Stockholms universitet

On the Linkage Between CMAC Age and its Morphological and Biological Features

Billy Pettersson Pablo

Kandidatuppsats i matematisk statistik
Bachelor Thesis in Mathematical Statistics

Kandidatuppsats 2012:8
Matematisk statistik
September 2012

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Mathematical Statistics
Stockholm University
Bachelor Thesis **2012:8**
http://www.math.su.se

On the Linkage Between CMAC Age and its Morphological and Biological Features

Billy Pettersson Pablo[*]

September 2012

Abstract

As metastasis is the primary cause of cancer lethality, understanding cell migration and cell adhesion plays a major roll in unraveling the mechanisms underlying cancer progression. In the laboratory of Staffan Strömblad at Karolinska Institutet, the coordinating unit of a EU network of excellence, researchers use a combination of advanced experimental setups, fluorescence microscopy and quantitative statistical analysis to investigate the systems of cell migration. As the behavior of cancer cells is partly affected by cell-matrix adhesion complexes (CMACs), attachments to the extracellular matrix, in this thesis we investigate the possibility of describing the relation between the age of a CMAC and a number of related features. Our findings demonstrate varying degrees of dependence due to the treatments used on the cellular level and due to the grouping of CMACs.

[*] Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: billy.pettersson@gmail.com . Supervisor: Mehrdad Jafari Mamaghani.

Acknowledgements

This paper constitutes a 15 ECTS thesis, which leads to a Bachelor's degree in Mathematical Statistics at the Department of Mathematics at Stockholm University.

I would like to express my sincere gratitude to my supervisor, Mehrdad Jafari Mamaghani, for guidance, patience, and invaluable support.

Stockholm, September 2012
Billy Pettersson Pablo

Contents

1 Introduction

1.1 Background

Cell migration and cell adhesion constitute the backbone of many physiological processes such as development, wound healing and disease progression. Specifically, cell migration and cell adhesion are fundamental to the process of metastasis, which itself is a major driver of cancer mortality. Thus, investigating cell migration and cell adhesion is of primary and principal importance in understanding the mechanisms behind cancer progression. In the laboratory of Staffan Strömblad at Karolinska Institutet, the coordinating laboratory of an EU network of excellence in Systems Microscopy, long-term efforts have been launched to conduct biological research aimed to illuminate the processes of cell migration and cell adhesion using an integrated platform of advanced microscopy, data mining and statistical modelling: a Systems Microscopy research platform[6]. More specifically, the aim of the platform is to focus on both the cellular and molecular level. In particular, the focus on the sub-cellular level has been on cell-matrix adhesion complexes (CMACs) and the F-actin network. CMACs act as communication hubs between the cell and its environment, the extra-cellular matrix. In microscopic analysis of cells, CMACs are detected and segmented through automated algorithms, recording measurements of a number of morphological and dynamical properties (see Figure 1).

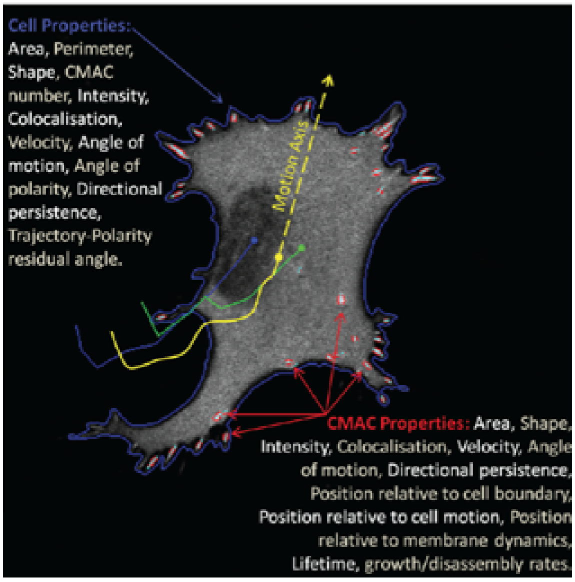
Cell Properties:
Area, Perimeter, Shape, CMAC number, Intensity, Colocalisation, Velocity, Angle of motion, Angle of polarity, Directional persistence, Trajectory-Polarity residual angle.
Motion Axis
CMAC Properties: Area, Shape, Intensity, Colocalisation, Velocity, Angle of motion, Directional persistence, Position relative to cell boundary, Position relative to cell motion, Position relative to membrane dynamics, Lifetime, growth/disassembly rates.

Figure 1: *Image display of a cell with examples of measured cell properties and CMAC properties.*

1.2 Cell-matrix adhesion complex

CMACs with their dynamic nature act as an information linkage between the extra-cellular matrix and parts of the cytoskeleton, and are thus interacting in most of the major cell functions[7]. With this in mind CMACs are an obvious target in the study of cancer cell migratory behaviour.

1.3 Description of data

The data set consists of 137605 observations of a number of CMAC units with a total number of 34 variables. Among these, one variable describes the speed of the cell, and 29 variables describe different features of CMACs (for a complete list, see 6.1.1 in the Appendix). We also have 4 indexing variables, that specify the cell ID, CMAC ID, perturbation, and time of each observation. The variable time is in itself not important, but it is used to create a variable *CMAC age*, that specifies the age of a CMAC for each time it has been observed.

However, CMACs have a wide variety of life expectancies, so as the function of a CMAC unit depends not only on how long it has existed, but rather on what stage of its life it is in, we will find it useful to construct a new variable, *b2d* (birth to death), ranging from 0 to 1, that tells us what proportion of its life a CMAC has passed at the time of the observation. Furthermore, using *b2d* we can now separate observations into, for example, three different groups, where in group one we have the observations where the CMACs are in an early stage of their life, group two characterizes middle-aged CMAC-s, and group three has CMACs in their final stages of their lives, close to disassembly. The exact number of groups shall be decided using statistical tests, which we will see in Section 3.4.

Now, as we do not know for how long the CMAC units that are observed at the onset of the experiment (i.e. time=0) have lived, we will have to remove these from our data. Neither do we know for how long the CMAC units at the end of the experiment will continue to live on, and therefore we will remove these as well. Hence, our data set is reduced to 115711 observations.

One may also note that, as we have four different perturbations, that are geared to manipulate the behaviour of the cells in different directions, we will have to treat our data as four different subsets, with the first one being a group with cells that have received a *DMSO* (control) treatment, and the other three containing observations of cells having received the three treatments of *Blebbistatin*, *RhoActivator*, or *RockInhibitor* respectively[1]. These four data sets will be named *pert0*, *pert1*, *pert2*, and *pert3* respectively.

2 Methods

Our main objective is to fit a linear model to describe the proportional age, *b2d*, of a CMAC unit using a number of variables. To do this we will perform linear regression and feature selection, using both stepwise regression and the elastic net. We will also, through multivariate analysis of variance, MANOVA, examine the possibilities of dividing CMAC units into different age categories depending on the variable *b2d*. As there is a risk that some methods lose effectiveness when background variables are too highly correlated, which we suspect to be the case with our data, the initial task is to counter this by the usage of principal component analysis, PCA. Another justification for using PCA is to transform the data set so as to meet the requirements of distribution symmetry (approximate normality).

2.1 Regression analysis

Regression analysis is a statistical procedure which includes a number of techniques for analyzing the relationship between one or more response variables and one or more background variables. In our case, regression is used for estimating the response variable given the background ones.

2.1.1 Linear regression

In a linear regression model, one has made an assumption, among others (see section 6.3.1in Appendix), claiming that the relationship between the response variable and the explanatory variables is linear. Through adding an extra term, ϵ, that will explain the disturbance we get the following linear model,

$$\mathbf{y} = \alpha + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1}$$

where α is the intercept, and $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$ are vectors/matrices of the observed data, the estimated parameters, and the errors respectively, as defined below.

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1m} \\ x_{21} & \cdots & x_{2m} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nm} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_m \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix},$$

[1] In crude terms specific to this study Blebbistatin and RockInhibitor lead to faster cells and decreased CMAC lifetimes while RhoActivator does the contrary on both levels.

where m is the number of variables, and n is the number of observations. This, (1), is called the regression function. There are a number of methods to decide the parameters in $\boldsymbol{\beta}$, and in our case the criterion of ordinary least squares (OLS) will be used. The variable $\mathbf{y}$ will represent observations on *b2d*, while $\mathbf{X}$ represents observations on remaining variables.

A linear model estimated by OLS is achieved by minimizing the squared orthogonal distances between observed values of the response variable and the values estimated by the regression function, as seen in 6.3.1.

2.1.2 Coefficient of determination

The coefficient of determination, R^2, is the proportion of the variance in the response variable, that can be explained by the variance in the explanatory variables of the model. It is defined as $R^2 = \frac{SS_{model}}{SS_{total}}$, and ranges from 0 to 1. It is commonly used when analyzing a fit of a model.

