OVERVIEW OF CORRESPONDENCE ANALYSIS

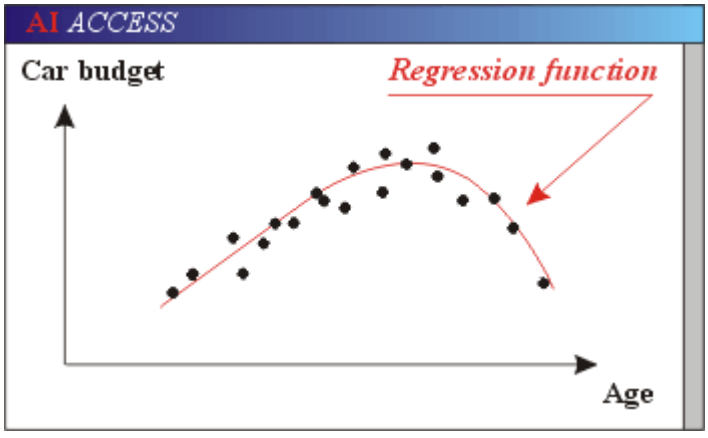
This first Tutorial is an overview of Correspondence Analysis. We show how contengency tables may be regarded as a numerical coding of the interaction between two categorical variables through frequencies of pairs of modalities.

A PCA-like transformation then allows the modalities of the variables to be represented as points in factorial planes. Visual analysis of these plots, and in particular of the proximities between modalities, will then give us a visual clue about whether the frequency profile of two modalities across other modalities are similar or not.

"Interaction" between categorical variables

Independent categorical variables

The concept of interaction between two numerical variables is quite easy to grasp. If nothing else, a simple scatter plot tells if "knowing the value of x gives a lot of information about the possible values of y". This is what regression is all about (illustration).



Interaction between two

categorical variables is not nearly as intuitive, if only because it does not lend itself to any obvious graphical representation.

As a matter of fact, the first approach to describing the interaction between two categorical variables, say V_1 and V_2, is through defining the **lack** of interaction, that is the **independence** between two such variables. As usual, "independence" is defined by the fact that knowing what modality an observation o has for V_1 gives absolutely no information as to what the modality of o for V_2 might be. In other words, the conditional probability distribution of V_2 does not depend at all on the modalities of V_1. This idea leads to the well known test known as the "Chi-square test of independence".

This test is very nice, but it is global : it only tells how likely it is that V_1 and V_2 are indeed independent, given the sample. If the hypothesis of independence is rejected as too unlikely, no detailed information about **how** V_1 and V_2 depart from idependence is provided by the test. In other words, "interaction" is then simply defined as "departure from independence".

Interaction between categorical variables

The objective of Correspondance Analysis is to finely analyze this departure from independence, and give a faithful **graphical** representation of the interaction between two categorical variables. Here is a very simple and classical example of what kind of details are expected. Say :

* V_1 is "Eye_color", with modalities (Blue, brown, green, grey).
* V_2 is "Hair_color", with modalities (Black, brown, blond).

It is a common cliché that "blond women tend to have blue eyes, and dark-haired men tend to have brown eyes". A Chi-square test will tell how likely it is that eye color and hair color are independent, but it won't address the above questions, that are beyond its reach.

Correspondence Analysis will answer the question **graphically** in a way that will answer these questions, as well as other questions of the same nature.

The mechanism of CA

The mechanism of CA is a bit cumbersome, but it seems that there is no way around getting somewhat into the details. We here only outline the general principles of CA, that are explained in more details in the next section.

Contingency tables

Although data is usually available as a table crossing observations with variables, CA will not work on this kind of representation. Rather it will work on **contingency tables**, that cross tabulate V_1 and V_2 on the sample.

Contengency tables are described in more detail in the next section.

PCA on rows and on columns

So a contengency table is a rectangular table of numbers, and we want to give some sort of 2D graphical representation of this table. This may ring a bell, as it it just what PCA does. So it seems quite natural to consider PCA as an adequate tool for generating graphical representations of contengency tables. But things are not that simple. For reasons that we will explain, ordinary PCA cannot be conducted on the raw contengency table.

1) The contingency table will be first "duplicated", because two different treatments will be applied to rows and to columns.

2) On the first of the two "twin" tables, rows are normalized so that numbers become proportions. A similar treatment is applied to the columns of the other table.

3) Each row (resp. column) will be ponderated in a way that accounts for its importance, that is, the population of the corresponding modality.

4) The ordinary euclidian distance will be replaced by the so called "Chi-square" distance.

5) PCA will be applied in turn to each of the two tables. Retaining for both the first two Principal Components yields two 2D plots, one for the modalities of V_1, the other one for the modalities of V_2. This may remind you of the "PCA on observations" and "PCA on variables" of the standard PCA procedure.

Of course, all these changes over standard PCA will require a bit of justification.

Simultaneous representation

The two plots are then overlaid, thus providing the final plot. You may remember that overlaying the two similar plots (observations and variables) in ordinary PCA was unjustified, but we will see that this restriction can, with some restrictions, be lifted in CA.

What is expected from a graphical representation ?

So CA will plot the modalities of **both** variables as **points on a plane** in a way that will **suggest** the following intuitive interpretations :

Axes

Just as PCA does, CA defines axes of decreasing importance. Any pair of axes may by used to define a projection plane, but of course, the most meaningful plane is defined by the two most important axes.

Again, just as in PCA, the practitioner will attempt to "interpret" the meaning of these axes in terms of how the modalities of each variable project on them.

Distance to the origin

A modalitiy whose distribution across the other variable's modalities is quite "average" is close to the center of the plot. To the contrary, a modality whose distribution across the other variables' modalities is quite "exotic" will lie at the periphery of the diagram.

Two modalities belonging to the same variable

Two modalities m_1 and m_2 of one same variable, say V_1, that are **close** to each other on the plot have nearly identical distributions across the other variable's modalities.

To the contrary, two modalities in opposite regions of the plot will have opposite types of distributions across the modalities of the other variable. Typically, cells of m_1 with a higher than average observations count will correspond to cells of m_2 with lower than average observations count.

Two modalities belonging to different variables

Two modalities that belong to different variables are close, or "attract each other" if it is more common than expected that an individual will possess both these modalities.

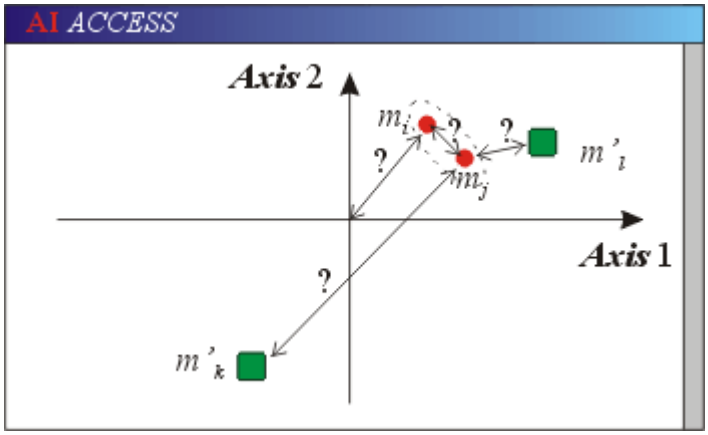
To the contrary, two modalities with a large distance from each other correspond to a lower-than-expected cell count.

Interpreting the distances between modalities belonging each to one of the two variables is a more chancy business than is interpreting the distances between modalities belonging to the same variable.

These questions are represented in a schematic way on this illustration. Each question mark represent two interrogations :

1) How can the relative positions of the various modalities be interpreted ?

2) Are these interpretations correct ?



We insist that these nice interpretations are only **suggested** by the plot, and that they require thorough confirmation. CA provides several tools that permit validating or not what plots suggest.

At any rate, all seasoned practitioners will deliver the same warning : casual interpretation of a CA plot will lead to erroneous interpretations, and that may very well be worse than no interpretation at all.

Validating the interpretation of a plot requires going into the mechanisms of CA, which this tutorial is doing in the next section.

MECHANISM OF CORRESPONDENCE ANALYSIS

Correspondence Analysis does not work on raw contengency tables. It first normalizes them so that cell counts are replaced by frequencies, and modalities of one variable are decribed by normalized "frequency profiles" across the modalities of the other variable.

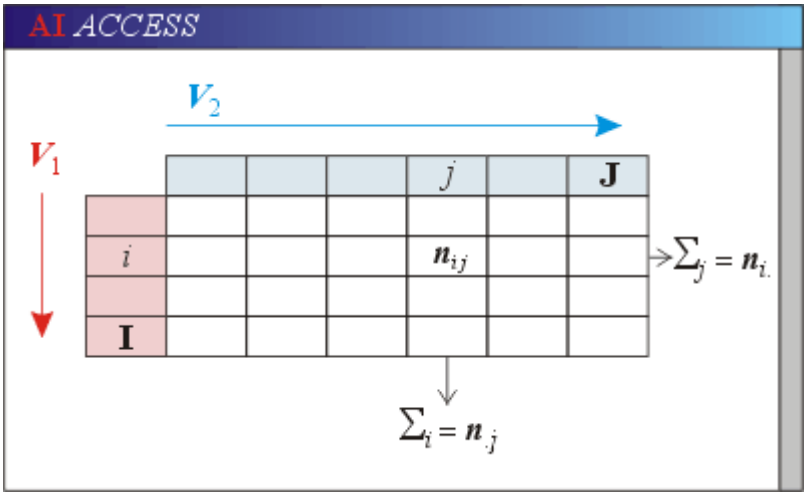
We then justify that the traditional euclidian distance in not appropriate in this setting for the purpose of measuring the similarity between modalities, and has to be replaced by the so-called "Chi-square distance". The upcoming PCAs will be performed with this newly defined distance.

Reformatting data

Data comes up usually in tables, with observations as rows, and variables (or attributes) in columns. Although this presentation is perfectly adequate for most models, we are here in a somewhat pecular situation. We want to analyze the relationship between two categorical variables, and there is no obvious graphical representation of this relationship, in contrast with what we have with numerical variables. So the first step of Correspondence Analysis is to reshape data so as to make it accessible to some sort of graphical representation.

Contengency tables

CA works on **contengency tables**.



Call "I" the number of modalities of V_1, and "J" the number of modalities of V_2. The contengency table of the sample is a I x J table. In cell (i, j) is the number n_{ij} of observations that have both modality i on V_1 and modality j on V_2.

