On the Outlier Detection for Standardized Tests

Nicolás Acevedo
Industrial Engineering Department
University of Chile
Santiago, Chile
nicacevedo@ug.uchile.cl

Charles Thraves
Industrial Engineering Department
University of Chile
Santiago, Chile
cthraves@dii.uchile.cl

María Leonor Varas
Department of Mathematical Engineering
University of Chile
Santiago, Chile
mlvaras@dim.uchile.cl

Abstract—This work focuses on the detection of outliers on standardized college admission test. We use four methods—Mahalanobis Distance, Isolation Forest, and variations of these—to detect outliers in multivariate data. Applying PCA, the outlier metrics are combined in a single scoring for each student. Students with the highest scores are clustered with k-means to get interpretable groups of outliers. The methodology is performed with real data of Chilean admission test results PTU.

Index Terms—Outliers, PCA, Standardized tests, k-means

I. Introduction

Standardized test are one of the most common methods used for college admission around the world. In addition, standardized tests are also used for several other purposes such as: certificate tests in a specific language, evaluation of students in intermediate school levels, etc. These tests are usually conformed by multiple sections where each of them has an individual score. Some of these sections might be similar to each other, and therefore results among these might have a positive correlation. On the contrary, there might be sections which scores do not correlate. As a result, not every combination of scores is equally likely to occur. For instance, if there a test with the sections arithmetic, geometry, and English reading; a score (in percentage units) of (90, 85, 70) is probably more likely than a score of (90, 10, 70). In this example, the latter score might either (i) be a "true" score, or (ii) be a score which does not corresponds to the "true" score as a result of different reasons such as: the pencil on the arithmetic test section did not mark the answers well, or the answer key used to correct the arithmetic test was not the adequate one, or there was cheating in one of the test, etc. Usually the number of people that take standardized tests is large enough (of the order of thousand, or hundred of thousand, or millions in some cases) so we should seek for more advanced and automatic techniques able find the tests results that show a sign to be unlikely to happen. Once these tests have been identified, we can look in more detail the reasons that underlie the obtained score, and in particular find whether or not there is a problem with the obtained test score.

The authors gratefully acknowledge financial support from CONICYT PIA/BASAL AFB180003.

We focus our work in the Chilean college admission test, Prueba de Transición Universitaria (PTU), also denoted as PSU before 2020, where students take three out of the four tests: Language, Mathematics, Science, and Social Sciences (the two former are mandatory for school applications). The test results together with high school grades [1] and a ranking score [2] (that depends on the relative position and grade of the student with respect to her class) are combined in a weighted average that depends on the degree and university the student applies to.

The research question is how can we identify outliers for standardized tests while providing a way to interpretate and classify the type of outlier of each respondent. In Section II we discuss on the data set used. Then, we explain the methodology in Section III. Results are shown in Section VI. Finally, in Section VII we provide conclusions.

II. Data

We use the results of the Chilean admission college test PTU of 2020. This test was taken by more than 150 thousand students. Since most students take three (of the four) sections of the test—either language, math, and sciences; or language, math, and social sciences—we divide the data and analysis in two groups. In case an applicant takes the four tests, we consider the student to be in the group corresponding to her highest score between the tests of Sciences and Social Sciences. Although we apply the methodology in both data sets (the one with the science and the one with the social science test), for the sake of brevity we only show the results applied to the group of language, math, and science tests.

A. Notation

The set of students (or applicant) is denoted by $\mathcal{I}$, and the set of tests (i.e. sections of the test) is denoted by $\mathcal{T}$. The set of tests of language, math, and science is denoted by $\mathcal{T}_c \subset \mathcal{T}$, whereas the set of tests of language, math, and social science is denoted by $\mathcal{T}_s \subset \mathcal{T}$. The score of student $i \in \mathcal{I}$ on test $t \in \mathcal{T}$ is denoted by x_{it}. $X^{(i)}$ will denote the test scores of the i-th student. μ denotes the mean scores of all tests among students.

III. Methodology

We first run four outlier methods, then apply PCA in order to obtain the first principal component labeled as

score, and then use k-means to cluster outliers for ease of interpretation. The four outlier methods are described below:

A. Mahalanobis Distance

We compute the Mahalanobis distance between each point (i.e. student test scores) and the average point. Unlike the regular Euclidean distance, Mahalanobis Distance takes into account correlations in the data. This distance computed is considered as the score, namely

$$d_M(X^{(i)}, \hat{\mu}) = \sqrt{(X^{(i)} - \hat{\mu})^\top \widehat{\Sigma}^{-1} (X^{(i)} - \hat{\mu})} \quad (1)$$

