Objectif

Travailler avec les corrélations partielles dans TANAGRA.

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 en médecine, 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 significativement 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.

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.

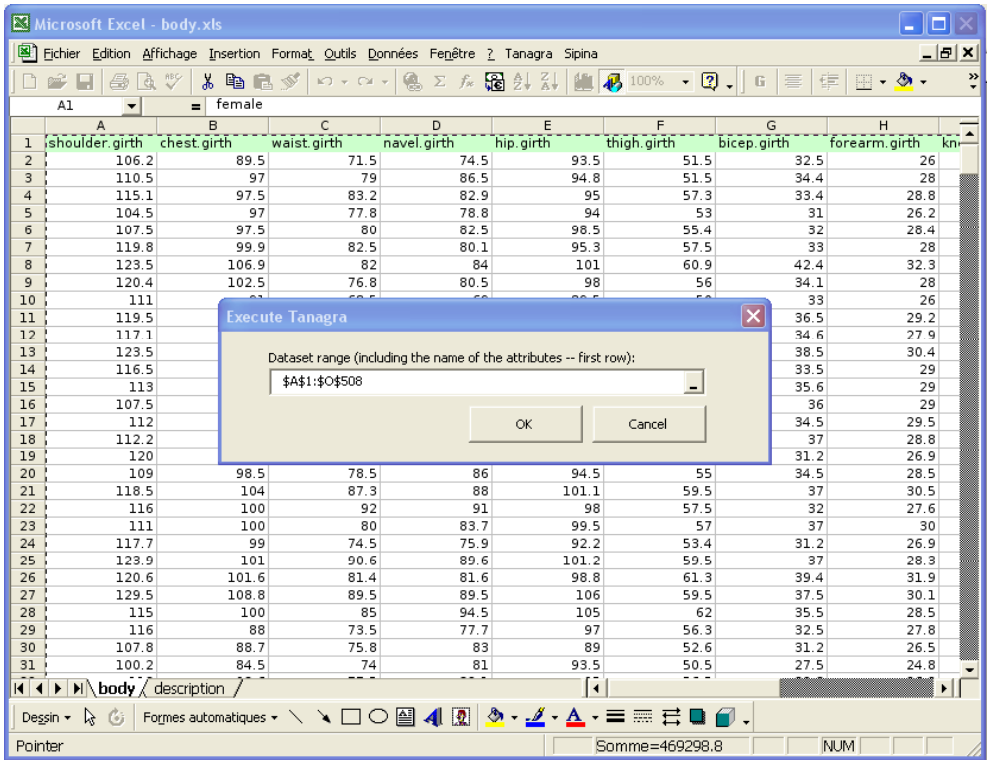
Analyse en composantes principales

Créer un diagramme

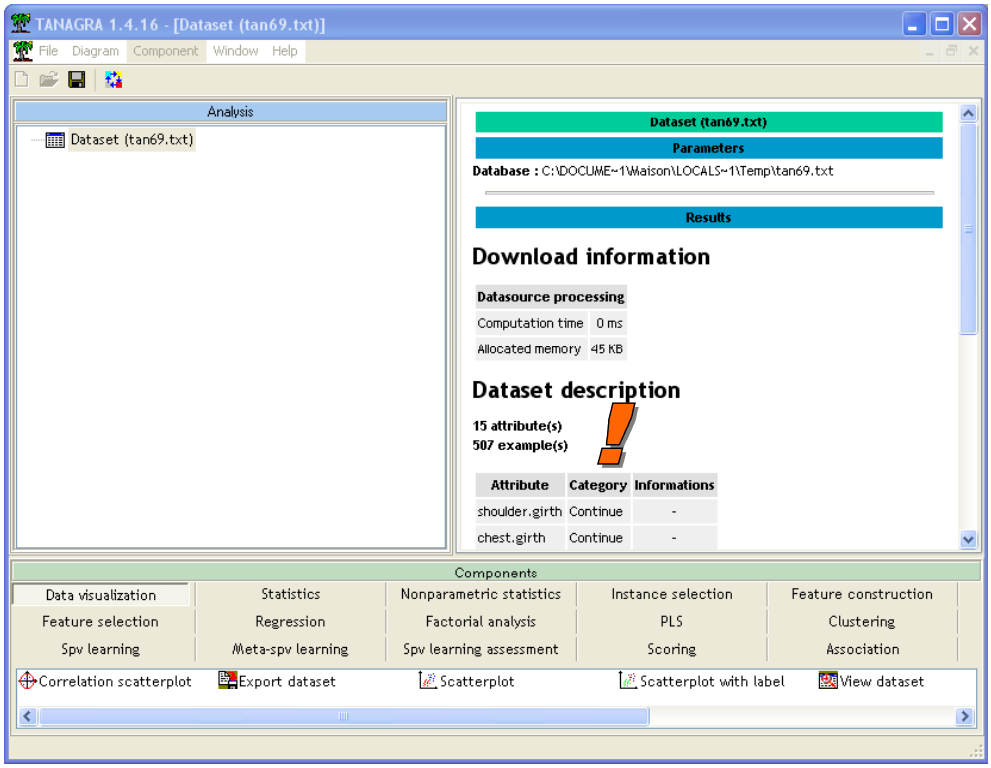
Dans un premier temps, nous réalisons une ACP « normale ». Pour ce faire, nous partons du fichier ouvert dans le tableur EXCEL, nous sélectionnons la plage de données et nous activons le menu TANAGRA / EXECUTE TANAGRA[1].

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

[1] 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 le didacticiel http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_Excel_AddIn.pdf pour la procédure d'installation et d'activation.

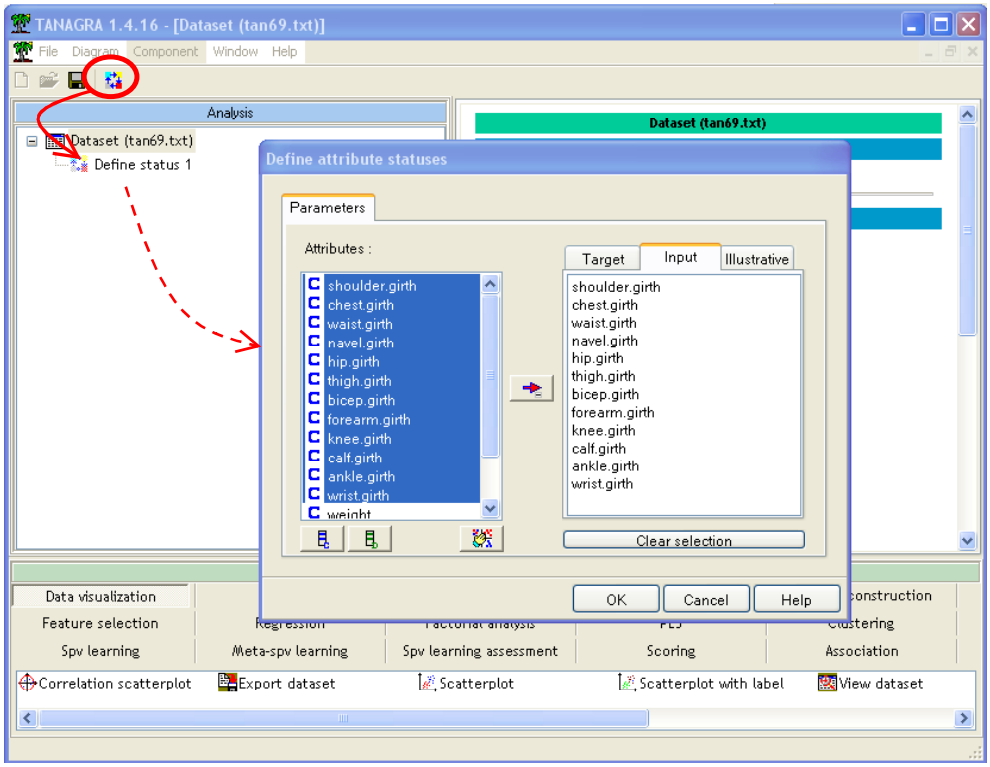


TANAGRA est automatiquement démarré. Un nouveau diagramme est créé avec les données de l'étude. Nous vérifions que nous disposons bien de 15 variables et 507 observations.

