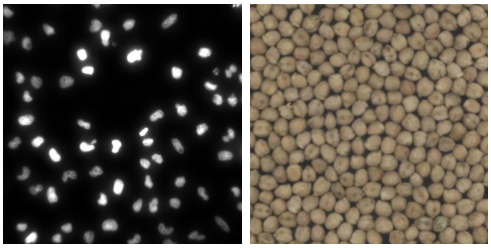
A Model-Free Approach for Bias Correction when Measuring Object Sizes in Images

Jiaping Zhang Jiaming Xie
University of California, Davis

May 30, 2015

1 Introduction

The truncated object problem is originally from a lot of scientific fields of study. The researchers are trying to use photos to capture the characteristics of the their research objects. However, usually due to the physical limitations, they can only capture the images for some samples instead of a panorama for the population. Under such a circumstance, researches would not be able to know all the information about the truncated objects, which are located around the edges of the images. The method developed in this research aims at resolving this problem by making reasonable predictions based on measured features of the completed objects in the images. It can be widely used in other researches fields as well. For example, it can help biologists to predict the sizes of the truncated part of cell objects in medical images(Fig.1a), so that they can have a better reasoning about the sample cell images of patients; in agricultural bean industries, this method can also be applied for quality control – since food companies harvest a large amount of beans, it is not feasible to capture images for all these beans, they can use this estimation method on some random samples to help them estimate the quality of the grains(Fig.1b). In this paper, we will propose a nonparametric approach to predict the sizes of truncated objects.



(a) (b)

Figure 1: Real Sample Images

[0]Research project for Winter and Spring 2015. Supervised by Prof.Thomas Lee.

2 Assumptions

We have the following assumptions for our specific research problem.

1. Multiple Binary images
2. Truncated objects on the images
3. Homogeneous objects with random locations
4. The shapes of objects are closed to ellipses
5. No overlapping in images

It's worth mentioning that we do not presume any distribution of the shapes of the objects. Thus, the method that we are going to use is nonparametric.

3 Methodology

The following figure shows the pattern of images that we are normally given. We can see that the sample image simulated thoroughly follows our assumption for our research purpose. In order to make statistically meaningful inferences

Figure 2: Sample Image from Simulation

about the truncated parts of the objects, we introduce the concept of artificial cut. The main goal of artificial cut is to uniformly choose an object, and cut at some random point vertically, as you can see in Fig.3.

Then we will measure various features separately for the left part, and right part of the object. We can use these features of the left parts of the objects to be the predictors, and the right part as response variables.After measuring these features of numerous partitioned objects, we can get a data set. In this way, we can apply statistical modeling methodologies on the data set to do predictions.

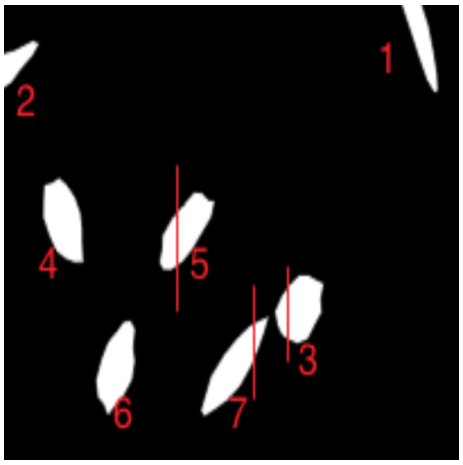


Figure 3: Artificial Cuts

After applying the artificial cut to the completed objects in the images, we intend to extract the features, or parameters. Totally, we defined nine features for the left part as the following:

- Left Area: the area size of the left part
- Cut Length: the length across the cut in a object
- Left Curve Length: the length of the left boundary of the object
- height: the maximum vertical distance in the object
- width: the maximum horizontal distance in the left part of object
- meanRadii: the mean of the distances between the point on the left boundary of the object to the mid point of the cut
- DistFromMid: the distance between the mid-point of the cut and the cutting tangent point on the left boundary
- Area ratio1: the ratio of the Left area divided by inner triangle area constructed by the middle point on the left boundary and the artificial cut
- Area ratio2: the ratio of the outer rectangle area divided by the Left area

4 Modeling

We use our simulation function to generate 500 sample images as our training data set and 200 sample images as our testing data set. Note that we also want to verify whether our functions of simulation and extracting feature are reliable and robust. We can look at the distribution of the total area calculated by the left area and the right area from the training data set. In Fig.4, we can see that the distribution is close to a normal distribution with a mean around 450,

which alligns with the value we designed in our simulation function. Thus, it proves that our functions work reliably.

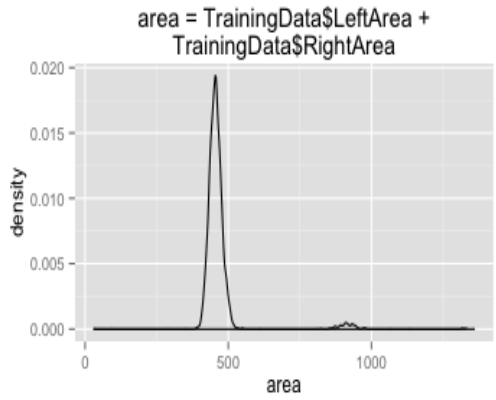


Figure 4: The Distribution of Area

Before we conduct the modeling, we do a Principle Component Analysis. From the Fig.5, we can see that there are only six main components in our data set, which indicates that there are some correlation between our variables.

