Dray, S. 2014. Analysing a pair of tables: coinertia analysis and duality diagrams. Pages 289–300 in J. Blasius and M. Greenacre, editors. Visualization and verbalization of data. CRC Press.

Chapter 18

Analysing a pair of tables: coinertia analysis and duality diagrams

Stéphane DRAY

In many fields (e.g., ecology, psychometrics, social science and marketing), researchers are faced with the challenge of summarizing the information contained in large data sets. In this context, multivariate analysis provides efficient tools for identifying the relationships between variables and the similarities between statistical units/individuals. Due to the natural boundaries between disciplines or schools of thought, several multivariate methods have been invented and reinvented by different groups in different countries for different purposes. This situation has resulted in a variety of apparently different methods that actually lead to the same equations for analyzing the same data. For instance, Greenacre (1984, chapter 1.3) detailed the history of correspondence analysis (CA) and showed how this method has been rediscovered several times in biometrics, psychometrics and linguistics. This process can be explained by the diversity of viewpoints adopted by researchers to describe a method (e.g., geometrical versus numerical or individual-centered versus variable-centered). Several authors have tried to provide a unifying mathematical framework to summarize the different properties of a given method and thus to identify analogies between existing methods. The *duality diagram* theory was first presented in Cazes (1970) and popularized by Cailliez and Pagès (1976) in a French book entitled "Introduction à l'Analyse des Données". Several French authors adopted this theory but I believe that it remains poorly known by statisticians outside France. Daniel Chessel, my PhD advisor, used the duality diagram as a formal way to develop new multivariate methods in ecology (e.g., Dolédec and Chessel, 1994; Dolédec et al., 1996). He implemented this framework in the ADE-4 software (Thioulouse et al., 1997) and several years later in the R package `ade4` (Chessel, Dufour and Thioulouse, 2004; Dray and Dufour, 2007; Dray, Dufour and Chessel, 2007). Hence, similarly to Obelix (Goscinny and Uderzo, 1989), I fell into the magical duality diagram when I was a little boy and then I have used it as a central framework in my further works.

Seven years after Cailliez and Pagès's book came out, Ramsay and de Leeuw

Dray, S. 2014. Analysing a pair of tables: coinertia analysis and duality diagrams. Pages 289–300 in J. Blasius and M. Greenacre, editors. Visualization and verbalization of data. CRC Press.

(1983) wrote a dithyrambic review and concluded that they "hope it will not be long before an English language counterpart appears". The book has never been translated into English, which is probably a major reason for its low impact on non-French readers. This lack is partially addressed by Escoufier (1987), Holmes (2006) and Dray and Dufour (2007), who provided a general overview of the duality diagram theory in English. More recently, a special section on modern multivariate analysis published in "The Annals of Applied Statistics" demonstrated the power of the duality diagram approach for analyzing data of different formats (De la Cruz and Holmes, 2011), including spatial (Dray and Jombart, 2011), temporal (Thioulouse, 2011) or phylogenetic (Purdom, 2011) information. To date, the most convincing application of the duality diagram is probably that found in Tenenhaus and Young (1985), which provided an overview of multiple correspondence analysis and related methods to quantify categorical multivariate data.

This chapter presents the duality diagram theory and its application to the analysis of a contingency table by correspondence analysis. Subsequently, I show how the framework can be generalized to the analysis of a pair of tables focusing on coinertia analysis (Dolédec and Chessel, 1994), and I conclude with several extensions.

18.1 The duality diagram

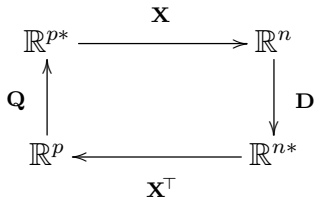
Definition

Let $\mathbf{X}$ be a matrix containing data for p variables (columns) collected from n individuals (rows). From a geometrical viewpoint, we can consider this information either as p points (the variables $\mathbf{x}^1, \cdots, \mathbf{x}^p$) in $\mathbb{R}^n$ or as n points (the individuals $\mathbf{x}_1, \cdots, \mathbf{x}_n$) in $\mathbb{R}^p$. These two viewpoints suggest two related objectives:

- the comparison of the variables. To conduct this study, it is necessary to define $\mathbf{D}$, an $n \times n$ positive symmetric matrix used as an inner product in $\mathbb{R}^n$ allowing the computation of relationships between the p variables.

- the comparison of individuals. In this case, a $p \times p$ positive symmetric matrix $\mathbf{Q}$ used as an inner product in $\mathbb{R}^p$ allows the quantification of the resemblances (distances) between the n individuals.

Multivariate analysis considers both questions simultaneously and leads to the definition of the triplet $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ represented in the following diagram:



In this diagram, we can see that four spaces are associated to the data table:

- the individual space ($\mathbb{R}^p$) which contains the n individuals;
- the variable space ($\mathbb{R}^n$) which contains the p variables;
- the variable coefficient space ($\mathbb{R}^{p*}$): an element $\mathbf{g} = [g_1, \cdots, g_p]^\top$ of $\mathbb{R}^{p*}$ is used to define a new synthetic variable $\sum_{i=1}^{p} g_i \mathbf{x}^i$. It is the space of linear functions on $\mathbb{R}^p$ and may be considerd as the dual space of $\mathbb{R}^p$.
- the individual coefficient space ($\mathbb{R}^{n*}$): an element $\mathbf{f} = [f_1, \cdots, f_n]^\top$ of $\mathbb{R}^{n*}$ is used to define a new individual $\sum_{i=1}^{n} f_i \mathbf{x}_i$. It is the dual space of $\mathbb{R}^n$.