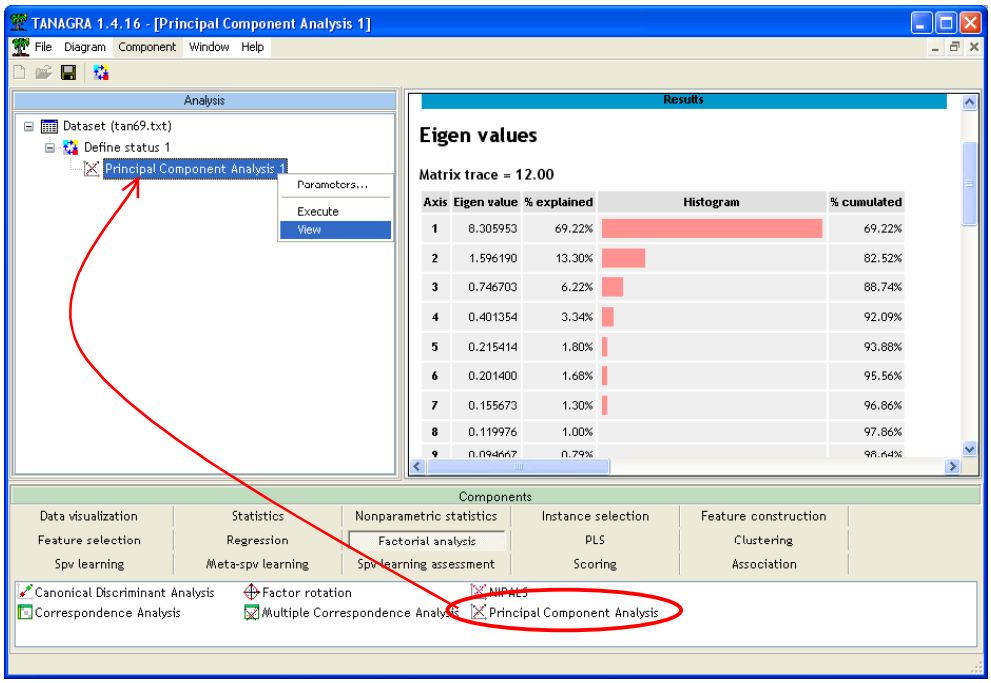


Définir une analyse

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'outil. Nous plaçons en INPUT les variables SHOULDER.GIRTH jusqu'à WRIST.GIRTH puis nous validons.

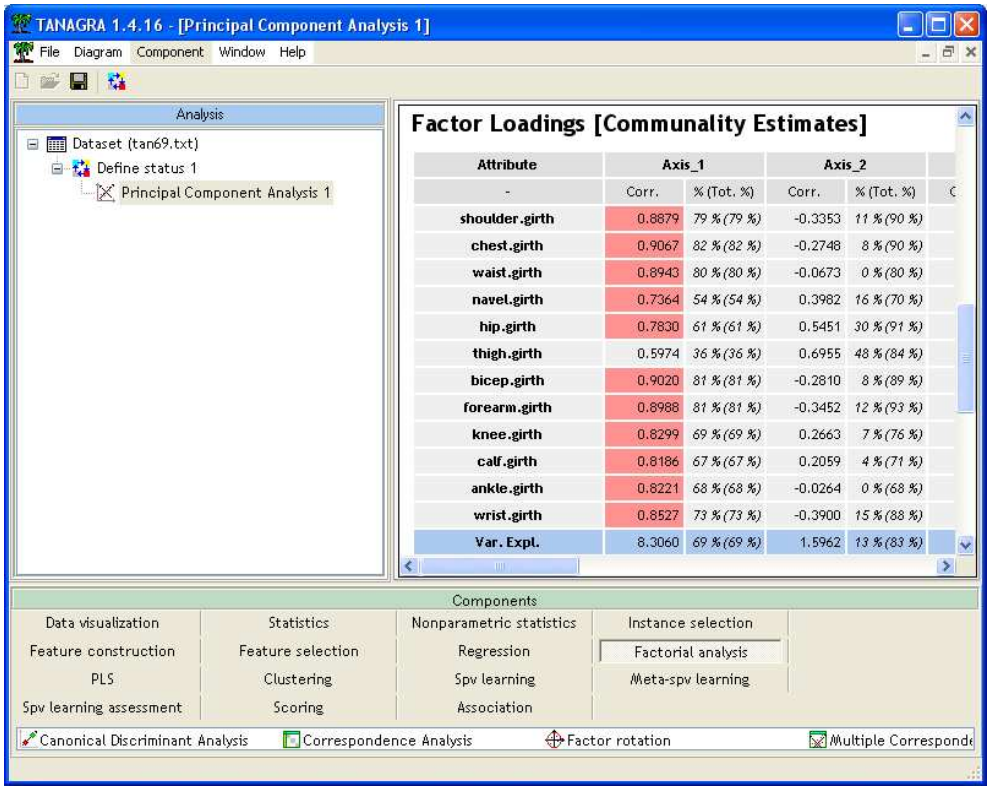


Puis nous insérons dans le diagramme le composant PRINCIPAL COMPONENT ANALYSIS situé dans l'onglet FACTORIAL ANALYSIS. Nous activons le menu VIEW pour accéder aux résultats.



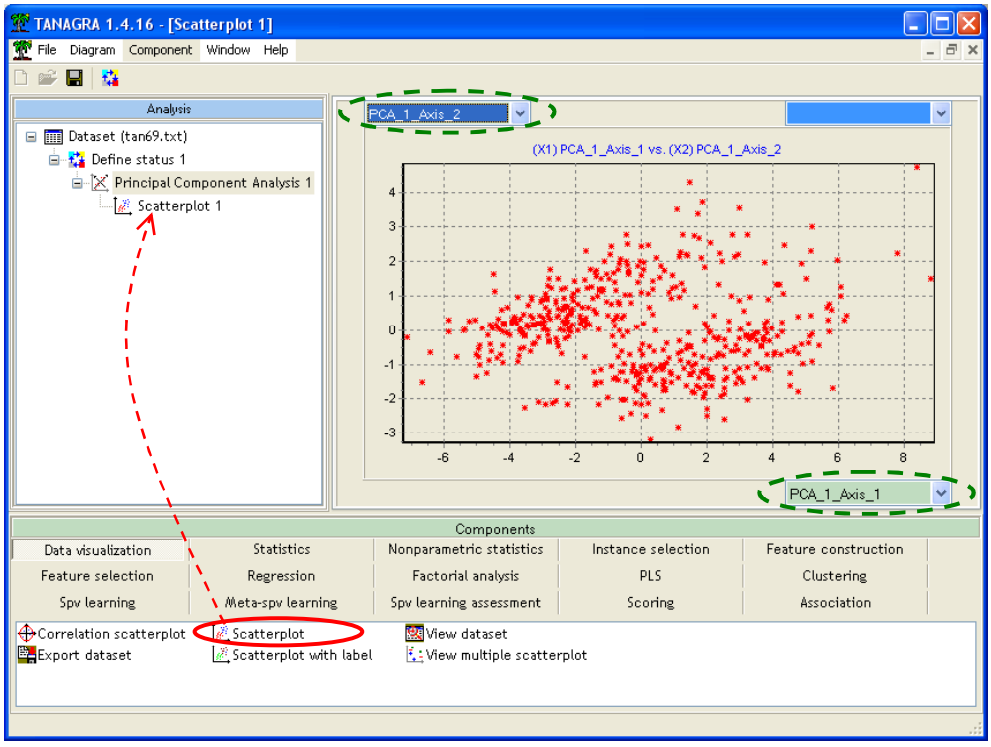
Voilà une belle ACP, les deux premiers axes résument 82,52% des informations. En consultant le tableau des corrélations, nous constatons que toutes les variables, exceptée THIGH (les cuisses),

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.

