doi.org/10.26398/IJAS.0029-006

LINKAGE INDEX OF VARIABLES AND ITS RELATIONSHIP WITH VARIANCE OF EIGENVALUES IN PCA AND MCA

Jean-Luc Durand[1]

Laboratoire d'Ethologie Expérimentale et Comparée, LEEC EA4443, Université Paris 13 (Sorbonne Paris Cité), Villetaneuse, France

Brigitte Le Roux

MAP5, UMR 8145, Université Paris Descartes (Sorbonne Paris Cité), Paris CEVIPOF, UMR 7048, Sciences Po, Paris, France

Abstract. *In the present article, we show that, in principal component analysis (PCA) on correlation matrix as well as in multiple correspondence analysis (MCA), the strength of the relationship between variables is linked to the variance of the eigenvalues, and indicates the axes to which the variables contribute the most. In PCA, we define the linkage index of a variable as the mean of the squared correlations between this variable and the others. We prove that the variance of eigenvalues is proportional to the mean linkage index and that, for each variable, the variance of eigenvalues weighted by the contributions of the variable to axes is proportional to the linkage index of the variable. In MCA, similar properties are proven regarding both categorical variables and categories. We illustrate these properties using two datasets coming from classical articles by Spearman (1904) for PCA and Burt (1950) for MCA.*

Keywords: *PCA, MCA, Variance of eigenvalues, Contributions to axes.*

1. INTRODUCTION

In the present article, we study the variance of eigenvalues in PCA and MCA, i.e., the mean of squared deviations from eigenvalues to their mean. The variance of eigenvalues can be seen as an index of departure from sphericity. We examine the relationship between the variance of eigenvalues and the correlations between variables in PCA or the contingency mean square coefficients (usually denoted Φ^2) between categorical variables in MCA. This article develops the properties presented in Durand (1998).

There are few studies on the variance of eigenvalues. However, we can find the expression of the sum of squares of eigenvalues in studies on the

[1] Corresponding author: Jean-Luc Durand, email: jean-luc.durand@univ-paris13.fr

number of axes to be used for interpretation, or on confidence interval (see e.g. Saporta, 2003; Karlis et al., 2003). The variance of eigenvalues is also used in applications, especially in biological studies (Pavlicev et al., 2009).

2. PRINCIPAL COMPONENT ANALYSIS

In this section we study the variance of eigenvalues in the case of PCA on correlation matrix.

2.1. BASIC PROPERTIES AND NOTATIONS

Let I denote a set of n individuals, K a set indexing p ($p > 1$) non–constant variables on I; the k–th variable is denoted $\boldsymbol{x}_k = \left[x_k^i\right]$, its mean $\overline{x}_k$ and its variance v_k.

Let $r_{kk'}$ be the correlation between variables $\boldsymbol{x}_k$ and $\boldsymbol{x}_{k'}$ and $\mathbf{R} = [r_{kk'}]$ the correlation matrix between the p variables. The calculation method of standard PCA is based on the diagonalization of the correlation matrix $\mathbf{R}$.

Let L be a set indexing the nonnull eigenvalues. If $\mathbf{\Lambda} = [\lambda_\ell]$ denotes the diagonal matrix of eigenvalues $(\lambda_\ell)_{\ell \in L}$ and $\mathbf{A} = [a_{k\ell}]$ the matrix of eigenvectors, then the PCA of variables $\boldsymbol{x}_k$ writes $\mathbf{RA} = \mathbf{A\Lambda}$ with $\mathbf{A}^\top \mathbf{A} = \mathbf{I}$. In the simple linear regression of the standardized initial variable $(\boldsymbol{x}_k - \overline{x}_k)/\sqrt{v_k}$ on the ℓ–th principal variable with variance 1, the *regression coefficient* is equal to the correlation coefficient $r_{k\ell}$ between the k–th initial variable and the ℓ–th principal variable. We have the properties:

$$r_{k\ell} = \sqrt{\lambda_\ell} a_{k\ell} \quad \text{and} \quad \sum_{\ell \in L} (r_{k\ell})^2 = 1.$$

The contribution of variable $\boldsymbol{x}_k$ to the variance λ_ℓ of axis ℓ, denoted Ctr_k^ℓ, is equal to $r_{k\ell}^2/\lambda_\ell$, with $\forall \ell \in L$, $\sum_{k \in K} \mathrm{Ctr}_k^\ell = 1$.

