OVERVIEW OF PCA

In this first Tutorial we review the main ideas behind Principal Components Analysis with no mathematics. We describe the three main steps of PCA :

* Identification of the axes on which observations should be projected for obtaining as faithful as possible a representation of the data in a low-dimension space.

* The same approach, but in the space of variables.

* Interpreting the projections. This phase is hard to formalize, and relies mostly on the analyst's know-how and experience.

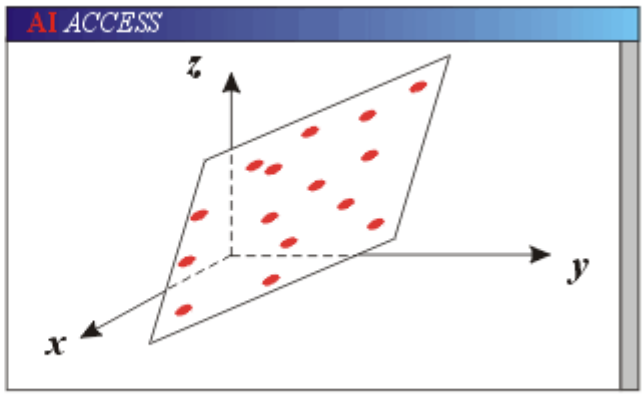
Except when explicitely mentioned, all variables are numerical and standardized (mean equal to 0 and variance equal to 1).

What does PCA do ?

An academic case

The 3 variable data set below can be represented as a "cloud" of points in a 3-D space. Upon examination, it turns out that, in this very particular case, all the points lie in a plane. We know that a plane is not truly 3D, but really just 2D : two coordinates (and not 3) are enough to locate a point unambiguously on a plane.

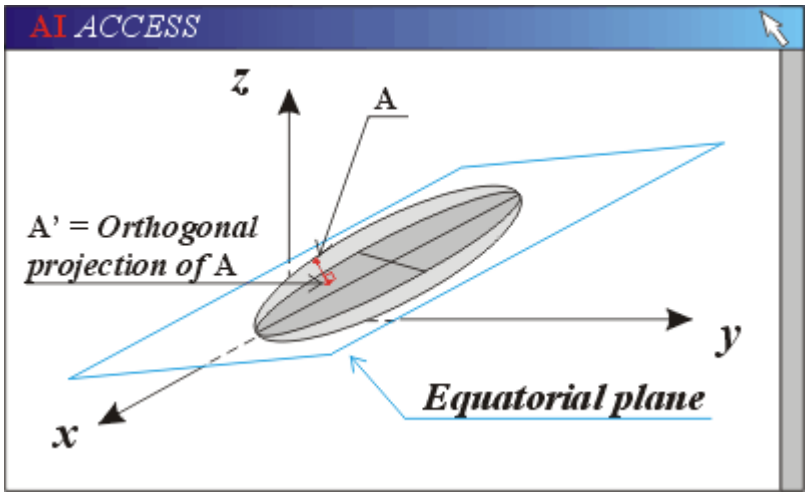
	x	y	z
1	0.453	1.297	0.846
2	0.056	2.052	1.583
3	1.027	0.065	0.739
4	0.995	1.065	2.981
----	-------	-------	-------
50	2.857	0.673	1.072



The conclusion is that our data does not **need** 3 variables to be fully described : 2 would have been enough. In fact, any pair of axes in the plane would do the job.

A barely more realistic case

Suppose that, instead of lying exactly in a plane, the data cloud had the shape of a flat pancake. Strictly speaking, we now do need 3 coordinates to identify a point unambiguously.



But consider the equatorial plane of the pancake. If we project all data points on this plane, we obtain a 2D representation of the cloud. And because the cloud is nearly flat, this representation is reasonably faithful. At any rate, it is certainly more faithful than would have been a 2D representation obtained by selecting any pair of the of the original 3 variables, and plotting the data set using these 2 variables only.

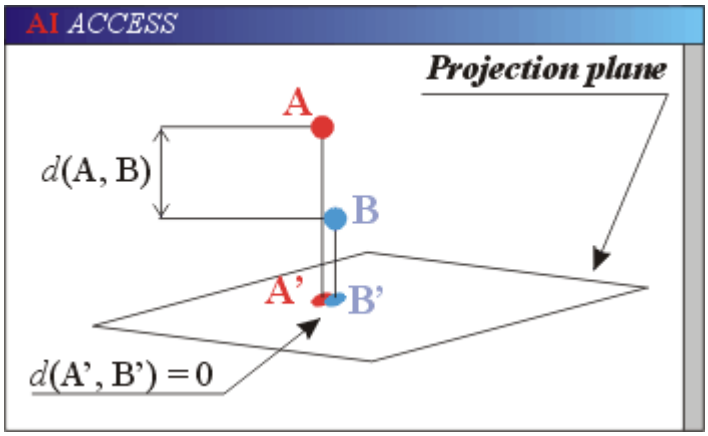
This remark is what originated PCA : if we want a 2D representation of a high dimensional data set for visual inspection, then not all planes will be equally suitable for that purpose : some planes will give a more faithful 2D picture of the cloud than some others. PCA will identify the "best" plane. In fact, as we will shortly see, PCA does a lot more than that.

What is a "faithful" representation ?

The first thing we have to do is to define more precisely what we mean by a "faithful 2D representation" of a cloud.
Let :

- A and B be any two points, and $d(A, B)$ their distance,
- and A' and B' their projections on a plane, and $d(A', B')$ their distance.

Here, "distance" usually means "Euclidian distance". If the coordinates of two points are very similar, then their distance is very small, and *vice-versa*.



Certainly, if we could achieve $d(A, B) = d(A', B')$ for **any** pair of points, we would consider the projection as a perfectly faithful 2D representation of the cloud. Clearly, this is not possible, so we have to accept the idea that our 2D map will create some distorsion. More specifically, some points may be separated in the original space, and yet have their projections on top of each other (zero distance).

So our goal is to identify the projection plane that will create the least distorted projection map of the cloud. This can be expressed mathematically in a somewhat cumbersome form. Fortunately, this condition turns out to be equivalent to another condition that is quite simple to express, and that we discuss now.

The "best" projection plane

In most applications, the original variables are standardized before running a PCA :

* They are centered, so that each variable has now a 0 mean,
* Each variable is squeezed or expanded so that its variance is now 1.

The origin of the new reference frame is now the barycenter G of the cloud.

The Principal Components

Take any straight line D_1 going through G, and project all observations on D. How much this set of projections "stretches" on D_1 is measured by its variance. The value of this variance is usually called the inertia of the cloud with respect to D_1.

Of all possible D_1s, pick the one with the largest value of this variance, and call it the **First Principal Component**.

Now, of all the straight lines D_2 **orthogonal** to D_1, pick the one with the largest variance for the set of projected observations, and call it the **Second Principal Component**.

The Principal Plane

D_1 and D_2 together define a plane, sometimes called the Principal Plane. This is the plane we were looking for, that is the plane that provides the most faithfull 2D representation of the original, many-dimensional cloud.

We will later generalize the Principal Plane to projection subspaces of higher dimensions.

A dual approach : PCA on variables

So far, we considered the rows of our data table as "observations", and the "columns" as variables. But the mathematical machinery behind PCA does not know "observations" and "variables", it only knows rectangular tables of numbers. So we could conceivably transpose our table, so that "rows" become "columns" and *vice versa*, and then feed the new table to the PCA algorithm. Does that make any sense ?

As a matter of fact, it does. We would then obtain a 2D plot where "points" are not the observations anymore, but rather the original variables. This plot is the "best 2D representation" of the set of variables, in a sense that will require some explaining.

The original data table and its transpose contain exactly the same information. So there has to be some sort of connection between the "observations plot" and the "variables plot". Of course, this connection exists, but it is somewhat tricky, and should be exploited with care.

Interpreting a PCA

Obtaining the two above mentioned plots is quite mechanical. Now, the practitioner may start **interpreting** these plots. This requires both practice, and a detailed understanding of the logics behind PCA. We will later develop the most important aspects of PCA interpretation :

1) Assessing the global and individual qualities of the representations of observations and of variables.

2) Interpreting the Principal Components in "business terms". That will usually require taking both plots (observations and variables) into consideration.

3) Considering other projection planes than just the Principal Plane.

4) Using additional observations and variables to further refine the analysis.

Other applications of PCA

By far the most widespread use of PCA is visual interpretation of high dimensional data. Yet, because its principles are so general, PCA has many other uses besides exploratory data analysis. We'll shortly mention :

* Data Compression and Data Reconstruction

* Data preprocessing before using some data modeling technique for Clustering, Regression or Classification.

INERTIA AND PROJECTION OF OBSERVATIONS

We now describe how the "best" projection subspaces are identified for projecting the cloud of individuals. These subspaces are **nested** : the best k-dimensional subspace is inside the best subspace of dimension k' for any $k' > k$.

We then explain why all observations are not equally well represented in a low-dimension projection subspace, and identify :

* The observations whose projections are reliable,
* And the observations that are particularly influential in defining the projection subspaces.

