INFERENCE SUR LES VALEURS PROPRES ET AUTRES INDICES EN ACP , AFC ET ACM

Gilbert Saporta

Conservatoire National des Arts et Métiers,Paris

gilbert.saporta@cnam.fr

Introduction

- Combien d'axes en ACP, AFC, ACM?
- Qu'est-ce qu'une grande contribution?
- Comment détecter les « outliers »?

La littérature fournit des tests, basés souvent sur des hypothèses irréalistes ou restrictives.

Les praticiens se servent de règles empiriques

Cet exposé:

- Un panorama des critères les plus importants
- Quelques nouvelles règles fondées sur des hypothèses faibles, ou utilisant une approche de type « carte de contrôle »

1.Composantes principales

- Approche « multivariate analysis »: hypothèses de lois. Intérêt pour les relations entre variables, individus anonymes et interchangeables.

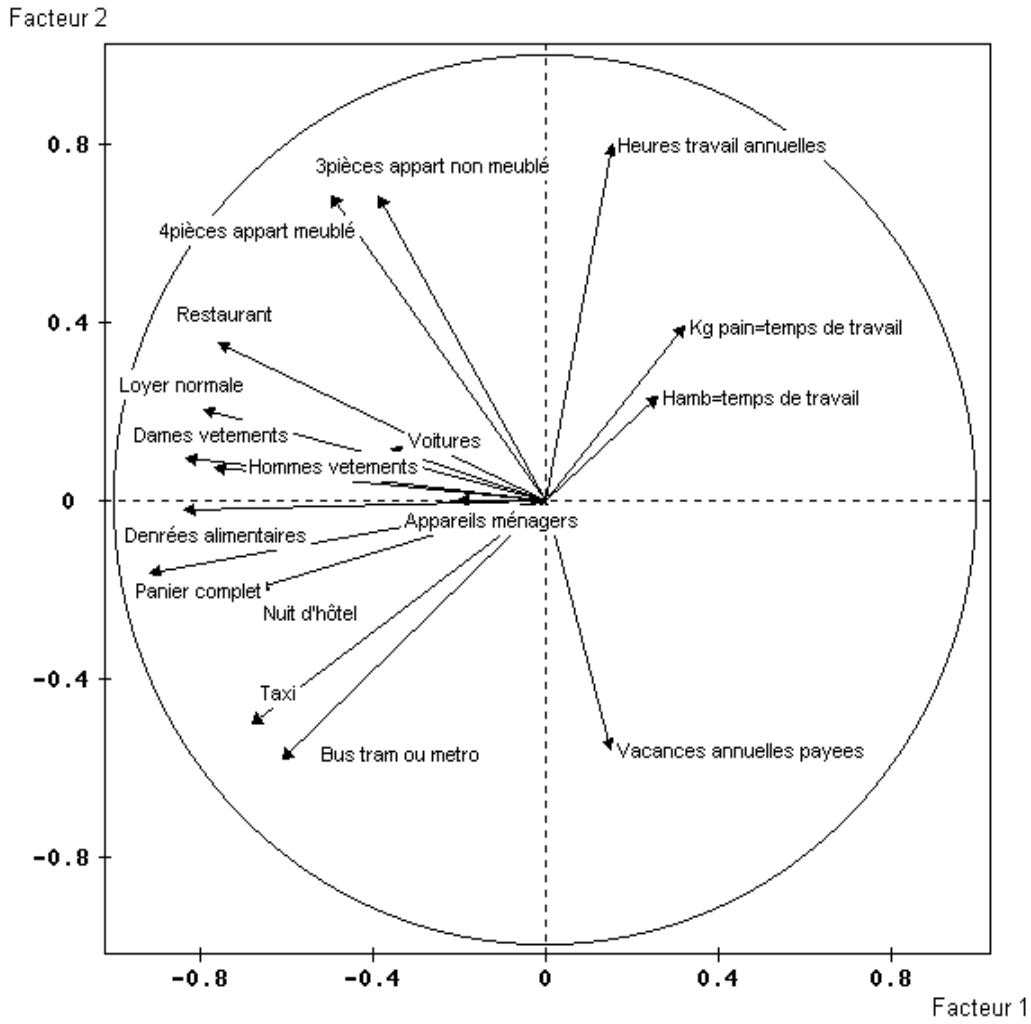
- Analyse exploratoire ou «analyse des données » réduction de dimension, projection des individus et des variables pour révéler des structures dans les données

