

# 1. Objectif

(11/03/2007) ACP sur corrélations partielles dans TANAGRA - Sections 1 à 6.

(16/06/2012, suite) Montrer l'implémentation de l'ACP sur corrélations partielles dans le logiciel SAS – Voir Section 7. Puis, programmation d'une procédure pour le logiciel R – Voir Section 8.

Le coefficient de corrélation linéaire mesure l'intensité de la liaison entre deux variables quantitatives. Il est utilisé dans de nombreuses situations, entre autres dans l'analyse en composantes principales (ACP) pour résumer les principales informations portées par un fichier de données.

Pour pratique qu'il soit, le coefficient de corrélation peut être trompeur. L'extrapolation de la corrélation à la causalité doit être faite avec précaution. Notamment parce qu'il peut y avoir une ou plusieurs variables supplémentaires, connues ou inconnues, qui influent sur les variations des variables étudiées, laissant à penser qu'il existe un lien entre ces variables. Ces tierces variables, on parle de facteurs confondants ou de variables modératrices, sont la cause de bien des problèmes dans les études réelles. Elles induisent des conclusions totalement faussées. Bien souvent, nous devons nous en remettre à l'expertise du domaine pour les circonscrire. Il importe alors de les traiter convenablement.

Dans ce tutoriel, nous montrons le fonctionnement du composant RESIDUAL SCORES de TANAGRA. Son rôle est d'enlever dans les variables cibles la variabilité causée par une série de variables annexes, qui ne semblent pas directement impliquées dans l'étude, mais qui en réalité pèsent énormément sur les résultats. Cela permet de mettre en oeuvre des études « toutes choses égales par ailleurs » où l'on ramène l'ensemble des variables à un référentiel commun.

## 2. Données

Nous travaillons sur le fichier BODY.XLS. Il recense les dimensions de différentes parties du corps (circonférence des chevilles, du coude, du genou, de la taille, des hanches, etc.). L'objectif est de vérifier s'il existe un lien entre les dimensions de ces différentes parties du corps. Trois variables supplémentaires sont disponibles : le poids, la taille et le sexe des individus étudiés. A priori, ces variables n'ont pas de rôle direct à jouer dans notre étude, nous verrons plus loin qu'elles tiennent en réalité une place considérable.

## 3. Analyse en composantes principales

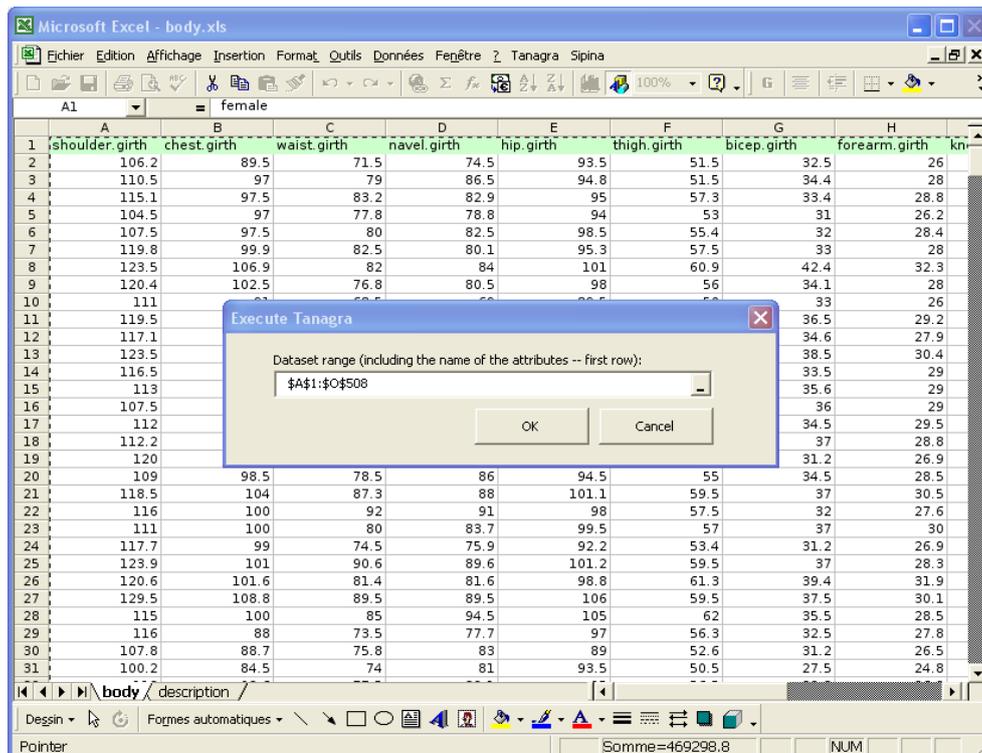
### Créer un diagramme

Dans un premier temps, nous réalisons une ACP « normale ». Pour ce faire, nous partons du fichier ouvert dans le tableur EXCEL, nous sélectionnons la plage de données et nous activons le menu TANAGRA / EXECUTE TANAGRA<sup>1</sup>.

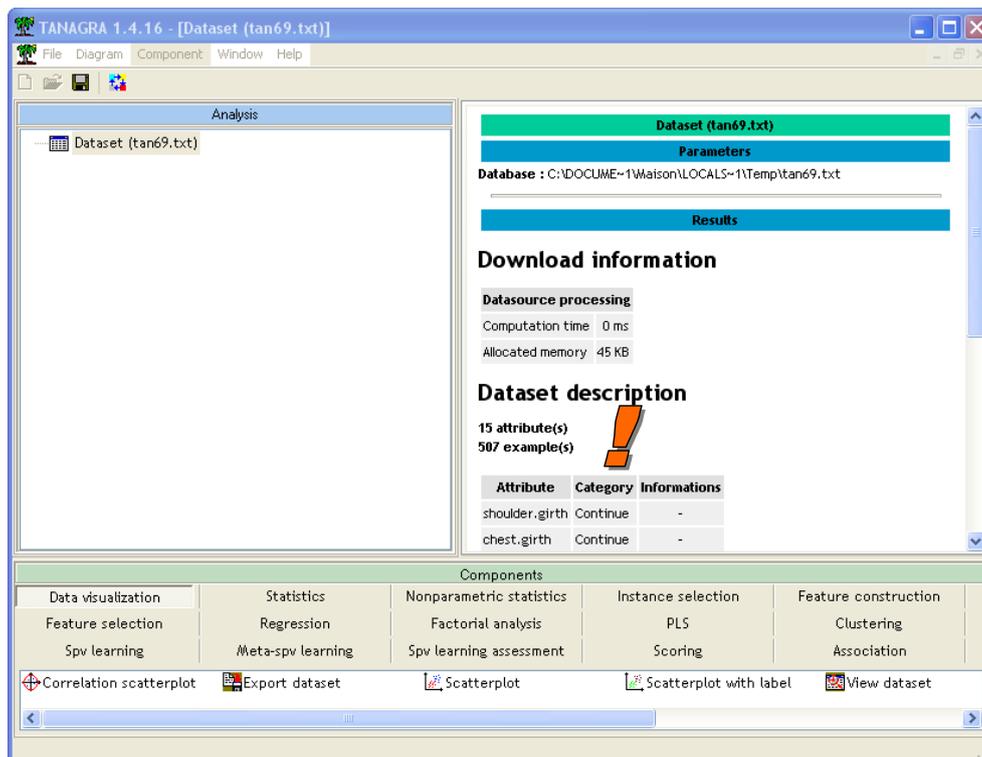
Une boîte de dialogue apparaît. Elle indique les coordonnées de la plage de cellules sélectionnée. Nous validons si la sélection est correcte.

---

<sup>1</sup> Ce nouveau menu a été installé dans le tableur EXCEL à l'aide de la macro complémentaire TANAGRA.XLA qui accompagne la distribution TANAGRA. Elle est copiée automatiquement sur le disque lors de l'installation. Voir <http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html> pour la procédure d'installation et d'activation sous Excel 2007 et 2010 ; et <http://tutoriels-data-mining.blogspot.fr/2008/03/importation-fichier-xls-excel-macro.html> pour les versions antérieures.

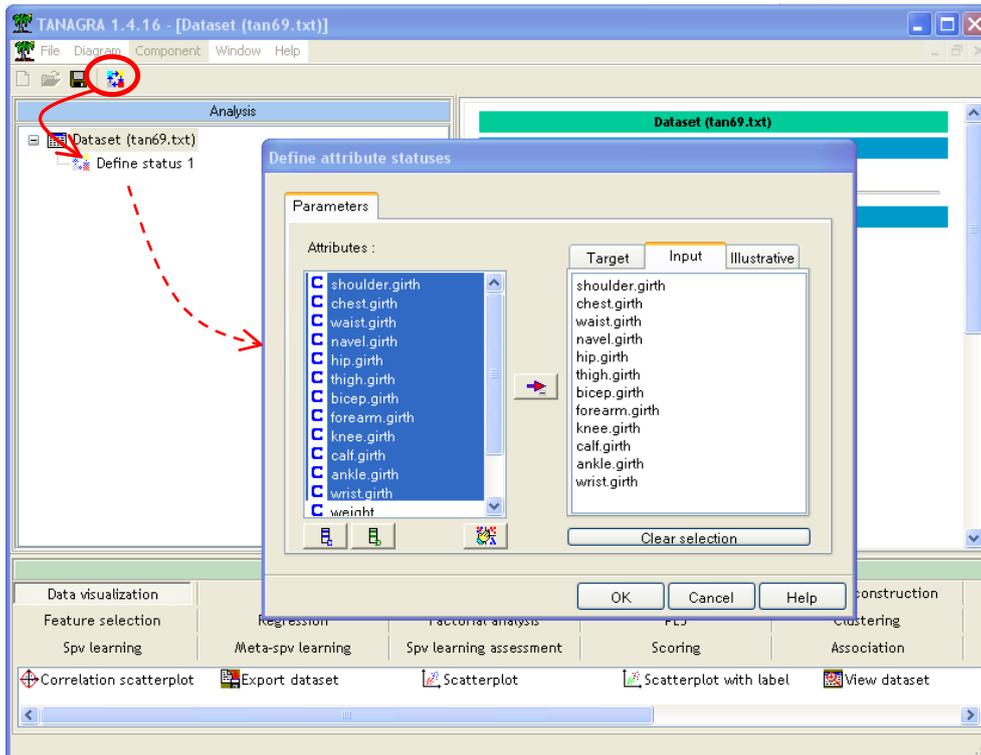


TANAGRA est automatiquement démarré. Un nouveau diagramme est créé avec les données de l'étude. Nous vérifions que nous disposons bien de 15 variables et 507 observations.