The concept of "inertia"

Inertia of a point

Let's forget PCA for a moment. In a plane, let O be a reference point, and A any point. By definition, the inertia of A with respect to O is simply the square of the distance d(A, O).

Inertia of a cloud of points

Again, let O be a reference point. Then, by definition, the inertia of the cloud with respect to O is simply the sum of the inertias of the points with respect to O.

The value of the cloud inertia depends on O. It can be shown that the cloud inertia reaches its smallest value when O is the center of gravity (or "barycenter") of the cloud. In what follows, O will be this barycenter, that we will call G.

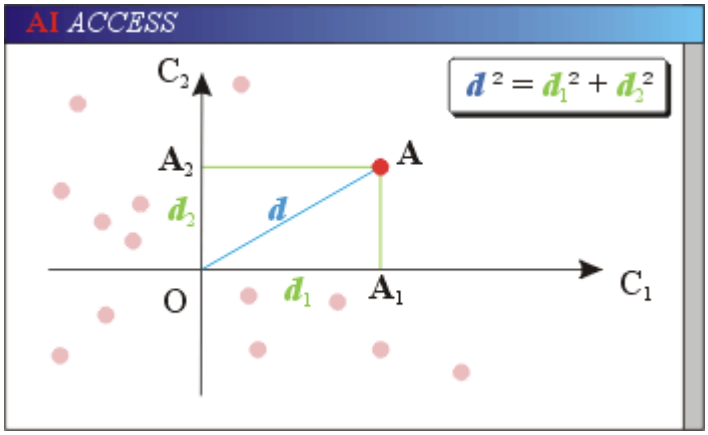
- When variables are standardized, it can be shown that the **inertia of a cloud with respect to its barycenter G is equal to the number of variables**.

Decomposition of inertia

Suppose we have a set of points in a plane, and let C_1 and C_2 be any pair of orthogonal straight lines in that plane going through point O. Let A be any point in the plane, and let :

* A_1 be the projection of A on C_1,

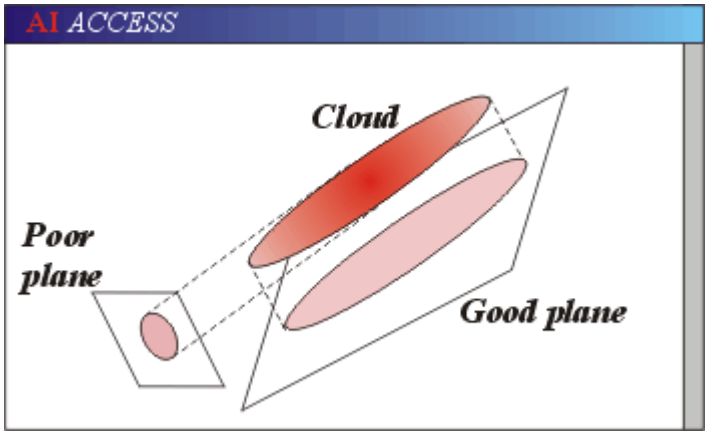
* A_2 be the projection of A on C_2.



A_1 and A_2 have their own inertias with respect to O. Then good old Pythagorean theorem tells us that the inertia of A is the sum of the inertias of A_1 and A_2. Therefore, the inertia of A can be "decomposed" along any pair of orthogonal lines.
The same applies for the inertia of a set of points. The sum of the inertias of the points on C_1 will be called the "inertia carried by C_1" or "inertia of C_1". The total inertia of the cloud is the sum of the inertias of each of the orthogonal lines.

In fact, this very important property is more general than that. If A and O are now in a space of any dimension *n* (and not just a plane) fitted with any set of *n* mutually orthogonal lines (axes), then it is still true that the inertia of A (with respect to O) is the sum of the inertias of each of the orthogonal lines.

Maximizing the projected inertia



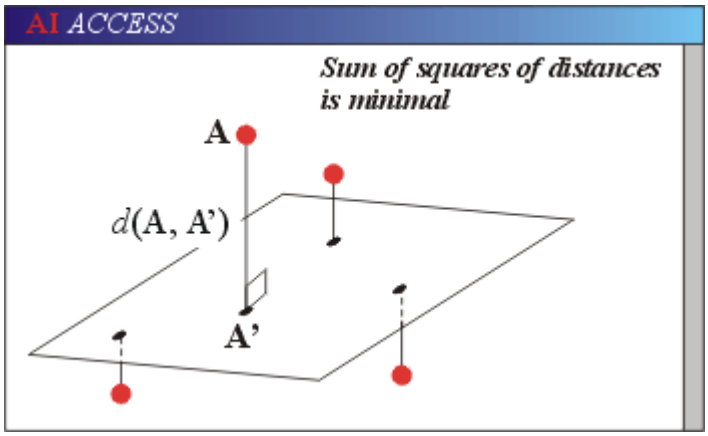
Back to PCA. It can be shown that the projection plane that will suffer the least amount of distorsion is that for which the inertia of the projected cloud with respect to its barycenter G is **largest**. It can be said

that this is the plane for which the projection of the cloud "spreads out" the most.

So the plane we are looking for is that for which the inertia of the cloud projection (with respect to G) is largest. We will call it the **Principal Plane**. In what follows, we will often use the word "best", and the meaning will always be "that maximizes the inertia of the cloud projection".

Minimizing the spread around the best plane

We will give now give another definition of "best", but that is equivalent to the one above.



Remember the image of a flat, pancake-like cloud. Our intuition told us that the best projection plane was the equatorial plane of the pancake. A consequence of this choice is that, in real space, points are never far away from this plane.

Let A be any point, and A' its (orthogonal) projection on a plane (any plane), and let's call d(A, A') the distance from A to A'. Let's now sum the squares of all such distances from a point to its projection. It can be shown that this quantity is **minimal** for the Principal Plane.

So the Principal Plane is also the plane that "fits" the data cloud best in the above sense.

The Principal Components

So far, our goal was to obtain the most faithful 2D representation of a cloud for visual examination purposes. We also restricted ourselves to a data set described by 3 variables only, so as to make the projection process accessible to the imagination. We now turn to the most general case :

- with a large number n of variables,
- and the choice of a "best" set of p variables, with $p < n$.

The special case $p = 2$ will answer our original problem (2D projection).

The First Two Principal Components

Let us consider a cloud in an n-dimensional space, and let G be its **barycenter.** Of all the straight lines going through G, one (and only one) makes the inertia of the line with respect to G largest. This is the **First Principal Component**. We will call it C_1. For mathematical reasons, the inertia carried by C_1 is called the first **eigenvalue** of the analysis.

More precisely, the amount of inertia carried by the first Principal Component is the highest value eigenvalue of the data correlation matrix.

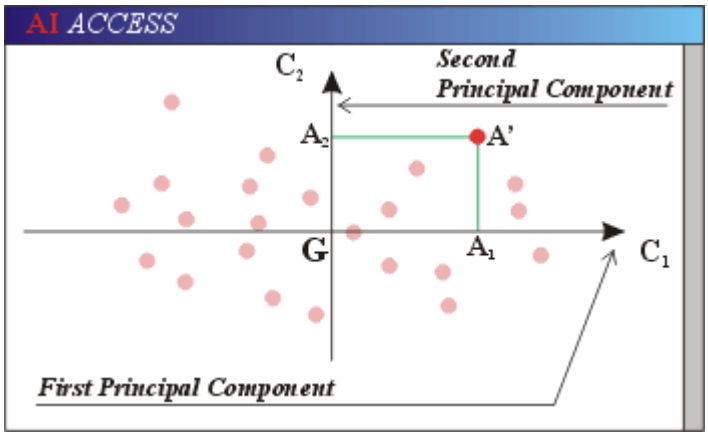
Of all the straight lines going through G and orthogonal to C_1, one (and only one) makes the inertia of the line with respect to G largest. This is the second Principal Component, and we call it C_2. Of course, the inertia of C_2 is less than that of C_1, (which means that the cloud spreads out less along

C_2 than it does along C_1). It is the second Principal Component, with the second largest eigenvalue of the analysis.

Now to the magics of PCA. C_1 and C_2 define a plane. It can be shown that this plane is just the Principal Plane, that is the best plane for data projection. In other words, we identified the best plane by :

* First identifying the "best line" (C_1),

* and then identifying the best line (C_2) orthogonal to C_1.



C_1 and C_2 make up a reference frame for the Principal Plane. Any point A of the original cloud projects on the Principal Plane as A', which in turn may be projected on C_1 and C_2 as A_1 and A_2. The abscissas of A_1 and A_2 may be understood as the **coordinates** of A' in the Principal ¨Plane.

All the Principal Components

What if we go on ? The magics goes on too. We now define C_3, the third Principal Component, as the line orthogonal to C_1 and C_2 (and going through G) that makes the inertia of C_3 largest (but still less than that of C_2). And the 3D space that is generated by (C_1, C_2, C_3) turns out ot be the best 3D space, in that no other 3D space will yield a projected cloud with a higher inertia (or less distance distorsion).