* The sum of all numbers n_{ij} in row i is denoted $n_{i.}$, and is the number of observations with modality i on V_1.

* Similarly, the sum of all numbers n_{ij} in column j is denoted $n_{.j}$, and is the number of observations with modality i on V_1.

Note that contengency tables make no reference whatsoever to observations, that will "disappear" from the analysis from now on.

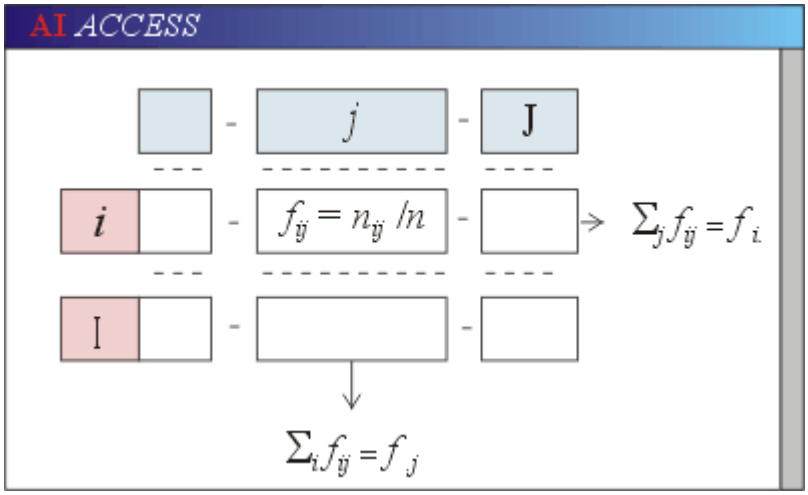
Frequencies

Take a contengency table, and multiply all the cell counts by 2. You get a new contengency table that is just the table you would have obtained with a sample twice as big as your original sample, everything else being equal. So, cell counts are not truly significant, but the **ratio** of a cell count to the total number of individuals is. This ratio is called the **frequency** of the cell, and is denoted f_{ij}. It is just the probability that an observation has both modality i on V_1 and modality j on V_2.

So, as a first step, all cell counts are going to be divided by the number of observations in the sample, to obtain a **frequency table**.

* The sum of all frequencies in row i is denoted by $f_{i.}$ (note the dot), and is called the **marginal frequency** of the row.

* The sum of the row frequencies is 1.



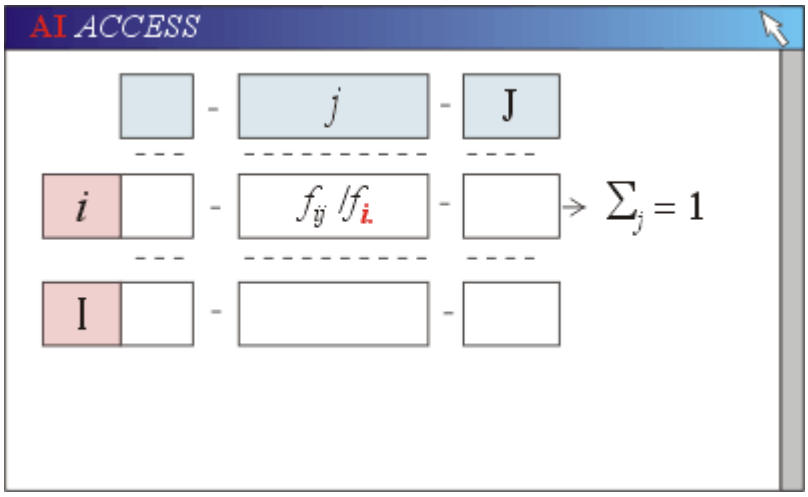
Of course, the same applies to columns :

* The sum of all frequencies in column j is denoted by $f_{.j}$ (note the dot), and is called the **marginal frequency** of the column.

* The sum of the column frequencies is 1.

Profiles

Let i_1 and i_2 denote two modalities of V_1, and therefore two rows of the frequency table. Suppose that the frequencies in row i_2 are exactly twice the frequencies in row i_1. This only means that modality i_2 is twice as populous as modality i, but other than that, both modalities are distributed exactly the **same way** across the modalities of V_2. So, even frequencies are not adequate to describe the interaction between categorical variables.



We now apply an additional transformation so that two rows that were **proportional** in the frequency table become now **identical.** In each row, divide each frequency by the row marginal frequency (top illustration). The content of each cell is now $f_{ij} / f_{i.}$. This is just the probability that an observation that is known to have modality i on V_1 has modality j on V_2. In other words, $f_{ij} / f_{i.}$ is the **conditional probability** of modality j given i.

Each row of frequencies is now tranformed into a **row profile**.

* By definition, the sum of the elements in a row profile is equal to 1.

* By construction, if two frequency-rows are proportional, then the corresponding row profiles are **identical**. Whatever the projection method used from there on, we know that two modalities of V_1 whose populations across the modalities of V_2 are proportional will project on the same point of the diagram, which is one of the central objectives of CA.

What we did with rows can just as well be done about columns. From the (common) frequency table, we define now column marginal frequencies $f_{.j}$ (note the dot). Each frequency is now divided by the corresponding column marginal frequency, to obtain column profiles (bottom illustration).

So from a single contengency table, we obtain **two** different "profile tables", one with row profiles, the other one with column profiles. Note the difference with PCA, where we considered one table and its transpose : both tables contained the same numbers in cells, just presented differently. Here,

row- and column-profiles tables contain different numbers.

CA will now perform two PCAs : one on row profiles, and the other one on column profiles.

Ponderation

Just as in PCA, rows of the row profiles table are going to be considered as points is a space with as many dimensions as variable V_1 has modalities. But a **major difference** with ordinary PCA is that now a **weight** is going to be attached to each point (modality). This weight is just the marginal frequency of the modality.

Because the determination of axes is based on maximizing inertias, and that inertias depend on weights (in addition to distances), modalities with large populations have a larger influence on determining the factors than scarcely populated modalities do. This will have to be kept in mind when we come to interpreting CA plots.

Exactly the same applies to columns : the column marginal frequencies will be used to ponderate column profiles when perforing PCA on the column profiles table.

The Chi-square distance

Definition of the Chi-square distance

We are now just about ready to perform PCA on the table of row profiles. Yet, one last issue needs to be adressed. The general definition of inertia is :

Inertia = weight.(distance to origin)²

so we need to specify what is meant by "distance" between two rows (or between two columns).

CA is going to use a modified version of the traditional euclidian distance, known as the "Chi-square distance". It is defined as follows :

$$d^2(i, l) = \sum_j (f_{ij} / f_{i.} - f_{lj} / f_{.l})^2 / f_{.j}$$

The summation is over the columns.

Let's go over this expression. In the parenthesis :

1) The first term is the coordinate of row profile *i* on the modality *j* (of V_2),

2) The second term is the coordinate of row profile *l*, also on the modality *j* (of V_2).

So the parenthesis is exactly what would be expected from the ordinary euclidian distance. But there is an extra "ponderation" coefficient, $1 / f_{.j}$, which is different for each term of the sum. The role of this coefficient is to equilibrate the influence of the populations of the columns (modalities of V_2) on the distance between rows : the contribution of a low population modality is thus artificially increased.

Why the Chi-square distance ?

This ponderation has an important practical consequence. Suppose that two columns of the original

contengency table are proportional. Then the corresponding column profiles are **identical**, which means that they are represented by the same point is the complete space, and also, of course, on the plot.

It can easily be shown that **merging** the two corresponding modalities of V_2 into one big modality will not change the distances between rows.

This is true because of the Chi-square distance, but is is not true with the ordinary euclidian distance.

Of course, a similar definition of the Chi-square distance applies to columns.

The 2 PCAs

This time, we are truly ready for PCA. The idea is in fact to perform PCA twice :

* Once on the table of rows. By extracting two axes from the analysis and projecting the modalities of V_1 of this plane, we will obtain a graphical representation of the modalities of V_1, from which we will later extract useful information.

* And then another time on the table of columns. By extracting two axes from the analysis and projecting the modalities of V_1 of this plane, we will obtain a graphical representation of the modalities of V_1, from which we will later extract useful information.

* With some reservations, we will later overlay these two plots to obtain a simultaneous representation of the modalities of both variables, and get useful information from analyzing the relative positions of these modalities.

We refer the reader to the Tutorial on PCA for some details about the mechanism of PCA. We want here to pinpoint two differences with standard PCA.

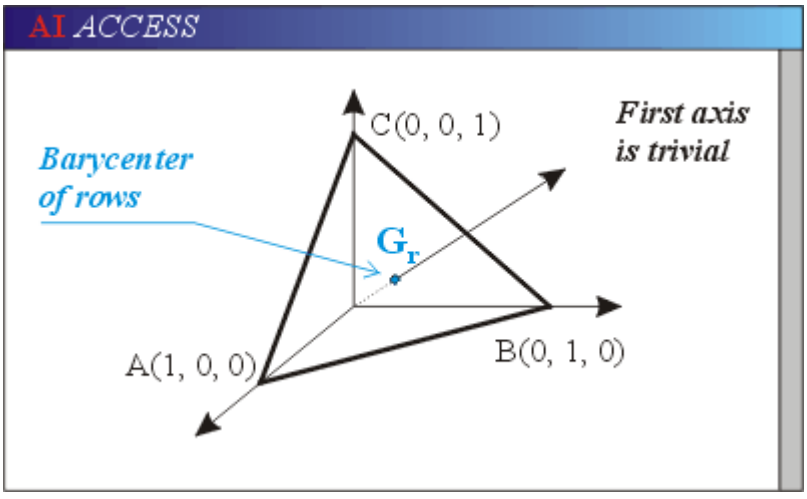
How many dimensions ?

In ordinary PCA, There are as many Principal Components as there are original variables. In CA, the situation is a bit different. We saw that for any row, the sum of the coordinates is always 1. This is expressed by :

$\sum_j x_j = 1$ Summation over the columns

which means that all the row points lie in a hyperplane of dimension p - 1, where q is the number of modalities of V_2 .

This hyperplane intersects the axes at the points (0, 0, ..., 1, ..., 0). It can be shown that the first Principal Component (or "factor") is orthogonal to this hyperplane, and that its intersection with the hyperplane is just G_r , the barycenter of the



cloud of row points. All points project on this factor in one point, namely G_r. So this first axis is trivial, and carries no information at all. Software always dismisses this first axis, which is not even mentioned.

