# Larger Variance Accounted for in Distance-Based Multivariate Analysis Compared to Classical Multivariate Analysis

**Patrick J.F. Groenen***
**Jacqueline J. Meulman**

**Data Theory Group**

**Leiden University**

* Department of Education, Data Theory Group, Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands (e-mail: groenen@rulfsw.fsw.LeidenUniv.nl). Supported by the Netherlands Organization for Scientific Research (NWO) by grant nr. 030-56403 for the 'PIONEER' project 'Subject Oriented Multivariate Analysis' to the second author.

**Abstract**

In this paper we compare the proportion of variance accounted for in classical and distance-based multivariate analysis. We show that distance-based multivariate analysis always can be made to yield a higher proportion of variance accounted for than classical multivariate analysis. This property is illustrated for principal components analysis, multiple correspondence analysis, multiple regression, and analysis of variance.

**Keywords:** distance-based multivariate analysis, variance accounted for, principal components analysis, multiple correspondence analysis, multiple regression, analysis of variance.

# 1 Introduction

The usual measure in determining the fit in multivariate analysis is to evaluate the proportion of variance of the variables that is accounted for. For example, in classical principal components analysis (PCA), the proportion of VAF is equal to the average squared correlation between objects and the principal components. An alternative approach to multivariate analysis, called distance-based multivariate analysis (Meulman, 1986, 1992), maximizes the Tucker's congruence coefficient between the distances of pairs of objects in the data space and distances in the representation space. It was unclear until now what the optimal representation space in distance-based PCA implied in terms of the fit of the original variables. Here we show that the proportion of total variance accounted for (VAF) can be increased by minimizing the loss function of distance-based multivariate analysis compared to classical multivariate analysis while estimating the same number of parameters.

It is well known that at a local minimum $x^*$, value of the classical multivariate analysis loss function $L_{\text{clas}}(x^*)$ is equal to one minus the proportion of VAF. We will prove that at a local minimum $x^{**}$, the distance-based multivariate analysis loss function $\sigma^2_{\text{db}}(x^{**})$ is also equal to one minus the proportion of VAF. Moreover, we prove that the proportion of VAF obtained by $\sigma^2_{\text{db}}(x^{**})$ is higher (or the same) as the one obtained by $L_{\text{clas}}(x^*)$ estimating the same number of parameters. These issues are visualized in Figure 1, where the value of loss functions of a classical and distance-based multivariate analysis model is plotted against the parameter to be estimated.

This paper is organized as follows. First, we prove the claims above applied to principal components analysis. It is shown that the VAF with respect to the variables in distance-based multivariate analysis is equivalent

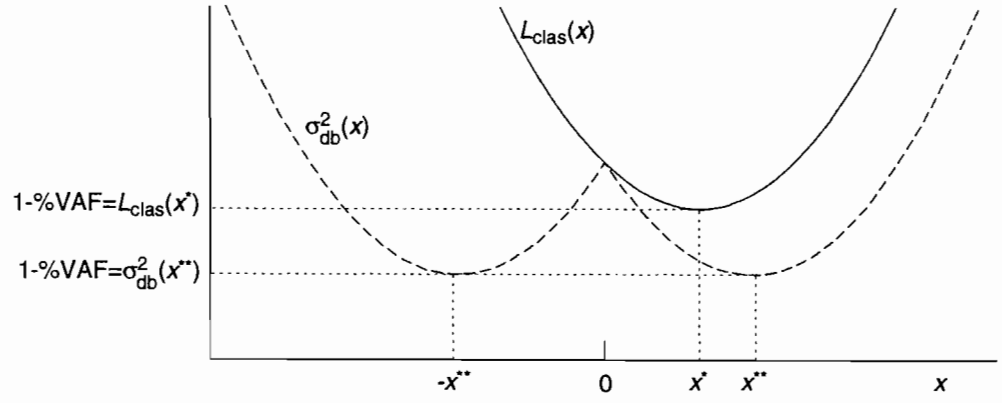


FIGURE 1. Function values of classical multivariate analysis, $L_{\text{clas}}(x)$, and distance-based multivariate analysis, $\sigma^2_{\text{db}}(x)$. $L_{\text{clas}}(x)$ is larger or equal to $\sigma^2_{\text{db}}(x)$ everywhere. At the local minimum $x^*$ of $L_{\text{clas}}(x)$ and the local minimum $x^{**}$ of $\sigma^2_{\text{db}}(x)$ the loss is equal to one minus the proportion of variance accounted for.

to the square of Tucker's coefficient of congruence between two vectors of distances. We discuss the requirement of distance-based models for this to hold. Then, these ideas are applied to multiple correspondence analysis, multiple regression, and analysis of variance. We present some examples and compare the difference in VAF, and end with a discussion and conclusions.

