On the Variance of Eigenvalues in PCA and MCA

Jean-Luc Durand

jean-luc.durand@univ-paris13.fr

Laboratoire d'Éthologie Expérimentale et Comparée - LEEC EA 4443
Université Paris 13 Sorbonne Paris Cité

Naples, CARME 2015
22/09/2015

CA, PCA and MCA

Which statistics measure
the overall magnitude of the relations between variables
in CA, PCA and MCA ?

- the sum of the eigenvalues in CA
- the variance of the eigenvalues in PCA and MCA.

Part 1

Principal Component Analysis (PCA on correlation matrix)

- Overall viewpoint: correlations and eigenvalues
- Local viewpoint: contributions of each variable to axes

PCA Measure of Overall Magnitude of Correlations

Definition

The "average linkage index" of p numerical variables is:

$$\text{ALI} = \frac{\sum_{k} \sum_{k' \neq k} r_{kk'}^2}{p(p-1)}$$

(mean of the squared correlations between variables).

$$0 \leq \min_{(k,k')} r_{kk'}^2 \leq \text{ALI} \leq \max_{(k,k' \neq k)} r_{kk'}^2 \leq 1$$

Variance of the p eigenvalues: $\mathrm{Var}(\boldsymbol{\lambda}) = \sum_{\ell=1}^{p} (\lambda_l - 1)^2 / p$

$$0 \leq \mathrm{Var}(\boldsymbol{\lambda}) \leq p - 1$$

Two extreme situations:

p uncorrelated variables	p perfectly correlated variables
$r_{kk'} = 0 \ \ (k \neq k')$	$r^2_{kk'} = 1$
$\lambda_1 = \ldots = \lambda_p = 1$	$\lambda_1 = p, \quad \lambda_2 = \ldots = \lambda_p = 0$
$\mathrm{Var}(\boldsymbol{\lambda}) = 0$	$\mathrm{Var}(\boldsymbol{\lambda}) = p - 1$
Spherical cloud	*Unidimensional* cloud

Theorem

In PCA of p standardized variables:

$$\mathrm{Var}(\boldsymbol{\lambda}) = (p - 1)\,\mathrm{ALI}$$

The Average Linkage Index is a measure:

- of the overall magnitude of correlations
- of the departure from sphericity of the cloud.

SPEARMAN C., "General Intelligence," Objectively Determined and Measured, *The American Journal of Psychology*, Vol. 15, No. 2, April 1904 (p. 291).

Correlation matrix

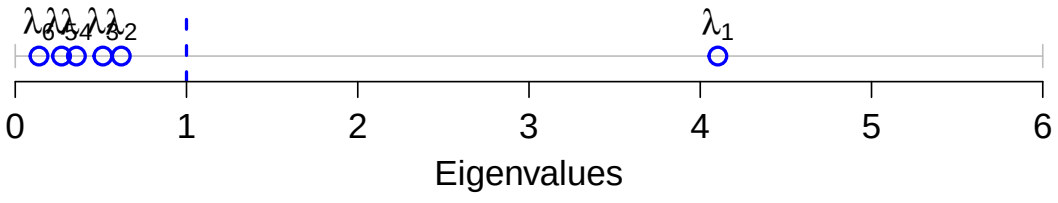
	Lit	Fre	Eng	Mat	Aud	Mus
Literature	1	.83	.78	.70	.66	.63
French		1	.67	.67	.65	.57
English			1	.64	.54	.51
Mathematics				1	.45	.51
Aud. disc.					1	.40
Music						1

$$\mathrm{ALI} = .3899 \qquad \sqrt{\mathrm{ALI}} = .62$$

Eigenvalues

$$\begin{aligned}\lambda_1 &= 4.1028\\ \lambda_2 &= 0.6191\\ \lambda_3 &= 0.5117\\ \lambda_4 &= 0.3570\\ \lambda_5 &= 0.2704\\ \lambda_6 &= 0.1389\end{aligned}$$

Sum 6



$$\mathrm{Var}(\boldsymbol{\lambda}) = 1.9496 = 5 \times \mathrm{ALI}$$

Definition

Given p numerical variables,
the "linkage index" of variable k is:

$$\mathrm{LI}_k = \frac{\sum_{k' \neq k} r^2_{kk'}}{p-1}$$

(mean of the squared correlations between variable k anf the others).

$$0 \leq \min_{k'} r^2_{kk'} \leq \mathrm{LI}_k \leq \max_{k' \neq k} r^2_{kk'} \leq 1$$

	LI_k
Lit	.5240
Fre	.4668
Eng	.4037
Mat	.3622
Aud	.3024
Mus	.2804
ALI	.3899

Mean: $\frac{\sum_k \mathrm{LI}_k}{p} = \mathrm{ALI}$

Definition

The "linkage ratio" of variable k is:

$$\mathrm{LR}_k = \frac{\mathrm{LI}_k}{\mathrm{ALI}}$$

	LR_k
Lit	1.34
Fre	1.20
Eng	1.04
Mat	0.93
Aud	0.78
Mus	0.72
Mean	1

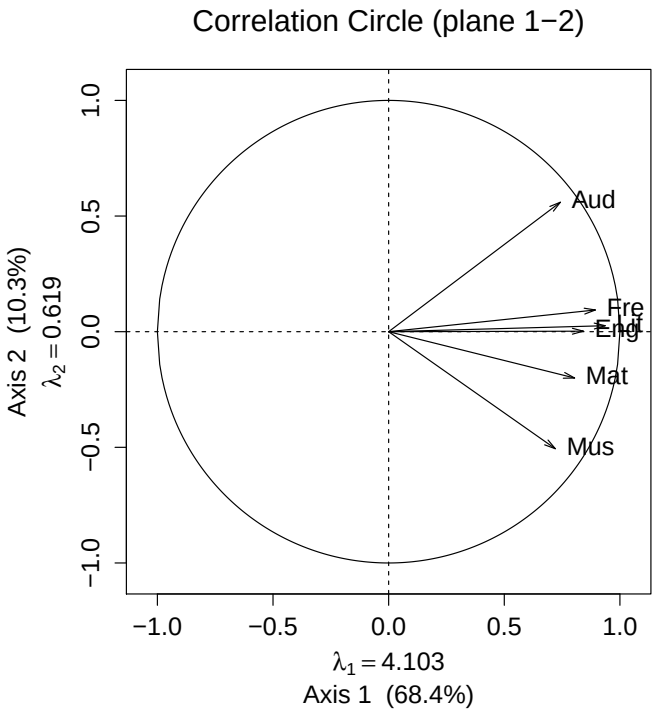
Mean: $\frac{\sum_k \mathrm{LR}_k}{p} = 1$

Regression coefficients: $r_{k\ell}$

	Axis 1	Axis 2
Literature	.94	.03
French	.89	.09
English	.84	.00
Mathematics	.80	-.20
Aud. disc.	.74	.56
Music	.72	-.51
λ_ℓ	4.103	0.619

$$\forall \ell \in (1, \ldots, p) \quad \sum_k r_{k\ell}^2 = \lambda_\ell$$

$$\forall k \in (1, \ldots, p) \quad \sum_\ell r_{k\ell}^2 = 1$$



Contributions of variables to axes (Ctr_k^ℓ)

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Sum
Literature	**.214**	.001	.001	.016	.071	**.697**	1
French	**.195**	.014	.000	.052	**.538**	**.201**	1
English	**.173**	.000	.117	**.585**	.036	.089	1
Mathematics	.158	.065	**.311**	**.321**	.144	.000	1
Aud. disc.	.134	**.506**	.150	.025	**.181**	.003	1
Music	.127	**.414**	**.420**	.000	.029	.010	1
Sum	1	1	1	1	1	1	
Eigenvalue	4.1028	0.6191	0.5117	0.3570	0.2704	0.1389	

Contribution of variable k to axis ℓ: $\mathrm{Ctr}_k^\ell = r_{k\ell}^2/\lambda_\ell$

$$\forall \ell \quad \sum_k \mathrm{Ctr}_k^\ell = 1 \qquad \forall k \quad \textstyle\sum_\ell \mathrm{Ctr}_k^\ell = 1$$

$\mathbf{w}_k$: vector of the contributions of variable k to axes

$$\forall k \quad \mathrm{Corr}(\mathbf{w}_k, \boldsymbol{\lambda}) = 0$$

Property of Uncorrelated Positive Variables

Let $\mathbf{x}$ and $\mathbf{y}$ be two *positive* variables with n values,

Mean of $\mathbf{x}$: $\frac{\sum x_i}{n}$, denoted $\bar{x}$

Weighted average of $\mathbf{x}$ by $\mathbf{y}$: $\frac{\sum y_i x_i}{\sum y_i}$, denoted $\mathrm{Avg}_y(\mathbf{x})$

(" $\mathbf{y}$-average of $\mathbf{x}$")

Property

$$r_{xy} = 0 \iff \bar{x} = \mathrm{Avg}_y(\mathbf{x}) \iff \bar{y} = \mathrm{Avg}_x(\mathbf{y})$$

In PCA on correlation matrix, $\forall k \quad \bar{\boldsymbol{\lambda}} = 1 = \mathrm{Avg}_{\mathbf{w}_k}(\boldsymbol{\lambda})$

Expansion Ratio

Let $\mathbf{x}$ and $\mathbf{w}$ be two *uncorrelated positive* variables,

Variance of $\mathbf{x}$: $\frac{\sum (x_i - \bar{x})^2}{n}$, denoted $\mathrm{Var}(\mathbf{x})$

Weighted variance of $\mathbf{x}$ by $\mathbf{w}$: $\frac{\sum w_i (x_i - \bar{x})^2}{\sum w_i}$, denoted $\mathrm{Var}_w(\mathbf{x})$

(" $\mathbf{w}$-variance of $\mathbf{x}$")

Definition

The "w-expansion ratio" of $\mathbf{x}$ is:

$$\mathrm{XR}_w(\mathbf{x}) = \frac{\mathrm{Var}_w(\mathbf{x})}{\mathrm{Var}(\mathbf{x})}$$

Theorem

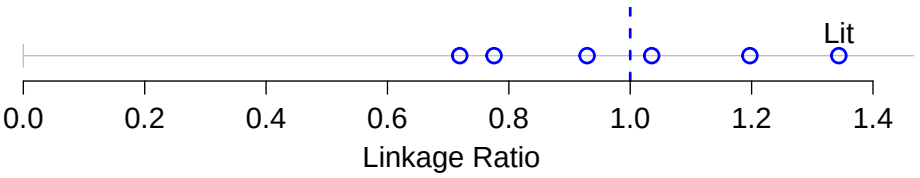
In PCA of p standardized variables:

$$\forall k \quad \mathrm{Var}_{\mathbf{w}_k}(\boldsymbol{\lambda}) = (p-1)\,\mathrm{LI}_k$$