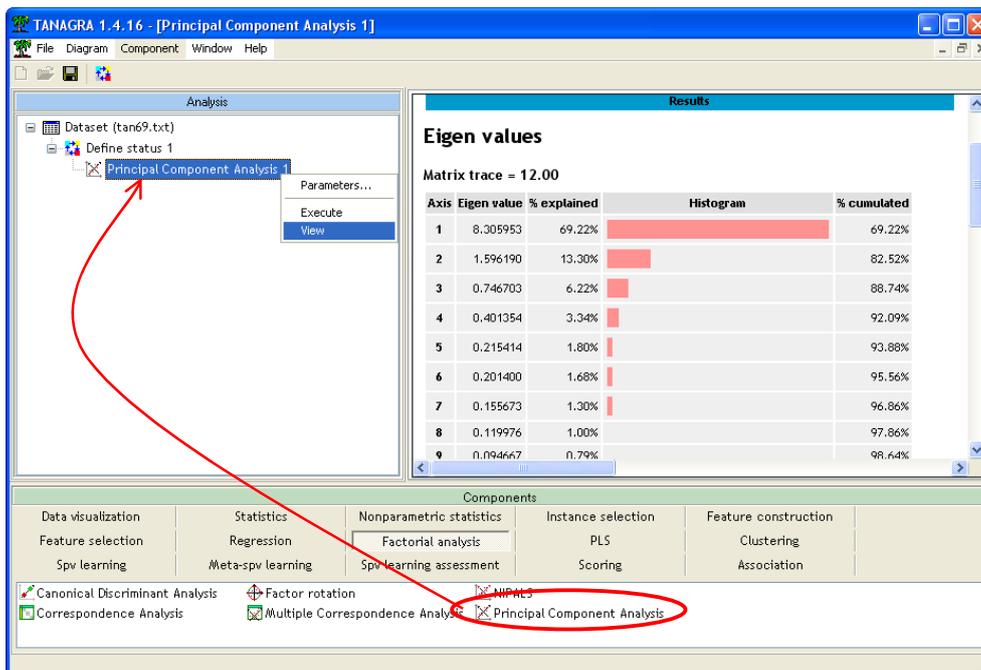


### Définir une analyse

Le composant DEFINE STATUS nous permet de choisir les variables de l'étude, nous l'insérons dans notre diagramme en utilisant le raccourci dans la barre d'outils. Nous plaçons en INPUT les variables SHOULDER.GIRTH jusqu'à WRIST.GIRTH puis nous validons.

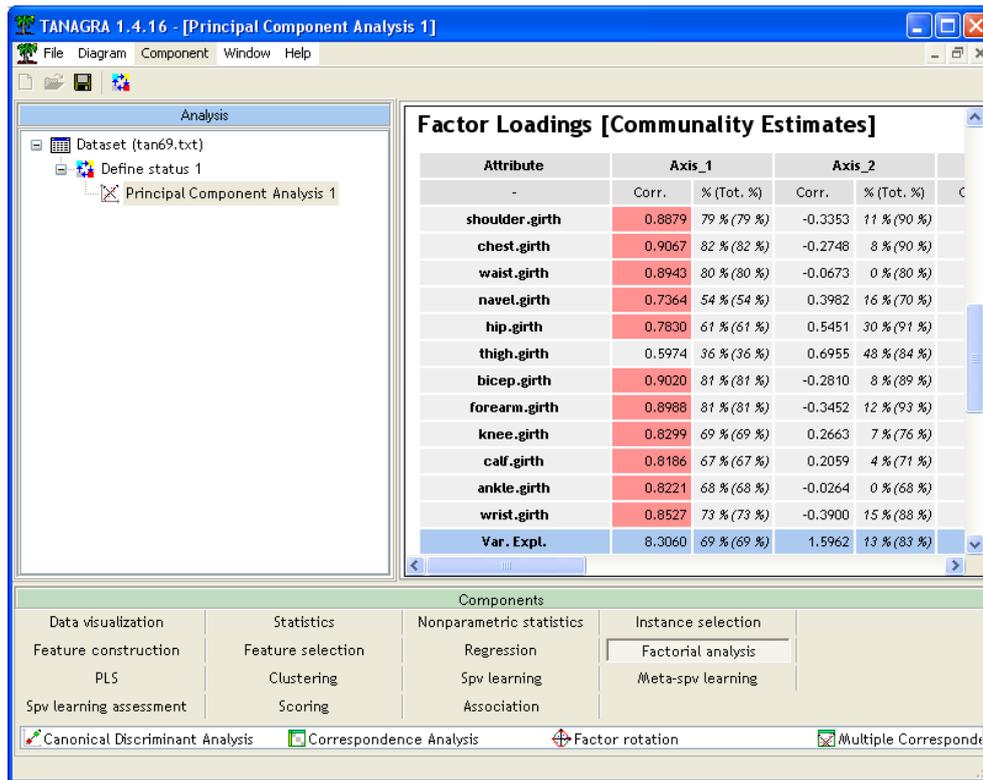


Nous insérons dans le diagramme le composant PRINCIPAL COMPONENT ANALYSIS situé dans l'onglet FACTORIAL ANALYSIS. Nous activons le menu VIEW pour accéder aux résultats.

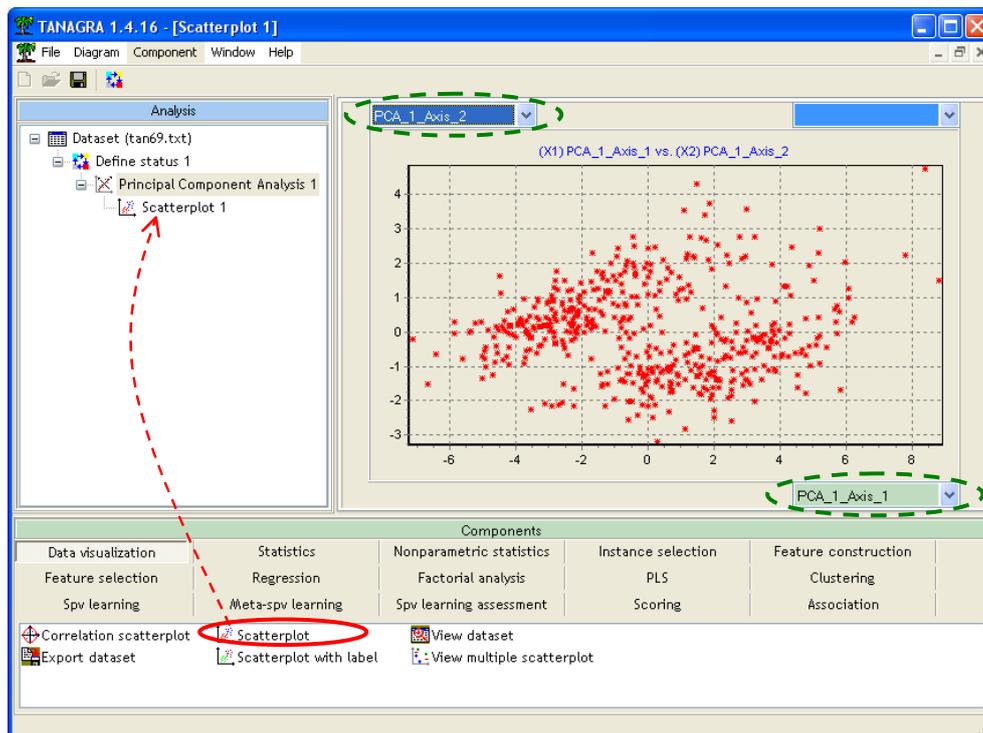


Voilà une belle ACP, les deux premiers axes résument 82,52% des informations. En consultant le tableau des corrélations, nous constatons que toutes les variables, exceptée THIGH (les cuisses), sont fortement corrélées (les corrélations supérieures à 0,7 sont surlignées en rouge) au premier axe. Quoi de plus normal en effet : les

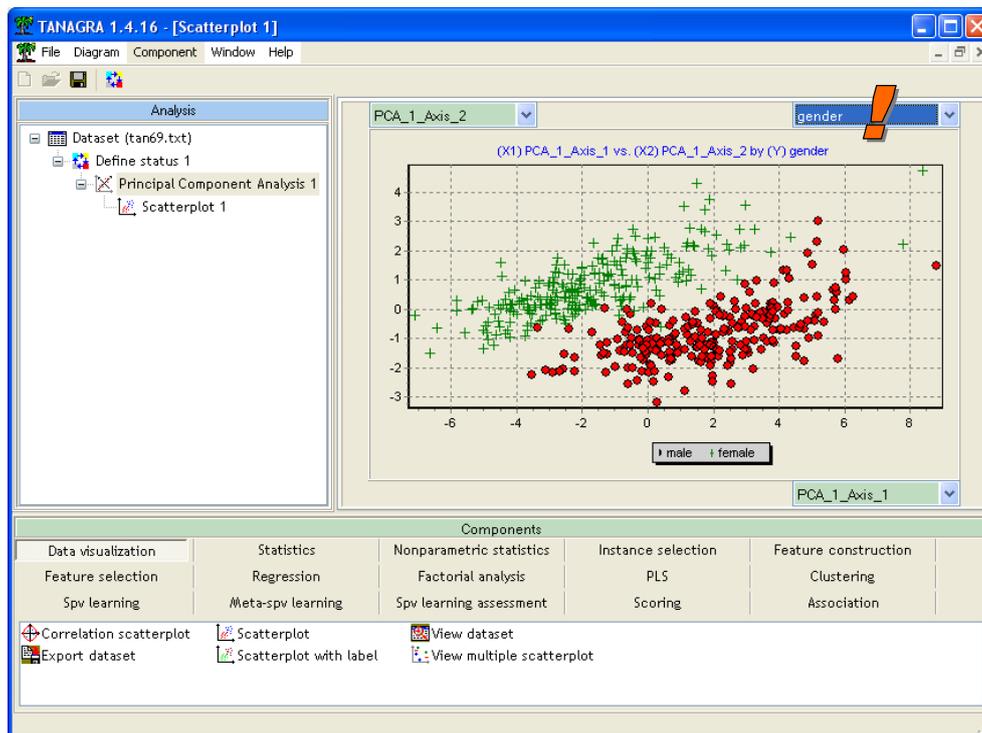
personnes avec des poignets plus épais, ont également des coudes plus épais, etc. Tout est pour le mieux apparemment.



Pour mieux apprécier les positions relatives des individus dans le premier plan factoriel, nous insérons le composant de visualisation SCATTERPLOT (onglet DATA VISUALIZATION). Nous mettons en abscisse la variable représentant le premier axe, calculée à l'aide de l'ACP, et en ordonnée le second axe. Nous obtenons le nuage de points suivant.



Soudain, un grand doute nous assaille. Il nous semble percevoir deux nuages de points distincts dans le plan factoriel. Y aurait-il deux situations ou groupes différents dans nos données ? Parmi les variables disponibles, nous voulons vérifier l'effet de la variable GENDER qui peut fausser nos résultats. Nous allons colorer les points selon cette variable, nous la sélectionnons comme variable illustrative dans ce composant de visualisation.