# 2 Principal Components Analysis

Principal components analysis aims at finding a low rank approximation $\widehat{\mathbf{Z}}$ of a data matrix $\mathbf{Z}$ of $n$ objects by $m$ variables. We assume throughout that each column of $\mathbf{Z}$ has a mean of zero and a sum of squares of one. The PCA model can be expressed as

$$L_{\text{PCA}}(\widehat{\mathbf{Z}}) = \|\mathbf{Z} - \widehat{\mathbf{Z}}\|^2, \tag{1}$$

where $\widehat{\mathbf{Z}}$ is of rank $p \leq m$ and $\|\mathbf{Z}\|^2$ denotes the sum of squared elements of $\mathbf{Z}$, i.e., tr $\mathbf{Z}'\mathbf{Z} = \sum_{i=1}^{n}\sum_{j=1}^{m} z_{ij}^2$. Suppose that $\widehat{\mathbf{Z}}^*$ is the least squares estimate that minimizes (1). A well known property of least squares estimates is that the residuals are orthogonal to the estimates, i.e., tr $\widehat{\mathbf{Z}}^{*\prime}(\mathbf{Z} - \widehat{\mathbf{Z}}^*) = 0$, so that tr $\widehat{\mathbf{Z}}^{*\prime}\mathbf{Z} = \|\widehat{\mathbf{Z}}^*\|^2$ and

$$L_{\text{PCA}}(\widehat{\mathbf{Z}}^*) = \|\mathbf{Z} - \widehat{\mathbf{Z}}^*\|^2 = \|\mathbf{Z}\|^2 - \|\widehat{\mathbf{Z}}^*\|^2. \tag{2}$$

Because $\mathbf{Z}$ is column centered, $n^{-1}\|\mathbf{Z}\|^2$ equals the sum of the variances over the variables. Dividing both sides of (2) by $n$ times the total variance, $\|\mathbf{Z}\|^2$,

gives the proportion of variance accounted for

$$\frac{L_{\text{PCA}}(\hat{\mathbf{Z}}^*)}{\|\mathbf{Z}\|^2} = 1 - \frac{\|\hat{\mathbf{Z}}^*\|^2}{\|\mathbf{Z}\|^2}. \tag{3}$$

In distance-based PCA (Meulman, 1986, 1992), the relationships between the objects rather than the variables is emphasized. In particular, the data matrix $\mathbf{Z}$ is interpreted as an $m$-dimensional coordinate matrix for $n$ objects, where each object corresponds to a point in $m$-dimensional space, with coordinates given by the scores of the objects on the variables $\mathbf{z}_k$, $k = 1, \ldots, m$. The aim of distance-based PCA is to reconstruct the distances between these objects in high-dimensional space as closely as possible in a low-dimensional space $\hat{\mathbf{Z}}$. In distance-based PCA, not the sum of squared error defined on $\mathbf{Z}$ itself is minimized, but the Stress loss function defined on the distances between the rows of $\mathbf{Z}$, i.e.,

$$\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}) = \|D(\mathbf{Z}) - D(\hat{\mathbf{Z}})\|^2, \tag{4}$$

where $D(\mathbf{Z})$ is the $n \times n$ matrix of Euclidean distances between the rows of $\mathbf{Z}$ with elements

$$d_{ij}(\mathbf{Z}) = \left(\sum_{s=1}^{m}(z_{is} - z_{js})^2\right)^{1/2}. \tag{5}$$

We need the observation (De Leeuw, 1977) that the sum of the squared Euclidean distances equals $n$ times the sum of squares of the centered coordinates, i.e.,

$$\|D(\mathbf{Z})\|^2 = 2n\|\mathbf{Z}\|^2. \tag{6}$$

Therefore, $\|D(\mathbf{Z})\|^2$ equals $2n^2$ times the sum of the variances of the variables in $\mathbf{Z}$.

Suppose that $\hat{\mathbf{Z}}^{**}$ is a local minimum solution of $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}})$. As in classical PCA, we have orthogonality of the residuals of (4) at a local minimum $\hat{\mathbf{Z}}^{**}$ (see, e.g., De Leeuw & Heiser, 1977).

**Lemma 1.** *Let $\hat{\mathbf{Z}} \in \Omega$ with $\Omega$ a cone. At a local minimum $\hat{\mathbf{Z}}^{**} \in \Omega$ of $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}})$, the residuals are orthogonal to the estimates of the distances, i.e.,* $\text{tr } D(\hat{\mathbf{Z}}^*)[D(\mathbf{Z}) - D(\hat{\mathbf{Z}}^*)] = 0$.

