1-16-2012

Detection of Non-Regular Galaxies at High Redshift

Rafael Izbicki
Carnegie Mellon University, rizbicki@andrew.cmu.edu

Ann B. Lee
Carnegie Mellon University, annlee@stat.cmu.edu

Peter Freeman
Carnegie Mellon University, pfreeman@stat.cmu.edu

Jeffrey A. Newman
University of Pittsburgh

Detection of Non-Regular Galaxies at High Redshift

Rafael Izbicki*, Ann B. Lee*, Peter Freeman* and Jeffrey A. Newman†

January 16, 2012

Abstract

Statistical analyses of the evolution of galactic appearance (or *morphology*) play an important role in cosmology in many circumstances, e.g. when trying to infer cosmological parameters. Automatic detection of some morphologies of galaxies can be done efficiently in the nearby universe using previously developed image-based summary statistics (see, e.g., [Conselice, 2003] and [Lotz et al., 2004]). However, to perform precise statistical inference, we must be able to classify galaxies at much greater distances, and the ability to detect different morphologies using current image-based summary statistics decreases with distance. The aim of this work is to develop new summary statistics and build classifiers that allow one to distinguish galaxies that are irregulars, mergers or interactions from other types even in the farther universe. First, we propose new more informative features. Then, we apply 4 different classification algorithms to the data set. We show that the results for all of them are consistently better than when using only traditional features. Finally, we show that the classifications can be further improved if the classifier is allowed to exclude a fraction of the data set which is specified by the user. In practice, this subset could be classified by experts. We also evaluate how the new features perform when trying to classify mergers/interactions only.

1 Introduction

Cosmological theories make predictions as to how the shapes of galaxies will evolve with time, and different cosmological parameters generate different possible evolutions. With information of how the volume of each morphological category of galaxies changes with time, one can compare cosmic simulations with the real universe, which allows one to see how reasonable different values of parameters are. Hence, galaxy morphology is a valuable data source by which theories can be tested, and key cosmological parameters estimated and therefore it is important to create accurate and efficient methods of classifying galaxies. There are three primary ways by which morphologies can be inferred: parametrized models ([Peng et al., 2002]), human classification ([Lintott et al., 2008]) and supervised learning via machine using nonparametric features ([Lotz et al., 2004]). Parametrized models are usually problematic because they have assumptions that do not hold for asymmetrical galaxies such as mergers or irregulars. Classifications based on humans have the obvious drawback of being costful and time consuming. On the other hand, creating nonparametric features and using machine learning methods to perform automatic classification is a promising framework, and it is the one we use in this work.

This project is based on the data set CANDELS - Cosmic Assembly Near-infrared Deep Extragalactic Legacy Survey. CANDELS is an imaging survey of the distant Universe being carried out with two cameras on the Hubble Space Telescope. In this work, we analyze data from the GOODS-S Deep Epoch 1 field (see, e.g., [Windhorst et al., 2011]). From these data have been extracted a catalog of 6178 objects, of which 1653 are galaxies that have been visually classified by astronomers according to different criteria (e.g., the Hubble classification scheme, which classifies galaxies according to their visual appearance). Most of the galaxies ($\approx 68.97\%$) were classified by

*Department of Statistics, Carnegie Mellon University.
†Department of Physics and Astronomy, University of Pittsburgh.

3 or 4 astronomers. Classification was based primarily on the H-band[1] data. For more details on how the visual classification was done see http://candels.ucolick.org/wiki/Structure/VisClassDescript. For each of the galaxies, the data set also includes the postage stamp image and an image mask that one can use to isolate the galaxy.

In addition to the images, the data set also includes statistics that were calculated based on them. Information on such statistics can be found in [Lotz et al., 2004] and [Conselice, 2003]. The ones that were used in this work are also described in section 2. More information about the CANDELS project can be found in http://candels.ucolick.org/About.html, where one can also download the data set.

One of the criteria used to classify galaxies in the data set is whether they have undergone a (recent) merge or not. The ones that have undergone a merge are called *mergers*. Unlike regular galaxies, mergers present a clumpy profile (for instance, there may be two nuclei instead of one). This clumpiness is due to tidal forces that affect the morphologies of the original galaxies as they interact and come together. Examples of mergers from the data set can be found in Figure 7. Before merging, galaxies first pass trough interactions that also change their morphology. Merging galaxies in the nearby universe ($\approx$ 2.5 billion light years) are efficiently detected via, e.g., the comparison of two summary statistics of a galaxy image: the Gini coefficient and the moment of light statistic M20 [Lotz et al., 2004]. M20 and Gini indices have also been shown to be useful in the detection of irregular galaxies, another morphology of interest to astronomy nowadays. Irregular galaxies essentially are galaxies that present disturbed morphologies. Examples of irregular galaxies from the data set can be found in Figure 7. We note that a galaxy can be both merger/interaction and irregular at the same time, and we use the term "non-regular" to designate a galaxy that is from at least one of these groups.

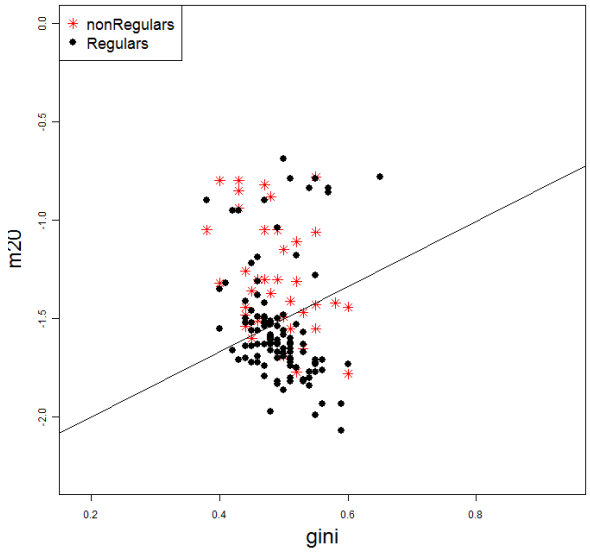


Figure 1: Gini index versus M20 index for a random subsample of CANDELS data set: separation of non-regular (galaxies that are mergers/interactions or irregulars) from regulars is not as clear as in [Lotz et al., 2004].

Although the detection of non-regular galaxies can be efficiently done in the nearby universe, these statistics alone do not allow one to detect these galaxies for farther universe as in our data set (see Figure 1, where separating non-regular galaxies from the ones that are regulars is not an easy task: the line presented is fitted according to section 3). This happens because in the farther universe galaxies get fainter, so the effect of pixel noise increases. Also, there is a loss of resolution. The deficiencies of Gini and M20 at high redshift were already predicted via simulations in [Lotz et al., 2004]. Our aim here is to develop new summary statistics that allow us to construct a good classifier for detecting non-regular galaxies.

[1]H-band is a window of width 307 nanometers centered at 1630 nanometers. It is in the infrared part of the electromagnetic spectrum.

In order to improve the previous methods (that is, using classifiers based only on the Gini index and on the moment of light statistics, M20) we created new covariates for capturing merger/interacting and irregular galaxies. They are described in section 2, which also presents a brief review of the other summary statistics that were used in our models. We then fitted several models that use the new covariates to best improve the automatic classifications. We finally checked the robustness of the new features for classification by using the different methods of classification.

Section 3 describes the different models that were fitted and the statistics that were used in order to evaluate their performances. Section 4 describes the numerical results of the fitted models. It also presents visual summaries of the results of the different models, as well as measures of importance of each covariate. We briefly describe common reasons of why some galaxies are misclassified. Moreover, we show how one can allow classifiers to interact with experts in order to improve the classifications. Finally, we apply the classifier to detect mergers/interactions only. Conclusions and future work can be found in section 5.

2 Old and new summary statistics

Let $I_{i,j}$ denote the pixel flux value at location (i, j) of image I, and let $f_{(i)}$, $i = 1, \ldots, n$ be the sorted pixel fluxes, where n is the number of pixels in the image. The following summary statistics that were already defined in other works were used here. For more on them, see [Lotz et al., 2004].

- *(Gini)* Gini coefficient.

$$\frac{1}{\bar{f}n(n+1)} \sum_{i=1}^{n} (2i - n - 1) f_{(i)},$$

  where $\bar{f}$ is the mean of $f_{(i)}$, $i = 1, \ldots, n$. For a flat surface brightness profile, $G \to 0$, while for galaxies with more concentrated light, $G \to 1$ (even if it is concentrated in different regions of the galaxy). This is a measure of concentration that originates from economics.

- *(M20)* Moment of Light.

$$log_{10} \left(\frac{\sum_{j \leq i_{20\%}} f_{(j)} [(x_j - x_c)^2 + (y_j - y_c)^2]}{\sum_{j \in mask} f_{(j)} [(x_j - x_c)^2 + (y_j - y_c)^2]} \right),$$

  where $i_{20\%}$ is the value i such that $\sum_{j=1}^{i} f_{(j)} = 0.2 \sum_{j=1}^{n} f_{(j)}$. x_c and y_c denote the center of the image. This is a measure of how spread out the light is in the image.

- *(A)* Asymmetry index.

$$\frac{\sum_{i,j} |I_{i,j} - I^{180}_{i,j}|}{\sum_{i,j} |I_{i,j}|} - B^{180},$$

  where B^{180} is the average asymmetry of the background and I^{180} is the image rotated by 180^o from the original image. That is, this index captures the asymmetry of the galaxy *with respect to a rotation of* 180^o.

To define the new summary statistics we first created a grid of different intensity quantiles (removing pixels with value 0). The following values were used: 50%, 52.5%, ..., 97%. Let $q_1, \ldots, q_m$ denote the quantiles. For each of these values, the following statistic was calculated:

- *(ratio^j)*

$$(\text{Area}_2^j / \text{Area}_1^j) * \text{Area}_2^j.$$

  where Area_i^j is the area of the i^{th} largest connected component (where 0/0 is defined to be 0) of the image $\mathbb{I}(I > q_j)$, that is, an image that is 1 for the pixels that have intensity larger than q_j in the original image and 0 otherwise.

Note that the idea of this statistic is to capture the presence of double nuclei. Its basis is $\text{Area}_2^j/\text{Area}_1^j$, which tends to be high (close to 1) in the presence of double nuclei and small (close to 0) when there is only one nucleus. We do not use this statistic alone because our experiments show that this quantity is extremely sensitive to spurious noise, specially for large values of q_j. In $ratio^j$, we multiply the term $\text{Area}_2^j/\text{Area}_1^j$ by Area_2^j. A small value of Area_2^j (especially for large values of q_j), usually indicates noise. Hence, by multiplying the original index by Area_2^j, the larger the second component is, the more weight we put on the original index. This is because we then believe that the component in fact represents a second nucleus and is not due to spurious noise.

Based on these statistics, we calculated the following new summary:

- *(maxRatio)*

$$\max_j \text{ ratio}^j.$$

Examples of this statistic for some real galaxies are shown in Figure 7. Examples 1 and 2 show galaxies with double nuclei, while Examples 3 to 6 show galaxies without double nuclei. We see that *maxRatio* is capturing the double nuclei for most cases (in the sense that it assumes higher values in these cases).