The adjusted R^2, $\bar{R}^2$, is a modified version of R^2, where the number of explanatory variables is taken into account. It is defined as follows,

$$\bar{R}^2 = 1 - \frac{SS_{error}df_{total}}{SS_{total}df_{error}},$$

where df_{total} and df_{error} are degrees of freedom, $n-1$ and $n-m-1$ respectively, and the summed squares are defined as

$$SS_{model} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2,$$

$$SS_{error} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

$$SS_{total} = \sum_{i=1}^{n}(y_i - \bar{y})^2,$$

where $\hat{y}_i,\ y_i,$ and $\bar{y}$ are i^{th} predicted, i^{th} observed, and mean of the response variable.

2.1.3 Feature selection: stepwise regression

In stepwise regression we start off with only an intercept and the response variable, and then alternate adding a significant variable, or removing a non-significant one. In each step an F-statistic is computed to test a model with or without a certain variable, with the null-hypothesis being that the variable has a zero coefficient if in the model.

2.1.4 Feature selection: the elastic net

In statistics and in particular in the fitting of linear regression models, the main trade-off is between the accuracy and simplicity of the model. The elastic net is a regularized regression technique which is used to reduce the number of predictors in a regression model by means of smooth shrinkage[8]. The method is useful when there are a number of highly correlated variables, where the elastic net's grouping effect leads to consistent results. The formulation of elastic net is to be found at 6.3.2 in the Appendix.

2.2 MANOVA

We will find ourselves in situations where we have to explore whether multivariate groups of data are different, and as we have multiple explanatory variables, the procedure of multivariate analysis of variance, MANOVA, is a suiting solution for this task. As Hotelling's T2 is a test of whether two vectors of means are sampled from the same distribution, MANOVA analogously compares means of two or more groups. The method provides a measure of the overall likelihood of two or more random vectors of means being derived from the same distribution, by using the covariance matrices of data[5]. For the mathematical definitions and procedure of MANOVA, see 6.3.3 in Appendix.

As in the case with ANOVA, where comparing groups pairwise increase the risk of type I errors [2], MANOVA also has problems of multiple post hoc comparisons. It does not specify which groups that differ from each other.

We will use MANOVA to explore how changes in explanatory variables affect the response variables. Mainly to find out how the four perturbations affect cells, and to test hypotheses of whether the explanatory variables can be used to predict the response ones. We will also use MANOVA in order to find a relation between CMAC age and different CMAC features, and thus create a way of age categorizing CMAC units.

2.3 The EM-algorithm

The expectation-maximization algorithm is an iterative procedure with the aim of finding maximum likelihood estimates of parameters in statistical models. The idea of the algorithm is to improve estimates through alternating between an expectation step and a maximization step, where the expectation step computes the log-likelihood of the model using the current parameter estimates, while in the maximization step new parameter estimates are obtained based on the result of the previous step by maximizing

[2] Type I errors occurs when the null hypothesis is rejected, even though it is true.

the log-likelihood of the model.[3] We will use the algorithm to cluster data into a given number of groups, and see if these groups are resemblant of the age categories.

2.4 Akaike information criterion

The Akaike information criterion, AIC, is used to measure the relative goodness of fit of a model, in comparison to other models, when both the complexity and precision of the models are taken into account. The criterion penalizes models as they increase in complexity[2]. Mathematical definition is found in appendix: 6.3.4. We will use AIC to determine the optimal number of groups given the groups' composition after having run the EM-algorithm.

2.5 Principal component analysis

Principal component analysis, PCA, is a method of linear orthogonal transformation, with the aim of transforming a data set into a linearly non-correlated set of variables called principal components, i.e. converting the covariance matrix into a diagonal matrix[1]. This is done either by an eigenvalue decomposition of the covariance matrix of data or a singular value decomposition of the normalized data matrix. The principal components are ordered in terms of their representative variance, thus leading to a reduction of inputs when the original data set consists of many correlated variables. Note that they do not affect the spatial distances between observations, but only rotate data in a preferable direction.

3 Statistical analysis

3.1 PCA Transformation

In order to solve the phenomenon of multicollinearity between our 29 explanatory variables, we transform our data into a new set, using principal component analysis. Figure 2 and the table below display correlations between variables and the variance accounted for when each new principal component is added in the PCA-transformed *pert0* data set.

As the figure shows, we have no correlation between the principal components of our data. Note, however, that we have not changed the spatial distances between observations, but only introduced new axes. In the table we find that approximately 95% of the variance is covered after adding only 14 principal components.

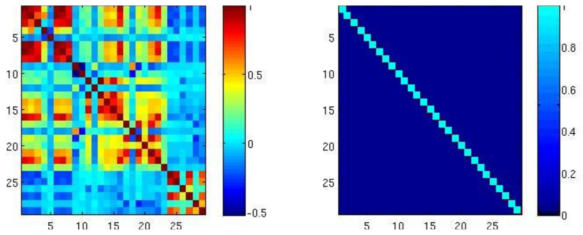

Figure 2: *Correlation between variables within data sets. To the left we have pert0 data, and to the right PCA transformed pert0 data.*

	pert0	*pert1*	*pert2*	*pert3*
1	0.3355	0.3096	0.3220	0.2924
2	0.4566	0.4797	0.4763	0.4774
3	0.5497	0.5621	0.5796	0.5680
4	0.6254	0.6381	0.6494	0.6352
5	0.6881	0.7035	0.7066	0.6968
6	0.7315	0.7586	0.7503	0.7492
7	0.7677	0.7999	0.7922	0.7903
8	0.8026	0.8348	0.8297	0.8248
9	0.8363	0.8682	0.8628	0.8541
10	0.8641	0.8944	0.8904	0.8789
11	0.8905	0.9163	0.9107	0.9027
12	0.9136	0.9358	0.9274	0.9202
13	0.9304	0.9488	0.9419	0.9372
14	0.9458	0.9604	0.9539	0.9508
15	0.9571	0.9703	0.9649	0.9635
16	0.9671	0.9759	0.9735	0.9739
17	0.9755	0.9812	0.9802	0.9806
18	0.9814	0.9852	0.9852	0.9852
19	0.9849	0.9887	0.9892	0.9891
20	0.9880	0.9914	0.9918	0.9922
21	0.9908	0.9935	0.9941	0.9940
22	0.9936	0.9951	0.9963	0.9955
23	0.9958	0.9965	0.9975	0.9969
24	0.9974	0.9977	0.9984	0.9981
25	0.9986	0.9987	0.9991	0.9991
26	0.9993	0.9994	0.9995	0.9996
27	0.9996	0.9998	0.9998	0.9998
28	0.9999	1.0000	0.9999	1.0000
29	1.0000	1.0000	1.0000	1.0000

Table 1: *Accumulated variances accounted for in a data set when adding 1 or up to 29 principal components.*

3.2 Grouping by perturbation

We are now interested to investigate the existence of any dependence between a perturbation and CMAC age. As we now have 29 variables depending on the variable *pert*, employing MANOVA would be a suitable choice. We want to find out whether the effects of the four perturbations differ or not, and therefore seek the dimension of the group means. This leads us to formalize and test a null hypothesis, H_0, stating that the expected means are the same for all groups.

After conducting the tests we obtain a vector of p-values, where the first p-value tests whether the dimension is 0, i.e. all means being equal, and the other two tests whether the dimension is 1 or 2. However, all p-values are approximately 0, indicating that the four multivariate means fall in a three dimensional room, and thus leading us to conclude that the four groups differ significantly, a conclusion supported by Figure 3, which displays a scatter plot of data represented by two canonical variables[3].

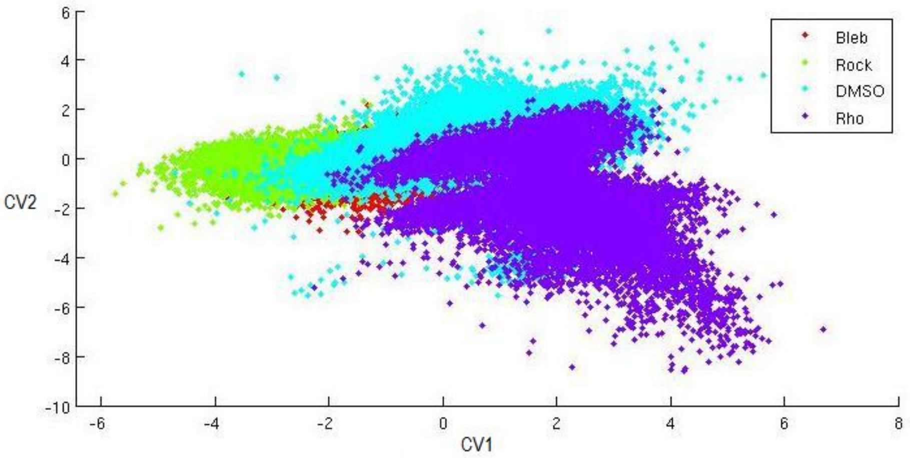

Figure 3: *This scatter plot of the full data set, represented with 2 canonical variables and colour marked by our four perturbations, suggests that the effects of the different perturbations differ quite significantly.*

With these results in mind, we reject the null hypothesis, and hence we will from here on consider our data as derived from four significantly different distributions. Our four data sets will be labelled *pert0* to *pert3*, and contain number of observations as according to the table below.

[3] As opposed to PCA, the resulting canonical vectors in MANOVA are selected in a manner to find linear combinations of the original variables that produce the largest separation between groups with respect to the within-group variations.