If a researcher is mainly interested in the first objective (relationships between variables), the analysis of the diagram in $\mathbb{R}^n$ consists of the eigendecomposition of $\mathbf{XQX}^\top\mathbf{D}$:

$$\mathbf{XQX}^\top\mathbf{DA} = \mathbf{A}\boldsymbol{\Lambda}_r \text{ and } \mathbf{A}^\top\mathbf{DA} = \mathbf{I}_r$$

The r nonzero eigenvalues $\lambda_1 > \lambda_2 > \cdots > \lambda_r > 0$ are stored in the diagonal matrix $\boldsymbol{\Lambda}_r$, and $\mathbf{A} = [\mathbf{a}^1, \cdots, \mathbf{a}^r]$ is an $n \times r$ matrix containing the associated eigenvectors (in columns). These eigenvectors are typically known as the *principal components* onto which the columns of $\mathbf{X}$ are projected to obtain scores for the variables ($\mathbf{C} = \mathbf{X}^\top\mathbf{DA}$).

In contrast, if the study aims to compare individuals, the analysis of the diagram in $\mathbb{R}^p$ consists of the eigendecomposition of $\mathbf{X}^\top\mathbf{DXQ}$. Left-multiplying the previous equation by $\mathbf{X}^\top\mathbf{D}$ leads to the following:

$$(\mathbf{X}^\top\mathbf{D})\mathbf{XQX}^\top\mathbf{DA} = (\mathbf{X}^\top\mathbf{D})\mathbf{A}\boldsymbol{\Lambda}_r$$

If $\mathbf{B} = \mathbf{X}^\top\mathbf{DA}\boldsymbol{\Lambda}_r^{-1/2}$, we obtain:

$$\mathbf{X}^\top\mathbf{DXQB} = \mathbf{B}\boldsymbol{\Lambda}_r \text{ and } \mathbf{B}^\top\mathbf{QB} = \mathbf{I}_r$$

The $p \times r$ matrix $\mathbf{B} = [\mathbf{b}^1, \cdots, \mathbf{b}^r]$ contains eigenvectors (in the columns) that are usually known as the *principal axes*. The rows of $\mathbf{X}$ are then projected onto the principal axes to produce scores for the individuals ($\mathbf{L} = \mathbf{XQB}$).

From this diagram, we can define two other operators, $\mathbf{QX}^\top\mathbf{DX}$ and $\mathbf{DXQX}^\top$, that can be diagonalized in $\mathbb{R}^{p*}$ and $\mathbb{R}^{n*}$, respectively. These decompositions produce the same eigenvalues, and the associated eigenvectors are the *principal factors* ($\mathbf{G}$) and the *principal cofactors* ($\mathbf{F}$), respectively. There are several close relationships between the four eigendecompositions; therefore only one system of axes is required to compute the three others. For instance, we have the following transition formulas:

$$\mathbf{G} = \mathbf{QB},\ \mathbf{A} = \mathbf{XG}\boldsymbol{\Lambda}_r^{-(1/2)},\ \mathbf{F} = \mathbf{DA} \text{ and } \mathbf{B} = \mathbf{X}^\top\mathbf{F}\boldsymbol{\Lambda}_r^{-(1/2)}$$

Using these transition formulas, the product $\mathbf{A}\boldsymbol{\Lambda}_r^{1/2}\mathbf{B}^\top$ can be rewritten as:

$$\mathbf{A}\boldsymbol{\Lambda}_r^{1/2}\mathbf{B}^\top = \mathbf{AA}^\top\mathbf{DX}$$

Left-multiplication by $\mathbf{A}^\top\mathbf{D}$ leads to:

$$\begin{aligned}\mathbf{A}^\top\mathbf{DA}\boldsymbol{\Lambda}_r^{1/2}\mathbf{B}^\top &= \mathbf{A}^\top\mathbf{DX}\\ \mathbf{A}\boldsymbol{\Lambda}_r^{1/2}\mathbf{B}^\top &= \mathbf{X}\end{aligned}$$

The diagonalization of a duality diagram is thus similar to the generalized singular value decomposition of $\mathbf{X}$ (Eckart and Young, 1936). The singular values are contained in the diagonal matrix $\boldsymbol{\Lambda}_r^{1/2}$ and the singular vectors stored in the matrices $\mathbf{A}$ and $\mathbf{B}$ are orthonormalized with respect to $\mathbf{D}$ and $\mathbf{Q}$ respectively ($\mathbf{A}^\top\mathbf{DA} = \mathbf{B}^\top\mathbf{QB} = \mathbf{I}_r$).

Properties

There are several properties linked to the diagonalization of a duality diagram:

- The vectors $\mathbf{a}^1, \mathbf{a}^2, \ldots, \mathbf{a}^r$ successively maximize, under the $\mathbf{D}$-orthogonality constraint, the quadratic form $\| \mathbf{X}^\top\mathbf{Da} \|^2_{\mathbf{Q}}$.
- The vectors $\mathbf{b}^1, \mathbf{b}^2, \ldots, \mathbf{b}^r$ successively maximize, under the $\mathbf{Q}$-orthogonality constraint, the quadratic form $\| \mathbf{XQb} \|^2_{\mathbf{D}}$.