Nous constatons que ces deux groupes reposent essentiellement sur la différenciation « homme - femme ». Nous observons également que la variabilité sur le premier axe est également due à cette différenciation, les centres de gravité des deux nuages sont bien différenciés sur ce premier axe. A posteriori, ce résultat semble évident. Les hommes ont en moyenne de plus grande taille que les femmes, ils ont des chevilles (*surtout les chevilles !*), les poignets, les genoux, etc. plus épais. Les corrélations traduisent cela. La vraie question est : « est-ce qu'elles ne traduisent que cela ? »

## 4. Travailler sur les corrélations partielles

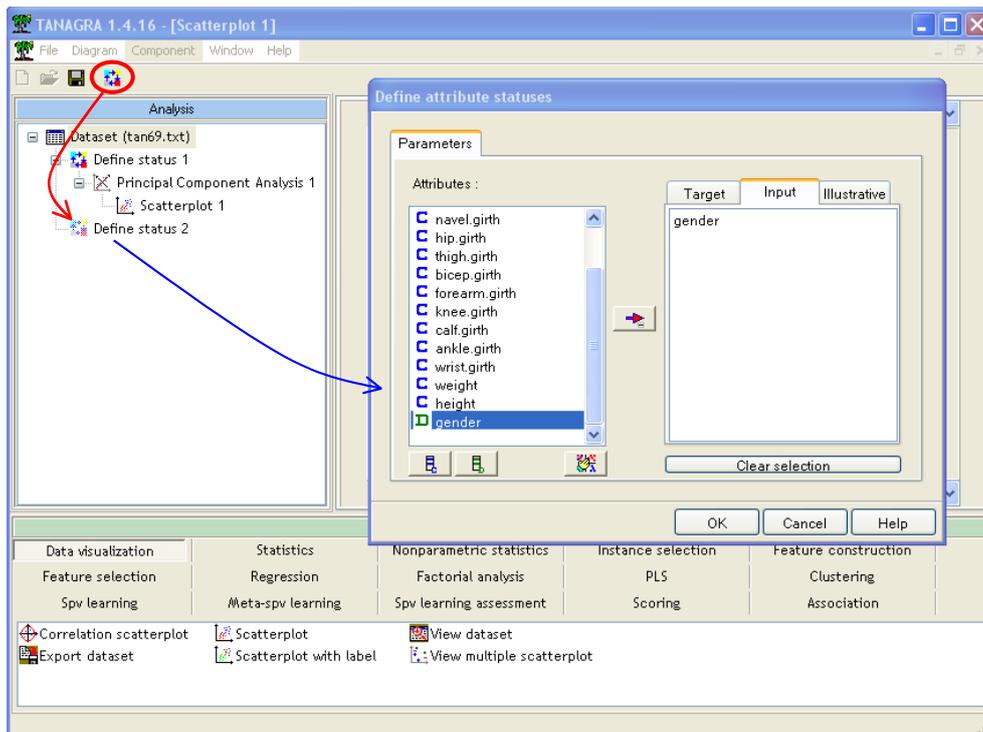
Il nous faut donc éliminer la différenciation selon le sexe dans notre étude. Une approche triviale serait de scinder le fichier en deux parties : d'un côté les hommes, de l'autre les femmes. Nous mènerions alors en parallèle les deux études, puis nous confrontons les résultats pour en extraire une synthèse.

Une autre approche possible serait de soustraire des variables de l'étude l'effet induit par le facteur confondant. Dans notre cas, il s'agirait de ramener les individus à un individu de référence qui peut être l'homme ou la femme, qu'importe, le plus important est de pouvoir les rendre comparable.

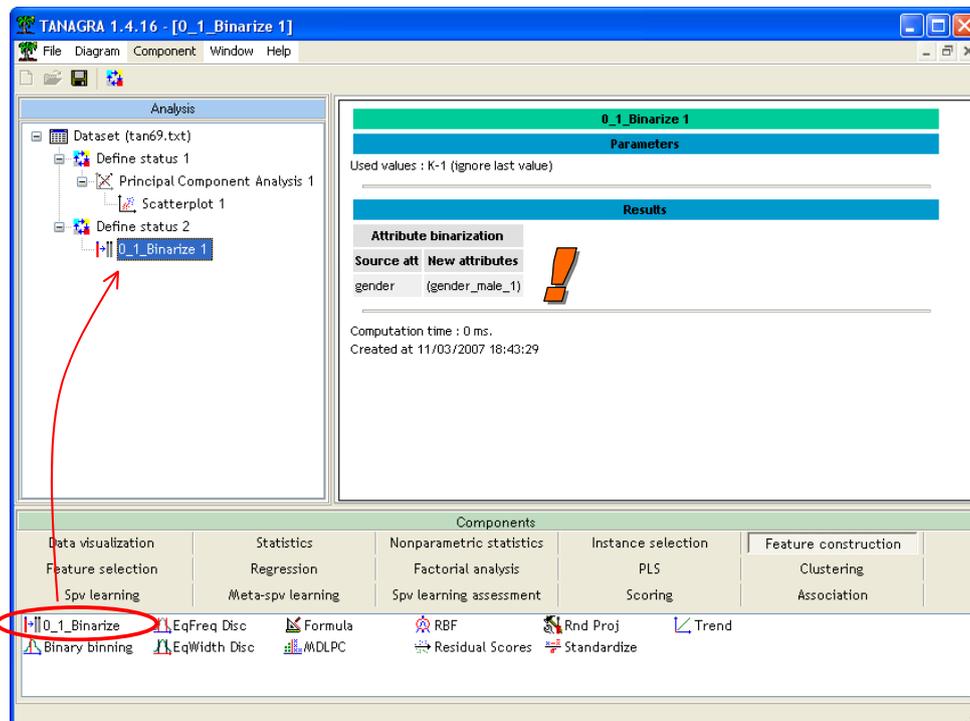
### Élimination de l'effet SEXE dans les variables

**Transformation de la variable GENDER.** Nous devons tout d'abord transformer la variable GENDER, catégorielle, en variable 0/1. Pour ce faire, nous sélectionnons le premier composant, le composant des données, dans le diagramme, puis nous insérons le composant DEFINE STATUS à l'aide du raccourci dans la barre d'outils.

Nous plaçons en INPUT la variable GENDER.

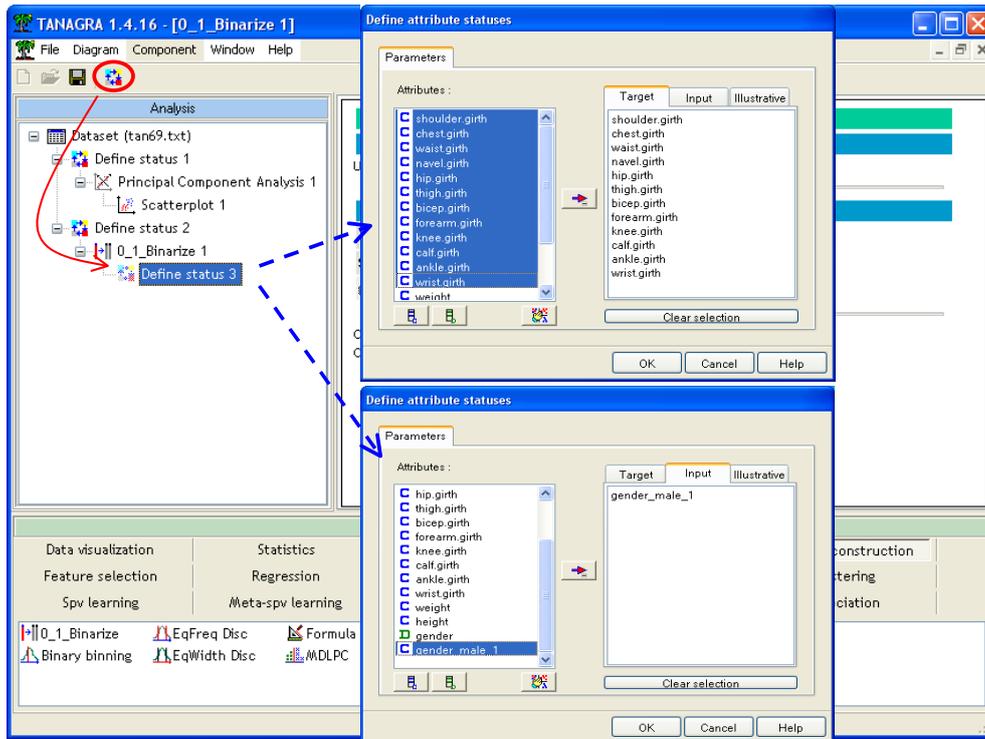


Nous insérons à la suite le composant o\_1\_BINARIZE (onglet FEATURE CONSTRUCTION). Nous actions le menu VIEW.

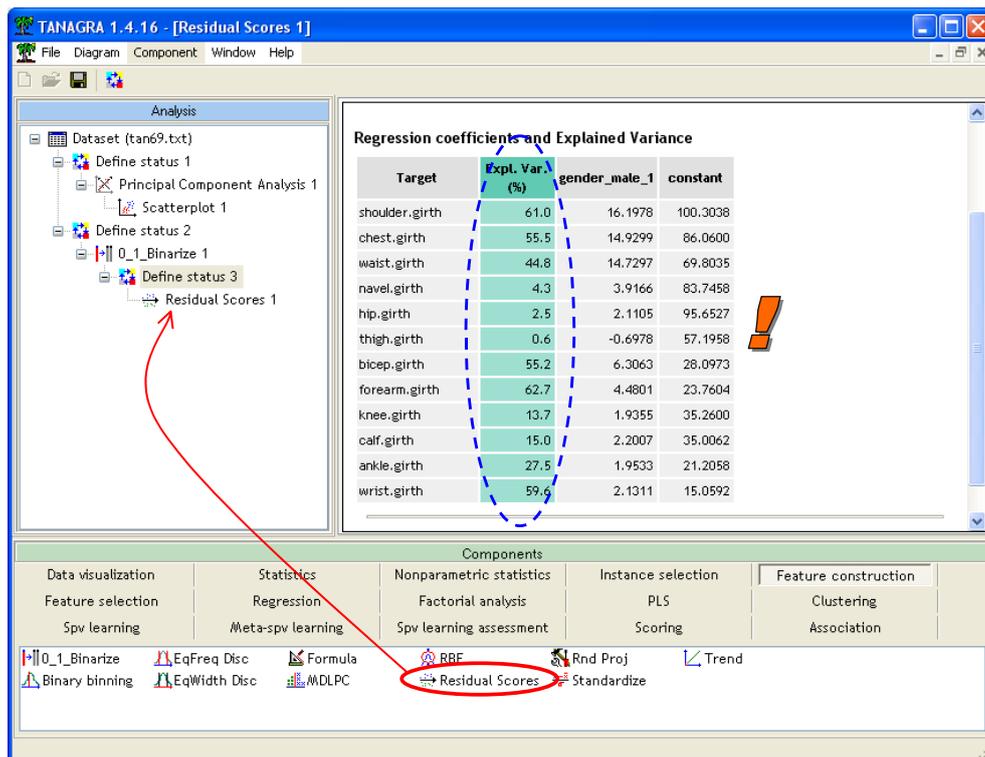


TANAGRA nous indique que la variable GENDER a été recodée en GENDER\_MALE\_1, c.-à-d. que les hommes sont codés 1, les femmes 0.

**Composant RESIDUAL SCORES.** Il nous faut maintenant éliminer des variables de notre étude l'effet induit par le genre. Nous introduisons de nouveau le composant DEFINE STATUS dans notre diagramme, puis nous plaçons en TARGET les variables SHOULDER.GIRTH à WRIST.GIRTH, et en INPUT la variable GENDER\_MALE\_1.



Nous insérons ensuite le composant RESIDUAL SCORES (onglet FEATURE CONSTRUCTION) dans le diagramme. Son rôle est simple, il calcule le résidu de la régression de chaque variable cible sur les variables INPUT. Le composant indique les coefficients de la régression, il indique surtout la proportion de variabilité expliquée par la régression. Cela nous permet de juger de l'importance de l'effet induit par les facteurs confondants.



Les résultats sont édifiants. Dans un cas sur deux, la différenciation « homme-femme » explique plus de la moitié de la variabilité des variables. D'autres en revanche sont très peu influencées par cette différenciation, la circonférence au niveau des cuisses (THIGH), du nombril (NAVEL) et des hanches (HIP)<sup>2</sup>.

