AFC sous Python avec scientisttools

Duvérier DJIFACK ZEBAZE

Ce tutoriel a pour objectif de présenter rapidement les principales fonctionnalités offertes par le package « scientisttools » pour réaliser une Analyse Factorielle des Correspondances.

Présentation des données

Les données sur lesquelles nous allons travailler proviennent du site http://factominer.free.fr/factomethods/analyse-factorielle-des-correspondances.html. Il s'agit des données issues d'un questionnaire réalisé sur des françaises en 1974.

Ces données sont issues d'une enquête du CREDOC publiée en 1974 par Nicole Tabard, intitulée Besoins et aspirations des familles et des jeunes. 1724 femmes ont répondu à différentes questions à propos du travail des femmes, parmi lesquelles :

1. Quelle est selon vous la famille parfaite ?
   — L'homme et la femme travaillent
   — L'homme travaille plus que la femme
   — Seul l'homme travaille
2. Quelle activité est la meilleure pour une mère quand les enfants vont à l'école ?
   — Rester à la maison
   — Travailler à mi - temps
   — Travailler à temps complet
3. Que pensez - vous de la phrase suivante : les femmes qui ne travaillent pas se sentent coupées du monde ?
   — Complètement d'accord
   — Plutôt d'accord
   — Plutôt en désaccord
   — Complètement en désaccord

Le tableau de données est formé de deux tableaux de contingence qui croisent les réponses de la première question à celles des deux autres.

Nous pouvons charger les données sur http://factominer.free.fr/factomethods/datasets/women_work.txt

```
# Chargement des données
import pandas as pd
url = "http://factominer.free.fr/factomethods/datasets/women_work.txt"
women_work = pd.read_table(url,header=0)
women_work.info()
```

```
## <class 'pandas.core.frame.DataFrame'>
## Index: 3 entries, both.man.and.woman.work to only.man.works
## Data columns (total 7 columns):
```

```
##  #   Column                                    Non-Null Count  Dtype
## ---  ------                                    --------------  -----
##  0   stay.at.home                              3 non-null      int64
##  1   part-time.work                            3 non-null      int64
##  2   full-time.work                            3 non-null      int64
##  3   housewives.cut.from.world.totally.agree   3 non-null      int64
##  4   housewives.cut.from.world.quite.agree     3 non-null      int64
##  5   housewives.cut.from.world.quite.disagree  3 non-null      int64
##  6   housewives.cut.from.world.totally.disagree 3 non-null     int64
## dtypes: int64(7)
## memory usage: 192.0+ bytes
```

Table 1 – Données d'enquête

	stay.at.home	part-time.work	full-time.work	housewives.cut.from.world.totally.agree	housewives.cut.from.world.quite.agree	housewives.cut.from.world.quite.disagree	housewives.cut.from.world.totally.disagree
both.man.and.woman.work	13	142	106	107	75	40	39
man.morks.more	30	408	117	192	175	100	88
only.man.works	241	573	94	140	215	254	299

Chaque valeur du tableau 1 correspond au nombre de femmes ayant donnée la réponse en ligne et la réponse en colonne.

Le point de départ de l'analyse est le tableau de contingence reproduit ci-dessous.

Table 2 – Données d'enquête

	stay.at.home	part-time.work	full-time.work
both.man.and.woman.work	13	142	106
man.morks.more	30	408	117
only.man.works	241	573	94

C'est ce type de données (les marges des totaux mis à part) que nous fournirons à la fonction de calcul de l'AFC.

Comme le souligne François Husson dans le MOOC Analyse des données multidimensionnelles sur la plateforme FUN, il est difficile de savoir à partir de ce tableau si les femmes sont favorables ou non au travail féminin. En effet, 908 femmes sur 1724, soit 52% ont répondu que la famille idéale est celle où « Only man works ». Elles sont néanmoins 1123 sur 1724 (65%) à avoir répondu que l'activité convenant le mieux à une mère de famille quand ses enfants vont à l'école est de travailler à mi-temps « part-time work ». L'AFC va nous permettre d'étudier le lien entre ces deux questions et de lever cette apparente contradiction. Elle va notamment nous permettre de visualiser la nature de la liaison entre les deux questions. Mais qu'est ce qu'une liaison ?

Une liaison entre deux variables est l'écart entre les données observées et le modèle d'indépendance. Mettons pour l'instant de côté cette notion, nous y reviendrons plus tard.

Objectifs

Les objectifs de l'AFC sont similaires à ceux de l'ACP : obtenir une typologie des lignes et des colonnes et étudier le lien entre ces deux typologies.

Cependant, le concept de similarité entre les lignes et les colonnes est diférent. Ici, la similarité entre deux lignes ou deux colonnes est complètement symétrique. Deux lignes (resp. colonnes) sont proches l'une de l'autre si elles s'associent aux colonnes (resp. lignes) de la même façon.

On recherche les lignes (resp. colonnes) dont la distribution est la plus différente de celle de la population. Celles qui semblent le plus ou le moins semblables.

Chaque groupe de lignes (resp. colonnes) est caractérisé par les colonnes (resp. lignes) auxquelles il est particulièrement ou particulièrement peu associé.

Nous travaillons d'abord avec seulement les 3 premières colonnes : « Stay at home », « Part time work » et « Full time work ».

```
# Selection des 3 premières colonnes
wfemmes = women_work.iloc[:,:3]
```

Table 3 – Données d'enquête - Tableau des données observées

	stay.at.home	part-time.work	full-time.work
both.man.and.woman.work	13	142	106
man.morks.more	30	408	117
only.man.works	241	573	94

Notons que nous pouvons calculer les marges lignes et les marges colonnes de ce tableau de contingence de la manière suivante :

```
# Ajout des marges lignes et colonnes
wfemmes_avec_marges = wfemmes.copy()
wfemmes_avec_marges.loc["Total",:] = wfemmes.sum(axis=0)
wfemmes_avec_marges.loc[:,"Total"] = wfemmes_avec_marges.sum(axis=1)
```

Table 4 – Données d'enquête avec marge ligne et colonne

	stay.at.home	part-time.work	full-time.work	Total
both.man.and.woman.work	13	142	106	261
man.morks.more	30	408	117	555
only.man.works	241	573	94	908
Total	284	1123	317	1724

Il est aussi intéressant de calculer les pourcentages en ligne et les pourcentages en colonne.

```
# Pourcentages en ligne
import numpy as np
wfemmes_pourcentage_en_ligne = wfemmes.copy()
wfemmes_pourcentage_en_ligne.loc["Profil ligne moyen",:] = wfemmes.sum(axis=0)
```

```
wfemmes_pourcentage_en_ligne = wfemmes_pourcentage_en_ligne.apply(
                                                lambda x : 100*x/np.sum(x),axis=1)
wfemmes_pourcentage_en_ligne.loc[:,"Total"] = wfemmes_pourcentage_en_ligne.sum(
                                                        axis=1)
```

Table 5 – Données d'enquête - Tableau des pourcentages en ligne

	stay.at.home	part-time.work	full-time.work	Total
both.man.and.woman.work	4.98	54.41	40.61	100
man.morks.more	5.41	73.51	21.08	100
only.man.works	26.54	63.11	10.35	100
Profil ligne moyen	16.47	65.14	18.39	100

Pour rappel, la ligne « Profil ligne moyen » correspond à la répartition en pourcentage des modalités à la question sur « l'activité qui convient le mieux à une mère de famille quand les enfants vont à l'école », quelque soit la réponse à la question sur la famille idéale. Le profil ligne moyen peut être comparé aux profils lignes (la répartition en pourcentages ou la distribution de probabilité d'une modalité en ligne). Ici, aucun des trois profils lignes n'est proche du profil ligne moyen.