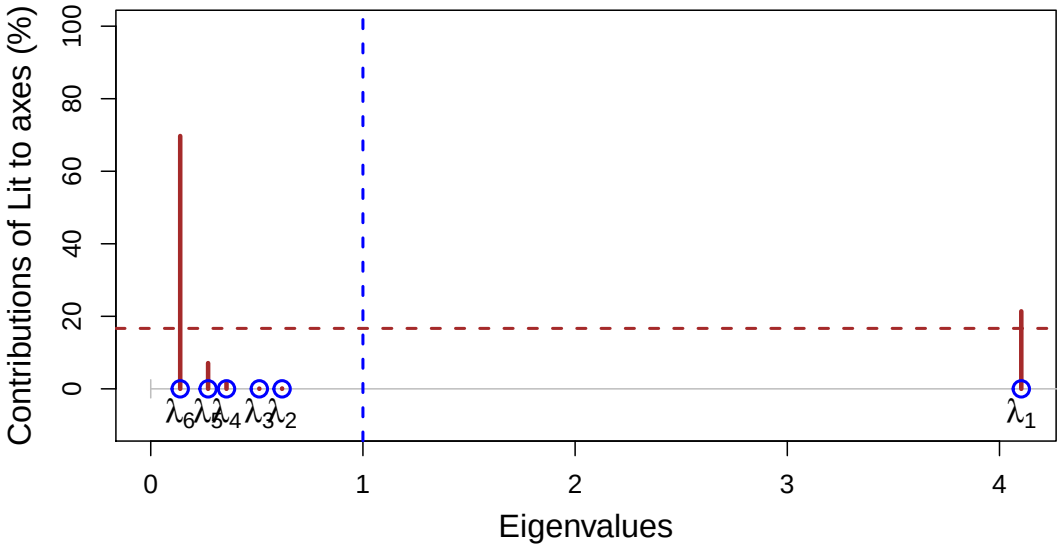
$$\mathrm{XR}_{\mathbf{w}_k}(\boldsymbol{\lambda}) = \mathrm{LR}_k$$

Interpretation

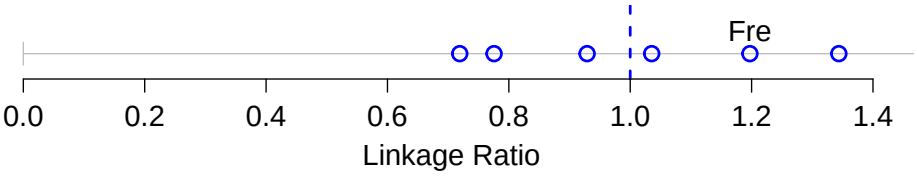
- The higher the linkage ratio of a variable, the higher the contributions of this variable to extreme eigenvalues.
- The lower the linkage ratio of a variable, the higher the contributions of this variable to central eigenvalues.