### Analyse en composantes principales sur les résidus

Ayant éliminé l'effet GENDER dans nos variables, nous pouvons maintenant lancer l'ACP. Le plus simple est de copier par glisser-déposer la séquence DEFINE STATUS 1 – PRINCIPAL COMPONENT ANALYSIS 1 – SCATTERPLOT 1 après le composant RESIDUAL SCORES 1.

The screenshot shows the TANAGRA 1.4.16 interface. The 'Analysis' tree on the left highlights 'Define status 1' with a red circle and a dashed red arrow pointing to 'Residual Scores 1'. The right panel displays the 'Regression coefficients and Explained Variance' table.

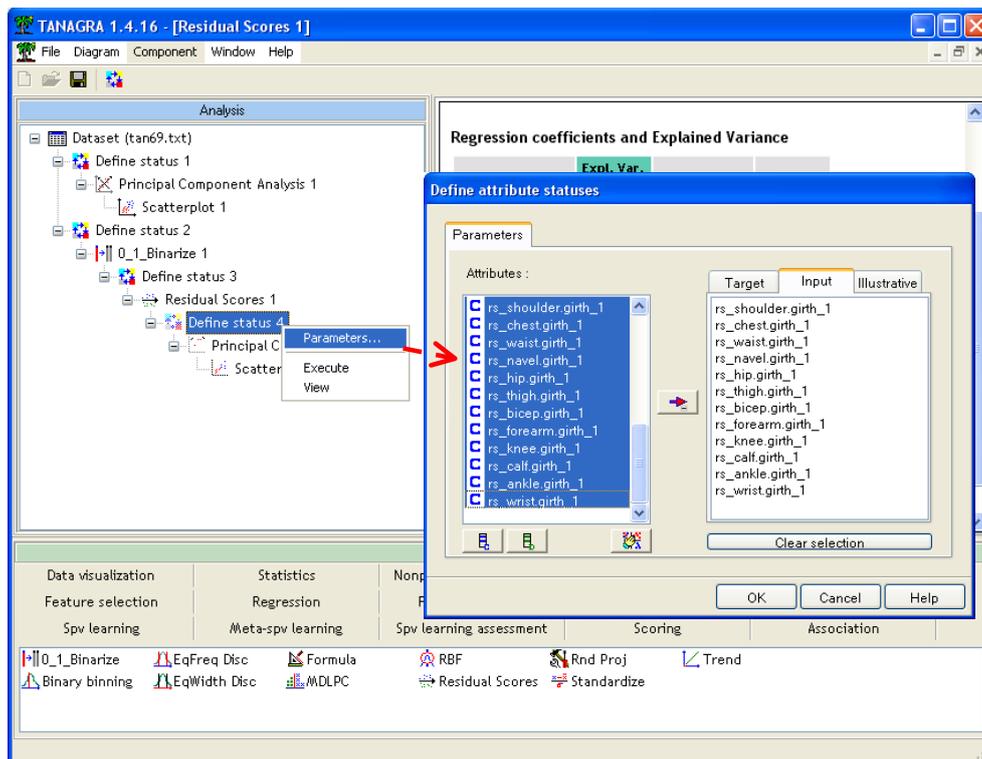
Target	Expl. Var. (%)	gender_male_1	constant
shoulder.girth	61.0	16.1978	100.3038
chest.girth	55.5	14.9299	86.0600
waist.girth	44.8	14.7297	69.8035
navel.girth	4.3	3.9166	83.7458
hip.girth	2.5	2.1105	95.6527
thigh.girth	0.6	-0.6978	57.1958
bicep.girth	55.2	6.3063	28.0973
forearm.girth	62.7	4.4801	23.7604
knee.girth	13.7	1.9355	35.2600
calf.girth	15.0	2.2007	35.0062
ankle.girth	27.5	1.9533	21.2058
wrist.girth	59.6	2.1311	15.0592

The bottom panel shows various components like Data visualization, Statistics, Nonparametric statistics, Instance selection, and Feature construction, along with a toolbar containing icons for operations like Binarize, Formula, RBF, Rnd Proj, Trend, Binary binning, EqWidth Disc, MDLPC, Residual Scores, and Standardize.

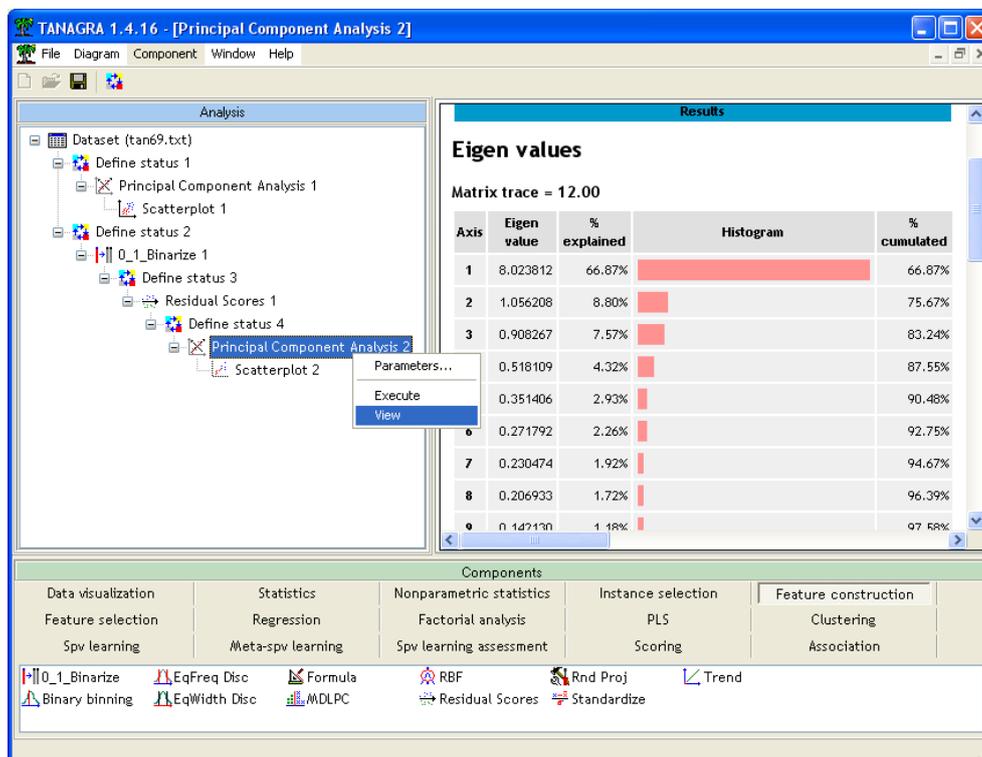
Nous paramétrons le composant DEFINE STATUS 4 (menu PARAMETERS) de manière à placer en INPUT les variables résiduelles RS\_SHOUDLER.GIRTH\_1 à RS\_WRIST.GIRTH\_1.

**Ce faisant, nous sommes en train de définir une ACP à partir de la matrice des corrélations partielles.** En effet, nous avons retiré de nos variables les effets de variation due à GENDER, la corrélation entre deux variables résiduelles est bien une corrélation partielle.

<sup>2</sup> A ce propos, on note que contrairement à ce qu'on pourrait penser, dans ce fichier en tous cas, en moyenne les hommes et les femmes ont des tours de hanches assez identiques (#95 cm). La vraie différence se fait au niveau du tour de taille (WAIST) où, en moyenne toujours, les hommes font 14 cm de plus.

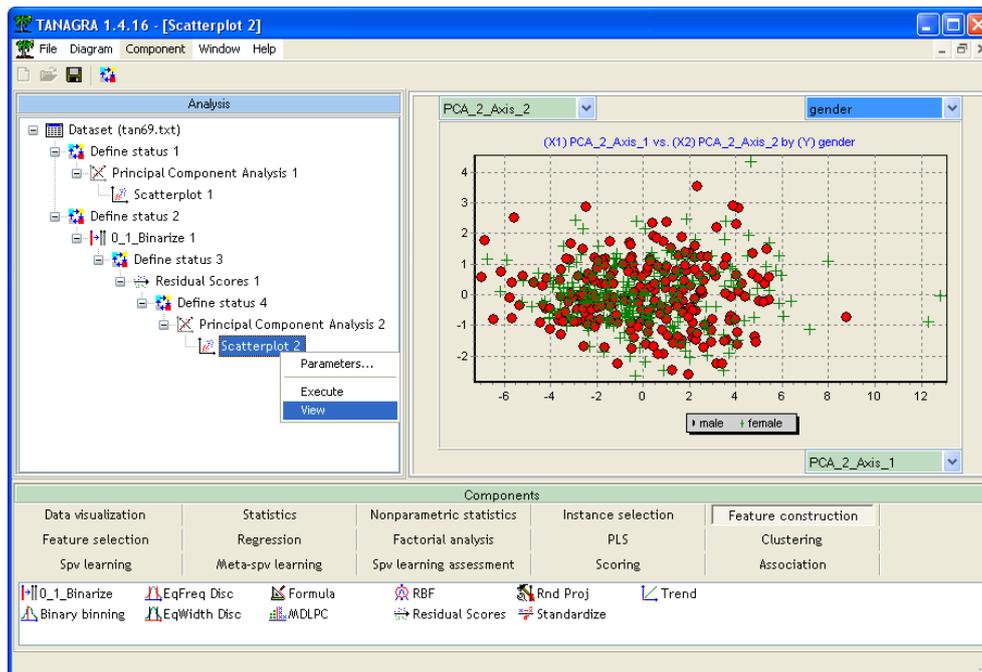


Nous activons le menu VIEW de l'ACP pour accéder aux résultats.



Les deux premiers axes factoriels expliquent 75,67% de la variabilité, nous avons un superbe « effet-taille », toutes les variables sont fortement corrélées avec le premier axe. Nous y reviendrons plus loin.

Pour évaluer la persistance de l'effet GENDER, le plus simple est de visualiser les individus dans le premier plan factoriel, et de vérifier si les groupes des hommes et des femmes sont toujours distincts ou pas.



Il n'y a plus de différenciation entre les hommes et les femmes.

