Composition of concrete and its influence on compressive strength

Filipe P. de Farias
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
filipepfarias@fisica.ufc.br

Yvo J. M. Sales
Department of Teleinformatics Engineering
Federal University of Ceará
Fortaleza, Brazil
yvo@gtel.ufc.br

***Abstract*—The compressive strength of concrete impacts directly on its application. The difference between the concrete for columns or beams and the concrete for pavements is mainly due the compressive strength it is able to resist. In this work, we perform an unconditional and a class-conditional mono-variate analysis as well as unconditional bi-variate and multi-variate analysis of the UCI concrete composition database.**

***Index Terms*—concrete, compresive, strength, machine, learning, pre-processing**

I. Introduction

A material formed by aggregates bonded together by a fluid material that hardens over time has been used by humans for construction since many years ago [3]. Nowadays this material is known as concrete and it's widely used in the construction field. The aggregates used in the concrete affect directly its compressive strength which highly impacts its applications. For instance, in general, the concrete for columns or beams needs to have a greater compressive strength than the one for pavement. On this paper, we carry out an statistical analysis on a dataset extracted from the UCI Machine Learning Repository (University of California, Irvine) [4] that collects information about the concentration of some aggregates used to form different types of concretes and their resulting compressive strenght.

This work is divided as follows. A description of the data is given in Section II. The Section **??** brings an unconditional and a class-conditional mono-variate analysis as well as unconditional bi-variate and multi-variate analysis of the data. Finally, the conclusions and considerations are exposed in Section IV.

II. Data description

The composition of each of the N concrete samples is given by the concentrations (kg/m^3) of D components: Cement, Blast Furnace Slag, Fly Ash, Water, Superplasticizer, Coarse Aggregate and Fine Aggregate. Each sample has its Age (day) and the measured Concrete compressive strength (MPa), as described in Table I.

TABLE I
Data description

Component	Description	Unit
D_1	Cement	kg/m^3
D_2	Blast Furnace Slag	kg/m^3
D_3	Fly Ash	kg/m^3
D_4	Water	kg/m^3
D_5	Superplasticizer	kg/m^3
D_6	Coarse Aggregate	kg/m^3
D_7	Fine Aggregate	kg/m^3
D_8	Age	days
Y	Compressive strength	MPa
N	1030 samples	

The concrete was stratified into $L = 3$ classes [1]. The concrete which is weak and not recommended for structures, the *Non-standard*, was labeled with L_1 and has 295 samples in it. The concrete whose strength is in a range that can be applied to structures is classified as L_2, or *Standard*, and has 525 samples. The high performance concrete L_3, *High-strength* and has 210 samples. The observations are the measured compressive strengths of each sample and, as the predictors $D_1 - D_7$, are continuous. The Age (D_8) of the concrete is extremely discrete. The output is the concrete strength

III. Methods

Regression models try to find relations between the *independent variables*, the predictors in this work, and the *dependent variables*, the outcomes. These relations can occur in different forms, but the more common one is when a linear increase in one represents a linear increase in the other. Relations of this type are comprised by the *Linear Regression*.

A. Ordinary Least Squares

$$Y = \begin{bmatrix}\beta_0 & \beta_1\end{bmatrix}\begin{bmatrix}1\\ X\end{bmatrix} + \varepsilon \tag{1}$$

$$\begin{bmatrix}\hat{\beta}_0\\ \hat{\beta}_1\end{bmatrix} = \left(\begin{bmatrix}1 & X\end{bmatrix}^\top \begin{bmatrix}1 & X\end{bmatrix}\right)^{-1}\begin{bmatrix}1 & X\end{bmatrix}^\top Y \tag{2}$$

This method tries to find the linear regression between the predictors and outcomes by fitting a line with linear coefficient β_0 and angular coefficient β_1, defined in (1), through the data. The fitting is done by minimising a cost function defined by the squared distance between the data and the line to be fitted. The sum of the distances will be minimum in the best fit, what

gives the name to the method. This results in coefficients as in (2). Regard the fact that β_1 can be a vector of the size such as the number of predictors in X.

