# Dimension Reduction - Comparison Script for Python

*Giménez Gredilla, Daniel*

*15 de diciembre de 2017*

# 1 - Introduction and general data set description

The current data set is a study on **normal**, **atypical**, and **variant** lymphocytes, in which 2867 numerical variables, based on colorimetric and geometric features have been measured. The aim of the study is more thoroughly explained both in the proposal and main report for this project.

The current data set has more than 13000 observations for 2874 variables, of which 7 (in which the response variable, *tipoCelula*, is included) are factors and 2867 are numeric predictor variables.

The response variable, *tipoCelula*, has three levels, being these **ATYPICAL_LYMPHOCYTE**, **VARIANT_LYMPHOCYTE** and **LYMPHOCYTE**. Two subsets will be generated with this data, a training set comprising 66% of all data, and a test set formed by the remaining entries. Both sets are generated by defining a random seed with the *train_test_split()* function, defined by the *random_state* parameter (the seed beeing 123).

# 2 - Goals and techniques

This script aims to assess the comparative performances of different dimension reduction techniques in the context of **Python** and also between **Python** and **R**. They will be measured for a common accuracy metric (*Cohen's Kappa*) and judged by their processing performance. Being **PCA** the most used, most assessed technique, it has been used, in the context of **R**, as a kind of touchstone in respect to requirements for the other techniques, and the generic parameters that were used with **R** have been extended to **Python** for the purpose of reproducibility and comparability.

For the effects of this project, the benchmark number of variables is that which satisfies one condition, being this that the extracted features account for an accumulated 95% of variance in the reference dimensionality reduction technique, **PCA**. In the **R** implementation of this algorithm, it was determined that 210 features extracted by **PCA** were the bare minimum to achieve this goal.

With this number of extracted features as an objective benchmark, the following techniques will be conducted:

- **PCA**
- **ICA**
- **Factor Analysis**
- **Linear Decomposition Analysis**

The resulting features will then be divided in training and test sets, and used as input, first, for the fitting and training of an **SVM** function and then as input of a prediction function, both from the *sklearn* module. The predicted classes will then be cross-tested with the actual test classes. A confusion matrix and a Cohen's Kappa weighted value will then be output, and used as that technique's entry in the final performance comparison.

**Cohen's Kappa** [@Luengo2009] is an alternative to **Classification Rate** that takes into account random correct hits. It was originally used to measure the degree of agreement between two subjects describing the same event. In the meantime it has been adapted for classification tasks, as it compensates for random hits in the same way that **AUC** does for **ROC**. The mathematical expression for **Cohen's Kappa** is applied to the contingency table of an event in the following way:

$$kappa = \frac{n\sum_{i=1}^{C} x_{ii} - \sum_{i=1}^{C} x_{i.}x_{.i}}{n^2 - \sum_{i=1}^{C} x_{i.}x_{.i}}$$

Where $x_{ii}$ is the cell count in the main diagonal, $n$ is the number of examples, $C$ is the number of class values and $x_{.i}x_{i.}$ are the total columns and rows counts, respectively.

Table 1: PCA observed versus predicted results

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 3685 | 346 | 415 |

Table 2: PCA observed versus predicted results - percentages

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 82.88 | 7.78 | 9.33 |

The value range of **Cohen's Kappa** goes from -1 (total disagreement) to 1 (perfect agreement).

The reasons for chosing **Cohen's Kappa** as accuracy metric are as follows:

- The data set upon which it is going to be used has a multiclass factor response variable.
- Those labels are not ascribable to a binary synthetic class system.
- **Cohen's Kappa** yields a scalar, simple value well suited for multiclass classification.
- It is more powerful than other multiclass accuracy metrics such as **Classification Rate**, because it takes into account random hits, scoring successes separately for each class and aggregating them.

Weighted and unweighted values of **Cohen's Kappa** differ, as their name implies, in that weighted scores take into account the differential weights of several levels of disagreement between observed and predicted classes. This is a level of information that is lost in binary classification, as all disagreements between observed and predicted classes share the same level of disagreement.

# 3 - Dimension Reduction Techniques

## 3.1 - PCA

PCA is the most used unsupervised, linear dimension reduction technique currently available. It is also the best, in the mean-square error sense [@Fodor2002]. Its central idea is the construction of a set of features from a number of initial variables [@Jolliffe2002]. The number of new features will be less than the initial variables, while retaining as much as possible of the initial variation. This is achieved by linear transformations of the original data, and then establishing a descending order of the new features attending to the amount of variation retained or explained by each of them.

For this project, 210 of those extracted features are retained and used in the fitting, training and prediction of classes. The following section depicts the results of this protocol.

After fitting and predicting, hit and accuracy values are extracted and represented in **Table 1** and **Table 2**. It is discovered that the predicted values all yield an **ATYPICAL_LYMPHOCYTE** result. This is unusual, and repeated in the next algorithm, but it provides an interesting result: this technique's *Cohen's Kappa* is **zero**, which is remarkable. In the observed test labels the **ATYPICAL_LYMPHOCYTE** result is registered a majority of times, which, if only raw accuracy was used as metric, would yield a high random hit rate. Instead, *Cohen's Kappa* takes into account the probability of this one being the result, which in the predicted values is of a 100%, and lowers the expected validity of such a prediction from the high random accuracy potential yield to a round zero. Even though the raw **PCA**-constructed prediction is of an underwhelming precision, the solidity of *Cohen's Kappa* as a metric is supported by this occurrence.