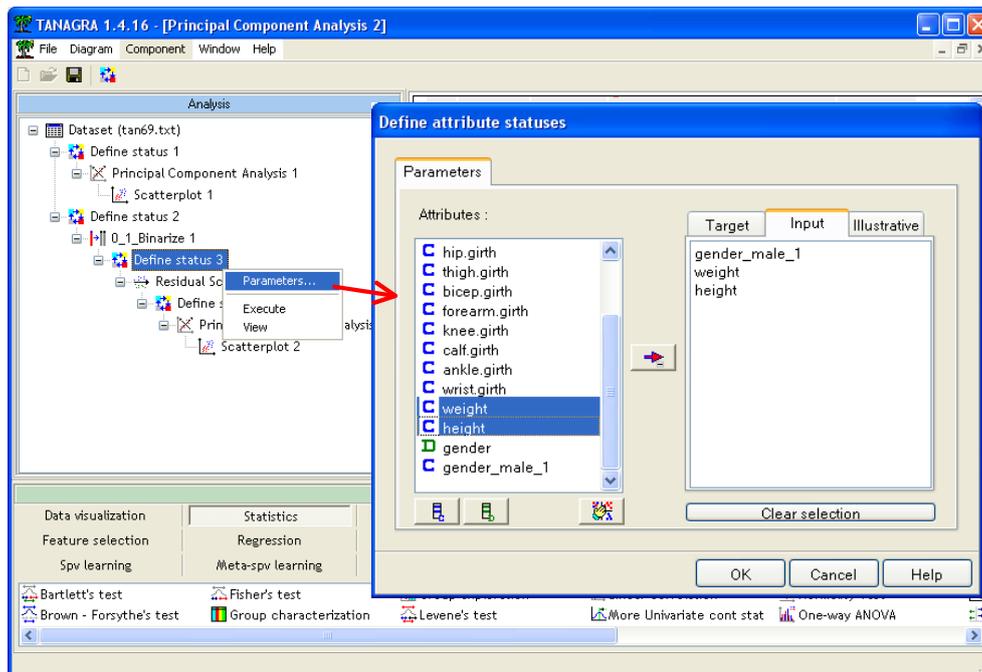
## 5. Élimination de l'effet taille

Malgré ce premier résultat, il reste une certaine insatisfaction. On se rend compte que toutes les variables sont fortement corrélées avec le premier axe factoriel, ce qu'on appelle communément « l'effet taille ». Dans notre cas, il s'agit bien de cela puisque l'on se rend compte que nous avons dans notre fichier des individus avec des tailles et des poids différents. Quoi de plus normal finalement que les individus de poids et de taille élevés aient des poignets et des chevilles plus épais que les autres ? De nouveau ici, nous n'utilisons pas un référentiel commun pour comparer les individus.

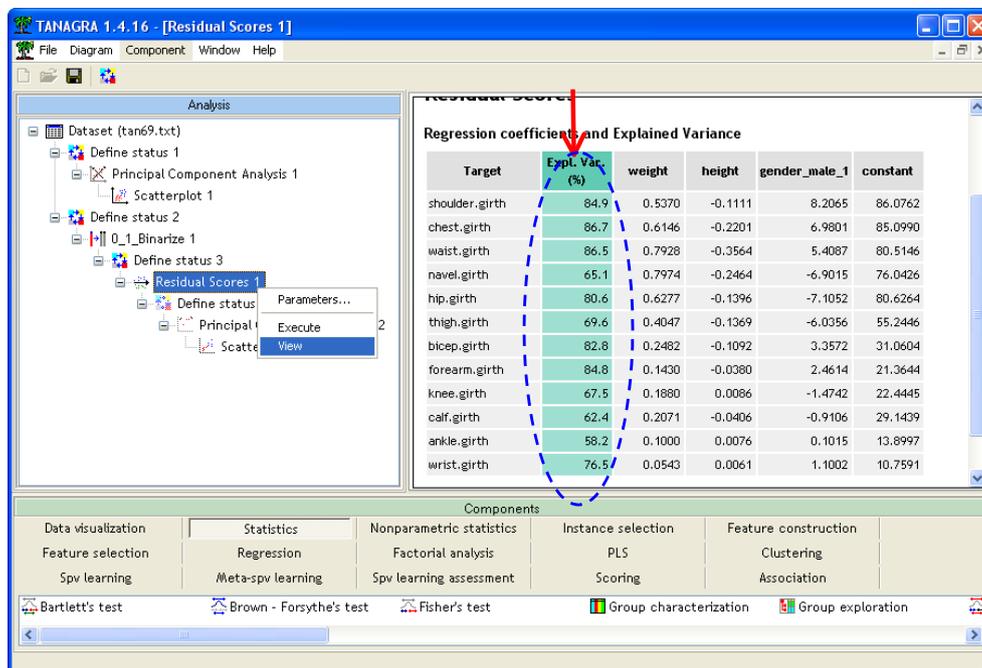
Tout dépend de l'étude que l'on veut mener. S'il nous satisfait et correspond à ce que nous cherchions, nous pouvons nous en tenir à ce résultat : les circonférences des différentes parties du corps humain sont positivement corrélées tout simplement parce qu'elles traduisent l'influence sous-jacente de la taille et du poids des individus.

Si, en revanche, il ne nous convient pas et que nous voulons ramener tous les individus dans un référentiel commun, sexe, poids et taille identiques, pour étudier les rapports de taille entre les différentes parties du corps, comment devons-nous procéder ?

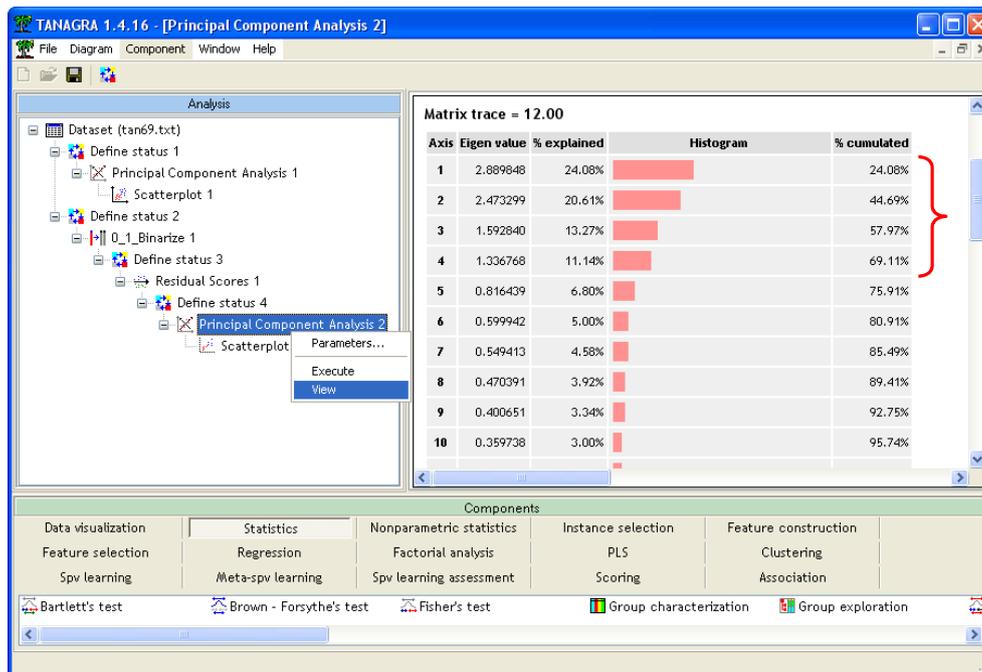
**Introduire le poids et la taille dans le calcul des résidus.** Nous revenons sur le composant DEFINES STATUS 3, et nous le paramétrons (menu PARAMETERS) de manière à introduire en INPUT, en plus de la variable GENDER\_MALE\_1, les variables WEIGHT et HEIGHT.



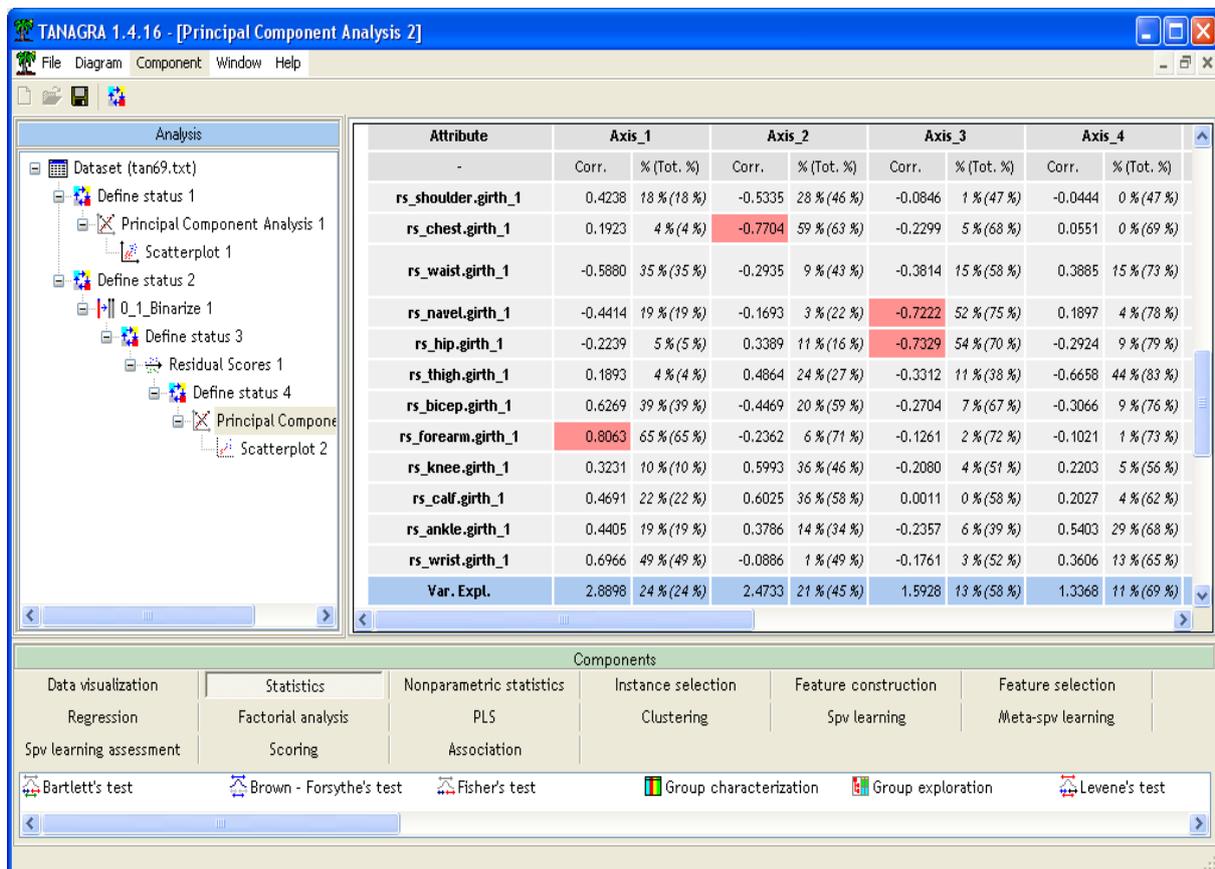
Nous activons le menu VIEW du composant RESIDUAL SCORES 1, nous observons que dans plusieurs cas, ce triplet de variables explique près de 80% de la variabilité. Ce qui confirme largement notre intuition ci-dessus.



Il ne nous reste plus qu'à relancer l'ACP (menu VIEW du composant PRINCIPAL COMPONENT ANALYSIS 2). Nous observons que l'information restituée est plus dispersée : les 4 premiers axes restituent 69,11% de l'information disponible.



L'étude des corrélations montre qu'en réalité les relations entre les dimensions des différentes zones du corps, **une fois que nous nous sommes ramenés à un individu « standard » (sexe, poids et taille égal)** sont un peu plus complexes que ce que nous pouvions imaginer dans un premier temps.

