MULTIDIMENSIONAL SCALING USING FACTOR SCORES

[1]Roberto N. Padua and [2]Joy M. Picar

Abstract

In this paper, we show how the modified factor scores obtained from factor analysis can be used as inputs to the classical multidimensional scaling problem. Some optimality theorems related to the Takane stress function are stated and proved. The use of factor scores in multidimensional scaling is justified on the basis of easy interpretability.

Keywords: *factor scores, multidimensional scaling, stress function.*

1.0 Introduction

The multidimensional scaling problem finds a representation of p-dimensional random vectors in low g-dimensional space such that the observed distances between pairs of vectors in the original p-dimensional space are most closely preserved. Scaling techniques were developed Shepard (1980), Kruskal (1978) and Takane (1977). Young et al. (1987) provides a good summary of the history and applications of multidimensional scaling.

The key objective of multidimensional scaling procedures is a low-dimensional picture because visual inspection can greatly aid interpretations. It is a technique for dimension reduction in statistics. When the multivariate observations are naturally numerical and Eulidean distances in R^p, $d_{ik}^{(p)}$, can be computed, the MDS seeks a q < p – dimensional representation by minimizing:

$$1.1 \quad E = \frac{\Sigma\Sigma_{i<k} (d_{ik}^{(p)} - d_{ik}^{(q)})^2}{\Sigma\Sigma_{i<k} d_{ik}^{(p)}} \quad \text{(Johnson et al., 1987)}$$

Equation (1.1) is called the stress function. Lower values of *E* denote better fit in the lower-dimensional space. Takane et al., (1977) introduced an alternative criterion to (1.1.):

$$1.2.\ \text{SStress} = \frac{\Sigma\Sigma_{i<k} (d_{ik}^{(p)2} - d_{ik}^{(q)2})^2}{\Sigma\Sigma_{i<k} d_{ik}^{(q)4}}$$

which is always a number between 0 to 1. Any value of SStress less than 0.10 is typically taken to mean there is a good representation of the objects by the points in the given configuration. Techniques for obtaining low-dimensional representations by minimizing (1.1) or (1.2) involve the use of nonlinear mappings and are often quite complex. For this reason, it is reasonable to search for techniques. That are easier and more intuitively appending. Such techniques usually allow for better interpretations of the results from MDS. In this paper, we investigate the stress characteristics of lower dimensional representations of multivariate observations obtained as a by-product of factor analysis. Factors are generally understood as groups of highly correlated variables that describe a construct being measured by the component variables. We propose to replace the p-dimensional factor score *F*.

[1] *Liceo de Cagayan University*
[2] *Mindanao University of Science and Technology*

2.0 Basic Concepts

Let $X = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{pmatrix}$ be a p – dimensional random vector with mean $\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$ and covariance matrix Σ. We define:

Definition 1. The random vector X has an orthogonal factor representation:

$$x_1 - \mu_1 = \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1$$
$$x_2 - \mu_2 = \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \varepsilon_2$$
$$\vdots$$
$$x_p - \mu_p = \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p \quad m < p$$

or

$$\underset{(px1)}{X - \mu} = \underset{(pxm)}{L}\ \underset{(mx1)}{F} + \underset{(px1)}{\varepsilon}$$

where:

$$E(X) = \mu \quad , \quad cov(X) = \Sigma$$
$$E(F) = 0 \quad , \quad cov(F) = I$$
$$E(\varepsilon) = \mu \quad , \quad cov(\varepsilon) = \Psi$$
$$cov(F, \varepsilon) = 0$$

Definition 2. The components of X are linearly dependent upon a few ($m < p$) unobservable random variables $F_1, F_2, \ldots, F_m$ called **common factors** and p random errors $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_p$, called **specific factors**. The matrix L is called the matrix of **factor loading** and ℓ_{ij} is the **loading** of the i^{th} variable on the j^{th} factor.

