PRINCIPAL COMPONENT ANALYSIS

A review of its theoretical foundation and a practical use case

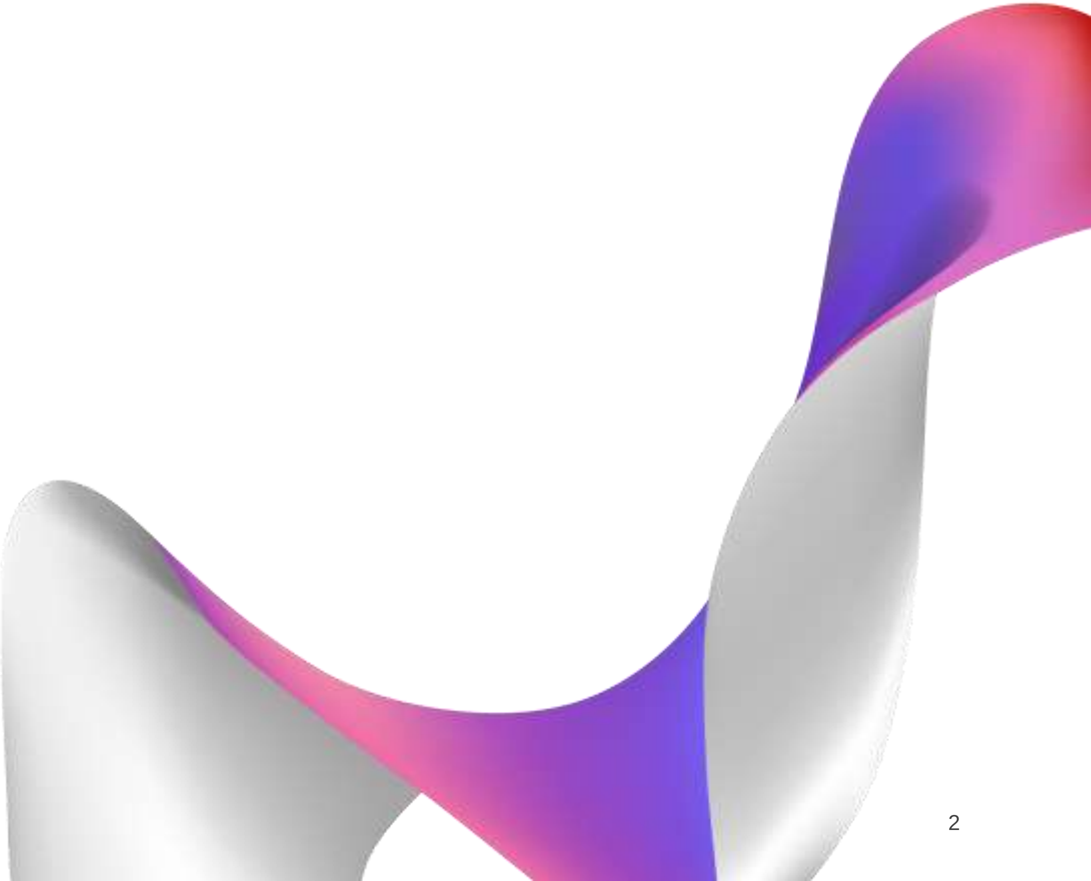
Anton Georgescu
Principal Product Architect, TEMS

December 2019

THE FUTURE OF
FINANCE IS OPEN

AGENDA

- What is Principal Component Analysis
- Mathematical foundation of PCA
- How PCA works – a set of steps
- PCA extraction
- PCA terminology: Eigenvectors vs Eigenvalues
- Practical use case with R packages
- Demo
- Other technologies that support PCA
- Conclusion



WHAT IS PRINCIPAL COMPONENT ANALYSIS

Principal Component Analysis (PCA) is a dimensionality-reduction technique that is often used to transform a high-dimensional dataset into a smaller-dimensional subspace prior to running a machine learning algorithm on the data.

PCA is mathematically defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on.

PCA, is a statistical procedure that is used prior to performing machine learning for a bunch of reasons:

- Reducing the dimensionality of the dataset reduces the number of degrees of freedom of the *hypothesis*, which reduces the risk of overfitting.
- Most algorithms will run significantly faster if they have fewer dimensions they need to look at.
- Reducing the dimensionality via PCA can simplify the dataset, facilitating description, visualization, and insight.
- Addresses the “**curse of dimensionality**” - a blanket term for an assortment of challenges presented by tasks in high-dimensional spaces;

PCA IN A NUTSHELL

In a nutshell, PCA is a transformation from a data “real rich space” into a “virtual simplified space.”

The new space can be read and analysed easier

The following slides are going to show us how we get from the real space to the virtual one.