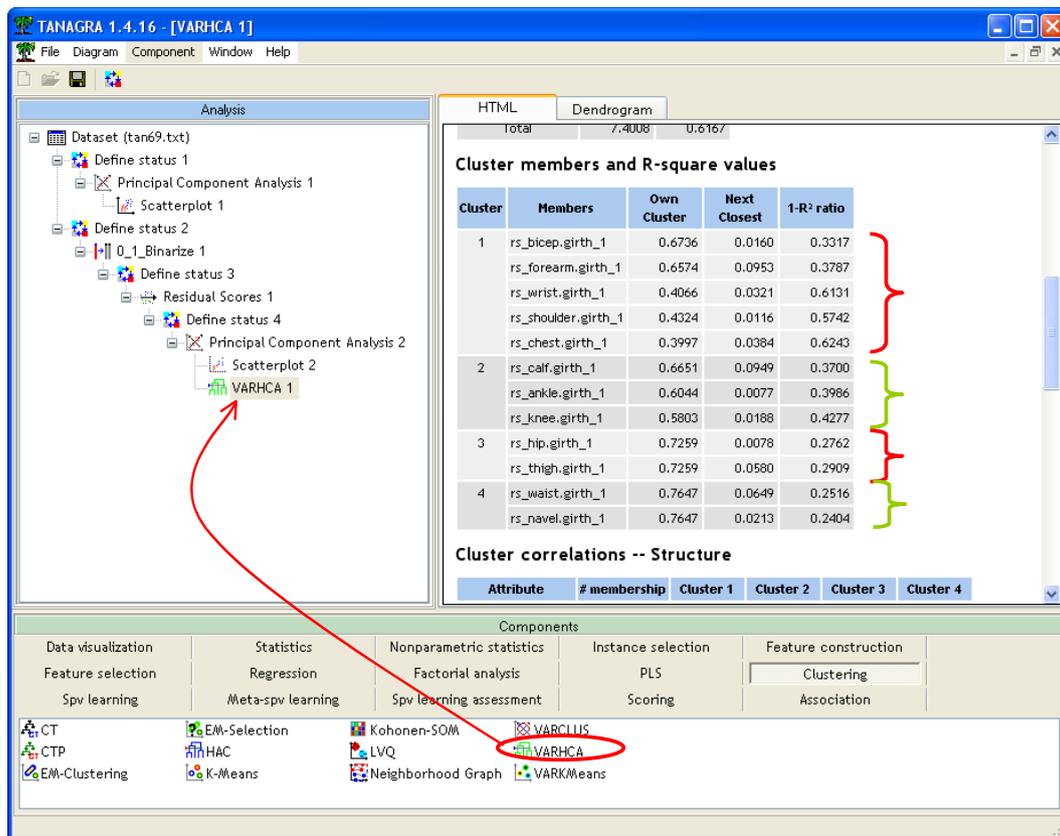


### Classification (clustering) de variables

Étudier simultanément des corrélations sur 4 axes peut se révéler rapidement fastidieux. Nous pouvons introduire à ce stade un autre composant qui permet de regrouper les variables selon leurs corrélations : les

composants de classification de variables. Nous pouvons d'ailleurs les considérer comme une variante de l'ACP où la contrainte d'orthogonalité stricte entre les facteurs a été relâchée.

Nous insérons donc le composant VARHCA<sup>3</sup> à la suite du composant DEFINE STATUS<sub>4</sub> dans le diagramme. Son exécution indique qu'il y a en réalité 4 groupes de variables corrélées selon les différentes zones du corps.



Les groupes de variables sont détaillés dans le tableau CLUSTERS MEMBERS. Nous retrouvons finalement 4 grandes zones du corps humain que nous résumons dans le tableau ci-dessous.

Variables	Zones
rs_bicep.girth_1 (biceps) rs_forearm.girth_1 (avant-bras) rs_wrist.girth_1 (poignet) rs_shoulder.girth_1 (épaule) rs_chest.girth_1 (poitrine)	Excepté la poitrine, nous retrouvons les membres supérieurs.
rs_calf.girth_1 (mollet) rs_ankle.girth_1 (cheville) rs_knee.girth_1 (genou)	Membres inférieurs.
rs_hip.girth_1 (hanche) rs_thigh.girth_1 (cuisse)	Sous le nombril et le genou

<sup>3</sup> L'utilisation des composants VARIABLE CLUSTERING (VARCLUS) est détaillée dans le didacticiel [http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr\\_Tanagra\\_VarClus.pdf](http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_VarClus.pdf) disponible en ligne.

rs\_waist.girth\_1 (taille)

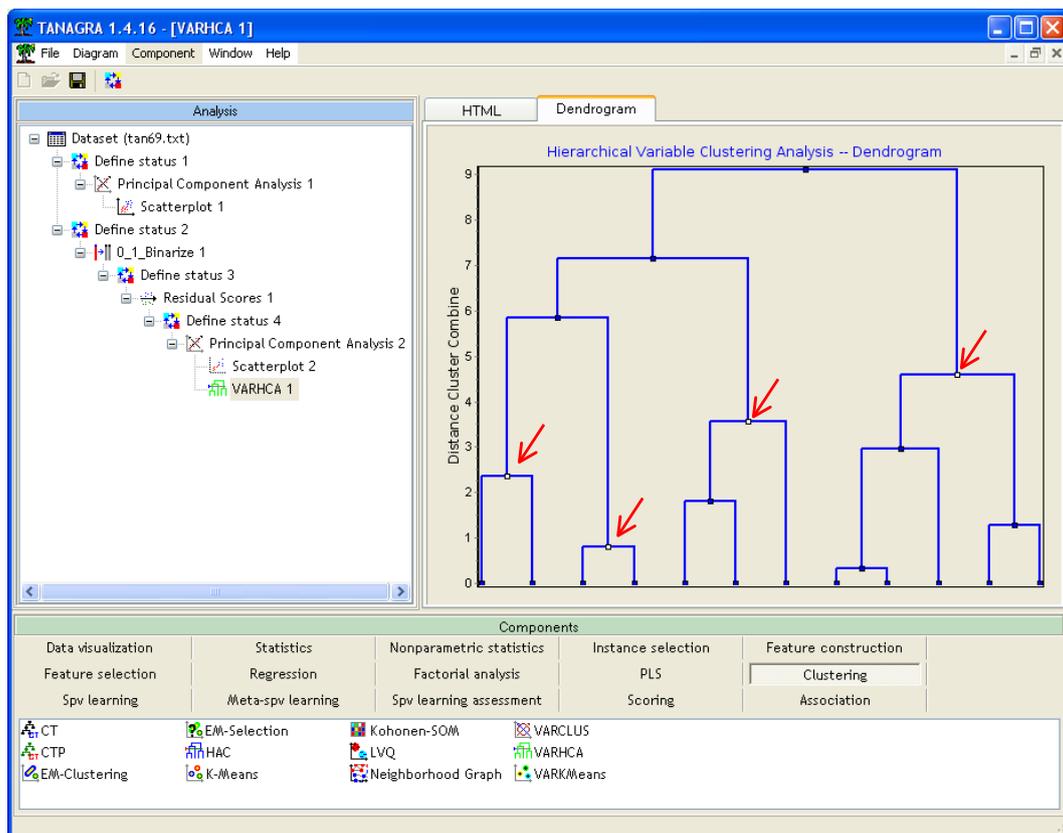
rs\_navel.girth\_1 (nombril)

Entre la poitrine et les hanches

Si l'on conçoit aisément que les circonférences du poignet et de l'avant-bras soient liées, on constate avec surprise -- pour moi c'en est une en tous les cas -- qu'à sexe, poids et taille identiques, il n'y a pas de liaison entre la circonférence du poignet et la circonférence de la cheville chez les individus<sup>4</sup>.

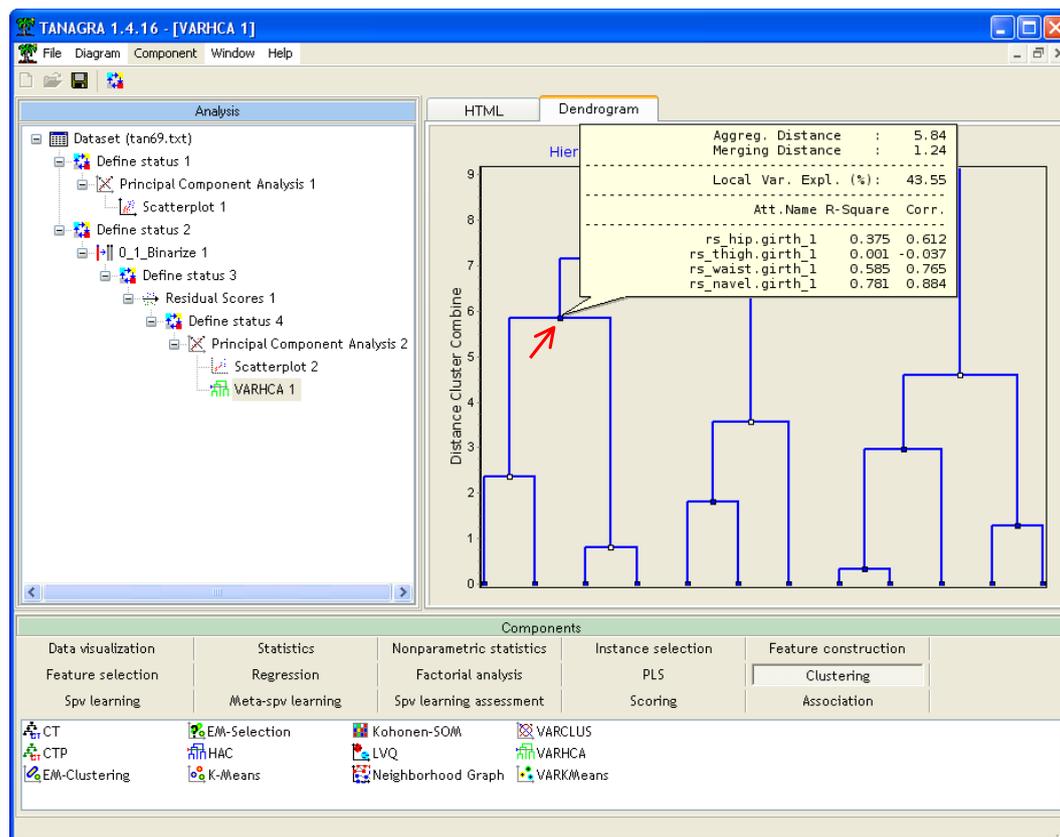
Cette subdivision en 4 groupes a été automatiquement déterminée par l'algorithme. Si elle ne cadre pas avec les connaissances du domaine, nous avons la possibilité de l'affiner en étudiant le dendrogramme associée au processus de typologie.

Les groupes détectés par la méthode sont en blanc dans le dendrogramme.



Nous constatons par exemple que si nous souhaitons passer à une partition en trois groupes, nous serons emmenés à fusionner les 3<sup>ème</sup> et 4<sup>ème</sup> groupes (hanche, cuisse, taille et nombril) situés à gauche dans le dendrogramme et correspondant à des zones adjacentes du corps. Nous avons accès à la liste des variables constituant le groupe fusionné en cliquant sur le sommet.

<sup>4</sup> Ici s'arrête le travail du statisticien ou du data miner et commence le travail de l'expert métier !!!



## 6. Conclusion

Les facteurs confondants font partie des plus gros écueils de l'analyse exploratoire des données. Ne pas les détecter nous expose à de gros problèmes d'interprétation des résultats.

Il serait illusoire de prétendre mettre en oeuvre des techniques statistiques pour détecter automatiquement les variables incriminées, si tant est qu'elles soient présentes dans le fichier. Dans la plupart des cas, nous devons nous en remettre à l'expertise du domaine pour les circonscrire. En revanche, une fois identifiées, la majorité des logiciels proposent un arsenal d'outils efficaces pour les traiter convenablement. C'est ce que nous avons essayé de montrer dans ce didacticiel.

