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 TANAGRA¹.

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 <u>http://tutoriels-data-mining.blogspot.fr/2010/08/ladd-in-tanagra-pour-excel-2007-et-2010.html</u> pour la procédure d'installation et d'activation sous Excel 2007 et 2010 ; et <u>http://tutoriels-data-mining.blogspot.fr/2008/03/importation-fichier-xls-excel-macro.html</u> pour les versions antérieures.

N 🔊	Microsoft Excel - body.xls										
	🐵 Eichier Edition Affichage Insertion Format Outlis Données Fenêtre ? Tanagra Sipina 📃 🗗 🗶										
	□ ☞ 🖩 帚 函 ♡ 🗼 🛍 砲 ダ い・マ・ 🌯 Σ 糸 🆓 싌 梨 🏭 🚜 100% ・ 🛛 - 🗴 🛱 田・参・ 🔅										
,	Al 👻	= female									
	A	В	С	D	E	F	G	Н			
1	shoulder.girth	chest.girth	waist.girth	navel.girth	hip.girth	thigh.girth	bicep.girth	forearm.girth	kn		
2	106.2	89.5	71.5	74.5	93.5	51.5	32.5	26	j i		
3	110.5	97	79	86.5	94.8	51.5	34.4	28			
4	115.1	97.5	83.2	82.9	95	57.3	33.4	28.8	3		
5	104.5	97	77.8	78.8	94	53	31	26.2	2		
6	107.5	97.5	80	82.5	98.5	55.4	32	28.4	1		
7	119.8	99.9	82.5	80.1	95.3	57.5	33	28	}		
8	123.5	106.9	82	84	101	60.9	42.4	32.3	1		
9	120.4	102.5	76.8	80.5	98	56	34.1	28			
10	111						33	26			
11	119.5	Execu	te ranagra				36.5	29.2			
12	117.1						34.0	27.9			
14	123.5		Dataset range (incl	uding the name of t	he attributes first	row):	30.5	30.4			
15	110.5		\$A\$1:\$O\$508				35.5	29			
16	107.5		1				36	29			
17	107.5				OF 1	Cancal	34.5	29.5			
18	112.2				OK	Cancer	37	28.8			
19	120						31.2	26.9			
20	109	98.5	78 5	86	94 5	55	34.5	28.5			
21	118.5	104	87.3	88	101.1	59.5	37	30.5			
22	116	100	92	91	98	57.5	32	27.6	5		
23	111	100	80	83.7	99.5	57	37	30)		
24	117.7	99	74.5	75.9	92.2	53.4	31.2	26.9	,		
25	123.9	101	90.6	89.6	101.2	59.5	37	28.3	3		
26	120.6	101.6	81.4	81.6	98.8	61.3	39.4	31.9			
27	129.5	108.8	89.5	89.5	106	59.5	37.5	30.1			
28	115	100	85	94.5	105	62	35.5	28.5	i i		
29	116	88	73.5	77.7	97	56.3	32.5	27.8	3		
30	107.8	88.7	75.8	83	89	52.6	31.2	26.5	i i		
31	100.2	84.5	74	81	93.5	50.5	27.5	24.8	· · · ·		
1	▶ N body (description /			[]	1					
Des	Dessin + 🔓 🕜 Formes automatiques + 🔪 🍡 🗆 🔿 🔛 🧟 🐗 💆 + 🚣 + 🚍 🚃 🛱 🛄 \iint +										
Poir	iter					Gomme=469298.	8	NUM			

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.

🛣 TANAGRA 1.4.16 - [Da	itaset (tan69.txt)]					
Tile Diagram Componen	t Window Help				-	đΧ
🗅 📽 🖪 🎇						
	Analysis			Dataset (tap)	1 4 v 4 V	
				Dataset (tarios	(
			Database : C:\DC	CLIME~1\Maison\LOCALS~1\	> Temn\tan69.txt	•
				Results		
			Download Datasource pro Computation tim Allocated memor Dataset du 15 attribute(s) 507 example(s) Attribute shoulder, girth C chest.girth C	information cessing e 0 ms y 45 KB escription Category Informations continue - continue -		
μ			Componente			
Data visualization	Statistics	Nonpara	metric statistics	Instance selection	Feature construction	
Feature selection	Regression	Fact	orial analysis	PLS	Clustering	
Spv learning	Meta-spv learning	Spv lear	ning assessment	Scoring	Association	
↔ Correlation scatterplot	Export dataset	🚀 So	atterplot	📈 Scatterplot wit	:h label 🛛 🧱 View dataset	
<						>
						.::

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 o/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.