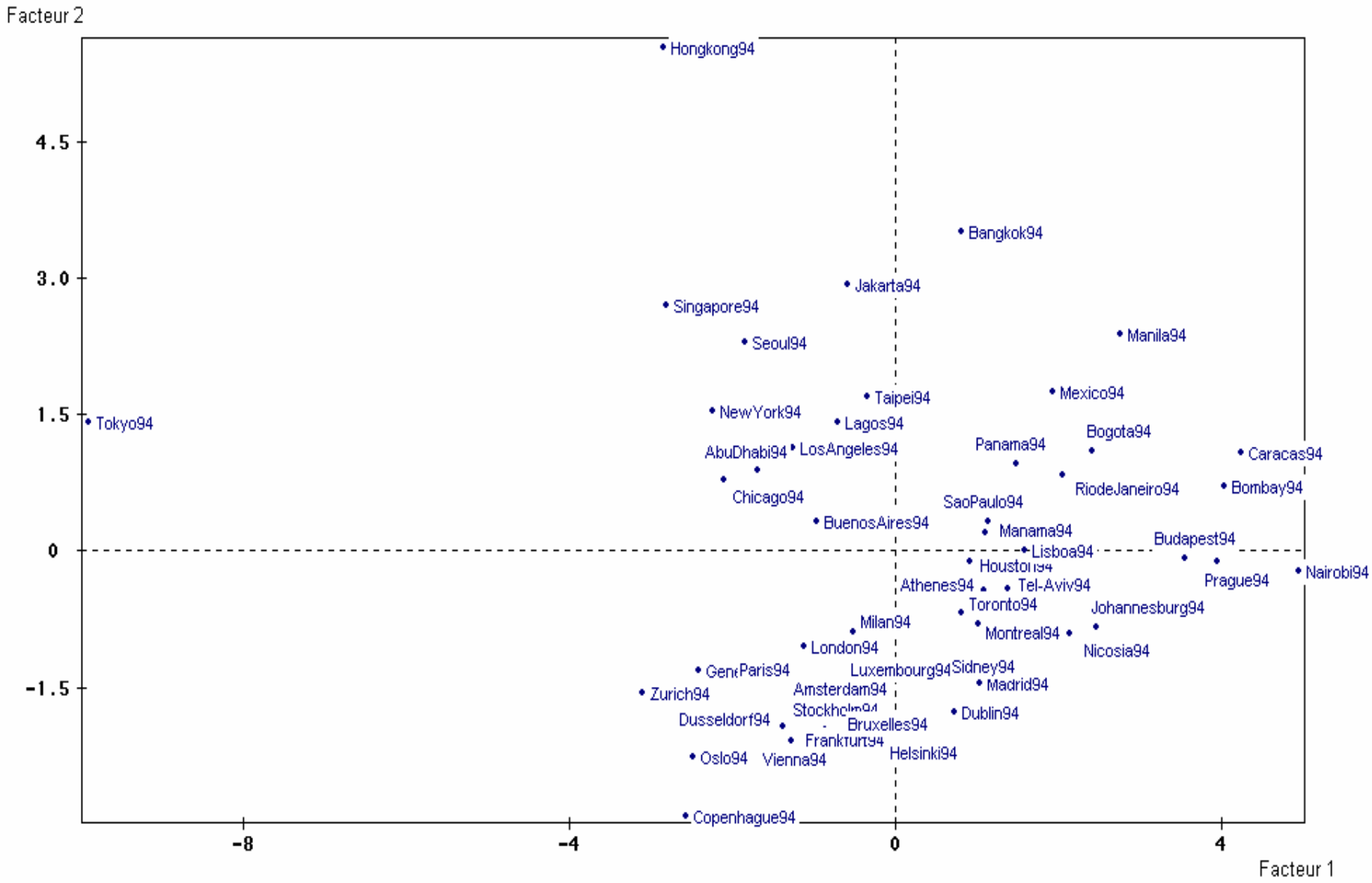
Un exemple: niveaux de vie dans 51 villes, 17 variables économiques

source : Union de Banques Suisses “ Prices and Earnings around the Globe ”, repris dans A.Morineau, T.Aluja-Banet “ Analyse en Composantes principales ”, Cisia.Ceresta, 1998.

ANALYSE EN COMPOSANTES PRINCIPALES
STATISTIQUES SOMMAIRES DES VARIABLES CONTINUES
EFFECTIF TOTAL : 51 POIDS TOTAL : 51.00

IDEN - LIBELLE	EFFECTIF	POIDS	MOYENNE	ECART-TYPE	MINIMUM	MAXIMUM
HEUR - Heures travail annuelles	51	51.00	1920.25	158.69	1669.00	2302.00
VACA - Vacances annuelles en jours	51	51.00	21.21	7.32	7.80	39.30
TPAI - Kg pain=temps de travail	51	51.00	23.29	21.56	6.00	117.00
THAM - Hamb=temps de travail	51	51.00	66.71	97.82	14.00	683.00
ALIM - Denrées alimentaires	51	51.00	347.43	141.31	112.00	938.00
PANI - Panier complet	51	51.00	1321.04	380.29	591.00	2569.00
DAME - Dames vetements	51	51.00	420.20	275.86	50.00	1910.00
HOMM - Hommes vetements	51	51.00	726.08	340.51	160.00	1800.00
4PIE - 4pièces appart meublé	51	51.00	1863.53	1319.33	500.00	8390.00
3PIE - 3pièces appart non meublé	51	51.00	1089.80	731.43	240.00	4260.00
RENT - Loyer normal	51	51.00	585.10	288.76	40.00	1440.00
MENA - Appareils ménagers	51	51.00	2280.20	619.62	1370.00	4250.00
BUS, - Bus tram ou metro	50	50.00	0.98	0.69	0.10	3.04
TAXI - Taxi	51	51.00	5.95	3.45	0.45	14.45
VOIT - Voitures	51	51.00	17758.82	7519.39	6100.00	54800.00
REST - Restaurant	51	51.00	27.80	14.24	8.00	76.00
HOTE - Nuit d'hôtel	51	51.00	226.78	60.95	97.00	382.00


Facteur 2
0.8
0.4
0
-0.4
-0.8
-0.8
-0.4
0
0.4
0.8
Facteur 1
3pièces appart non meublé
Heures travail annuelles
4pièces appart meublé
Restaurant
Kg pain=temps de travail
Loyer normale
Hamb=temps de travail
Dames vetements
Voitures
Hommes vetements
Appareils ménagers
Denrées alimentaires
Panier complet
Nuit d'hôtel
Taxi
Bus tram ou metro
Vacances annuelles payees


Facteur 2
Facteur 1
4.5
3.0
1.5
0
-1.5
-8
-4
0
4
Hongkong94
Bangkok94
Jakarta94
Singapore94
Manila94
Seoul94
Taipei94
Mexico94
Tokyo94
NewYork94
Lagos94
Bogota94
Panama94
Caracas94
AbuDhabi94
LosAngeles94
RiodeJaneiro94
Bombay94
Chicago94
SaoPaulo94
BuenosAires94
Manama94
Budapest94
Lisboa94
Houston94
Nairobi94
Prague94
Athenes94
Tel-Aviv94
Toronto94
Johannesburg94
Milan94
Montreal94
London94
Nicosia94
Paris94
Luxembourg94
Sidney94
Zurich94
Amsterdam94
Madrid94
Dusseldorf94
Dublin94
Bruxelles94
Oslo94
Vienna94
Helsinki94
Copenhague94

1.1 Valeurs propres

- Hypothèses intéressantes

pour données centrées-réduites: $\lambda_i \neq 0$ or $\lambda_i > 1$?

