1 Objectif

Calculer les corrélations partielles avec Tanagra.

Le coefficient de corrélation est une mesure statistique destinée à quantifier l'intensité d'un lien (linéaire) entre 2 variables. Il est possible de mettre en place un test de significativité qui cherche à établir l'existence de la relation dans la population.

Le coefficient de corrélation est un instrument très populaire. Mais comme tout outil numérique, il a ses faiblesses. La plus criante étant certainement la corrélation factice : 2 variables semblent fortement liées, on se rend compte après coup que la liaison repose sur l'intervention d'une troisième variable. Par exemple, la corrélation entre la longueur des jambes et la longueur des avant-bras est très forte. Elle repose en réalité sur la taille des personnes : les grands ont tendance à avoir des jambes et des avant-bras longs, inversement chez les petits.

La corrélation partielle corrige cet inconvénient. Elle mesure la liaison en annulant l'effet de la troisième variable, dite variable de contrôle. Dans notre exemple, il s'agit de mesurer, à taille de personne égale, la relation entre les longueurs des jambes et des bras. Nous pouvons faire intervenir plusieurs variables de contrôle[1].

Dans ce didacticiel, nous montrons comment mettre en œuvre le composant PARTIAL CORRELATION dans Tanagra. Nous reprenons un exemple décrit sur un excellent site de cours en ligne[2]. Outre une présentation théorique de la technique, le détail des calculs est disponible. Nous pouvons retracer les étapes de construction de la mesure, le test de significativité et l'élaboration des intervalles de confiance[3]. Nous pouvons aussi nous comparer avec les résultats établis à l'aide d'autres logiciels de statistique.

2 Données

Les données proviennent d'un test d'intelligence (QI) basé sur la méthode WAIS (Wechsler Adult Intelligence Scale)[4]. Nous disposons de 37 observations mesurées sur 4 dimensions : « Information », le degré de connaissance associée à la culture ; « Similarities », la capacité d'abstraction verbale ; « Arithmetic », le calcul mental ; « Picture.Completion », la capacité à percevoir les détails visuel. Nous cherchons à caractériser la liaison entre INFORMATION et SIMILARITIES, les variables de contrôle seront ARITHMETIC et PICTURE.COMPLETION.

3 Calculer la corrélation partielle

3.1 Importer les données et créer un diagramme

Le plus simple pour lancer Tanagra et charger les données est d'ouvrir le fichier XLS[5] dans le tableur EXCEL. Nous sélectionnons la plage de données. La première ligne doit correspondre au

[1] Voir le support en ligne http://eric.univ-lyon2.fr/~ricco/cours/cours/Analyse_de_Correlation.pdf

