DMFA: Dual Multiple Factor Analysis

Sébastien Lé, François Husson, and Jérôme Pagès

Agrocampus Rennes
CNRS UMR 6625 Mathématiques appliquées
65 rue de Saint-Brieuc - CS 84215
35042 RENNES Cedex, France
(e-mail: sebastien.le@agrocampus-rennes.fr)

Abstract. In this article, we propose a new method called Dual Multiple Factor Analysis (DMFA), which is an extension of MFA in the case where individuals are structured according to a partition. The heart of the method rests on a factor analysis known as internal, in reference to the internal correspondence analysis, for which data are systematically centered by group. This analysis is an internal PCA when all the variables are quantitative. DMFA provides the classic results of a PCA as well as additional outputs induced by the consideration of a partition on individuals, such as the superimposed representation of the L scatter plots of variables associated with the L groups of individuals and the representation of the scatter plot of the correlations matrices associated each one with a group of individuals.

Keywords: Dual Multiple Factor Analysis, groups of individuals, common structures, geometrical interpretation.

1 Introduction

The analysis of data comprising several sets of individuals described by a same set of variables is a problem frequently encountered. Examples are provided by international surveys where groups of individuals coming from different countries are questioned according to a same set of questions. In this article, we propose a new method called Dual Multiple Factor Analysis (DMFA), which is an extension of Multiple Factor Analysis [Escofier and Pagès, 1988] (MFA) in the case where individuals are structured according to a partition. The guiding principles of DMFA (a factor analysis of the whole set of individuals where variables are centered by group), its specificities as a factor analysis (weighting of the groups of individuals) as well as the types of results it provides (a scatter plot of the partial variables, a representation of the groups of individuals) are presented herein.

2 Data, notations

The data comprise several groups of individuals I described by the same set of variables K. To each group I corresponds a data table X of dimension (I × K). All groups of individuals being described by the same set of variables, all the tables can be column-wised juxtaposed to form only one table X
of dimension \((I \times K)\) crossing individuals and variables: one denotes by \(I\) the union of the \(I_l\). The initial data tables appear as one single table structured in sub-tables. One denotes by \(x_{ik}\) the value taken by the individual \(i\) for the variables \(k\) and \(k^l\), where \(k^l\) indicates the restriction of \(k\) to the individuals of the group \(l\).

Let \(m_k\) be the weight of the variable \(k\) and \(p_i\) the weight of the individual \(i\), in the analysis of the whole data table \(X\); one supposes that \(\sum_{i \in I} p_i = 1\). Let \(q_l = \sum_{i \in I} p_i\) be the sum of the weights of the individuals of \(I^l\) and \(p_i^l\) the weight of the individual \(i \in I^l\) in the separate analysis of the data table \(X_i\); one also supposes that \(\sum_{i \in I^l} p_i^l = 1\), \(\forall l = 1, \ldots, L\).

Within the framework of this article, we consider the case where all the variables are quantitative.

### 3 Problematic

In classic factor analysis problems, one studies the principal variability factors between individuals on the one hand and to draw up an assessment of the correlations between variables on the other hand.

A first suggestion consists in: studying the total individuals variability with a classic principal components analysis (PCA) of the entire data table \(X\); taking into account in the interpretation of each axis the partition on the individuals by splitting up the total variability into between and within groups variability. One denotes by \(N^l_I\) the scatter plot associated with the group of individuals \(I^l\). From a graphical point of view an individual is represented using a distinctive sign, function of the scatter plot \(N^l_I\) with which the individual is associated.

A second suggestion consists in carrying out a discriminant analysis on the data table \(X\) and consequently in studying the between groups variability.

Lastly, considering a partition structure on the individuals rises the problem of the comparison of the within groups variability, \(i.e.\) of the comparison of the scatter plots associated with the \(N^l_I\).

The preceding suggestions result in working on variables defined on the whole data table \(X\) and in analyzing their correlations based on all the individuals. However, these correlations are not necessary the same from one group of individuals to another and arises the problem of studying the correlations induced by the each group. This implies to get interested in the various clouds of variables defined by each group of individuals: one wishes to compare these various scatter plots, to study the evolution of each variable with through the groups of individuals, \(i.e.\) the evolution of its coefficients of correlation with the other variables (from one group of individuals to the other).

This comparison of the correlations between variables would be facilitated if one had a simultaneous representation of the scatter plots of variables defined for each group of individuals.
4 Variables space, representation of the clouds of variables

4.1 Cloud of the variables defined on all the individuals

A variable $k$, measured on all the $I$ individuals, is considered as an element of $R^I$ (each dimension of $R^I$ is associated with one and only one individual). With the space of variables $R^I$ one associates the diagonal metric $D$, where $D$ denotes the matrix whose $i^{th}$ diagonal term is worth $p_i$, i.e. the weight associated with the individual $i$.