*Proof.* If $\hat{\mathbf{Z}}$ is in a cone, then $\alpha\hat{\mathbf{Z}} \in \Omega$ for $\alpha > 0$ (Rockafellar, 1970, p. 13). $\Omega$ in PCA defines a cone, because if $\hat{\mathbf{Z}}$ is of rank $p$ then $\alpha\hat{\mathbf{Z}}$ is also of rank $p$. If $\hat{\mathbf{Z}}^{**} \in \Omega$ is a local minimum of $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}})$ then $\alpha\hat{\mathbf{Z}}^{**}$ must also be a local

minimum of $\sigma^2_{\text{PCA}}(\alpha\hat{\mathbf{Z}}^{**})$ for $\alpha = 1$. To find the minimum of $\sigma^2_{\text{PCA}}$ over $\alpha$, the fact is used that the Euclidean distance is positively homogeneous, i.e., $D(\alpha\hat{\mathbf{Z}}^{**}) = \alpha D(\hat{\mathbf{Z}}^{**})$ for any $\alpha \geq 0$. Thus,

$$\begin{aligned}\sigma^2_{\text{PCA}}(\alpha\hat{\mathbf{Z}}^{**}) &= \|D(\mathbf{Z}) - \alpha D(\hat{\mathbf{Z}}^{**})\|^2 \\ &= \|D(\mathbf{Z})\|^2 + \alpha^2\|D(\hat{\mathbf{Z}}^{**})\|^2 - \alpha\text{tr } D(\mathbf{Z})D(\hat{\mathbf{Z}}^{**}).\end{aligned}$$

Differentiating with respect to $\alpha$ and setting the derivative equal to zero, gives

$$\alpha^* = \frac{\text{tr } D(\mathbf{Z})D(\hat{\mathbf{Z}}^{**})}{\|D(\hat{\mathbf{Z}}^{**})\|^2}.$$

Because at this local minimum $\alpha^* = 1$, we have tr $D(\mathbf{Z})D(\hat{\mathbf{Z}}^{**}) = \|D(\hat{\mathbf{Z}}^{**})\|^2$, so that tr $D(\hat{\mathbf{Z}}^{**})[D(\mathbf{Z}) - D(\hat{\mathbf{Z}}^{**})] = 0$, which proves orthogonality of the residuals to the estimates of the distances. □

**Theorem 1.** *At a local minimum $\hat{\mathbf{Z}}^{**} \in \Omega$, $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}^{**})/(2n\|\mathbf{Z}\|^2)$ is equal to one minus the proportion of VAF.*

*Proof.* Using the orthogonality of the residuals from Lemma 1, the Stress at a local minimum can be expressed as

$$\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}^{**}) = \|D(\mathbf{Z})\|^2 - \|D(\hat{\mathbf{Z}}^{**})\|^2.$$

Using (6), $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}^{**})$ can also be expressed as

$$\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}^{**}) = 2n\|\mathbf{Z}\|^2 - 2n\|\hat{\mathbf{Z}}^{**}\|^2.$$

Dividing $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}^{**})$ by $\|D(\mathbf{Z})\|^2 = 2n\|\mathbf{Z}\|^2$ gives

$$\frac{\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}}^{**})}{\|D(\mathbf{Z})\|^2} = 1 - \frac{2n\|\hat{\mathbf{Z}}^{**}\|^2}{2n\|\mathbf{Z}\|^2} = 1 - \frac{\|\hat{\mathbf{Z}}^{**}\|^2}{\|\mathbf{Z}\|^2}$$

which is one minus the proportion of VAF. □

At their local minima $L_{\text{PCA}}$ and $\sigma^2_{\text{PCA}}$ can both be expressed as a proportion of VAF with respect to the variables. The next theorem tells how these proportions are related.

**Theorem 2.** *There always exists a local minimum of distance-based PCA with VAF larger than or equal to the VAF attained at the classical PCA solution.*

*Proof.* Let us denote $(z_{is} - z_{js}) = \mathbf{z}_s'(\mathbf{e}_i - \mathbf{e}_j)$ with $\mathbf{e}_i$ column $i$ of the identity matrix $\mathbf{I}$ and $\mathbf{z}_s$ column $s$ of $\mathbf{z}$. Then the squared distance can be expressed as

$$\begin{aligned} d_{ij}^2(\mathbf{Z}) &= \sum_{s=1}^{m} (\mathbf{z}_s'(\mathbf{e}_i - \mathbf{e}_j))^2 \\ &= \sum_{s=1}^{m} \mathbf{z}_s'(\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'\mathbf{z}_s = \text{tr } \mathbf{Z}'\mathbf{A}_{ij}\mathbf{Z}, \end{aligned}$$

where $\mathbf{A}_{ij} = (\mathbf{e}_i - \mathbf{e}_j)(\mathbf{e}_i - \mathbf{e}_j)'$.

Now consider the Cauchy-Schwarz inequality for the Euclidean distance, i.e.,

$$-d_{ij}(\mathbf{Z})d_{ij}(\hat{\mathbf{Z}}) = -(\text{tr } \mathbf{Z}'\mathbf{A}_{ij}\mathbf{Z})^{1/2}(\text{tr } \hat{\mathbf{Z}}'\mathbf{A}_{ij}\hat{\mathbf{Z}})^{1/2} \leq -\text{tr } \mathbf{Z}'\mathbf{A}_{ij}\hat{\mathbf{Z}}.$$