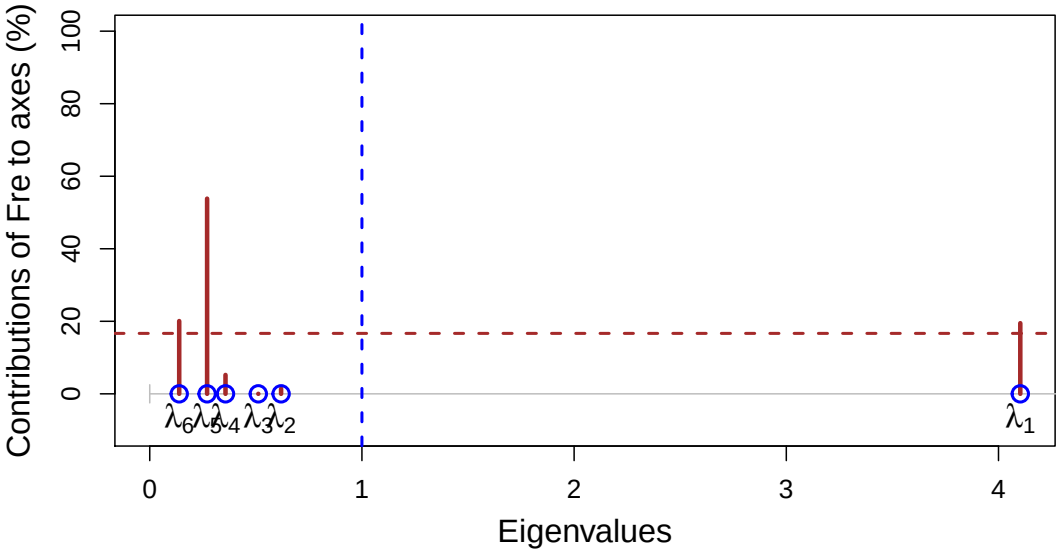
$$\mathrm{XR} = 1.34$$



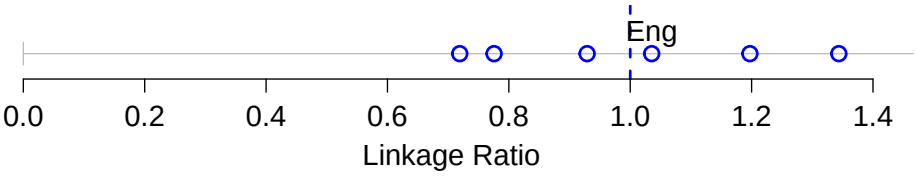
Spearman Example — French



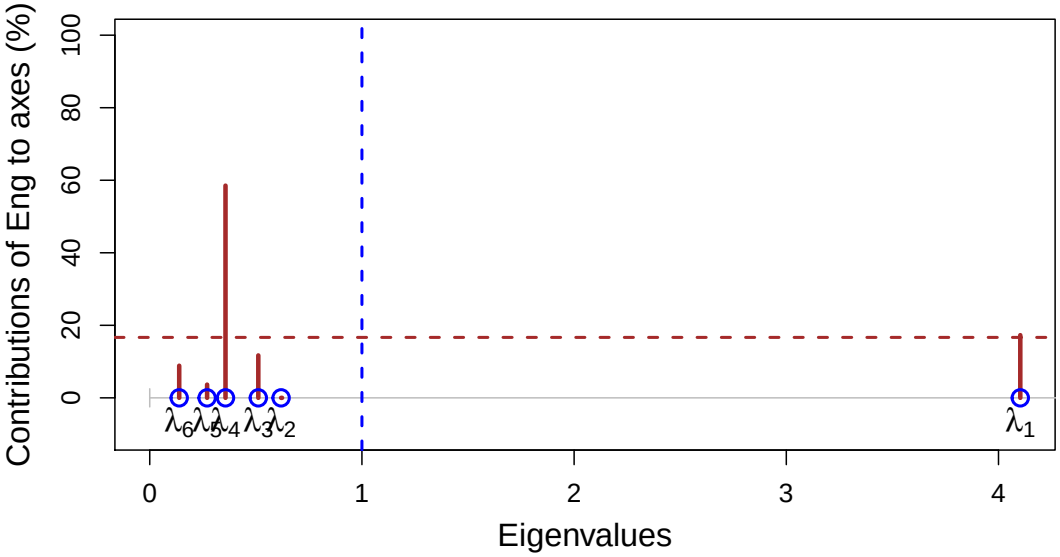
$$\mathrm{XR} = 1.20$$

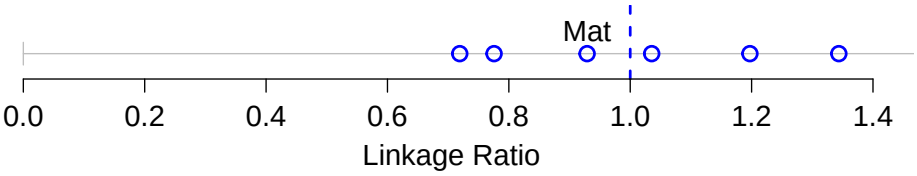


Spearman Example — English

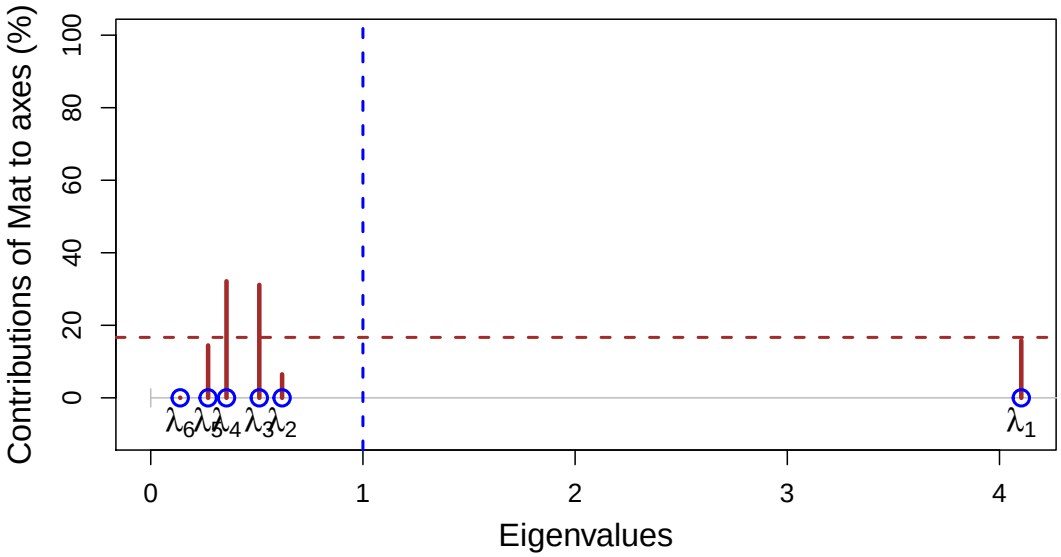


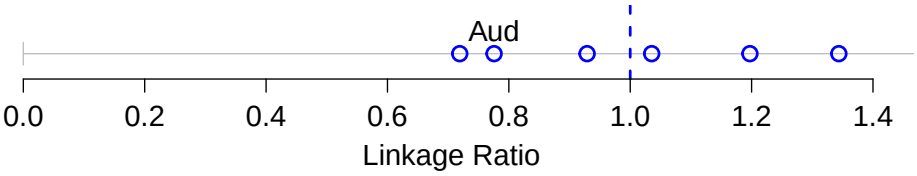
$$\mathrm{XR} = 1.04$$



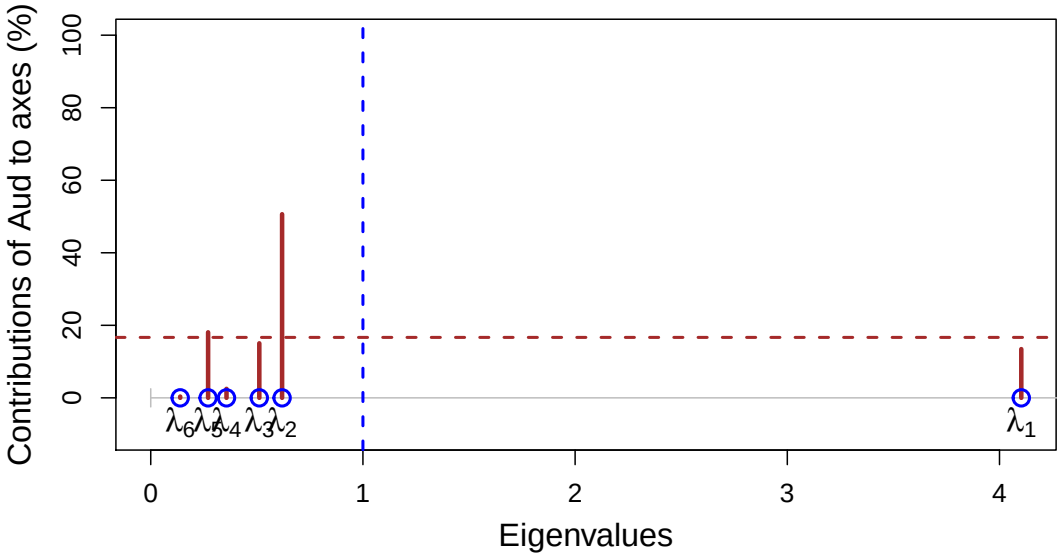


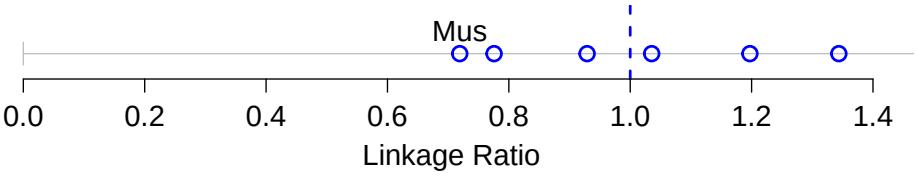
$$\mathrm{XR} = 0.93$$



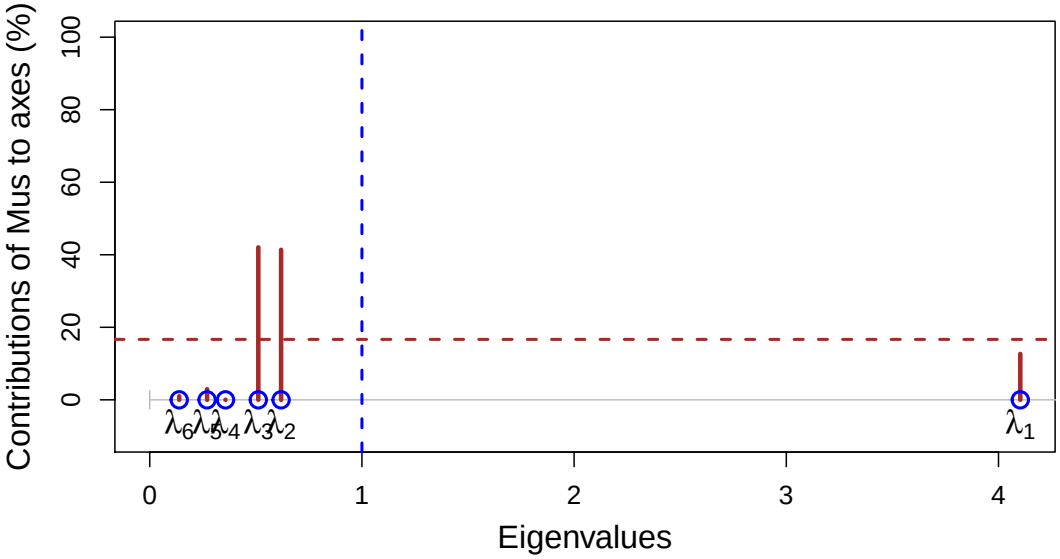


$$\mathrm{XR} = 0.78$$





$$\mathrm{XR} = 0.72$$



PCA

Summary

In PCA on correlation matrix:

- The variance of eigenvalues is proportional to the average linkage index.
- The distribution of contributions of a variable to axes depends on the linkage ratio of this variable.

Part 2

Multiple Correspondence Analysis

- Overall viewpoint: Φ^2 coefficients and eigenvalues
- Local viewpoint: contributions of each question to axes

MCA Burt's Data

BURT C., The Factorial Analysis of Qualitative Data, *British Journal of Statistical Psychology*, Vol. 3, Issue 3, November 1950 (p. 177).