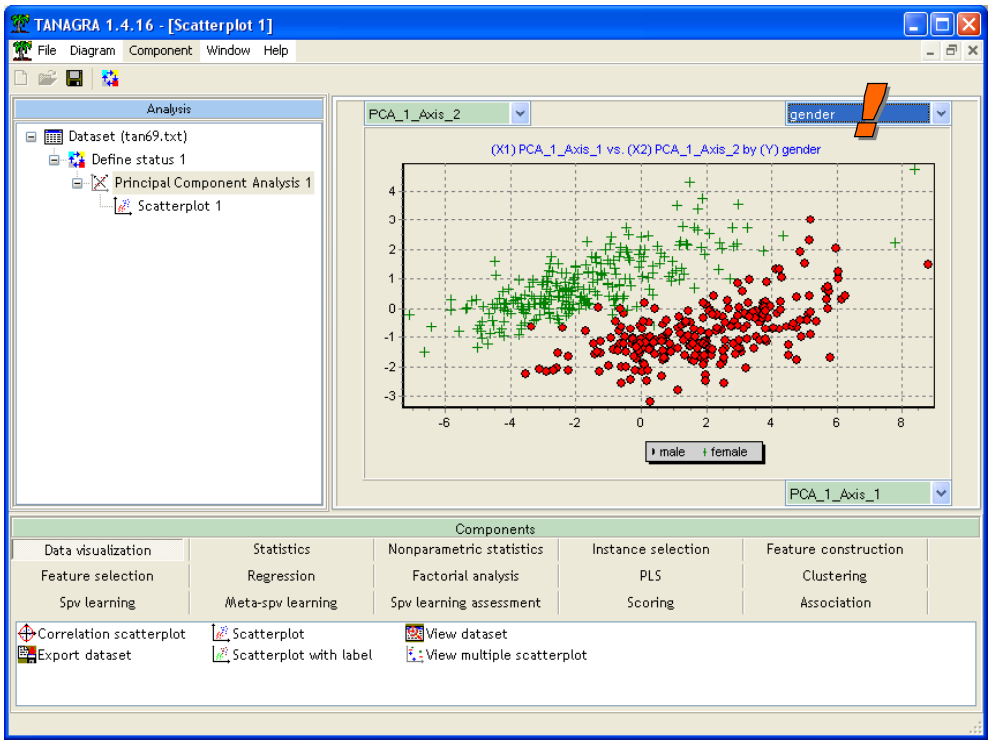


Attribute	Axis_1		Axis_2	
-	Corr.	% (Tot. %)	Corr.	% (Tot. %)
shoulder.girth	0.8879	79 % (79 %)	-0.3353	11 % (90 %)
chest.girth	0.9067	82 % (82 %)	-0.2748	8 % (90 %)
waist.girth	0.8943	80 % (80 %)	-0.0673	0 % (80 %)
navel.girth	0.7364	54 % (54 %)	0.3982	16 % (70 %)
hip.girth	0.7830	61 % (61 %)	0.5451	30 % (91 %)
thigh.girth	0.5974	36 % (36 %)	0.6955	48 % (84 %)
bicep.girth	0.9020	81 % (81 %)	-0.2810	8 % (89 %)
forearm.girth	0.8988	81 % (81 %)	-0.3452	12 % (93 %)
knee.girth	0.8299	69 % (69 %)	0.2663	7 % (76 %)
calf.girth	0.8186	67 % (67 %)	0.2059	4 % (71 %)
ankle.girth	0.8221	68 % (68 %)	-0.0264	0 % (68 %)
wrist.girth	0.8527	73 % (73 %)	-0.3900	15 % (88 %)
Var. Expl.	8.3060	69 % (69 %)	1.5962	13 % (83 %)

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 ? »

Travailler sur les corrélations partielles

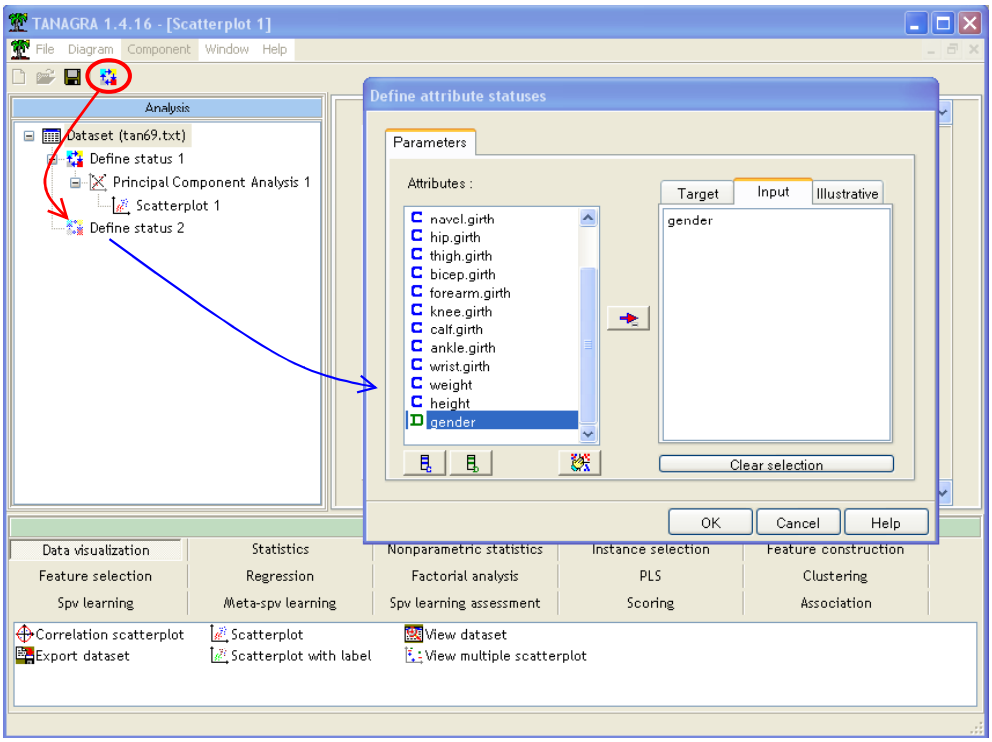
Il nous faut donc éliminer la différenciation selon le sexe dans notre étude. Une approche triviale serait de scinder le fichier en deux parties : d'un côté les hommes, de l'autre les femmes. Nous mènerions alors en parallèle les deux études, puis nous confrontons les résultats pour en extraire une synthèse.

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.

