Author’s response to reviews

Title: Identification of influential observations in high-dimensional cancer survival data through the rank product test

Authors:
Eunice Trigueirão (eunice.trigueirao@tecnico.ulisboa.pt)
Susana Vinga (susanavinga@tecnico.ulisboa.pt)
Marta Lopes (marta.lopes@ist.utl.pt)
André Veríssimo (andre.verissimo@tecnico.ulisboa.pt)

Version: 1 Date: 23 Nov 2017

Author’s response to reviews:

BIDM-D-17-00040

Identification of influential observations in high-dimensional cancer survival data through the rank product test

Eunice Carrasquinha Trigueirão, Ph.D.; Susana Vinga, Ph.D; Marta B. Lopes, Ph.D; André Veríssimo

BioData Mining

23 November 2017

Dear Editor,

The authors wish to thank the Editor and Reviewers for their time and effort in reviewing our manuscript. We have addressed all the concerns raised point-by-point, which implied adding more results to demonstrate the robustness of our method. In particular, we now added a resampling strategy to illustrate the consistency of the resulting identification of the outliers. To ease the reviewer process, we are also submitted a DIFF file with the differences between both versions.

We believe the changes listed below have made the manuscript clearer and suitable for publication in BioData Mining.

Looking forwards to hearing from you soon.
Sincerely,

Eunice Carrasquinha

On behalf of the authors

-----

Reviewer reports:

Reviewer #1: In this manuscript the authors proposed a Rank Product method to rank the residuals of each sub-model based on different number of selected genes on the same group of samples to identify the outliers for the survival analysis. The work may important to those with closely related research interests, however, more experiments are required to evaluate the proposed model.

1) Both Rank Product and regularized survival model have been proposed before, the contribution of the model is limited.

The Rank Product (RP) has been used before, along with the description of its statistical properties under the null hypothesis, and the corresponding approximation methods, as cited. Also, regularized optimization has already been proven to be an effective and successful strategy to deal with high-dimensional data. Although we agree with Reviewer#1 that both approaches have been applied before, the core of our manuscript is precisely the combination of these two techniques to deal with a key problem in survival analysis. The idea of using ranks of some deviance or residuals derived from a regression model and then use these ranks to obtain an ensemble list is, to the best of our knowledge, original and constitutes the main novelty of the approach. As illustrated, this confers robustness to the outlier detection task.

We agree that our description was not clear enough and, in the new version, we highlight what is our major contribution in this context, page 2, line -7:

“The aim of this work is, therefore, given a high-dimensional dataset, to find outliers (or influential observations) from different sub-models, which are obtained from distinct techniques for variable selection. The method proposed is based on the Rank Product (RP) test, a non-parametric method, to identify the outliers that are consistently highly ranked in each of the sub-models. The ovarian cancer dataset, with gene expressions as covariates, was chosen to illustrate the applicability of the proposed method. Three gene expression sub-models are presented, and the RP test is applied as a consensus or ensemble test that combines the results obtained by each model, often distinct and sometimes contradictory. Notice that each sub-model has different baselines, since for this particular dataset there is no groundtruth to start from.
Although the rank product and the deviances measures for survival models were already proposed previously in different contexts, the combination of RP-based statistical tests as a means of conferring robustness to outlier detection tasks represents the main novelty of this work.

2) The experiment is not comprehensive to show the importance of the model and lack of baseline method to compare with.

When dealing with high-dimensional datasets, the dimension reduction has to be taken in consideration. The solution to this issue is not straightforward, since there are a variety of techniques to overcome this. For different methods to reduce the dimensionality, different models are obtained and, consequently, different outliers are detected. The Rank Product (RP) test is applied here to give us a consensual answer when different results are obtained depending on the model chosen. For the three sub-models used, different results for the outliers, were obtained. However, by using the RP test, we identify the outliers that are consistently highly ranked in each sub-models.

Regarding the lack of a baseline method to compare the results, there is no groundtruth for this particular dataset regarding the identity of the outliers. In order to fulfill this gap, a sampling strategy was now designed to determine whether resampling the data using a subset of covariates would identify the same outlier observations. The following paragraph was added in the manuscript (page 6, line 7):

“To overcome the fact that the results obtained for each of the analysis are model-based, a sampling strategy was also implemented in order to determine whether resampling the data using a sub-model of covariates (genes) would recognize the outliers previously identified. The resampling algorithm randomly picked 1000 genes (without replacement) from the ovarian cancer dataset. The Cox regression model with elastic-net regularization was then fitted (using glmnet), resulting in a reduced set of selected genes. In order to calculate the corresponding martingale residuals, a Cox regression is then performed on this reduced gene set (using coxph). The resulting residuals allow to sort the observations accordingly to their outlyingness level. This procedure is repeated 100 times, resulting in 100 models to feed the RP test.”