To create two other covariates, we used the mean shift algorithm [Cheng, 1995] to detect peaks of intensity. The idea is to view the normalized image as a bivariate histogram. We then use the algorithm to find the modes of the distribution of this data. For computational efficiency, we only considered pixels larger than the 80% quantile. Also, we deleted all connected components that were not the main one (that is, that had area smaller than the most important component). By doing this, we could correct problems in some of the masks. The bandwidth of the Gaussian kernel used for smoothing in the mean shift algorithm was chosen via cross-validation. When multiple modes were found in the same bin, they were merged. Based on the modes that were found, we calculated:

- *(intensityRatio)*

$$\text{intensity}_2/\text{intensity}_1.$$

  where intensity_i is the sum of the intensities in a neighbourhood (here defined as a square with 7 bins on the sides, value that minimizes the estimated risk, see section 3) of the i-th mode in the normalized image. Modes are sorted according to this sum (the first being the one in which this quantity is the largest), that is, $\text{intensity}_1 \geq \text{intensity}_2$. 0 in case of only one mode.

- *(intensitySecondLargest)*

$$\text{intensity}_2.$$

The main idea of these summary statistics is also to capture double nuclei, but this time by first finding the peaks of intensity and then measuring their importance by inspecting a neighborhood around them in two different ways. Note that here we do not need several different level sets in order to define them. Again, larger values of these statistics reflect the presence of double nuclei, while galaxies with only one nucleus usually have smaller values of both of them. Examples of these can be found in Figures 7. Modes that were found are also shown in the figures. Examples 1 and 2 show galaxies with double nuclei, while Examples 3 to 6 show galaxies without double nuclei. We see that both *intensityRatio* and *intensitySecondLargest* are capturing the two nuclei for most cases (in the sense that they assume higher values in these cases).

In order to try to capture irregularity of galaxies, we use the statistic

- *(dModeMass)*

$$d(centerMass, MSFirstMode),$$

where centerMass is the center of mass of the galaxy, MSFirstMode is the most important mode (according to the sum of intensities, as described before) found by the mean shift algorithm. Essentially, this statistic tries

to capture the asymmetry of the galaxy: if the distance between the center of mass (mean) and the mode is large, the galaxy is asymmetrical. We use the center of the level sets as an alternative way to try to find the mode. In Figure 7, one can see that in Examples 1 to 4 dModeMass is larger than in Examples 5 to 6. In fact, these are the galaxies which are non-regulars. Note that A, the asymmetry index defined above, captures only asymmetry with respect to a rotation of 180^0. On the other hand, *dModeMass* captures other kinds of asymmetries.

A summary of how these statistics perform in non-regulars and regulars can be found in Table 1, in which we present the 1st, 2nd and 3rd quantiles of the distributions on galaxies that were considered to be non-regular or not (according to majority vote, as described in Section 3). The violin plots [Hintze et al., 1998] in Figure 2 also compare these distributions by showing the kernel density estimates for each of the covariates in each of the groups as well as their boxplots. Figure 3 contains the scatter plots of the covariates.

Table 1: Statistics of covariates for non-regular (1) and regular galaxies

Variable	nonRegular	1st Quantile	Median	3rd Quantile
Gini	0	0.47	0.50	0.53
	1	0.46	0.48	0.52
m20	0	-1.71	-1.63	-1.51
	1	-1.53	-1.34	-1.07
A	0	0.11	0.15	0.19
	1	0.18	0.25	0.33
maxRatio	0	0.00	0.01	0.06
	1	0.01	0.15	1.92
intensitySecondLargest	0	0.00	0.00	0.00
	1	0.00	0.00	0.14
intensityRatio	0	0.00	0.00	0.00
	1	0.00	0.00	0.47
dModeMass	0	0.28	0.48	0.85
	1	0.99	1.56	2.43

Figure 2 indicates that the distribution of all covariates, except Gini index, seem to be different among non-regular and regular galaxies. However, we see that none of them alone are enough to allow us to distinguish between these two classes, once there is also an intersection on the support of their distributions. P-values obtained by Mann-Whitney tests were also used to compare the distribution of each covariate across the groups: in all cases they were smaller than 0.0001. It is also interesting to note that most non-regular galaxies have zero maxRatio, intensitySecondLargest and intensityRatio.

Figure 3 shows that most pairs of covariates are not highly correlated, except both features related to the intensities of the modes. The asymmetry index A and dModeMass also present some correlation.

In the next section we describe the models that were used based on these covariates in order to distinguish non-regular galaxies from the regular ones.

3 Models

In this analysis, we considered only galaxies that had at least two votes (1639 is the total remaining sample size, excluding the 8 galaxies with only one vote and 6 galaxies considered unclassifiable by all astronomers). 65% of the dataset was selected as training data set, and the remaining was used as a test dataset. The category of each galaxy was defined by majority vote, with the number of votes for non-regular being defined as

$$\text{votes for non-regular} \equiv \text{mean}(\underline{\text{votes for non-regular}}, \overline{\text{votes for non-regular}}),$$

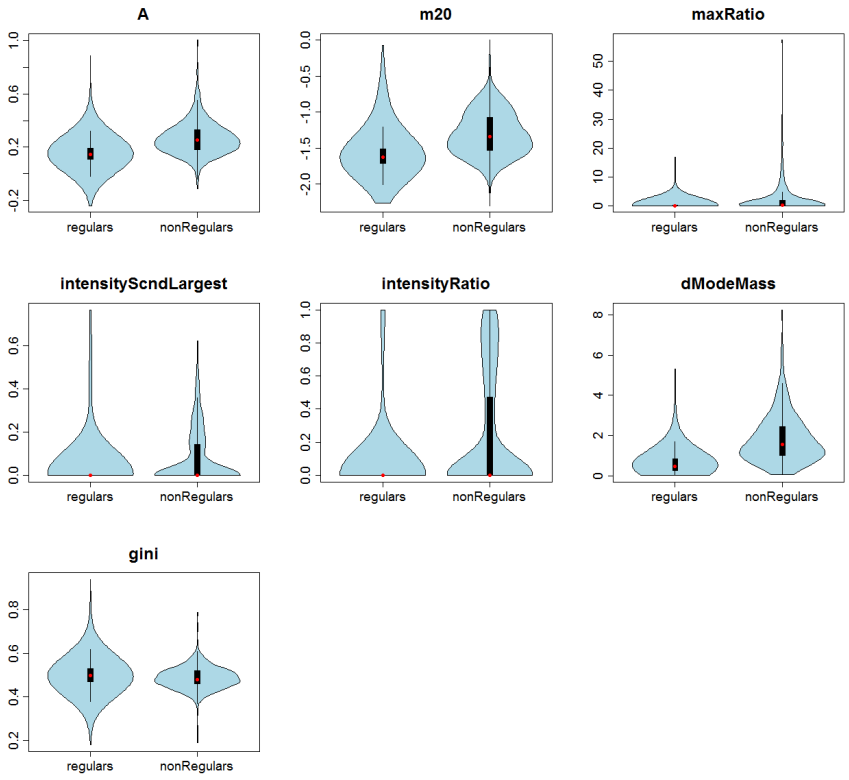


Figure 2: Violin plots comparing the covariates in the two groups

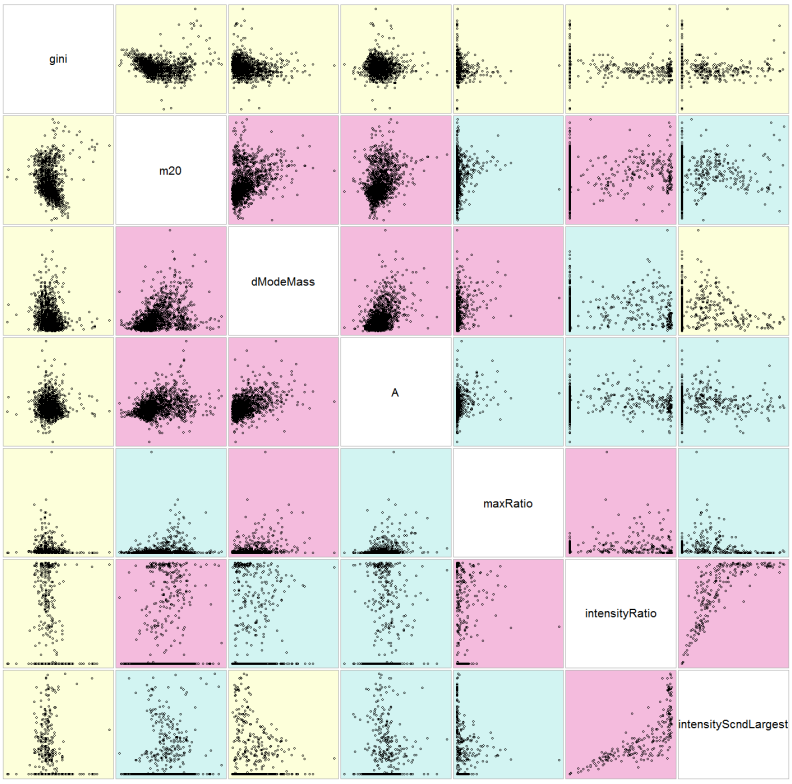


Figure 3: Scatterplots of all pairs of covariates, colored according to their linear correlation.

where

$$\underline{\text{votes for non-regular}} = \max\{\text{votes for irregular}, \text{votes for merger} + \text{votes for interaction}\}$$

and

$$\overline{\text{votes for non-regular}} = \min\{\text{votes for irregular} + \text{votes for merger} + \text{votes for interaction}, 1\}$$

are lower and upper bounds on the number of astronomers that voted for merger/interaction or irregular. Note that we do not know which astronomer voted for each category, so that we cannot know how many of the votes for irregular are also votes for merger/interaction. For the cases in which there were the same number of votes for non-regulars and regulars, a lasso (as described below) based on the other observations was applied to define its category. There were 373 (25.56%) non-regular galaxies. We note that in about 41.09% of the galaxies, at least one astronomer disagrees about if the galaxy is a regular or not.

We fitted models with the following sets of covariates:

- *(set1)* Gini, m20,
- *(set2)* Gini, m20, A,
- *(set3)* m20, m20^2, A, maxRatio, intensitySecondLargest, intensityRatio, dModeMass.

In all fitted models, covariates were first scaled. The response was the proportion of votes for non-regulars (that is, $\frac{\text{votes for non-regular}}{\text{total number of votes}}$), rather than majority vote. In this way, we try to account for the fact that we are more sure about the class of some galaxies that of others. The fitted models were the following (see [Hastie et al., 2001] for more on these models):

- (*Random Forest*) Random Forest for regression. In each tree, we allowed up to 8 leaves.
- (*Lasso*) Standard lasso algorithm. Tuning parameter chosen so as to minimize the cross-validation risk.
- (*SVM*) Support vector machine for regression. A radial basis kernel was used.
- (*PC Regression*) Principal Components Regression.

