uc3m | Universidad Carlos III de Madrid

Master Degree in Statistics for Data Science
Academic Year 2019-2020

Master Thesis

"Explained classification of high-dimensional tabular data"

Javier López

Carlo Sguera
Iñaki Úcar

Madrid Julio 2020

SUMMARY

Explaining the predictions of machine learning models in high-dimensional data is particularly challenging due to the large number of features to handle in the interpretations.

This thesis presents LASSO as an interpretable model well suited for high-dimensional data. The model is put to the test in two case studies where tabular data is classified, using a neural network as prediction model in one of them. The interpretations are examined from a global perspective (explaining the prediction model as a whole) and from a local perspective (explaining particular observations), using LIME in the later case, a method to fit linear models to local regions of the data. During the analysis, the trade-off between complexity and fidelity in the explanations is addressed and the interpretations from the different methods are compared. In all the scenarios, correlation is found between different interpretation methods but with some noticeable differences.

Dimensionality reduction through PCA is proposed as an alternative way to get a data representation with fewer elements in the interpretations. The tests with PCA components show that local interpretations tend to select fewer components with higher explained variance compared to global interpretations.

Keywords: machine learning, interpretability, LASSO, LIME, PCA

With appreciation to Carlo Sguera and Iñaki Úcar for their guidance throughout the development of this project.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

One of the strong points of machine learning is its potential to add value in almost every known domain we know. Wherever there is data, machine learning can step in and help to take better decision, from improving the personal assistant in your mobile to helping in the prevention of diseases. This is why data is sometimes referred to as the "new electricity", the prospect of a future society that will greatly depend on consuming data through AI technology in every aspect of the everyday life. This feature of ubiquity of machine learning highlights the importance of understanding how decisions are taken, specially in applications involving sensitive areas such as health, privacy or social fairness, where trust needs to be built before the technology is fully adopted. In that spirit, the EU introduced recently what is commonly known as the "right to explanation" in the GDPR regulation [1]:

(71) The data subject should have the right not to be subject to a decision, which may include a measure, evaluating personal aspects relating to him or her which is based solely on automated processing and which produces legal effects concerning him or her or similarly significantly affects him or her, such as automatic refusal of an online credit application or e-recruiting practices without any human intervention.

Being able to explain the decisions taken by ML models is not just a matter of complying with regulation, it can also be an engineering tool to detect faults that otherwise would go unnoticed. For instance, lack of diversity in the data (or just not enough data) could make overfitting detection a hard task. The examination of the decisions could help to identify the problem in a qualitative way (e.g. is your image classifier focusing on the object related to the predicted label, or is the decision actually based on something in the background that happens to be correlated to the label in the data). Lack of generalization could also be exploited in adversarial attacks, a concern in fields like cybersecurity or automated driving. It is clear that the more interpretable the model is, the easier these problems can be avoided.

1.1. Scenario

When it comes to choosing and tuning a model to predict / explain data, we always face the predictability vs interpretability trade-off, naturally impossible to avoid. Sometimes a simpler model easy to explain is good enough, in other cases we don't want to forgo the prediction performance of a more opaque model. In this project we will aim to close the gap in the predictability vs interpretability trade-off as much as we can and will focus on the following scenario:

- Classification problems with tabular data. Tabular data is very common in many

different domains.

- Data with a large number of regressors. Having to deal with high-dimensional data is specially challenging in terms of interpretability. The higher the number of independent variables, the higher the variance in the universe of explanations and therefore the more valuable would be a robust interpretation method.
- The aim of maximizing prediction accuracy. In general opaque models predict better for complex data than simple models, but it comes at a price, their outcomes are harder to explain.

1.2. Goals

To try to achieve this, we will resort to an interpretable model that is known for performing well with high-dimensional data, LASSO [2], and different ways to integrate it:

- As an interpretable model in charge of making predictions.
- As a global surrogate model that tries to explain the mechanics of an opaque prediction model (a neural network).
- As a local surrogate model that tries to explain the predictions for particular observations done by an opaque prediction model (a neural network). LIME [3] will be the method to adapt LASSO to local regions of the data, dealing with nonlinearities and sparsity.

The goals of this project are the assessment of the different implementations of LASSO to explain classifications and how they are related to each other. In particular we want to compare the results from two different perspectives, a global perspective describing the classifications in a broad way and a local perspective that takes into account the peculiarities of given instances of interest, usually the case in practice. Understanding the classifications from both perspectives and their relationship is key to get a complete vision of the explanations and gain trust in the prediction model.

After presenting with more detail LASSO and LIME, we'll study two cases applying different interpretation methods and analysing the results. Dimensionality reduction will also be introduced with the objective of explaining the classifications with fewer variables through PCA components that will replace the original features.

2. INTERPRETABLE MODELS

Interpretable models in the context of high-dimensional tabular data are models that manage to explain outcomes with a reduced number of meaningful features. LASSO and LIME fall in this category.

2.1. LASSO

LASSO (least absolute shrinkage and selection operator) is a well known linear model that adds regularization by introducing a coefficient penalization component. The amount of penalization is controlled by a parameter λ. The quantity that is minimized is:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

It also performs feature selection. An intuitive way to understand how LASSO makes the selection of features is to see it as a linear optimization problem as illustrated in figure 2.1. The vertices of the constraint shape lie on axis intersections where some variables are set to zero due to the absolute value function in the penalization.

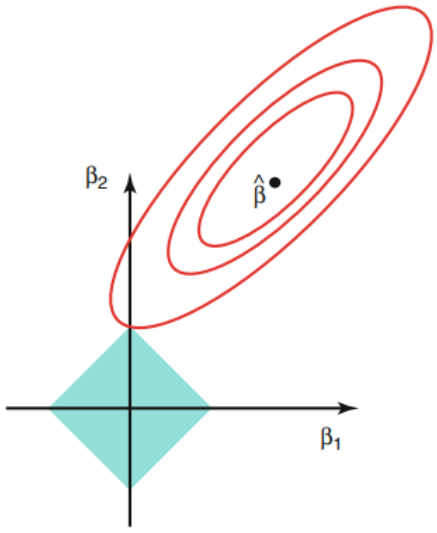

Fig. 2.1. LASSO approached as an optimization problem [4].

LASSO improves accuracy in high-dimensional data by greatly decreasing variance at the expense of introducing some bias, while being very interpretable thanks to the reduced number of selected features.

Another application of LASSO are surrogate models. A surrogate model is an interpretable model that tries to mimic a non-interpretable model (which we will call black box from now on). The surrogate model is fit to the features of the training data and the outputs of the black box, its role is not to make predictions but to provide a way to understand how the black box makes predictions. The result is an approximation of the black box that helps to identify the features in the data that drive the predictions at a global level (i.e. for no particular prediction).

When the surrogate model is fit, overfitting is not something to avoid but rather welcomed since we want to get a good approximation of the black box (which already should have took care of properly generalizing the data). However there is a trade-off between how well the surrogate model fits the black box and the degree of interpretability of the surrogate model. This is known as the fidelity vs interpretability trade-off [3].

2.2. LIME

LIME (Local interpretable model-agnostic explanations) is a model-agnostic interpretable model that aims to close the fidelity vs interpretability gap at a local level. The method was proposed in the *Why Should I Trust You?: Explaining the Predictions of Any Classifier (Ribeiro, Singh, and Guestrin 2016)* paper [3]. The intuition behind this technique is to approximate the black box locally with an interpretable surrogate model (like LASSO) by assuming that the data structure is linear around particular inputs, hence allowing for a model "tailored" to the region of interest.