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	6.0506	35.59	35.59	**********
2	2.9257	17.21	52.80	*****
3	2.4905	14.65	67.45	****
4	1.2691	7.47	74.92	**
5	0.9784	5.76	80.67	*
6	0.6530	3.84	84.51	*
7	0.6170	3.63	88.14	*
8	0.4878	2.87	91.01	*
9	0.4215	2.48	93.49	*
10	0.3269	1.92	95.41	*
11	0.2257	1.33	96.74	*
12	0.1843	1.08	97.83	*
13	0.1574	0.93	98.75	*
14	0.0955	0.56	99.31	*
15	0.0555	0.33	99.64	*
16	0.0481	0.28	99.92	*
17	0.0131	0.08	100.00	*

- test de sphéricité:

$$(n-\frac{2p+11}{6})(p-k)\ln(\frac{a}{g}) \approx \chi^2_{(p-k+2)(p-k+1)/2}$$

- a et g moyennes arithmétiques et géométriques des dernières p-k valeurs propres
- distributions asymptotiques d'Anderson

$$\sqrt{n-1}(\hat{\lambda}_i - \lambda_i) \approx N(0;\lambda_i\sqrt{2})$$

- La méthode « delta »
  - Soit T tel que $T \to N\left(\theta; \frac{\sigma(\theta)}{\sqrt{n}}\right)$

    alors $g(T) \to N\left(g(\theta); \frac{g'(\theta)\sigma(\theta)}{\sqrt{n}}\right)$

- Application aux valeurs propres:

$$\ln\left(\hat{\lambda}_i\right) \sim N\left(\ln(\lambda_i); \sqrt{\frac{2}{n-1}}\right)$$

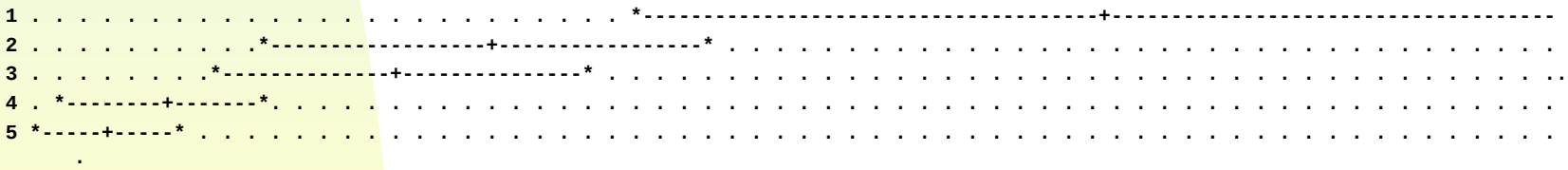
Intervalles de confiance d'Anderson

$$\hat{\lambda}_i \exp(-1.96\sqrt{\frac{2}{n-1}}) < \lambda < \hat{\lambda}_i \exp(1.96\sqrt{\frac{2}{n-1}})$$

INTERVALLES LAPLACIENS D'ANDERSON INTERVALLES AU SEUIL 0.95

NUMERO	BORNE INFERIEURE	VALEUR PROPRE	BORNE SUPERIEURE
1	3.6788	6.0506	8.4224
2	1.7788	2.9257	4.0726
3	1.5142	2.4905	3.4668
4	0.7716	1.2691	1.7667
5	0.5949	0.9784	1.3619

ETENDUE ET POSITION RELATIVE DES INTERVALLES



- Valables seulement pour des matrices de covariance, non robustes à la non-normalité
- Loi d'une matrice de covariance
  - **M** (p,p) suit une loi de Wishart $W_p(n,\Sigma)$ si **M=X'X** où **X** est une matrice de n observations indépendantes d'une $N_p(0;\Sigma)$
  - nV suit alors une $W_p(n-1,\Sigma)$
- Remarque: l 'hypothèse de normalité implique une notion d'homogénéité des données, en contradiction avec les objectifs de la statistique exploratoire.

Critères empiriques

- Critère de Kaiser $\hat{\lambda}_i > 1$

- Critère du « coude »: saut dans le diagramme de décroissance des valeurs propres

Exemple villes: Kaiser ⇨ 4 , coude ⇨ 3

Version analytique du coude: critère de Cattell

- Par analogie avec la recherche d'un point d'inflexion: changement de signe dans les différences secondes

$$\delta_1 = \hat{\lambda}_1 - \hat{\lambda}_2$$

$$\delta_2 = \hat{\lambda}_2 - \hat{\lambda}_3 \qquad \varepsilon_1 = \delta_1 - \delta_2$$

$$.... \qquad \varepsilon_2 = \delta_2 - \delta_3$$

- Problème: ne marche en général pas!

Le modèle à effets fixes (Besse et al.1988)

- Pour chaque observation :

$$\mathbf{x}_i = \mathbf{m}_i + \sigma\boldsymbol{\varepsilon}_i$$

$$E(\mathbf{x}_i) = \mathbf{m}_i \qquad \mathbf{V}(\mathbf{x}_i) = \frac{\sigma^2}{p_i}\boldsymbol{\Gamma}$$

- Hypothèse: $\mathbf{m}_i \in$ sous-espace de dimension k

$\hat{\mathbf{x}}_{\mathbf{i}}^{\mathbf{k}}$ projection sur les k premiers vecteurs de $\mathbf{VM} = \mathbf{V}\boldsymbol{\Gamma}^{-1}$

- Minimiser le critère d'ajustement:

$$f_k = E\left[\sum_{i=1}^{n} p_i \left\|\hat{\mathbf{x}}_i^k - \mathbf{x}_i\right\|^2_{\boldsymbol{\Gamma}^{-1}}\right]$$

Un nouveau critère:

- Moyenne et dispersion des valeurs propres

$$\overline{\hat{\lambda}} = 1 \qquad \sum_{i=1}^{p} \hat{\lambda}_i^2 = p + \sum\sum_{i \neq j} r_{ij}^2$$

- Pour un couple de variables indépendantes $E(R^2)=1/(n-1)$

$$E(\sum \hat{\lambda}_i^2) = p + \frac{p(p-1)}{n-1}$$

- Dispersion espérée des valeurs propres:

$$E(\frac{1}{p}\sum_{i=1}^{p}(\hat{\lambda}_i - 1)^2) = \frac{p-1}{n-1}$$

- Une borne inférieure:

$$\hat{\lambda}_i > 1 + 2\sqrt{\frac{p-1}{n-1}}$$

- Dans l'exemple, la borne est 2.13 d'où 3 axes sélectionnés.