Each of these models outputs an estimate of the conditional (on the explanatory variables) probability of a galaxy being a non-regular. After fitting the models, we defined a class of classifiers by $h_c(X) = I(\hat{P}(Y = 1|x) > c)$, $c \in [0, 1]$, where X represents the explanatory variables (features) and Y is 1 in case the galaxy is a non-regular and 0 otherwise. The estimate of the risk of each of these classifiers was

$$\hat{P}(h_c(X) = 0|Y = 1) + \hat{P}(h_c(X) = 1|Y = 0)$$

that is, the sum of the estimates of the two conditional probabilities of error. Note that the risk is $P(h_c(X) = 0|Y = 1) + P(h_c(X) = 1|Y = 0)$ when the loss incurred when misclassifying one galaxy is inversely proportional to the (unconditional) probability of its class. For more details on this derivation, see the Appendix A. The estimate of the risk was obtained by using the training dataset and estimating $P(h_c(X) = 0|Y = 1)$ and $P(h_c(X) = 1|Y = 0)$ as

$$\hat{P}(h_c(X) = 0|Y = 1) = \frac{\sum_{i:Y_i=1} I(\hat{Y}_i = 0)}{\sum_{i:Y_i=1} 1}, \; \hat{P}(h_c(X) = 1|Y = 0) = \frac{\sum_{i:Y_i=0} I(\hat{Y}_i = 1)}{\sum_{i:Y_i=0} 1}.$$

After calculating this quantity for a grid of values c between 0 and 1, we fitted a kernel regression (with bandwidth=0.05) to the function $f(c) = \hat{P}(h_c(X) = 0|Y = 1) + \hat{P}(h_c(X) = 1|Y = 0)$ with the aim of smoothing it. We then selected the cutoff c that minimizes the smoothed version of f. This procedure is illustrated in Figure 6.

The previous idea of changing the usual of $c = \frac{1}{2}$ for the traditional plug in methods was adopted because in this problem the cost (loss) of classifying a galaxy as non-regular when it is regular and the cost of classifying

a galaxy as regular when it is a non-regular are not the same. It is worse to misclassify a non-regular galaxy: they are the important cases that should be detected. See [Lin et al., 2000] for more on changing the cutoff of logistic regressions and other plug in methods in order to correct for this fact.

Table 2 contains the confusion matrix of a classifier. TN represents True Negatives, while FN stands for False Negatives, FP for False Positives and TP for true positives. The measures of performance we use are the following:

- *1-Sensitivity*: proportion of non-regular galaxies that was classified as regular ($\frac{FN}{FN+TP}$).
- *1-Specificity*: proportion of regular galaxies that was classified as non-regular ($\frac{FP}{TN+FP}$).
- *Estimated risk*: sum of 1-Sensitivity and 1-Specificity (which, as explained, is what we are trying to minimize).
- *Total error*: proportion of misclassified galaxies.
- *1-PPV*: (1-positive predictive value) proportion of regular galaxies among the galaxies that were classified as non-regulars ($\frac{FP}{FP+TP}$).
- *1-NPV*: (also known as 1-negative predictive value) proportion of non-regular galaxies among the galaxies that were classified as regulars ($\frac{FN}{TN+FN}$).

Table 2: Confusion matrix of a fictitious classifier

	Predicted Regular	Predicted Non-regular
Real Regular	TN (true negatives)	FP (false positives)
Real Non-regular	FN (false negatives)	TP (true positives)

The next section presents the results of the fits for the 4 different models on the 3 different sets of covariates, as well as the analyses of them.

4 Results and Analysis

4.1 Comparison of different models

Figures 4 and 5 show the results for the fit of all models in all 3 sets of covariates in 100 different random splits training vs test datasets. Table 3 contains the median of each of the measures of errors in each model.

Table 3: Median of errors for each of the models in each set of covariates

Set	1-Sensitivity	1-Specificity	Estimated Risk	Total error	1-PPV	1-NPV
			—Random Forest—			
set1	0.28	0.27	0.55	0.27	0.52	0.11
set2	0.24	0.23	0.47	0.23	0.47	0.10
set3	0.19	0.18	0.37	0.18	0.40	0.07
			—Lasso—			
set1	0.26	0.30	0.56	0.29	0.54	0.11
set2	0.27	0.21	0.49	0.23	0.46	0.11
set3	0.17	0.18	0.35	0.18	0.39	0.07
			—SVM—			
set1	0.27	0.29	0.57	0.29	0.54	0.12
set2	0.28	0.22	0.50	0.24	0.47	0.11
set3	0.16	0.20	0.36	0.19	0.41	0.07
			—PC Regression—			
set1	0.26	0.30	0.56	0.29	0.54	0.11
set2	0.28	0.22	0.50	0.24	0.47	0.11
set3	0.17	0.27	0.43	0.24	0.48	0.07

We see that in all different classification methods, the minimum estimated expected risk is consistently obtained when using covariates from set3. The other summary statistics also present better performances for features of set3. Therefore, we see that the new covariates are able to significantly improve the traditional classifications obtained using only Gini and M20 coefficients. We also see that they are robust to the classification scheme, in the sense that for all methods we used (Random Forests, Lassos, SVMs and PC Regressions), they improve the classifications. It is worth noting that the asymmetry index A also improves the classifications based on only Gini and M20.

We now consider one classifier from the framework that showed best performance in most cases: Lasso, set3. When we fit one model (with 65% of the data randomly assigned to the training set), we have the results in Tables 4 and 5.

Table 4: Measures of performances for the fitted model using a lasso with covariates from set3

1-Sensitivity	1-Specificity	Estimated Risk	Total error	1-PPV	1-NPV
0.112	0.208	0.324	0.184	0.404	0.048

Figure 6 (a) shows the process of choosing the cutoff (in this case, 0.270) for this fit. Figure 6 (c) shows the Cross-Validated ROC curves for the three sets of covariates when using the lasso. The areas under the curve (AUC) are 0.765, 0.837 and 0.901 respectively. The third set of covariates performs significantly better than the second ($p<0.0001$). Figure 6 (b) shows a histogram on how many models each galaxy was misclassified. It shows that most of the misclassified galaxies are most of the times misclassified by all 4 models. Note that the classifiers that were used are of completely different nature and use different heuristics to perform the

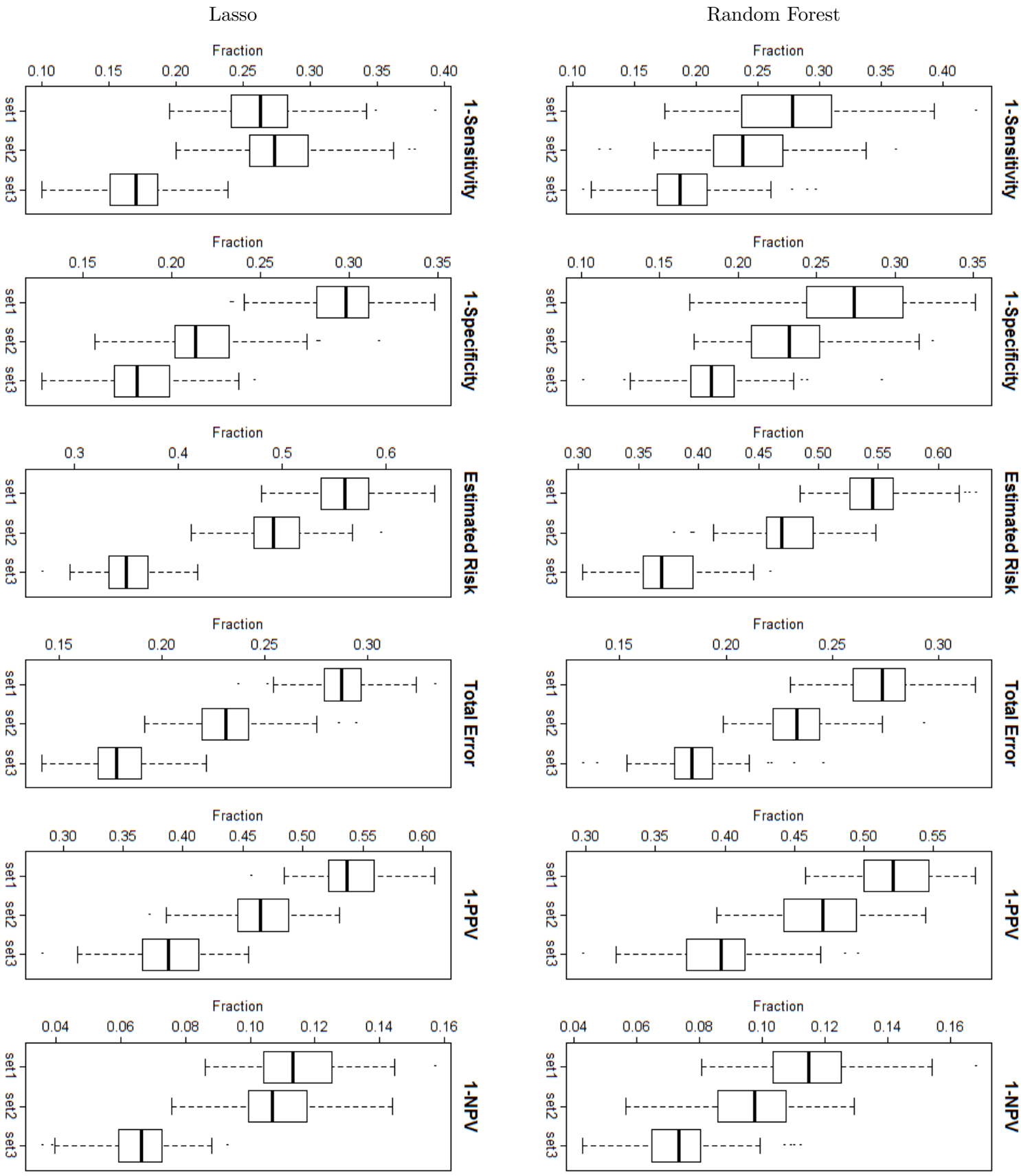


Figure 4: Boxplots with estimates of errors for different sets of covariates for Random Forest (right) and Lasso (left). Boxplots are obtained from a sample of errors using different splits training/testing datasets.