Data set	*pert0*	*pert1*	*pert2*	*pert3*
Number of observations	60518	7844	18251	29098

Table 2: *Number of observations in the subgroups of data.*

3.3 Linear regression

3.3.1 An initial model

An initial linear regression model per data set, will give us coefficients of determination as presented in the table below. As the models contain all explanatory variables, these are the maximum values of R^2 achievable by linear regression analysis on our four complete data sets. The parameter estimates are to be found in Table 9 in appendix. Note that we have used the principal components in this case due to their relatively more symmetric distributions and thus, superior suitability for OLS regression modelling compared to our original data.

Condition	pert0	pert1	pert2	pert3	full data set
R^2	0.1555	0.0752	0.1071	0.0334	0.1065

Table 3: *Determination coefficients derived from linear regression analyses on our data sets.*

Though the R^2 values might be considered quite small we can not exclude the hypothesis of some of the regression coefficients being statistically significant. There might be important relationships between the response variable and explanatory variables, even though they do not explain much of the variance in the response variable.

In order for the models to fit, we require for the residuals to be normally distributed with a constant variance. To find out whether this is the case or not, we will examine a few plots. In Figure 4 below, we find the normal probability plots of the residuals of our four data sets. We want the residuals to be normally distributed, as oppose to following some trend. Normally distributed observations are expected to be scattered tightly along the diagonal line, which is not completely our case.

However, we will in Section 3.3.2 try to improve this fit, by excluding some observations of our data due to limitations regarding measurement precision.

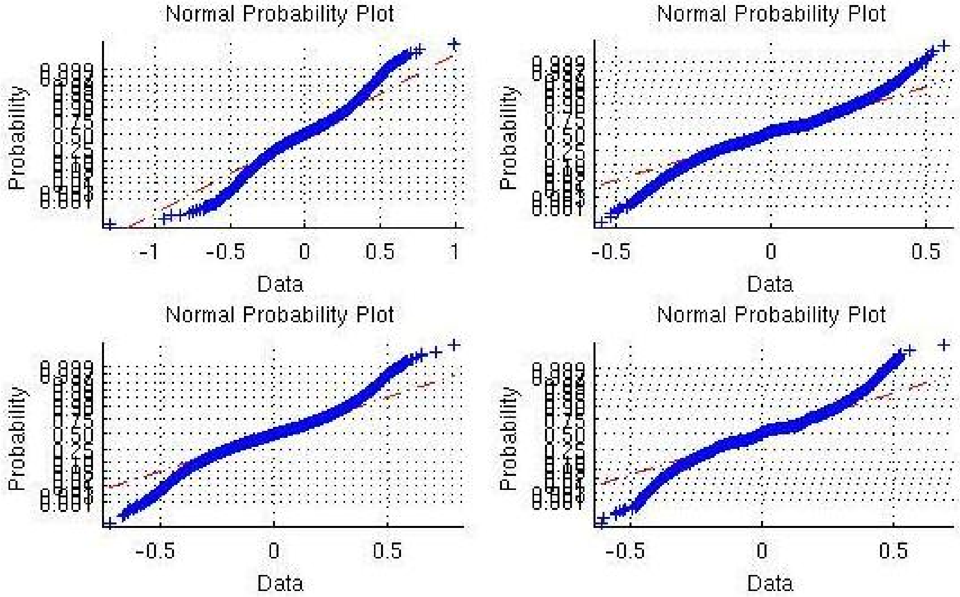

Figure 4: *Normal probability plots of the residuals of our four regression models.*

3.3.2 Reduced data set

The precision of the variable *b2d* is dependent on the length of the time intervals between observations. Let us say, for example, that a CMAC is observed only once. Then its whole lifetime, with its different features, will be summed up in only one observation. This gives us reason to analyse a reduced data set, where such observations have been excluded.

We exclude observations where *b2d*=0 from our data, and investigate the outcome. First thing we find is that all R^2 values have increased (Table 4 below), though not remarkably. Neither in the normal probability plots of the residuals, Figure 9 in Appendix, can we spot any significant improvements. However, for the reasons stated in the beginning of this section, continued analysis will be performed on reduced data sets.

Data set	pert0	pert1	pert2	pert3	full data set
Number of Observations	59241	6824	17992	25244	109301
R^2	0.1577	0.0841	0.1086	0.0372	0.1096

Table 4: *Number of observations per reduced data set, and determination coefficients of linear regression models.*

3.3.3 Feature selection

Let us now find out if we can reduce the number of inputs through different means of feature selection. With stepwise regression we can keep R^2 almost the same size as of full models, though the numbers of explanatory variables have decreased (see Table 5). For parameter estimates, see Table 10.

Figure 10 presents the contributing variables based on the elastic net. With even more stringent requirements, elastic net in combination with tenfold cross validation, only the variables presented in Table 17 pass the tests of significance and contribute to our EN-models. Figure 5 displays a summary of the elastic net procedure. It shows how the R^2 values change when excluding variables from the models. It appears as if we can remove a number of explanatory variables while keeping the coefficients of determination somewhat unchanged.

Condition	pert0	pert1	pert2	pert3	full data set
R^2	0.1553	0.0728	0.1056	0.0325	0.0687
$\bar{R}^2$	0.1550	0.0716	0.1049	0.0320	0.0685
Number of variables	24	10	14	17	24

Table 5: *Results from stepwise regression modelling on data.*

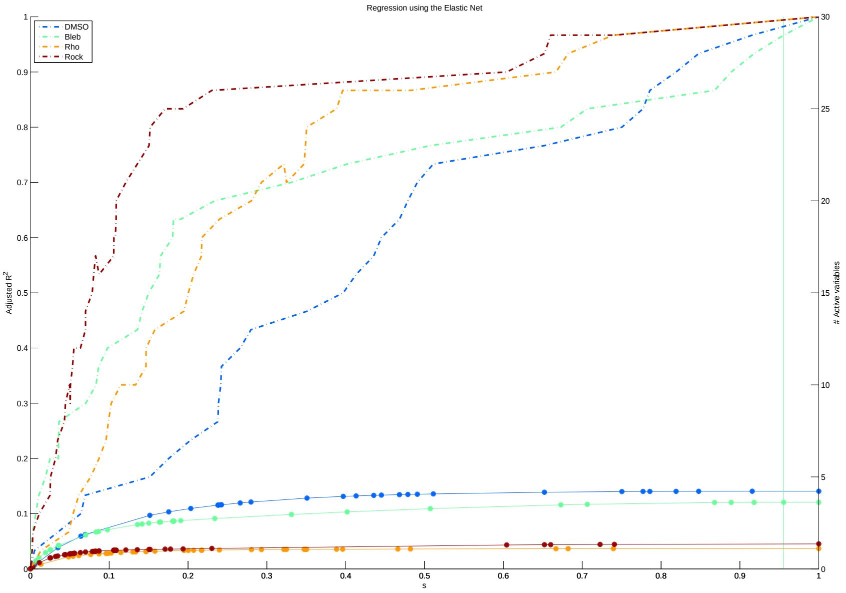

Figure 5: *Summary of the variable selection process of elastic net, with R^2 represented as dotted lines.*

3.4 Categorization by age

In this chapter we will try to define the groups in which to divide our observations. This will be done through using on one hand MANOVA, and on the other hand the EM-algorithm.

3.4.1 MANOVA

As mentioned, we will divide observations into groups, depending on what stage of its life the CMAC unit is currently in at the time of the observation. At first we need to decide the preferable number of age categories, and therefore use the method of multivariate analysis of variance, MANOVA.

Table 18 in appendix presents p-values for testing whether the means lie in a space of dimension 0, 1, and so on. In conclusion we can say that for the majority of our data sets we can, on a 5% significance level, reject the hypothesis of the means lying in a three dimensional space or less. Unfortunately, MANOVA does not reveal which groups that are having vectors of means that differ from the rests, but we hope to get some clarifications using the expectation-maximization algorithm.

3.4.2 EM-algorithm

Results presented in this chapter are achieved through the employment of the EM-algorithm. Our intention is to decide how to split each data set into different groups, and hope that these groups will be resemblant of different age categories. In Table 19, we have used the Akaike information criterion to select preferable ones among EM-algorithm groupings. However, if we look at the *b2d* means of the subgroups (Figure 6) we can spot no obvious distinctions in sample distribution, although some means appear to be higher/lower than the rest.