All applications of PCA rely entirely on this property : once you have identified the best *p*-dimensional space, then the best (*p*+1)-dimensional space on which to project the data set is just and extension of the best *p*-space, augmented by the $(p+1)^{th}$ Principal Component. In a bit more technical terms, best *p*-dimensional spaces are **nested** in one another.

From now on, we will write note Principal Components "PC" for short.

What have we gained ?

Our original goal was visual examination. So why should we bother identifying best 3D, 4D etc... data projections ? After all, if we pursue the process all the way to the end, then we have simply subsituted a new set of *p* coordinates to the original set of *p* coordinates. Besides its cute "optimality" property, what is this new set of coordinates good for ?

We'll see that PCA has uses other than data visualization. In general terms, it may be used any time one wants to replace a large set of variables with a smaller set of (new) variables, and yet lose as little information as possible in the process. This is called "dimensionality reduction", and is a key step in obtaining stable and reliable models. We'll come back to that later.

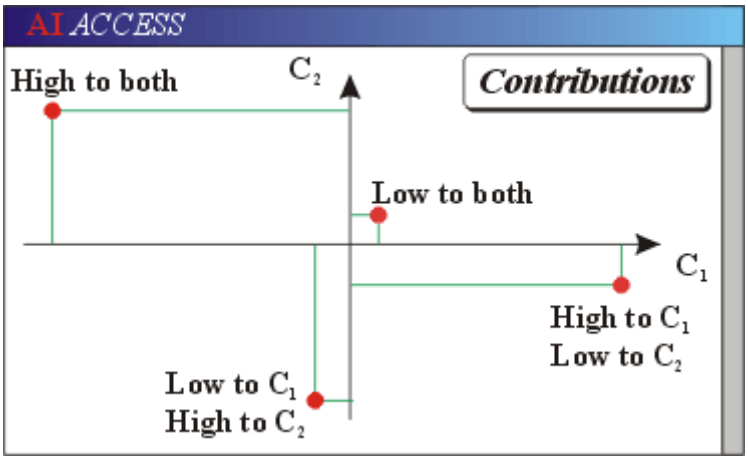
Projection of the observations

The barycenter

Remember that we made all the Principal Components go through the barycenter of the cloud. Therefore, in the Principal Plane, the origin represents a perfectly "average" observation" : in the original frame of reference, any of its coordinates is equal to the average of the same coordinate on all the observations.

As you move away from the origin, you meet observations that differ more and more from the average. One of the main goals of PCA is to identify directions away from the average observation that can be described in "business terms". Of course, interpreting the meaning of the Principal Components will be on top of the list.

Contribution of an observation to a Principal Component



Remember that C_1 is the component with the largest inertia, which is equal to the sum of the inertias of the projections of the observations on this component.

* Some observations project "far away" on the axis, hence contribute a great deal to the inertia, and therefore have a strong influence on defining the orientation of C_1.

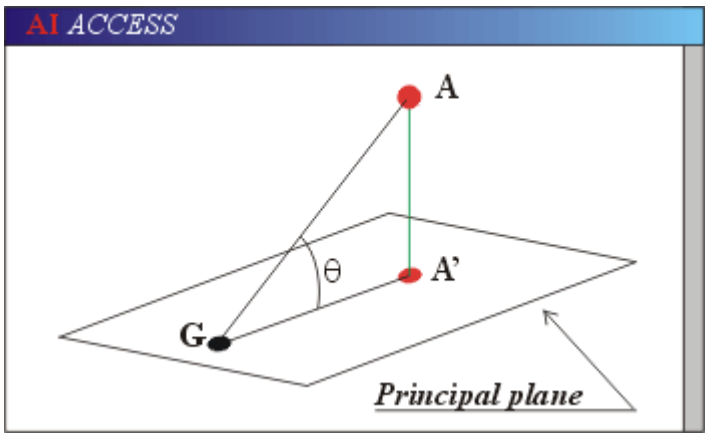
* On the other hand, observation whose projections on C_1 are close to the origin have very low inertias with respect to C_1.

How much influence a particular observation has on a Component is measured by the fraction of the inertia carried by this Component that is due to the observation. It is called the **Contribution** of this observation, and is usually displayed by software. Observations greatly contributing to a Component will be dutifully identified during the interpretation phase. But an observation may have a high contribution to a Component, and yet have a negligeable contribution to another Component.

Generalization of the concept of "Contribution to the inertia of a PC" to that of "Contribution to the inertia of the Principal Plane" is straightforward, because the inertia of a point in the plane is just the sum of its inertias with respect to C_1 and to C_2.

Quality of representation, "Squared Cosine"

Is the distance between the origin G and the projection of a point in the Principal Plane a good approximation of the true distance of the observation to the origin ? Of course, "it

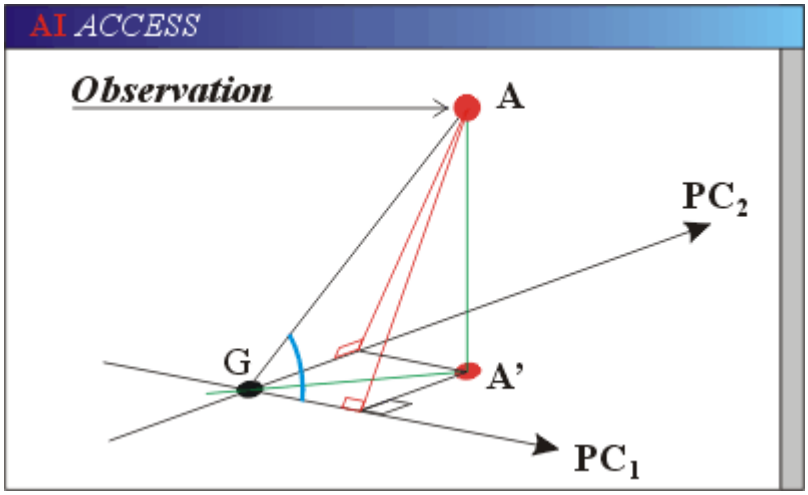


depends". For a given projection (with a fixed distance to G) a point may have any distance above that plane, and therefore its true distance to the origin may be arbitrarily large. Only those points that are close to the plane have their distance to the origin represented reasonably faithfully.

It is not customary to display the distance of an observation to the Principal Plane. Rather, software often posts the ratio of the projected distance to the true distance (to G). This quantity is just the square of the cosine of the angle between :

- the straight line going through the origin and the observation in the complete space.
- and the straight line going through the origin and the projection of the point on the Principal Plane.

It is called the "**Squared cosine**" of the projected point. Only observations with a high Squared Cosine (i.e., close to 1) are well represented on the Principal Plane.



But the converse is not true : just because a point on the Principal Plane is close to a factor does not necessarily mean that it is close to it in the complete space. So when you see an observation close to a PC on a plot, check out the squared cosine of the observation with the PC before you believe what you see.

Why use the squared cosine, instead of just the cosine ? It is because this definition extends easily to any Principal subspace. For example, the Squared Cosine of a projected point in the Principal Plane is just the sum of the Squared Cosines of this same point over the first two PCs. In this illustration, we show in blue the angle whose squared cosine represents the quality of the representation of the observation by the first PC. Note that this is **not** the (squared) cosine of the projection of GA' with PC_1 (this last quantity has no meaning).

Software always displays the squared cosine of observations with every PC.

One is often tempted to intepret the distance between two points on a PCA plot as a genuine measure of their true distance. We already insisted on the fact that this "projected distance" may be quite misleading. Yet, when **both** points have high squared cosines, their distance on the plot is a good approximation of their true distance in space.

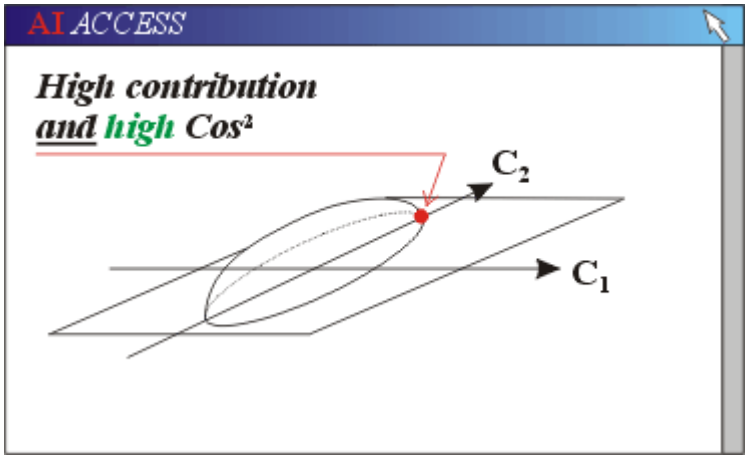
Are "high Contribution" and "high Squared Cosine" equivalent ?