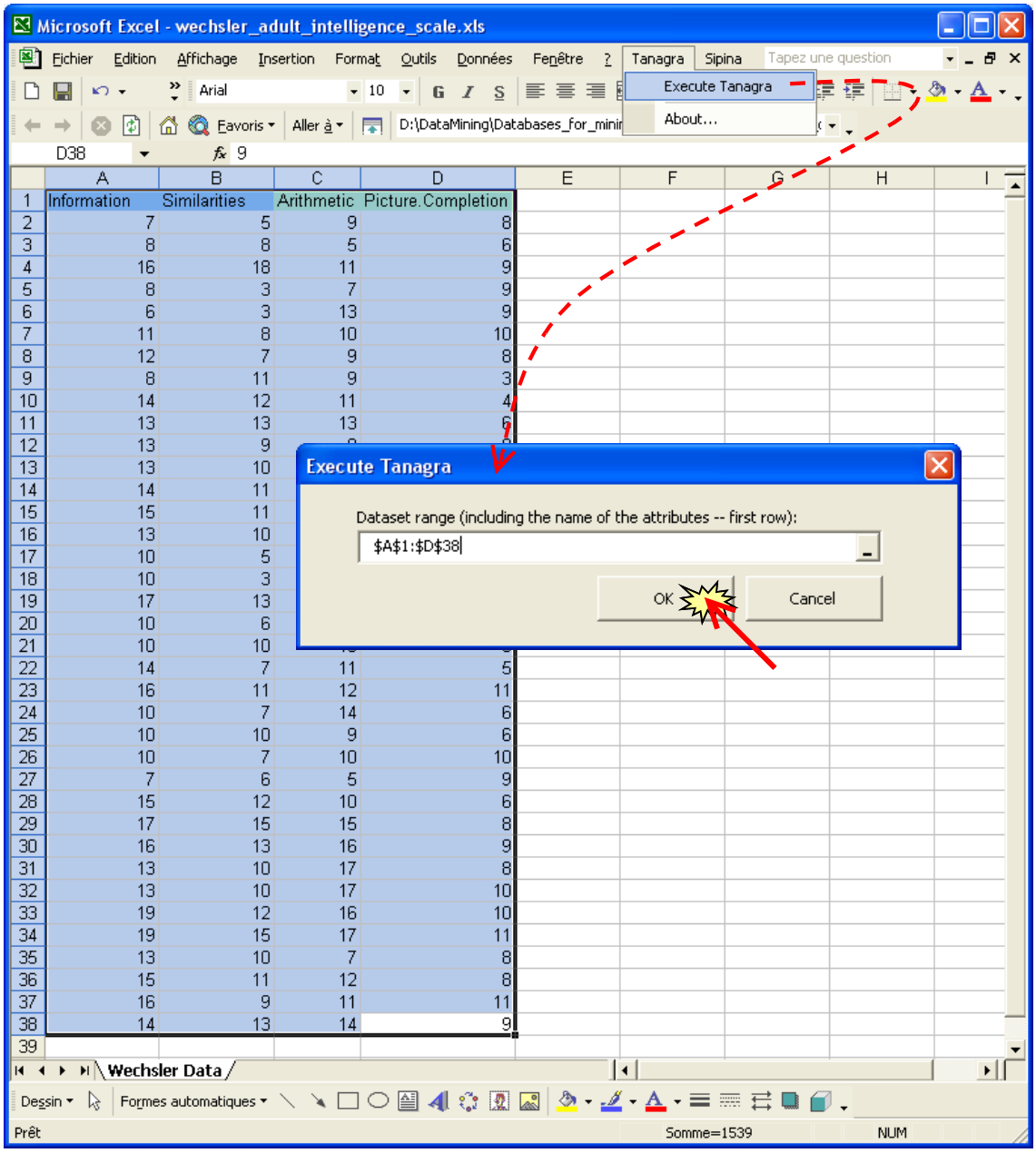
[2] http://www.stat.psu.edu/online/development/stat505/07_partcor/01_partcor_intro.html

[3] http://www.stat.psu.edu/online/development/stat505/07_partcor/06_partcor_partial.html

[4] http://en.wikipedia.org/wiki/Wechsler_Adult_Intelligence_Scale

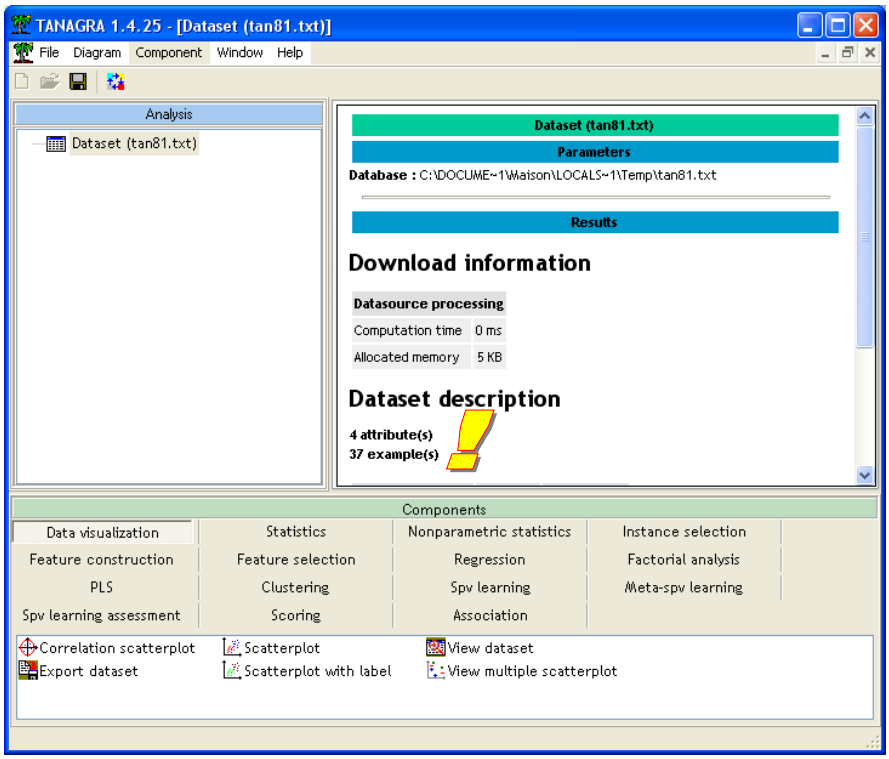
[5] http://eric.univ-lyon2.fr/~ricco/tanagra/fichiers/wechsler_adult_intelligence_scale.xls

nom des variables. Puis nous activons le menu TANAGRA / EXECUTE TANAGRA qui a été installé avec la macro complémentaire TANAGRA.XLA[6]. Une boîte de dialogue apparaît. Nous vérifions la sélection. Si tout est en règle, nous validons en cliquant sur le bouton OK.