If the correlation matrix has full rank (all eigenvalues are strictly positive), one has the property:

$$\forall k \in K, \sum_{\ell \in L} \mathrm{Ctr}_k^\ell = 1. \tag{1}$$

This property comes from $\mathrm{Ctr}_k^\ell = r_{k\ell}^2/\lambda_\ell = a_{k\ell}^2$, with $\sum_{\ell \in L} a_{k\ell}^2 = 1$ since matrix $\mathbf{A}$ is orthogonal ($\mathbf{AA}^\top = \mathbf{A}^\top \mathbf{A} = \mathbf{I}$).

2.2. VARIANCE OF EIGENVALUES

Theorem 2.1. *The variance of eigenvalues, denoted* $V(\lambda)$*, is such that*

$$V(\lambda) = \frac{1}{p} \sum_{k \in K} \sum_{\substack{k' \in K \\ k' \neq k}} r_{kk'}^2 .$$

Proof. The mean of eigenvalues is $\overline{\lambda} = 1$, the variance is $V(\lambda) = \frac{1}{p} \sum_{\ell \in L} \lambda_\ell^2 - 1$, $\sum_{\ell \in L} \lambda_\ell^2$ being equal to the trace of matrix $\mathbf{R}^2$. The entries of $\mathbf{R}^2$ are equal to $\sum_{k' \in K} r_{kk'} r_{k'k''}$ and the diagonal entries are $(\sum_{k' \in K} r_{kk'}^2)_{k \in K}$, hence the trace of $\mathbf{R}^2$ is equal to $\sum_{k \in K} \left(1 + \sum_{k' \neq k} r_{kk'}^2\right) = p + \sum_{k \in K} \sum_{\substack{k' \in K \\ k' \neq k}} r_{kk'}^2$.

This expression of the variance invites us to consider the mean of the squared correlations between one variable and the others. As we will see later on, this quantity is an index of the strength of the link between one variable and the others. Hence the following definition:

Definition 2.1 (Linkage index of a variable)**.** *We call linkage index of variable* $\boldsymbol{x}_k$*, denoted* LI_k*, the mean of the squared correlations between the variable* $\boldsymbol{x}_k$ *and the* $(p-1)$ *others:*

$$\mathrm{LI}_k = \frac{1}{p-1} \sum_{\substack{k' \in K \\ k' \neq k}} r_{kk'}^2 .$$

Note that the linkage index of a variable is between zero and one.

Proposition 1. *The* mean linkage index *of the p variables, denoted* $\overline{\mathrm{LI}}$*, is equal to the variance of eigenvalues divided by (p-1):*

$$\overline{\mathrm{LI}} = \frac{1}{p-1} V(\lambda) .$$

Comment. The mean linkage index is a measure of both the global magnitude of correlations and of the departure from sphericity. In the particular case of an equicorrelation matrix, all off–diagonal elements of which are equal to r (Morrison, 1976, p. 331), the mean linkage index is equal to r^2.

2.3. VARIABLES AND EIGENVALUES

From now on, we assume that the correlation matrix has full rank.

We will now express the linkage index of a variable as a function of the eigenvalues and the contributions of this variable to axes.

Considering the contributions of the initial variable $\boldsymbol{x}_k$ to axes (namely $(\mathrm{Ctr}_k^\ell)_{\ell\in L}$ with $\sum_{\ell\in L}\mathrm{Ctr}_k^\ell = 1$, see Equation 1), we define the weighted mean $(\widehat{\lambda}_k)$ and the weighted variance $(\widehat{V}_k(\lambda))$ of eigenvalues:

$$\widehat{\lambda}_k = \sum_{\ell\in L}\mathrm{Ctr}_k^\ell\,\lambda_\ell \quad \text{and} \quad \widehat{V}_k(\lambda) = \sum_{\ell\in L}\mathrm{Ctr}_k^\ell(\lambda_\ell - \widehat{\lambda})^2\,.$$

The following two properties can easily be shown:

$$\forall k\in K,\ \widehat{\lambda}_k = 1 \quad \text{and} \quad \widehat{V}_k(\lambda) = \sum_{\ell\in L}\lambda_\ell r_{k\ell}^2 - 1\,.$$

Theorem 2.2. *The linkage index of variable $\boldsymbol{x}_k$ is proportional to $\widehat{V}_k(\lambda)$:*

$$\mathrm{LI}_k = \tfrac{1}{p-1}\widehat{V}_k(\lambda)\,.$$

Proof. By the reconstitution formula of the correlation matrix (see Le Roux and Rouanet, 2004, p. 153), one has $\mathbf{R} = \mathbf{A}\boldsymbol{\Lambda}\mathbf{A}^\top$, hence $\mathbf{R}^2 = \mathbf{A}\boldsymbol{\Lambda}^2\mathbf{A}^\top$. The k–th diagonal entry of $\mathbf{R}^2$ is equal to $\sum_{k'\in K} r_{kk'}^2 = 1 + \sum_{k'\neq k} r_{kk'}^2$ (see the proof of Theorem 2.1) and also to $\sum_{\ell\in L}\lambda_\ell^2 a_{k\ell}^2 = \sum_\ell \lambda_\ell r_{k\ell}^2$ (since $r_{k\ell} = \sqrt{\lambda_\ell}a_{k\ell}$). Hence $\sum_{\ell\in L}\lambda_\ell r_{k\ell}^2 = \widehat{V}_k(\lambda) + 1 = 1 + \sum_{\substack{k'\in K\\ k'\neq k}} r_{kk'}^2$.