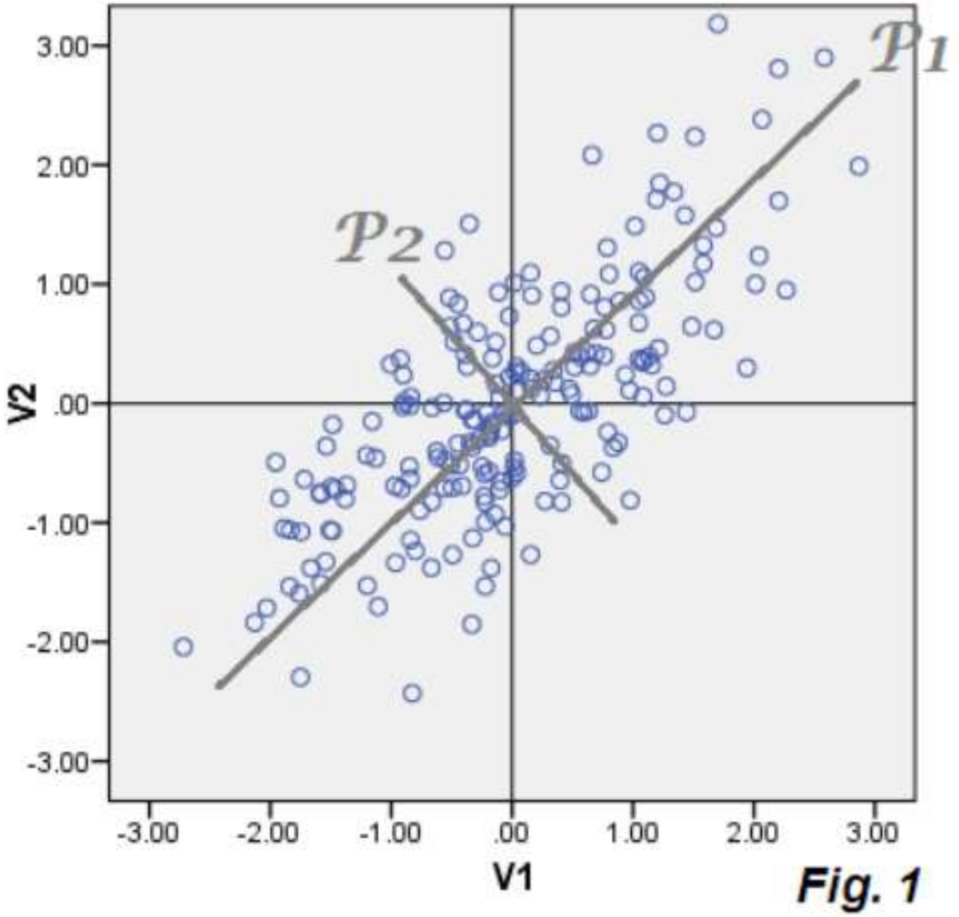


Fig. 1

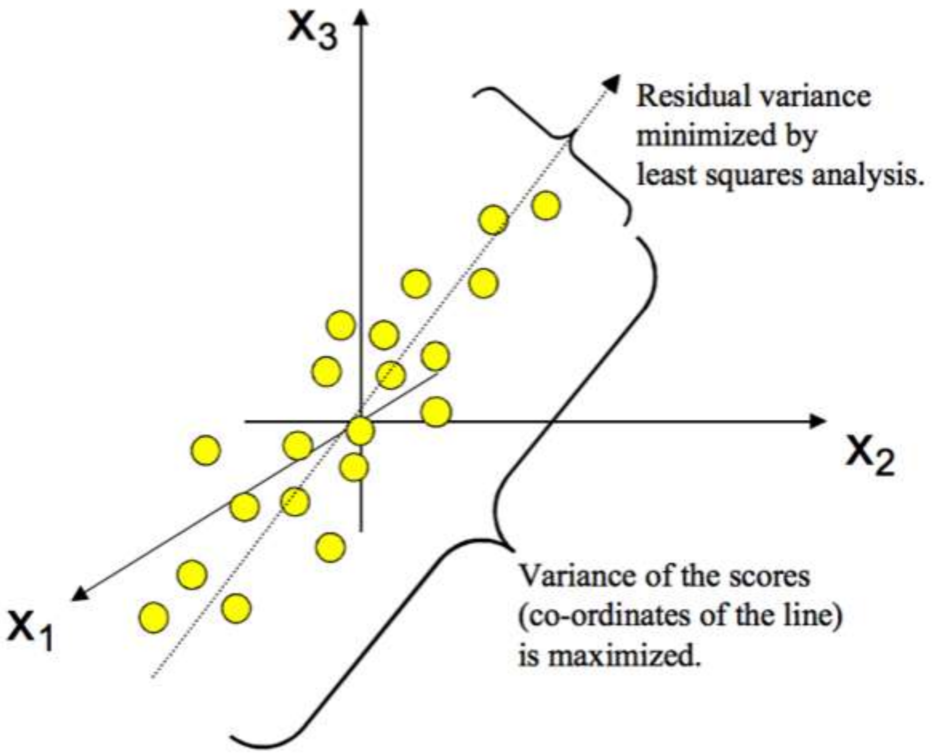
MATHEMATICAL FOUNDATION OF PCA

Statistically, PCA finds lines, planes and hyper-planes in the K-dimensional space that approximate the data as well as possible in the *least squares sense.

A K-dimensional space is where the location of a data point (observation) can be completely described with exactly k orthogonal axes.

A line or plane that is the least squares approximation of a set of data points makes the variance of the coordinates on the line or plane as large as possible.

PCA creates a visualization of data that minimizes residual **variance in the least squares sense and maximizes the variance of the projection coordinates.


X_3
X_2
X_1
Residual variance minimized by least squares analysis.
Variance of the scores (co-ordinates of the line) is maximized.

----

* The **least squares method** is a statistical procedure to find the best fit for a set of data points by minimizing the sum of the offsets or residuals of points from the plotted curve.

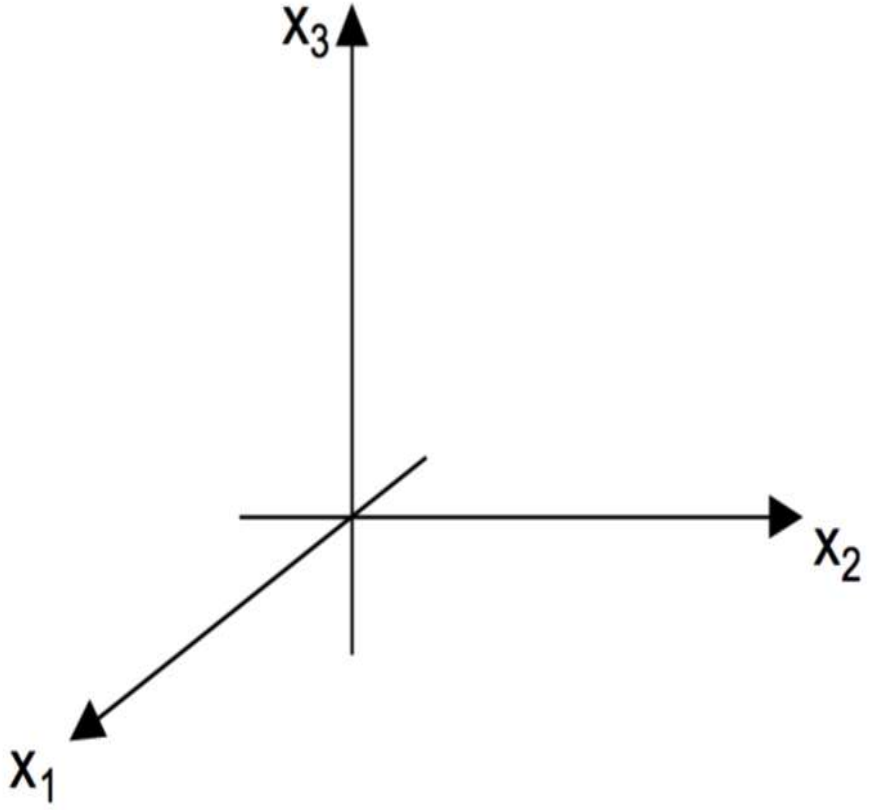
** The **variance** (σ^2) is a measure of how far each value in the **data** set is from the mean. ... Subtract the mean from each value in the **data**. This gives you a measure of the distance of each value from the mean.