The goal is to find the local model that maximizes, in a balanced way, both the local fidelity and the interpretability. The idea is formalized as:

$$g_x = \arg\min_{g \in G} \mathcal{L}(f, g, \pi_x) + \Omega(g)$$

Where, in our classification context:

x is the observation of interest for which we want to explain the classification.

g belongs to the universe G of all possible interpretable models (LASSO, decision trees, ...).

g_x is the model explaining the classification of x that is optimal in terms of both interpretability and fidelity.

$\Omega(g)$ is a function that measures the complexity of an interpretable model (for instance the number of selected features by LASSO).

f represents the black box we are trying to explain. The response of the model are the probabilities than an observation belongs to certain classes.

π_x is a function that measures the distance between any observation in the data space an x.

$\mathcal{L}(f, g, \pi_x)$ is a function that measures how unfaithful g is in approximating f in the locality defined by π_x

In the example presented in figure 2.2 below, each axis represents an explanatory variable of the data, whereas the background color represents a binary classification. The observation of interest is the big red cross. If we zoom in enough, the frontier between different categories is linear and therefore the assumption of LIME is correct and we can fit a linear model with high fidelity. More data is simulated around the observation to compensate for the sparsity of data in the "zoomed in" region.

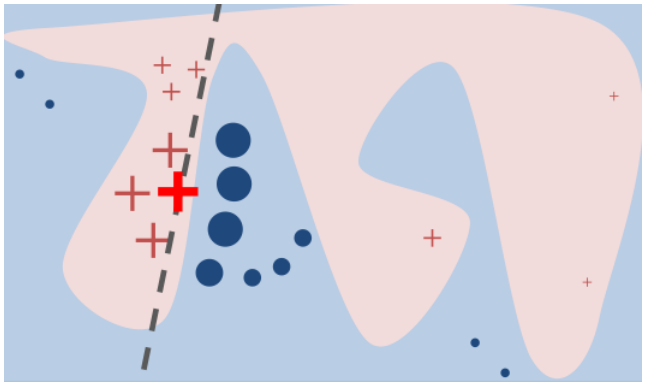

Fig. 2.2. LIME in a binary classification problem with 2-dimensional data [3].

A big advantage with LIME is that the data representation for the interpretations can be decoupled from the original data fed to the black box. We could use whatever is more convenient for the interpretations while still using the original data to get the best possible predictions, as long as we keep a mapping between both data representations. This indirection property, in addition to the model-agnostic property, makes LIME very flexible.

A usual example of interpretable data representation are superpixels in the field of image classification. A superpixel represents a segment of an image that groups pixels that are interconnected and share similar colors. As opposed to individual pixels, this representation is natural for human understanding and simplifies the identification of specific regions that could have high influence in the classification of an image. For instance, the picture in figure 2.3 could be labelled as both dog and guitar.

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Fig. 2.3. LIME explaining the predicted labels for a dog playing guitar [3].

LIME, by fitting a surrogate model in a dataset that consists on copies of the original picture where only some superpixels are left enabled and their corresponding classifications done by the black box, can identify which superpixels have an impact in the classification of the image for a particular label. The black box still uses the pixels to label the images, but the interpretation is done at a higher level, easier to understand for us than individual pixels.

3. CASE STUDIES

Two case studies are carried out. A first analysis is done on the data to set the expectations and gather the relevant information. Then the implementation of LIME and its parameters to tune is discussed, and finally the models are fit and the predictions and explanations are analysed and contrasted.

3.1. Data

The datasets used in the case studies come from fields where high-dimensional data is common: genomics and protein sequencing. Both datasets contain a large number of features, however one contains very correlated features while the other shows little correlation.

3.1.1. Genes dataset

The data consists of 801 patients with tumors occurring in different parts of the body. The tumor types covered include: lung adenocarcinoma (LUAD), breast carcinoma (BRCA), kidney renal clear-cell carcinoma (KIRC), colon adenocarcinoma (COAD) and prostate adenocarcinoma (PRAD) [5]. The absolute frequencies are shown in figure 3.1.

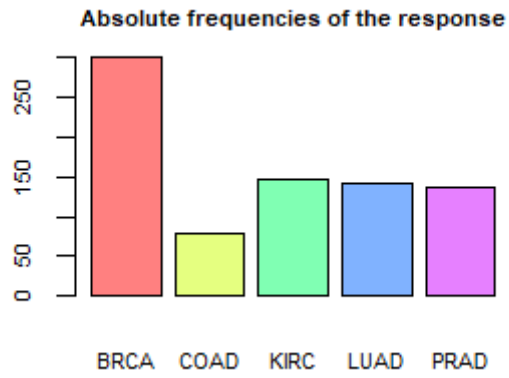

Fig. 3.1. Number of observations for each category in the genes dataset.

Among more than 20,000 RNA sequencing gene expressions, the goal is to identify which gene expressions could have been altered through mutation causing the condition.

The distribution of means shown in figure 3.2 suggests there are two categories of gene expressions.

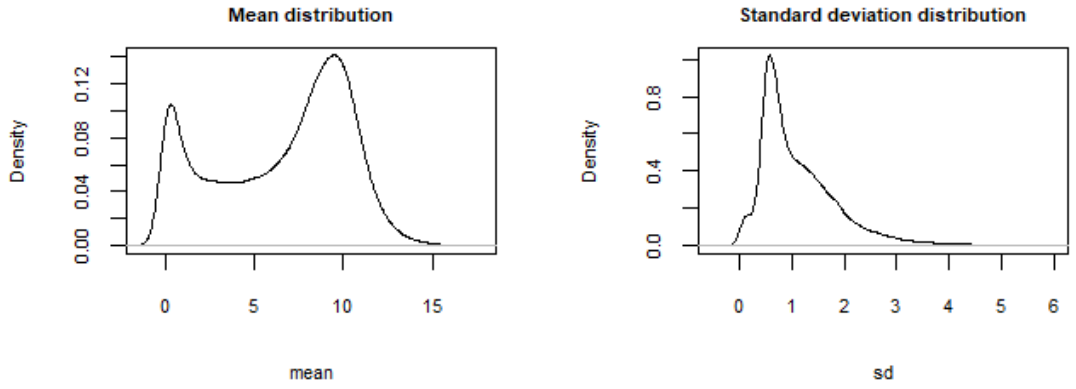

Fig. 3.2. Kernel density estimations of mean and standard deviation of genes.

The data is standardized to help with the training of the prediction model and the interpretations.

The explained variance concentration in a few PCA components shown in figure 3.3, plus having more than 20,000 gene expressions and only 5 categories suggests high multicollinearity.