🐨 TANAGRA 1.4.16 - [So	atterplot 1]						
File Diagram Componen	t Window Help						_ 8 ×
Analysi Analysi Analysi Analysi Analysi Analysi Define status 1 Analysi Analysi Define status 2 Analysi Analysi Analysi Define status 2 Analysi Analysi Define status 2 Analysi Analysi Analysi Define status 1 Analysi Analysi Define status 1 Analysi Analysi Define status 1 Analysi Analysi Define status 1 Analysi An	s Imponent Analysis 1 plot 1	Define attribute statuses		Target ender Cle OK	Input Illus	trative Help	
Data visualization	Statistics	Nonparametric statistics	Instance sele	ction	Feature cor	nstruction	T
Feature selection	Regression	Factorial analysis	PLS		Cluste	ering	
Spv learning	Meta-spv learning	Spv learning assessment	Scoring		Associ	ation	
Correlation scatterplot	k Scatterplot k Scatterplot with label	View dataset ≟: View multiple scatter	rplot				

Nous insérons à la suite le composant o_1_BINARIZE (onglet FEATURE CONSTRUCTION). Nous actions le menu VIEW.

TANAGRA 1.4.16 - [0]	_1_Binarize 1] t Window Help								
) 📽 🔲 😘									
Analysi	s		0_1_Binarize 1						
Dataset (tan69.txt)		Parameters							
B Venne status I	use Use	Used values : K-1 (ignore last value)							
L 🧖 Scatter	plot 1	Pecults							
🖃 🎎 Define status 2		Attribute binarization							
Source att New attributes									
gender (gender_male_1)									
	_	4	<u></u>		-				
	Com	Computation time : 0 ms.							
	Crea	ated at 11/03/2007 18:43:29							
		Components							
Data visualization	Statistics	Nonparametric statistics	Instance selection	Feature construction					
Feature selection	Regression	Factorial analysis	PLS	Clustering					
Spv learning	Meta-spv learning	Spv learning assessment	Scoring	Association					
→∥0_1_BinarizeEqI	Freq Disc 🛛 🕍 Formula	🙊 RBF 🛛 🖏	Rnd Proj 🗌 📿 Trend						
🔥 Binary binning 🛛 🕂 Eq	Width Disc 👬 MDLPC	🕀 Residual Scores 👘	Standardize						

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)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.

💯 TANAGRA 1.4.16 - [Re	TANAGRA 1.4.16 - [Residual Scores 1]												
💇 File Diagram Component	t Window Help							- 8 ×					
	Analysis mponent Analysis 1 plot 1 e 1 tatus 3 dual Scores 1 Jefine status 4 Principal Component Analy Lefi: Scatterplot 2	Regression coeff Target shoulder.girth chest.girth waist.girth hip.girth thigh.girth bicep.girth forearm.girth knee.girth calf.girth ankle.girth	icients and Expt. Var. (%) 61.0 55.5 44.8 4.3 2.5 0.6 55.2 62.7 13.7 15.0 27.5 56.4 (%)	Explained Vari gender_male_1 16.1978 14.9299 14.7297 3.9166 2.1105 -0.6978 6.3063 4.4801 1.9355 2.2007 1.9533 2.4344	Constant 100.3038 86.0600 69.8035 83.7458 95.6527 57.1958 28.0973 23.7604 35.2600 35.2000 35.0062 21.0582								
				57.0	2.1311	15.0572							
			Componente										
Data visualization	Statistics	Nonpa	rametric statistics	Instance	e selection	Feature	construction						
Feature selection	Regression	Fa	ctorial analysis	ſ	PLS	, Clu:	stering						
Spv learning	Meta-spv learning	Spv le:	arning assessment	Sc	oring	Asso	ciation						
┝║0_1_Binarize ⚠️EqF ⚠️Binary binning ⚠️EqV	Freq Disc 🔟 Formula Width Disc 🏦 MDLPC	x	RBF 😽	Rnd Proj Standardize	📿 Trend								

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.