- The vectors $\mathbf{f}^1, \mathbf{f}^2, \ldots, \mathbf{f}^r$ successively maximize, under the $\mathbf{D}^{-1}$-orthogonality constraint, the quadratic form $\| \mathbf{X}^\top \mathbf{f} \|^2_{\mathbf{Q}}$.
- The vectors $\mathbf{g}^1, \mathbf{g}^2, \ldots, \mathbf{g}^r$ successively maximize, under the $\mathbf{Q}^{-1}$-orthogonality constraint, the quadratic form $\| \mathbf{X}\mathbf{g} \|^2_{\mathbf{D}}$.
- If we search for a pair of vectors $\mathbf{b}$ (a $\mathbf{Q}$-normalized vector of $\mathbb{R}^p$) and $\mathbf{a}$ (a $\mathbf{D}$-normalized vector of $\mathbb{R}^n$) that maximize the inner product $\langle \mathbf{XQb} | \mathbf{a} \rangle_{\mathbf{D}} = \langle \mathbf{X}^\top \mathbf{Da} | \mathbf{b} \rangle_{\mathbf{Q}}$, the solution is unique. It is obtained for $\mathbf{b} = \mathbf{b}^1$ and $\mathbf{a} = \mathbf{a}^1$, and the maximum is equal to $\sqrt{\lambda_1}$. Under the orthogonality constraint, these results can be extended for the other pairs.

If $\mathbf{D}$ is diagonal, we can compute the total inertia for the cloud of row vectors (in $\mathbb{R}^p$) as follows:

$$\text{inertia}(\mathbf{X}, \mathbf{Q}, \mathbf{D}) = \sum_{i=1}^{n} d_{ii} \| \mathbf{x}_i \|^2_{\mathbf{Q}} = \text{trace}(\mathbf{XQX}^\top \mathbf{D}) = \sum_{i=1}^{r} \lambda_i$$

where $\mathbf{d}_{ij}$ is the element in the i-th row and j-th column of $\mathbf{D}$, and $\mathbf{x}_i$ is the i-th row of the matrix $\mathbf{X}$. The rows of $\mathbf{X}$ can be projected onto a $\mathbf{Q}$-normalized vector $\mathbf{b}$, and the projected inertia is then equal to:

$$\text{inertia}(\mathbf{b}) = \mathbf{b}^\top \mathbf{QX}^\top \mathbf{DXQb} = \| \mathbf{XQb} \|^2_{\mathbf{D}}$$

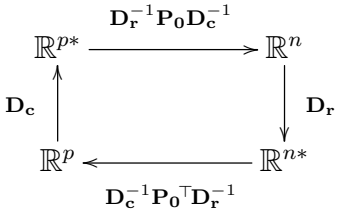
Hence, it appears that the diagonalization of the diagram consists of identifying a set of $\mathbf{Q}$-normalized vectors (the principal axes) that maximize the projected inertia. The inertia projected onto the principal axis $\mathbf{b}^k$ is equal to λ_k.

18.2 Playing with correspondence analysis

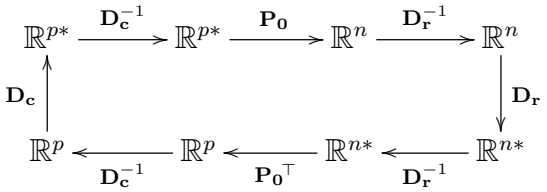
The duality diagram is very general, which enables each analysis to be defined as a particular choice for matrices $\mathbf{X}$, $\mathbf{Q}$ and $\mathbf{D}$. To illustrate its use, I consider the case of the correspondence analysis of an $n \times p$ contingency table $\mathbf{N} = [n_{ij}]$, where n_{ij} is the count for the i-th row and j-column. From the correspondence matrix $\mathbf{P} = \mathbf{N}/n_{++}$ (where n_{++} is the grand total of the contingency table), two vectors $\mathbf{r} = \mathbf{P1}_p$ $(n \times 1)$ and $\mathbf{c} = \mathbf{P}^\top \mathbf{1}_n$ $(p \times 1)$ of row and column masses are derived. The diagonal matrices of the row and column weights are:

$$\mathbf{D}_{\mathbf{r}} = diag(\mathbf{r}) \text{ and } \mathbf{D}_{\mathbf{c}} = diag(\mathbf{c})$$

Lastly, the matrix $\mathbf{P}$ is doubly centered, such that $\mathbf{P_0} = \mathbf{P} - \mathbf{D_r}\mathbf{1}_n\mathbf{1}_p{}^\top\mathbf{D_c}$. Correspondence analysis is the analysis of the triplet $(\mathbf{D_r}^{-1}\mathbf{P_0}\mathbf{D_c}^{-1}, \mathbf{D_c}, \mathbf{D_r})$, and the associated diagram is:



This diagram is equivalent to:



Applying the general formulas of the duality diagram to the CA triplet allows the definition of several properties. This analysis searches for a principal axis $\mathbf{b}$ maximizing

$$\| \mathbf{D_r}^{-1}\mathbf{P_0}\mathbf{D_c}^{-1}\mathbf{D_c}\mathbf{b} \|^2_{\mathbf{D_r}} = \| \mathbf{D_r}^{-1}\mathbf{P_0}\mathbf{b} \|^2_{\mathbf{D_r}}$$

