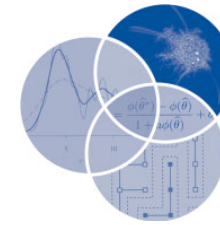
Overview



Principal component analysis

Hervé Abdi[1]* and Lynne J. Williams[2]

Principal component analysis (PCA) is a multivariate technique that analyzes a data table in which observations are described by several inter-correlated quantitative dependent variables. Its goal is to extract the important information from the table, to represent it as a set of new orthogonal variables called principal components, and to display the pattern of similarity of the observations and of the variables as points in maps. The quality of the PCA model can be evaluated using cross-validation techniques such as the bootstrap and the jackknife. PCA can be generalized as *correspondence analysis* (CA) in order to handle qualitative variables and as *multiple factor analysis* (MFA) in order to handle heterogeneous sets of variables. Mathematically, PCA depends upon the eigen-decomposition of positive semi-definite matrices and upon the singular value decomposition (SVD) of rectangular matrices. © 2010 John Wiley & Sons, Inc. *WIREs Comp Stat* 2010 2 433–459

Principal component analysis (PCA) is probably the most popular multivariate statistical technique and it is used by almost all scientific disciplines. It is also likely to be the oldest multivariate technique. In fact, its origin can be traced back to Pearson[1] or even Cauchy[2] [see Ref 3, p. 416], or Jordan[4] and also Cayley, Silverster, and Hamilton, [see Refs 5,6, for more details] but its modern instantiation was formalized by Hotelling[7] who also coined the term *principal component*. PCA analyzes a data table representing observations described by several dependent variables, which are, in general, inter-correlated. Its goal is to extract the important information from the data table and to express this information as a set of new

orthogonal variables called *principal components*. PCA also represents the pattern of similarity of the observations and the variables by displaying them as points in maps [see Refs 8–10 for more details].

PREREQUISITE NOTIONS AND NOTATIONS

Matrices are denoted in upper case bold, vectors are denoted in lower case bold, and elements are denoted in lower case italic. Matrices, vectors, and elements

*Correspondence to: herve@utdallas.edu

[1]School of Behavioral and Brain Sciences, The University of Texas at Dallas, MS: GR4.1, Richardson, TX 75080-3021, USA

[2]Department of Psychology, University of Toronto Scarborough, Ontario, Canada

DOI: 10.1002/wics.101

from the same matrix all use the same letter (e.g., **A**, **a**, *a*). The transpose operation is denoted by the superscript^{T}. The identity matrix is denoted **I**.

The data table to be analyzed by PCA comprises I observations described by J variables and it is represented by the $I \times J$ matrix **X**, whose generic

element is $x_{i,j}$. The matrix $\mathbf{X}$ has rank L where $L \leq \min\{I, J\}$.

In general, the data table will be preprocessed before the analysis. Almost always, the columns of $\mathbf{X}$ will be centered so that the mean of each column is equal to 0 (i.e., $\mathbf{X}^{\mathsf{T}}\mathbf{1} = \mathbf{0}$, where $\mathbf{0}$ is a J by 1 vector of zeros and $\mathbf{1}$ is an I by 1 vector of ones). If in addition, each element of $\mathbf{X}$ is divided by $\sqrt{I}$ (or $\sqrt{I-1}$), the analysis is referred to as a *covariance* PCA because, in this case, the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is a covariance matrix. In addition to centering, when the variables are measured with different units, it is customary to standardize each variable to unit norm. This is obtained by dividing each variable by its norm (i.e., the square root of the sum of all the squared elements of this variable). In this case, the analysis is referred to as a *correlation* PCA because, then, the matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ is a correlation matrix (most statistical packages use correlation preprocessing as a default).

The matrix $\mathbf{X}$ has the following singular value decomposition [SVD, see Refs 11–13 and Appendix B for an introduction to the SVD]:

$$\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} \tag{1}$$

where $\mathbf{P}$ is the $I \times L$ matrix of left singular vectors, $\mathbf{Q}$ is the $J \times L$ matrix of right singular vectors, and $\mathbf{\Delta}$

is the diagonal matrix of singular values. Note that $\mathbf{\Delta}^2$ is equal to $\mathbf{\Lambda}$ which is the diagonal matrix of the (nonzero) eigenvalues of $\mathbf{X}^T\mathbf{X}$ and $\mathbf{X}\mathbf{X}^T$.

The *inertia of a column* is defined as the sum of the squared elements of this column and is computed as

$$\gamma_j^2 = \sum_i^I x_{i,j}^2 \,. \tag{2}$$

The sum of all the γ_j^2 is denoted $\mathcal{I}$ and it is called the *inertia* of the data table or the *total inertia*. Note that the total inertia is also equal to the sum of the squared singular values of the data table (see Appendix B).

The *center of gravity of the rows* [also called centroid or barycenter, see Ref 14], denoted $\mathbf{g}$, is the vector of the means of each column of $\mathbf{X}$. When $\mathbf{X}$ is

centered, its center of gravity is equal to the $1 \times J$ row vector $\mathbf{0}^{\mathsf{T}}$.

The (Euclidean) distance of the i-th observation to $\mathbf{g}$ is equal to

$$d^2_{i,\mathbf{g}} = \sum_j^J \left(x_{i,j} - g_j\right)^2 . \tag{3}$$

When the data are centered Eq. 3 reduces to

$$d^2_{i,\mathbf{g}} = \sum_j^J x^2_{i,j}. \tag{4}$$

Note that the sum of all $d^2_{i,\mathbf{g}}$ is equal to $\mathcal{I}$ which is the inertia of the data table .

GOALS OF PCA

The goals of PCA are to

(1) extract the most important information from the data table;

(2) compress the size of the data set by keeping only

this important information;

(3) simplify the description of the data set; and

(4) analyze the structure of the observations and the variables.

In order to achieve these goals, PCA computes new variables called *principal components* which are obtained as linear combinations of the original variables. The first principal component is required to have the largest possible variance (i.e., inertia and therefore this component will 'explain' or 'extract' the largest part of the inertia of the data table). The second component is computed under the constraint of being orthogonal to the first component and to have the largest possible inertia. The other components are computed likewise (see Appendix A for proof). The values of these new variables for the observations are called *factor scores*, and these factors scores can be interpreted geometrically as the *projections* of the observations onto the principal components.

Finding the Components

In PCA, the components are obtained from the SVD

of the data table $\mathbf{X}$. Specifically, with $\mathbf{X} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}$ (*cf.* Eq. 1), the $I \times L$ matrix of factor scores, denoted $\mathbf{F}$, is obtained as:

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta}. \tag{5}$$

The matrix $\mathbf{Q}$ gives the coefficients of the linear combinations used to compute the factors scores. This matrix can also be interpreted as a *projection* matrix because multiplying $\mathbf{X}$ by $\mathbf{Q}$ gives the values of the *projections* of the observations on the principal components. This can be shown by combining Eqs. 1 and 5 as:

$$\mathbf{F} = \mathbf{P}\boldsymbol{\Delta} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{X}\mathbf{Q}. \tag{6}$$

The components can also be represented geometrically by the rotation of the original axes. For example, if $\mathbf{X}$ represents two variables, the length of a word (Y) and the number of lines of its dictionary definition (W), such as the data shown in Table 1, then PCA represents these data by two orthogonal factors. The geometric representation of PCA is shown in Figure 1. In this figure, we see that the factor scores give the length (i.e., distance to the origin) of the

projections of the observations on the components. This procedure is further illustrated in Figure 2. In this context, the matrix $\mathbf{Q}$ is interpreted as a matrix of direction cosines (because $\mathbf{Q}$ is orthonormal). The matrix $\mathbf{Q}$ is also called a *loading* matrix. In this context, the matrix $\mathbf{X}$ can be interpreted as the product of the factors score matrix by the loading matrix as:

$$\mathbf{X} = \mathbf{F}\mathbf{Q}^{\mathsf{T}} \quad \text{with } \mathbf{F}^{\mathsf{T}}\mathbf{F} = \boldsymbol{\Delta}^2 \text{ and } \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}. \qquad (7)$$

This decomposition is often called the *bilinear* decomposition of $\mathbf{X}$ [see, e.g., Ref 15].

s of the Observations to the Components, Squared Distances to the Center of

	F_1^2	F_2^2	d^2	$\cos_1^2 \times 100$	$\cos_2^2 \times 100$
	44.52	0.48	45	99	1
	0.71	0.29	1	71	29
	21.89	3.11	25	88	12
	0.71	0.29	1	71	29
	8.95	8.05	17	53	47
	24.85	0.15	25	99	1
	0	0.00	0	0	0
	9.41	0.59	10	94	6
	17.15	0.85	18	95	5
	1.15	2.85	4	29	71
	31.35	5.65	37	85	15

m the Mean, Coordinates, Squared Coordinates on the Components, Contribution
ervations for the Example Length of Words (Y) and Number of Lines (W)

y	w	F_1	F_2	$\text{ctr}_1 \times 100$	$\text{ctr}_2 \times 100$
−3	6	6.67	0.69	*11*	1
0	−1	−0.84	−0.54	0	1
−4	3	4.68	−1.76	*6*	**6**
0	1	0.84	0.54	0	1
−4	1	2.99	−2.84	2	**15**
3	−4	−4.99	0.38	**6**	0
0	0	0.00	0.00	0	0
−1	3	3.07	0.77	3	1
3	−3	−4.14	0.92	5	2
−2	0	1.07	−1.69	0	5
1	−6	−5.60	−2.38	**8**	**11**

TABLE 1 | Raw Scores, Deviations fro
Gravity, and Squared Cosines of the Obse

	Y	*W*	
Bag	3	14	
Across	6	7	
On	2	11	
Insane	6	9	
By	2	9	
Monastery	9	4	
Relief	6	8	
Slope	5	11	
Scoundrel	9	5	
With	4	8	
Neither	7	2	

5.71	4.29	41	90	10
5.30	1.70	17	90	10
3.68	1.32	5	74	26
2.59	6.41	9	29	71
5.49	1.51	58	97	3
0.49	1.51	32	95	5
2.61	3.39	26	87	13
3.71	4.29	53	92	8
4.71	5.29	20	74	26
392	52	444		
λ_1	λ_2	$\mathcal{I}$		

e squared cosines are multiplied by 100 for ease of reading. The *positive*

−6.06	2.07	**9**	*8*	36
3.91	1.30	4	3	15
1.92	−1.15	1	3	3
1.61	−2.53	1	**12**	2
−7.52	−1.23	**14**	3	56
−5.52	1.23	**8**	3	30
4.76	1.84	*6*	*7*	22
6.98	2.07	*12*	*8*	48
−3.83	2.30	4	*10*	14
0	0	100	100	3

abel the columns: $w = (W - M_W)$; $y = (Y - M_Y)$. The contributions and the
tant contributions are represented in bold.

Overview
Volume 2, July/August 2010

Pretentious	11	4	5	−4
Solid	5	12	−1	4
This	4	9	−2	1
For	3	8	−3	0
Therefore	9	1	3	−7
Generality	10	4	4	−4
Arise	5	13	−1	5
Blot	4	15	−2	7
Infectious	10	6	4	−2
$\sum$	120	160	0	0

$M_W = 8$, $M_Y = 6$. The following abbreviations are used to l
important contributions are italicized , and the **negative** impor

© 2010 John Wiley & Sons, Inc.

Components

Equation 6 shows that matrix $\mathbf{Q}$ is a projection matrix which transforms the original data matrix into factor scores. This matrix can also be used to compute factor scores for observations that were not included in the PCA. These observations are called *supplementary* or *illustrative* observations. By contrast, the observations actually used to compute the PCA are called *active* observations. The factor scores for supplementary observations are obtained by first positioning these observations into the PCA space and then projecting them onto the principal components. Specifically a $1 \times J$ row vector $\mathbf{x}_{\text{sup}}^{\mathsf{T}}$, can be projected into the PCA space using Eq. 6. This gives the $1 \times L$ vector of factor scores, denoted $\mathbf{f}_{\text{sup}}^{\mathsf{T}}$, which is computed as:

$$\mathbf{f}_{\text{sup}}^{\mathsf{T}} = \mathbf{x}_{\text{sup}}^{\mathsf{T}} \mathbf{Q}. \tag{8}$$

If the data table has been preprocessed (e.g., centered or normalized), the same preprocessing should be applied to the supplementary observations *prior* to the computation of their factor scores.

As an illustration, suppose that—in addition to

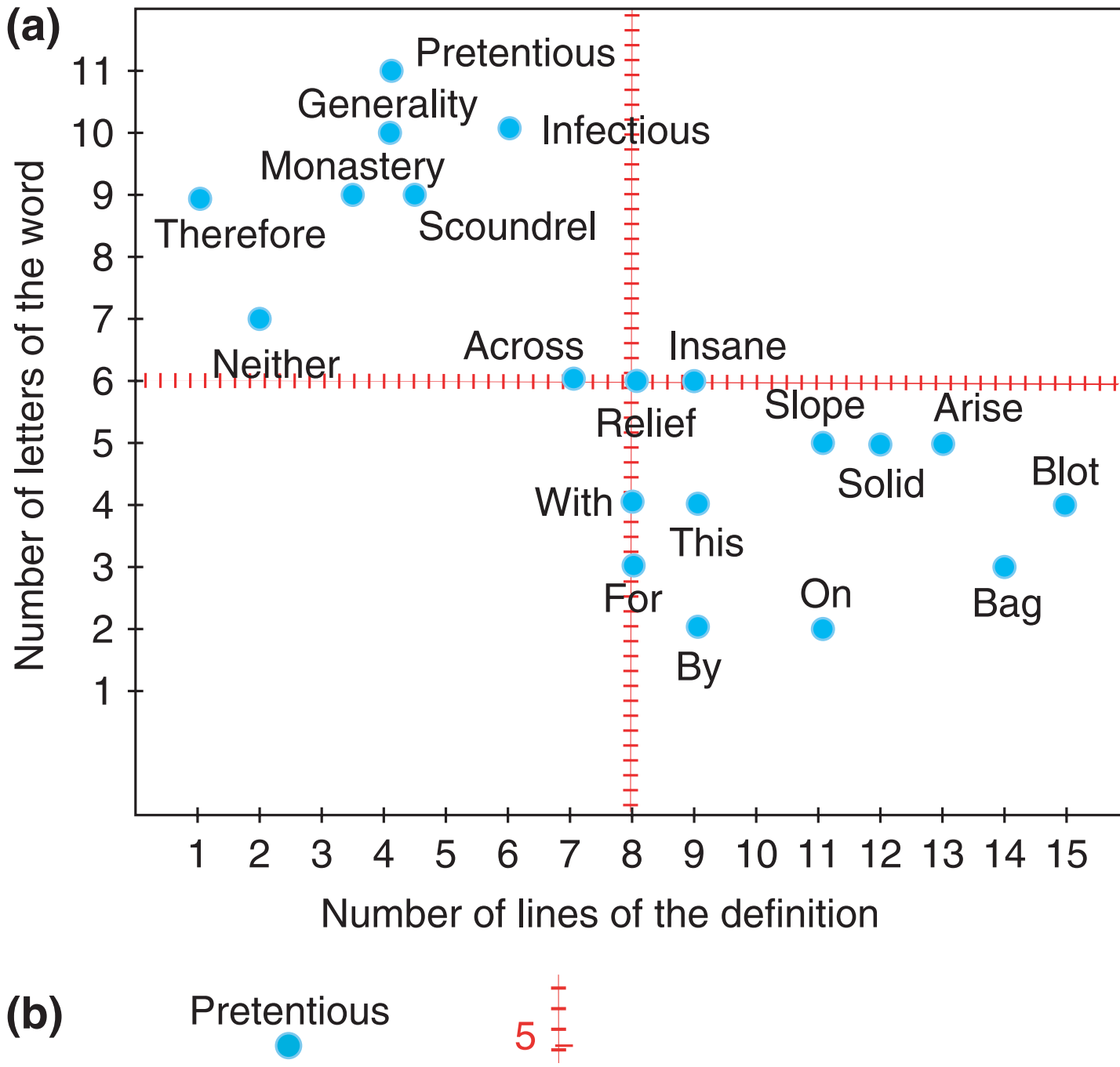
the data presented in Table 1—we have the French word *'sur'* (it means 'on'). It has $Y_{\text{sur}} = 3$ letters, and our French dictionary reports that its definition has $W_{\text{sur}} = 12$ lines. Because *sur* is not an English word, we do not want to include it in the analysis, but we would like to know how it relates to the English vocabulary. So, we decided to treat this word as a supplementary observation.

The first step is to preprocess this supplementary observation in a identical manner to the active observations. Because the data matrix was centered, the values of this observation are transformed into deviations from the English center of gravity. We find the following values:

$$\begin{aligned} y_{\text{sur}} &= Y_{\text{sur}} - M_Y = 3 - 6 = -3 \quad \text{and} \\ w_{\text{sur}} &= W_{\text{sur}} - M_W = 12 - 8 = 4. \end{aligned}$$

Then we plot the supplementary word in the graph that we have already used for the active analysis. Because the principal components and the original variables are in the same space, the projections of the supplementary observation give its coordinates (i.e., factor scores) on the components. This is shown in

Figure 3. Equivalently, the coordinates of the projections on the components can be directly computed



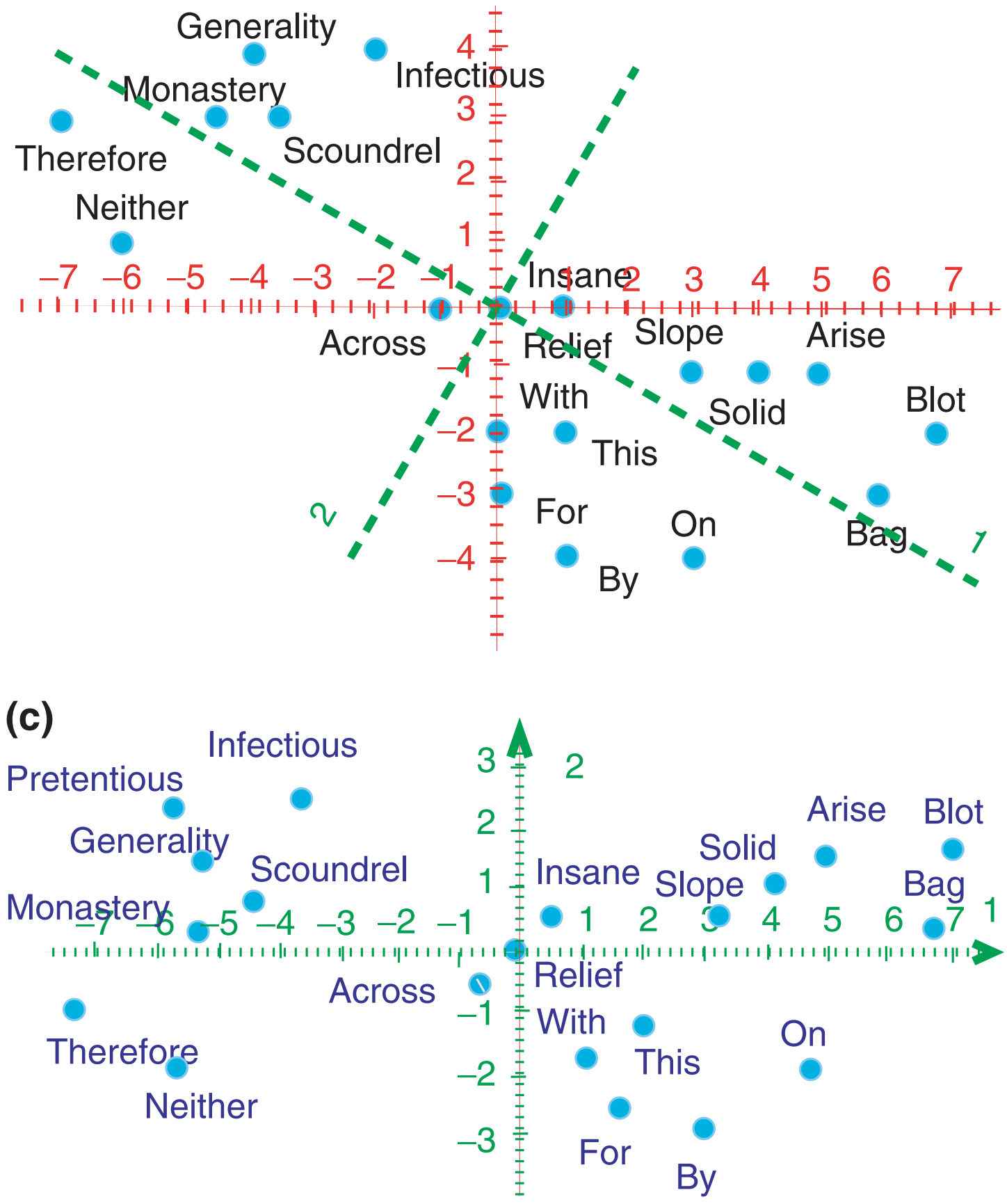


FIGURE 1 | The geometric steps for finding the components of a principal component analysis. To find the components (1) center the variables then plot them against each other. (2) Find the main direction

(called the first component) of the cloud of points such that we have the minimum of the sum of the squared distances from the points to the component. Add a second component orthogonal to the first such that the sum of the squared distances is minimum. (3) When the components have been found, rotate the figure in order to position the first component horizontally (and the second component vertically), then erase the original axes. Note that the final graph could have been obtained directly by plotting the observations from the coordinates given in Table 1.

© 2010 John Wiley & Sons, Inc.