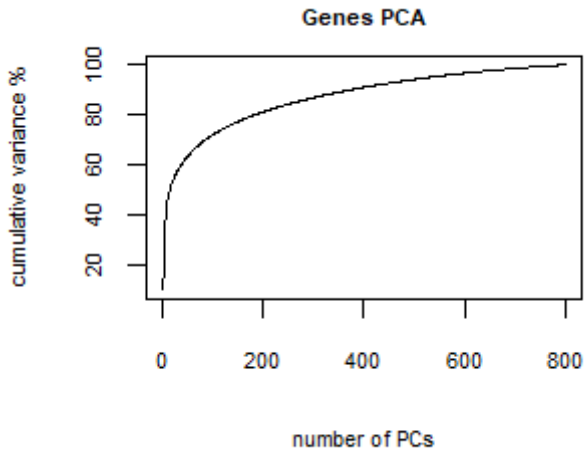

Fig. 3.3. Cumulative explained variance of PCA components in the genes dataset.

First 4 PCs look enough to classify the types of cancer despite only accounting for a third of the variability of the data as it can be observed in figure 3.4. LUAD is the category that overlaps the most with other categories.

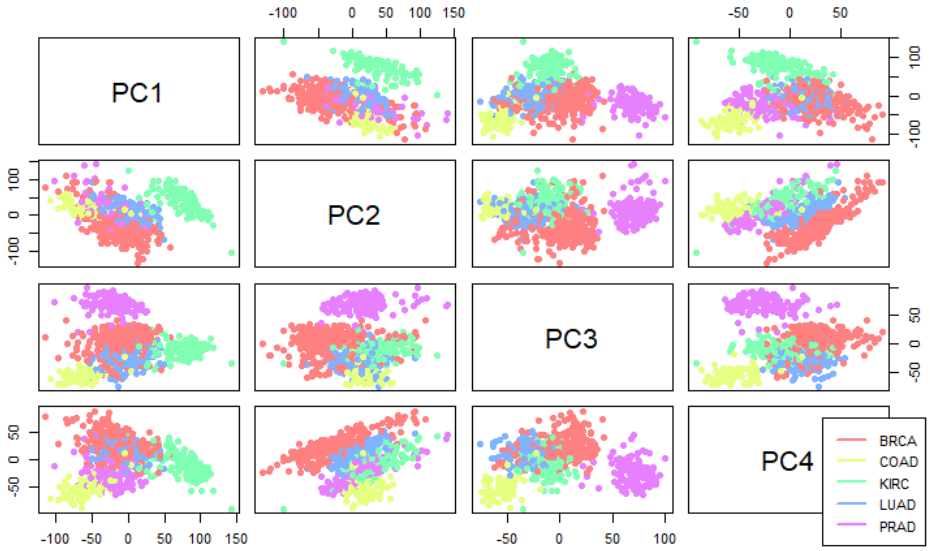

Fig. 3.4. 2-dimensional data representations with the first 4 PCA components.

3.1.2. Proteins dataset

The second dataset consists of 9972 proteins that are candidates for having anti-freezing properties. The goal is to identify which observations are antifreeze proteins (AFPs), which are important for the survival of animals in extreme cold environment conditions. The data is imbalanced, only 1.8% of proteins are AFPs.

The 841 features measure the amino acid and di-peptide composition of the proteins [6]. The variances of the mean and standard deviation of the features are low as shown in figure 3.5.

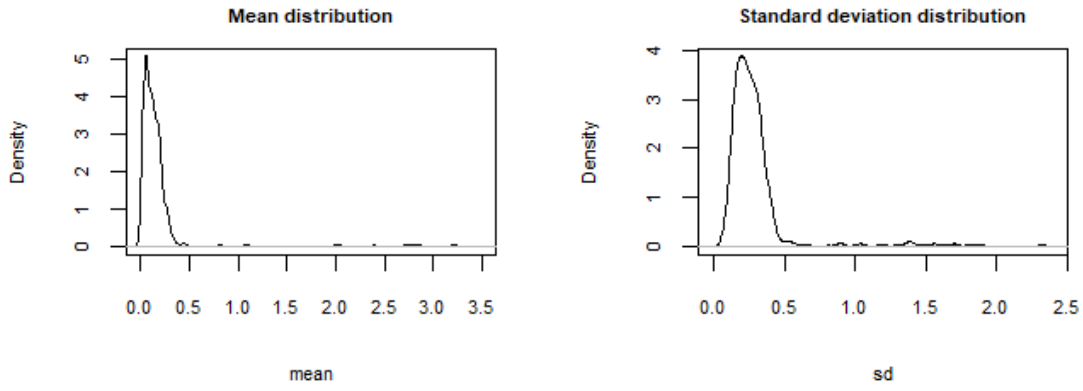

Fig. 3.5. Kernel density estimations of mean and standard deviation of protein features.

The data is also standardized to help with the training of the prediction model and the interpretations. In contrast to the genes dataset, the variance is not concentrated in a small number of PCA components but rather the opposite as shown in figure 3.6:

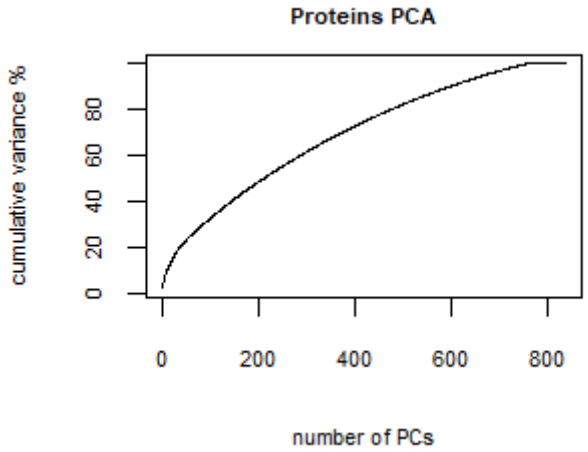

Fig. 3.6. Cumulative explained variance of PCA components in the proteins dataset.

The correlation between features is very low. In the correlation plot in figure 3.7, 50 features picked at random are almost uncorrelated.

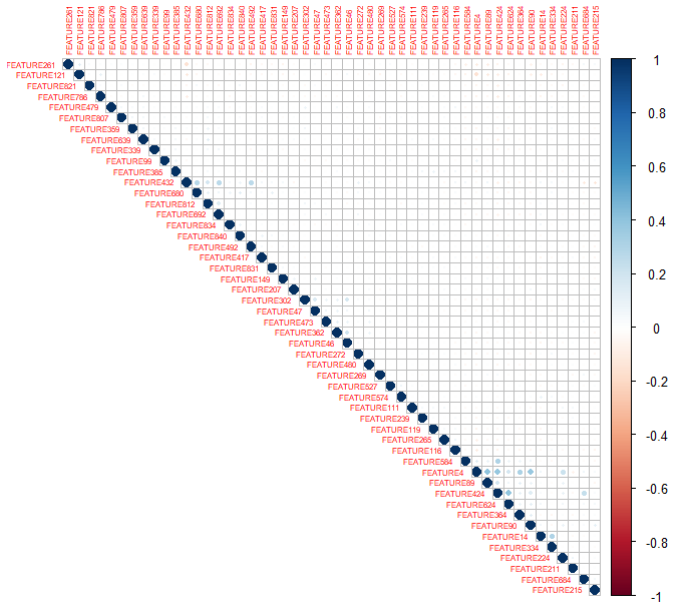

Fig. 3.7. Correlation plot of 50 features from the proteins dataset picked at random with hierarchical clustering order.

3.2. Implementation

Different implementations of LIME exist. In this project we will use the *lime* package in R, which implements the following steps:

- Before a local linear model is fit around an observation which classification we want to explain, new data points are simulated in the surroundings of the observation of interest. To achieve that, kernel density estimations are computed for each feature and sampled to simulate new data, reducing therefore the sparsity.