Although interpretation of a PCA insists on :

* observations that contribute much to the inertia (of a PC, or of the Principal Plane),

* observations that are well represented on a PC, or on the Principal Plane (Squared Cosine close to 1),

these two concepts are not equivalent.



Take a point that contributes much to the inertia of the Principal Plane. It is far from the origin G, and therefore at the periphery of the cloud. Its high contribution to the inertia makes it an important point. Does that imply that the point is also well represented ?

It will be the case if (and only if) the point's Squared Cosine is high, that is if the point is close to the Principal Plane. Far out points being close to the plane mean that the cloud has a roughly convex shape , a bit like a multidimensional football (top illustration). Certainly, multinormal distributions do meet this condition, and it is quite common for real life clouds to assume a somewhat convex shape.

But this is not a strict rule, and it is advisable to identify those points that have a large contribution to the Principal Plane (or to a Component), and yet have a low Squared Cosine.

- If there are many such points, then the cloud has a funny "diabolo" shape, and it is dubious that PCA interpretation can be very meaningful (bottom illustration).
- If there are only few of these points, then they are " outliers". As usual, outliers should clearly be identified :
    - either as *bona fide* exceptional observations,
    - or as observations whose attributes are corrupted by some kind of error. They should then be removed before another PCA is attempted, as PCA is quite sensitive to outliers.

Points around the center usually have both high and low squared cosines because the cloud usually spreads continuously across the Principal Plane.

PCA ON VARIABLES

The analyst is just as interested in variables as in observations. In particular, it is expected that analyzing data should allow the easy discovery of groups of variables that are strongly pairwise correlated. Such groupings may be detected by a cautious and tedious examination of the correlation matrix of the data, but Princpal Components Analysis allows the detection of such groupings visually.

For this purpose, the same rearch that was conducted in the space of observations is now conducted in the **space of variables**, which is sort of dual of the space of observations. Variables will the be represented as points in projection planes, and, provided that the quality of the projection is satisfactory, close "variable points" will represent strongly correlated variables. Anti-correlated and uncorrelated pairs of variables may also be visualized, and therefore easily detected.

The space of variables

Consider a set of *n* observations described by *p* (numerical) variables. It is quite natural to interpret this data set as a cloud of *n* points in a *p* dimensional space, and it's just what we did since the beginning of this tutorial. The coordinates of the points are placed in a table that we call X. The maths behind PCA will work on this table.

In fact, PCA will work on the data correlation matrix, that can be readily obtained from X. Software usually allows PCA to accept either data table or data correlation matrix as input.

Now transpose the table so that the "old" lines become the "new" columns and *vice versa*.

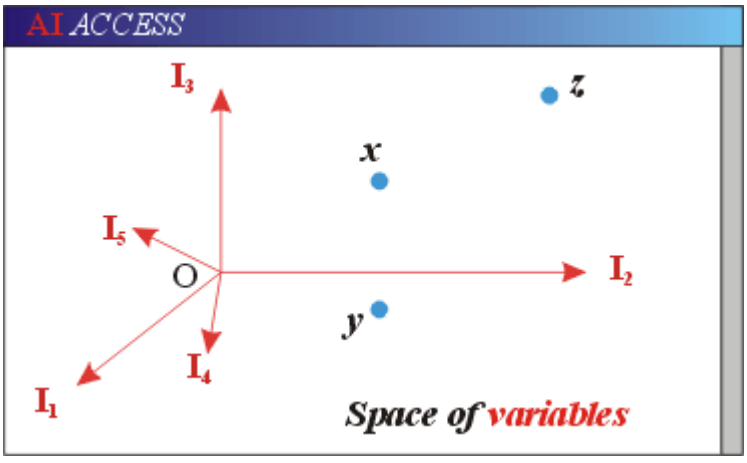
	x	*y*	*z*
I_1	0.243	-1.65	-----
I_2	2.564	-----	-----
I_3	-----	-----	-----
I_4	-----	-----	-----
I_5	-----	-----	-----
I_6	-----	-----	-----

→

	I_1	I_2	I_3	I_4	I_5	I_6
x	0.243	2.564	-----	-----	-----	-----
y	-1.65	-----	-----	-----	-----	-----
z	-----	-----	-----	-----	-----	-----

We now define a new space :

- Its dimension is the number of observations in the data set (*n*).
- Each dimension is labeled with the "name" of one observation.



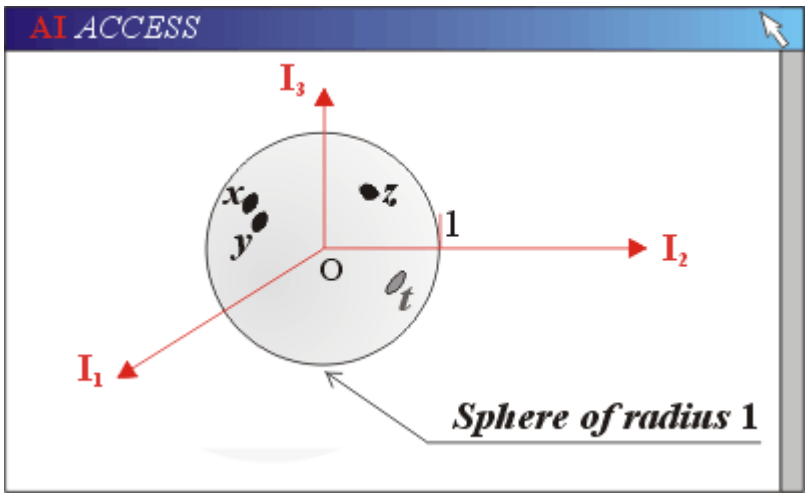
This space is called the "**space of variables**". It depends on the sample. By contrast, we'll call the "ordinary" space the "**space of observations**". So, observations sit in the "space of variables" and variables sit in the "space of variables".

Now take a variable, and list the values that this variable takes over the sample. There are *n* such values, just the dimension of our new space. Therefore, this variable may be plotted as a point in the space of variables. If you do that for every variable, then you end up with a cloud of *p* points in a *n*-dimensional space. So by transposing the original data table, we have constructed a new representation of our sample.

In general, your sample has many more observations than variables. Therefore, the cloud in the variables space has far fewer points than the space it lives in has dimensions.
On the bottom illustration (space of variables), all axes (observations) are to be understood as being orthogonal to each other.

Why the space of variables ?

Why bother with this new representation, as it obviously brings about no new information about our data set ? The reason is that some properties of the original variables have a nice **geometrical** interpretation in the space of variables. We assume that variables have been standardized (0 mean and unit variance).



* First, all variables being standardized, all variable-points are at distance "1" form the origin. Therefore, all the *p* points lie on a "hypershere" of radius 1.

* It is very easy to show that the observed value of the **correlation coefficient** between two variables is just the **cosine** of the angle θ between the two lines going from the origin O to their respective representative points (bottom illustration). So :

- If two points (representing two variables) are very close to each other, then the two variables are strongly positively correlated (correlation coefficient close to +1).

- If two lines joining the origin O with two points are orthogonal, then the corresponding two variables are uncorrelated (correlation coefficient close to 0).

- If two points are symmetrical with respect to the origin O, then the two variables are strongly negatively correlated (correlation coefficient close to -1).

"Distance" between variables

We introduced PCA as a way of projecting a cloud of points on a plane so that relative distances between points would be as closely respected as possible. By "distances", we meant the ordinary Euclidian distance.

We are now facing a new cloud of points. How should we define the distance between two points (variables) ? The following quantity :

$d(\boldsymbol{x}, \boldsymbol{y}) = 2.(1 - \cos^2(\theta))$

has a satisfactory behavior :

* It is "0" when the variables have maximum correlation, either positive or negative,

* and it is largest when the two variables are uncorrelated.

It is this definition of "distance" that PCA uses when working on variables.

PCA on variables

Much of what we said about PCA in the space of observations is still valid in the space of variables. Now that we have defined a suitable distance between points in the space of variables, we may unambiguously define the "best" plane on which to project the cloud of variables.

The Principal Components for variables (or "axes")

Inertia of the cloud of variables

The cloud of variables has a certain intertia. It is easily shown that this inertia is equal to the number of variables, and is therefore the **same** as the inertia of the cloud of observations.

PCA on variables define Principal Components in the space of variables (that we will call "axes"), just as it did in the space of observations. Each axis will carry as much inertia as possible, under the constraint of being orthogonal to all the previously defined axes.

The first two axes

On the best projection plane, the first two axes, call them F_1 and F_2, make an orthogonal reference frame.

F_1 is the unique synthetic variable that represents best the set of the original variables in the following sense :

It is the variable for which the sum of the squares of the correlations with all other variables is maximal.

* It can be shown that the amount of inertia carried by F_1 (largest "eigenvalue") is the **same** as the amount of inertia carried by C_1, the first PC in the space of observations.

F_2 is the variable that is :

* orthogonal to F_1,

* and that maximizes the sum of the squares of its correlation coefficients with all the original variables.

As for F_1, F_2 carries the same amount of inertia as does C_2, the second PC in the space of variables.