- Mais aussi: avoir des axes interprétables
  - Peut se formaliser en testant les corrélations avec des variables supplémentaires

1.2 contributions à l'inertie

- Les composantes principales c peuvent souvent être approchées par une N(0; $\lambda^{1/2}$) si *p et n* sont grands
- D 'où: $\frac{c_{ik}^2}{\lambda_k}$ distribué comme χ_1^2
- La contribution $\frac{1}{n}\frac{c_{ik}^2}{\lambda_k}$ sera jugée significativement grande au seuil α = 0.05

  si elle dépasse 3.84/n

1.3 Distance au centre de gravité

- Pour des observations normalement distribuées le carré D^2 (DISTO) de la distance à *0* est une somme pondérée de *p* chi-deux indépendants:

$$\sum_{i=1}^{p} \lambda_i \chi^2_{i.1}$$

- $E(D^2)=p$ $V(D^2)= 2\sum_{i=1}^{p} \lambda_i^2$

- Les observations avec un D^2 plus grand que:

$$p + 2\sqrt{2\sum_{i=1}^{p} \lambda_i^2}$$

sont des outliers potentiels.

- Exemple villes:

Contribution significative si >7.7%
outliers potentiels si D^2>38

COORDONNEES, CONTRIBUTIONS ET COSINUS CARRES DES INDIVIDUS

INDIVIDUS			COORDONNEES					CONTRIBUTIONS				
IDENTIFICATEUR	P.REL	DISTO	1	2	3	4	5	1	2	3	4	5
AbuDhabi94	1.96	20.94	-1.68	0.88	-1.41	-1.89	-1.06	0.9	0.5	1.6	5.5	2.3
Amsterdam94	1.96	5.90	-0.66	-1.79	-0.61	0.32	-0.70	0.1	2.2	0.3	0.2	1.0
Athenes94	1.96	6.09	1.09	-0.43	-0.30	0.66	0.14	0.4	0.1	0.1	0.7	0.0
Bangkok94	1.96	27.29	0.81	3.50	2.31	1.47	1.46	0.2	8.2	4.2	3.4	4.3
Bogota94	1.96	9.02	2.42	1.10	0.06	-0.34	0.12	1.9	0.8	0.0	0.2	0.0
Bombay94	1.96	27.32	4.04	0.71	-1.29	1.17	-2.29	5.3	0.3	1.3	2.1	10.5
Bruxelles94	1.96	6.75	-0.98	-1.86	0.15	0.57	-0.18	0.3	2.3	0.0	0.5	0.1
Budapest94	1.96	13.51	3.55	-0.07	0.16	-0.34	0.22	4.1	0.0	0.0	0.2	0.1
BuenosAires94	1.96	16.71	-0.95	0.32	2.07	-0.29	0.48	0.3	0.1	3.4	0.1	0.5
Caracas94	1.96	38.31	4.24	1.07	2.32	-1.43	-0.99	5.8	0.8	4.3	3.2	1.9
Chicago94	1.96	13.76	-2.10	0.77	-1.35	-1.67	1.15	1.4	0.4	1.4	4.3	2.7
Copenhague94	1.96	20.96	-2.57	-2.93	0.57	0.63	1.11	2.1	5.7	0.3	0.6	2.5
Dublin94	1.96	6.77	0.73	-1.78	-0.69	0.43	1.00	0.2	2.1	0.4	0.3	2.0
Dusseldorf94	1.96	8.14	-1.38	-1.93	-0.55	0.99	-0.37	0.6	2.5	0.2	1.5	0.3
Frankfurt94	1.96	7.57	-0.85	-1.98	-0.43	1.08	-0.57	0.2	2.6	0.1	1.8	0.7
Geneve94	1.96	11.27	-2.41	-1.32	0.58	0.31	-0.48	1.9	1.2	0.3	0.1	0.5
Helsinki94	1.96	6.01	-0.21	-1.98	0.49	0.54	0.30	0.0	2.6	0.2	0.4	0.2
Hongkong94	1.96	56.68	-2.83	5.54	-2.87	1.23	-1.18	2.6	20.6	6.5	2.4	2.8

Houston94	1.96	9.82	0.92	-0.11	-0.70	-2.45	1.06	0.3	0.0	0.4	9.3	2.2
Jakarta94	1.96	23.85	-0.58	2.92	-0.89	2.40	-0.34	0.1	5.7	0.6	8.9	0.2
Johannesburg94	1.96	10.20	2.47	-0.83	-0.30	-0.75	0.59	2.0	0.5	0.1	0.9	0.7
Lagos94	1.96	87.24	-0.70	1.41	8.92	0.15	-1.26	0.2	1.3	62.7	0.0	3.2
Lisboa94	1.96	4.38	1.59	0.00	-0.70	0.45	0.35	0.8	0.0	0.4	0.3	0.2
London94	1.96	10.36	-1.12	-1.06	-1.05	0.94	-0.58	0.4	0.8	0.9	1.4	0.7
LosAngeles94	1.96	9.33	-1.26	1.13	-1.22	-0.97	0.96	0.5	0.9	1.2	1.5	1.8
Luxembourg94	1.96	7.82	-0.62	-1.44	-0.82	0.54	-0.51	0.1	1.4	0.5	0.5	0.5
Madrid94	1.96	6.14	1.04	-1.45	-0.57	0.69	-0.82	0.4	1.4	0.3	0.7	1.3
Manama94	1.96	6.81	1.11	0.19	-0.46	-0.40	-0.74	0.4	0.0	0.2	0.2	1.1
Manila94	1.96	15.92	2.77	2.38	0.50	-0.21	0.99	2.5	3.8	0.2	0.1	2.0
Mexico94	1.96	8.68	1.93	1.74	-0.46	-0.97	-0.22	1.2	2.0	0.2	1.5	0.1
Milan94	1.96	4.45	-0.51	-0.89	-0.48	0.17	-0.98	0.1	0.5	0.2	0.0	1.9
Montreal94	1.96	5.81	1.02	-0.81	-0.45	-1.38	1.00	0.3	0.4	0.2	2.9	2.0
Nairobi94	1.96	26.22	4.95	-0.22	0.79	0.06	-0.05	7.9	0.0	0.5	0.0	0.0
NewYork94	1.96	13.55	-2.25	1.54	-1.35	-1.49	-0.22	1.6	1.6	1.4	3.4	0.1
Nicosia94	1.96	9.03	2.14	-0.91	0.27	0.12	0.82	1.5	0.6	0.1	0.0	1.3
Oslo94	1.96	15.50	-2.49	-2.26	1.06	0.17	1.34	2.0	3.4	0.9	0.0	3.6
Panama94	1.96	9.13	1.49	0.95	-0.77	0.43	-0.90	0.7	0.6	0.5	0.3	1.6
Paris94	1.96	8.42	-2.00	-1.29	0.47	0.41	-1.14	1.3	1.1	0.2	0.3	2.6
Prague94	1.96	19.39	3.96	-0.11	0.00	-0.67	-0.74	5.1	0.0	0.0	0.7	1.1
RiodeJaneiro94	1.96	7.17	2.06	0.83	-0.17	0.85	-0.83	1.4	0.5	0.0	1.1	1.4
SaoPaulo94	1.96	8.76	1.15	0.32	-1.06	1.38	-1.22	0.4	0.1	0.9	2.9	3.0
Seoul94	1.96	18.71	-1.84	2.29	0.30	-2.51	0.25	1.1	3.5	0.1	9.7	0.1
Singapore94	1.96	42.43	-2.81	2.70	-0.20	3.33	3.15	2.6	4.9	0.0	17.2	19.9
Stockholm94	1.96	7.47	-0.86	-1.96	0.49	0.27	0.53	0.2	2.6	0.2	0.1	0.6
Sidney94	1.96	6.57	0.60	-1.26	-0.81	-0.99	0.21	0.1	1.1	0.5	1.5	0.1
Taipei94	1.96	8.49	-0.34	1.69	-0.62	-0.68	0.82	0.0	1.9	0.3	0.7	1.3
Tel-Aviv94	1.96	10.58	1.39	-0.41	0.77	0.14	1.90	0.6	0.1	0.5	0.0	7.3
Tokyo94	1.96	109.65	-9.90	1.42	1.05	-1.51	-1.33	31.8	1.3	0.9	3.5	3.6
Toronto94	1.96	7.07	0.81	-0.68	-0.96	-1.32	0.76	0.2	0.3	0.7	2.7	1.2
Vienna94	1.96	7.93	-1.28	-2.08	-0.10	0.18	-0.55	0.5	2.9	0.0	0.1	0.6
Zurich94	1.96	17.16	-3.10	-1.57	0.32	0.17	-0.46	3.1	1.6	0.1	0.0	0.4