Calculons maintenant le tableau des pourcentages en colonne

```
# Pourcentage en colonne
wfemmes_pourcentage_en_colonne = wfemmes.copy()
wfemmes_pourcentage_en_colonne.loc[:,"Profil colonne moyen"] = wfemmes.sum(axis=1)
wfemmes_pourcentage_en_colonne = wfemmes_pourcentage_en_colonne.apply(
                                                lambda x : 100*x/np.sum(x),axis=0)
wfemmes_pourcentage_en_colonne.loc["Total",:] = wfemmes_pourcentage_en_colonne.sum(
                                                        axis=0)
```

Table 6 – Données d'enquête - Tableau des pourcentages en colonne

	stay.at.home	part-time.work	full-time.work	Profil colonne moyen
both.man.and.woman.work	4.58	12.64	33.44	15.14
man.morks.more	10.56	36.33	36.91	32.19
only.man.works	84.86	51.02	29.65	52.67
Total	100.00	100.00	100.00	100.00

Ce tableau permet de constater que la répartition des réponses sur la famille idéale pour la modalité « Part-time work » est le plus proche de la répartion des réponses à la question sur la famille idéale. Autrement dit, le profil colonne « Part-time work » est le profil colonne le plus proche du profil colonne moyen. Cette similitude se traduira sur le graphe de l'AFC comme nous le verrons plus loin.

Nous verrons également que l'on passera en paramètre à la fonction Python de calcul de l'AFC, le tableau de contingence. Mais l'AFC travaille en réalité sur le tableau de probabilités que l'on peut calculer en divisant les valeurs du tableau de contingence par le nombre d'individus (on effectue le calcul sur le tableau de contingence avec marge pour mieux constater que l'effectif total du tableau de probabilité est bien égal à 1, ce qui est la marque d'une distribution de probabilités) :

```
# Tableau des probabilités
wfemmes_tableau_de_probabilite = wfemmes_avec_marges/1724
```

Table 7 – Données d'enquête - Tableau de probabilité

	stay.at.home	part-time.work	full-time.work	Total
both.man.and.woman.work	0.00754	0.08237	0.06148	0.15139
man.morks.more	0.01740	0.23666	0.06787	0.32193
only.man.works	0.13979	0.33237	0.05452	0.52668
Total	0.16473	0.65139	0.18387	1.00000

Rappellons que notre objectif est de visualiser la nature de la liaison entre deux variables qualitatives. Mais faut-il encore que cette liaison soit significative. Pour ce faire, nous réalisons un test du Khi2.

Test du χ^2

Le test du χ^2 mesure la significativité d'une liaison mais pas son intensité. Afin de réaliser ce test du χ^2, nous utilisons la fonction chi_contingency de scipy.

```
# Test de contingence du chi2
import scipy.stats as st
stat, pvalue, dof, expected = st.chi2_contingency(wfemmes)
chisq_test = pd.DataFrame({"statistic":stat,"dof":dof,"pvalue":pvalue},
                          index=["chi2 - test"])
print(chisq_test)
```

```
##               statistic  dof        pvalue
## chi2 - test  233.430417    4  2.410248e-49
```

La fonction chi_contingency nous donne, entre autres, la valeur du χ^2 qui est un indicateur de la significativité de la liaison. Mais ce qui nous interesse ici est la p-value. Nous voyons ici que la p-value est égale à $2.4102475 \times 10^{-49}$. Cela signifie que la probabilité que les variables soient indépendantes est égale à $2.4102475 \times 10^{-49}$. Ce qui nous permet de rejeter l'hypothèse d'indépendance entre les deux variables. Pour autant, cela ne veut pas dire que les variables soient dépendantes. Les réponses à la question sur la famille idéale sont probablement liées aux réponses concernant l'activité convenant le mieux à une mère de famille dont les enfants vont à l'école.

Test de χ^2 - Explications

Le test du χ^2 permet de déterminer la probabilité que les deux variables d'un tableau de contingence soient indépendantes, c'est-à-dire qu'il n'existe pas de relation entre les modalités en ligne et les modalités en colonne (les unes ne conditionnent pas les autres, et réciproquement). Dit autrement et comme le rappelle très clairement Julien Barnier, cela veut dire que le « fait d'appartenir à une modalité de la première variable n'a pas d'influence sur la modalité d'appartenance de la deuxième variable ». Dans ce test, l'hypothèse nulle (H0) suppose qu'il y a indépendance entre les deux variables. Si nous acceptons l'hypothèse d'indépendance (H0), nous n'aurons pas d'utilité à réaliser une AFC car les points projetés seront extrêmement proches ou

confondus avec le centre de gravité, confondus avec le centre du graphe. Si nous rejetons l'hypothèse d'indépendance (p-value $< 0,05$), l'hypothèse alternative (H1) suppose que la liaison entre les deux variables est significative sans que nous puissions définir l'intensité de la liaison.

Rappelons que pour que le test du χ^2 soit opératoire, il doit respecter un certain nombre de conditions (pour reprendre les propos de Claude Grasland) :

— L'effectif total du tableau de contingence doit être supérieur ou égal à 20.
— L'effectif marginal du tableau de contingence doit toujours être supérieur ou égal à 5.
— L'effectif théorique des cases du tableau de contingence doit être supérieur à 5 dans 80% des cases du tableau de contingence.

Du fait que nous ayons obtenu une p-value égale à $2.4102475 \times 10^{-49}$ et, par extension, inférieure au seuil de $0,05$, nous rejetons l'hypothèse d'indépendance entre les deux variables.

Test du χ^2 - Aide à l'interprétation

Le test du χ^2 est symétrique. Les lignes et les colonnes du tableau croisé sont interchangeables. Le résultat du test sera exactement le même. Il n'y a pas de « sens de lecture » du tableau.