Theorem 1. Given the orthogonal factor representation of a p-variate random vector X, then:

$$\Sigma\ LL^T + \psi$$

Proof: From $X - \mu = LF + \varepsilon$, we obtain:

$$cov(X - \mu) = cov\ LF + \varepsilon$$
$$E(X - \mu)(X - \mu)^T = L\ cov(F)L^T + cov(\varepsilon)$$
since $cov(F, \varepsilon) = 0$

$$\sum = L.IL^T + \psi$$

$$\sum = LL^T + \psi \text{ as desired.}$$

Corollary 1. From $\Sigma = LL^T + \psi$, then:

(i.) $Var(X_i) = \ell_{i1}^2 + \ell_{i2}^2 + \ldots + \ell_{im}^2 + \psi_i$

$cov(X_i X_k) = \ell_{i1}\ell_{k1} + \ell_{i2}\ell_{k2} + \ldots + \ell_{im}\ell_{km}$

(ii.) $cov(X_i F_j) = \ell_{ij}$

Proof: (i). The variance of X_i , $Var(X_i)$, is the ith diagonal element of Σ. The ith diagonal element of $LL^T + \psi$ is: is:

$$Var(X_i) = \sum^{m} \ell_{i1}^2 + \psi_i \ , \quad j = 1,2,\ldots, m$$

The covariance between X_i and X_j , $cov(X_i, X_j)$ is the ij^{th} element of Σ. The i^{th} element of $LL^T + \psi$ is:

$$cov(X_i, X_j) = \sum^{m} \ell_{ik}\ell_{jk} + 0 = \ell_{i1}\ell_{j1} + \ell_{i2}\ell_{j2} + \ldots + \ell_{im}\ell_{jm}$$

(iii.) $cov(X, F) = E[(X - \mu)F^T]$

$= E(LFF^T + \varepsilon F^T)$

$= LE(FF^T) + 0$ since $cov(\varepsilon, F) = 0$

$= L.I$ since $var(F) = I$

$= L$

Hence, $cov(X_i, F_j) = \ell_{ij}$

Factor analysis aims to reconstruct the covariance matrix Σ by finding L and ψ that satisfy Theorem 1. A particularly useful result in linear algebra is stated below for reference:

Theorem 2. Spectral Decomposition Theorem. Let Σ be a positive p x p matrix with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_p$. Let $e_1, e_2, \ldots, e_p$ be the orthogonal eigenvectors corresponding to these eigenvalues. Let P be the orthogonal matrix whose diagonals are the eigenvalues of Σ. Then:

$$\sum = P\,D\,P^T$$

We can write:

$$2.1.\quad \Sigma = \left[\sqrt{\lambda_1}e_1 : \sqrt{\lambda_2}e_2 : \ldots : \sqrt{\lambda_p}e_p\right]\begin{bmatrix}\sqrt{\lambda_1}e_1\\ \sqrt{\lambda_2}e_2\\ \vdots\\ \sqrt{\lambda_m}e_m\end{bmatrix}.$$

If we now take $L = \left[\sqrt{\lambda_1}e_1 : \sqrt{\lambda_2}e_2 : \ldots : \sqrt{\lambda_p}e_p\right]$, we have the exact representation:

$$2.2\quad \Sigma = LL^T + 0 \text{ where } \psi = 0.$$

The factor analysis representation (2.2) is exact yet is not useful in practice because there are as many factor representations as there are variables and does not allow for specific variations ε. One remedy is to consider only the *m* largest eigenvalues and to neglect the $p - m$ smalls eigenvalues in their contribution to Σ, e.i. neglect λ_{m+1}, e_{m+1}, $e^T_{m+1} + \lambda_{m+2}\ e_{m+2}e^T_{m+2} + \ldots + \lambda_p e_p e_p{}^T$. Hence, we obtain the approximation:

$$\Sigma \simeq \left[\sqrt{\lambda_1}e_1 : \sqrt{\lambda_2}e_2 : \ldots : \sqrt{\lambda_m}e_m\right]\begin{bmatrix}\sqrt{\lambda_1}e_1\\ \sqrt{\lambda_2}e_2\\ \vdots\\ \sqrt{\lambda_m}e_m\end{bmatrix} + \begin{bmatrix}\psi_1 & 0 & 0 & \ldots & 0\\ 0 & \psi_2 & 0 & \ldots & 0\\ \ldots & \ldots & \ldots & \ldots & \psi_p\end{bmatrix}$$

$$\simeq LL^T + \psi$$

The elements of ψ are obtained from:

$$2.4\quad \psi_i = diag_i\,(\Sigma - LL^T).$$

The spectral decomposition approach or principal components method to factor analysis is adopted in this study. We next define what we mean by factor scores.

Definition 3. The estimated values of the common factors in the orthogonal factor model are called factor scores.