Corollary 2.2.1. *The ratio of the linkage index of variable $\boldsymbol{x}_k$ and the mean linkage index is equal to the ratio of the weighted variance of eigenvalues $\widehat{V}_k(\lambda)$ to the variance of eigenvalues $V(\lambda)$:*

$$\frac{\mathrm{LI}_k}{\overline{\mathrm{LI}}} = \frac{\widehat{V}_k(\lambda)}{V(\lambda)}.$$

Comments

1) The more the linkage index of a variable is *superior* to the mean, the more this variable contributes to *extreme* axes (first and last axes).

2) The more the linkage index of a variable is *inferior* to the mean, the more this variable contributes to *central* axes (axes with variance near 1).

2.4. APPLICATION TO SPEARMAN'S DATA

Table 1 (see Spearman, 1904, p.291) gives the correlations between performance variables of English pupils in the following subjects: *Classics* ($k1$), *French* ($k2$), *English* ($k3$), *Mathematics* ($k4$), *Pitch Discrimination* ($k5$), and *Music* ($k6$).

Table 1: Correlations, linkage indexes LI_k and ratios $\mathrm{LI}_k/\overline{\mathrm{LI}}$.

		$k1$	$k2$	$k3$	$k4$	$k5$	$k6$	LI_k	$\mathrm{LI}_k/\overline{\mathrm{LI}}$
$k1$	*Classics*	1	0.83	0.78	0.70	0.66	0.63	0.524	1.34
$k2$	*French*	0.83	1	0.67	0.67	0.65	0.57	0.467	1.20
$k3$	*English*	0.78	0.67	1	0.64	0.54	0.51	0.404	1.04
$k4$	*Mathematics*	0.70	0.67	0.64	1	0.45	0.51	0.362	0.93
$k5$	*Pitch discrim.*	0.66	0.65	0.54	0.45	1	0.40	0.302	0.78
$k6$	*Music*	0.63	0.57	0.51	0.51	0.40	1	0.280	0.72

$\overline{\mathrm{LI}} = 0.390$

We notice (see Table 1) that *Classics* is the most correlated with other variables with a linkage index equal to 0.524, which is 34% higher than the average. As we can see in Table 2, this variable is the one that contributes the most to axes $\ell 1$ and $\ell 6$, for which variances are the farthest from 1 ("extreme" axes). In sharp contrast, *Music* is the least correlated with other variables (linkage index equal to 0.280) and contributes heavily to axes $\ell 2$ and $\ell 3$, for which variances are the closest to 1 ("central" axes).

Table 2: Eigenvalues λ_ℓ, contributions of variables to axes (in %), variance of eigenvalues weighted by contributions $\widehat{V}_k(\lambda)$ and variance ratios $\widehat{V}_k(\lambda)/V(\lambda)$.

	$\ell 1$	$\ell 2$	$\ell 3$	$\ell 4$	$\ell 5$	$\ell 6$		
λ_ℓ	4.103	0.619	0.512	0.357	0.270	0.139	$\widehat{V}_k(\lambda)$	$\widehat{V}_k(\lambda)/V(\lambda)$
$k1$	21	0	0	2	7	70	2.62	1.34
$k2$	19	1	0	5	54	20	2.33	1.20
$k3$	17	0	12	59	4	9	2.02	1.04
$k4$	16	6	31	32	14	0	1.81	0.93
$k5$	13	51	15	2	18	0	1.51	0.78
$k6$	13	41	42	0	3	1	1.40	0.72

$V(\lambda) = 1.95$

3. MULTIPLE CORRESPONDENCE ANALYSIS

We will now adopt the same approach for MCA.

3.1. BASIC PROPERTIES AND NOTATIONS

Let I denote the set of n individuals and Q the set of categorical variables (questions). The table analyzed by MCA is an $I\times Q$ table such that the entry in cell (i,q) is the category of variable q chosen by individual i. The set of categories of variable q is denoted by K_q and its cardinal by K_q; the overall set of categories is denoted by K and its cardinal by K.

The number of individuals who have chosen category k is denoted by n_k (with $n_k > 0$) and the corresponding relative frequency by $f_k = n_k/n$.