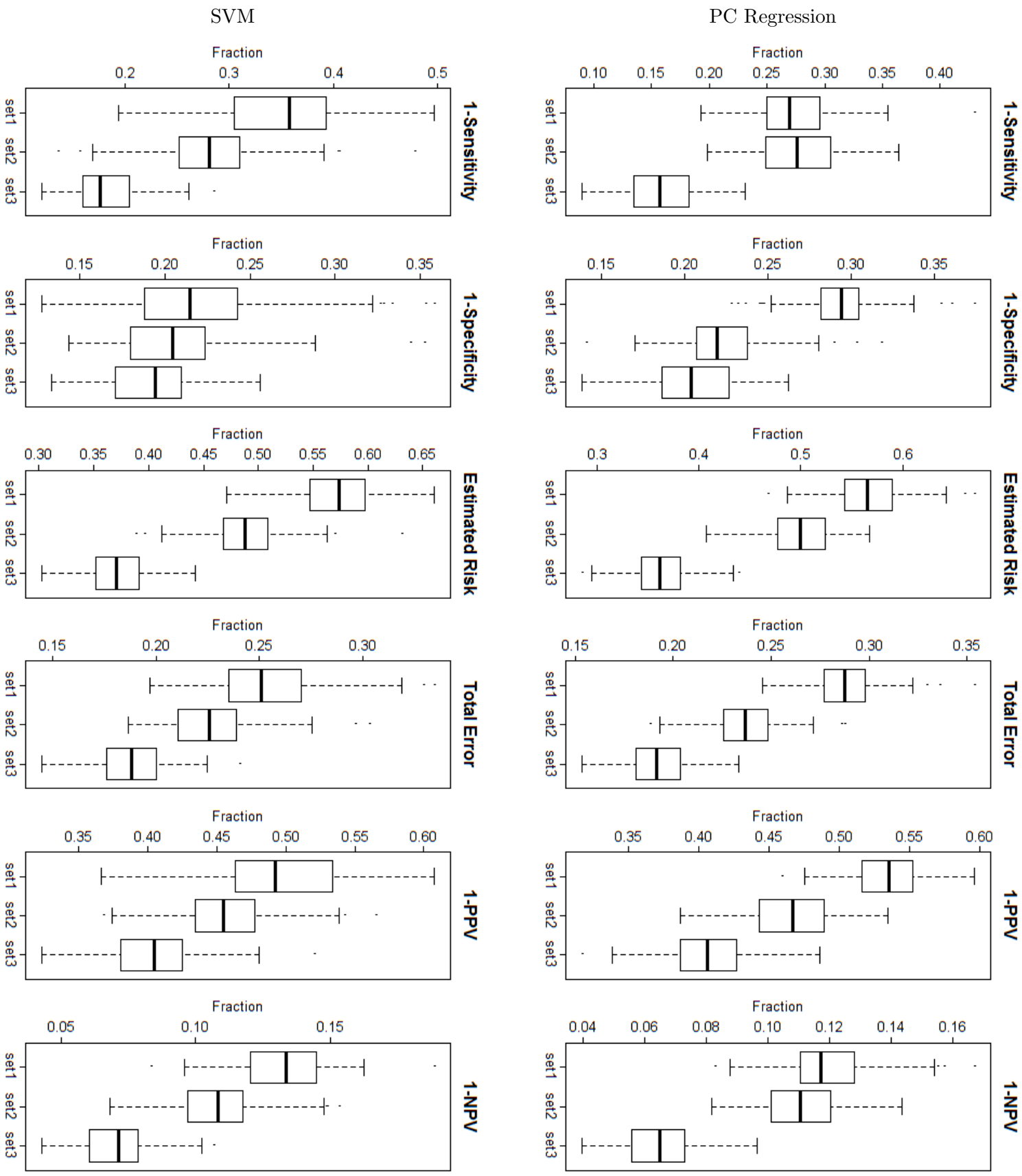
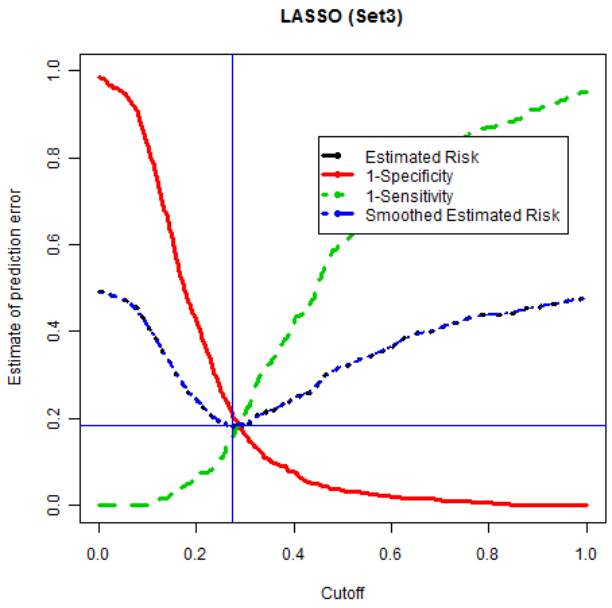
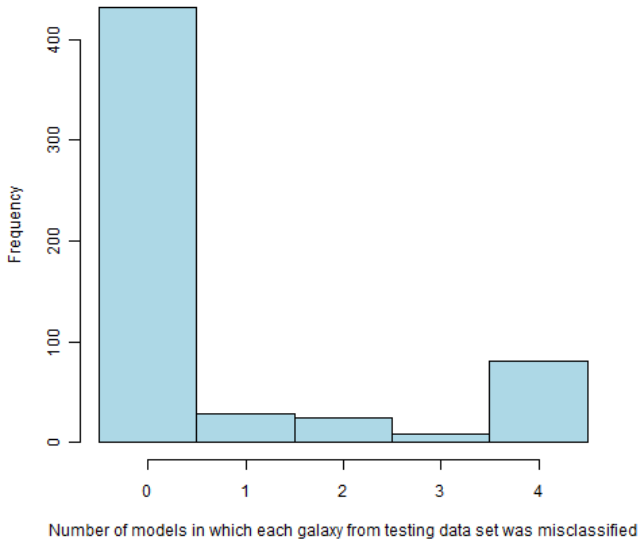


Figure 5: Boxplots with estimates of errors for different sets of covariates for PC Regression (right) and SVM (left). Boxplots are obtained from a sample of errors using different splits training/testing datasets.

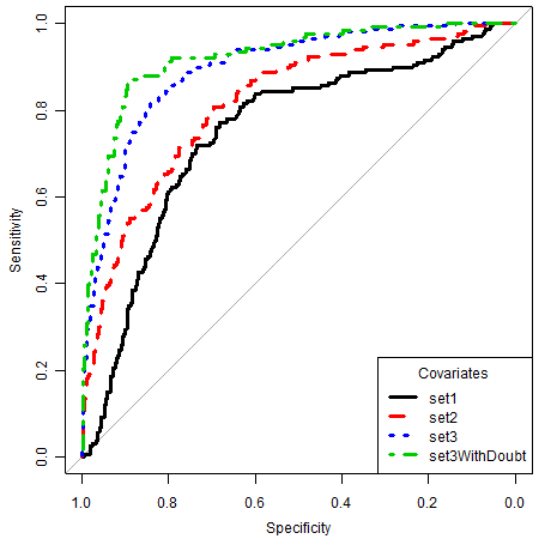
classifications. Hence, we see that either the galaxies that were misclassified are too difficult to be classified, or we need more covariates to capture their class. Another possibility is that some of the labels are not correct.



(a) Misclassification errors versus cutoff



(b) Number of models in which each galaxy was misclassified



(c) ROC curves

Figure 6: (a) Estimates of predictions errors and risk vs cutoff. Note that the curve corresponding to the risk is actually scaled (divided by 2) for visualization purposes. (b) Histogram of the number of models in which each galaxy (from test data set) was misclassified. (c) Cross-Validated ROC curves for the three sets of covariates when using a lasso. Third set of covariates performs significantly better than the others ($p<0.0001$). Set3Withdoubt is explained in Section 4.5.

4.2 Examples of performance: successful cases and failures

Figure 7 shows some examples of galaxies of the test data set in which the final classifier presented good results, with the number of votes for each morphological category, number of astronomers that voted and final probability

of non-regular. It also includes the value of some features, as well as the modes that were found when using the mean shift algorithm and the center of mass of the galaxies.

Figure 8 contains a summary of which kinds of galaxies were misclassified. The following problems are common:

- Problem in the image or in the mask (example a).
- Using the mean shift algorithm is not useful because the image is very "spread", and therefore the kernel density estimate is very wiggly and contains many local modes (example b).
- Mean shift algorithm fails in finding the modes (example c).
- Some disks are asymmetrical to a degree we cannot distinguish them from non-regular galaxies (examples d and e).
- There is some information on the other bands that is not in H band (example f). This happens specially among the false negatives.

4.3 Importance of covariates

Figure 9 contains the importance of each covariate in a fit of a random forest model. These are measured by both the percentage of increase in the Mean Square Error as well as the increase in the Node Purity [Hastie et al., 2001]. The greater the values are, the higher the importance of the covariate is.

We see from this figure that the most important feature in predicting non-regulars is dModeMass. The asymmetry index A and the maxRatio also play a very important role. The index dModeMass works better than A because it is more general: A only measures asymmetry with respect to a rotation of 180^o.

The importance of the covariates was also determined by fitting a traditional linear regression and calculating the p-values for each of the coefficients. Results are in Table 6. All covariates except maxRatio have very small p-values. These results are reasonably consistent with results from Figure 9. Note that dModeMass has a very high t value, also indicating a high importance.

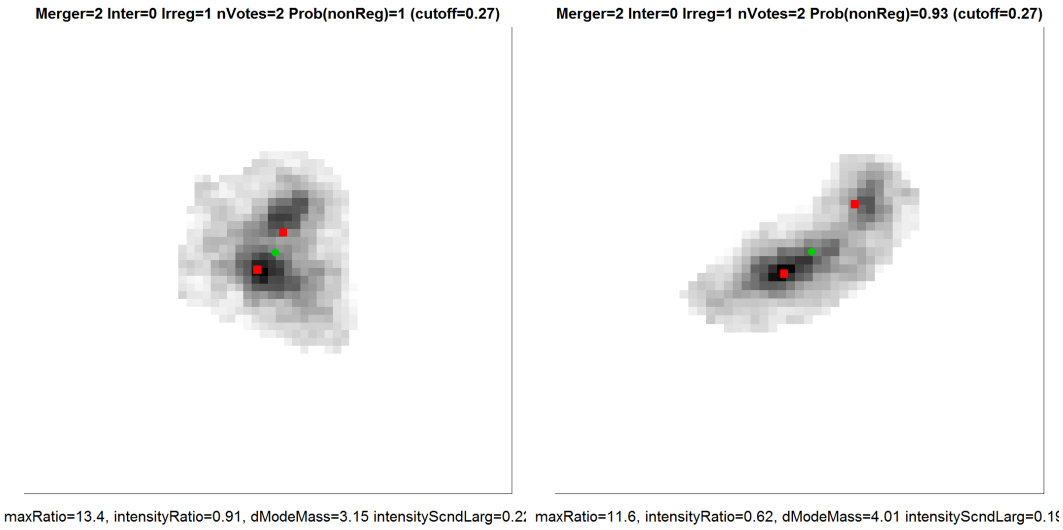
4.4 Visualization of results

Figure 10 shows a visual summary of the results. In all figures, the y axis present the predicted value $\hat{y}$ given by the lasso regression (using the covariates in set1 - Gini index and M20 - in (b) and the covariates in set3 in (a), (c) and (d)). The horizontal line refers to the optimal cutoff derived. In (a) and (b), we use the same random subset of size 150 as in Figure 1. Figures (c) and (d) contain non-regulars and regulars, respectively. One can see that Figure (a) distinguishes better non-regulars from regulars than Figure (b). In particular, the proportion of regulars that was classified as non-regulars and the proportion of regulars among the galaxies that were classified as non-regulars is much lower in the classifier that used set3 as also shown in Table 3. Also, Figures (c) and (d) emphasize that the line separates the two categories.

We note that the probability that astronomers disagree about one galaxy or at least one astronomer thinks that the galaxy is unclassifiable when it is misclassified is 1.31 times greater than the same probability when when it is correctly classified, indicating that the misclassified galaxies are in fact hard cases. Also, the proportion of disks among the galaxies incorrectly classified as non-regulars is 78.82% (while in the total sample it is 55.79%). On the other hand, the proportion of spheroids among the galaxies incorrectly classified as non-regulars is 17.64% (while in the total sample it is 31.94%). In fact, one can verify that many of the misclassifications occur in edge-on disks.