  The parameter to tune for this step is the number of simulated data points (*n_permutations*). The value of the parameter should be high enough to avoid high variance in the simulation between interpretations for a same observation. In this project, the value is set to 5,000 for all the interpretations, except for the case of individual genes where the value is set to 500[1]. In all the cases, the robustness of the simulation is checked by verifying that the important features selected by LIME in different interpretations are always very close for a same observation.

- Of data points simulated by sampling from the features kernel density estimations, we are specially interested in those in the surroundings to the observation of interest, as we are fitting a local model. So we need to give more weight to the data close to the instance of interest and this is done with a smoothing exponential kernel. The width of the kernel is a parameter of LIME (*kernel_width*) and probably the more difficult one to tune as shown in figure 3.8.

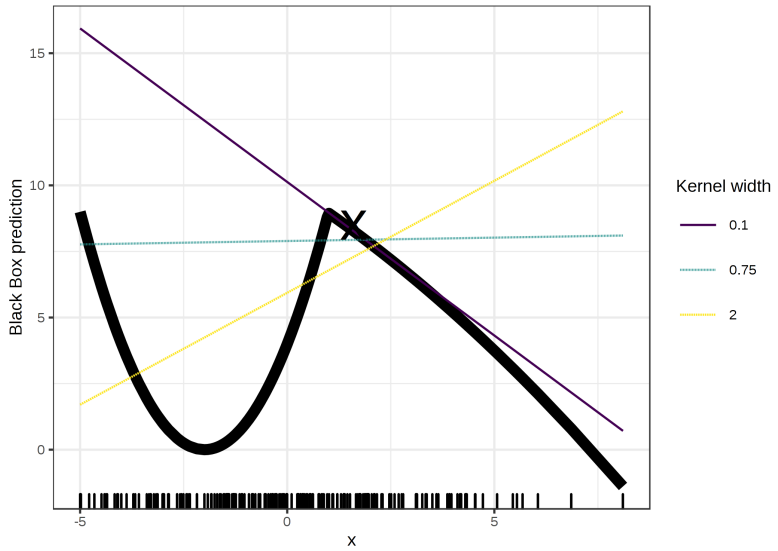

Fig. 3.8. LIME fitting in 1-dimensional data with different kernel widths [7].

[1]It takes about 90 minutes to compute the interpretations of 601 observations in the genes training data with an Intel Core i7 processor and 32 GB RAM. For each observation, a 500 permutations x 20253 genes matrix is simulated and a LASSO local model fit to it.

In this example changes in the kernel width lead to drastic changes in the local linear fit for the instance of interest (the cross in the plot). Adding more dimensions would increase the sensitivity of the width parameter even more [7].

The default value in the *lime* package is $0.75\sqrt{p}$, the more number of dimensions (p), the more the data is sparse and therefore the kernel width should be increased accordingly. The appropriate value seems to depend on the surroundings of the data point being explained so there is not a clear rule of thumb to follow for all the cases. We keep the default value for both case studies.

Another parameter is the distance function that measures the proximity between the instance of interest and the simulated data points (*dist_fun*). The default option is Gower's distance but others like Euclidean or Manhattan can also be used. We keep the default value for both case studies.

- The next step is to feed the black box with the simulated data to get its response required to fit the local model.

- With the simulated data and the corresponding predictions, a local model is fit (*feature_select*). Several models are available, we will choose LASSO for two reasons: the high-dimensionality of our datasets and to get a better comparison with the global surrogate LASSO model.
  The local LASSO model will use the the output of the smoothing kernel to give more weight to the data points in the neighbourhood of the observation to explain.

- Finally the coefficients with higher absolute values are selected (*n_features*) to explain the output (*n_labels*, 1 if we are just interested in the selected category).

3.3. Results

3.3.1. Explaining tumor classification with LASSO

The aim is to reduce the number of genes to a bunch we can handle. It turns out that by fitting a prediction LASSO model with 80% of the data and with optimal λ, the number of selected genes is indeed small, ranging between 5 and 13, depending on the category. LASSO managed to get rid of a vast amount of redundant information in form of multi-collinearity by introducing some bias (the selection of genes).

The accuracy of the model with the remaining 20% of the data is very close to 1, only two observations of category LUAD in the test data were misclassified:

		Actual				
		BRCA	COAD	KIRC	LUAD	PRAD
Predicted	BRCA	61	0	0	2	0
	COAD	0	15	0	0	0
	KIRC	0	0	31	0	0
	LUAD	0	0	0	27	0
	PRAD	0	0	0	0	25

Table 3.1. CONFUSION MATRIX IN GENES CLASSIFICATION.

LASSO classified with high accuracy and explained the data with a small manageable set of genes for each category, no opaque prediction model is required for this dataset. Something worth to mention too is how computationally efficient is LASSO, crucial when dealing with high-dimensional data.

3.3.2. Explaining tumor classification with LIME

LIME purpose is usually to explain non-interpretable models, nevertheless because the goal in this project is to explore the relationships between global explanations and local explanations, we now explain with LIME the classification done by the prediction LASSO model for each of the 640 observations in the training data. We keep the number of selected features low (10 features, in the same range as with the prediction LASSO model) for the sake of interpretability.

Note that LIME is not explaining just the classifications done by the prediction LASSO model, but its outputs (probabilities). On average, the response variance explained by LIME is high, but LIME does not fit to the inputs and outputs of the prediction LASSO model evenly for all the categories. In figure 3.9. we can find more unexplained response for category KIRC.

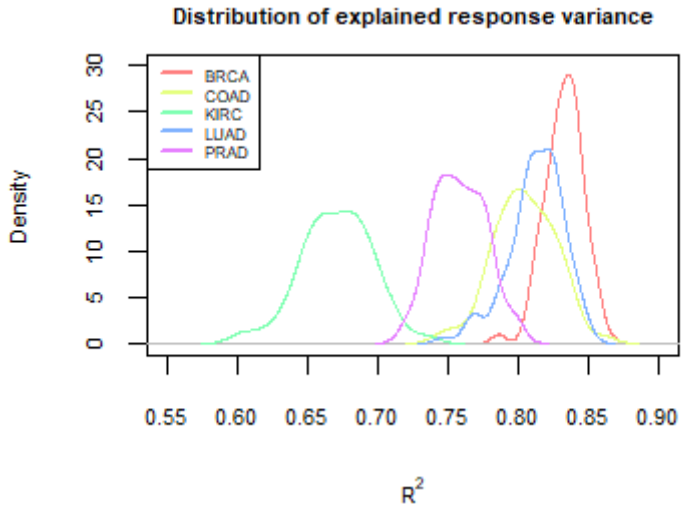

Fig. 3.9. Kernel distribution estimations of the response variance explained by LIME in tumor classifications done by LASSO.

To compare the genes selected globally by the prediction LASSO model with those selected locally by LIME, since LIME interpretations for different observations in a specific category will potentially include different selected genes, we have to make a single selection of genes summarizing all the LIME interpretations for each category.

To identify the most influential genes selected by LIME, the sums of the absolute values of the coefficients of all the selected genes are computed. Figure 3.10 displays a heatmap of the 30 most influential genes according to LIME.