Multiple correspondence on $I\times K$. Let us denote $\delta_{IK} = (\delta_{ik})_{i\in I, k\in K}$ the multiple correspondence on $I\times K$ defined by

$$\delta_{ik} = \begin{cases} 1 & \text{if individual } i \text{ has chosen category } k \\ 0 & \text{if not} \end{cases}.$$

Performing the MCA of the $I\times Q$ table is equivalent to proceeding to Correspondence Analysis of the $I\times K$ table δ_{IK} (Benzécri, 1977; Greenacre, 1984). The solution is given by the diagonalization of the symmetric matrix $\mathbf{S} = [s_{kk'}]$ with $s_{kk'} = \frac{1}{Q}\frac{n_{kk'} - n_k n_{k'}/n}{\sqrt{n_k n_{k'}/n}}$ ($n_{kk'}$ is the number of individuals who have chosen both categories k and k').

We denote L the set indexing the $\mathsf{K}-\mathsf{Q}$ nonnull eigenvalues and $(y_\ell^k)_{\ell\in L}$ the principal coordinates of the category point k. The sum of eigenvalues $(\lambda_\ell)_{\ell\in L}$ is equal to $(\mathsf{K}-\mathsf{Q})/\mathsf{Q}$, hence the mean is $\overline{\lambda} = 1/\mathsf{Q}$.

Burt table and mean square contingency coefficients. The Burt table associated with δ_{IK} is the symmetric $K\times K$ table defined by:

$$b_{kk'} = \sum_{i\in I} \delta_{ik}\,\delta_{ik'} = \begin{cases} n_k & \text{if} \quad k = k' \\ 0 & \text{if} \quad k \neq k' \text{ with } k, k' \in K_q \\ n_{kk'} & \text{if} \quad k \in K_q \text{ and } \; k' \in K_{q'} \text{ with } q \neq q' \end{cases}.$$

Denoting $\Phi^2_{qq'}$ the mean square contingency coefficient of the contingency table crossing variables q and q', one has: $\Phi^2_{qq} = \mathsf{K}_q - 1$ and for $q' \neq q$, $\Phi^2_{qq'} = \sum_{k\in K_q}\sum_{k'\in K_{q'}} \frac{(f_{kk'} - f_k f_{k'})^2}{f_k\, f_{k'}}$.

The Φ^2 of the *Burt table*, denoted Φ^2_{Burt}, is the average of the Φ^2 of the Q^2 subtables of the Burt table. Denoting $\overline{\Phi}^2$ the mean of the Φ^2 of the

$\mathsf{Q}(\mathsf{Q}-1)$ non–diagonal subtables, one has:

$$\Phi^2_{\text{Burt}} = \frac{1}{\mathsf{Q}^2}\sum_{q\in Q}\sum_{q'\in Q}\Phi^2_{qq'} = \frac{1}{\mathsf{Q}}\Big(\frac{\mathsf{K}-\mathsf{Q}}{\mathsf{Q}} + \frac{1}{\mathsf{Q}}\sum_{q\in Q}\sum_{\substack{q'\in Q\\ q'\neq q}}\Phi^2_{qq'}\Big) = \frac{1}{\mathsf{Q}}\frac{\mathsf{K}-\mathsf{Q}}{\mathsf{Q}} + \frac{\mathsf{Q}-1}{\mathsf{Q}}\overline{\Phi}^2.$$

Contributions of categories and of variables. The squared distance between the *category point* k and the mean point of the cloud is equal to $\frac{1-f_k}{f_k} = \sum_{\ell\in L}(y^k_\ell)^2$.

The *contribution of category* k to axis ℓ, denoted Ctr^ℓ_k, is equal to $\frac{f_k}{\mathsf{Q}}\frac{(y^k_\ell)^2}{\lambda_\ell}$. We have the two following properties:

$$\forall \ell\in L,\ \sum_{k\in K}\text{Ctr}^\ell_k = 1 \quad\text{and}\quad \forall k\in K,\ \sum_{\ell\in L}\text{Ctr}^\ell_k = 1-f_k\ .$$

The first property follows the definition of contribution to axes. The second one can be proven as follows: The ℓ-th unit eigenvector of $\mathbf{S}$ associated with nonnull eigenvalue λ_ℓ is $(c_{k\ell})_{k\in K}$ with $c_{k\ell} = \sqrt{f_k/\mathsf{Q}}(y^k_\ell/\sqrt{\lambda_\ell})$ and the Q ones associated with null eigenvalue are $(c_{kq})_{q\in Q}$ with $c_{kq} = \sqrt{f_k}$ for $k\in K_q$ and 0 for $k\notin K_q$. Hence $\sum_{\ell\in L}c^2_{k\ell} + \sum_{q\in Q}c^2_{kq} = \frac{f_k}{\mathsf{Q}}\sum_{\ell\in L}\frac{(y^k_\ell)^2}{\lambda_\ell} + f_k = 1$.