Factor scores are not estimates of unknown parameters in the usual sense. Rather, they are estimates of values for the unobservable random vectors *Fj, j* = 1, …, n. That is,

$$2.5\quad \hat{f_j} = \text{estimate of the values } f_j \text{ attained } F_j \text{ for the } j^{th} \text{ case.}$$

Theorem 3. Suppose that the mean vector μ, factor loading L and specific variance ψ are known for:

$$X - \mu = LF + \varepsilon$$

Then the least-squares estimate of the scores for the jth case is:

$$\hat{f_j} = (L^TL)^{-1}L^T(x_j - \bar{x}).$$

Proof: Note that:

$$\varepsilon = X - \mu - Lf$$

So that:

$$2.6\quad SSE = \varepsilon^T\varepsilon = (X - \mu - Lf)^T(X - \mu - Lf)$$

We find ***f*** that minimizes (2.6). taking the derivative with respect to *f* yields:

$$2.7\quad L^TLf = L^T(X - \mu) \text{ or:}$$

$$2.8\quad \hat{f} = (L^TL)^{-1}L^T(X - \mu).$$

The least-squares estimate of μ is $\bar{x}$, hence, for the jth case:

$$2.9\quad \hat{f} = (L^TL)^{-1}L^T(x_j - \bar{x}).$$

The regression approach yields a different formula for f_j:

$$2.10\quad \hat{f_j} = L^TS^{-1}(X_j - \bar{X}),$$ where *S* is an estimate of Σ.

2.2 Classical Multidimensional Scaling

Let x_i ,1,2,…,n be p-dimensional vectors and denote the n x p data matrix by x. Define $\delta_{ij}^2 = (x_i - x_j)^T (x_i - x_j)$ and let A = (α_{ij}) where $\alpha_{ij} = -\frac{1}{2} d_{ij}^2$. Define the matrix by $H = I_n - \frac{1}{n} II^T$ and $B = HAH = (b_{ij})$, $(b_{ij}) = (x_i - \bar{x})^T (x_i - \bar{x})$. Following the Lemma are now stated and proved.

Lemma 1. Let $H = I_n - \frac{1}{n} II^T$, then

$$H^2 = H \text{ and } H^T = H$$

so that H is symmetric idempotent.

Proof:

$$H^2 = (I - \tfrac{1}{n} II^T)(I - \tfrac{1}{n} II^T)$$

$$= (I - \tfrac{2}{n} II^T + \tfrac{n}{n^2} II^T)$$

$$= I - \tfrac{1}{n} II^T = H$$

To show symmetry,

$$H^T = (I^T - \tfrac{1}{n} II^T = I - \tfrac{1}{n} II^T = H$$

A useful alternative expression for B that uses the original data matrix is $B = (HX)(HX)^T$.

Lemma 2. The matrix B is a real, symmetric positive-semidefinite matrix.

Proof: Note that B is an n x n matrix. To prove symmetry, we have:

$$B^T = [(HX)(HX)^T]^T = (HX)(HX)^T = B.$$

Let y be an n x 1 vector and let $Z = Hy$. Then $Z^TBZ = Z^T(HX)(HX)^TZ$. Thus:

$$Z^TBZ = (HY)^T (HX)(HX)^T (HY).$$
$$= Y^TH^2 XX^T H^2 Y$$
$$= Y^T H(XX^T) HY = Z^T (XX^T) Z \geq 0.$$

It follows from Lemma 2 that the eigenvalues of B are either positive or zero. If pxn and the pxp sub-matrix Σ_p of B is non-singular, there will be $n - p$ zero eigenvalues. Let the spectral decomposition of B be $B = VDV^T$ where D is a diagonal matrix:

$$\begin{matrix} \lambda_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \square & \vdots \\ 0 & \cdots & \lambda_{p_0} & \vdots \\ 0 & \cdots & \cdots & 0 \end{matrix}$$

$V = [V_1 : V_2 : \ldots : V_p : 0 \ldots : 0]$ is the matrix whose columns are the eigenvectors corresponding to these eigenvalues. Re-arrange the eigenvalues so that $\lambda_1 \geq \lambda_2, \geq \ldots \geq \lambda_p \geq 0$, and let V_p be the n x p matrix of properly orthonormalized eigenvectors. Write

$$B = \underset{nxp}{V_p} \underset{nxp}{V_p} \underset{nxp}{V_p^T} \quad \text{or} \quad B = \left(V_p D_p^{\frac{1}{2}}\right)\left(V_p D_p^{\frac{1}{2}}\right)^T.$$

The reconstructed coordinates of X can be written as: $\hat{X} = V_p D_p^{1/2}$. The origin of the reconstructed coordinates is at the mean, $\bar{X}$, of the n points and the axes are the principal axes of the $\hat{X}_j$ configuration.

The optimality theorem below is due to Madria et al. (1979).