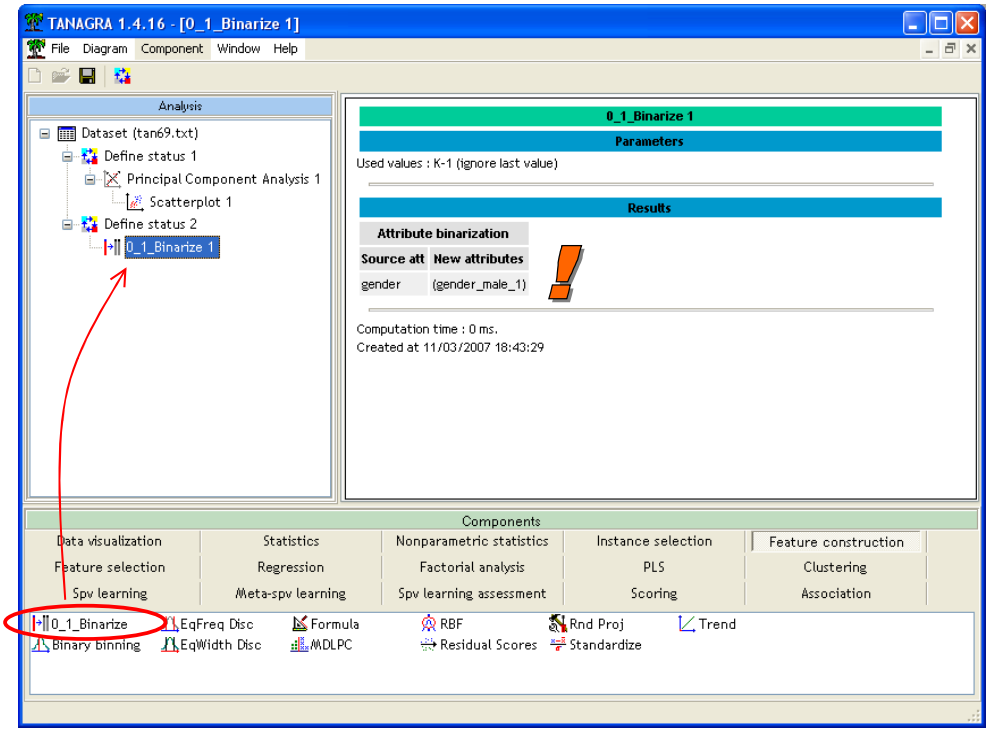
Élimination de l'effet SEXE dans les variables

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'outil.

Nous plaçons en INPUT la variable GENDER.



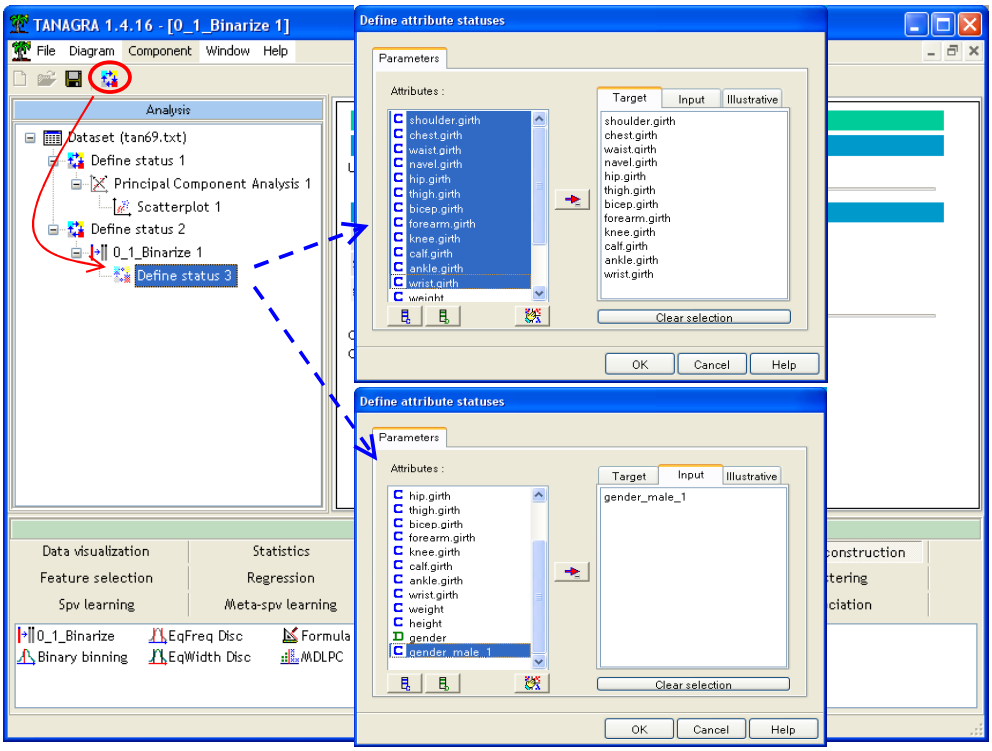
Nous insérons à la suite le composant 0_1_BINARIZE (onglet FEATURE CONSTRUCTION). Nous actions le menu VIEW.



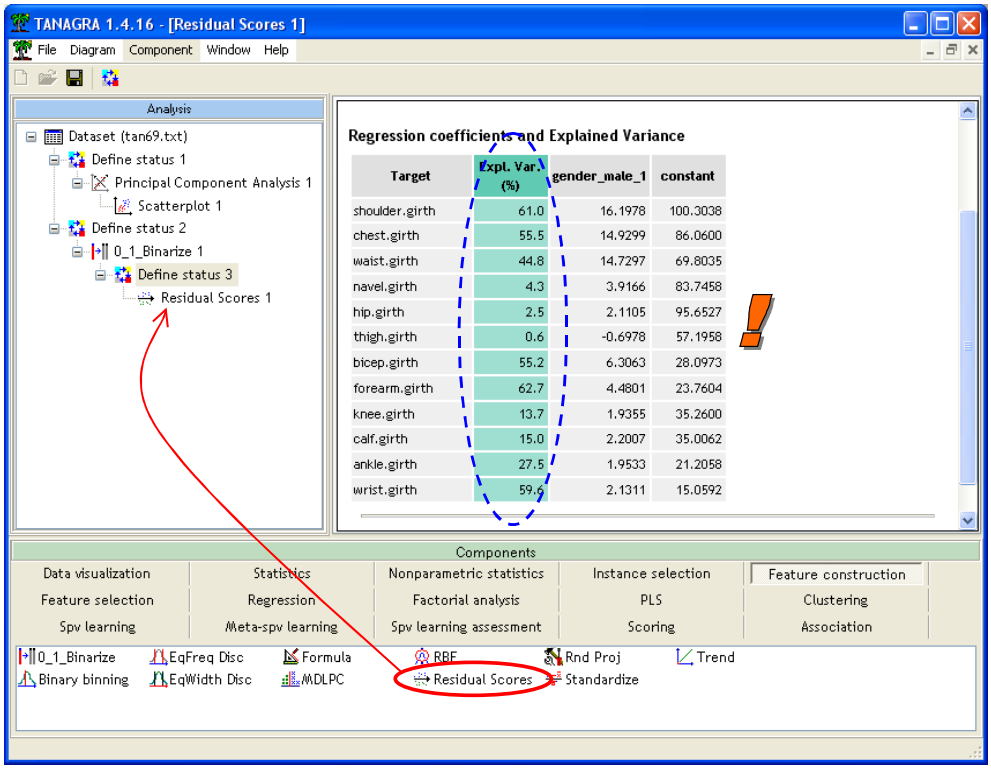
TANAGRA nous indique que la variable GENDER a été recodée en GENDER_MALE_1, c.-à-d. que les hommes sont codés 1, les femmes 0.

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.