HOW PCA WORKS – VARIABLE SPACE

Consider a matrix X with N rows (aka "observations") and K columns (aka "variables").

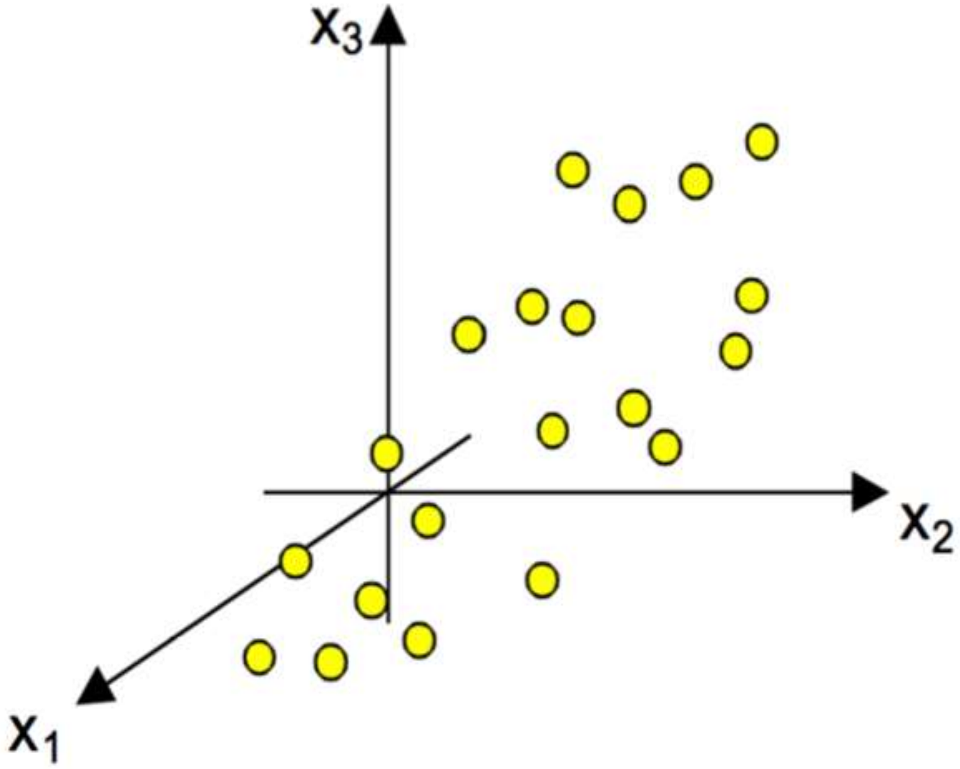
For this matrix we construct a variable space with as many dimensions as there are variables (see figure on the right).

Each variable represents one coordinate axis. For each variable, the length has been *standardized according to a scaling criterion*, normally by scaling to unit variance.



HOW PCA WORKS: OBSERVATIONS SPACE

In the next step, each observation (row) of the X-matrix is placed in the K-dimensional variable space. Consequently, the rows in the data table form a swarm of points in this space.

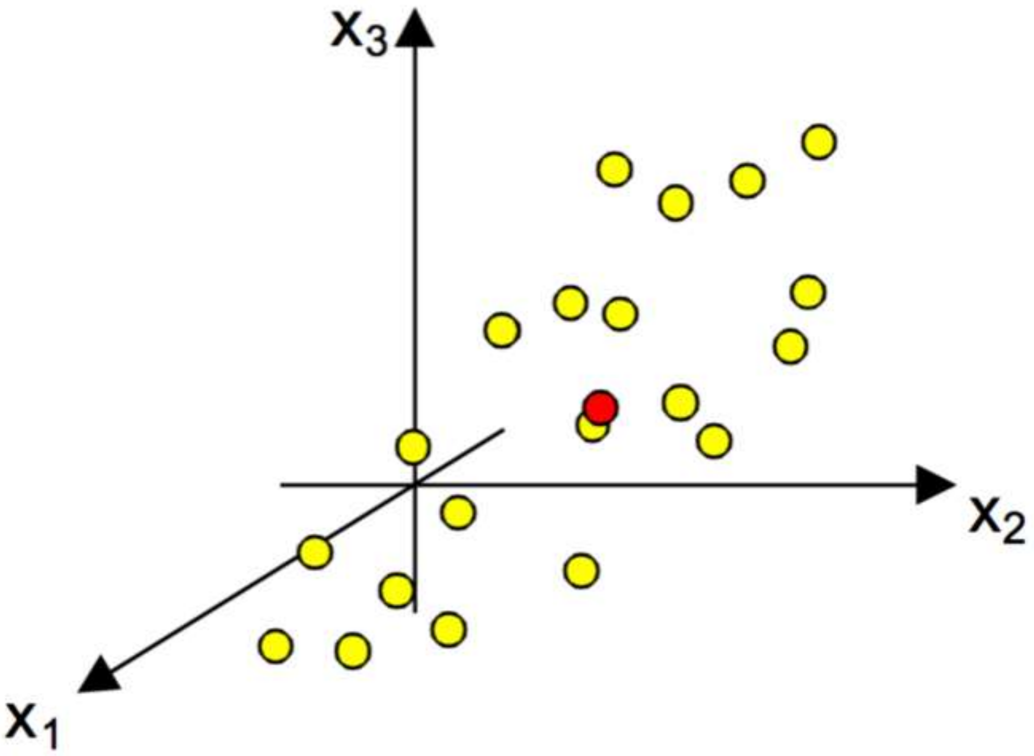


HOWS PCA WORKS: MEAN CENTERING

In this step we do *mean centering

Mean-centering involves the subtraction of the variable averages from the data. The vector of averages corresponds to a point in the K-space.