As a result, the maximum number of axes is only q - 1 for the PCA on the cloud of rows.

Similarly, the maximum number of axes for the PCA on the cloud of columns is p - 1, with p the number of modalities of V_1.

So, refering to PCA, and because rows and columns play symmetrical roles, the maximum number of axes will be the smallest of (I - 1) and (J - 1).

The barycenters

What are the coordinates of the barycenter G_r of the cloud of rows ?

In general terms, the definition of the **barycenter** G of a set of points (P_1, ..., P_I) with weights (w_1, ..., w_I) is :

$G = (w_1.P_1 + ... + w_I.P_I) / (w_1 + ...+ w_I)$

In our case :

* the "points" are the row profiles i, whose coordinates on V_2 are $f_{ij} / f_{i.}$.

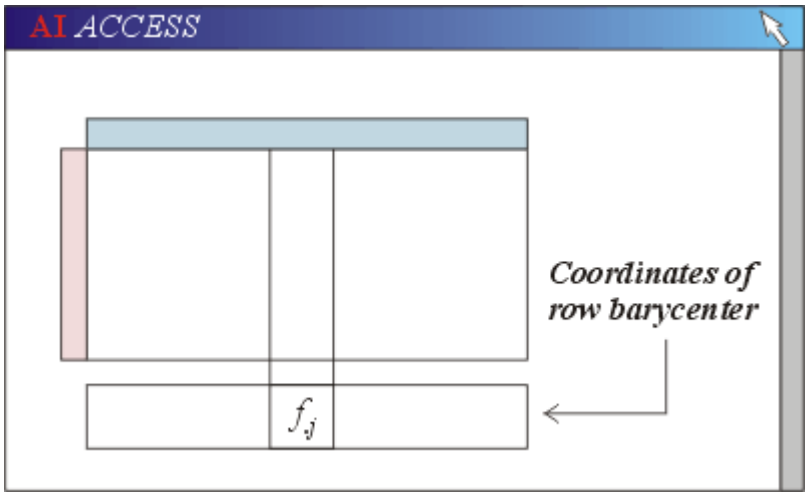
* the weights are the marginal frequencies $f_{i.}$.

So, the coordinate of G_r on modality j (of V_2) is :

$g_j = (\sum_i f_{i.} f_{ij} / f_{i.}) / (\sum_i f_{i.})$

* The numerator is just $\sum_i f_{ij} = f_{.j}$

* The denominator is 1.



and

$g_j = f_{.j}$

So, the coordinates of G_r (on V_2) are just the column marginal frequencies.

The row barycenter may be thought of as an additional row which is the ponderated average of the rows. It represents a fictitious modality whose distribution across the modalities of V_2 is quite "average". We will keep this remark in mind when we come to interpreting a CA plot.

Of course, the coordinates of the barycenter of the columns are the rows marginal frequencies (bottom illustration).

Chi-square and total inertia

Now that we know the coordinates of the barycenter G_r, we can calculate the inertia of the cloud of rows with respect to its barycenter. The general formula is :

I = Inertia = $\sum_i$ Weight.(Distance to barycenter)²

We now make this equation explicit. Remember that we are not using the euclidian distance, but the so-called "Chi-square" distance. So :

$$I = \sum_i f_i \sum_j (f_{ij} / f_{i.} - f_{.j})^2 / f_{.j} = \sum_{ij} (f_{ij} - f_{i.} \cdot f_{.j})^2 / f_{i.} \cdot f_{.j}$$

Let us now forget about CA for a moment.

We might consider running a Chi-square test of independence on the pair of variables (V_1, V_2). For that purpose, we would calculate the quantity :

$$Z^2 = \sum_{ij} (\mathbf{o}_{ij} - \mathbf{e}_{ij})^2 / \mathbf{e}_{ij}$$

where :

* is the "observed" population of cell *ij*,
* is the "expected" population of cell *ij*.

But, calling *n* the population of the sample, we have :

* $\mathbf{o}_{ij} = n.f_{ij}$ by definition of f_{ij}.
* $\mathbf{e}_{ij} = n.f_{i.}.f_{.j}$ by definition of independence.

So, the final and important result is :

$$Z^2 = n.\mathrm{I}$$

The total inertia of the cloud of rows (with respect to its barycenter) is 1/*n* times the Z^2 of the cloud. This is the reason why the "distance" we used is called the "Chi-square distance".

Of course, we have the same result for the cloud of columns, and both clouds carry the same inertia. So from now on, we will simply mention "the" inertia, without refering to either rows or columns.

Note an important difference with ordinary PCA :

* In PCA, the inertia of the cloud with respect to its barycenter (and for standardized variables) is *p*, the number of variables.

* In CA, this inertia is **not** directly related to the number of modalities of the variables.

INTERPRETATION OF CORRESPONDENCE ANALYSIS

At this stage, we have performed two PCAs :

1) One on row profiles,
2) One on column profiles.

We are ready to proceed with the interpretation of the results. This interpretation will be inspired by the interpretation procedure of regular PCA, with some changes because of the specifics of CA : ponderation of the modalities, Chi-square distance and the ensuing changes in interpreting inertias.

We review here the elements that will be needed for interpreting a CA. Later on, we will interpret a simple, but realistic example of CA, and we will need to keep the elements below in mind.

Plots

CA yields plots for row profiles and column profiles. We mention them first because they are very popular, but we will later insist that that they should not be considered as the primary ingredient of CA interpretation, at least in the early phase of the interpretation.

As we already mentioned, CA is essentially a double PCA on row-profiles and column profiles. The row profiles are changed in the process into new sets of coordinates on the factors, and so are column profiles.

The central idea is then the same as in PCA : if the first factors carry enough inertia, then 2D plots of the profiles (modalities) are faithful enough representations of the clouds of modalities to venture interpretations of the factors, and of the relative positions of the modalities of the variables.

We will later come back at length on the interpretation of modality plots. The only point we want to stress here is that whereas in PCA, overlaying the two plots (observations and variables) was not justified, it is perfectly justified in CA for reasons that we will not develop here. Suffice it to say that on the combined plot of modalities, we will make :

* Both barycenters coincide with the origin of the plot.

* Factors of the same rank in the space of rows and the space of columns coincide.

Note an important theoretical difference between PCA and CA.

* In PCA, the second space ("Space of variables") is obtained by transposing the original table.

* CA operates on two tables (row profiles, and column profiles) that are **not** transpose of each other because of the different ponderations.

Interpretation of the total inertia

In PCA, the total inertia is of just the number of variables : it is therefore meaningless in terms of interpretation of the PCA process. In Correspondence Analysis, the situation is quite different : remember that the total inertia is proportional to the value of the Chi-square of the contengency table, so the value of the inertia carries some meaning in terms of interpreting the CA :

* A small inertia means a low Chi-square, that is a situation of near independence between the variables. In each of their respective spaces, the clouds of the modalities are small and compact around the common barycenter.

* A large inertia means a situation of strong dependence between the two variables. The clouds spread out to a large distance from the barycenter.

What does "small" or "large" mean ? Remember also that we were able to establish the largest possible value for the Chi-square of a contengency table, and that we found :

$Z^2 / n \leq \min(I - 1, J - 1)$

So we have a scale against which we can evaluate how large the inertia is :

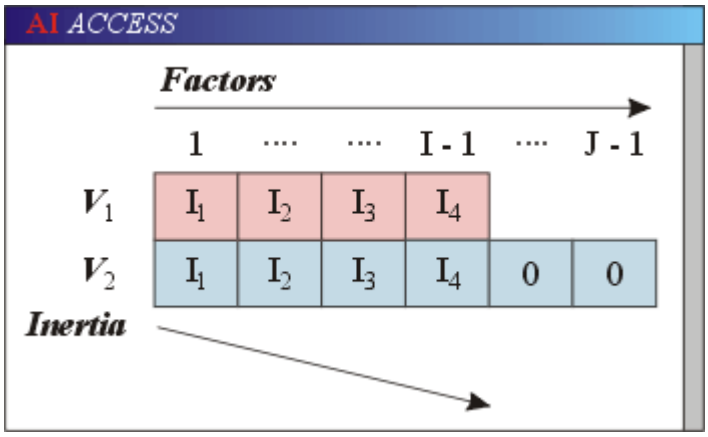
The inertia lies between 0 (independence) and min(I - 1, J - 1)

Finally, remember that we noted that the maximum value of the Chi-square occurs when there is a functional relationship between the two variables : for any modality of the variable with the larger number of modalities, there was only one modality of the other variable with a non-zero population. So, a high inertia is definitely an indication of a near-functional relationship between the two variables.

Yet, the total inertia, just as the Chi-square, is only a **global** estimation of the degree to which the variables interact. To go further into the details, we need to consider now the individual inertias of the factors : they will tell us in which directions the clouds stretch out most, not just the "average" amount of spreading.

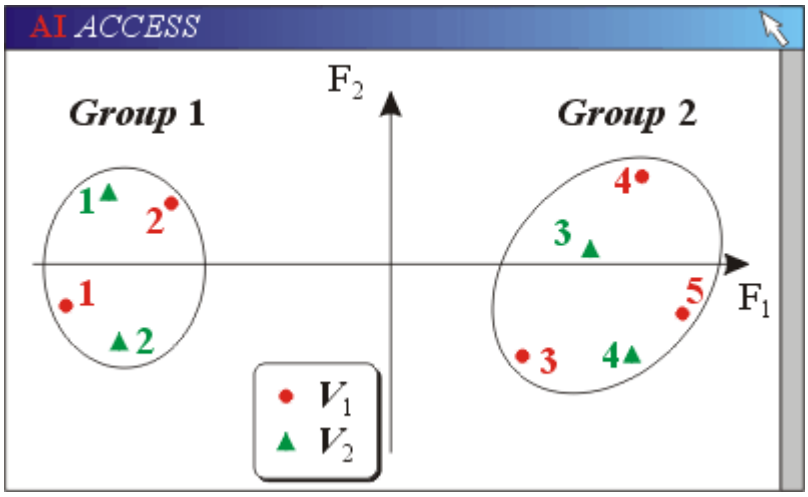
Eigenvalues

First of all, let us state that it can be shown that the inertia carried by any factor is always **less than** 1.



Then let us state that the inertia carried by the first factor for the first variable is the **same** as the inertia carried by the first factor for the second variable, and that the same happens for all factors (PCA had the same property).

