ON SOME PROPERTIES ON CHI-SQUARE DISTANCE

Khaled Moussa Aludaat
Department of Statistics, Yarmouk University,
Yarmouk, Jordan.

ABSTRACT

In this paper we have described the properties of the distance between two multi dimensional points, which is called Chi-square distance. It is a weighted Euclidean distance.

INTRODUCTION

A large number of multivariate problem can be viewed in terms of "distances" between two observations, or between samples of observations, or between populations of observations. Many distance measures have been proposed and used in multivariate analysis. Here we introduce the distance used in factor analysis of correspondence (method of a multivariate analysis proposed by J.P. Benzercri in (1962)).

This distance has many properties which deserves our attention.

Let K_{IJ} be a contingency table of positive numbers k (i,j) with I rows and J columns:

$$K_{IJ} = \{k\,(i,j) \mid i \in I,\ j \in J\}$$

We define the marginals k_I, k_J and the total K over K_{IJ}; also the profiles f_j^i and f_i^j of the rows and the columns of K_{IJ}. The definitions are given below:

$$\forall\, i \varepsilon I,\ k(i) = \sum_j k(i,j), K_I = \{k(i) \mid i \varepsilon I\}$$

$$\forall\, j \varepsilon J,\ k(i) = \sum_i k(i,j), K_J = \{k(j) \mid j \varepsilon J\}$$

$$K = \sum_i \sum_j k(i,j).$$

$$f_{ij} = \frac{k(i,j)}{K}, f_{IJ} = \{f_{ij} \mid i \varepsilon I, j \varepsilon J\}$$

$$f_i = \frac{k(i)}{K}, f_I = \{f_i \mid i \varepsilon I\}$$

$$f_j = \frac{k(j)}{K}, f_J = \{f_j \mid j \varepsilon J\}$$

$$f_j^i = \frac{k(i,j)}{K(i)} = \frac{f_{ij}}{f_i}, f_i^j = \{f_i^j \mid j \varepsilon J\}$$

$$f_j^i = \frac{k(i,j)}{K(j)} = \frac{f_{ij}}{f_j}, f_i^j = \{f_I^j \mid i \varepsilon J\}$$

The class of the profiles of the rows is defined as:

$$N_I = \{ (f_J^i, f_i) \mid i \in I\}$$

and similarly for the columns we define

$$N_J = \{ (f_I^j, f_j) \mid j \in j\}$$

we can say that the profile f_J^i of the rows i is weighted by the marginal probabilities f_i of i.

Definition: The chi-square distance d^2 (i,i′) between the profiles of two rows i′ and i is given by

$$d^2(i,i') = \sum_j \left\{ \frac{1}{f_j} \left(f_j^i - f_j^{i'} \right)^2 \right\}$$

$$=\sum_j \{\frac{1}{f_j}(\frac{f_{ij}}{f_i} - \frac{f_{i'j}}{f_{i'}})^2\}$$

Similarly the chi-square distance $d^2(j,j')$ between the conditional probabilities f_j^i and $f_j^{i'}$ profiles of the columns j and j′ is given by

$$d^2(j,j') = \sum_i \{\frac{1}{f_i}(f_i^j - f_i^{j'})^2\}$$

we can say that the chi-square distance is a weighted Euclidean distance.

This chi-square has the following properties:

Lemma (1):

If the two discrete random variables x and y are both independent of another discrete random variable z then the chi-square distance between their conditional probabilities is zero.

Proof: *Consider*

$$d^2(x,y) = \sum_z \{\frac{1}{f_z}(f_z^x - f_z^y)^2 \mid z \varepsilon Z\}$$

$$= \sum_z \{\frac{1}{f_z}(\frac{f_{xz}}{f_x} - \frac{f_{yz}}{f_y})^2 \mid z \varepsilon Z\}$$

$$= \sum_z \{\frac{1}{f_z}(\frac{f_x f_z}{f_x} - \frac{f_y f_z}{f_y})^2 \mid z \varepsilon Z\}$$

$$= \sum_z \{\frac{1}{f_z}(f_z - f_z)^2 \mid z \varepsilon Z\} = 0$$

If we have a contingency table with two rows i and i′ such that $\forall\ j \varepsilon k(i',j) = \alpha\, k(i,j)$ for some $\alpha \varepsilon R^+$ (proportional rows) then it follows that $K(i') = \alpha k(i)$. Therefore adding i_o is such that $\forall\ j \varepsilon J$: we have

$$k(i_o,j) = k(i,j) + k(i',j)$$

$$= k(i,j) + \alpha k(i,j)$$

$$= (1+\alpha)\, k(i,j).$$

Furthermore,

$$f_{i_o}j = \frac{k(i_o,j)}{K} = \frac{(1+\alpha)k(i,j)}{K} = \frac{k(i,j)}{K} + \frac{\alpha\, k(i,j)}{K}$$

i.e.,

$$f_{i_o} = f_{ij} + f_{i'}j$$

Lemma (2):

(preservation of the addition). The chi-square distance between two columns j and j′, is not changed by the addition of the two rows which become one row.