Afterward, one denotes by $N_K$ the cloud of the variables defined on all the individuals. $N_K$ can be obtained on the basis of the data table $X$ transformed the following manners:

1. variables are mean-centered, which is always the case in a usual PCA framework. If $\bar{x}_k = \sum_{i \in I} p_i x_{ik}$ denotes the mean of the $k$ variable, then $\forall k \in K, \bar{x}_k = 0$.
2. variables are standardized, as in a PCA performed on a correlation matrix. $N_K$ is then included in a hypersphere of radius 1. If $s^2_k = \sum_{i \in I} p_i (x_{ik} - \bar{x}_k)^2$ denotes the variance of the $k$ variable, then $\forall k \in K, s^2_k = 1$.
3. variables are mean-centered by group. If $\bar{x}_{lk} = \sum_{i \in I} p_{li} x_{ik}$ denotes the mean of the variable $k_l$, then $\forall l = 1, \ldots, L, \forall k \in K, \bar{x}_{lk} = 0$.
4. variables are standardized by group. If $s^2_{lk} = \sum_{i \in I} p_{li} (x_{ik} - \bar{x}_{lk})^2$ denotes the variance of the $k_l$ variable, then $\forall l = 1, \ldots, L, \forall k \in K, s^2_{lk} = 1$.

Thereafter, we will suppose the data centered by group; the variables are generally standardized (according to one of the two transformations considered previously) and belong in this case to the hypersphere of $R^I$ of radius 1.

4.2 Clouds of the variables defined for each group of individuals

One denotes by $N^l_K = \{k^l, k = 1, \ldots, K\}$ the cloud made up of the $k^l$ variables associated with the individuals of the $l$ group. By definition, each cloud of variables $N^l_K$ belongs to a space of dimension $I^l$, denoted $R^{I^l}$. According to a reasoning similar to the one used in MFA, it is possible to represent simultaneously the $L$ scatter plots of variables $N^l_K$ in the space $R^I$ by noticing that $R^I$ can be split up into a direct sum of $L$ subspaces orthogonal two by two and isomorphs with the $R^{I^l}$ spaces respectively:

$$R^I = \bigoplus_{l \in L} R^{I^l}$$

$R^{I^l}$ can be perceived as a subspace of $R^I$ or as a space in itself. Depending on the situation we will consider different metrics. In each subspace $R^{I^l}$ (i.e. when it is considered as a space in itself), one considers the diagonal metric...
D^I\) whose diagonal terms are the \(p_i^l\) defined previously such as \(\sum_{i \in I} p_i^l = 1, \forall l = 1, \ldots, L\). The coordinates of the variables of the cloud \(N_{k}^I\) are contained in the data table \(X_l\). The coordinates of these variables in the space \(R^I\) are contained in the data table \(\tilde{X}_l\) whose elements are worth \(x_{ik}\) if \(i \in I^l\) and 0 if not; one denotes by \(\tilde{k}^l\) the variable of \(K\) on \(I\), associated with the \(k^{th}\) column of table \(\tilde{X}_l\).

In the space \(R^I\), one denotes by \(\tilde{N}_{k}^I = \{\tilde{k}^l, k = 1, \ldots, K\}\) the cloud of the variables \(k^l\) associated with group \(l\). By construction,

\[||k^l||_{D^I}^2 = 1 \Rightarrow ||\tilde{k}^l||_{D}^2 = q_l.\]

Thus, so that the clouds of variables \(N_{k}^I\) and \(\tilde{N}_{k}^I\) are isomorphic (i.e. \(||k^l||_{D^I}^2 = 1\), each vector \(\tilde{k}^l\) must be multiplied by a coefficient \(1/\sqrt{q_l}\); one denotes by \(N_{k}^I\) the cloud of variables thus obtained, on the hypersphere of \(R^I\) of radius 1: thereafter, the same notation will be used to denote the cloud of the columns of \(X_l\) in \((R^I, D^I)\) as well as the cloud of the columns of \(\tilde{X}_l/\sqrt{q_l}\) in \((R^I, D)\).

Geometrically, the study of the correlations between variables amounts to the analysis of the shape of the scatter plot \(N_K\) (correlations based on all the individuals) and to the comparison of the shapes of the scatter plots \(N_{k}^I\) (correlations by group of individuals).

One of the objectives of the method is to obtain a representation of each set of variables \(\{k^l, k = 1, \ldots, K\}\), in other words of each scatter plot \(N_{k}^I\), comparable from one group to another, where \(k^l\) denotes the restriction of the variable \(k\) on the subpopulation \(I^l\) \((k^l \in R^I)\).

