Chapter 10 Decomposition of Data Matrices by Factors

Applied Multivariate Statistical Analysis

Course Instructor: Shuen-Lin Jeng

Department of Statistics
National Cheng Kung University

March, 10, 2021

10.1 The Geometric Point of View

In this chapter, we take a descriptive perspective and show how using a geometrical approach provides the "best" way of reducing the dimension of a data matrix.

It is derived with respect to a least squares criterion.

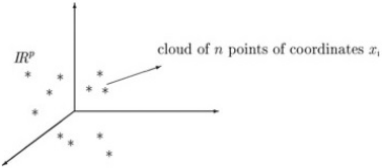
As a matter of introducing certain ideas , assume that the data matrix $\mathcal{X}(n \times p)$ is composed of n observations (or individuals) of p variables.

10.1 The Geometric Point of View

There are in fact two ways of looking at $\mathcal{X}$, row by row or column by column:

1. Each row (observations) is a vector $x_i^\top = (x_{i1}, \dots, x_{ip}) \in \mathbb{R}^p$.

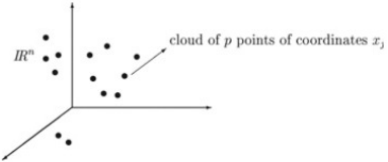
Fig. 10.1 Cloud of n points in $\mathbb{R}^p$



10.1 The Geometric Point of View

2. Each column (variable) is a vector $x_{[j]} = (x_{1j}, \ldots, x_{nj})^T \in \mathbb{R}^n$

Fig. 10.2 Cloud of p points in $\mathbb{R}^n$



When n and/or p are large(larger than 2 or 3), we cannot produce interpretable graphs of these clouds of points, We shall try to simultaneously approximate the column space $C(\mathcal{X})$ and the row space $C(\mathcal{X}^T)$ with smaller subspaces.

10.2 Fitting the p-Dimensional Point Cloud

Subspaces of Dimension 1

In this section, $\mathcal{X}$ is represented by a cloud of n points in $\mathbb{R}^p$ (considering each row). The question is how to project this point cloud onto a space of lower dimension. The problem boils down to finding a straight line F_1 through the origin.The direction of this line can be defined by a unit vector $u_1 \in \mathbb{R}^p$.

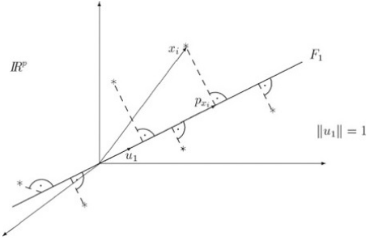


Fig. 10.3 Projection of point cloud onto u space of lower dimension

10.2 Fitting the p-Dimensional Point Cloud

The representation of the i-th individual $x_i \in \mathbb{R}^p$ on this line is obtained by the projection of the corresponding point onto u_1, i.e., the projection point p_{x_i}. We know from (2.42) that the coordinate of x_i on F_1 is given by

$$p_{x_i} = x_i^\top \frac{u_1}{\|u_1\|} = x_i^\top u_1. \tag{10.1}$$

We define the *best line* F_1 in the following“”least-squares” sense: Find $u_1 \in \mathbb{R}^p$ which minimizes

$$\sum_{i=1}^{n} \left\| x_i - p_{x_i} \right\|^2 \tag{10.2}$$

10.2 Fitting the p-Dimensional Point Cloud

Since $\left\|x_i - p_{x_i}\right\|^2 = \|x_i\|^2 - \left\|p_{x_i}\right\|^2$ by Pythagora's theorem, the problem of minimizing (10.2) is equivalent to maximizing $\sum_{i=1}^{n} \left\|p_{x_i}\right\|^2$. Thus the problem is to find $u_1 \in \mathbb{R}^p$ that maximizes $\sum_{i=1}^{n} \left\|p_{x_i}\right\|^2$ under the constraint $\|u_1\| = 1$. With (10.1) we can write

$$\begin{pmatrix} p_{x_1} \\ p_{x_2} \\ \vdots \\ p_{x_n} \end{pmatrix} = \begin{pmatrix} x_1^\top u_1 \\ x_2^\top u_1 \\ \vdots \\ x_n^\top u_1 \end{pmatrix} = \mathcal{X} u_1$$

and the problem can finally be reformulated as find $u_1 \in \mathbb{R}^p$ with $\|u_1\| = 1$ that maximizes the quadratic form $(\mathcal{X} u_1)^\top (\mathcal{X} u_1)$ or

$$\max_{u_1^T u_1 = 1} u_1^\top \left(\mathcal{X}^\top \mathcal{X}\right) u_1. \tag{10.3}$$