The other axes

Higher order axes may be determined in the same manner. But now, we are facing a major difference with what we had with PCs in the space of observations.

* First, remember that it is very likely that you have many more observations than you have variables.

* As each axis is carrying exactly the same amount of inertia as the corresponding PC in the space of observations, and as both clouds (observations and variables) have the same total inertia, we will run out of inertia after the first *p* axes !

So :

* the first *p* eigenvalues associated with the *p* first axes are the same as that associated with the PCs.

* and the last *n* - *p* eigenvalues are all "0".

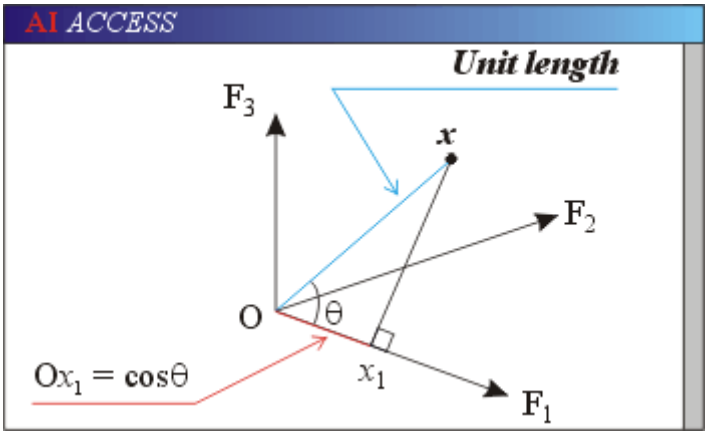
Axes and PCs

Now a very **important** result. It can be shown that the variable F_1 is no other than the first

PC in the space of observations (except for a multiplicative constant). In other words, F_1 defines the same direction as C_1 does in the space of observations. This result will prove important when interpreting the plot of the projection of observations.

Similar considerations apply to all F_2, and to all the other axes in the space of variables.

Coordinates of the variables, loadings



Just as observations have coordinates on the PCs, variables also have coordinates on the Principal Components in the space of variables (or "axes"). Vvariables are assumed to be standardized, and therefore be of of length "1". So the coordinate of variable ***x*** on F_1, is just the **cosine** of the angle between F_1 and this variable. Because the direction of F_1 is just the same as that of the first PC in the space of variables (see previous paragraph), we have the important result :

The coordinate of a variable on one of the axes is just the **correlation coefficient** of this variable with the corresponding PC is the space of variables.

The coordinates of the variables on the PCs are also called **loadings**.

Contribution of a variable to an axis

Consider axis F_1, whose inertia is I_1.Variable ***x*** projects on F_1 in x_1.

Ox_1^2 is the inertia of x_1 on F_1. The inertias of all the variables on F_1 add up to I_1. The proportion of I_1 that is due to variable ***x*** is called the **contribution** of ***x*** to F_1 :

Contribution of ***x*** = Ox_1^2 / I_1

Saying that "***x*** has a larger contribution to F_1 than ***y***" is the same as saying that "***x*** projects on F_1 at a larger distance from the origin than ***y***". We'll soon make use of this remark.

Plot of variables

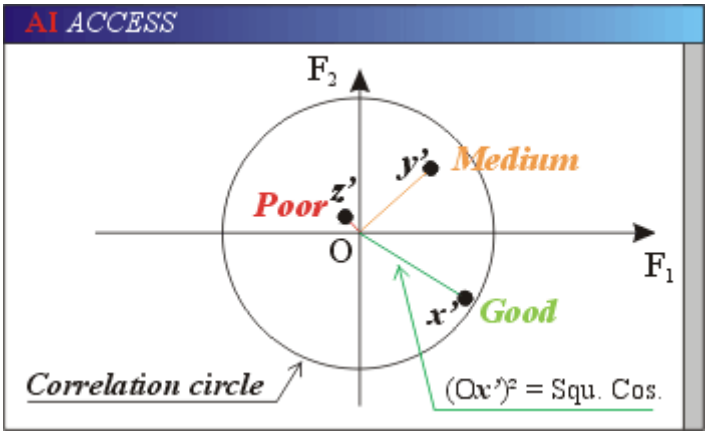
The Correlation Circle

Remember that all variables have unit length. Therefore, their projections on the Principal Plane always lie inside a circle of unit length. This circle is called the "**Correlation Circle**".

Note that the origin of the plot of the projection of the observations on the (C_1, C_2) plane was the barycenter of the projected observations. Not so with variables : in the space of variables, the origin has no reason for being some sort of barycenter of the variable-points.

Quality of the representation of a variable

On the projection plane



The projections of the variables on the best Projection Plane (for variables) is called the **plot of variables**.

* If the point representing a variable is close to the correlation circle, the corresponding point in the space of variables is close to the projection plane (again, remember that variables lie on a unit sphere). Therefore, it is faithfully represented.

* On the other hand, if a point is close to the origin O, then it is as far as can be from the projection plane in the space of variables, and its projection is therefore not meaningful.

So we are going to concentrate our attention on those points that are **close to the correlation circle** : only then can we draw conclusions about the relative positions of the projected variables.

The global quality of the representation of a variable in the plane may be estimated by a number : the square of the distance of this projection to the origin O. Because the true distance of a variable to the origin is always 1, this quantity is just the **square of the cosine** of the angle between :

* the straight line going through the origin and the point ***x*** representing the variable in the space of variables,

* and the staight line going from the origin to the projection ***x'*** (note the prime) of the variable in the plane.

One may also say that it is the square of the cosine of O***x*** with the plane (F_1, F_2).

Good represention is equivalent to high squared cosine (close to 1).

On an axis

The global quality of the projection of ***x*** (square of the cosine of O***x*** with the plane (F_1, F_2)) is the sum of :

* The square of the cosine of O***x'*** with the axis F_1,

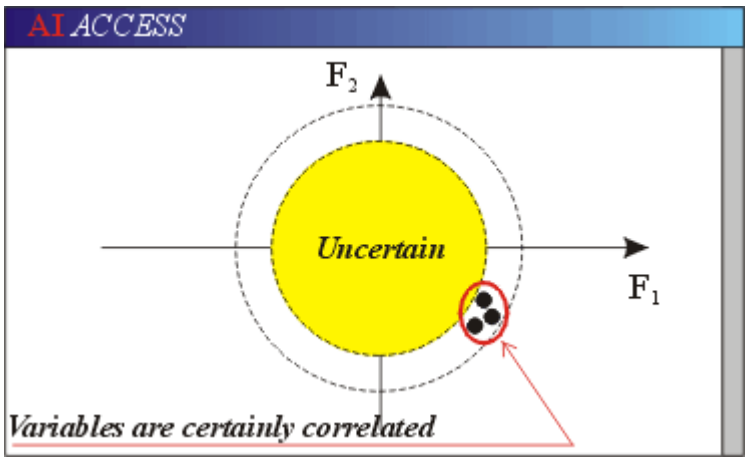
* and the square of the cosine of O***x'*** with the axis F_2.

So the global quality of the projection is the sum of two terms, each being the quality of the representation of the variable by an axis. A variable is well represented by an axis if the squared cosine with this axis is large (close to 1).

Correlation of variables

In the space of variables, all variables are distributed on the unit sphere, and we noted that highly correlated variables are represented by points on the sphere that are very close to each other. Does this property translates in projection ?

Certainly, if two variable-points are close to each other on the sphere, their projections will also be close to each other. But unfortunately, the converse is not necessarily true. Only if **both** variables are well represented, that is if their projections are close to the correlation circle, can it be ascertained that the variables are indeed highly correlated.



It may very well be that the projections of two poorly represented variables are very close to each other, but then nothing can be inferred about their correlation.

We will shortly see that practitioners are particularly interested in **groups** of highly correlated variables, as they probably represent closely related quantities. Visual examination of the plot of variables may occasionally allow quick detection of such groups, if the corresponding points are :

* All close to the correlation circle,

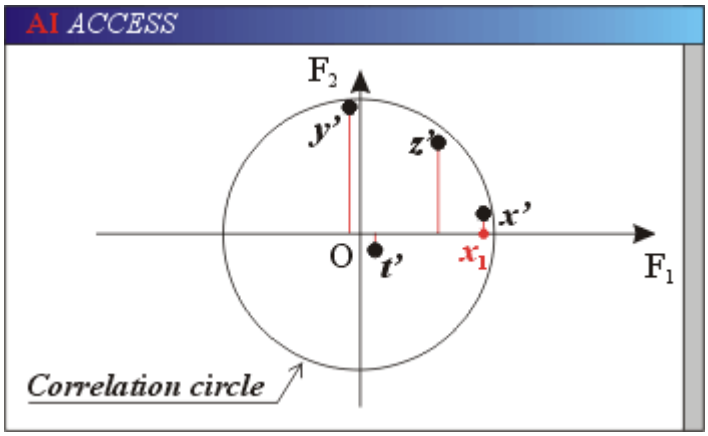
* and close to each other.