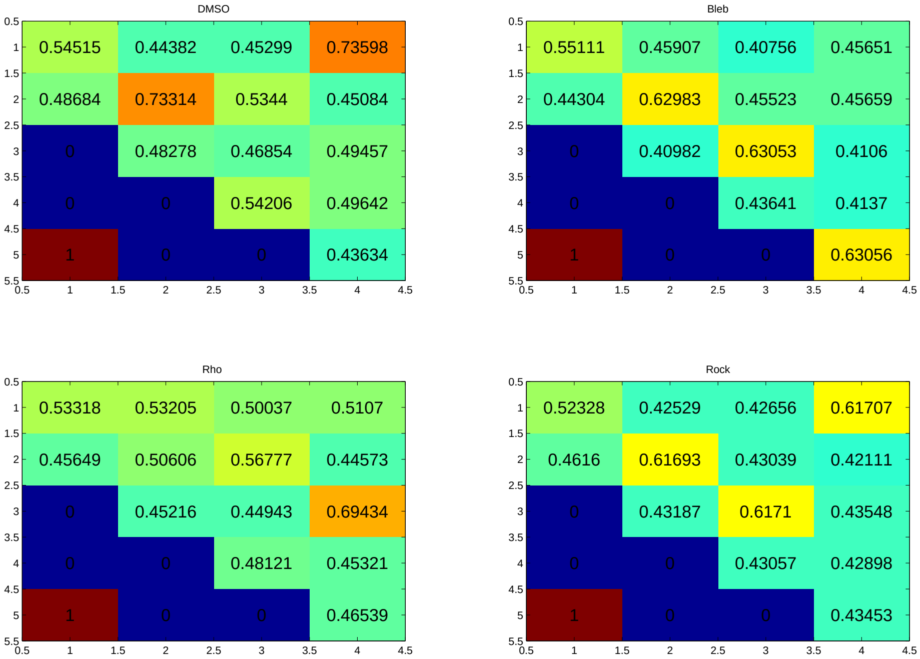

Figure 6: *Means of the values observed on the b2d variable in the subgroups given by the EM-algorithm. On the x axis we have 4 different groupings of each data set, from 2-5 subgroups, and on the y axis we can see the b2d means of each subgroup.*

4 Extra analysis of *pert0* data

As we can observe a great variety of lengths on CMAC life-span, we decide to split the data set *pert0* into a number of subgroups, and perform an extra analysis. In this case, a subgroup contains all observations of each CMAC unit that will reach a certain age. Data is named from *pert00* to *pert05* and contain observations as described in the table below.

Data set	pert00	pert01	pert02	pert03	pert04	pert05
Age at disassembly	< 5	≤ 10 > 5	≤ 20 > 10	≤ 40 > 20	≤ 50 > 40	> 50
Number of observations	19935	19433	14984	5646	333	187

Table 6: *Distribution of observations from pert0 data into subgroups.*

4.1 Linear regression

We will now go through results of regression analysis performed on data sets *pert00-pert05*, in a similar fashion as we did with the *pert0-pert3* data. To begin with, after computing OLS parameter estimates (Table 14), we notice an immense difference in R^2 numbers compared to what we have observed earlier, as well as well fitted residual plots (Figure 8), especially in the cases of *pert03-05*. However, we assume that the reason for this is that in the case of short life periods, the CMACs go rapidly through different stages in a manner that lets us treat observations as independent (random), while in the case of long life CMACs the changes they go through between observations are less palpable. Therefore it might be a good idea to disregard results based on those data sets. Another legitimate reason for cautionary interpretations is the number of unique CMAC units involved in the analys. On both of the counts mentioned above the groups *pert04* and *pert05* fail to meet our criteria.

Regarding the elastic net, R^2 based selection gives contributing variables listed in Figure 11, and Figure 7 summarizes the procedure. For the same reasons as stated in the above section, we can see instability in the variable selection for especially the *pert05* data set.

If we instead combine with the more stringent criteria of cross validation we get a model of variables listed in Table 16, with estimated parameters displayed in Table 15. However, due to the sparse number of variables passing the criteria, we lose great proportions of the R^2.

Condition	pert00	pert01	pert02	pert03	pert04	pert05
R^2_{full}	0.1049	0.1996	0.1882	0.1746	0.4166	0.7394
$R^2_{stepwise}$	0.1049	0.1993	0.1879	0.1723	0.3750	0.5501
$R^2_{e\_net,CV}$	0.0651	0.1686	0.1518	0.0535	0.0205	0.0473

Table 7: *Coefficients derived from stepwise regression modeling on subgroups of pert0. Estimated parameter coefficients are to be found in Table 13 and Table 14 of appendix.*

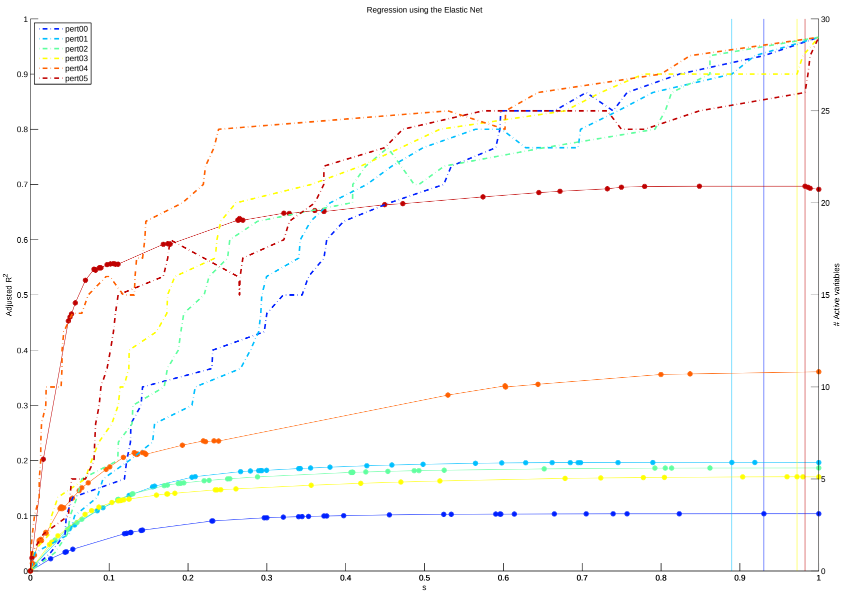

Figure 7: *Summary of the variable selection process of elastic net, with R^2 represented as dotted lines.*

5 Discussion

5.1 Biological aspects

As the role of cell migration and cell adhesion is fundamental to many physiological phenomena such as cancer progression and metastasis, studying cell migration and cell adhesion has been of primary importance in the molecular biology branch of cancer research. In this study the particular aim has been focused on cell-matrix adhesion complexes (CMACs) as they constitute the means through which the cell communicates with its environment.

Our analysis shows that the relationship between the age of the CMACs, and other morphological and biological features of CMACs undergoes palpable changes as the cells within which the CMACs are embedded are perturbed with different treatments.

5.2 Statistical aspects

Our main objective was to describe a relationship between the age of the CMACs and various explanatory variables. Though few things appear as naturally linear as oppose to, for example, logarithmic or of higher order, our approach was to try, with various methods, to find linear connections since those are the simplest and most understandable types to deal with when having a large number of explanatory variables, and therefore constitute a reasonable first analysis.

Regarding assumptions of independent observations, which is required when constructing a linear model, we have two cases. For a large data set of observations of short lived CMACs, we can assume approximate independence between observations. The reason for this is that those CMACs will rapidly develop and change their features. CMACS that do not disassemble quickly, on the other hand, keep their features for longer periods of time, and we are thus having a dependence between observations of the same CMAC.

However, in cases where multiple linear regression models do not appear to have a good fit, one still can not reject the hypothesis of some of the regression coefficients being statistically significant. There might be important relationships between the response variable and explanatory variables, even though they do not explain much of the variance in the response variable. Through the employment of feature selection methods we can present a number of selected variables that have passed the selection algorithm's criteria. Whether or not they are actually reasonable from a biological perspective is for the biologists to determine.

We use MANOVA to test if multivariate means are derived from the same sampling distribution, since MANOVA is an adequate method for the task. However, MANOVA does not tell which groups (if any) that differ from the rest. That is why we also incorporate the EM-algorithm. Though the indexing given by the EM-algorithm did not clearly match age any categories we can not reject the hypothesis that there is an efficient method of determining CMAC ages given certain explanatory variables.

Possible improvements upon our presented models can potentially be achieved by *i*) means of non-parametric/non-linear (*in silico*) models, and *ii*) expansion of biological experiments and inclusion of extended variable sets (*in vitro*).