By definition, the *contribution of a variable* to axis ℓ is the sum of the contributions of its categories: $\text{Ctr}^\ell_q = \sum_{k\in K_q}\text{Ctr}^\ell_k$, and we have the two following properties:

$$\forall \ell\in L,\ \sum_{q\in Q}\text{Ctr}^\ell_q = 1 \quad\text{and}\quad \forall q\in Q,\ \sum_{\ell\in L}\text{Ctr}^\ell_q = \mathsf{K}_q - 1\,.$$

Burt cloud. The mean point of the subcloud of individuals who have chosen category k is called *category mean point*. Its profile (obtained from the Burt table) is equal to $\frac{1}{Q}(f_{kk'}/f_k)_{k'\in K}$; its squared distance to the mean point is equal to $\sum_{k'\in K}\frac{1}{\mathsf{Q}^2}\frac{(f_{kk'}/f_k - f_{k'})^2}{f'_k/\mathsf{Q}} = \frac{1}{\mathsf{Q}f_k}\sum_{k'\in K}\frac{(f_{kk'}-f_kf_{k'})^2}{f_kf_{k'}}$. Letting $\phi^2_{q'}(k) = \sum_{k'\in K_{q'}}\frac{(f_{kk'}-f_kf_{k'})^2}{f_kf_{k'}}$ if $k\in K_q$ and $q'\neq q$, the squared distance writes:

$$\frac{1}{\mathsf{Q}f_k}(1-f_k) + \frac{1}{\mathsf{Q}f_k}\sum_{\substack{q'\in Q\\ q'\neq q}}\phi^2_{q'}(k), \quad\text{with}\quad \Phi^2_{qq'} = \sum_{k\in K_q}\phi^2_{q'}(k). \tag{2}$$

The K category mean points define the Burt cloud (see Le Roux and Rouanet, 2004, pp. 199-200). The principal coordinates of the category mean point k on axis ℓ are equal to $y_\ell^k/\sqrt{\lambda_\ell}$, hence its squared distance to the mean point is also equal to $\sum_{\ell \in L} (y_\ell^k)^2/\lambda_\ell$.

The eigenvalues verify the following property: $\sum_{\ell=1}^{L} \lambda_\ell^2 = \Phi^2_{\text{Burt}}$.

3.2. VARIANCE OF EIGENVALUES

Theorem 3.1. *The variance of eigenvalues, denoted $V(\lambda)$, is such that:*

$$V(\lambda) = \frac{1}{\mathsf{K}-\mathsf{Q}} \frac{\mathsf{Q}-1}{\mathsf{Q}} \overline{\Phi}^2 .$$

Proof. $\frac{1}{\mathsf{K}-\mathsf{Q}} \sum_{\ell \in L} (\lambda_\ell - \overline{\lambda})^2 = \frac{1}{\mathsf{K}-\mathsf{Q}} \Phi^2_{\text{Burt}} - \overline{\lambda}^2$ with $\overline{\lambda} = 1/\mathsf{Q}$. Hence the variance is $V(\lambda) = \frac{1}{\mathsf{K}-\mathsf{Q}} \Big(\frac{1}{\mathsf{Q}} \frac{\mathsf{K}-\mathsf{Q}}{\mathsf{Q}} + \frac{\mathsf{Q}-1}{\mathsf{Q}} \overline{\Phi}^2 \Big) - \frac{1}{\mathsf{Q}^2}$.

This expression of the variance of eigenvalues leads us to consider the mean of the Φ^2 between one categorical variable and the others, that is, it leads us to the following definition.

Definition 3.1 (Linkage index of categorical variable)**.** *The linkage index of categorical variable q, denoted* LI_q*, is such that:*

$$\text{LI}_q = \frac{1}{\mathsf{K}_q - 1} \Big(\frac{1}{\mathsf{Q}-1} \sum_{q' \in Q, q' \neq q} \Phi^2_{qq'} \Big).$$

Note that the linkage index of a categorical variable is between zero and one, since $\Phi^2_{qq'} \leq \mathsf{K}_q - 1$.