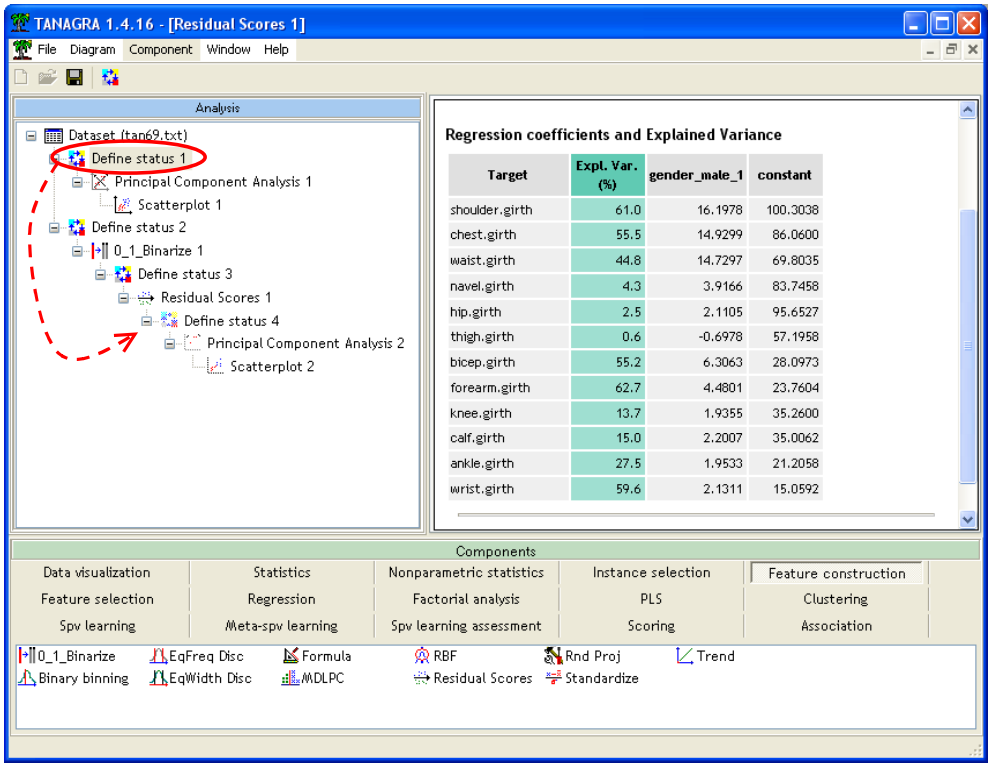
Nous insérons alors le composant RESIDUAL SCORES (onglet FEATURE CONSTRUCTION) dans le diagramme. Son rôle est simple, il calcule le résidu de la régression de chaque variable cible sur les variables INPUT. Le composant indique les coefficients de la régression, il indique surtout la proportion de variabilité 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)[2].

Analyse en composantes principales sur les résidus

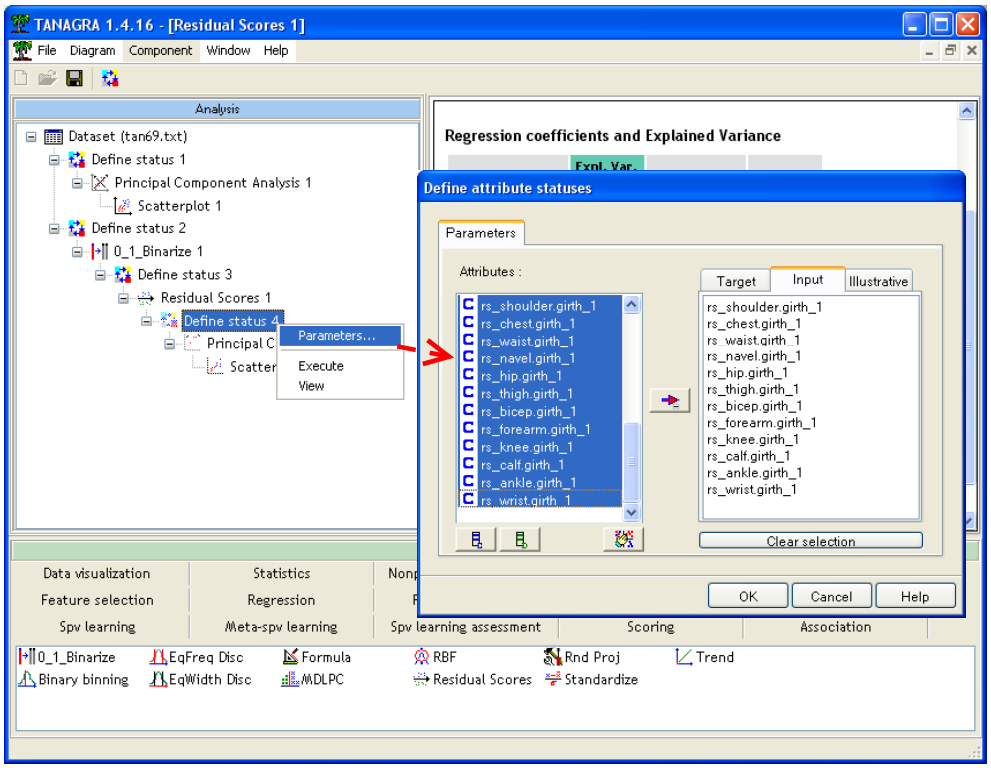
Ayant éliminé l'effet GENDER dans nos variables, nous pouvons maintenant lancer l'ACP. Le plus simple est de copier par glisser-déposer la séquence DEFINE STATUS 1 – PRINCIPAL COMPONENT ANALYSIS 1 – SCATTERPLOT 1 après le composant RESIDUAL SCORES 1.



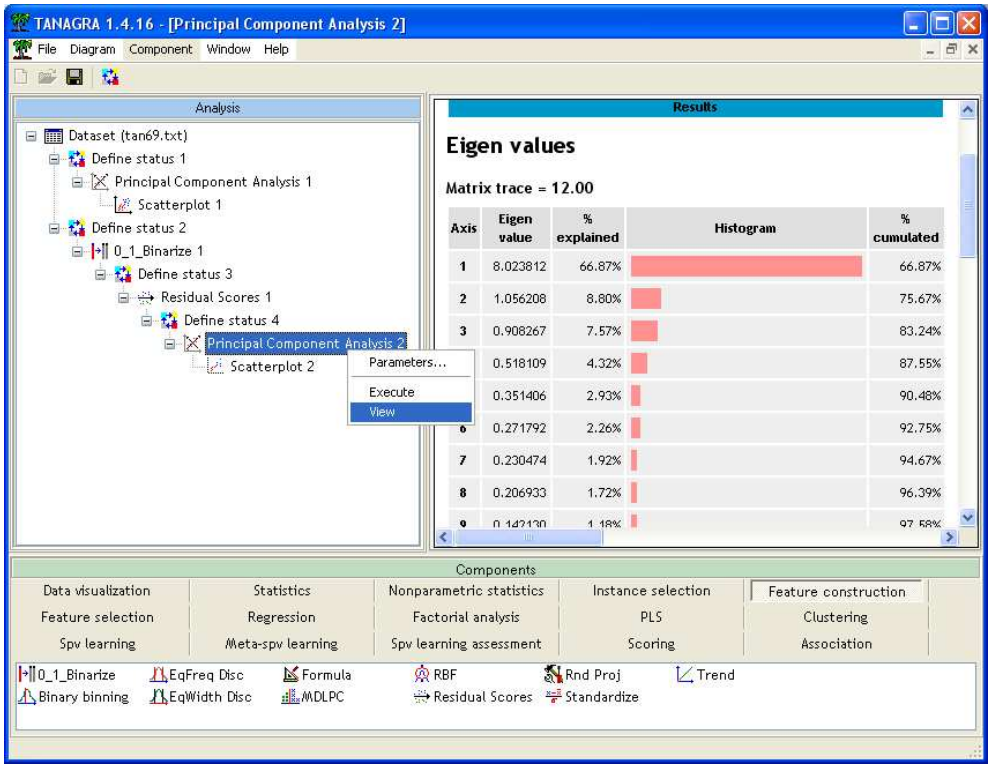
Nous paramétrons le composant DEFINE STATUS 4 (menu PARAMETERS) de manière à placer en INPUT les variables résiduelles RS_SHOUDLER.GIRTH_1 à RS_WRIST.GIRTH_1.

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.

[2] 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.