Theorem 4. Let X denote a configuration of points in R^p with interitem distances $\delta_{ij}^2 = (X_i - X_j)^T (X_i - X_j)$. Let L be a p x p rotation matrix and set $L = (L_1, L_2)$, where L_1 is a p x k matrix for $k < p$. Let $\hat{X} = XL_1$ be the projection onto a k – dimensional subspace of R^p and let $d_{ij}^2 = (\hat{X}_i - \hat{X}_j)^T (\hat{X}_i - \hat{X}_j)T$. Then, among all projections $\hat{X} = XL_1$, the quantity

$$\Phi = \sum_{i,j} (\delta ij - \hat{d}_{ij}^2)$$

is minimized when X is projected onto its coordinates in k dimensions. Moreover, $d_{ij} \leq \delta_{ij}$ for all i, j.

Proof: Let $B = (HX)(HX)^T = b_{ij}$, where, $b_{ij} = (x_i - x_j)^T (x_i - x_j)$ and let $\hat{B} = (H\hat{X})(H\hat{X})^T = \hat{b}_{ij}$ where $\hat{b}_{ij} = (\hat{x}_i - \hat{x}_j)^T (\hat{x}_i - \hat{x}_j)$. Then Φ can be expressed as:

$$\Phi = \| B - \hat{B} \| = \| H (XX^T - \hat{X}\hat{X}^T) H \|$$

Thus, minimizing Φ is equivalent to minimizing $\hat{X}\hat{X}^T$. Let $\hat{X} = XL_1$ be an arbitrary projection onto the k-dimensional subspace of R^p. We have:

$$Max: \hat{X}\hat{X}^T = X(LL_1^T)X$$
$$\text{Subject to: } X_j^T Xj \le 1 \text{ for } j = 1, \ldots, n.$$

Using Lagrange multiplier, we from the Lagrangian function:

$$L = X_j^T (L_1^T L_1) Xj - \lambda (X_j^T X_j - 1)$$

Taking the derivative and setting it equal to zero yields:

$$\frac{\partial L}{\partial x_j} = 2\Sigma_1 x_j - 2\lambda x_j = 0$$

where

$$\Sigma_1 = L_1^T L_1$$

Hence:

$$(\Sigma_1 - \lambda I)x_j = 0 \text{ for } j = 1, 2, \ldots, n$$

Which will have non-trivial solutions iff det $(\Sigma_1 - \lambda I)x_j = 0$. The values of λ are, therefore, precisely the eigenvalues of Σ_1 and the solutions are the orthonormalized eigenvectors of Σ_1 .

3.0 Main Results

The proposed reconstructed coordinates of X in n-dimension consists of the factor scores $\hat{f}$ which are obtained essentially from the principal components decomposition of the covariance matrix $\sum$. We attempt to find a relationship between the classical scaling solution and principal component analysis. This relationship is well-established (Cox and Cox, 1994).

Theorem 5. Let S be the sample covariance matrix of X. then, the projection of X onto the eigenvectors of nS returns the classical scaling solution.

Proof: We show that the eigenvalues of nS are the p non-zero eigenvalues of B. Let:

$$S = \frac{1}{n} X^T HX$$

Since $H^2 = H$, we can be write $nS = (HX)^T(HX)$. Let v_i be an orthonormal vector of B such that $Bv_i = \lambda_i v_i$, that is: $Bv_i = (HX)(HX)^T v_i = \lambda_i v_i$. Pre-multiply by $(HX)^T$:

$$(HX)TBvi = (HX)^T)(HX)(HX)^T v_i = \lambda_i (HX)$$

or

$$(nS)(HX)^T v_i = \lambda (HX)^T v_i = \lambda_i (HX)$$

Hence, λ_i is an eigenvalue of nS also. Here, Y_i is the corresponding eigenvector of n S. it is easy to show that $Y_i^T Y_i = \lambda_i$.