Nous pouvons afficher le tableau d'indépendance (tableau des effectifs théoriques) en sélectionnant la valeur `expected`. Dans ce contexte, nous calculons le tableau des pourcentages théoriques, en multipliant pour chaque case la proportion observée dans la population des deux modalités correspondantes. Puis, le tableau des effectifs théoriques se calcule en multipliant le tableau des pourcentages théoriques par l'effectif total.

```
# Tableau des effectifs théoriques
effectif_theorik = pd.DataFrame(expected,index=wfemmes.index,
                                columns=wfemmes.columns)
```

Table 8 – Données d'enquête - Tableau des effectifs théoriques

	stay.at.home	part-time.work	full-time.work
both.man.and.woman.work	42.99536	170.0133	47.9913
man.morks.more	91.42691	361.5226	102.0505
only.man.works	149.57773	591.4640	166.9582

Le tableau des effectifs théoriques n'a que peu d'intérêt en lui-même mais en a davantage comparativement au tableau des données observées.

Nous pouvons aussi afficher le tableau des résidus standardisés (tableau des écarts à l'indépendance). Un résidu standardisé positif signifie que les effectifs dans la case sont supérieur à ceux attendus sous l'hypothèse d'indépendance. Et l'inverse pour un résidu standardisé négatif.

```
# Residus standardisés
standardized_residuals = (wfemmes - effectif_theorik)/np.sqrt(effectif_theorik)
```

Table 9 – Données d'enquête - Résidus standardisés

	stay.at.home	part-time.work	full-time.work
both.man.and.woman.work	-4.57450	-2.14844	8.37359
man.morks.more	-6.42424	2.44441	1.47986
only.man.works	7.47513	-0.75921	-5.64638

Exprimé d'une autre manière, l'écart à l'indépendance représente l'écart entre l'effectif observé et l'effectif théorique, et ceci pour chacune des cases du tableau de contingence. D'ailleurs, comme le note Philippe Cibois, l'écart à l'indépendance « est un effectif et c'est un invariant, indépendant du choix des lignes et des colonnes (c'est la différence entre l'effectif observé et l'effectif théorique : le résultat est donc un effectif). » Par ailleurs,

— Un écart à l'indépendance positif correspond à une attraction entre les deux modalités pour la case observée.
— À l'inverse, un écart à l'indépendance négatif correspond à une opposition entre les deux modalités pour la case observée.

Plus la valeur de l'écart à l'indépendance est importante, plus l'attraction/opposition entre les modalités est forte.

AFC

Notre objectif est bien de visualiser la nature de la liaison entre les deux variables qualitatives. Sachant qu'une liaison correspond à l'écart entre les données observées et le modèle d'indépendance, nous souhaitons donc visualiser la nature de l'écart à l'indépendance entre deux variables qualitatives.

Par ailleurs, il y a trois façons de caractériser la liaison entre les deux variables qualitatives.

— La significativité de la liaison (qui se mesure avec le test du χ^2).
— L'intensité de la liaison (qui se mesure, entre autre, avec le ϕ^2).
— La nature de la liaison (qui correspond à l'association entre les modalités et qui est représentée par le biais de l'AFC).

Le test du χ^2 a permis d'écarter l'hypothèse d'indépendance. Il y a donc une liaison entre les modalités des deux variables. De fait, nous pouvons faire une AFC pour visualiser la nature de la liaison. Pour notre part, nous avons choisi d'utiliser le package « scientisttools » (dédié à l'analyse multidimensionnelle de données).

On utilisera les trois première colonnes (correspondant aux réponses de la deuxième question) comme variables actives et les quatre dernières (correspondant à la troisième question) comme variables illustratives.

Nous chargeons donc la librairie « scientisttools »

```
# Chargement de la librairie
from scientisttools.decomposition import CA
```

Lignes et colonnes actives seulement

Lors du précédent test du χ^2, nous avons obtenu une p-value égale à $2.4102475 \times 10^{-49}$. Nous avons donc rejeté l'hypothèse d'indépendance entre les deux variables et admis que la liaison entre ces deux variables est significative. Nous sommes en droit de réaliser une AFC afin de visualiser la nature de la liaison. Pour ce faire, nous allons employer la fonction `CA`, fournie par le package « scientisttools ».

On crée une instance de la classe CA, en lui passant ici des étiquettes pour les lignes et les colonnes. Ces paramètres sont facultatifs ; en leur absence, le programme détermine automatiquement des étiquettes.

```
# Instanciation du modèle
my_ca = CA(n_components=None,
           row_labels=wfemmes.index,
           col_labels=wfemmes.columns,
           row_sup_labels=None,
           col_sup_labels=None)
```

On estime le modèle en appliquant la méthode `fit` de la classe CA sur le jeu de données.

```
# Entraînement - Estimation du modèle
my_ca.fit(wfemmes)
```

```
## CA(col_labels=Index(['stay.at.home', 'part-time.work', 'full-time.work'], dtype='object')
##    row_labels=Index(['both.man.and.woman.work', 'man.morks.more', 'only.man.works'], dtyp
```

Valeurs propres

L'exécution de la méthode `my_ca.fit(wfemmes)` provoque le calcul des attributs parmi lesquels `my_ca.eig_` pour les valeurs propres.

```
# Valeurs propres
print(my_ca.eig_)
```

```
## [[1.16840024e-01 1.85604492e-02]
##  [9.82795752e-02            nan]
##  [8.62921829e+01 1.37078171e+01]
##  [8.62921829e+01 1.00000000e+02]]
```

L'attribut `my_ca.eig_` contient :

- en 1ère ligne : les valeurs propres en valeur absolue
- en 2ème ligne : les différences des valeurs propres
- en 3ème ligne : les valeurs propres en pourcentage de la variance totale (proportions)
- en 4ème ligne : les valeurs propres en pourcentage cumulé de la variance totale.

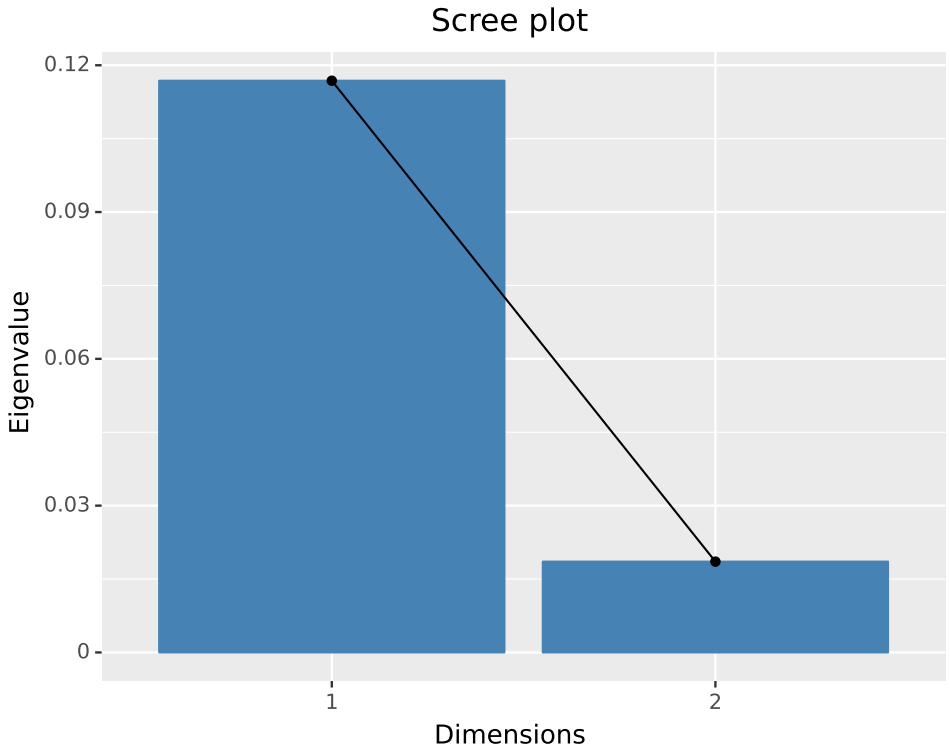
La fonction `get_eig` retourne les valeurs propres sous forme de tableau de données.

```
# Valeurs propres
from scientisttools.extractfactor import get_eig
print(get_eig(my_ca))
```