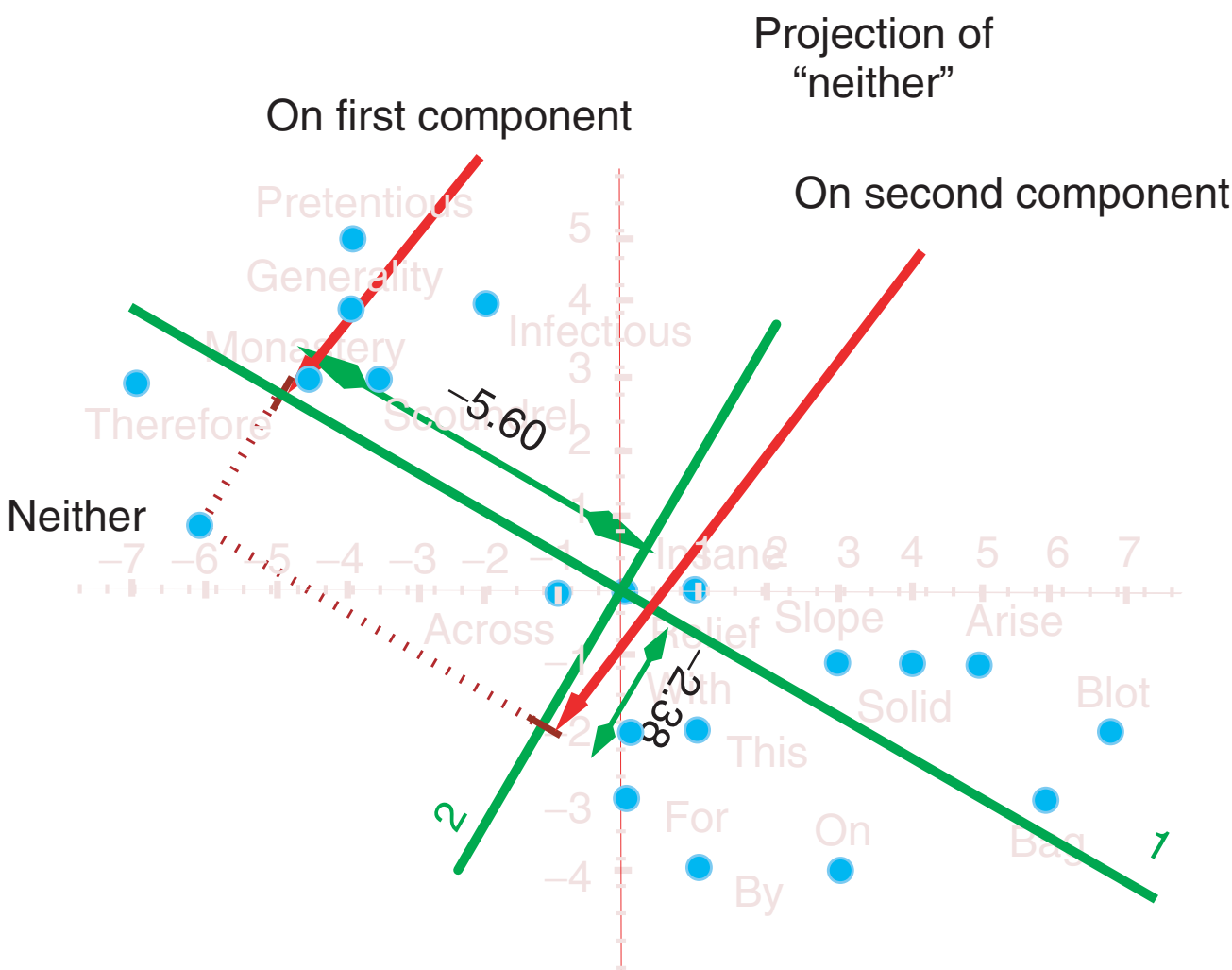


FIGURE 2 | Plot of the centered data, with the first and second components. The projections (or coordinates) of the word 'neither' on the first and the second components are equal to −5.60 and −2.38.

from Eq. 8 (see also Table 3 for the values of Q) as:

$$\mathbf{f}_{\text{sup}}^{\mathsf{T}} = \mathbf{x}_{\text{sup}}^{\mathsf{T}}\mathbf{Q} = \begin{bmatrix} -3 & 4 \end{bmatrix} \times \begin{bmatrix} -0.5369 & 0.8437 \\ 0.8437 & 0.5369 \end{bmatrix} = \begin{bmatrix} 4.9853 & -0.3835 \end{bmatrix}. \quad (9)$$

INTERPRETING PCA

Inertia explained by a component

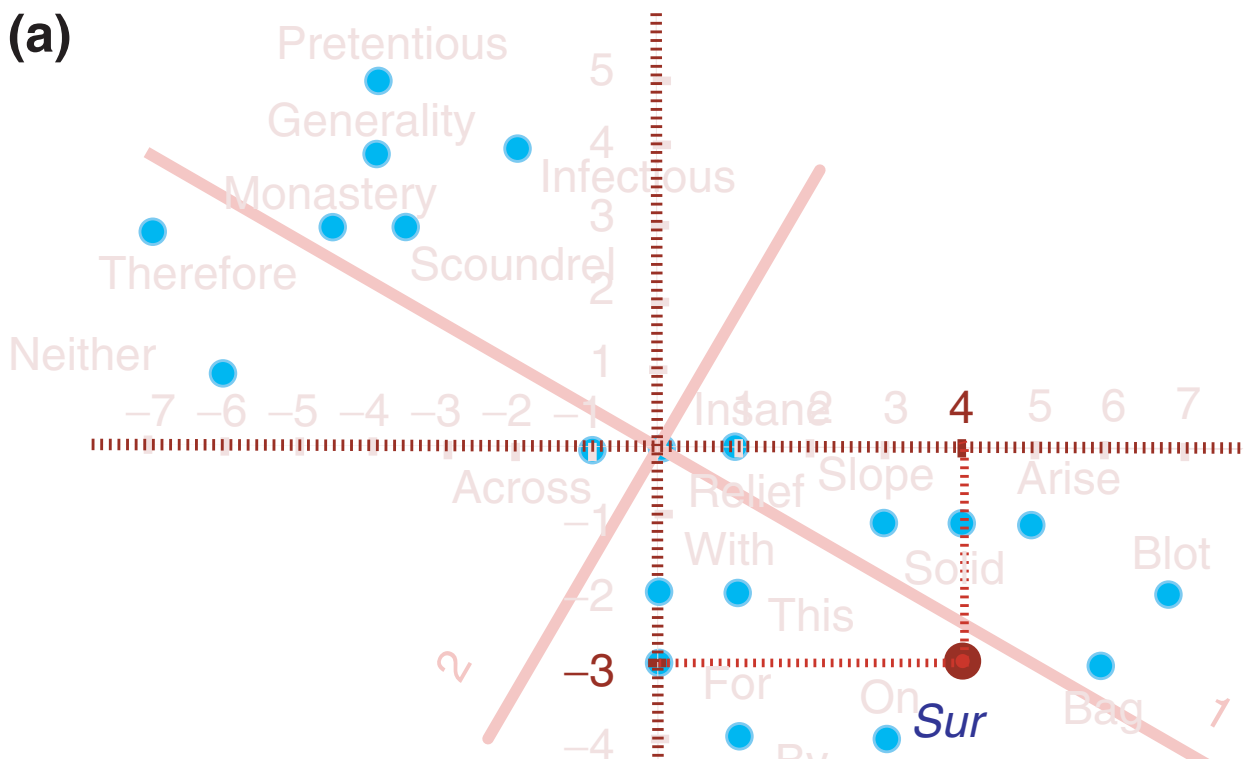
The importance of a component is reflected by its inertia or by the proportion of the total inertia "explained" by this factor. In our example (see Table 2) the inertia of the first component is equal to 392 and this corresponds to 83% of the total inertia.

Contribution of an Observation to a

Component

Recall that the eigenvalue associated to a component is equal to the sum of the squared factor scores for this component. Therefore, the importance of an observation for a component can be obtained by the ratio of the squared factor score of this observation by the eigenvalue associated with that component. This ratio is called the *contribution* of the observation to the component. Formally, the contribution of observation i to component ℓ is, denoted $\text{ctr}_{i,\ell}$, obtained as:

$$\text{ctr}_{i,\ell} = \frac{f_{i,\ell}^2}{\sum_i f_{i,\ell}^2} = \frac{f_{i,\ell}^2}{\lambda_\ell} \tag{10}$$



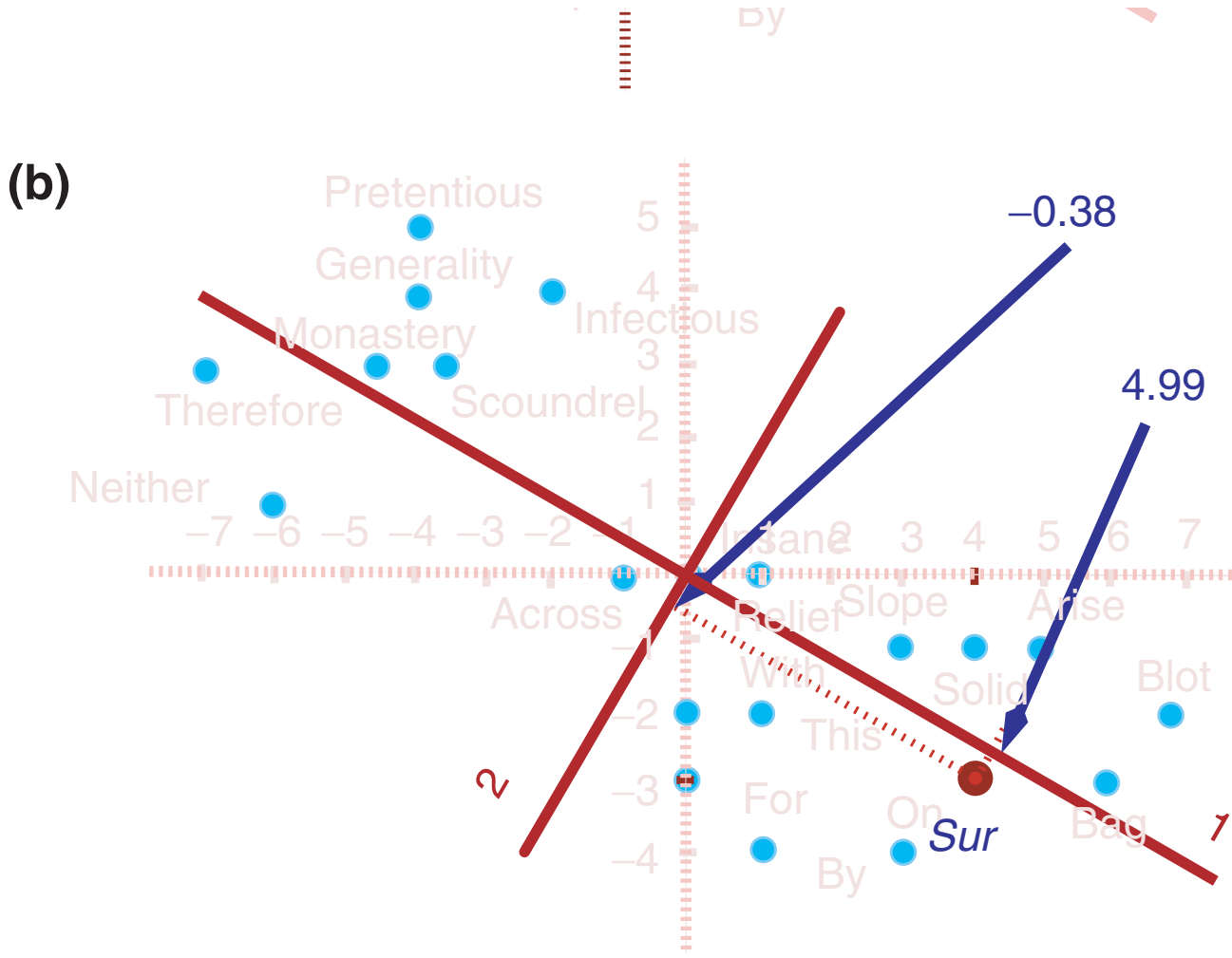


FIGURE 3 | How to find the coordinates (i.e., factor scores) on the principal components of a supplementary observation: (a) the French word *sur* is plotted in the space of the active observations from its deviations to the W and Y variables; and (b) The projections of the *sur* on the principal components give its coordinates.

where λ_ℓ is the eigenvalue of the ℓ-th component. The value of a contribution is between 0 and 1 and, for a given component, the sum of the contributions of all observations is equal to 1. The larger the value of the contribution, the more the observation

contributes to the component. A useful heuristic is to base the interpretation of a component on the observations whose contribution is larger than the average contribution (i.e., observations whose contribution is larger than $1/I$). The observations with high contributions and different signs can then be opposed to help interpret the component because these observations represent the two endpoints of this component.

The factor scores of the supplementary observations are not used to compute the eigenvalues and therefore their contributions are generally not computed.

Squared Cosine of a Component with an Observation

The *squared cosine* shows the importance of a component for a given observation. The squared

© 2010 John Wiley & Sons, Inc.

TABLE 2 | Eigenvalues and Percentage of Explained Inertia by Each Component

Component	λ_i (eigenvalue)	Cumulated (eigenvalues)	Percent of of Inertia	Cumulated (percentage)
1	392	392	83.29	83.29
2	52	444	11.71	100.00

cosine indicates the contribution of a component to the squared distance of the observation to the origin. It corresponds to the square of the cosine of the angle from the right triangle made with the origin, the observation, and its projection on the component and is computed as:

$$\cos^2_{i,\ell} = \frac{f^2_{i,\ell}}{\sum_{\ell} f^2_{i,\ell}} = \frac{f^2_{i,\ell}}{d^2_{i,\mathbf{g}}} \tag{11}$$

where $d^2_{i,\mathbf{g}}$ is the squared distance of a given observation to the origin. The squared distance, $d^2_{i,\mathbf{g}}$, is computed (thanks to the Pythagorean theorem) as the sum of the squared values of all the factor scores of this observation (*cf.* Eq. 4). Components with a large value of $\cos^2_{i,\ell}$ contribute a relatively large portion to the total distance and therefore these components are important for that observation.

The distance to the center of gravity is defined for supplementary observations and the squared cosine can be computed and is meaningful. Therefore, the value of $\cos^2$ can help find the components that are important to interpret both active and supplementary

observations.

Loading: Correlation of a Component and a Variable

The correlation between a component and a variable estimates the information they share. In the PCA framework, this correlation is called a *loading*. Note that the sum of the *squared* coefficients of correlation between a variable and all the components is equal to 1. As a consequence, the *squared* loadings are easier to interpret than the loadings (because the squared loadings give the proportion of the variance of the variables explained by the components). Table 3 gives the loadings as well as the squared loadings for the word length and definition example.

It is worth noting that the term 'loading' has several interpretations. For example, as previously mentioned, the elements of matrix **Q** (*cf.* Eq. B.1) are also called loadings. This polysemy is a potential source of confusion, and therefore it is worth checking what specific meaning of the word 'loadings' has been chosen when looking at the outputs of a program or when reading papers on PCA. In general, however, different meanings of 'loadings' lead to equivalent interpretations of the components. This happens

because the different types of loadings differ mostly by their type of normalization. For example, the correlations of the variables with the components are normalized such that the sum of the squared correlations of a given variable is equal to one; by contrast, the elements of Q are normalized such that the sum of the squared elements of a given component is equal to one.

Plotting the Correlations/Loadings of the Variables with the Components

The variables can be plotted as points in the component space using their loadings as coordinates. This representation differs from the plot of the observations: The observations are represented by their *projections*, but the variables are represented by their *correlations*. Recall that the sum of the squared loadings for a variable is equal to one. Remember, also, that a circle is defined as the set of points with the property that the sum of their squared coordinates is equal to a constant. As a consequence, when the data are perfectly represented by only two components, the sum of the squared loadings is equal to one, and therefore, in this case, the loadings will be positioned on a circle which is called the *circle of correlations*. When more than two components are

needed to represent the data perfectly, the variables will be positioned *inside* the circle of correlations. The closer a variable is to the circle of correlations, the better we can reconstruct this variable from the first two components (and the more important it is to interpret these components); the closer to the center of the plot a variable is, the less important it is for the first two components.

Figure 4 shows the plot of the loadings of the variables on the components. Each variable is a point whose coordinates are given by the loadings on the principal components.

We can also use *supplementary variables* to enrich the interpretation. A supplementary variable should be measured for the same observations

TABLE 3 | Loadings (i.e., Coefficients of Correlation between Variables and Components) and Squared Loadings

	Loadings		Squared Loadings		Q	
Component	*Y*	*W*	*Y*	*W*	*Y*	*W*
1	−0.9927	−0.9810	0.9855	0.9624	−0.5369	0.8437
2	0.1203	−0.1939	0.0145	0.0376	0.8437	0.5369
$\sum$			1.0000	1.0000		

The elements of matrix Q are also provided.

TABLE 4 | Supplementary Variables for the Example Length of Words and Number

of lines

	Frequency	# Entries
Bag	8	6
Across	230	3
On	700	12
Insane	1	2
By	500	7
Monastery	1	1
Relief	9	1
Slope	2	6
Scoundrel	1	1
With	700	5
Neither	7	2
Pretentious	1	1
Solid	4	5
This	500	9
For	900	7
Therefore	3	1
Generality	1	1
Arise	10	4
Blot	1	4
Infectious	1	2

'Frequency' is expressed as number of occurrences per 100,000 words, *'# Entries'* is obtained by counting the number of entries for the word in the dictionary.

used for the analysis (for all of them or part of them, because we only need to compute a coefficient of correlation). After the analysis has been performed, the coefficients of correlation (i.e., the loadings) between the supplementary variables and the components are computed. Then the supplementary variables are displayed in the circle of correlations using the loadings as coordinates.

For example, we can add two supplementary variables to the word length and definition example.

TABLE 5 | Loadings (i.e., Coefficients of Correlation) and Squared Loadings between Supplementary Variables and Components

	Loadings		Squared Loadings	
Component	Frequency	# Entries	Frequency	# Entries
1	−0.3012	0.6999	0.0907	0.4899
2	−0.7218	−0.4493	0.5210	0.2019
∑			.6117	.6918

These data are shown in Table 4. A table of loadings for the supplementary variables can be computed from the coefficients of correlation between these variables and the components (see Table 5). Note that, contrary to the active variables, the squared loadings of the supplementary variables do *not* add up to 1.

STATISTICAL INFERENCE: EVALUATING THE QUALITY OF THE MODEL

Fixed Effect Model

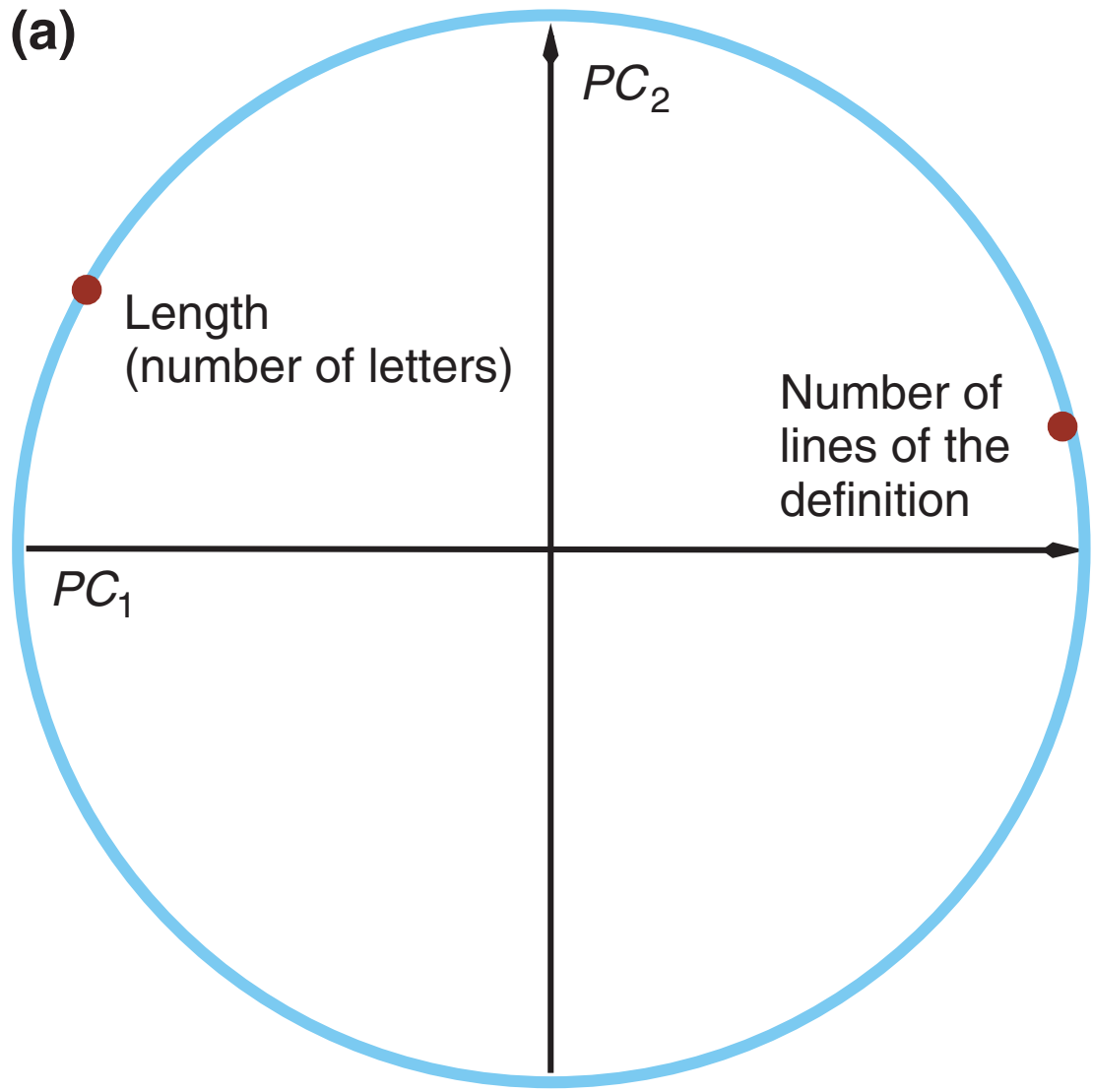
The results of PCA so far correspond to a fixed effect model (i.e., the observations are considered to be the population of interest, and conclusions are limited to these specific observations). In this context, PCA is descriptive and the amount of the variance of **X** explained by a component indicates its importance.