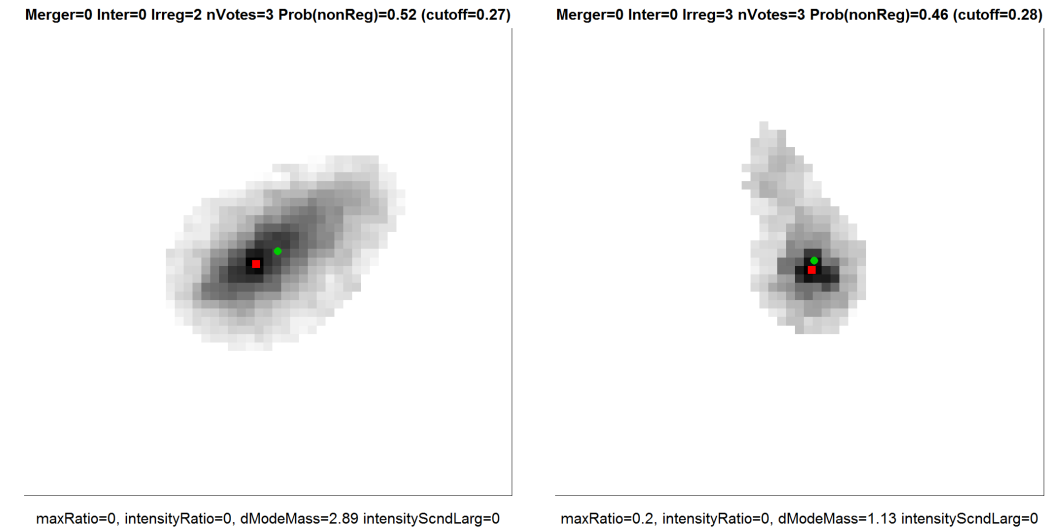
4.5 Allowing the classifier to interact with experts

Another idea that can be used to improve the classifications is the following: whenever the automatic classifier believes a galaxy is too difficult to be classified, it decided that it is better for an astronomer to classify it.



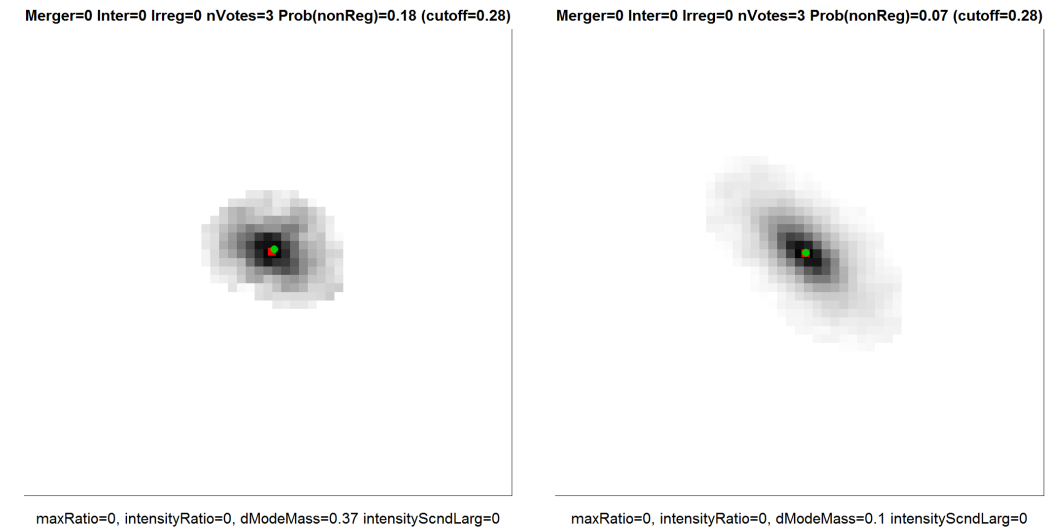
(a) Example 1: merger galaxy

(b) Example 2: merger galaxy



(c) Example 3: irregular non-merger

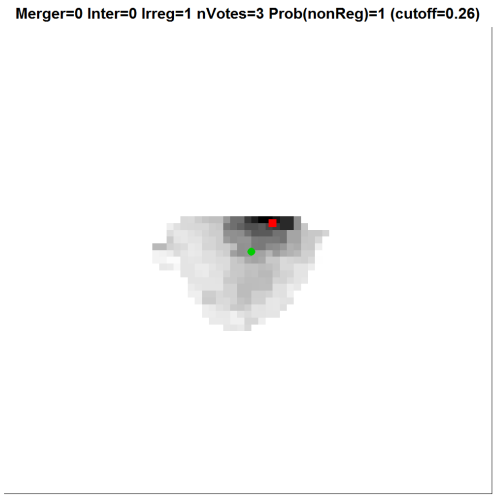
(d) Example 4: irregular non-merger



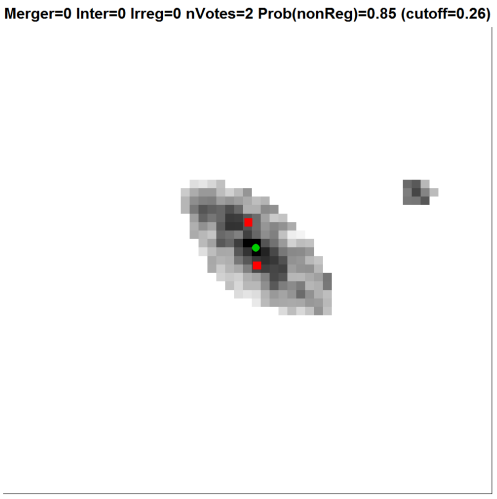
(e) Example 5: regular

(f) Example 6: regular

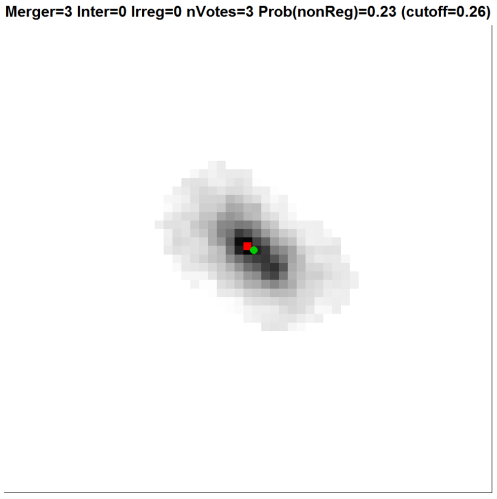
Figure 7: Examples of successful classifications in the test data set, with statistics that were calculated. Modes found by the mean shift algorithm are in red (square), and centers of mass are in green (circle).



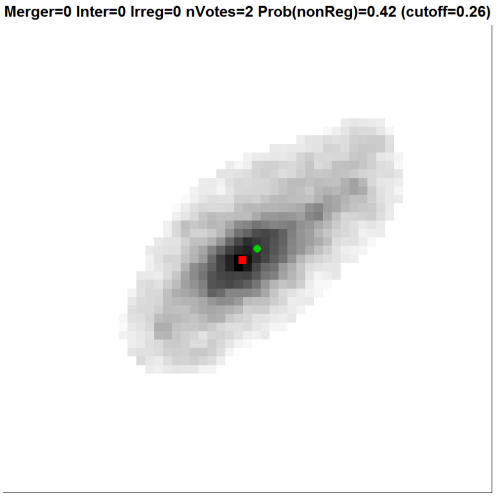
(a)



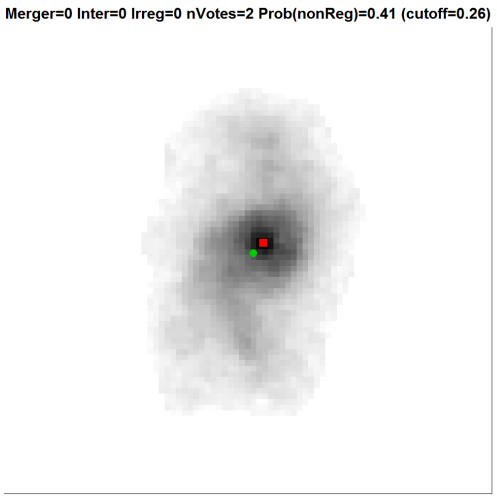
(b)



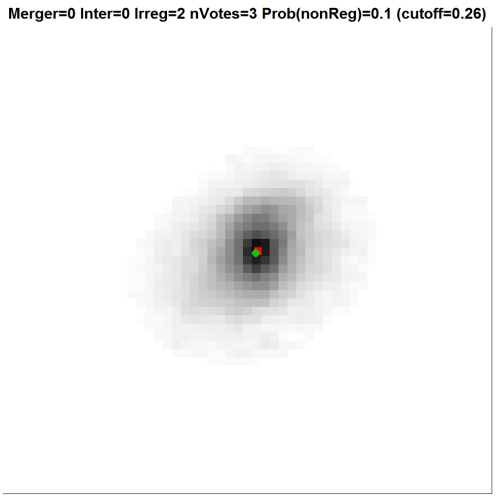
(c)



(d)



(e)



(f)

Figure 8: Examples of failure classifications in the test data set, with statistics that were calculated. Modes found by the mean shift algorithm are in red (square), and centers of mass are in green (circle). See text to a description of why the misclassifications happen.

Table 5: Confusion matrix of the lasso classified in the test data set

	Predicted Regulars	Predicted Non-Regulars
Real Regulars	337	89
Real Non-Regulars	17	131

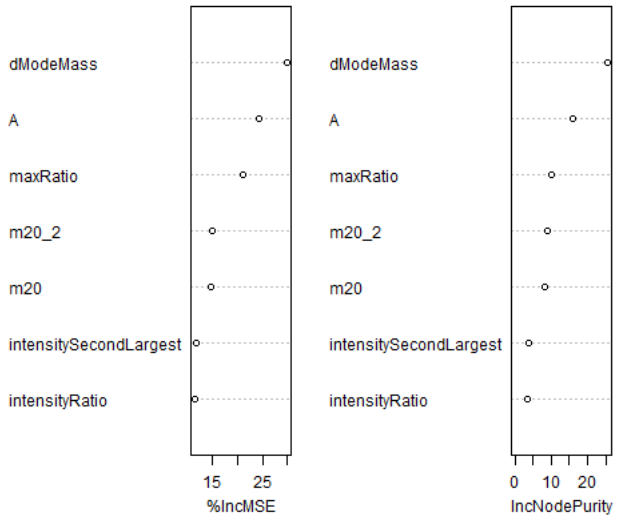
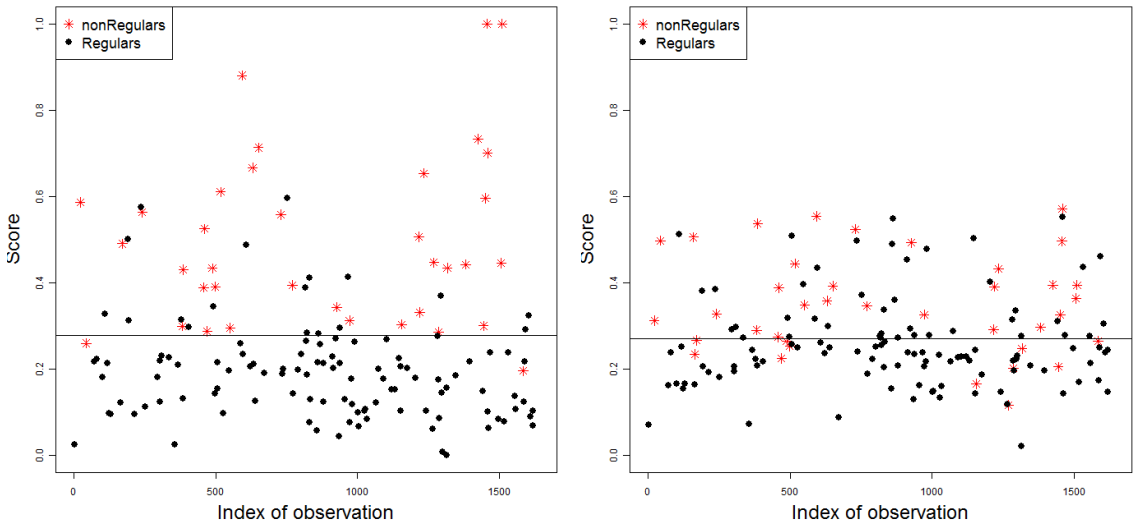