10.2 Fitting the p-Dimensional Point Cloud

Theorem 10.1 *The vector u_1 which minimizes (10.2) is the eigenvector of $\mathcal{X}^T\mathcal{X}$ associated with the largest eigenvalue λ_1 of $\mathcal{X}^T\mathcal{X}$.*

Representation of the Cloud on F_1

The coordinates of the n individuals on F_1 are given by $\mathcal{X}u_1$. $\mathcal{X}u_1$ is called the *first factorial variable* or the *first factor* and u_1 the *first factorial axis*. The n individuals, x_i, are now represented by a new factorial variable $z_1 = \mathcal{X}u_1$. This factorial variable is a linear combination of the original variables $(x_{[1]}, \ldots, x_{[p]})$ whose coefficients are given by the vector u_1, i.e.,

$$z_1 = u_{11}x_{[1]} + \ldots + u_{p1}x_{[p]} \tag{10.4}$$

10.2 Fitting the p-Dimensional Point Cloud

Subspaces of Dimension 2

If we approximate the n individuals by a plane (dimension 2), it can be shown via Theorem 2.5 that this space contains u_1. The plane is determined by the best linear fit (u_1) and a unit vector u_2 orthogonal to u_1 which maximizes the quadratic form $u_2^\top \left(\mathcal{X}^\top \mathcal{X} \right) u_2$ under the constraints

$$\|u_2\| = 1, \text{ and } u_1^\top u_2 = 0.$$

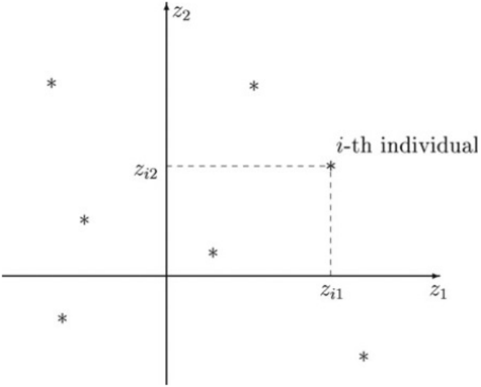
10.2 Fitting the p-Dimensional Point Cloud

Theorem 10.2 *The second factorial axis, u_2, is the eigenvector of $\mathcal{X}^T\mathcal{X}$ corresponding to the second largest eigenvalue λ_2 of $\mathcal{X}^T\mathcal{X}$.*

The unit vector u_2 characterizes a second line, F_2, on which the points are projected. The coordinates of the n individuals on F_2 are given by $z_2 = \mathcal{X}u_2$. The variable z_2 is called the *second factorial variable or the second factor.* The representation of the n individuals in two-dimensional space ($z_1 = \mathcal{X}u_1$ vs. $z_2 = \mathcal{X}u_2$) is shown in Fig. 10.4.

10.2 Fitting the p-Dimensional Point Cloud

Fig. 10.4 Representation of the individuals $x_1, \dots, x_n$ as a two-dimensional point cloud



10.2 Fitting the p-Dimensional Point Cloud

Subspaces of Dimension q ($q \leq p$)

Following the same argument as above, it can be shown via Theorem 2.5 that this best subspace is generated by $u_1, u_2, \ldots, u_q$, the orthonormal eigenvectors of $\mathcal{X}^\top\mathcal{X}$ associated with the corresponding eigenvalues $\lambda_1 \geq \lambda_2 \ldots \geq \lambda_q$. The coordinates of the n individuals on the k-th factorial axis, u_k, are given by the k-th factorial variable $z_k = \mathcal{X}u_k$ for $k = 1, \ldots, q$. Each factorial variable $z_k = (z_{1k}, z_{2k}, \ldots, z_{nk})^\top$ is a linear combination of the original variables $x_{[1]}, x_{[2]}, \ldots, x_{[p]}$ whose coefficients are given by the elements of the k-th vector $u_k : z_{ik} = \sum_{m=1}^{p} x_{im}u_{mk}$.