For a fixed effect model, the quality of the PCA model using the first M components is obtained

by first computing the estimated matrix, denoted $\widehat{\mathbf{X}}^{[M]}$, which is matrix $\mathbf{X}$ reconstituted with the first M components. The formula for this estimation is obtained by combining Eqs 1, 5, and 6 in order to obtain

$$\mathbf{X} = \mathbf{F}\mathbf{Q}^{\mathsf{T}} = \mathbf{X}\mathbf{Q}\mathbf{Q}^{\mathsf{T}} . \tag{12}$$

© 2010 John Wiley & Sons, Inc.



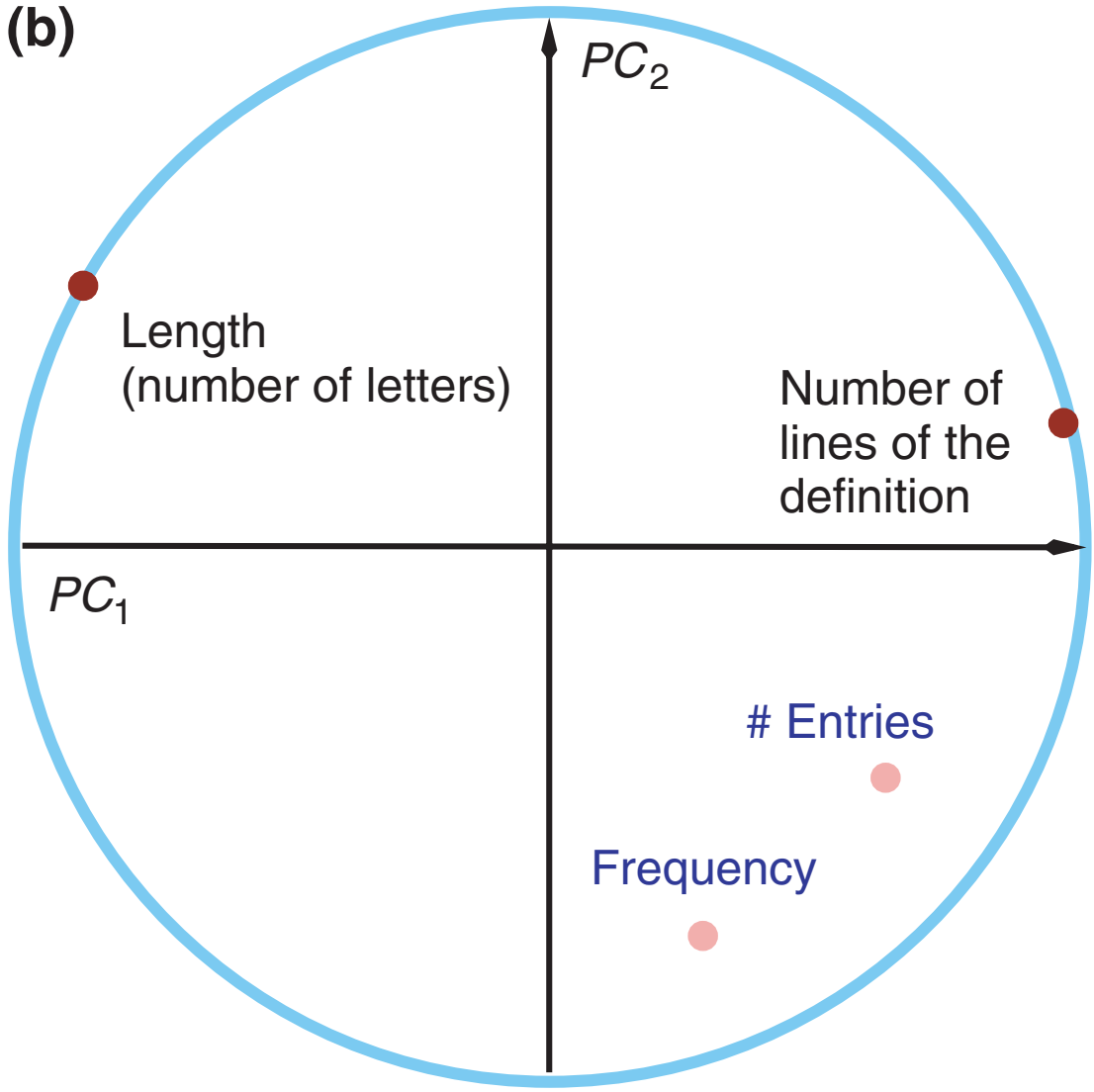


FIGURE 4 | Circle of correlations and plot of the loadings of (a) the variables with principal components 1 and 2, and (b) the variables and supplementary variables with principal components 1 and 2. Note that the supplementary variables are not positioned on the unit circle.

Then, the matrix $\hat{\mathbf{X}}^{[M]}$ is built back using Eq. 12 keeping only the first M components:

$$\hat{\mathbf{X}}^{[M]} = \mathbf{P}^{[M]} \boldsymbol{\Delta}^{[M]} \mathbf{Q}^{[M]\mathsf{T}}$$
$$= \mathbf{F}^{[M]} \mathbf{Q}^{[M]\mathsf{T}}$$

$$= \mathbf{X}\mathbf{Q}^{[M]}\mathbf{Q}^{[M]\mathsf{T}} \tag{13}$$

where $\mathbf{P}^{[M]}$, $\mathbf{\Delta}^{[M]}$, and $\mathbf{Q}^{[M]}$ represent, respectively the matrices $\mathbf{P}$, $\mathbf{\Delta}$, and $\mathbf{Q}$ with only their first M components. Note, incidentally, that Eq. 7 can be rewritten in the current context as:

$$\mathbf{X} = \widehat{\mathbf{X}}^{[M]} + \mathbf{E} = \mathbf{F}^{[M]}\mathbf{Q}^{[M]\mathsf{T}} + \mathbf{E} \tag{14}$$

(where $\mathbf{E}$ is the error matrix, which is equal to $\mathbf{X} - \widehat{\mathbf{X}}^{[M]}$).

To evaluate the quality of the reconstitution of $\mathbf{X}$ with M components, we evaluate the similarity between $\mathbf{X}$ and $\widehat{\mathbf{X}}^{[M]}$. Several coefficients can be used for this task [see, e.g., Refs 16–18]. The squared coefficient of correlation is sometimes used, as well as the R_V coefficient.[18,19] The most popular coefficient, however, is the *residual sum of squares* (RESS). It is computed as:

$$RESS_M = \|\mathbf{X} - \widehat{\mathbf{X}}^{[M]}\|^2$$

$$= \mathsf{trace}\left\{\mathbf{E}^{\mathsf{T}}\mathbf{E}\right\}$$

$$M$$

$$= \mathcal{I} - \sum_{\ell=1} \lambda_\ell \tag{15}$$

where $\| \|$ is the norm of $\mathbf{X}$ (i.e., the square root of the sum of all the squared elements of $\mathbf{X}$), and where the trace of a matrix is the sum of its diagonal elements. The smaller the value of RESS, the better the PCA model. For a fixed effect model, a larger M gives a better estimation of $\widehat{\mathbf{X}}^{[M]}$. For a fixed effect model, the matrix $\mathbf{X}$ is always perfectly reconstituted with L components (recall that L is the rank of $\mathbf{X}$).

In addition, Eq. 12 can be adapted to compute the estimation of the supplementary observations as

$$\widehat{\mathbf{x}}_{\text{sup}}^{[M]} = \mathbf{x}_{\text{sup}} \mathbf{Q}^{[M]} \mathbf{Q}^{[M]\mathsf{T}}. \tag{16}$$

Random Effect Model

In most applications, the set of observations represents a sample from a larger population. In this case, the goal is to estimate the value of *new* observations from this population. This corresponds to a *random effect* model. In order to estimate the generalization capacity of the PCA model, we cannot use standard parametric procedures. Therefore, the performance of the PCA model is evaluated using computer-based resampling

techniques such as the bootstrap and cross-validation techniques where the data are separated into a learning and a testing set. A popular cross-validation technique is the jackknife (*aka* 'leave one out' procedure). In the jackknife,[20–22] each observation is dropped from the set in turn and the remaining observations constitute the learning set. The learning set is then used to estimate (using Eq. 16) the left-out observation which constitutes the testing set. Using this procedure, each observation is estimated according to a random effect model. The predicted observations are then stored in a matrix denoted $\tilde{\mathbf{X}}$.

The overall quality of the PCA random effect model using M components is evaluated as the similarity between $\mathbf{X}$ and $\tilde{\mathbf{X}}^{[M]}$. As with the fixed effect model, this can also be done with a squared coefficient of correlation or (better) with the R_V

coefficient. Similar to RESS, one can use the *predicted residual sum of squares* (PRESS). It is computed as:

$$PRESS_M = \|\mathbf{X} - \tilde{\mathbf{X}}^{[M]}\|^2 \tag{17}$$

The smaller the *PRESS* the better the *quality* of the estimation for a random model.

Contrary to what happens with the fixed effect model, the matrix **X** is not always perfectly reconstituted with all L components. This is particularly the case when the number of variables is larger than the number of observations (a configuration known as the 'small N large P' problem in the literature).

How Many Components?

Often, only the important information needs to be extracted from a data matrix. In this case, the problem is to figure out how many components need to be considered. This problem is still open, but there are some guidelines [see, e.g.,Refs 9,8,23]. A first procedure is to plot the eigenvalues according to their size [the so called "scree," see Refs 8,24 and Table 2] and to see if there is a point in this graph (often called an 'elbow') such that the slope of the graph goes from 'steep' to "flat" and to keep only the components which are before the elbow. This procedure, somewhat subjective, is called the *scree* or *elbow* test.

Another standard tradition is to keep only the components whose eigenvalue is larger than the

average. Formally, this amount to keeping the ℓ-th component if

$$\lambda_\ell > \frac{1}{L}\sum_{\ell}^{L}\lambda_\ell = \frac{1}{L}\mathcal{I} \tag{18}$$

(where L is the rank of $\mathbf{X}$). For a correlation PCA, this rule boils down to the standard advice to 'keep only the eigenvalues larger than 1' [see, e.g., Ref 25]. However, this procedure can lead to ignoring important information [see Ref 26 for an example of this problem].

Random Model

As mentioned earlier, when using a random model, the quality of the prediction does not always increase with the number of components of the model. In fact, when the number of variables exceeds the number of observations, quality typically increases and then decreases. When the quality of the prediction decreases as the number of components increases this is an indication that the model is overfitting the data (i.e., the information in the learning set is not useful to fit the testing set). Therefore, it is important to determine the optimal number of components to keep when the

goal is to generalize the conclusions of an analysis to new data.

A simple approach stops adding components when PRESS decreases. A more elaborated approach [see e.g., Refs 27–31] begins by computing, for each component ℓ, a quantity denoted Q_ℓ^2 is defined as:

$$Q_\ell^2 = 1 - \frac{\mathrm{PRESS}_\ell}{\mathrm{RESS}_{\ell-1}} \tag{19}$$

with PRESS_ℓ (RESS_ℓ) being the value of PRESS (RESS) for the ℓ-th component (where RESS_0 is equal to the total inertia). Only the components with Q_ℓ^2 greater or equal to an arbitrary critical value (usually $1 - 0.95^2 = 0.0975$) are kept [an alternative set of critical values sets the threshold to 0.05 when $I \leq 100$ and to 0 when $I > 100$; see Ref 28].

Another approach—based on cross-validation—to decide upon the number of components to keep uses the index W_ℓ derived from Refs 32 and 33. In contrast to Q_ℓ^2, which depends on RESS and PRESS, the index W_ℓ, depends only upon PRESS. It is computed for the ℓ-th component as

$$W_\ell = \frac{PRESS_{\ell-1} - PRESS_\ell}{PRESS_\ell} \times \frac{df_{\text{residual},\,\ell}}{df_\ell}, \qquad (20)$$

where PRESS_0 is the inertia of the data table, df_ℓ is the number of degrees of freedom for the ℓ-th component equal to

$$df_\ell = I + J - 2\ell, \qquad (21)$$

and $df_{\text{residual},\,\ell}$ is the residual number of degrees of freedom which is equal to the total number of degrees of freedom of the table [equal to $J(I-1)$] minus the number of degrees of freedom used by the previous components. The value of $df_{\text{residual},\,\ell}$ is obtained as:

$$\begin{aligned} df_{\text{residual},\,\ell} &= J(I-1) - \sum_{k=1}^{\ell}(I + J - 2k) \\ &= J(I-1) - \ell(I + J - \ell - 1). \end{aligned} \qquad (22)$$

Most of the time, Q^2_ℓ and W_ℓ will agree on the number of components to keep, but W_ℓ can give a more conservative estimate of the number of components

to keep than Q_ℓ^2. When J is smaller than I, the value of both Q_L^2 and W_L is meaningless because they both involve a division by zero.

Bootstrapped Confidence Intervals

After the number of components to keep has been determined, we can compute confidence intervals for the eigenvalues of $\tilde{\mathbf{X}}$ using the bootstrap.[34–39] To use the bootstrap, we draw a large number of samples (e.g., 1000 or 10,000) with replacement from the learning set. Each sample produces a set of eigenvalues. The whole set of eigenvalues can then be used to compute confidence intervals.

ROTATION

After the number of components has been determined, and in order to facilitate the interpretation, the analysis often involves a rotation of the components that were retained [see, e.g., Ref 40 and 67, for more details]. Two main types of rotation are used: *orthogonal* when the new axes are also orthogonal to each other, and *oblique* when the new axes are

not required to be orthogonal. Because the rotations are always performed in a subspace, the new axes will always explain less inertia than the original components (which are computed to be optimal). However, the part of the inertia explained by the total subspace after rotation is the same as it was before rotation (only the partition of the inertia has changed). It is also important to note that because rotation always takes place in a subspace (i.e., the space of the retained components), the choice of this subspace strongly influences the result of the rotation. Therefore, it is strongly recommended to try several sizes for the subspace of the retained components in order to assess the robustness of the interpretation of the rotation. When performing a rotation, the term loadings almost always refer to the elements of matrix **Q**. We will follow this tradition in this section.

Orthogonal Rotation

An orthogonal rotation is specified by a rotation matrix, denoted **R**, where the rows stand for the original factors and the columns for the new (rotated) factors. At the intersection of row m and column n we have the cosine of the angle between the original axis and the new one: $r_{m,n} = \cos\theta_{m,n}$. A rotation matrix

has the important property of being orthonormal because it corresponds to a matrix of direction cosines and therefore $\mathbf{R}^{\mathsf{T}}\mathbf{R} = \mathbf{I}$.

Varimax rotation, developed by Kaiser,[41] is the most popular rotation method. For varimax a simple solution means that each component has a small number of large loadings and a large number of zero (or small) loadings. This simplifies the interpretation because, after a varimax rotation, each original variable tends to be associated with one (or a small number) of the components, and each component represents only a small number of variables. In addition, the components can often be interpreted from the opposition of few variables with positive loadings to few variables with negative loadings. Formally varimax searches for a linear combination of the original factors such that the variance of the squared loadings is maximized, which amounts to maximizing

$$v = \sum (q_{j,\ell}^2 - \overline{q}_\ell^2)^2 \tag{23}$$

with $q_{j,\ell}^2$ being the squared loading of the j-th variable of matrix $\mathbf{Q}$ on component ℓ and $\overline{q}_\ell^2$ being the mean of the squared loadings.

Oblique Rotations

With oblique rotations, the new axes are free to take any position in the component space, but the degree of correlation allowed among factors is small because two highly correlated components are better interpreted as only one factor. Oblique rotations, therefore, relax the orthogonality constraint in order to gain simplicity in the interpretation. They were strongly recommended by Thurstone,[42] but are used more rarely than their orthogonal counterparts.

For oblique rotations, the promax rotation has the advantage of being fast and conceptually simple. The first step in promax rotation defines the target matrix, almost always obtained as the result of a varimax rotation whose entries are raised to some power (typically between 2 and 4) in order to force the structure of the loadings to become bipolar. The second step is obtained by computing a least square fit from the varimax solution to the target matrix. Promax rotations are interpreted by looking at the correlations—regarded as loadings—between the rotated axes and the original variables. An interesting recent development of the concept of oblique rotation corresponds to the technique of

independent component analysis (ICA) where the axes are computed in order to replace the notion of orthogonality by statistical independence [see Ref 43,for a tutorial].

When and Why Using Rotations

The main reason for using rotation is to facilitate the interpretation. When the data follow a model (such as the psychometric model) stipulating (1) that each variable load on only one factor and (2) that there is a clear difference in intensity between the relevant

TABLE 6 | An (Artificial) Example of PCA using a Centered and Normalized Matrix

	Hedonic	For Meat	For Dessert	Price	Sugar	Alcohol	Acidity
Wine 1	14	7	8	7	7	13	7
Wine 2	10	7	6	4	3	14	7
Wine 3	8	5	5	10	5	12	5
Wine 4	2	4	7	16	7	11	3
Wine 5	6	2	4	13	3	10	3

Five wines are described by seven variables [data from Ref 44].

factors (whose eigenvalues are clearly larger than one) and the noise (represented by factors with eigenvalues clearly smaller than one), then the rotation is likely to provide a solution that is more reliable than the original solution. However, if this model does not

accurately represent the data, then rotation will make the solution less replicable and potentially harder to interpret because the mathematical properties of PCA have been lost.

EXAMPLES

Correlation PCA

Suppose that we have five wines described by the average ratings of a set of experts on their hedonic dimension, how much the wine goes with dessert, and how much the wine goes with meat. Each wine is also described by its price, its sugar and alcohol content, and its acidity. The data [from Refs 40,44] are given in Table 6.

A PCA of this table extracts four factors (with eigenvalues of 4.76, 1.81, 0.35, and 0.07, respectively). Only two components have an eigenvalue larger than 1 and, together, these two components account for 94% of the inertia. The factor scores for the first two components are given in Table 7 and the corresponding map is displayed in Figure 5.

TABLE 7 | PCA Wine Characteristics Factor scores, contributions

of the observations to the components, and squared cosines of the observations on principal components 1 and 2.

	F_1	F_2	ctr_1	ctr_2	$\cos^2_1$	$\cos^2_2$
Wine 1	−1.17	−0.55	**29**	17	77	17
Wine 2	−1.04	0.61	**23**	*21*	69	24
Wine 3	0.08	0.19	0	2	7	34
Wine 4	0.89	−0.86	17	**41**	50	46
Wine 5	1.23	0.61	*32*	*20*	78	19

The *positive* important contributions are italicized, and the **negative** important contributions are represented in bold. For convenience, squared cosines and contributions have been multiplied by 100 and rounded.

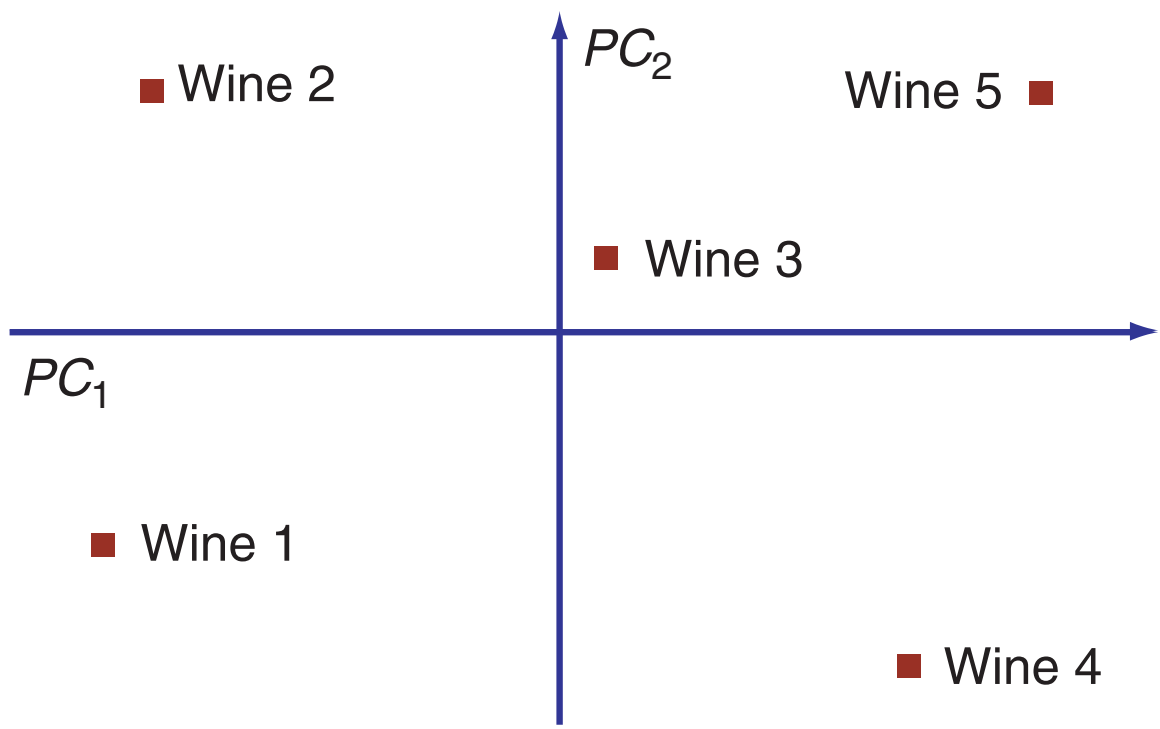


FIGURE 5 | PCA wine characteristics. Factor scores of the observations plotted on the first two components. $\lambda_1 = 4.76$, $\tau_1 = 68\%$; $\lambda_2 = 1.81$, $\tau_2 = 26\%$.