An interesting fact to observe is if the predictors correlates with the outcome, we can observe the error will be small. This is because the correlation evaluates linear variations between the variables as such as the linear regression.

TABLE II
STATISTICS SUMMARY OF THE PREDICTORS

	Mean	STD	Skewness
D_1	281.16	104.506	0.50873
D_2	73.89	86.279	0.79955
D_3	54.18	63.997	0.53657
D_4	181.56	21.354	0.07451
D_5	6.20	5.973	0.90588
D_6	972.91	77.754	0.04016
D_7	773.58	80.176	0.25264
D_8	45.66	63.169	3.26441

In Table II it is possible to see that the data are not *highly skewed* except by the Age data, discussed in the next section. This can be easier verified in Figure 1.

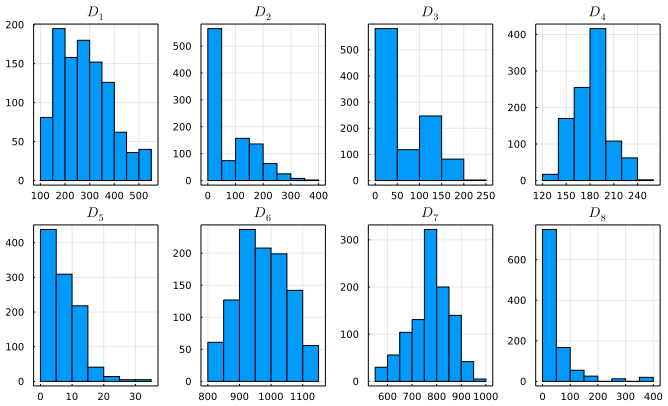

Fig. 1. Predictors' histograms for the unconditional monovariate analysis.

An interesting fact is that the components that, in general, are not absent in the concrete composition has the lower skewness, as Cement, Water, Coarse and Fine Aggregates. The other components have higher skewness due to the fact they are optional in concrete mixture, thus they are allowed to have zero values in some concrete samples. Another seeable fact is the discreteness of the Age values. By standard, the concrete is submitted to *cure* in the period of 28 days and that explains the large number of samples concentrated on the first bin of the histogram D_8 on Figure 1, which comprises the values between 0 and 50. The other cure periods are sampled in dataset to test its influence in compressive strength.

B. Class-conditional mono-variate analysis

As mentioned before, a Gaussian-like distribution can be noticed on the components except the Superplasticizer and the Age. On the other ones, as in Fly Ash, there's a large number of zeros. This behaviour is repeated on the three

TABLE III
STATISTICS SUMMARY FOR CLASS 1

	Mean	STD	Skewness
D_1	227.649	77.9188	0.813198
D_2	59.1078	85.7506	1.15493
D_3	58.1166	69.7541	0.558909
D_4	185.405	15.7587	-0.933526
D_5	3.96847	4.77879	0.781783
D_6	994.762	70.8093	-0.143029
D_7	794.88	66.6208	-0.116007
D_8	13.9322	15.9189	4.75222

TABLE IV
STATISTICS SUMMARY FOR CLASS 2

	Mean	STD	Skewness
D_1	277.673	97.8943	0.480469
D_2	75.5691	87.3597	0.793451
D_3	57.6739	61.9693	0.330694
D_4	184.16	21.6533	0.16274
D_5	6.04743	5.55232	0.745442
D_6	967.124	75.0478	-0.114811
D_7	767.236	82.4853	-0.332702
D_8	53.2952	69.2818	3.02584

TABLE V
STATISTICS SUMMARY FOR CLASS 3

	Mean	STD	Skewness
D_1	365.087	100.273	-0.0967353
D_2	90.4862	81.1218	0.403134
D_3	39.9562	58.619	0.978704
D_4	169.694	23.2573	0.97391
D_5	9.73905	6.82741	0.756237
D_6	956.722	87.007	0.367005
D_7	759.52	86.0634	0.162043
D_8	71.1524	70.9618	2.53224