6 Appendix

6.1 Tables

6.1.1 List of variables

	Descriptive	Indexing	Perturbations
1	CMAC Area	CMAC ID	DMSO
2	CMAC Major Axis	Cell ID	Bleb
3	CMAC Minor Axis	Pert	Rho
4	CMAC Eccentricity	Time	Rock
5	CMAC Angle		
6	CMAC Perimeter		
7	CMAC Convex Area		
8	CMAC Solidity		
9	CMAC DistBorder		
10	CMAC DistCenter		
11	CMAC MeanIRaw-ch1		
12	CMAC Local BG-ch1		
13	CMAC MeanI-ch1		
14	CMAC StdevI-ch1		
15	CMAC MaxI-ch1		
16	CMAC Integrated Intensity-ch1		
17	CMAC MeanIRaw-ch2		
18	CMAC Local BG-ch2		
19	CMAC MeanI-ch2		
20	CMAC StdevI-ch2		
21	CMAC MaxI-ch2		
22	CMAC Integrated Intensity-ch2		
23	CMAC Pearson-ch1vs2		
24	CMAC Growth of Area		
25	CMAC Length Growth		
26	CMAC Delta Intensity-ch1		
27	CMAC Delta Integrated Intensity-ch1		
28	CMAC Delta Intensity-ch2		
29	CMAC Delta Integrated Intensity-ch2		
30	Cell Speed		

Table 8: *List of measured biological and morphological variables and indexing variables.*

6.1.2 Regression models

	Data set			
Variables	pert0	pert1	pert2	pert3
intercept	0.6532	0.5423	0.603	0.5811
CMAC Area	-0.1083	-0.0599	0.0061	-0.1531
CMAC Major Axis	0.0302	0.037	0.0214	0.049
CMAC Minor Axis	-0.0537	-0.2031	-0.1591	-0.1799
CMAC Eccentricity	-0.0179	-0.0071	-0.0915	-0.007
CMAC Angle	0	0	0	0
CMAC Perimeter	0.0014	-0.0015	0.001	-0.0087
CMAC Convex Area	0.0411	-0.0022	-0.0066	0.1143
CMAC Solidity	-0.0163	0.0598	0.0216	0.0253
CMAC DistBorder	-0.0086	-0.0019	-0.0088	-0.0018
CMAC DistCenter	-0.0003	0.0005	0.0012	0.0002
CMAC MeanIRaw-ch1	0.1086	0.6087	-0.0927	-0.2232
CMAC Local BG-ch1	0.0015	-0.02	0.0032	-0.0005
CMAC MeanI-ch1	-0.1526	-0.5693	0.045	0.258
CMAC StdevI-ch1	-0.0002	0.0017	-0.0011	0.0026
CMAC MaxI-ch1	-0.0164	-0.0469	-0.0047	-0.1004
CMAC Integrated Intensity-ch1	0.0011	-0.0024	0.0002	0.0047
CMAC MeanIRaw-ch2	0.0706	0.0392	-0.0717	-0.0661
CMAC Local BG-ch2	0.001	0.0005	0.0042	0.0036
CMAC MeanI-ch2	-0.0641	-0.044	0.0303	0.0254
CMAC StdevI-ch2	-0.0007	0.0001	0.0011	-0.0011
CMAC MaxI-ch2	0.0052	-0.0049	-0.0199	0.0137
CMAC Integrated Intensity-ch2	0.0003	0.0008	0.0001	-0.0003
CMAC Pearson-ch1vs2	-0.0018	0.0061	-0.0105	-0.0028
CMAC Growth of Area	-0.1331	-0.1277	-0.0736	-0.174
CMAC Length Growth	-0.0254	0.0546	-0.0309	0.054
CMAC Delta Intensity-ch1	-0.0053	-0.0053	-0.0032	-0.0045
CMAC Delta Integrated Intensity-ch1	0	-0.0003	0	0
CMAC Delta Intensity-ch2	0.0003	0.0005	-0.0001	0.0005
CMAC Delta Integrated Intensity-ch2	0	0	0	0

Table 9: *Parameter estimates of linear regression models based on PCA data.*

	Data set			
Variables	pert0	pert1	pert2	pert3
intercept	0.6542	0.5540	0.5964	0.5704
CMAC Area	-0.1041	0	0	0
CMAC Major Axis	0.0288	0	0.0314	0
CMAC Minor Axis	-0.0699	-0.1568	-0.1396	-0.2222
CMAC Eccentricity	-0.0193	0	-0.0838	0
CMAC Angle	0	0	0	0
CMAC Perimeter	0	0	0	0
CMAC Convex Area	0.0413	0	0	0
CMAC Solidity	0	0	0	0
CMAC DistBorder	-0.0086	-0.0024	-0.0078	-0.0018
CMAC DistCenter	-0.0003	0	0.0017	0.0002
CMAC MeanIRaw-ch1	0.1041	0	0	-0.2402
CMAC Local BG-ch1	0.0015	0	0	0
CMAC MeanI-ch1	-0.1477	0	-0.0373	0.2963
CMAC StdevI-ch1	0	0	-0.0011	0.0026
CMAC MaxI-ch1	-0.0188	0	0	-0.1092
CMAC Integrated Intensity-ch1	0.001	-0.0028	0	0.0031
CMAC MeanIRaw-ch2	0.0742	0.0365	0	-0.0675
CMAC Local BG-ch2	0.0009	0.0005	0.0034	0.0037
CMAC MeanI-ch2	-0.0629	-0.037	-0.0192	0.0249
CMAC StdevI-ch2	-0.0006	0	0	-0.0011
CMAC MaxI-ch2	0	0	0	0.0119
CMAC Integrated Intensity-ch2	0.0003	0	0	0
CMAC Pearson-ch1vs2	0	0	0	0
CMAC Growth of Area	-0.1339	-0.0797	-0.0705	-0.1682
CMAC Length Growth	-0.0249	0	-0.0306	0.0425
CMAC Delta Intensity-ch1	-0.0053	-0.0055	-0.0032	-0.0045
CMAC Delta Integrated Intensity-ch1	0	-0.0002	0	0
CMAC Delta Intensity-ch2	0.0003	0.0007	0	0.0005
CMAC Delta Integrated Intensity-ch2	0	0	0	0

Table 10: *Parameter estimates of stepwise regression models.*

	Data set			
Variables	pert0	pert1	pert2	pert3
CMAC Area	0	0	0	0
CMAC Major Axis	0	0	0	0
CMAC Minor Axis	0	0	0	0
CMAC Eccentricity	0	0	0	0
CMAC Angle	0	0	0	0
CMAC Perimeter	0	0	0	0
CMAC Convex Area	0	0	0	0
CMAC Solidity	0	0	0	0
CMAC DistBorder	0	0	0	0
CMAC DistCenter	0	0	-0.0106	0
CMAC MeanIRaw-ch1	0	0	0	0
CMAC Local BG-ch1	0	0	0	0.0239
CMAC MeanI-ch1	0	0	0	0
CMAC StdevI-ch1	-0.0136	-0.0187	0	-0.0060
CMAC MaxI-ch1	0	0	0	0
CMAC Integrated Intensity-ch1	0	0	0	0
CMAC MeanIRaw-ch2	0	0	0	0.0025
CMAC Local BG-ch2	0	-0.0275	0	0
CMAC MeanI-ch2	0	0	0	0
CMAC StdevI-ch2	0	-0.0029	0	0
CMAC MaxI-ch2	0	0	0	0
CMAC Integrated Intensity-ch2	0	0	0	0
CMAC Pearson-ch1vs2	0	0	0	0
CMAC Growth of Area	0	0	0	0
CMAC Length Growth	0	0	0	0
CMAC Delta Intensity-ch1	0	0	0	0
CMAC Delta Integrated Intensity-ch1	0	0	0	0
CMAC Delta Intensity-ch2	0	0	0	0
CMAC Delta Integrated Intensity-ch2	0	0	0	0

Table 11: *Parameter estimates of the models constructed through employment of elastic net.*

	Data set			
Variables	pert0	pert1	pert2	pert3
intercept	0.4991	0.4983	0.4988	0.4994
CMAC Area	-0.0937	-0.0123	-0.0100	-0.0086
CMAC Major Axis	0.0155	0.0080	0.0176	0.0066
CMAC Minor Axis	-0.0054	-0.0108	-0.0183	-0.0219
CMAC Eccentricity	-0.0043	-0.0004	-0.0181	-0.0039
CMAC Angle	-0.0022	0.0026	-0.0027	-0.0003
CMAC Perimeter	0.0074	-0.0005	0.0070	-0.0127
CMAC Convex Area	0.0317	-0.0059	0.0021	0.0084
CMAC Solidity	-0.0069	0.0137	-0.0044	0.0112
CMAC DistBorder	-0.0281	-0.0015	-0.0188	-0.0057
CMAC DistCenter	-0.0023	0.0014	0.0080	0.0040
CMAC MeanIRaw-ch1	0.0497	0.1534	-0.0204	0.0070
CMAC Local BG-ch1	0.0045	-0.0260	0.0267	-0.0061
CMAC MeanI-ch1	-0.0678	-0.1372	0.0068	0.0046
CMAC StdevI-ch1	-0.0030	0.0120	-0.0095	0.0141
CMAC MaxI-ch1	-0.0143	-0.0248	-0.0088	-0.0333
CMAC Integrated Intensity-ch1	0.0339	-0.0146	0.0005	0.0161
CMAC MeanIRaw-ch2	0.0519	0.0156	-0.0645	-0.0452
CMAC Local BG-ch2	0.0223	0.0128	0.0631	0.0314
CMAC MeanI-ch2	-0.0715	-0.0380	0.0375	0.0244
CMAC StdevI-ch2	-0.0096	-0.0001	0.0074	-0.0082
CMAC MaxI-ch2	0.0062	-0.0016	-0.0116	0.0162
CMAC Integrated Intensity-ch2	0.0216	0.0096	0.0037	-0.0073
CMAC Pearson-ch1vs2	-0.0006	0.0001	-0.0026	0.0005
CMAC Growth of Area	-0.0626	-0.0241	-0.0312	-0.0329
CMAC Length Growth	-0.0057	0.0051	-0.0110	0.0095
CMAC Delta Intensity-ch1	-0.0360	-0.0270	-0.0210	-0.0239
CMAC Delta Integrated Intensity-ch1	-0.0102	-0.0274	-0.0205	-0.0022
CMAC Delta Intensity-ch2	0.0051	0.0091	-0.0019	0.0099
CMAC Delta Integrated Intensity-ch2	0.0198	0.0113	0.0119	-0.0030