		Hair			Eyes			Head		Stature	
		fair	red	dark	light	mixed	brown	narrow	wide	tall	short
Hair	fair	22	0	0	14	6	2	14	8	13	9
	red	0	15	0	8	5	2	11	4	10	5
	dark	0	0	63	11	25	27	44	19	20	43
Eyes	light				33	0	0	27	6	29	4
	mixed				0	36	0	20	16	10	26
	brown				0	0	31	22	9	4	27
Head	narrow							69	0	30	39
	wide							0	31	13	18
Stature	tall									43	0
	short									0	57

$\mathrm{Q} = 4$ questions $\qquad \mathrm{K} = \sum \mathrm{K}_q = 3 + 3 + 2 + 2 = 10$ categories

Average number of categories per question: $\frac{\mathrm{K}}{\mathrm{Q}} = 2.5$

Φ^2 table

	Hair	Eyes	Head	Stature
Hair	2	0.2125	0.0045	0.0901
Eyes		2	0.0563	0.4198
Head			1	0.0002
Stature				1

Definition

The "average linkage index" of Q questions with K categories is:

$$\mathrm{ALI} = \frac{\overline{\Phi^2}}{\frac{\mathsf{K}}{\mathsf{Q}} - 1}$$

$$\text{with } \overline{\Phi^2} = \frac{\sum_q \sum_{q' \neq q} \Phi^2_{qq'}}{\mathsf{Q}(\mathsf{Q}-1)}$$

(mean of the $\Phi^2_{qq'}$ coefficients $(q \neq q')$).

$$0 \leq \mathrm{ALI} \leq 1$$

Burt example: $\overline{\Phi^2} = 0.1306,$ $\mathrm{ALI} = \frac{0.1306}{1.5} = 0.0870$