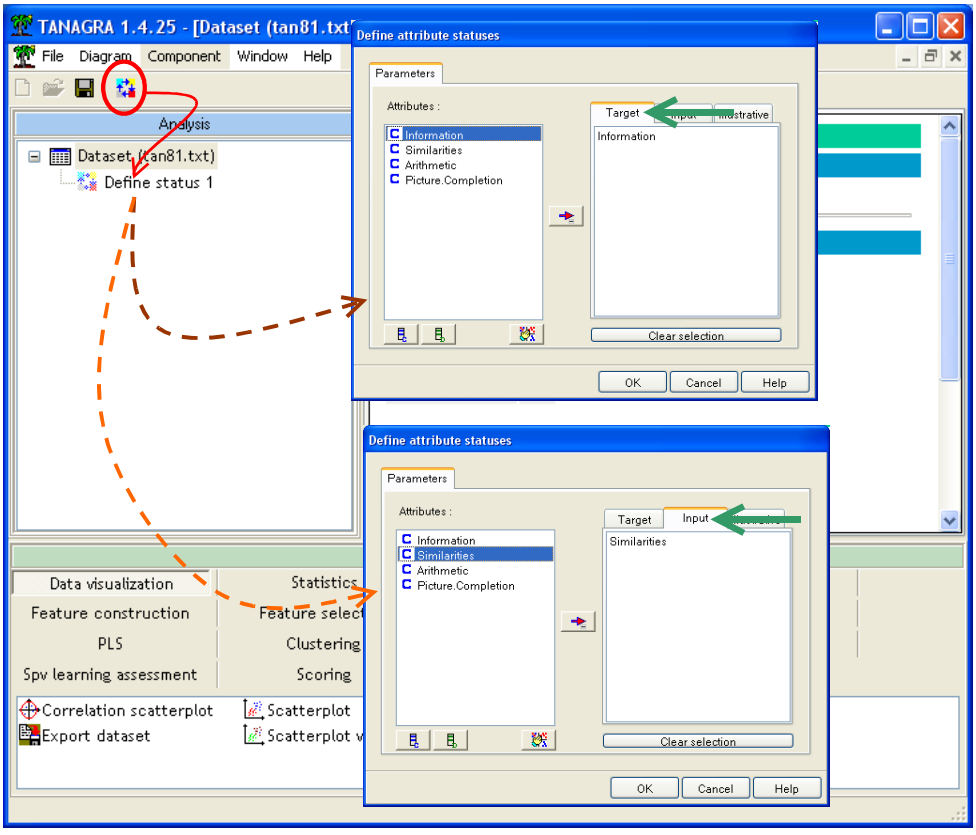
Tanagra est automatiquement démarré, les données chargées et un nouveau diagramme de traitements est créé. Nous disposons de 37 observations décrites à l'aide de 4 variables.

[6] Voir http://tutoriels-data-mining.blogspot.com/2008/03/importation-fichier-xls-excel-macro.html concernant l'installation et l'utilisation de la macro complémentaire TANAGRA.XLA.