Figure 9: Importance of the covariates in set3 for detection of non-regular galaxies

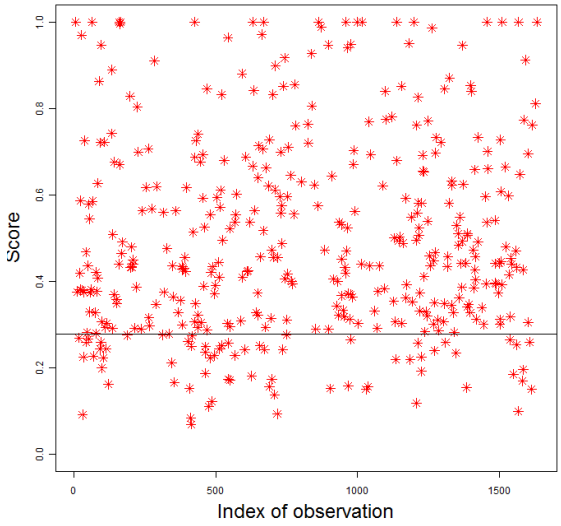
Table 6: Results of the fit of a linear regression model using the proportion of votes for non-regulars as the response

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.2754	0.0070	39.50	0.0000
A	0.0728	0.0083	8.80	0.0000
m20	-0.1352	0.0385	-3.52	0.0005
m20_2	-0.1668	0.0389	-4.29	0.0000
maxRatio	0.0133	0.0084	1.58	0.1134
intensitySecondLargest	-0.0814	0.0202	-4.04	0.0001
intensityRatio	0.1217	0.0215	5.66	0.0000
dModeMass	0.1070	0.0090	11.93	0.0000

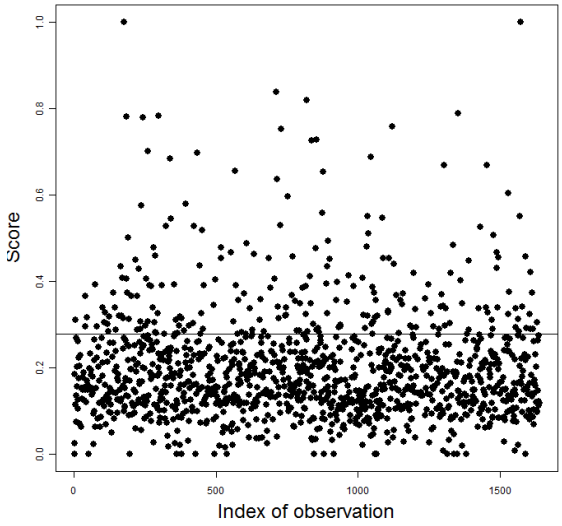


(a) Set3

(b) Set1



(c) Set3, non-regulars



(d) Set3, regulars

Figure 10: Visualization of the results for set3 and set1. Y axis is the predicted value given by the lasso regression and the horizontal line is the optimal cutoff. Figures (a) and (b) show a random subsample from the data set, (a) is a fit using set3, while (b) is a fit using set1. Figures (c) and (d) show non-regular galaxies and regular galaxies, respectively. They are both obtained using set3.

In Figure 11 we implement this idea. The data presented there is the same from Figure 10 (a). Whenever a galaxy falls in the hatched area, the classifier does not output a morphological class, but instead it says that an astronomer should look at it. The size of the region chosen in this procedure is defined according to the number of galaxies we allow the classifier not to classify. In our case, we chose it to be 20%. The derivation of such region can be found in Appendix A, and follows an idea presented in [Ripley, 1996] to derive an optimal way to do this procedure. We note that, in this case, the misclassification errors decrease to 0.156 (1-sensitivity), 0.101 (1-specificity), 0.258 (estimated risk), 0.115 (total error), 0.265 (1-PPV) and 0.055 (1-NPV). Compare this results with the ones from Table 4. Note that, in practice, this performance measures are even better, because the 20% of the data set would be manually classified would have no errors in theory. The ROC curve with the results of this method is presented in Figure 6 (c). When analysing this curve, one should take into account that there is an extra cost to this methods, namely that one person is needed in order to do the classification of some of the galaxies.

Note that this procedure could be used interactively: in principle, an expert could classify one of the difficult cases and train the classifier again with the new sample now labelled. The process then follows by finding new difficult cases until the astronomer decides to stop and use the final classifier. This is an active learning procedure similar to Uncertainty Sampling [Settles, 2009], but that also takes into account the different misclassification costs.

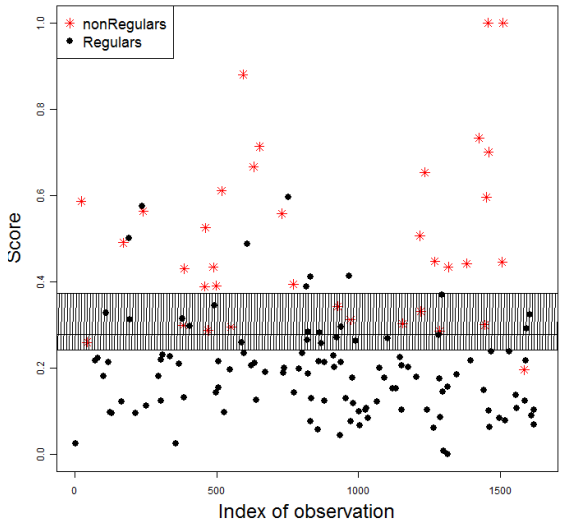


Figure 11: Same results from Figure 10 (a), but now allowing the classifier to reject to classify galaxies that it considers to be difficult. The hatched area, that contains 20% of the galaxies, indicates which of them experts should manually classify.

4.6 Performance in mergers

It is also interesting to analyze how these covariates perform when trying to classify mergers/interactions as a single class. The results are shown in Figure 12, for fits of a Lasso model. The visualization of the results for this model can be found in Figure 13. Note that the proportion of mergers in the data set is of 9.88% (162 mergers). We see that our covariates can in fact increase the performance of the automatic classification. It is also important to note that the proportion of galaxies that are considered to be mergers but are actually not mergers (1-PPV) is very high. One can see that many of these misclassified galaxies are actually irregulars.

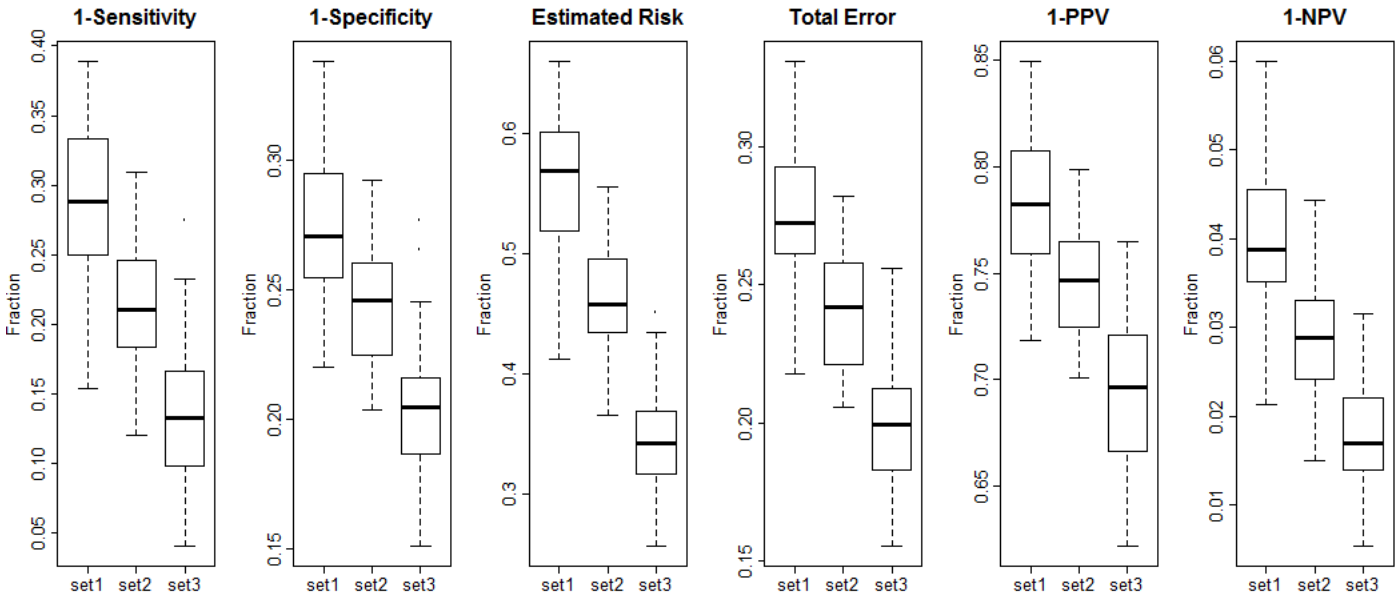
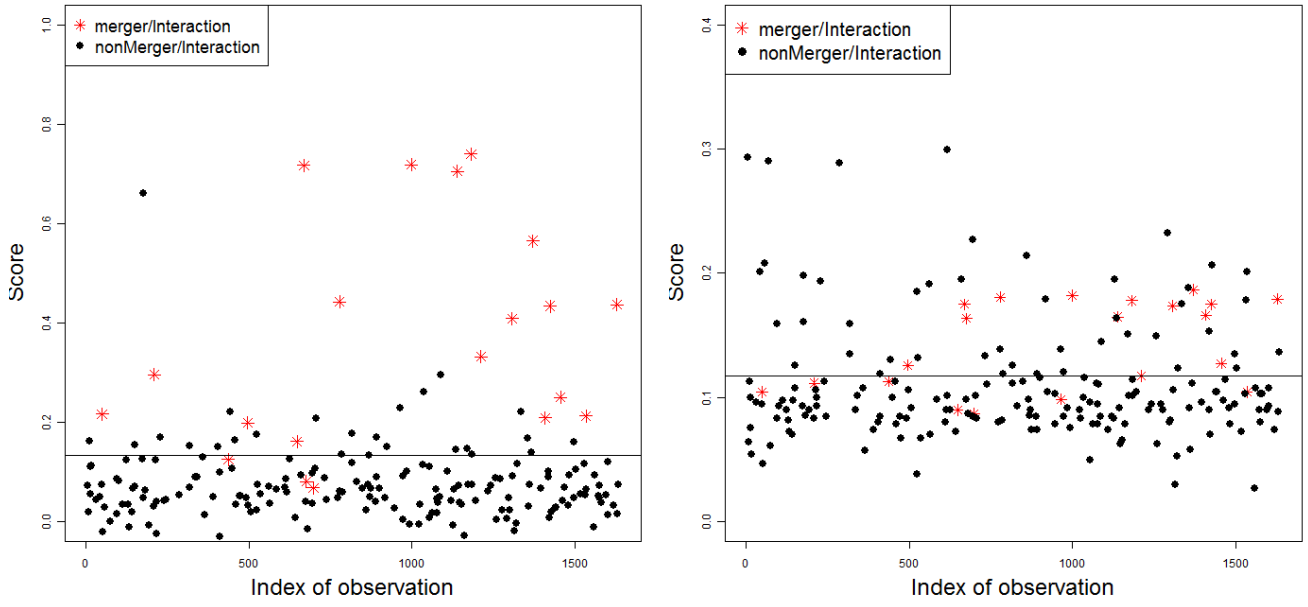


Figure 12: Boxplots for lasso models with estimates of errors for mergers/interactions. Each point corresponds to a different random split training/testing datasets.