Property 1 (Mean linkage index)**.** *The mean linkage index of the Q categorical variables weighted by $(\mathsf{K}_q - 1)_{q \in Q}$, denoted $\overline{\text{LI}}$, is such that:*

$$\overline{\text{LI}} = \frac{\mathsf{Q}^2}{\mathsf{Q}-1} V(\lambda) = \overline{\Phi}^2 / \Big(\frac{\mathsf{K}-\mathsf{Q}}{\mathsf{Q}} \Big) .$$

Proof. $\sum_{q \in Q} (\mathsf{K}_q - 1) = \mathsf{K} - \mathsf{Q}$, hence the weighted mean of linkage indexes is $\frac{1}{\mathsf{K}-\mathsf{Q}} \sum_{q \in Q} (\mathsf{K}_q - 1) \text{LI}_q = \frac{1}{\mathsf{K}-\mathsf{Q}} \sum_{q \in Q} \frac{1}{\mathsf{Q}-1} \sum_{q' \in Q, q' \neq q} \Phi^2_{qq'} = \frac{\mathsf{Q}^2}{\mathsf{Q}-1} V(\lambda)$.

Definition 3.2 (Linkage index of category)**.** *Given a category k of the categorical variable q, the linkage index of category k, denoted* LI_k*, is defined as follows:*

$$\mathrm{LI}_k = \tfrac{1}{1-f_k} \times \tfrac{1}{\mathsf{Q}-1} \sum_{q' \in Q, q' \neq q} \phi^2_{q'}(k)$$

with for $q' \neq q$, $\phi^2_{q'}(k) = \sum_{k' \in K_{q'}} \frac{(f_{kk'} - f_k f_{k'})^2}{f_k f_{k'}}$.

One deduces from $\sum_{k \in K_q} \phi^2_{q'}(k) = \Phi^2_{qq'}$ that the linkage index of variable q is equal to the mean of the linkage indexes of its categories weighted by $(1-f_k)_{k \in K_q}$: $\mathrm{LI}_q = \frac{1}{\mathsf{K}_q - 1} \sum_{k \in K_q} (1-f_k)\mathrm{LI}_k$.

3.3. CATEGORIES, CATEGORICAL VARIABLES AND EIGENVALUES

In order to explain the link between categorical variables or categories and eigenvalues, we will now express the linkage indexes in terms of eigenvalues weighted by contributions to axes.

Lemma 3.1. *The mean of eigenvalues weighted by the contributions of category k to axes, denoted $\widehat{\lambda}_k$, is equal to $1/\mathsf{Q}$.*

Proof. $\widehat{\lambda}_k = \sum_{\ell \in L} \mathrm{Ctr}^\ell_k \, \lambda_\ell / (\sum_{\ell \in L} \mathrm{Ctr}^\ell_k) = \frac{1}{1-f_k} \frac{f_k}{\mathsf{Q}} \sum_{\ell \in L} (y^k_\ell)^2 = \frac{1}{1-f_k} \frac{f_k}{\mathsf{Q}} \frac{1-f_k}{f_k} = \frac{1}{\mathsf{Q}}$.

Lemma 3.2. *The variance of eigenvalues weighted by the contributions of category k of variable q to axes is denoted $\widehat{V}_k(\lambda)$ and called k–variance of eigenvalues; it is equal to* $\frac{1}{\mathsf{Q}^2(1-f_k)} \sum_{q' \in Q, q' \neq q} \phi^2_{q'}(k)$.

Proof. The weighted sum of the squared eigenvalues is equal to $\sum_{\ell \in L} \frac{f_k}{Q} (y^k_\ell)^2 \lambda_\ell = \sum_{\ell \in L} \frac{f_k}{Q} \big(\frac{y^k_\ell}{\sqrt{\lambda_\ell}}\big)^2 = \frac{f_k}{\mathsf{Q}} \Big[\frac{1}{\mathsf{Q} f_k} (1 - f_k) + \frac{1}{\mathsf{Q} f_k} \sum_{q' \neq q} \phi^2_{q'}(k) \Big]$ (Equation 2). Hence $\widehat{V}_k(\lambda) = \frac{1}{\mathsf{Q}^2} \big[(1-f_k) + \sum_{q' \neq q} \phi^2_{q'}(k) \big] / (1-f_k) - \frac{1}{\mathsf{Q}^2} = \frac{1}{\mathsf{Q}^2(1-f_k)} \sum_{q' \neq q} \phi^2_{q'}(k)$.