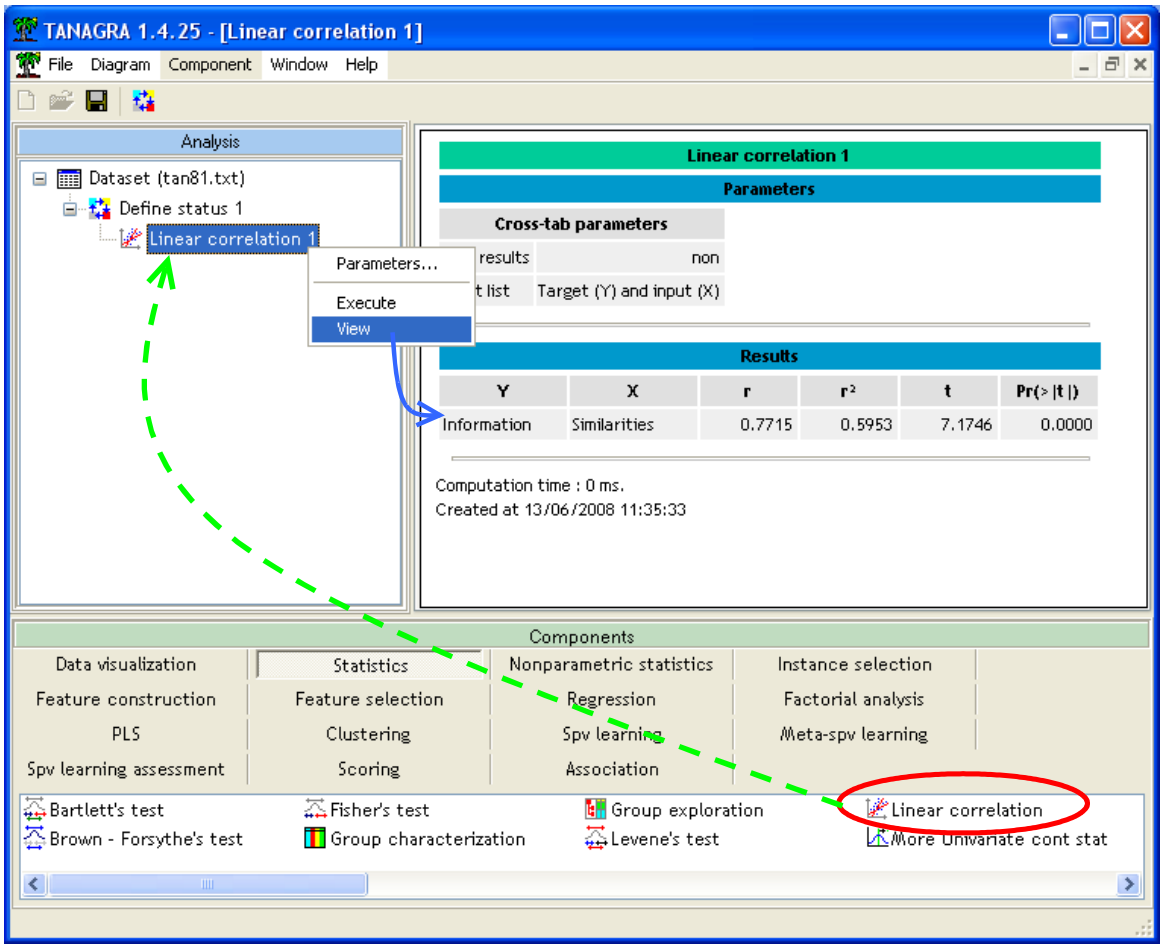


3.2 Corrélation brute

Dans un premier temps, nous désirons mesurer la corrélation de Pearson entre INFORMATION et SIMILARITIES. Nous introduisons tout d'abord le composant DEFINES STATUS dans le diagramme, accessible via le raccourci dans la barre d'outils. Nous plaçons en TARGET la variable INFORMATION, en INPUT, SIMILARITIES. L'analyse est symétrique, nous aurions pu faire l'inverse.



Nous insérons ensuite le composant LINEAR CORRELATION (onglet STATISTICS) dans le diagramme. Nous activons directement le menu VIEW pour accéder aux résultats.