## 7. Mise à jour 16/06/2012 – Analyse avec SAS

### Données

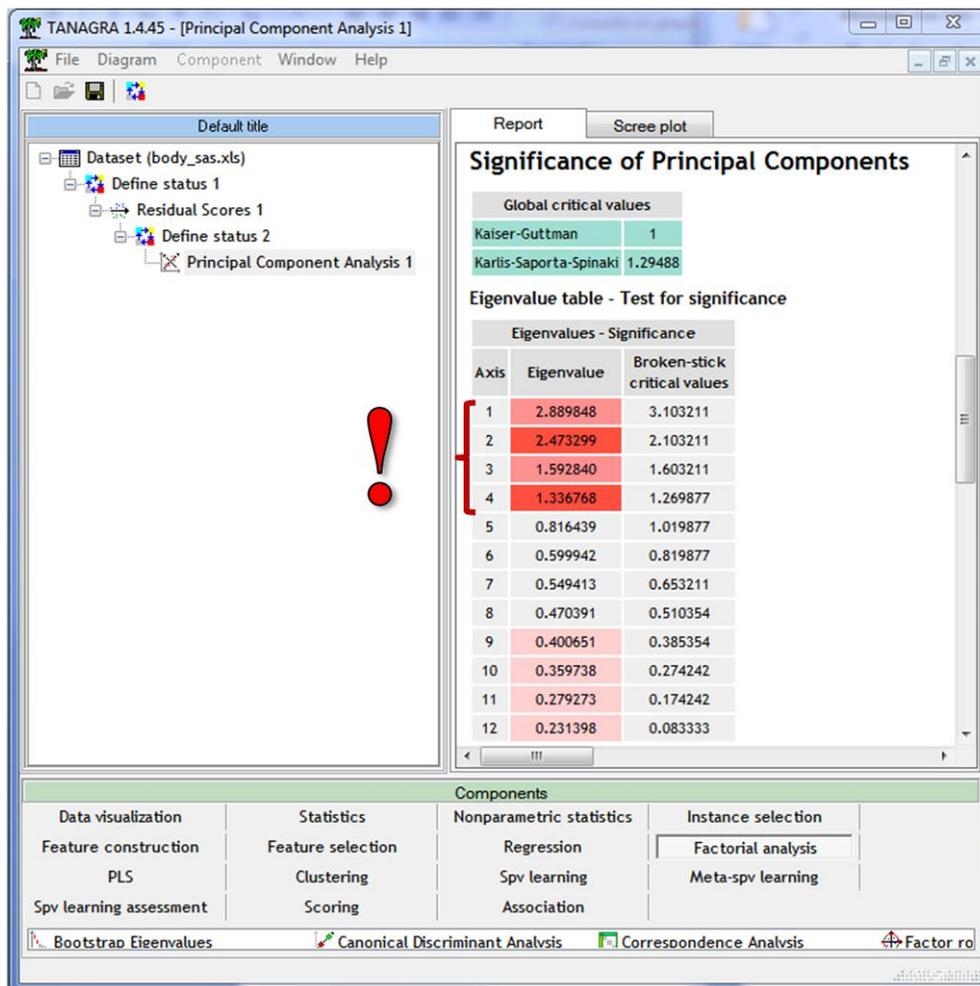
Pour nous faciliter la vie, la variable GENRE a été recodée en une indicatrice 1/0 dans le tableur Excel. Le nom du nouveau fichier est « body\_sas.xls ».

### Compléter les résultats de Tanagra

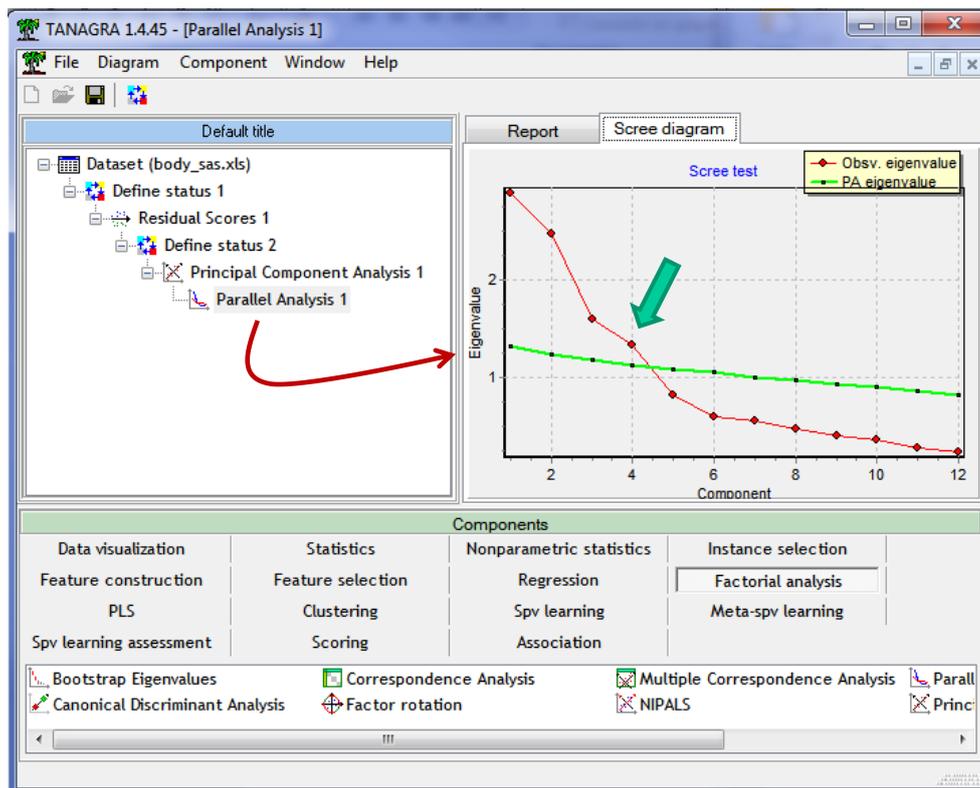
**Détection du nombre de composantes.** Depuis la première version de ce document, Tanagra a évolué. Nous sommes passés de la 1.4.16 (01/03/2007) à la 1.4.45 (12/06/2012). L'analyse en composantes principales comprend maintenant de nouveaux outils<sup>5</sup> pour la détection (*aide à la détection*) du nombre adéquat de facteurs. Nous les avons mis à contribution pour notre ACP sur les corrélations partielles.

Les tests de Kaiser, Karlis-Saporta-Spinaki et Legendre-Legendre convergent pour une analyse en 4 facteurs comme nous pouvons le constater dans la copie d'écran ci-dessous.

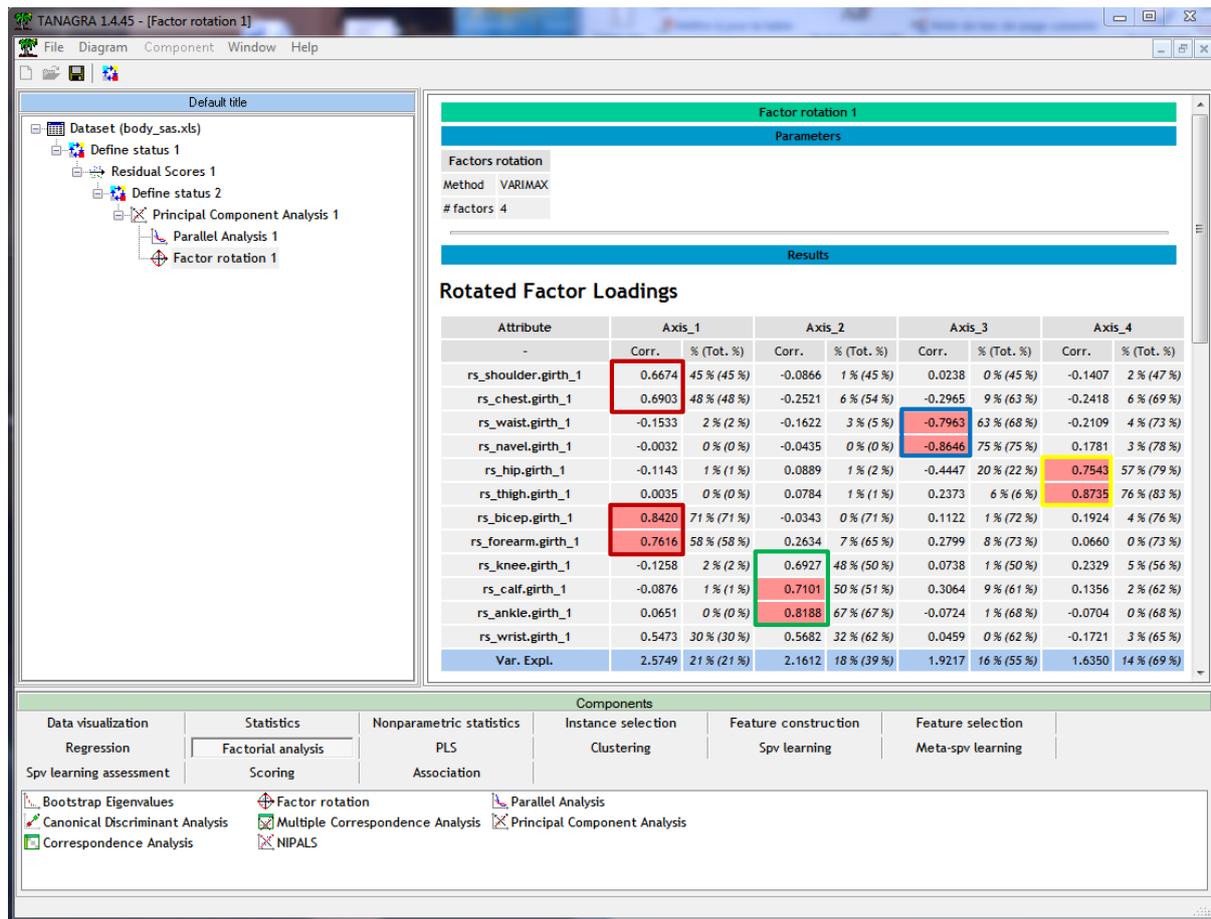
<sup>5</sup> <http://tutoriels-data-mining.blogspot.fr/2012/06/acp-avec-tanagra-nouveaux-outils.html> pour leur utilisation et la lecture des résultats.



L'analyse parallèle accrédite ce résultat.



**Rotation VARIMAX.** Pour « forcer le trait », nous pouvons faire pivoter les axes de manière à ce que les corrélations avec les variables soient plus marquées. Nous spécifions 4 facteurs (menu PARAMETERS) dans la rotation VARIMAX<sup>6</sup>.



Nous retrouvons (à quelques détails près, WRIST est à cheval sur 2 axes) les résultats de la classification de variables (page 12 et suivantes).