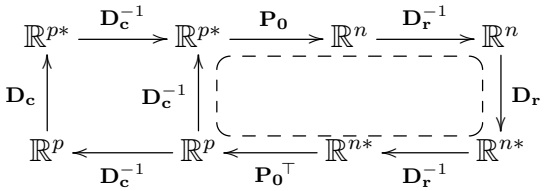
The matrix $\mathbf{D_r}^{-1}\mathbf{P_0}$ contains the centered row profiles such that the product $\mathbf{D_r}^{-1}\mathbf{P_0}\mathbf{b}$ places rows at the barycenters (weighted averages) of the column points, and thus the quantity maximized is simply a variance between rows. Hence, in $\mathbb{R}^p$, columns have a unit-variance score $\mathbf{b}$ that maximizes the variance between the row barycenters. In $\mathbb{R}^n$, CA searches for a principal axis $\mathbf{a}$ maximizing

$$\| \mathbf{D_c}^{-1}\mathbf{P_0}^\top\mathbf{D_r}^{-1}\mathbf{D_r}\mathbf{a} \|^2_{\mathbf{D_c}} = \| \mathbf{D_c}^{-1}\mathbf{P_0}^\top\mathbf{a} \|^2_{\mathbf{D_c}}$$

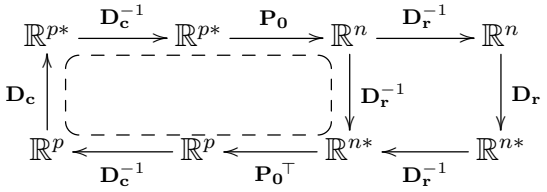
By symmetry, the matrix $\mathbf{D_c}^{-1}\mathbf{P_0}^\top$ contains the centered column profiles such that the product $\mathbf{D_c}^{-1}\mathbf{P_0}^\top\mathbf{a}$ places columns at the barycenters (weighted averages) of the row points ($\mathbf{a}$). Hence, the rows are placed by a unit-variance score $\mathbf{a}$ that maximizes the variance between column barycenters ($\| \mathbf{D_c}^{-1}\mathbf{P_0}^\top\mathbf{a} \|^2_{\mathbf{D_c}}$). It can be demonstrated (see e.g., Greenacre, 1984, p. 92) that replacing $\mathbf{P_0}$ with $\mathbf{P}$ produces the same results, except that one trivial dimension with an eigenvalue equal to one is produced.

These two viewpoints show that CA treats the rows and columns of the contingency table simultaneously and in a symmetric manner. Hence, analyzing $\mathbf{N}$ or $\mathbf{N}^\top$ produces the same results. Manipulating the original duality

diagram allows the highlighting of two different geometrical interpretations for rows and columns. We can rewrite the CA diagram and thus obtain a new statistical triplet (represented by the dashed rectangle):

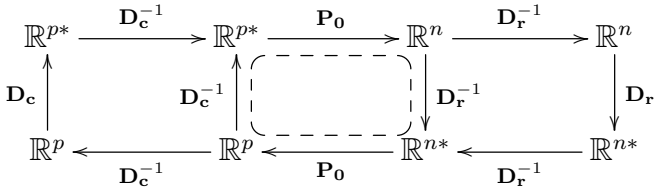


Hence, CA also corresponds to the triplet $(\mathbf{D_r^{-1}P_0}, \mathbf{D_c^{-1}}, \mathbf{D_r})$. In this case, the analysis considers the centered row profiles $(\mathbf{D_r^{-1}P_0})$, with weights $(\mathbf{D_r})$ and χ^2 metrics $(\mathbf{D_c^{-1}})$. Note that in $\mathbb{R}^n$, the analysis of the row profiles returns exactly the same principal components as the analysis of the original diagram. CA can also be rewritten as follows:

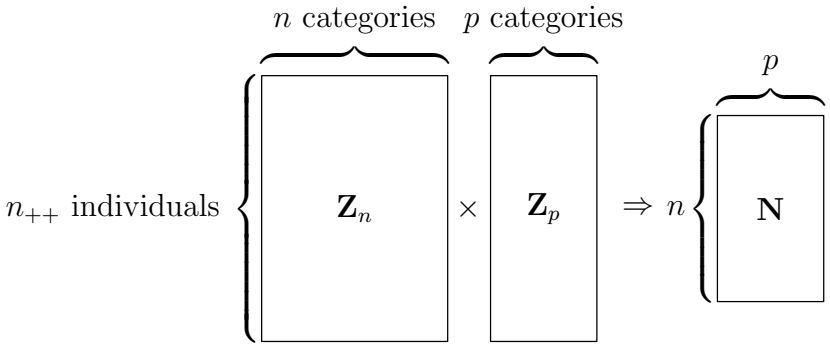


In this case, CA corresponds to the analysis of the centered column profiles $(\mathbf{D_c^{-1}P_0^\top})$, with weights $(\mathbf{D_c})$ and χ^2 metrics $(\mathbf{D_r^{-1}})$. Note that, in $\mathbb{R}^p$, the analysis of the column profiles returns exactly the same principal axes as the analysis of the original diagram. Hence, manipulating the original CA diagram shows that it corresponds to two analyses of two sets of points (row or column) with two different metrics and weighting matrices. These two viewpoints correspond to two discriminant analyses that identify linear combinations of columns (or rows) that maximize the separation of the rows (or columns).