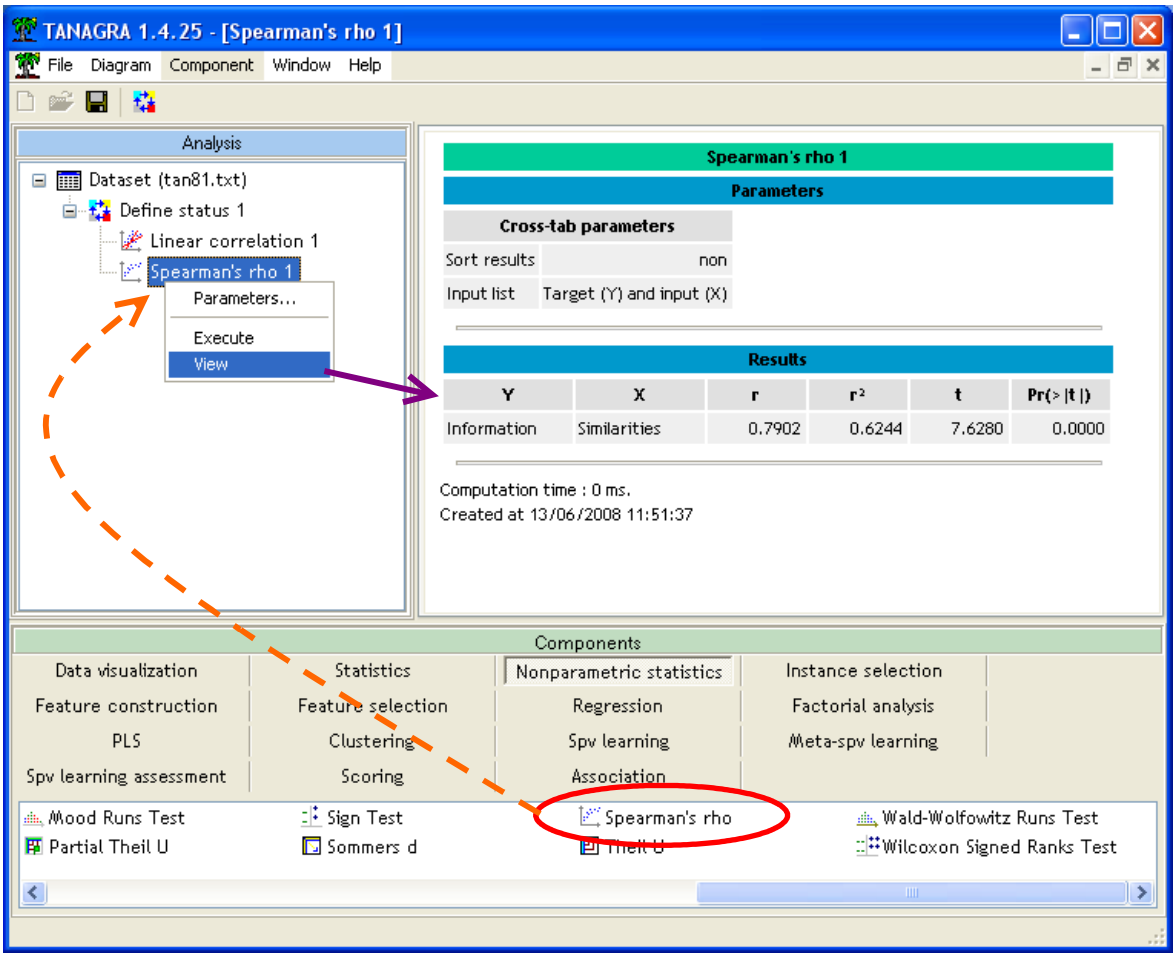
La corrélation entre INFORMATION et SIMILARITIES est 0.7715. La statistique du test de significativité est t = 7.1746. Sous H0, elle suit une loi de Student. La corrélation mesurée est très significative, avec une probabilité critique inférieure à 0.0001. Les personnes ayant un score élevé sur les tests relatifs à INFORMATION sont également performants pour ce qui est de SIMILARITIES.

Note : Ces calculs sont détaillés sur notre site de référence http://www.stat.psu.edu/online/development/stat505/06_propmean/03_propmean_infercorr.html. Le nombre de degrés de liberté est (37 – 2 = 35) pour le test de significativité de la corrélation brute. Nous y reviendrons plus tard lorsque l'on étudiera les corrélations partielles.

3.3 Corrélation de rangs : coefficient de Spearman

Le coefficient de Pearson est un bon indicateur, très populaire. Mais il présente des fragilités qu'il importe de circonscrire : il ne caractérise que les liaisons linéaires, la présence de points atypiques perturbe les calculs. Pour éviter ces deux écueils, une solution simple est de transformer les données en rangs, puis de calculer la corrélation sur ces nouvelles données, on vient de définir le coefficient de corrélation de rangs de Spearman. Il peut caractériser des liaisons non linéaires, qui doivent être monotones néanmoins. Il est peu sensible aux points aberrants.

Pour calculer le coefficient de Spearman, nous insérons le composant SPEARMAN'S RHO (onglet NONPARAMETRIC STATISTICS) dans le diagramme. Nous obtenons les résultats suivants :

