Project Deliverable

Project acronym: **SOUND**  
GA number: **633974**

**Project title:** Statistical Multi-Omics Understanding of Patient Data  
Funding Scheme: Collaborative Project (H2020-PHC-2014-2015/H2020-PHC-2014-two-stage)  
Health, novel medical developments

**Project start date:** 01 September 2015  
**Duration:** 36 months  
**Project’s coordinator:** Dr Wolfgang Huber (European Molecular Biology Laboratory, Heidelberg)

**D4.2 Technical report on approaches to outlier detection in patient 'omics data**

Due date of deliverable: Month 18 - 28.02.2017  
Actual submission date: 23.02.2017

Organization name of lead contractor for this deliverable: Instituto de Engenharia Mecânica (IDMEC)  
Organization name of other involved partners: EMBL, ETH.  
Personnel involved: Susana Vinga, Eunice Carrasquinha, André Veríssimo, Marta Lopes and Wolfgang Huber.

**Project co-funded by the European Commission within the H2020 Program (2015-2018)**

<table>
<thead>
<tr>
<th>Dissemination Level</th>
<th>PU</th>
<th>PP</th>
<th>RE</th>
<th>CO</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Public</td>
<td>Restricted to other program participants (including the Commission Services)</td>
<td>Restricted to a group specified by the consortium (including the Commission Services)</td>
<td>Confidential, only for members of the consortium (including the Commission Services)</td>
</tr>
</tbody>
</table>

x
D4.2 - Technical report on approaches to outlier
detection in patient ‘omics data

February 2017

Contents

1 Introduction 2

2 Model-based outlier detection 3
   2.1 Robust statistics and regression 3
       2.1.1 Survival analysis 3
       2.1.2 Supervised classification and logistic regression 5
   2.2 Model selection and dimensionality reduction 7
   2.3 Outlier diagnostics and influential observations 8
       2.3.1 Residuals 8
       2.3.2 Measuring influential observations 11
       2.3.3 The concordance c-index 12
       2.3.4 Quantile regression 13
   2.4 RANdom SAmple Consensus (RANSAC) 16
   2.5 Meta-analysis and the rank product test 16

3 Results and Discussion 18
   3.1 Myeloma 19
   3.2 Ovarian cancer 20
   3.3 Breast invasive carcinoma (BRCA) 23
   3.4 Chronic lymphocytic leukemia (CLL) 30

4 Conclusions and Future Work 34

5 Annex 35
   5.1 Prototype R packages 35
   5.2 R Markdown files 36

*SOUND project - European Union’s Horizon 2020 research and innovation program
grant agreement No. 633974.
1 Introduction

The main goal of Workpackage 4 – *Cross-cutting Methods for Oncology and Genetics*, is to address overarching challenges posed by multi-omics patient data through the development of novel mathematical and statistical methods for data analysis, which will ultimately support diagnosis, improve prognosis and guide treatment decisions.

In particular, Deliverable 4.2 – *Technical report on best practices for outlier detection in patient’omics data*, concerns one of the challenges arising from the analysis of high-dimensional and heterogeneous data, which is the detection of outliers.

This deliverable is strongly associated with Task 4.2 - *Improved automated outlier detection*, where the major methodologies to identify unusual observations are reviewed and expanded. In this context, robust regression seems a promising approach to identify outliers in a model-based context. In particular model-based outlier detection methods provide a structured framework to separate abnormal cases, i.e., those who significantly deviate from what would be expected. The definition of outlier becomes, therefore, highly coupled with the model statistical learning process, with the obvious interpretability advantage: an outlier is a case that deviates from what would be expected given the corresponding covariates. This aspect also brings a new challenge since model selection has a very high impact on the obtained results that is addressed through the meta-analysis of methods and models.

This report will overview several outlier detection methodologies and will be focused on two major modeling approaches, adequate for different kinds of data: 1) Cox proportional hazards models for survival data; and 2) Logistic regression for binary response variables. These are expected to be crucial under the scope of the project SOUND, in particular for WP1-3 as well as WP5-8, and therefore the theoretical description and computational examples will be focused on these two categories.

The theoretical description of the methods will be complemented by its application to several real clinical case-studies. These include a well-studied clinical dataset on myeloma, the analysis of breast and ovarian cancers datasets obtained from The Cancer Genome Atlas (TCGA) and, finally, the analysis of the Chronic lymphocytic leukemia (CLL) dataset.

All the implementations described are available as R Markdown documents, which allows full reproducibility and adaptation to new datasets. Several prototype R packages were developed and will be fully integrated in the SOUND pipelines during the project.
2 Model-based outlier detection

2.1 Robust statistics and regression

Patient’omics data usually contain abnormal variable measurements arriving from many sources, from experimental errors to clinical outliers, i.e., unusual observations regarding molecular phenotypes, mutations or disease progression patterns. Failing to identify these potential cases may negatively impact the correct understanding of the biology of the disease and prevent the characterization of features such as novel interesting biomarkers.

Robust statistic’s goal is to produce consistent and efficient estimators even if the underlined model is misspecified or approximate [1]. An alternative and complementary approach is to construct outlier diagnostics for the regression, which can pinpoint these observations in order to analyzed them more in detail (and decide for their correction or deletion) [2]. The ultimate aim is to have stable learning methods that can cope with changes in the data or model deviations, such as the presence of outliers, leverage points and influential observations.

This report is focused on parametric methods for outlier detection and robust regression, which are based on learning statistical models from the data. However, standard non-parametric approaches should always complement this analysis in practical cases. These model-free strategies include, for example, statistical methods based on the Mahalanobis distance (and the corresponding version with mean and covariance matrix estimated in a robust way), on clustering (groups with one or few elements can be considered outliers), and on the distance distributions to the k-nearest neighbors. This preliminary strategy can support the immediate correction of abnormal cases, since it can identify leverage points and potential influential observations.

2.1.1 Survival analysis

Survival analysis is a statistical technique widely used in many fields of science, in particular in the medical area, and which studies the time until an event of interest occurs. The event may be death, the relapse of a tumour, or the development of a disease. The response variable is the time until that event, called survival or event time, which can be censored, i.e. not observed on all individuals present in the study.

There are different ways of modelling this type of data, one of most widely used due to its flexibility is the Cox proportional hazards regression model [3]. This is a semi-parametric model because the baseline hazard does not need to be specified. One of the problems of this technique is the fact that a single abnormal observation can affect the parameter estimates. To overcome this issue it is important to study mechanisms to identify individuals who lived too long or too short, given their covariates.
The detection of outliers in survival data has gained great importance due to the fact that the identification of individuals with survival time too high or too short can lead in the medical field to the detection of new prognostic factors \[4\]. The first attempt to analyse and identify outliers was based on residuals. In this context, graphical methods based on the analysis of martingale, score and deviance residuals were proposed \[5\]. Nardi and colleagues \[4\] proposed two new types of residuals: the log-odds and normal deviate residuals.

More recently, three outlier detection algorithms for censored data were presented \[6\]: the residual-based, boxplot, and scoring algorithms, all based on quantile regression, which is robust to outliers \[1\]. Alternative methods for outlier detection based on the concordance c-index were also proposed \[7\]: the one-step deletion (OSD) and bootstrap hypothesis testing (BHT) strategy, subsequently improved with a dual bootstrap hypothesis testing (DBHT) version \[8\].

Cox regression model

The Cox regression model is one of the most used methods in Survival analysis \[3\]. It is based on a semi-parametric likelihood, which is able to deal with censored data and assumes that the hazard function \(h(t)\) at time \(t\) is:

\[
h(t; x) = h_0(t) \exp(x^T \beta),
\]

where \(\beta = (\beta_1, \ldots, \beta_p)\) are the unknown regression coefficients, which represent the covariate effect in the survival, \(h_0(t)\) represents the baseline hazards and \(x = (x_1, \ldots, x_p)\) is the covariate vector associated to an individual.

The Cox regression model is called a semi-parametric regression model, because the baseline hazard function, \(h_0(t)\) is not specified. This contributes for the flexibility of the model. The unknown regression coefficients, \(\beta\) are determined by maximizing the partial likelihood function

\[
L(\beta) = \prod_{i=1}^{n} \left[ \frac{\exp(x_i^T \beta)}{\sum_{j \geq i} \exp(x_j^T \beta)} \right]^{\delta_i},
\]

where \(\delta_i\) is the censored indicator.

Although the Cox regression model is a widely used method due to its simplicity, the corresponding estimator has a breakdown point of \(1/n\) \[9\], which means that the presence of outlying observations may have extreme influence on the estimation of the model parameters. In order to handle this problem, a robust version of the Cox regression model has been proposed \[10\] and will briefly be presented.
Cox Robust regression model

The robust version of the Cox regression model \[10\] is based on doubly weighting the partial likelihood function of the Cox regression model.

The solution for the unknown coefficients $\beta$, based on the partial likelihood function Eq.(2), can be obtained by the solution of the first order equation

$$\sum_{i=1}^{n} \delta_i \left[ x_i - \frac{\sum_{j \geq i} \exp(x_j^T \beta) x_j}{\sum_{j \geq i} \exp(x_j^T \beta)} \right] = 0.$$ \hspace{1cm} (3)

The robust alternative for the Cox regression model is obtained by introducing weights in Eq.(3).

Let $w(t, x)$ be a weight function, were $w_{ij} = w(t_i, x_j)$ and $w_i = w_{ii} = w(t_i, x_i)$ are the weights for all $1 \leq i \leq j \leq n$. The solution of the unknown parameters $\beta$ for the robust case of the partial likelihood function, presented by \[10\] and \[11\], is given by

$$\sum_{i=1}^{n} w_i \delta_i \left[ x_i - \frac{\sum_{j \geq i} w_{ij} \exp(x_j^T \beta) x_j}{\sum_{j \geq i} w_{ij} \exp(x_j^T \beta)} \right] = 0,$$ \hspace{1cm} (4)

where $w(t, x)$ appears in the main sum, down-weighting the uncensored observations, and in the inner sums. This allows that outlying observations will have a lower weight in the likelihood function, thus also down-weighting their influence on the parameter estimations. In this way, the most outlying observations will contribute less to the inference of the $\beta$. The weight function, $w(t, x)$, can be linear, exponential or quadratic. The choice of $w(t, x)$, is presented in \[11\] along with more details regarding the robust version of the Cox regression model.

The robust Cox is presented here as an alternative method to the Cox regression model estimation, as a framework that allows to infer the parameters in a more robust way when outlying observations are present, i.e. individuals that lived to long or died too early when compared to others with the same clinical conditions. Furthermore, the weights obtained with this method can give information about which observations are more influential and therefore can be considered as putative outliers.