In the illustration, V_1 is mentioned as having only I - 1 inertias, whereas it has I modalities. This is because, as we mentioned, the first factor is in fact trivial, and is not mentioned



The first factor is the one that carries the largest inertia.

What if this inertia is close to 1 (largest possible value) ? This is an indication that the projections of the modalities of each variable are sitting far away from the origin. Because this origin is the barycenter of the projections, these have to be distributed on either side of the origin. So, the clouds of modalities break up into two clouds sitting on opposite ends of the first factor (top illustration).

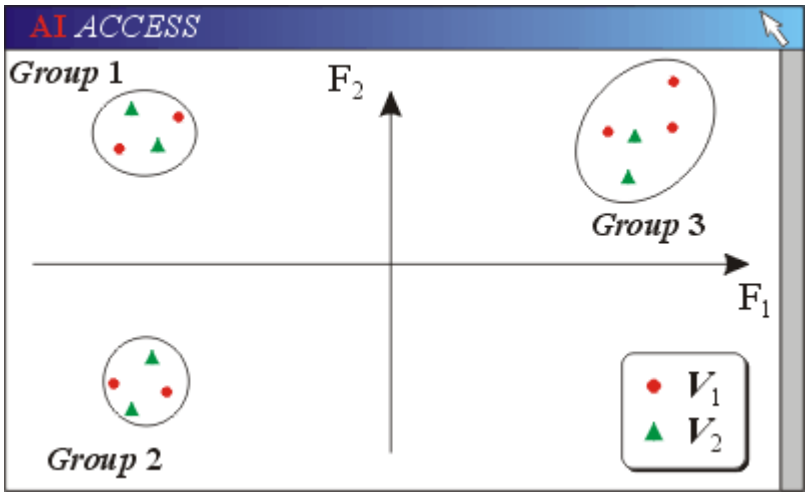
This is a weak form of functional relationship. To make this more apparent, one may reorder the lines and columns of the contengency table as follows :

1) First, columns are ordered by increasing values of the coordinates of the modalities of the column variable on the first factor.

2) Then, the lines are ordered by increasing values of the coordinates of the modalities of the line variable on the first factor.

The result is as shown on the bottom illustration. High value cells bundle up along the diagonal of the new contengency table, which may be perceived as broken up into two blocks.

If there are many modalities, then each group of modalities may be worth being analyzed by a CA of its own.



If both the first and the second factor carry large inertias (close to 1), then the projections are at both ends of the first and second factor, for example because the clouds split up into three groups (see top illustration). The reordering is now a two-step process :

* The first step is as 1) + 2) above.

* Then within each group, the same type of reordering is done with respect to coordinates on the second factor.

This is a new step in the direction of a true functional relationship (bottom illustration)..

Inertia of the modalities

Weights of the modalities

A major difference between CA and standard PCA is that the "observations" (modalities) have **weights** : rows are ponderated by their marginal frequencies, and so are columns. This will have to be kept in mind when interpreting the factors.

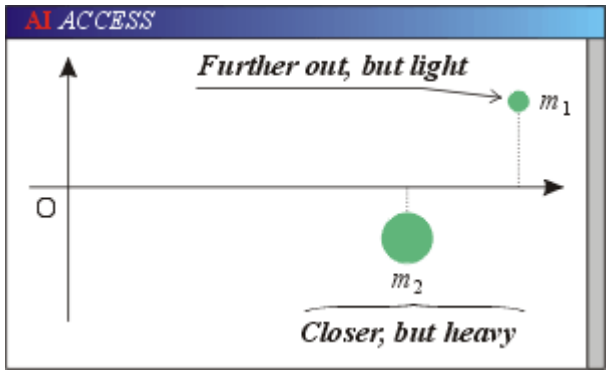
Coordinates, weight and inertia

A modality is just a point in space, and its inertia with respect to the origin (G_r if we consider rows) is defined by :

Inertia of a modality = **weight of the modality**.(Distance to the origin)²

Note that here "distance" means the "Chi-square distance", not the euclidian distance.

Remember that the weight of a modality is its marginal frequency, that is, the proportion of observations with this modality.

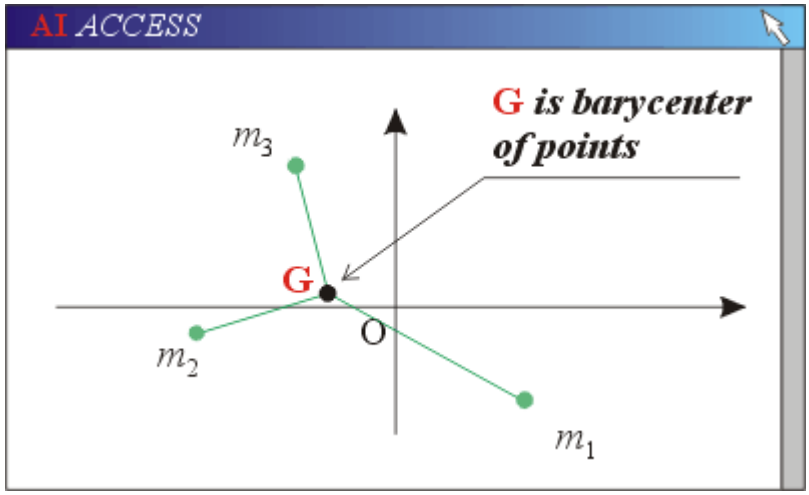


Visual analysis of a projection plot, if not supported by a carfeul scutiny of the contributions, may be quite misleading in this respect : in this illustration, modality m_1 projects further out on the first

factor than modality m_2 does, but because it is also "lighter" than m_2, visual examination cannot tell which modality has the highest inertia on the first factor.

Software sometimes code the weight of a modality as the size of its symbol on the plot.

Barycenters and origin



In the complete space of rows, we called G_r the barycenter of rows. A nice general property of barycenters is that the barycenter of the projections is the same as the projection of the barycenter. Thus, the origin of the plot is always the barycenter of the projected modalities, whether the modalities are well represented on the projection plane or not.

As a consequence, modalities of one variable are never all on one side of a factor, or of any straight line running through the origin.

This property remains true when one considers projections on any factor, rather than the complete plane : the ponderated sum of the coordinates of one variable on one factor is always 0.

Because of the weights, this may not be quite visible on a plot. For example, on the top illustration, the origin O (that is, G_r), is not the barycenter G of the plots of the modalities m_1, m_2, m_3. But in fact, it is so if weight are taken into account (bottom illustration).

Of course, the same applies to the plot of column modalities.

Contribution of a modality to a factor

For any of the two variables, the inertia I_i of a factor i is the sum of the inertias of the modalities of the variable on this factor.

The **contribution** of a modality to a factor i is thus simply the ratio :

Contribution = $(\text{Inertia of the modality})_i / I_i$

It is the **proportion** of the inertia of the factor that is carried by the modality.

The sum of the contributions of the modalities on one factor is, by definition, 1.

Quality of representation of the modalities

Just as in ordinary PCA, a modality may or may not be well represented in projection, depending on whether it is close to the projection plane, or far from it. If you need a little brush-up on the concept of "Quality of representation" and "Squared cosine", we suggest that you take a look here.

As CA interpretation relies heavily on visual interpretation of the projection plots, one has to constantly keep that in mind, and constantly check out the quality of representation of the modalities that are being considered at any one time.

Inertia of the factors

The inertia of a factor is the sum of the inertias of the projections of the modalities of V_1 on this factor. As in ordinary PCA, the factors are ranked and labeled by decreasing order of inertia.

Again, just as in ordinary PCA, it can be shown that if two factors :

* The first one relative to V_1,

* The other one relative to V_2,

have the same rank in their respective space, then they also carry the same inertia. So, in fact, we can talk about the inertia of a factor without making any reference to a variable.

Of course, if both variables do not have the same number of modalities, say I < J, then the J - I last factors of V_2 carry 0 inertia.

We now move on to treating a complete example. Although simpler than most real life problems, its analysis will follow standard guidelines that can be followed for any problem.

INTERPRETING THE FACTORS

We now treat a simple but realistic example. Although real life problems are usually quite a bit more complex, the step-by-step interpretation procedure that we demonstrate here would be very much the same. The treatment of this example covers the next three sections.

The first section covers the interpretation of the factors.

The data

The contingency table

We address the haunting problem of a possible relationship between women's eye color and hair color. Do blond women tend to have blue eyes, dark-haired women black eyes, and red-haired women green eyes ?

Hair color and eye color of 1000 women have been noted. The result is the following contingency table :

Contengency table

	Blond	Red	Dark	Auburn	Total
Blue	159	29	34	142	*364*
Brown	12	44	115	200	*371*
Green	27	24	8	49	*108*
Chestnut	17	24	25	91	*157*
Total	*215*	*121*	*182*	*482*	**1000**

The bottom line contains the populations of the modalities of "Hair_color", while the rightmost column contains the populations of the modalities of "Eye_color". Divide these number by 1000 (the total population of the sample), and you obtain the marginal frequencies $f_{.j}$ and $f_{i.}$.

We note that the most frequent hair color is "Auburn", while the least frequent is "Red". We also note that the most common eye color is "Brown", with "Blue" a close runner-up. "Green" and "Chestnut" are far less frequent.

The Chi-square

If you are not familiar with the "Chi-square test of independence", you may first read here.

* The Chi-square of the contengency table is 234.

* There are 9 degrees of freedom (if this not clear to you, please see here).

* The *p*-value is less than 0.001. Therefore, the $\mathbf{H}_0$ hypothesis "The two variables are independent" can be safely rejected. Yet, at this point, we have no details about the interaction between the variables.

The inertia

Total inertia

Each of the variables have 4 modalities, so there are 4 - 1 = 3 factors.

Factor	Eigenvalue	Proportion	Cumulated
1	0.2079	88.85	88.85
2	0.0235	10.05	98.90
3	0.0026	1.10	100.00

The largest possible value of an eigenvalue is 1, so the largest possible value of the inertia is 3. Here, the total inertia is 0.234, very far from the maximum. We should have expected this result, as we already noted that the value of the inertia is χ^2/n, with n the number of observations in the sample.

So although there is certainly a substantial relationship between the two variables, it is still far from being a functional relationship.

How many factors ?

Recall that every eigenvalue is the amount of inertia carried by the corresponding factor. We see that the first two factors carry almost all of the inertia, and that almost no additional information is expected from the third factor.

We can anticipate, and expect all modalities to be well represented in the (F_1, F_2) plane.