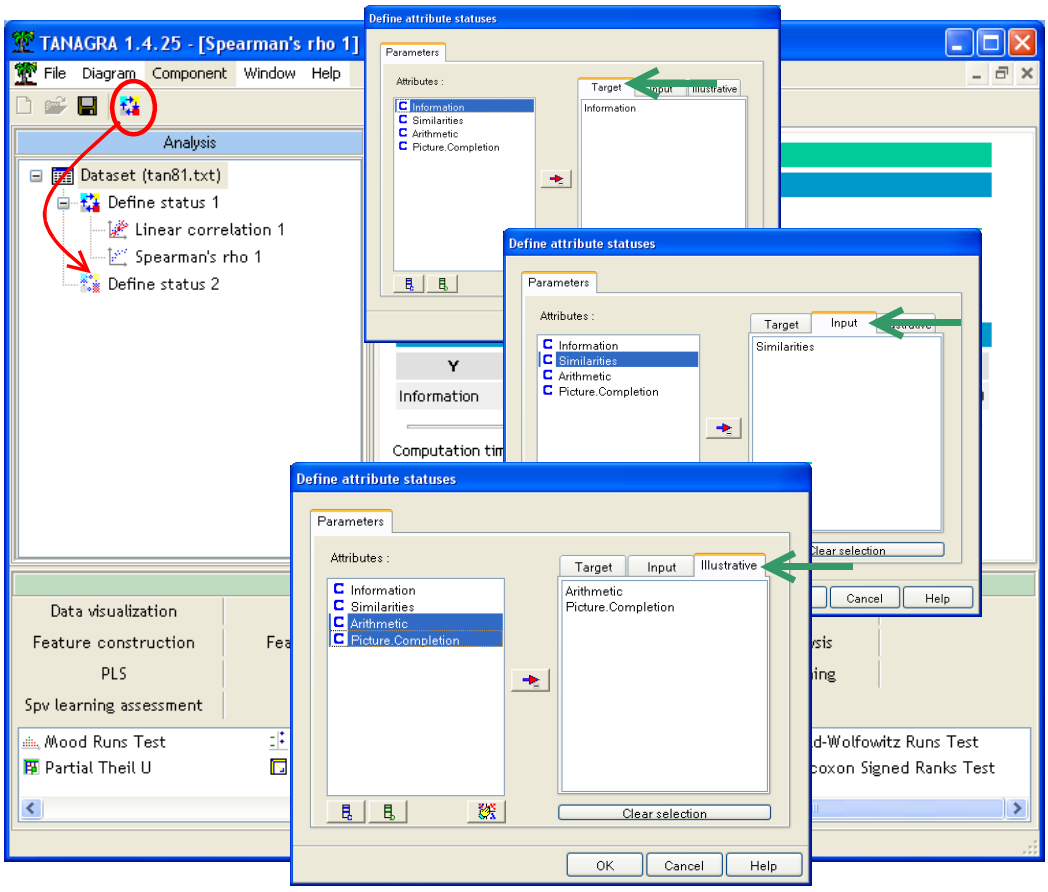


Rho = 0.7902, il est aussi très significatif. La proximité entre les valeurs de « Rho » et « r » laisse à penser que la liaison est effectivement linéaire. Dans le cas contraire, lorsque « Rho » est nettement plus élevé (en valeur absolue) que « r », il faut s'en inquiéter, la liaison est certes monotone, mais elle n'est pas linéaire. Un graphique nuage de points précise généralement l'idée que l'on doit se faire de la forme de la relation.