10.3 Fitting the n-Dimensional Point Cloud

Subspaces of Dimension 1

Suppose that $\mathcal{X}$ is represented by a cloud of p points (variables) in $\mathbb{R}^n$ (considering each column). We have to find a straight line G_1, which is defined by the unit vector $v_1 \in \mathbb{R}^n$, and which gives the best fit the initial cloud of p points. Algebraically, this is the same problem as above (replace $\mathcal{X}$ by $\mathcal{X}^\top$ and follow Sect.10.2): the representation of the j-th variable $x_{[j]} \in \mathbb{R}^n$ is obtained by the projection of the corresponding point onto the straight line G_1 or the direction v_1. Hence we have to find v_1 such that $\sum_{j=1}^{p} \left\| p_{x_{[j]}} \right\|^2$ is maximized, or equivalently, we have to find the unit vector v_1 which maximizes $\left(\mathcal{X}^\top v_1\right)^\top (\mathcal{X}^\top v_1) = {v_1}^\top \left(\mathcal{X}\mathcal{X}^\top\right) v_1$. The solution is given by Theorem 2.5.

10.3 Fitting the n-Dimensional Point Cloud

Theorem 10.3 *v_1 is the eigenvector $\mathcal{X}\mathcal{X}^T$ corresponding to the largest eigenvalue μ_1 of $\mathcal{X}\mathcal{X}^T$.*

Representation of the Cloud on G1

The coordinates of the p variables in G_1 are given by $w_1 = \mathcal{X}^T v_1$. The p variables are now represented by a linear combination of the original individuals $x_1, \dots, x_n$, whose coefficients are given by the vector v_1, i.e., for $j = 1, \dots, p$

$$w_{1j} = v_{11}x_{1j} + \dots + v_{1n}x_{nj} \tag{10.5}$$

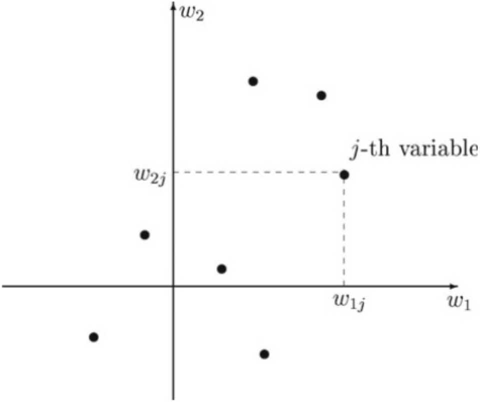
10.3 Fitting the n-Dimensional Point Cloud

Subspaces of Dimension q ($q \leq n$)

The representation of the p variables in a subspace of dimension q is done in the same manner as for the n individuals above. The best subspace is generated by the orthonormal eigenvectors $v_1, v_2, \ldots, v_q$ of $\mathcal{X}\mathcal{X}^\top$ associated with the eigenvalues $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_q$. The coordinates of the p variables on the k-th factorial axis are given by the factorial variables $w_k = \mathcal{X}^\mathcal{T} v_k$, $k = 1$,..., q. Each factorial variable $w_k = (w_{k1}, w_{k2}, \ldots, w_{kp})^T$ is a linear combination of the original individuals $x_1, x_2, \ldots, x_n$ whose coefficients are given by the elements of the k-th vector $v_k : w_{kj} = \sum_{m=1}^{n} v_{km} x_{mj}$.

10.3 Fitting the n-Dimensional Point Cloud

Fig. 10.5 Representation of the variables $x_{[1]}, \ldots, x_{[p]}$ as a two-dimensional point cloud



10.4 Relations Between Subspaces

The aim of this section is to present a duality relationship between the two approaches shown in Sects 10.2 and 10.3.

Theorem 10.4(Duality Relations) *Let r be the rank of $\mathcal{X}$. For $k \leq r$, the eigenvalues λ_k of $\mathcal{X}^T\mathcal{X}$ and $\mathcal{X}\mathcal{X}^T$ are the same and the eigenvectors (u_k and v_k, respectively) are related by*

$$u_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}^{\top}v_k \tag{10.11}$$

$$v_k = \frac{1}{\sqrt{\lambda_k}}\mathcal{X}u_k \tag{10.12}$$

10.4 Relations Between Subspaces

Note that u_k and v_k provide the SVD of $\mathcal{X}$ (see Theorem 2.2). Letting $U = \begin{bmatrix} u_1 & u_2 & \cdots & u_r \end{bmatrix}$, $V = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix}$ and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_r)$ we have