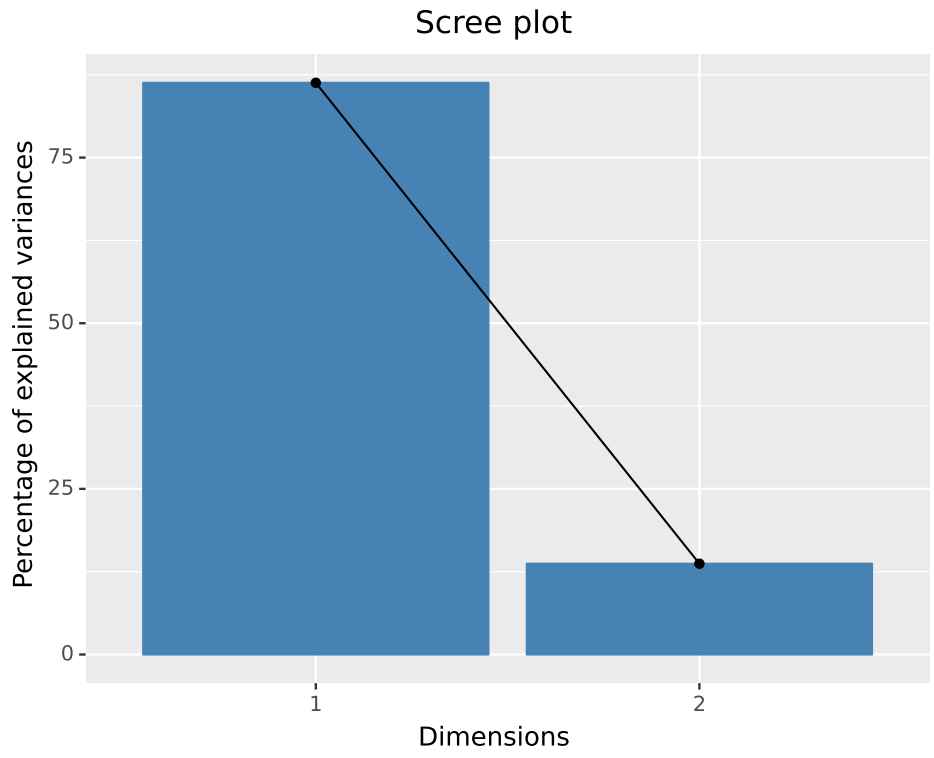
```
##        eigenvalue  difference  proportion  cumulative
## Dim.1     0.11684     0.09828   86.292183   86.292183
## Dim.2     0.01856          NaN   13.707817  100.000000
```

Les valeurs propres peuvent être représentées graphiquement

```
from scientisttools.ggplot import fviz_eigenvalue
print(fviz_eigenvalue(my_ca,choice="eigenvalue"))
```




```
print(fviz_eigenvalue(my_ca,choice="proportion"))
```



On peut obtenir un résumé des principaux résultats en utilisant la fonction `summaryCA`.

```
from scientisttools.extractfactor import summaryCA
summaryCA(my_ca)
```

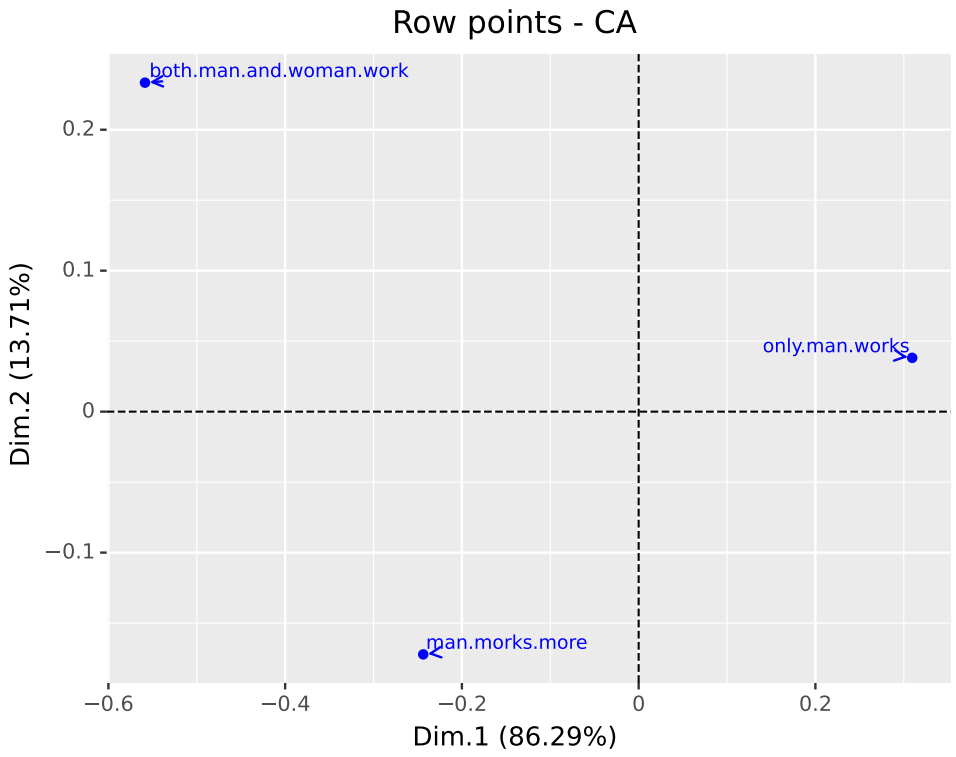
```
##                       Correspondence Analysis - Results
##
## Importance of components
##                             Dim.1    Dim.2
## Variance                    0.117    0.019
## Difference                  0.098      NaN
## % of var.                  86.292   13.708
## Cumulative of % of var.    86.292  100.000
##
## Rows
##
##                           d(i,G)   p(i)  I(i,G)  ...  Dim.2     ctr   cos2
## both.man.and.woman.work    0.367  0.151   0.055  ...  0.233  44.429  0.149
## man.morks.more             0.089  0.322   0.029  ... -0.172  51.436  0.333
## only.man.works             0.097  0.527   0.051  ...  0.038   4.135  0.015
##
## [3 rows x 9 columns]
##
## Columns
##
##                  d(k,G)   p(k)  I(k,G)  Dim.1  ...   cos2  Dim.2     ctr   cos2
## stay.at.home      0.416  0.165   0.068  0.618  ...  0.920  0.183  29.613  0.080
## part-time.work    0.010  0.651   0.006 -0.004  ...  0.001 -0.100  34.853  0.999
## full-time.work    0.329  0.184   0.060 -0.541  ...  0.891  0.189  35.533  0.109
##
## [3 rows x 9 columns]
```