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 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 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.

💯 TANAGRA 1.4.16 - [Pr	rincipal Componer	nt Analysis I	2]						
Tile Diagram Component	t Window Help								- 8 ×
D 📽 🖪 🎎									
	Analysis			atrix trace - 1	2.00				^
😑 🏢 Dataset (tan69.txt)									
🖃 🙀 Define status 1			A	xis Eigen value	% explained	He	stogram	% cumut	ned
😑 🔀 Principal Co	omponent Analysis 1			2.889848	24.08%			24.	.08%
🔤 📈 Scatterp	plot 1			2 2.473299	20.61%			44.	.69%
🖃 🏠 Define status 2				1 592840	13 27%			57	97%
⊡ → 0_1_Binarize	e 1				1012770	_		57.	
🖃 🎲 Define st	tatus 3			4 1.336768	11.14%			69.	.11%
	dual scores 1			5 0.816439	6.80%			75.	.91%
	Principal Compo	nent Analysis	2	6 0.599942	5.00%			80.	.91%
	Scatterplot	Parameters.		7 0 549412	4 50%	-		0.0	49%
		Execute		0.547415	4,50%	-		05.	47/0
		View		B 0.470391	3.92%			89.	.41%
	L			9 0.400651	3.34%			92.	75%
			1	0 0.359738	3.00%			95.	.74%
						-			~
						J			>
				Componer	ıts				
Data visualization	Statistics	N	lonparame	tric statistics	Instanc	e selection	Feature o	construction	
Feature selection	Regression	ו ו	Factoria	al analysis		PLS	Clu	stering	
Spv learning	Meta-spv lear	ning S	opv learning	g assessment	So	oring	Asso	ociation	
🕁 Bartlett's test	🔆 Brown - For	rsythe's test	🚠 Fist	ner's test		Group characte	rization	🚮 Group explora	tion 🕁
<									>

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.

🕅 TANAGRA 1.4.16 - [Principal Component Analysis 2]												
🝸 File Diagram Component Window Help 🔤 🖉												
Analysis	Attribute	Axis	5_1	Axi	s_2	Axis_3		Axis	;_4	^		
🖃 🏢 Dataset (tan69.txt)	-	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)	Corr.	% (Tot. %)			
🖃 🚰 Define status 1	rs_shoulder.girth_1	0.4238	18 % (18 %)	-0.5335	28 % (46 %)	-0.0846	1 % (47 %)	-0.0444	0 % (47 %)			
🖶 🔀 Principal Component Analysis 1	rs_chest.girth_1	0.1923	4 % (4 %)	-0.7704	59 % (63 %)	-0.2299	5 % (68 %)	0.0551	0 % (69 %)			
⊡– <mark>∦</mark> Scatterplot 1 ⊒– 🎦 Define status 2	rs_waist.girth_1	-0.5880	35 % (35 %)	-0.2935	9 % (43 %)	-0.3814	15 % (58 %)	0.3885	15 % (73 %)			
□ • 0_1_Binarize 1 rs_navel.girth_1 -0.4414 19 % (19 %) -0.1693 3 % (22 %) -0.7222 52 % (75 %) 0.1897 4 % (78 %)												
□ 🚰 Define status 3 rs_hip.girth_1 -0.2239 5 % (5 %) 0.3389 11 % (16 %) -0.7329 54 % (70 %) -0.2924 9 % (79 %)												
Residual Scores 1	rs_thigh.girth_1	0.1893	4 % (4 %)	0.4864	24 % (27 %)	-0.3312	11 %(38 %)	-0.6658	44 % (83 %)			
□ t ₄ Denne status 4	rs_bicep.girth_1	0.6269	39 % (39 %)	-0.4469	20 % (59 %)	-0.2704	7 % (67 %)	-0.3066	9 % (76 %)			
Scatterplot 2	rs_forearm.girth_1	0.8063	65 % (65 %)	-0.2362	6 % (71 %)	-0.1261	2 % (72 %)	-0.1021	1 % (73 %)			
	rs_knee.girth_1	0.3231	10 % (10 %)	0.5993	36 % (46 %)	-0.2080	4 % (51 %)	0.2203	5 % (56 %)			
	rs_calf.girth_1	0.4691	22 %(22 %)	0.6025	36 % (58 %)	0.0011	0 % (58 %)	0.2027	4 % (62 %)			
	rs_ankle.girth_1	0.4405	19 % (19 %)	0.3786	14 % (34 %)	-0.2357	6 % (39 %)	0.5403	29 % (68 %)			
	rs_wrist.girth_1	0.6966	49 % (49 %)	-0.0886	1 % (49 %)	-0.1761	3 % (52 %)	0.3606	13 % (65 %)			
	Var. Expl.	2.8898	24 % (24 %)	2,4733	21 % (45 %)	1.5928	13 %(58 %)	1.3368	11 % (69 %)	~		
		Ш								>		
		Compone	ents									
Data visualization Statistics	Nonparametric statistics	s Ins	tance select	ion	Feature co	nstruction	Feat	ure selectio	n			
Regression Factorial analysis	PLS		Clustering		Spv lea	arning	Meta	-spv learnin	g			
Spv learning assessment Scoring	Association											
🛱 Bartlett's test 🛛 🕌 Brown - Forsythe'	's test 🛛 🖾 Fisher's test		🚺 Group d	characteriz	ation 👪	Group expl	oration	∰Leve	ene's test			
<										>		