### 2.1.2 Supervised classification and logistic regression

Frequently, the outcome variable to be estimated based on patient’omics data is categorical, as is the case of a ‘cancer’ versus ‘normal’ condition, or different subtypes of cancer. An accurate classification of individuals allows going through the patterns of the independent variables in each class, a key aspect when the goal is to find biomarkers for a particular disease condition.
Logistic regression model

Logistic regression is a popular classification method describing the relationship between one or more independent variables when the dependent variable is categorical. Logistic regression belongs to a family of regression models called Generalized Linear Models (GLM) [12], which extends traditional linear models and allows the mean of a population to depend on a linear predictor through a nonlinear link function. In GLM the response probability distribution can be any member of an exponential family of distributions. Binary logistic regression refers to logistic regression when the dependent outcome variable $Y$ is binary, assuming only two possible values, zero and one. In the patient’omics context, a binary clinical outcome can be predicted by the logistic function, where $p$ is the probability of $Y = 1$:

$$p = \frac{\exp(x^T\beta)}{1 + \exp(x^T\beta)}$$

which is equivalent to fitting a linear model in which the dependent variable is replaced by the logarithm of the odds ratio (defined as the ratio of the probability of success, and the probability of failure), through the \textit{logit} transformation given by

$$\log\left(\frac{p}{1-p}\right) = x^T\beta.$$  

It is therefore assumed that the \textit{logit} transformation of the outcome variable ($Y$) has a linear relationship with the predictor variables ($x$). The parameters of the logistic model are estimated by maximizing the log-likelihood function of the logistic model

$$l(\beta) = \sum_{i=1}^{n} \{y_i x_i^T \beta - \log(1 + \exp(x_i^T \beta))\}.$$  

Robust logistic regression model

Graphical and diagnostic tools before model fitting, as well as a residual analysis after model fitting, are common practices for detecting outliers, culminating in their removal from the dataset. However, the strategy of removing observations, although apparently simple, can be not only unpractical, but also very misleading [1], depending on the complexity of the dataset (in which concerns dimensionality), the appropriateness of the measures of outlyingness chosen and the masking effects produced. In many cases it could be more efficient to weight observations differently depending on how well behaved these observations are which constitutes the rational for robust regression methods.

In the binary response setting, deviations in the response space take the form of misclassification (a zero instead of a one, or vice-versa), and
the difference between an outlier and a leverage point is less clearcut. To address the potential problem of deviating points, a general class of robust M-estimators can be used, as proposed for the GLM class [1].

2.2 Model selection and dimensionality reduction

Outlier detection methods are highly dependent on the statistical models learnt. The inherent high-dimensionality and multi-collinearity of patient’omics data, with variables very often outnumbering the cases enrolled, constitutes a challenge to identify an interpretable model since it usually lead to ill-posed inverse problems. Regularized optimization is a promising strategy and is being addressed in SOUND’s Task 4.1 – Sparse models for high-dimensional multi-omics patient data.

At this stage, and in order to analyze the proposed outlier detection pipelines, we will apply classical sparse and dimensionality reduction techniques. These will be further developed to incorporate group-based regularization and network information, and integrated into the final deliverables.

Variable selection and feature extraction

Variable selection can support the identification of biomarkers associated to a disease or its subcategories. For high-dimensional datasets, several regularization methods have been proposed, namely, the least absolute shrinkage and selection operator (Lasso) [13], which uses an $l_1$ regularizer, and the elastic net [14] that uses a linear combination of $l_1$ and $l_2$ penalties. While in the presence of highly correlated variables the Lasso tends to arbitrarily select one of them, the elastic net, due to the $l_2$ Ridge term, encourages smaller $\beta$. More recently, and when the features can be represented through graphs, network-based regularizers have also emerged as a strategy to include a priori information in the optimization process. These include, for example, NetCox, which promotes smoothness of the features coefficients across the network [15] and DEGREECox, which accounts for the node degrees [16]. Model and features selection are being comprehensively explored in Task 4.1.

The problem of multi-collinearity can also be approached by feature extraction methods like Partial Least Squares (PLS) regression [17, 18]. In PLS regression an orthogonal basis of latent (not directly observed or measured) variables (LV) is constructed in such a way that they are maximally correlated with the response variable. Formally, PLS expresses $x$ ($n \times p$) and $y$ ($n \times m$) according to

$$
\begin{align*}
    x &= tp^T + e \\
    y &= uq^T + f
\end{align*}
$$

(8)
where \( t \) and \( u \) are the \((n \times l)\) matrices of the \( l \) extracted score (latent) vectors \((l \ll p)\), whereas \( p \) \((p \times l)\) and \( q \) \((m \times l)\) are the matrices of orthogonal loadings, and \( e \) \((n \times p)\) and \( f \) \((n \times m)\) are matrices of residuals. Given \( t \) and \( u \), the PLS estimate of the regression coefficients vector \( \beta = (\beta_1, ..., \beta_p) \) is

\[
\hat{\beta} = x^T u (t^T xx^T u)^{-1} t^T y.
\]  

The basic assumptions of PLS regression is that the relationship between \( x \) and \( y \) is linear and that this linearity assumption still holds for the relationship between the latent variables. The projection of the observed data onto a subspace of orthogonal LVs, typically of small number, has been shown to be a powerful technique when the observed variables are highly correlated, noisy, and the ratio between the number of observations and variables is small, which justifies its use for the analysis of genomic data [19]. PLS can also be applied to classification problems through Partial Least Squares Discriminant Analysis (PLS-DA), when the response variable is categorical and expresses a class membership. PLS regression has also been combined to variable selection for classification problems [20].

2.3 Outlier diagnostics and influential observations

An outlier can be defined as an observation with a large residual, whose dependent variable value is unusual given its value on the predictor variables; an outlier may indicate a sample peculiarity or a data entry error. A leverage observation, on the other hand, is an observation with extreme distance values on the predictor variables. An influential observation severely affects the parameters estimates, i.e., the regression coefficients change when they are removed from the dataset.

Three types of outlier detection algorithms in survival analysis and logistic regression are described next. They are based on the residuals, the concordance c-index and, finally, on quantile regression.

2.3.1 Residuals

Outlier diagnostics can be firstly approached via inspection of the residuals. An appropriate definition of residual is fundamental to evaluate a regression model. For censored data, several types of residuals have been proposed, such as Cox-Snell, Schoenfeld, Martingale and Deviance, adequate to analyse Cox proportional hazards models. Others such as Pearson and Deviance, along with their standardized versions, commonly targeted for the analysis under the logistic regression framework.

- Cox-Snell residuals
Cox-Snell residuals were the first to be proposed for the proportional hazard regression model [21]. If the model is well adjusted, then the residuals should follow a known distribution.

For a given individual, $\mathbf{x} = (x_1, ..., x_p)$ represents the covariate vector and $\mathbf{\beta} = (\beta_1, ..., \beta_p)$ the regression coefficients. Let $T$ be a continuous r.v. with distribution function $F$, which follows an Uniform distribution $(0, 1)$. The survival function, $S = 1 - F$, also follows the same distribution, $S \sim U(0, 1)$. Since $H(T) = - \log S(T)$, $H(T)$ follows an exponential distribution with parameter $\lambda = 1$, where $H(T)$ represents the cumulative hazard function.

From the Cox regression model,

$$H(t; \mathbf{x}) = \int_0^t h_0(u) \exp(\mathbf{\beta}^T \mathbf{x}) du = \exp(\mathbf{\beta}^T \mathbf{x}) H_0(t),$$

where $H_0(t)$ corresponds to the cumulative baseline function. The residual for the $i^{th}$ individual, is defined as

$$r_i = \hat{H}(t_i) = \exp(\mathbf{x}_i^T \hat{\mathbf{\beta}}) \hat{H}_0(t_i), \quad i = 1, ..., n, \quad (11)$$

where $\hat{\mathbf{\beta}}$ and $\hat{H}_0(t_i)$ are the estimates obtained by the partial maximum likelihood of the Cox regression model. The residuals will follow an exponential distribution with parameter $\lambda = 1$, if the estimate values of $\hat{H}(t_i)$ are similar to the real values of $H(t_i)$. That is, the residuals $r_i$ should behaviour as an exponential distribution with mean value 1. Note that expression Eq. (11) does not take into account censored data. The modified Cox-Snell residuals that consider censoring are given by

$$r_i^* = \begin{cases} r_i, & \text{if } t_i \text{ is not a censored observation} \\ r_i + 1, & \text{if } t_i \text{ is a censored observation}. \end{cases} \quad (12)$$

The Cox-Snell residuals are positive, and if they are higher than 1 the $t_i$ observation is censored, and if the graphical representation is approximately a straight line with gradient 1 and $y$-intercept null, the model is appropriate. For more details, see [21].

• Schoenfeld residuals

Schoenfeld residuals are very useful on the evaluation of the proportional hazards assumption, after adjustment of a Cox model to a given dataset [22].

Consider that $n$ is the number of individuals and that $k$ different survival times $t_{(1)} < ... < t_{(k)}$ were observed, for $k < n$. The number at risk in $t_{(i)}$ is given by $R_i = \{ j : t_j \geq t_{(i)} \}$, and represents the individuals still in the cohort immediately before $t_{(i)}$. The Schoenfeld residual for the $i^{th}$ individual with covariate vector $\mathbf{x}_j$, $j = 1, ..., p$, is given by
\[ r_{ij} = x_{ij} - \hat{E}[X_{ij}|R_i], \]  
where

\[ \hat{E}[X_{ij}|R_i] = \frac{\sum_{k \in R_i} x_{kj} \exp(\hat{\beta}^T x_k)}{\sum_{k \in R_i} \exp(\hat{\beta}^T x_k)}. \]

If the survival time is not observed, for a given individual, \( r_{ij} \) will be zero. In order to distinguish the true zero residuals from the ones corresponding to censored observations, the later are usually considered as missing values.

For a given individual whose death was observed in \( t_i \), the residual is the difference between the value of the covariate \( x_j \) and the weight mean of the values of that covariate for all individuals in risk at \( t_i \). The weight associated to an individual \( l \in R_i \) is \( \exp(\hat{\beta}^T x_l) \).

The partial likelihood function, the \( i \)th portion of the sum Eq.(15), determined in \( \hat{\beta} \), is therefore the Schoenfeld residual for covariate \( x_j \) for the \( i \)th individual. Notice that \( \hat{\beta}_j \) verifies:

\[ \frac{\partial \log L}{\partial \beta_j} \Bigg|_{\hat{\beta}} = 0. \]

Graphically, the Schoenfeld residuals should be a random cloud of points around zero when the model was well adjust to the data. Therefore, these residuals are useful in the evaluation of the proportional hazard hypothesis, after the data are adjust to the Cox’s regression model. Other version of these residuals can be seen in [23].

- Martingale residuals

The Martingale residuals arise from a linear transform of the Cox-Snell residuals [3] and are also very useful for outlier detection. Let all the covariates be fixed, the martingale residual for the \( i \)th individual is given by

\[ r_{Mi} = \delta_i - H_0(t_i) \exp(\hat{\beta}^T x_i). \]

These residuals are asymmetric and take values in \((-\infty, 1)\).

The martingale residuals are the difference between the observed number of the events for the \( i \)th individual in \((0, t_i)\) and the corresponding expected number, obtained by the adjusted model. The observed number of “deaths” is one if \( t_i \) is not censored, i.e., is equal to \( \delta_i \). On the other hand, \( r_i \) is the estimate of \( H(t_i) \), which can be interpret as the expected number of “deaths” in \((0, t_i)\), since it is only considered an individual.
The martingale residual will reveal the individuals that are not well adjusted to the model, i.e., those that lived too long (large negative values) or died too soon (values near one), when compared to other individuals with the same covariate pattern.