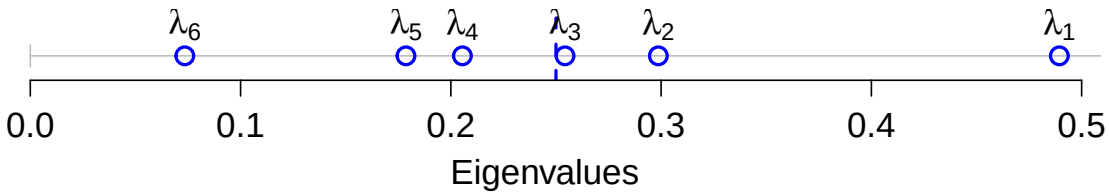
Eigenvalues

$\lambda_1 = 0.4894$
$\lambda_2 = 0.2987$
$\lambda_3 = 0.2543$
$\lambda_4 = 0.2055$
$\lambda_5 = 0.1787$
$\lambda_6 = 0.0735$

$$\sum_\ell \lambda_\ell = \frac{K}{Q} - 1 = 1.5$$

$$\text{Mean}(\boldsymbol{\lambda}) = \frac{1}{Q} = 0.25$$

$$\text{Var}(\boldsymbol{\lambda}) = 0.01632$$



$$\text{Var}(\boldsymbol{\lambda}) = \frac{\sum_{\ell=1}^{K-Q} \left(\lambda_\ell - \frac{1}{Q}\right)^2}{K - Q}$$

$$0 \leq \text{Var}(\boldsymbol{\lambda}) \leq \frac{Q - 1}{Q^2}$$

Two extreme situations:

Q independent questions

$$\Phi^2_{qq'} = 0 \ \ (q \neq q')$$

$$\lambda_1 = \ \ldots \ = \lambda_p = \frac{1}{Q}$$

$$\text{Var}(\boldsymbol{\lambda}) = 0$$

Spherical cloud

Q equivalent questions

$$\Phi^2_{qq'} = K_q - 1 = K_{q'} - 1$$

$$\lambda_\ell = \begin{cases} 1 & \text{if } 1 \leq \ell \leq \frac{K}{Q} - 1 \\ 0 & \text{if } \frac{K}{Q} \leq \ell \leq K - Q \end{cases}$$

$$\text{Var}(\boldsymbol{\lambda}) = \frac{Q - 1}{Q^2}$$

$(\frac{K}{Q} - 1)$-*dimensional* cloud

Theorem

In MCA on a table with Q *questions:*

$$\mathrm{Var}(\boldsymbol{\lambda}) = \frac{\mathrm{Q}-1}{\mathrm{Q}^2}\,\mathrm{ALI}$$

The Average Linkage Index is a measure:

- of the overall magnitude of relations between questions
- of the departure from sphericity for the cloud.

Definition

Given Q questions,
the "linkage index" of a question q with K_q categories is:

$$\mathrm{LI}_q = \frac{\overline{\Phi^2_q}}{\mathrm{K}_q - 1}$$

$$\text{with } \overline{\Phi^2_q} = \frac{\sum_{q' \neq q} \Phi^2_{qq'}}{\mathrm{Q}-1}$$

(mean the Φ^2 coefficients between question q and the others).

$$0 \leq \mathrm{LI}_q \leq 1$$

Average of linkage indexes weighted by number of categories minus 1

$$\frac{\sum_q (\mathrm{K}_q - 1)\mathrm{LI}_q}{\mathrm{K}-\mathrm{Q}} = \mathrm{ALI}$$

Φ^2 table

	Hair	Eyes	Head	Stature	$\overline{\Phi^2_q}$	$K_q - 1$	LI_q
Hair		0.2125	0.0045	0.0901	0.1224	2	0.0512
Eyes	0.2125		0.0563	0.4198	0.2295	2	0.1148
Head	0.0045	0.0563		0.0002	0.0203	1	0.0203
Stature	0.0901	0.4198	0.0002		0.1700	1	0.1700

Average of linkage indexes weighted by number of categories minus 1

$$ALI = 0.0870$$

Definition

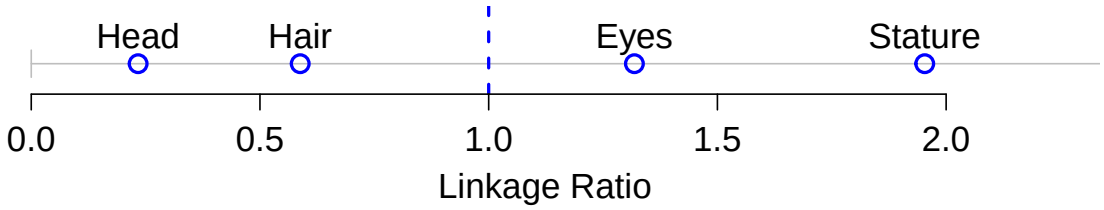
The "linkage ratio" of question q is:

$$LR_q = \frac{LI_q}{ALI}$$

Average of linkage ratios weighted by number of categories minus 1

$$\frac{\sum_q (K_q - 1) LR_q}{K - Q} = 1$$

	LR_q	$\mathrm{K}_q - 1$
Hair	0.784	2
Eyes	1.758	2
Head	0.156	1
Stature	1.302	1
Weighted average	1	



Contributions of questions to axes (Ctr_q^ℓ)

	Axis 1	Axis 2	Axis 3	Axis 4	Axis 5	Axis 6	Sum
Hair	.233	.059	**.911**	.054	**.674**	.069	2
Eyes	**.413**	**.447**	.080	**.468**	.040	**.551**	2
Head	.011	**.493**	.005	**.447**	.009	.035	1
Stature	**.343**	.000	.004	.031	**.277**	**.345**	1
Sum	1	1	1	1	1	1	
Eigenvalue	0.4894	0.2987	0.2543	0.2055	0.1787	0.0735	

$$\forall \ell, \quad \sum_q \mathrm{Ctr}_q^\ell = 1 \qquad \forall q, \quad \sum_\ell \mathrm{Ctr}_q^\ell = \mathrm{K}_q - 1$$