Table 12: *Parameter estimates of stepwise regression models constructed with PCA transformed data sets.*

	Data set					
Variables	pert00	pert01	pert02	pert03	pert04	pert05
intercept	0.5485	0.716	0.8268	0.815	1.0303	0.7658
CMAC Area	-0.1295	-0.1319	-0.1142	-0.1724	-0.161	0.0082
CMAC Major Axis	0.1229	0.0294	-0.0388	0.0753	-0.8042	1.0903
CMAC Minor Axis	0.1309	-0.0716	-0.2417	0.0303	1.0945	0.6629
CMAC Eccentricity	-0.0151	-0.0246	-0.0583	-0.0121	0.4436	-0.6199
CMAC Angle	0	-0.0001	0	-0.0001	0	-0.0003
CMAC Perimeter	-0.0103	-0.0065	0.0124	-0.0014	0.2322	-0.2139
CMAC Convex Area	0.0383	0.0486	0.0655	0.1107	-0.2034	-0.123
CMAC Solidity	-0.0866	-0.0319	-0.0072	-0.1423	-0.5067	0.3556
CMAC DistBorder	-0.0036	-0.0078	-0.0123	-0.0138	-0.0017	-0.0489
CMAC DistCenter	0.0005	-0.0001	-0.0013	-0.0016	-0.013	-0.0305
CMAC MeanIRaw-ch1	-0.0043	0.0914	0.222	0.5216	2.5776	0
CMAC Local BG-ch1	0	0.0011	0.0037	0.0072	0.0115	0.0487
CMAC MeanI-ch1	-0.0171	-0.1393	-0.315	-0.6508	-3.1479	0.0715
CMAC StdevI-ch1	-0.0004	0	-0.0008	-0.0003	0.0019	-0.018
CMAC MaxI-ch1	0.0021	-0.0204	-0.0207	-0.0168	-0.017	0.534
CMAC Integrated Intensity-ch1	0.0004	0.0014	0.002	0.0017	0.004	-0.0096
CMAC MeanIRaw-ch2	0.0282	0.0505	0.0691	0.1419	0.6195	0
CMAC Local BG-ch2	0.001	0.0012	0.0011	-0.0001	-0.0049	0.0236
CMAC MeanI-ch2	-0.0349	-0.0733	-0.0579	-0.0869	-0.2948	-0.1461
CMAC StdevI-ch2	-0.0007	-0.0008	-0.0005	0.0007	-0.0093	0.0008
CMAC MaxI-ch2	0.0047	0.0142	0.0043	-0.0023	0.0431	0.0333
CMAC Integrated Intensity-ch2	0.0004	0.0007	0.0002	-0.0002	0.0004	0.0031
CMAC Pearson-ch1vs2	-0.0024	-0.0066	0.0107	-0.0195	0.0807	0.0193
CMAC Growth of Area	-0.1398	-0.1718	-0.1206	-0.1211	-0.024	0.304
CMAC Length Growth	0.0113	-0.0498	-0.0496	-0.0531	-0.1728	-0.0721
CMAC Delta Intensity-ch1	-0.0023	-0.0067	-0.0068	-0.0064	-0.0063	0.0081
CMAC Delta Integrated Intensity-ch1	0	0	0	0	0	-0.0002
CMAC Delta Intensity-ch2	0.0006	0	-0.0001	-0.0004	0.0009	-0.0049
CMAC Delta Integrated Intensity-ch2	0	0	0	0	0.0001	0.0001

Table 13: *Parameter estimates in linear regression models based on the split pert0 data set.*

	Data set					
Variables	pert00	pert01	pert02	pert03	pert04	pert05
intercept	0.5432	0.7161	0.4984	0.4989	0.5015	0.4991
CMAC Area	-0.1197	-0.1243	-0.1117	-0.1402	0	0
CMAC Major Axis	0.0771	0	0	0	-0.9016	0
CMAC Minor Axis	0.107	-0.136	-0.2047	-0.1655	0	0
CMAC Eccentricity	0	-0.0329	-0.0545	0	0	0
CMAC Angle	0	-0.0001	0	0	0	0
CMAC Perimeter	0	0	0	0	0.2709	0
CMAC Convex Area	0.036	0.0445	0.0673	0.1045	-0.2936	0
CMAC Solidity	-0.0825	0	0	0	0	0
CMAC DistBorder	-0.0036	-0.0072	-0.0123	-0.0142	0	0
CMAC DistCenter	0.0005	0	-0.0013	-0.0017	-0.0082	-0.0247
CMAC MeanIRaw-ch1	0	0.0805	0.2168	0.5161	2.4868	0
CMAC Local BG-ch1	0	0	0.0038	0.0071	0	0
CMAC MeanI-ch1	-0.0172	-0.1254	-0.311	-0.6651	-3.0142	-0.2708
CMAC StdevI-ch1	0	0	-0.0007	0	0	0
CMAC MaxI-ch1	0	-0.0211	-0.0199	0	0	0
CMAC Integrated Intensity-ch1	0	0.0013	0.002	0.0015	0.0054	0
CMAC MeanIRaw-ch2	0.0318	0.0509	0.0719	0.1408	0.5895	0
CMAC Local BG-ch2	0.0009	0.0012	0.001	0	-0.0044	0.0231
CMAC MeanI-ch2	-0.035	-0.0738	-0.0561	-0.0908	-0.2654	0
CMAC StdevI-ch2	-0.0006	-0.0008	-0.0004	0.0006	-0.0068	0
CMAC MaxI-ch2	0	0.014	0	0	0	0
CMAC Integrated Intensity-ch2	0.0005	0.0007	0.0002	0	0	0
CMAC Pearson-ch1vs2	0	0	0	-0.0265	0	0
CMAC Growth of Area	-0.1316	-0.1669	-0.12	-0.1403	0	0
CMAC Length Growth	0	-0.0485	-0.049	0	0	0
CMAC Delta Intensity-ch1	-0.0023	-0.0065	-0.0068	-0.0065	0	0
CMAC Delta Integrated Intensity-ch1	0	0	0	0	-0.0001	0
CMAC Delta Intensity-ch2	0.0006	0	0	0	0	0
CMAC Delta Integrated Intensity-ch2	0	0	0	0	0.0001	0

Table 14: *Parameter estimates in stepwise regression models based on the split pert0 data set.*

	Data set					
Variables	pert00	pert01	pert02	pert03	pert04	pert05
CMAC Area	0	0	0	0	0	-0.0146
CMAC Major Axis	0	0	0	0	-0.0325	0
CMAC Minor Axis	0	-0.0379	-0.0106	0	0	0
CMAC Eccentricity	0	0	0	0	0	0
CMAC Angle	0	0	0	0	0	0
CMAC Perimeter	0	0	0	0	0	0
CMAC Convex Area	0	0	0	0	0	0
CMAC Solidity	0	-0.0162	-0.0135	0	0	0
CMAC DistBorder	0	0	0	0	0	0
CMAC DistCenter	0	0	0	0	0	0
CMAC MeanIRaw-ch1	0	0	0	0	0	0
CMAC Local BG-ch1	0	0	0	0	0	0
CMAC MeanI-ch1	0	0	0	0	0	0
CMAC StdevI-ch1	0	0	-0.0113	0	-0.0387	0
CMAC MaxI-ch1	0	-0.0657	-0.1178	-0.0676	0	0
CMAC Integrated Intensity-ch1	0	0	0	0	0	0
CMAC MeanIRaw-ch2	0	0	0	0	0	0
CMAC Local BG-ch2	0.0355	0.0601	0.0679	0.0017	0	0.0424
CMAC MeanI-ch2	-0.0488	-0.0692	-0.0332	0	0	0
CMAC StdevI-ch2	0	0	0	0	0	0
CMAC MaxI-ch2	0	0	0	0	0	0
CMAC Integrated Intensity-ch2	0	0	0	0	0	0
CMAC Pearson-ch1vs2	0	0	0	0	0	0
CMAC Growth of Area	-0.1136	-0.2381	-0.1112	-0.0043	0	0
CMAC Length Growth	0	-0.0191	-0.0267	0	0	0
CMAC Delta Intensity-ch1	0	-0.1312	-0.1199	0	0	0
CMAC Delta Integrated Intensity-ch1	-0.0316	-0.0270	-0.1009	-0.0799	-0.0138	0
CMAC Delta Intensity-ch2	0.0476	0	0	0	0	0
CMAC Delta Integrated Intensity-ch2	0	0	0	0	0	0

Table 15: *Parameter estimates of the models gained through employment of elastic net on the subgroups of the pert0 data.*

