway Analysis*, 2009. Statistical Research Project.

[21] Ken C. Lo, Leighton C. Stein, Jenniffer A. Panzarella, John K. Cowell, and Lesleyann Hawthorn. Identification of genes involved in squamous cell carcinoma of the lung using synchronized data from dna copy number and transcript expression profiling analysis. *Lung Cancer*, 59(3):315 – 331, 2008.

[22] Eric F. Lock, Katherine A. Hoadley, J.S. Marron, and Andrew B. Nobel. Joint and individual variation explaind (jive) for integrated analysis of mutiple datatypes. Technical report, University of North Carolina at Chapel Hill, https://genome.unc.edu/jive/, 02 2011.

[23] Karl Pearson. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572, 1901.

[24] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2011. ISBN 3-900051-07-0.

[25] William M. Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, 66(336):846–850, 1971.

[26] Martin Schumacher and Gabi Schulgen. *Methodik klinischer Studien*, volume 77-89. Springer, 2nd edition, 2006.

[27] Ronglai Shen. *iCluster: Integrative clustering of multiple genomic data types*, 2010. R package version 1.2.0.

[28] Ronglai Shen, Adam B. Olshen, and Marc Ladanyi. Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis. *Bioinformatics/computer Applications in The Biosciences*, 25:2906–2912, 2009.

[29] Ronglai Shen, Sijian Wang, and Qianxing Mo. Sparse integrative clustering of multiple omics data sets. Technical report, University of Wisconsin Department of Biostatistics and Medical Informatics, 06 2011.

[30] Matt Shotwell. *profdpm: Profile Dirichlet Process Mixtures*, 2011. R package version 3.0.

[31] Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha, Sayan Mukherjee, Benjamin L. Ebert, Michael A. Gillette, Amanda Paulovich, Scott L. Pomeroy, Todd R. Golub, Eric S. Lander, and et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

[32] Adi Laurentiu L. Tarca, Sorin Draghici, Purvesh Khatri, Sonia S. Hassan, Pooja Mittal, Jung-Sun S. Kim, Chong Jai J. Kim, Juan Pedro P. Kusanovic, and Roberto Romero. A novel signaling pathway impact analysis. *Bioinformatics (Oxford, England)*, 25(1):75–82, January 2009.

[33] Barry S. Taylor, Jordi Barretina, Nicholas D. Socci, Penelope DeCarolis, Marc Ladanyi, Matthew Meyerson, Samuel Singer, and Chris Sander. Functional copy-number alterations in cancer. *PLoS ONE*, 3(9):e3179, 09 2008.

[34] Terry Therneau. *survival: Survival analysis, including penalised likelihood*, 2011. R package version 2.36-9.

[35] Robert J. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

[36] Virginia G. Tusher, Robert J. Tibshirani, and Gilbert Chu. Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, 98(9):5116–5121, April 2001.

[37] David L. Wallace. A method for comparing two hierarchical clusterings: comment. *Journal of the American Statistical Association*, 78:569–576, 1983.

[38] Ernst Wit and John McClure. *Statistics for Microarrays*, volume 1-265. Wiley, 1st edition, 2004.

[39] Daniela M. Witten and Robert J. Tibshirani. A comparison of fold-change and the t-statistic for microarray data analysis. *Analysis*, pages 1–15, 2007.

[40] Daniela M. Witten and Robert J. Tibshirani. Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1), 2009.

[41] Hongyuan Zha, Xiaofeng He, Chris Ding, Horst Simon, and Ming Gu. Spectral relaxation for k-means clustering. pages 1057–1064. MIT Press, 2001.

# Declaration of academic honesty

I hereby declare that I wrote this diploma thesis on my own. Concerning this I have not used any other documents and aids than those stated above.

New York City, $30^{th}$ November 2011

Oliver S Kühnle

# Eidesstattliche Erklärung

Hiermit erkläre ich, dass ich die vorliegende Diplomarbeit selbstständig verfasst habe. Hierzu habe ich keine anderen als die angegebenen Quellen und Hilfsmittel benutzt.

New York City, 30. November 2011

Oliver S Kühnle