In the mean-centering procedure, you first compute the variable averages. This vector of averages is interpretable as a point (here in red) in space. The point is situated in the middle of the point swarm (at the center of gravity).

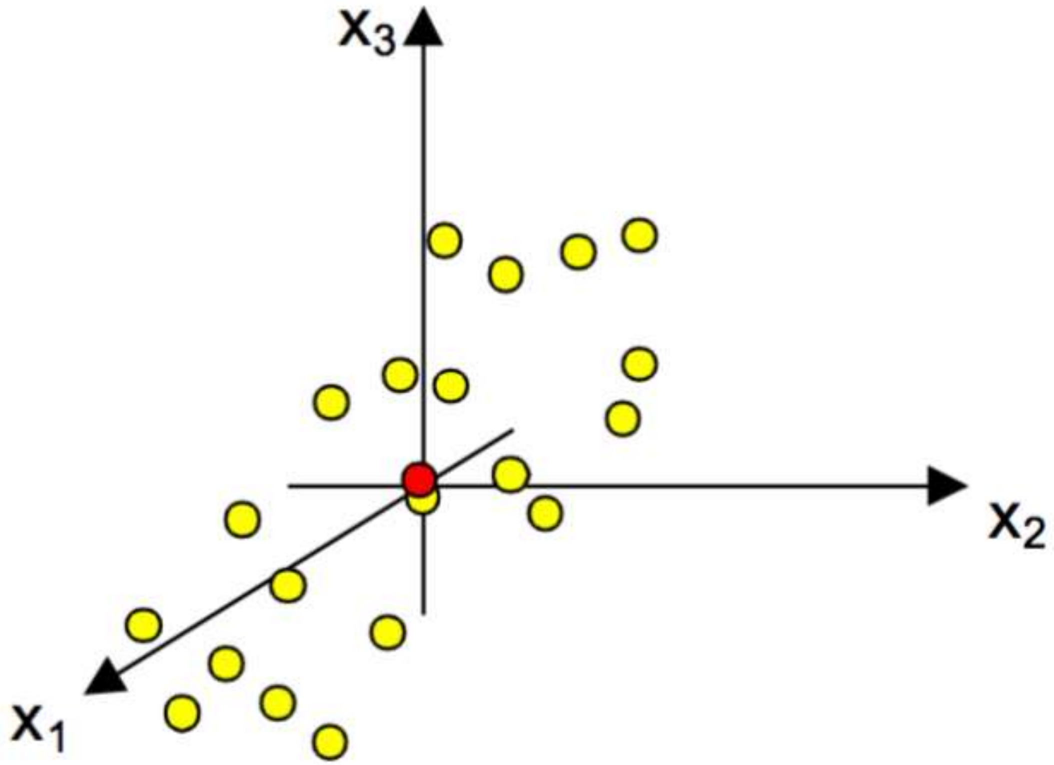


* In statistics in a general sense, to find the **mean**, add up the **values** in the data set (or population) and then divide by the number of **values** that you added.

HOW PCA WORKS: DATA REPOSITIONING

The subtraction of the averages from the data corresponds to a re-positioning of the coordinate system, such that the average point now is the origin.

The mean-centering procedure corresponds to moving the origin of the coordinate system to coincide with the average point (here in red).



PCA EXTRACTION: FIRST PRINCIPAL COMPONENT

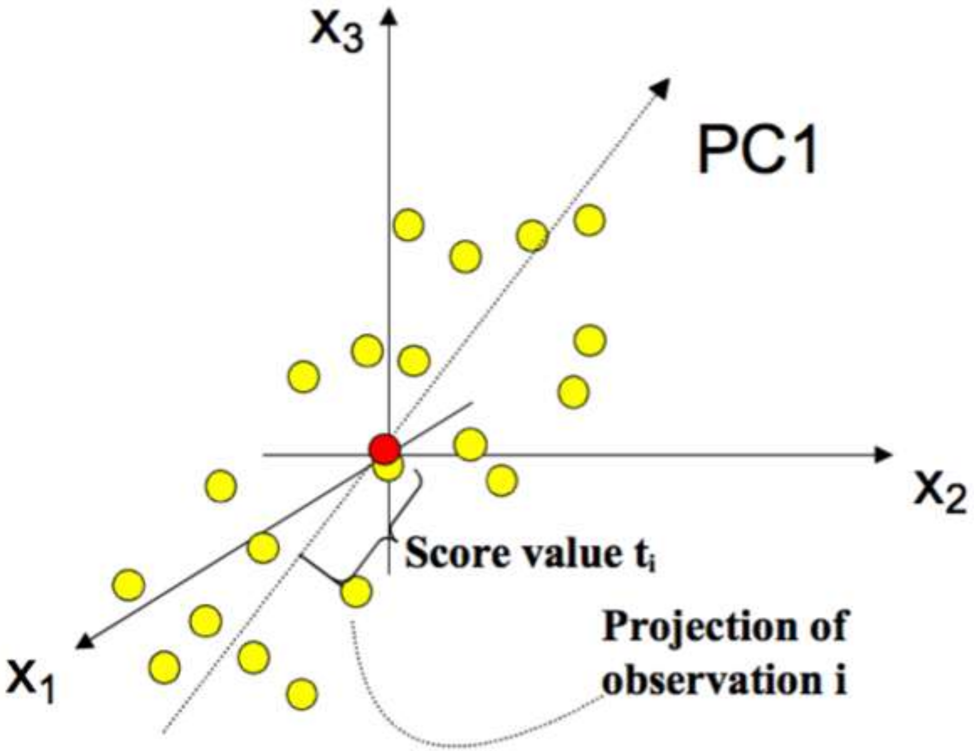
After mean-centering and **scaling to unit variance*, the data set is ready for computation of the first summary index, the first principal component (PC1).

This component is the line in the K-dimensional variable space that best approximates the data in the least squares sense.

This line goes through the average point. Each observation (yellow dot) may now be projected onto this line in order to get a coordinate value along the PC-line.

This new coordinate value is also known as the *score*.

* This function provides a data pretreatment approach called Autoscaling (also known as **unit variance scaling**). The data for each variable (metabolite) is mean centered and then divided by the standard deviation of the variable.