Summation over $i$ and $j$ gives

$$\begin{aligned} -\text{tr } D(\mathbf{Z})D(\hat{\mathbf{Z}}) &= -\sum_{i,j} d_{ij}(\mathbf{Z})d_{ij}(\hat{\mathbf{Z}}) \\ &\leq -\sum_{i,j} \text{tr } \mathbf{Z}'\mathbf{A}_{ij}\hat{\mathbf{Z}} = -\text{tr } \mathbf{Z}'(\sum_{i,j} \mathbf{A}_{ij})\hat{\mathbf{Z}} \\ &= -\text{tr } \mathbf{Z}'(2n\mathbf{I} - 2\mathbf{1}\mathbf{1}')\hat{\mathbf{Z}}, = -2n\text{tr } \mathbf{Z}'\hat{\mathbf{Z}}, \end{aligned} \tag{7}$$

since it is assumed that $\mathbf{Z}$ and $\hat{\mathbf{Z}}$ are column centered. Using (7) we get the inequality $\sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}) \leq 2nL_{\text{PCA}}(\hat{\mathbf{Z}})$, i.e.,

$$\begin{aligned} \sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}) &= \|D(\mathbf{Z}) - D(\hat{\mathbf{Z}})\|^2 = \|D(\mathbf{Z})\|^2 + \|D(\hat{\mathbf{Z}})\|^2 - 2\text{tr } D(\mathbf{Z})D(\hat{\mathbf{Z}}) \\ &\leq 2n\|\mathbf{Z}\|^2 + 2n\|\hat{\mathbf{Z}}\|^2 - 2n\text{tr } \mathbf{Z}'\hat{\mathbf{Z}} \\ &= 2n\|\mathbf{Z} - \hat{\mathbf{Z}}\|^2 = 2nL_{\text{PCA}}(\hat{\mathbf{Z}}). \end{aligned} \tag{8}$$

The majorization theory for multidimensional scaling (see, e.g., De Leeuw, 1988; Groenen, Mathar, & Heiser, 1995) guarantees a monotonically nonincreasing sequence of Stress values, implying the inequality

$$\sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}^{**}) \leq \sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}^{*}), \tag{9}$$

where as before $\hat{\mathbf{Z}}^{**}$ is a local minimum of $\sigma_{\text{PCA}}^2$ and $\hat{\mathbf{Z}}^{*}$ a local minimum of $L_{\text{PCA}}$. Combining (8) and (9) and dividing by $\|D(\mathbf{Z})\|^2$ gives

$$\frac{\sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}^{**})}{\|D(\mathbf{Z})\|^2} \leq \frac{\sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}^{*})}{\|D(\mathbf{Z})\|^2} \leq \frac{2nL_{\text{PCA}}(\hat{\mathbf{Z}}^{*})}{2n\|\mathbf{Z}\|^2}.$$

Combining this result with Theorem 1 and (3) gives

$$\frac{\sigma_{\text{PCA}}^2(\hat{\mathbf{Z}}^{**})}{\|D(\mathbf{Z})\|^2} = 1 - \frac{\|\hat{\mathbf{Z}}^{**}\|^2}{\|\mathbf{Z}\|^2} \leq 1 - \frac{\|\hat{\mathbf{Z}}^{*}\|^2}{\|\mathbf{Z}\|^2} = \frac{L_{\text{PCA}}(\hat{\mathbf{Z}}^{*})}{\|\mathbf{Z}\|^2},$$

which proves Theorem 2. □

Let us summarize the results above. We have shown the well known fact of expressing the results of classical PCA at a local minimum in terms of VAF. Theorem 1 extends this result to distance-based PCA: at a local minimum of $\sigma^2_{\text{PCA}}$ the loss can also be expressed in terms of VAF with respect to the variables. Moreover, Theorem 2 says that the VAF obtained by distance-based PCA is never lower than that of classical PCA whenever the majorization algorithm for distance-based PCA is started at the classical PCA solution.

Without loss of generality we may express $\hat{\mathbf{Z}} = \mathbf{X}\mathbf{A}'$ with $\mathbf{X}$ an orthonormal $n \times p$ matrix and $\mathbf{A}$ an $m \times p$ matrix. PCA can also be expressed as maximizing the sum of squared correlations between $\mathbf{Z}$ and $\mathbf{X}$, which is (globally) maximized by the minimum of $L_{\text{PCA}}(\hat{\mathbf{Z}})$. Therefore, this measure is higher than the one obtained at a minimum of $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}})$. Conversely, a minimum of $\sigma^2_{\text{PCA}}(\hat{\mathbf{Z}})$ maximizes (locally) the square of Tucker's congruence coefficient between $D(\hat{\mathbf{Z}})$ and $D(\mathbf{Z})$, i.e.,

$$\tau^2(D(\hat{\mathbf{Z}}), D(\mathbf{Z})) = \frac{\left(\text{tr } D(\hat{\mathbf{Z}})D(\mathbf{Z})\right)^2}{\|D(\hat{\mathbf{Z}})\|^2\|D(\mathbf{Z})\|^2}, \tag{10}$$

see, e.g., De Leeuw (1977), and this measure will be generally lower at a minimum of $L_{\text{PCA}}(\hat{\mathbf{Z}})$. In distance-based PCA the VAF is no longer equal to the squared correlation between $\mathbf{Z}$ and $\mathbf{X}$.