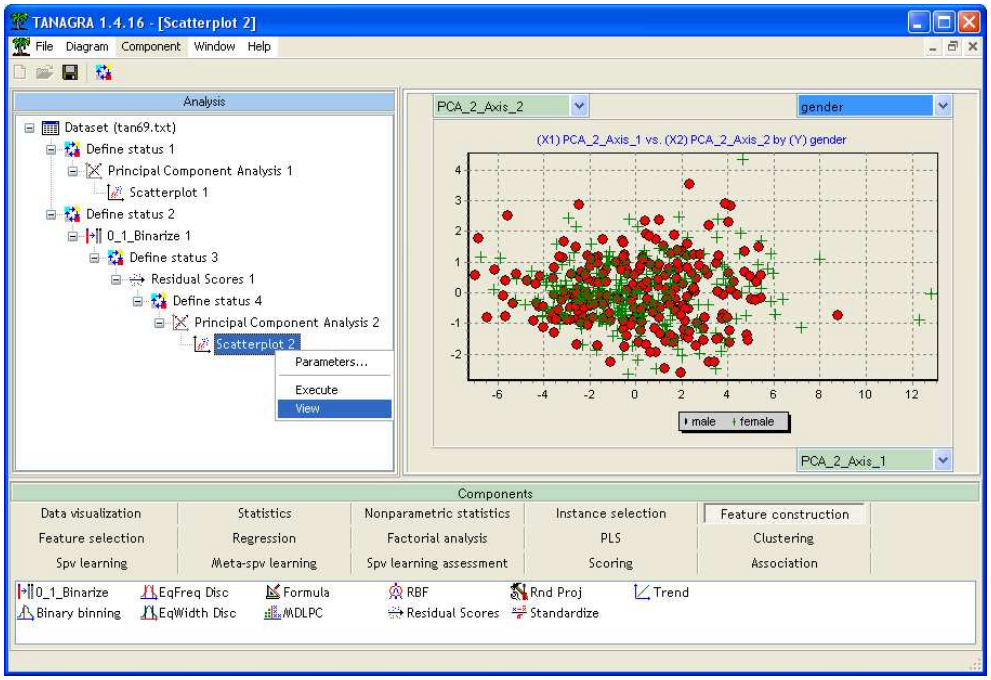


Nous activons le menu VIEW de l'ACP pour accéder aux résultats.



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.

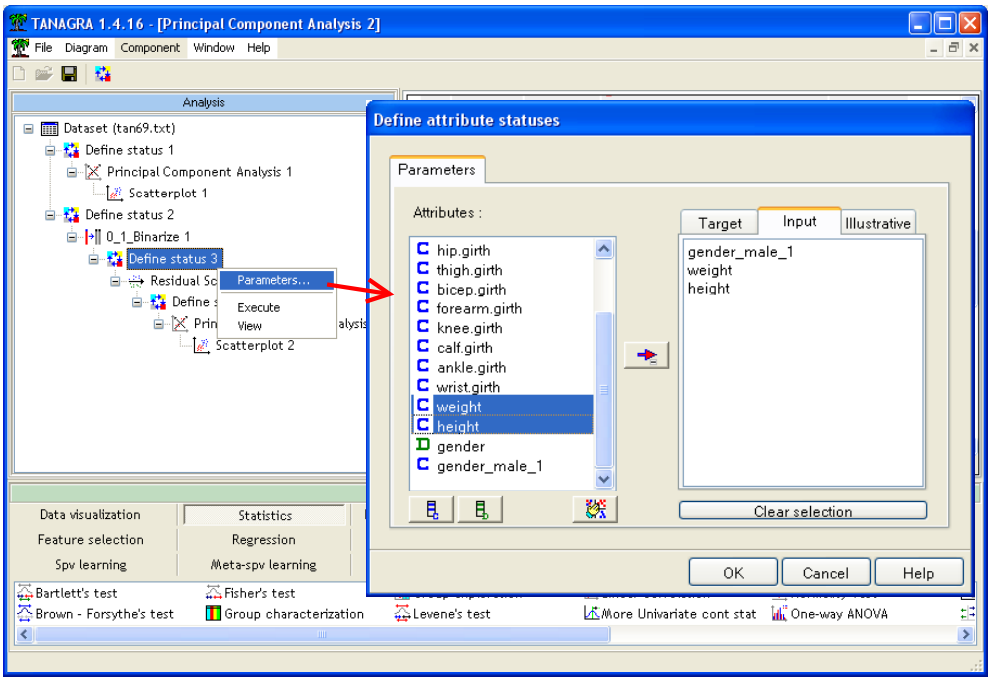
Élimination de l'effet taille

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.

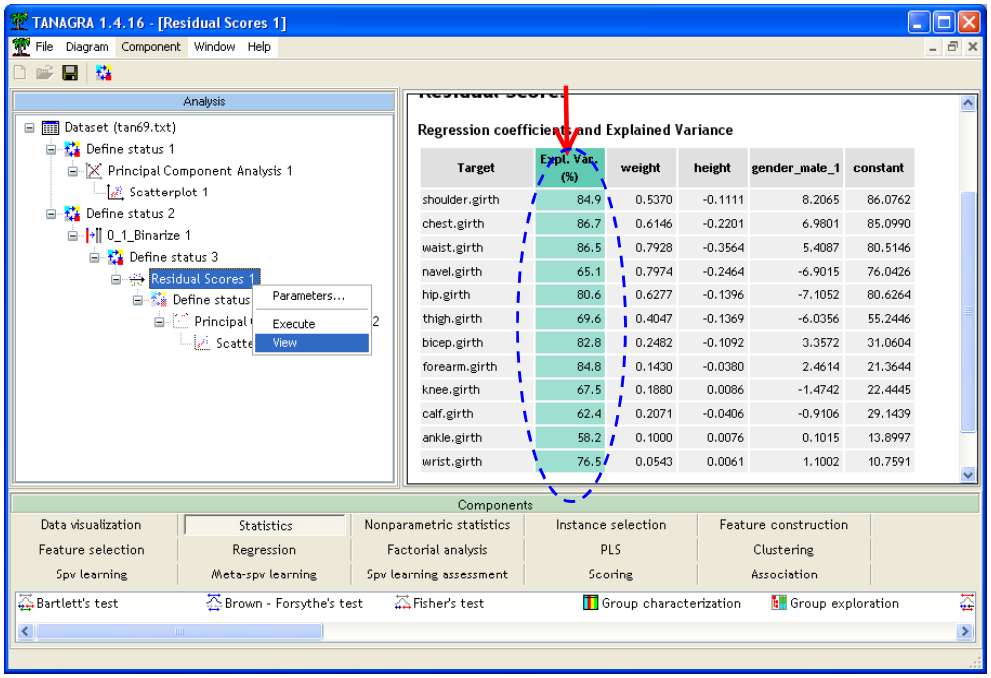
Tout dépend de l'étude que l'on veut mener. Si ce résultat nous satisfait et correspond à ce que nous cherchions, nous pouvons nous en tenir à ce résultat : les circonférences des différentes parties du corps humain sont positivement corrélées tout simplement parce qu'elles traduisent l'influence sous-jacente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

Si, en revanche, ce résultat 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 devrons nous procéder ?