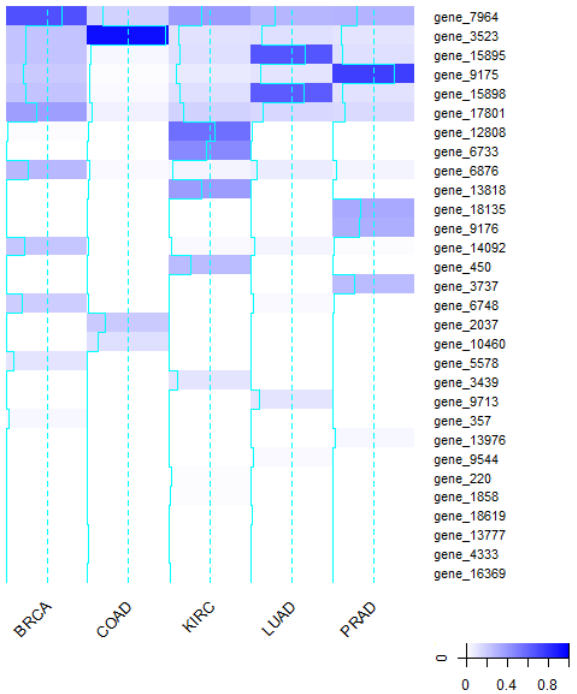

Fig. 3.10. Heatmap of the sum of the coefficient absolute values of the 30 most influential genes selected by LIME for all the observations in the training data (normalized).

When the sum of the coefficient absolute values of a gene is very close to 0, the gene came up in the explanations rarely and with negligible coefficients. Those are left out and the remaining ones are summarized with the median (the mean yields similar values).

In figure 3.11, the plot at the left displays the genes selected by the prediction LASSO model, the plot at the right displays the genes selected by LIME on aggregate to explain the classifications done by the prediction LASSO model.

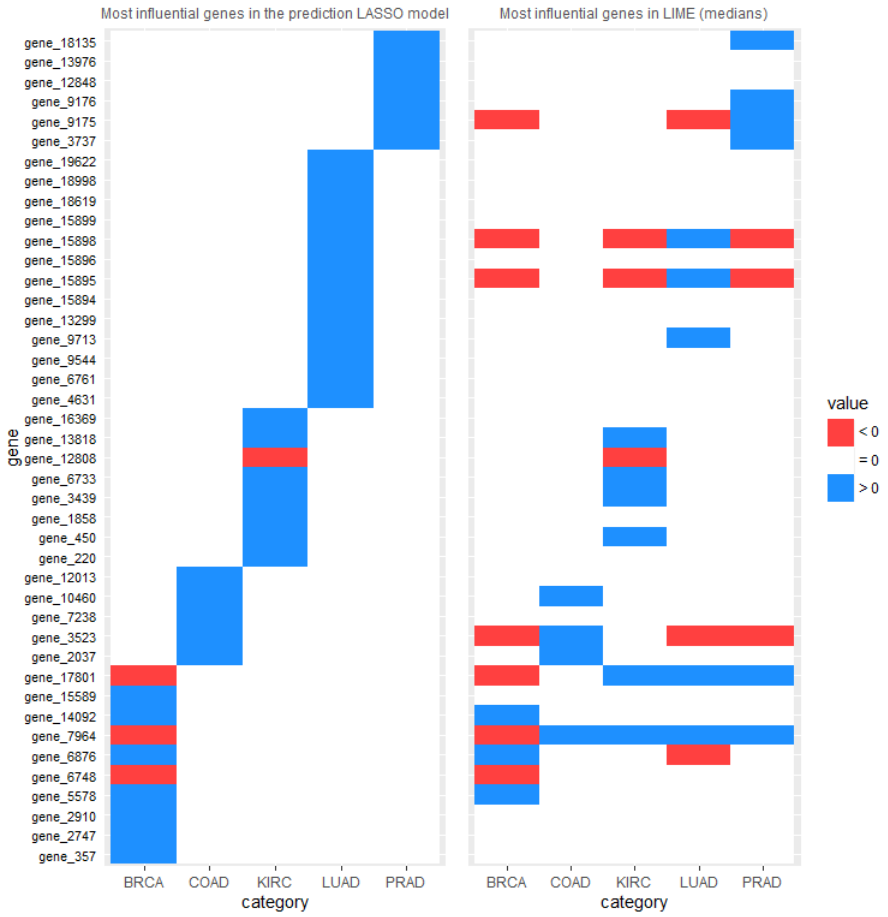

Fig. 3.11. Comparison between LASSO and LIME explanations (medians) in tumor classification. The color shows the sign of the gene coefficient.

This comparison is a good way to analyse where LIME, with around 78% fidelity depending on the category, diverges from the features that the prediction LASSO model is using to classify the tumors. LIME uses a subset (around two thirds) of the genes selected by the prediction LASSO model. The genes selected by the prediction LASSO model for each category are mutually exclusive sets, whereas in LIME explanations, some genes are important in most of the categories.

3.3.3. Explaining antifreeze protein classification with LASSO

To tackle the imbalanced data, the training data is down-sampled, ending up with 600 observations, half of them antifreeze proteins.

A LASSO model is fit to identify antifreeze proteins. It achieves 0.84 of balanced accuracy with a selection of 51 features through cross-validation.

		Actual	
		AFP	non-AFP
Predicted	AFP	150	1301
	non-AFP	31	7890

Table 3.2. CONFUSION MATRIX IN PROTEINS CLASSIFICATION WITH LASSO.

Alternatively another prediction model is fit for the classifications, this time a fully connected neural network with one hidden layer, a non-interpretable model with tens of thousands of parameters.

This model brings a small accuracy improvement of 0.02 with respect to the prediction LASSO model. Even if the improvement is modest it could be relevant in terms of research cost savings. Moreover, further improvements in the parameter tuning of the neural network are still possible while maintaining the interpretability.

		Actual	
		AFP	non-AFP
Predicted	AFP	156	1069
	non-AFP	25	8122

Table 3.3. CONFUSION MATRIX IN PROTEINS CLASSIFICATION WITH NEURAL NETWORK.

To explain the predictions of the neural network, a global surrogate LASSO model is fit with the features of the training data and the corresponding outputs of the neural network. This approach involves the possibility of adjusting the number of selected features explaining the classifications without compromising the accuracy in the predictions that are still performed by the neural network. The fidelity vs interpretability trade-off shown in figure 3.12 has to be considered in the choice.

We will target at least 80% of fidelity to explain the classifications done by the neural network, which requires 173 selected features.

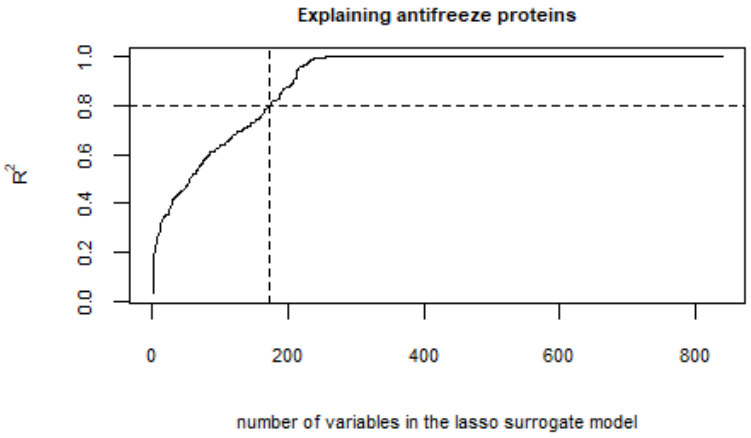

Fig. 3.12. Trade-off between how well the surrogate model explains the neural network predictions and the number of selected features in the explanations.