Theorem 3.2. *The linkage index of category k is proportional to the variance of eigenvalues weighted by contributions of k to axes.*

$$\mathrm{LI}_k = \frac{\mathsf{Q}^2}{\mathsf{Q}-1} \widehat{V}_k(\lambda).$$

Proof. $\widehat{V}_k(\lambda) = \frac{\mathsf{Q}-1}{\mathsf{Q}^2}\big[\frac{1}{(\mathsf{Q}-1)(1-f_k)}\sum_{q'\neq q}\phi^2_{q'}(k)\big] = \frac{\mathsf{K}_q-1}{\mathsf{Q}^2}\mathrm{LI}_k.$

From Property 1 and Property 3.2 we deduce that:

Corollary 3.2.1. *The ratio of the linkage index of category k to the mean linkage index is equal to the ratio of the k–variance of eigenvalues to the variance of eigenvalues:*

$$\frac{\mathrm{LI}_k}{\overline{\mathrm{LI}}} = \frac{\widehat{V}_k(\lambda)}{V(\lambda)}.$$

The *properties about categorical variables* follows the property of average of linkage indexes of categories ($\mathrm{LI}_q = \frac{1}{\mathsf{K}_q-1}\sum_{k\in K_q}(1-f_k)\mathrm{LI}_k$) and of the property of sum of contributions ($\mathrm{Ctr}^\ell_q = \sum_{k\in K_q}\mathrm{Ctr}^\ell_k$). We denote $\widehat{V}_q(\lambda)$ the q–variance of eigenvalues (variance of eigenvalues weighted by the contributions of variable q): $\widehat{V}_q(\lambda) = \frac{1}{\mathsf{K}_q-1}\sum_{\ell\in L}\mathrm{Ctr}^\ell_q(\lambda_\ell - \frac{1}{\mathsf{Q}})^2$. One has the following property: $\mathrm{LI}_q = \frac{\mathsf{Q}^2}{\mathsf{Q}-1}\widehat{V}_q(\lambda)$. Hence:

$$\frac{\mathrm{LI}_q}{\overline{\mathrm{LI}}} = \frac{\widehat{V}_q(\lambda)}{V(\lambda)}.$$

Comments

1) The more the linkage index of a category (or a categorical variable) is *superior* to the mean, the more this category (or this variable) contributes to *extreme* axes (first and last axes).

2) The more the linkage index of a category (or a categorical variable) is *inferior* to the mean, the more this category (or this variable) contributes to *central* axes (axes with variance near $1/\mathsf{Q}$).

3.4. APPLICATION TO BURT'S DATA

Burt's data (Table 3), reproduced from Burt (1950, p.171), gives, for 100 individuals (men living in Liverpool), the observed response patterns and their absolute frequencies for four attributes (categorical variables), that is, *A Hair* ($a1$: *fair*, $a2$: *red*, $a3$: *dark*), *B Eyes* ($b1$: *light*, $b2$: *mixed*, $b3$: *brown*), *C Head* ($c1$: *narrow*, $c2$: *wide*), *D Stature* ($d1$: *tall*, $d2$: *short*).

As we can see in Table 4, the categories having the highest linkage indexes are the category $b1$ (*light*) of *Eyes* and the two categories of *Stature*.

Table 3: Observed response patterns with their absolute frequencies.

	Abs.freq
$a1b1c1d1$	8
$a1b1c1d2$	4
$a1b1c2d1$	2
$a1b2c1d1$	1
$a1b2c1d2$	1
$a1b2c2d1$	2
$a1b2c2d2$	2
$a1b3c2d2$	2

	Abs.freq
$a2b1c1d1$	6
$a2b1c2d1$	2
$a2b2c1d1$	2
$a2b2c1d2$	1
$a2b2c2d2$	2
$a2b3c1d2$	2

	Abs.freq
$a3b1c1d1$	9
$a3b1c2d1$	2
$a3b2c1d1$	3
$a3b2c1d2$	12
$a3b2c2d1$	2
$a3b2c2d2$	8
$a3b3c1d1$	1
$a3b3c1d2$	19
$a3b3c2d1$	3
$a3b3c2d2$	4

$a1$	$a2$	$a3$	$b1$	$b2$	$b3$	$c1$	$c2$	$d1$	$d2$
22	15	63	33	36	31	69	31	43	57

Table 4: Eigenvalues λ_ℓ, linkage index LI_k, ratio $\mathrm{LI}_k/\overline{\mathrm{LI}}$ and contributions of categories to axes (in %).

				$\ell 1$	$\ell 2$	$\ell 3$	$\ell 4$	$\ell 5$	$\ell 6$
			λ_ℓ	0.489	0.299	0.254	0.206	0.179	0.073
		LI_k	$\mathrm{LI}_k/\overline{\mathrm{LI}}$						
$a1$	*fair*	.054	0.63	9	5	36	1	23	5
$a2$	*red*	.027	0.31	6	0	55	4	20	0
$a3$	*dark*	.099	1.14	9	1	0	0	25	2
$b1$	*light*	.211	2.42	26	2	1	0	1	36
$b2$	*mixed*	.037	0.43	3	26	5	24	0	5
$b3$	*brown*	.093	1.07	12	16	2	23	3	14
$c1$	*narrow*	.020	0.23	0	15	0	14	0	1
$c2$	*wide*	.020	0.23	1	34	0	31	1	2
$d1$	*tall*	.170	1.95	20	0	0	2	16	20
$d2$	*short*	.170	1.95	15	0	0	1	12	15

$\overline{\mathrm{LI}} = 0.087$

Table 5: Eigenvalues λ_ℓ, linkage index LI_q, ratio $\mathrm{LI}_q/\overline{\mathrm{LI}}$ and contributions of categorical variables to axes (in %).