Interpretation of the first factor

The basic principle

We carry over to CA the same methodology that we worked out for PCA. A standard interpretation of a factor is :

"Factor 1 **opposes** "xxx" to the left, to "yyy" to the right" (resp. "Top" and "Bottom")

where "xxx" and "yyy" are expressions in plain language that summarize our understanding of the projections of modalities on the factor. One usually concentrates on a small number of modalities, those whose contributions are largest in absolute value, but are at opposite ends of the factor. Because the plot does not tell about inertias, but only about distances to the origin (coordinates), it is even more **imperative** than in ordinary PCA to constantly refer to the **contributions** of the modalities before proposing an interpretation of the factors.

Which modalities determine the first factor ?

Hair color

We resist the temptation to look at the plot of modalities right away, because it does not show inertias, and therefore may lead to erroneous interpretations of the factors. Rather we refer to the **tables of contributions**. Here is the table of contributions of the variable "Hair_color" relative to the **first factor** :

1st factor	Frequency	Coordinate	Inertia	Contribution
Blond	21.5%	-0.83	0.148	**71.5%**
Red	12.1%	0.13	0.00204	0.9%
Dark	18.2%	0.51	0.0473	**22.6%**
Auburn	48.2%	0.15	0.0108	5.0%

"**Frequencies**" are just the frequencies that we found in the contengency table. The frequencies are the weights used to calculate the inertias of the modalities on the first factor. For example, you may check that the inertia of "Blond" is :

$0.215.(-0.83)^2 = 0.148$

The column labled "Inertia" is often not displayed by software, as the interpretation of a factor will be based of Contributions (percentages), not values..

Each **contribution** is the ratio of the corresponding modality's inertia and the factor's inertia (eigenvalue). For example, the contribution of "Blond" is :

$C(\text{Blond}) = 0.148 / 0.2079 = 0.715$

The contributions, being percentages, add up to 1.

Note that although "Red" and "Auburn" have about equal coordinates on F_1, "Auburn" 's contribution is more than five times larger than that's of "Red". This is because "Auburn" carries a much larger weight, that is, "Auburn" is much more frequent than "Red".

We boldfaced the most important contributions : "Blond" in **red**, because it projects on the negative side of the first factor, and "Dark" in **green**, because it projects on the positive side of the factor.

Eye color

We now show the table of contributions of the other variable ("Eye_color") to the **first factor**.

1st factor	Frequency	Coordinate	Inertia	Contribution
Blue	36.4%	-0.54	0.106	**51.8%**
Brown	37.1%	0.49	0.089	**43.4%**
Green	10.8%	-0.17	0.00312	1.4%
Chestnut	15.7%	0.21	0.00692	3.3%

Note that "Brown" projects about twice as far as "Chestnut" does. With equal frequencies, this would grant "Brown" an inertia about **four times** that of "Chestnut", because inertias are based on the **square** of coordinates. But on top of that, "Brown" is about twice as frequent as "Chestnut", and this accounts for "Brown" 's inertia being over 10 times that of "Chestnut".

Interpretation of the first factor

Each variable can be used separately to interpret the first factor.

* "Hair_color" clearly opposes "Blond" to the left, and "Dark" to the right. The other two modalities ("Red" and "Auburn") have little to say about the meaning of the factor, because their contributions are very low. So we can venture the following suggestion :

"The First Factor opposes light colored hair to the left, to dark colored hair to the right"

It is certainly not shocking to see "Red" and "Auburn" fall somewhere in between.

* "Eye_color" clearly opposes "Blue" to the left, to "Brown" to the right. The other two modalities ("Brown" and "Green") have little to say about the meaning of the factor, because their contributions are very low. So we can venture the following suggestion :

"The First Factor opposes brightly colored eyes to the left, to dull colored eyes to the right"

The second factor

Interpreting the second factor goes exactly along the same lines. Here are the Tables of Contributions.

2nd factor	Frequency	Coordinate	Inertia	Contribution
Blond	21.5%	-0.07	$1.15_{*}10^{-3}$	4.9%
Red	12.1%	0.33	$12.8_{*}10^{-3}$	**54.6%**
Dark	18.2%	-0.22	$9.0_{*}10^{-3}$	**38.1%**
Auburn	48.2%	0.03	$0.56_{*}10^{-3}$	2.4%
2nd factor	**Frequency**	**Coordinate**	**Inertia**	**Contribution**
Blue	36.4%	-0.09	$2.94_{*}10^{-3}$	11.4%
Brown	37.1%	-0.09	$3.01_{*}10^{-3}$	**12.9%**
Green	10.8%	0.35	$13.1_{*}10^{-3}$	**55.6%**

Chestnut	15.7%	0.17	$4.72_{*}10^{-3}$	20.1%

We first notice that the values of the inertias of the modalities are very small. This is not surprising, as the inertia of the second factor is almost ten times smaller that the inertia of the first factor.

* On the "Hair_color" side, "Red" finally takes its revenge : it has the largest responsibility in determining the second factor. This factor may be interpreted as opposing "Brightly colored hair" to the top, to "Dull colored hair" to the bottom. "Auburn" and "Blond" are somewhere in the middle, which is not very important anyway owing to their low contributions.

* On the "Eye_color" side, "Green" is by far the dominant modality. All other modalities have average to low contributions. The negative side is hard to interpret, with "Blue" and "Brown" acting together to balance out "Green" and "Chestnut". But we notice that the frequencies of "Eye_color" project on F_2 in reverse order of their frequencies, with the rarest modality (Green) at the top, the two most frequent modalities (Blue and Brown) at the bottom, and Chestnut somewhere in between. So we may venture the following interpretation : "F_2 opposes rare eye colors at the top, to common eye colors at the bottom".

Other factors

We will not consider the third factor, because it carries so little intertia. Yet, in real life, more complicated problems often require considering higher order factors, should they carry a substantial fraction of the inertia.

Summary of the interpretation of the factors

We now summarize the interpretion of the factors.

1) The Chi-square of the contingency table is 234, with 9 degrees of freedom, and the *p*-value is very low, confirming that "Hair_color" and "Eye_color" are definitely not independent.

2) Yet, the total inertia is a low 0.234 (maximum is 3), discarding the possibility of a functional relationship between the two variables.

3) The first two factors carry almost all of the inertia, with the first factor carrying about 10 times as much inertia as the second factor.

4) The first factor opposes "Light colored hair, brightly colored eyes" to the left, to "Dark colored hair, dull colored eyes" to the right.

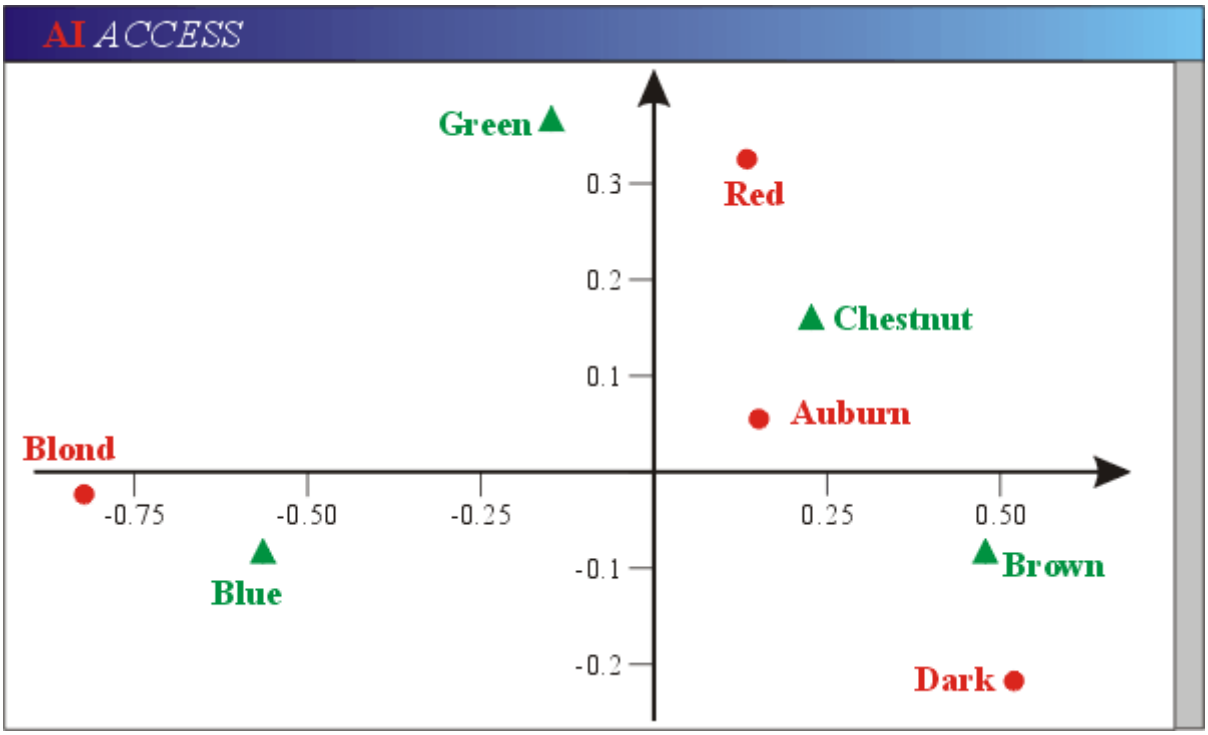
5) The second factor opposes "Brightly colored hair, rare color eyes" to the top, to "Dulled colored hair, common color eyes".

Note that this interpretation was conducted without any reference to a graphical representation. This is because we can't trust a graphical representation for interpreting the factors (contrary to PCA). Interpreting a factor relies on inertias, not coordinates, and inertia does not show on a plot.

Interpreting the plot will be useful for detecting opposing, or associating modalities, and this is what we are doing in the next section.

INTERPRETING THE MODALITIES

We now finally draw the plot of modalities on the first two factors, as determined by the list of coordinates.



The role of the plot of modalities is to **suggest** associations of modalities by pair, belonging :

* either to the same variable,

* or to different variables.

In this section, we address the issue of interpreting each variable individually. So the above "combined" plot of modalities should be thought of as two different overlayed plots : one for

"Hair_color", and the other one for "Eye_color". Software sometimes (but not always) allow displaying these individual plots.

"Quality" or "Square Cosines"

We highlighted "suggest" because such associations cannot be taken at face value just looking at the plot. The reason is that the modalities are seen in projection, and that two modalities may project on two points that are very close, and yet be far from each other in the complete (here, 3D) space. Please see here for more details.