Cette fonction `summaryCA` nous permet d'obtenir :

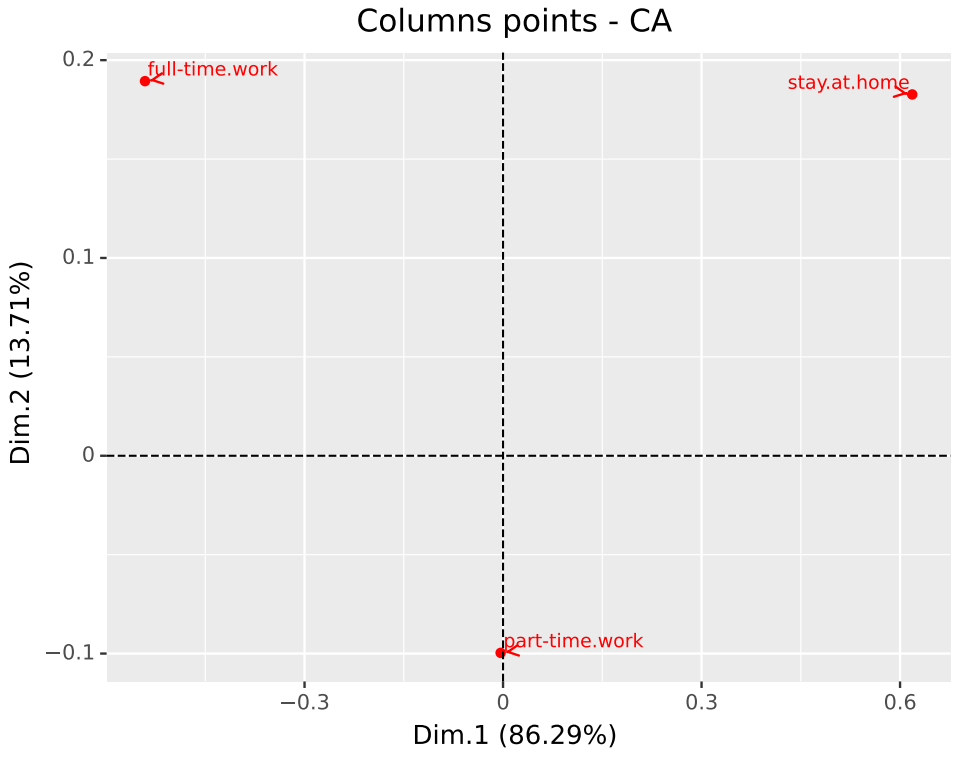
- Un tableau avec les valeurs propres, les différences, les pourcentages et les pourcentages cumulés d'inertie associés à chaque dimension.
- Un tableau avec les résultats sur les lignes actives avec leur coordonnées (Dim.n) sur chaque dimension, leur contribution à la construction (ctr) de chaque dimension et leur qualité de représentation (cos2) sur chaque dimension.
- Un tableau avec les résultats sur les colonnes actives (Dim.n, ctr, cos2)

Représentation graphique

```
# Carte des points lignes
from scientisttools.ggplot import fviz_ca_row
print(fviz_ca_row(my_ca,color="blue",repel=True))
```




```
# Carte des points colonnes
from scientisttools.ggplot import fviz_ca_col
print(fviz_ca_col(my_ca,color="red",repel=True))
```

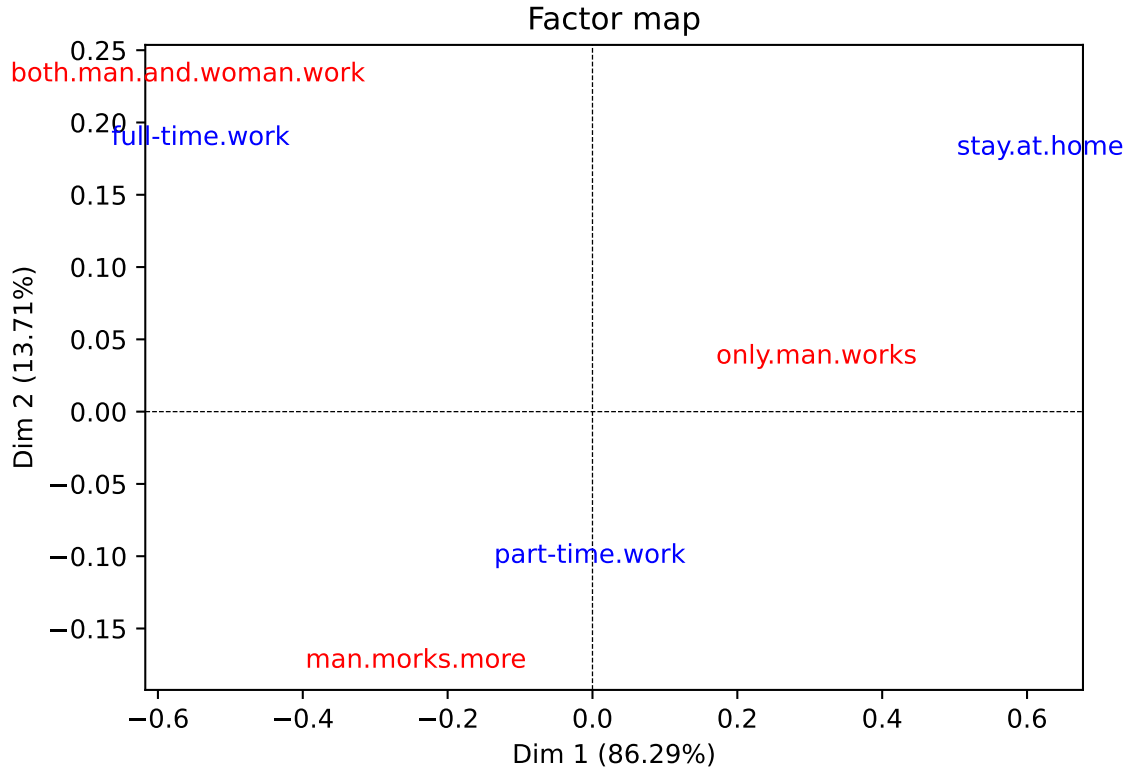


Le nuage des colonnes montre que le premier axe oppose « Stay at home » et « Full-time work », ce qui signifie qu'il oppose deux profils de femmes. Les femmes qui ont répondu « Stay at home » ont répondu « Only husband works » plus souvent que l'ensemble de la population et « Both husband and wife work » moins souvent que l'ensemble de la population.