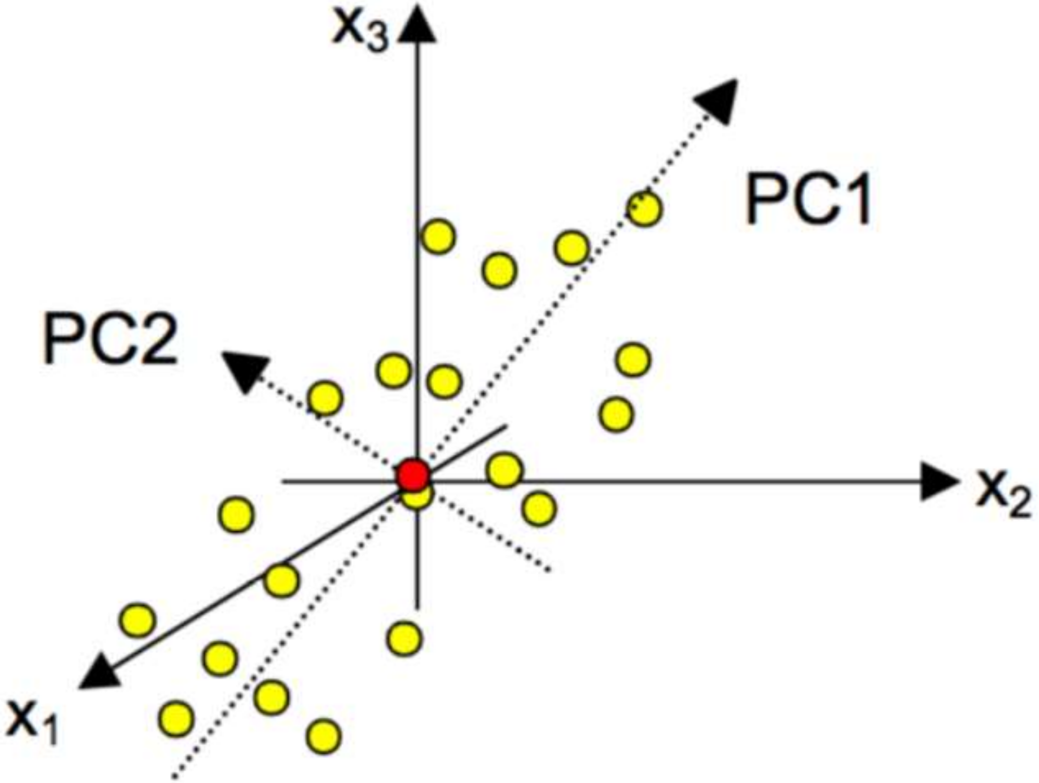


PCA EXTRACTION: SECOND PRINCIPAL COMPONENT

Usually, one summary index or principal component is insufficient to model the systematic variation of a data set. Thus, a second summary index – a second principal component (PC2) – is calculated.

The second PC is also represented by a line in the K-dimensional variable space, which is orthogonal to the first PC.

This line also passes through the average point, and improves the approximation of the X-data as much as possible.

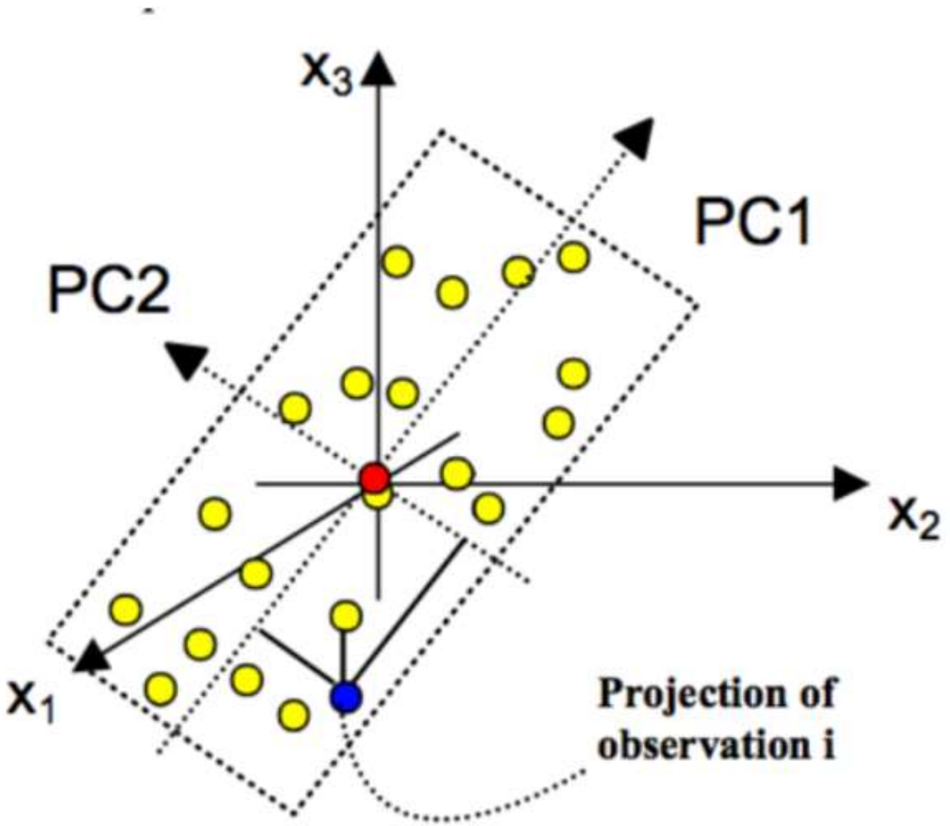


TWO PRINCIPAL COMPONENTS DEFINE A MODEL PLANE

When two principal components have been derived, they together define a place, a window into the K-dimensional variable space.

By projecting all the observations onto the low-dimensional sub-space and plotting the results, it is possible to visualize the structure of the investigated data set.

The coordinate values of the observations on this plane are called scores, and hence the plotting of such a projected configuration is known as a score plot.



HOW MANY PRINCIPAL COMPONENTS YOU NEED TO EXTRACT

There is no strict rule-based answer to how many principal components we need to extract.

In general, the data will tend to follow the 80/20 rule. Most of the variance (interesting part of data) will be explained by a very small number of principal components. You might be able to explain 95% of the variance in your dataset using only 10% of the original number of attributes. However, this is entirely dependent on the dataset.