$\mathbf{w}_q$: vector of the contributions of question q to axes

$$\forall q, \quad \mathrm{Corr}(\mathbf{w}_q, \boldsymbol{\lambda}) = 0$$

Theorem

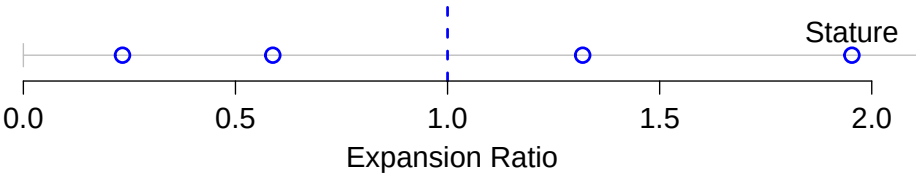
In MCA on a table with Q questions:

$$\forall q \quad \mathrm{Var}_{\mathbf{w}_q}(\boldsymbol{\lambda}) = \frac{\mathrm{Q}-1}{\mathrm{Q}^2}\,\mathrm{LI}_q$$

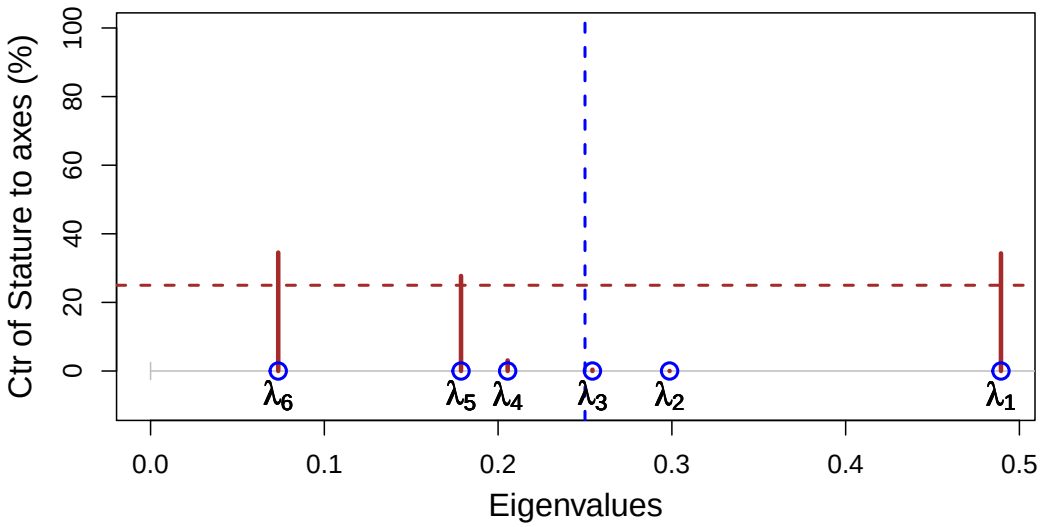
$$\mathrm{XR}_{\mathbf{w}_q}(\boldsymbol{\lambda}) = \mathrm{LR}_q$$

Interpretation

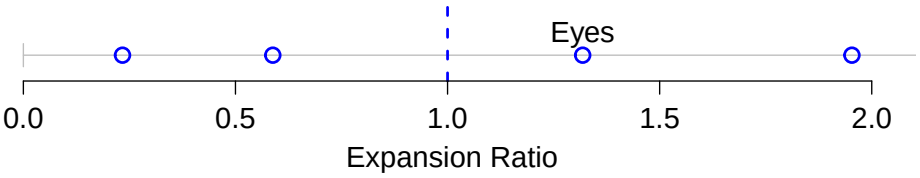
- The higher the linkage ratio of a variable, the higher the contributions of this variable to extreme eigenvalues.
- The lower the linkage ratio of a variable, the higher the contributions of this variable to central eigenvalues.