The increase in VAF by distance-based PCA can also be understood by geometrical arguments. The object space $\mathbf{X}$ in classical PCA is restricted to be a linear subspace of the high-dimensional space of $\mathbf{Z}$. In distance-based PCA, no such restrictions are imposed, so that there is more freedom to fit the data.

# 3 Multiple Correspondence Analysis

Multiple correspondence analysis (MCA) can be seen as PCA for categorical variable (for a recent account, see, e.g., Greenacre, 1984; Gifi, 1990; Heiser & Meulman, 1994). Each category is coded as a zero-one variable indicating presence (one) or absence (zero) of an object in the category. The categories of all variables are collected in the $n \times K$ matrix $\mathbf{G}$ of zero-one variables, where $K$ is the total number of categories. One way to specify classical MCA is to minimize

$$L_{\text{MCA}}(\hat{\mathbf{Z}}) = \|\mathbf{J}\mathbf{G}\mathbf{D}^{-1/2} - \hat{\mathbf{Z}}\mathbf{D}^{1/2}\|^2, \tag{11}$$

where $\mathbf{D} = \text{Diag}(\mathbf{1}'\mathbf{G})$ is the $K \times K$ diagonal matrix with the marginal frequencies of all categories, $\mathbf{J} = \mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}'$ centers the columns of $\mathbf{G}\mathbf{D}^{-1/2}$,

and $\hat{\mathbf{Z}}$ is a rank $p < K$ matrix. Without loss of generality, the low rank matrix $\hat{\mathbf{Z}}$ is factored as $\mathbf{XY}'$ with $\mathbf{X}$ and $n \times p$ matrix and $\mathbf{Y}$ an orthonormal $m \times p$ matrix, so that $\mathbf{Y}'\mathbf{DY} = \mathbf{I}$. The centering of $\mathbf{GD}^{-1/2}$ by $\mathbf{J}$ is required to avoid the uninteresting solution of $\mathbf{x}_1 = \mathbf{1}$ and $\mathbf{y}_1 = \mathbf{1}$, where $\mathbf{x}_1$ denotes the first column in $\mathbf{X}$, and $\mathbf{y}_1$ the first column in $\mathbf{Y}$.

There exists a nice relation between the sum of squares of $\mathbf{JGD}^{-1/2}$ and the $\chi^2$-statistic of $\mathbf{G}$, where $\mathbf{G}$ is considered as a degenerated contingency table (Meulman, 1986). Let $\mathbf{E}$ be the $n \times K$ matrix of expected values under the independence model, i.e., with elements $e_{ij} = d_j/n$, where $d_j$ denotes the marginal frequency of category $j$. The sum of squares of $\mathbf{JGD}^{-1/2}$ equals

$$\begin{aligned} \|\mathbf{JGD}^{-1/2}\|^2 &= \|(\mathbf{I} - n^{-1}\mathbf{1}\mathbf{1}')\mathbf{GD}^{-1/2}\|^2 = \|(\mathbf{G} - n^{-1}\mathbf{1}\mathbf{1}'\mathbf{G})\mathbf{D}^{-1/2}\|^2 \\ &= \|(\mathbf{G} - \mathbf{E})\mathbf{D}^{-1/2}\|^2 = \sum_{i,j} \frac{(g_{ij} - e_{ij})^2}{d_j} \\ &= n^{-1} \sum_{i,j} \frac{(g_{ij} - e_{ij})^2}{e_{ij}} = n^{-1}\chi^2. \end{aligned} \tag{12}$$

Therefore, in MCA we do not use the term VAF but $\chi^2$ accounted for (CAF).

The distance-based counterpart of MCA can be expressed as

$$\sigma^2_{\text{MCA}}(\mathbf{X}) = \|D(\mathbf{GD}^{-1/2}) - D(\hat{\mathbf{Z}}\mathbf{D}^{1/2})\|^2, \tag{13}$$

Note that (11) needs column centering of $\mathbf{GD}^{-1/2}$ by $\mathbf{J}$, whereas this is not necessary in (13) because distances do not change under translation, i.e., $D(\mathbf{JGD}^{-1/2}) = D(\mathbf{GD}^{-1/2})$. Also, $D(\hat{\mathbf{Z}}\mathbf{D}^{1/2}) = D(\mathbf{XY}'\mathbf{D}^{1/2}) = D(\mathbf{X})$ since $\mathbf{Y}'\mathbf{DY} = \mathbf{I}$, and Euclidean distances also do not change under rotation.

From (6) and (12) it can be derived that the sum of squared distances $\|D(\mathbf{GD}^{-1/2})\|^2 = 2\chi^2$.

The lemmas and theorems given in the previous section for PCA also hold for MCA. Thus, at a local minimum $\sigma^2_{\text{MCA}}/\|D(\mathbf{GD}^{-1/2})\|^2$ equals one minus the proportion of CAF. Furthermore, whenever the majorization algorithm for distance-based MCA is started from the classical MCA solution, its CAF at a local minimum is larger or equal to that of the CAF of classical MCA.