Here are the tables of the Squared Cosines, or "Qualities" of the modalities on the first two factors.

Table of Squared Cosines (or Qualities)

Quality	F_1	F_2
Blond	0.99	0.01
Red	0.12	0.82
Dark	0.83	0.16
Auburn	0.86	0.05

Quality	F_1	F_2
Blue	0.98	0.02
Brown	0.97	0.03
Green	0.17	0.78
Chestnut	0.52	0.36

We do not bother displaying the qualities on the third factor, because we deemed it negligeable already. This is the reason why the sum of the qualities of any of the modalities over the factor falls short of 1.

The qualities play no role in interpreting the factors : only the contributions do. The qualities will be used only when we come to interpreting the relative positions of the modalities, and this is why we delayed introducing them until now.

It is good practice to first examine the variables separately, and software sometimes allow displaying the above plot for either one of the two variables.

Distance to the origin

"Near center" modalities

Remember that the origin represent a "fictitious" modality whose profile is an average profile across the modalities of the other variable.

Let's make this idea a bit more concrete. The bottom line of the contengency table (see here) provides the proportion of each hair color in the sample :

Blond	**Red**	**Dark**	**Auburn**
21.5%	12.1%	18.2%	48.2%

Now suppose a certain fictitious eye color has exactly this profile across hair colors : it would be considered an "average" eye color in that it behaves exactly as the average of all real eye colors.

So an eye color close to the origin has a profile (across hair colors) that is close to this average profile.

Here, no eye color projects near the origin. This mean that there is no such thing as an "average" eye color (at least, as far as hair colors are concerned). Each eye color has a it own specific "signature" across hair colors.

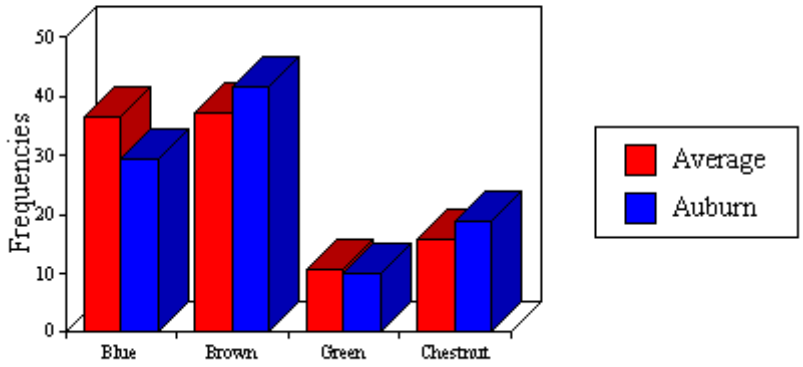
What about an "average hair color" ? The proportions of each eye color in the sample are :

	Blue	**Brown**	**Green**	**Chestnut**
Average	36.4%	37.1%	10.8%	15.7%

If a "hair color" is close to the origin, it should have a profile across eye colors similar to this profile. "Auburn" is the closest hair color to the origin. Its profile is :

Auburn	29.5%	41.5%	10.2%	18.9%

which is indeed a close match.



A word of caution : just because a modality projects near the origin does **not** necessarily mean that the modality is indeed close to the origin. It will be so if and only if, in the complete space, it is close to the projection plane. This can be checked out by adding the qualities of the modality on F_1 and F_2, and making sure that this sum is close to 1. In this case, the quality of "Auburn" is :

$(0.86)^2 + (0.05)^2 = 0.742$

which is reasonably high, and allows us to consider the distance of the "Auburn" point to the origin on the plot as a reasonably faithful representation of the true distance from "Auburn" to the origin in the complete space. At any rate, the above histogram confirms the small distance from the origin.

"Remote" modalities

If a modality projects far from the origin, you may be sure that it is far from the origin in the complete space, for this true distance is at least as large as the projected distance.

A "remote" modality has a profile that departs significantly from average. For example, "Blue" is very far from average. Let us compare its profile with the average profile :

	Blond	**Red**	**Dark**	**Auburn**
Blue	43.7%	8.0%	9.3%	39.0%
Average	21.5%	12.1%	18.2%	48.2%

The pattern of departure from average is clear : in "Blue", there are proportionally many more "Blonds" than average, and fewer other hair color than average.

Each direction from the origin is the signature of a certain type of departure from average (that is, rate of increase or decrease of individual frequencies of modalities of the other variable). So, if several modalities are more or less lined up with a direction from the origin, departures from average distribution are more and more pronouced as you move out along this direction, but the

"pattern" remains the same as long as the qualities of the modalities remain reasonably large. Here, because we have only few modalities, there is no such modality "line up" pattern on any of the two variables.

Heavy modalities

The first idea when interpreting a one-variable plot is that the origin is the barycenter of the modalities weighted by their marginal frequencies. This is also true about the projections of the modalities of the factors : the origin is still the barycenter of the projections of the modalities on a factor (or on any staight line going through the origin, for that matter).

This is true whether the modalities are well represented (high quality) or not.

Now take a heavy modality.

* If it is close to the barycenter, it places little constraint on the position of the other modalities. This is the case of "Auburn", which is both heavy and close to the center.

* But if its distance to the origin is large, the other modalities will be positioned so as to balance out the dominant influenceof the heavy modality. For example, "Blue" is both heavy (36.4%), and a long distance from the origin. So, "Brown", the other heavy modality (37.1%), has to project far out to the right to counterbalance the influence of "Blue".

The situation is even more striking if one considers both "Blue" and "Brown" simultaneously. Both are on the same side of F_2, and even though they are pretty close to F_2, "Chestnut" and "Green" have to be far "North" on the diagram, because to their small weights.

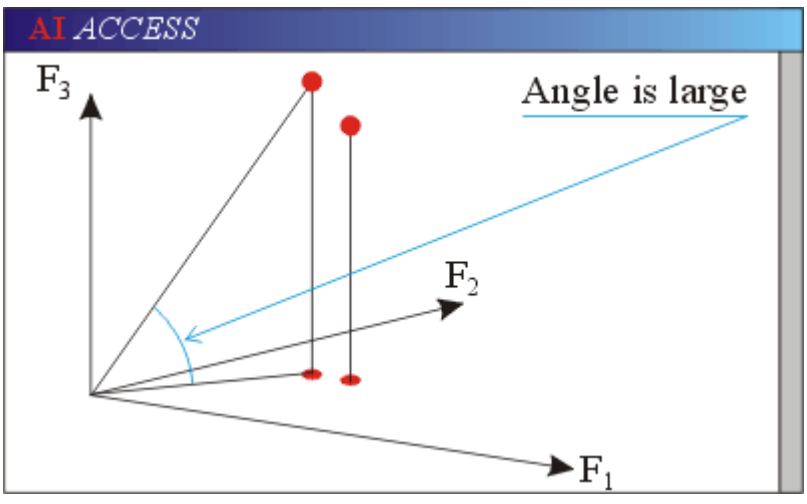
Neighboring modalities

What if two modalities project very close to each other ? Remember that a fundamental property of CA is that if two modalities have identical profiles, then they will coincide in the complete space, and therefore also in projection. It is then possible to merge the two modalities, and in so doing, not change the distances between the modalities of the other variable. This may prove advantageous, as it reduces the number of modalitites, and also may raise useful questions as to **why** these two modalities are so similar.

In this simple example, no pair of modalities are close enough to envision merging them into a "super-modality".

At any rate, the first thing to do when the projections of two modalities (of the same variable) are very close to each other is to verify that the two modalities are indeed close to each other in the complete space.

* If both qualities on the plane are high, then the two modalities are very likely to be close to each other. A quick check on higher order projection planes (e.g. including F_3 or even F_4) will be enough to convince oneself that the modalities are indeed close to each other.



* But it is possible that both modalities have poor qualities with respect to the plane, and yet be close to each

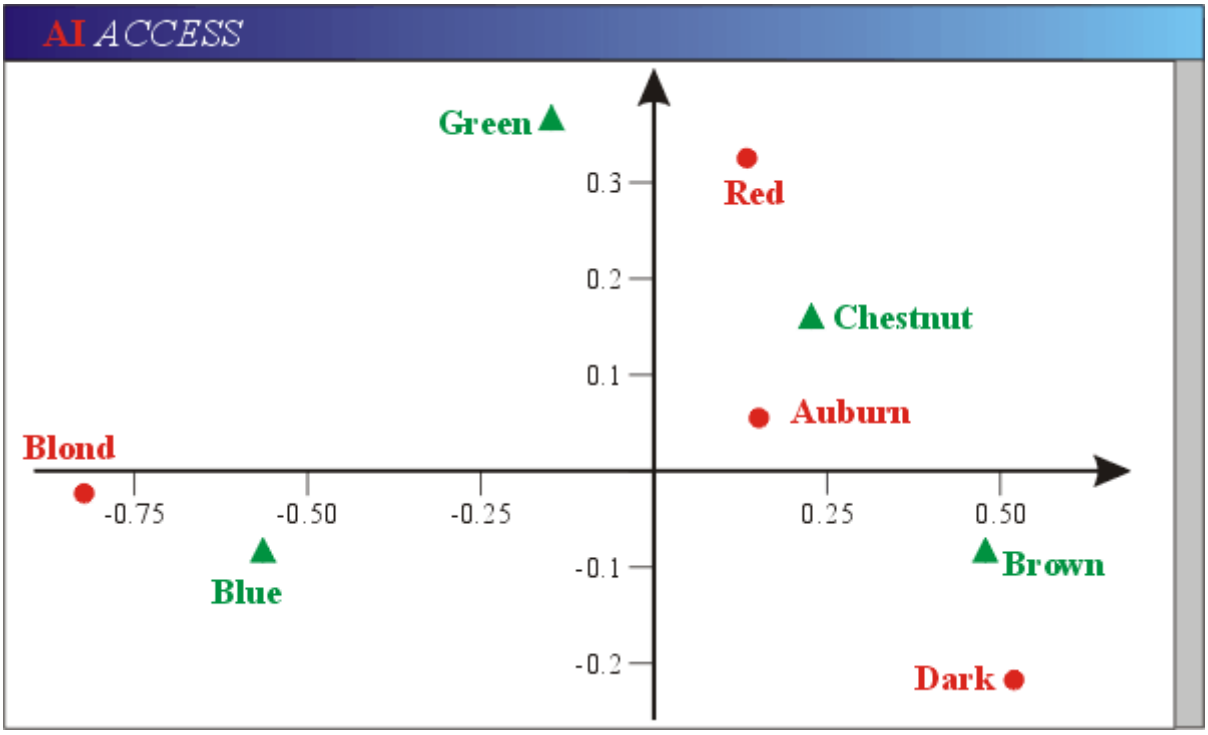
other in reality, as on this illustration.