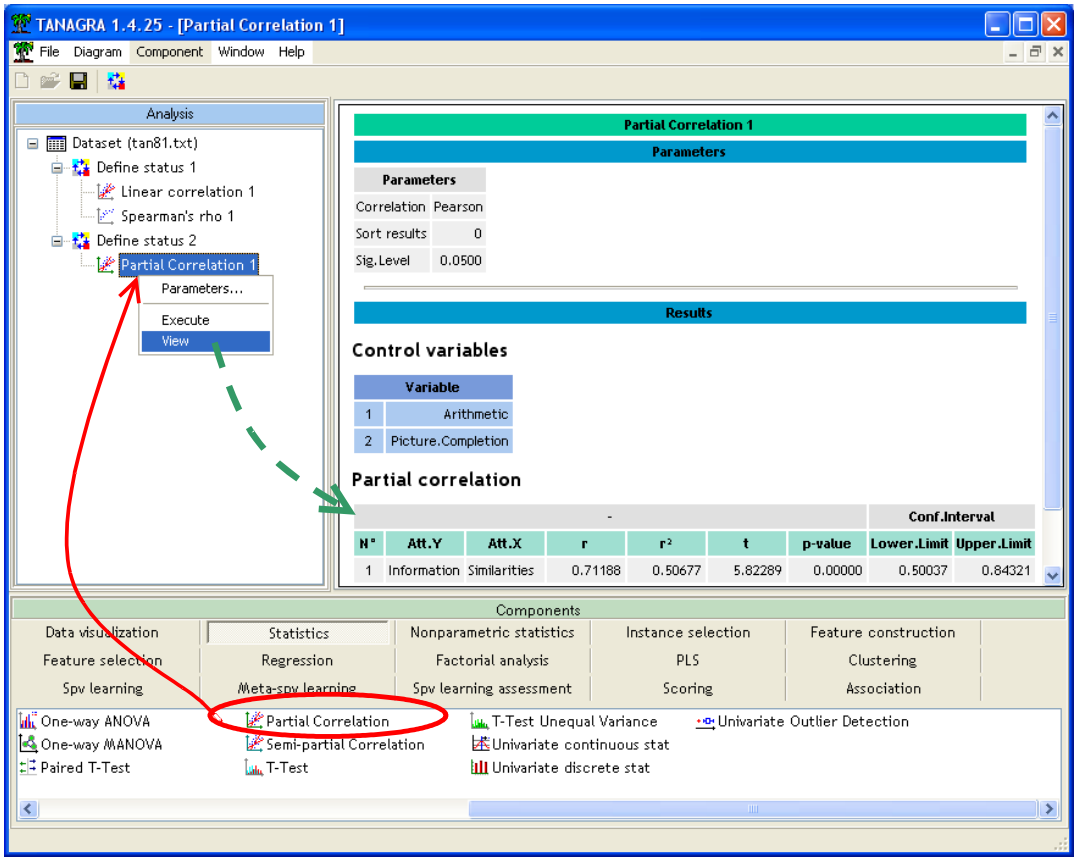
3.4 Corrélation partielle

L'idée de la corrélation partielle est d'évaluer dans quelle mesure la corrélation brute n'est pas influencée par les variables de contrôle. Dans notre cas, nous avons 2 variables de contrôle, ARITHMETIC et PICTURE.COMPLETION, on parle de corrélation partielle d'ordre 2.

Insérons de nouveau le composant DEFINE STATUS dans le diagramme. Nous plaçons INFORMATION en TARGET, SIMILARITIES en INPUT et, principale nouveauté dans cette section, le couple ARITHMETIC – PICTURE.COMPLETION en ILLUSTRATIVE.



Nous introduisons le composant PARTIAL CORRELATION (onglet STATISTICS). Nous cliquons sur le menu VIEW.



La corrélation partielle est r.part = 0.71188. Elle est significative également avec une p-value < 0.0001. Notons que les degrés de liberté sont (37 – 2 – 2 = 33) maintenant.

En nous appuyant sur la transformation de Fisher, nous pouvons déduire l'intervalle de confiance : il y a 95% de chances que l'intervalle [0.50037 ; 0.84321] recouvre la « vraie » valeur du coefficient.

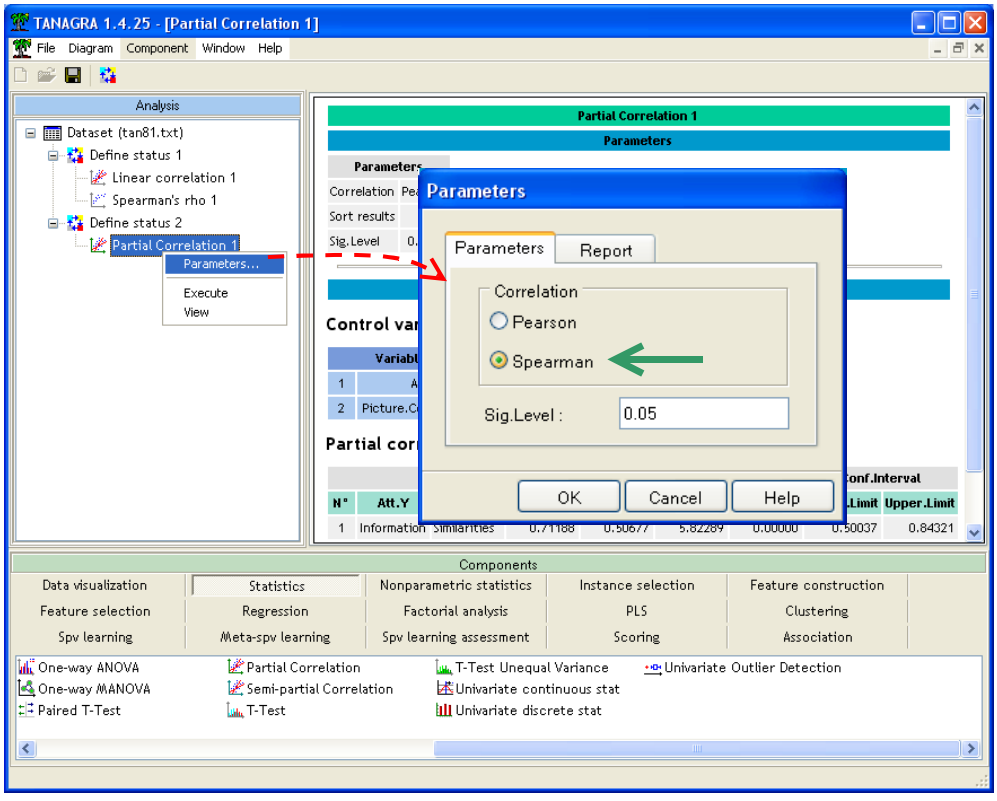
On notera la similitude du coefficient partiel (0.71188) avec le coefficient brut (0.7715). Cela indique que dans notre exemple, la corrélation entre INFORMATION et SIMILARITIES n'est pas due à l'influence conjointe des variables de contrôle.