Classical MCA can also be expressed as the maximization of the sum of the so-called discrimination measures (Gifi, 1990) over dimensions and variables, that is,

$$n \sum_{s=1}^{p} \sum_{j=1}^{m} \eta^2_{js} = n \sum_{s=1}^{p} \frac{\mathbf{x}_s'\mathbf{GD}^{-1}\mathbf{G}'\mathbf{x}_s}{\|\mathbf{x}_s\|^2} \tag{14}$$

which is for classical MCA equal to the $\chi^2$ in a full dimensional solution, but not for distance-based MCA.

There is another way to generalize classical MCA to distance-based MCA where each categorical variable defines a separate set as in generalized canonical correlation analysis (Meulman, 1986, 1992), but for this generalization Theorem 2 cannot be proven anymore.

# 4 Multiple Regression and Analysis of Variance

The same idea discussed in the previous sections can be used in multiple regression analysis and analysis of variance (ANOVA) as well. The model becomes $\mathbf{y} = \hat{\mathbf{y}} + \mathbf{e} = \mathbf{X}\mathbf{b} + \mathbf{e}$, where $\mathbf{y}$ represents the dependent variable and $\mathbf{X}$ is the matrix of independent variables which is assumed to be of full rank. Again it is assumed that all variables have a mean equal zero.

In classical multiple regression analysis, the sum of squared error is minimized, i.e.,

$$L_{\text{reg}}(\mathbf{b}) = \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2, \tag{15}$$

which has the well known solution $\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$. At this minimimum, we have $\|\hat{\mathbf{y}}\|^2 = \hat{\mathbf{y}}'\mathbf{y}$, so that

$$L_{\text{reg}}(\mathbf{b}^*) = \|\mathbf{y}\|^2 - \|\hat{\mathbf{y}}\|^2,$$

or, dividing by $\|\mathbf{y}\|^2$,

$$\frac{L_{\text{reg}}(\mathbf{b}^*)}{\|\mathbf{y}\|^2} = 1 - \frac{\|\hat{\mathbf{y}}\|^2}{\|\mathbf{y}\|^2}.$$

In the distance-based version of multiple regression analysis, we minimize

$$\sigma^2_{\text{reg}}(\mathbf{b}) = \|D(\mathbf{y}) - D(\mathbf{X}\mathbf{b})\|^2. \tag{16}$$

In the notation of Section 2, we have $\mathbf{Z} = \mathbf{y}$ and $\hat{\mathbf{Z}} = \mathbf{X}\mathbf{b}$. To guarantee orthogonality of the residuals in Lemma 1, $\mathbf{X}\mathbf{b}$ must be in a cone. The substitution $\mathbf{b} = \alpha\tilde{\mathbf{b}}$ gives $\mathbf{X}\mathbf{b} = \alpha\mathbf{X}\tilde{\mathbf{b}}$ so that $\mathbf{X}\mathbf{b}$ is in a cone indeed. Therefore, the relations found in Section 2 hold here too.

# 5 Examples

## 5.1 Principal Components Analysis

To illustrate the difference in VAF, we analyze data from Mardia, Kent, and Bibby (1979, p. 3-4) of the marks of 88 students on five examinations.

TABLE 1. Comparison of classical PCA and distance-based PCA on data of Mardia et al. (1979). Reported are proportion of VAF, the average squared correlation $r^2_{\text{tot}}$ of variables and components, and the squared Tucker's coefficient $\tau^2$ between $D(\widehat{\mathbf{Z}})$ and $D(\mathbf{Z})$).

| | Classical PCA | | | Distance-based PCA | | |
|---|---|---|---|---|---|---|
| Dimensions | % VAF | $r^2_{\text{tot}}$ | $\tau^2$ | % VAF | $r^2_{\text{tot}}$ | $\tau^2$ |
| 1 | 63.6 | 63.6 | 89.6 | 91.4 | 61.7 | 91.4 |
| 2 | 78.4 | 78.4 | 97.0 | 97.9 | 77.0 | 97.9 |
| 3 | 87.3 | 87.3 | 99.0 | 99.4 | 86.5 | 99.4 |
| 4 | 95.1 | 95.1 | 99.8 | 99.8 | 94.8 | 99.8 |
| 5 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 | 100.0 |

Table 1 compares classical PCA and distance-based PCA in several solutions with a different number of dimensions with respect to the proportion of VAF, the average squared correlation between variables and components (which is equal to the VAF in classical PCA, but not in distance-based PCA), and the squared Tucker's coefficient $\tau^2$ between $D(\widehat{\mathbf{Z}})$ and $D(\mathbf{Z})$ (which is equal to the VAF with respect to the variables in distance-based PCA, but not in classical PCA).