We can see from Figure 5 that the first component separates Wines 1 and 2 from Wines 4 and 5, while the second component separates Wines 2 and 5 from Wines 1 and 4. The examination of the values of the contributions and cosines, shown in Table 7, complements and refines this interpretation because the contributions suggest that Component 1 essentially contrasts Wines 1 and 2 with Wine 5 and that Component 2 essentially contrasts Wines 2 and 5 with Wine 4. The cosines show that Component 1 contributes highly to Wines 1 and 5, while Component 2 contributes most to Wine 4.

To find the variables that account for these differences, we examine the loadings of the variables on the first two components (see Table 8) and the circle of correlations (see Figure 6 and Table 9). From these, we see that the first component contrasts price with the wine's hedonic qualities, its acidity, its amount of alcohol, and how well it goes with meat (i.e., the wine tasters preferred inexpensive wines). The second component contrasts the wine's hedonic qualities, acidity, and alcohol content with its sugar content and how well it goes with dessert. From this, it appears that the first component represents characteristics that

are inversely correlated with a wine's price while the second component represents the wine's sweetness.

To strengthen the interpretation, we can apply a varimax rotation, which gives a clockwise rotation

TABLE 8 | PCA Wine Characteristics. Loadings (i.e., $\mathbf{Q}$ matrix) of the Variables on the First Two Components

	Hedonic	For Meat	For Dessert	Price	Sugar	Alcohol	Acidity
PC 1	−0.40	−0.45	−0.26	0.42	−0.05	−0.44	−0.45
PC 2	0.11	−0.11	−0.59	−0.31	−0.72	0.06	0.09

TABLE 9 | PCA Wine Characteristics. Correlation of the Variables with the First Two Components

	Hedonic	For meat	For dessert	Price	Sugar	Alcohol	Acidity
PC 1	−0.87	−0.97	−0.58	0.91	−0.11	−0.96	−0.99
PC 2	0.15	−0.15	−0.79	−0.42	−0.97	0.07	0.12

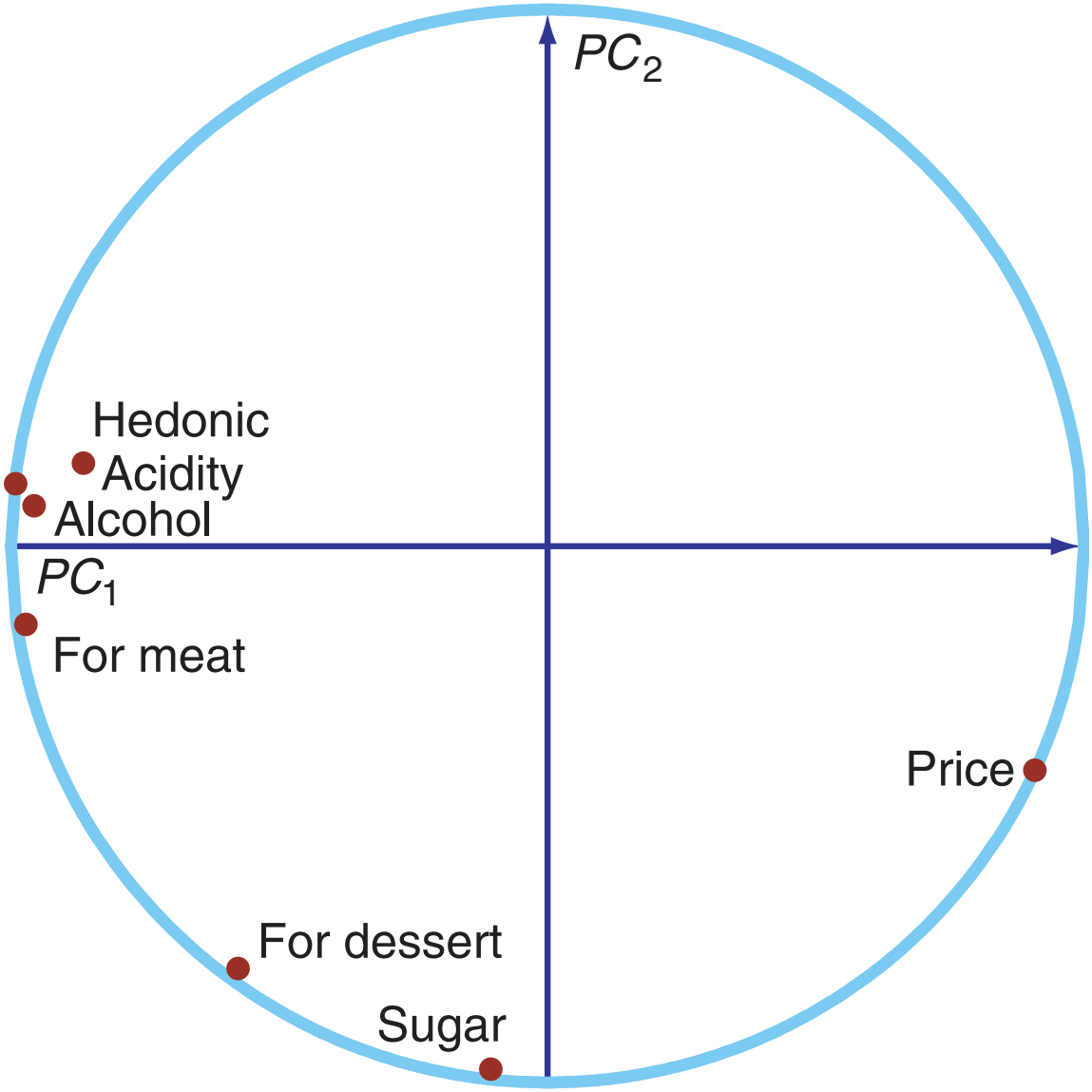


FIGURE 6 | PCA wine characteristics. Correlation (and circle of correlations) of the Variables with Components 1 and 2. $\lambda_1 = 4.76$, $\tau_1 = 68\%$; $\lambda_2 = 1.81$, $\tau_2 = 26\%$.

of 15° (corresponding to a cosine of 0.97). This gives the new set of rotated loadings shown in Table 10. The rotation procedure is illustrated in Figure 7. The improvement in the simplicity of the interpretation is marginal, maybe because the component structure

of such a small data set is already very simple. The first dimension remains linked to price and the second dimension now appears more clearly as the dimension of sweetness.

Covariance PCA

Here we use data from a survey performed in the 1950s in France [data from Ref 45]. The data table gives the average number of Francs spent on several categories of food products according to social class and the number of children per family. Because a Franc spent on one item has the same value as a Franc spent on another item, we want to keep the same unit of measurement for the complete space. Therefore we will perform a covariance PCA, rather than a correlation PCA. The data are shown in Table 11.

A PCA of the data table extracts seven components (with eigenvalues of 3,023,141.24, 290,575.84, 68,795.23, 25,298.95, 22,992.25, 3,722.32, and 723.92, respectively). The first two components extract 96% of the inertia of the data table, and we will keep only these two components for further consideration (see also Table 14 for the choice of the number of components to keep). The factor scores for the first two components are given in Table 12 and

the corresponding map is displayed in Figure 8.

We can see from Figure 8 that the first component separates the different social classes, while the second component reflects the number of children per family. This shows that buying patterns differ both by social class and by number of children per family. The contributions and cosines, given in Table 12, confirm this interpretation. The values of the contributions of the observations to the components indicate that Component 1 contrasts blue collar families with three children to upper class families with three or more children whereas Component 2 contrasts blue and white collar families with five children to upper class families with three and four children. In addition, the cosines between the components and the variables show that Component 1 contributes to the pattern of food spending seen by the blue collar and white collar families with two and three children and to the upper class families with three or more children while Component 2 contributes to the pattern of food spending by blue collar families with five children.

To find the variables that account for these differences, we refer to the squared loadings of the variables on the two components (Table 13) and to

© 2010 John Wiley & Sons, Inc.

TABLE 10 | PCA Wine Characteristics: Loadings (i.e., $\mathbf{Q}$ matrix), After Varimax Rotation, of the Variables on the First Two Components

	Hedonic	For Meat	For Dessert	Price	Sugar	Alcohol	Acidity
PC 1	−0.41	−0.41	−0.11	0.48	0.13	−0.44	−0.46
PC 2	0.02	−0.21	−0.63	−0.20	−0.71	−0.05	−0.03

TABLE 11 | Average Number of Francs Spent (per month) on Different Types of Food According To Social Class and Number of Children [dataset from Ref 45]

		Type of Food						
		Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine
Blue collar	2 Children	332						428
White collar	2 Children	293						559
Upper class	2 Children	372						767
Blue collar	3 Children	406						563
White collar	3 Children	386						608
Upper class	3 Children	438						843
Blue collar	4 Children	534						660
White collar	4 Children	460						699
Upper class	4 Children	385						789
Blue collar	5 Children	655						776

White collar	5 Children		584			995		
354	1437		526		247		427	
388	1527		567		239		258	
562	1948		927		235		433	
341	1507		544		324		407	
396	1501		558		319		363	
689	2345		1148		243		341	
367	1620		638		414		407	
484	1856		762		400		416	
621	2366		1149		304		282	
423	1848		759		495		486	
548	2056		893		518		319	
Upper class	5 Children	515	1097	887	2630	1167	561	284
Mean		447	732	505	1887	803	358	369

$\hat{S}$	107	189	165	396	250	117	72

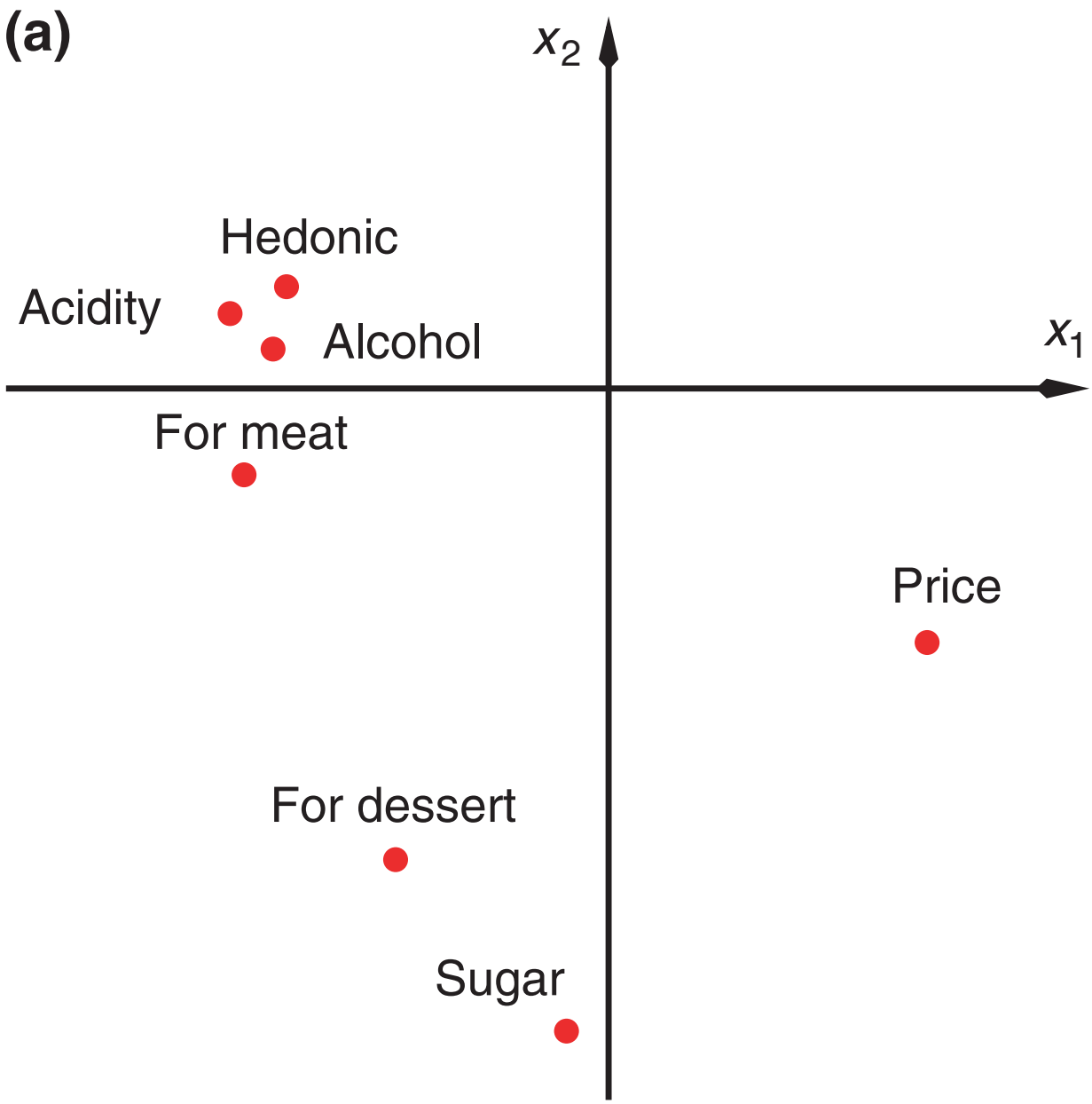
TABLE 12 | PCA Example. Amount of Francs Spent (per month) by Food Type, Social Class, and Number of Children. Factor scores, contributions of the observations to the components, and squared cosines of the observations on principal components 1 and 2

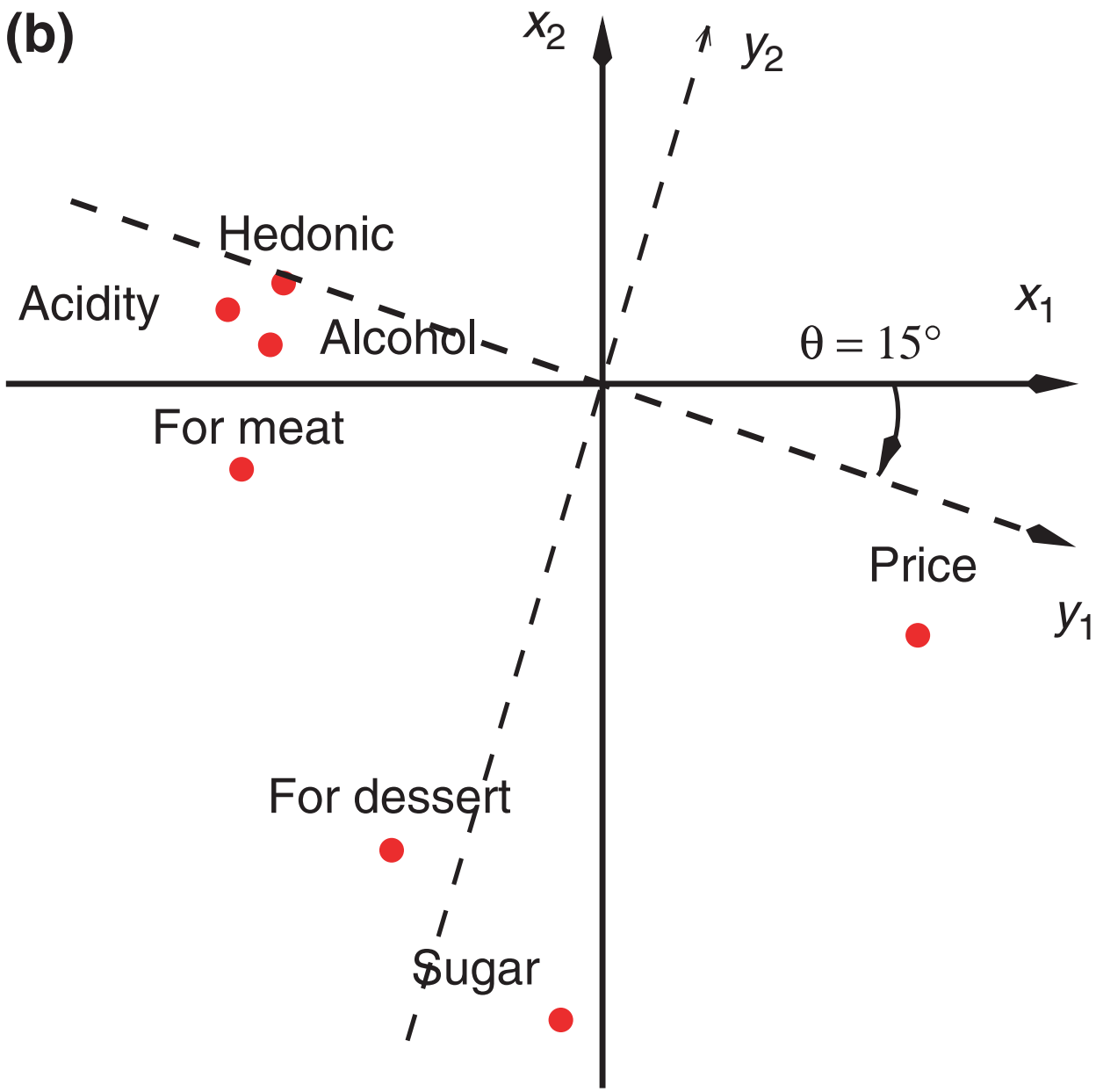
		F_1	F_2	ctr_1	ctr_2	$\cos^2_1$	$\cos^2_2$
Blue collar	2 Children	635.05					
White collar	2 Children	488.56					
Upper class	2 Children	−112.03					
Blue collar	3 Children	520.01					
White collar	3 Children	485.94					
Upper class	3 Children	−588.17					
Blue collar	4 Children	333.95					
White collar	4 Children	57.51					
Upper class	4 Children	−571.32					
Blue collar	5 Children	39.38					

White collar	5 Children	−296.04		
−120.89	13	5	95	3
−142.33	8	7	86	7
−139.75	0	7	26	40
12.05	*9*	0	100	0
1.17	8	0	98	0
−188.44	**11**	**12**	89	9
144.54	4	7	83	15
42.86	0	1	40	22
−206.76	**11**	**15**	86	11
264.47	0	*24*	2	79
235.92	3	*19*	57	36
Upper class	5 Children	−992.83	97.15	**33**	3	97	1

The *positive* important contributions are italicized, and the **negative** important contributions are represented in bold. For convenience, squared cosines and contributions have been multiplied by 100 and rounded.

© 2010 John Wiley & Sons, Inc.




(b)
x_2
y_2
Hedonic
Acidity
Alcohol
x_1
$\theta = 15°$
For meat
Price
y_1
For dessert
Sugar

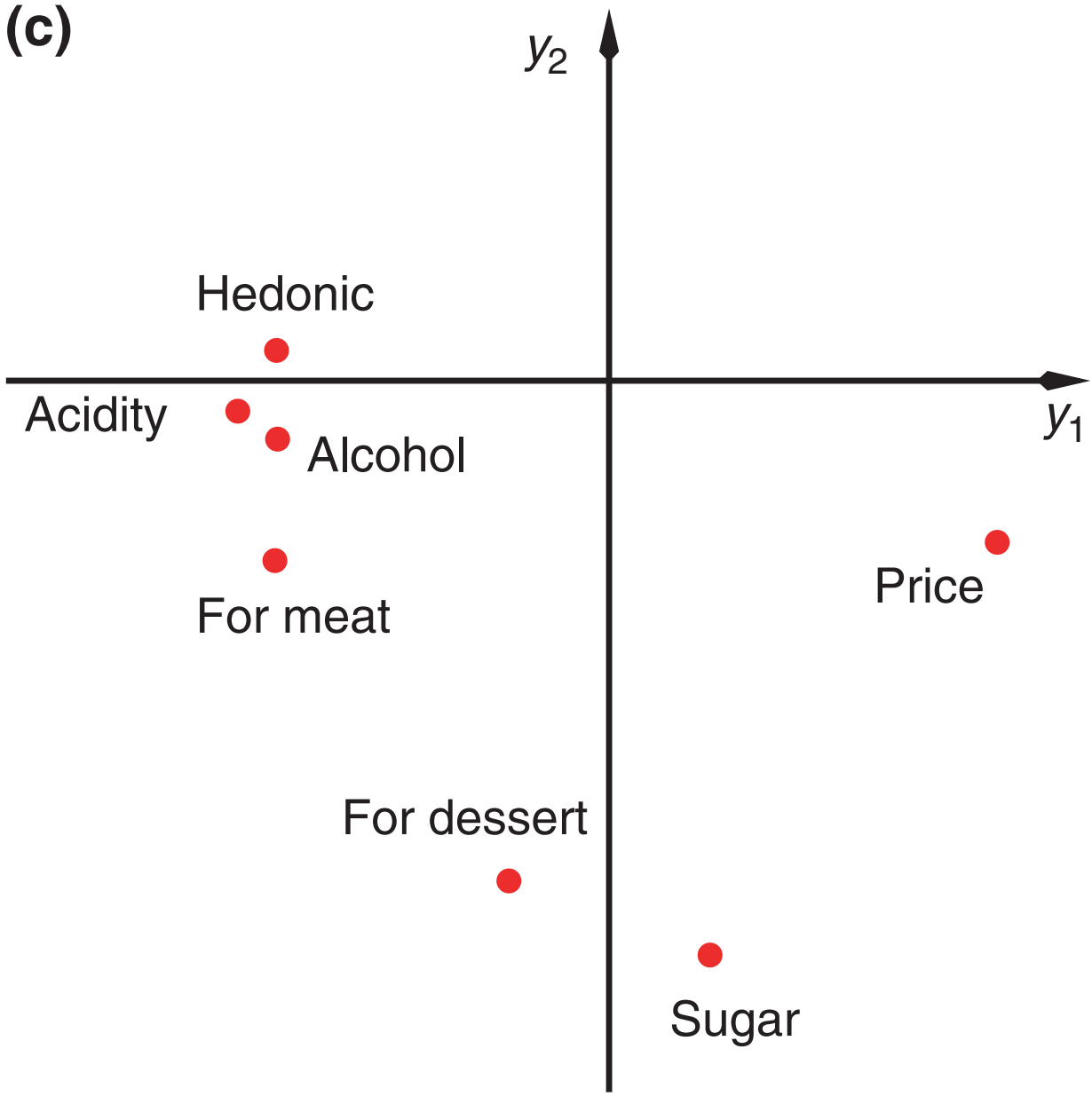


FIGURE 7 | PCA wine characteristics. (a) Original loadings of the seven variables. (b) The loadings of the seven variables showing the original axes and the new (rotated) axes derived from varimax. (c) The loadings after varimax rotation of the seven variables.