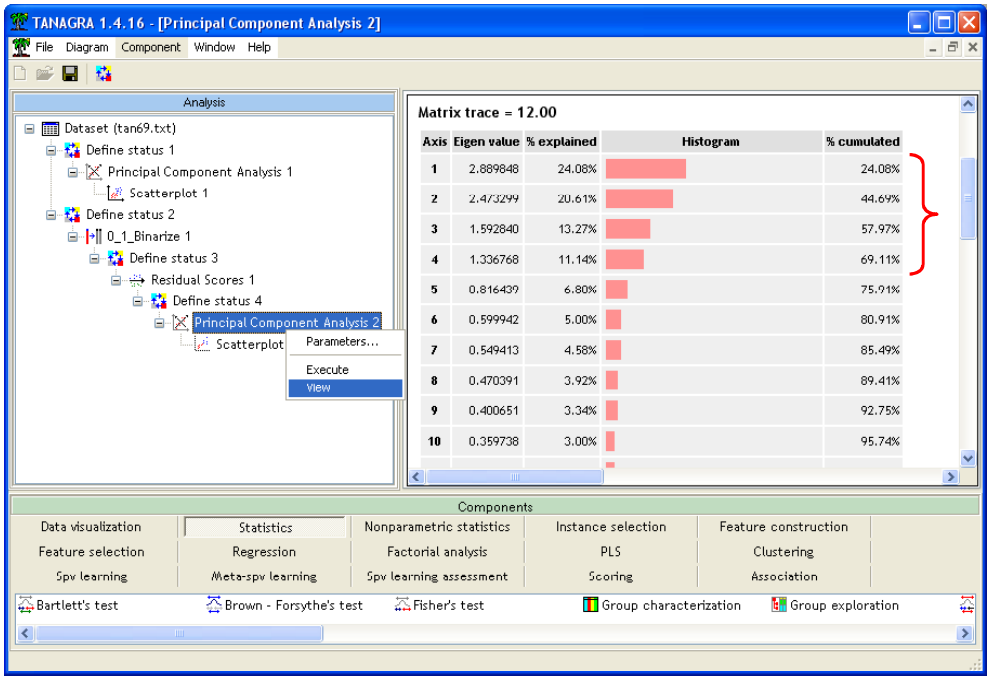
Introduire le poids et la taille dans le calcul des résidus. Nous revenons sur le composant DEFINES STATUS 3, et nous le paramétrons (menu PARAMETERS) de manière à introduire en INPUT, en plus de le variable GENDER_MALE_1, les variables WEIGHT et HEIGHT.



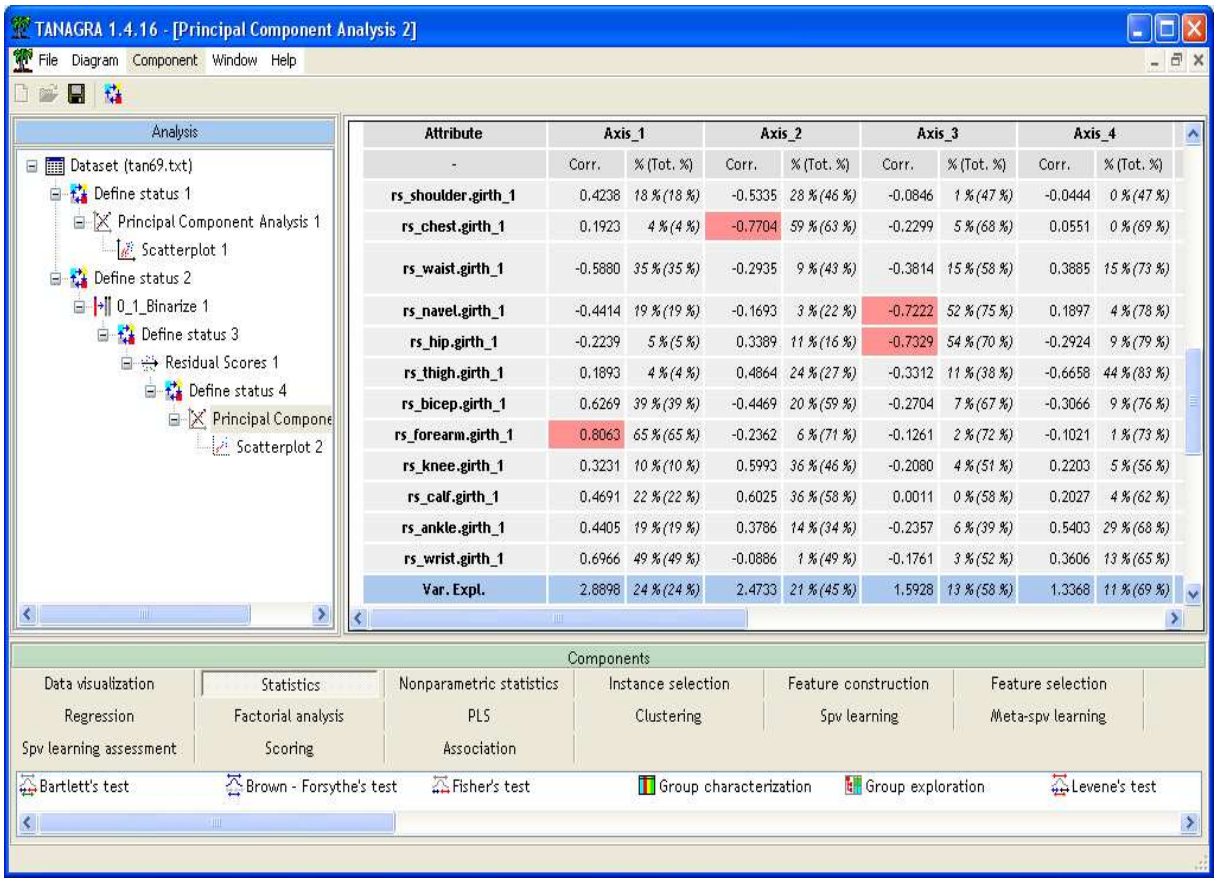
Nous activons alors le menu VIEW du composant RESIDUAL SCORES 1, nous observons que dans plusieurs cas, ce triplet de variables explique près de 80% de la variabilité. Ce qui confirme largement notre intuition ci-dessus.



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.



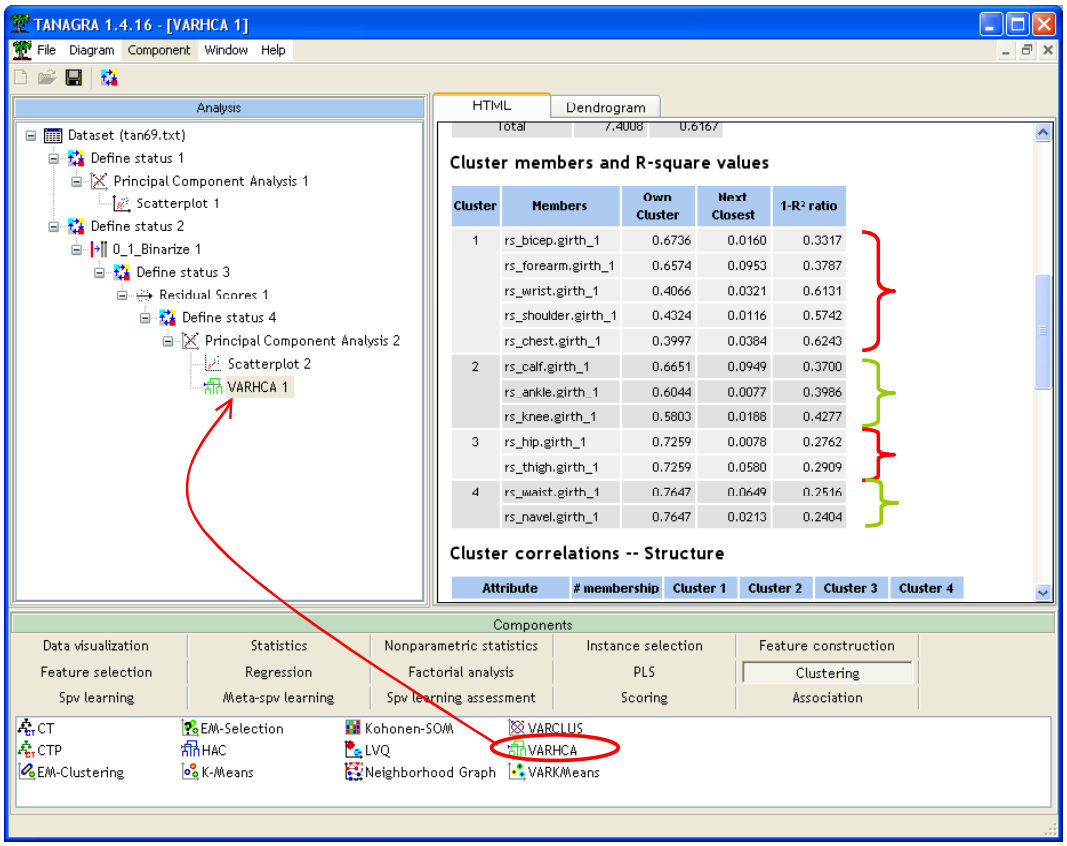
L'étude des corrélations montre qu'en réalité les relations entre les dimensions des différentes zones du corps, une fois que nous nous sommes ramenés à un individu « standard » (sexe, poids et taille égal) sont un peu plus complexes que ce que nous pouvions imaginer dans un premier temps.