3.3.4. Explaining antifreeze protein classification with LIME

Global surrogate models provide a global sense of the influence of the features in the response. However if the data structure is complex, it might not explain well some predictions if the global surrogate model does not fit the black box well enough. Some particular regions of the data space could be dominated by specific features that got overlooked by the global surrogate model. In other words, the fidelity of the global surrogate model is not constant across the data space.

The classifications of the 600 proteins in the training data done by the neural network are now explained with LIME, with the same number of selected features as with the surrogate LASSO model.

The response variance percentage of the output of the neural network explained by LIME is around 0.52 on average, as shown in figure 3.13. Worse than with the global surrogate LASSO model.

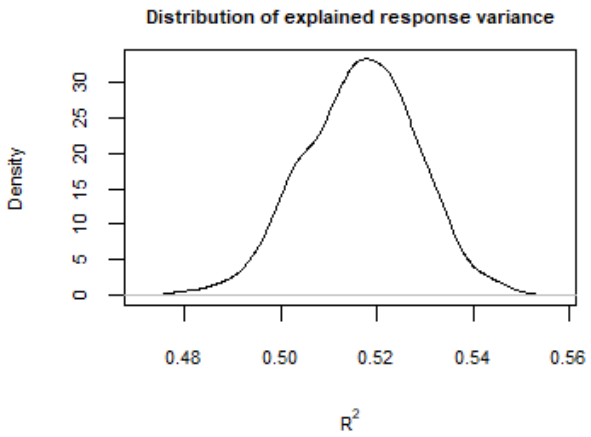

Fig. 3.13. Kernel distribution estimations of the response variance explained by LIME in protein classification.

In figure 3.14 below, the influence of the features selected by the three interpretability methods implemented (straight LASSO, global surrogate LASSO to explain the neural network and LIME to explain the neural network) are compared.

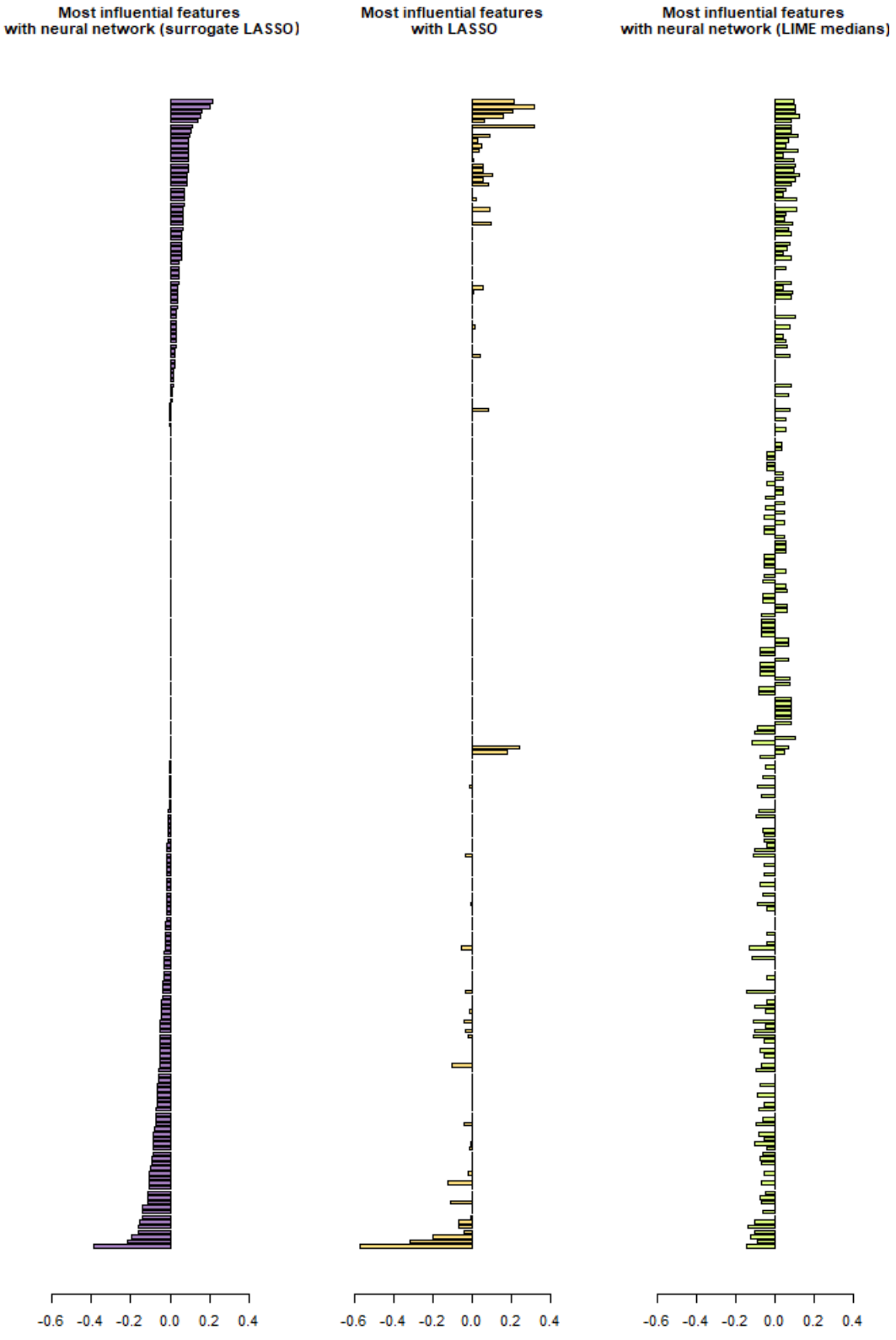

Fig. 3.14. Levels of influence of the most influential features selected by the three interpretation methods side by side. There are more than 200 features, ordered by level of influence in the surrogate LASSO model.

There is a clear correlation between the features selected by the prediction LASSO model and the features selected by the surrogate LASSO model explaining the neural network, the main difference being some features that are ignored by the neural network but are among the most 10 important features in the LASSO model prediction and vice versa. This is something expected, the neural network is able to fit nonlinearities and predicts differently from LASSO.

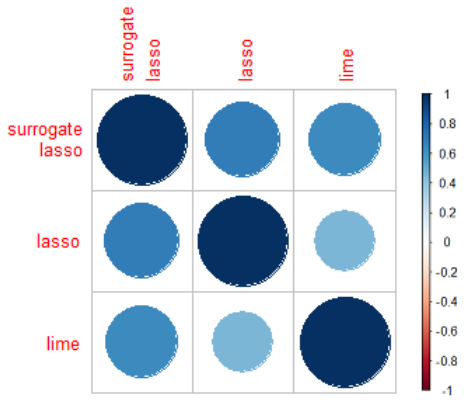

Fig. 3.15. Correlation plot of the coefficients of the selected features (normalized) between different interpretability methods with the proteins dataset.