We see in Figure 14 that maxRatio becomes the most important covariate when trying to identify mergers. Also, dModeMass and the asymmetry index are still very important. On the other hand, m20 loses its importance.



(a) Set3

(b) Set1

Figure 13: Visualization of the results for the classification of mergers/interactions only. Y axis in (a) is the predicted value of the class on the lasso regression using all data set, and for (b) using just Gini and M20 indices.

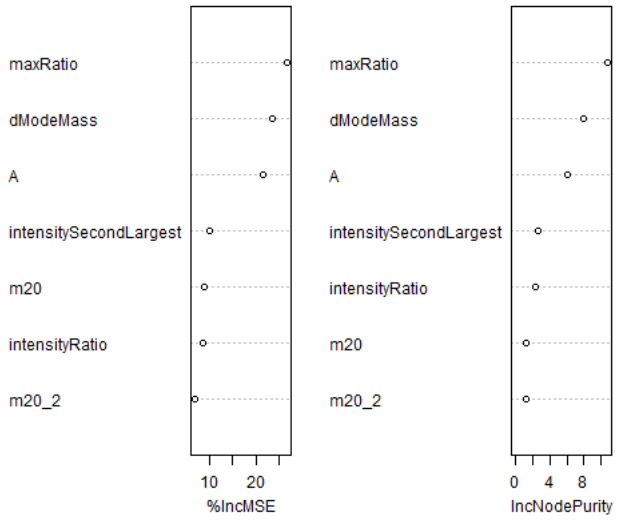


Figure 14: Importance of the covariates in set3 for detection of merger/interaction galaxies only

5 Summary and Future Directions

In this work, we first showed that two standard statistics to detection of non-regular galaxies (Gini index and second moment of light statistic) do not work well for distant galaxies (Figure 1). We then show that by introducing new covariates that capture structures that are characteristic of non-regular galaxies, one can improve the accuracy of classifiers for detecting them (Figure 10). In particular, we can get relatively high levels of sensitivity, specificity and negative predictive values. On the other hand, positive predictive values are still not very high, even though they are much better than compared to classifiers that only use Gini and M20 indexes. We also saw that the misclassified galaxies are more difficult cases, in the sense that more astronomers disagree on them.

The new covariate dModeMass was shown to be of great importance in the detection of non-regular galaxies, mainly because it better captures asymmetries than the traditional asymmetry index A. On the other hand, maxRatio, an index that detects double nuclei, was important in the detection of merger/interaction galaxies. We also showed that this improvement in the accuracy was present in a broad set of different classification methods (Section 4). Finally we found evidence that, by excluding a fraction of the data to be classified later by an expert, one can build automatic classifiers that have even higher accuracy.

In future work, we intend to:

- Incorporate information on which astronomer classified which galaxy to improve the classifications. More information on how one can do that can be found in Appendix B.
- Figure 13 (a) indicates that most misclassified mergers/interactions are near the cutoff line. Therefore, applying techniques such as the one described in Section 4.5 to allow the classifier to identify difficult cases may be useful. We intend to implement this.
- Change the size of the neighbourhood in intensityRatio and intensitySecondLargest in a way that depends on the size of the galaxy. Also, make dModeMass invariant to the size of the galaxy.
- Our classifier classifies galaxies that are mergers/interactions or irregulars well. However, we saw that the covariates are not sufficient to predict mergers alone due to the large number of galaxies falsely classified as mergers. Can we apply a different classifier once non-regulars are already selected to identify merger

galaxies? Recent papers indicate that by looking at differences between bands for the same galaxy one could detect mergers. We intend to investigate this issue.

- One could also try to improve the classifications by using a two stage classification process: first, we identify which galaxies are easy and which are difficult to classify, using a process such as the one in Section 4.5. We then classify the easy cases, and train another classifier (possibly with different covariates) for the difficult cases.
- There are many galaxies in our data set that were not visually inspected by any astronomer, and hence one cannot tell if they are mergers or not. We intend to develop the ideas presented in section 4.5 and use active learning methods in order to guide astronomers in the task of choosing which galaxies they should inspect.

References

Y. Cheng. Mean shift, mode seeking, and clustering. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 17:790 – 799, 1995. 4

C. J. Conselice. The relationship between stellar light distributions of galaxies and their formation histories. *The Astrophysical Journal Supplement Series*, 147(1), 2003. 1, 2

T. Hastie, R. Tibshirani, and J. H. Friedman. *The elements of statistical learning: data mining, inference, and prediction: with 200 full-color illustrations*. New York: Springer-Verlag, 2001. 7, 13

J. L. Hintze, K.S. Narendra, and D. N. Ray. Violin plots: A box plot-density trace synergism. *The American Statistician*, 52(2):181–84, 1998. 5

Y. Lin, Y. L. Yoonkyung, Y. Lee, and G. Wahba. Support vector machines for classification in nonstandard situations. In *Machine Learning*, pages 191–202, 2000. 8

C. J. Lintott, K. Schawinski, A. Slosar, K. Land, S. Bamford, D. Thomas, M J. Raddick, R. C Nichol, A. Szalay, D. Andreescu, P. Murray, and J. van D. Berg. Galaxy zoo : Morphologies derived from visual inspection of galaxies from the sloan digital sky survey. *Monthly Notices of the Royal Astronomical Society*, 389, 2008. 1

J. M. Lotz, J. Primack, and P. Madau. A new nonparametric approach to galaxy morphological classification. *The Astronomical Journal*, 128(1):163, 2004. 1, 2, 3

C. Y. Peng, L. C. Ho, C. D. Impey, and H. W. Rix. Detailed structural decomposition of galaxy images. *The Astronomical Journal*, 124(1):266, 2002. 1

D. R. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press, 1996. 18, 22

B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009. 18

R. A. Windhorst, S. H. Cohen, N. P. Hathi, P. J. McCarthy, R. E. Jr Ryan, H. Yan, I. K. Baldry, S. P. Driver, J. A. Frogel, D. T. Hill, L. S. Kelvin, A. M. Koekemoer, M. Mechtley, R. W. O'Connell, A. S. G. Robotham, M. J. Rutkowski, M. Seibert, A. N. Straughn, R. J. Tuffs, B. Balick, H. E. Bond, H. Bushouse, D. Calzetti, M. Crockett, M. J. Disney, M. A. Dopita, D. N. B. Hall, J. A. Holtzman, S. Kaviraj, R. A. Kimble, MacKenty. J. W., M. Mutchler, F. Paresce, A. Saha, J. I. Silk, J. T. Trauger, A. R. Walker, B. C. Whitmore, and E. T. Young. The hst wfc3 early release science data: Panchromatic faint object counts for 0.2-2 μm wavelength. *The Astrophysical Journal Supplement Series*, 193(2), 2011. 1

A Loss functions and risk

Let l_1 be the cost (loss) of classifying a galaxy that is a non-regular as a regular and l_0 be the cost (loss) of classifying a galaxy that is a regular as a non-regular. Also, denote by p_0 (p_1) the proportion of regular (non-regular) galaxies in the population.

Given a classifier $h : X \to Y$, the expected loss function (risk) for a new observation X is

$$R(h) = l_1 P(h(X) = 0, Y = 1) + l_0 P(h(X) = 1, Y = 0) = l_1 p_1 P(h(X) = 0|Y = 1) + l_0 p_0 P(h(X) = 1|Y = 0),$$

where Y is 1 if the galaxy is a non-regular.

If we consider that $l_1 = 1/p_1$ and $l_0 = 1/p_0$, the risk is given by

$$R(h) \propto P(h(X) = 0|Y = 1) + P(h(X) = 1|Y = 0).$$

In our problem, if we estimate p_0 and p_1 by the proportions in our sample, we get $l_0 = 1.343$ and $l_1 = 3.911$. Hence, these values agree with what is argued in section 3: it is worse to misclassify a non-regular than a regular galaxy. Therefore, this motivates the risk used in that section.

The Bayes classifier classifies a galaxy as non-regular if, and only if,

$$P(Y = 1|x)l_1 > P(Y = 0|x)l_0 \iff \frac{P(Y = 1|x)}{P(Y = 0|x)} > \frac{l_0}{l_1} \iff P(Y = 1|x) > \frac{l_0}{l_0 + l_1}.$$

Using the loss functions that were proposed here, we have that $\frac{l_0}{l_0+l_1} = p_1 = 0.256$. Rather than using this value, we decided to use the value that minimizes the empirical risk, however we note that for most cases they were close to each other. We denote the optimal cutoff by c^*.

[Ripley, 1996] derives the Bayes classifier when one is allowed to classify galaxies in the category "Doubt", that is the classifier can refuse to classify certain galaxies. According to this classifier, one should classify a galaxy with covariates x as

$$\begin{cases} \text{Regular, if } P(Y = 1|x)l_1 < P(Y = 0|x)l_0 \text{ and } P(Y = 1|x)l_1 < d \\ \text{Non-Regular, if } P(Y = 0|x)l_0 < P(Y = 1|x)l_1 \text{ and } P(Y = 0|x)l_0 < d \\ \text{Doubt, if } d/l_1 < P(Y = 1|x) < 1 - d/l_0 \end{cases}$$

where d is the cost of a misclassification. In this spirit, we used the same rule, but using $l_1^* = \frac{1}{c^*}$ and $l_0^* = \frac{1}{1-c^*}$. By doing this, we recover the original classifier we proposed when $d \longrightarrow \infty$. Figure 15 shows how, by doing this procedure, the risk decreases while the proportion of doubts increases.