Note : Le détail de la mise en place du test et le calcul de l'intervalle de confiance sont décrits sur le site http://www.stat.psu.edu/online/development/stat505/07_partcor/06_partcor_partial.html

3.5 Corrélation de Spearman partiel

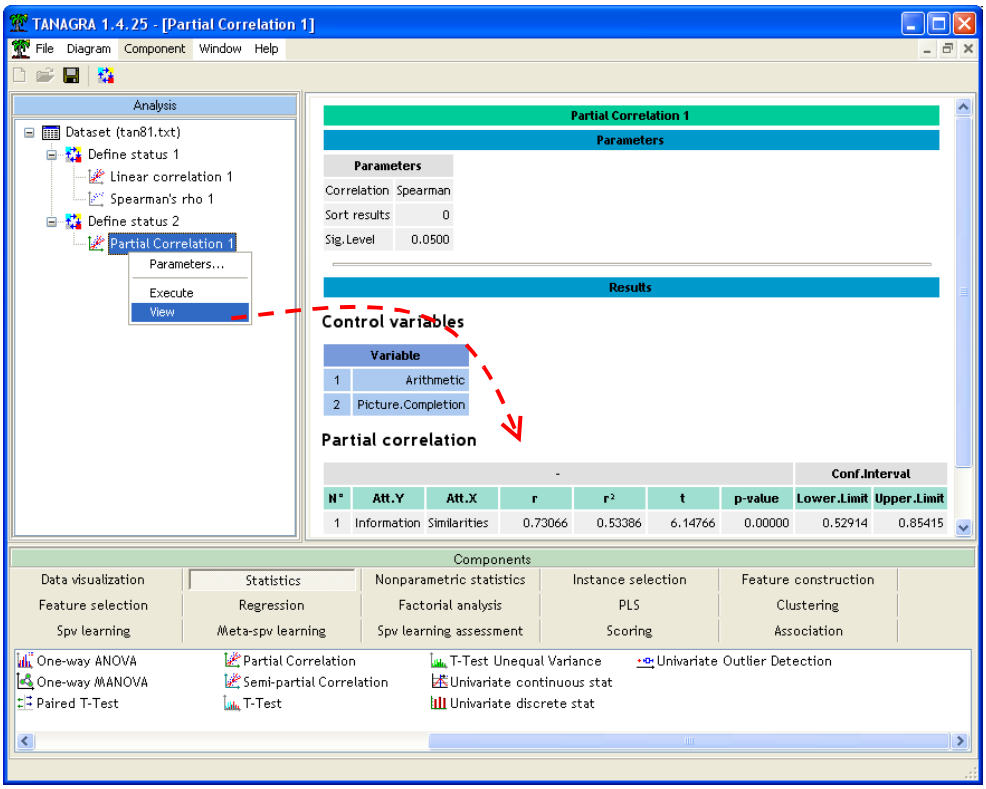
De la même manière que précédemment, il est possible de baser le calcul de la corrélation partielle sur les rangs, nous obtenons ainsi le coefficient de Spearman partiel. Mise à part la transformation des données en rangs, la procédure est exactement la même[7].

Dans notre cas, il s'agit de re-paramétrer simplement le composant. Nous activons le menu contextuel PARAMETERS. Nous choisissons l'option SPEARMAN puis nous validons. Un clic sur VIEW fait apparaître les nouveaux résultats.



Les conclusions émises dans le paragraphe précédent ne sont pas remises en cause. La relation entre INFORMATION et SIMILARITIES n'est pas imputable aux variables de contrôle, le rho de Spearman partiel est 0.73066.

[7] http://support.sas.com/documentation/cdl/en/procstat/59629/HTML/default/procstat_corr_sect017.htm



4 Conclusion

L'analyse des corrélations partielles est complexe. Les techniques numériques sont là pour nous guider, pour valider les intuitions. Mais il est évident que sans connaissances du domaine étudié, nous aurions vite fait de tourner en rond. Un excellent site décrit les différentes interactions qu'il peut y avoir entre les variables, nous reproduisons volontiers le schéma qui les résume (voir http://www2.chass.ncsu.edu/garson/PA765/partialr.htm)