As could be expected from the theoretical results, the VAF is higher for distance-based PCA than for classical PCA in all solutions. The PCA and distance-based PCA solutions are not very different, because the values of $\tau^2$ are only slightly higher for distance-based PCA compared to classical PCA, and conversely $r^2_{\text{tot}}$ for classical PCA is only slightly better than for distance-based PCA.

Note that in contrast to classical PCA, the dimensions in distance-based PCA are not nested, i.e., the first dimension in a two dimensional solution may be different from the first dimension in a three or higher-dimensional solution.

## 5.2 Multiple Correspondence Analysis

To illustrate the difference in reconstructed $\chi^2$, 36 different types of cetacea (whales, dolphins, and porpoises) reported by Vescia (1985) were analyzed (see also, Burg, 1985; Meulman, 1986) by MCA and distance-based MCA. The morphology, osteology, and behavior of the cetacea were specified in 15 categorical variables yielding a total of 59 different categories.

The proportion of CAF for these data is reported in Table 2. Also reported is the sum of the discrimination measures (14) as a proportion of the total $\chi^2$, i.e.,

$$\eta^2_{\text{tot}} = \frac{n}{n(K-m)} \sum_{s=1}^{p} \frac{\mathbf{x}_s' \mathbf{G} \mathbf{D}^{-1} \mathbf{G}' \mathbf{x}_s}{\|\mathbf{x}_s\|^2}. \tag{17}$$

TABLE 2. Comparison of classical MCA and distance-based MCA on the cetacea data of Vescia (1985). Reported are proportion of CAF, $\eta^2_{\text{tot}}$, and the squared Tucker's coefficient $\tau^2$ between $D(\widehat{\mathbf{Z}})$ and $D(\mathbf{Z})$).

| | Classical MCA | | | Distance-based MCA | | |
|---|---|---|---|---|---|---|
| Dimensions | % CAF | $\eta^2_{\text{tot}}$ | $\tau^2$ | % CAF | $\eta^2_{\text{tot}}$ | $\tau^2$ |
| 1 | 20.3 | 20.3 | 45.2 | 78.7 | 18.1 | 78.7 |
| 2 | 34.7 | 34.7 | 55.2 | 92.1 | 28.9 | 92.1 |
| 3 | 45.4 | 45.4 | 61.7 | 96.1 | 40.0 | 96.1 |
| 4 | 54.3 | 54.3 | 63.3 | 97.7 | 48.5 | 97.7 |
| 5 | 62.4 | 62.4 | 63.1 | 98.6 | 56.6 | 98.6 |

TABLE 3. Hypothetical data of $2 \times 3$ factorial design for ANOVA from Stevens (1992).

| Method A | Method B | y | Method A | Method B | y |
|---|---|---|---|---|---|
| 1 | 1 | 3 | 2 | 1 | 9 |
| 1 | 1 | 5 | 2 | 1 | 14 |
| 1 | 1 | 6 | 2 | 1 | 5 |
| 1 | 2 | 2 | 2 | 2 | 6 |
| 1 | 2 | 4 | 2 | 2 | 7 |
| 1 | 2 | 8 | 2 | 2 | 7 |
| 1 | 3 | 11 | 2 | 3 | 9 |
| 1 | 3 | 7 | 2 | 3 | 8 |
| 1 | 3 | 8 | 2 | 3 | 10 |

It is clear that distance-based MCA achieves much higher CAF than classical MCA. In one dimension, classical MCA reconstructs 20% CAF, whereas distance-based MCA reconstructs 78%. Distance-based MCA reaches high proportions of CAF much faster than classical MCA. Since $\eta^2_{\text{tot}}$ is globally maximized by classical MCA, its values are higher for classical than for distance-based MCA. Similarly, $\tau^2$ is maximized in distance-based MCA, so its values are higher than the $\tau^2$ values for classical MCA.

The reason that distance-based MCA yields such high CAF in low-dimensional solutions is that the dimensionality of $\mathbf{G}$ is very high (in this instance 35), and the classical projection techniques are much more restrictive than the nonlinear mappings obtained by the distance-based approach.

## 5.3 Analysis of Variance

Consider a small hypothetical example of ANOVA taken from Stevens (1992). The data of this $2 \times 3$ factorial design are reported in Table 3. The decomposition of sum of squares of classical ANOVA and distance-based ANOVA is reported in Table 4. The proportion VAF by distance-based ANOVA is 63% and by classical ANOVA is 48%.

Because the design is factorial, the reconstructed sum of squares can be

TABLE 4. Decomposition of sum of squares for the data from Stevens (1992) by classical ANOVA and distance-based ANOVA

| Effect | SSQ classic | SSQ distance-based |
|---|---|---|
| Main A | 24.500 | 40.500 |
| Main B | 30.333 | 32.350 |
| Interaction A×B | 14.333 | 17.905 |
| VAF | 69.167 | 90.755 |
| Within cells error | 75.333 | 53.745 |
| Total | 144.500 | 144.500 |
| Proportion VAF | .479 | .628 |
| $R^2(\mathbf{y}, \hat{\mathbf{y}})$ | .479 | .437 |
| $\tau^2(D(\mathbf{y}), D(\hat{\mathbf{y}}))$ | .573 | .628 |

decomposed according to the two factors and their interaction. The decomposition of VAF shows higher values for all effects in the distance-based ANOVA solution when compared to the classical ANOVA solution.