Lastly, a third viewpoint can be identified that corresponds to the triplet $(\mathbf{P_0}, \mathbf{D_c^{-1}}, \mathbf{D_r^{-1}})$:



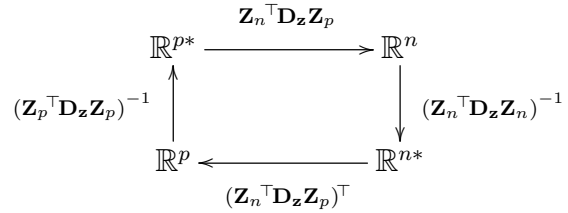
To simplify the presentation, I will consider the triplet $(\mathbf{P}, \mathbf{D_c}^{-1}, \mathbf{D_r}^{-1})$, which is equivalent but produces an additional trivial dimension associated with the eigenvalue $\lambda_1 = 1$ and eigenvectors $\mathbf{a}_1 = \mathbf{1}_n$, $\mathbf{b}_1 = \mathbf{1}_p$. The contingency table $\mathbf{N}$ is the result of the crossing of two qualitative variables. The information is encoded as dummy variables indicating to which categories of the two variables each individual belongs:



Let $\mathbf{D_z} = diag(1/n_{++}, \cdots, 1/n_{++})$ be a weighting matrix for the n_{++} individuals. We have the following relationships:

$$\mathbf{P} = {\mathbf{Z}_n}^\top \mathbf{D_z} \mathbf{Z}_p,\ \mathbf{D_r} = {\mathbf{Z}_n}^\top \mathbf{D_z} \mathbf{Z}_n \text{ and } \mathbf{D_c} = {\mathbf{Z}_p}^\top \mathbf{D_z} \mathbf{Z}_p$$

Using these relationships, we can then rewrite the analysis of $(\mathbf{P}, \mathbf{D_c}^{-1}, \mathbf{D_r}^{-1})$ as the analysis of $\left({\mathbf{Z}_n}^\top \mathbf{D_z} \mathbf{Z}_p, ({\mathbf{Z}_p}^\top \mathbf{D_z} \mathbf{Z}_p)^{-1}, ({\mathbf{Z}_n}^\top \mathbf{D_z} \mathbf{Z}_n)^{-1}\right)$.



In this analysis, the orthogonality constraint on the principal cofactor $\mathbf{f}$ leads to:

$$\| \mathbf{f} \|^2_{((\mathbf{Z}_p{}^\top \mathbf{D_z} \mathbf{Z}_p)^{-1})^{-1}} = \mathbf{f}^\top \mathbf{Z}_p{}^\top \mathbf{D_z} \mathbf{Z}_p \mathbf{f} = \| \mathbf{Z}_p \mathbf{f} \|^2_{\mathbf{D_z}} = 1$$

Hence, the principal cofactors can be viewed as coefficients for the p dummy variables, allowing the computation of a linear combination of unit variance $\mathbf{Z}_p\mathbf{f}$. In contrast, we obtain a linear combination of the n dummy variables ($\mathbf{Z}_n\mathbf{g}$) due to the orthogonality constraint on the principal factor $\mathbf{g}$:

$$\| \mathbf{g} \|^2_{((\mathbf{Z}_n{}^\top \mathbf{D_z} \mathbf{Z}_n)^{-1})^{-1}} = \mathbf{g}^\top \mathbf{Z}_n{}^\top \mathbf{D_z} \mathbf{Z}_n \mathbf{g} = \| \mathbf{Z}_n \mathbf{g} \|^2_{\mathbf{D_z}} = 1$$

Using the transition formulas, the inner product maximized by the analysis can be rewritten as:

$$\begin{aligned} \left\langle \mathbf{Z_p}{}^\top \mathbf{D_z} \mathbf{Z_n} (\mathbf{Z}_n{}^\top \mathbf{D_z} \mathbf{Z}_n)^{-1} \mathbf{a} | \mathbf{b} \right\rangle_{(\mathbf{Z}_n{}^\top \mathbf{D_z} \mathbf{Z}_n)^{-1}} &= \mathbf{b}^\top (\mathbf{Z}_n{}^\top \mathbf{D_z} \mathbf{Z}_n)^{-1\top} \mathbf{Z_p}{}^\top \mathbf{D_z} \mathbf{Z_n} (\mathbf{Z}_n{}^\top \mathbf{D_z} \mathbf{Z}_n)^{-1} \mathbf{a} \\ &= \mathbf{g}^\top \mathbf{Z_p}{}^\top \mathbf{D_z} \mathbf{Z_n} \mathbf{f} = cor(\mathbf{Z}_p \mathbf{g}, \mathbf{Z}_n \mathbf{f}) \end{aligned}$$

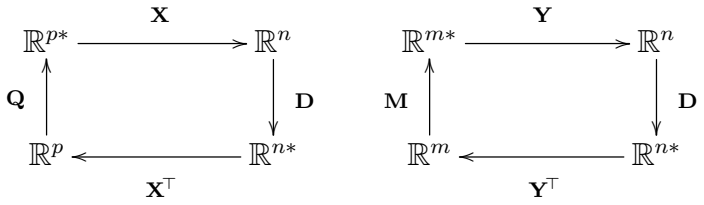
Hence, it appears that CA is a particular case of canonical correlation analysis (Hotelling, 1936) that searches for a linear combination of rows ($\mathbf{Z}_n\mathbf{f}$) and a linear combination of columns ($\mathbf{Z}_p\mathbf{g}$) of maximal correlation.

The duality diagram appears to be a powerful and unifying tool to easily describe the various properties of an analysis. This tool provides a mathematical framework that facilitates the development and comparison of multivariate methods. In this study, we identify four diagrams associated with CA that were completely described by Cazes, Chessel and Dolédec (1988). The canonical correlation viewpoint is explicit in Williams (1952) and is used by Thioulouse and Chessel (1992) and Gimaret-Carpentier, Dray and Pascal (2003) in an ecological context. The discriminant analysis viewpoint is used in ecology by Hill (1973, 1974) and extended by ter Braak (1987) to introduce a table of explanatory variables in canonical correspondence analysis.