- Deviance residuals

The deviance residuals were introduced in [5], in order to overcome the fact that the martingale residuals are asymmetric distributed. The deviance residuals are defined as

\[ r_{Di} = \text{sign}(\hat{M}_i)\sqrt{-2\left(\hat{M}_i + \delta_i \log(\delta_i - \hat{M}_i)\right)}. \]  

(17)

where \( \hat{M}_i \) is the martingale residual for the \( i \)th individual and \( \text{sign}(.) \) is the sign function. Notice that the deviance residuals are components of the deviance statistics, \( D \), given by

\[ D_i = -2(\ln \hat{L}_a - \ln \hat{L}_s), \]  

(18)

where \( \hat{L}_a \) is partial likelihood of the adjusted model and \( \hat{L}_s \) is partial likelihood of the model containing all the covariates. Lower values of \( D \) correspond to better adjusted models. In this way, \( D = \sum_{i=1}^{n} r_{Di}^2 \), i.e., observations with high values of residuals in absolute value, are observations that are not well adjust by the model and potential outliers. The adjusted model (where only the significant covariates are included) is obtained from a variable selection method applied to a complete, saturated model.

In the context of logistic regression, deviance residuals are the signed square roots of the individual observations of the overall deviance, given by

\[ r_{di} = \text{sign}(y_i - \hat{y}_i)\sqrt{2y_i \log\left(\frac{y_i}{\hat{y}_i}\right) + 2(n_i - y_i) \log\left(\frac{n_i - y_i}{n_i - \hat{y}_i}\right)}. \]  

(19)

- Pearson residuals

The Pearson residuals are commonly used for regression diagnostics and inspection of outliers in binary data, defined as

\[ r_{Pi} = \frac{y_i - n_i\hat{p}_i}{\sqrt{n_i\hat{p}_i(1 - \hat{p}_i)}}. \]  

(20)

2.3.2 Measuring influential observations

The Cook’s distance [24], \( D_{\text{Cook}} \), is a widely used estimate of the influence of a data point when performing logistic regression, combining the information of leverage and residual. For each observation \( i \), \( D_{\text{Cook}_i} \) measures the

\[ D_{\text{Cook}_i} = \sum_{j=1}^{n} \frac{r_{Di}^2}{MSE} \]  

where \( MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n - p} \) is the mean squared error.

11
change in $\hat{y}$ for all observations with $(\hat{y}_j)$ and without $(\hat{y}_j(i))$ the presence of observation $i$, so that we know how much the observation $i$ impacted the fitted values:

$$D_{\text{Cook}, i} = \frac{\sum_{j=1}^{n}(\hat{y}_j - \hat{y}_j(i))^2}{p \times MSE},$$

with MSE the mean squared error. Other diagnostic measures including DFFITS and DFBETAS can also be used for evaluating the influence on the fit when an observation is deleted \cite{25}.

### 2.3.3 The concordance c-index

In survival analysis, the concordance c-index \cite{26} denotes the probability that a randomly selected subject who experienced the outcome will have a higher predictive probability of having the outcome occur compared to a randomly selected subject who did not experience the event. Here the concordance c-index is used as a test statistics that is sensitive to the presence of outliers, i.e., the larger the number of outliers, the lower the performance of the model adjusted. Thus, the concordance c-index measures how well predicted values are concordant with the rank ordered response variables. Three alternative methods for outlier detection in survival data based on the c-index were proposed \cite{7, 8}, based on the rational that an outlier is “an observation that when absent from the data, will likely decrease the prediction error of the fitted model”. These algorithms are implemented in the prototype R package BSCOD – see Annex.

- **One-step deletion (OSD)**

  The One-Step Deletion (OSD) is an algorithm that tries to identify which observation, if removed, leads to the best improvement of the concordance c-index of the obtained model.

  In the first step of the procedure, each observation of the dataset is removed temporarily, and the difference between the original c-index and the one obtained is calculated. In the following step, the most outlying case is definitively removed, and the procedure is repeated again for the remaining data.

  The process ends when the quantity of erased observations reaches a pre-defined proportion of expected outliers. In the end of the algorithm the ranked list of the observations removed constitutes the putative outliers.

- **Bootstrap hypothesis test (BHT)**

  The main idea behind the bootstrap hypothesis test (BHT) for outlier detection is to perform $B$ hypothesis tests for the concordance variation on bootstrap samples *without* the target observation $i$.  

Let $C_{-i}$ be the c-index of the fitted model of the data without the $i^{th}$ observation and $C_0$ the c-index corresponding to the full dataset model. The hypothesis test for a certain observation $i$ is:

$$H_0 : \Delta C_i \leq 0 \quad vs. \quad H_1 : \Delta C_i > 0$$

where $\Delta C_i = C_{-i} - C_0$. The null hypothesis states that there are no improvements on the concordance when removing observation $i$. The algorithm starts by computing $C_0$. For each observation $i$ under test, the c-index from the model fitted without $i$, $C_{-i}$, is calculated and also $\Delta C_i$. Then $B$ bootstrap samples are generated from the data without observation $i$. The p-value is determine based on the proportion of samples having $\Delta C_i \leq 0$. The lower the p-value, the more outlying the observation is considered to be.

- Dual bootstraps hypothesis testing (DBHT)

The dual bootstrap hypothesis testing (DBHT) is an improvement of the BHT described before. The BHT removes one observation from the dataset, and then evaluates the impact of each removal on the concordance c-index. Notice that the model has less observations than the original dataset, and therefore the concordance c-index has the tendency to increase, which may hamper the evaluation of the statistical significance and may increase the number of false positives.

The DBHT method generates two histograms from two opposite versions of the bootstrap procedure and compares them. In one of the samples, the observation under test is *always* removed ($A$), whereas in the other resampling scheme it is forced to be in all the ($P$) bootstrapped samples. The null hypothesis is that the expected value of $\Delta C_i^A$ is larger than $\Delta C_i^P$ (see [7, 8] for more details).

### 2.3.4 Quantile regression

Censored quantile regression was introduced by [27] as an alternative to the Cox’s regression model for survival data and is based on a generalization of the Kaplan-Meier one sample estimator. Due to the fact that the Cox’s regression model assumes proportional hazards, the censored quantile regression can provide more flexibility.

Three different algorithms to detect outlying observations in survival data based on censored quantile regression have been proposed [8] and are implemented in R in the *OutlierDC* package.

Let $Y_i = \min(T_i, C_i)$ represent the observed response variable. The quantile regression model is defined by

$$T_i = \mathbf{x}_i^T \beta(\tau) + \varepsilon_i(\tau), \quad (22)$$
where $\beta(\tau)$ is a $p$-dimensional quantile coefficient vector for $\tau \in (0, 1)$, and $\varepsilon_i(\tau)$ is a random error whose $\tau$th conditional quantile equals zero.

Consider $Q_{T_i}(\tau|x_i) = \inf \{ t : F(t|x_i) \geq \tau \}$, the $\tau$th conditional quantile of $T_i$ given $x_i$, and $F(t|x_i)$ the conditional cumulative distribution function of the survival time $t$ given $x_i$. Then the conditional quantile is given by

$$Q_{T_i}(\tau|x_i) = x_i^T \beta(\tau).$$

(23)

Several techniques are available in the literature to estimate the conditional quantile coefficients $\beta(\tau)$.

The most used technique for outlier detection is proposed by [28]. Their work focuses on locally weighted censored quantile regression which is based on the local Kaplan-Meier estimator.

The three algorithms considered by [6] for outlier detection based on censored quantile regression are: residual-based, boxplot and scoring. The residual-based and boxplot algorithms were developed by modifying existing ones, [4] and [29] respectively, and the scoring algorithm was introduced to provide the outlying magnitude of each point from the distribution of observations and to enable the determination of a threshold by visualizing the scores.

- Residual algorithm

The residual algorithm arises from the outlier detection algorithm for the Cox’s regression model for censored data proposed by [4] but now based on quantile regression. The residual for the $i$th individual is given by

$$r_i = Y_i - Q_{0.50}(x_i),$$

(24)

where $Q_{0.50}(x_i)$ is the 50th conditional quantile for the $i$th individual by censored quantile regression.

Let $k_r$ be a resistant parameter in order to control the cut-offs, and

$$\hat{\sigma} = \text{median} \left\{ |r_{q_i}|, i = 1, ..., n \right\},$$

(25)

with $\hat{\beta}_0 = \Phi^{-1}(p)$ the inverse cumulative distribution of the Normal distribution for quantile $p$. Then the indicator function $O_{r}^{q}$ which gives the information if the $i$th individual is or is not an outlier is defined as

$$O_{r}^{q} = \begin{cases} 1, & \text{if } r_{q_i} > k_{rq} \hat{\sigma} \\ 0, & \text{otherwise} \end{cases},$$

(26)

which mean that if the residual for the $i$th individual, $r_{q_i}$, is higher than a threshold, $k_{rq} \hat{\sigma}$, observation $i$ is considered to be an outlier.
• Boxplot algorithm

The boxplot algorithm is a modification of the algorithm used by \cite{29} using quantile regression for censored data.

Obtaining the outlying individuals involves two steps. First, the censored quantile regression models are fitted for $\tau = 0.25$ and $\tau = 0.75$ in order to obtain the conditional quantile estimates $Q(0.25|x_i)$ and $Q(0.75|x_i)$, respectively. Based on those, the inter-quantile range (IQR) is determined for observation $i$, and an upper fence $UF_i$ is defined as:

$$UF_i = Q(0.75|x_i) + k_{q} IQR(x_i),$$

where $k_{q}$ is a resistant parameter to control the tightness of cut-offs. The indicator function to declare if the $i^{th}$ individual is an outlier is given by

$$O^{by}_i = \begin{cases} 1, & \text{if } Y_i > UF_i \\ 0, & \text{otherwise}. \end{cases}$$

which means that an observation is considered to be an outlier if it is located above the upper fence.

• Score algorithm

In both the residual-based and the boxplot algorithms a threshold should be specified \textit{a priori}. To overcome this limitation, the scoring algorithm was proposed, which is able to determine the deviations from the distribution of the individuals given the covariates using a flexible cut-off, $k_{s}$.

In order to obtain the outlying individuals, first the censored quantile regression model has to be fitted for $\tau = 0.25, 0.50, 0.75$ in order to obtain the conditional quantile estimates, $Q(0.25|x_i)$, $Q(0.50|x_i)$ and $Q(0.75|x_i)$, respectively. By considering those, the outlying score for the $i^{th}$ individual is determined by

$$s_{qi} = \begin{cases} Y_i - Q(0.50|x_i), & \text{if } Y_i > Q(0.50|x_i) \\ Q(0.75|x_i) - Q(0.50|x_i), & \text{if } Y_i \leq Q(0.50|x_i). \end{cases}$$

The indicator function to declare if the $i^{th}$ individual is an outlier is given by

$$O^{sq}_i = \begin{cases} 1, & \text{if } s_{qi} > k_{sq} \\ 0, & \text{otherwise}. \end{cases}$$

where $k_{sq}$ can be determined \textit{a posteriori} by graphical visualization of the Q-Q plot of the scores.
2.4 RANdom SAmple Consensus (RANSAC)

Random sample consensus (RANSAC) was first introduced by [30] as a new paradigm for robust parameter estimation. This method is capable of fitting experimental data when in presence of errors or outliers, and has become a powerful method in the area of computer vision.