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

Nous insérons donc le composant VARHCA³ 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.



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
<pre>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)</pre>	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

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

rs_waist.girth_1 (taille)	Entre la poitrine et les hanches
rs_navel.girth_1 (nombril)	

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 individus⁴.

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.

TANAGRA 1.4.16 - [VARHCA 1] Tile Diagram Component Window Help E. 🔲 🔛 pril. Dendrogram HTML Analysis 🖃 🏢 Dataset (tan69.txt) Hierarchical Variable Clustering Analysis -- Dendrogram 🚊 🚰 Define status 1 🗄 🔀 Principal Component Analysis 1 🧟 Scatterplot 1 -🔁 Define status 2 🛓 👬 Define status 3 🛓 🕀 Residual Scores 1 Distance Cluster Combine 🚊 🔂 Define status 4 🛓 🔀 Principal Component Analysis 2 🧭 Scatterplot 2 TH VARHCA 1 Component Nonparametric statistics Data visualization Statistics Instance selection Feature construction Feature selection Regression Factorial analysis PLS Clustering Association Spv learning Meta-spy learning Spy learning assessment Scoring A, CT 🔯 VARCLUS 🙀 Kohonen-SOA Selection **T**VARHCA 🚓 СТР **船**HAC No. EM-Clustering 🔣 Neighborhood Graph 🛛 🛃 VARKMeans 🔥 K-Means

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^{eme} et 4^{eme} 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 !!!



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 outils⁵ 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.

⁵ <u>http://tutoriels-data-mining.blogspot.fr/2012/06/acp-avec-tanagra-nouveaux-outils.html</u> 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 VARIMAX⁶.