Contribution of a variable to an axis

Earlier, we considered the projection of a variable ***x*** in the space of variables directly on the axis F_1. But when we look at the plots of variables, we see the projections of the variables on the projection plane. Because of the properties of inertia, it is euivalent to :

* Project ***x*** directly on F_1, or

* First project ***x*** on the (F_1, F_2) plane in ***x'***, and then project ***x'*** on F_1 in x_1.



So in this illustration, Ox1 is the **true** projection of ***x*** on F_1, whether ***x*** is well represented or not in the plane. The primes ' denote the projection of the variables.

In this illustration, we see that :

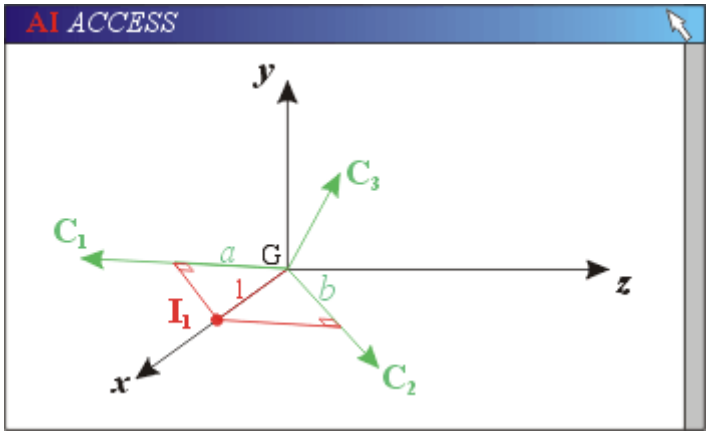
* ***x*** contributes highly to F_1, and almost nothing to F_2.

* ***y*** contributes highly to F_2, and almost nothing to F_1.

* ***z*** has a substantial contribution to both F_1 and F2.

* ***t*** contributes almost nothing to either F_1 or F_2.

Note in passing that as ***x*** and ***y*** are both well represented, and are orthogonal, we conclude that these two variables are almost uncorrelated.



We now mention a result of practical interest. Let's go back to the space of observations (the "ordinary" space), and imagine a fictitious observation "$\mathbf{I}_1$" whose coordinates on ***x*** is "1", and its coordinates on all other variables are "0". Therefore, its coordinates on the original variables are (1, 0, 0, ..., 0). On the PCs, the coordinates of $\mathbf{I}_1$ are (*a, b, c..., l*).

On the top illustration to the left, projection on C_3 has not been shown for clarity.
On the bottom illustration, we've drawn a line from the origin G that also passes through I_1. It materializes the direction of increasing values for ***x***.

Then it can be shown that the **contribution of variable *x* to axis $\mathbf{F_1}$ is equal to a^2**.

This result is valid only if all original variables are standardized.

Of course, similar relationships exist for all other variables.

Simultaneous projections ?

Now look at what we have.

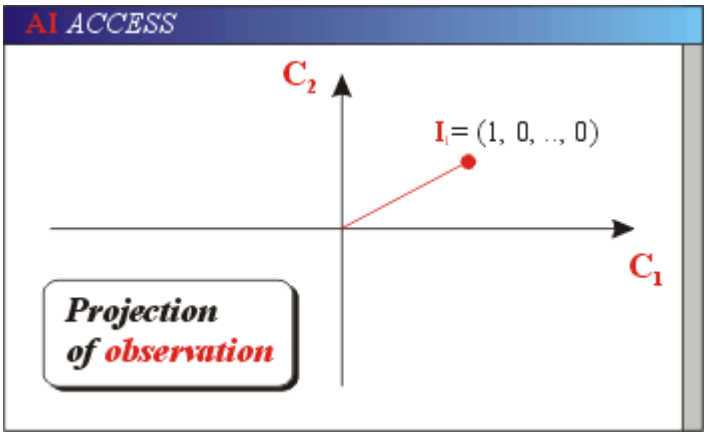
1) We have a Principal Components plane, with two orthogonal PCs, C_1, and C_2.

2) We have a (F_1, F_2) plane in the space of variables, and we stated that F_1 considered as a variable, is just C_1 (same with F_2 and C_2).

It is quite tempting to superimpose the two planes so that :

* F_1 coincides with C_1,
* and F_2 coincides with C_2.

so we would have a simultaneous representation of observations **and** variables on the same plot.



Is this justified ? NO, it is not, or rather not quite, and it is easy to see why.

1) We just mentioned the fictitious observation $\mathbf{I}_1$ whose coordinates are are (1, 0, 0, ..., 0). The coordinates of $\mathbf{I}_1$ are "1" on x_1, and "0" on all the other variables. Draw line $\mathbf{D}_1$ from the origin to this point.

2) Now consider variable x_1 in the space of variables, and project it on the (F_1, F_2) plane. Call $\mathbf{D}'_1$ this projection.

If you try to superimpose the "Observations" and "Variables" plots, $\mathbf{D}_1$ and $\mathbf{D}'_1$ do **not** coincide. This ruins our hopes of making the superimposition of both plots meaningful. Yet, both plots are strongly related : for any i, the directions of $\mathbf{D}_i$ and $\mathbf{D}'_i$ are pretty close, but they never coincide.

There exists a simple mathematical relationship between the directions $\mathbf{D}_i$ and $\mathbf{D}'_i$.

Software often allows a simultaneous representation of observations and "variables" on the same plot. It is to be understood that the"variables" are then the projections of the $\mathbf{I}_i$ unit-observations (see above), and the set of vectors is just the original reference frame "squashed" on the Principal Plane, and **not** the plot of variables. What that means is that no interpretation of the angles between "variables" can be safely conducted on the simultaneous plot.

The same plot usually displays a circle of radius 1, which is then abusively called the "Correlation Circle".

INTERPRETATION OF THE RESULTS

The goal of Exploratory Analysis is to allow the analyst to understand the underlying structure of data just as if he could "see" directly the data in its natural high-dimension space. As this is not possible, PCA will project the data (observations or variables) on **factorial planes**, each plane being defined by two factors as determined by PCA.

The best projection planes have been identified by PCA : they are spanned by the low order factors.

A good deal of experience and know-how are needed to extract valuable information from numbers and projection diagrams as determined by PCA : this is the **interpretation** phase.

The data

It is about time that we introduce an example to make this interpretation more tangible. So we are going to consider a bank with, say, 50 local agencies (the "observations") in various countries that are suspected to have different savings cultural traditions. Each agency is asked to partition the population of its customers who own a savings account by profession :

- *Engineer*
- *Teacher*
- *Farmer*
- *Employee*
- *Worker*
- *Secretary*

- *Liberal profession (Doctor, lawyer, consultant)*
- *etc...*

	Engin.	*Teacher*	*Farmer*		*Worker*
1	0.153	0.297	0.146	-----	0.046
2	0.256	0.052	0.383	-----	0.183
3	0.027	0.065	0.339	-----	0.139
4	0.195	0.265	0.281	-----	0.081
	-----	----	-----	-----	-----
50	0.257	0.373	0.172	-----	0.103
Average	0.283	0.173	0.272	-----	0.095

At the intersection of an agency (observation) and a profession (variables), is the ratio of the average savings to the average income.

These numbers are totally fictitious, and are not supposed to reflect any reality.

The bank decides to conduct a PCA on these numbers in the hope of detecting some meaningful trend.

Quality of the Principal Components

Before attempting to interpret the plot, it should be asked whether it is a sufficiently faithful representation of reality to be worth spending any time it. What is hoped for is that the first two Principal Components will amount for as large as possible a fraction of the total inertia of the cloud. Suppose that the first PC carries 60% of the inertia, and the second PC carries 25% of the inertia. Then the Principal Plane could account for 85% of the inertia, which is certainly very satisfactory.

As the total amount of inertia of the Principal Plane goes down, the PCA analysis will become more and more difficult, and will require more and more care in interpreting the wisual suggestions rpovided by the plot.

Interpretation of the Principal Components

The central issue is the interpretation of the Principal Components. Recall that they are linear combinations of the original variables, and that they are therefore genuine variables. We knew the meaning of the original variables, as they were attributes of the observations. Can a similar "meaning" be given to the PCs ?

Recall also that the first two Principal Components in the space of variables define directions that are identical to those defined by the first two PCs in the space of observations.

Interpreting the Principal Components will resort to the plot of variables.