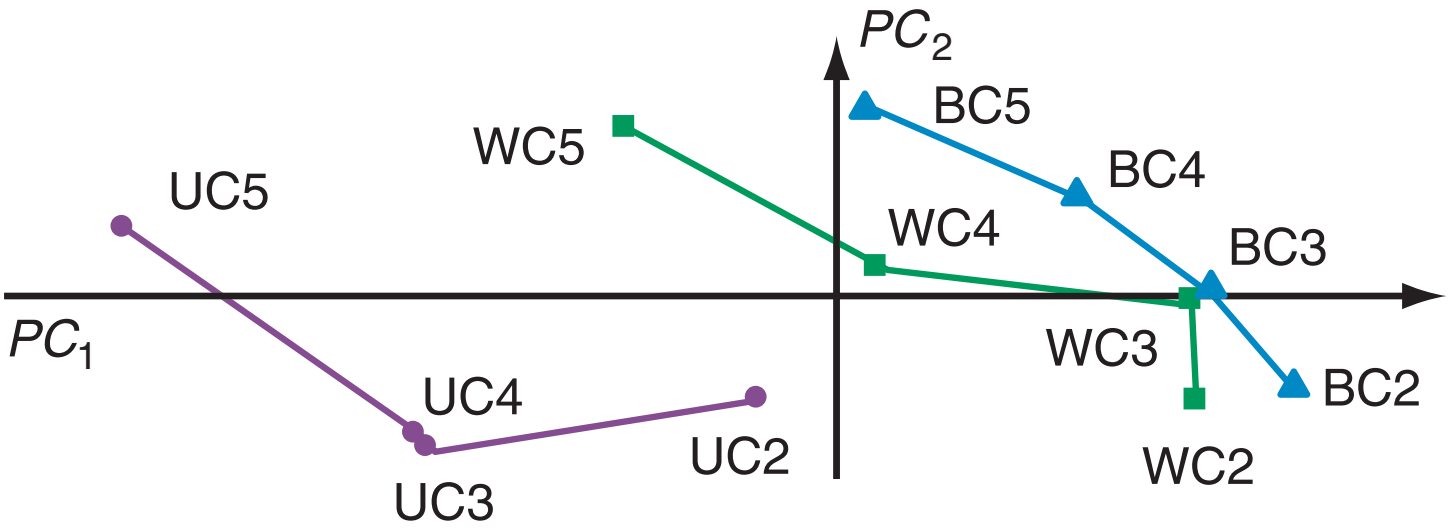


FIGURE 8 | PCA example. Amount of Francs spent (per month) on food type by social class and number of children. Factor scores for principal components 1 and 2. $\lambda_1 = 3{,}023{,}141.24$, $\tau_1 = 88\%$; $\lambda_2 = 290{,}575.84$, $\tau_2 = 8\%$. BC = blue collar; WC = white collar; UC = upper class; 2 = 2 children; 3 = 3 children; 4 = 4 children; 5 = 5 children..

TABLE 13 | PCA example: Amount of Francs Spent (per month) on Food Type by Social Class and Number of Children. Squared Loadings of the Variables on Components 1 and 2

	Bread	Vegetables	Fruit	Meat	Poultry	Milk	Wine
PC 1	0.01	0.33	0.16	0.01	0.03	0.45	0.00
PC 2	0.11	0.17	0.09	0.37	0.18	0.03	0.06

the circle of correlations (see Figure 9). From these, we see that the first component contrasts the amount

spent on wine with all other food purchases, while the second component contrasts the purchase of milk and bread with meat, fruit, and poultry. This indicates that wealthier families spend more money on meat, poultry, and fruit when they have more children, while white and blue collar families spend more money on bread and milk when they have more children. In addition, the number of children in upper class families seems inversely correlated with the consumption of wine (i.e., wealthy families with four or five children consume *less* wine than all other types of families). This curious effect is understandable when placed in the context of the French culture of the 1950s, in which wealthier families with many children tended to be rather religious and therefore less inclined to indulge in the consumption of wine.

Recall that the first two components account for 96% of the total inertia [i.e., $(\lambda_1 + \lambda_2)/\mathcal{I} = (3{,}023{,}141.24 + 290{,}575.84)/3{,}435{,}249.75 = 0.96$]. From Table 14 we find that RESS_2 is equal to 4% and this value represents the error when $\widehat{\mathbf{X}}$ is estimated from Components 1 and 2 together. This means that for a fixed effect model, a two-component solution represents $\mathbf{X}$ well. PRESS_2, the error of estimation using a random effect model with two components, is equal to 8% and this value indicates that $\widetilde{\mathbf{X}}$ represents

X adequately. Together, the values of RESS and PRESS suggest that only the first two components should be kept.

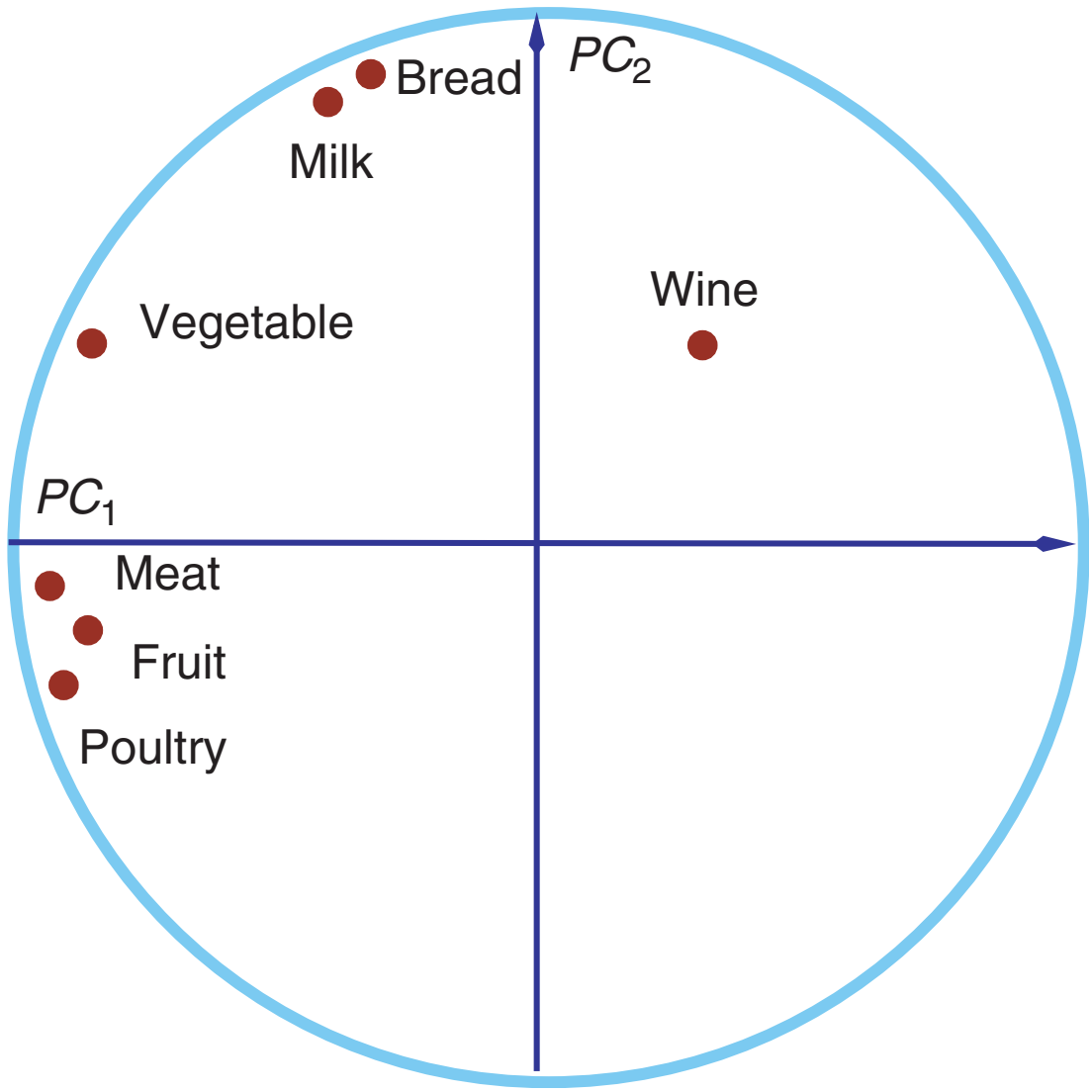


FIGURE 9 | PCA example: Amount of Francs spent (per month) on food type by social class and number of children. Correlations (and circle of correlations) of the variables with Components 1 and 2. $\lambda_1 = 3{,}023{,}141.24$, $\tau_1 = 88\%$; $\lambda_2 = 290{,}575.84$, $\tau_2 = 8\%$.

To confirm the number of components to keep, we look at Q^2 and W. The Q^2 values of 0.82 and 0.37 for Components 1 and 2 both exceed the critical value of 0.095, indicating that both components should be kept. Note that a negative Q^2 value suggests that a component should not be kept. In contrast, the W values of 1.31 and 0.45 for the first two components suggest that only the first component should be kept because only W_1 is greater than 1.

SOME EXTENSIONS OF PCA

Correspondence Analysis

Correspondence analysis (CA; see Refs 46–51) is an adaptation of PCA tailored to handle nominal variables. It can be interpreted as a particular case of generalized PCA for which we take into account masses (for the rows) and weights (for the columns). CA analyzes a contingency table and provides factor scores for both the rows and the columns of the contingency table. In CA, the inertia of the contingency table is proportional to the χ^2 which can be computed to test the independence of the

rows and the columns of this table. Therefore, the factor scores in CA decompose this independence χ^2 into orthogonal components (in the CA tradition, these components are often called *factors* rather than components, here, for coherence, we keep the name component for both PCA and CA).

Notations

The $I \times J$ contingency table to be analyzed is denoted $\mathbf{X}$. CA will provide two sets of factor scores: one for the rows and one for the columns. These factor scores are, in general, scaled such that their inertia is equal to their eigenvalue (some versions of CA compute row or column factor scores normalized to unity). The grand total of the table is denoted N.

Computations

The first step of the analysis is to compute the probability matrix $\mathbf{Z} = N^{-1}\mathbf{X}$. We denote $\mathbf{r}$ the vector of the row totals of $\mathbf{Z}$, (i.e., $\mathbf{r} = \mathbf{Z1}$, with $\mathbf{1}$ being a conformable vector of 1s), $\mathbf{c}$ the vector of the columns totals, and $\mathbf{D_c} = \mathbf{diag}\{\mathbf{c}\}$, $\mathbf{D_r} = \mathbf{diag}\{\mathbf{r}\}$. The factor scores are obtained from the following *generalized* SVD (see Appendix):

$$\left(\mathbf{Z} - \mathbf{r}\mathbf{c}^{\mathsf{T}}\right) = \tilde{\mathbf{P}}\tilde{\boldsymbol{\Delta}}\tilde{\mathbf{Q}}^{\mathsf{T}} \text{ with } \tilde{\mathbf{P}}^{\mathsf{T}}\mathbf{D}_{\mathbf{r}}^{-1}\tilde{\mathbf{P}} = \tilde{\mathbf{Q}}^{\mathsf{T}}\mathbf{D}_{\mathbf{c}}^{-1}\tilde{\mathbf{Q}} = \mathbf{I} \tag{24}$$

The row and (respectively) column factor scores are obtained as:

$$\mathbf{F} = \mathbf{D}_{\mathbf{r}}^{-1}\tilde{\mathbf{P}}\tilde{\boldsymbol{\Delta}} \qquad \text{and} \qquad \mathbf{G} = \mathbf{D}_{\mathbf{c}}^{-1}\tilde{\mathbf{Q}}\tilde{\boldsymbol{\Delta}} \tag{25}$$

In CA, the rows and the columns of the table have a similar role, and therefore we have contributions and cosines for both sets. These are obtained in a similar way as for standard PCA, but the computations of the contributions need to integrate the values of the masses (i.e., the elements of $\mathbf{r}$) and weights (i.e., the elements of $\mathbf{c}$). Specifically, the *contribution* of row i to component ℓ and of column j to component ℓ are obtained respectively as:

$$\text{ctr}_{i,\ell} = \frac{r_i f_{i,\ell}^2}{\lambda_\ell} \qquad \text{and} \qquad \text{ctr}_{j,\ell} = \frac{c_j g_{j,\ell}^2}{\lambda_\ell} \tag{26}$$

(with r_i begin the ith element of $\mathbf{r}$ and c_j being the j-th element of $\mathbf{c}$). As for standard PCA, contributions

help locating the observations or variables important for a given component.

The vector of the squared (χ^2) distance from the rows and columns to their respective barycenter are obtained as:

$$\mathbf{d_r} = \text{diag}\left\{\mathbf{F}\mathbf{F}^{\mathsf{T}}\right\} \quad \text{and} \quad \mathbf{d_c} = \text{diag}\left\{\mathbf{G}\mathbf{G}^{\mathsf{T}}\right\} \tag{27}$$

As for PCA, the total inertia in CA is equal to the sum of the eigenvalues. By contrast with PCA, the total inertia can also be computed equivalently as the weighted sum of the squared distances of the rows *or*

TABLE 14 | PCA Example: Amount of Francs Spent (per month) on Food Type by Social Class and Number of children. Eigenvalues, cumulative eigenvalues, RESS, PRESS, Q^2, and W values for all seven components

	λ	$\lambda/\mathcal{I}$	$\sum\lambda$	$\sum\lambda/\mathcal{I}$	RESS	RESS/$\mathcal{I}$	PRESS	PRESS/$\mathcal{I}$	Q^2	W
PC 1	3,023,141.24	0.88	3,023,141.24	0.88	412,108.51	0.12	610,231.19	0.18	0.82	1.31
PC 2	290,575.84	0.08	3,313,717.07	0.96	121,532.68	0.04	259,515.13	0.08	0.37	0.45
PC 3	68,795.23	0.02	3,382,512.31	0.98	52,737.44	0.02	155,978.58	0.05	−0.28	0.27
PC 4	25,298.95	0.01	3,407,811.26	0.99	27,438.49	0.01	152,472.37	0.04	−1.89	0.01
PC 5	22,992.25	0.01	3,430,803.50	1.00	4,446.25	0.00	54,444.52	0.02	−0.98	1.35
PC 6	3,722.32	0.00	3,434,525.83	1.00	723.92	0.00	7,919.49	0.00	−0.78	8.22
PC 7	723.92	0.00	3,435,249.75	1.00	0.00	0.00	0.00	0.00	1.00	∞
$\sum$	3,435,249.75	1.00								
	$\mathcal{I}$									

the columns to their respective barycenter. Formally,

the inertia can be computed as:

$$\mathcal{I} = \sum_{l}^{L} \lambda_\ell = \mathbf{r}^\mathsf{T}\mathbf{d_r} = \mathbf{c}^\mathsf{T}\mathbf{d_c} \tag{28}$$

The squared *cosine* between row i and component ℓ and column j and component ℓ are obtained respectively as:

$$\cos^2_{i,\ell} = \frac{f^2_{i,\ell}}{d^2_{r,i}} \qquad \text{and} \qquad \cos^2_{j,\ell} = \frac{g^2_{j,\ell}}{d^2_{c,j}} \tag{29}$$

(with $d^2_{r,i}$ and $d^2_{c,j}$, being respectively the i-th element of $\mathbf{d_r}$ and the j-th element of $\mathbf{d_c}$). Just like for PCA, squared cosines help locating the components important for a given observation or variable.

And just like for PCA, supplementary or illustrative elements can be projected onto the components, but the CA formula needs to take into account masses and weights. The projection formula, is called the *transition* formula and it is specific to CA. Specifically, let $\mathbf{i}^\mathsf{T}_{\text{sup}}$ being an illustrative row and $\mathbf{j}_{\text{sup}}$ being an illustrative column to be projected (note that in CA, prior to projection, a illustrative row or column

is re-scaled such that its sum is equal to one). Their coordinates of the illustrative rows (denoted $\mathbf{f}_{sup}$) and column (denoted $\mathbf{g}_{sup}$) are obtained as:

$$\mathbf{f}_{sup} = \left(\mathbf{i}_{sup}^{\mathsf{T}}\mathbf{1}\right)^{-1}\mathbf{i}_{sup}^{\mathsf{T}}\mathbf{G}\tilde{\boldsymbol{\Delta}}^{-1} \text{ and } \mathbf{g}_{sup} = \left(\mathbf{j}_{sup}^{\mathsf{T}}\mathbf{1}\right)^{-1}\mathbf{j}_{sup}^{\mathsf{T}}\mathbf{F}\tilde{\boldsymbol{\Delta}}^{-1} \quad (30)$$

[note that the scalar terms $\left(\mathbf{i}_{sup}^{\mathsf{T}}\mathbf{1}\right)^{-1}$ and $\left(\mathbf{j}_{sup}^{\mathsf{T}}\mathbf{1}\right)^{-1}$ are used to ensure that the sum of the elements of $\mathbf{i}_{sup}$ or $\mathbf{j}_{sup}$ is equal to one, if this is already the case, these terms are superfluous].

Example

For this example, we use a contingency table that gives the number of punctuation marks used by the French writers Rousseau, Chateaubriand, Hugo, Zola, Proust, and Giraudoux [data from Ref 52]. This table indicates how often each writer used the period, the comma, and all other punctuation marks combined (i.e., interrogation mark, exclamation mark, colon, and semicolon). The data are shown in Table 15.

A CA of the punctuation table extracts two components which together account for 100% of the inertia (with eigenvalues of .0178 and .0056, respectively). The factor scores of the observations (rows) and variables (columns) are shown in Tables 16 and the corresponding map is displayed in Figure 10.

We can see from Figure 10 that the first component separates Proust and Zola's pattern of punctuation from the pattern of punctuation of the other four authors, with Chateaubriand, Proust, and Zola contributing most to the component. The squared cosines show that the first component accounts for all of Zola's pattern of punctuation (see Table 16).

The second component separates Giraudoux's pattern of punctuation from that of the other authors. Giraudoux also has the highest contribution indicating that Giraudoux's pattern of punctuation is important for the second component. In addition, for Giraudoux the highest squared cosine (94%), is obtained for Component 2. This shows that the second component is essential to understand Giraudoux's pattern of punctuation (see Table 16).

In contrast with PCA, the variables (columns) in CA are interpreted identically to the rows. The

factor scores for the variables (columns) are shown in

TABLE 15 | The Punctuation Marks of Six French Writers (from Ref 52).

Author's Name	Period	Comma	Other	x_{i+}	**r**
Rousseau	7,836	13,112	6,026	26,974	0.0189
Chateaubriand	53,655	102,383	42,413	198,451	0.1393
Hugo	115,615	184,541	59,226	359,382	0.2522
Zola	161,926	340,479	62,754	565,159	0.3966
Proust	38,177	105,101	12,670	155,948	0.1094
Giraudoux	46,371	58,367	14,299	119,037	0.0835
x_{+j}	423,580	803,983	197,388	$N = 142{,}4951$	1.0000
$\mathbf{c}^T$	0.2973	0.5642	0.1385		

The column labeled x_{i+} gives the total number of punctuation marks used by each author. N is the grand total of the data table. The vector of mass for the rows, **r**, is the proportion of punctuation marks used by each author ($r_i = x_{i+}/N$). The row labeled x_{+j} gives the total number of times each punctuation mark was used. The centroid row, $\mathbf{c}^T$, gives the proportion of each punctuation mark in the sample ($c_j = x_{+j}/N$).

TABLE 16 | CA punctuation. Factor scores, contributions, mass, mass × squared factor scores, inertia to barycenter, and squared cosines for the rows.