The algorithm is based on the following steps: first, a subset $S$ of inliers (size $n$) of the original data $D$ (size $N$) is sampled, and used to fit a model $M$. Given this model, all the points in $D \setminus S$ are tested and all the observations within some pre-defined error tolerance $\epsilon$ are included in a consensus set $S^*$, becoming also inliers. This error tolerance can be, for example, the difference between the real and the predicted value under $M$. If $S^*$ has more than $t$ elements, i.e., the consensus set has cardinality larger than $t$, the model is considered valid and a new model $M^*$ is then estimated based on all these (inliers) points.

The iterative procedure of RANSAC is based on the repetition of the previous procedure $B$ times, which eventually will lead to an accurate consensus model for a large subset of the original data. The advantage of this method is to have a final model that fits well a subset of the data while identifying simultaneously the outlying elements. This strategy is somehow the opposite of other smoothing techniques since it starts from a subset of the data and tries to expand this set with observations that are consistent with the estimated model.

RANSAC has been used in mass spectrometry analysis [31] but their extensive testing and applications to clinical data analysis is still lacking. We have tested the application of the RANSAC paradigm to classification of omics patient data using the logistic model to breast cancer data, see Section 3.3. Our results show that this strategy is very promising for the analysis of high-dimensional data and the concomitant tasks of outlier detection and robust regression.

A prototype R package ransac was developed and will be expanded in the future with other glm supported models.

2.5 Meta-analysis and the rank product test

The outlier detection methods analysed, given their distinct assumptions and rationales, usually lead to distinct sets of solutions and outlyingness rankings. In addition, different estimated models will also significantly influence the obtained results regarding the identification of these discrepant cases.

We propose to adopt a consensus strategy to cope with this expected variability of the results, in order to delineate a more robust regression method and accurate outlier detection framework. The rationale is that, if a given observation is systematically classified as an outlier, independently
of the chosen method and/or model, then our trust on the accuracy of that particular classification should increase.

Statistically, one possibility of performing a consensus ranking of the observations in terms of their relative outlyingness is to use Rank Products (RP). The required input is to have a list of all the observations ranked by their level of outlyingness, which can be based on the previously described residuals and influential measures. This non-parametric statistical technique gained great importance in detecting differentially regulated genes in replicated microarray experiments [32] and can support the meta-analysis of independent studies [33].

Let $n$ be the number of observations and $k$ the number of different methods for outlier detection presented before. Consider $Z_{ij}$ a measure of the deviance (or outlyingness) of the $i$th observation in the $j$th outlier detection method, with $1 \leq i \leq n$ and $1 \leq j \leq k$. The deviance rank for each $Z_{ij}$ considering method $j$ is defined by

$$R_{ij} = \text{rank}(Z_{ij}), \quad 1 \leq R_{ij} \leq n.$$ 

In the case of outlier detection, the lowest ranks indicate that the observation is more outlier than the others, i.e., exhibits larger deviances.

The rank product is defined by:

$$RP_i = \prod_{j=1}^{k} R_{ij}.$$ 

Several methods were proposed in order to estimate the statistical significance of $RP_i$ under the null hypothesis of random rankings (discrete uniform distribution for each method).

In [32] the distribution of $RP_i$ was based on a permutation approach. An alternative formulation that is less computational intensive was described more recently, based on an approximation of the logarithm of these values using the gamma distribution with parameters $(k, 1)$ [34]. In [35] the exact probability distribution for the rank product was derived. However, this approach for large $n$ is increasingly expensive, which motivated another solution based on the geometric mean of upper and lower bounds, defined recursively [36]. The results shows that the algorithm provides accurate approximate $p$-values for the rank product when compared to the exact ones.

Another key issue when performing these tests is related with the multiple testing problem. In fact, since many observations are tested, type-I errors (false positives) will increase. Several correction methods exist that usually adjust $\alpha$ so that the probability of observing at least one significant result due to chance remains below a desired significance level. The Bonferroni correction is one classical choice, with less conservative options
also available, such as the False Discovery Rate (FDR)\footnote{37} through the calculation of q-values.

Another option multiple testing is the Independent hypothesis weighting (IHW)\footnote{38}. This proposal increases power when compared to the Benjamini and Hochberg\footnote{39} method by assigning data-driven weights to each hypothesis. However the IHW is not adequate in the case-studies presented since no covariate is chosen.

We have used the function \texttt{Heskes.pvalues.R}\footnote{36} for testing the RP and the package \texttt{qvalue} for multiple testing correction.

The Rank Product can also be used as a consensus technique to aggregate the distinct results obtained from different models as a meta-analysis strategy to cope with different feature selection methods.

\section{Results and Discussion}

In this section several real clinical case-studies with omics data will be analysed to illustrate the performance of the outlier detection methods reviewed. The objective is evaluate the different strategies to be used in future applications. The chosen datasets try to span common challenges encountered when analysing patients data, namely survival analysis and classification tasks in high-dimensional spaces. All the analysis were performed in R\footnote{40} and are fully documented in the Rmd files – see Annex.

The libraries used for survival analysis were: \texttt{survival}, for the Cox regression model, \texttt{OutlierDC}, outlier detection method based on quantile regression, and \texttt{qvalue}, to determine the q-values. Two versions of the robust Cox regression model were considered: the one proposed in\footnote{10} is available in R, library \texttt{coxrobust}, and an improvement of this method is available in\footnote{1}. The algorithm implementation to obtain the p-values for the rank product, based on the geometric mean, is provided by the authors\footnote{36}.

For outlier detection in logistic regression the following libraries used were: \texttt{glmnet}, \texttt{caret} and \texttt{spls}, to perform dimensionality reduction via the Lasso and elastic net regularization, PLS and sparse PLS classification, respectively; \texttt{binomTools} for graphical inspection of residuals, hat-values and Bonferroni p-values; \texttt{robustbase} for robust logistic regression and \texttt{pROC} for the c-index calculation.

Additionally, several prototype R packages were developed (see Section\footnote{5} for full description): \texttt{brca.data}, to read TCGA data, \texttt{BCGSD}, for outlier detection based on the c-index and \texttt{ransac}, to test the corresponding algorithm.

In the survival analysis examples and for the Cox’s robust regression model\footnote{1}, an exponential weight was chosen. The number of bootstraps used for the BHT and the DBHT was 100.
3.1 Myeloma

The myeloma dataset is composed by clinical information on 16 covariates of 65 patients with multiple myeloma [41]. All the covariates were considered to the Cox regression model and subsequently reduced using the stepwise method in order to decrease the data dimensionality.

The results regarding the Cox regression model and the robust version are presented in Table 1. From the results, only covariate blood urea nitrogen (\texttt{bun}) is statistically significant across all the models tested, whereas gender (\texttt{sex}) and total serum protein (\texttt{total.serum.prot}) are always non-significant. However, proteinuria at diagnosis (\texttt{proteinuria}) and protein in urine (\texttt{protein.urine}) are statistically significant for the Cox regression model and Cox robust proposed by [1] but not significant when using the methodology of [10]. For the hemoglobin (\texttt{hgb}), a \textit{p}-value= 0.032 was obtained for the Cox model, but in the robust version this variable is no longer found to be significant.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cox</th>
<th>CoxRobust</th>
<th>CoxR (Heritier)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>estimate [se]</td>
<td>p-value</td>
<td>estimate [se]</td>
</tr>
<tr>
<td>hgb</td>
<td>-0.136 (-0.063)</td>
<td>0.032</td>
<td>-0.138 (-0.100)</td>
</tr>
<tr>
<td>sex</td>
<td>-0.512 (-0.363)</td>
<td>0.158</td>
<td>-0.546 (-0.639)</td>
</tr>
<tr>
<td>proteinuria</td>
<td>0.066 (-0.028)</td>
<td>0.020</td>
<td>0.060 (-0.055)</td>
</tr>
<tr>
<td>protein.urine</td>
<td>0.925 (-0.381)</td>
<td>0.015</td>
<td>0.860 (-0.520)</td>
</tr>
<tr>
<td>total.serum.prot</td>
<td>0.133 (-0.069)</td>
<td>0.054</td>
<td>0.122 (-0.063)</td>
</tr>
<tr>
<td>bun</td>
<td>0.024 (-0.006)</td>
<td>0.000</td>
<td>0.025 (-0.007)</td>
</tr>
</tbody>
</table>

Table 1: Results for the Cox’s regression model and Cox’s robust (both proposals) for the myeloma dataset.

In order to identify influential observations that may explain those differences, a plot of the robust estimates with log-transformed exponential weights was performed (Figure 1). Observations 40, 44 and 48 are identified as influential observations, since they have the lowest weights.

The results regarding the martingale and deviance residuals are presented in Figure 2. From the martingale residuals, observations 40, 44 and 48 exhibit the lowest values and for the deviance observations 3, 2, 5 and 15 have the highest absolute values.

It is noteworthy that we are analysing the martingale residuals based on their lowest values (long term survivals). A similar analysis should be performed for values near one.

The top-10 outliers obtained for each method are presented in Table 2. As we can see observation 40 appears in the first ten observations in almost all the methods (except in the deviance), observation 44 and 48 appears in the first ten observations in the martingale residual, and on the outlier methods based on quantile regression (residual and score).
Figure 1: Plot of robust estimates with log-transformed exponential weight versus case number for the myeloma data with six covariates.

Figure 2: Plot of the martingale and deviance (absolute value) residuals for the myeloma dataset with six covariates.

As expected, different results are obtained for each of the methods. The application of rank product tests allowed to combine them in consensus ranking using the \(q\)-values.

The results in Table 3 show that for a 1% and 5% level of significance, the observations that are considered outliers are: 40 and 44. Notice that patient 40 was considered an outlier by the diagnostic techniques in the previous study of [41], which confirms the present approach.

### 3.2 Ovarian cancer

The ovarian cancer dataset is based on gene expression data of oncological patients and is constituted by 517 observations over 12042 covariates. This data was obtained from The Cancer Genome Atlas (TCGA). It comprises the follow-up time, survival status and microarray gene expression of 517 patients. The microarray data was obtained using the HG-U133A platform and contains 12042 gene expression levels [42]. The dataset is publicly available through R package curatedOvarianData and it was normalized and aggregated by the TCGA consortium allowing for the analysis to be
Table 2: Top-10 outlying observations for the myeloma dataset with six covariates.