- Si les variables sont indépendantes, on utilisera l ’espérance pour obtenir une borne inférieure plus simple:

$$p + 2\sqrt{2p(1 + \frac{p-1}{n-1})}$$

et si n est grand:

$$p + 2.8\sqrt{p}$$

ce qui est généralement trop petit (ici 28 au lieu de 38)

1.4 qualité des projections

- Les cosinus carrés (CO2) sont peu fiables
- Un critère meilleur: le carré de la distance à un sous-espace. Pour le premier plan principal:

$$\sum_{j=3}^{p} (c_i^j)^2 = \sum_{j=3}^{p} \lambda_j \chi^2_{j.1}$$

- La borne supérieure à 5% est donnée par:

$$\sum_{i=3}^{p} \lambda_i + 2\sqrt{2\sum_{i=3}^{p} \lambda_i^2}$$

2. L’analyse des correspondances

- N table de contingence avec m_1 lignes et m_2 colonnes
- Les résultats précédents basés sur des approximations normales pour les distances et les contributions ne s’appliquent que pour de grands tableaux.
- Sélection des valeurs propres
  - Hypothèse d’intérêt :$\lambda_i=0$

2.1 Distributions des valeurs propres

- Contexte: N est une réalisation d'une distribution multinomiale $M(n;p_{ij})$ d'espérance nP. Les λ_j sont les valeurs propres de nP. Dans le cas de l'indépendance $\lambda_j = 0 \quad \forall j$
- Lebart et O'Neill ont montré que :

  si $\lambda_i = 0$, $\hat{\lambda}_i$ a la même loi que la ième valeur propre d'une matrice de Wishart $W_{(m_1-1)(m_2-1)}(r, I)$

  avec $r=\min(m_1-1 ; m_2-1)$.

  Sinon , lorsque $\lambda_i \neq 0$, $\sqrt{\hat{\lambda}_i}$ est normalement distribuée

- Comme les tests sont délicats, certains auteurs (Tostado-Torres, Reiczigel) ont proposé des techniques bootstrap pour tester la non-normalité (ie $\lambda_i = 0$)ou pour obtenir des intervalles de confiance pour $\sqrt{\hat{\lambda}_i}$

- Cependant le bootstrap échoue ici:

  les distributions basées sur la loi de Wishart pour les valeurs propres ne peuvent être observées, car on ré-échantillonne dans N qui est presque sûrement de rang r.

Modèle à effets fixes

- Dans le contexte multinomial

$$\mathbf{x}_i = (.... \frac{n_{ij}}{n_{i.}} .....)$$

$$\sigma^2 = \frac{1}{n} \qquad p_i = \frac{n_{i.}}{n}$$

$$\mathbf{\Gamma} = diag(\frac{n_{.j}}{n}) \qquad \mathbf{\Gamma}^{-1} = \text{métrique du chi-deux}$$

- Critère à minimiser:

$$\hat{f}_k = \sum_{k+1}^{p} \hat{\lambda}_i + \frac{1}{n}\left[2k(p+k+q) + 4\sum_{j=1}^{k}\sum_{l=k+1}^{q} \frac{\hat{\lambda}_j}{\hat{\lambda}_j - \hat{\lambda}_l} \right]$$

2.2 Le test de Malinvaud

- Basé sur la formule de reconstitution

$$n_{ij} = \left(n_{i.} n_{.j} / n\right)\left(1 + \sum_{l=1}^{r} a_{il} b_{jl} / \sqrt{\hat{\lambda}_l}\right)$$

- Sous l'hypothèse H_k où seules k valeurs propres sont non-nulles les $\tilde{n}_{ij}$ sont les effectifs espérés

$$\tilde{n}_{ij} = \left(n_{i.} n_{.j} / n\right)\left(1 + \sum_{l=1}^{k} a_{il} b_{jl} / \sqrt{\hat{\lambda}_l}\right)$$