	F_1	F_2	ctr_1	ctr_2	r_i	$r_i \times F_1^2$	$r_i \times F_2^2$	$r_i \times d_{r,i}^2$	$\cos_1^2$	$\cos_2^2$
Rousseau	−0.24	−0.07	6	2	0.0189	0.0011	0.0001	0.0012	91	9
Chateaubriand	−0.19	−0.11	**28**	**29**	0.1393	0.0050	0.0016	0.0066	76	24
Hugo	−0.10	0.03	15	4	.2522	0.0027	0.0002	0.0029	92	8
Zola	0.09	−0.00	*19*	0	.3966	0.0033	0.0000	0.0033	100	0
Proust	0.22	−0.06	*31*	8	.1094	0.0055	0.0004	.0059	93	7
Giraudoux	−0.05	0.20	1	*58*	0.0835	0.0002	0.0032	0.0034	6	94
$\sum$	—	—	100	100	—	.0178	.0056	.0234		
						λ_1	λ_2	$\mathcal{I}$		
						76%	24%			
						τ_1	τ_2			

The *positive* important contributions are italicized, and the **negative** important contributions are represented in bold. For convenience, squared cosines and contributions have been multiplied by 100 and rounded.

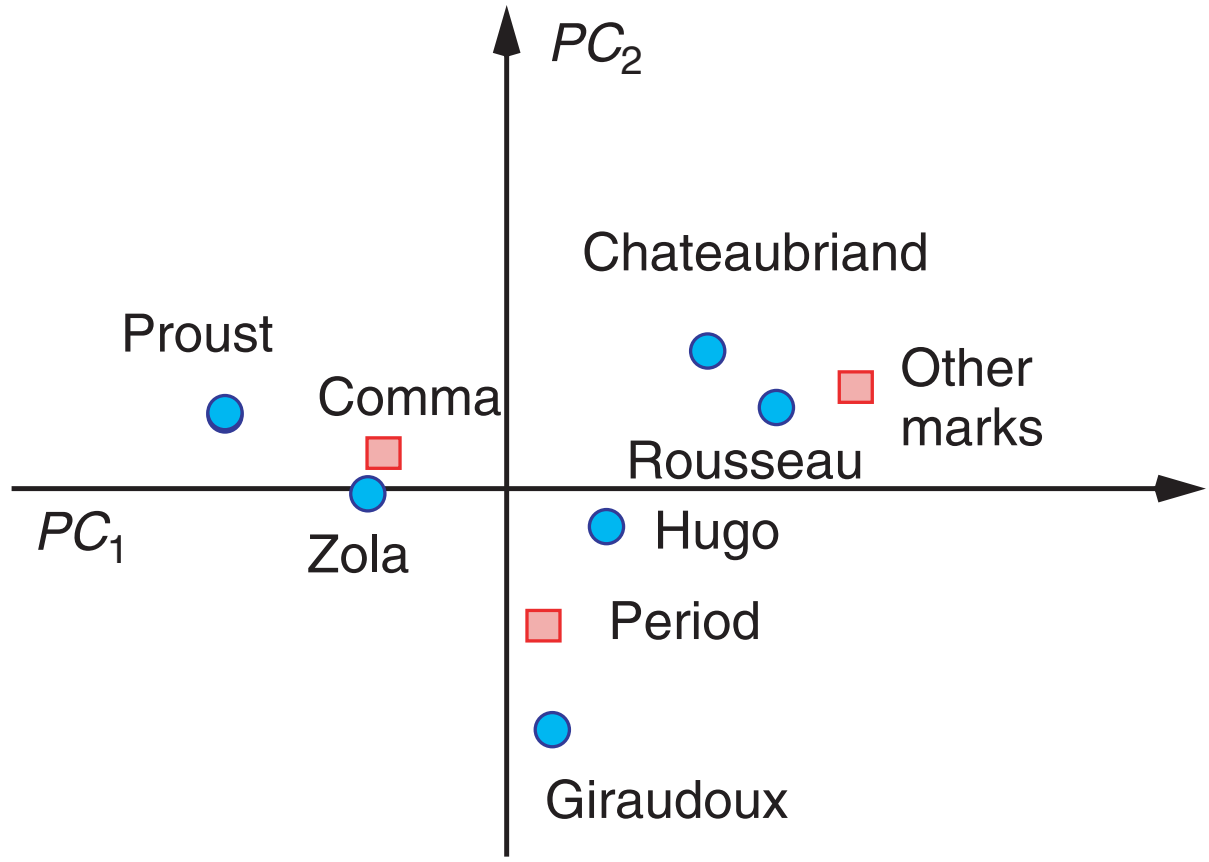


FIGURE 10 | CA punctuation. The projections of the rows and the columns are displayed in the same map. $\lambda_1 = 0.0178$, $\tau_1 = 76.16$; $\lambda_2 = 0.0056$, $\tau_2 = 23.84$.

Table 17 and the corresponding map is displayed in the same map as the observations shown in Figure 10.

From Figure 10 we can see that the first component also separates the comma from the 'others' punctuation marks. This is supported by the high contributions of 'others' and comma to the component. The cosines also support this interpretation because the first component accounts for 88% of the use of the comma and 91% of the use of the 'others'

punctuation marks (see Table 17).

The second component separates the period from both the comma and the 'other' punctuation marks. This is supported by the period's high contribution to the second component and the component's contribution to the use of the period (see Table 17).

TABLE 17 | CA punctuation. Factor scores, contributions, mass, mass × squared factor scores, inertia to barycenter, and squared cosines for the columns

	F_1	F_2	ctr_1	ctr_2	c_j	$c_j \times F_1^2$	$c_j \times F_2^2$	$c_j \times d_{c,j}^2$	$\cos_1^2$	$\cos_2^2$
Period	−0.05	0.11	4	*66*	.2973	0.0007	0.0037	0.0044	16	84
Comma	0.10	−0.04	30	14	.5642	0.0053	0.0008	0.0061	88	12
Other	−0.29	−0.09	**66**	20	.1385	0.0118	0.0011	0.0129	91	9
$\sum$	—	—	100	100	—	.0178	0.0056	0.0234		
						λ_1	λ_2	$\mathcal{I}$		
						76%	24%			
						τ_1	τ_2			

The *positive* important contributions are italicized, and the **negative** important contributions are represented in bold. For convenience, squared cosines, and contributions have been multiplied by 100 and rounded.

Together, the pattern of distribution of the points representing the authors and the punctuation marks suggest that some of the differences in the authors' respective styles can be attributed to differences in their use of punctuation. Specifically, Zola's œuvre is characterized by his larger than average use of the comma, while Chateaubriand's is characterized by his larger than average use of other types of

punctuation marks than the period and the comma. In addition, Giraudoux's œuvre is characterized by a larger than average use of the period.

Multiple Factor Analysis

Multiple factor analysis [MFA; see Refs 53–55] is used to analyze a set of observations described by several groups of variables. The number of variables in each group may differ and the nature of the variables (nominal or quantitative) can vary from one group to the other but the variables should be of the same nature in a given group. The analysis derives an integrated picture of the observations and of the relationships between the groups of variables.

Notations

The data consists of T data sets. Each data set is called a *subtable*. Each subtable is an $I \times_{[t]} J$ rectangular data matrix, denoted ${}_{[t]}\mathbf{Y}$, where I is the number of observations and ${}_{[t]}J$ the number of variables of the t-th subtable. The total number of variables is equal to J, with:

$$J = \sum_{t} {}_{[t]}J \ . \tag{31}$$

Each subtable is preprocessed (e.g., centered and normalized) and the preprocessed data matrices actually used in the analysis are denoted ${}_{[t]}\mathbf{X}$.

The ℓ-th eigenvalue of the t-th subtable is denoted ${}_{[t]}\varrho_{\ell}$. The ℓ-th singular value of the t-th subtable is denoted ${}_{[t]}\varphi_{\ell}$.

Computations

The goal of MFA is to integrate different groups of variables (i.e., different subtables) describing the same observations. In order to do so, the first step is to make these subtables comparable. Such a step is needed because the straightforward analysis obtained by concatenating all variables would be dominated by the subtable with the strongest structure (which would be the subtable with the largest first singular value). In order to make the subtables comparable, we need to normalize them. To normalize a subtable, we first compute a PCA for this subtable. The first singular value (i.e., the square root of the first eigenvalue) is the normalizing factor which is used to divide the elements of this subtable. So, formally, The normalized subtables are computed as:

$$[t]\mathbf{Z} = \frac{1}{\sqrt{[t]\varrho_1}} \times [t]\mathbf{X} = \frac{1}{[t]\varphi_1} \times [t]\mathbf{X} \qquad (32)$$

The normalized subtables are concatenated into an $I \times J$ matrix called the *global data matrix* and denoted $\mathbf{Z}$. A PCA is then run on $\mathbf{Z}$ to get a global solution. Note that because the subtables have previously been centered and normalized with their first singular value, $\mathbf{Z}$ is centered but it is *not* normalized (i.e., columns from different subtables have, in general, different norms).

To find out how each subtable performs relative to the global solution, each subtable (i.e., each $[t]\mathbf{X}$) is projected into the global space as a supplementary element.

As in standard PCA, variable loadings are correlations between original variables and global

factor scores. To find the relationship between the variables from each of the subtables and the global solution we compute loadings (i.e., correlations) between the components of each subtable and the

components of the global analysis.

Example

Suppose that three experts were asked to rate six wines aged in two different kinds of oak barrel from the same harvest of Pinot Noir [example from Ref 55]. Wines 1, 5, and 6 were aged with a first type of oak, and Wines 2, 3, and 4 with a second type of oak. Each expert was asked to choose from 2 to 5 variables to describe the wines. For each wine, the expert rated the intensity of his/her variables on a 9-point scale. The data consist of $T = 3$ subtables, which are presented in Table 18.

The PCAs on each of the three subtables extracted eigenvalues of ${}_1\varrho_1 = 2.86$, ${}_2\varrho_1 = 3.65$, and ${}_3\varrho_1 = 2.50$ with singular values of ${}_1\varphi_1 = 1.69$, ${}_2\varphi_1 = 1.91$, and ${}_3\varphi_1 = 1.58$, respectively.

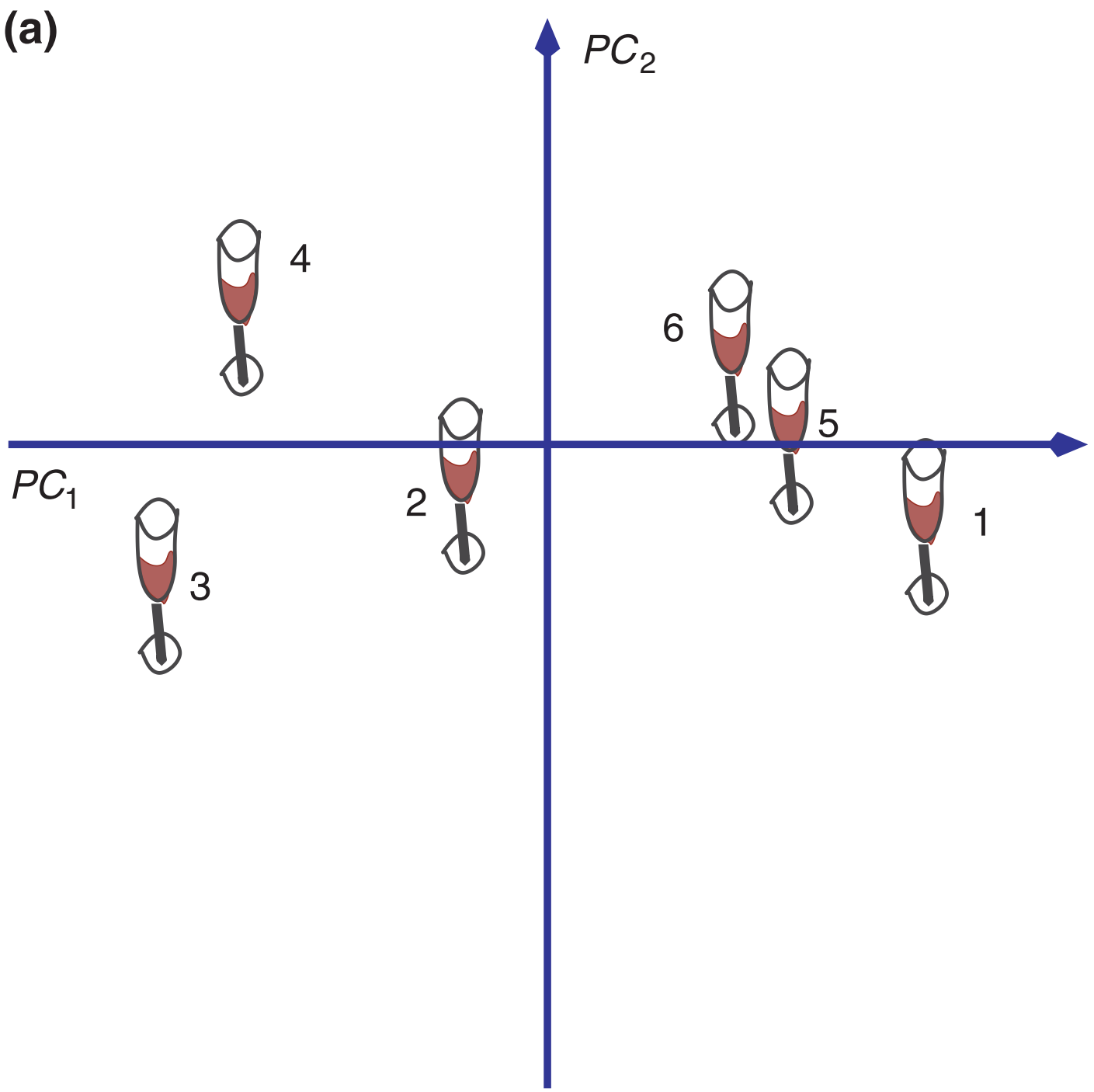
Following normalization and concatenation of the subtables, the global PCA extracted five components (with eigenvalues of 2.83, 0.36, 0.11, 0.03, and 0.01). The first two components explain 95% of the inertia. The factor scores for the first two components of the global analysis are given in Table 19 and the corresponding map is displayed in Figure 11a.

We can see from Figure 11 that the first

component separates the first type of oak (Wines 1, 5, and 6) from the second oak type (Wines 2, 3, and 4).

In addition to examining the placement of the wines, we wanted to see how each expert's ratings fit into the global PCA space. We achieved this by projecting the data set of each expert as a supplementary element [see Ref 18 for details of the procedure]. The factor scores are shown in Table 19. The experts' placement in the global map is shown in Figure 11b. Note that the position of each wine in the global analysis is the center of gravity of its position for the experts. The projection of the experts shows that Expert 3's ratings differ from those of the other two experts.

The variable loadings show the correlations between the original variables and the global factor scores (Table 20). These loadings are plotted in Figure 12. This figure also represents the loadings (Table 21) between the components of each subtable and the components of the global analysis as the 'circle of correlations' specific to each expert. From this we see that Expert 3 differs from the other experts, and is mostly responsible for the second component of the global PCA.


(a)
PC_2
PC_1
4
6
5
2
1
3

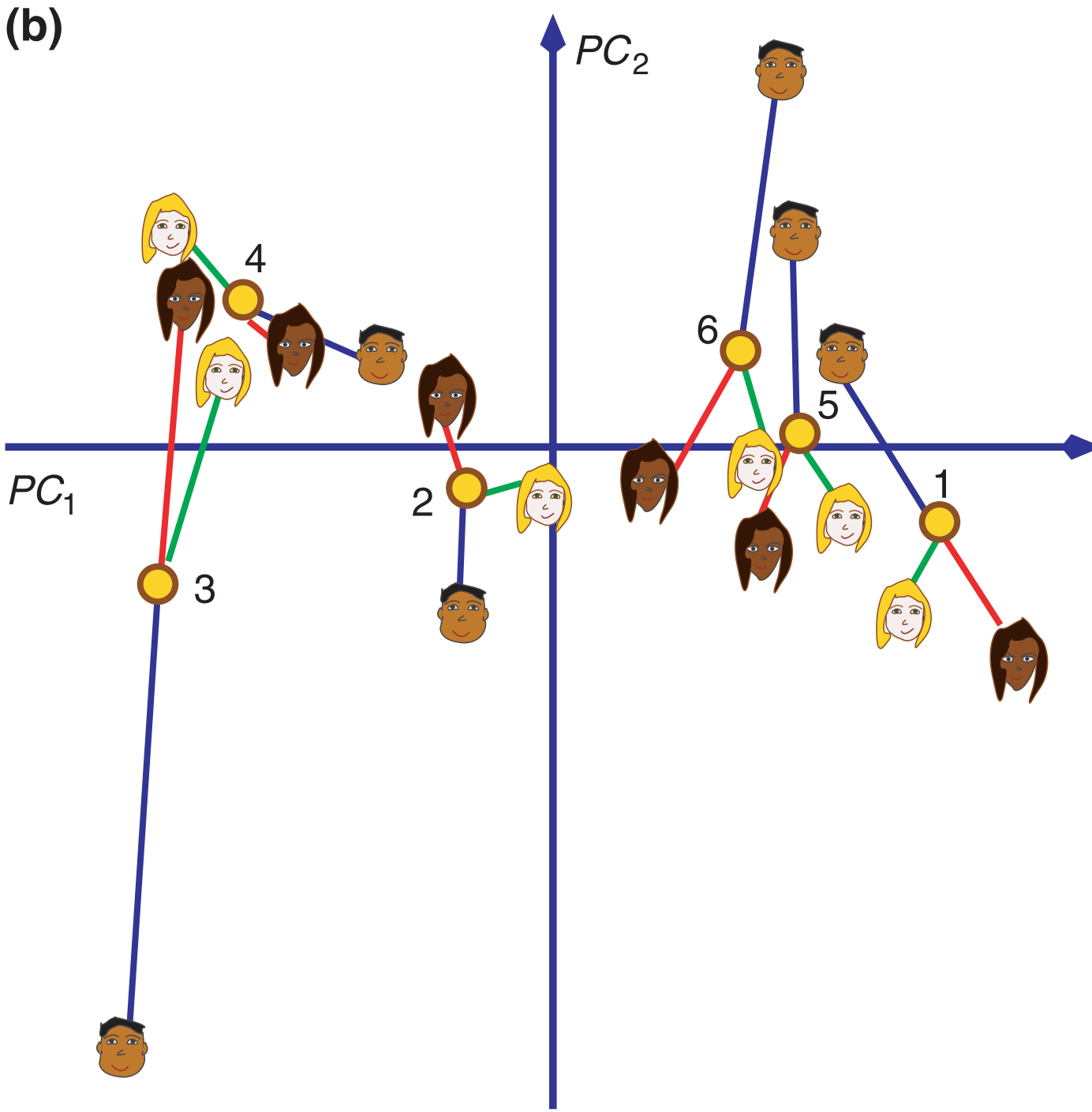


FIGURE 11 | MFA wine ratings and oak type. (a) Plot of the global analysis of the wines on the first two principal components. (b) Projection of the experts onto the global analysis. Experts are represented by their faces. A line segment links the position of the wine for a given expert to its global position. $\lambda_1 = 2.83$, $\tau_1 = 84\%$; $\lambda_2 = 2.83$, $\tau_2 = 11\%$.

CONCLUSION

PCA is very versatile, it is the oldest and remains the most popular technique in multivariate analysis. In addition to the basics presented here, PCA can also be interpreted as a neural network model [see, e.g., Refs 56,57]. In addition to CA, covered in this paper, generalized PCA can also be shown to incorporate a very large set of multivariate techniques such as

TABLE 18 | Raw Data for the Wine Example [from Ref 55

		Expert 1			Expert 2				Expert 3		
Wines	Oak Type	Fruity	Woody	Coffee	Red Fruit	Roasted	Vanillin	Woody	Fruity	Butter	Woody
Wine 1	1	1	6	7	2	5	7	6	3	6	7
Wine 2	2	5	3	2	4	4	4	2	4	4	3
Wine 3	2	6	1	1	5	2	1	1	7	1	1
Wine 4	2	7	1	2	7	2	1	2	2	2	2
Wine 5	1	2	5	4	3	5	6	5	2	6	6
Wine 6	1	3	4	4	3	5	4	5	1	7	5

canonical variate analysis, linear discriminant analysis [see, e.g., Ref 47], and barycentric discriminant analysis techniques such as discriminant CA [see e.g., Refs 58–61].

APPENDIX A: EIGENVECTORS AND

EIGENVALUES

Eigenvectors and *eigenvalues* are numbers and vectors associated to square matrices. Together they provide the *eigen-decomposition* of a matrix, which analyzes the structure of this matrix. Even though the eigen-decomposition does not exist for all square matrices, it has a particularly simple expression for matrices such as correlation, covariance, or cross-product matrices. The eigen-decomposition of this type of matrices is important because it is used to find the maximum (or minimum) of functions involving these matrices. Specifically PCA is obtained from the eigen-decomposition of a covariance or a correlation matrix.

Notations and Definition

There are several ways to define eigenvectors and eigenvalues, the most common approach defines an eigenvector of the matrix $\mathbf{A}$ as a vector $\mathbf{u}$ that satisfies the following equation:

$$\mathbf{Au} = \lambda\mathbf{u} \; . \tag{A.1}$$

When rewritten, the equation becomes:

$$(\mathbf{A} - \lambda\mathbf{I})\mathbf{u} = 0, \tag{A.2}$$

TABLE 19 | MFA Wine Ratings and Oak Type. Factor Scores for the Global Analysis, Expert 1, Expert 2, and Expert 3 for the First Two Components

	Global		Expert 1_{sup}		Expert 2_{sup}		Expert 3_{sup}	
	F_1	F_2	$_{[1]}F_1$	$_{[1]}F_2$	$_{[2]}F_1$	$_{[2]}F_2$	$_{[3]}F_1$	$_{[3]}F_2$
Wine 1	2.18	−0.51	2.76	−1.10	2.21	−0.86	1.54	0.44
Wine 2	−0.56	−0.20	−0.77	0.30	−0.28	−0.13	−0.61	−0.76
Wine 3	−2.32	−0.83	−1.99	0.81	−2.11	0.50	−2.85	−3.80
Wine 4	−1.83	0.90	−1.98	0.93	−2.39	1.23	−1.12	0.56
Wine 5	1.40	0.05	1.29	−0.62	1.49	−0.49	1.43	1.27
Wine 6	1.13	0.58	0.69	−0.30	1.08	−0.24	1.62	2.28

sup, supplementary element.