Note that the squared multiple correlation is equal to VAF for classical ANOVA, but for distance-based ANOVA it drops to .44. Similarly, the squared Tucker's coefficient is equal to the VAF for distance-based ANOVA, but for classical ANOVA it decreases to .57.

# 6 Discussion and Conclusions

In this paper we have shown that not only classical multivariate analysis models can be expressed in terms of the proportion of variance of the variables accounted for (VAF) at the minimum, but that this also holds for distance-based multivariate analysis models. Moreover, it was shown that the VAF of classical models can be increased by switching to distance-based models without estimating additional parameters. This theory was illustrated for principal components analysis, multiple correspondence analysis, multiple regression and analysis of variance.

This paper has shown that the solutions of classical and distance-based multivariate analysis models can be compared in terms of the neutral measure of the proportion of variance of the variables accounted for. Note that VAF itself is not maximized by either two models. Maximizing VAF itself is trivial, because it is not bounded from above. Thus, without a model like (1) or (4), the VAF has no meaning. However, it is true that classical PCA globally maximizes VAF under the constraint that (1) is minimal, and that distance-based PCA (locally) maximizes VAF with respect to the variables under the constraint that (4) is (locally) minimal. We always have to realize that the

VAF in distance-based multivariate analysis is not equal to the (average) squared correlation.

The theory discussed in this paper can be extended to models with more than one set of data. As long as the requirements outlined in Section 2 are fulfilled, the classical MVA model will account for less variance than its distance-based counterpart.

Though we proved that distance-based MVA attains higher VAF than classical MVA, there may well exist other loss functions that obtain even higher values of VAF at their minimum.

A program called PIONEER (Groenen, Commandeur, & Meulman, 1997) for performing distance-based multivariate analysis can be found at

http://www.fsw.leidenuniv.nl/www/w3_data/pioneer/pioneer.htm

The current version of the program (version 3.1) contains an implementation of distance-based PCA.

# References


Burg, E. van der. (1985). HOMALS classification of whales, porpoises and dolphins. In J.-F. Marcotorchino, J.-M. Proth, & J. Janssen (Eds.), *Data analysis in real life environment: Ins and outs of solving problems* (pp. 25–36). Amsterdam: North-Holland.

De Leeuw, J. (1977). Applications of convex analysis to multidimensional scaling. In J. R. Barra, F. Brodeau, G. Romier, & B. van Cutsem (Eds.), *Recent developments in statistics* (pp. 133–145). Amsterdam, The Netherlands: North-Holland.

De Leeuw, J. (1988). Convergence of the majorization method for multidimensional scaling. *Journal of Classification, 5*, 163–180.

De Leeuw, J., & Heiser, W. J. (1977). Convergence of correction-matrix algorithms for multidimensional scaling. In J. C. Lingoes, E. E. Roskam, & I. Borg (Eds.), *Geometric representations of relational data* (pp. 735–752). Ann Arbor, MI: Mathesis Press.

Gifi, A. (1990). *Nonlinear multivariate analysis.* Chichester: Wiley.

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

Groenen, P. J. F., Commandeur, J. J. F., & Meulman, J. J. (1997). PIONEER: A program for distance-based multivariate analysis. In W. Bandilla & F. Faulbaum (Eds.), *Softstat'97: Advances in statistical software* (pp. 83–90). Stuttgart: Lucius & Lucius.

Groenen, P. J. F., Mathar, R., & Heiser, W. J. (1995). The majorization approach to multidimensional scaling for Minkowski distances. *Journal of Classification, 12*, 3–19.

Heiser, W. J., & Meulman, J. J. (1994). Homogeneity analysis:Exploring the distribution of variables and their nonlinear relationships. In M. J. Greenacre & J. Blasius (Eds.), *Correspondence analysis in the social sciences* (pp. 179–209). Londen: Academic Press.

Mardia, K. V., Kent, J. T., & Bibby, J. M. (1979). *Multivariate analysis.* London: Academic Press.

Meulman, J. J. (1986). *A distance approach to nonlinear multivariate analysis.* Leiden, The Netherlands: DSWO Press.

Meulman, J. J. (1992). The integration of multidimensional scaling and multivariate analysis with optimal transformations. *Psychometrika, 57*, 539–565.

Rockafellar, R. T. (1970). *Convex analysis.* Princeton, NJ: Princeton University Press.

Stevens, J. (1992). *Applied multivariate statistics for the social sciences* (Second ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Vescia, G. (1985). Descriptive classification of cetacea: Whales, porpoises and dolphins. In J.-F. Marcotorchino, J.-M. Proth, & J. Janssen (Eds.), *Data analysis in real life environment: Ins and outs of solving problems* (pp. 7–13). Amsterdam: North-Holland.