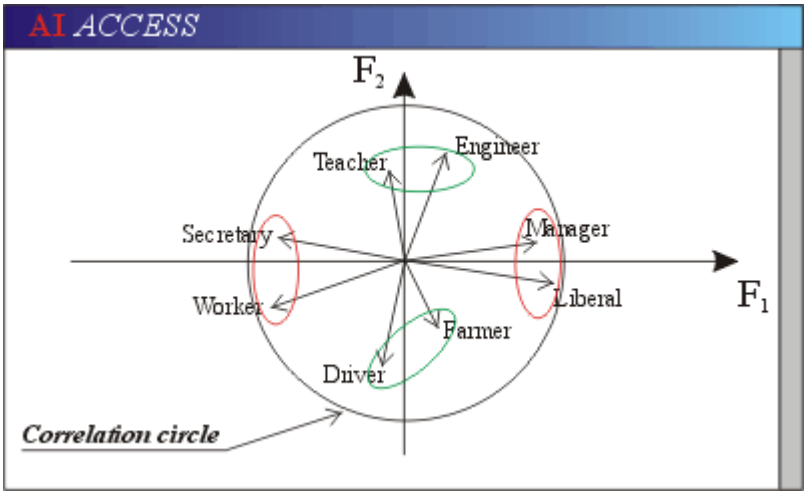
We are paying a particular attention to those variables that are :

- Close to the correlation circle (well represented variables)
- and close to the F_1 axis.

as they will define a trend in some quantity from left to right. We want to identify the nature of this quantity.

The general strategy is to look at both ends of the PC, and identify clusters of well represented variables. The standard sentence of identification of a PC will then be :

* "PC_1 **opposes** this-and-that on the left, to this-and-that on the right, and therefore **represents** the quantity xxx".



In this example, on the right hand side of the plot are the variables : (Manager, Liberal), whereas on the left hand side we find (Worker, Secretary). So we will interpret the First Component as something akin to "Level of Education". Note that PCA does not tell you that : you have to make this deduction yourself, which may require getting back to the values of other variables that were not used in the analysis.

Now we proceed to the Second Component of the plot of variables. Of course, because this Second Component carries less inertia than the First one did, interpretation is not as clear cut as before : the variables that contribute the most to the determination of F_2 carry, on the average, less inertia than those instrumental in defining F_1.

Yet, we see that this second axis contrasts (Teacher, Engineer) on top with (Driver, Farmer) at the bottom. After some thinking, we will interpret the second Component as indicative of " Office_Work" vs. "Outdoors_Work".

In a more realistic situation, there would be many more variables, whose projections would be closer to the center of the correlation circle. These variables would therefore be poorly represented, and left out of the interpretation of the PCs.

The plot of observations

We can move on to the plot of observations, whose PCs have just been interpreted.

Origin of the plot

The origin of the plot is the barycenter of the cloud's projection. It represents a ficticious "average" agency. For this agency, every professional activity (the original variables) would have the "average" savings level. As you move away from the center to reach peripheral regions, one

finds agencies that depart from this average behavior one way or another.

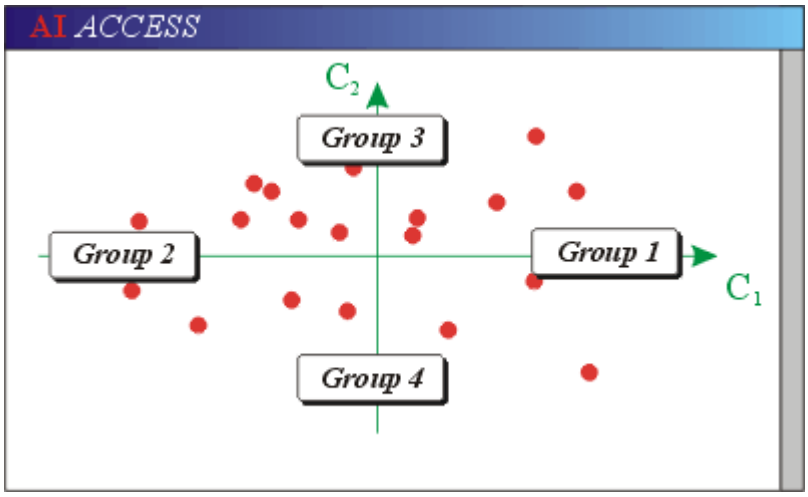
Moving along the Principal Components

Starting from the origin of the plot, we'll "slide along" the PCs that have have just been interpreted by the plot of variables. Remember that the origin represents the "average" agency, so we'll conduct the interpretation in terms of how observations encountered during this slide differ from average.

Region 1

We define "Group 1" as the set of observations that are close to the first Principal Component C_1, and to the right of the plot. Agencies in this regions are characterized by higher than average values of "deposits" in the columns "Manager" and "Liberal". We'll therefore interpret Group 1 as the set of agencies where highly educated people tend to save more than average.

Region 2



As one slides to the left, although "Level of Education" (that is, C_1), and therefore "Income" decreases, savings of Workers and Secretaries increase, and become proportionally higher than average : agencies to the left of the diagram and close to C_1 are located in areas with a local culture of "saving" for lower income people.

Region 3

The same kind of analysis may be conducted on C_2, the second Principal Component. At the top of C_2, we find agencies which departs from the average in that "Office Professions" tend to save more than average.

Region 4

At the bottom of C_2 are agencies where "Non Office Professions" tend to save more than average (and therefore Office Professions less than average).

Of course, for all regions, these conclusions are valid only for those agencies which are well represented on the Principal Plane, that is agencies with a high Squared Cosine for the plane.

Half plots

We can also interpret the plot by halfs :

- The left part concerns agencies whose customers project on C_1 as "less than average Level of Education". The right part contains agencies whose customers have more than average Level of Education.
- In the upper part, we find agencies whose populations are mostly "Non_office" professions, whereas the lower part contains agencies whose customers hold mostly Office positions.

General distribution of the observations

Some general features of the distribution may now be examined. More particularly, one will (visually) look for :

* Clusters of observations in the Principal Plane. Here, they will be groups of agencies whose populations of customers share a common profile.

* Whether the cloud extends uniformely in all directions and is well centered around the rigin, or tends to "stretch" out in regions of low density.

* Whether observations at the periphery of the cloud are well represented (high squared cosine), which is to be expected from a well behaved, convex cloud. Peripheral observations (outliers) with low squared cosines may either be genuinely exceptional observations, or else be observations whose attributes have erroneous values.

Higher order Principal Components

The same kind of analysis may be conducted for pairs of higher order PCs, like (C_1, C_3) or (C_2, C_3). Although higher order PCs carry less and less inertia, and therefore the corresponding planes display less and less faithful 2D images of the cloud, some additional information about PCs interpretation and cloud distribution may still be gathered, using the same lines of thinking as for (C_1, C_2).

Quality of the PCA

Although we hinted at the fact that different clouds may have shapes that are more or less propitious to a faithful 2D projection, we did not try to quantify the quality of a PCA. We'll do that now.

Eigenvalues

The amount of inertia carried by a PC is called the "**eigenvalue**" associated with the rank of the PC.

The amount of inertia carried by the projection plane is the sum of the eigenvalues associated with the 2 PCs that determine the plane.

Recall also that the total inertia of the cloud of observations is just *p*, the number of original variables (when variables are standardized).

So, the quantity :

$$(\text{Eigenvalue 1} + \text{Eigenvalue 2}) / p$$

is a natural indicator of the global quality of the projection in the Principal Plane. A similar quantity may be used for any pair of PCs.

This criterion is straightforwardly extended to any Principal Subspace with more than 2 dimensions.

Communalities

The above quantity is an obvious criterion that tells how faithful the global representation of observations is. Now how about the global quality of representation of the variables ?

We describe here a popular criterion that is quite useful for expressing the quality of representation of one particular variable (besides the length of its projection), say $\boldsymbol{x}$. The axes of the Principal Plane for variables are honest, genuine variables. So it is only natural to ask whether they contain enough information to reconstruct $\boldsymbol{x}$ by linear combination. In other words, on may think of trying a Multiple Linear Regression with $\boldsymbol{x}$ as the dependent variable, and (F_1, F_2) as the independent variables. The quality of this regression is measured by a quantity named R^2, or "coefficient of determination", which is just the proportion of the variance of $\boldsymbol{x}$ that is "explained" by (F_1, F_2). In the context of PCA, this quantity is called the "**Communality**" of $\boldsymbol{x}$ with respect to (F_1, F_2).

The concept of Communality extends to more than 2 axes. As you add new axes, the communality of $\boldsymbol{x}$ increases, to reach 1 when all axes are taken in the Regression.

PCA : OTHER APPLICATIONS

We are now briefly reviewing some popular applications of PCA beyond visual examinations of data.

* Because it reduces the number of variables needed to describe data, PCA can be regarded as a (lossy) data compression technique. Data may by summarized by the coordinates of the observations on some few first factors, and therefore be "compressed". It is possible to reconstuct the complete data from this partial description. The reconstruction is of course not perfect.

* All models are sensitive to the bias-variance tradeoff, which requires the model to incorporate as few variables as possible (for a given amount of information injected into the model). Because PCA reduces the number of variables with minimum loss of information, it can be used as a data pre-processing tool before building another model.

Data Compression

Although Data Mining is not concerned by the concept of Data Compression, many applications are. Besides, we suggest that you read this anyway, as some points are of general interest.

Why can PCA compress data ?