$$\mathcal{X} = V \Lambda^{1/2} U^\top$$

so that

$$x_{ij} = \sum_{k=1}^{r} \lambda_k^{1/2} v_{ik} u_{jk} \tag{10.14}$$

10.5 Practical Computation

- We consider the data set in Sect. B.6 which gives the food expenditures of various French families (manual workers = MA, employees = EM, managers = CA) with varying numbers of children (2, 3, 4, or 5 children). We are interested in investigating whether certain household types prefer certain food types.
- We observe a rather high correlation (0.98) between meat and poultry, whereas the correlation for expenditure for milk and wine (0.01) is rather small. Are there household types that prefer, say, meat over bread?

10.5 Practical Computation

- First, note that in this particular problem the origin has no specific meaning (it represents a “zero” consumer). So it makes sense to compare the consumption of any family to that of an “average family” rather than to the origin. Therefore, the data is first centered (the origin is translated to the center of gravity, $\bar{x}$).
- Furthermore, since the dispersions of the seven variables are quite different each variable is standardized so that each has the same weight in the analysis (mean 0 and variance 1).
- Finally, for convenience, we divide each element in the matrix by $\sqrt{n} = \sqrt{12}$(This will only change the scaling of the plots in the graphical representation.)

10.5 Practical Computation

- $$\mathcal{X}_* = \frac{1}{\sqrt{n}} \mathcal{H} \mathcal{X} \mathcal{D}^{-1/2},$$

  where $\mathcal{H}$ is the centering matrix and $\mathcal{D} = diag(s_{X_i X_i})$.

- Eigenvalues

  $$\lambda = (4.33, 1.83, 0.63, 0.13, 0.06, 0.02, 0.00)^T$$

  shows that the directions of the first two eigenvectors play a dominant role (sum of first two eigenvalues contribute 88% of sum of all eigenvalues), whereas the other directions contribute less than 12% of inertia. A two dimensional plot should suffice for interpreting this data set.

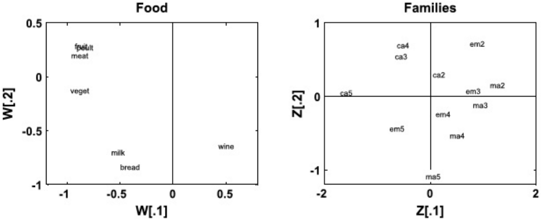
10.5 Practical Computation

- The coordinates of the projected data points are given in the two lower windows of Fig. 10.6. Let us first examine the food expenditure window. In this window we see the representation of the $p = 7$ variables given by the first two factors. The plot show the factorial variables w_1 and w_2 in the same fashion as Fig. 10.4.
- We see that the points for meat, poultry, vegetables, and fruits are close to each other in the upper left of the graph. The expenditures for bread and milk can be found in the lower left, whereas wine stands alone in the lower right. The first factor, w_1, may be interpreted as the vege/meat factor of consumption, the second factor, w_2, as the bread/milk component.

10.5 Practical Computation

- Note that by the Duality Relations of Theorem 10.4, the factorial variables z_j are linear combinations of the factors w_k from the left window. The points displayed in the consumer window (graph on the right) are plotted relative to an average consumer represented by the origin.
- The manager families (CA) are located in the upper left corner of the graph whereas the manual workers (MA) and employees (EM) tend to be in the lower right.The factorial variables for CA5 (managers with five children) lie close to the vege/meat factor. Relative to the average consumer this household type is a large consumer of vegetables/fruits and meat/poultry.

10.5 Practical Computation



bread	-0.49873	-0.84162
veget	-0.96975	-0.13310
fruit	-0.92913	0.27791
meat	-0.96210	0.19107
poult	-0.91125	0.26590
milk	-0.58434	-0.70690
wine	0.42820	-0.64815

ma2	1.12853	0.14353
em2	0.74582	0.70752
ca2	0.04654	0.28641
ma3	0.80563	-0.12763
em3	0.66887	0.06423
ca3	-0.66886	0.53487
ma4	0.36826	-0.54151
em4	0.09955	-0.24969
ca4	-0.63184	0.68533
ma5	-0.08726	-1.09621
em5	-0.77026	-0.44656
ca5	-1.70498	0.03971

Fig. 10.6 Representation of food expenditures and family types in two dimensions MVAdecofood