6.1.3 Elastic net

pert00	pert01
CMAC Local BG-ch2	CMAC Minor Axis
CMAC MeanI-ch2	CMAC Solidity
CMAC Growth of Area	CMAC MaxI-ch1
CMAC Delta Integrated Intensity-ch1	CMAC Local BG-ch2
CMAC Delta Intensity-ch2	CMAC MeanI-ch2
	CMAC Growth of Area
	CMAC Length Growth
	CMAC Delta Intensity-ch1
	CMAC Delta Integrated Intensity-ch1

pert02	pert03
CMAC Minor Axis	CMAC MaxI-ch1
CMAC Solidity	CMAC Local BG-ch2
CMAC StdevI-ch1	CMAC Growth of Area
CMAC MaxI-ch1	CMAC Delta Integrated Intensity-ch1
CMAC Local BG-ch2	
CMAC MeanI-ch2	
CMAC Growth of Area	
CMAC Length Growth	
CMAC Delta Intensity-ch1	
CMAC Delta Integrated Intensity-ch1	

pert04	pert05
CMAC Major Axis	CMAC Area
CMAC StdevI-ch1	CMAC Local BG-ch2
CMAC Delta Integrated Intensity-ch1	

Table 16: *Contributing variables in the models created with elastic net in combination with ten-fold cross validation.*

pert0	pert1
CMAC StdevI-ch1	CMAC StdevI-ch1 CMAC Local BG-ch2 CMAC StdevI-ch2
pert2	pert3
CMAC DistCenter	CMAC StdevI-ch1 CMAC Local BG-ch2 CMAC StdevI-ch2

Table 17: *Contributing variables in the models created with elastic net in combination with ten-fold cross validation.*

6.1.4 MANOVA

Number of groups, i	*pert0*	*pert1*	*pert2*	*pert3*
1	0	0	0	0
2	0	0	0	0
3	0	0	0.2726	0
4	0	0.4279	0.9182	0
5	0.1300	0.7571	0.9644	0.0264
	5	4	3	6

Table 18: *P-values derived from tests of whether data consists of 1 or up to 5 age categories. If the ith p value is close to zero, then we doubt the hypothesis of the group means lying on a space of i-1 dimensions.*

6.1.5 EM-algorithm

	AIC			
Number of groups	*pert0*	*pert1*	*pert2*	*pert3*
2	3.875	0.587	44.125	-2.569
3	3.559	0.386	38.673	-1.930
4	3.275	0.764	32.830	-3.842
5	3.032	0.239	30.334	-2.467
	5	5	4	4

Table 19: *The AIC values achieved with different EM-models.*

6.2 Figures

6.2.1 Residual analysis

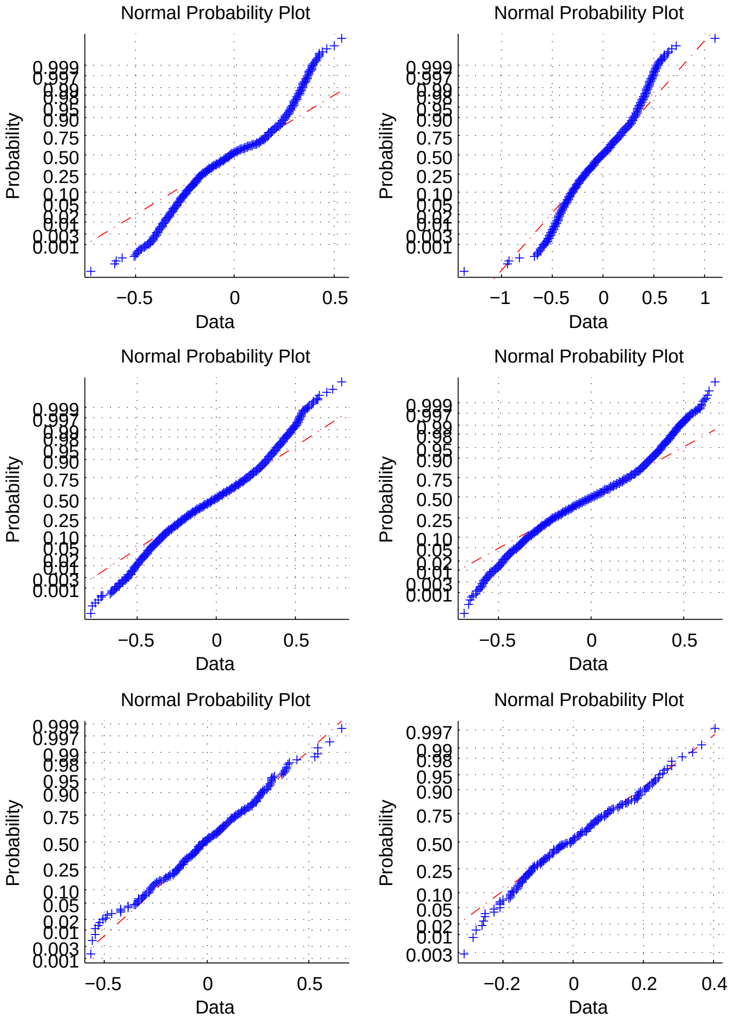

Figure 8: *Residual plots of regression models from subgroups of pert0.*

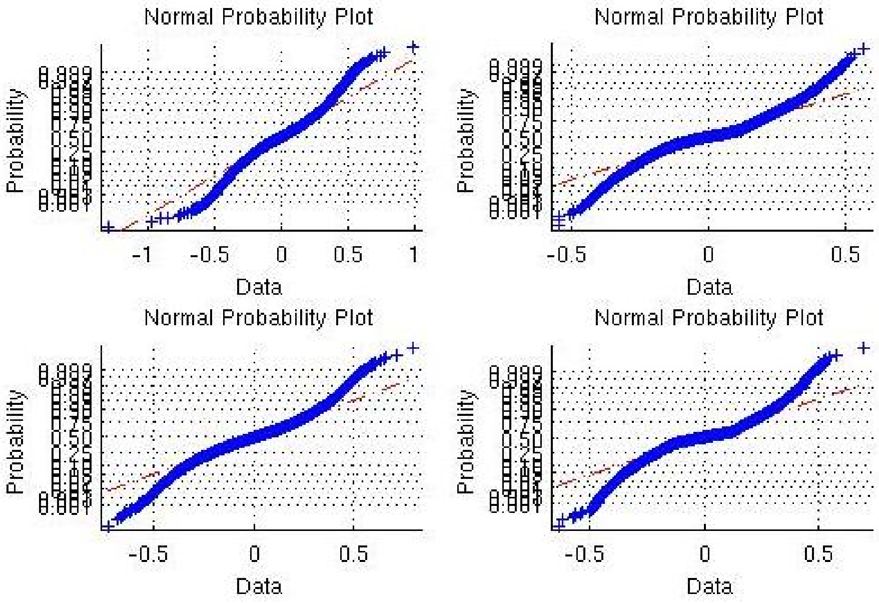

Figure 9: *Normal probability plots of the residuals from our four models of data sets where CMAC units observed only once have been excluded.*