As seen in figure 3.15, there is also correlation between the features selected in the LIME explanations and the features selected in the other two interpretation methods, however there are no dominant features and about one third of the features that came up in the LIME explanations are not considered important in the LASSO global models, the difference could be explained from the low fidelity in the features selection and the lack of homogeneity in the influence of the features across all the observations.

3.3.5. Explaining with alternative interpretable data representations

173 features in the explanations of protein classifications with the neural network is still an overwhelming number of features to handle. Next, we will try to find an alternative data representation to better explain the classification of proteins done by the neural network. The idea is similar to the use of superpixels to explain labels in image classification presented in 2.2, but for high dimensional tabular data.

A way to get a more parsimonious data representation with tabular data would be dimensionality reduction. We will use classic PCA which could help in two ways:

- To reduce the number of features to deal with while maintaining or increasing fidelity.
- To potentially get hidden meaningful features from linear combinations of the original features.

The usefulness of this method will depend on the practitioner, expert in the domain, being able to make sense of the main components involved in the explanations.

The global surrogate LASSO model is fit again with the first 150 PCA components (accounting for 41% of the variance in the data). For the same level of fidelity we targeted with the original features (0.8), the number of components required in the explanations is reduced to 115 (from 173 with the original features). We should take into account how uncorrelated are the features in the proteins dataset in particular. With other datasets the reduction in the number of required components could be much higher.

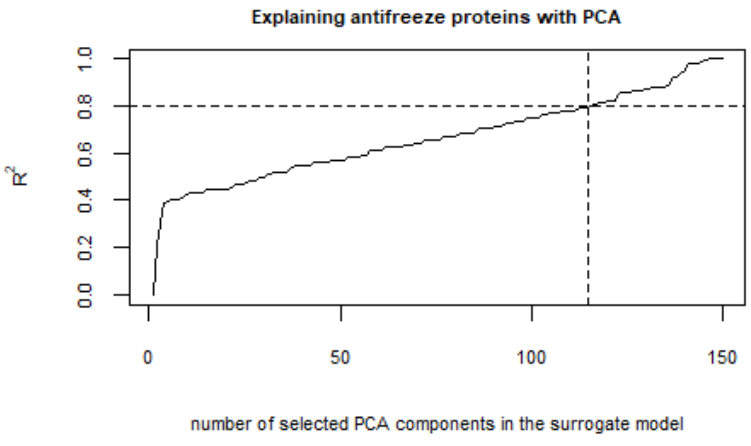

Fig. 3.16. Trade-off between how well the surrogate model explains the neural network predictions and the number of selected PCA components in the explanations.

The LIME interpretations are repeated for all the observations in the training data, this time using the first 150 PCA components as interpretable data representation and selecting 115 components to explain each observation.

Figure 3.17 shows that the response variance of the neural network explained by LIME improves from around 0.52 with the original features to around 0.78 on average.

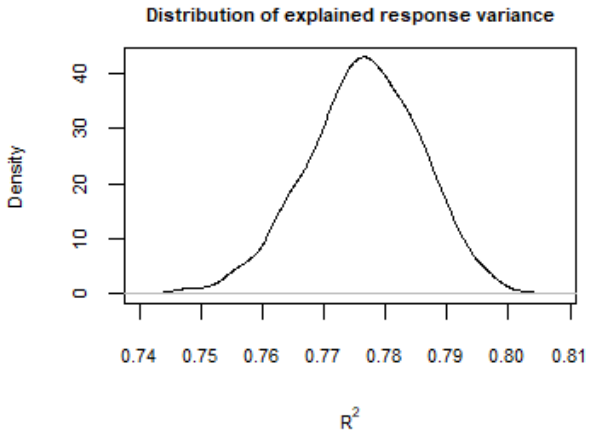

Fig. 3.17. Kernel distribution estimations of the response variance explained by LIME to explain protein classification with PCA components.

The global surrogate LASSO model and the LIME interpretations with PCA components are compared side by side in figure 3.18.

There is high correlation in the value of the coefficients, but the explanations in LIME tend to put more weight on the PCA components with more explained variance (PC1, PC2, ...) whereas the global surrogate LASSO model spreads the weights more evenly across all the components. Interestingly enough, some components with low explained variance like PC93 are very influential in the global surrogate LASSO model.

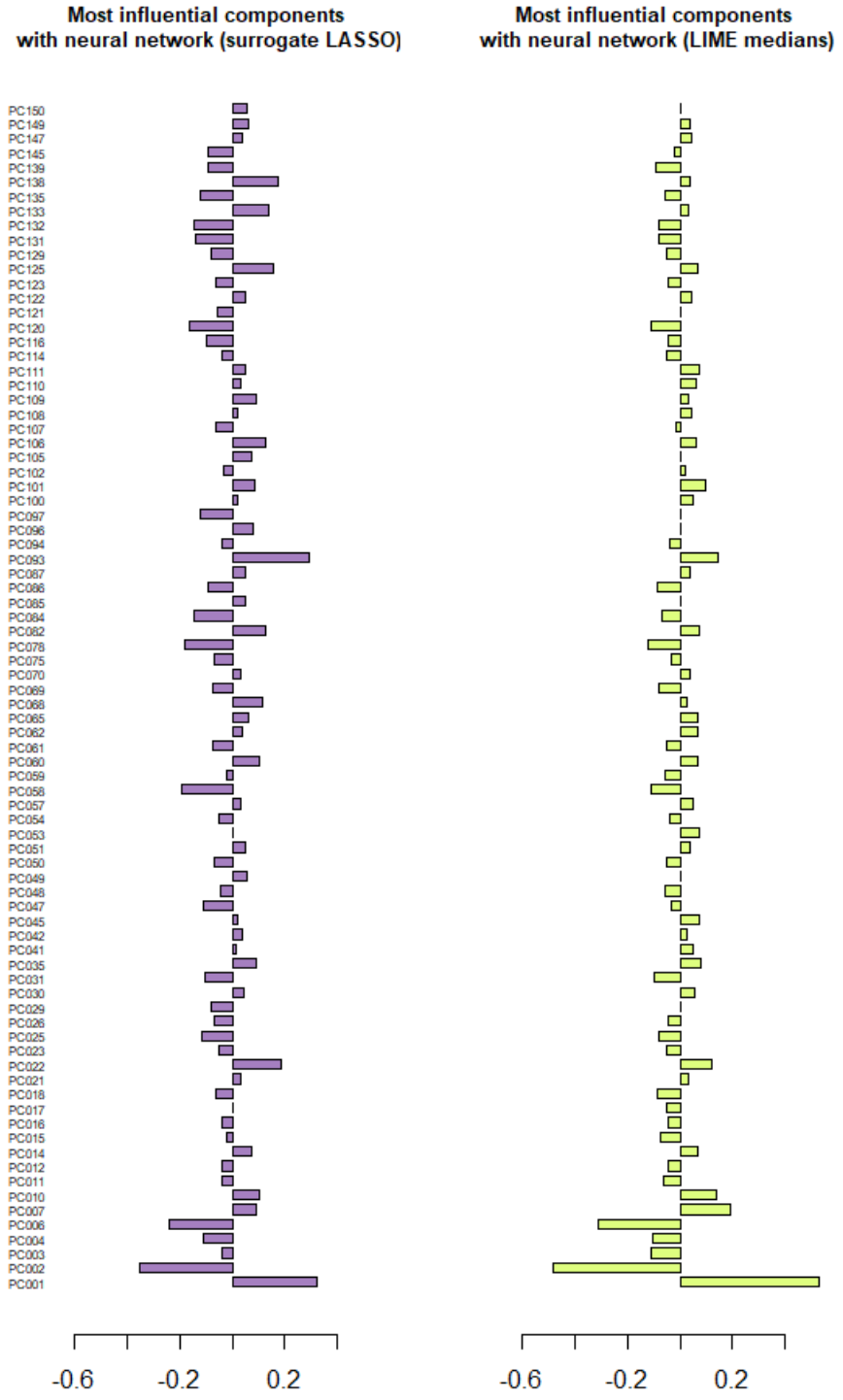

Fig. 3.18. Levels of influence of the most important PCA components selected by global surrogate LASSO and LIME side by side. The components are ordered by amount of explained variance.