De même, les femmes qui ont répondu « Full-time work » ont répondu « Only husband works » moins souvent que l'ensemble de la population et « Both husband and wife work » plus souvent que l'ensemble de la population. Le premier axe ordonne les modalités de la deuxième question de la moins à la plus en faveur du travail des femmes.

La même interprétation peut être faite pour le premier axe du nuage des lignes. Les modalités sont triées de la moins (« Only husband works ») à la plus (« Both husband and wife work ») en faveur du travail des femmes.

On peut représenter à la fois les lignes et les colonnes.



« Stay at home » est associé à « Only husband works » et peu associé aux deux autres modalités.

« Both husband and wife work » est associé à « Full-time work » et opposé à « Stay at home ».

Revenons un instant sur les données du tableau 1, issu de l'enquête de Nicole Tabard, croisant les deux variables qualitatives (questions) :

— Quelle est la famille idéale pour vous ?
— Quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ?

Il est important de rappeler que les résultats de cette enquête ont été publiés en 1974. Il est fort à parier que la répartition des réponses serait totalement, si ce n'est en grande partie, différente aujourd'hui.

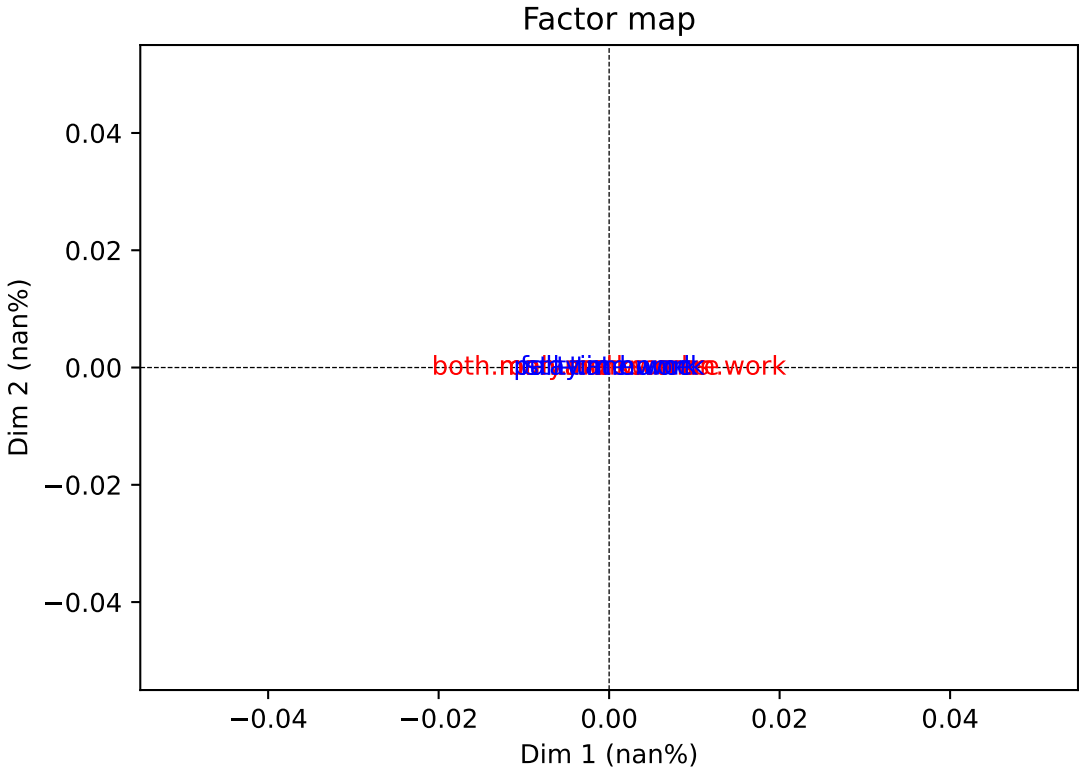
Lors d'une première lecture de ce tableau de contingence, François Husson soulève une apparente contradiction. À la question « Quelle est la famille idéale pour vous ? », nous voyons que 908 femmes sur 1724 (visible dans la marge colonne), soit environ 53% des répondantes, déclarent « Only man works » et seulement 261 femmes sur 1724 (environ 15%) déclarent « Both man and woman work ». Sur la base de ces premières réponses, nous pouvons émettre l'hypothèse, qu'à cette époque, une majorité était en faveur d'un modèle familial où seul le mari travaille.

À côté de ça, à la question « Quelle activité convient le mieux à une mère de famille quand ses enfants vont à l'école ? », elles sont 1440 sur 1724 (visible dans la marge ligne), soit environ 84%, à être en faveur du travail à mi-temps « Part time work » ou à plein-temps « Full time work ». Les réponses à cette question semblent indiquer que les femmes sont moins hostiles au travail féminin (bien au contraire).

Du coup, à ce stade de l'interprétation, nous nous retrouvons a priori face une contradiction. De cela, nous pouvons dire que le tableau de contingence ne permet pas de savoir si les femmes des années 70 sont favorables ou non à l'activité féminine. Par contre, Une première lecture du graphe de l'AFC nous permet de dire que les modalités des réponses s'associent entre elles des plus favorables au travail féminin aux plus défavorables au travail féminin.

Avant d'approfondir, plus en détail, l'interprétation de cette AFC, nous allons faire un pas de côté et voir ce qui se passe dans le cas où il y aurait indépendance entre les deux variables.

Si nous réalisons une AFC avec les données du modèle d'indépendance, on obtient la figure suivante :



La lecture de ce graphique nous permet de voir que les points sont quasiment tous confondus avec le centre de gravité, correspondant au profil moyen. La représentation graphique est trompeuse mais l'échelle des axes va dans le sens de notre interprétation. Simplement, ce qu'il y

a retenir de ce graphe, c'est que, lorsqu'il y a indépendance entre les deux variables, tous les points sont confondus avec l'origine. Du fait qu'il n'y ait pas d'écarts à l'indépendance, il n'y a graphiquement rien à exploiter, rien à interpréter, rien à analyser. Ce graphe donne à voir ce que nous avons précédemment énoncé, à savoir que :