The original data was represented by a table with m lines and p columns. A crude way of compressing data would be to reject some of the original variables on the basis of the fact that they carry little information about data distribution. In doing so, we would lose information. This loss would translate into having some distinct observations being indistinguishable on the remaining variables, because they have almost identical coordinates on these variables. How much information is lost in the process is hard to assess, and is probably large anyway.

PCA will salvage this basic idea, and make it operational.

After PCA, data is still represented by a table with m lines and p columns. But the last PCs carry

little information in terms of distinguishing between observations, so they can be removed and yet create little confusion between observations. Besides, because of the special properties of PCs, we know that after removing the last q PCs, the remaining $(p - q)$ PCs carry as much information as any set of $(p - q)$ variables will ever do.

So, when expressed in the first p' ($< p$) PCs, data comes as a $m.p'$ table, and has therefore been compressed (with loss), and this compression causes the smallest loss of information possible.

How many Principal Components should be kept ?

How many PCs should kept, and how many should be discarded ?

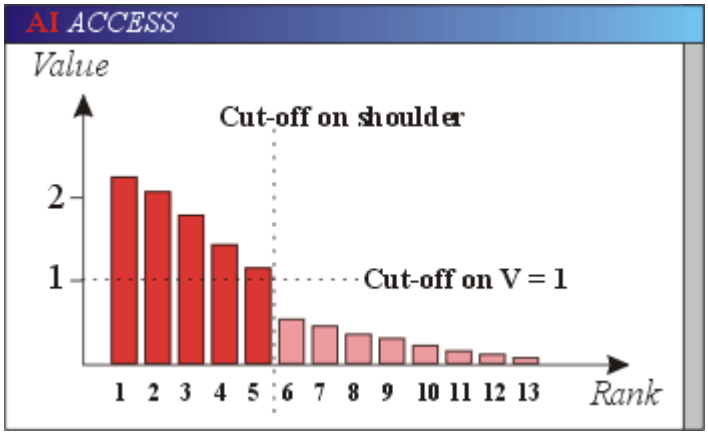
PCA tells you how much information you're throwing away together with the last $(p - p')$ PCs : it is just the sum of the inertias carried by the discarded PCs.

So :

1) If you know how much information loss you are accepting to incur, PCA tells you how many PCs you should retain.

2) Conversely, if you know how many new variables you want to keep, PCA tells you how much information you are going to throw away.

In practice, software displays a graph of the (decreasing) values of the inertias, or "eigenvalues" carried by the PCs.



* The most favorable case is when the eigenvalues diagram shows a sharp "shoulder" for some PC rank. Then you know that discarding all PCs after this rank will result only in a small loss of information. Some simple formulas may help you identify a good cut-off point.

* Another rule of thumb is to discard all PCs that are carrying less inertia than the average amount of inertia per PC (that is the total inertia p, divided by the number of variables, that is p, that is an average inertia of "1").

Both rules do not always define the same number of PCs.

Data reconstruction

It is possible to reconstruct a cloud of points in the complete space from its projection on one of the PCA subspaces. This cloud is an approximation of the original cloud. The maths behind this reconstruction is beyond the scope of this short tutorial.

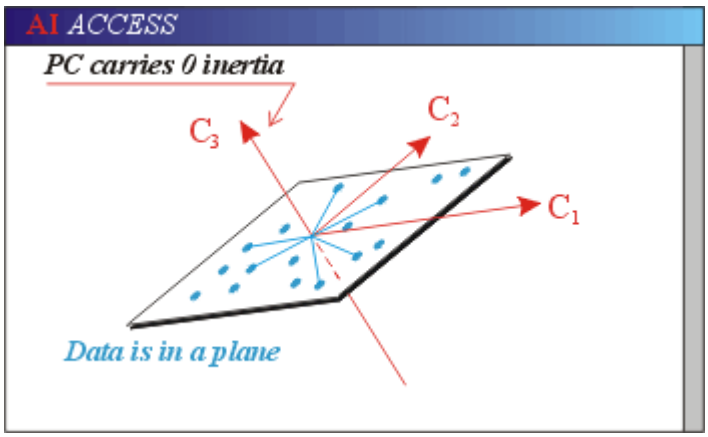
Data Pre-processing

Dimensionality reduction

The possible existence of a cut-off in the eigenvalues diagram suggests that some high order eigenvalues might in fact be "0". Is that possible ? That would mean that the inertia of the data

cloud with respect to the corresponding PC is "0", and therefore that all observations would project on the origin of that PC.

Try to visualize this situation in the ordinary 3D space : all points projecting on the same point of a line mean that the points are distributed in a plane that is orthogonal to the line. So the data is actually 2D (and not 3D), and PCA has been able to detect the 2D space in which the data lies. The same line of reasoning applies to any number of "0" eigenvalues : if, say, the last 3 eigenvalues are "0", then the true dimensionality of data is p - 3, and PCA has identified the (p - 3)-dimensional space in which the data lies.



Another way of looking at the "eigenvalue = 0" phenomenon is to say that the original variables are not linearly independent. In our 3D example with variables (***x, y, z***), the third eigenvalue being equal to 0 **implies** that there is some sort of linear relationship between the variables, for example :

$\boldsymbol{x} = b.\boldsymbol{y} + c.\boldsymbol{z}$

In the space of observations, this translates into the following statement :

- Point ***x*** lies in the plane defined by the origin and the points ***y*** and ***z***.

So our data is redundant, as either ***y*** or ***z*** could be selected to describe the data set with no loss of information. This kind of situation is very common, and is usually disastrous for the stability and interpretability of all sorts of models (see bias-variance tradeoff).

Now, back to the real world. Real life data never lies in the high-dimension equivalent of a plane, so real life eigenvalues are never quite "0". But when these values are small enough, one can think that :

- data always stays "close" to such a flat subspace of the total space,
- and that, equivalently, some variables are close to entertaining a linear relationship among themselves.

PCs are uncorrelated

Before being fed to a model, data has to undergo quite a bit of conditioning. One important step is to reduce the number of variables while losing as little information as possible in the process : this is the "Dimensionality reduction" phase. Many techniques are available for Dimensionality reduction, and PCA appears now as a convenient way of squeezing redundancy out of the original set of variables. As a matter of fact, one easily shows that PCs are **uncorrelated** variables. So, in a Dimensionality reduction perspective, PCA has two benefits :

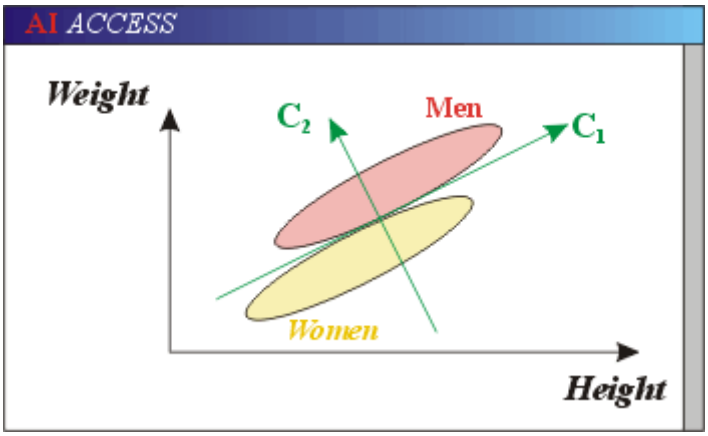
- A small number of PCs contains most of the information,
- and the new variables (the PCs) are uncorrelated.

These two advantages are closely related : should the new variables not be uncorrelated, then further dimensionality reduction could be achieved by "PCA-ing" them.

Models built after an appropriate Dimensionality reduction technique like PCA has been used to pre-process the data are more **reliable** than models built with the complete set of variables. A side effect may be a small loss is accuracy (because of the small loss of information), but this is often a small price to pay for obtaining a model with good generalization capacity.

How many Principal Components should be kept ? (revisited)

Above, "information" is a general term that refers to "information about data density in space". It may be that this is **not** the kind of information that is of interest to you.
For example, suppose your data base contains information about "***Height***" and "***Weight***" of "Men" and "Women". Data distribution and Principal Components look something like this.



Your problem is to build a model such that, given the "***Height***" and the "***Weight***" of an observation, you can predict whether this person is a Man or a Woman. Ordinary Discriminant Analysis would do very well.

Now, suppose that you want to build another model, but this time with with **one** variable only. Naturally, you would think of the first PC as your best choice for the job, as you know that, among all possible variables, this is the one that wastes the least amount of information about your data.
But take another look : once projected onto C_1, the two classes ("Men" and "Women") are hopelessly overlapping, making any attempt at separating them futile.

Things are quite different with the second Principal Component : once projected onto C_2, the two classes overlap only a little. It is indeed possible to build a good classification model with one variable only, but then C_1 appears to be a poor choice, whereas C_2 is a much better choice.

Note that the same problem would happen for Multiple Linear Regression.

So PCA is not the miracle cure for variable selection in predictive modeling. We discuss this issue further in the context of Multiple Linear Regression.

In summary, blindly retaining the first PCs in Classification or Regression is a poor strategy, as the information that is useful for reaching your goal might be concealed in high order PCs.