<table>
<thead>
<tr>
<th>id</th>
<th>Martingale</th>
<th>Deviance</th>
<th>BHT</th>
<th>DBHT</th>
<th>Residual</th>
<th>Score</th>
<th>p-values</th>
<th>q-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>3</td>
<td>15</td>
<td>15</td>
<td>15</td>
<td>44</td>
<td>40</td>
<td>(\approx 0)</td>
<td>0.0001</td>
</tr>
<tr>
<td>40</td>
<td>3</td>
<td>23</td>
<td>11</td>
<td>3</td>
<td>13</td>
<td>16</td>
<td>6</td>
<td>0.0042</td>
</tr>
<tr>
<td>65</td>
<td>10</td>
<td>37</td>
<td>8</td>
<td>4</td>
<td>11</td>
<td>17</td>
<td>6</td>
<td>0.0138</td>
</tr>
<tr>
<td>46</td>
<td>5</td>
<td>30</td>
<td>52</td>
<td>52</td>
<td>4</td>
<td>2</td>
<td>6</td>
<td>0.0387</td>
</tr>
<tr>
<td>63</td>
<td>9</td>
<td>5</td>
<td>31</td>
<td>3</td>
<td>45</td>
<td>47</td>
<td>0.0071</td>
<td>0.1147</td>
</tr>
<tr>
<td>57</td>
<td>6</td>
<td>11</td>
<td>3</td>
<td>13</td>
<td>16</td>
<td>9</td>
<td>0.0271</td>
<td>0.1497</td>
</tr>
<tr>
<td>15</td>
<td>62</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>65</td>
<td>53</td>
<td>0.0131</td>
<td>0.1497</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>22</td>
<td>26</td>
<td>25</td>
<td>3</td>
<td>5</td>
<td>0.0071</td>
<td>0.1147</td>
</tr>
<tr>
<td>65</td>
<td>7</td>
<td>12</td>
<td>55</td>
<td>47</td>
<td>2</td>
<td>8</td>
<td>0.0407</td>
<td>0.2939</td>
</tr>
<tr>
<td>35</td>
<td>21</td>
<td>21</td>
<td>17</td>
<td>17</td>
<td>40</td>
<td>47</td>
<td>0.1420</td>
<td>0.8386</td>
</tr>
</tbody>
</table>

Table 3: Ranks for outlier detection methods (Martingale, Deviance, BHT, DBHT, Residual and Score) sorted by q-value. Myeloma dataset.

<table>
<thead>
<tr>
<th>id</th>
<th>Martingale</th>
<th>Deviance</th>
<th>BHT</th>
<th>DBHT</th>
<th>Residual</th>
<th>Score</th>
<th>p-values</th>
<th>q-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>44</td>
<td>1</td>
<td>15</td>
<td>2</td>
<td>16</td>
<td>1</td>
<td>4</td>
<td>(\approx 0)</td>
<td>0.0001</td>
</tr>
<tr>
<td>40</td>
<td>1</td>
<td>15</td>
<td>11</td>
<td>2</td>
<td>10</td>
<td>1</td>
<td>(\approx 0)</td>
<td>0.0067</td>
</tr>
<tr>
<td>57</td>
<td>3</td>
<td>11</td>
<td>3</td>
<td>13</td>
<td>16</td>
<td>6</td>
<td>0.0042</td>
<td>0.0909</td>
</tr>
<tr>
<td>48</td>
<td>2</td>
<td>22</td>
<td>26</td>
<td>25</td>
<td>3</td>
<td>5</td>
<td>0.0071</td>
<td>0.1147</td>
</tr>
<tr>
<td>15</td>
<td>62</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>65</td>
<td>53</td>
<td>0.0131</td>
<td>0.1497</td>
</tr>
<tr>
<td>35</td>
<td>10</td>
<td>37</td>
<td>8</td>
<td>4</td>
<td>11</td>
<td>7</td>
<td>0.0138</td>
<td>0.1497</td>
</tr>
<tr>
<td>64</td>
<td>4</td>
<td>7</td>
<td>44</td>
<td>28</td>
<td>6</td>
<td>9</td>
<td>0.0251</td>
<td>0.2327</td>
</tr>
<tr>
<td>46</td>
<td>5</td>
<td>30</td>
<td>52</td>
<td>52</td>
<td>4</td>
<td>2</td>
<td>0.0387</td>
<td>0.2939</td>
</tr>
<tr>
<td>65</td>
<td>7</td>
<td>12</td>
<td>55</td>
<td>47</td>
<td>2</td>
<td>8</td>
<td>0.0407</td>
<td>0.2939</td>
</tr>
<tr>
<td>35</td>
<td>10</td>
<td>37</td>
<td>8</td>
<td>4</td>
<td>11</td>
<td>7</td>
<td>0.0138</td>
<td>0.1497</td>
</tr>
<tr>
<td>63</td>
<td>9</td>
<td>5</td>
<td>31</td>
<td>3</td>
<td>45</td>
<td>47</td>
<td>0.0071</td>
<td>0.1147</td>
</tr>
<tr>
<td>57</td>
<td>6</td>
<td>11</td>
<td>3</td>
<td>13</td>
<td>16</td>
<td>6</td>
<td>0.0042</td>
<td>0.0909</td>
</tr>
</tbody>
</table>

The clinical data was cleaned using Days to last follow-up and Days to death attributes to detect inconsistencies between them. Only the cases where the number of days matched were included in the analysis. The same process was performed for the attributes Days to death and Vital status, where some cases had as status deceased, but a missing Days to death.

For the analysis 18 genes were considered, based on those selected in a previous study [15], so no model selection was performed. The dataset is a matrix of size 517 × 18, and, in this case, the only genes statistically significant after fitting Cox’s and Cox’s robust models were: CRYAB and SPARC – see Table 4.

The CRYAB gene codes for the crystallin alpha B chain, a protein that acts as a molecular chaperone. Its function is to bind misfolded proteins and, interestingly, some defects associated to this protein and gene have already been associated with cancer, among other diseases. In particular, a recent study [43] analysed which molecular factors could affect ovarian cancer cell...
apoptosis and the authors found out that there was a statistical significant association between the expression of crystallin B (CRYAB) with survival. This protein has, indeed, a negative regulation of tumor necrosis, which may explain these results.

The SPARC gene codes for Secreted protein acidic and rich in cysteine, a protein that appears to be a regulator of cell growth, by interaction with cytokines, the extracellular matrix and also binding calcium, copper, and several others biochemical compounds. This protein is overexpressed in ovarian cancer tissues [44], playing a central role in growth, apoptosis and metastasis. It also has been identified as a candidate therapeutic target [45].

Table 4 shows that observations 113 and 219 are identified as influential observations (lowest weights).

Observation 219 has the lowest martingale residual – see Figure 4. Regarding the deviance residuals, the observations with the highest absolute values are 346, 415 and 202. The results for the top-10 most outlying observations, for the outlier detection methods, are presented in Table 5. Based on the p-values obtained through the rank product test, the observations that are considered outliers, considering the results of the q-values, for a 1% level of significance were: 113, 219, 221, 452 and 455 (see Table 6). Notice that most of these observations corresponds to patients that have a survival time higher than 3000 days. For instance, for observation 452 the survival time was 5481 days, which corresponds to the maximum value in the sample.
A cluster analysis based on the k-means algorithm was also performed in order to establish if there were any kind of relation between those observations. By considering 4 initial clusters, observations 113, 221, 452 and 455, were all in the same cluster. The link between cluster analysis and interpreting the results will be expanded in future work.

### 3.3 Breast invasive carcinoma (BRCA)

To illustrate the application of the outlier detection tools described above for logistic regression of high-dimensional genomic data, the Breast Invasive Carcinoma (BRCA) dataset publicly available (https://cancergenome.nih.gov/) from the Cancer Genome Atlas (TCGA) Data Portal was used. The BRCA
Table 5: Top-10 outlying observations for the ovarian cancer with 18 genes.

<table>
<thead>
<tr>
<th>id</th>
<th>Martingale</th>
<th>Deviance</th>
<th>BHT</th>
<th>DBHT</th>
<th>Residual</th>
<th>Score</th>
<th>p-values</th>
<th>q-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>113</td>
<td>4</td>
<td>79</td>
<td>1</td>
<td>1</td>
<td>7</td>
<td>1</td>
<td>≈ 0</td>
<td>≈ 0</td>
</tr>
<tr>
<td>455</td>
<td>2</td>
<td>11</td>
<td>94</td>
<td>5</td>
<td>6</td>
<td>3</td>
<td>≈ 0</td>
<td>≈ 0</td>
</tr>
<tr>
<td>221</td>
<td>3</td>
<td>14</td>
<td>128</td>
<td>2</td>
<td>18</td>
<td>2</td>
<td>≈ 0</td>
<td>≈ 0</td>
</tr>
<tr>
<td>219</td>
<td>1</td>
<td>5</td>
<td>177</td>
<td>19</td>
<td>4</td>
<td>19</td>
<td>≈ 0</td>
<td>≈ 0</td>
</tr>
<tr>
<td>452</td>
<td>7</td>
<td>28</td>
<td>199</td>
<td>186</td>
<td>1</td>
<td>5</td>
<td>≈ 0</td>
<td>0.00155</td>
</tr>
<tr>
<td>372</td>
<td>6</td>
<td>23</td>
<td>139</td>
<td>201</td>
<td>13</td>
<td>8</td>
<td>0.00016</td>
<td>0.01370</td>
</tr>
<tr>
<td>211</td>
<td>5</td>
<td>18</td>
<td>513</td>
<td>512</td>
<td>3</td>
<td>7</td>
<td>0.00020</td>
<td>0.01430</td>
</tr>
<tr>
<td>115</td>
<td>14</td>
<td>47</td>
<td>20</td>
<td>48</td>
<td>23</td>
<td>70</td>
<td>0.00037</td>
<td>0.02420</td>
</tr>
<tr>
<td>114</td>
<td>11</td>
<td>42</td>
<td>42</td>
<td>139</td>
<td>15</td>
<td>38</td>
<td>0.00054</td>
<td>0.03100</td>
</tr>
<tr>
<td>405</td>
<td>16</td>
<td>148</td>
<td>196</td>
<td>7</td>
<td>20</td>
<td>30</td>
<td>0.00067</td>
<td>0.03440</td>
</tr>
<tr>
<td>55</td>
<td>8</td>
<td>34</td>
<td>329</td>
<td>222</td>
<td>11</td>
<td>16</td>
<td>0.00110</td>
<td>0.04740</td>
</tr>
<tr>
<td>516</td>
<td>10</td>
<td>39</td>
<td>323</td>
<td>405</td>
<td>5</td>
<td>13</td>
<td>0.00105</td>
<td>0.04740</td>
</tr>
<tr>
<td>120</td>
<td>25</td>
<td>63</td>
<td>44</td>
<td>54</td>
<td>35</td>
<td>40</td>
<td>0.00155</td>
<td>0.05710</td>
</tr>
<tr>
<td>155</td>
<td>9</td>
<td>37</td>
<td>116</td>
<td>308</td>
<td>16</td>
<td>27</td>
<td>0.00152</td>
<td>0.05710</td>
</tr>
<tr>
<td>60</td>
<td>12</td>
<td>44</td>
<td>158</td>
<td>453</td>
<td>9</td>
<td>17</td>
<td>0.00168</td>
<td>0.05790</td>
</tr>
</tbody>
</table>

Table 6: Ranks for outlier detection methods (Martingale, Deviance, BHT, DBHT, Residual and Score) sorted by q-value. TCGA 18 genes dataset.

gene expression data were imported using the R prototype package **brca.data** (see Annex).