The results concerning the resampling strategy are presented in the Rank Product results. By creating new models based on different subsets of covariates, the same outliers could be identified, thus showing the robustness of the RP test as a consensual procedure when the results depend on the method used for variable selection.

The best baseline is provided by the individual sub-models results. In fact, each of the Cox model obtained leads to different rankings. For this particular TCGA ovarian dataset we do not have a groundtruth information to know which observation is considered outlier. Nevertheless, the resampling technique shows the robustness of the RP test.

3) The authors claimed the method can identify the outliers in the high-dimensional space, however, 517 samples cannot be considered as a very large number in the survival
analysis. In addition, the sub-models assume that the important genes have been identified.

When we refer to “high-dimensional”, we meant the space of the covariates (gene expression data), which corresponds to p=12,042 values/genes, and not the sample size (n=517), following the definitions in machine learning and biostatistics (Bühlmann and S. van de Geer, 2011). In fact, it is precisely in this setting that the estimation problem becomes quite challenging, due to the inherent ill-posed inverse problem.

To fully clarify this key aspect, we added the following paragraph in the Background (page 2, line 24):

“One of the challenges arising when dealing with patient’ omics data is the high-dimensionality problem. In this type of data, the number of covariates (p) is often much larger than the number of observations (n), i.e., p>>n. In this context, the usual statistical techniques for the estimation of the parameters cannot be applied, due to the inherent ill-posed inverse problem (Bühlmann and S. van de Geer, 2011).”

The following reference was also added:


Regarding the second part of the question, only in two of the chosen sub-models we have used prior information. The first analysis consisted in the (automatic) gene selection using regularization techniques. This was to illustrate that our method is very general, in the sense that we can use rankings derived from multiple strategies simultaneously and build an ensemble rank.

We agree that our description was not entirely clear and the following sentence was added in the Results section (page 6, line 4):

“It is noteworthy that, although we have pursued these three analyses, we can indeed include many others, for example, using different feature selection methods or prior clinical information.”

Additionally, with the new resampling strategy, we also partially overcome this problem.

Reviewer #2: In the manuscript 'Identification of influential observations in high-dimensional cancer survival data through the rank product test', Carrasquinha et al propose the application of the Rank Product test as a consensus method to prioritize outlying observations reported by different sub-models in survival analysis. This is an interesting report, however, below points need to be addressed.

1. For the datasets where the Cox regression model is applied, are the data normally distributed? Regression models, including the Cox model, generally generate more reliable results with normally distributed data. The authors should provide information on the data distribution.
Indeed, regression models, in general, produce more consistent results with normally distributed data, however, in the Cox’s regression model, the only assumption is the proportional hazard hypothesis, which was comprehensively assessed for all the estimated models (data not shown in our first version of the manuscript). To clarify this issue, the results of the proportional hazard hypothesis for each analysis are now included in the new version.

We added the following paragraph in the Results (see page 6, line 25), and an additional reference:

“ The proportional hazard assumption (Smith, 2002) for the Cox's regression model was tested, and the results showed that this hypothesis was never violated. The p-values for each of the sub-model presented are the following: 0.1932 (63 genes), 0.3795 (18 genes) and 0.3868 (22 genes).


Despite the fact that the assumption of the normal distribution of the data is not required for the Cox’s regression model, we analysed each gene for every sub-model. The Shapiro test was used, and the results revealed that for the 18 genes sub-model, none of them was normally distributed; for the 22 genes sub-model, five genes were normal distributed (BARD1, MLH1, NBN, PALB2, PMS2) and for the 63 genes sub-model, eight genes were normal distributed (UBE2J1, FLJ20323, NDUFA3, FJX1, RAB40B, PPM2C, WTAP, EHMT1). This result does not affect the resulting models’ validity.

We also added the following, page 6, line 29:

“The majority of gene expression do not have a normal distribution (see Supplementary files for the Shapiro tests conducted) although this fact does not affect the resulting Cox models’ validity.”