6.2.2 Elastic net

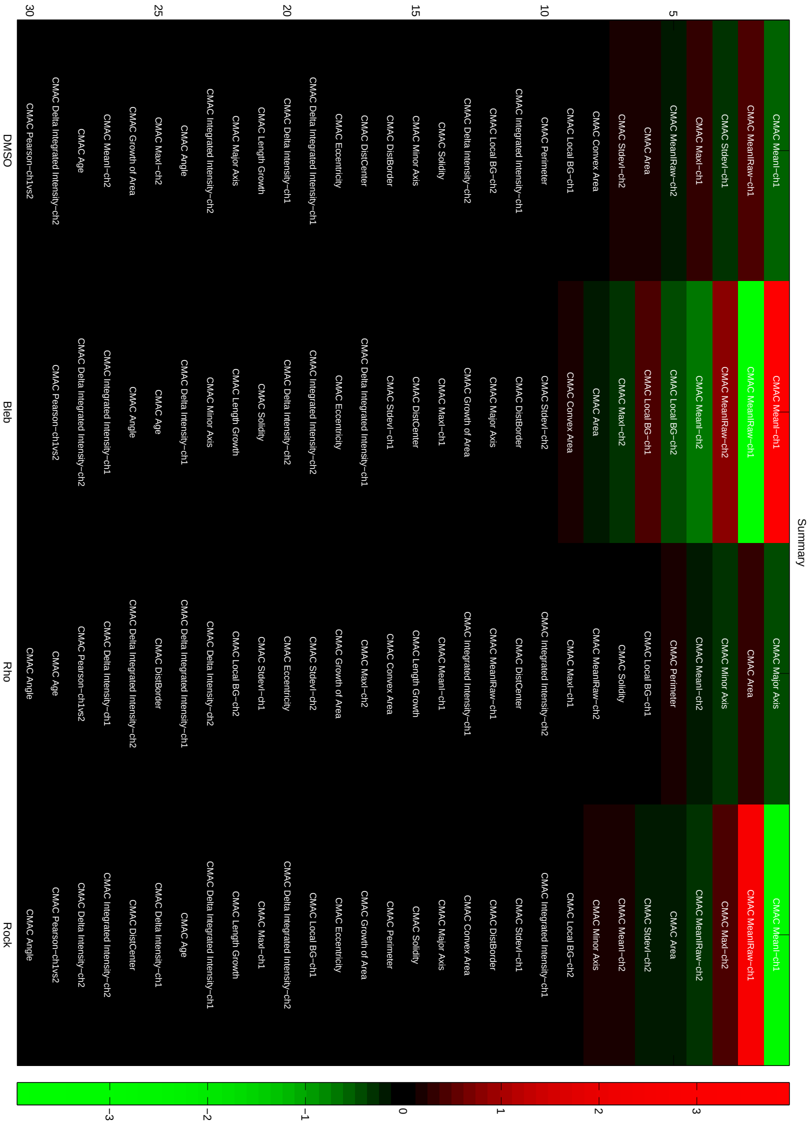

Figure 10: *Significant variables according to elastic net using R^2.*

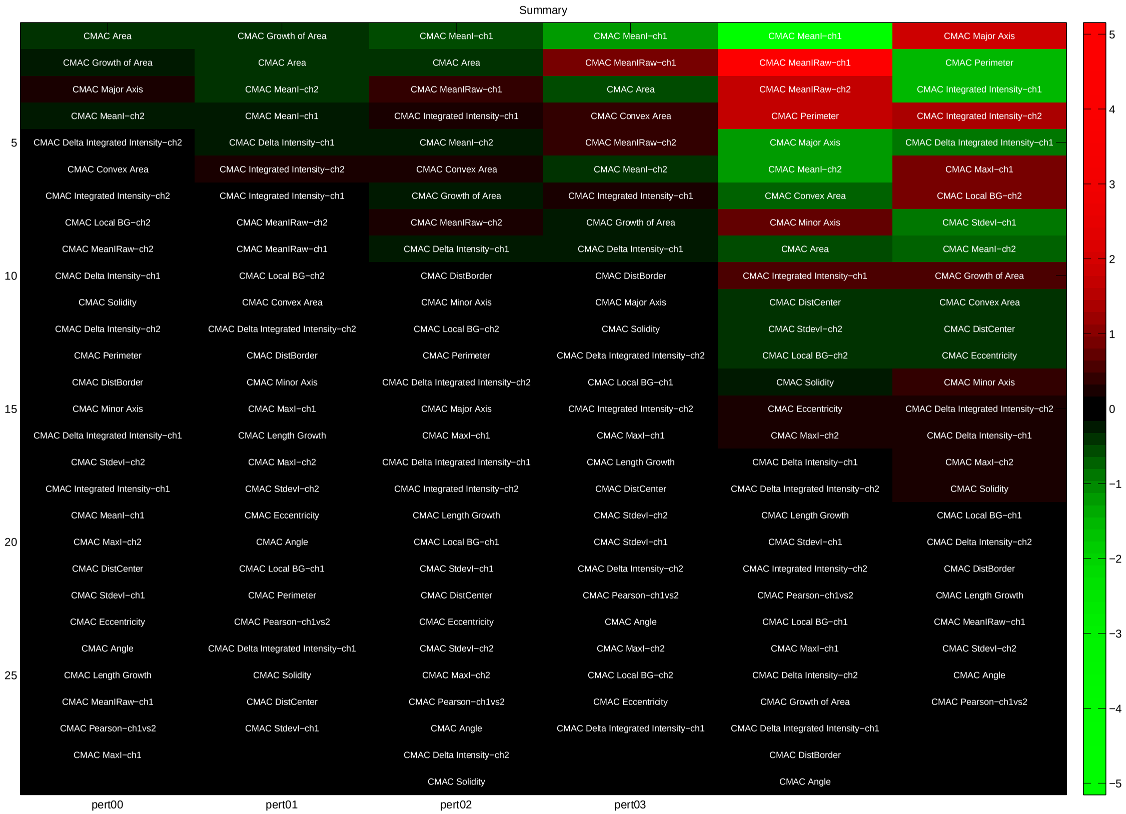

Figure 11: *Significant variables according to elastic net using R^2.*

6.3 Mathematical formulae

6.3.1 Linear regression

A classical linear regression consists of the following assumptions.

- The regression model is linear in the parameters.
- The disturbance terms, ϵ_i, are independent and $N(0, \sigma^2)$-distributed.
- The columns in the data matrix, $\mathbf{X}$, are linearly independent.

Among other methods to estimate the parameters in a linear regression model, we choose OLS. An ordinary least squares model is achieved by minimizing the sum, S, of squared residuals

$$S = \sum_{i=1}^{n} r_i^2,$$

where

$$r_i = y_i - f(x_i, \boldsymbol{\beta}),$$

are the residuals of each observations, and with $f(x, \boldsymbol{\beta})$ being the regression function.

6.3.2 Elastic net

For an α between 0 and 1, and $\lambda \geq 0$, elastic net solves the following problem,

$$\min_{\beta_0, \beta} \left(\frac{1}{2N} \sum_{i=1}^{N} \left(y_i - \beta_0 - x_i^T \beta\right)^2 + \lambda P_\alpha(\beta) \right),$$

where

$$P_\alpha(\beta) = \frac{1-\alpha}{2} \|\beta\|_2^2 + \alpha \|\beta\|_1 = \sum_{j=1}^{P} \left(\frac{1-\alpha}{2} \beta_j^2 + \alpha |\beta_j| \right),$$

and

- N is the number of observations.
- y_i is the response at observation i.
- x_i is data, a vector of p values at observation i.
- λ is a positive regularization parameter corresponding to one value of *Lambda*.
- The parameters β_0 and β are scalar and p-vector respectively.

6.3.3 One-way MANOVA

For a matrix, $\boldsymbol{X}$, of data, MANOVA tests make the following assumptions:

- Multivariate normality.
- Multivariate homogeneity.
- Observations are mutually independent.

		Group			
		1	2	$\dots$	p
	Sample size	n_1	n_2	$\dots$	n_p
Mean vector	Observed	$\hat{\mathbf{X}}_1$	$\hat{\mathbf{X}}_2$	$\dots$	$\hat{\mathbf{X}}_p$
	Expected	$\boldsymbol{\mu}$	$\boldsymbol{\mu}$	$\dots$	$\boldsymbol{\mu}$
Covariance matrix	Observed	$\hat{\mathbf{S}}_1$	$\hat{\mathbf{S}}_2$	$\dots$	$\hat{\mathbf{S}}_p$
	Expected	$\boldsymbol{\Sigma}$	$\boldsymbol{\Sigma}$	$\dots$	$\boldsymbol{\Sigma}$

Table 20: *Observed and expected statistics for the mean vectors and the covariance matrices of four groups in a one-way MANOVA, under the null hypothesis.*

In testings of the dimension of the group means, there is a statistic, Wilks' lambda, defined as

$$\text{Wilks' lambda: } \boldsymbol{\Lambda} = \frac{|\boldsymbol{W}|}{|\boldsymbol{T}|},$$

where

$$\mathbf{B} = SSCP_{between} = \mathbf{Q}_{group} - \mathbf{Q}_{total},$$
$$\mathbf{W} = SSCP_{within} = \mathbf{Q}_{data} - \mathbf{Q}_{group},$$
$$\mathbf{T} = SSCP_{total} = \mathbf{Q}_{data} - \mathbf{Q}_{total},$$

are the matrices of sums of squares in the diagonal entries and cross products off the diagonals, and

$$\mathbf{Q}_{data} = \mathbf{X}'\mathbf{X},$$
$$\mathbf{Q}_{group} = \mathbf{T}_j'\mathbf{M}_j,$$
$$\mathbf{Q}_{total} = \mathbf{T}\mathbf{M}',$$

where $\mathbf{X}$ is the matrix of raw data, $\mathbf{T}_j$ represents sums of the individual groups, $\mathbf{T}$ grand sums, $\mathbf{M}_j$ means per group, and $\mathbf{M}$ grand means.

For Wilks' lambda, smaller is better, so to be significant, our obtained lambda must be smaller than the tabled value. Regarding degrees of freedom, we have k groups, so df_b=$k-1$, and there are n observations, so $df_w = n - k$.[4] Also note that $R^2 = 1 - \Lambda$.

There are four useful statistics:

- Wilks' lambda,
- The Pillai trace,
- Lawley-Hotellings treace,
- Roy's largest root,

however in our analyses we have used only the Wilks' lambda, since it is considered to be the most stable one.

6.3.4 Akaike information criterion

The Akaike information criterion is defined as follows,

$$AIC = 2k - 2ln(L),$$

where k is the number of parameters, and L is the maximum of the likelihood function of the model. Over a number of different models with different AIC-values, we choose the one with the smallest AIC-value.

7 References

[1] Hervé Abdi and Lynne J. Williams. Principal component analysis. *WIREs Comp Stat*, 2(4):433–459, July 2010.

[2] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974.

[3] Jeff A. Bilmes. A Gentle Tutorial on the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models. Technical report, 1997.

[4] Bruce Brown. *Multivariate analysis for the biobehavioral and social sciences.* Oxford : Wiley-Blackwell, 2012.

[5] Richard A. Johnson and Dean W. Wichern. *Applied multivariate statistical analysis.* Pearson Prentice Hall, 2007.

[6] John G. Lock and Staffan Strömblad. Systems microscopy: an emerging strategy for the life sciences. *Experimental cell research*, 316(8):1438–1444, May 2010.

[7] John G. Lock, Bernard Wehrle-Hallerb, and Staffan Strömblad. Cell-matrix adhesion complexes: master control machinery of cell migration. 2007.

[8] Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, April 2005.