Often, a good rule of thumb is to identify the principal components that explain 80% of the variance in the data. By projecting all the observations onto the low-dimensional sub-space and plotting the results, it is possible to visualize the structure of the investigated data set.

You cannot have more principal components than the number of attributes in the original dataset.

TERMINOLOGY: EIGENVECTORS

Mathematically, the principal components are the eigenvectors of the covariance matrix of the original dataset.

Covariance tell you how much two random variables vary together. It’s similar to variance, but where variance tells you how a single variable varies, covariance tells you how two variables vary together.

Because the covariance matrix is symmetric, the eigenvectors are orthogonal.

The principal components (eigenvectors) correspond to the direction (in the original n-dimensional space) with the greatest variance in the data.

TERMINOLOGY: EIGENVALUES

Each eigenvector has a corresponding eigenvalue. An eigenvalue is a scalar. Recall that an eigenvector corresponds to a direction. The corresponding eigenvalue is a number that indicates how much variance there is in the data along that eigenvector (or principal component).

In other words, a larger eigenvalue means that that principal component explains a large amount of the variance in the data.

A principal component with a very small eigenvalue does not do a good job of explaining the variance in the data.

In the extreme case, if a principal component had an eigenvalue of zero, then it would mean that it explained none of the variance in the data.

When PCA is used for dimensionality reduction, you will typically want to discard any principal components with zero or near-zero eigenvalues.

A PCA USE CASE WITH R AND FACTOMINER PACKAGE

The second part of this presentation will go through a practical case of PCA implementation

The subject of this use case is “Calculating the risk of student loan repayment delinquency with Principal Component Analysis”

The implementation is using two technologies:

- .NET Framework’s WinForms for easy visualization of the PCA extraction procedure and its results
- R language with a bunch of packages (most important FactoMinerR and factoextra) implementing the PCA methods

PCA FLAVOURS WITH FACTOEXTRA

factoextra is an R package making easy to *extract* and *visualize* th output of exploratory **multivariate data analyses**, including:

Principal Component Analysis (PCA), which is used to summarize the information contained in a continuous (i.e, quantitative) multivariate data by reducing the dimensionality of the data without loosing important information.

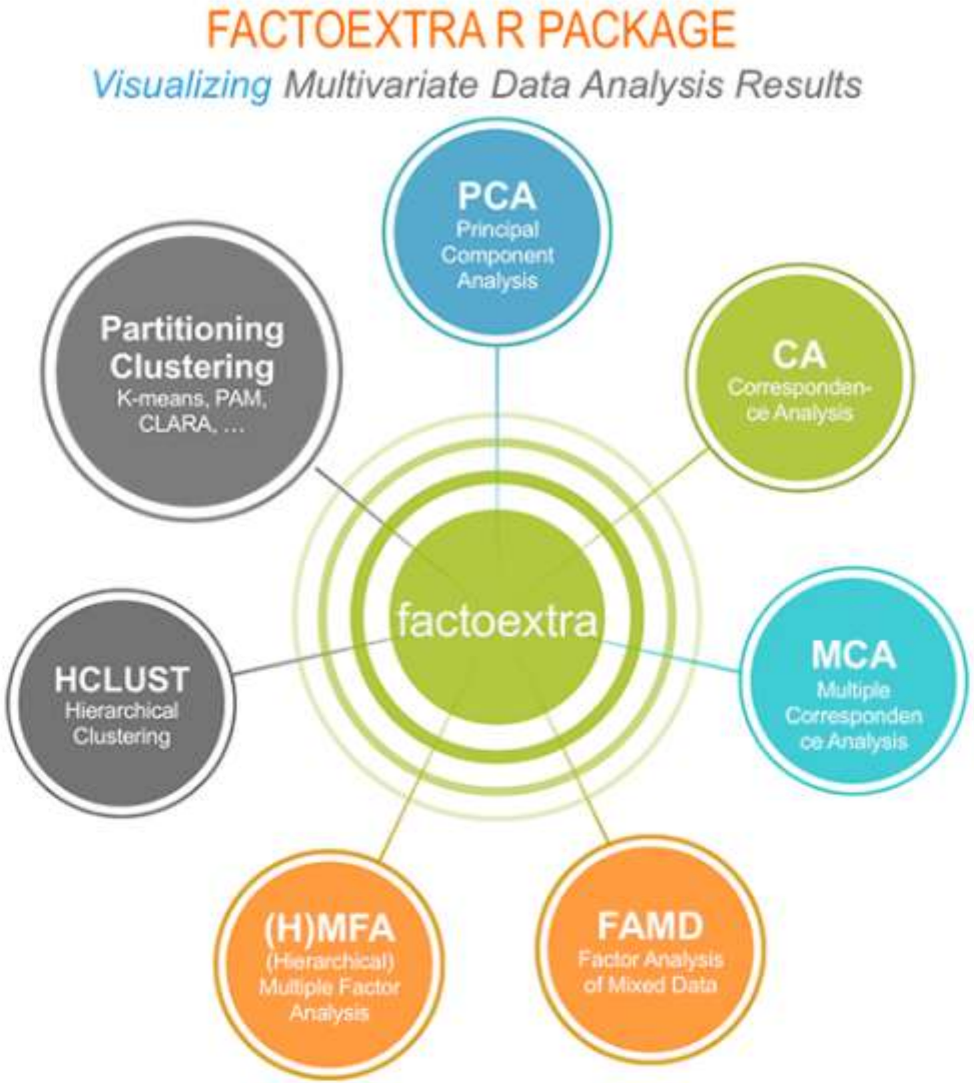
Correspondence Analysis (CA), which is an extension of the principal component analysis suited to analyse a large contingency table formed by two *qualitative variables* (or categorical data).

Multiple Correspondence Analysis (MCA), which is an adaptation o CA to a data table containing more than two categorical variables.

Multiple Factor Analysis (MFA) dedicated to datasets where variables are organized into groups (qualitative and/or quantitative variables).

Hierarchical Multiple Factor Analysis (HMFA): An extension of MFA in a situation where the data are organized into a hierarchical structure.

Factor Analysis of Mixed Data (FAMD), a particular case of the MFA dedicated to analyze a data set containing both quantitative and qualitative variables.