Table 3: ICA observed versus predicted results

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 3685 | 346 | 415 |

Table 4: ICA observed versus predicted results - percentages

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 82.88 | 7.78 | 9.33 |

## 3.2 - ICA

Independent Component Analysis (*ICA*) is a statistical method for transforming an observed multidimensional random vector into components that are statisticaly as independent from each other as possible, this is, a tendency to **redundancy reduction** [@Tobergte2013]. In its linear approach, as with other dimension reduction algorithms, its goal is to take a zero-mean, $m$-dimensional variable, and by means of a linear transformation, find its $n$-dimensional transform, such that $n \leq m$, this transformation having some suitable properties. The vectors obtained from this transformation are neither orthogonal nor ranked in order.

Feature extraction is a prominent application of *ICA*. It is originally motivated by results in neuroscience that suggest that the same cited principle of redundancy reduction is applied by the brain for the early processing of sensory data.

*ICA* is a generative model (it describes how the observed data are generated by describing the components), and it seeks the minimization of mutual information between the transformed variables. It depends on the supposition of nongaussianity for the data; gaussian data is independent and of mean zero, it has no skewness and as such can only be estimated up to an orthogonal transformation [@Hyvarinen2000].

Applying it to the current data set, the results are as follows:

Hit and accuracy values are extracted and represented in **Table 3** and **Table 4**.

This results in a *Cohen's Kappa* of **zero**, as in this technique, just as in *sklearn*'s **PCA** implementation, a label value of **ATYPICAL_LYMPHOCYTE** is predicted for every observation. Once again, this metric succeeds in catching the random ratio of hits that could create an artifact when measuring raw accuracy, while suggesting that a further fine tuning is needed in this technique to bring it up to its **R** analogue's performance.

## 3.3 - Factor Analysis

The basic idea underlying Factor Analysis is that $p$ observed random variables, $\mathbf{x}$, can be expressed, except for an error term, as linear functions of $m(< p)$ hypothetical (random) variables or *common factors* [@Jolliffe2002]. The aim of Factor Analysis is to group variables that share a "common theme" under the same grouping, such that the dimensionality of the dataset is decreased.

Factor Analysis has been applied in psychology to identify groups of inter-related variables, as those components of intelligence that can be placed under a single factor $g$ or *general intelligence*, grouping factors such as *broad visual perception* (it includes all the intelligence variables related to visual tasks), or *broad auditory percention* (same as before, but with auditory tasks). This is interpreted as someone with a high $g$ having good *broad auditory and visual perceptions*, and $g$ sinthetically explaining the behaviour of the factors and variables "contained" within itself.

This technique is applied here to the given data set.

After fitting and predicting, hit and accuracy values are extracted and represented in **Table 5** and **Table 6**. The confusion matrix and *Cohen's Kappa* score show a definite improvement over **PCA** and **ICA**.

Table 5: Factor Analysis observed versus predicted results

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 3636 | 240 | 95 |
| LYMPHOCYTE | 24 | 106 | 0 |
| VARIANT_LYMPHOCYTE | 25 | 0 | 320 |

Table 6: Factor Analysis observed versus predicted results - percentages

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 81.78 | 5.40 | 2.14 |
| LYMPHOCYTE | 0.54 | 2.38 | 0.00 |
| VARIANT_LYMPHOCYTE | 0.56 | 0.00 | 7.20 |

This technique gives a weighted Cohen's Kappa Value of 0.76. This value is the first one in the **Python** implementation to yield a positive result, in some degree of accordance to the **R** results, in which, surprisingly, **Factor Analysis** ended up giving a better prediction accuracy than both **PCA** and **ICA**.

### 3.4 - LDA

**Linear Discriminant Analysis**, or **LDA**, is a generalization of *Fisher's Linear Discriminant*. It is a well-known technique for feature extraction, and it has been widely used for such uses as facial recognition, image retrieval or microarray data classification. **LDA** focuses on the response variable classes. It projects the data onto a lower-dimensional vector space such that the ratio of the between-class distance to the within-class distance is maximized, thus achieving maximum discrimination.

Mathematically, given a data matrix, classical **LDA** aims to find a transformation that maps each column $a_i$ of $A$, for $1 \leq i \leq n$ in the $N$-dimensional space to a vector $b_i$ in the $l$-dimensional space. It creates clusters, such that the quality of each cluster is high if it is well-separated from other clusters and tightly grouped [@Klecka1980].

This technique is applied here.

After fitting and predicting, hit and accuracy values are extracted and represented in **Table 7** and **Table 8**. **LDA** yields the highest *Cohen's Kappa* score, 0.93, in accordance to the results from the **R** implementation of this technique. The fact that in both languages the best-ranking technique is **LDA** is significant, and will be explored in more depth in the final report for this project.

## 4 - Conclussions

For the time being, **no definitive conclussions** can be extracted from this data. The protocols still need to be refined for the final report of this project, so that possible artifacts and misleading parameters are avoided. If any partial conclussions have to be extracted from this data, it is that the precise area in which

Table 7: LDA observed versus predicted results

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 3632 | 61 | 18 |
| LYMPHOCYTE | 42 | 285 | 0 |
| VARIANT_LYMPHOCYTE | 11 | 0 | 397 |

Table 8: LDA observed versus predicted results - percentages

| | ATYPICAL_LYMPHOCYTE | LYMPHOCYTE | VARIANT_LYMPHOCYTE |
|---|---|---|---|
| ATYPICAL_LYMPHOCYTE | 81.69 | 1.37 | 0.40 |
| LYMPHOCYTE | 0.94 | 6.41 | 0.00 |
| VARIANT_LYMPHOCYTE | 0.25 | 0.00 | 8.93 |

these techniques are applied is of an utmost importance to the final utility of each of them, and that no generalisation (e.g. "*PCA is always the best solution*") can be made without an exhaustive application.

# 5 - References