2. Neither of the two outliers (39 and 350) from the 63-gene sub-model is on the final outlier list defined by the RP method. Also, in Table3, it looks the ranks reported by the other two sub-models (18 genes and 22 genes) are more similar (e.g., the observation that ranks high in one of the two sub-models is also in the high-rank bucket in the other sub-model), but are quite different from the rank reported by the 63-gene sub-model. Please elaborate a bit more on the rationale for choosing each of these sub-models, and why different sub-models might generate contradictory results.

Indeed the two sub-models concerning the 18 and 22 genes presented more similar results, when compared to the analysis of the 63 genes sub-model. For different sub-models from the original dataset, different results are usually obtained. This is one problem that often arises when dealing with high-dimensional datasets.

To solve the fact that depending on the sub-model chosen different results are obtained, we have significantly improved the Results section by adding a new analysis based on a resampling strategy.
We fully agree with the Reviewer that the consistency and robustness of the method are key aspects that should be more deeply addressed.

Therefore, we have expanded the analysis to evaluate how the variability of the results is affected by the choice of distinct models, answering this very relevant concern raised. In particular, the question is what is the impact on the overall outlier identification if different gene subsets are chosen.

This strategy was designed to determine whether resampling the data using a subset of genes would identify the same outlier observations when compared to using the three sub-models. In this sense, the main reason to use the RP test in this context is to obtain a consensual result regardless the model chosen.

The following paragraphs were added to improve and solve the issues regarding the robustness of the RP test:

· abstract:

“Additionally, a resampling strategy was conducted to demonstrate the methods' consistency and robustness.”

· page 2, line -7:

“The aim of this work is, therefore, given a high-dimensional dataset, to find outliers (or influential observations) from different sub-models, which are obtained from distinct techniques for variable selection. The method proposed is based on the Rank Product (RP) test, a non-parametric method, to identify the outliers that are consistently highly ranked in each of the sub-models. The ovarian cancer dataset, with gene expressions as covariates, was chosen to illustrate the applicability of the proposed method. Three gene expression sub-models are presented, and the RP test is applied as a consensus or ensemble test that combines the results obtained by each model, often distinct and sometimes contradictory. Notice that each sub-model has different baselines, since for this particular dataset there is no groundtruth to start from.

Although the rank product and the deviances measures for survival models were already proposed previously in different contexts, the combination of RP-based statistical tests as a means of conferring robustness to outlier detection tasks represents the main novelty of this work.”

· page 6, line 7:

“To overcome the fact that the results obtained for each of the analysis are model-based, a sampling strategy was also implemented in order to determine whether resampling the data using a sub-model of covariates (genes) would recognize the outliers previously identified. The resampling algorithm randomly picked 1000 genes (without replacement) from the ovarian cancer dataset. The Cox regression model with elastic-net regularization was then fitted (using
glmnet), resulting in a reduced set of selected genes. In order to calculate the corresponding martingale residuals, a Cox regression is then performed on this reduced gene set (using coxph). The resulting residuals allow to sort the observations accordingly to their outlyingness level. This procedure is repeated 100 times, resulting in 100 models to feed the RP test.”

With this new analysis we have illustrated the robustness of the rank product test to particular choices of the sub-models considered, which significantly strengthens the rationale for the application of this strategy to real clinical data.

3. The RP method proposed in this work combines the results from all sub-models considered. However, with the fact that, some sub-models generate similar results (ranks), while other sub-models can result in very different ranks, it is unclear to the reviewer whether some of the sub-models might indeed create noises that may undermine the RP results, and to what extent this influences the results. The authors should clarify that and discuss the limitations of the proposed strategy.

This point is highly related with the previous comment #2. In fact, this relevant question is now addressed through the described resampling strategy, which allows to answer the main concerns raised (see previous point).

Again, we fully agree with the Reviewer that the consistency and robustness of the method are key aspects and we added the following:

· page 6, line 4

“It is noteworthy that, although we have pursued these three analyses, we can indeed include many others, for example, using different feature selection methods or prior clinical information.”

· page 8, line 24

“ To illustrate the robustness of the RP test, a resampling technique was performed as described above. The results displayed in Table (5) show that the observations considered outliers for the three different sub-models are also outlying observations for the 100 different models obtained. This includes all the observations considered outliers in Table (4). Indeed, there are individuals that consistently appear with larger residuals, irrespectively of the model. It is noteworthy that, although the genes selected in each model are different, there is a set of patients that always exhibit discrepant values for their survivals, as would be predicted by their covariates. This illustrates the robustness of the method to a particular choice of the model.“