FACTOEXTRA R PACKAGE
Visualizing Multivariate Data Analysis Results
PCA
Principal Component Analysis
Partitioning Clustering
K-means, PAM, CLARA, ...
CA
Correspondence Analysis
factoextra
HCLUST
Hierarchical Clustering
MCA
Multiple Correspondence Analysis
(H)MFA
(Hierarchical) Multiple Factor Analysis
FAMD
Factor Analysis of Mixed Data

MORE ON FACTOMINER R PACKAGE

FactoMineR is an **R** package dedicated to multivariate Exploratory Data Analysis. It is developed and maintained by François Husson, Julie Josse, Sébastien Lê, d'Agrocampus Rennes, and J. Mazet.

It performs classical principal component methods: Principal Components Analysis (PCA), Correspondence analysis (CA), Multiple Correspondence Analysis (MCA), clustering

The packages contains advanced methods that take into account a **structure on the data** (groups of variables, hierarchy on the variables, groups of individuals).

It allows to **add supplementary informations** such as supplementary individuals and/or variables.

It provides a geometrical point of view, a lot of graphical outputs, helps to interpret (automatic description of the dimensions, various indicators, ...).

It handles missing values with **missMDA** (see here).

It has a **a GUI with a Shiny interface that draws interactive graphs** with **Factoshiny** (see here)

It gives **automatic interpretation** of the results with **FactoInvestigate** (see here).

FACTOMINER FUNDAMENTALS

When individuals are described by one set of variables, several methods are available depending on the types of variables considered (numerical or categorical variables):

- When variables are numericals one can perform a PCA (Principal components analysis).
- With a contingency table, one can perform a CA (Correspondence Analysis).
  - In statistics, a contingency table (also known as a **cross tabulation or crosstab) is a type of table in a matrix format that displays the (multivariate) frequency distribution of the variables**.
  - Contingency tables are heavily used in survey research, business intelligence, engineering and scientific research.
- When individuals are described by a set of categorical variables one can perform a MCA (Multiple Correspondence Analysis).

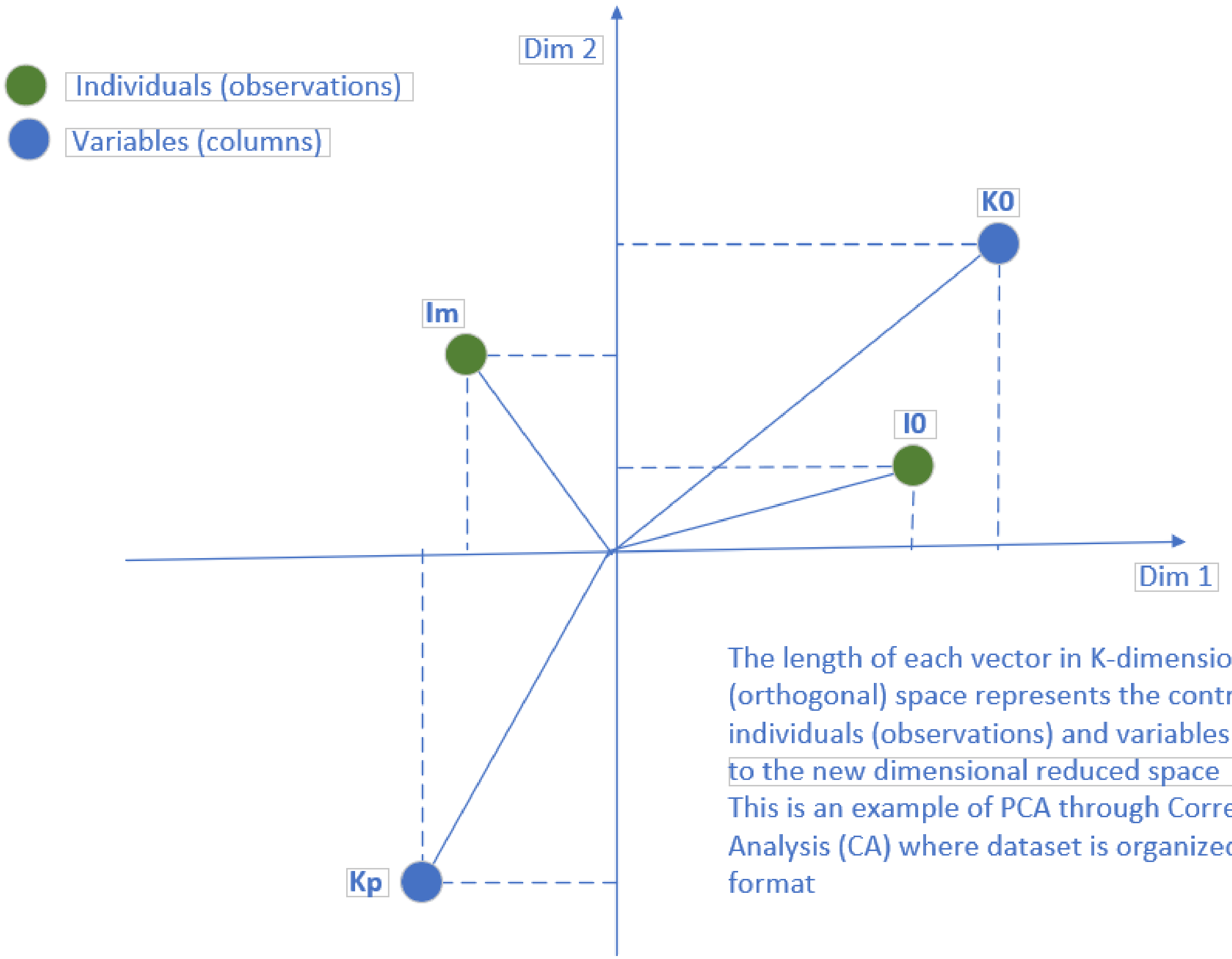
FACTOMINER – BIPLOT REPRESENTATION

- Individuals study: similarity between individuals for all the variables (Euclidian distance) → partition between individuals
- Variables study: are there linear relationships between variables? Visualization of the correlation matrix; find some synthetic variables
- Link between the two studies: characterization of the groups of individuals by the variables; particular individuals to better understand the links between variable
- There are 3 views in relating both rows (individuals or observations) and columns (variables) to PCA:
  - Rows/cols coordinates (in the new orthogonal space)
  - Rows/cols squared correlation (a.k.a. goodness of fit) – The correlation (denoted r) **measures** the amount of linear association between two variables. r is always between -1 and 1 inclusive. The R-squared value, denoted by R ^{2}, is the **square of the correlation**.
  - Rows/cols contribution - refers to the length of vector that projects orthogonally on all eigenvectors. The bigger the length, the bigger the contribution (reflected in *data inertia or variability*.)