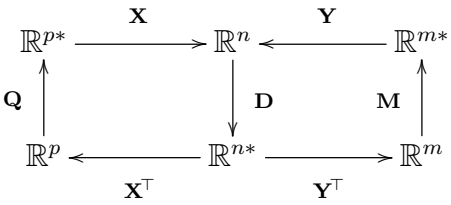
18.3 Relating two diagrams

In many situations, two sets of variables are measured on the same set of n individuals. This information is stored in two matrices, $\mathbf{X}$ ($n \times p$) and $\mathbf{Y}$ ($n \times m$). Each set of variables can be treated by a multivariate analysis

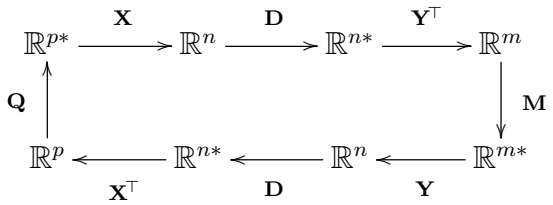
defining two statistical triplets $(\mathbf{X}, \mathbf{Q}, \mathbf{D})$ and $(\mathbf{Y}, \mathbf{M}, \mathbf{D})$:



Note that $\mathbf{D}$ is common to the two triplets because we consider the same individuals in the two analyses. The individuals can be represented as a cloud of n points in $\mathbb{R}^p$ (rows of $\mathbf{X}$) or as n points in $\mathbb{R}^m$ (rows of $\mathbf{Y}$). Although separate analysis allows the independent study of the structures in each table, a relevant question is the evaluation of the concordance between these two configurations of individuals. To achieve this goal, the two duality diagrams must be combined into a single analysis to identify which structures are common to both data sets (i.e., co-structures):



The above diagram can be rewritten as follows:



Coinertia analysis (Dolédec and Chessel, 1994) is the analysis of this diagram and thus is defined by the triplet $\left(\mathbf{Y}^\top\mathbf{DX}, \mathbf{Q}, \mathbf{M}\right)$. The total inertia associated with this triplet is equal to

$$\text{inertia}(\mathbf{Y}^\top\mathbf{DX}, \mathbf{Q}, \mathbf{M}) = \text{trace}(\mathbf{Y}^\top\mathbf{DXQX}^\top\mathbf{DYM})$$

This quantity is a measure of the concordance between the two data sets and is equal to the numerator of the RV coefficient (Escoufier, 1973, - see chapter by Pagès in this book), a multivariate generalization of the

squared correlation coefficient. Coinertia analysis decomposes this vectorial covariance onto orthogonal axes, and the general properties of the diagram lead to the maximization of the following inner product:

$$\left\langle \mathbf{Y}^\top\mathbf{DXQb}|\mathbf{a}\right\rangle_{\mathbf{M}} = \mathbf{a}^\top\mathbf{MY}^\top\mathbf{DXQb} = \left\langle \mathbf{XQb}|\mathbf{YMa}\right\rangle_{\mathbf{D}} = \sqrt{\lambda}$$

If $\mathbf{X}$ an $\mathbf{Y}$ contain centered variables, the total inertia is simply a sum of squared covariances between all combinations of variables of the two data sets ($\sum_{i=1}^{p}\sum_{j=1}^{m} cov^2(\mathbf{x}^i, \mathbf{y}^j)$). In this case, coinertia analysis finds two vectors of coefficients $\mathbf{b}$ and $\mathbf{a}$ to obtain linear combinations of the variable of $\mathbf{X}$ and $\mathbf{Y}$ of maximal covariance ($\left\langle \mathbf{XQb}|\mathbf{YMa}\right\rangle_{\mathbf{D}} = cov(\mathbf{XQb}, \mathbf{YMa})$). This covariance can be decomposed as a product of three factors:

$$cov(\mathbf{XQb}, \mathbf{YMa}) = cor(\mathbf{XQb}, \mathbf{YMa}) \cdot \parallel \mathbf{XQb} \parallel_{\mathbf{D}} \cdot \parallel \mathbf{YMa} \parallel_{\mathbf{D}}$$

The first term ($cor(\mathbf{XQb}, \mathbf{YMa})$) is optimized by canonical correlation analysis. The second ($\parallel \mathbf{XQb} \parallel_{\mathbf{D}}$) is maximized by the analysis of $\mathbf{X}$ that aims to identify the main structures in this data set. The last term ($\parallel \mathbf{YMa} \parallel_{\mathbf{D}}$) corresponds to the simple analysis of $\mathbf{Y}$. Hence, coinertia analysis can be viewed as a compromise between the three analyses aiming to find linear combinations of the two data sets with maximal co-structure. Unlike canonical correlation analysis, which requires many more individuals than variables, coinertia analysis is based on covariances and thus allows us to deal with tables in which the number of individuals is less than the number of variables. In this context, it shares certain similarities with the partial least squares methods (Burnham et al., 1996; Krishnan et al., 2010).