TABLE 20 | MFA Wine Ratings and Oak Type. Loadings (i.e., correlations) on the Principal Components of the Global Analysis of the Original Variables. Only the First Three Dimensions are Kept

			Loadings with Original Variables									
			Expert 1			Expert 2				Expert 3		
PC	λ	τ (%)	Fruity	Woody	Coffee	Fruit	Roasted	Vanillin	Woody	Fruity	Butter	Woody
1	2.83	85	−0.97	0.98	0.92	−0.89	0.96	0.95	0.97	−0.59	0.95	0.99
2	0.36	11	0.23	−0.15	−0.06	0.38	−0.00	−0.20	0.10	−0.80	0.19	0.00
3	0.12	3	0.02	−0.02	−0.37	−0.21	0.28	−0.00	−0.14	0.08	0.24	−0.11

TABLE 21 | MFA Wine Ratings and Oak Type. Loadings (i.e., correlations) on the Principal Components of the Global Analysis of the Principal Components of the Subtable PCAs

			Loadings with First Two Components from Subtable PCAs					
			Expert 1		Expert 2		Expert 3	
PC	λ	τ (%)	$_{[1]}PC_1$	$_{[1]}PC_2$	$_{[2]}PC_1$	$_{[2]}PC_2$	$_{[3]}PC_1$	$_{[3]}PC_2$
1	2.83	85	.98	0.08	0.99	−0.16	0.94	−0.35
2	0.36	11	−0.15	−0.28	−0.13	−0.76	0.35	0.94
3	0.12	3	−0.14	0.84	0.09	0.58	0.05	−.01

Only the first three dimensions are kept.

where λ is a scalar called the *eigenvalue* associated to the *eigenvector*.

In a similar manner, we can also say that a vector $\mathbf{u}$ is an eigenvector of a matrix $\mathbf{A}$ if the length of the vector (but not its direction) is changed when it is multiplied by $\mathbf{A}$.

For example, the matrix:

$$\mathbf{A} = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \tag{A.3}$$

has for eigenvectors:

$$\mathbf{u}_1 = \begin{bmatrix} 3 \\ 2 \end{bmatrix} \qquad \text{with eigenvalue} \quad \lambda_1 = 4 \tag{A.4}$$

and

$$\mathbf{u}_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix} \qquad \text{with eigenvalue} \quad \lambda_2 = -1 \tag{A.5}$$

For most applications we normalize the eigenvectors (i.e., transform them such that their length is equal to one), therefore

$$\mathbf{u}^{\mathsf{T}}\mathbf{u} = 1 \tag{A.6}$$

Traditionally, we put the set of eigenvectors of $\mathbf{A}$ in a matrix denoted $\mathbf{U}$. Each column of $\mathbf{U}$ is an eigenvector of $\mathbf{A}$. The eigenvalues are stored in a diagonal matrix (denoted $\mathbf{\Lambda}$), where the diagonal elements gives the eigenvalues (and all the other values are zeros). We can rewrite the Eq. A.1 as:

$$\mathbf{AU} = \mathbf{\Lambda U}; \tag{A.7}$$

or also as:

$$\mathbf{A} = \mathbf{U\Lambda U}^{-1}. \tag{A.8}$$

For the previous example we obtain:

$$\begin{aligned}\mathbf{A} &= \mathbf{U\Lambda U}^{-1} \\ &= \begin{bmatrix} 3 & -1 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & -1 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ -4 & 6 \end{bmatrix} \\ &= \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix}\end{aligned} \tag{A.9}$$

Together, the eigenvectors and the eigenvalues of a matrix constitute the *eigen-decomposition* of this

matrix. It is important to note that not all matrices have an eigen-decomposition. This is the case, e.g., of the matrix $\begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}$. Also some matrices can have imaginary eigenvalues and eigenvectors.

Positive (semi-)Definite Matrices

A type of matrices used very often in statistics are called *positive semi-definite*. The eigen-decomposition of these matrices always exists, and has a particularly convenient form. A matrix is said to be positive semi-definite when it can be obtained as the product of a matrix by its transpose. This implies that a positive semi-definite matrix is always symmetric. So, formally, the matrix $\mathbf{A}$ is positive semi-definite if it can be obtained as:

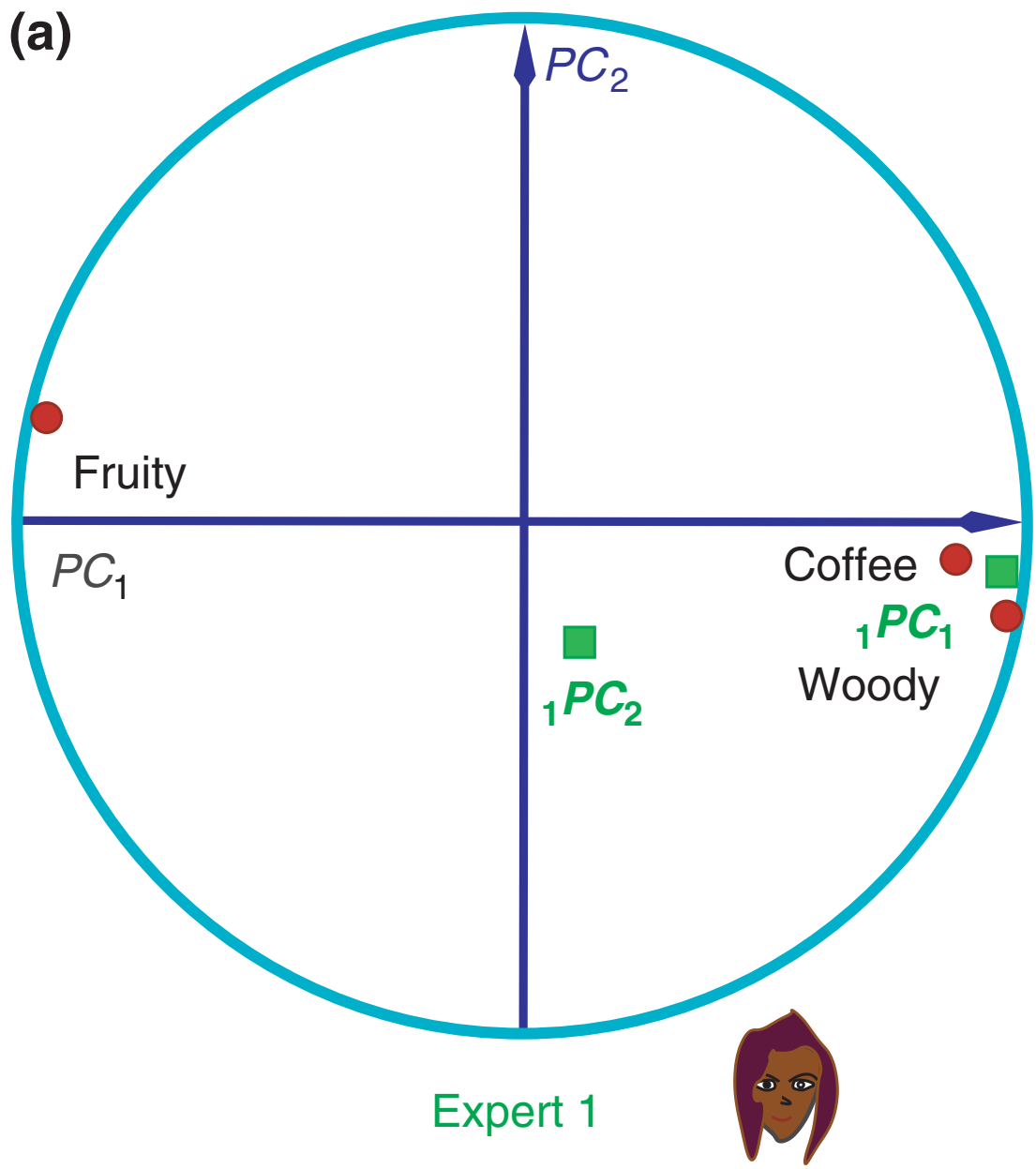
$$\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathsf{T}} \tag{A.10}$$

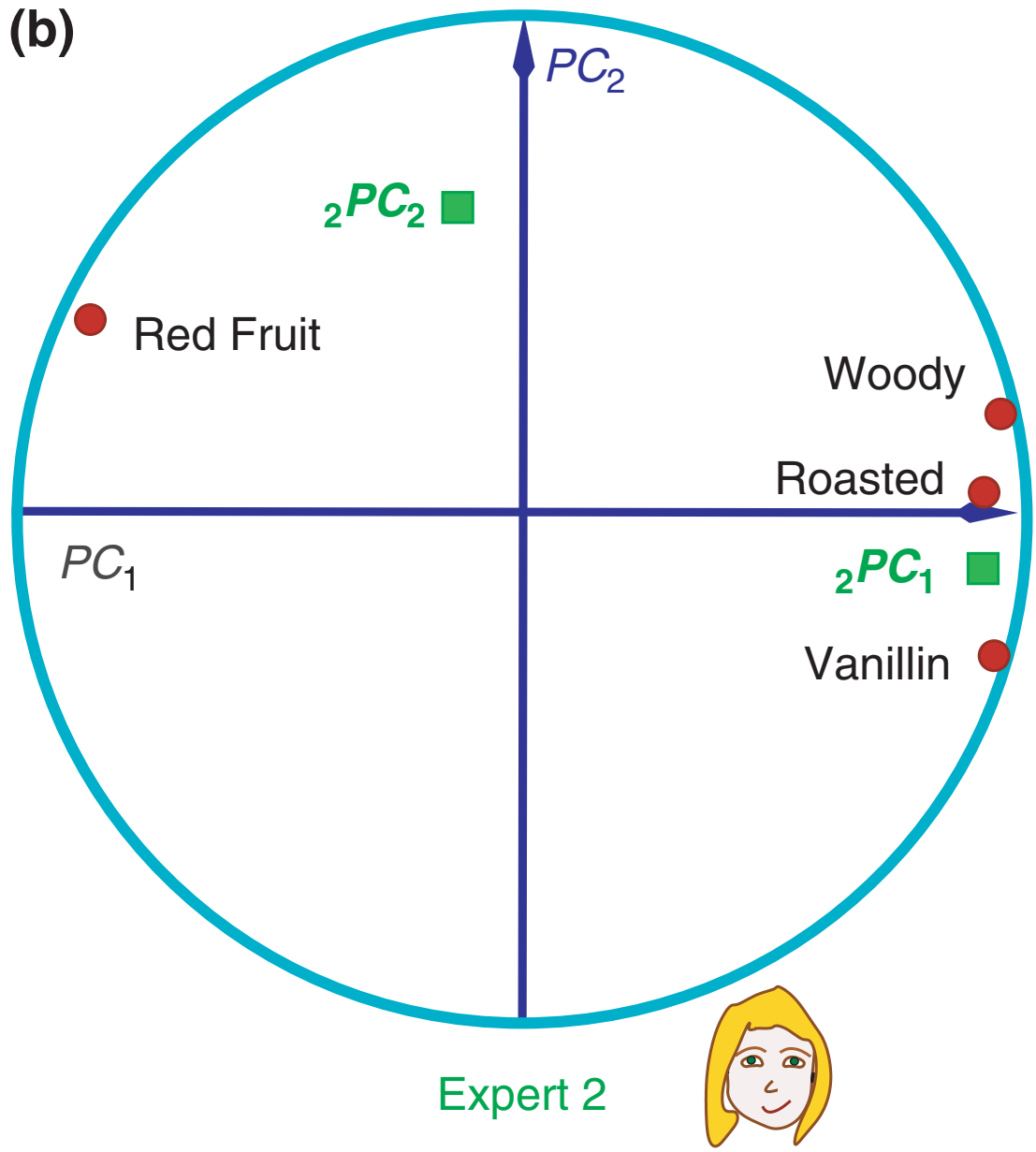
for a certain matrix $\mathbf{X}$ (containing real numbers). In particular, correlation matrices, covariance, and cross-product matrices are all positive semi-definite matrices.

The important properties of a positive semi-definite matrix is that its eigenvalues are always

positive or null, and that its eigenvectors are pairwise orthogonal when their eigenvalues are different. The eigenvectors are also composed of real values (these last two properties are a consequence of the symmetry of the matrix, for proofs see, e.g.,

© 2010 John Wiley & Sons, Inc.




(b)
PC2
2PC2
Red Fruit
Woody
Roasted
PC1
2PC1
Vanillin
Expert 2

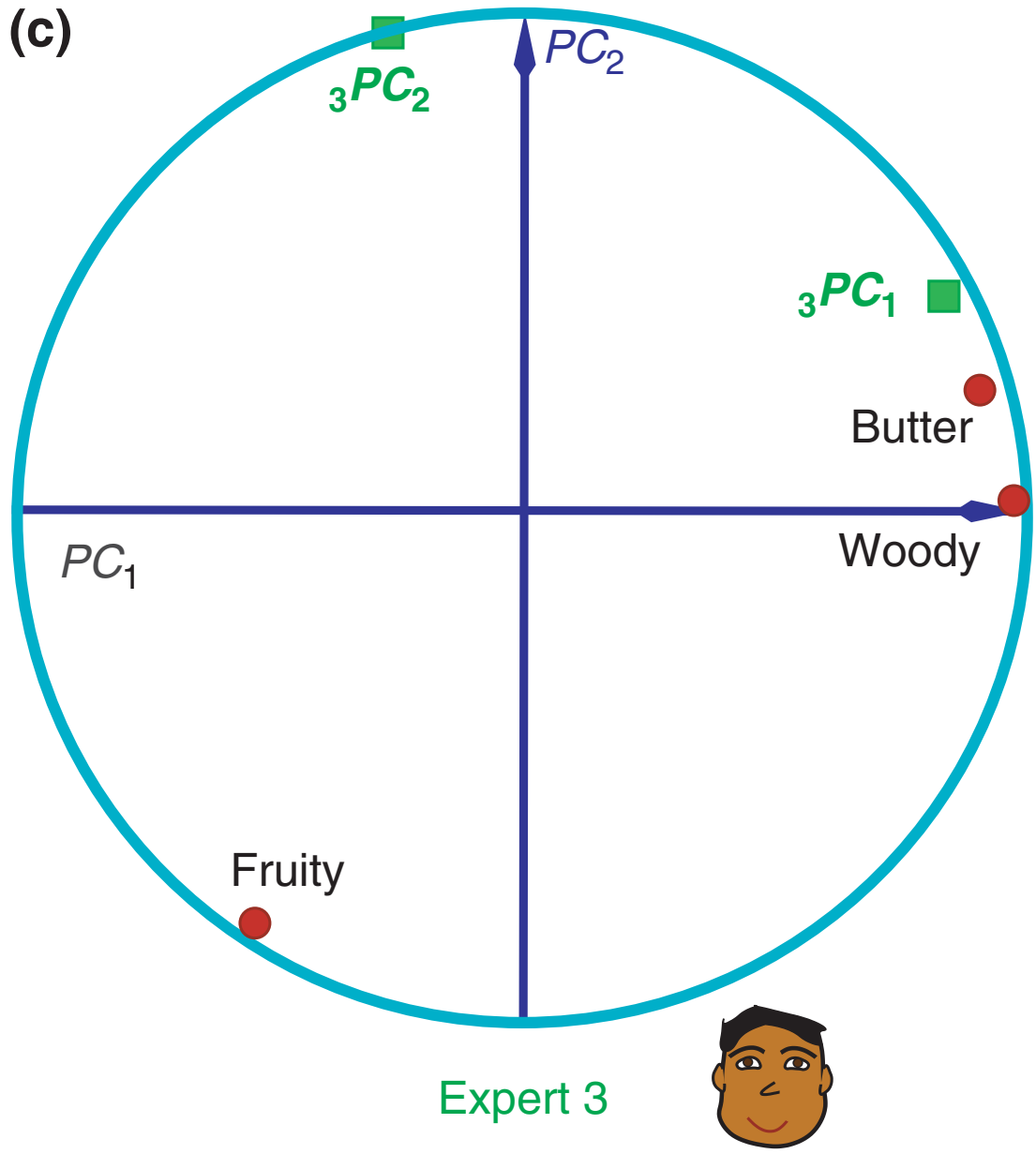


FIGURE 12 | MFA wine ratings and oak type. Circles of correlations for the original variables. Each experts' variables have been separated for ease of interpretation. .

Refs 62,63). Because eigenvectors corresponding to different eigenvalues are orthogonal, it is possible to store all the eigenvectors in an orthogonal matrix (recall that a matrix is orthogonal when the product of this matrix by its transpose is a diagonal matrix).

This implies the following equality:

$$\mathbf{U}^{-1} = \mathbf{U}^{\mathsf{T}}. \tag{A.11}$$

We can, therefore, express the positive semi-definite matrix $\mathbf{A}$ as:

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}} \qquad \text{with} \quad \mathbf{U}^{\mathsf{T}}\mathbf{U} = \mathbf{I} \tag{A.12}$$

where $\mathbf{U}$ is the matrix storing the normalized eigenvectors; if these are not normalized then $\mathbf{U}^{\mathsf{T}}\mathbf{U}$ is a diagonal matrix.

For example, the matrix:

$$\mathbf{A} = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \tag{A.13}$$

can be decomposed as:

$$\begin{aligned} \mathbf{A} &= \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^{\mathsf{T}} \\ &= \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \\ &= \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \end{aligned} \tag{A.14}$$

with

$$\begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} \begin{bmatrix} \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} \\ \sqrt{\frac{1}{2}} & -\sqrt{\frac{1}{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{(A.15)}$$

Statistical Properties of the Eigen-decomposition

The eigen-decomposition is important because it is involved in problems of optimization. Specifically, in PCA, we want to find row *factor scores*, obtained as linear combinations of the columns of $\mathbf{X}$ such that these factor scores 'explain' as much of the variance of $\mathbf{X}$ as possible and such that the sets of factor scores are pairwise orthogonal. We impose as a constraint that the coefficients of the linear combinations are finite and this constraint is, in general, expressed as imposing to the sum of squares of the coefficients of each linear combination to be equal to unity. This amounts to defining the factor score matrix as:

$$\mathbf{F} = \mathbf{XQ}, \quad \text{(A.16)}$$

© 2010 John Wiley & Sons, Inc.

(with the matrix **Q** being the matrix of coefficients of the 'to-be-found' linear combinations) under the constraints that

$$\mathbf{F}^{\mathsf{T}}\mathbf{F} = \mathbf{Q}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{Q} \tag{A.17}$$

is a diagonal matrix (i.e., **F** is an orthogonal matrix) and that

$$\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I} \tag{A.18}$$

(i.e., **Q** is an orthonormal matrix). The solution of this problem can be obtained with the technique of the Lagrangian multipliers where the constraint from Eq. A.18 is expressed as the multiplication with a diagonal matrix of Lagrangian multipliers, denoted $\mathbf{\Lambda}$, in order to give the following expression

$$\mathbf{\Lambda}\left(\mathbf{Q}^{\mathsf{T}}\mathbf{Q} - \mathbf{I}\right) \tag{A.19}$$

(see Refs 62,64; for details). This amount for defining the following equation

$$\begin{aligned}\mathcal{L} &= \mathsf{trace}\left\{\mathbf{F}^{\mathsf{T}}\mathbf{F} - \boldsymbol{\Lambda}\left(\mathbf{Q}^{\mathsf{T}}\mathbf{Q} - \mathbf{I}\right)\right\} \\ &= \mathsf{trace}\left\{\mathbf{Q}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{Q} - \boldsymbol{\Lambda}\left(\mathbf{Q}^{\mathsf{T}}\mathbf{Q} - \mathbf{I}\right)\right\}\end{aligned} \tag{A.20}$$

(where the $\mathsf{trace}\{\}$ operator gives the sum of the diagonal elements of a square matrix). In order to find the values of $\mathbf{Q}$ which give the maximum values of $\mathcal{L}$, we first compute the derivative of $\mathcal{L}$ relative to $\mathbf{Q}$:

$$\frac{\partial \mathcal{L}}{\partial \mathbf{Q}} = 2\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{Q} - 2\mathbf{Q}\boldsymbol{\Lambda}, \tag{A.21}$$

and then set this derivative to zero:

$$\mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{Q} - \mathbf{Q}\boldsymbol{\Lambda} = 0 \Longleftrightarrow \mathbf{X}^{\mathsf{T}}\mathbf{X}\mathbf{Q} = \mathbf{Q}\boldsymbol{\Lambda}. \tag{A.22}$$

This implies also that

$$\mathbf{X}^{\mathsf{T}}\mathbf{X} = \mathbf{Q}\boldsymbol{\Lambda}\mathbf{Q}^{\mathsf{T}}. \tag{A.23}$$

Because $\boldsymbol{\Lambda}$ is diagonal, this is clearly an eigen-decomposition problem, and this indicates that $\boldsymbol{\Lambda}$ is the matrix of eigenvalues of the positive semi-definite matrix $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ ordered from the largest to the smallest

and that $\mathbf{Q}$ is the matrix of eigenvectors of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$ associated to $\mathbf{\Lambda}$. Finally, we find that the factor matrix has the form

$$\mathbf{F} = \mathbf{XQ}. \tag{A.24}$$

The variance of the factors scores is equal to the eigenvalues because:

$$\mathbf{F}^{\mathsf{T}}\mathbf{F} = \mathbf{Q}^{\mathsf{T}}\mathbf{X}^{\mathsf{T}}\mathbf{XQ} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{\Lambda}. \tag{A.25}$$

Taking into account that the sum of the eigenvalues is equal to the trace of $\mathbf{X}^{\mathsf{T}}\mathbf{X}$, this shows that the first factor scores 'extract' as much of the variance of the original data as possible, and that the second factor scores extract as much of the variance left unexplained by the first factor, and so on for the remaining factors. Incidently, the diagonal elements of the matrix $\mathbf{\Lambda}^{\frac{1}{2}}$ which are the standard deviations of the factor scores are called the *singular values* of matrix $\mathbf{X}$ (see Section on Singular Value Decomposition).