				$\ell 1$	$\ell 2$	$\ell 3$	$\ell 4$	$\ell 5$	$\ell 6$
			λ_ℓ	0.489	0.299	0.254	0.206	0.179	0.073
		LI_q	$\mathrm{LI}_q/\overline{\mathrm{LI}}$						
$q1$	*Hair*	.051	0.59	23	6	91	5	67	7
$q2$	*Eyes*	.115	1.32	41	45	8	47	4	55
$q3$	*Head*	.020	0.23	1	49	0	45	1	3
$q4$	*Stature*	.170	1.95	34	0	0	3	28	35

$\overline{\mathrm{LI}} = 0.087$

Their linkage indexes are about twice the mean ($\mathrm{LI}_{b1}/\overline{\mathrm{LI}} = 2.42$ and $\mathrm{LI}_{d1}/\overline{\mathrm{LI}} = \mathrm{LI}_{d2}/\overline{\mathrm{LI}} = 1.95$). These three categories contribute heavily to "extreme" axes $\ell 1$ and $\ell 6$ (together, they account for 61% of axis 1 and 71% of axis 2). In contrast, both categories of *Head* and the category $a2$ (*red*) of *Hair* have the smallest linkage indexes, less than a third of the mean ($\mathrm{LI}_{c1}/\overline{\mathrm{LI}} = \mathrm{LI}_{c2}/\overline{\mathrm{LI}} = 0.23$ and $\mathrm{LI}_{a2}/\overline{\mathrm{LI}} = 0.31$). The contributions of these three categories to both "extreme" axes ($\ell 1$ and $\ell 6$) are very small (7% and 3%, respectively) but they contribute heavily to "central" axes $\ell 2$, $\ell 3$ and $\ell 4$ (49%, 55% and 50%, respectively).

In Table 5, we see that *Head* has a very small linkage index; this variable does not contribute to the first axis (neither to the 5th and the 6th axes).

4. CONCLUSION

In this paper, we emphasize that the higher the mean of the linkage indexes of (numerical or categorical) variables, the higher the variance of eigenvalues, that is, the larger the departure of clouds from sphericity.

In addition, further analysis shows that the more the linkage index of a variable is superior to the mean, the more this variable contributes to extreme axes (first and last axes); otherwise this variable contributes to *central* axes (whose variances are close to the mean). So, if the range of linkage indexes of variables is large, one can predict that the variables with the greatest linkage indexes will play a preponderant role in the interpretation of first axes. Then, if we decide to reduce the number of active variables in the analysis, linkage indexes will be a useful tool: if a variable with a weak linkage index is discarded, the proportion of variance associated with the first axes will increase and the interpretation will remain the same.

REFERENCES

Benzécri, J.-P. (1977). Sur l'analyse des tableaux binaires associés à une correspondance multiple. *Les cahiers de l'analyse des données,* 2(1): 55–71, from a mimeographed note of 1972.

Burt, C. (1950). The factorial analysis of qualitative data. *British Journal of Statistical Psychology,* 3: 166–185.

Durand, J.-L. (1998). Taux de dispersion des valeurs propres en ACP, AC et ACM. *Mathématiques Informatique et Sciences humaines,* 144: 15–28.

Greenacre, M. (1984). *Theory and Applications of Correspondence Analysis.* London: Academic press.

Karlis, D., Saporta, G. and Spinakis, A. (2003). A simple rule for the selection of principal components. *Communications in Statistics-Theory and Methods,* 32(3): 643–666.

Le Roux, B. and Rouanet, H. (2004). *Geometric Data Analysis. From Correspondence Analysis to Structured Data Analysis.* Dordrecht: Kluwer.

Morrison, D. (1976). *Multivariate Statistical Methods.* New York: McGraw-Hill Publ. Co.

Pavlicev, M ., Cheverud, J. and Wagner, G . (2009). M easuring morphological integration using eigenvalue variance. *Evolutionary Biology*, 36(1): 157–170.

Saporta, G. (2003). A control chart approach to select eigenvalues in principal component and correspondence analysis. *54th Session of the International Statistical Institute-Berlin.*

Spearman, C. (1904). 'General intelligence', objectively determined and measured. *American Journal of Psychology*, 15: 201–292.