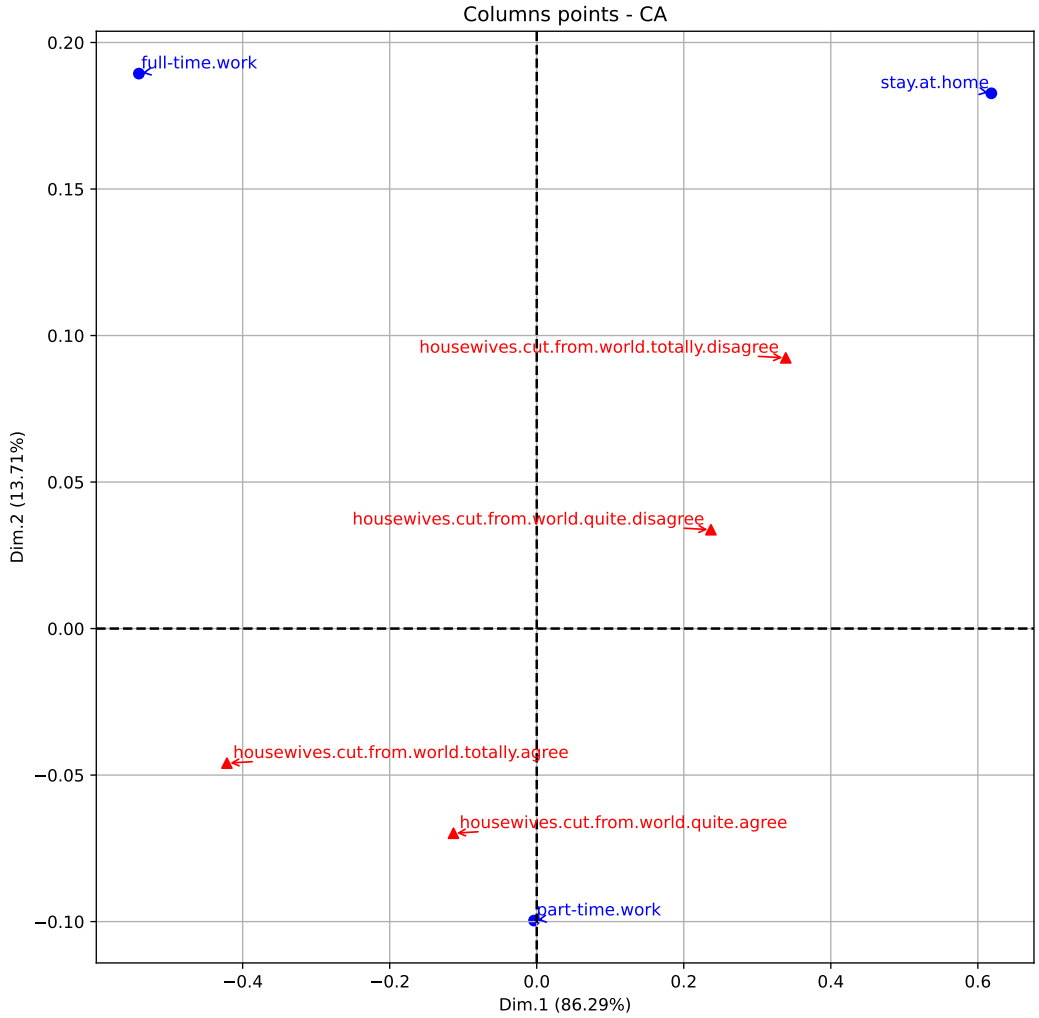
— Si nous acceptons l'hypothèse d'indépendance (p-value > 0.05 dans le cas d'un test du χ^2), nous n'aurons pas d'utilité à réaliser une AFC car les points projetés seront extrêmement proches ou confondus avec le centre de gravité, confondus avec le centre du graphe.
— La réalisation d'un test du χ^2 est donc fortement conseillée avant la réalisation d'une AFC.
— Plus précisément, le test du χ^2 conditionne l'éventuelle réalisation d'une AFC.

Addition de colonnes illustratives

On ajoute les colonnes qui correspondent à la troisième question en tant que variables illustratives. Tapez :

```
# Modèle avec colonnes supplémentaires
my_ca2 = CA(n_components=None,
            row_labels=women_work.index.values,
            col_labels=women_work.columns[:3].values,
            row_sup_labels=None,
            col_sup_labels=women_work.columns[3:].values).fit(women_work)
```

```
# Carte de modalités colonnes
import matplotlib.pyplot as plt
from scientisttools.pyplot import plotCA
fig,axe = plt.subplots(figsize=(10,10))
plotCA(my_ca2,choice ="col",add_sup=True,color="blue",repel=True,ax=axe)
plt.show()
```



« Totally agree » et « Quite agree » pour « Women who do not work feel cut off from the world » sont proches des modalités en faveur du travail des femmes.

« Quite disagree » et « Totally “disagree » sont proches des modalités opposées au travail des femmes.

Pour ajouter des points lignes illustratifs, utilisez l'argument suivant de la fonction PCA :

```
row_sup_labels
```

Tous les résultats détaillés peuvent être vus dans l'objet `my_pca2`. On peut récupérer les valeurs propres, les résultats des points lignes actifs et illustratifs, les résultats des points colonnes actifs et supplémentaires en tapant :

```
from scientisttools.extractfactor import get_ca_row,get_ca_col,get_eig
eig = get_eig(my_ca2)
```

```
row = get_ca_row(my_ca2)
print(row.keys())
```

```
## dict_keys(['coord', 'cos2', 'contrib', 'dist', 'res.dist', 'infos'])
```

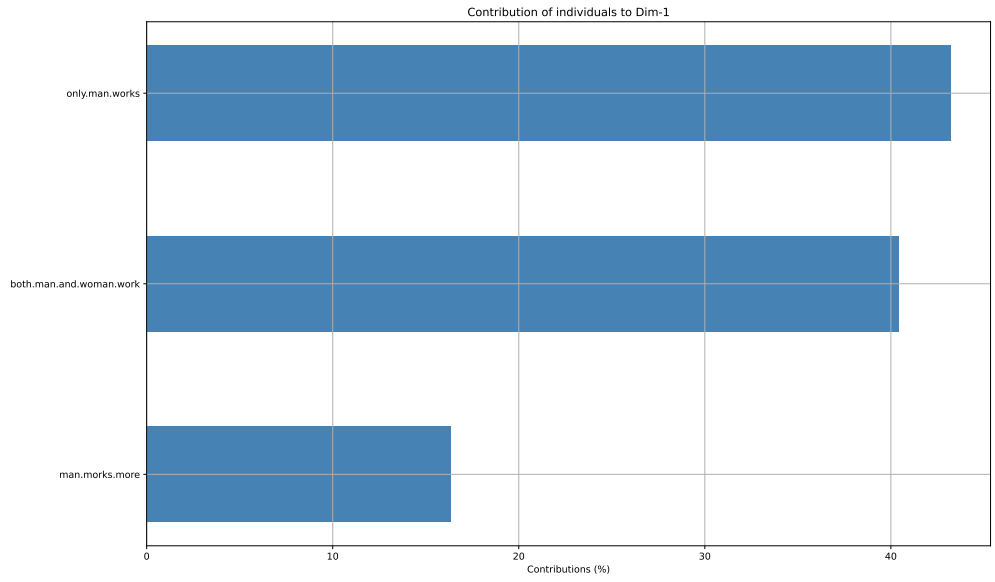
```
col = get_ca_col(my_ca2)
print(col.keys())
```

```
## dict_keys(['coord', 'cos2', 'contrib', 'dist', 'res.dist', 'infos', 'col_sup'])
```