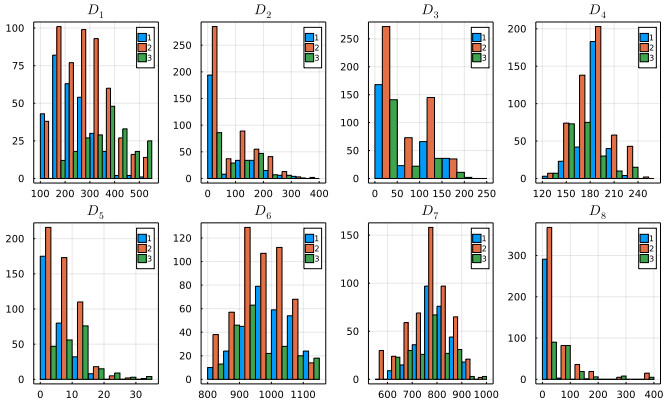

Fig. 2. Histograms of the predictors for the class-conditional mono-variate analysis with each color representing a different class.

classes. There's no apparent distinction between the classes too. In the trial to adjust the data of each component class into a Gaussian distribution, it was noticed that a mixture model would fits better given we have exact zeros and a distribution of non zero values. This is reasonable because the zero values are the representative absence of the component.

C. Unconditional bi-variate analysis

In the pairwise comparison of the predictors, only in (D_4, D_5), the Superplasticizer and Water concentrations, there is a prominent negative correlation and the other predictors seem to have no numerical correlations, as can be seen in the Figures 4 and 3. Therefore, it was not possible to simplify the problem by eliminating predictors. In the same manner mentioned before, visually it was not possible to elect a predictor that might separate the classes. The statistics of each predictor are at Tables III, IV and V.

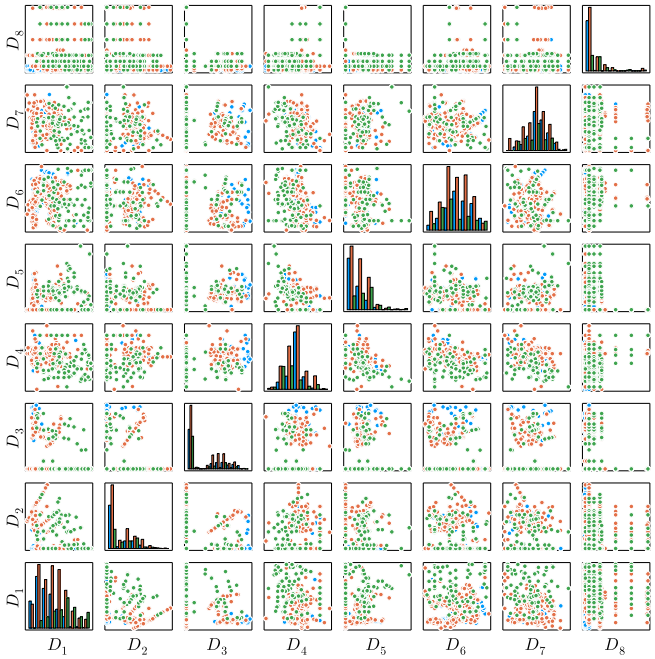

Fig. 3. Pairwise scatter plot of each concrete component. The color code is the same from Figure 2

D. Unconditional multivariate analysis (PCA)

How the PCs comprises the original variance is plotted on Figure 6. The original predictors are projected onto the PCs space, Figure 5. Note the first component, the cement D_1 has less information in the first two PCs components. An interesting fact about this is to know that the cement can even turns the concrete weaker to compression as long its concentration grows. The compression strength will be given by the aggregate and some chemicals, which are the predictors of major norm in the projection.