- On les compare aux n_{ij} avec un test du chi-deux

- Au lieu du classique $Q_k = \sum_{i,j} \frac{\left(n_{ij} - \tilde{n}_{ij}\right)^2}{\tilde{n}_{ij}}$

  qui peut être négatif, Malinvaud a suggéré d'utiliser:

$$Q_k^{'} = \sum_{i,j} \frac{(n_{ij} - \tilde{n}_{ij})^2}{\frac{n_{i.} n_{.j}}{n}}$$

- qui est égal à:

$$n(I - \hat{\lambda}_1 - \hat{\lambda}_2 - \ldots - \hat{\lambda}_k) = n(\hat{\lambda}_{k+1} + \hat{\lambda}_{k+2} + \ldots + \hat{\lambda}_r)$$

et peut être comparé à un chi-deux à *(m_1-k-1)(m_2-k-1) degrés de liberté*

- Tests successifs
  - k=0 ? (indépendance) $p_{ij} = p_{i.}p_{.j}$
  - Si indépendance rejetée k=1?

$$p_{ij} = p_{i.}p_{.j}\left(1+\frac{\alpha_{i1}\beta_{j1}}{\sqrt{\lambda_1}}\right)$$

  - Si k=1 rejeté, k=2? $p_{ij} = p_{i.}p_{.j}\left(1+\frac{\alpha_{i1}\beta_{j1}}{\sqrt{\lambda_1}}+\frac{\alpha_{i2}\beta_{j2}}{\sqrt{\lambda_2}}\right)$
  - Etc.
  - Modèle saturé: $p_{ij} = p_{i.}p_{.j}\left(1+\sum_{m=1}^{r}\frac{\alpha_{im}\beta_{jm}}{\sqrt{\lambda_m}}\right)$

Un exemple: beurres allégés

'Contient du beurre	269	70	69	223	14	21	153	118	165	168	23	36	89
'Bon goût	178	74	46	138	12	13	128	90	158	131	20	23	82
'Authentique	124	22	25	84	6	7	70	46	86	61	6	7	22
'Pour gens modernes	184	95	74	184	12	26	158	96	162	229	20	31	138
'Pour la famille	214	80	59	192	18	25	168	114	177	172	21	31	102
'Ressemble au beurre	200	65	32	153	15	17	115	90	138	130	13	22	76
'Qualité/Prix	110	58	30	105	8	13	98	55	114	105	12	15	55
'Se tartine facilem.	243	115	68	217	20	21	231	138	227	247	33	43	113
'Moins de calories	303	137	95	286	24	39	271	165	251	327	36	51	146
'Attention (manger)	253	117	77	244	20	31	210	132	217	282	26	43	124
'Pour les jeunes	121	60	35	117	8	18	98	65	101	134	15	21	95
'Utilisé en cuisson	73	20	12	61	11	5	88	31	44	54	6	2	23
'Composition végétale'	86	46	29	88	9	12	146	38	82	112	11	15	49
'Présentation jolie	158	74	39	127	10	13	121	85	149	175	18	19	84
'Attention (forme)	240	113	98	216	21	33	196	134	197	276	26	45	124
'Aimer la nature	76	38	20	92	7	13	60	46	70	75	9	13	54
'Pour la santé	215	93	55	193	17	26	173	110	173	194	27	34	92
'Pour gens sains	167	76	49	162	16	22	130	93	142	155	17	29	82
'Economique	85	51	27	82	7	10	77	43	87	83	12	13	49

1 . BRIDEL
2 . EFFI
3 . ELLE ET FINE
4 . ELLE ET VIRE
5 . FINE LIGNE
6 . FINESSE
7 . LESIEUR TARTINE
8 . SAINT FLORIDAY
9 . SAINT HUBERT 41
10 . SYLPHIDE
11 . TARTIFINE
12 . TARTINE LEGERE
13 . VIVE LA VIE

EDITION DES VALEURS PROPRES

APERCU DE LA PRECISION DES CALCULS : TRACE AVANT DIAGONALISATION .... .0162
SOMME DES VALEURS PROPRES ..... .0162

HISTOGRAMME DES 12 PREMIERES VALEURS PROPRES

NUMERO	VALEUR PROPRE	POURCENT.	POURCENT. CUMULE	
1	.0064	39.37	39.37	**********
2	.0045	27.93	67.30	*******
3	.0017	10.24	77.55	***
4	.0014	8.32	85.86	***
5	.0008	4.65	90.52	**
6	.0006	3.45	93.96	**
7	.0004	2.21	96.17	*
8	.0003	1.82	97.99	*
9	.0001	.80	98.79	*
10	.0001	.73	99.53	*
11	.0001	.44	99.97	*
12	.0000	.03	100.00	*

Test de Malinvaud

n=21900 p=19 q=13 ⇒ r=12

- Test d'indépendance $n(\hat{\lambda}_1 + \hat{\lambda}_2 + \ldots + \hat{\lambda}_{12}) = 356.28$