FACTOMINER – BIPLOT REPRESENTATION

A Principal Components Analysis Biplot (or PCA Biplot for short) is a two-dimensional chart that represents the relationship between the rows and columns of a table.

Contingency tables are tools used by statisticians **when they need to make sense of data that has more than one variable**. Contingency tables are also called cross tabulation tables or cross tab. Contingency tables are displayed in matrix, or grid, form. The numbers displayed give the frequency of each data point



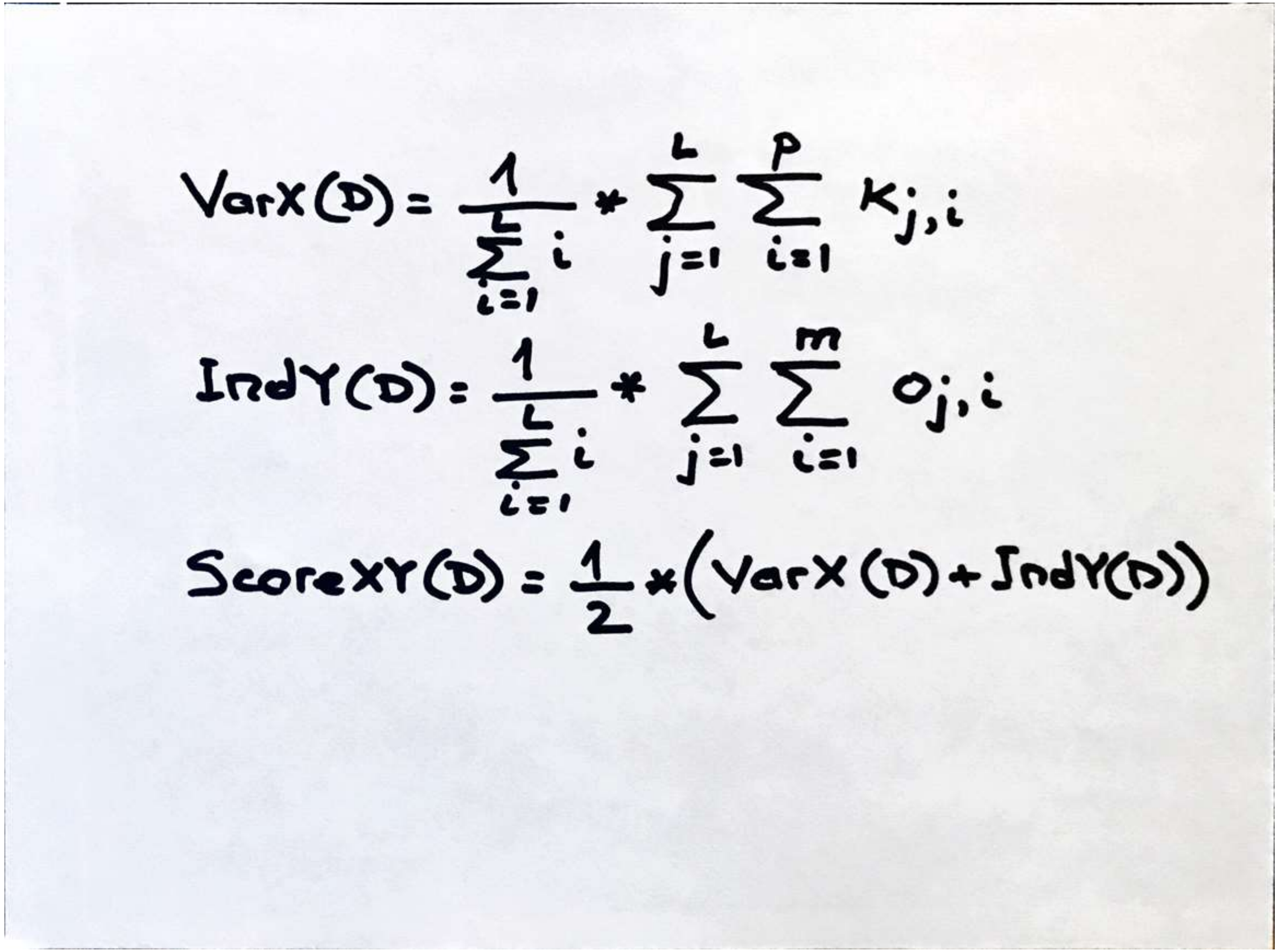
The length of each vector in K-dimensional (orthogonal) space represents the contribution of individuals (observations) and variables (columns) to the new dimensional reduced space
This is an example of PCA through Correspondence Analysis (CA) where dataset is organized in a tabular format

PCA FLAVOURS WITH FACTOEXTRA

The score calculated in the formula on the right side represents the total contributions of both individuals (observations) and variables (columns in a dataset organized in a 2-dimensional tabular format)

Since all the rows are observations linked to the phenomenon to be observed (true hypothesis) the score can be also equated to the “risk of occurrence” for the respective phenomenon

In the particular implementation we are bout to present this is the risk of having a loan going default (delinquency)



$$VarX(D) = \frac{1}{\sum_{i=1}^{L} i} * \sum_{j=1}^{L} \sum_{i=1}^{p} k_{j,i}$$

$$IndY(D) = \frac{1}{\sum_{i=1}^{L} i} * \sum_{j=1}^{L} \sum_{i=1}^{m} o_{j,i}$$

$$ScoreXY(D) = \frac{1}{2} * \left(VarX(D) + IndY(D)\right)$$

DEMO

The code under https://github.com/antongeorgescu/risk-analysis-with-pca

OTHER TECHNOLOGIES THAT SUPPORT PCA

- Python scikit-learn packages
- Accord.NET Framework (NuGet)
- .NET Extreme.Statistics.Multivariate classes
- Machine Learning Studio (part of Azure cloud offering)
- IBM Watson (part of IBM cloud offering)
- Other R packages (eg pcaMethods)
- etc.

Thank you

anton.georgescu@finastra.com

- @FinastraFS
- Finastra LinkedIn
- Finastra YouTube