A first idea consists in carrying out a PCA by group of individuals \(I^l\). This strategy is “optimal” regarding the representation of each cloud \(N_{k}^I\), but is not directed towards their comparison. Indeed, the clouds \(N_{k}^I\) are then represented in spaces generated by principal components whose interpretation can be different from one group to the other and de facto the relative positions of the variables compared to these components are not easily comparable.

Rather than carrying out separate PCA on the entire data table \(X\) and as in MFA in representing all \(\{k^l, k = 1, \ldots, K, l = 1, \ldots, L\}\) within a single framework. In PCA, principal components are linear combinations of the original variables \(k \in K\), what can be written \(F_s = \sum_{k} u_s(k)k\) (supposing that the weight of each variable is equal to 1), where \(u_s = \{u_s(k), k = 1, \ldots, K\}\) denotes the vector of the coordinates of the variables \(k\) on the principal component of rank \(s\). One denotes by \(F_s^l\) the restriction of \(F_s\) on the subpopulation \(I^l\). By definition, \(F_s^l\) is obtained from the same linear combination of variables than \(F_s\), in other words \(F_s^l = \sum_{k} u_s(k)k^l\). If the variables are centered and standardized by group, their “contribution” in the construction of the partial component \(F_s^l\) is the same from one group to the other. In this case, for a given \(s\), each linear combination \(F_s^l\) has the same interpretation as its associated principal component \(F_s\). This argument justifies the representation of the \(k^l\) in a...
common framework, namely the space generated by the principal components $F_s$.

In a PCA performed on the correlation matrix, the coordinate $G_s(k)$ of the variable $k$ on the axis of rank $s$ is at the same time equal to the correlation coefficient between $k$ and $F_s$ and to the coefficient of the linear combination associated with the variable $k$ in the construction of $F_s$, in other words $G_s(k) = r(k, F_s) = \sqrt{\lambda_s} u_s(k)$. It can be interesting to represent the $k^l$ from one or the other from its two properties. As it is impossible to benefit from the two properties simultaneously, one has to choose one among the two.

Since the coefficients of the linear combination inducing the $F^l_s$ are the same from one group to another their use to represent the $k^l$ is without interest. Hence the idea of representing the $k^l$ using their correlation coefficients with the factors.

Suppose all the individuals have the same weight. By definition, 

$$G^l_i = \frac{1}{\sqrt{\text{Var}(F^l_i)}} X^l_i' D^l F^l_i,$$

where $G^l_i$ denotes the vector of the coordinates of the variables restricted to the individuals $I^l$. Hence, we can show that the coordinate of a variable $k$ restricted to the individuals $I^l$ can also be written 

$$G^l_s(k) = \frac{1}{\sqrt{\lambda_s}} \frac{1}{\sqrt{\text{Var}(F^l_s)}} \sum_{z \in K} r(z^l, k^l) G_s(z).$$

Thus, the coordinate of $k^l$ is a linear combination of coordinates of the variables on the factor of rank $s$ resulting from the total analysis; which is in itself a second justification for a superimposed representation since the coordinates of the $k^l$ are deduced from the same points through the $G_s(z)$. It takes into account, within the group of individuals $I^l$, the correlations between the variable $k$ and all the others, by granting as much importance to a variable $z$ than this one is correlated with the axis of rank $s$. Thus, proximity between two homologous variables (of two different groups of individuals) can be interpreted like a resemblance measurement between the two groups of individuals associated each one with one of both variables.

### 5 Space of the groups of individuals and cloud of correlation matrices

Let $R^{K^2}$ denotes the vectorial space of dimension $K^2$. In this space, each group of individuals $I^l$ is represented by its correlations matrix $C_l = X^l_i' D^l X_l$, where $D^l$ denotes the metric induced by $D$ on each subspace $R^{I^l}$. As a set of $K^2$ scalars, $C_l$ can be considered as a vector of $R^{K^2}$; in that sense, $R^{K^2}$ will be often called the space of the groups of individuals. In this space, $\{C_l, l = 1, \ldots, L\}$ can be considered as a cloud that will be denoted $N_L$. 

On this space $R^{K^2}$ we will use the scalar product introduced by Escoufier [Escoufier, 1973] the following way for two matrices $C_l$ and $C_{l'}$ :

$$< C_l, C_{l'} >_{R^{K^2}} = \text{trace}(C_l M C_{l'} M),$$

where $M$ denotes the metric associated with $R^K$. In order to simplify the notations we will consider that all the variables have the same weight and that $M$ is worth the identity matrix of dimension $K \times K$.