The BRCA gene expression data is composed of 54,820 variables for a total of 1222 samples (1102 with primary solid tumor, 7 metastatic and 113 with normal tissue) from 1097 individuals. For this study only paired samples were considered, i.e., individuals for whom both normal and cancer tissue were sampled, corresponding to a total of 113 individuals and 226 samples. Two thirds of the individuals were assigned to training samples, whereas the remaining individuals were assigned to testing samples. Normalization of the data was performed prior to data analysis, by subtracting each variable vector by its mean and dividing by its standard deviation.

Given the high dimensionality of the BRCA data and the fact that most data variability might be contained in a small subset of the original variables,
three strategies for data dimension reduction were considered before outlier detection:

1. variable selection by sparse logistic regression based on Lasso and elastic net regularization, using the \texttt{glmnet} library;

2. feature extraction by PLS-DA, using the \texttt{caret} library;

3. sparse variable selection and feature extraction by sparse PLS-DA, using the \texttt{spls} library.

The first strategy, based on Lasso and elastic net ($\alpha = 0.3$) regularization, enabled the selection of 39 and 80 variables, respectively, out of 54,480 original variables. PLS-DA extracted 8 latent variables, which are based on linear combinations of the regional ones, summarising 40% of the data variance. Finally, sparse PLS-DA based on elastic net regularization extracted 7 latent variables built based on 42 original variables. The three models were estimated on a training set and the predictive performance evaluated for a testing set. For both training and test datasets, the three models yielded no misclassifications.

When searching for real and simulated outliers in the entire dataset, following the above dimensionality reduction strategies, numerical problems occurred, with model fitting producing large standard deviations and zero values for the $z$ statistic. Model fitting via logistic regression is sensitive to collinearities among the independent variables in the datasets. Numerical problems associated to complete separation and collinearity are manifested by extraordinarily large estimated standard errors and sometimes by a large estimated coefficients as well \cite{46}. These corresponds to the problems encountered for the BRCA dataset, which yielded non-trusted models that compromised the performance of further detection of outliers and influential observations. Ongoing work on alternative methods for logistic regression is being conducted under Task 4.1 to address these challenges.

At this stage and in order to illustrate the suitability of the methods described for detection of outliers and influential observations, a subset of the features was used. This new dataset was based on two variables selected (ENSG72778 and ENSG235505) by stepwise logistic regression after dimensionality reduction using Lasso regularization.

Logistic regression (LR) applied to this dataset yielded 8 misclassifications: 4 false negatives (observations 22, 56, 60 and 85) and 4 false positives (observations 134, 157, 206 and 220). All these observations were found to be influential by the Cook’s distance (Fig. 5), with observation 56 being the most influential one. However, the Bonferroni outlier test \cite{47} for observation 56, which has largest Studentized residual, yielded a $p$-value of 0.86. Therefore, observation 56, the most influential observation, was not classified as an outlier observation.
Figure 5: The Cook’s distance for all observations obtained after model fitting BRCA and BRCA with simulated outliers (noisy BRCA) datasets with logistic regression; simulated outlier observations are highlighted in red.

The c-index was also used for outlier detection. For each observation \(i\), 1000 bootstrap samples were created by sampling with replacement \(n - 1\) observations from the BRCA dataset. The concordance for each bootstrap sample was computed and the proportion of bootstrap samples having \(C_i - C_{\text{original}} \leq 0\) was taken as the \(p\)-value for observation \(i\). The \(p\)-values represented in Fig. 6 indicate that the above observations, identified as influential observations by the Cook’s distance, are the more outlying observations based on the c-index, as expected, since both methods identify influential observations.

When applying robust logistic regression (RLR) to the BRCA dataset, 7 misclassifications were obtained: 3 false negatives (observation 56, 60 and 85) and 3 false positives (134, 157, 206 and 220). The most down-weighted observations were observations 56, 85 and 134 (Fig. 7), with observation 56 the most down-weighted, thus corroborating the results found by LR.

The introduction of outliers by randomly replacing the response variable (‘tumour’ vs. ‘normal’) by the opposite class in 20 observations was also investigated in order to evaluate the effectiveness of the methods used for the identification of outliers and influential observations. The application of LR to the resulting noisy BRCA dataset yielded 27 misclassifications,
Figure 6: p-values based on the c-index obtained from model fitting with logistic regression 1000 bootstrap samples from BRCA and BRCA with simulated outliers (noisy BRCA) datasets; simulated outlier observations are highlighted in red.

Figure 6: p-values based on the c-index obtained from model fitting with logistic regression 1000 bootstrap samples from BRCA and BRCA with simulated outliers (noisy BRCA) datasets; simulated outlier observations are highlighted in red.

Corresponding to all simulated outlier observations (classified in their original class) plus observations 56, 60, 75, 85, 134, 157 and 206, previously identified as influential observations by LR on the BRCA dataset without simulated outliers. From these observations, all were found as influential by the Cook’s distance (Fig. 5; outliers highlighted in red), except observations 26, 84, 129, 136, and 157 (Fig. 5). Observations 155 and 161 were the most influential observations, provided by their largest Cook’s distances (Fig. 5). The results obtained from measuring the outlierness for the noisy BRCA dataset based on the c-index (Fig. 6) were concordant with the results by the Cook’s distance (Fig. 5), as expected, for the majority of observations.

Robust logistic regression applied to the noisy data resulted in 28 misclassifications, including all simulated outlier observations, except observation 79, plus observations 22, 56, 60, 85, 89, 134, 206 and 220. Down-weighted observations corresponded to all outlier observations, except observation 79 (Fig. 7; outliers highlighted in red), plus observation 56, a consistently found influential observation in both BRCA datasets (with and without outliers).

Both LR and RLR were able to identify the 20 simulated outliers. While with logistic regression the outlier detection measures enabled the identification of the outliers as influential observations and misclassified them into the
original classes, robust logistic regression down-weighted simulated outlier observations and also classified them into the original classes. Overall, the outlier detection methods are consistent and allow to retrieve misclassifications in perturbed datasets. The main challenge to address is the problem of model selection and overfitting, shown to have an impact on the detection of outliers and influential observations.

RANSAC

The RANSAC algorithm was implemented in a prototype R package to support logistic regression using either the \texttt{glmnet} or the \texttt{stats} package. The full description of the package is detailed in Section 5. The BRCA dataset was used in these experiments, comprising 226 cases and using the same model with two features (gene expression levels) described above. As before, the cases are separated in two classes that represent the type of tissue were the sample was retrieved from: 1) primary location of tumor; and 2) normal tissue. RANSAC was then applied to this BRCA dataset using logistic regression in order to perform supervised classification of the two classes.

The workflow to run the RANSAC experiments in the original dataset and in a perturbed version is described by:
1. Select the parameters for RANSAC;

   (a) Number of RANSAC iterations: \( B = 5000 \);
   (b) Minimum number of observations to estimate the initial model: \( n = 30 \);
   (c) Error threshold to accept an observation as an inlier, given by the Pearson residual: \( \epsilon = 0.04 \);
   (d) Minimum percentage of inliers (over all observations) to accept that consensus model: 80%;

2. Estimate a model using RANSAC algorithm for the logistic regression without any regularization based on only the two genes previously selected (ENSG72778 and ENSG235505);

3. Create a baseline model using all observations using the same two features;

4. Run the last two steps again with a small perturbation of the dataset, by manually changing the class of 10 random observations in each group (tumor and normal).

   The results are discussed using two metrics: 1) the root mean square error (RMSE) loss function; and 2) the accuracy in classifying the observed classes.

   Both models perform with high precision in this task, having only 4 misclassification in the results from RANSAC and 8 in the baseline model (out of 226 observations). The good results are also supported by low RMSEs for both models, with RANSAC showing an improvement over the baseline, with values of 1.7799e−2 and 4.1113e−2, respectively. These results show a very good separation of the two classes with the identification of 2–4% of outliers, considering that there are only 2 covariates in the model. When studying the distribution of the response probability we observe that the predicted values in the RANSAC model have a higher confidence, being either 0 or 1, while the baseline model is more uncertain with several observation in the border of the two classes as Figure 8 shows.

   In this dataset RANSAC performs with a higher accuracy and is able to select a subset of observations that are potential outliers. These observations are consistent with previous results are all 4 potential outliers are common to the results of the baseline model and of the robust logistic regression. The 4 potential outliers are observations 56, 75, 85 and 157.

   The BRCA dataset was modified by switching the class of 10 observations to test the robustness of RANSAC, and how it performed compared to the baseline model and the previous section, as the same noisy dataset was used. We observe a degradation of performance of the baseline model while RANSAC classifies the observation with high precision as Figure 9 shows.
Both methods are able to identify all the observations that were modified, however, RANSAC also identifies observations 56, 75, 85 and 157 that were also potential outliers in the original non-perturbed dataset. The baseline model identifies 7 additional outliers, where 5 are common with the outliers flagged in the non-noisy dataset.

**Figure 9:** Classification obtained from the noisy BRCA dataset with the RANSAC and baseline model (Logistic regression). The noisy dataset was obtained by changing the class of 10 random observations in each group (normal and tumour).

RANSAC’s results indicate that this method is robust to outliers and has consistent results with increasing noisy data but it is not without caveats, as parameter tuning is required and it is a computationally heavy method. Extensive testing of RANSAC with other parameters and using different datasets is currently being performed.

### 3.4 Chronic lymphocytic leukemia (CLL)

The Chronic lymphocytic leukemia (CLL) dataset is composed by 92 observations and 15,366 gene expression values, as well as follow-up times and
survival status. It is part of the PACE rpackage developed by the SOUND consortium.

The original dataset dimensionality was reduced using a two-step procedure. Firstly, model fittings was performed using the following regularization methods: 1) Lasso; 2) Lasso+DegreeCox; 3) Lasso+elasticNet; and 4) Lasso+elasticNet+DegreeCox, leading to four different sets of selected genes. The union of all these sets was then considered - this allowed reduce the dimensionality from 15,366 to 64 features and retrieve several sets of genes that may be associated with the outcome. Secondly, these 64 features were further reduced using elastic net regularization [48], which leads to 19 selected genes. Since the obtained results led to a model with some non statistically significant parameters (Table 7), a stepwise selection was further applied with the step function leading to a model with 10 genes.

In the presented results the covariates (genes) names were changed to simplify the outcome - for example, gene ENSG00000001561 is coded as ENSG1561.

The dataset is matrix $92 \times 19$ and the results obtained by the Cox’s and Cox’s robust are presented in Table 7. Notice that for the Cox’s regression model the genes that are significant (5% level of significance) are: ENSG1561, ENSG3400, ENSG5022, ENSG5156 and ENSG3402. In the first approach of the robust version of Cox’s, only genes ENSG3400 and ENSG5156 were considered significant. Regarding the second robust approach, genes ENSG1561, ENSG1629, ENSG2330, ENSG3400, ENSG4838, ENSG4897, ENSG5022, ENSG5156, ENSG3402, were considered significant. The reason of different results for the three different regression models could be the presence of influential observations.

Figure 10 shows that observations $H051$, $H078$, $H148$ and $H014$ are identified as influential observations (lowest weights).