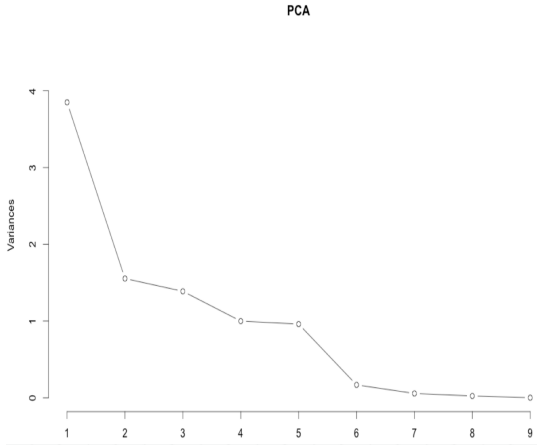


Figure 5: Principle Component Analysis

We will do the predictive modeling using the statistical methods that can deal with dependent variables and nonlinearities. We use the MSE as our metric to compare our models. From the table below, we can see that generally those methods work well.

Models	MSE
Penalized Linear Regression(LASSO)	0.259
Multivariate Adaptive Regression Splines	0.182
Support Vector Machine Regression (SVM)	0.145
Bagging	0.176

Let's look at the prediction performance. We plot the true value versus the predictions to check the prediction accuracy. There are 8000 data points in the plots. Ideally, we would see most of the points are around the straight line with the 45 degrees. Thus, it seems that for the methods we tried, we have relatively good prediction accuracy although there are some extreme cases.

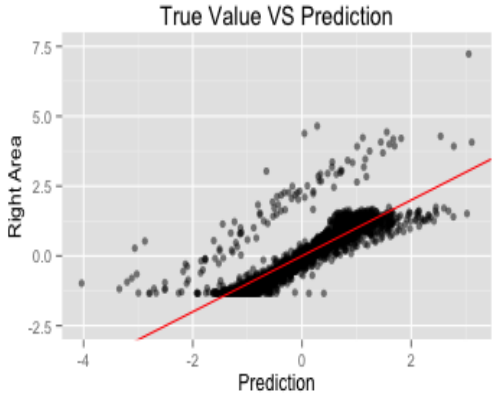


Figure 6: Prediction Accuracy for LASSO model

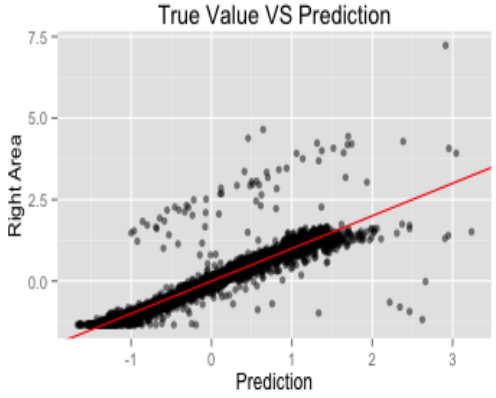


Figure 7: Prediction Accuracy for Nonlinear model(spline)

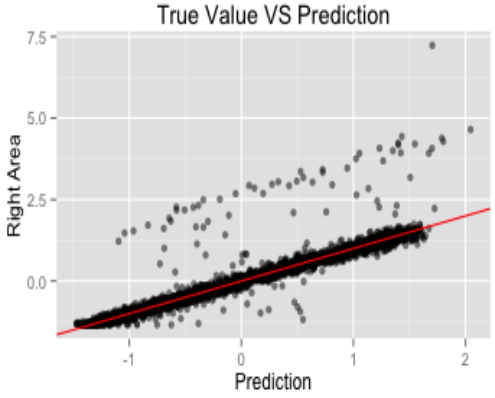


Figure 8: Prediction Accuracy for SVM Regression model

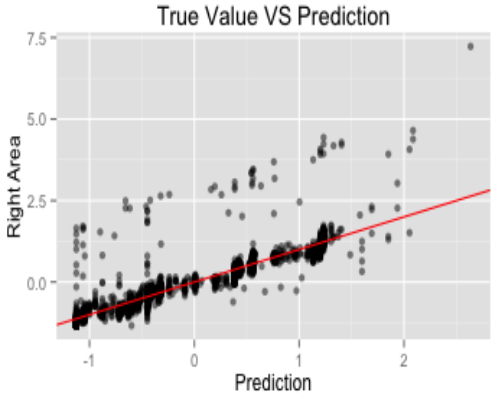


Figure 9: Prediction Accuracy for Bagging model

Through our investigation, we found out there are about 14% extreme cases in all the data points. Furthermore, those extreme cases mostly come from an unideal situation that a random cut happens to be close to the boundary of the object, which would enlarge the bias when modeling.

5 Conclusion

In this project, we have successfully validated the methodology through simulation and modeling. We also observe that this approach has a robust and good performance in predictions with less images by increasing the number of cuts.

So the realistic significance is that this approach can reduce the cost of imaging and computation. However, the disadvantage of this methodology is that some extreme cases/outliers are sensitive to be generated due to random cuts, which would affect the prediction accuracy of modeling.

When considering further study, we expect to extend our approach to three dimension objects and improve the algorithm to control the behavior of random cuts for better bias correction and prediction accuracy.