B. Robust Mahalanobis Distance

Since there could be problems with the previous method because of outliers, Robust Mahalanobis Distance, proposed by [3], computes a variation of the covariance matrix so that is less sensible to these points. In particular, the covaraince matrix is computed as

$$\widehat{\Sigma}_R = \sum_{i=1}^{n} \frac{K(d_M(X^{(i)}, \widetilde{X})^2)(X^{(i)} - \widetilde{X})(X^{(i)} - \widetilde{X})^\top}{\sum_{j=1}^{n} K(d_M(X^{(j)}, \widetilde{X})^2)} \quad (2)$$

where $K(u) = e^{-0.1u}$, and the distance then is computed as

$$d_{RM}(X^{(i)}, \widetilde{X}) = \sqrt{(X^{(i)} - \widetilde{X})^\top \widehat{\Sigma}_R^{-1} (X^{(i)} - \widetilde{X})} \quad (3)$$

IV. Isolation Forest

Isolation Forest selects randomly a column of the data, and samples uniformly a point in which to cut the data in two. This process continues until the first point is singled out. The process is performed several times. If a point gets isolated in an early branching rather than a later one, it is more likely to be an outlier. Then, the scoring for each point is computed as

$$s(X^{(i)}, n) = 2^{-\frac{\mathbb{E}(h(X^{(i)}))}{c(n)}} \quad (4)$$

where $h()$ denotes the depth of the tree in which the i-th point is left as an outlier. $c(n)$ is the average depth of the tree in an instance with n elements. For more details see [4].

V. Extended Isolation Forest

This method works similar to Isolation Forest, except that the sample cuts are not necessarily orthogonal to the axis.

VI. Results

A. Outlier Detection

The four outlier methods applied are: Mahalanobis Distance, Robust Mahalanobis Distance, Isolation Forest, and Extended isolation Forest. For each of these methods, a score is computed for each of applicant. We observe from Figure 1 the scoring for 50 random applicants in $\mathcal{T}_2$. As expected, the outlier scores obtained are highly correlated among the methods. In order to consolidate these scorings in a single value, we apply PCA.

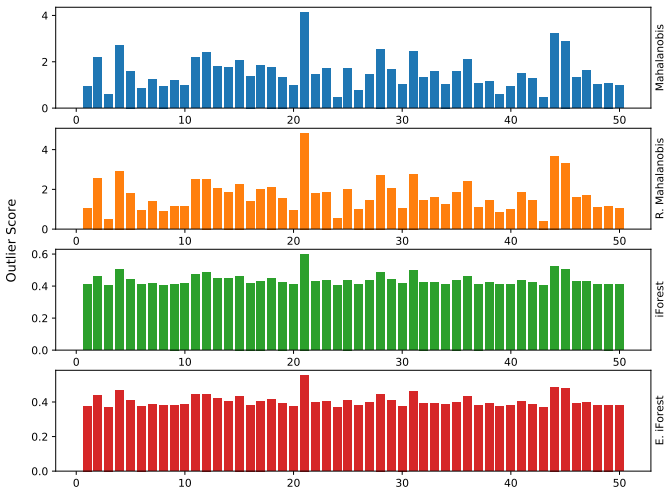

Fig. 1. Outlier score for each method on a sample of $n = 50$ observations.

B. PCA

The goal is to get the first principal component while using as input the scores computed with each of the four methods. Table I shows the variance obtained from each component. It is observed that the first component explains almost 97% of the variability of the data. Thus, there is very little information lost when considering the principal component as the score.

TABLE I
PCA Explained Variance

	PC1	PC2	PC3	PC4
Value	3.88	0.10	0.01	0.01
Percentage	96.96%	2.52%	0.35%	0.16%

We also compute the correlation matrix with (i) the Kendall's τ [5] and the Spearman's ρ [6] coefficients. The objective of this is to visualize how similar are the scores obtained between each pair of methods, and also how similar are these with respect to the first principal component.

TABLE II
Correlation Matrix with Kendall's τ coefficient

	Mahala.	R. Mahal.	iFor.	E. iFor	PC1
Mahalanobis	1.00				
R. Mahalanobis	0.90	1.00			
iForest	0.88	0.86	1.00		
E. iForest	0.86	0.88	0.88	1.00	
PC1	0.93	0.93	0.91	0.92	1.00

We observe from Tables II and III that scores obtained with all methods are very similar, and more so when comparing each method and the first principal component (PC1).