Figure 10: Plot of robust estimates with log-transformed exponential weight versus case number for the CLL data with 19 genes.
Table 7: Results for the Cox’s regression model and Cox’s robust (both proposals) for the CLL data with 19 genes.

<table>
<thead>
<tr>
<th>variables</th>
<th>coef</th>
<th>se(coef)</th>
<th>p-value</th>
<th>coef</th>
<th>se(coef)</th>
<th>p-value</th>
<th>estimate</th>
<th>SE</th>
<th>p-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>ENSG1461</td>
<td>0.6574</td>
<td>1.4890</td>
<td>0.6589</td>
<td>1.4375</td>
<td>2.5990</td>
<td>0.5802</td>
<td>1.4739</td>
<td>1.4577</td>
<td>0.3120</td>
</tr>
<tr>
<td>ENSG1497</td>
<td>2.2182</td>
<td>2.1403</td>
<td>0.3000</td>
<td>1.4975</td>
<td>3.9440</td>
<td>0.7042</td>
<td>1.5068</td>
<td>2.9402</td>
<td>0.6083</td>
</tr>
<tr>
<td>ENSG1561</td>
<td>-1.2797</td>
<td>0.4770</td>
<td>0.0073</td>
<td>-1.4258</td>
<td>0.8050</td>
<td>0.0764</td>
<td>-1.4377</td>
<td>0.4833</td>
<td>0.0029</td>
</tr>
<tr>
<td>ENSG1629</td>
<td>2.7768</td>
<td>1.5785</td>
<td>0.0785</td>
<td>3.7171</td>
<td>2.2000</td>
<td>0.0911</td>
<td>3.8285</td>
<td>1.5510</td>
<td>0.0136</td>
</tr>
<tr>
<td>ENSG2330</td>
<td>1.6498</td>
<td>1.1763</td>
<td>0.1607</td>
<td>1.1590</td>
<td>1.0759</td>
<td>0.0779</td>
<td>1.1714</td>
<td>0.8900</td>
<td>0.0174</td>
</tr>
<tr>
<td>ENSG2745</td>
<td>-1.2797</td>
<td>0.4770</td>
<td>0.0073</td>
<td>-1.4258</td>
<td>0.8050</td>
<td>0.0764</td>
<td>-1.4377</td>
<td>0.4833</td>
<td>0.0029</td>
</tr>
<tr>
<td>ENSG2834</td>
<td>4.5999</td>
<td>2.5950</td>
<td>0.0763</td>
<td>4.3210</td>
<td>4.1000</td>
<td>0.3272</td>
<td>4.3235</td>
<td>2.7546</td>
<td>0.1165</td>
</tr>
<tr>
<td>ENSG3056</td>
<td>-0.5465</td>
<td>1.4933</td>
<td>0.7144</td>
<td>-1.1100</td>
<td>2.7560</td>
<td>0.6872</td>
<td>-1.0928</td>
<td>1.4734</td>
<td>0.4583</td>
</tr>
<tr>
<td>ENSG3249</td>
<td>0.0083</td>
<td>0.5996</td>
<td>0.9817</td>
<td>0.0090</td>
<td>0.3530</td>
<td>0.9866</td>
<td>0.0181</td>
<td>0.3462</td>
<td>0.9582</td>
</tr>
<tr>
<td>ENSG3400</td>
<td>1.4622</td>
<td>0.5696</td>
<td>0.0102</td>
<td>1.5674</td>
<td>0.7940</td>
<td>0.0482</td>
<td>1.5584</td>
<td>0.4604</td>
<td>0.0007</td>
</tr>
<tr>
<td>ENSG4139</td>
<td>-2.2865</td>
<td>1.4957</td>
<td>0.1263</td>
<td>-2.4972</td>
<td>2.4040</td>
<td>0.2990</td>
<td>-2.5310</td>
<td>1.4221</td>
<td>0.0751</td>
</tr>
<tr>
<td>ENSG4468</td>
<td>-0.0267</td>
<td>0.1670</td>
<td>0.8729</td>
<td>-0.0344</td>
<td>0.2880</td>
<td>0.9024</td>
<td>-0.0367</td>
<td>0.1750</td>
<td>0.8338</td>
</tr>
<tr>
<td>ENSG4660</td>
<td>-0.0677</td>
<td>0.6659</td>
<td>0.9191</td>
<td>-0.0552</td>
<td>0.9780</td>
<td>0.9557</td>
<td>-0.0153</td>
<td>0.5229</td>
<td>0.9767</td>
</tr>
<tr>
<td>ENSG4766</td>
<td>-1.3016</td>
<td>2.3235</td>
<td>0.5754</td>
<td>-1.6067</td>
<td>3.9420</td>
<td>0.6385</td>
<td>-1.4804</td>
<td>2.2818</td>
<td>0.5165</td>
</tr>
<tr>
<td>ENSG4838</td>
<td>-1.8786</td>
<td>1.5239</td>
<td>0.1341</td>
<td>-2.1790</td>
<td>1.5509</td>
<td>0.1622</td>
<td>-2.1145</td>
<td>0.8380</td>
<td>0.0116</td>
</tr>
<tr>
<td>ENSG4897</td>
<td>-2.2522</td>
<td>1.5713</td>
<td>0.1567</td>
<td>-2.5574</td>
<td>2.2250</td>
<td>0.2504</td>
<td>-2.6445</td>
<td>1.3268</td>
<td>0.0462</td>
</tr>
<tr>
<td>ENSG5022</td>
<td>-5.6807</td>
<td>2.1267</td>
<td>0.0076</td>
<td>-6.1094</td>
<td>3.1530</td>
<td>0.0526</td>
<td>-6.0593</td>
<td>1.9548</td>
<td>0.0019</td>
</tr>
<tr>
<td>ENSG5156</td>
<td>5.7170</td>
<td>2.0522</td>
<td>0.0053</td>
<td>6.2066</td>
<td>2.8060</td>
<td>0.0270</td>
<td>6.2195</td>
<td>2.0714</td>
<td>0.0027</td>
</tr>
<tr>
<td>ENSG5156</td>
<td>1.6500</td>
<td>1.5674</td>
<td>0.0070</td>
<td>4.1792</td>
<td>2.8880</td>
<td>0.1462</td>
<td>4.2245</td>
<td>1.8487</td>
<td>0.0223</td>
</tr>
</tbody>
</table>

The results regarding the residuals are shown in Figure 11. Observations H014, H078 and H148 in the martingale residuals have the lowest values when compared to the all the others. Regarding the deviance residuals, the observations with the highest absolute values are H045, H230 and H023. From the results presented for these residuals it seems that there are not matches of the observations. However observation H078, which is the second lowest value for the martingale residual, has the fourth highest absolute value for the deviance residual.

Table 8: Top-10 outlying observations for the CLL data with 19 genes.

<table>
<thead>
<tr>
<th>Martingale</th>
<th>Deviance</th>
<th>BHT</th>
<th>DBHT</th>
<th>Residual</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>H078</td>
<td>H045</td>
<td>H005</td>
<td>H045</td>
<td>H078</td>
<td>H044</td>
</tr>
<tr>
<td>H148</td>
<td>H230</td>
<td>H016</td>
<td>H115</td>
<td>H051</td>
<td>H101</td>
</tr>
<tr>
<td>H014</td>
<td>H023</td>
<td>H019</td>
<td>H171</td>
<td>H014</td>
<td>H102</td>
</tr>
<tr>
<td>H029</td>
<td>H078</td>
<td>H020</td>
<td>H029</td>
<td>H056</td>
<td>H105</td>
</tr>
<tr>
<td>H051</td>
<td>H115</td>
<td>H021</td>
<td>H013</td>
<td>H111</td>
<td>H187</td>
</tr>
<tr>
<td>H005</td>
<td>H029</td>
<td>H023</td>
<td>H005</td>
<td>H042</td>
<td>H049</td>
</tr>
<tr>
<td>H096</td>
<td>H013</td>
<td>H036</td>
<td>H078</td>
<td>H173</td>
<td>H021</td>
</tr>
<tr>
<td>H012</td>
<td>H005</td>
<td>H042</td>
<td>H027</td>
<td>H010</td>
<td>H073</td>
</tr>
<tr>
<td>H027</td>
<td>H096</td>
<td>H045</td>
<td>H051</td>
<td>H083</td>
<td>H236</td>
</tr>
<tr>
<td>H017</td>
<td>H042</td>
<td>H051</td>
<td>H099</td>
<td>H040</td>
<td>H039</td>
</tr>
</tbody>
</table>

Table 8: Top-10 outlying observations for the CLL data with 19 genes.

The results for the top-10 most outlying observations, for the outlier detection methods, are presented in Table 8. Based on the rank product
test, the corresponding $q$-values were obtained. The results in Table 9 show that for a 5% level of significance, the observations that are considered outliers were $H078$, $H005$ and $H051$.

![Plot of the martingale and deviance (absolute value) residuals for the CLL data with 19 genes](image)

Table 9: Ranks for outlier detection methods (Martingale, Deviance, BHT, DBHT, Residual and Score) sorted by $q$-value. CLL dataset with 19 covariates.

<table>
<thead>
<tr>
<th>id</th>
<th>Martingale</th>
<th>Deviance</th>
<th>BHT</th>
<th>DBHT</th>
<th>Residual</th>
<th>Score</th>
<th>p-values</th>
<th>q-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>H078</td>
<td>1</td>
<td>4</td>
<td>12</td>
<td>7</td>
<td>1</td>
<td>21</td>
<td>$\approx$0</td>
<td>0.00106</td>
</tr>
<tr>
<td>H005</td>
<td>6</td>
<td>8</td>
<td>1</td>
<td>6</td>
<td>14</td>
<td>27</td>
<td>0.00031</td>
<td>0.01417</td>
</tr>
<tr>
<td>H051</td>
<td>5</td>
<td>19</td>
<td>10</td>
<td>9</td>
<td>2</td>
<td>18</td>
<td>0.00093</td>
<td>0.02852</td>
</tr>
<tr>
<td>H014</td>
<td>3</td>
<td>16</td>
<td>29</td>
<td>12</td>
<td>3</td>
<td>17</td>
<td>0.00255</td>
<td>0.05864</td>
</tr>
<tr>
<td>H029</td>
<td>4</td>
<td>6</td>
<td>32</td>
<td>4</td>
<td>18</td>
<td>33</td>
<td>0.00517</td>
<td>0.09518</td>
</tr>
<tr>
<td>H045</td>
<td>92</td>
<td>1</td>
<td>9</td>
<td>1</td>
<td>74</td>
<td>56</td>
<td>0.00902</td>
<td>0.13831</td>
</tr>
<tr>
<td>H027</td>
<td>9</td>
<td>17</td>
<td>30</td>
<td>8</td>
<td>17</td>
<td>38</td>
<td>0.04155</td>
<td>0.34748</td>
</tr>
<tr>
<td>H042</td>
<td>87</td>
<td>10</td>
<td>8</td>
<td>11</td>
<td>6</td>
<td>42</td>
<td>0.03575</td>
<td>0.34748</td>
</tr>
<tr>
<td>H096</td>
<td>7</td>
<td>9</td>
<td>14</td>
<td>89</td>
<td>11</td>
<td>25</td>
<td>0.03881</td>
<td>0.34748</td>
</tr>
<tr>
<td>H115</td>
<td>89</td>
<td>5</td>
<td>50</td>
<td>2</td>
<td>13</td>
<td>35</td>
<td>0.03704</td>
<td>0.34748</td>
</tr>
<tr>
<td>H148</td>
<td>2</td>
<td>15</td>
<td>20</td>
<td>38</td>
<td>20</td>
<td>37</td>
<td>0.03237</td>
<td>0.34748</td>
</tr>
<tr>
<td>H019</td>
<td>80</td>
<td>38</td>
<td>3</td>
<td>31</td>
<td>12</td>
<td>36</td>
<td>0.12243</td>
<td>0.75088</td>
</tr>
<tr>
<td>H021</td>
<td>34</td>
<td>48</td>
<td>5</td>
<td>60</td>
<td>33</td>
<td>7</td>
<td>0.11690</td>
<td>0.75088</td>
</tr>
<tr>
<td>H081</td>
<td>16</td>
<td>24</td>
<td>13</td>
<td>34</td>
<td>31</td>
<td>19</td>
<td>0.10845</td>
<td>0.75088</td>
</tr>
<tr>
<td>H171</td>
<td>19</td>
<td>28</td>
<td>21</td>
<td>3</td>
<td>72</td>
<td>46</td>
<td>0.11559</td>
<td>0.75088</td>
</tr>
</tbody>
</table>

Table 10 contains the results of the lowest $q$-values obtained. In this approach, case $H078$ was again identified as outlier.