Proof: Let

$$S = \frac{1}{f_i}(f_j^i - f_j^{i'})^2 + \frac{1}{f_{i'}}(f_i^{j'} - f_i^{j'})^2$$

$$= \frac{1}{f_i}\left(\frac{f_{ij}}{f_j} - \frac{f_{ij'}}{f_{j'}}\right)^2 + \frac{1}{f_{i'}}\left(\frac{f_{i'j}}{f_j} - \frac{f_{i'j'}}{f_{j'}}\right)^2$$

$$= f_i\left(\frac{f_{ij}}{f_i f_j} - \frac{f_{ij'}}{f_i f_{j'}}\right)^2 + f_{i'}\left(\frac{f_{i'j}}{f_{i'} f_j} - \frac{f_{i'} f_{j'}}{f_i f_{j'}}\right)^2$$

$$= f_i\left(\frac{f_j^i}{f_j} - \frac{f_j^{i'}}{f_{j'}}\right)^2 + f_{i'}\left(\frac{f_j^{i'}}{f_j} - \frac{f_j^{i'}}{f_{j'}}\right)^2$$

$$= \left(f_i + f_{i'}\right)\left(\frac{f_j^{io}}{f_j} - \frac{f_j^{i_o'}}{f_{j'}}\right)^2$$

$$= f_{io}\left(\frac{f_j^{io}}{f_j} - \frac{f_j^{i_o'}}{f_{j'}}\right)^2$$

Since

$$f_j^i = f_j^{i'} = f_j^{io} \text{ and } f_{j'}^{i} = f_{j'}^{i'} = f_j^{i_o'}$$

$$= \frac{1}{f_{io}}\left(f_{io}^{j} - f_{io}^{j'}\right)^2$$

Therefore,

$$d^2(j,j') = \sum_i \{\frac{1}{f_i}\left(f_i^j - f_i^{j'}\right)^2 \mid i \varepsilon I\}$$

$$= \sum_i \{\frac{1}{f_i}\left(f_i^j - f_i^{j'}\right)^2 \mid i \varepsilon I^*\}$$

Note that Card I^* = Card I-1

Lemma (3):

Let (X,Y) be the bivariate random variable and A be a vector whose elements are the conditional probability of x with respect to y then $\| A \|^2 \leq \frac{1}{f_X(x)}$ with f_X is the marginal probability of x

Proof:

$$\| A \|^2 = \| f_Y^X \|^2 = \sum_y \{\frac{1}{f_y}\left(f_y^x\right)^2 \mid y \varepsilon Y\}$$

$$\sum_y \{\frac{1}{f_y}\left(\frac{k(x,y)}{k(x)}\right)^2 \mid y \varepsilon Y\}$$

$$\sum_y \{\frac{1}{f_y}\frac{k(x,y)}{k(x)} \cdot \frac{k(x,y)}{k(x)} \mid y \varepsilon Y\}$$

$$\sum_y \{\frac{K}{k(y)}\frac{k(x,y)}{k(x)} \cdot \frac{k(x,y)}{k(x)} \mid y \varepsilon Y\}$$

$$= \frac{K}{k(x)}\sum_y \frac{k(x,y)}{k(y)} \cdot \frac{k(x,y)}{k(x)} \mid y \varepsilon Y\}$$

we know that $\forall y \varepsilon Y: \left|\frac{k(x,y)}{k(y)}\right| \leq 1$

$$\| a \|^2 \leq \frac{K}{k(x)}\sum_y \{\frac{k(x,y)}{k(x)}\}$$

$$= \frac{K}{k(x)}\frac{k(x)}{k(x)}$$

$$= \frac{K}{k(x)} = \frac{1}{f_x(x)}$$

REFERENCES

[1] J.P. Benzecri et Collaborateur, *L'analyse des données*, Dunod, 3e édition. 1980.

[2] J.P. Benzecri et Collaborateur, *Pratique de l' analyse des données 2*. Abrégé théorique: Etude de cas modèle Dunod, 1980.

[3] J.P. Nakache, A. Chevalier et V. Maurice, *Exercises Commentés de Mathématiques pour l'analyse stastistique des données*, Dunod, Paris, 1981.