Then one must check out the set of coordinates of both modalities up to high orders, and verify that these coordinates are never very different from each other. Again, looking at higher order projection planes may prove an easy way to ascertain that the modalities are indeed close to each other, as they will then project on nearby points for all projection planes.

INTERPRETING THE COMBINED PLOT

In the previous section, we interpreted each variable individually. We could thus discover some properties of the modalities that could certainly have been dug out of the contengency table, but that the plot of modalities made lot easy to identify.

We now come to interpreting the combined plot of modalities in order to analyze the interactions between the two variables. For this purpose, we display again the same "combined" plot as we did in the previous section, but this time we will consider both variables simultaneously.



The basic idea

In a nutshell, the idea is that if two modalitites m_{1i} and m_{2j} (belonging respectively to variable V_1 and V_2) are close to each other on the plot, then they exhibit a strong "positive" interaction, which, in turns, means that the (i, j) cell of the contengency table contains a much larger population than the assumption of independence between V_1 and V_2 would lead us to expect.

So, for example, because "Blond" and "Blue" are close to each other on the plot, we are lead to **assume** that there are many more women with both blond hair **and** blue eyes than expected if "Hair_color" and "Eye_color" were truly independent.

So things look nice and simple. In fact, they ar not, for at least two reasons :

1) As for any other Factor Analysis technique, it remains to be determined that this "visual association" is not an artefact due to projection, for fear we draw unjustified (a soft word for "wrong") conclusions from our analysis.

2) But there is a more fundamental reason. We defined the distance between modalities that belong to the **same** variable. But we did not define anything resembling a distance between modalities belonging to **different** variables, and indeed, there is no such a distance. So how can we possibly say that a modality of V_1 is "close" to a modality of V_2 ?

We hope to be able to get back to this point in a later tutorial. At this point, let us just say that we will proceed "as if" the distances on the plot could be interpreted as "distances" between modalities belonging to different variables, but we will not venture into calculating any of these (meaningless) distances.

At any rate, any interpretation made from the plot, and based on the proximity of modalities

belonging to different variables will be thoroughly checked out by refering to the contengency table, that never lies.

Neighboring modalities

So "Blond" and "Blue" are "close" to each other. A rule of thumb of CA is : "Never believe what you see. The plot only suggests, but you have to check out whether the suggestion is appropriate or deceptive".

The most favorable case is when both modalities are well represented on the plane. This means that, for both modalities, the sum of the qualities on the 2 factors defining the plane is high. Here, we have :

* Quality(Blond) = $(0.99)^2 + (0.01)^2 = 0.98$
* Quality (Blue) = $(0.98)^2 + (0.02)^2 = 0.96$

Both modalities are very well represented on the plane, and their perceived distance can be trusted as being very close to their real distance. We therefore can safely state that "Blond" and "Blue" are strongly positively associated. We will in a moment confirm this on the contengency table.

.

What about "Auburn" and "Chestnut" ? They are even closer to each other than "Blond" and "Blue" are. What about their qualities ?

* Quality(Auburn) = $(0.86)^2 + (0.05)^2 = 0.74$
* Quality(Chestnut) = $(0.52)^2 + (0.36)^2 = 0.40$

The quality of "Auburn" is still relatively high, but that of "Chestnut" is definitely poor. This suggests that "Auburn" is near the (F_1, F_2) plane, but that "Chestnut" is further away from the plane.

Why "suggest" and "probably" ? Because "quality" is **not** a measure of the distance to the plane, but only of the (square of the) ratio of the projected distance to the true distance. But because the projected distances are similar, we think that quality is probably a good indicator of the true squared distances to the plane.

So the poor quality of representation of "Chestnut" leaves a doubt about the reality of the association suggested by the plot. So for the time being, we'll just keep this association as a mere possibility that needs to be verified.

The same kind of analysis would suggest that :

* "Brown" is positively associated with "Dark", and to a lesser extent with "Auburn".
* "Green" is strongly associated with "Red".

Confirming with the contingency table

It si good practice to confirm the previous analysis by refering to the contengency table (that contains more complete information than the plot of modalities).

Expected populations

A convenient way to do that is to use the marginal frequencies of the modalities to construct the table of **expected populations** under the assumption that the two variables are independent. If such were the case, then the frequency f_{ij} of cell (i, j) would just be the product of the marginal frequency $f_{i.}$ with the marginal frequency $f_{.j}$. Multiplying these f_{ij} by the total population of the sample yields the table of expected populations. For additional explanations, please refer to the Tutorial on the Chi-square test if independence.

Here is the table of expected populations of our contengency table :

Table of expected populations

	Blond	Red	Dark	Auburn	Total
Blue	78	44	66	175	*364*
Brown	80	45	68	179	*371*
Green	23	13	20	52	*108*
Chestnut	34	19	29	76	*157*
Total	*215*	*121*	*182*	*482*	**1000**

Of course, the marginal populations are the same as for the true contengency table.

A simple way to use this table is, for each cell, to calculate the ratio of the actual population to the expected population. Here is what we get :

Ratios Expected / Actual

	Blond	Red	Dark	Auburn
Blue	**2.04**	*.660*	*0.52*	*0.81*
Brown	*0.15*	*0.98*	**1.69**	**1.12**
Green	**1.17**	**1.85**	*0.42*	*0.94*
Chestnut	*0.50*	**1.26**	*0.86*	**1.20**

We highlighted the cells pointing to positive associations between modalities (numbers larger than 1). A number close to 1 means "Very small interaction between the modalities".

Another way of displaying the same information is through the table of the "Contributions to Chi-square" (please see here for more information).

Contributions to Chi-square

	Blond	Red	Dark	Auburn
Blue	**0.260**	**0.028**	**0.057**	**0.009**
Brown	**0.175**	**0.000**	**0.115**	**0.004**
Green	**0.007**	**0.173**	**0.081**	**0.001**
Chestnut	**0.060**	**0.017**	**0.005**	**0.009**

We highlighted in green those contributions corresponding to an excess of population in the cell with respect to the expected cell count, and in red those corresponding to a deficit. A number close to 0 means "Very small deviation from the expected cell count under the assumption of independence.

We clearly see that :

* The "Blond hair-Blue eyes" combination contributes highly to the Chi-square, and is therefore very significant.

* And so do "Red hair-Green eyes" and the "Dark hair-Brown eyes" combinations, although to a lesser extent.

Note the very low contributions of "Auburn". This confirms that "Auburn" is not strongly associated with any specific eye color.

Software will not always do these little auxiliary calculations. It is our experience that they provide additional enlightment of the CA analysis and are well worth the (little) time spent constructing these extra tables.

An association is not symmetrical

Let us consider the "Blond hair-Blue eyes" association.

What is the proportion of blond-haired women that also have blue eyes ? To answer this question, we go back to the contingency table. There are 215 women with blond hair, of which 159 have blue eyes. So :

159 / 215 = 74%

of blond women also have blue eyes.

Another question is : what is the proportion of blue-eyed women that also have blond hair ? There are 364 women with blue eyes, of which 159 also have blond hair. So :

159 / 364 = 44%

of blue-eyed women are also blond.

These two percentages are quite different. Clearly, "Blond hair" is much better at predicting "Blue eyes" than the converse. This clearly due to the fact that there are many more women with blue eyes than there are blond women.

So the relationship between two modalities is **not** symmetrical.

Summary of the analysis

At this point, we may consider the Analysis as finished. We summarize here its conclusions :

Analysis of the cloud

1) The Chi-square test of independence clearly shows that "Hair_color" and "Eye_color" are not independent.

2) Yet, the value of the total inertia is low, pointing to a large degree of residual randomness in the association.

3) Two factors are enough to account for almost all of the inertia.

Interpretation of the factors

4) The first factor carries about ten times more inertia than the second factor does.

5) The first factor is interpreted as opposing "Light colored hair, brightly colored eyes" to "Dark colored hair, dull colored eyes" to the right.

6) The second factor opposes "Brightly colored hair, rare color eyes" to the top, to "Dull colored hair, common color eyes" to the bottom.

Interpretation of individual variables

7) No eye color distributes across Hair_color in an "average" way. Every eye color has a distinctive pattern of distribution of hair colors.

8) On the other hand, "Auburn" has a distribution across "Eye_color" that is pretty close to average. So knowing that Hair_color is Auburn gives no clue to what "Eye_color" might be.

9) "Blue" is a "remote" modality : its pattern of distribution across "Hair_color" is very different from average. This is mostly due to the large proportion of "Blond hair" in the "Blue eyes" population. Similar statements can be made for other remote modalities as well.

Interpretation of the combined plot

10) "Blond hair" and "Blue eyes" are strongly associated. So are "Red hair" and "Green eyes", and also "Dark hair" and "Brown eyes", but to a lesser extent.

In the next section, we address the issues of :

* Supplementary variables, or modalities.

COMPLEMENTS ON CORRESPONDENCE ANALYSIS

We finally address some additional questions pertaining to the interpretation of the plots :

* Supplemetary variables, which are variables that were not taken into account for building the model, but that are diplayed on the plots and may facilitate their interpretation.

* Ordinal variables, which are categorical variables whose modalities are naturally ordered. In particular, we show how non linear interactions between variables may then be detected by a fundamentally linear technique.

Supplementary variables

Data bases usually have a large number of variables, and choosing the pair of categorical variables that is going to be submitted to a Correspondence Analysis requires some thinking.

Among the variables that are being left out of the analysis, one often finds variables that seem to be closely related to the pair that is being retained. Although they will not feed the algorithm behind CA, they may still be used as a help for interpreting the plots. These additional variables are called **supplementary**, or **passive** variables.

We will work again on our "Hair/Eye color" problem. Suppose there is one other variable, "Country", that we could have used for the analysis instead of say, Hair_color. It has many modalities, but we do not have to take all modalitites into account, as "Country" does not participate actively to the analysis (see below).

Here are the two columns describing the distribution of Eye_colors for two countries.