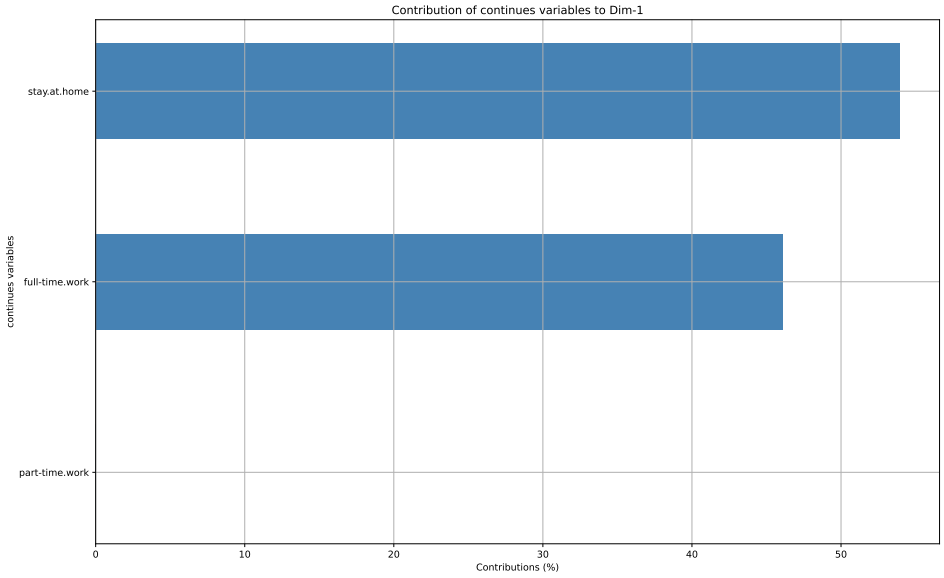
Interprétation des axes

Des graphiques qui permettent d'interpréter rapidement les axes : on choisit un axe factoriel (le 1er axe dans notre exemple) et on observe quels sont les points lignes et colonnes qui présentent les plus fortes contributions et cos2 pour cet axe.

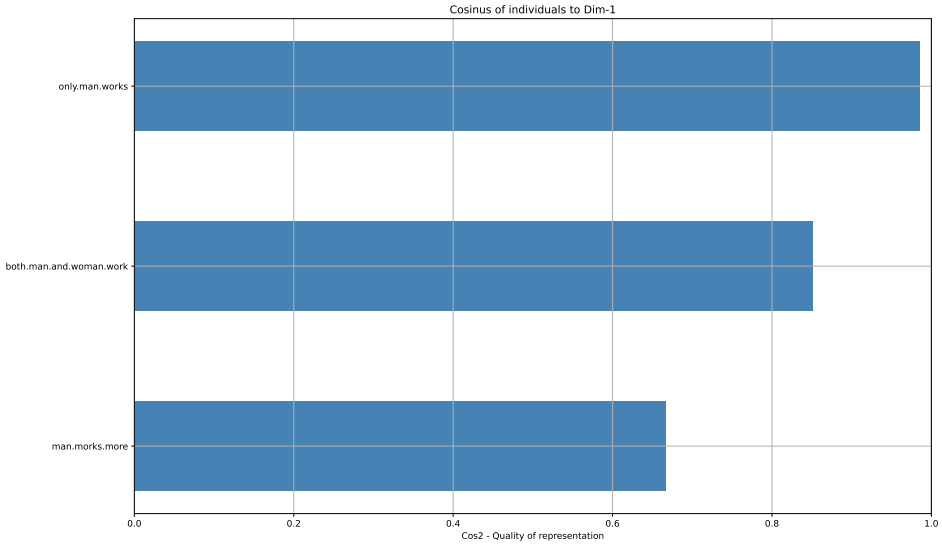
```
# Classement des points lignes en fonction de leur contribution au 1er axe
from scientisttools.pyplot import plot_contrib, plot_cosines
fig,axe = plt.subplots(figsize=(16,10))
plot_contrib(my_ca,choice="ind",axis=0,top_contrib=10,ax=axe)
plt.show()
```



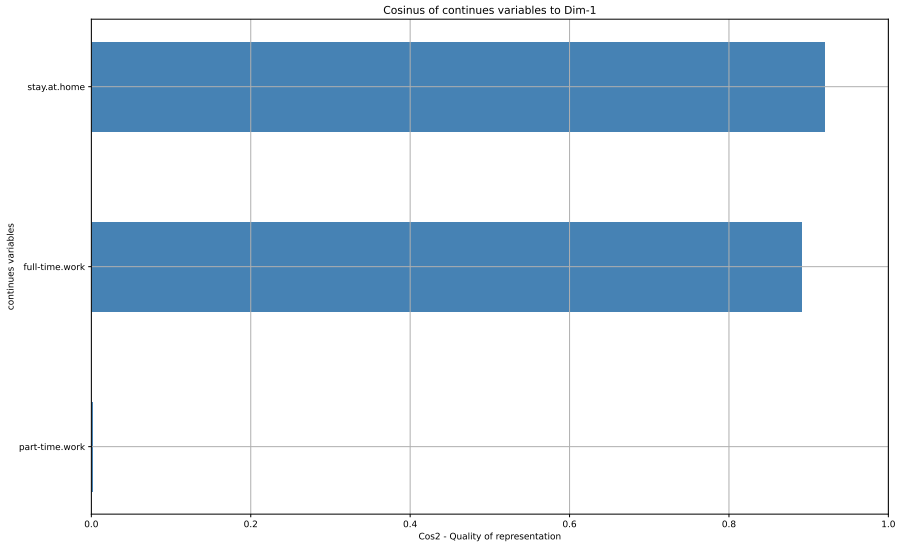

```
# Classement des points colonnes en fonction de leur contribution au 1er axe
fig,axe = plt.subplots(figsize=(16,10))
plot_contrib(my_ca2,choice="var",axis=0,ax=axe)
plt.show()
```




```
# Classement des points lignes en fonction de leur cos2 sur le 1er axe
fig,axe = plt.subplots(figsize=(16,10))
plot_cosines(my_ca2,choice="ind",axis=0,top_cos2=10,ax=axe)
plt.show()
```




```
# Classement des points colonnes en fonction de leur cos2 sur le 1er axe
fig,axe = plt.subplots(figsize=(16,10))
plot_cosines(my_ca2,choice="var",axis=0,ax=axe)
plt.show()
```



Pour plus d'informations sur l'ACF sous scientisttools, consulter le notebook https://github.com/enfantbenidedieu/scientisttools/blob/master/notebooks/ca_example2.ipynb