The same comparison is done with the genes dataset in figure 3.19. The plot at the left shows the components selected by a prediction LASSO model with PCA components as inputs, the plot at the right shows the components selected by LIME on aggregate to explain the prediction LASSO model. The same difference is observed, LIME tends to concentrate the selection of influential features to a smaller set that corresponds to the PCA components with more explained variance.

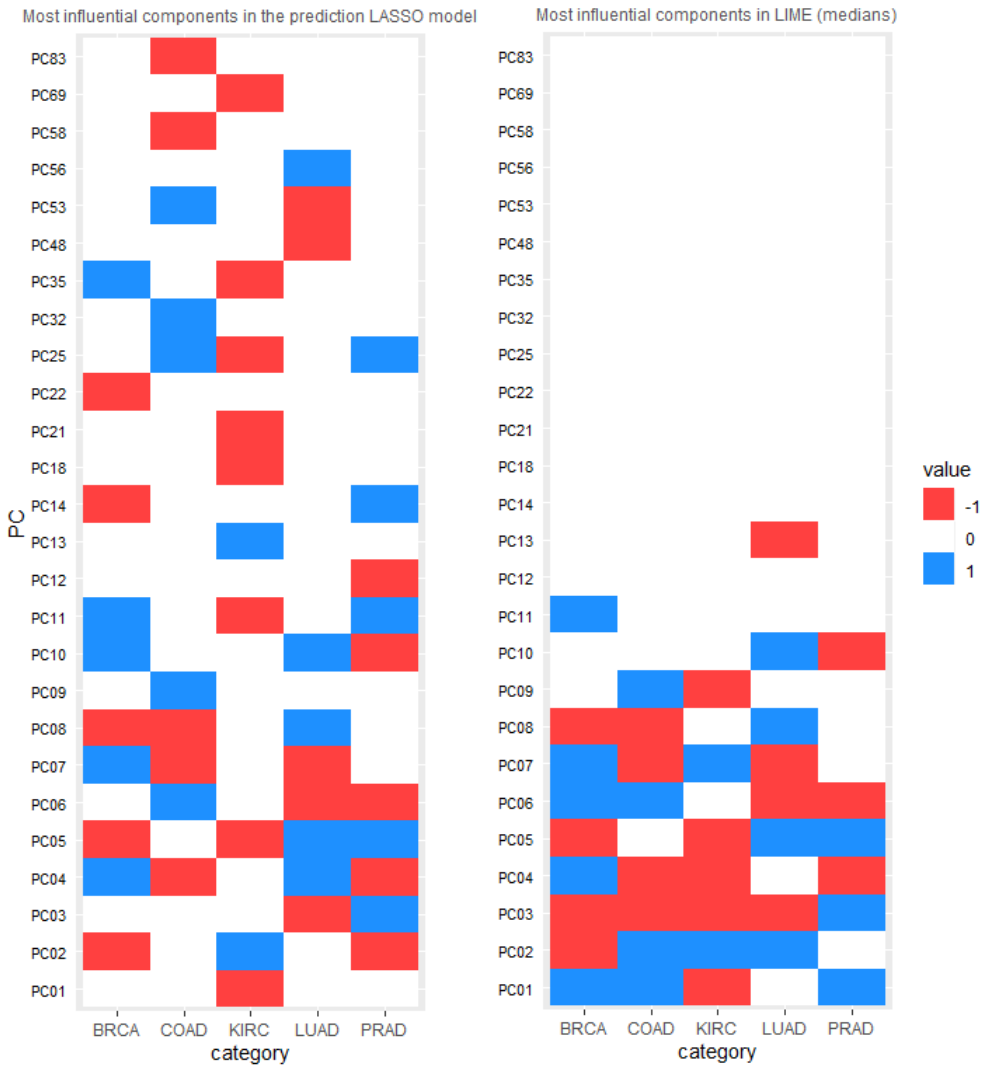

Fig. 3.19. Comparison of the selected PCA components between LASSO and LIME explanations (medians) in tumor classification.

4. CONCLUSIONS

In the two case studies carried out in this project, LASSO has proven to be a very effective model performing prediction and interpretation of high-dimensional data.

When LASSO predictions can be beaten by more sophisticated models, we can still benefit from its interpretability capabilities by means of surrogate models that perform feature selection either globally or locally, while still keeping the accuracy of the sophisticated model intact.

We compared the feature selection of two models to explain the identification of anti-freezing proteins, a prediction LASSO model, and a neural network with a surrogate LASSO model. Both models yield similar accuracies, yet the interpretations of the predictions contain some differences, some features that are very influential in one model are ignored in the other one and vice versa. Those features are key to explain the difference in how both models work.

PCA was introduced as an alternative interpretable data representation to reduce the number of elements in LASSO and LIME explanations while keeping or increasing the their fidelity. The result in both cases, with the genes dataset and the proteins dataset, show a tendency of LIME to select fewer more influential features compared to LASSO. In addition, the selected features in LIME tend to correspond to PCA components with more explained variance, a possible bias in the LIME method.

Other interpretable data representations could be more meaningful than PCA components. For instance, the features in the proteins dataset that measure the level of presence of certain amino acids and dipeptides in the molecular composition of the proteins, could be grouped into feature clusters based on a classification of the amino acids and dipeptides being measured. This way a cluster of features could represent a type of dipeptide that could be influential in the anti-freezing property of some proteins. These feature clusters could be used as an interpretable data representation for LASSO and LIME.

BIBLIOGRAPHY

[1] EU. (). "Eu general data protection regulation (eu-gdpr)", [Online]. Available: https://www.privacy-regulation.eu/en/recital-71-GDPR.htm.

[2] R. Tibshirani, "Regression shrinkage and selection via the lasso.", *Journal of the Royal Statistical Society*, 1996.

[3] Ribeiro, M. Tulio, S. Singh, and C. Guestrin, "Why should i trust you?: Explaining the predictions of any classifier.", *In Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, 1135–44. ACM.*, 2016.

[4] G. James, D. Witten, TrevorHastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*. Springer, 2013.

[5] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project.", *Nature Genetics volume 45, pages1113–1120*, 2013.

[6] S. Khan, I. Naseem, R. Togneri, and M. Bennamoun, "Robust prediction of antifreeze proteins using localized analysis of n-peptide compositions.", *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2018.

[7] P. Biecek and T. Burzykowski, *Explanatory Model Analysis: Explore, Explain and Examine Predictive Models.* 2020.