In order to look at the similarities between matrices of correlations, one studies the cloud of the groups of individuals $N_L$ in the space $R^{K^2}$. To do so, one projects this scatter plot on subspaces of $R^{K^2}$. This type of problems was already solved by the Statis method [Lavit, 1988] which projects the scatter plot $N_L$ on its principal inertia directions as in PCA. This method has the advantage of optimizing the quality of representation of the groups of individuals. However, it presents the disadvantage of being based on axes that are impossible to interpret. In DMFA, the representation of the correlation matrices is obtained in a way similar to the representation of the scalar products matrices in $R^{I^2}$ in MFA: one projects the correlations matrices on the axes $U_s = u_s u_s'$ induced by the factor axes $u_s$ of $R^K$ resulting from DMFA.

The projection of the correlation matrix $C_l$ on $U_s$ can be written

$$< U_s, C_l >_{R^{K^2}} = \text{Trace}(u_s u_s' C_l)$$

$$= u_s' C_l u_s$$

$$= \sum_{h,k} u_s(k) u_s(h) r(k^l, h^l)$$

$$< U_s, C_l >_{R^{K^2}} = \frac{1}{\lambda_s} \sum_{h,k} r(F_s, k) r(F_s, h) r(k^l, h^l).$$

This coordinate, which is always positive, is all the more high that, for each couple of variables $k$ and $h$ :

1. their correlation restricted to the subpopulation $I^l$ is high;
2. their correlation on the entire population $I$ is highlighted by the axis of rank $s$.

### 6 Example

We consider herein two groups of individuals with the following correlations structures. For the first group, $V_1$ and $V_2$ (resp. $V_3$ and $V_4$) are closely correlated ; $V_5$, $V_6$ and $V_7$ are almost mutually orthogonal on the one hand and in an orthogonal space to the one spanned by $\{V_1, V_2, V_3, V_4\}$ on the other hand (cf. table 1).

Variables of group 2 are such as $V_5$, $V_6$, $V_7$ (resp. $V_1$ and $V_2$ ; resp. $V_3$ and $V_4$) are closely correlated ; the triplet $\{V_5, V_6, V_7\}$ as well as the
Table 1. Correlations structure between variables of group 1

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1.000</td>
<td>0.980</td>
<td>0.246</td>
<td>-0.049</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V2</td>
<td>1.000</td>
<td>0.049</td>
<td>-0.246</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V3</td>
<td>1.000</td>
<td>0.956</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V4</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V5</td>
<td>1.000</td>
<td>0.050</td>
<td>-0.002</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V6</td>
<td>1.000</td>
<td>-0.050</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V7</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Table 2. Correlations structure between variables of group 2

<table>
<thead>
<tr>
<th></th>
<th>V1</th>
<th>V2</th>
<th>V3</th>
<th>V4</th>
<th>V5</th>
<th>V6</th>
<th>V7</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>1.000</td>
<td>0.969</td>
<td>0.171</td>
<td>-0.171</td>
<td>0.124</td>
<td>0.484</td>
<td>-0.294</td>
</tr>
<tr>
<td>V2</td>
<td>1.000</td>
<td>0.171</td>
<td>-0.171</td>
<td>-0.124</td>
<td>0.253</td>
<td>-0.520</td>
<td>0.000</td>
</tr>
<tr>
<td>V3</td>
<td>1.000</td>
<td>0.941</td>
<td>0.000</td>
<td>0.064</td>
<td>-0.071</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V4</td>
<td>1.000</td>
<td>0.000</td>
<td>-0.064</td>
<td>0.071</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V5</td>
<td>1.000</td>
<td>0.928</td>
<td>0.912</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V6</td>
<td>1.000</td>
<td>0.694</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>V7</td>
<td>1.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
<td>0.000</td>
</tr>
</tbody>
</table>

Fig. 1. Factorial planes (1-2) and (1-3) provided by DMFA on both groups 1 and 2
Fig. 2. Partial representation of the variables; underlined variables correspond to group 1. For the variables of group 1 (resp. 2), the hachured surface (resp. in grey) is proportional to the quality of representation on the planes spanned by \((v_1^1, v_1^2)\) (resp. \((v_2^2, v_2^3)\)) on the one hand, \((v_1^1, v_1^3)\) (resp. \((v_2^2, v_2^3)\)) on the other hand.

couples \{V1, V2\} on the one hand, \{V3, V4\} on the other hand, are almost mutually orthogonal (cf. table 2).

On the principal factorial plane (1-2) provided by \(\text{DMFA}\), one can see the strong positive correlation between variables \(V1\) and \(V2\) on the one hand, \(V3\) and \(V4\) on the other hand, for both groups of individuals (cf. figure 2).

On the factorial plane (1-3), one can see the strong positive correlation between variables \(V5\), \(V6\) and \(V7\), for the only individuals of group 2; for group 1, the same variables are not sufficiently well represented to conclude.

References