Centering X (by HX) and projecting onto the unit vector $\hat{Y}_i = \frac{Y_i}{\sqrt{\lambda_i}}$ yields:

$$(HX) = \hat{Y}_i = \lambda_i^{-\frac{1}{2}} (HX)(HX)^T v_i = \lambda_i^{-\frac{1}{2}} v_i \quad \blacksquare$$

Let $\hat{f}_i$ be given by Equation (2.10), then the reconstructed data matrix in m-dimension using a modified version of (2.10) is:

$$3.1 \quad \hat{f}_j = LTS^{-½}(X - \bar{X})$$

It follows that $\hat{f}^T = (X - \bar{X})^T S^{-½} L$ where S^{-1} is the estimated covariance matrix of X. We write in the more familiar notation:

$$3.2 \quad \hat{f}^T = (HX)^T S^{-½} L_m$$

Let Φ be defined by theorem 4 and let $B = (HX)(HX)^T = b_{ij}$ where $b_{ij} = (x_i - x_j)^T (x_i - x_j)$. Define $\hat{B} = \hat{f}\hat{f}^T = (\hat{b}_{ij})$ where $b_{ij} = (\hat{f}_i - \hat{f}_j)^T(\hat{f}_i - \hat{f}_j)$

The next theorem shows that the PCA approach is equivalent to the factor scores approach.

Theorem 6. Let X be an $n \times p$ matrix of random vectors and let the factor scores be defined by (3.1) where S is the estimated covariance matrix of X. Let B, $\hat{B}$ and Φ be as

defined. Then, the factor scores $\hat{f}$ minimizes Φ and attains the same minimum as the principal components projection of the reconstructed data on an m-dimensional subspace of R^p.

Proof: Let $L = V_p D^{1/2}$ and let $L = (L_m, L_{p-m})$ denote the first m components and second p-m components of L. Then, $\hat{f}^T = (HX)^T S^{-\frac{1}{2}} L_m^T$ is the standardized value of X projected onto the subspace generated by the first m principal components of the covariance matrix. By Theorem 4, Φ is minimized by the representation $\hat{f}$.

Let $\hat{f} = L_m S^{-\frac{1}{2}} H(X)$.

Then:

$$3.3 \quad \hat{f}\hat{f}^T = L_m S^{-\frac{1}{2}} H(X) H(X)^T S^{-\frac{1}{2}} L_m^T$$
$$= L_m S^{-\frac{1}{2}} B S^{-\frac{1}{2}} L_m^T$$
$$= L_m L_m^T = \Sigma_m$$

Diagonalizing Σ_m, we have: $D_m = V_m^T \Sigma_m V_m$. Hence:

$$3.4 \quad \Phi = \| D - Dm \| = 2n (\lambda_{m+1} + \ldots + \lambda_p)$$

as desired. The $\Phi = \Sigma_{i,j} (\delta_{ij}^2 - \hat{d}_{ij}^2)$ function is recognized as a variant of Takane's (1987) stress function:

$$3.5 \quad \Phi = \frac{\Sigma_{i,j} (\delta_{ij}^2 - \hat{d}_{ij}^2)^2}{\Sigma_{i,j} \delta_{ij}^4}$$

So that if Φ is minimized so is (3.5).

Theorem 7: The factor score representation of X minimizes the Takane stress function among all projections onto the $k < p$ dimensional subspace of R^p.

Proof: Take the derivative of (3.5) with respect to δ_{ij}^2, we have:

$$\frac{dS}{d\delta_{ij}^2} = \frac{\Sigma(\delta_{ij}^2 - \hat{d}_{ij}^2)}{\Sigma \delta_{ij}^2} = \frac{\Phi}{\Sigma \delta_{ij}^2}$$

which is close to zero when Φ is minimum.

4.0 Concluding Discussions

The classical scaling solution to the metric multidimensional scaling problem is seen to be equivalent to the approach of projecting the data onto an $m < p$ dimensional subspace of R^p. In this paper, we investigated a variant to this approach by considering the factor scores $\hat{f}$ obtained in the usual factor analysis procedure. It is shown that the standardized data are, in fact, projected onto the subspace of $m < p$ dimension through the first m principal components axes. This latter proposal has the advantage of being more readily interpretable when the factors are identified correctly in the preliminary factor analysis.

References:

Anderson, T.W. (1984). *An introduction to multivariate statistical methods*. 2nd ed. New York, John Wiley.

Bartlett, M.S. (1937). The statistical conception of mental factors. *British Journal of Psychology, Vol. 18, 97-104.*

Kruskel, J.B. (1978). Multidimensional scaling. *Sage University Paper Series on Quantitative Application in Social Sciences, 07-011.* Beverly Hills, London: Sage Publications.

Takane, Y., Young, F.W. and de Leeuw, J. (1977). Non-metric individual differences multidimensional scaling: alternating least squares with optimal scaling feature. *Psycometrika, Vol. 42, 7-67.*

Young, F.W. and Hamer, R.M. (1987). *Multidimensional scaling: History, theory and Application.* Hillsdale, N.J. Lawrence Erlbaum Association Publishers.