The duality diagram of coinertia analysis is very general and encompasses several existing methods as particular cases (Chessel and Mercier, 1993). If $\mathbf{X}$ and $\mathbf{Y}$ are analyzed by a normed principal component analysis, coinertia analysis corresponds to Tucker's (1958) inter-battery analysis. It is also similar to Procrustes rotation (Dray, Chessel and Thioulouse, 2003b; Gower, 1971) and two-block partial least-squares (Rohlf and Corti, 2000). If $\mathbf{M} = (\mathbf{Y}^\top\mathbf{DY})^{-1}$ and $\mathbf{Q} = (\mathbf{X}^\top\mathbf{DX})^{-1}$, coinertia analysis is equivalent to canonical correlation analysis. If only $\mathbf{Q} = (\mathbf{X}^\top\mathbf{DX})^{-1}$, it corresponds to principal component analysis with instrumental variables (Rao, 1964), also known as redundancy analysis (van den Wollenberg, 1977), which aims to study the variation in $\mathbf{Y}$ explained by $\mathbf{X}$. When $\mathbf{Y}$ is a contingency table analyzed by correspondence analysis and $\mathbf{Q} = (\mathbf{X}^\top\mathbf{DX})^{-1}$, it is similar to canonical correspondence analysis (ter Braak, 1987). If both $\mathbf{X}$ and $\mathbf{Y}$ contain qualitative variables and are analyzed by multiple correspondence analysis, coinertia analysis corresponds to the correspondence analysis of the Burt matrix. The only difference is that coinertia analysis preserved the

Dray, S. 2014. Analysing a pair of tables: coinertia analysis and duality diagrams. Pages 289–300 in J. Blasius and M. Greenacre, editors. Visualization and verbalization of data. CRC Press.

original structure of the data whereas the correspondence analysis viewpoint does not consider the individuals (i.e. the rows of $\mathbf{X}$ and $\mathbf{Y}$). It is thus linked to the works of Leclerc (1975) and Benzécri (1982) if a table contains only one categorical variable, whereas Cazes (1980) provides insights into positioning the rows as supplementary points. Lastly, we consider a situation that is often encountered in community ecology: the abundance of q species (the columns of $\mathbf{Y}$) is sampled in n sites (rows) for which qualitative environmental variables (columns of $\mathbf{X}$) are measured. A common practice is to construct a table of ecological profiles summarizing the distribution of species in the different environmental classes (e.g., Sabatier et al., 1997). Correspondence analysis of the ecological profiles table (Bonin and Roux, 1978; Romane, 1972) is equivalent to coinertia analysis in which $\mathbf{Y}$ is analyzed by a correspondence analysis and $\mathbf{X}$ is treated by multiple correspondence analysis (Mercier, Chessel and Dolédec, 1992).

Coinertia analysis is based on a very general principle that has been extended to several situations (Dray, Chessel and Thioulouse, 2003a) to analyze a series of tables (Chessel and Hanafi, 1996) or to link external information on both rows and columns of a contingency table (Dolédec et al., 1996). The presentation based on the duality diagram allows to summarize the various properties of a method and thus to simplify the comparison among methods. These abilities would be helpful in identifying concordances between methods that have been developed in different fields with few connections but similar methodological questions. For instance, I recently discovered the SVD method described by Bretherton, Smith and Wallace (1992) in a review of two-tables methods in climatology. This method is similar to a coinertia analysis between two centered principal component analyses. The use of a common mathematical language, as provided by the duality diagram theory, would probably improve exchanges between statisticians working in psychometrics, chemometrics, ecology, climatology and other fields.

Dray, S. 2014. Analysing a pair of tables: coinertia analysis and duality diagrams. Pages 289–300 in J. Blasius and M. Greenacre, editors. Visualization and verbalization of data. CRC Press.

References

Benzécri, J. (1982). Sur la généralisation du tableau de burt et son analyse par bandes. *Les Cahiers de l'Analyse des Données*, *7*, 33–43.

Bonin, G., and Roux, M. (1978). Utilisation de l'analyse factorielle des correspondances dans l'étude phyto-écologique de quelques pelouses de l'Apennin lucano-calabrais. *Oecologia Plantarum*, *13*, 121–138.

Bretherton, C., Smith, C., and Wallace, J. (1992). An intercomparison of methods for finding coupled patterns in climate data. *Journal of climate*, *5*(6), 541–560.

Burnham, A. J., Viveros, R., and MacGregor, J. F. (1996). Frameworks for latent variable multivariate regression. *Journal of Chemometrics*, *10*(1), 31–45.

Cailliez, F., and Pagès, J. P. (1976). *Introduction à l'analyse des données*. Paris: SMASH.

Cazes, P. (1970). *Application de l'analyse des données au traitement de problèmes géologiques*. Thèse de 3ème cycle, Faculté des Sciences de Paris.

Cazes, P. (1980). L'analyse de certains tableaux rectangulaires décomposé en blocs : généralisation des propriétés rencontrées dans l'étude des correspondances multiples. I. Définitions et applications à l'analyse canonique des variables qualitatives. *Les Cahiers de l'Analyse des Données*, *5*, 145–161.

Cazes, P., Chessel, D., and Dolédec, S. (1988). L'analyse des correspondances internes d'un tableau partitionné: son usage en hydrobiologie. *Revue de Statistique Appliquée*, *36*, 39–54.

Chessel, D., Dufour, A.-B., and Thioulouse, J. (2004). The ade4 package - I: One-table methods. *R News*, *4*, 5–10.

Chessel, D., and Hanafi, M. (1996). Analyse de la co-inertie de {K} nuages de points. *Revue de Statistique Appliquée*, *44*(2), 35–60.

Chessel, D., and Mercier, P. (1993). Couplage de triplets statistiques et liaisons espèces-environnement. In J. D. Lebreton, and B. Asselain (Eds.) *Biométrie et Environnement*, (pp. 15–43). Paris: Masson.

De la Cruz, O., and Holmes, S. (2011). The duality diagram in data analysis: examples of modern applications. *The Annals of Applied Statistics*, *5*(4), 2266–2277.

Dolédec, S., and Chessel, D. (1994). Co-inertia analysis: an alternative method for studying species-environment relationships. *Freshwater Biology*, *31*, 277–294.