  ddl = 18x12=216 $P(\chi^2_{216} > 356) = 0$

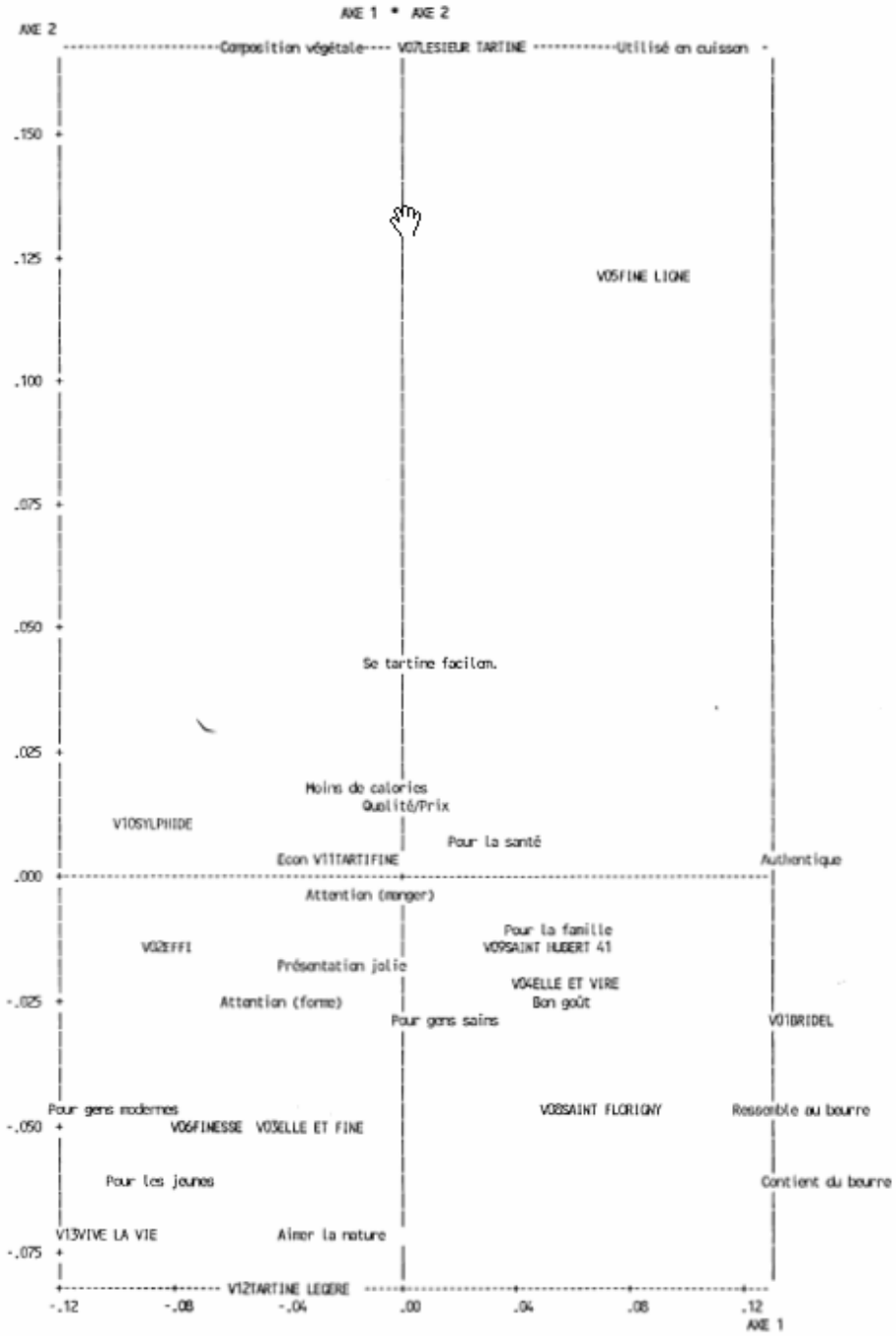
- Test d'unidimensionalité $n(\hat{\lambda}_2 + \ldots + \hat{\lambda}_{12}) = 214.84$

  ddl=17x11=187 $P(\chi^2_{187} > 215) = 0.08$

- Test de bidimensionalité $n(\hat{\lambda}_3 + \ldots + \hat{\lambda}_{12}) = 115.33$

  ddl=16x10=160 $P(\chi^2_{160} > 115) = 0.99$

Deux axes sont retenus


AXE 1 * AXE 2
AXE 2
Composition végétale
V07LESIEUR TARTINE
Utilisé en cuisson
.150
.125
V05FINE LIGNE
.100
.075
.050
Se tartine facilem.
.025
Moins de calories
Qualité/Prix
V10SYLPHIDE
Pour la santé
Econ V11TARTIFINE
Authentique
.000
Attention (manger)
Pour la famille
V09SAINT HUBERT 41
V02EFFI
Présentation jolie
V04ELLE ET VIRE
-.025
Attention (forme)
Bon goût
Pour gens sains
V01BRIDEL
Pour gens modernes
V08SAINT FLORIGNY
Ressemble au beurre
-.050
V06FINESSE
V03ELLE ET FINE
Pour les jeunes
Contient du beurre
V13VIVE LA VIE
Aimer la nature
-.075
V12TARTINE LEGERE
-.12
-.08
-.04
.00
.04
.08
.12
AXE 1

3. L'analyse des correspondances multiples

- Analyse des correspondances du tableau disjonctif $\mathbf{X} = \left(\mathbf{X_1} \middle| \mathbf{X_2} \middle| \ldots \middle| \mathbf{X_{p_p}}\right)$ des p variables qualitatives de rang $q = \sum_{i=1} m_i - p$

- Le test de Malinvaud n'est pas applicable, car valeurs 0 et 1, et non effectifs.

- Rappel de quelques formules :

$$\sum_{i=1}^{q} \hat{\lambda}_i = \frac{1}{p} \sum_{i=1}^{p} m_i - 1$$

$$\sum_{i=1}^{q} \hat{\lambda}_i^2 = \frac{1}{p^2} \sum_{i=1}^{p} (m_i - 1) + \frac{1}{p^2} \sum_{i \neq j} \sum \varphi_{ij}^2$$

3.1 Critères usuels

- Les parts d'inertie ne sont pas pertinentes.
- L'AFC de X et celle du tableau de Burt X'X sont équivalentes, mais $\lambda_i \rightarrow \lambda_i^2$
- Puisque la moyenne des valeurs propres vaut 1/p, on peut éliminer $\hat{\lambda}_i < \frac{1}{p}$
- Le critère du coude

3.2 Le cas de l'indépendance deux à deux

- Chaque $n\varphi_{ij}^2$ est distribué comme $\chi^2_{(m_i-1)(m_j-1)}$

et a pour espérance $(m_i\text{-}1)(m_j\text{-}1)$

d 'où $$E(\sum_{i=1}^{q}\hat{\lambda}_i^2) = \frac{q}{p^2} + \frac{1}{p^2}\frac{1}{n}\sum\sum_{i\neq j}(m_i-1)(m_j-1)$$

en posant $$S_\lambda^2 = \frac{1}{q}\sum_{i=1}^{q}(\hat{\lambda}_i - \frac{1}{p})^2$$

on a:

$$E(S_\lambda^2) = \frac{1}{p^2}\frac{1}{n}\frac{1}{q}\sum\sum_{i\neq j}(m_i-1)(m_j-1) = \sigma^2$$

- On peut considérer que l'intervalle $\frac{1}{p} \pm 2\sigma$ devrait contenir environ 95% des valeurs propres quand les variables sont indépendantes On gardera les valeurs propres qui dépassent la borne supérieure. S.Ben Ammou, G.Saporta (1998)

3.3 Une étude par simulation

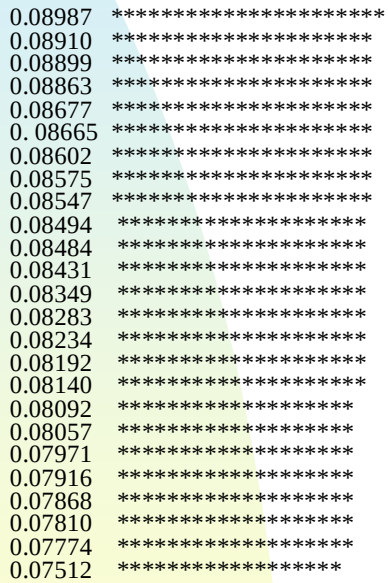
12 variables indépendantes avec un total de 48 catégories: 26 valeurs propres de moyenne =1/12.