All the results obtained for this analysis are available in .Rmd file.
### Table 10: Ranks for outlier detection methods (Martingale, Deviance, BHT, DBHT, Residual and Score) sorted by q-value. CLL dataset with 10 covariates.

<table>
<thead>
<tr>
<th>id</th>
<th>Martingale</th>
<th>Deviance</th>
<th>BHT</th>
<th>DBHT</th>
<th>Residual</th>
<th>Score</th>
<th>p-values</th>
<th>q-values</th>
</tr>
</thead>
<tbody>
<tr>
<td>H014</td>
<td>1</td>
<td>11</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>11</td>
<td>≈ 0</td>
<td>≈ 0</td>
</tr>
<tr>
<td>H078</td>
<td>2</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>1</td>
<td>10</td>
<td>≈ 0</td>
<td>≈ 0</td>
</tr>
<tr>
<td>H045</td>
<td>92</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>61</td>
<td>60</td>
<td>0.00102</td>
<td>0.03127</td>
</tr>
<tr>
<td>H056</td>
<td>9</td>
<td>15</td>
<td>8</td>
<td>8</td>
<td>5</td>
<td>14</td>
<td>0.00183</td>
<td>0.04212</td>
</tr>
<tr>
<td>H017</td>
<td>4</td>
<td>24</td>
<td>38</td>
<td>12</td>
<td>3</td>
<td>12</td>
<td>0.00453</td>
<td>0.08332</td>
</tr>
<tr>
<td>H005</td>
<td>7</td>
<td>13</td>
<td>24</td>
<td>11</td>
<td>15</td>
<td>20</td>
<td>0.01673</td>
<td>0.02322</td>
</tr>
<tr>
<td>H051</td>
<td>6</td>
<td>29</td>
<td>16</td>
<td>63</td>
<td>4</td>
<td>13</td>
<td>0.02019</td>
<td>0.23219</td>
</tr>
<tr>
<td>H238</td>
<td>89</td>
<td>4</td>
<td>4</td>
<td>2</td>
<td>38</td>
<td>72</td>
<td>0.01782</td>
<td>0.23219</td>
</tr>
<tr>
<td>H083</td>
<td>10</td>
<td>16</td>
<td>21</td>
<td>59</td>
<td>21</td>
<td>5</td>
<td>0.03779</td>
<td>0.38631</td>
</tr>
<tr>
<td>H096</td>
<td>3</td>
<td>8</td>
<td>29</td>
<td>58</td>
<td>24</td>
<td>30</td>
<td>0.04802</td>
<td>0.40162</td>
</tr>
<tr>
<td>H148</td>
<td>8</td>
<td>42</td>
<td>22</td>
<td>13</td>
<td>14</td>
<td>19</td>
<td>0.04383</td>
<td>0.40162</td>
</tr>
<tr>
<td>H230</td>
<td>91</td>
<td>2</td>
<td>6</td>
<td>9</td>
<td>60</td>
<td>58</td>
<td>0.05381</td>
<td>0.41252</td>
</tr>
<tr>
<td>H012</td>
<td>5</td>
<td>27</td>
<td>44</td>
<td>23</td>
<td>16</td>
<td>21</td>
<td>0.06577</td>
<td>0.46548</td>
</tr>
<tr>
<td>H027</td>
<td>23</td>
<td>33</td>
<td>18</td>
<td>6</td>
<td>23</td>
<td>36</td>
<td>0.08504</td>
<td>0.48244</td>
</tr>
<tr>
<td>H030</td>
<td>16</td>
<td>23</td>
<td>15</td>
<td>41</td>
<td>36</td>
<td>8</td>
<td>0.08285</td>
<td>0.48244</td>
</tr>
</tbody>
</table>

### 4 Conclusions and Future Work

The goal of this report was to revisit different methods for outlier detection and applied them to real patient’omic data, in order to identify problems that may hamper the correct identification of discrepant observations. The focus was on a model or parametric framework, warranted by the desired interpretability of the high number of features associated with omics data. In particular, Cox and logistic models were applied to survival and binary multi-dimensional data, respectively, as representatives of common case-studies related with Sound’s WPs. These include TCGA ovarian and breast cancer and CLL datasets.

As expected, the results are usually dependent on the method chosen. Furthermore, different model selection strategies also lead to distinct results. To cope with this problem, we propose a consensus method based on the Rank Product test, which allows to combine the results obtained by each method (and/or model) and to identify the observations systematically highly ranked in terms of their outlyingness level. The proposed application of the Rank Product test illustrates that it is possible to combine disparate methods and to obtain a consensus list of putative outliers that can further be explored from a clinical point of view. It remains a question for future work how to combine the uncertainty between different models and retrieve observations that are systematically classified as outliers (independently of the model considered).

RANSAC is also proposed as a robust method for parameter estimation and outlier detection in this scope, which allows to fit a model to a consensus
set and simultaneously identify discrepant observations. Preliminary results under the logistic regression illustrate the potential of this strategy, that will be further explored and expanded.

A crucial issue that remains to be solved is the $N \ll p$ or high-dimensional problem, which occurs very frequently in patient’omics data where the number of features ($p$) is much larger that the sample size ($N$). Classical estimation procedures tend to fail and the resulting models may lack interpretability, which may severely affect outlier detection. This challenge is now being addressed in Task 4.1 and it is expected that the solutions will be coupled with the described pipelines of the present report.

5 Annex

5.1 Prototype R packages

Several prototype R packages have been developed and are temporarily available in the URLs described below.

**BRCA.Data** The **BRCA.Data** R package is a data package that contains a snapshot of TCGA breast invasive carcinoma (BRCA) clinical, expression levels and mutation data. The gene expression levels (HTSeq - FPKM) are organized by sample origin type, which can be primary solid tumor tissue, normal solid tissue or metastatic tissue. The purpose of the package is to make this data easily available and ready to use as matrices.

This package takes advantage of the TCGAbiolinks [49] bioconductor’s package to download and prepare the data directly from TCGA. Both the original data and the post-processed are available in the package as data objects, see [http://sels.tecnico.ulisboa.pt/gitlab/SOUND/brca.data/repository/archive.zip](http://sels.tecnico.ulisboa.pt/gitlab/SOUND/brca.data/repository/archive.zip)

**RANSAC** The **RANSAC** R package is a prototype package that has implemented the RANSAC algorithm to fit a model to a set of predictors and responses. It currently supports logistic regression from stats and glmnet rpackages but can be extended to other parameter estimation methods by adding a list of functions.

The **RANSAC** package is available at [http://sels.tecnico.ulisboa.pt/gitlab/SOUND/ransac/repository/archive.zip](http://sels.tecnico.ulisboa.pt/gitlab/SOUND/ransac/repository/archive.zip) and is licensed under GPLv3.

**BCSOD** The **BCSOD** – bootstrapped concordance survival outlier detection, is a prototype R package that implements the algorithms described in [7][8] and is available at [http://sels.tecnico.ulisboa.pt/gitlab/SOUND/bcsod/repository/archive.zip](http://sels.tecnico.ulisboa.pt/gitlab/SOUND/bcsod/repository/archive.zip)
5.2 R Markdown files

The presented results are available as R Markdown (.Rmd) and html documents, along with the original data files used in the analysis performed. Table 11 lists all the folders accompanying the present Deliverable 4.2 report. All these files are available at [http://web.tecnico.ulisboa.pt/~susanavinga/SOUND/Deliverable_4.2/](http://web.tecnico.ulisboa.pt/~susanavinga/SOUND/Deliverable_4.2/)

Table 11: Annex - Files

<table>
<thead>
<tr>
<th>Folders and Files</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>README.txt</td>
<td></td>
</tr>
<tr>
<td>Dataset1-myeloma</td>
<td>Chapter7.rdata</td>
</tr>
<tr>
<td></td>
<td>myeloma.txt</td>
</tr>
<tr>
<td></td>
<td>myeloma_dataset.Rmd</td>
</tr>
<tr>
<td></td>
<td>myeloma_dataset.html</td>
</tr>
<tr>
<td>Dataset2-ovarian</td>
<td>TCGA-18genes.Rmd</td>
</tr>
<tr>
<td></td>
<td>TCGA-18genes.html</td>
</tr>
<tr>
<td></td>
<td>tega18.txt</td>
</tr>
<tr>
<td></td>
<td>timestatus.txt</td>
</tr>
<tr>
<td>Dataset3-BRCA</td>
<td>BRCA.data.RData</td>
</tr>
<tr>
<td></td>
<td>BRCA.data.norm.RData</td>
</tr>
<tr>
<td></td>
<td>BRCA_OutlierDetection.Rmd</td>
</tr>
<tr>
<td></td>
<td>BRCA_OutlierDetection.html</td>
</tr>
<tr>
<td></td>
<td>aucLOGIT.boot.R</td>
</tr>
<tr>
<td></td>
<td>ransac.Rmd</td>
</tr>
<tr>
<td></td>
<td>ransac.html</td>
</tr>
<tr>
<td>Dataset4-CLL</td>
<td>CLL-19covariates.Rmd</td>
</tr>
<tr>
<td></td>
<td>CLL-19covariates.html</td>
</tr>
<tr>
<td></td>
<td>CLL-19covariatesStep10.Rmd</td>
</tr>
<tr>
<td></td>
<td>CLL-19covariatesStep10.html</td>
</tr>
<tr>
<td></td>
<td>functions</td>
</tr>
<tr>
<td></td>
<td>Chapter7.r</td>
</tr>
<tr>
<td></td>
<td>Chapter7_functions.r</td>
</tr>
<tr>
<td></td>
<td>Heskes_pvalues.R</td>
</tr>
</tbody>
</table>

References