Dolédec, S., Chessel, D., ter Braak, C. J. F., and Champely, S. (1996). Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics*, *3*, 143–166.

Dray, S., Chessel, D., and Thioulouse, J. (2003a). Co-inertia analysis and the linking of ecological data tables. *Ecology*, *84*, 3078–3089.

Dray, S., Chessel, D., and Thioulouse, J. (2003b). Procrustean co-inertia analysis for the linking of multivariate data sets. *Ecoscience*, *10*(1), 110–119.

Dray, S., and Dufour, A. B. (2007). The ade4 package: implementing the duality diagram for ecologists. *Journal of Statistical Software*, *22*(4), 1–20.

Dray, S., Dufour, A. B., and Chessel, D. (2007). The ade4 package - II: Two-table and K-table methods. *R News*, *7*(2), 47–52.

Dray, S., and Jombart, T. (2011). Revisiting Guerry's data: introducing spatial constraints in multivariate analysis. *The Annals of Applied Statistics*, *5*(4), 2278–2299.

Eckart, C., and Young, G. (1936). The approximation of one matrix by another of lower rank. *Psychometrika*, *1*(3), 211–218.

Escoufier, Y. (1973). Le traitement des variables vectorielles. *Biometrics*, *29*, 750–760.

Escoufier, Y. (1987). The duality diagram : a means of better practical applications. In P. Legendre, and L. Legendre (Eds.) *Developments in numerical ecology*, vol. 14, (pp. 139–156). Berlin: Springer Verlag.

Gimaret-Carpentier, C., Dray, S., and Pascal, J.-P. (2003). Broad-scale biodiversity pattern of the endemic tree flora of the Western Ghats (India) using canonical correlation analysis of herbarium records. *Ecography*, *26*, 429–444.

Goscinny, R., and Uderzo, A. (1989). *How Obelix fell into the magic potion when he was a little boy*. London: Hodder and Stoughton.

Gower, J. C. (1971). Statistical methods of comparing different multivariate analyses of the same data. In F. R. Hodson, D. G. Kendall, and P. Tautu (Eds.) *Mathematics in the archaeological and historical sciences*, (pp. 138–149). Edinburgh: Edinburgh University Press.

Greenacre, M. J. (1984). *Theory and applications of correspondence analysis*. London: Academic Press.

Hill, M. O. (1973). Reciprocal averaging : an eigenvector method of ordination. *Journal of Ecology*, *61*, 237–249.

Hill, M. O. (1974). Correspondence analysis : a neglected multivariate method. *Applied Statistics - Journal of the Royal Statistical Society Series C*, *23*, 340–354.

Holmes, S. (2006). Multivariate analysis: the French way. In D. Nolan, and T. Speed (Eds.) *Festschrift for David Freedman*. Beachwood, OH: IMS.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, *28*, 321–377.

Krishnan, A., Williams, L. J., McIntosh, A. R., and Abdi, H. (2010). Partial least squares (PLS) methods for neuroimaging: A tutorial and review. *NeuroImage*, *56*(2), 455–475.

Leclerc, A. (1975). L'analyse des correspondances sur juxtaposition de tableaux de contingence. *Revue de Statistique Appliquée*, *23*(3), 5–16.

Mercier, P., Chessel, D., and Dolédec, S. (1992). Complete correspondence analysis of an ecological profile data table: a central ordination method. *Acta Oecologica - International Journal of Ecology*, *13*(1), 25–44.

Purdom, E. (2011). Analysis of a data matrix and a graph: metagenomic data and the phylogenetic tree. *The Annals of Applied Statistics*, *5*(4), 2326–2358.

Ramsay, J., and de Leeuw, J. (1983). Review. *Psychometrika*, *48*(1), 147–151.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. *Sankhya A*, *26*, 329–359.

Rohlf, F. J., and Corti, M. (2000). Use of two-block partial least-squares to study covariation in shape. *Systematic Biology*, *49*(4), 740–753.

Romane, F. (1972). Utilisation de l'analyse multivariable en phytoécologie. *Investigacion Pesquera*, *36*, 131–139.

Sabatier, D., Grimaldi, M., Prévost, M. F., Guillaume, J., Godron, M., Dosso, M., and Curmi, P. (1997). The influence of soil cover organization on the floristic and structural heterogeneity of a Guianan rain forest. *Plant Ecology*, *131*, 81–108.

Tenenhaus, M., and Young, F. W. (1985). An analysis and synthesis of multiple correspondence analysis, optimal scaling, dual scaling, homogeneity analysis and other methods for quantifying categorical multivariate data. *Psychometrika*, *50*(1), 91–119.

ter Braak, C. J. F. (1987). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, *69*, 69–77.

Thioulouse, J. (2011). Simultaneous analysis of a sequence of paired ecological tables: A comparison of several methods. *The Annals of Applied Statistics*, *5*(4), 2300–2325.

Thioulouse, J., and Chessel, D. (1992). A method for reciprocal scaling of species tolerance and sample diversity. *Ecology*, *73*, 670–680.

Thioulouse, J., Chessel, D., Dolédec, S., and Olivier, J. M. (1997). ADE-4: a multivariate analysis and graphical display software. *Statistics and Computing*, *7*, 75–83.

Tucker, L. R. (1958). An inter-battery method of factor analysis. *Psychometrika*, *23*(2), 111–136.

van den Wollenberg, A. L. (1977). Redundancy analysis, an alternative for canonical analysis. *Psychometrika*, *42*(2), 207–219.

Williams, E. J. (1952). Use of scores for the analysis of association in contingency tables. *Biometrika*, *39*, 274–289.