N=100 et 10000

- Pour n=100 σ=0.0403, toutes les valeurs propres appartiennent à l'intervalle sauf la plus grande.

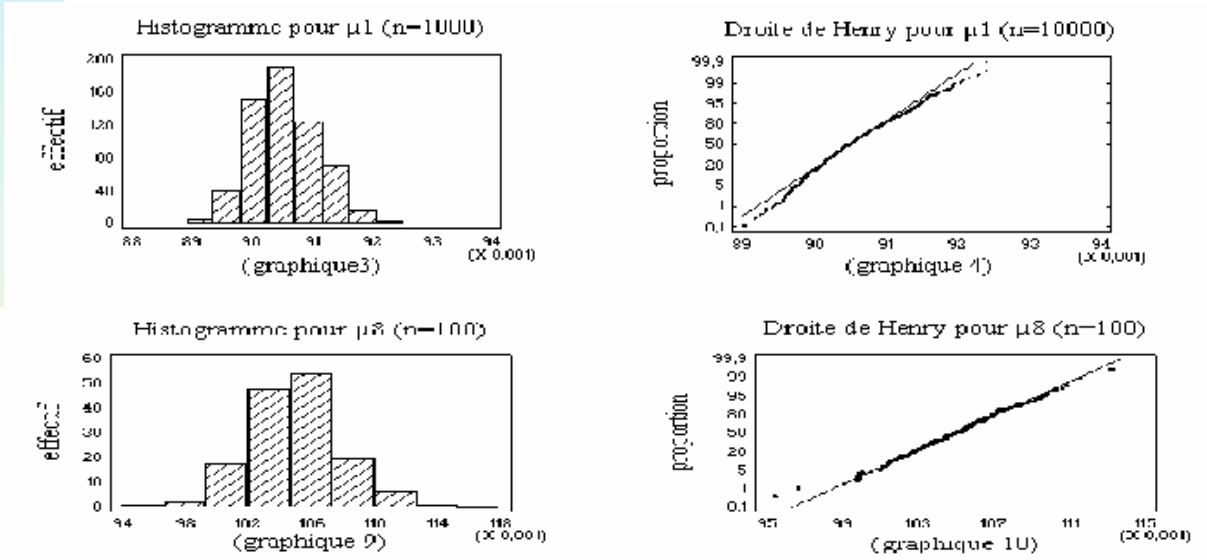
```
0.16902  ********************
0.15126  ******************
0.14483  *****************
0.12939  ****************
0.12245  ***************
0.11691  **************
0.11256  **************
0.11021  *************
0.09771  ************
0.09411  ***********
0.08849  ***********
0.08450  **********
0.07451  *********
0.06908  ********
0.06630  ********
0.06114  *******
0.05762  *******
0.05535  *******
0.05187  ******
0.04927  ******
0.04284  *****
0.04211  *****
0.03724  ****
0.02799  ***
0.02659  ***
```

- Pour n=10000, σ= 0.004199, l'intervalle [0.07527 ; 0.09139] contient toutes les valeurs propres sauf la dernière



- Distribution des valeurs propres.

Les approximations normales ne marchent pas pour les valeurs propres extrêmes (600 simulations)



En guise de conclusion

- Quelques formules simples pour détecter les valeurs propres, les contributions significatives et les outliers en ACP.

  Certaines peuvent s'appliquer en AFC et ACM
- Pour les très grands fichiers les tests sont sans intérêt car toujours significatifs.
- Mais chercher si $\lambda>1$ en ACP, $\lambda> 0$ enAFC, $\lambda>1/p$ en ACM est-elle la bonne question?

Références

- S.Ben Ammou, G.Saporta (1998) Sur la normalité asymptotique des valeurs propres en ACM sous l'hypothèse d'indépendance des variables. *Revue de Statistique Appliquée* , Vol. XLVI, n°3, p.21-35,
- Karlis, D., Saporta, G. and Spinakis A. (2003) A Simple Rule for the Selection of Principal Components, Communications in Statistics, Theory and Applications, 32, 3, 643-666
- L.Lebart (1976). The significance of Eigenvalues issued from Correspondence Analysis *COMPSTAT*, Physica Verlag, Vienna, p 38-45 .
- L.Lebart, A.Morineau, M.Piron (2006). *Statistique exploratoire multidimensionnelle*. 4ème édition, Dunod, Paris
- E.Malinvaud, (1987) Data analysis in applied socio-economic statistics with special consideration of correspondence analysis, *Marketing Science Conference*, Jouy en Josas, France, 1987
- M.E.O'Neill. (1978). Asymptotic distributions of the canonical correlations from contingency tables. *Australian Journal of. Statistics*. 20(1) p 75-82.
- M.E.O'Neill (1978). Distributional expansion for canonical correlations from contingency tables . *Journal of the Royal. Statistical Society. B*. 40, n°3 p 303-312.
- G.Saporta, N.Tambrea (1993): About the selection of the number of components in correspondence analysis in J.Janssen et C.H.Skiadas, eds. *Applied Stochastic Models and Data Analysis*, World Scientific, p. 846-856,
- G. Saporta (1999) Some simple rules for interpreting outputs of principal components and correspondence analysis . In ASMDA99, IX International Symposium on Applied Stochastic Models and Data Analysis, Lisbonne, Portugal, 14-17 Juin 1999
- L.Zater (1989). Contribution a l'étude de la variabilité des valeurs propres et du choix de la dimension en analyse factorielle des correspondances. *Thèse de l'Université Paris IX Dauphine.*