TABLE III
Correlation Matrix with Spearman's ρ coefficient

	Mahala.	R. Mahal.	iFor.	E. iFor	PC1
Mahalanobis	1.00				
R. Mahalanobis	0.99	1.00			
iForest	0.98	0.97	1.00		
E. iForest	0.97	0.98	0.98	1.00	
PC1	0.99	0.99	0.99	0.99	1.00

TABLE IV
k-means centroids

Cluster	Leng	Math	Scie
1	94.61	-121.06	26.45
2	-119.95	127.74	-7.79
3	-4.94	109.84	-104.90
4	-9.23	-10.94	20.17
5	133.31	-70.44	-62.87
6	90.84	70.79	-161.63
7	-118.62	38.36	80.26

C. k-means

Up to this point, we can sort students decreasing in the scoring obtained, and look to those with the highest values. Unfortunately, this provides little intuition of why an applicant is an outlier, and also is difficult to analyze, since some identified outliers could be of interest while others do not. For example, although someone that scores very high on all tests is probably an outlier according to the mathematical definition, this is not that interesting. (For space constraint we do not provide such a table of results). Then, a possible solution is to apply k-means in order to cluster the outliers in groups that follow a similar pattern. Note that more important than the absolute value of the test scores of a student, is the relative difference between these values. In other words, an outlier could be someone whose math score is high while his science score is low. Yet, we do not care how high or low are the respective scores. For each student, we subtract the average score among the three tests to each of the tests. Then, we apply k-means. In order to choose the value of k, we note that there are 7 patterns of outliers. From these 7 patterns, three patterns correspond to the cases when the student performs much better in one test, three patterns correspond to the cases when the student performs much better in two tests, and the last pattern is when the student performs similar in all three tests (this could be the case for a students that did either well or terrible in all tests). This can be see from the clusters' centroids shown in Table IV.

The five highest outliers of each cluster are shown in Table V. We observe that each group is clearly defined by a pattern. For example, group 1 corresponds to the applicants whose scores in Math are significantly lower than their scores in Language and Science. Group 4 has students which perform extremely well in all three tests. So on, we can characterize each cluster with the pattern observed.

TABLE V
k-means centroids

Cluster	Lang	Math	Scie	PC1
1	770	210	626	21.65
1	669	304	547	14.21
1	755	370	530	13.58
1	646	304	486	12.98
1	657	326	552	12.50
2	299	663	509	13.51
2	336	626	403	11.87
2	492	831	618	11.35
2	436	756	704	11.11
2	498	831	670	10.55
3	395	737	233	20.10
3	404	552	211	13.53
3	471	597	253	12.60
3	626	850	587	11.38
3	420	548	233	11.36
4	834	831	850	9.81
4	834	775	850	9.67
4	834	850	796	9.56
4	834	850	814	9.53
4	834	728	832	9.46
5	755	370	444	13.91
5	802	578	477	11.78
5	755	453	457	11.25
5	657	349	431	10.98
5	850	609	622	10.95
6	755	570	204	20.11
6	616	585	233	15.22
6	531	476	150	14.88
6	755	521	332	14.83
6	531	465	150	14.80
7	598	850	850	11.10
7	386	578	675	10.59
7	531	602	778	10.21
7	598	704	850	10.03
7	367	609	639	9.93

VII. Conclusions

The present work shows a methodology to search for outliers on standardized tests and characterize these depending on the pattern of their scores. The results of this work are currently being used and applied with the official organization in charge of the PTU, DEMRE in order to detect unlike score cases which are worth looking in more detail in order to confirm that the scores are actually the ones that correspond to the applicant. Therefore, we believe there is a high impact in this research since this allows to fix scores that have been wrongly assigned due to unobservable or undetected elements.

References

[1] DEMRE. Notas de Enseñanza Media (NEM). DEMRE - Departamento de evaluación, medición y registro educacional. Retrieved June 27, 2021, from https://demre.cl/proceso-admision/factores-seleccion/notas-ensenanza-media

[2] DEMRE. Puntaje Ranking. DEMRE - Departamento de evaluación, medición y registro educacional. Retrieved June 27, 2021, from https://demre.cl/proceso-admision/factores-seleccion/puntaje-ranking

[3] Penny, K. I.,& Jolliffe, I. T. (2001). A Comparison of Multivariate Outlier Detection Methods for Clinical Laboratory Safety Data. Journal of the Royal Statistical Society: Series D (The Statistician), 50(3), 295–307. https://doi.org/10.1111/1467-9884.00279

[4] Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation Forest. 2008 Eighth IEEE International Conference on Data Mining. Published. https://doi.org/10.1109/icdm.2008.17
[5] Kendall's Tau. (2017). The SAGE Encyclopedia of Communication Research Methods. Published. Los Angeles, Calif: Sage Publications. https://doi.org/10.4135/9781483381411.n283
[6] Spearman Rank Correlation Coefficient. (2008). The Concise Encyclopedia of Statistics, 502–505. Springer, New York, NY. https://doi.org/10.1007/978-0-387-32833-1_379