Classification des 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[3] à la suite du composant DEFINE STATUS 4 dans le diagramme. Son exécution indique qu'il y a en réalité 4 groupes de variables corrélées selon les différentes zones du corps humain.



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)	Sous le nombril et le genou

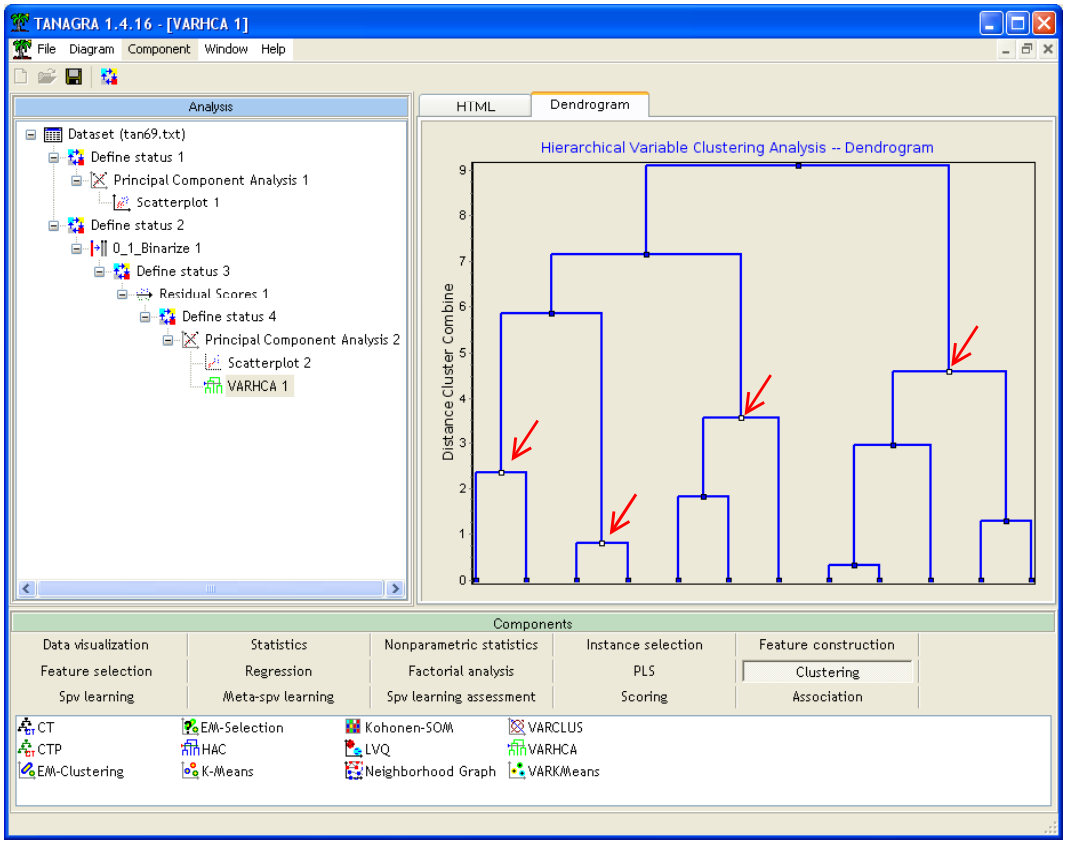
[3] L'utilisation des composants VARIABLE CLUSTERING (VARCLUS) est détaillée dans le didacticiel http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/fr_Tanagra_VarClus.pdf disponible en ligne.

rs_thigh.girth_1 (cuisse)	
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[4].

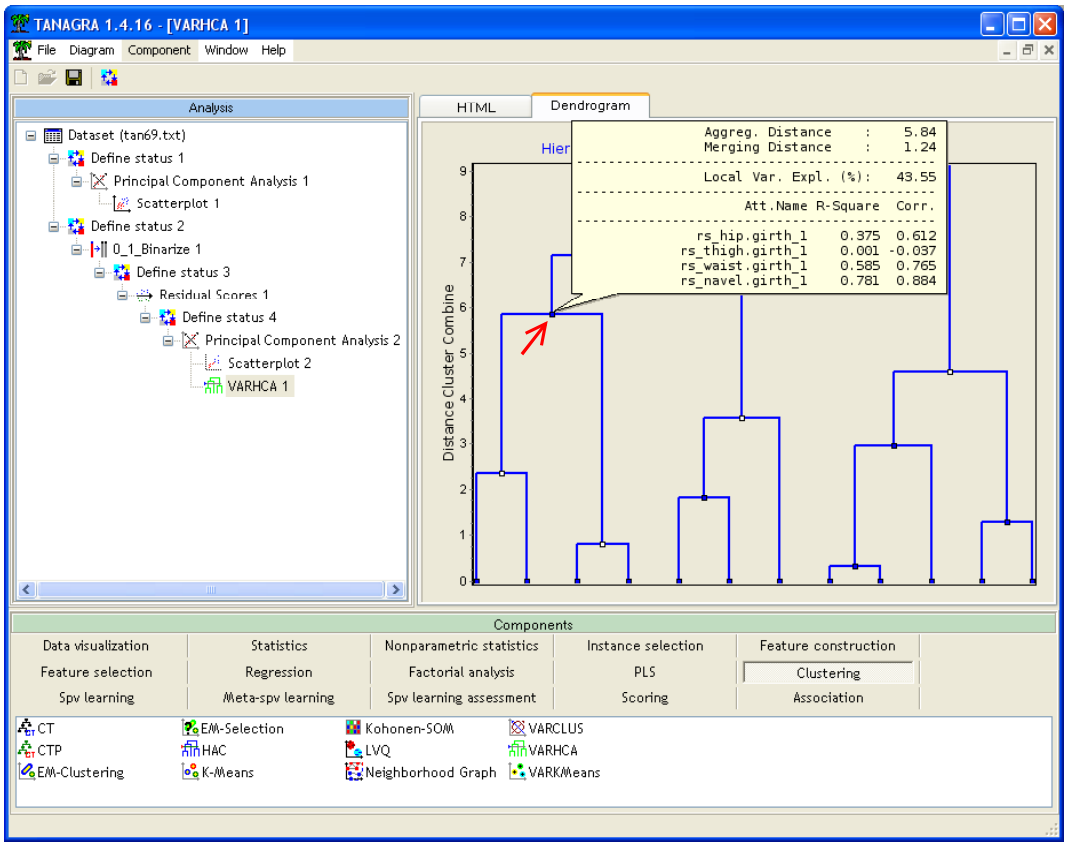
Cette subdivision en 4 groupes a été automatiquement déterminée par l'algorithme. Si elle ne cadre pas avec les connaissances du domaine, nous avons la possibilité de l'affiner en étudiant le dendrogramme associée au processus de typologie.

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 3ème et 4ème 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.

[4] Ici s'arrête le travail du statisticien ou du data miner et commence le travail de l'expert métier !!!



Conclusion

Les facteurs confondants font partie des plus gros écueils de l'analyse exploratoire des données. Ne pas les détecter nous exposent 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.