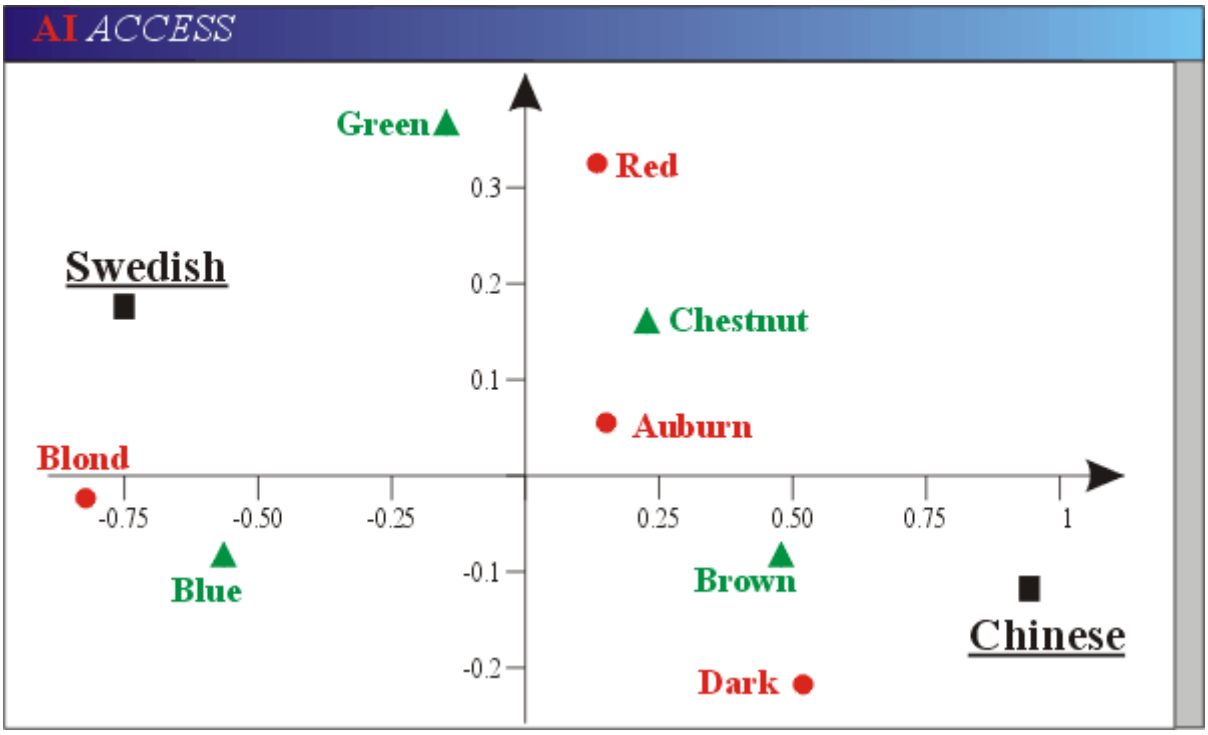
Supplementary modalities

	Swedish	**Chinese**
Blue	65	0
Brown	5	75
Green	20	5
Chestnut	10	20

1) You may select more than just one supplementary variable. Actually, any group of variables may be used as supplementary variables.
2) In our sample, there are as many "Chinese" are there are "Swedish" (100). This means that our sample is not representative of the population of the world at large. So any conclusion we draw is relative to the sample only.

What CA did on "Hair_color" and "Eye_color" was essentially to change the reference frame in the space of "Hair_color" (and in the space of "Eye_color" as well). This change comes about as formulas that work on the cell contents of the contengency table in order to produce coordinates of the modalities on the factors. These formulas depend on the contengency table, but once they are determined, they can be used for **any** other column to produce coordinates of these modalities on the factors. This is what CA does when "Supplementary" (or "Passive") variables are specified. The selected modalities of the supplementary variable(s) appear as new points on the plot :

Supplementary modalities



* Supplementary variables do **not** participate to the determination of the factors, and therefore the concept of "Contribution" is meaningless for the modalities of a supplementary variable.

* For the same reason, the origin is **not** the barycenter of the modalities of a supplementary variable

On the other hand, because a CA is just a rotation of a reference frame, distances are preserved :

* The distances between the modalities of a supplementary variable are meaningful.

* The distance of a modalitiy of a supplementary variable to the origin is meaningful.

* The concept of "Quality" of the representation of a modality of a supplementary variable on a factor is valid. Here, we skip the table of qualities, and ask you to believe that both "Chinese" and "Swedish" are very well represented on the first factor.

Here, we see that "Chinese" is close to both "Brown" and "Dark", which we interpret as meaning that Chinese women are more commonly both dark-haired and brown-eyed that the average of the population of our sample. Similarly, "Swedish" is close to Blond and to Blue, which we interpret in a similar and obvious way.

Besides, both these modalities are at a large distance form the origin. We have to be a little bit careful here, as the origin has two different meanings :

1) It represents a fictitious hair color, that distributes across "Eye_color" as the average of the sample.

2) It represents a fictitious eye color, that distributes across "Hair_color" as the average of the sample.

Now take "Chinese".

1) It is far from the origin, and therefore has a hair color distribution across eye colors that is very different from the average.

2) It is far from the origin, and therefore an eye color distribution across hair colors that is very different from the average.

The same conclusions can be drawn for "Swedish".

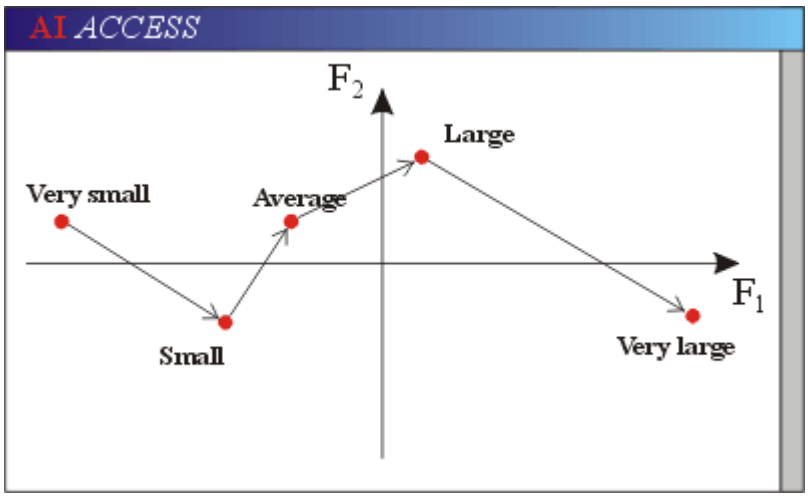
Finally, "Chinese" and "Swedish" are almost opposite to each other, with the origin in between. This means that their distributions across Hair_color (or Eye_color) are very different. For example :

* When "Chinese" is high on "Brown eyes", "Swedish" is low.

* When "Chinese" is high on "Blue eyes", "Chinese" is low.

Ordinal variables

Categorical variables are often "ordinal", that is, there modalities exhibit a natural order. An example is "Size", with modalities "Very large, Large, Average, Small, Very small". Ordinal variables often come as the result of discretizing a continuous variable into contiguous classes.



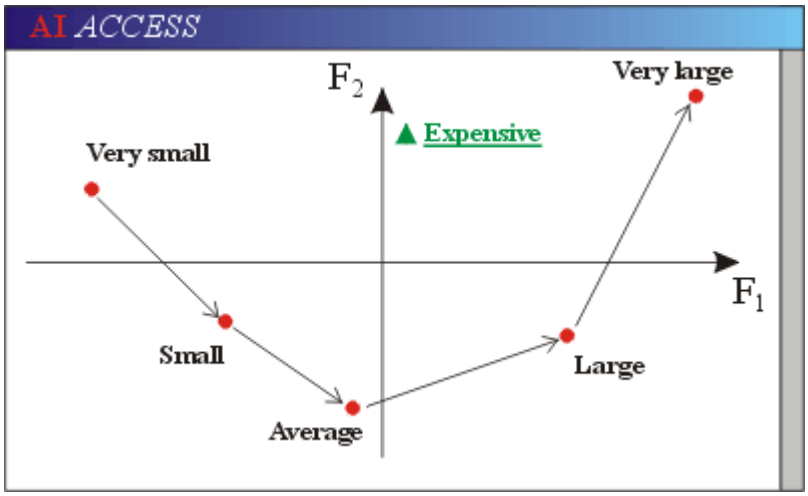
Interpreting the factors

It is good practice to draw a segmented line between the modalities of an ordinal variable, whether active or

supplementary (software will usually do that for you). Although this line contains no additional information, it is a convenient aid to the interpretation of the factors.

In this illustration, the first factor would clearly be interpreted as representing "Size".

The Guttman effect



A particularly interesting situation is when the modalities of an ordinal variable distribute according to a somewhat parabolic shape, as on this illustration. Then the variable :

* Helps interpreting both factors,

* In a particularly instructive way.

* Here, the first factor can be clearly interpreted as "Size".

* The interpretation of the second factor is more subtle : it opposes "Extremes" (Very large, and Very small) to the top, to "Average" at the bottom.

Now consider a modality of the other variable that falls somewhere in between the points representing "Very large" and "Very small", in the "fork" of the parabola. For example, the CA might be conducted by an insurance company who is trying to analyze the relationship between the size of stolen goods and their price.

Assuming that "Expensive" is well represented on the plane, one sees that Expensive stolen goods are not attracted by any particular modality of "Size", and therefore may come in any size, even very small (think of jewlery) or very large (think of cars).

Another interesting consequence of the Guttman effect is the following. The modality "Expensive" ia part of a variable, say "Price", whose other modalities are "Inexpensive", "Average". What CA discovers then is a **non linear** relationship between the two (initially) numerical variables "Price" and "Size". This is quite remarkable, owing to the fact that CA is essentially a linear technique. This non linear relationship could be discovered because of the appropriate discretization of the numerical variables.

The series of steps that we described in this tutorial is very typical of CA interpretation. Yet, we should insist that this analysis was particularly simple :

* Each of the variables has only few modalities. Variables in real life problems often have 6 or more modalities, making the plots a lot harder to analyze.

* Only two factors were enough to account for almost all of the inertia. In real life, it is common that 3 or 4 factors have to be taken into account, especially for the purpose of lifting ambiguities

about close-by projections.

* We were lucky to find a small number of well defined associations. It is often the case that associations are both numerous and loose, and require more careful scrutiny from the practitioner.

* Supplementary variables (or modalities) do not always make the interpretation more clear.

* Ordinal variables do not always nicely line up, or distribute as a clean Guttman effect.