### Résultats des traitements sous SAS

Les données ont été importées et intégrées dans SAS. Nous avons utilisé la commande FACTOR. L'option PARTIAL est primordiale dans notre analyse.

```

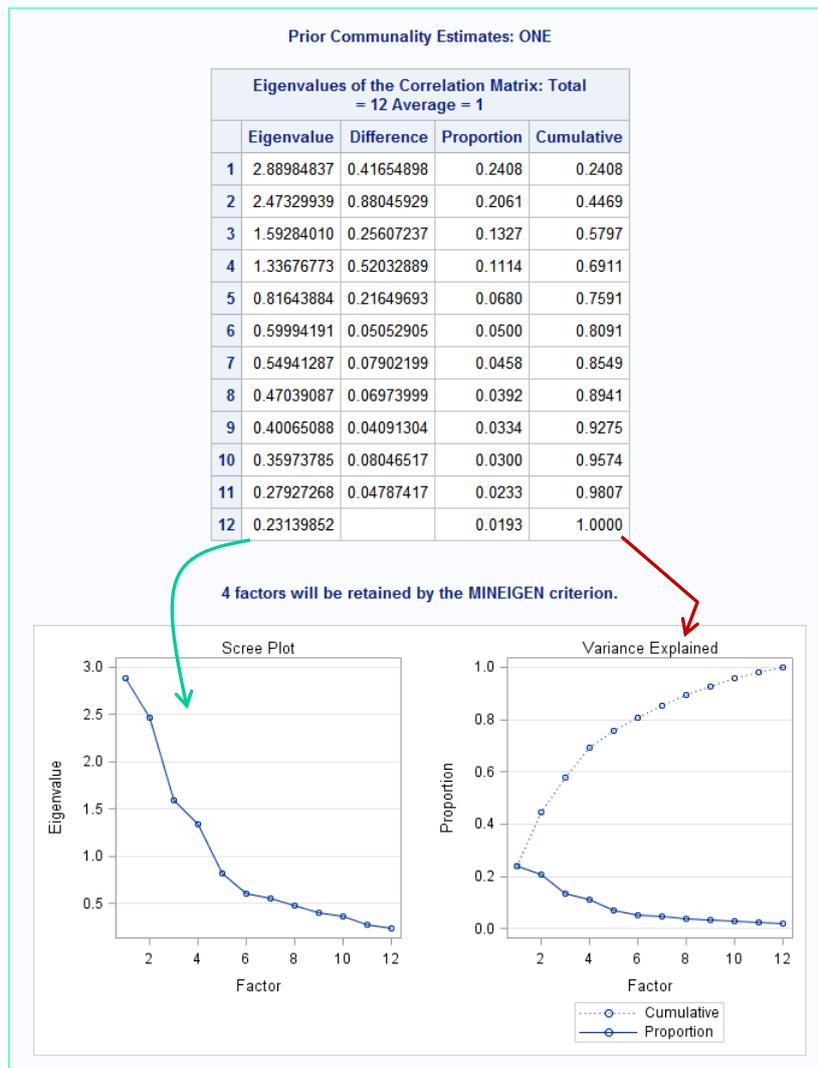
sas_proc_factor.sas
proc factor data = body method = principal corr msa plots=(scree);
var shoulder_girth chest_girth waist_girth navel_girth
hip_girth thigh_girth bicep_girth forearm_girth knee_girth
calf_girth ankle_girth wrist_girth;
partial weight height gender;
run;

```

La séquence des valeurs propres est exactement la même que celle de Tanagra (Section 5). Ce qui confirme l'idée, au cas où on en douterait, que travailler à partir des résidus de la régression des variables actives de l'ACP avec les 3 modératrices (poids, taille et genre) équivaut bien à une analyse en composantes principales sur la matrice des corrélations partielles.

<sup>6</sup> Voir <http://tutoriels-data-mining.blogspot.fr/2008/04/rotation-varimax-en-acp.html>

Voici le graphique « éboulis des valeurs propres » fourni par SAS (cf. page 12 pour Tanagra)



SAS a choisi 4 axes parce que leurs valeurs propres sont supérieures à MINEIGEN qui vaut 1 par défaut.

Nous avons ensuite les corrélations des variables avec les composantes (cf. page 12 pour Tanagra).

Factor Pattern					
		Factor1	Factor2	Factor3	Factor4
<b>shoulder_girth</b>	shoulder#girth	0.42375	0.53350	0.08461	-0.04438
<b>chest_girth</b>	chest#girth	0.19231	0.77044	0.22987	0.05507
<b>waist_girth</b>	waist#girth	-0.58803	0.29350	0.38144	0.38852
<b>navel_girth</b>	navel#girth	-0.44136	0.16934	0.72225	0.18971
<b>hip_girth</b>	hip#girth	-0.22387	-0.33888	0.73290	-0.29239
<b>thigh_girth</b>	thigh#girth	0.18928	-0.48642	0.33122	-0.66584
<b>bicep_girth</b>	bicep#girth	0.62690	0.44689	0.27038	-0.30659
<b>forearm_girth</b>	forearm#girth	0.80626	0.23622	0.12608	-0.10205
<b>knee_girth</b>	knee#girth	0.32313	-0.59929	0.20801	0.22028
<b>calf_girth</b>	calf#girth	0.46907	-0.60249	-0.00108	0.20272
<b>ankle_girth</b>	ankle#girth	0.44053	-0.37858	0.23571	0.54033
<b>wrist_girth</b>	wrist#girth	0.69657	0.08861	0.17615	0.36056

Et les  $\text{COS}^2$  des variables sur les 4 composantes sélectionnées (cf. page 12 pour Tanagra).

Final Commuality Estimates: Total = 8.292756											
shoulder_girth	chest_girth	waist_girth	navel_girth	hip_girth	thigh_girth	bicep_girth	forearm_girth	knee_girth	calf_girth	ankle_girth	wrist_girth
0.47331904	0.68643578	0.72836252	0.78110888	0.78759324	0.82548235	0.75981050	0.73216577	0.55535517	0.62411688	0.68490294	0.65410250

Si nous souhaitons introduire la rotation VARIMAX, voici la commande adéquate.

```
proc factor data = body method = principal corr rotate = varimax n = 4;
var shoulder_girth chest_girth waist_girth navel_girth
hip_girth thigh_girth bicep_girth forearm_girth knee_girth
calf_girth ankle_girth wrist_girth;
partial weight height gender;
run;
```

## Conclusion

Lorsque j'ai écrit la première version de ce tutoriel, mon idée était de montrer qu'il était possible de trouver des solutions pour contrecarrer l'influence exagérée de certaines variables qui induisent un effet taille fort dans l'ACP, masquant ainsi par des évidences les informations intéressantes que recèlent les données. Cela revenait à réaliser une analyse en composantes sur une matrice de corrélations partielles. Je montrais alors pas à pas les étapes, assez ardues je le reconnais, pour obtenir le résultat souhaité.

La nouveauté de ce tutoriel est que je me suis rendu compte que l'approche était implémentée de manière native dans SAS par le biais d'une options de la PROC FACTOR.

## 8. Mise à jour 16/06/2012 – Analyse avec R

### Programmation sous R

Si c'est possible avec SAS, ça l'est sûrement avec R. J'ai un peu (beaucoup) cherché un équivalent de la PROC FACTOR avec l'option PARTIAL pour le logiciel R. J'avoue ne pas avoir trouvé un package répondant au cahier des charges ci-dessus<sup>7</sup>. Je me suis dit qu'il y avait là une opportunité d'écrire un programme qui, somme toute, est relativement simple. Il faut suivre pas à pas les étapes retranscrites dans le diagramme de traitements de Tanagra (Section 5) c.-à-d. réaliser une ACP à partir des résidus des régressions des variables actives sur les modératrices.

Voici le code correspondant :

```
rm(list=ls())

##### BEGIN FUNCTIONS #####

#résidu de la régression d'une variable avec les modératrices
get_one_residual <- function(x, variables, moderators){
  numero <- as.numeric(gsub("\\D","", deparse(substitute(x)), perl=T))
  str.formula <- paste(names(variables)[numero],paste(names(moderators),collapse="+"), sep=" ~ ")
  reg <- lm(formula = as.formula(str.formula),data=cbind(variables,moderators))
  return(reg$residuals)
}

#résidus de toutes les variables par rapport aux modératrices
get_residuals <- function(variables,moderators){
  new.data <- lapply(variables,get_one_residual,variables,moderators)
  return(as.data.frame(new.data))
}
```

<sup>7</sup> Peut être que ça existe mais que je n'ai pas trouvé simplement. J'ai quand même pas mal cherché avant de me résoudre à programmer la méthode. Finalement, le code source est très simple, 3 courtes fonctions suffisent amplement.

```

#fonction : ACP sur la matrice des corrélations partielles
pca_partial <- fonction(variables, moderators){
  new.scores <- get_residuals(variables,moderators)
  pca.object <- princomp(new.scores,cor=T)
  return(pca.object)
}

##### END FUNCTIONS #####

#charger les données
library(xlsx)
body.data <- read.xlsx(file="body_sas.xls",sheetIndex=1,header=T)

#subdivision : variables actives, variables modératrices
body.variables <- subset(body.data, select=shoulder.girth:wrist.girth)
body.moderators <- subset(body.data, select=weight:gender)

#appel de la fonction
body.pca <- pca_partial(body.variables,body.moderators)

#affichage des valeurs propres
print(body.pca$sdev^2)

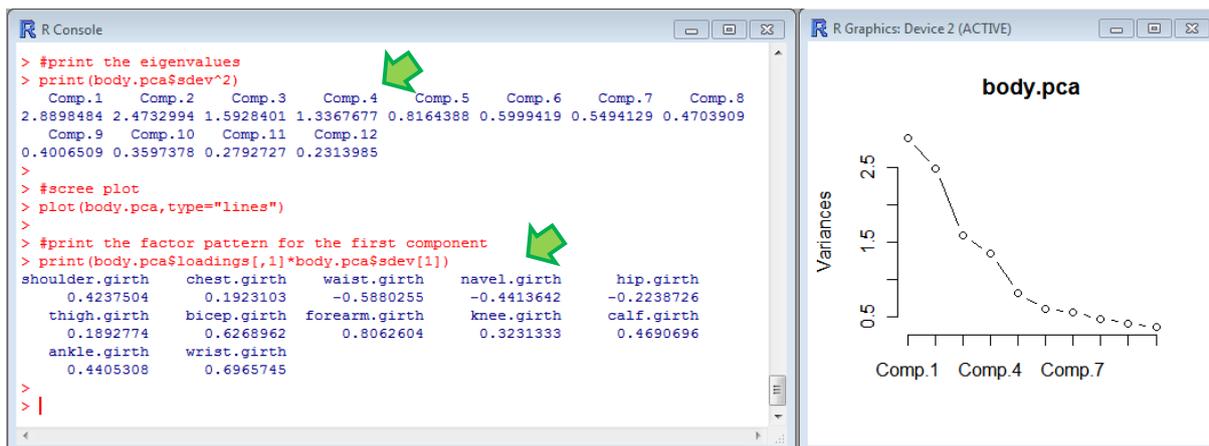
#éboulis des valeurs propres
plot(body.pca,type="lines")

#affichage des corrélations des variables avec la première composante
print(body.pca$loadings[,1]*body.pca$sdev[1])

```

## Résultats

Sans surprise, nous retrouvons exactement les résultats de SAS et de Tanagra.



## Conclusion

Il est rare quand même que l'on ne trouve pas – comme ici, l'ACP sur matrices de corrélations partielles – de package permettant de résoudre un problème précis sous R. Heureusement, on peut quasiment tout implémenter avec le langage R, c'est ce qui fait son charme. Avec un programme de taille relativement réduite, on obtient rapidement les résultats souhaités.