TANAGRA 1.4.45 - [Factor Tile Diagram Compo	rotation 1] nent Window Help				U y	-		18	9			_ O Σ
Default title												
🕞 🎹 Dataset (body_sas.x	ls)						Parameter					
😑 🚰 Define status 1			For the sector	N			Paramete	#FS				
🖃 🔅 Residual Sco	res 1		Factors rota	tion								
🖃 🎇 Define sta	atus 2		Method VA	(IMAX								
Component Analysis 1 A parallel Analysis 1												
→ Parallel Analysis 1 Results												
Factor rotation 1												
Rotated Factor Loadings												
Attribute Axis_1 Axis_2 Axis_3 Axis_4												
- <u>Corr.</u> % (Tot. %) Corr. % (Tot. %) Corr. % (Tot. %)												
- Corr. % (1ot. %) Corr. % (1ot. %) Corr. % (Tot. %) Corr. % (Tot. %) Corr. % (Tot. %) rs_shoulder.girth_1 0.6674 45 % (45 %) -0.0866 1 % (45 %) 0.0238 0 % (45 %) -0.1407 2 % (47 %)												
			rs che	st.girth 1	0.6903	48 % (48 %)	-0.2521	6 % (54 %)	-0.2965	9 % (63 %)	-0.2418	6 % (69 %)
			rs waist.girt		-0.1533	2 % (2 %)	-0.1622	3 % (5 %)	-0.7963	63 % (68 %)	-0.2109	4 % (73 %)
			rs_navel.girth_1		-0.0032	0%(0%)	-0.0435	0 % (0 %)	-0.8646	75 % (75 %)	0.1781	3 % (78 %)
			rs_hip.girth_1		-0.1143	1%(1%)	0.0889	1 % (2 %)	-0.4447	20 % (22 %)	0.7543	57 % (79 %)
			rs_thig	h.girth_1	0.0035	0 % (0 %)	0.0784	1%(1%)	0.2373	6 % (6 %)	0.8735	76 % (83 %)
			rs_bice	p.girth_1	0.8420	71 % (71 %)	-0.0343	0 % (71 %)	0.1122	1 % (72 %)	0.1924	4 % (76 %)
			rs_forea	rm.girth_1	0.7616	58 % (58 %)	0.2634	7 % (65 %)	0.2799	8 % (73 %)	0.0660	0 % (73 %)
			rs_kne	e.girth_1	-0.1258	2 % (2 %)	0.6927	48 % (50 %)	0.0738	1 % (50 %)	0.2329	5 % (56 %)
			rs_cal	f.girth_1	-0.0876	1%(1%)	0.7101	50 % (51 %)	0.3064	9 % (61 %)	0.1356	2 % (62 %)
			rs_ank	le.girth_1	0.0651	0% (0%)	0.8188	67 % (67 %)	-0.0724	1 % (68 %)	-0.0704	0 % (68 %)
			rs_wri	st.girth_1	0.5473	30 % (30 %)	0.5682	32 % (62 %)	0.0459	0 % (62 %)	-0.1721	3 % (65 %)
			Var	. Expl.	2.5749	21 % (21 %)	2.1612	18 % (39 %)	1.9217	16 % (55 %)	1.6350	14 % (69 %)
<u> </u>												
Data davaliantian	Charlinting	Nerra		Com	ponents				Frates			
Data visualization	Statistics	Nonpara	metric statistics	instand	e selection	reat	ure constru	cuon	reature	selection		
Regression	Factorial analysis		PLS	Clu	istering		Spv learnin	g	Meta-spv	learning		
Spv learning assessment	Scoring	As	ssociation									
Bootstrap Eigenvalues	+ Factor rotation	on .	<u>}</u> ₽	arallel Analysis								
Canonical Discriminant A	Anatysis 🔛 Multiple Corr	espondenc	e Analysis 🛛 🕅 F	rincipal Compo	nent Analysis							
	> KINALS											

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.



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.

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



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

Et les COS² des variables sur les 4 composantes sélectionnées (cf. page 12 pour Tanagra).

	Final Communality Estimates: Total = 8.292756											
shoulder_girth	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	0.47331904 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-dessus⁷. 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))
}</pre>
```

⁷ 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 <- function(variables, moderators) {</pre>
 new.scores <- get residuals(variables,moderators)</pre>
 pca.object <- princomp(new.scores,cor=T)</pre>
 return(pca.object)
#**** END FUNCTIONS ****#
#charger les données
library(xlsx)
body.data <- read.xlsx(file="body sas.xls",sheetIndex=1,header=T)</pre>
#subdivision : variables actives, variables modératrices
body.variables <- subset(body.data, select=shoulder.girth:wrist.girth)</pre>
body.moderators <- subset(body.data, select=weight:gender)</pre>
#appel de la fonction
body.pca <- pca partial(body.variables,body.moderators)</pre>
#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.