APPENDIX B: SINGULAR VALUE DECOMPOSITION (SVD)

The SVD is a generalization of the eigen-decomposition. The SVD decomposes a rectangular matrix into three simple matrices: two orthogonal matrices and one diagonal matrix. If $\mathbf{A}$ is a rectangular matrix, its SVD gives

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}}, \tag{B.1}$$

with

- $\mathbf{P}$: The (normalized) eigenvectors of the matrix $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ (i.e., $\mathbf{P}^{\mathsf{T}}\mathbf{P} = \mathbf{I}$). The columns of $\mathbf{P}$ are called the *left singular vectors* of $\mathbf{A}$.
- $\mathbf{Q}$: The (normalized) eigenvectors of the matrix $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ (i.e., $\mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}$). The columns of $\mathbf{Q}$ are called the *right singular vectors* of $\mathbf{A}$.
- $\boldsymbol{\Delta}$: The diagonal matrix of the *singular values*, $\boldsymbol{\Delta} = \boldsymbol{\Lambda}^{\frac{1}{2}}$ with $\boldsymbol{\Lambda}$ being the diagonal matrix of the eigenvalues of matrix $\mathbf{A}\mathbf{A}^{\mathsf{T}}$ and of the matrix $\mathbf{A}^{\mathsf{T}}\mathbf{A}$ (as they are the same).

The SVD is a straightforward consequence of the eigen-decomposition of positive semi-definite matrices [see, e.g., Refs 5,11,47,65].

Note that Eq. B.1 can also be rewritten as:

$$\mathbf{A} = \mathbf{P}\boldsymbol{\Delta}\mathbf{Q}^{\mathsf{T}} = \sum_{\ell=1}^{L} \delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^{\mathsf{T}}, \tag{B.2}$$

with L being the rank of $\mathbf{X}$ and δ_ℓ, $\mathbf{p}_\ell$, and $\mathbf{q}_\ell$ being (respectively) the ℓ-th singular value, left and right singular vectors of $\mathbf{X}$. This shows that $\mathbf{X}$ can be

© 2010 John Wiley & Sons, Inc.

reconstituted as a sum of L rank one matrices (i.e., the $\delta_\ell \mathbf{p}_\ell \mathbf{q}_\ell^{\mathsf{T}}$ terms). The first of these matrices gives the best reconstitution of $\mathbf{X}$ by a rank one matrix, the sum of the first two matrices gives the best reconstitution of $\mathbf{X}$ with a rank two matrix, and so on, and, in general, the sum of the first M matrices gives the best reconstitution of $\mathbf{X}$ with a matrix of rank M.

Generalized Singular Value Decomposition

The generalized singular value decomposition (GSVD) decomposes a rectangular matrix and takes into account constraints imposed on the rows and the columns of the matrix. The GSVD gives a weighted

generalized least square estimate of a given matrix by a lower rank matrix For a given $I \times J$ matrix $\mathbf{A}$, generalizing the SVD, involves using two positive definite square matrices with size $I \times I$ and $J \times J$. These two matrices express constraints imposed on the rows and the columns of $\mathbf{A}$, respectively. Formally, if $\mathbf{M}$ is the $I \times I$ matrix expressing the constraints for the rows of $\mathbf{A}$ and $\mathbf{W}$ the $J \times J$ matrix of the constraints for the columns of $\mathbf{A}$. The matrix $\mathbf{A}$ is now decomposed into:

$$\mathbf{A} = \widetilde{\mathbf{P}}\widetilde{\mathbf{\Delta}}\widetilde{\mathbf{Q}}^{\mathsf{T}} \qquad \text{with: } \widetilde{\mathbf{P}}^{\mathsf{T}}\mathbf{M}\widetilde{\mathbf{P}} = \widetilde{\mathbf{Q}}^{\mathsf{T}}\mathbf{W}\widetilde{\mathbf{Q}} = \mathbf{I}. \quad \text{(B.3)}$$

In other words, the generalized singular vectors are orthogonal under the constraints imposed by $\mathbf{M}$ and $\mathbf{W}$.

This decomposition is obtained as a result of the standard SVD. We begin by defining the matrix $\widetilde{\mathbf{A}}$ as:

$$\widetilde{\mathbf{A}} = \mathbf{M}^{\frac{1}{2}}\mathbf{A}\mathbf{W}^{\frac{1}{2}} \iff \mathbf{A} = \mathbf{M}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{W}^{-\frac{1}{2}}. \quad \text{(B.4)}$$

We then compute the standard singular value decomposition as $\widetilde{\mathbf{A}}$ as:

$$\widetilde{\mathbf{A}} = \mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathsf{T}} \qquad \text{with: } \mathbf{P}^{\mathsf{T}}\mathbf{P} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}. \quad \text{(B.5)}$$

The matrices of the generalized eigenvectors are obtained as:

$$\widetilde{\mathbf{P}} = \mathbf{M}^{-\frac{1}{2}}\mathbf{P} \quad \text{and} \quad \widetilde{\mathbf{Q}} = \mathbf{W}^{-\frac{1}{2}}\mathbf{Q}. \tag{B.6}$$

The diagonal matrix of singular values is simply equal to the matrix of singular values of $\widetilde{\mathbf{A}}$:

$$\widetilde{\mathbf{\Delta}} = \mathbf{\Delta}. \tag{B.7}$$

We verify that:

$$\mathbf{A} = \widetilde{\mathbf{P}}\widetilde{\mathbf{\Delta}}\widetilde{\mathbf{Q}}^{\mathsf{T}}$$

by substitution:

$$\begin{aligned}\mathbf{A} &= \mathbf{M}^{-\frac{1}{2}}\widetilde{\mathbf{A}}\mathbf{W}^{-\frac{1}{2}} \\ &= \mathbf{M}^{-\frac{1}{2}}\mathbf{P}\mathbf{\Delta}\mathbf{Q}^{\mathsf{T}}\mathbf{W}^{-\frac{1}{2}} \\ &= \widetilde{\mathbf{P}}\mathbf{\Delta}\widetilde{\mathbf{Q}}^{\mathsf{T}} \qquad \text{(from Eq. B.6).}\end{aligned} \tag{B.8}$$

To show that Condition B.3 holds, suffice it to show that:

$$\widetilde{\mathbf{P}}^{\mathsf{T}}\mathbf{M}\widetilde{\mathbf{P}} = \mathbf{P}^{\mathsf{T}}\mathbf{M}^{-\frac{1}{2}}\mathbf{M}\mathbf{M}^{-\frac{1}{2}}\mathbf{P} = \mathbf{P}^{\mathsf{T}}\mathbf{P} = \mathbf{I} \tag{B.9}$$

and

$$\widetilde{\mathbf{Q}}^{\mathsf{T}}\mathbf{W}\widetilde{\mathbf{Q}} = \mathbf{Q}^{\mathsf{T}}\mathbf{W}^{-\frac{1}{2}}\mathbf{W}\mathbf{W}^{-\frac{1}{2}}\mathbf{Q} = \mathbf{Q}^{\mathsf{T}}\mathbf{Q} = \mathbf{I}. \tag{B.10}$$

Mathematical Properties of the Singular Value Decomposition

It can be shown that the SVD has the important property of giving an optimal approximation of a matrix by another matrix of smaller rank [see, e.g., Ref 62,63,65]. In particular, the SVD gives the best approximation, in a least square sense, of any rectangular matrix by another rectangular matrix of same dimensions, but smaller rank.

Precisely, if $\mathbf{A}$ is an $I \times J$ matrix of rank L (i.e., $\mathbf{A}$ contains L singular values that are not zero), we denote it by $\mathbf{P}^{[M]}$ (respectively $\mathbf{Q}^{[M]}$, $\mathbf{\Delta}^{[M]}$) the matrix made of the first M columns of $\mathbf{P}$ (respectively $\mathbf{Q}$, $\mathbf{\Delta}$):

$$\mathbf{P}^{[M]} = \left[\mathbf{p}_1, \ldots, \mathbf{p}_m, \ldots, \mathbf{p}_M\right] \tag{B.11}$$

$$\mathbf{Q}^{[M]} = \left[\mathbf{q}_1, \ldots, \mathbf{q}_m, \ldots, \mathbf{q}_M\right] \tag{B.12}$$

$$\boldsymbol{\Delta}^{[M]} = \mathrm{diag}\left\{\delta_1, \ldots, \delta_m, \ldots, \delta_M\right\} \ . \tag{B.13}$$

The matrix $\mathbf{A}$ reconstructed from the first M eigenvectors is denoted $\mathbf{A}^{[M]}$. It is obtained as:

$$\mathbf{A}^{[M]} = \mathbf{P}^{[M]}\boldsymbol{\Delta}^{[M]}\mathbf{Q}^{[M]\mathsf{T}} = \sum_{m}^{M} \delta_m \mathbf{p}_m \mathbf{q}_m^{\mathsf{T}}, \tag{B.14}$$

(with δ_m being the m-th singular value).

The reconstructed matrix $\mathbf{A}^{[M]}$ is said to be optimal (in a least squares sense) for matrices of rank M because it satisfies the following condition:

$$\begin{aligned} \left\|\mathbf{A} - \mathbf{A}^{[M]}\right\|^2 &= \mathrm{trace}\left\{\left(\mathbf{A} - \mathbf{A}^{[M]}\right)\left(\mathbf{A} - \mathbf{A}^{[M]}\right)^{\mathsf{T}}\right\} \\ &= \min_{\mathbf{X}} \left\|\mathbf{A} - \mathbf{X}\right\|^2 \end{aligned} \tag{B.15}$$

for the set of matrices $\mathbf{X}$ of rank smaller or equal to M [see, e.g., Ref 65,66]. The quality of the reconstruction is given by the ratio of the first M eigenvalues (i.e., the

squared singular values) to the sum of all the eigenvalues. This quantity is interpreted as *the reconstructed proportion* or *the explained variance*, it corresponds to the inverse of the quantity minimized by Eq. B.14. The quality of reconstruction can also be interpreted as the squared coefficient of correlation [precisely as the R_v coefficient,[18] between the original matrix and its approximation.

The GSVD minimizes an expression similar to Eq. B.14, namely

$$\mathbf{A}^{[M]} = \min_{\mathbf{X}} \left[\mathsf{trace} \left\{ \mathbf{M} \left(\mathbf{A} - \mathbf{X} \right) \mathbf{W} \left(\mathbf{A} - \mathbf{X} \right)^{\mathsf{T}} \right\} \right], \tag{B.16}$$

for the set of matrices $\mathbf{X}$ of rank smaller or equal to M.

ACKNOWLEDGEMENT

We would like to thank Yoshio Takane for his very helpful comments on a draft of this article.

REFERENCES

1. Pearson K. On lines and planes of closest fit to systems of points in space. *Philos Mag A* 1901, 6:559–572.

2. Cauchy AL. *Sur l'équation à l'aide de laquelle on détermine les inégalités séculaires des mouvements des planètes*, vol. 9. Œuvres Complètes (IIème Série); Paris: Blanchard; 1829.

3. Grattan-Guinness I. *The Rainbow of Mathematics*. New York: Norton; 1997.

4. Jordan C. Mémoire sur les formes bilinéaires. *J Math Pure Appl* 1874, 19:35–54.

5. Stewart GW. On the early history of the singular value decomposition. *SIAM Rev* 1993, 35:551–566.

6. Boyer C, Merzbach U. *A History of Mathematics*. 2nd ed. New York: John Wiley & Sons; 1989.

7. Hotelling H. Analysis of a complex of statistical variables into principal components. *J Educ Psychol* 1933, 25:417–441.

8. Jolliffe IT. *Principal Component Analysis*. New York: Springer; 2002.

9. Jackson JE. *A User's Guide to Principal Components*.

New York: John Wiley & Sons; 1991.

10. Saporta G, Niang N. Principal component analysis: application to statistical process control. In: Govaert G, ed. *Data Analysis*. London: John Wiley & Sons; 2009, 1–23.

11. Abdi H. Singular Value Decomposition (SVD) and Generalized Singular Value Decomposition(GSVD). In: Salkind NJ, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks: Sage Publications; 2007, 907–912.
12. Abdi H. Eigen-decomposition: eigenvalues and eigenvectors. In: Salkind NJ, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks: Sage Publications; 2007, 304–308.

13. Takane Y. Relationships among various kinds of eigenvalue and singular value decompositions. In: Yanai H, Okada A, Shigemasu K, Kano Y, Meulman J, eds. *New Developments in Psychometrics*. Tokyo: Springer Verlag; 2002, 45–56.

14. Abdi H. Centroid. *Wiley Interdisciplinary Reviews: Computational Statistics,* 2009, 1:259–260.

15. Kruskal JB. Factor analysis and principal component analysis: Bilinear methods. In: Kruskal WH, Tannur JM, eds. *International Encyclopedia of Statistics*. New York: The Free Press; 1978, 307–330.

16. Gower J. Statistical methods of comparing different multivariate analyses of the same data. In: Hodson F, Kendall D, Tautu P, eds. *Mathemematics in the Archæological and Historical Sciences*. Edinburgh: Edingburh University Press; 1971, 138–149.

17. Lingoes J, Schönemann P. Alternative measures of fit for the Schönemann-Carrol matrix fitting algorithm. *Psychometrika* 1974, 39:423–427.

18. Abdi H. RV cofficient and congruence coefficient. In: Salkind NJ, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks: Sage Publications; 2007, 849–853.

19. Dray S. On the number of principal components: a test of dimensionality based on measurements of similarity between matrices. *Comput Stat Data Anal* 2008, 52:2228–2237.

20. Quenouille M. Notes on bias and estimation. *Biometrika* 1956, 43:353–360.

© 2010 John Wiley & Sons, Inc.

21. Efron B. *The Jackknife, the Bootstrap and other Resampling Plans*, vol. 83, CMBF-NSF Regional Conference Series in Applied Mathematics: New York SIAM; 1982.

22. Abdi H, Williams LJ. Jackknife. In: Salkind NJ, ed.

Encyclopedia of Research Design. Thousand Oaks: Sage Publications; 2010, (In press).

23. Peres-Neto PR, Jackson DA, Somers KM. How many principal components? stopping rules for determining the number of non-trivial axes revisited. *Comput Stat Data Anal* 2005, 49:974–997.

24. Cattell RB. The scree test for the number of factors. *Multivariate Behav Res* 1966, 1:245–276.

25. Kaiser HF. A note on Guttman's lower bound for the number of common factors. *Br J Math Stat Psychol* 1961, 14:1–2.

26. O'Toole AJ, Abdi H, Deffenbacher KA, Valentin D. A low dimensional representation of faces in the higher dimensions of the space. *J Opt Soc Am [Ser A]* 1993, 10:405–411.

27. Geisser S. A predictive approach to the random effect model. *Biometrika* 1974, 61:101–107.

28. Tennenhaus M. *La régression* PLS. Paris: Technip; 1998.

29. Stone M. Cross-validatory choice and assessment of statistical prediction. *J R Stat Soc [Ser A]* 1974, 36:111–133.

30. Wold S. PLS for multivariate linear modeling. In: van de Waterbeemd H, ed. *Chemometric Methods*

in Molecular Design. Weinheim: Wiley-VCH Verlag; 1995, 195–217.

31. Malinowski ER. *Factor Analysis in Chemistry*. 3rd ed. New York: John Wiley & Sons; 2002.

32. Eastment HT, Krzanowski WJ. Cross-validatory choice of the number of components from a principal component analysis. *Technometrics* 1982, 24:73–77.

33. Wold S. Cross-validatory estimation of the number of components in factor and principal component analysis. *Technometrics* 1978, 20:397–405.

34. Diaconis P, Efron B. Computer intensive methods in statistics. *Sci Am* 1983, 248:116–130.

35. Holmes S. Using the bootstrap and the R_v coefficient in the multivariate context. In: Diday E, ed. *Data Analysis, Learning, Symbolic and Numeric Knowledge*. New York: Nova Science; 1989, 119–132.

36. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman and Hall; 1993.

37. Jackson DA. Stopping rules in principal aomponents analysis: a comparison of heuristical and statistical approaches. *Ecology* 1993, 74:2204–2214.

38. Jackson DA. Bootstrapped principal components analysis: a reply to Mehlman et al. *Ecology* 1995, 76:644–645.

39. Mehlman DW, Sheperd UL, Kelt DA. Bootstrapping principal components analysis: a comment. *Ecology* 1995, 76:640–643.

40. Abdi H. Multivariate analysis. In: Lewis-Beck M, Bryman A, Futing T, eds. *Encyclopedia for Research Methods for the Social Sciences*. Thousand Oaks, CA: Sage Publications 2003, 669–702.

41. Kaiser HF. The varimax criterion for analytic rotation in factor analysis. *Psychometrika* 1958, 23:187–200.

42. Thurstone LL. *Multiple Factor Analysis*. Chicago, IL: University of Chicago Press; 1947.

43. Stone JV. *Independent Component Analysis: A Tutorial Introduction*. Cambridge: MIT Press; 2004.

44. Abdi H. Partial least square regression, Projection on latent structures Regression, PLS-Regression. *Wiley Interdisciplinary Reviews: Computational Statistics, 2* 2010, 97–106.

45. Lebart L, Fénelon JP. *Statistique et informatique appliquées*. Paris: Dunod; 1975.

46. Benzécri J-P. *L'analyse des données*, Vols. 1 and 2. Paris: Dunod; 1973.

47. Greenacre MJ. *Theory and Applications of Correspondence Analysis*. London: Academic Press; 1984.

48. Greenacre MJ. *Correspondence Analysis in Practice*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC; 2007.

49. Abdi H, Valentin D. Multiple correspondence analysis. In: Salkind NJ, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications; 2007, 651–657.

50. Hwang H, Tomiuk MA, Takane Y. Correspondence analysis, multiple correspondence analysis and recent developments. In: Millsap R, Maydeu-Olivares A, eds. *Handbook of Quantitative Methods in Psychology*. London: Sage Publications 2009, 243–263.

51. Abdi H, Williams LJ. Correspondence analysis. In: Salkind NJ, ed. *Encyclopedia of Research Design*. Thousand Oaks: Sage Publications; 2010, (In press).

52. Brunet E. Faut-il pondérer les données linguistiques. *CUMFID* 1989, 16:39–50.

53. Escofier B, Pagès J. *Analyses factorielles simples et multiples: objectifs, méthodes, interprétation*. Paris: Dunod; 1990.

54. Escofier B, Pagès J. Multiple factor analysis. *Comput Stat Data Anal* 1994, 18:121–140.

55. Abdi H, Valentin D. Multiple factor analysis (mfa). In: Salkind NJ, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications; 2007, 657–663.

56. Diamantaras KI, Kung SY. *Principal Component Neural Networks: Theory and Applications*. New York:

John Wiley & Sons; 1996.

57. Abdi H, Valentin D, Edelman B. *Neural Networks*. Thousand Oaks, CA: Sage; 1999.

58. Nakache JP, Lorente P, Benzécri JP, Chastang JF. Aspect pronostics et thérapeutiques de l'infarctus myocardique aigu. *Les Cahiers de, Analyse des Données*, 1977, 2:415–534.

59. Saporta G, Niang N. Correspondence analysis and classification. In: Greenacre M, Blasius J, eds. *Multiple Correspondence Analysis and Related Methods*. Boca Raton, FL: Chapman & Hall; 2006, 371–392.

60. Abdi H. Discriminant correspondence analysis. In: Salkind NJ, ed. *Encyclopedia of Measurement and Statistics*. Thousand Oaks, CA: Sage Publications; 2007, 270–275.

61. Abdi H, Williams LJ. Barycentric discriminant analyis (BADIA). In: Salkind NJ, ed. *Encyclopedia of Research Design*. Thousand Oaks, CA: Sage; 2010, (In press).

FURTHER READING

62. Abdi H, Valentin D. *Mathématiques pour les sciences cognitives*. Grenoble: PUG; 2006.
63. Strang G. *Introduction to Linear Algebra*. Cambridge, MA: Wellesley-Cambridge Press; 2003.
64. Harris RJ. *A Primer of Multivariate Statistics*. Mahwah, NJ: Lawrence Erlbaum Associates; 2001.
65. Good I. Some applications of the singular value decomposition of a matrix. *Technometrics* 1969, 11:823–831.
66. Eckart C, Young G. The approximation of a matrix by another of a lower rank. *Psychometrika* 1936, 1:211–218.
67. Abdi H. Factor Rotations. In Lewis-Beck M, Bryman A., Futimg T., eds. *Encyclopedia for Research Methods for the Social Sciences*. Thousand Oaks, CA: Sage Publications; 2003, 978–982.

Baskilevsky A. *Applied Matrix Algebra in the Statistical Sciences*. New York: Elsevier; 1983.

© 2010 John Wiley & Sons, Inc.