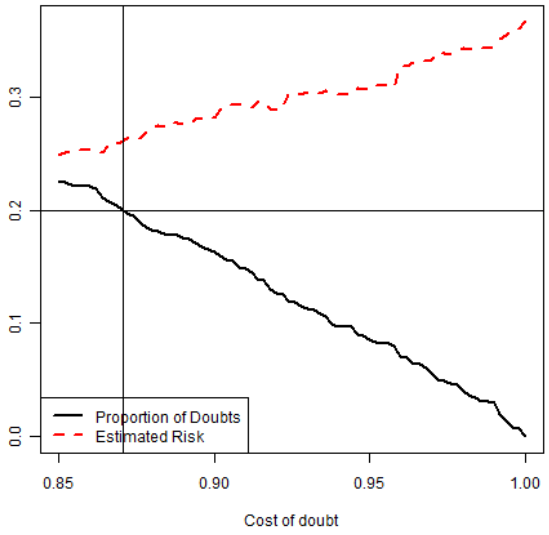


Figure 15: Proportion of galaxies that are not classified and misclassification errors (on test data set) versus estimated risk on test data set.

B Taking into account divergence of astronomers

B.1 Model to determine the classes

If we have the identification of who the votes are from, we could take this into account in order to decide which galaxies are non-regulars. One way to do it is the following:

Let Y_{ik} denote the class of galaxy i attributed by astronomer k. Let Z_i be the (latent) variable that is 1 if galaxy i is a non-regular and 0 otherwise. We make the following assumptions:

- $Y_{ik}|Z_i = 1 \sim Ber(\frac{exp\{\alpha+\alpha_k+\gamma_i\}}{1+exp\{\alpha+\alpha_k+\gamma_i\}})$.
- $Y_{ik}|Z_i = 0 \sim Ber(\frac{1}{1+exp\{\alpha+\alpha_k+\gamma_i\}})$.
- $Z_i \sim Ber(\frac{\exp\{\beta_0+\sum_{j=i}^{J}\beta_j x_{ij}\}}{1+\exp\{\beta_0+\sum_{j=i}^{J}\beta_j x_{ij}\}})$, where x represents the covariates.
- (Y_{ik}, Z_i) independent of (Y_{jk}, Z_j) for all i, j and k.
- (Y_{ik}, Z_i) independent of (Y_{jl}, Z_j) for all i, j, k and l
- Y_{ik} and Y_{ij} independent given Z_i.

The higher α_k is the better astronomer k is. The higher γ_i is the easier it is to classify galaxy i. We impose that $\alpha_1 = \gamma_1 = 0$ to avoid trivial identifiability problems. The model assumes that there is no interaction between galaxy and astronomer. Note that this is a mixture model of products of independent Bernoulli variables.

Denoting

$$\mu_i = \frac{\exp\{\beta_0 + \sum_{j=i}^{J}\beta_j x_{ij}\}}{1+\exp\{\beta_0 + \sum_{j=i}^{J}\beta_j x_{ij}\}},$$

the complete likelihood is given by

$$L(y, z; \theta) = \prod_i P(Y_{ik} = y_{ik}, Z_i = z_i \ \forall k) = \prod_i P(Y_{ik} = y_{ik} \ \forall k | Z_i = z_i) P(Z_i = z_i) =$$

$$\prod_i \left(\prod_{k \in A_i} P(Y_{ik} = y_{ik} | Z_i = z_i) \right) P(Z_i = z_i) =$$

$$= \prod_i \mu_i^{z_i}(1-\mu_i)^{1-z_i} \times b_i,$$

where

$$b_i = \prod_{k \in A_i} \left[\left(\frac{exp\{\alpha + \alpha_k + \gamma_i\}}{1 + exp\{\alpha + \alpha_k + \gamma_i\}} \right)^{y_{ik}} \left(\frac{1}{1 + exp\{\alpha + \alpha_k + \gamma_i\}} \right)^{1-y_{ik}} \right]^{z_i} \times$$

$$\times \left[\left(\frac{1}{1 + exp\{\alpha + \alpha_k + \gamma_i\}} \right)^{y_{ik}} \left(\frac{exp\{\alpha_k + \gamma_i\}}{1 + exp\{\alpha_k + \gamma_i\}} \right)^{1-y_{ik}} \right]^{1-z_i}$$

and A_i is the set of all astronomers that vote for galaxy i. (Note that as not all astronomers evaluate all galaxies necessarily). Hence, the log-likelihood is given by

$$l(y, z; \theta) =$$

$$\sum_i \left(\sum_{k \in A_i} d_{ik} log \left(\frac{exp\{\alpha + \alpha_k + \gamma_i\}}{1 + exp\{\alpha + \alpha_k + \gamma_i\}} \right) + (1 - d_{ik}) log \left(\frac{1}{1 + exp\{\alpha + \alpha_k + \gamma_i\}} \right) \right) +$$

$$+ z_i log(\mu_i) + (1 - z_i) log(1 - \mu_i),$$

where

$$d_{ik} = 1 + 2z_i y_{ik} - z_i - y_{ik}.$$

In order to introduce sparsity on the coefficients β_j, instead of maximizing the (marginal) log-likelihood, we can solve the following maximization problem:

$$\arg\sup_\theta \; l(y_{ij} \; \forall i, j; \theta) - \lambda \sum_{j=1}^{J} |\beta_j| \; .$$

In order to solve this problem, we will first reframe it in terms of a Bayesian problem that leads to the same results. Let's imagine that we assign a prior probability for θ as follows:

$$\pi(\theta) \propto \exp(-\lambda \sum_{j=1}^{J} |\beta_j|).$$

This is the product of (improper) uniform priors for γ's and α's with Laplacian distributions for the β's.

Now, suppose we want to find the Maximum a Posteriori estimate (MAP) of θ. We then have to find

$$\arg\sup_\theta \; L(y_{ij} \; \forall i, j; \theta) \exp(-\lambda \sum_{j=1}^{J} |\beta_j|) = \arg\sup_\theta \; l(y_{ij} \; \forall i, j; \theta) - \lambda \sum_{j=1}^{J} |\beta_j|,$$

which is exactly the same problem. Hence, we can use a MAP-EM approach to solve the original problem. More specifically, it could be implemented as follows. Let

$$g(\theta, z) = l(y, z; \theta) - \lambda \sum_{j=1}^{J} |\beta_j| \; .$$

We attribute initial values to θ. Then we interact until convergence:

- (Expectation Step) Find the expected value of the $g(\theta, Z)$ (denoted by $E[g(\theta, Z)]$), conditional on the current estimates of the parameters θ and on y_{ij}.
- (Maximization Step) Maximize $E[g(\theta, Z)]$ with respect to θ.

To perform the Expectation Step, we note that $g(\theta, Z)$ is linear in the latent variables Z, so we just need to calculate (for all i's)

$$E[Z_i|Y_{ik} = y_{ik}\ \forall i,k] = P(Z_i = 1|Y_{ik} = y_{ik}\ \forall i,k) = P(Z_i = 1|Y_{ik} = y_{ik}\ \forall k) =$$

$$\frac{\prod_{k\in A_i}\left(exp\{\alpha+\alpha_k+\gamma_i\}\right)^{y_{ik}}\mu_i}{\prod_{k\in A_i}\left(exp\{\alpha+\alpha_k+\gamma_i\}\right)^{y_{ik}}\mu_i + \prod_{k\in A_i}\left(exp\{\alpha+\alpha_k+\gamma_i\}\right)^{1-y_{ik}}(1-\mu_i)} =$$

$$\frac{\mu_i exp\{\sum_{k\in A_i} y_{ik} * (\alpha+\alpha_k+\gamma_i)\}}{\mu_i exp\{\sum_{k\in A_i} y_{ik} * (\alpha+\alpha_k+\gamma_i)\} + (1-\mu_i)exp\{\sum_{k\in A_i}(1-y_{ik}) * (\alpha+\alpha_k+\gamma_i)\}}.$$

We can then perform the expectation step by plugging this values in $g(\theta, Z)$.

For the maximization step, we have that

$$\arg\sup_{\theta}\ l(y_{ij}\ \forall i,j;\theta) - \lambda\sum_{j=1}^{J}|\beta_j| =$$

$$\arg\sup_{\gamma' s,\alpha' s}\ \sum_i\left(\sum_{k\in A_i} d_{ik} log\left(\frac{exp\{\alpha+\alpha_k+\gamma_i\}}{1+exp\{\alpha+\alpha_k+\gamma_i\}}\right) + (1-d_{ik})log\left(\frac{1}{1+exp\{\alpha+\alpha_k+\gamma_i\}}\right)\right) +$$

$$\arg\sup_{\beta}\ \sum_i\left(z_i log(\mu_i) + (1-z_i)log(1-\mu_i)\right) - \lambda\sum_{j=1}^{J}|\beta_j|.$$

Hence, we have two independent maximization problems. From here on I assume all a astronomers evaluate all g galaxies for simplicity of notation. The first maximization problem can be rewritten as

$$\arg\sup_{\gamma' s,\alpha' s}\ \sum_{i=1}^{2ga} w(i)log(\mu_i'),$$

where

$$w_i = \begin{cases} i\text{-th element of the vectorization of the matrix } (d_{jl})_{1\le j\le g, 1\le l\le a} \text{ for } 1\le i\le ga \\ 1 - w_{i-ga} \text{ for } ga+1\le i\le 2ga \end{cases}$$

and

$$\mu_i' = \begin{cases} i\text{-th element of the vectorization of the matrix } (\mu_{jl})_{1\le j\le g, 1\le l\le a} \text{ for } 1\le i\le ga \\ 1 - \mu_{i-ga}' \text{ for } ga+1\le i\le 2ga \end{cases}$$

Here,

$$\mu_{ik} = \left(\frac{exp\{\alpha+\alpha_k+\gamma_i\}}{1+exp\{\alpha+\alpha_k+\gamma_i\}}\right)$$

This is just a Weighted Logistic Regression, and can be solved using functions such as glm in R. Note that the observations related to this maximization problem are

$$\overbrace{1,\ldots,1,}^{\text{ga times}}\ \overbrace{0,\ldots,0}^{\text{ga times}}$$

The second maximization problem can be rewritten as

$$\arg\sup_{\beta} \sum_{i=1}^{2g} w_i log(\mu_i') - \lambda \sum_{j=1}^{J} |\beta_j|,$$

where

$$w_i = \begin{cases} z_i & \text{for } 1 \leq i \leq g \\ 1 - z_{i-g} & \text{for } g+1 \leq i \leq 2g \end{cases}$$

and

$$\mu_i' = \begin{cases} \mu_i & \text{for } 1 \leq i \leq g \\ 1 - \mu_{i-g} & \text{for } g+1 \leq i \leq 2g \end{cases}$$

This is just a Weighted L1-Regularized Logistic Regression, and can be solved using libraries such as glmnet in R. Note that the observations related to this maximization problem are

$$\overbrace{1, \ldots, 1}^{\text{g times}}, \overbrace{0, \ldots, 0}^{\text{g times}}$$

How to pick λ or competing models? One way is to think about the prior. Another can be minimizing a cross validated version of

$$\sum_i \sum_{j \in A_i} |y_{ij} - z_i^{\lambda}|.$$

We can also see (by cross validation) how well the model predicts y.