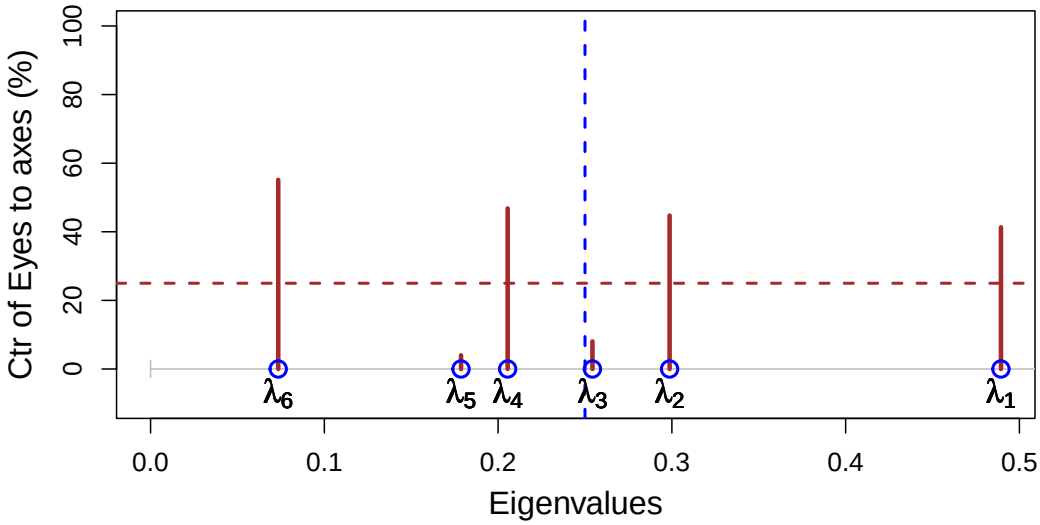
$$\mathrm{XR} = 1.95$$



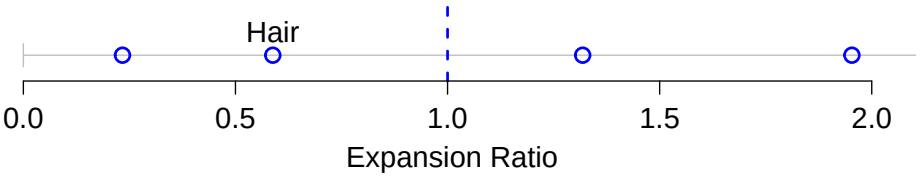
Burt Example — Eyes



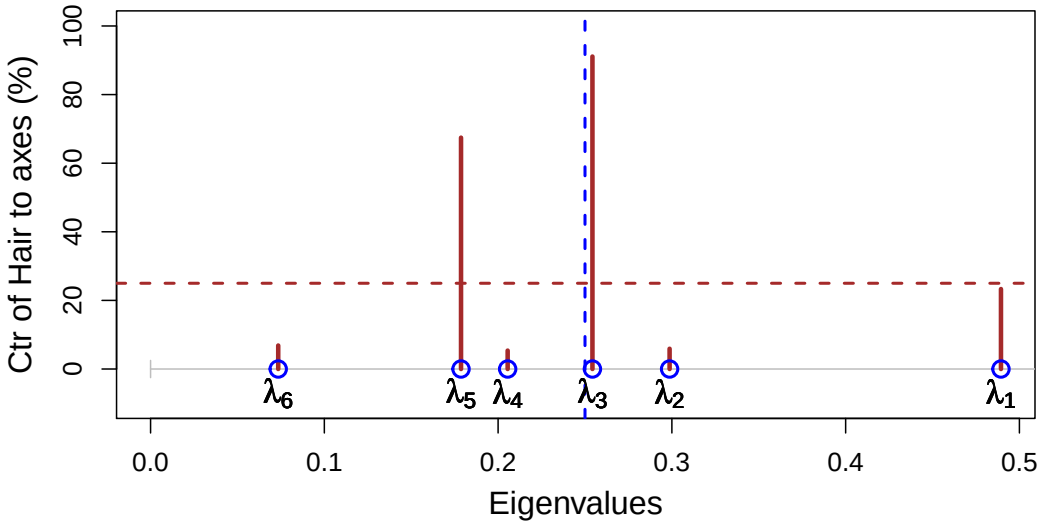
$$\mathrm{XR} = 1.32$$

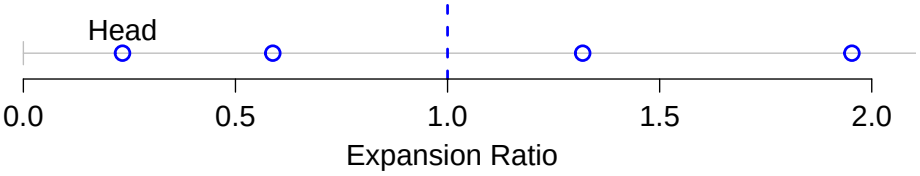


Burt Example — Hair

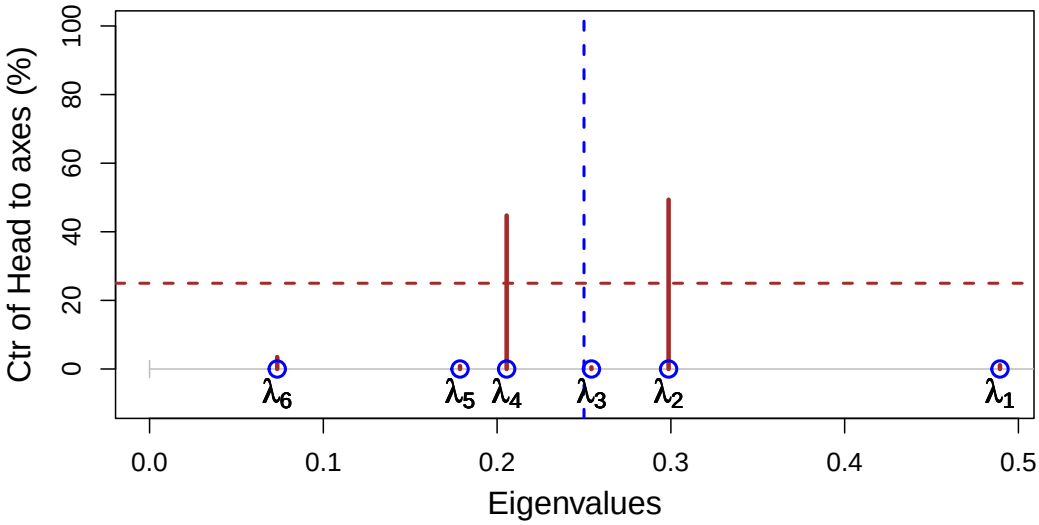


$$\mathrm{XR} = 0.59$$





$$\mathrm{XR} = 0.23$$



Summary

In MCA:

- The variance of eigenvalues is proportional to the average linkage index.
- The distribution of contributions of a question to axes depends on the linkage ratio of this question.

Suggestions

- in PCA results: report ALI and $\mathrm{LI}s$ or $\mathrm{LR}s$ of variables.
- in MCA results: report Φ^2 table, ALI and $\mathrm{LI}s$ or $\mathrm{LR}s$ of questions.
- in agglomerative hierarchical clustering of variables: use the within-class ALI as an aggregation index.