In the Figure 5 is shown the projection of the original data onto de first two PCs. The classes are highly overlapping

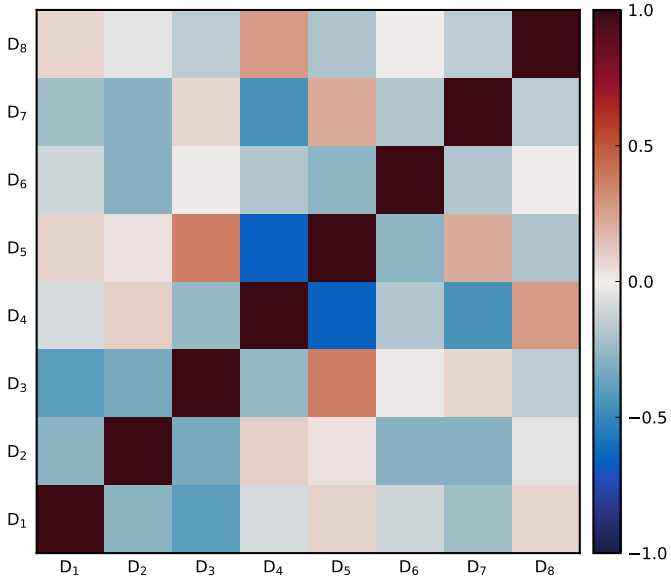

Fig. 4. Correlation matrix of the predictors.

and another methods are needed to allow the classes to be separeted.

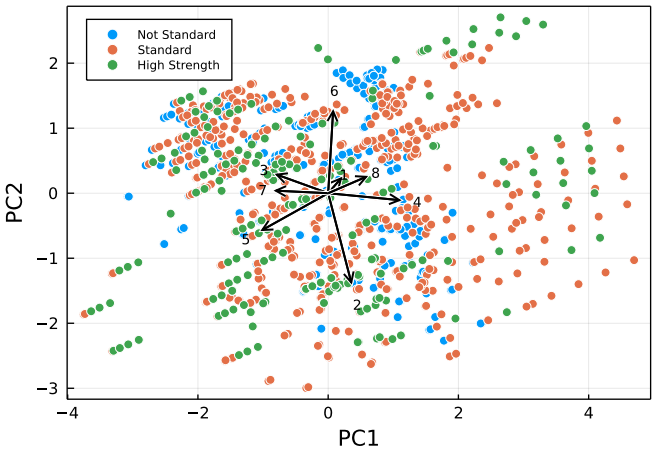

Fig. 5. The first two principal components with $46,2\%$ of the total variance. The loadings were scale by a factor of 2 for better viewing.

IV. Conclusion

The database of concrete is not easy to analyse if there is no previous knowledge about the problem of the components mixture. In none of the analysis the data have been shown as separable on the initially determined classes. The next step is to try to perform regression to model the compressive strength itself before trying to classify the samples.

References

[1] ACI Manual of Concrete Practice 2000, Part 1: Materials and General Properties of Concrete. American Concrete Institute. Farmington Hills, MI.

[2] Tibshirani, Robert, et al. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Alemanha, Springer New York, 2009.

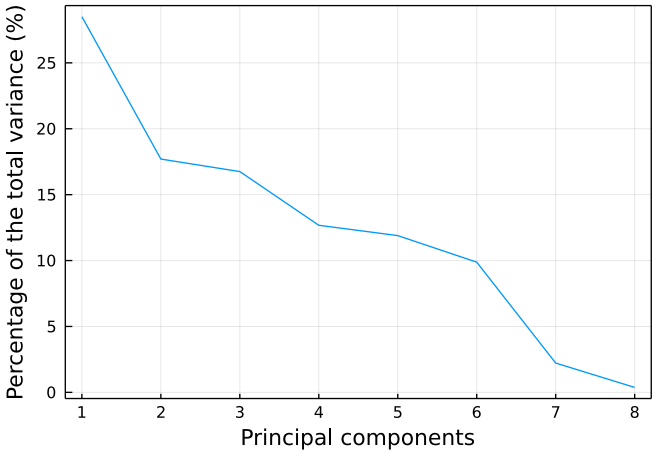

Fig. 6. Percentagem of the total variance of the original space of the data in each of the principal components and the respective loadings of each component.

[3] Mindess, S., and Young, J.F. Concrete. Prentice-Hall, Inc., Englewood Cliffs, NJ, 1981.

[4] I-Cheng Yeh, "Modeling of strength of high performance concrete using artificial neural networks," Cement and Concrete Research, Vol. 28, No. 12, pp. 1797-1808 (1998)